ITHEA

International Journal
INFORMATION THEORIES & APPLICATIONS

2017 Volume 24 Number 3

International Journal

INFORMATION THEORIES & APPLICATIONS
Volume 24 / 2017, Number 3

International Journal "INFORMATION THEORIES & APPLICATIONS" (IJ ITA)
is official publisher of the scientific papers of the members of
the ITHEA International Scientific Society

IJ ITA welcomes scientific papers connected with any information theory or its application.
IJ ITA rules for preparing the manuscripts are compulsory.
The **rules for the papers** for IJ ITA are given on *www.ithea.org*.
Responsibility for papers published in IJ ITA belongs to authors.

International Journal "INFORMATION THEORIES & APPLICATIONS" Vol. 24, Number 3, 2017

Edited by the **Institute of Information Theories and Applications FOI ITHEA**, Bulgaria, in collaboration with:
V.M.Glushkov Institute of Cybernetics of NAS, Ukraine,
Universidad Politécnica de Madrid, Spain,
Hasselt University, Belgium,
University of Perugia, Italy,
St. Petersburg Institute of Informatics, RAS, Russia,
Institute for Informatics and Automation Problems, NAS of the Republic of Armenia,
University of Telecommunications and Post, Bulgaria,
Institute of Mathematics and Informatics, BAS, Bulgaria

Printed in Bulgaria
Publisher ITHEA®
Sofia, 1000, P.O.B. 775, Bulgaria. www.ithea.org, e-mail: info@foibg.com
Technical editor: **Ina Markova**

ISSN 1310-0513 (printed) **ISSN 1313-0463 (online)**

CUBE-SPLIT TECHNIQUE IN QUANTITATIVE ASSOCIATION RULE MINING

Levon Aslanyan, Hasmik Sahakyan

***Abstract**: In this paper we consider association rules mining in tables with quantitative attributes. The chain split technique of finding frequent itemsets, known in Boolean association rule mining domain, is extended to the "cube-split" technique being used for finding frequent itemsets for the case of specific quantitative attributes.*

***Keywords**: Data mining, frequent itemsets, quantitative association rules, chain-split technique, cube-split technique.*

***ITHEA Keywords**: I.5 Pattern recognition: I.5.1 Models*

1. Introduction

The goal of the data mining and knowledge discovery is to extract high level information relationships in raw big data of applications. Logical-Statistical Association rules (simply association rules) are one of the key data mining techniques. An association rule is a logical-statistical relation of the form $A \Rightarrow B$, where A and B are events. The rule states that with a certain probability, called the ***confidence*** of the rule, when A occurs in the given database so does B. The second important characterization of rules is the probability of occurrence of the event $A\&B$ over the database records, called the rule ***support***. The problem is to find all association rules that satisfy the requirement of minimum support and minimum confidence. The main bottlenecks of the rule generation are the speed of the algorithms used, the limitation on subset of the rule collections obtained, and the applied value of these rules – their interpretation. A well-known interpretation of association rules comes with the problem market-basket-data-analysis.

The structure of this article is as follows: Boolean association rule mining problem is considered in Section 2, and a brief description of the APRIORI algorithm, the de facto standard approach for the case of binary attributes, - is introduced. Section 3 describes the necessary structural knowledge on the discrete grids that is necessary for the functionality of the chain split technique which is the main instrument considered in this work. Section 4 is devoted to the analysis of rule mining with quantitative attributes. The chain split technique is extended to the case of quantitative attribute rule mining.

2. Boolean Association Rule Mining

About the problem

Association rule mining in terms of confidence/support was first introduced by [Agrawal et al., 1993], and later is addressed in [Agrawal et al., 1994]. These first papers consider databases consisting of categorical attributes only (a *categorical attribute* is one which contains discrete, and typically, unordered data). Thus the events A and B on both sides of the rule $A \Rightarrow B$, are logical expressions of categorical variables. *The aim is to find all rules with confidence and support above the user-defined thresholds of these parameters* (**minconf** and **minsup**). It is desirable that the set of constructed rules is sharply limited in size. Then the rules may have a better interpretation.

Earlier, for the purpose of discovering exact or almost exact rules Piatetsky-Shapiro [Piatetsky-Shapiro, 1991] introduced three principles for rule interestingness (0 if the variables are statistically independent, monotonically increasing if the variables occur more often together, and monotonically decreasing if one of the variables alone occurs more often) defining in this way the base concepts of contemporary rule mining. In structural data analysis rules are likely to be exact. Exact rules appear also as functional dependencies in relational databases [Armstrong, 1998]. On the other hand, in business databases and in very large databases (similar to supermarket transaction databases), rules may be descriptive even in approximation, with the confidence much less than 100%. A number of efficient algorithms for mining binary association rules have been developed (see [Agrawal et al., 1994], [Mannila et al., 1994], [Toivonen, 1996] for just a few examples). And the APRIORI [Agrawal et al., 1996] algorithm is known as the de facto standard of Boolean association rule mining.

Description

Consider a set $I = \{x_1, x_2, \dots, x_n\}$ consisting of n items x_i, and their subsets (itemsets) $X \subseteq I$. We say that it is given a k-itemset, when $|X| = k$. Let D be a database of records (transactions) that are itemsets.

We say that the record $T \in D$ is contributing to the itemset X, if $X \subseteq T$. Association rule is an "if-then" type logical rule $X \Longrightarrow Y$, the fulfillment of which is related to the certain (statistical) conditions. Let X and Y be itemsets where $X \cap Y = 0$. The ratio of the number of all records of D contributing to X and the overall number of records of D - is called support of X in D:

$$\sup(X) = |\{T \in D,\ X \subseteq T\}|/|D|.$$

Next to this is the concept of support for the rule $X \Longrightarrow Y$ itself:

$$\sup(X \Longrightarrow Y) = \sup(X \cup Y) = |\{T \in D, X \cup Y \subseteq T\}|/|D|.$$

Another important property for the rules is the confidence that is defined as:

$$\mathrm{conf}(X \Longrightarrow Y) = \sup(X \cup Y)\,/\sup(X),$$

which is the conditional probability that a record contains Y when it is known that it contains X.

APRIORI algorithm

Practical implementation of association rule mining techniques is a subject of intensive theoretical and algorithmic studies. It is well known that the problem splits naturally into two stages [Agrawal et al., 1996]. The first stage is constructing the so called **frequent fragments** (itemsets), those that occur in the database with frequencies above the predetermined value of support. The second stage is actually the phase of **synthesizing the rules** with a given confidence from the set of frequent subsets constructed during the first stage of the algorithm.

The well recognized algorithm in association rules mining APRIORI [Agrawal et al., 1996] builds the set of frequent subsets (first stage mentioned above) with so-called *building up* method. APRIORI first considers one-element subsets, and computes their frequencies. Next to this, it considers all two-element subsets one-element subsets of which are frequent, and verifies their proper occurrences in the database. Thus the frequent subsets can be building up to the state when it includes subsets that all are not frequent enough. This procedure is known as ***growing*** of frequent itemsets. Computational complexity here is significant and it is especially important because of algorithms will be used over the very large data volumes.

Alternative approaches

Are there any alternative approaches for building rules? There is a huge number of approaches, ideas and algorithms that address this issue. The case of quantitative attributes is much harder. General approaches applied are: rule interpretation as a union of parts – "population-subset" and "extraordinary-behavior" with consecutive optimization (maximization of support and/or confidence); design and consideration of interpretable regions such as attribute and multi-attribute "convex" regions; and - application of fast algorithms, for example randomized algorithms, OPUS, Kadane's, Elias's and other algorithms.

In this paper we propose a new approach to rule mining which connects with the well-known results from the geometry of the n-dimensional unit cube that corresponds to the Hansel's algorithm for

monotone Boolean function identification. We give an extension of this approach to the multidimensional multivalued grid [Aslanyan et.al., 2017]. The extension will cover the case of rule mining by the sets of numerical attributes.

Monotonicity property

Monotonicity is the important property of function $\sup(\mathrm{X})$,

$$(\mathrm{X} \subseteq \mathrm{Y}) \Rightarrow \sup(\mathrm{X}) \geq \sup(\mathrm{Y}).$$

Monotonicity is the key property of the Boolean rule mining models, but it is not simply interpretable in the case of numerical attributes. Things changed when we try to speak in terms of attribute negations [Boulicaut, 2001]. This is when we aim at knowing the frequencies of conjunctions of all literals – that means attributes and their negations. In computational layer this doubles the number of attributes and forces the flooding of area of frequent subsets. [Boulicaut, 2001] and other publications derived formulas that introduce frequencies for negations in terms of formulas of positive frequencies. The issue of positive-negative attributes appears in case of quantitative attributes as the monotone-anti-monotone behavior. Quantitative-monotone, in its turn, means that the increase of attribute values increases probability of the target event. For a particular target event we suppose that all individual attributes are monotonic and that they, in integration, also monotonically depend on the target event.

Isoareas

Isoareas of $\sup(\mathrm{X})$ in Figure 1 can be presented by the sequences of embedded monotone Boolean functions. But such sequences cannot behave arbitrarily. For example, if we consider a simple 2-row database then these rows correspond to 2 vertices of n-cube, let they be denoted by v_1 and v_2. Now consider the subcubes $I_1(v_1, \tilde{1})$ and $I_1(v_2, \tilde{1})$ formed by these vertices and by the $\tilde{1} = (1,1,\cdots,1)$ vertex of the cube. All points of these subcubes 1-support one of the rows but the points in intersection support both of them. So it is not possible to design a database that gives homogeneous support value over the area covered by $I_1(v_1, \tilde{1})$ and $I_1(v_2, \tilde{1})$. So the mentioned "sequences of embedded monotone Boolean functions" are a specific inclusion-exclusion type objects to be described and studied in deep.

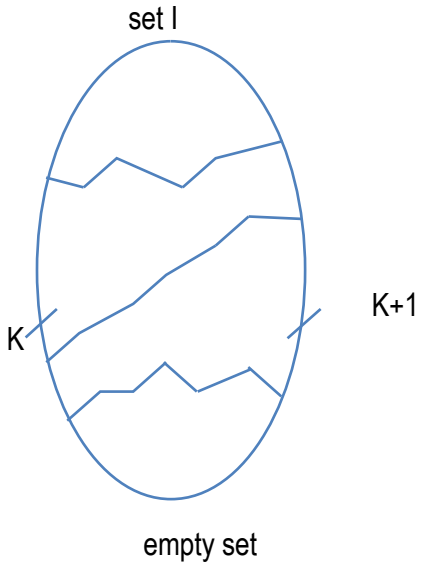

Figure 1. Hasse type diagram of power set of set I of attributes with isoareas of the values of function $\sup(X)$

***Frequent Subsets Mining* (FSSM) Problem**

Given: $\mathcal{D}$ – the database of subsets of some set A of attributes; and a frequency threshold $t \geq 0$.
Provide: the t-frequent subsets in $\mathcal{D}$.
For the FSSM Problem several approaches are proposed: variations to the original APRIORI algorithm [Agrawal et al., 1996], as well as other approaches like [Aumann et al., 2003]. The most efficient algorithms are based on the observation that the frequent subsets are determined by the closed frequent subsets, and such algorithms need to mine the closed frequent subsets at first [Li et al., 2006].

A subset $X \subseteq A$ is called *closed* (maximal subset) if for each $Y \subseteq A$, $Y \supset X$, it holds $\sup(Y) < sup(X)$.

For $X \subseteq A$ we define $\rho(X) := \bigcap\{Y \subseteq A : Y \supseteq X \text{ and } \mathcal{D}(Y) > 0\}$ to be the closure of X in $\mathcal{D}$. It can be checked that $X \subseteq \rho(X)$, $X \subseteq Y \Rightarrow \rho(X) \subseteq \rho(Y)$ and $\rho\big(\rho(X)\big) = \rho(X)$. It also can be checked that the closed subsets are the closures.

Obviously $\mathcal{D}: 2^A \to \mathbb{N}_0$ defines a monotonic decreasing integer function when X increases by the set-inclusion, over the Boolean cube $(2^A, \subseteq)$, and it can be checked that the frequent subsets correspond to the subsets of the closed frequent subsets, and thereby closed frequent subsets determine the frequent subsets.

3. On the Geometry of the n-Dimensional Unit Cube

n dimensional unite cube

Variable with the only values 0 and 1 (false and true) is called a *Boolean variable*. n-dimensional *Boolean function* is a single-valued transformation of the set of all vectors composed by n Boolean variables on to the Boolean set $B = \{0,1\}$. The domain where the Boolean function is given is known as the set of vertices of the n *dimensional unite cube* B^n that is the n-th Cartesian degree of the set B. B^n is the set of all binary vectors $\tilde{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_n)$, which are called *vertices* or points. Usually B^n is presented geometrically via the Hasse diagram, in which vertices of B^n are placed in horizontal layers; the k-th layer $(0 \leq k \leq n)$ contains all the vertices with k number of ones. The layers are arranged vertically starting from the zero layer (at the bottom) to the layer with number . The k-th layer consists of C_n^k vertices. Two vertices $\tilde{\alpha}$ and $\tilde{\beta}$ are called adjacent if they differ in exactly one coordinate. These neighboring vertices are connected by an edge.

Vertices of B^n are organized as follows: a point $\tilde{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_n)$ of B^n precedes the point $\tilde{\beta} = (\beta_1, \beta_2, \dots, \beta_n)$ of B^n, if $\alpha_i \leq \beta_i$, $1 \leq i \leq n$. The fact that a point $\tilde{\alpha}$ precedes the point $\tilde{\beta}$ is denoted by $\tilde{\alpha} \preccurlyeq \tilde{\beta}$. If $\tilde{\alpha} \preccurlyeq \tilde{\beta}$ and $\tilde{\alpha} \neq \tilde{\beta}$ then we write $\tilde{\alpha} \prec \tilde{\beta}$. Two different points $\tilde{\alpha}$ and $\tilde{\beta}$ are called *comparable* if one of the following conditions occur: $\tilde{\alpha} \preccurlyeq \tilde{\beta}$ or $\tilde{\beta} \preccurlyeq \tilde{\alpha}$. Otherwise they are *incomparable*.

It is evident, in general, that to uniquely identify a Boolean function it is necessary to know its values at all points of the n-dimensional unit cube. But if the function belongs to some specific class that is narrower than the set of all Boolean functions, then for the unique determination of its values at all points of B^n is not necessary to know in advance the values of function at all points of B^n, and sometimes it is enough to know the values on a very small subset of vertices of B^n. For example, to uniquely identify a symmetric Boolean function of n variables (these functions possess the same value on each layer of B^n) it is enough to know its values on the set of points from B^n which is intersecting all layers of B^n.

Monotone Boolean functions

Boolean function $f(x_1, x_2, \dots, x_n)$ is called *monotone* if from the fact that $\tilde{\alpha} \prec \tilde{\beta}$ it implies that $f(\alpha_1, \alpha_2, \dots, \alpha_n) \leq f(\beta_1, \beta, \dots, \beta_n)$. The class of all monotone Boolean functions of n variables is denoted by M_n . Some geometric properties of monotone Boolean functions are evident. To each function there is a unique set $\hat{f}^0$ of incomparable vertices of B^n, so that $f(\tilde{\alpha}) = 0$ iff $\tilde{\alpha}$ precedes one

of the vertices of $\hat{f}^0$. Geometrically the area of 0 assignments of f is a union of *subcubes*, composed by the vertex $\tilde{0}$ and the vertices of $\hat{f}^0$. Another important property is that on growing chains of vertices in B^n, the function values - 0's and 1's fills two different intervals at most.

Two type of *recognition problems* about the monotone Boolean functions are rising in different applications. One is the recognition whether the given $f(\tilde{x})$ ***belongs to*** $\boldsymbol{M_n}$, the class of all monotone Boolean functions; and the second is in ***deciphering of*** $\boldsymbol{f(\tilde{x})}$ itself given that $f(\tilde{x}) \in M_n$. We address the second topic because of its identity to the problem of frequent itemset mining.

Chain Split

A separate group of algorithms for Boolean association rule mining is introduced in [Aslanyan et al., 2008].

Suppose that an arbitrary (unknown to us) function $f(\tilde{x}) \in M_n$ is given by an operator A_f, which returns the value $f(\tilde{\alpha})$ by the given input $\tilde{\alpha} \in B^n$. Given the operator A_f it is required to fully restore the set of values of the function $f(\tilde{x})$. After each call to the operator which resumes the value $f(\tilde{\alpha})$ for the point $\tilde{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_n) \in B^n$ other points of B^n become determined through the extension by the monotonicity property. It is clear that we should strive for optimality of these algorithms that is to minimize the steps of applying to A_f.

Consider the set R of all algorithms that solve this problem. That is, for a monotone Boolean function $f(x_1, x_2, \dots, x_n)$ an algorithm from R exploiting the operator A_f restores the complete table of values of $f(\tilde{x})$. Obviously the work of algorithms consists of several stages. Algorithm selects a point $\tilde{\alpha} \in B^n$ and with help of operator A_f computes the value $f(\alpha_1, \alpha_2, \dots, \alpha_n)$ (*selection*). The resulting value of the function at the point $\tilde{\alpha}$ is inserted into the table of computed values of the function. The table is extended by monotonicity, which includes determination of all points that can't have 0 or 1 values arbitrarily after knowing the value at $\tilde{\alpha}$ (*extension*). For example if $f(\tilde{\alpha}) = 1$ then for all points $\tilde{\beta}$ that are higher that $\tilde{\alpha}$ (according to the order of vertices defined above) $f(\tilde{\beta}) = 1$ and the table of values of f is filled in accordingly. Next step is the rule that selects another input for operator A_f and the table of values of f is filled again by monotonicity. This process is repeated until the table of values is filled completely.

Obviously a pair <algorithm $r \in R$ and monotone function $f(x_1, x_2, \dots, x_n)$> can be associated with the number $\varphi(r, f)$ of calls to the operator A_f during the recovery of table of values of the function $f(x_1, x_2, \dots, x_n)$ by the algorithm r.

It is appropriate to evaluate the quality of the algorithms R using function $\varphi(R, f) = \min_r \varphi(r, f)$. We have a condition: $f \in M_n$. The complexity of the recognition of a class of n-

dimensional monotone functions can be characterized by the function $\varphi(n) = \varphi(R, M_n) = \max_f \varphi(R, f)$, where the maximum is taken over all monotone functions.

Let us introduce some general terms on function deciphering [KOR,1965]. Suppose we are given a certain class N of Boolean functions and a function f, belonging to this class. The set of points $G(f, N)$ from B^n is called resolving set for the pair (f, N), if from the fact that

a) the function g belongs to N,
b) values of f and g are the same on the set $G(f, N)$ it follows that $f = g$.

To restore the table of values of functions it is sufficient to determine the values of function on some of its resolving sets. Resolving set $G(f, N)$ is called a deadlock resolving set for (f, N), if no subset of it is resolving for the pair (f, N).

Let us denote by $H(\tilde{\alpha})$ the set of points $\tilde{\beta}$ satisfying the condition $\tilde{\alpha} \prec \tilde{\beta}$, and by $L(\tilde{\alpha})$ - the set of points $\tilde{\gamma}$ such that $\tilde{\gamma} \prec \tilde{\alpha}$.

The upper zero of monotone function $f(x_1, x_2, \dots, x_n)$ is the point $\tilde{\alpha}$ from B^n such that $f(\tilde{\alpha}) = 0$ and $f(\tilde{\beta}) = 1$ for all points $\tilde{\beta} \in H(\tilde{\alpha})$.

The lower one of a monotone function $f(x_1, x_2, \dots, x_n)$ is a point $\tilde{\alpha}$ such that $f(\tilde{\alpha}) = 1$ and $f(\tilde{\gamma}) = 0$ for any point $\tilde{\gamma} \in L(\tilde{\alpha})$.

Let $Z(f)$ denotes the set of all upper zeros of a monotone function $f(x_1, x_2, \dots, x_n)$, and $O(f)$, - the set of all lower ones. Each monotone Boolean function has a unique deadlock resolving set that is included in its all resolving sets (mention that this is not the case for other classes, for instance in class of symmetric Boolean functions that we mentioned above). This deadlock resolving set for a monotone Boolean function is the set $G(f) = Z(f) \cup O(f)$.

A brief characterization of the chain split approach is as follows. The n-dimensional unit cube B^n is a binary lattice consisting of 2^n vertices that correspond to binary strings of length *n*, which are usually arranged in layers in the way that on the *k*-th layer there are all those vertices that have *k* units (1 values). Vertices that differ in one coordinate are called adjacent and are connected by an edge. Chain in B^n is a sequence of adjacent vertices. A chain is called growing if it contains at most one vertex in one layer.

G. Hansel [Hansel, 1966] showed that B^n can be split into growing chains under certain conditions. Further, he considered the monotone Boolean functions and built an algorithm of optimal recognition of these functions using the constructed chains. Relationship of these constructions with the association rules are that frequent subsets with given parameters correspond to a set of zero value vertices of a monotone Boolean function.

Direct use of this technique of Boolean function recognition is difficult because the constructing and storing the Hansel chains is a problem of algorithmic exponential complexity – in computation, and in memory used.

G. Tonoyan [Tonoyan, 1976] offered a computational approach to the work with chains. This is fundamentally and significantly simplifying the recognition algorithm although the complexity is still very high. The idea is in selecting one particular chain split in the collection of Hansel splits. Then a number of functions are introduced that map chains and their elements to each other. In total, this provides the necessary information to recognize monotone Boolean functions and eliminates the need in storing the complete structure of Hansel chains. This means sensitive economy of memory versus a small additional computation over the chain split.

The global aim of this paper is to introduce the necessary chain split and computation technique in terms of problems of search of association rules in large databases and extend this technique to the mining problems of numerical attributes. Additionally, it is to take into account one more important feature of the problem for mining association rules. It is known that in data mining the number of considered elements, *n,* - is very large. It is also knowing that frequent subsets consist of relatively small number of elements. According to this an assumption occurs that there exists a value *k* such that all subsets above this power are not frequent. It turns out that the problem of search of frequent subsets is equivalent to decoding of a special class of monotone Boolean functions, which in turn requires an expansion of the results mentioned above for general Boolean functions, according to some restrictions of the set of functions considered. Extended results are introduced in terms of problem of frequent subsets synthesizing, thus providing the way of determining the set of all maximal (largest by inclusion) frequent subsets, without considering and constructing their sub-subsets. This avoids the part that particularly complicates the building up process.

4. Quantitative Association Rule Mining

In practice, many, if not most, databases contain quantitative data. ***Unfortunately, the definition of categorical association rules does not translate directly to the quantitative case.*** This initiates intensive search for a definition or model of association rules for the case of databases with quantitative attributes. [Srikant et al., 1996] made an approach to extend the categorical rule definition to include quantitative data. They used some kind of grouping and discretization in intervals of values of attributes. Thus, each basic event becomes either a categorical item or a range of numerical values. This way, although the base approach is powerful, raises a number of drawbacks in form of correct interval composition or as the exponential blowup of the number of the rules generated.

[Fukuda et al., 1996a] considered a different perspective to quantitative association rule mining problem, providing efficient algorithms by the given values of ***minconf*** and ***minsup.*** They use computational geometry methods achieving efficiency even for very large size databases, but the rules

considered are plain: with one categorical attribute at the right side. The left side event of the rule have the form $A \in [v_1, v_2]$ with an attribute A and its values v_1, v_2. The technique applied use randomized construction of intervals, combining them into the equi-depth buckets, then joining consecutive buckets into the regions of high support. The main target is to provide the required minconf, maximize the region support, and to keep the algorithms at linear complexity.

Next approach to the problem of quantitative association rules is derived by [Aumann et al., 2003]. The idea used is to compute and apply databases statistical values to increase rule interestingness and to combat the flooding of the number of generated rules. This paper well summarizes several data mining concepts. First of all, it brings the following useful description: An association rule indicates association between ***a subset of the population*** described by the left-hand side of the rule, and an ***extraordinary behavior of this subset*** described by the right-hand side of rule. [Webb, 2001] later ([Aumann et al., 2003] appeared as a conference publication at 1999) labelled these rules as impact rules to outline the difference to the approach [Srikant et al., 1996].

Thus, the introduced general structure of an association rule is of form:

population-subset ⇒***extraordinary behavior.*** ***(1)***

In summary, an association rule considered is a rule of the form: "***population-subset"⇒"mean of values for the subset",*** where population-subset is large enough and the mean of the subset is significantly different to the mean of its complement in the database to form an extraordinary event. And we indeed foresee the efficient algorithms that are able to generate the required rules of this type. The general structure (1) gives rise to many different concrete rule types, determined by the subset class used on the left-hand side, and the description of extraordinary used for the right-hand side. Continuing in this way [Aumann et al., 2003] constructed two types of rules: Categorical to Quantitative rules with an unlimited number of attributes on each side, and Quantitative to Quantitative where both sides contain a single attribute only. It is evident that still these are narrow rule sets. Information on other work on quantitative rule mining may be found at [Aumann et al., 2003] and [Hahsler, 2017]. In particular, [Fukuda et al., 1996b] and [Yoda et al., 1997] consider rules, from 2 numerical to one Boolean attributes, where numerical attributes construct a connected and x-monotone, rectangular, or rectilinear areas (in terms of discrete tomography x-monotone means v-convex, and rectilinear means hv-convex). The whole diversity of studies, as we see, consider a large number of very restricted types of rules, and models, that rarely provide effective computation of the limited number of properly optimized rules. Another concern is that these quantitative approaches loose the main essence of the Boolean rule mining as is the frequent itemset growing.

Let us mention also QuantMiner [Salleb-Aouissi et al., 2007] that is a Quantitative Association Rules Mining tool available online. It takes into consideration a set of numerical attributes in the mining process without a prior binning/discretization of the data. It exploits a recent and innovative research in genetic algorithms.

Concluding,

(*) In this context we aim at constructing ***(multi&numerical)*** $\Longrightarrow$ ***Binary*** rules under the unique supposition of attribute monotonicity. Monotonicity is a natural phenomenon not absolute but it can be supposed for any kind of extraordinary behavior. In exceptional cases the attribute value domain can be split in several intervals that will support monotonicity. All we suppose is that these cases of split for monotonicity are only exceptions and that the individual attribute monotonicity integrates in a collective monotonicity of the model. As a consequence, we will receive a model based on frequent itemset growing technique. And we will extend the Hansel's chain split technique and the monotone Boolean function recognition as the algorithmic basis of our newly quantitative association rule mining model.

Multi-valued cube splitting to unit-cubes

Let $\widetilde{m} = (m_1, m_2, \dots, m_n)$ be an integer vector of n dimensions, and $\Xi_{\widetilde{m}}^n$ be the set of vertices of the n coordinate discrete grid defined as the Cartesian product of sets $\Xi_{m_i} = \{0,1, \dots, m_i - 1\}$:

$$\Xi_{\widetilde{m}}^n = \Xi_{m_1} \times \Xi_{m_2} \times \dots \times \Xi_{m_n} = \{(a_1, a_2, \dots, a_n): a_i \in \Xi_{m_i}, i \in \overline{1,n}\}.$$

In this section we introduce a special decomposition of $\Xi_{\widetilde{m}}^n$ into the structures isomorphic to binary cubes.

Binary cubes may have different dimensions but their distribution by the cube-size is canonical and the cubes in total cover the $\Xi_{\widetilde{m}}^n$ disjointly and entirely. Before descriptions we distinguish several type of special vertices in $\Xi_{\widetilde{m}}^n$.

Middle vertices

Vertices $\widetilde{m}_{mid+} = (\left\lceil \frac{m_1}{2} \right\rceil, \left\lceil \frac{m_2}{2} \right\rceil, \dots, \left\lceil \frac{m_n}{2} \right\rceil)$ and $\widetilde{m}_{mid-} = (\left\lfloor \frac{m_1}{2} \right\rfloor, \left\lfloor \frac{m_2}{2} \right\rfloor, \dots, \left\lfloor \frac{m_n}{2} \right\rfloor)$ we call middle vertices of $\Xi_{\widetilde{m}}^n$. These two vectors coincide when all m_i are even values. Being skewed because

of the possible differences of values $m_1, m_2, \dots, m_n$, $\Xi_{\tilde{m}}^n$ has exactly the mentioned one or two special points at the center. Even with one odd m_i the central points are different. Let $n_{\neq}$ denote the number of all odd m_i values. The set of all points with coordinates $m_{imid+} = \left\lceil \frac{m_i}{2} \right\rceil$ and/or $m_{imid-} = \left\lfloor \frac{m_i}{2} \right\rfloor$ are allocated between the $\tilde{m}_{mid+}$ and $\tilde{m}_{mid-}$. Number of such vertices is equal to $2^{n_{\neq}}$. These points fill a structure isomorphic to $B^{n_{\neq}}$. Symmetrically, let us denote $n - n_{\neq}$ by $n_{=}$.

Upper vertices, lower vertices

A vertex $(a_1, a_2, \dots, a_n)$ of $\Xi_{\tilde{m}}^n$ is called upper vertex if $(a_1, a_2, \dots, a_n) \geq \tilde{m}_{mid+}$. Similarly, vertex $(a_1, a_2, \dots, a_n)$ of $\Xi_{\tilde{m}}^n$ is called lower vertex if $(a_1, a_2, \dots, a_n) \leq \tilde{m}_{mid-}$. $\hat{\Xi}$ and $\check{\Xi}$ denote the sets of all upper and lower vertices of $\Xi_{\tilde{m}}^n$, correspondingly. It is easy to check, that

$$\left|\hat{\Xi}\right| = \left|\check{\Xi}\right| = \prod_{i=1}^{n}(\lfloor m_i/2 \rfloor + 1).$$

Vertical equivalence

Vertices $\tilde{a} = (a_1, a_2, \dots, a_n)$ and $\tilde{b} = (b_1, b_2, \dots, b_n)$ of $\Xi_{\tilde{m}}^n$ are called vertically equivalent if $a_i \in \{b_i, m_i - b_i\}$ for $1 \leq i \leq n$. It is easy to check that this condition is to symmetrically apply to $\tilde{a}$ and $\tilde{b}$, creating a structure of equivalence classes over the $\Xi_{\tilde{m}}^n$. Let $V(\tilde{a})$ denote the class of V-equivalence of vertex $\tilde{a}$. In $V(\tilde{a})$ we distinguish two vertices $\hat{a}$ and $\check{a}$ with coordinates defined as follows:

$$\hat{a}_i = \begin{cases} a_i & if\ a_i \geq m_{imid+} \\ m_i - a_i & if\ a_i \leq m_{imid-} \end{cases}$$

$$\check{a}_i = \begin{cases} m_i - a_i & if\ a_i \geq m_{imid+} \\ a_i & if\ a_i \leq m_{imid-} \end{cases}$$

Vertices $\hat{a}$ and $\check{a}$ are the only "two" vertices for an arbitrary $V(\tilde{a})$ that belong to $\hat{\Xi}$ and $\check{\Xi}$, respectively. Thus all vertices of sets $V(\tilde{a})$ can be extended from the upper and/or lower elements of the class of V-equivalency by component subset inversions (in respect to values m_i). It is evident that the equivalence classes of different vertices of $\hat{\Xi}$ (or $\check{\Xi}$) are disjoint. This construction provides partitioning of $\Xi_{\tilde{m}}^n$ into $\left|\hat{\Xi}\right|$ equivalence classes uniquely defined by the elements of $\hat{\Xi}$.

For a given $\tilde{a} \in \Xi_{\tilde{m}}^{n}$ define an integer $k_{\neq} = |\{a_i : a_i \neq m_i/2\}|$. Then $|V(\tilde{a})| = 2^{k_{\neq}}$. We identify each vertex $\tilde{\beta} \in V(\tilde{a})$ with an n-dimensional binary sequence γ, such that $\tilde{\gamma}_i = 1$ if and only if $\beta_i = \hat{\alpha}_i$. In this manner, $V(\tilde{a})$ becomes isomorphic to the $k_{\neq}$-dimensional binary cube $B^{k_{\neq}}$: the 0-th level contains the lower vertex of $V(\tilde{a})$ belonging to $\check{\Xi}$; the i-th level consists of all vertices of $V(\tilde{a})$ which can be obtained from the lower vertex by applying i number of component inversions.

Thus $\Xi_{\tilde{m}}^{n}$ is partitioned into $\left|\hat{\Xi}\right|$ disjoint equivalence classes - that are identical in structure to the binary cubes. It is worth to mention that in usual chain-split (as is the partition of the binary cube in [Hansel, 1966]) vertices in chains are composed of neighbor vertices, whereas in the case of cube-split edges of the chains connect, in general, vertices that do not belong to the neighbor layers of $\Xi_{\tilde{m}}^{n}$.

Let us obtain the general description of the collection of all V-equivalence clusters of $\Xi_{\tilde{m}}^{n}$. If all m_i are odd, sizes of all $\left|\hat{\Xi}\right|$ subcubes of partition are equal to 2^n.

When all m_i are even, this is the case of the unique middle point, and the arbitrary vertex $\tilde{a} \in \Xi_{\tilde{m}}^{n}$ may have any given number $k_{\neq} \leq n$ of coordinates that are different from $m_i/2$. Volumes of subcubes corresponding to such points $\tilde{a} \in \Xi_{\tilde{m}}^{n}$ is $2^{k_{\neq}}$. The number of all mentioned points with $k_{\neq}$ "un-concentrated" coordinates may be calculated in the following way. It is to construct all different n-vectors that have $n - k_{\neq}$ coordinates equal to $m_i/2$ and the reminder ones accept all feasible assignments. For i and m_i number of such independent evaluations equals to $m_i/2$ taking into account that it is to consider the part of upper vertices, those from $\left|\hat{\Xi}\right|$. For one collection of fixed $n - k_{\neq}$ coordinates we receive production of terms $m_i/2$ by the set of coordinates out of the $n - k_{\neq}$ that accepted values $m_i/2$. The total number of $k_{\neq}$ different upper vertices is equal to the sum of products by all elements of $k_{\neq}$, products of terms $m_i/2$, and it involves all selections of the $k_{\neq}$ collections. Denote this number by $\hat{\varphi}(\tilde{m}, k_{\neq})$,

$$\hat{\varphi}(\tilde{m}, k_{\neq}) = \prod_{I \subseteq \tilde{m},\, |I| = k_{\neq}} \hat{\varphi}(\tilde{m}, I) = \prod_{I \subseteq \tilde{m},\, |I| = k_{\neq}} \prod_{m_i \in I} m_i/2.$$

It is easy to check that the combinatorial generating function of these numbers is

$$\hat{g}(\tilde{m}) = \prod_{i=1}^{n} (1 + (m_i/2)x). \qquad \textbf{(2)}$$

To obtain the value of $\hat{\varphi}(\tilde{m}, k_{\neq})$ it is to maintain the expression of $\hat{g}(\tilde{m})$ analytically, taking the coefficient at the $x^{k_{\neq}}$.

Let us also bring the formula

$$\sum_{k_{\neq}} \hat{\varphi}(\tilde{m}, k_{\neq}) 2^{k_{\neq}} = \prod_i (m_i + 1), \qquad \textbf{(3)}$$

which is some kind of check of the structural description of cube split $\hat{\varphi}(\tilde{m}, k_{\neq})$. $\hat{\varphi}(\tilde{m}, k_{\neq})$ is the coefficient at $x^{k_{\neq}}$ of $\hat{g}(\tilde{m})$ in formula (2). Taking $x = 2$ in (2) we multiply the coefficients $\hat{\varphi}(\tilde{m}, k_{\neq})$ by $2^{k_{\neq}}$, that gives the individual summand of the left side formula of (3). Substitution $x = 2$ in (2) directly, gives the right side formula of (3) proving the check.

Thus we considered two special cases of $\tilde{m}$, one with all odd and second with all even coordinates. The general case that may have as odd as well even coordinates, may be analyzed by a simple integration of these two subcases.

Cube-Split Association Rule Mining (CSARM) Algorithm

At this point we suppose that we are given a relational table T with n attributes: $A_1, A_2, \dots, A_n$. The row collection composes a stream of transactions where Boolean and quantitative attributes are applied together. For an arbitrary categorical attribute, when it is necessary, an additional set of Boolean attributes may be generated as indicators of the different value domains of that attribute, but we suppose that the table T is already in mixed form of Boolean and quantitative attributes and it will not be changed structurally. As we mentioned above, we will consider rules of type $A \Rightarrow B$ where the left side attributes are monotonically related to the target event B. Our next supposition (secondary but important) is the notion that all transactions are sparse, which means that attributes involved with positive (none empty) values are a very limited share of the whole set of attributes. In traditional problem of supermarket basket analysis, the entire number of items in supermarket is of course very large, but each individual purchase basket contains only a very limited number of items. In these suppositions our association rule mining procedure reduces to the monotone binary function recognition algorithm.

To complete the descriptions of CSARM model, it reminds to redirect the reader to the works [Aslanyan et al., 2017] and [Aslanyan et al., 2008] that describe the Boolean monotone recognition in details. The basic application of the framework of the Cube-and-Chain Split Data Mining was with help

of the new versions of Intrusion Detection system SPARTA [Aslanyan et al., 2011]. SPARTA is mining sets of LOG files in a computer system to determine the nonstandard and extraordinary behavior such as the intrusion into the system. We just need to bring the general algorithmic framework of the CSARM system.

Let $f: \Xi_{\widetilde{m}}^{n} \to \{0,1\}$ be a monotone function defined with the help of an oracle Ω_{f}.

CSARM algorithm at first splits $\Xi_{\widetilde{m}}^{n}$ into $\left|\widehat{\Xi}\right|$ vertical equivalence classes. At second stage CSARM maps the monotone binary function f onto the constructed binary cubes. This procedure produces a large set of monotone Boolean functions that will be recognized with the help of Hansel's algorithm.

The final step of CSARM is for integration of the fragments of recognition into a general structure of the frequent subsets of the attributes.

CSARM Algorithm.

- Partition of $\Xi_{\widetilde{m}}^{n}$ into the set of vertical equivalence classes: $V_1, V_2, \cdots, V_{|\widehat{\Xi}|}$, and compose the corresponding binary subcubes: $E(V_1), E(V_2), \cdots, E(V_{|\widehat{\Xi}|})$ as it is described in previous sections (cube split).
- In every cube $E(V_i)$ consider the binary function $f_i: E(V_i) \to \{0,1\}$, defined as follows: $f_i(\beta) = 1$ if and only if $f(b) = 1$, for every $\beta \in E(V_i)$, where b is the origin of β in $\Xi_{\widetilde{\mathrm{m}}}^{\mathrm{n}}$. f_i is monotone and is given with the help of Ω_F. Apply the Hansel's chain split method for the recognition f_i,
- Integrate the results of $\left|\widehat{\Xi}\right|$ binary recognitions procedures to obtain the target function f.

Let $\phi_A(n)$ be the minimal number of queries which is sufficient for recognizing arbitrary monotone function of n variables defined on Ξ_{m+1}^{n} by the CSARM Algorithm. Then:

$$\phi_A(n) = \sum_{k=0}^{n}\left(\hat{\varphi}(\widetilde{m},k)\cdot\left(C_k^{\left\lfloor\frac{k}{2}\right\rfloor} + C_k^{\lfloor k/2\rfloor+1}\right)\right) \text{ for all } m_i \text{ even, and}$$

$$\phi_A(n) = \prod_i\left(\frac{m_i+1}{2}\right)\cdot\left(C_n^{\left\lfloor\frac{n}{2}\right\rfloor} + C_n^{\lfloor n/2\rfloor+1}\right) \text{ for all } m_i \text{ odd.}$$

Thus we obtained formulas for $\phi_A(n)$ in two special cases of $\widetilde{m}$, one with all odd and second with all even coordinates. The general case that may have as odd as well even coordinates, may be analyzed by a simple integration of these two subcases.

Bibliography

[Agrawal et al., 1993] Agrawal R., Imielinski T., and Swami A.. Mining association rules between sets of items in large databases. Proceedings of the ACM SIGMOD International Conference on Management of Data, pages 207-216, Washington D.C., May 1993.

[Agrawal et al., 1994] Agrawal R., Srikant R.. Fast algorithms for mining association rules. Proceedings of 20th International Conference Very Large Data Bases, VLDB, Santiago, Chile, pp. 487-499.

[Agrawal et al., 1996] Agrawal R., Mannila H., Srikant R., Toivonen H., Verkamo I.. Fast Discovery of Association Rules. Advances in knowledge discovery and data mining 12 (1), pp. 307-328, 1996.

[Armstrong, 1998] Armstrong T., Marriott K., Schachte P., Sondergaard H.. Two Classes of Boolean Functions for Dependency Analysis. Science of Computer Programming, 31(1): 3-45, 1998.

[Aslanyan et al., 2017] Aslanyan L., Sahakyan H.. The Splitting technique in monotone recognition. Discrete Applied Mathematics, 216 (2017), pp. 502–512.

[Aslanyan et al., 2008] Aslanyan L., Khachatryan R.. Association rule mining enforced by the chain decomposition of an n-cube. Mathematical Problems of Computer Science, XXX, 2008, ISSN 0131-4645.

[Aslanyan, 1976] Aslanyan L.. Isoperimetry problem and related extremal problems of discrete spaces. Problemy Kibernetiki, 36, pp. 85-126 (1976).

[Aumann et al., 2003] Aumann Y., Lindell Y.. A Statistical Theory for Quantitative Association Rules. Journal of Intelligent Information Systems, vol. 20, 255-283, 2003.

[Boulicaut, 2001] Boulicaut JF., Bykowski A., Jeudy B.. Towards the Tractable Discovery of Association Rules with Negations. In Larsen H.L., Andreasen T., Christiansen H., Kacprzyk J., Zadrożny S. (eds) Flexible Query Answering Systems, Advances in Soft Computing, vol 7, 2001, Physica, Heidelberg.

[Fukuda et al., 1996a] Fukuda T., Morimoto Y., Morishita S., Tokuyama T. Mining Optimized Association Rules for Numeric Attributes. PODS '96 Proceedings of the fifteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems, pp. 182-191, Montreal, Quebec, Canada, June 04 - 06, 1996; Journal of Computer and System Sciences, Volume 58, Issue 1, pp. 1-12, February 1999.

[Fukuda et al., 1996b] Fukuda T., Morimoto Y., Morishita S., Tokuyama T.. Data mining using two-dimensional optimized association rules: Scheme, algorithms, and visualization. Proceedings of the ACM SIGMOD Conference on Management of Data,' pp. 13-23, June 1996.

[Hahsler, 2017] Annotated Bibliography on Association Rule Mining by Michael Hahsler, http://michael.hahsler.net/research/bib/association_rules

[Han et al., 2007] Han J., Cheng H., Xin D., Yan X., Frequent pattern mining: Current status and future directions, Data Mining and Knowledge Discovery, 14(1), 2007.

[Hansel, 1966] Hansel G.. Sur le nombre des functions booleennes monotones de n variables, C.R. Acad. Sci. Paris, 262, serie A (1966), 1088.

[Iberman et al. 2001] Imberman S., Domanski B., Finding Association Rules from Quantitative Data Using Data Booleanization, Americas Conference on Information Systems, Proceedings, 2001.

[KOR,1965] V. Korobkov, On monotone functions of algebra of logic, Prob. Cyb. 13 (1965).

[Li et al.] Li H. F., Lee S. Y., Shan M. K.. DSM-PLW: single-pass mining of path traversal patterns over streaming web click-sequences. Proc. of Computer Networks on Web Dynamics, pp. 1474–1487, 2006.

[Mannila et al., 1994], Mannila H., Toivonen H., Verkamo I., Efficient algorithms for discovering association rules, In AAAI Workshop on Knowledge Discovery in Databases, pp. 181-192, Seattle, Washington D.C., AAAI Press, 1994.

[Piatetsky-Shapiro, 1991] Piatetsky-Shapiro G., Discovery, analysis, and presentation of strong rules, In Knowledge Discovery in Databases, pp. 229-248, 1991.

[Rastogi, 2002] Rastogi R., Shim K., Mining Optimized Association Rules with Categorical and Numeric Attributes, IEEE Transactions on Knowledge and Data Engineering, vol. 14, no. 1, 2002.

[Salleb-Aouissi et al., 2007] Salleb-Aouissi A., Vrain C., Nortet C.. QuantMiner: A Genetic Algorithm for Mining Quantitative Association Rules. Proceedings of the 20th International Conference on Artificial Intelligence, 2007, pp. 1035-1040, India, http://www1.ccls.columbia.edu/~ansaf/QuantMiner

[Srikant et al., 1996] Srikant R., Agrawal R.. Mining Quantitative Association Rules in Large Relational Databases. Proc. of ACM SIGMOD Conference on Management of Data, Montreal, Canada, 1996.

[Toivonen, 1996] Toivonen H., Sampling Large Databases for Association Rules, In Proceedings of the 22th International Conference on Very Large Data Bases, pp. 134-145, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1996.

[Tonoyan, 1976] Tonoyan G.. Chain decomposition of n dimensional unit cube and reconstruction of monotone Boolean functions. JVM&F, v. 19, No. 6 (1976), 1532-1542.

[Yoda et al., 1997] Yoda K., Fukuda T., Morimoto Y., Morishita S., Tokuyama T., Computing optimized rectilinear regions for association rules, in Proceedings of the Third International Conference on Knowledge Discovery and Data Mining,' pp. 96-103, Aug. 1997.

[Webb, 2001] Webb G. I. Discovering associations with numeric variables. Proceedings of the 7th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 383-388, New York, 2001, ACM Press.

[Aslanyan et al., 2011] L. Aslanyan, H. Sahakyan et al, Managing Risk and Safety (chapter 1), Intelligent Data Processing in Global Monitoring for Environment and Security, ITHEA, Sofia (Bulgaria) and Kiev (Ukraine), Editors: K. Markov and V. Velichko, ISBN: 978-954-16-0045-0 (printed), 410 p., 2011.

Authors' Information

Levon Aslanyan *– Institute for Informatics and Automation Problems of the National Academy of Sciences of Armenia, head of department; 1 P.Sevak str., Yerevan 0014, Armenia; e-mail: lasl@sci.am*

Major Fields of Scientific Research: Discrete analysis – algorithms and optimization, pattern recognition theory, information technologies.

Hasmik Sahakyan *– Institute for Informatics and Automation Problems of the National of Science of Armenia; Scientific Secretary. 1 P.Sevak str., Yerevan 0014, Armenia; e-mail: hsahakyan@sci.am*

Major Fields of Scientific Research: Combinatorics, Discrete tomography, Data Mining

EYE EVOLUTION SIMULATION WITH A GENETIC ALGORITHM BASED ON THE HYPOTHESIS OF NILSSON AND PELGER

R. Salas Machado, A. Castellanos, R. Lahoz-Beltra

Abstract: *The present work addresses for the first time the simulation of the evolution of an elemental eye by means of a simple genetic algorithm. The problem of the gradual evolution of a structure as complex as the eye was raised by Darwin, being still at the beginning of the 21st century a source of controversy between creationists and evolutionists. Taking as a starting point the paper of Nilsson and Pelger and their hypothesis that the evolution of the eye can be studied if we limit ourselves to its optical geometry, we show how eye evolution could take place gradually applying the principle of natural selection. Our model is limited to studying how an array of photosensitive epithelial cells is bent gradually to achieve a camera obscura.*

Keywords: *Eye evolution, intelligent design theory, genetic algorithms, Nilsson and Pelger.*

ACM Classification Keywords: *I.6 Simulation and Modeling*

Introduction

Currently there are scientific problems that go beyond an area of knowledge, being studied in different disciplines. For example, in biology the study of the evolution of certain organs or appendages in living beings raises serious problems in finding a satisfactory explanation about the stages through which they passed during their evolution. In our opinion many of the ideas and methods used in this work will also be useful in industrial design problems. Coming back again to the field of biology, at present two examples illustrating these difficulties are the evolution of the eye and the evolution of what is known as the bacterial flagellar motor. As it is known the eye is the organ of the visual system whose purpose is the vision. A first primitive version of the present eye goes back to the Cambrian explosion, 600 million years ago [Breitmeyer, 2010]. From this explosion of life, took place the evolution of an ancestral proto-eye originating the eye that today have very different animals such as mollusks, vertebrates, cephalopods, annelids, crustaceans and some cnidarian (i.e. cubozoa). In evolutionary biology the accepted explanation is that all modern eyes come from the same eye, species sharing some common facts: (i) the 'basic organ' would be a layer of photoreceptor cells attached to an optic nerve; (ii) eye development depends of a common gene shared by different species, called PAX6 gene, and (iii) all the improvements that a modern eye presents would have emerged evolutionarily in just a few million years.

Interestingly, when Darwin introduced his theory of evolution by natural selection, he recognized [Darwin, 1859] some difficulties in explaining satisfactorily the evolution of the eye:

"Reason tells me, that if numerous gradations from a simple and imperfect eye to one complex and perfect can be shown to exist, each grade being useful to its possessor, as is certainly the case; if further, the eye ever varies and the variations be inherited, as is likewise certainly the case and if such variations should be useful to any animal under changing conditions of life, then the difficulty of believing that a perfect and complex eye could be formed by natural selection, though insuperable by our imagination, should not be considered as subversive of the theory."

But where does the problem arise from Darwinian theory with a simple explanation of the evolution of the eye? The problem of Darwinism comes up in the idea of the gradual development of complex structures [Conway-Morris, 1998], e.g. the eye. In 1996 Michael J. Behe writes the book *Darwin's Black Box: The Biochemical Challenge to Evolution*: A book that introduces a new version of creationism based on what is known as 'intelligent design theory'. According to this hypothesis certain features of living systems are explained by the action of intelligent agents resulting in complex and specified information. The book introduces a criticism to the mechanism of gradual changes when such mechanism is applied to the evolution of complex organs. The theory of intelligent design is based on two principles. On the one hand, the principle called (i) *irreducible complexity* states that certain organs cannot evolve by small and successive changes of a precursor ancient organ. On the other hand, another principle known as (ii) *complex specified information* states that anything with a less than 1 in 10^{150} probability of occurring is a consequence of the intervention of intelligent agents [Dembsky et al., 2007]. In summary, the *problem of design* arises in those living systems composed of several interacting components that contribute to the basic function. In such systems if any one of the components is removed then the system stops working correctly. That is, the system seems to be designed for the purpose of performing a certain task or specific function, such as engineers when designing a device or machine.

Now, what does a biological problem, such as the evolution of the eye, have to do with natural computing? Genetic algorithms as well as other evolutionary algorithms are methods of optimization inspired by Darwin's natural selection mechanism. As such an optimization method, a GA uses an objective measure that guides the search for an optimal value, whether it is a maximum or a minimum. The problem arises when GAs are applied to the simulation of biological evolution: GAs use an evaluation function, objective function or fitness function that measures the goodness of a design. It is precisely at this point of reasoning that lies the main problem that we try to solve in this paper: to what extent does the fitness function contains information about the solution we are looking for? How much information does the fitness function contains about the optimum design? [Salas Machado et al., 2016].

In this work adopting a functional definition of design and using a simple genetic algorithm [Lahoz-

Beltra, 2016], we simulate the evolution of an elementary eye. This elementary eye is just a camera obscura (Figure 1), thus a closed box in one of whose ends there is small hole to the outside. When from the outside the light from a scene enters the hole and is received by the layer of photosensitive cells, then the image is reproduced (inverted and reversed). In the model we assume that we have a layer of photoreceptor cells. By means of natural selection, the layer of cells is progressively curved until a camera obscura is obtained. The model defines what a design is based on [Ralph-Wand 2009] definition: a design is the (i) specification of an object, (ii) manifested by an agent, (iii) intended to accomplish goals, in a (iv) particular environment. In our case (i) it is an eye, (ii) an algorithm, (iii) vision and (iv) a multicellular organism, specifically an animal. Next, we define the model used in the simulation experiments.

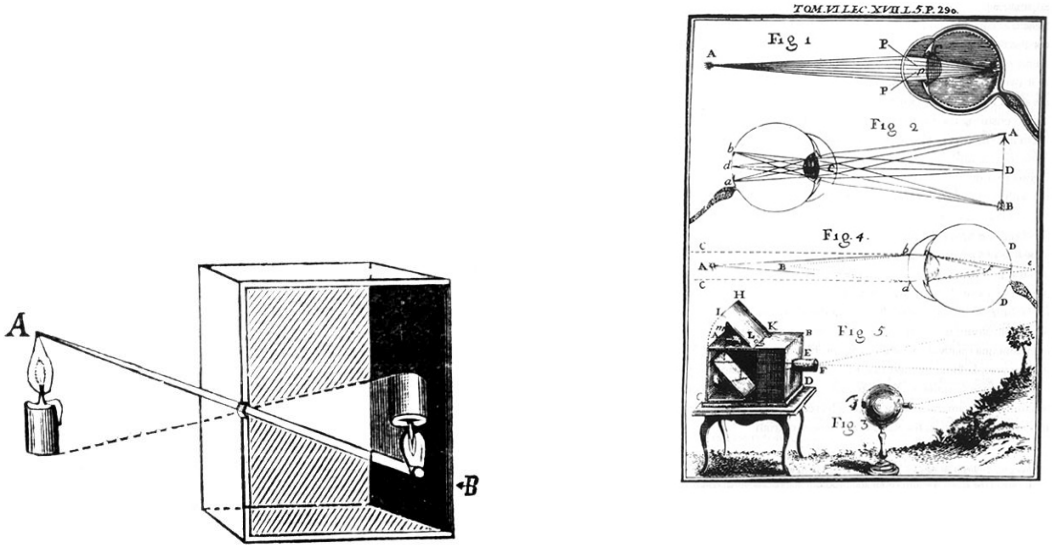

Figure 1. (Left) Camera obscura. (Right) Comparison of eye and camera obscura, early eighteenth century

Model of the evolution of an elementary eye

The model basically generate an array of 3x3 for each elementary eye. In this matrix the elements represent cells such that a value of 1 simulates the presence of epithelium and 0 its absence, i.e. vitreous body filling the cavity. All eyes begin evolution in a primitive state (t=0) in which the eye is a matrix of cells without any vision composed by a layer of cells or epithelium:

111
000
000

In the model we have assumed the existence of an 'eye population', all eyes being in the initial state shown above. The optimal eye, that is to say that eye with some degree of vision, would be similar to a camera obscura, being represented by the matrix that is shown next:

111
101
101

The model is based on the following assumptions:

(a) We consider only a layer of epithelial cells sensitive to light, without considering the problem of the evolution of photoreceptor cells [Nilsson and Pelger, 1994].

(b) We evolve an elementary eye, without going into physiological or zoological considerations.

(c) Unlike the work [Nilsson and Pelger, 1994] we do not estimate the number of generations required but instead we do simulation experiments.

A main feature of our model is that is based on the hypothesis of Nilsson and Pelger [Nilsson and Pelger, 1994] which we could summarize in the following sentence "the problem of the evolution of the eye is solvable if the evolution is limited to its optical geometry".

In other words, a complex structure or its design could evolve through the mechanics of natural selection when the structure is transformed into another equivalent that is the sum of simple quantitative characteristics. For example, the Dawkins model of biomorphs illustrates a complex structure that is the result of the sum of simple quantitative features [Guil López et al., 2016].

Using a simple genetic algorithm [Lahoz-Beltra, 2016] the evolution of the 'eye population' towards the optimal eye is simulated by evaluating each 'eye design' by means of a parameter V_{max} related to vision and which is termed as maximum detectable spatial frequency. This frequency is a measure related with the resolution of the perceived image in the eye. From a computational point of view, e.g. in a photograph, the resolution of an image is defined as the total number of pixels it contains. For instance, assuming that human eye is like a video stream, [Clark, 2016] has been estimated for the human eye an image resolution of 576 megapixels. In Nilsson and Pelger's model [Nilsson and Pelger, 1994] the resolution of the image captured by an eye is evaluated based on the theory of [Snyder, 1979] and [Warrant and McIntyre, 1993]. According to this theory in an eye the frequency V_{max} is calculated by the following expression:

$$V_{max} = (0.375 . \frac{P}{A}) \ln \left[\left(0.746 A^2 \sqrt{I} \right) \right]^{\frac{1}{2}} \quad (1)$$

where A represents the aperture (diameter) of the eye and P is the nodal distance or pit depth. In the model we assume that the intensity of the light I is kept constant during a simulation experiment, being equal to 1.

In our model *A* value is given by the number of 0s in the last row of the matrix that conforms it. In the model, and in order to ensure a minimum aperture, a penalty P_A parameter is declared and used in case the middle position in the last row is set to 1. Therefore, the final equation we used to ensure a minimal aperture is:

$$A = (m_w - ep + P_A)\frac{es}{m_w} \tag{2}$$

where m_w, *ep*, *es* are the arrays maximum width, is the epithelium in the last row and the size in mm of each part of the epithelium in the last row, respectively. P_A was set to 10 in the simulation experiments. Thus, in the spatial frequency the bigger the aperture the worst the resolution of the image. Also, by adding the penalty P_A parameter to aperture measurement we avoid obtaining an aperture equal to 0 avoiding in expression (1) a zero division. The size of the eye (*es*) is assumed equal to 10 mm.

Following, *P* value was obtained by measuring the epithelium thickness in the middle column of the array. The rest of the columns are inspected in the same way and in case we found a gap in the epithelium shape we use a gap penalty parameter *G* to count them. Therefore, the final equation we used is:

$$P = es - (mce_{end} - mce_{begin})\frac{es}{m_d} \tag{3}$$

In the above expression (3) and for each one of the columns in the eye matrix, we track the starting mce_{begin} and ending mce_{end} points of the epithelium, being m_d the arrays maximum depth.

In order to illustrate the most peculiar steps of our algorithm, suppose the following example:

111
101
111

First, we analyze each individual column of the matrix storing the mce_{begin} and mce_{end} values. In this case we found a gap in the second column, therefore the gap penalty *G* is increased by 1. Afterwards, we measure the epithelium in the last row in order to calculate the aperture. In this case, the aperture is 0 since m_w - *A* = 0, but considering the penalty due to the epithelium placed in the minimal aperture position we have an aperture of $10\frac{es}{m_w}$. With all these information, we are ready to calculate the fitness for a single eye for a later comparison.

Finally, the fitness value *f* was calculated as follows:

$$f = \alpha V_{\max} - \beta G \tag{4}$$

setting in (4) the values $\alpha = 100$ and $\beta = 10$.

The evolution of this ancestral eye is governed by some elementary rules of biological inspiration. In relation to the mutation operator, only the empty cells (0) are allowed to be occupied by epithelium (1) if they are surrounded by other cells with epithelium (1). This rule simulates a growth in the epithelium. Also, a cell occupied with epithelium (1) is not allowed to change to a 0 state, since this would simulate the rupture of the epithelium previously formed.

Simulation results

The main conclusion of this work is the possibility to successfully evolve an ancestral eye gradually by Darwinian natural selection when the goal is to obtain an organ equivalent to a camera obscura (Figure 2).

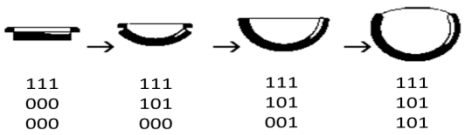

Figure 2. Elementary eye evolution.

Moreover, a simple genetic algorithm is sufficient to emulate the evolution of an elementary eye. Figure 3 shows the characteristic performance graph obtained in four experimental conditions. As the rate of mutation is reduced the average fitness of the population becomes more stable which means that the variation between the generations is lesser. However, we observed in the simulation experiments that a higher rate of mutation it is possible to reach the maximum fitness faster, i.e. better vision faster. For instance, in the 200 generation the overall fitness is already almost the maximum fitness for the higher mutation rate. In the other side, with a lower mutation rate the population seems to increase the fitness through generational steps: 1st region (0-100 generations), 2nd region (100-320 generations), 3rd region (320–450 generations), 4th region (450–500 generations). Once a region has been reached this average fitness behavior reflects a lesser variation between generations. In general, we appreciate that a lower crossover softens the performance graph what means that the propagation of individuals with smaller fitness, i.e. worse vision, is reduced.

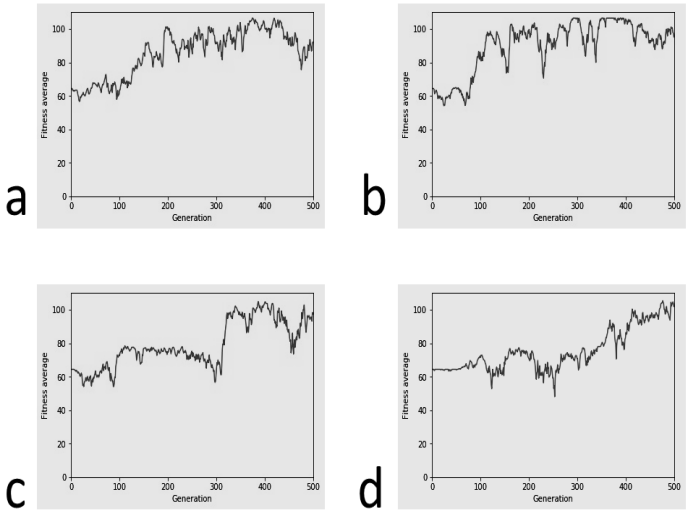

Figure 2. Simulation experiments with a population composed of 100 elementary eyes that evolved over 500 generations. (a) Crossover rate=0.65, eye mutation rate=0.1 and cell mutation rate =0.05. (b) Crossover rate=0.65, eye mutation rate=0.07 and cell mutation rate=0.05. (c) Crossover rate =0.65, eye mutation rate=0.07 and cell mutation rate=0.03. (d) Crossover rate=0.40, eye mutation rate=0.07 and cell mutation rate=0.03.

Conclusion

A model of eye evolution is proposed adopting as a starting point the Nilsson and Pelger hypothesis establishing that the evolution of the eye can be studied if we limit ourselves to its optical geometry. We show how eye evolution could take place gradually applying the principle of natural selection. Our model is limited to studying how an array of photosensitive epithelial cells is bent gradually to achieve a camera obscura.

Bibliography

[Breitmeyer, 2010] B. Breitmeyer. 2010. Blindspots: The Many Ways We Cannot See. New York: Oxford University Press.

[Clark, 2016] R.N. Clark. 2016. ClarkVision.com. http://www.clarkvision.com/articles/eye-resolution.html

[Conway-Morris, 1998] S. Conway-Morris. 1998. The Crucible of Creation. Oxford: Oxford University Press.

[Darwin, 1859] C. Darwin. 1859. On the origin of species by means of natural selection. London: John Murray.

[Dembsky et al., 2007] W.A. Dembsky, W. Ewert, R.J. Marks II. The Evolutionary Informatics Lab. http://www.evoinfo.org/index/

[Guil López et al., 2016] S. Guil López, P. Cuesta Alvaro, S. Cano Alsua, R. Salas Machado, J. Castellanos, R. Lahoz-Beltra. 2016. Towards a Dawkins' genetic algorithm: Transforming an interactive evolutionary algorithm into a genetic algorithm. International Journal "Information Technologies & Knowledge" 10(3): 234-249.

[Lahoz-Beltra, 2016] R. Lahoz-Beltra. 2016. Simple genetic algorithm (SGA). figshare. dx.doi.org/10.6084/ m9.figshare.3397714.v2

[Nilsson and Pelger, 1994] D-E. Nilsson, S. Pelger. 1994. A pessimistic estimate of the time required for an eye to evolve. Proc. R. Soc. Lond. B 256: 53-58.

[Salas Machado et al., 2016] R. Salas Machado, J. Castellanos, R. Lahoz-Beltra. 2016. Evolutionary synthesis of QCA circuits: A critique of evolutionary search methods based on the Hamming oracle. International Journal "Information Technologies & Knowledge" 10(3): 203-215.

[Snyder, 1979] A.W. Snyder. 1979. Physics of vision in compound eyes. In: Handbook of sensory physiology vii/6A, Berlin: Springer (Ed. H-J. Austrum): 225-313.

[Warrant and McIntyre, 1993] E.J. Warrant, P.D. McIntyre. 1993. Arthropod eye design and the physical limits to spatial resolving power. Prog. Neurobiol. 40: 413-461.

Authors' Information

Ramses Salas Machado *– Research Scholar at Carlos III University of Madrid and member of the Grupo de Computación Natural, Departamento de Inteligencia Artificial, Facultad de Informática, Universidad Politécnica de Madrid, Spain; e-mail: ramsesjsm@gmail.com*

Major Fields of Scientific Research: Natural computing.

Angel Castellanos *– Applied Mathematics Department. Universidad Politécnica de Madrid, Madrid; Spain; Natural Computing Group, e-mail: angel.castellanos@upm.es*

Major Fields of Scientific Research: Artificial Intelligence, applied mathematics

Rafael Lahoz–Beltra *– Department of Applied Mathematics, Faculty of Biological Sciences, Complutense University of Madrid, 28040 Madrid, Spain; e-mail: lahozraf@ ucm.es*

Major Fields of Scientific Research: Evolutionary computation, bioinspired algorithms.

ON EFFICIENCY OF P SYSTEMS WITH SYMPORT/ANTIPORT AND MEMBRANE DIVISION

**Luis F. Macías-Ramos, Bosheng Song,
Tao Song, Linqiang Pan, Mario J. Pérez-Jiménez**

Abstract: Classical membrane systems with symport/antiport rules observe the *conservation law*, in the sense that they compute by changing the places of objects with respect to the membranes, and not by changing the objects themselves. In these systems the environment plays an active role because the systems not only send objects to the environment, but also bring objects from the environment. In the initial configuration of a system, there is a special alphabet whose elements appear in an arbitrary large number of copies. The ability of these computing devices to have infinite copies of some objects has been widely exploited in the design of efficient solutions to computationally hard problems.

This paper deals with computational aspects of P systems with symport/antiport and membrane division rules where there is not an environment having the property mentioned above. Specifically, we establish the relationships between the polynomial complexity class associated with P systems with symport/antiport, membrane division rules, and with or without environment. As a consequence, we prove that the role of the environment is irrelevant in order to solve **NP**–complete problems in an efficient way.

Keywords: *Membrane Computing, P System with Symport/Antiport, Membrane Division, Computational Complexity.*

Preliminaries

An *alphabet* Γ is a non–empty set whose elements are called *symbols*. An ordered finite sequence of symbols is a *string* or *word*. If u and v are strings over Γ, then so is their *concatenation* uv, obtained by juxtaposition, that is, writing u and v one after the other. The number of symbols in a string u is the *length* of the string and it is denoted by $|u|$. As usual, the empty string (with length 0) will be denoted by λ. The set of all strings over an alphabet Γ is denoted by Γ^*. In algebraic terms, Γ^* is the free monoid generated by Γ under the operation of concatenation. Subsets of Γ^* are referred to as *languages* over Γ. The set of symbols occurring in a string $u \in \Gamma^*$ is denoted by $alph(u)$.

The *Parikh vector* associated with a string $u \in \Gamma^*$ with respect to the alphabet $\Sigma = \{a_1, \ldots, a_r\} \subseteq \Gamma$ is $\Psi_\Sigma(u) = (|u|_{a_1}, \ldots, |u|_{a_r})$, where $|u|_{a_i}$ denotes the number of occurrences of symbol a_i in string u. The application Ψ_Σ is called the *Parikh mapping* associated with Σ. Notice that, in this definition, the ordering of the symbols from Σ is relevant. If $\Sigma_1 = \{a_{i_1}, \ldots, a_{i_r}\} \subseteq \Gamma$, then we define $\Psi_{\Sigma_1}(u) = (|u|_{a_{i_1}}, \ldots, |u|_{a_{i_r}})$, for each $u \in \Gamma^*$.

A *multiset* m over a set A is a pair (A, f) where $f : A \to \mathbb{N}$ is a mapping. If $m = (A, f)$ is a multiset then its *support* is defined as $supp(m) = \{x \in A \mid f(x) > 0\}$. A multiset is empty (resp. finite) if its support is the empty set (resp. a finite set). If $m = (A, f)$ is a finite multiset over A and $supp(m) = \{a_1, \ldots, a_k\}$, then it will be denoted as $m = \{a_1^{f(a_1)}, \ldots, a_k^{f(a_k)}\}$. That is, superscripts indicate the multiplicity of each element, and if $f(x) = 0$ for $x \in A$, then element x is omitted. A

finite multiset $m = \{a_1^{f(a_1)}, \ldots, a_k^{f(a_k)}\}$ can also be represented by the string $a_1^{f(a_1)} \ldots a_k^{f(a_k)}$ over the alphabet $\{a_1, \ldots, a_k\}$. Nevertheless, all permutations of this string identify the same multiset m precisely. Throughout this paper, we speak about "the finite multiset m" where m is a string, meaning "the finite multiset represented by the string m". If $m_1 = (A, f_1)$, $m_2 = (A, f_2)$ are multisets over A, then we define the union of m_1 and m_2 as $m_1 + m_2 = (A, g)$, where $g = f_1 + f_2$, that is, $g(a) = f_1(a) + f_2(a)$, for each $a \in A$.

For any sets A and B the *relative complement* $A \setminus B$ of B in A is defined as follows: $A \setminus B = \{x \in A \mid x \notin B\}$. For any set A we denote $|A|$ the cardinal (number of elements) of A, as usual.

In what follows, we assume the reader is already familiar with the basic notions and terminology of P systems. For details, see [Păun, 2002].

P systems with symport/antiport rules and membrane division

Cell division is an elegant process that enables organisms to grow and reproduce. Mitosis is a process of cell division which results in the production of two daughter cells from a single parent cell. Daughter cells are identical to one another and to the original parent cell. Through a sequence of steps, the replicated genetic material in a parent cell is equally distributed to two daughter cells. While there are some subtle differences, mitosis is remarkably similar across organisms.

Before a dividing cell enters mitosis, it undergoes a period of growth where the cell replicates its genetic material and organelles. Replication is one of the most important functions of a cell. DNA replication is a simple and precise process that creates two complete strands of DNA (one for each daughter cell) where only one existed before (from the parent cell).

Next, we introduce an abstraction of these operation in the framework of P systems with symport/antiport rules. In these models, the membranes are not polarized; the membranes obtained by division have the same labels as the original membrane, and if a membrane is divided, its interaction with other membranes or with the environment is locked during the division process. In some sense, this means that while a membrane is dividing it closes its communication channels.

Definition 1. *A P system with symport/antiport rules and membrane division of degree $q \geq 1$ is a tuple $\Pi = (\Gamma, \mathcal{E}, \mu, \mathcal{M}_1, \ldots, \mathcal{M}_q, \mathcal{R}_1, \cdots, \mathcal{R}_q, i_{out})$, where:*

1. Γ *is a finite* alphabet.
2. $\mathcal{E} \subseteq \Gamma$.
3. μ *is a membrane structure (a rooted tree) whose nodes are injectively labelled with* $1, 2 \ldots, q$.
4. $\mathcal{M}_1, \ldots, \mathcal{M}_q$ *are multisets over* Γ.
5. $\mathcal{R}_1, \cdots, \mathcal{R}_q$ *are finite set of rules of the following forms:*
   (a) Communication rules*:* $(u, out), (u, in), (u, out; v, in)$*, for* u, v *multisets over* Γ *and* $|u| + |v| > 0$*;*
   (b) Division rules*:* $[a]_i \to [b]_i[c]_i$*, where* $i \neq i_{out}$ *and* $a, b, c \in \Gamma$*;*
6. $i_{out} \in \{0, 1, \ldots, q\}$.

A P system with symport/antiport rules and membrane division

$$\Pi = (\Gamma, \mathcal{E}, \mu, \mathcal{M}_1, \ldots, \mathcal{M}_q, \mathcal{R}_1, \cdots, \mathcal{R}_q, i_{out})$$

of degree q can be viewed as a set of q membranes, labelled by $1, \ldots, q$, arranged in a hierarchical structure, such that: (a) $\mathcal{M}_1, \ldots, \mathcal{M}_q$ represent the finite multisets of objects initially placed in the q membranes of the system; (b) $\mathcal{E}$ is the set of objects initially located in the environment of the system, all of them available in an arbitrary number of copies; and (c) i_{out} represents a distinguished *region* which will encode the output of the system. We use the term *region* i ($0 \leq i \leq q$) to refer to membrane i in the case $1 \leq i \leq q$ and to refer to the environment in the case $i = 0$.

A rule of the type (u, out) or (u, in) is called a *symport* rule. A rule of the type $(u, out; v, in)$, where $|u| + |v| > 0$, is called an *antiport* rule. A P system with symport rules (resp. with antiport rules) is a P system with only symport rules (resp. only antiport rules) as communication rules. The length of rule (u, out) or (u, in) (resp. $(u, out; v, in)$) is defined as $|u|$ (resp. $|u| + |v|$).

An *instantaneous description* or a *configuration* at an instant t of a P system with symport/antiport and membrane division is described by all multisets of objects over Γ associated with all the membranes present in the system, and the multiset of objects over $\Gamma - \mathcal{E}$ associated with the environment at that moment. Recall that there are infinitely many copies of objects from $\mathcal{E}$ in the environment, and hence this set is not properly changed along the computation. The *initial configuration* is $(\mathcal{M}_1, \cdots, \mathcal{M}_q; \emptyset)$.

A rule $(u, out) \in \mathcal{R}_i$ is *applicable* to a configuration $\mathcal{C}$ at an instant t if membrane i is in $\mathcal{C}$ and multiset u is contained in such membrane. When applying a rule $(u, out) \in \mathcal{R}_i$, the objects specified by u are sent out of membrane i into the region immediately outside (its father), this can be the environment in the case of the skin membrane.

A rule $(u, in) \in \mathcal{R}_i$ is *applicable* to a configuration $\mathcal{C}$ at an instant t if membrane i is in $\mathcal{C}$ and multiset u is contained in the immediately upper region (its father), this is the environment in the case when the rule is associated with the skin membrane (the root of the tree μ). When applying a rule $(u, in) \in \mathcal{R}_i$, the multiset of objects u enters the region defined by the membrane i from the immediately upper region (its father), this is the environment in the case when the rule is associated with the skin membrane (the root of the tree μ).

A rule $(u, out; v, in) \in \mathcal{R}_i$ is *applicable* to a configuration $\mathcal{C}$ at an instant t if membrane i is in $\mathcal{C}$ and multiset u is contained in such membrane, and multiset v is contained in the immediately upper region (its father). When applying a rule $(u, out; v, in) \in \mathcal{R}_i$, the objects specified by u are sent out of membrane i into the region immediately outside (its father), at the same time bringing the objects specified by v into membrane i.

A rule $[a]_i \rightarrow [b]_i[c]_i \in \mathcal{R}_i$ is *applicable* to a configuration $\mathcal{C}$ at an instant t if the following holds: (a) membrane i is in $\mathcal{C}$; (b) object a is contained in such membrane; and (c) membrane i is neither the skin membrane nor the output membrane (if $i_{out} \in \{1, \ldots, q\}$). When applying a division rule $[a]_i \rightarrow [b]_i[c]_i$, under the influence of object a, the membrane with label i is divided into two membranes with the same label; in the first copy, object a is replaced by object b, in the second one, object a is replaced by object c; all the other objects residing in membrane i are replicated and copies of them are placed in the two new membranes. The output membrane i_{out} cannot be divided.

The rules of a P system with symport/antiport rules and membrane division are applied in a non-deterministic maximally parallel manner (at each step we apply a multiset of rules which is maximal, no further applicable rule can be added), with the following important remark: if a membrane divides, then the division rule is the only one which is applied for that membrane at that step; the objects inside that membrane do not

evolve by means of communication rules. In other words, before division a membrane interrupts all its communication channels with the other membranes and with the environment. The new membranes resulting from division will interact with other membranes or with the environment only at the next step – providing that they do not divide once again. The label of a membrane precisely identifies the rules which can be applied to it.

Let us fix a P system with symport/antiport rules and membrane division Π. We say that configuration $\mathcal{C}_1$ yields configuration $\mathcal{C}_2$ in one *transition step*, denoted by $\mathcal{C}_1 \Rightarrow_\Pi \mathcal{C}_2$, if we can pass from $\mathcal{C}_1$ to $\mathcal{C}_2$ by applying the rules from $\mathcal{R}_1 \cup \cdots \cup \mathcal{R}_q$ following the previous remarks. A *computation* of Π is a (finite or infinite) sequence of configurations such that:

1. the first term of the sequence is the initial configuration of the system;
2. each non-initial configuration of the sequence is obtained from the previous configuration by applying rules of the system in a maximally parallel manner with the restrictions previously mentioned; and
3. if the sequence is finite (called *halting computation*) then the last term of the sequence is a *halting configuration* (a configuration where no rule of the system is applicable to it).

All computations start from an initial configuration and proceed as stated above; only halting computations give a result, which is encoded by the objects present in the output region i_{out} in the halting configuration.

If $\mathcal{C} = \{\mathcal{C}_t\}_{t<r+1}$ of Π ($r \in \mathbb{N}$) is a halting computation, then the *length of* $\mathcal{C}$, denoted by $|\mathcal{C}|$, is r, that is, $|\mathcal{C}|$ is the number of non-initial configurations which appear in the finite sequence $\mathcal{C}$. We denote by $\mathcal{C}_t(i), 1 \leq i \leq q$, the multiset of objects over Γ contained in all membranes labelled by i (by applying division rules different membranes with the same label can be created) at configuration $\mathcal{C}_t$. We denote by $\mathcal{C}_t(0)$ the multiset of objects over $\Gamma \setminus \mathcal{E}$ contained in the environment at configuration $\mathcal{C}_t$. Finally, we denote by $\mathcal{C}_t^*$ the multiset $\mathcal{C}_t(0) + \mathcal{C}_t(1) + \cdots + \mathcal{C}_t(q)$.

Definition 2. *A P system with symport/antiport rules and membrane division*

$\Pi = (\Gamma, \mathcal{E}, \mu, \mathcal{M}_1, \ldots, \mathcal{M}_q, \mathcal{R}_1, \cdots, \mathcal{R}_q, i_{out})$,

where $\mathcal{E} = \emptyset$, is called a P system with symport/antiport rules, membrane division and without environment.

Usually, we omit the alphabet of the environment in the tuple describing such P system.

Polynomial complexity classes of P systems with symport/antiport

Let us recall that a *decision problem* is a pair (I_X, θ_X) where I_X is a language over a finite alphabet (whose elements are called *instances*) and θ_X is a total Boolean function over I_X. Many abstract problems are not decision problems. For example, in *combinatorial optimization problems* some value must be optimized (minimized or maximized). In order to deal with such problems, they can be transformed into roughly equivalent decision problems by supplying a target/threshold value for the quantity to be optimized, and then asking whether this value can be attained.

There exists a correspondence between decision problems and formal languages. So that, the solvability of decision problems is defined through the recognition of the languages associated with them.

In order to study the computing efficiency of membrane systems, the notions from classical *computational complexity theory* are adapted for membrane computing, and a special class of cell-like P systems is

introduced in [Pérez-Jiménez, Romero-Jiménez and Sancho-Caparrini, 2006]: *recognizer P systems* (called *accepting P systems* in a previous paper [Pérez-Jiménez, Romero-Jiménez and Sancho-Caparrini, 2003]).

Definition 3. *A* recognizer P system *with symport/antiport rules and membrane division of degree* $q \geq 1$ *is a tuple*

$$\Pi = (\Gamma, \mathcal{E}, \Sigma, \mathcal{M}_1, \ldots, \mathcal{M}_q, \mathcal{R}_1, \cdots, \mathcal{R}_q, i_{in}, i_{out})$$

where:

- $(\Gamma, \mathcal{E}, \mathcal{M}_1, \ldots, \mathcal{M}_q, \mathcal{R}_1, \cdots, \mathcal{R}_q, i_{out})$ *is a P system with symport/antiport rules and membrane division of degree* $q \geq 1$*, as defined in the previous section;*
- *The working alphabet* Γ *has two distinguished objects* `yes` *and* `no`*, at least one copy of them present in some initial multisets* $\mathcal{M}_1, \ldots, \mathcal{M}_q$*, but none of them is present in* $\mathcal{E}$*;*
- Σ *is an (input) alphabet strictly contained in* Γ *such that* $\mathcal{E} \cap \Sigma = \emptyset$*;*
- $\mathcal{M}_1, \ldots, \mathcal{M}_q$ *are multisets over* $\Gamma \setminus \Sigma$*;*
- $i_{in} \in \{1, \ldots, q\}$ *is the input membrane;*
- *The output region* i_{out} *is the environment;*
- *All computations halt;*
- *If* $\mathcal{C}$ *is a computation of* Π*, then either object* `yes` *or object* `no` *(but not both) must have been released into the output region (the environment), and only at the last step of the computation.*

Definition 4. *A* recognizer P system *with symport/antiport rules, membrane division and without environment of degree* $q \geq 1$ *is a tuple*

$$\Pi = (\Gamma, \mathcal{E}, \Sigma, \mathcal{M}_1, \ldots, \mathcal{M}_q, \mathcal{R}_1, \cdots, \mathcal{R}_q, i_{in}, i_{out})$$

where:

- $(\Gamma, \mathcal{E}, \Sigma, \mathcal{M}_1, \ldots, \mathcal{M}_q, \mathcal{R}_1, \cdots, \mathcal{R}_q, i_{out})$ *is a P system with symport/antiport rules and membrane division.*
- *The working alphabet* Γ *has two distinguished objects* `yes` *and* `no`*, at least one copy of them present in some initial multisets* $\mathcal{M}_1, \ldots, \mathcal{M}_q$*, but none of them is present in* $\mathcal{E}$*.*
- $\mathcal{E} = \emptyset$*.*
- Σ *is an (input) alphabet strictly contained in* Γ *such that* $\mathcal{E} \cap \Sigma = \emptyset$*.*
- $\mathcal{M}_1, \ldots, \mathcal{M}_q$ *are multisets over* $\Gamma \setminus \Sigma$*.*
- $i_{in} \in \{1, \ldots, q\}$ *is the input membrane.*
- $i_{out} \in \{1, \ldots, q\}$ *is the output membrane.*
- *All computations halt.*

- *If $\mathcal{C}$ is a computation of Π, then either object* $\mathtt{yes}$ *or object* $\mathtt{no}$ *(but not both) must have been released into the output region, and only at the last step of the computation.*

For each multiset $m \in \Sigma^*$, the *computation of the system* Π *with input* $m \in \Sigma^*$ starts from the configuration of the form $(\mathcal{M}_1, \ldots, \mathcal{M}_{i_{in}} + m, \ldots, \mathcal{M}_q; \emptyset)$, that is, the input multiset m has been added to the contents of the input membrane i_{in}, and we denote it by $\Pi + m$. Therefore, we have an initial configuration associated with each input multiset m (over the input alphabet Σ) in this kind of systems.

Given a recognizer P system with symport/antiport rules (with or without environment) and a halting computation $\mathcal{C} = \{\mathcal{C}_t\}_{t<r+1}$ of Π ($r \in \mathbb{N}$), we define the result of $\mathcal{C}$ as follows:

$$Output(\mathcal{C}) = \begin{cases} \mathtt{yes}, & \text{if } \Psi_{\{\mathtt{yes},\mathtt{no}\}}(M_{r,i_{out}}) = (1,0) \wedge \\ & \quad \Psi_{\{\mathtt{yes},\mathtt{no}\}}(M_{t,i_{out}}) = (0,0) \text{ for } t = 0, \ldots, r-1 \\ \mathtt{no}, & \text{if } \Psi_{\{\mathtt{yes},\mathtt{no}\}}(M_{r,i_{out}}) = (0,1) \wedge \\ & \quad \Psi_{\{\mathtt{yes},\mathtt{no}\}}(M_{t,i_{out}}) = (0,0) \text{ for } t = 0, \ldots, r-1 \end{cases}$$

where Ψ is the Parikh mapping, and $M_{t,i_{out}}$ is the multiset over $\Gamma \setminus \mathcal{E}$ associated with the output region at the configuration $\mathcal{C}_t$, in particular, $M_{r,i_{out}}$ is the multiset over $\Gamma \setminus \mathcal{E}$ associated with the output region at the halting configuration $\mathcal{C}_r$.

We say that a computation $\mathcal{C}$ is an *accepting computation* (respectively, *rejecting computation*) if $Output(\mathcal{C}) = \mathtt{yes}$ (respectively, $Output(\mathcal{C}) = \mathtt{no}$), that is, if object $\mathtt{yes}$ (respectively, object $\mathtt{no}$) appears in the output region associated with the corresponding halting configuration of $\mathcal{C}$, and neither object $\mathtt{yes}$ nor $\mathtt{no}$ appears in the output region associated with any non–halting configuration of $\mathcal{C}$.

Let us notice that if a recognizer P system

$$\Pi = (\Gamma, \mathcal{E}, \Sigma, \mathcal{M}_1, \ldots, \mathcal{M}_q, \mathcal{R}_1, \cdots, \mathcal{R}_q, i_{in}, i_{out})$$

has a symport rule of the type $(i, \lambda/u, 0)$ then $alph(u) \cap (\Gamma \setminus \mathcal{E}) \neq \emptyset$, that is, the multiset u must contains some object from $\Gamma \setminus \mathcal{E}$ because on the contrary, all computations of Π would be not halting.

For each natural number $k \geq 1$, we denote by $\mathbf{CDC}(k)$ (respectively, $\mathbf{CDS}(k)$ or $\mathbf{CDA}(k)$) the class of recognizer P systems with membrane division and with symport/antiport rules (respectively, allowing only symport or antiport rules) of length at most k. In the case of P systems without environment, we denote by $\widehat{\mathbf{CDC}}(k)$ ($\widehat{\mathbf{CDS}}(k)$ or $\widehat{\mathbf{CDA}}(k)$ respectively) the class of recognizer P systems with membrane division without environment and with symport/antiport rules (allowing only symport or only antiport rules respectively) of length at most k.

Polynomial complexity classes of P systems with symport/antiport

In this section, we define what solving a decision problem in the framework of P systems with symport/antiport rules in a uniform and efficient way, means. Bearing in mind that they provide devices with a finite description, a numerable family of membrane systems will be necessary in order to solve a decision problem.

Definition 5. *We say that a decision problem $X = (I_X, \theta_X)$ is solvable in a uniform way and polynomial time by a family $\mathbf{\Pi} = \{\Pi(n) \mid n \in \mathbb{N}\}$ of recognizer P systems with symport/antiport rules and membrane division (with or without environment) if the following holds:*

- *The family* $\mathbf{\Pi}$ *is* polynomially uniform *by Turing machines, that is, there exists a deterministic Turing machine working in polynomial time which constructs the system* $\Pi(n)$ *from* $n \in \mathbb{N}$.
- *There exists a pair* (cod, s) *of polynomial-time computable functions over* I_X *such that:*
  - *for each instance* $u \in I_X$, $s(u)$ *is a natural number, and* $cod(u)$ *is an input multiset of the system* $\Pi(s(u))$*;*
  - *for each* $n \in \mathbb{N}$, $s^{-1}(n)$ *is a finite set;*
  - *the family* $\mathbf{\Pi}$ *is* polynomially bounded *with regard to* (X, cod, s)*, that is, there exists a polynomial function* p*, such that for each* $u \in I_X$ *every computation of* $\Pi(s(u))$ *with input* $cod(u)$ *is halting and it performs at most* $p(|u|)$ *steps;*
  - *the family* $\mathbf{\Pi}$ *is* sound *with regard to* (X, cod, s)*, that is, for each* $u \in I_X$*, if there exists an accepting computation of* $\Pi(s(u))$ *with input* $cod(u)$*, then* $\theta_X(u) = 1$*;*
  - *the family* $\mathbf{\Pi}$ *is* complete *with regard to* (X, cod, s)*, that is, for each* $u \in I_X$*, if* $\theta_X(u) = 1$*, then every computation of* $\Pi(s(u))$ *with input* $cod(u)$ *is an accepting one.*

From the soundness and completeness conditions above we deduce that every P system $\Pi(n)$ is *confluent*, in the following sense: every computation of a system with the *same* input multiset must always give the *same* answer.

Let $\mathbf{R}$ be a class of recognizer P systems with symport/antiport rules. We denote by $\mathbf{PMC}_{\mathrm{R}}$ the set of all decision problems which can be solved in a uniform way and polynomial time by means of families of systems from R. The class $\mathbf{PMC}_{\mathrm{R}}$ is closed under complement and polynomial–time reductions [Pérez-Jiménez, Romero-Jiménez and Sancho-Caparrini, 2003].

In what follows, we prove two technical results concerning recognizer P systems.

Proposition 1. *Let* $\Pi = (\Gamma, \mathcal{E}, \Sigma, \mathcal{M}_1, \ldots, \mathcal{M}_q, \mathcal{R}_1, \cdots, \mathcal{R}_q, i_{in}, i_{out})$ *be a recognizer P systems with symport/antiport rules with length at most* k, $k \geq 2$*, and without membrane division. Let* $M = |\mathcal{M}_1 + \cdots + \mathcal{M}_q|$ *and let* $\mathcal{C} = (\mathcal{C}_0, \mathcal{C}_1, \ldots, \mathcal{C}_m)$ *be a computation of* Π *Then,* $|\mathcal{C}_0^*| = M$*, and for each* t, $0 \leq t < m$*, we have*

$$|\mathcal{C}_{t+1}^*| \leq |\mathcal{C}_t^*| \cdot k, \text{ and } |\mathcal{C}_{t+1}^*| \leq M \cdot k^t$$

Proof: Obviously, $|\mathcal{C}_0^*| = |\mathcal{C}_0(0) + \mathcal{C}_0(1) + \cdots + \mathcal{C}_0(q)| = |\mathcal{M}_1 + \cdots + \mathcal{M}_q| = M$. Suppose $0 \leq t < m$, and let us compute $\mathcal{C}_{t+1}^* = \mathcal{C}_{t+1}(0) + \mathcal{C}_{t+1}(1) + \cdots + \mathcal{C}_{t+1}(q)$. Bearing in mind that only the skin membrane can send and receive objects from the environment, we have

$$\mathcal{C}_{t+1}(0) + \mathcal{C}_{t+1}(2) + \mathcal{C}_{t+1}(3) + \cdots + \mathcal{C}_{t+1}(q) \subseteq \mathcal{C}_t(0) + \mathcal{C}_t(1) + \cdots + \mathcal{C}_t(q)$$

Next, let us see what objects membrane 1 can receive at step $t + 1$.

- On the one hand, membrane 1 can receive objects from $\mathcal{C}_t(0)$.
- On the other hand, membrane 1 can receive objects from $\mathcal{E}$ by means of rules in the skin membrane of the types:
  - $(a\, e_{i_1} \ldots\, e_{i_r}, in)$ with $a \in \mathcal{C}_t(0)$ and $e_{i_1}, \ldots, e_{i_r} \in \mathcal{E}$, $r \leq k - 1$.
  - $(a, out; e_{i_1} \ldots\, e_{i_r}, in)$ with $a \in \mathcal{C}_t(1)$ and $e_{i_1}, \ldots, e_{i_r} \in \mathcal{E}$, $r \leq k - 1$.

Then, $|\mathcal{C}_{t+1}(1)| \leq |\mathcal{C}_t(0) + \mathcal{C}_t(1)| \cdot (k-1)$. So, we have

$$\begin{array}{rcl} |\mathcal{C}^*_{t+1}| & = & |\mathcal{C}_{t+1}(0) + \mathcal{C}_{t+1}(2) + \mathcal{C}_{t+1}(3) + \cdots + \mathcal{C}_{t+1}(q)| + |\mathcal{C}_{t+1}(1)| \\ & \leq & |\mathcal{C}_t(0) + \mathcal{C}_t(1) + \cdots + \mathcal{C}_t(q)| + |\mathcal{C}_t(0) + \mathcal{C}_t(1)| \cdot (k-1) \\ & \leq & |\mathcal{C}^*_t| + |\mathcal{C}^*_t| \cdot (k-1) \leq |\mathcal{C}^*_t| \cdot k \end{array}$$

Finally, let us see that $|\mathcal{C}^*_{t+1}| \leq M \cdot k^t$ by induction on t. For $t = 1$ the result is trivial because of $|\mathcal{C}^*_1| \leq (|\mathcal{C}^*_0| + M) \cdot (k-1) = 2M \cdot (k-1)$.

Let t be such that $1 < t < m$ and the result holds for t. Then,

$$|\mathcal{C}^*_{t+1}| \leq |\mathcal{C}^*_t| \cdot k \overset{h.i}{\leq} M \cdot k^{t-1} \cdot k \leq M \cdot k^t$$

□

Proposition 2. *Let $\mathbf{\Pi} = \{\Pi(n) \mid n \in \mathbb{N}\}$ a family of recognizer P systems from $\mathbf{CDC}(k)$, where $k \geq 2$, solving a decision problem $X = (I_X, \theta_X)$ in polynomial time according to Definition 5. Let (cod, s) be a polynomial encoding associated with that solution. There exists a polynomial function $r(n)$ such that for each instance $u \in I_X$, $2^{r(|u|)}$ is an upper bound of the number of objects in all membranes of the system $\Pi(s(u)) + cod(u)$ along any computation.*

Proof: Let $p(n)$ be a polynomial function such that for each $u \in I_X$ every computation of $\Pi(s(u)) + cod(u)$ is halting and it performs at most $p(|u|)$ steps.

Let $u \in I_X$ be an instance of X and

$$\Pi(s(u)) + cod(u) = (\Gamma, \mathcal{E}, \Sigma, \mathcal{M}_1, \ldots, \mathcal{M}_q, \mathcal{R}_1, \ldots, \mathcal{R}_q, i_{in}, i_{out})$$

Let $M = |\mathcal{M}_1 + \cdots + \mathcal{M}_q|$. Let $\mathcal{C} = (\mathcal{C}_0, \mathcal{C}_1, \ldots, \mathcal{C}_m)$, $0 \leq m \leq p(|u|)$, be a computation of Π.

First, let us suppose that we apply only communication rules at m consecutive transition steps. From Proposition 1 we deduce that $|\mathcal{C}^*_0| = M$ and $|\mathcal{C}^*_{t+1}| \leq M \cdot k^t$, for each t, $0 \leq t < m$. Thus, if we apply in a consecutive way the maximum possible number of communication rules (without applying any division rules) to the system $\Pi(s(u)) + cod(u)$, in any instant of any computation of the system, $M \cdot k^{p(|u|)}$ is an upper bound of the number of objects in the whole system.

Now, let us consider the effect of applying in a consecutive way the maximum possible number of division rules (without applying any communication rules) to the system $\Pi(s(u)) + cod(u)$ when the initial configuration has $M \cdot k^{p(|u|)}$ objects. After that, an upper bound of the number of objects in the whole system by any computation is $M \cdot k^{p(|u|)} \cdot 2^{p(|u|)} \cdot p(|u|)$. Then, we consider a polynomial function $r(n)$ such that $r(|u|) \geq \log(M) + p(|u|) \cdot (1 + \log k) + \log(p(|u|))$, for each instance $u \in I_X$. The polynomial function $r(n)$ fulfills the property required.

□

Corollary 1. *Let $\mathbf{\Pi} = \{\Pi(n) \mid n \in \mathbb{N}\}$ a family of recognizer P systems with symport/antiport rules and membrane division, solving a decision problem $X = (I_X, \theta_X)$ in polynomial time according to Definition 5. Let (cod, s) a polynomial encoding associated with that solution. Then, there exists a polynomial function $r(n)$ such that for each instance $u \in I_X$, $2^{r(|u|)}$ is an upper bound of the number of objects from $\mathcal{E}$ which are moved from the environment to all membranes of the system $\Pi(s(u)) + cod(u)$ along any computation.*

Proof: It suffices to note that from Proposition 2 there exists a polynomial function $r(n)$ such that for each instance $u \in I_X$, $2^{r(|u|)}$ is an upper bound of the number of objects in all membranes of the system $\Pi(s(u)) + cod(u)$.

□

Simulating systems from $\mathbf{CDC}(k)$ by means of systems from $\widehat{\mathbf{CDC}}(k)$

The goal of this section is to show that any P system with symport/antiport rules and membrane division can be simulated by a P system symport/antiport rules, membrane division and without environment, in an efficient way.

First of all, we define the meaning of efficient simulations in the framework of recognizer P systems with symport/antiport rules.

Definition 6. *Let Π and Π' be recognizer P systems with symport/antiport rules. We say that Π'* simulates *Π in an efficient way if the following holds:*

1. *Π' can be constructed from Π by a deterministic Turing machine working in polynomial time.*
2. *There exists an injective function, f, from the set* **Comp**(Π) *of computations of Π onto the set* **Comp**(Π') *of computations of Π' such that:*
   - ⋆ *There exists a deterministic Turing machine that constructs computation $f(\mathcal{C})$ from computation $\mathcal{C}$ in polynomial time.*
   - ⋆ *A computation $\mathcal{C} \in$* **Comp**(Π) *is an accepting computation if and only if $f(\mathcal{C}) \in$* **Comp**(Π') *is an accepting one.*
   - ⋆ *There exists a polynomial function $p(n)$ such that for each $\mathcal{C} \in$* **Comp**(Π) *we have $|f(\mathcal{C})| \leq p(|\mathcal{C}|)$.*

Now, for every family of recognizer P system with symport/antiport rules and membrane division solving a decision problem, we design a family of recognizer P systems with symport/antiport rules, membrane division and *without environment* efficiently simulating it, according to Definition 6.

In what follows throughout this Section, let $\mathbf{\Pi} = \{\Pi(n) \mid n \in \mathbb{N}\}$ a family of recognizer P systems with symport/antiport rules and membrane division solving a decision problem $X = (I_X, \theta_X)$ in polynomial time according to Definition 5, and let $r(n)$ be a polynomial function such that for each instance $u \in I_X$, $2^{r(|u|)}$ is an upper bound of the number of objects from $\mathcal{E}$ which are moved from the environment to all membranes of the system by any computation of $\Pi(s(u)) + cod(u)$.

Definition 7. *For each $n \in \mathbb{N}$, let*

$$\Pi(n) = (\Gamma, \mathcal{E}, \Sigma, \mu, \mathcal{M}_1, \ldots, \mathcal{M}_q, \mathcal{R}_1, \ldots, \mathcal{R}_q, i_{in}, i_{out})$$

an element of the previous family $\mathbf{\Pi}$, and for the sake of simplicity we denote r instead of $r(n)$ and 1 is the label of the skin membrane. Let us consider the recognizer P system with symport/antiport rules of degree $q_1 = 1 + q \cdot (r+2) + |\mathcal{E}|$, with membrane division and without environment

$$\mathbf{S}(\Pi(n)) = (\Gamma', \Sigma', \mu', \mathcal{M}'_0, \mathcal{M}'_1, \ldots, \mathcal{M}'_{q_1}, \mathcal{R}'_0, \mathcal{R}'_1, \ldots, \mathcal{R}'_{q_1}, i'_{in}, i'_{out})$$

defined as follows:

- $\Gamma' = \Gamma \cup \{\alpha_i : 0 \leq i \leq r-1\}$.
- $\Sigma' = \Sigma$.
- *Each membrane* $i \in \{1, \ldots, q\}$ *of* Π *provides a membrane of* $\mathbf{S}(\Pi(n))$ *with the same label. In addition,* $\mathbf{S}(\Pi(n))$ *has:*
  - $r+1$ *new membranes, labelled by* $(i,0),(i,1),\ldots,(i,r)$*, respectively, for each* $i \in \{1,\ldots,q\}$.
  - *A distinguished membrane labelled by 0.*
  - *A new membrane, labelled by* l_b*, for each* $b \in \mathcal{E}$.
- μ' *is the rooted tree obtained from* μ *as follows:*
  - *Membrane* 0 *is the root of* μ' *and it is the father of the root of* μ.
  - *For each* $b \in \mathcal{E}$*, membrane* 0 *is the father of membrane* l_b.
  - *We consider a linear structure whose nodes are* $(i,0),(i,1),\ldots,(i,r)$ *and such that* (i,j) *is the father of* $(i,j-1)$*, for each* $1 \leq i \leq q$ *and* $1 \leq j \leq r$.
  - *For each membrane* i *of* μ *we add the previous linear structure being membrane* i *the father of membrane* (i,r).
- *Initial multisets:* $\mathcal{M}'_0 = \emptyset$, $\mathcal{M}'_{l_b} = \{\alpha_0\}$*, for each* $b \in \mathcal{E}$*, and*

$$(1 \leq i \leq q) \begin{cases} \mathcal{M}'_{(i,0)} & = & \mathcal{M}_i \\ \mathcal{M}'_{(i,1)} & = & \emptyset \\ \ldots & & \ldots \\ \mathcal{M}'_{(i,r)} & = & \emptyset \\ \mathcal{M}'_i & = & \emptyset \end{cases}$$

- *Set of rules:*

$$\mathcal{R}'_0 \cup \mathcal{R}'_1 \cup \cdots \cup \mathcal{R}'_q \cup \{\mathcal{R}'_{(i,j)} : 1 \leq i \leq q, 0 \leq j \leq r\} \cup \{\mathcal{R}'_{l_b} : b \in \mathcal{E}\}$$

  where $\mathcal{R}'_0 = \emptyset$, $\mathcal{R}'_i = \mathcal{R}_i$ *for* $1 \leq i \leq q$*, and*

$$\begin{array}{lcl} \mathcal{R}'_{(i,j)} & = & \{\big(a, out; \lambda, in\big) : a \in \Gamma\},\ \textit{for}\ 1 \leq i \leq q \ \wedge\ 0 \leq j \leq r\} \\ \mathcal{R}'_{l_b} & = & \{[\alpha_j]_{l_b} \to [\alpha_{j+1}]_{l_b}\,[\alpha_{j+1}]_{l_b} : 0 \leq j \leq r-2\} \cup \\ & & \{[\alpha_{r-1}]_{l_b} \to [b]_{l_b}\,[b]_{l_b}, (l_b, out; \lambda, in)\},\ \textit{for}\ b \in \mathcal{E} \end{array}$$

- $i'_{in} = (i_{in}, 0)$, *and* $i'_{out} = 0$.

Let us notice that $\mathbf{S}(\Pi(n))$ can be considered as an extension of $\Pi(n)$ without environment, in the following sense:

- $\Gamma \subseteq \Gamma', \Sigma \subseteq \Sigma'$ and $\mathcal{E} = \emptyset$.
- Each membrane in Π is also a membrane in $\mathbf{S}(\Pi(n))$.

- ⋆ There is a distinguished membrane in $\mathbf{S}(\Pi(n))$ labelled by 0 which plays the role of environment of $\Pi(n)$.
- ⋆ μ is a subtree of μ'.
- ⋆ $\mathcal{R} \subseteq \mathcal{R}'$, and now 0 is the label of a "ordinary membrane" in $\mathbf{S}(\Pi(n))$.

Next, we analyze the structure of the computations of system $\mathbf{S}(\Pi(n))$ and we compare them with the computations of $\Pi(n)$.

Lemma 1. *Let $\mathcal{C}' = (\mathcal{C}'_0, \mathcal{C}'_1, \dots)$ be a computation of $\mathbf{S}(\Pi(n))$. For each t $(1 \leq t \leq r)$ the following holds:*

- *$\mathcal{C}'_t(i) = \emptyset$, for $0 \leq i \leq q$.*
- *For each $1 \leq i \leq q$, and $0 \leq j \leq r$ we have:*

$$\mathcal{C}'_t(i,j) = \begin{cases} \mathcal{M}_i, & \text{if} \quad j = t \\ \emptyset, & \text{if} \quad j \neq t \end{cases}$$

- *For each $b \in \mathcal{E}$, there exist 2^t membranes labelled by l_b whose father is membrane 0 and their content is:*

$$\mathcal{C}'_t(l_b) = \begin{cases} \{\alpha_t\}, & \text{if} \quad 1 \leq t \leq r-1 \\ \{b\}, & \text{if} \quad t = r \end{cases}$$

Proof: By induction on t.

Let us start with the basic case $t = 1$. The initial configuration of system $\mathbf{S}(\Pi(n))$ is the following:

- $\mathcal{C}'_0(i) = \emptyset$, for $0 \leq i \leq q$.
- For each $1 \leq i \leq q$ we have $\mathcal{C}'_0(i,0) = \mathcal{M}_i$, and $\mathcal{C}'_0(i,j) = \emptyset$, for $1 \leq j \leq r$.
- For each $b \in \mathcal{E}$, there exists only one membrane labelled by l_b whose contents is $\{\alpha_0\}$.

At configuration $\mathcal{C}'_0$, only the following rules are applicable:

- $[\alpha_0]_{l_b} \to [\alpha_1]_{l_b}\, [\alpha_1]_{l_b}$, for each $b \in \mathcal{E}$.
- $\big(a, out; \lambda, in\big) \in \mathcal{R}_{(i,0)}$, for each $a \in supp(\mathcal{M}_i)$.

Thus,

(a) For each i $(1 \leq i \leq q)$ we have:

$$\begin{cases} \mathcal{C}'_1(i) & = & \emptyset \\ \mathcal{C}'_1(0) & = & \emptyset \\ \mathcal{C}'_1(i,0) & = & \emptyset \\ \mathcal{C}'_1(i,1) & = & \mathcal{M}_i \\ \mathcal{C}'_1(i,j) & = & \emptyset, \text{ for } 2 \leq j \leq r \end{cases}$$

(b) For each $b \in \mathcal{E}$, there are 2 membranes labelled by l_b whose father is membrane 0 and their content is $\{\alpha_1\}$.

Hence, the result holds for $t = 1$.

By induction hypothesis, let t be such that $1 \leq t < r$, and let us suppose the result holds for t, that is,

- $\mathcal{C}'_t(i) = \emptyset$, for $0 \leq i \leq q$.
- For each $1 \leq i \leq q$, and $0 \leq j \leq r$ we have:

$$\mathcal{C}'_t(i,j) = \begin{cases} \mathcal{M}_i, & \text{if} \quad j = t \\ \emptyset, & \text{if} \quad j \neq t \end{cases}$$

- For each $b \in \mathcal{E}$, there exist 2^t membranes labelled by l_b whose father is membrane 0 and their content is $\mathcal{C}'_t(l_b) = \{\alpha_t\}$ (because $t \leq r-1$).

Then, at configuration $\mathcal{C}'_t$ only the following rules are applicable:

(1) If $t \leq r-2$, the rules $[\alpha_t]_{l_b} \rightarrow [\alpha_{t+1}]_{l_b}\,[\alpha_{t+1}]_{l_b}$, for each $b \in \mathcal{E}$.

(2) If $t = r-1$, the rules $[\alpha_t]_{l_b} \rightarrow [b]_{l_b}\,[b]_{l_b}$, for each $b \in \mathcal{E}$.

(3) $\big(a, out; \lambda, in\big) \in \mathcal{R}_{(i,t)}$, for each $a \in supp(\mathcal{M}_i)$.

From the application of rules of types (1) or (2) at configuration $\mathcal{C}'_t$, we deduce that there are 2^{t+1} membranes labelled by l_b in $\mathcal{C}'_{t+1}$, for each $b \in \mathcal{E}$, whose father is membrane 0 and their content is $\{\alpha_{t+1}\}$, if $t \leq r-2$, or $\{b\}$, if $t = r-1$.

From the application of rules of type (3) at configuration $\mathcal{C}'_t$, we deduce that

$$\mathcal{C}'_{t+1}(i,j) = \begin{cases} \mathcal{M}_i, & \text{if} \quad j = t+1 \\ \emptyset, & \text{if} \quad 0 \leq j \leq r \ \wedge \ j \neq t+1 \end{cases}$$

Bearing in mind that no other rule of system $\mathbf{S}(\Pi(n))$ is applicable, we deduce that $\mathcal{C}'_{t+1}(i) = \emptyset$, for $0 \leq i \leq q$.

This completes the proof of this Lemma.

□

Lemma 2. *Let $\mathcal{C}' = (\mathcal{C}'_0, \mathcal{C}'_1, \dots)$ be a computation of the P system $\mathbf{S}(\Pi(n))$. Configuration $\mathcal{C}'_{r+1}$ is the following:*

(1) $\mathcal{C}'_{r+1}(0) = b_1^{2^r} \dots b_\alpha^{2^r}$, where $\mathcal{E} = \{b_1, \dots, b_\alpha\}$.

(2) $\mathcal{C}'_{r+1}(i) = \mathcal{M}_i = \mathcal{C}_0(i)$, for $1 \leq i \leq q$.

(3) $\mathcal{C}'_{r+1}(i,j) = \emptyset$, for $1 \leq i \leq q$, $0 \leq j \leq r$.

(4) For each $b \in \mathcal{E}$, there exist 2^r membranes labelled by l_b whose father is membrane 0 and their content is empty.

Proof: From Lemma 1, the configuration $\mathcal{C}'_r$ is the following:

- $\mathcal{C}'_r(i) = \emptyset$, for $0 \leq i \leq q$.
- For each i $(1 \leq i \leq q)$ we have

$$\mathcal{C}'_r(i,j) = \begin{cases} \mathcal{M}_i, & \text{if} \quad j = r \\ \emptyset, & \text{if} \quad j \neq r \end{cases}$$

- For each $b \in \mathcal{E}$, there exist 2^r membranes labelled by l_b whose father is membrane 0 and their content is $\{b\}$.

At configuration $\mathcal{C}'_r$ only the following rules are applicable:

- $\big(a, out; \lambda, in\big) \in \mathcal{R}_{(i,r)}$, for each $a \in \Gamma \cap supp(\mathcal{M}_i)$.
- $\big(b, out; \lambda, in\big) \in \mathcal{R}_{l_b}$, for each $b \in \mathcal{E}$.

Thus,

- $\mathcal{C}'_{r+1}(0) = b_1^{2^r} \ldots b_\alpha^{2^r}$, where $\mathcal{E} = \{b_1, \ldots, b_\alpha\}$.
- $\mathcal{C}'_{r+1}(i) = \mathcal{M}_i = \mathcal{C}_0(i)$, for $1 \leq i \leq q$.
- $\mathcal{C}'_{r+1}(i,j) = \emptyset$, for $1 \leq i \leq q$ and $0 \leq j \leq r$.
- For each $b \in \mathcal{E}$, there exist 2^r membranes labelled by l_b whose father is membrane 0 and their content is empty.

□

Definition 8. *Let $\mathcal{C} = (\mathcal{C}_0, \mathcal{C}_1, \ldots, \mathcal{C}_m)$ be a halting computation of $\Pi(n)$. Then we define the computation $\mathbf{S}(\mathcal{C}) = (\mathcal{C}'_0, \mathcal{C}'_1, \ldots, \mathcal{C}'_r, \mathcal{C}'_{r+1}, \ldots, \mathcal{C}'_{r+1+m})$ of $\mathbf{S}(\Pi(n))$ as follows:*

(1) *The initial configuration is:*

$$\begin{cases} \mathcal{C}'_0(i) & = & \emptyset, \text{ for } 0 \leq i \leq q \\ \mathcal{C}'_0(i,0) & = & \mathcal{C}_0(i), \text{ for } 1 \leq i \leq q \\ \mathcal{C}'_0(i,j) & = & \emptyset, \text{ for } 1 \leq i \leq q \text{ and } 1 \leq j \leq r \\ \mathcal{C}'_0(l_b) & = & \alpha_0, \text{ for each } b \in \mathcal{E} \end{cases}$$

(2) *The configuration $\mathcal{C}'_t$, for $1 \leq t \leq r$, is described by Lemma 1.*

(3) *The configuration $\mathcal{C}'_{r+1}$ is described by Lemma 2.*

(4) *The configuration $\mathcal{C}'_{r+1+s}$, for $0 \leq s \leq m$, coincides with the configuration $\mathcal{C}_s$ of Π, that is, $\mathcal{C}_s(i) = \mathcal{C}'_{r+1+s}(i)$, for $1 \leq i \leq q$. The content of the remaining membranes (excluding membrane 0) at configuration $\mathcal{C}'_{r+1+s}$ is equal to the content of that membrane at configuration $\mathcal{C}'_{r+1}$, that is, these membranes do not evolve after step $r+1$.*

That is, every computation $\mathcal{C}$ of $\Pi(n)$ can be "reproduced" by a computation $\mathbf{S}(\mathcal{C})$ of $\mathbf{S}(\Pi(n))$ with a delay: from step $r+1$ to step $r+1+m$, the computation $\mathbf{S}(\mathcal{C})$ restricted to membranes $1,\ldots,q$ provides the computation $\mathcal{C}$ of $\Pi(n)$.

From Lemma 1 and Lemma 2 we deduce the following:

(a) $\mathbf{S}(\mathcal{C})$ is a computation of $\mathbf{S}(\Pi(n))$.

(b) $\mathbf{S}$ is an injective function from $\mathbf{Comp}(\Pi(n))$ onto $\mathbf{Comp}(S(\Pi(n)))$.

Proposition 3. *The P system $\mathbf{S}(\Pi(n))$ defined in Definition 7 simulates $\Pi(n)$ in an efficient way.*

Proof: In order to show that $\mathbf{S}(\Pi(n))$ can be constructed from $\Pi(n)$ by a deterministic Turing machine working in polynomial time, it is enough to note that the amount of resources needed to construct $\mathbf{S}(\Pi(n))$ from $\Pi(n)$ is polynomial in the size of the initial resources of $\Pi(n)$. Indeed,

1. The size of the alphabet of $\mathbf{S}(\Pi(n))$ is $|\Gamma'| = |\Gamma| + r$.
2. The initial number of membranes of $\mathbf{S}(\Pi(n))$ is $1 + q \cdot (r+2) + |\mathcal{E}|$.
3. The initial number of objects of $\mathbf{S}(\Pi(n))$ is the initial number of objects of $\Pi(n)$ plus $|\mathcal{E}|$.
4. The number of rules of $\mathbf{S}(\Pi(n))$ is $|\mathcal{R}'| = |\mathcal{R}| + (r+1)\cdot|\mathcal{E}| + |\Gamma|\cdot q\cdot(r+1)$.
5. The maximal length of a communication rule of $\mathbf{S}(\Pi(n))$ is equal to the maximal length of a communication rule of $\Pi(n)$.

From Lemma 1 and Lemma 2 we deduce that:

(a) Every computation $\mathcal{C}'$ of $\mathbf{S}(\Pi(n))$ has associated a computation $\mathcal{C}$ of $\Pi(n)$ such that $\mathbf{S}(\mathcal{C}) = \mathcal{C}'$ in a natural way.

(b) The function $\mathbf{S}$ is injective.

(c) A computation $\mathcal{C}$ of $\Pi(n)$ is an accepting computation if and only if $\mathbf{S}(\mathcal{C})$ is an accepting computation of $\mathbf{S}(\Pi(n))$.

Finally, let us notice that if $\mathcal{C}$ is a computation of $\Pi(n)$ with length m, then $\mathbf{S}(\mathcal{C})$ is a computation of $\mathbf{S}(\Pi(n))$ with length $r+1+m$.

□

Computational complexity classes of P systems with membrane division and without environment

In this Section, we analyze the role of the environment in the efficiency of P systems with membrane division. That is, we study the ability of these P systems with respect to the computational efficiency when the alphabet of the environment is an empty set.

Theorem 1. *For each $k \in \mathbb{N}$ we have* $\mathbf{PMC}_{\mathbf{CDC}(k+1)} = \mathbf{PMC}_{\widehat{\mathbf{CDC}}(k+1)}$.

Proof: Let us recall that $\mathbf{PMC}_{\mathbf{CDC}(1)} = \mathbf{P}$ (see [Macías-Ramos, Song, Pan, and Pérez-Jiménez, 2017] for details). Then,

$$\mathbf{P} \subseteq \mathbf{PMC}_{\widehat{\mathbf{CDC}}(1)} \subseteq \mathbf{PMC}_{\mathbf{CDC}(1)} = \mathbf{P}$$

Thus, the result holds for $k = 0$. Let us show the result holds for $k \geq 1$. Since $\widehat{\mathbf{CDC}}(k+1) \subseteq \mathbf{CDC}(k+1)$ it suffices to prove that $\mathbf{PMC}_{\mathbf{CDC}(k+1)} \subseteq \mathbf{PMC}_{\widehat{\mathbf{CDC}}(k+1)}$. For that, let $X \in \mathbf{PMC}_{\mathbf{CDC}(\mathbf{k+1})}$.

Let $\{\Pi(n) \mid n \in N\}$ be a family of P systems from $\mathbf{CDC}(k+1)$ solving X according to Definition 5. Let (cod, s) be a polynomial encoding associated with that solution. Let $u \in I_X$ be an instance of the problem X that will be processed by the system $\Pi(s(u)) + cod(u)$. According to Proposition 2, let $r(n)$ be a polynomial function that $2^{r(|u|)}$ is an upper bound of the number of objects from $\mathcal{E}$ which are moved from the environment to all membranes of the system by any computation of

$$\Pi(s(u)) + cod(u) = (\Gamma, \mathcal{E}, \Sigma, \mathcal{M}_1, \ldots, \mathcal{M}_{i_{in}} + cod(u), \ldots, \mathcal{M}_{q_1}, \mathcal{R}, i_{in}, i_{out})$$

Then, we consider the P system without environment

$$\mathbf{S}(\Pi(s(u))) + cod(u) = (\Gamma', \Sigma', \mathcal{M}'_0, \mathcal{M}'_1, \ldots, \mathcal{M}'_{i_{in}} + cod(u), \ldots, \mathcal{M}'_{q_1}, \mathcal{R}', i'_{in}, i'_{out})$$

according to Definition 7, where $q_1 = 1 + q \cdot (r(|u|) + 2) + |\mathcal{E}|$.

Therefore, $\mathbf{S}(\Pi(s(u))) + cod(u)$ is a P system from $\widehat{\mathbf{CDC}}(k+1)$ such that verifies the following:

- A distinguished membrane labelled by 0 has been considered, which will play the role of the environment at the system $\Pi(s(u)) + cod(u)$.
- At the initial configuration, it has enough objects in membrane 0 in order to simulate the behaviour of the environment of the system $\Pi(s(u))) + cod(u)$.
- After $r(n) + 1$ step, computations of $\Pi(s(u)) + cod(u)$ are reproduced by the computations of $\mathbf{S}(\Pi(s(u))) + cod(u)$ exactly.

Let us suppose that $\mathcal{E} = \{b_1, \ldots, b_\alpha\}$. In order to simulate $\Pi(s(u)) + cod(u)$ by a P system without environment in an efficient way, we need to have enough objects in the membrane of $\mathbf{S}(\Pi(s(u))) + cod(u)$ labelled by 0 available. Specifically, $2^{r(n)}$ objects in that membrane are enough.

In order to start the simulation of any computation $\mathcal{C}$ of $\Pi(s(u)) + cod(u)$, it would be enough to have $2^{r(n)}$ copies of each object $b_j \in \mathcal{E}$ in the membrane of $\mathbf{S}(\Pi(s(u))) + cod(u)$ labelled by 0. For this purpose,

- For each $b \in \mathcal{E}$ we consider a membrane in $\mathbf{S}(\Pi(s(u))) + cod(u)$ labelled by l_b which only contains object α_0 initially. We also consider the following rules:
  - $[\alpha_j]_{l_b} \rightarrow [\alpha_{j+1}]_{l_b}\ [\alpha_{j+1}]_{l_b}$, for $0 \leq j \leq r(|u|) - 2$,
  - $[\alpha_{p(n)-1}]_{l_b} \rightarrow [b]_{l_b}\ [b]_{l_b}$,
  - $(l_b, b/\lambda, 0)$.
- By applying the previous rules, after $r(|u|)$ transition steps we get $2^{r(|u|)}$ membranes labelled by l_b, for each $b \in \mathcal{E}$ in such a way that each of them contains only object b. Finally, by applying the third rule we get $2^{r(|u|)}$ copies of objects b in membrane 0, for each $b \in \mathcal{E}$.

Therefore, after the execution of $r(|u|)+1$ transition steps in each computation of $\mathbf{S}(\Pi(s(u)))+cod(u)$ in membrane 0 of the corresponding configuration, we have $2^{r(|u|)}$ copies of each object $b \in \mathcal{E}$. This number of copies is enough to simulate any computation $\mathcal{C}$ of $\Pi(s(u)) + cod(u)$ through the system $\mathbf{S}(\Pi(s(u)) + cod(u))$.

From Proposition 3 we deduce that the family $\{\mathbf{S}(\Pi(n))|\, n \in N\}$ solves X in polynomial time according to Definition 5. Hence, $X \in \mathbf{PMC}_{\widehat{\mathbf{CDC}}(k+1)}$.

□

Conclusions and Further Works

Initial configurations of ordinary P systems with symport/antiport rules have an arbitrarily large amount of copies of some kind of objects belonging to a distinguished alphabet which specifies the *environment* of the system.

The previous condition is no too nice from the computational complexity point of view. In this paper, we show that in P systems with with symport/antiport rules and membrane division the environment can be "removed" without a loss of efficiency.

Acknowledgements

The authors acknowledge the support received by National Natural Science Foundation of China (Grant No. 61320106005).

Bibliography

Díaz-Pernil, D., Gutiérrez-Naranjo, M.A., Pérez-Jiménez, M.J., Riscos- Núñez, A. Romero-Jiménez. Computational efficiency of cellular division in tissue-like P systems. *Romanian Journal of Information Science and Technology*, **11**, 3 (2008), 229-241.

Gutiérrez-Escudero, R., Pérez-Jiménez, M.J. and Rius-Font, M. Characterizing tractability by tissue-like P systems. *Lecture Notes in Computer Science* **5957**,

Macías-Ramos, L.F., Pérez-Jiménez, M.J., Riscos-Núñez, A., Rius-Font, M. and Valencia-Cabrera, L. The efficiency of tissue P systems with cell separation rely on the environment. *Lecture Notes in Computer Science*, **7762** (2013), 243-256.

Macías-Ramos, L.F., Song, B., Pan, L., Pérez-Jiménez, M.J., Limits on efficient computation in P systems with symport/antiport rules, submitted BWMC 2017.

Pan, L. and Ishdorj, T.-O. P systems with active membranes and separation rules. *Journal of Universal Computer Science*, **10**, 5, (2004), 630–649. (2010), 289-300.

Păun, Gh. Attacking **NP**-complete problems. In *Unconventional Models of Computation, UMC'2K* (I. Antoniou, C. Calude, M. J. Dinneen, eds.), Springer-Verlag, 2000, 94-115.

Păun, Gh. *Membrane Computing. An Introduction*. Springer–Verlag, Berlin, (2002).

Păun, Gh., Pérez-Jiménez, M.J. and Riscos-Núñez, A. Tissue P System with cell division. *In. J. of Computers, communications & control*, **3**, 3, (2008), 295–303.

Pérez-Jiménez, M.J., Romero-Jiménez, A. and Sancho-Caparrini, F. Complexity classes in models of cellular computing with membranes. *Natural Computing*, **2**, 3 (2003), 265–285.

Pérez-Jiménez, M.J., Romero-Jiménez, A. and Sancho-Caparrini, F. A polynomial complexity class in P systems using membrane division. *Journal of Automata, Languages and Combinatorics*, **11**, 4, (2006), 423-434.

Porreca, A.E., Murphy, N. Pérez-Jiménez, M.J. An efficient solution of Ham Cycle problem in tissue P systems with cell division and communication rules with length at most 2. In M. García-Quismondo, L.F. Macías-Ramos, Gh. Păun, I. Pérez Hurtado, L. Valencia-Cabrera (eds.) *Proceedings of the Tenth Brainstorming Week on Membrane Computing*, Volume II, Seville, Spain, January 30- February 3, 2012, Report RGNC 01/2012, Fénix Editora, 2012, pp. 141-166.

Authors' Information

Luis F. Macías-Ramos *- Research Group on Natural Computing - Dpt. Computer Science and Artificial Intelligence. Universidad de Sevilla - Avda. Reina Mercedes s/n, 41012 Sevilla, Spain. e-mail: lfmaciasr@us.es*

Bosheng Song *- Key Laboratory of Image Information Processing and Intelligent Control, School of Automation. Huazhong University of Science and Technology - Wuhan 430074, Hubei, China e-mail: boshengsong@163.com*

Tao Song *- Key Laboratory of Image Information Processing and Intelligent Control, School of Automation. Huazhong University of Science and Technology - Wuhan 430074, Hubei, China e-mail: songtao0608@hotmail.com*

Linqiang Pan *- Key Laboratory of Image Information Processing and Intelligent Control, School of Automation. Huazhong University of Science and Technology - Wuhan 430074, Hubei, China e-mail: lqpan@mail.hust.edu.cn*

Mario J. Pérez-Jiménez *- Research Group on Natural Computing - Dpt. Computer Science and Artificial Intelligence. Universidad de Sevilla - Avda. Reina Mercedes s/n, 41012 Sevilla, Spain. e-mail: marper@us.es*

DISCRETE TOMOGRAPHY OVERVIEW: CONSTRAINTS, COMPLEXITY, APPROXIMATION

Hasmik Sahakyan, Ani Margaryan

***Abstract**: In this paper we give an overview of discrete tomography problems, addressing constraints/properties; complexity and approximation issues. We present also some notes on existence/reconstruction of binary images from the given horizontal and diagonal projections.*

***Keywords**: Discrete tomography, constraints, approximation.*

***ITHEA Keywords**: G.2. Discrete Mathematics: G.2.3 Applications*

Introduction

Reconstruction of discrete sets from given projections, - is one of the main tasks of *Discrete Tomography*. Discrete sets can be presented as binary images. The line sum of a line through the image is the number of points on this line. The projection of the image in a certain direction consists of all the line sums of the parallel lines passing through the image in this direction. Any binary image with exactly the same projections as the original image represents a reconstruction of that image.

Reconstruction algorithms that are intended to solve particular inverse problems, have many applications in areas such as the image processing, medicine, computer tomography assisted engineering and design, etc. A large number of well-known medical problems require discrete reconstruction technique ([PrauseOnnasch, 1996], [SlumpGerbrands, 1982]). For example, in angiography, the values 0 and 1 can represent the absence or presence of a contrast agent in heart chambers.

Opposite to methods of Computerized Tomography, which use several hundreds of projections, in Discrete Tomography a few projections are available. The main problem arising here is that different binary images may appear with the same projections; and in case of small number of projections the problem in this form can have large number of solutions ([GardGrizmPran, 1999]). For exactly two directions, the horizontal and vertical ones, it is possible to reconstruct an image in polynomial time ([Ryser, 1957]). But in general, if only the horizontal and vertical projections are given, then the number of solutions can be exponentially large ([DLungo, 1994]).

On the other hand, for any set of more than two directions, the problem of reconstructing a binary image from its projections in those directions is NP-complete.([GardGrizmPran, 1999]).

One way to eliminate these problems is to suppose that there is some prior knowledge on the image to be reconstructed and this can reduce the search space of the possible solutions. It can be assumed that the image has some geometrical properties.

Using geometrical knowledge about the discrete sets, such as convexity and connectedness, is a well-studied area. The existence problem for convex matrices, as well as the existence problem for connected matrices are NP-complete ([BarcDLungoNivatPinz, 1996], [Woeginger, 2001]. In the meantime, the existence problem for horizontally and vertically convex and connected matrices can be solved in polynomial time ([DurrChrobak, 1999]).

Another property of discrete sets, which is new and specific for the domain of discrete tomography, is that the rows of the matrix to be reconstructed are distinct [Sahakyan, 2009], [Sahakyan, 2013], Sahakyan, 2014]. This constraint comes from applications, such as design of experiments; it is also related to known problems of other domains (discrete isoperimetry problem ([Aslanyan, 1979], [AslanyanDanoyan, 2013]), hypergraph degree sequence problem ([Sahakyan, 2015]), and others).

Another strategy can be: to find a possibly good but not necessarily the exact solution. Approximation algorithms with greedy approach are introduced in [Sahakyan, 2010], [SahakyanAslanyan, 2011].

In this paper we give some notes/overview on discrete tomography problems: addressing constraints/properties, complexity and approximation issues. We consider also the existence/reconstruction of binary image from the given horizontal and diagonal projections.

Orthogonal projections

Consider T, a finite set in the two-dimensional integer grid Z^2. A projection of T in any direction calculates the number of points of T on the lines parallel to the projection direction. Given a finite set of projections, it is required to reconstruct T or to construct any set matching these projections. T may be presented as a binary matrix, where ones correspond to the points of T.

In the simplest case of the orthogonal projections the existence and construction problem is solved by Gayle and Ryser in combinatorial terms in 1957. Let $A = \{a_{i,j}\}$ be a binary matrix with m rows and n columns. Let $R = (r_1, \cdots, r_m)$ and $S = (s_1 \cdots, s_n)$ denote the row and column sums of A, correspondingly, where $r_i = \sum_{j=1}^{n} a_{i,j}$, $i = 1, \cdots, m$ and $s_j = \sum_{i=1}^{m} a_{i,j}$, $j = 1, \cdots, n$. $U(R,S)$ denotes the set of all binary matrices with row sum R and column sum S.

Theorem 1. [Ryser, 1957].

Let $R = (r_1, r_2, \cdots, r_m)$ and $S = (s_1, s_2, \cdots, s_n)$ be vectors with non-negative integer components arranged in decreasing order. $S^* = (s_1^*, s_2^*, \cdots, s_n^*)$ is the conjugate vector of R: $s_i^* = |\{r_i : r_i \geq j,\ j=1,\cdots,m\}$. Then the class $U(R,S)$ is not empty if and only if S is majorised by $S*$: $\sum_{i=1}^{k} s_i \leq \sum_{i=1}^{k} s_i^*$, $k = 1, \cdots, n-1$, and $\sum_{i=1}^{n} s_i = \sum_{i=1}^{n} s_i^*$.

For a given finite set of projections there may exist different sets with the same projections. Any property of the recovering object, if such property exists, can narrow the class of possible solutions.

Geometrical properties

Definition 1. A binary matrix is h-convex, if the ones in every row form an interval; and is v-convex if the ones in every column form an interval. A binary matrix is hv-convex if it is both h-convex and v-convex.

Definition 2. A binary matrix is connected, if the ones are connected with respect to the adjacency relation.
Connected by 4-adjacency (vertical and horizontal) matrix is called polyomino.

Complexity

If there is no additional restriction, then according to Theorem 1, the existence problem of a binary matrix with given orthogonal projections has polynomial complexity. The existence problem of a binary matrix is NP-complete for h-convex, v-convex matrices, and h-convex, v-convex polyominos ([BarcDLungoNivatPinz, 1996]. But in case of hv-convex polyiominos there exists a polynomial time algorithm ([DurrChrobak, 1999]). NP-completeness of a number of other cases is proven in [Woeginger, 2001].

Distinct rows

Let $A = \{a_{i,j}\}$ be a binary matrix with m rows and n columns, and let $S = (s_1 \cdots, s_n)$ denote the column sum vector of A.

$U(S)$ denotes the class of all binary matrices of size $m \times n$, with the column sum vector S. Let $\overline{U}(S)$ denote the subclass of $U(S)$ where all matrices consist of all distinct rows.

For a given integer vector S the problem of existence/reconstruction of a binary $m \times n$ matrix in the class $\overline{U}(S)$ is investigated in [Sahakyan, 2009] –[Sahakyan 2014]. The complete structural characterization of the set of column sum vectors of all binary $m \times n$ matrices with distinct rows is given in [Sahakyan, 2009]–[Sahakyan, 2014]. It is worth mentioning that these problems have

counterparts in terms of hypergraphs and degree sequences, which are long standing open problems in the graph theory ([Berge, 1989]).

Approximation

A strategy to solve hard discrete tomography problems can be: to search for possibly good but not necessarily the exact solutions of the problem.

A relaxed version of the existence problem (in the class $\overline{U}(S)$) is addressed in [SahakyanAslanyan, 2017], where some constant number of repeated rows is allowed; the complexity of the relaxed problem is investigated, and several properties/results are obtained.

Another approach to obtain approximate solutions is applied in [Sahakyan, 2010]-[SahakyanAslanyan, 2011]. Greedy algorithm is proposed which constructs a matrix from the given column sum $S = (s_1, \cdots, s_n)$. The strategy is the following: to construct the matrix column-by-column in such a way that in each step the number of different pairs of rows is maximized. A schematic picture of the greedy partitioning is given in Figure 1:

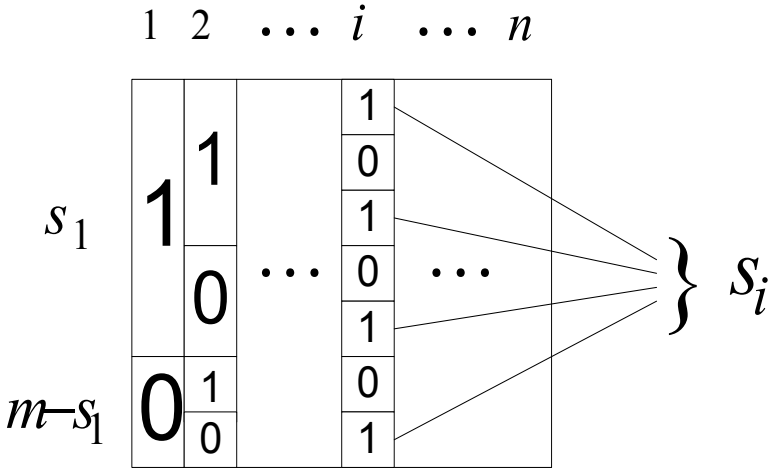

Figure 1

Diagonal projections

The problem of reconstructing binary images from given orthogonal and diagonal projections is studied in [GardGrizmPrang, 1999], [BarBrunDLunNivat, 2001]. In general the problem of reconstructing binary images, from given orthogonal and diagonal projections, is NP-complete [GardGrizmPrang, 1999].

The case of horizontal-vertical-diagonal connected and convex sets is studied in [BarBrunDLunNivat, 2001], and a polynomial-time algorithm is provided for reconstructing these sets.

Diagonal and anti-diagonal projections were studied in [VermaShriv, 2014], and an approach for reconstruction of binary images from diagonal and anti-diagonal projections is provided in [SrivastavaVermaPatel, 2012]; a comparison is done with the existed methods.

The uniqueness of solution for reconstruction problem with the diagonal and anti-diagonal projections is discussed in [SrivastavaVerma, 2013].

Horizontal and Diagonal projections

In this section we consider existence/reconstruction of binary matrices from the given horizontal and diagonal projections.

Consider a binary matrix $A = \{a_{i,j}\}$ with m rows and n columns. Let $R = (r_1, \cdots, r_m)$ denote the row sum, and $D = (d_1, \dots, d_{m+n-1})$ denote the diagonal sum vector of A, where $r_i = \sum_{j=1}^{n} a_{i,j}$, $i = 1, \cdots, m$, and $d_k = \sum_{i+j=k+1} a_{ij}$, $k = 1, \dots, m+n-1$.

For example, the image given in the Figure 2 has the following row and diagonal sums: $R = (3,4,5,4,5,5,6,2,1)$ and $D = (0,0,0,3,5,5,4,2,5,5,2,2,2,0,0,0)$.

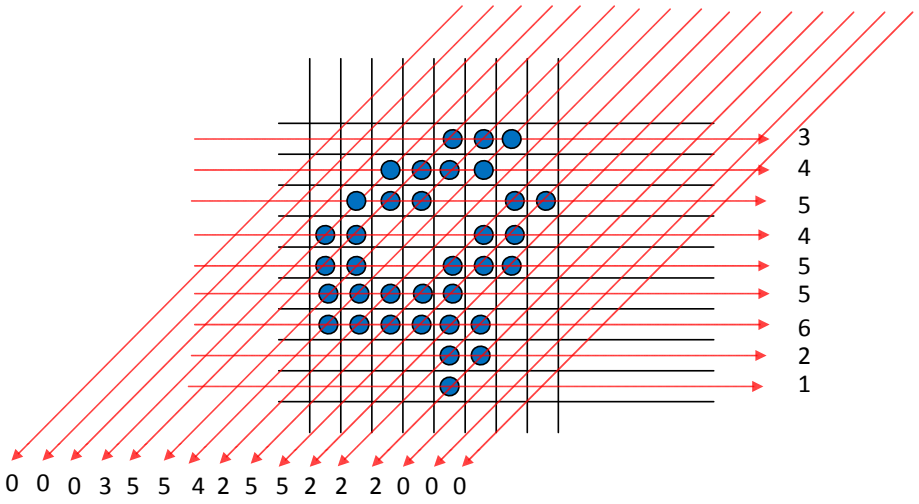

Figure 2

We say that a pair (R, D) is compatible if the following conditions hold:

$\sum_{k=1}^{m+n-1} d_k = \sum_{j=1}^{m} r_j$

$r_i \leq n$, and for $1 \leq i \leq m$;

$d_j \leq m_j$;

where m_j is the j-th component of the following $(m+n-1)$-length vector:

$$M = (1,2,3, \dots, \overbrace{\min(m,n), \min(m,n), \min(m,n)}^{m+n-1-2*(\min(m,n)-1)}, \dots, 3,2,1)$$

For a given vector $R = (r_1, r_2, \dots, r_m)$ we compose the maximal matrix of size $m \times n$ (denoted by $\bar{A}$), where each row has the following structure:

$$\overbrace{1,1,\cdots,1}^{r_i}\overbrace{0,0,\cdots,0}^{n-r_i}.$$

Let $R = (r_1, r_2, \dots, r_m)$ and $D = (d_1, \dots, d_{m+n-1})$ be a pair of compatible vectors.

We propose an algorithm for constructing a matrix A with the row sum R and diagonal sum D from the maximal matrix $\bar{A}$. Let R^i denote the collection of rows of $\bar{A}$ that intersect with the i-th diagonal line. For each i the algorithm shifts d_i ones from the rows of R^i and locates them in the i- th diagonal line of A.

To provide the performance of the algorithm we use/define several fragments in the maximal matrix and require majorization conditions/properties for each of them. These are necessary conditions for existing the matrix, and on the other hand they provide the construction of the matrix in case when such matrix exists.

Below is one of such fragments in the maximal matrix $\bar{A} = \{\bar{a}_{i,j}\}$.

Fragment 1.

For every i, $1 \le i \le min\,(m, n)$ we denote by $F1^i$ the left part of $\bar{A}$, bounded by the i-th diagonal line as shown in the Figure 3. $F1^i$ has i rows and i columns. $S^{F1\_i} = (s_1^{F1_i}, s_2^{F1_i}, \cdots, s_i^{F1\_i})$ is the column sum vector of $F1^i$, where $s_j^{F1_i} = \sum_{k=1}^{i-(j-1)} \bar{a}_{k,j}$ for each $j, 1 \le j \le i$.

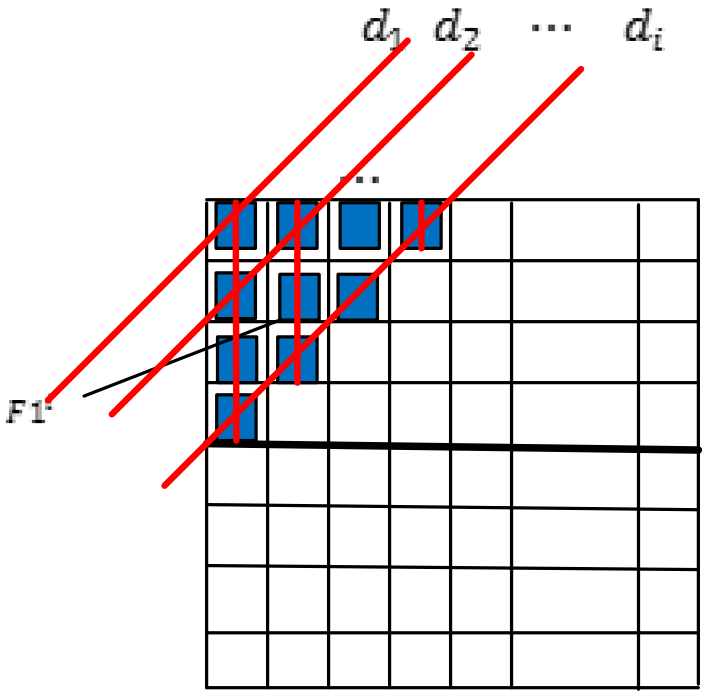

Figure 3.

Let D^{1_j} denote the initial part of D: $D^{1\_j} = (d_1, \dots, d_j)$.

Majorization 1.

For a given i, $1 \leq i \leq min\ (m,n)$ we say that S^{F1_i} majorizes D^{1_i}: $S^{F1\_i} \succcurlyeq^1 D^{1\_i}$ if for each $1 \leq j \leq i$ the following conditions hold:

$d_j \leq s_1^{F1_i}$; $d_j + d_{j-1} \leq s_1^{F1_i} + s_2^{F1_i}$; ... $d_j + d_{j-1} + \cdots + d_1 \leq s_1^{F1_i} + s_2^{F1_i} + \cdots + s_i^{F1\_i}$.

Conclusion

In this paper we give an overview of discrete tomography problems, addressing constraints/properties; complexity and approximation issues. We present also some notes on existence/reconstruction of binary images from the given horizontal and diagonal projections.

Bibliography

[PrauseOnnasch, 1996] G. Prause and D. Onnasch, Binary reconstruction of the heart chambers from biplane angiographic image sequence, IEEE Transactions Medical Imaging, 15, pp. 532-559, 1996.

[SlumpGerbrands, 1982] C. Slump, J. Gerbrands, A network flow approach to reconstruction of the left ventricle from two projections, Comput. Gr. Im. Proc., 18, pp.18-36, 1982.

[GardGrizmPran, 1999] Gardner R. J., Gritzmann P., Prangenberg D., On the computational complexity of reconstructing lattice sets from their X-rays, Discrete Mathematics, 202, pp. 45-71, (1999).

[Ryser, 1957] H. Ryser, Combinatorial properties of matrices of zeros and ones, Canad. J. Math. 9, pp.371–377, 1957.

[DLungo, 1994] A. Del Lungo, Polyominoes deffined by two vectors, Theoretical Computer Science, 127, pp.187-198, 1994.

[BarcDLungoNivatPinz, 1996] E. Barcucci, A. Del Lungo, M. Nivat, and R. Pinzani, Reconstructing convex polyominoes from horizontal and vertical projections, Theoret. Comput. Sci., 155, pp.321-347, 1996.

[Woeginger, 2001] G. Woeginger, The reconstruction of polyominoes from their orthogonal projections, Inform. Process. Lett., 77, pp. 225-229, 2001.

[DurrChrobak, 1999] Ch. Durr, M. Chrobak, Reconstructing hv-convex polyominoes from orthogonal projections, Information Processing Letters, 69, pp. 283-291, 1999.

[Berge, 1989] Berge C., Hypergraphs: Combinatorics of Finite Sets, North-Holland, 1989.

[Aslanyan, 1979] L. Aslanyan, The discrete isoperimetry problem and related extremal problems of discrete spaces, Problemy. Kibernetiki. 36, pp. 85–127, 1979 (in Russian).

[AslanyanDanoyan, 2013] L. Aslanyan, H. Danoyan, On the optimality of the Hash-Coding type nearest neighbour search algorithm, CSIT 2013 - 9th International Conference on Computer Science and Information Technologies, Revised Selected Papers, 2013, 7 pages.

[Sahakyan, 2009] H. Sahakyan, Numerical characterization of n-cube subset partitioning, Discrete Applied Mathematics, 157, 9, pp. 2191-2197, 2009.

[Sahakyan, 2014] H. Sahakyan, Essential points of n-cube subsets partitioning characterization, Discrete Applied Mathematics, 163, 2, pp. 205-213, 2014.

[Sahakyan 2013] H. Sahakyan, (0,1)-matrices with different rows, CSIT 2013 Revised Selected Papers, IEEE conference proceedings, DOI: 10.1109/CSITechnol.2013.6710342.

[Sahakyan, 2015] Sahakyan H., On the set of simple hypergraph degree sequences, Applied Mathematical Sciences, v. 9, 2015, no. 5, pp. 243-253.

[SahakyanAslanyan, 2010] H. Sahakyan, L. Aslanyan, Linear Program Form for ray different discrete tomography, International Journal "Information technologies&Knowledge", volume 4. Number 1, 41-50, 2010.

[Sahakyan, 2010] H. Sahakyan, Approximation greedy algorithm for reconstructing of (0.1)-matrices with different rows, Information Theories and Applications, Vol. 17, Number 2, pp. 124-137, 2010.

[SahakyanAslanyan, 2011] H. Sahakyan, L. Aslanyan, Evaluation of Greedy algorithm of constructing (0,1)-matrices with different rows, Information Technologies & Knowledge Vol.5, Number 1, pp. 55-66, 2011.

[SahakyanAslanyan, 2017] H.Sahakyan, L.Aslanyan On Discrete Tomography with the Constraint of Distinct Rows: Relaxation, submitted to the CSIT 2017 conference.

[BarcBrunDELunNivat, 2001] Elena Barcuccia, Sara Brunettib; Alberto Del Lungob, Maurice Nivat

Reconstruction of lattice sets from their horizontal, vertical and diagonal X-rays, Discrete Mathematics 241 (2001) 65–78

[VermaShriv, 2014] Verma S.K., Shrivastava T., Patel D. (2014) Efficient Approach for Reconstruction of Convex Binary Images Branch and Bound Method, Proceedings of the Third International Conference on Soft Computing for Problem Solving. Advances in Intelligent Systems and Computing, vol 259. Springer, New Delhi.

[SrivastavaVermaPatel, 2012];Tanuja Srivastava, Shiv Kumar Verma and Divyesh Patel, Reconstruction of Binary Images from Two OrthogonalProjections, International Journal of Tomography and Simulation, 2012, Volume: 21, Issue Number: 3.

[SrivastavaVerma, 2013], [Tanuja Srivastava, Shiv Kumar Verma , "Uniqueness algorithm with diagonal and anti-diagonal projections", International Journal of Tomography and Simulation, 2013, Volume 23, Issue number 2.

Authors' Information

Hasmik Sahakyan *– Institute for Informatics and Automation Problems of the National of Science of Armenia; Scientific Secretary. 1 P.Sevak str., Yerevan 0014, Armenia; e-mail: hsahakyan@sci.am*

Major Fields of Scientific Research: Combinatorics, Discrete tomography, Data Mining

Ani Margaryan *– Institute for Informatics and Automation Problems of the National of Science of Armenia; PhD student. 1 P.Sevak str., Yerevan 0014, Armenia; e-mail: ani.margaryan1991@gmail.com>*

Major Fields of Scientific Research: Discrete tomography algorithms

MULTICLASS DETECTOR FOR MODERN STEGANOGRAPHIC METHODS

Dmytro Progonov

Abstract: *Creation of advanced steganalysis methods for reliably detection of hidden messages in widespread multimedia files, such as digital images, is topical task today. One of the key requirements to such methods is ability to reveal the stego files even in case of limited or absent information relating to used embedding methods. For solving this task there was proposed the multiclass stegdetector, based on applying the powerful methods of digital image structural analysis. Obtained earlier results confirmed the high efficiency of proposed stegdetector by message hiding in cover image's transformation domain. There is conducted analysis of stegdetector performance in case of message hiding according to advanced adaptive steganogaphic methods, such as HILL, MiPOD and Synch algorithms. It is shown that usage the "extended" cover image model, includes not only statistical, but also correlation and fractal features, gives opportunity to improve the detection accuracy of stegdetector in most difficult cases of image steganalysis.*

Keywords: *digital images steganalysis, adaptive embedding methods, multiclass stegdetector.*

ITHEA Keywords: *K.6.5 Management of computing and information systems. Security and Protection; I.4.10 Image processing and computer vision. Image representation.*

Introduction

Protection of private as well as state-owned sensitive information is urgent problem today. Considerable quantity of freely available malware, ransomware and operation system's backdoors packets allow any users of Internet to create the personal toolbox for attacking not only private computers, but also the information infrastructures systems of governmental agencies as well as private corporations. Distinctive feature of such attacks is wide usage of complicated methods for creation the hidden communication [Cisco, 2015; Cisco, 2016; FireEye, 2015]. These channels are integrated into information flows in telecommunication systems, like email, social networks, file sharing networks, which complicates the issue of theirs detection and counteraction by state security analytics agencies.

It is worth noting that in most cases information relating to data embedding process is limited or even absent. Therefore, applying of known signature or statistical steganalysis methods does allow providing the high accuracy of stego files detection. That is why development of new steganalysis approaches,

which allow detecting the hidden messages in case of limitation or absence the advance information regard used steganography technique, are required to be developed.

Related work

For revealing the hidden communication channels there are proposed considerable numbers of targeted steganalysis methods, based on usage the signature database and statistical models of cover files, such as digital images [Fridrich, 2009; Cox et al, 2008; Böhme, 2010]. Advantage of these methods is high accuracy of hidden messages (stego files), but only when embedding method is a priory known. For improvement the performance of signature and statistical stegdetectors in case of limited information relating to used steganalysis technique, there was proposed to use the rich cover model [Fridrich and Kodovsky, 2012a], obtained by merging of several statistical models in spatial as well as JPEG domains. Nevertheless practical usage of proposed stegdetectors is limited due to ample quantity of cover's model parameters which should be computed, for instance 34,671 parameters for SRM [Fridrich and Kodovsky, 2012a] and 35,263 features for J+SRM [Fridrich and Kodovský, 2012b] models.

Alternative approach to stego image detection is based on usage the simplified or approximate cover models [Avcibas et al, 2003; Farid, 2001]. Obtained universal (blind) stegdetectors give opportunity to overcome mentioned drawbacks of targeted steganalysis methods and reveal the hidden messages when there is no information about embedding process. But usage of approximate cover model makes unfeasible elicitation of slight changes of parameters the sophisticated cover models, which are widely used in modern embedding algorithm. It leads to deterioration of stegdetectors performance, especially in case of usage the adaptive steganographic techniques, such as HUGO algorithm, MiPOD algorithm and UNIWARD family of embedding methods.

For overcome mentioned drawbacks of well-known steganalysis methods, there was proposed to use the powerful methods of structural analysis for revelation the slight changes the cover image fine structure, caused by message hiding [Progonov and Kushch, 2014a; Progonov and Kushch, 2014b; Progonov and Kushch, 2015a; Progonov and Kushch, 2015b]. Based on developed methods of structural steganalysis it was proposed the multiclass stegdetector (MCS), which gives opportunity not only reveal the stego images, but also determinate the class of steganographic methods used for theirs creation. Results of comparative analysis the performance of MCS in case of stegodata hiding in transformation domains [Progonov, 2016] confirmed the high efficiency of proposed approach. Therefore it is of interest further examination of multiclass stegdetector performance by stego image formation according to advance embedding methods.

Task and challenges

Our purpose is investigation the performance of proposed multiclass stegdetector in case of usage the modern adaptive methods for data embedding in digital images.

Advanced methods for data embedding in digital images

For message hiding in digital images, there was proposed significant number of steganographic methods. Such methods can be divided into four groups [Fridrich, 2009; Cox et al, 2008; Böhme, 2010]:

1. Model preserving methods – are designed to preserve the simplified model of the cover source. The examples of such methods are MBS1 and MBS2 algorithms.
2. Mimicking natural image processing methods – the goal of such methods is to masquerade the embedding as some natural process of images, such as noise superposition during image acquisition. In this group of steganographic methods can be included the stochastic modulation method.
3. Steganalysis-aware methods – use known steganalysis attacks as guidance for the design the embedding process. As examples it should be mentioned (±1) algorithm, F5 algorithm and HUGO algorithms.
4. Minimal-impact (adaptive) methods – are based on minimizing the total cost (impact) of data hiding during formation of stego image. The total cost is measured as sum of embedding changes at each cover image element during hiding the separate stegobit. The most well-known adaptive methods are WOW method, UNIWARD family of steganographic algorithms, Synch algorithm.

The stego scheme, based on model-preserving principle, will be undetectable as long as the chosen model completely describes the cover images. Due to lack of accurate models for real images, there are used simplified model of image, for instance based on preserving its first-order statistics or histogram [Fridrich, 2009]. Applying by stego analytic more precise model of cover image source, for example, including the high-order statistics, allows reliably detecting the stego images, formed according to such schemes.

By usage the stego methods from the second group, even if the effect of embedding were in-distinguishable from some natural processing, the obtained stego images should stay compatible with the distribution of cover images. Distinction between the cover image dataset used by steganographer and steganalytics can be used by the latter for reliably detection the formed stego images.

The most secure stego schemes for message hiding today are related to groups of steganalysis-aware and minimal-impact methods. Such methods are typically realized in two steps: firstly, compute the cost of changing each cover image's pixel with usage of predefined distortion function. Secondly, secret

message is embedding while minimizing the sum of cost of all changed pixels. Such approach gives opportunity to create high robust embedding methods, which are most challenging to steganalysis. Well-known examples of such methods are WOW [Holub and Fridrich, 2012], UNIWARD method's family [Holub et al, 2014], HILL [Li et al, 2014] and MiPOD [Sedighi et al, 2016] embedding algorithms. Let us consider such algorithm in more details.

Peculiarity of first adaptive embedding methods was usage of heuristic-defined function $\rho(\mathbf{x})$ for estimation the cover image $\mathbf{x}$ distortion due to message hiding. Applying of simplified image model, which does not capture the interpixels dependences, allows represent $\rho(\mathbf{x})$ as superposition of local disturbances ρ_{ij} of cover image's characteristics due to stegodata embedding. One of the well-known examples of such distortion function was proposed in the WOW embedding algorithm [Holub and Fridrich, 2012]:

$$\rho_{ij} = \sum_{l=1}^{L} \frac{1}{\left| \sum\limits_{(m,n)\in M\times N} \left|\mathbf{R}_{mn}^{(l)}\right| \cdot \left|\mathbf{R}_{mn}^{(l)} - \mathbf{R}_{[ij]mn}^{(l)}\right| \right|},$$

where $\mathbf{R}^{(l)} = \mathbf{x} * \mathbf{K}^{(l)}$ – lth residuals, obtained by convolution of cover image $\mathbf{x}$ and lth direction filter $\mathbf{K}^{(l)}$; $\mathbf{R}_{[ij]}^{(l)} = \mathbf{x}_{[ij]} * \mathbf{K}^{(l)}$ – lth residuals, calculated for cover image after hiding separate stegobit by altering the pixel brightness at position $\mathbf{x}_{[ij]}$; $\aleph_L = \left\{\mathbf{K}^{(1)}, \mathbf{K}^{(2)}, \ldots, \mathbf{K}^{(L)}\right\}$ – bank of directional filters; M, N – size of cover image $\mathbf{x}$. For additional decreasing the number of disturbed pixels stegodata is preprocessed with usage of syndrome-trellis codes. WOW algorithm forces the distortion to be high where the content is predictable in at least one direction (smooth areas and clean edges) and low where the content is unpredictable in every direction (as in textures).

Modification of WOW's distortion function was proposed in HILL algorithm [Li et al, 2014]:

$$\boldsymbol{\rho} = \frac{1}{\left|\mathbf{x} * \mathbf{H}\right| * \mathbf{L}_1} * \mathbf{L}_2,$$

where $\mathbf{H}$ – high-pass filter (Ker–Böhme kernel); $\mathbf{L}_1, \mathbf{L}_2$ – correspondingly, low-pass (averaging) filter of support 3×3 and 15×15 pixels. Low-pass filtering of the costs $\boldsymbol{\rho}$ allows improving empirical security

due to increasing the entropy of embedding changes in highly textured regions and, therefore, reducing the distortion for the same payload.

Further development of WOW's distortion function is universal wavelet relative distortion (UNIWARD) [Holub et al, 2014]:

$$\rho(\mathbf{x},\mathbf{y}) = \sum_{k=1}^{3}\sum_{u=1}^{n_1}\sum_{v=1}^{n_2}\frac{\left|W_{uv}^{(k)}(\mathbf{x}) - W_{uv}^{(k)}(\mathbf{y})\right|}{\sigma + \left|W_{uv}^{(k)}(\mathbf{x})\right|},$$

where $\mathbf{x},\mathbf{y}$ – correspondingly cover and stego images; $W_{uv}^{(k)}(\mathbf{x})$ – uvth wavelet coefficient in the kth subband of the first level the two-dimensional discrete wavelet transformation the cover image; $\sigma(\sigma > 0)$ – constant stabilizing the numerical calculations. Usage of proposed distortion function allows create the state-of-art uniform approach to cover image parameters disturbances regardless of the message embedding domain [Holub et al, 2014].

It should be noted, that considered embedding algorithms allow minimize the distortion of cover image parameters by message hiding, but do not taking into account the statistical detectability of obtained stego images. Design of distortion functions that measure cover image distortions as well as statistical detectability of formed stego images is one of open problems in digital image steganography today [Ker et al, 2013]. For solve this problem there were proposed various approaches, based on usage only pixels, which have the smallest impact on the empirical statistical distribution of pixels groups [Pevný, 2010] or usage the distortion functions, which are optimized to minimize the empirical detectability in terms of the margin between cover and stego images represented using low-dimensional features [Filler and Fridrich, 2011]. These approaches are limited to empirical "models" that need to be learned from a database of images and, therefore, may become highly detectable should the Warden choose a different feature representation [Filler and Fridrich, 2011]. For overcome mentioned drawback there was proposed to model the cover pixels as a sequence of independent Gaussian random variables with unequal variances (multivariate Gaussian or MVG). It gives opportunity to achieve the empirical security of the embedding methods, which was subpar with respect to state-of-the-art steganographic methods [Holub and Fridrich, 2012; Holub et al, 2014]. Example of steganographic techniques, based on such approach, is MiPOD embedding method, which uses the locally-estimated multivariate Gaussian cover image model.

Message hiding in grayscale cover image $\mathbf{x}$ with size $M \times N$ (pixels) according to MiPOD method is carried out in several steps [Sedighi et al, 2016]:

1. Suppress the image content $\mathbf{x} = (x_1, x_2, \ldots, x_L)$, $L = M \cdot N$, using a denoising filter F:

$$\mathbf{r} = \mathbf{x} - F(\mathbf{x}),$$

where $\mathbf{x}$ is represented in column-wise order;

2. Measure pixels residual variance σ_l^2 using Maximum Likelihood Estimation and local parametric linear model:

$$\mathbf{r}_l = \mathbf{G}\mathbf{a}_l + \xi_l, \tag{1}$$

where $\mathbf{r}_l$ – represents the value of the residual $\mathbf{r}$ inside the $p \times p$ block surrounding the l[th] residual put into a column vector of size $p^2 \times 1$; $\mathbf{G}$ – a matrix if size $p^2 \times p$ that defines the parametric model of remaining expectation; $\mathbf{a}_l$ – a vector of $q \times 1$ of parameters; ξ_l – the signal whose variance is need to be estimated.

The pixels residual variance σ_l^2 is estimated according to further formula:

$$\sigma_l^2 = \frac{\left\| \mathbf{P}_{\mathbf{G}}^{\perp} \mathbf{r}_l \right\|^2}{p^2 - q},$$

where $\mathbf{P}_{\mathbf{G}}^{\perp} = \mathbf{I}_l - \mathbf{G}\left(\mathbf{G}^T\mathbf{G}\right)^{-1}\mathbf{G}^T$ – the orthogonal projection of the residual $\mathbf{r}_l$, estimated according to (1), onto the $p^2 - q$ dimensional subspace spanned by the left null space of $\mathbf{G}$; $\mathbf{I}_l$ – the $l \times l$ unity matrix.

3. Determine the probability of l[th] embedding change $\beta_l, I \in \{1, 2, \ldots, L\}$ that minimize the deflection coefficient ς^2 between cover and stego image distributions:

$$\varsigma^2 = 2\sum_{l-1}^{L} \beta_l^2 \sigma_l^{-4}, \tag{2}$$

under payload constrain

$$R = \sum_{l=1}^{L} H(\beta_l),$$

where $H(z) = -2z\log z - (1-2z)\log(1-2z)$ – is ternary entropy function; R – cover image payload in nats.

Minimization of (2) can be achieved by using the method of Lagrange multipliers. The change rate β_l and the Lagrange multiplier λ can be determined by numerically solving of further $(I+1)$ equations:

$$\beta_l \sigma_l^{-4} = \frac{1}{2\lambda} \ln\left(\frac{1-2\beta_l}{\beta_l}\right), l \in \{1,2,\ldots,L\},$$

$$R = \sum_{l=1}^{L} H(\beta_l).$$

4. Convert the change rate β_l to cost ρ_l:

$$\rho_l = \ln(1/\beta_l - 2); \quad (3)$$

5. Embed the desired payload R using syndrome-trellis codes (STCs) with pixel costs determined according to (3).

Applying the locally-estimated multivariate Gaussian cover model in MiPOD algorithm gives opportunity to derive a closed-form expression for the performance of the detector but, at the same time, complex enough to capture the non-stationary character of natural images [Sedighi et al, 2016].

Mentioned additive distortion functions use simple assumption that cost of not making a change is always zero. It does not take into account the influence of surrounding pixels on analyzed pixel's brightness value, which leads to underestimate the cover image distortion by message hiding. Therefore it was proposed to use the non-additive distortion functions for improving the empirical security of embedding schemes [Denemark and Fridrich, 2015].

In the work there was also investigated the case of usage the Synch scheme [Denemark and Fridrich, 2015] for improving the MiPOD embedding algorithm. The main steps of stegodata $\mathbf{m}$ embedding in in grayscale cover image $\mathbf{x}$ with size $M \times N$ (pixels) according to Synch-MiPOD algorithm are:

1. Divide message into two equal size parts:

$$\mathbf{m} = \mathbf{m}_1 \cup \mathbf{m}_2;$$

2. Compute the cost $\rho_{ij}, i \in \{1,2,\ldots,N\}, j \in \{1,2,\ldots,M\}$ from the cover image $\mathbf{x}$ according to formula (3);

3. Set stego image $\mathbf{y}$ is equal to cover image $\mathbf{x}$;

4. For each stego image pixel compute the cost of its modification in range $\Delta \in \{-1;0;+1\}$:

$$\rho_{ij}^{(+1)} = D_A\left(\mathbf{y}, x_{ij} + 1\mathbf{y}_{\sim ij}\right), \tag{4}$$

$$\rho_{ij}^{(0)} = D_A\left(\mathbf{y}, x_{ij}\mathbf{y}_{\sim ij}\right), \tag{5}$$

$$\rho_{ij}^{(-1)} = D_A\left(\mathbf{y}, x_{ij} - 1\mathbf{y}_{\sim ij}\right), \tag{6}$$

where

$$D_A\left(\mathbf{x},\mathbf{y}\right) = \sum_{x_{ij}\neq y_{ij}} D\left(\mathbf{x}, \mathrm{y}_{ij}\ \mathbf{x}_{\sim ij}\right)$$

is additive approximation of the distortion function

$$D\left(\mathbf{x},\mathbf{y}\right) = \sum_{((i,j),(k,l))\in\wp} S_{\wp}\left(x_{ij} - y_{ij}, x_{kl} - y_{kl}\right),$$

$$S_{\wp}\left(a,b\right) = \begin{cases} 0 & when \quad a = b, \\ A_{\wp} & when \quad |a| + |b| = 1, \\ \nu A_{\wp} & when \quad (a \neq b) \wedge \left(|a| + |b| = 2\right), \end{cases}$$

$\wp$ – index set of all two-pixels cliques formed by two vertically and horizontally adjacent pixels; $A_{\wp} = \left(\rho_{ij} + \rho_{kl}\right)/2$ – average clique cost; $\nu\,(\nu \geq 0)$ – parameter controlling the strength of penalizing desynchronized changes; $y_{ij}\mathbf{x}_{\sim ij}$ – shorthand for $\mathbf{x}$ in which only the (i,j) pixel x_{ij} was changed to y_{ij};

5. Embed i[th] element of message $\mathbf{m}_q$ into cover image, by taking into consideration the computed costs (4)-(6), with usage of STCs;

6. Repeat steps #4-5 q times ($q = 2$);

7. Repeat step #6 k times ($k \in \{1,2,\ldots,K\}$).

Embedding with different costs of all three possibilities $\{-1;0;+1\}$ requires the use of the so-called multi-layer STCs [Filler et al, 2011]. It should be mentioned that the costs A_{ij} are computed only once before the embedding starts and are kept the same throughout the embedding, i.e., they are not recomputed after every k sweep. Finally, the recipient reads the secret message using the same STCs applied to each sublattice and concatenating both parts.

Structural steganalysis of digital images

The most common approach to revealing the stego image is based on analysis the alteration of cover image's statistical characteristics, such as first-order statistics, second-order statistics and so on [Fridrich, 2009; Böhme, 2010]. There was proposed considerable number of powerful statistical steganalysis methods, based on applying the rich models if cover image in spatial (SPAM, SRM models) as well as JPEG (CC-PEV, CC-JRM models) domains. Despite of high accuracy the stego image detection, there is significant limitation of practical usage of mentioned methods, connected with great number of model's parameters, for instance 22,510 parameters for CC-JRM [Fridrich and Kodovský, 2012b] model, 34,671 parameters for SRM model [Fridrich and Kodovsky, 2012a]. It leads to sizeable increasing the stegdetector tuning and image processing times, which is inappropriate for real-time detection systems.

For overcome mentioned drawback of statistical steganalysis methods, there was proposed to use the powerful methods of digital image structural analysis, such as variogram analysis [Progonov and Kushch, 2014b], multifractal detrended fluctuation analysis [Progonov and Kushch, 2014a; Progonov and Kushch, 2015a] and multifractal analysis [Progonov and Kushch, 2015b].

Variogram analysis is widely used for investigation the correlation characteristics of time series $\mathbf{I}(\mathbf{s})$ and based on usage the variogram function [Cressie and Wikle, 2011]:

$$2\gamma_{\mathbf{I}}(h) = 2\left(C_{\mathbf{I}}(0) - C_{\mathbf{I}}(h)\right),$$

where $C_{\mathbf{s}}(h) = \mathrm{cov}\left(\mathbf{I}(\mathbf{s}), \mathbf{I}(\mathbf{s}+h)\right)$ – covariation of values the time series adjacent elements; h – time shift (lag). In most applications further approximation of variogram $2\gamma_{\mathbf{I}}(h)$ is used [Cressie and Wikle, 2011]:

$$2\hat{\gamma}_{\mathbf{I}}(h) = \frac{1}{|N_h|}\sum_{i,j\in N_h}\left(\mathbf{I}(s_i) - \mathbf{I}(s_j)\right)^2, N_h = \left\{(i,j): s_i - s_j = h\right\},$$

where N_h – set of possible pairs of position the elements, when distance between them is equal to h. Usage of variogram approximation $2\hat{\gamma}_{\mathbf{I}}(h)$ allows estimate such correlation characteristics of time series [Cressie and Wikle, 2011]:

1. Nugget-effect – the value of correlation between adjacent elements if time series:

$$N_{\mathbf{I}} = 2\hat{\gamma}_{\mathbf{I}}(h)\big|_{h=1},$$

2. Sill – the value of maximal variance the time series element's values:

$$S_{\mathbf{I}} = 2\hat{\gamma}_{\mathbf{I}}(h)\big|_{h\to+\infty},$$

3. Range – the interval of correlation between values of adjacent elements of time series:

$$R_{\mathbf{I}} = \max\left\{h : \left(1 - \frac{2\hat{\gamma}_{\mathbf{I}}(h)}{S_{\mathbf{I}}}\right) \geq \varepsilon_R\right\}, \varepsilon_R \in \mathbb{R}_+.$$

Value of range $R_{\mathbf{I}}$ usually is determined when correlation between adjacent elements is not less than 10% [Cressie and Wikle, 2011]. Despite of high accuracy estimation of image correlation parameters by usage of variogram analysis, this approach has limited opportunity to investigate the parameters of separate image components like intrinsic noise, contours etc. It requires applying the specialized processing methods, such as multifractal detrended fluctuation analysis (MF-DFA).

MF-DFA is generalization of well-known detrended fluctuation analysis and allows not only estimate the Hurst coefficient H values, but also investigate the multifractal nature of intrinsic noise of time series [Kantelhardt et al, 2002] – spectrum of generalized Hurst exponents $h(q)$ as well as multifractal spectrum $f_h(\alpha_h)$. Variation of scaling parameter q gives opportunity to estimate the generalized Hurst exponent $h(q)$ for time series element's value fluctuation with small ($q < 0$) and ($q > 0$) large amplitude. On the other hand, discrete values of multifractal spectrum $f_h(\alpha_h)$ correspond to Hausdorff dimension of the analyzed signal subset, which exponent of Hölder condition is equal to α_h. Values of α_h are varied between $\alpha_h = \alpha_h^{\min}$, which corresponds of signal components with minimal fluctuations between adjacent pixels, to $\alpha_h = \alpha_h^{\max}$, which corresponds to most "irregular" components.

Increasing of stegdetector performance requires improving the used model of cover source. Besides the widely used statistical and correlation characteristics of cover images, it is also of interest to include the

cover-specific features, such as fractality – preserving the statistical characters on the different scales [Peitgen et al, 2014]. Multifractal analysis allows extend the opportunity of "classical" fractal analysis – gives opportunity to investigate the fractal properties of image components with usage of spectrum the generalized fractal dimensions (Renie spectrum) D_q as well as multifractal spectrum $f(\alpha)$. Spectrum D_q allows not only estimate the Hausdorff dimensions of image components with various average brightness, in particular case minimal and maximal, which are correspond to D_q^{MIN} and D_q^{MAX}, but also information (D_1) and correlation (D_2) dimensions. The former characterizes the growth rate of the Shannon entropy given by successively finer discretizations of the space, while the latter is a measure of the dimensionality of the space occupied by a set of random points.

Variogram analysis, multifractal detrended fluctuation analysis and multifractal analysis of digital images was performed according to algorithms, described in [Progonov, 2016].

Multiclass stegdetector for digital images

Joint use of mentioned methods the structural steganalysis allows not only detect the stego images, but also carry out the forensic steganalysis – ascertains the domain, where message has been hidden, estimates the payload, and determines the processing chain of cover image as well as stegodata [Progonov, 2016]. Based on these results there was developed the multiclass stegdetector, which structural scheme is shown at Fig. 1.

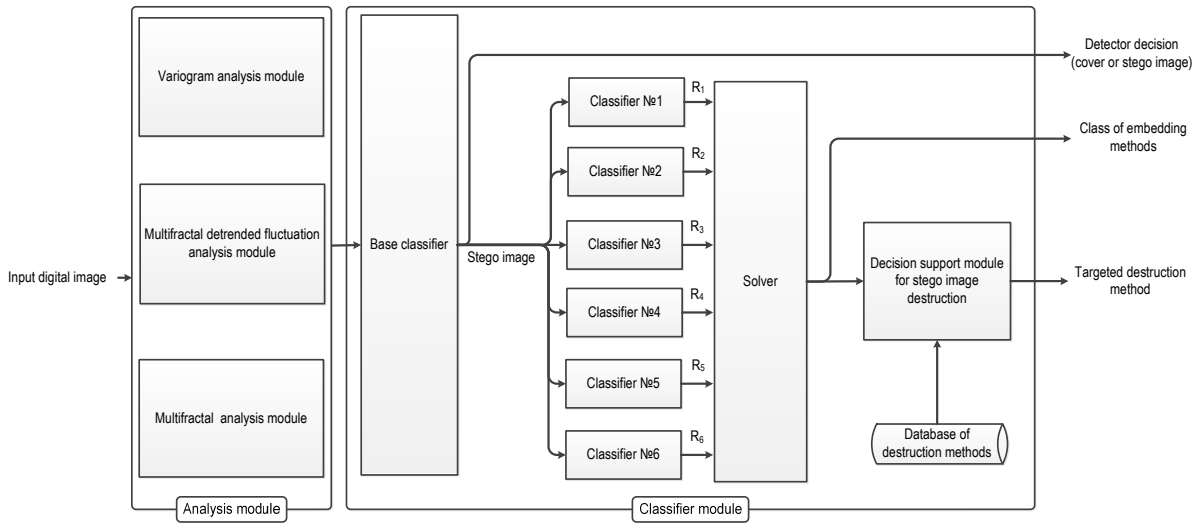

Figure 1. The flowchart of digital image processing by proposed generalized multiclass stegdetector

The stegdetector consists of two parts (Fig. 1) – analysis and classifier modules. Former module is subdivided into three modules, namely variorgam, multifractal detrended fluctuation and multifractal analyses, which are used for determination the statistical, correlation and fractal characteristics of inputted image. Obtained features are transferred to classifier module (Fig. 1). At first stage, base classifier map processing image to class of covers or stegos, depending on obtained feature values. If image is classified as containing the hidden messages (stego image, Fig. 1), additional classifier's submodules are applied for determine the class of steganographic techniques used for stego creation (Table 1).

Table 1. Classifiers for determination the class of used steganographic technique

Classifier number	Cover processing chain	Stegodata processing chain
1	One-stage, common transformation (Fourier, cosine or wavelet discrete transformations)	–
2	One-stage, uncommon transformation (for instance Singular Value Decomposition)	–
3	Two-stage, composition of common and uncommon transformations	–
4	Three-stage, composition of common and uncommon transformations	–
5	One-stage, common transformation (Fourier, cosine or wavelet discrete transformations)	Scrambling transformation
6	Two-stage, common transformation (Fourier, cosine or wavelet discrete transformations)	Scrambling transformation

Each classifier (Fig. 1) calculates the probability $P_i, i \in \{1,2,\ldots,6\}$ that analyzed image has been modified according to embedding method, belonging to corresponding class of steganographic techniques. Identification of most probable class the steganographic methods, used for creation of analyzed image, is carried out in decision support module by comparison of obtained probabilities P_i and determination of maximum probability P_i^{MAX}. According to decision of MCS there are also shown

recommendation for choosing the most effective (targeted) method for destruction the revealed stego image.

Experiments

For estimation the accuracy of stego image revealing by usage of proposed multiclass stegdetecor there were conducted the tuning and testing of MCS on test packet of 2,500 digital images from MIRFlickr-25k dataset [Huiskes and Lew, 2008]. Test packet was divided into training (1,250 images) and testing (1,250 images) subpacket in a pseudorandom manner. All images were scaled to the same size 512×512 pixels with usage of Lanczos kernel and saved in lossless JPEG format (Image Quality Factor is equal to 100%).

Payload of cover image was varied from 5% to 25% with step 5% and from 25% to 95% with step 10%.

Training of MCS was conducted with usage of image characteristics, obtained by applying the variogram analysis (39 parameters), multifractal detrended fluctuation analysis (182 parameters) and multifractal analysis (14 parameters). Estimation of mentioned features was carried out according to developed algorithms, represented in [Progonov, 2016]. Total number of used image features is equal to 235.

Testing of tuned MCS was repeated 10 times with reinitialize of training and testing subpackets. The averaged probabilities of cover and stego images attribution to steganographic technique's classes (Table 1) are shown in Table 2. For sake of convenience, the largest values of probabilities $P_i, i \in \{1,2,\ldots,6\}$ for each embedding methods are marked in thick print and underlined.

It should be mentioned that usage of proposed multiclass stegdetector allows correctly determine the cover image processing chain in case of usage the WOW and S-UNIWARD embedding methods (Table 2) – applying of common (two-dimensional discrete wavelet transformation) and specific (minimizing the distortion function value) processing methods. On the other hand, minor changes of WOW embedding scheme in HILL algorithm leads to misclassify the obtained stego images by MCS as formed according to simple embedding methods in frequency domain (class #1, please see Table 2). Obtained classification results for mentioned embedding methods remain permanent even in case of high cover image payload (more than 50%, Table 2).

In case of applying the modern adaptive steganographic schemes like MiPOD and Synch-MiPOD, multiclass stegdetector misclassify incorrectly classify obtained stego images as formed according to multistage embedding methods (Table 2), despite any cover transformations have not been applied. Misclassification of stego image in such case can be explained by disparity of used cover image model – multivariate Gaussian image model in MiPOD algorithm and union of Markov and fractal models in proposed MCS.

Table 2. Averaged probabilities of stego images attribution to considered steganographic technique's classes in case of low (10%) and high (85%) payload of cover image

Cover image payload	Embedding method	Steganographic technique's class					
		#1	#2	#3	#4	#5	#6
10%	WOW	0.39	0.17	***0.47***	0.28	0.09	0.08
	HILL	***0.42***	0.28	0.33	0.06	0.01	0.12
	S-UNIWARD	0.08	0.07	***0.58***	0.01	0.38	0.01
	MiPOD	0.02	0.15	0.34	***0.41***	0.22	0.19
	Synch-MiPOD	0.27	0.01	0.21	***0.46***	0.31	0.07
85%	WOW	0.63	0.41	***0.77***	0.22	0.11	0.18
	HILL	***0.91***	0.66	0.74	0.13	0.10	0.23
	S-UNIWARD	0.07	0.02	***0.98***	0.03	0.14	0.01
	MiPOD	0.01	0.28	***0.71***	0.68	0.07	0.11
	Synch-MiPOD	0.18	0.02	0.12	***0.89***	0.22	0.02

Conclusion

On the basis of conducted comprehensive analysis of performance the proposed multiclass stegdetector it is established that:

– It is confirmed the high efficiency of stegdetector even in case of investigation the stego images, formed according to a priory unknown embedding methods. Ability of stegdetector to determine the class of steganographic techniques, used for stego image creation, allows choose the targeted methods for hidden message destruction with minimal impact on cover image visual quality;

– Applying of adaptive embedding methods, based on usage the uncommon (multivariate Gaussian) cover model for stego image creation, allows significantly decrease the accuracy of it detection by usage of multiclass stegdetector. It is explained by usage of steganalytic "simplified" digital image model, which capture the most general features (fractality, correlation of brightness the adjacent

pixels) and has limited opportunity to represent the complicated local dependences in high-textured area of image. Overcome the revealed limitation requires creation the generalized image model for accurate capture the various features of real images, such as non-stationarity and heterogeneity.

Acknowledgement

The paper is published with partial support by the project ITHEA XXI of the ITHEA ISS (www.ithea.org) and the ADUIS (www.aduis.com.ua).

Bibliography

[Avcibas et al, 2003] Avcibas I., Memon N., Sankur B. Steganalysis using image quality metrics. IEEE Transactions on Image Processing. Volume 12, Issue 2, 2003. pp. 221–229. DOI 10.1109/TIP.2002.807363;

[Böhme, 2010] Böhme R. Advanced Statistical Steganalysis. Springer, 2010. 285 p. ISBN (eBook) 978-3-642-14313-7. ISBN (Hardcover) 978-3-642-14312-0. DOI: 10.1007/978-3-642-14313-7;

[Cisco, 2015] Cisco Systems, Inc., Annual Security Report. http://www.cisco.com/c/dam/assets/-about/ar/pdf/2015-cisco-annual-report.pdf;

[Cisco, 2016] Cisco Systems, Inc., Annual Security Report. http://www.cisco.com/c/dam/en_us/about/-annual-report/2016-annual-report-full.pdf;

[Cox et al, 2008] Cox I. J., Miller M. L., Bloom J. A., Fridrich J., Kalker T. Digital Watermarking and Steganography. Elsevier, 2008. 593 p.;

[Cressie and Wikle, 2011] Cressie N., Wikle C. Statistics for Spatio-Temporal Data. Wiley, 2011. 624 p.

[Denemark and Fridrich, 2015] Denemark T., Fridrich J. Improving Steganographic Security by Synchronizing the Selection Channel. Proceedings of the 3rd ACM Workshop on Information Hiding and Multimedia Security, 2015. DOI 10.1145/2756601.2756620;

[Farid, 2001] Farid H. Detecting Steganographic Messages in Digital Images. Technical Report. Dartmouth College Hanover, 2001. p.9;

[Filler and Fridrich, 2011] T. Filler, J. Fridrich. Design of adaptive steganographic schemes for digital images. Proceedings SPIE, Electronic Imaging, Media Watermarking, Security and Forensics III. Edited by A. Alattar, N. D. Memon, E. J. Delp, and J. Dittmann. 2011. pp. 1-14;

[Filler et al, 2011] T. Filler, J. Judas, J. Fridrich. Minimizing additive distortion in steganography using syndrome-trellis codes. IEEE Transactions on Information Forensics and Security. Volume 6(3), 2011. pp.920–935;

[FireEye, 2015] FireEye, Inc., HAMMERTOSS: Stealthy Tactics Define a Russian Cyber Threat Group. https://www2.fireeye.com/rs/848-DID-242/images/rpt-apt29-hammertoss.pdf;

[Fridrich, 2009] J. Fridrich Steganography in Digital Media: Principles, Algorithms, and Applications. 1st Edition. Cambridge University Press, 2009. p. 437. ISBN 978–0–521–19019–0;

[Fridrich and Kodovský, 2012a] Fridrich J., Kodovský J. Rich Models for Steganalysis of Digital Images. IEEE Transactions on Information Forensics and Security. Volume 7, Issue 3, 2012. pp. 868-882.

[Fridrich and Kodovský, 2012b] Fridrich J., Kodovský J. Steganalysis of JPEG images using rich models. Proceedings SPIE 8303, Media Watermarking, Security, and Forensics Edited by Memon Nasir D., Alattar Adnan M., Delp Edward J. doi:10.1117/12.907495;

[Holub and Fridrich, 2012] Holub V., Fridrich J. Designing Steganographic Distortion Using Directional Filters. Proceedings of IEEE Workshop on Information Forensic and Security. 2012;

[Holub et al, 2014] Holub V., Fridrich J., Denemark T. Universal distortion function for steganography in an arbitrary domain. EURASIP Journal on Information Security. 2014.

[Huiskes and Lew, 2008] Huiskes M.J., Lew M.S. The MIR Flickr Retrieval Evaluation. Proceedings of ACM International Conference on Multimedia Information Retrieval. 2008;

[Kantelhardt et al, 2002] Kantelhardt J. W., Zschiegner S. A., Koscielny-Bunde E., Bunde A., Havlin S., Stanley H. E. Multifractal detrended fluctuation analysis of nonstationary time series. Cornell University Library. Electronic Archive, 2002. https://arxiv.org/abs/physics/0202070;

[Ker et al, 2013] Ker A. D., Bas P., Böhme R., Cogranne R., Craver S., Filler T., Fridrich J., Pevný T. Moving steganography and steganalysis from the laboratory into the real world. Proceedings of the first ACM workshop on Information hiding and multimedia security (IH&MMSec '13). New York, 2013;

[Li et al, 2014] B. Li, M. Wang, J. Huang. A new cost function for spatial image steganography. Proceedings of IEEE International Conference on Image Processing (ICIP-2014);

[Peitgen et al, 2014] H.-O. Peitgen, J. Hartmut, D. Saupe. Chaos and Fractals. New Frontiers of Science. 2nd Edition. Springer, 2004. 864 p.;

[Pevný, 2010] Pevný T., TFiller T., Bas P. Using High-Dimensional Image Models to Perform Highly Undetectable Steganography. Proceedings of International Workshop on Information Hiding (IH 2010). Edited by Böhme R., Fong P.W.L., Safavi-Naini R. Springer, Berlin, Heidelberg;

[Progonov and Kushch, 2014a] Progonov D.O., Kushch S.M. Revealing of stego images with data, embedded in cover image transformation domain [In Ukrainian]. Bulletin of National Technical University of Ukraine. Series Radiotechnique. Radioapparatus Building. Vol. 57, 2014. pp. 128-142;

[Progonov and Kushch, 2014b] Progonov D.O., Kushch S.M. Variogram analysis of steganograms forme accrding to complex embedding methods. Bulletin of National Technical University of Ukraine "Lviv Polytechnic". Series Information systems and networks. Volume 806, 2014. pp.226-232;

[Progonov and Kushch, 2015a] Progonov D.O., Kushch S.M. Multifractal Detrended Fluctuation Analysis of steganograms. System research and information technologies. Volume 4, 2015. pp. 39-47;

[Progonov and Kushch, 2015b] Progonov D.O., Kushch S.M. Spectral analysis of steganograms. Radio Electronics, Computer Science, Control. Volume 2 (33), 2015. pp. 71-81;

[Progonov, 2016] Progonov D.O. Structural methods of digital image passive steganalysis. PhD Thesis. National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", 2017. 293 p.;

[Sedighi et al, 2016] Sedighi V., Cogranne R., Fridrich J. Content-Adaptive Steganography by Minimizing Statistical Detectability. IEEE Transactions on Information Forensics and Security. Vol. 11, Iss. 2., 2016. pp. 221-234.

Authors' Information

Dmytro Progonov – *The Institute of Physics and Technology, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"; PhD, Associate Professor; 37, ave. Peremohy, Solomenskiy district, Kyiv, Postcode 03056, Ukraine; e-mail: progonov@gmail.com*

Major Fields of Scientific Research: digital media steganalysis, digital image forensics, machine learning, advanced signal processing

Conjunctive Boolean Query as a logic-objective recognition problem

Tatiana Kosovskaya

Abstract: *A well-known NP-complete problem Conjunctive Boolean Query is considered as the one of the logic-objective recognition problems. Both these problems have the same formulation but their implementations are rather different. It is offered to adapt the technique which involves a decreasing of computational complexity for a logic-objective recognition problem by means of construction of a level class description to the solution of Conjunctive Boolean Query.*

Keywords: *NP-completeness, Conjunctive Boolean Query, logic-objective recognition, multi-level description*

ACM Classification Keywords*: F.1.3 Complexity Measures and Classes, Reducibility and completeness; F.2.1 Analysis of Algorithms and Problem Complexity, Numerical Algorithms and Problems, Number-theoretic computations*

Introduction

Many recent scientific investigations are devoted to the analysis of algorithms solving different NP-complete problems. Essential attention is given to repeatedly solved ones with very big input data. Two problems "Conjunctive Boolean Query" and "Satisfiability in a Finite Interpretation" (used for the solving of a logic-objective recognition problem) having the same formulations but essentially different implementations are under consideration in the paper. In particular, the input data of these problems may be divided into two parts. While solving the first problem one part of input data remains practically fixed and the other changes while every query. While solving the other problem the first part changes while every query and the other remains practically fixed.

This difference does not allow to use directly the technique which involves a decreasing of computational complexity for a logic-objective recognition problem by means of construction of a level class description Kosovskaya, [2008] to the solution of Conjunctive Boolean Query.

While creating and the use of a data base the time of data processing is one of the most important parameters. It is essentially significant because of huge volume of information stored in contemporary data bases. Conjunctive Boolean Query is one of NP-complete problems concerning data bases. Here is its formulation in the form as it is done in Garey, Johnson, [1979].

Conjunctive Boolean Query (Garey, Johnson, [1979])

Instance: Finite domain set D, a collection $R = \{R_1, R_2, \ldots, R_n\}$ of relations, where each R_i consists of a set of d_i-tuples with entries from D, and a conjunctive Boolean query Q over R and D, where such a query Q is of the form

$$\exists y_1, y_2, \ldots, y_l (A_1 \,\&\, A_2 \,\&\, \ldots \,\&\, A_r)$$

with each A_i of the form $R_j(u_1, u_2, \ldots, u_{d_j})$ where each $u \in \{y_1, y_2, \ldots, y_l\} \cup D$.

Question: Is Q, when interpreted as a statement about R and D, true?

As far as NP-complete problems are the problems of the form $\exists Y \; P(X, Y)$, where X is input data, let's give another formulation of the above mentioned problem.

Conjunctive Boolean Query

Instance: Finite domain set D, a collection $R = \{R_1, R_2, \ldots, R_n\}$ of predicates, where each R_i defines a d_i-ary relation between entries from D, a set $S(D)$ of all atomic formulas with predicates from R which are true on D, and a conjunctive Boolean query Q over R and D, where such a query Q is of the form $A_1 \,\&\, A_2 \,\&\, \ldots \,\&\, A_r$ with each A_i of the form $R_j(u_1, u_2, \ldots, u_{d_j})$ where each $u \in \{y_1, y_2, \ldots, y_l\} \cup D$.

Question: Is $\exists y_1, y_2, \ldots, y_l Q$, when interpreted as a statement about R and D, true?

That is whether

$$S(D) \Rightarrow \exists y_1, y_2, \ldots, y_l (A_1 \,\&\, A_2 \,\&\, \ldots \,\&\, A_r)?$$

Such setting of the problem Conjunctive Boolean Query is very similar to the earlier investigated in Kosovskaya, [2007, 2008] problem Satisfiability in a Finite Interpretation appeared while recognition of an object in the frameworks of logic-objective approach to the pattern recognition.

Satisfiability in a Finite Interpretation

Instance: *A set* $\omega = \{\omega_1, \ldots, \omega_t\}$,

a collection of predicates $\{p_1, \ldots, p_n\}$, *setting properties of elements from* ω *and relations between them,*

a collection $S(\omega)$ *of true constant atomic predicate formulas of the form* $p_i(\overline{\tau})$, *where* $i = 1, \ldots, n$, $\tau \subseteq \omega$,

quantifier-less formula $A(\overline{y})$, *presented in the form of disjunction of elementary conjunctions of atomic predicate formulas*[1].

Question: *Is there exist a list of values for* $\overline{y}$ *from* ω^a, *such that the formula* $A(\overline{y})$ *is true?*

That is whether

$$S(\omega) \Rightarrow \exists \overline{y} A(\overline{y})?$$

Essential difference in implementation of these problems consists in the following:

— data base may be not changeable at all or have very small changes, but queries may differ every time ($S(D)$ is fixed, but the query Q often may be changed);

— while pattern recognition the set of goal formulas (description of classes) may be not changeable at all or has changes very rarely, but the recognized objects may be different every time (the set of all possible formulas $A(\overline{y})$ is fixed, but the object ω and its description $S(\omega)$ often may be changed).

Problems of logic-objective recognition

Investigated objects in many Artificial Intelligence problems may be described in the terms of properties of their parts and relations between them. In such a case an investigated object ω may be represented as a set of its parts $\omega = \{\omega_1, ..., \omega_t\}$ and the properties and relations between these parts are described by predicates $p_1, ..., p_n$ defined on them. Facts which are known for an investigator over an object ω are defined by the set of constant atomic formulas $S(\omega)$ which is called the description of ω.

Below the notation $\overline{x}$ will be used to designate an ordered list of variables $\overline{x} = (x_1, ..., x_m)$. In particular, $\overline{\omega}$ designate some ordered list of elements of ω corresponding to some permutation of its elements.

The goal condition of the problem may be represented by such a formula $A(\overline{x})$ of a formalized language that if the formula $A(\overline{\omega})$ is valid for an investigated object ω then the problem has a positive solution. Moreover the goal condition may be represented by a quantifier-free formula in the form of disjunction of elementary conjunctions of atomic predicate formulas.

The following problem may be considered as formalization for an essential example of Artificial Intelligence problems.

Identification problem. *To extract a part of the object* ω *satisfying the goal condition* $A(\overline{x})$.

This problem may be reduced to checking the formula Kosovskaya, [2007]

$$S(\omega) \Rightarrow \exists \overline{x}_{\neq} A(\overline{x}),$$

(where $\exists \overline{x}_{\neq}$ means "$\exists \overline{x}$ and its elements are distinct in pairs") and is an NP-complete one.

[1] $\overline{y} = (y_1, \ldots, y_a)$ be a list of objective variables of the formula

If one can solve the problem $S(\omega) \Rightarrow \exists \overline{x}_{\neq} A(\overline{x})$, where $A(\overline{x})$ is a conjunction of atomic formulas then he can solve the problem with $A(\overline{x})$ be a disjunction of elementary conjunctions, and the number of steps of its solution would differ from the first one polynomially. That is why the complexity bounds of algorithms will be done for this problem with $A(\overline{x})$ be a conjunction of atomic formulas.

The **exhaustive search method** is one which allows not only to prove the sequent but also to finds values for variables $\overline{x}$. It is proved in Kosovskaya, [2007] that its number of steps is

$$O(t^m),$$

where t is the number of elements in ω, m is the number of variables in the formula $A(\overline{x})$.

Logical methods (namely logical derivation in a sequent calculus or resolution method) also allow to finds values for variables $\overline{x}$. Both these methods has the number of steps

$$O(\sum_{i:p_i \; is \; in \; A(\overline{x})} s_i^{a_i}),$$

where s_i and a_i are the numbers of occurrences of the predicate p_i in the description $S(\omega)$ and in the formula $A(\overline{x})$ respectively.

One can see that these upper bounds of number of steps of the algorithms have different parameters in the exponent of the power. So a researcher may choose the method in applications in dependence of the structure of the initial predicates and the goal conditions.

It is evident that the algorithms solving the problem Conjunctive Boolean Query have the same estimates.

Level description of classes

To decrease the obtained step number estimates a level description of goal formulas was offered in Kosovskaya, [2008]. The procedure of a level description construction uses the following definition.

Definition. *Two elementary conjunctions of atomic predicate formulas A and B are called isomorphic if there exists such an elementary conjunction C and such substitutions $\lambda_{A,C}$ and $\lambda_{B,C}$ of objective variables from C instead of objective variables from A and B respectively that the results of these substitutions $A\lambda_{A,C}$ and $B\lambda_{B,C}$ coincide with the formula C up to the order of literals.*

These substitutions $\lambda_{A,C}$ and $\lambda_{B,C}$ are called the unifiers of formulas A and B respectively with the formula C.

Let the set of all possible objects ω is divided to some classes and these classes have descriptions consisting of conjunctions of atomic predicate formulas $A_1(\overline{x}_1)$, ..., $A_K(\overline{x}_K)$.

Find all sub-formulas $P_i^1(\overline{y}_i^1)$ with a "small complexity" such that sub-formulas isomorphic to them "frequently" appear in goal formulas $A_1(\overline{x}_1)$, ..., $A_K(\overline{x}_K)$ and denote them by atomic formulas with new first-level predicates p_i^1 and new first-level arguments z_i^1 for lists $\overline{y}_i^1$ of initial variables. Write down a system of equivalences

$$p_i^1(z_i^1) \Leftrightarrow P_i^1(\overline{y}_i^1), \quad i = 1, \ldots, n_1.$$

Let $A_k^1(\overline{x}_k^1)$ be a formula received from $A_k(\overline{x}_k)$ by substitution of $p_i^1(z_{i,j}^1)$ instead of the jth appearance of the formula isomorphic to $P_i^1(\overline{y}_i^1)$. Here $\overline{x}_k^1$ is a list of all variables in $A_k(\overline{x}_k^1)$ including both some (may be all) initial variables of $A_k(\overline{x}_k)$ and first-level variables appeared in the formula $A_k^1(\overline{x}_k^1)$.

A set of all atomic formulas of the type $p_i^1(\omega_i^1)$ where ω_i^1 denotes some ordered list $\overline{\tau}_i^1$ of elements from ω for which the formula $P_i^1(\overline{\tau}_i^1)$ is valid is called a first-level object description and denoted by $S^1(\omega)$. Such a way extracted subsets $\overline{\tau}_i^1$ are called first-level objects.

Repeat the above described procedure with formulas $A_k^1(\overline{x}_k^1)$. After L repetitions L-level goal conditions in the following form will be received.

$$\left\{\begin{array}{ccc} & A_k^L(\overline{x}_k^L) & \\ p_1^1(z_1^1) & \Leftrightarrow & P_1^1(\overline{y}_1^1) \\ & \vdots & \\ p_{n_1}^1(z_{n_1}^1) & \Leftrightarrow & P_{n_1}^1(\overline{y}_{n_1}^1) \\ & \vdots & \\ p_i^l(z_i^l) & \Leftrightarrow & P_i^l(\overline{y}_i^l) \\ & \vdots & \\ p_{n_L}^L(z_{n_L}^L) & \Leftrightarrow & P_{n_L}^L(\overline{y}_{n_L}^L) \end{array}\right. .$$

Procedure of a level description use for the identification problem consists in the following Kosovskaya, [2014].

1. For every i check $S(\omega) \Rightarrow \exists \overline{y}_{i\neq}^1 P_1^1(\overline{y}_i^1)$ and find all unifiers of $P_1^1(\overline{y}_i^1)$ with true first-level predicates. Add these first-level true atomic formulas to the object description and form $S^1(\omega)$. $l := 1$.

2. If an l-level ($l = 1, \ldots, L-1$) object description $S^l(\omega)$ is formed then for every i check $S^l(\omega) \Rightarrow \exists \overline{y}_{i\neq}^l P_1^l(\overline{y}_i^l)$ and find all values for true $(l+1)$-level predicate arguments.

3. Add these $(l+1)$-level true atomic formulas to the object description $S^l(\omega)$ and receive $S^{l+1}(\omega)$.

4. Substitute $p_i^l(y_{i,j}^l)$ instead of the jth appearance of $P_i^l(\overline{y}_i^l)$ into $A_k^l(\overline{y}_k^l)$.

5. Repeat the previous steps for $l = 1, \ldots, L$.

6. Check $S^L(\omega) \Rightarrow \exists \overline{y}_{k\neq}^L A_k^L(\overline{y}_k^L)$.

Such L-level goal conditions may be used for efficiency of an algorithm solving a problem formalized in the form of logical sequent. To decrease the number of steps of an exhaustive algorithm (for every t greater than some t_0) with the use of 2-level goal description it is sufficient

$$n_1 \cdot t^r + t^{s_1+n_1} < t^m,$$

where r is a maximal number of arguments in the formulas $P_i^1(\overline{y}_i^1)$, n_1 is the number of first-level predicates, s_1 is the number of atomic formulas in the first-level description, m is the number of variables in the initial goal condition.

Similar condition for decreasing the number of steps of a logical algorithm solving the problem is

$$\sum_{k=1}^{K} s^{1\,a_k^1} + \sum_{j=1}^{n_1} s^{\rho_j^1} < \sum_{k=1}^{K} s^{a_k},$$

where a_k and a_k^1 are maximal numbers of atomic formulas in $A_k(\overline{x}_k)$ and $A_k^1(\overline{x}_k^1)$ respectively, s and s^1 are numbers of atomic formulas in $S(\omega)$ and $S^1(\omega)$ respectively, ρ_j^1 is the number of atomic formulas in $P_i^1(\overline{y}_i^1)$.

Extraction of such sub-formulas $P_i^1(\overline{y}_i^1)$ with a "small complexity" which "frequently" appear in goal formulas $A_1(\overline{x}_1)$, ..., $A_K(\overline{x}_K)$ is described in Kosovskaya, [2014]. The procedure of the extraction of the maximal sub-formula which is isomorphic to some sub-formulas of two elementary conjunctions is described in Petrov, [2016].

Level description of classes construction.

1) For every $i = 1, \ldots, n-1$, $j = i+1, \ldots, n$ extract maximal sub-formulas $Q_{i,j}$ which are isomorphic to some sub-formulas of elementary conjunctions $A_i(\overline{x}_i)$ and $A_j(\overline{x}_j)$ ($i \neq j$) and find common unifiers $\lambda_{i,ij}$ and $\lambda_{j,ij}$ of these conjunctions with $Q_{i,j}$.

l) The procedure of maximal sub-formulas extraction and obtaining of common unifiers is repeated with the formulas received earlier.

The process will end for some $l = L$, because the lengths of the extracted formulas decrease from step to step. Sub-formulas consisting of one literal are not under consideration. That's why the item l is repeated for $l = 2, \ldots, L$.

$L+1$) The extracted sub-formulas that haven't another common sub-formulas are denoted by $P_i^1(\overline{y}_i^1)$ $(i = 1,\ldots,n_1)$.

$L+2$) For every $l = 1,\ldots,L-1$ sub-formulas having common sub-formulas $P_i^l(\overline{y}_i^l)$ are denoted by $P_i^{l+1}(\overline{y}_i^{l+1})$ $(i = 1,\ldots,n_{l+1})$. At the same time instead of the jth occurrence of $P_i^l(\overline{y}_i^l)$ substitute $p_i^l(y_{i,j}^l)$ where $y_{i,j}^l$ is an l-level variable for a list of variables $\overline{y}_{i,j}^l$ taking into account the corresponding unifier.

An approach to the construction of a level data base for solving the Conjunctive Boolean Query problem

While creation a data base we are not sure what queries may make a user. But the data base itself remains practically fixed. That's why patterns (formulas isomorphic to some conjunctions of atomic formulas from the data base) must be searched in the data base itself (in the set of constant atomic formulas $S(D)$).

In order to receive complexity estimates regard the case when the query has the form of one elementary conjunction, as well as it was done in the problem of logic-objective recognition.

Remind, that the estimates of the number of steps needed for checking the logical consequence of the form $C(\overline{x}) \Rightarrow \exists y_1, y_2, \ldots, y_l A(y_1, y_2, \ldots, y_l, d_1, \ldots, d_r)$, where $C(\overline{x})$ is a set of atomic formulas or their conjunction, are the following:

— $O(t^l)$, where t is the number of elements in $\overline{x}$, l is the number of variables in $A(\overline{y})$, while using an exhaustive algorithm;

— $O(\sum_{i:p_i\ is\ in\ A(\overline{y})} s_i^{a_i})$, where s_i and a_i are the numbers of occurrences of the predicate p_i in the $C(\overline{x})$ and in the formula $A(\overline{y})$ respectively, while using a logic algorithm.

Two-level data base construction

1. Let for $i = 1,\ldots,n_1$ groups of mutually disjoint sub-sets of $S(D)$ such that all conjunctions of elements of a sub-set from the ith group are isomorphic to each other and to some formula $P_i^1(\overline{y}_i^1)$ are extracted. For every group unifiers of the conjunctions of a sub-set elements with $P_i^1(\overline{y}_i^1)$ are found.

2. Introduce first-level predicates p_i^1 $(i = 1,\ldots,n_1)$ defined by the equivalence $p_i^1(y_i^1) \Leftrightarrow P_i^1(\overline{y}_i^1)$, where y_i^1 are the first-level variables for the lists of initial variables $\overline{y}_i^1$.

3. Supplement the set $S(D)$ by the set of atomic first-level formulas in the form $p_i^1(d_i^{1,j})$, where $d_i^{1,j}$ is the notation of a list of constants from D which is included into unifier of some conjunction of a sub-set with $P_i^1(\overline{y}_i^1)$.

A two-level data base $S^1(D)$ is constructed.

Note that the construction of a two-level data base is an NP-hard problem with huge input data, because in the item 1 it is solved an NP-hard problem of extraction of groups of mutually disjoint sub-sets with small capacity of $S(D)$ such that all conjunctions of elements of a sub-set from the a group are isomorphic to each other from all formulas of the data base.

But this NP-hard problem with huge input data is solved only once.

Two-level data base implementation

Let we have a conjunctive Boolean query $\exists y_1, y_2, \ldots, y_l A(y_1, y_2, \ldots, y_l, d_1, \ldots, d_r)$, where $d_1,\ldots,d_r$ are constants from D.

1. Checking the logical sequent

$$A(y_1, y_2, \ldots, y_l, d_1, \ldots, d_r) \Rightarrow \exists \overline{y}_i^1 P_i^1(\overline{y}_i^1)$$

for $i = 1,\ldots,n_1$ allows (if it is fulfilled) to find all sub-formulas of the query, which are isomorphic to $P_i^1(\overline{y}_i^1)$, and their unifiers with the corresponding lists of constants.

In spite of the fact that the problem in this item is NP-hard, its input data have not large length and the estimates of number of steps have not large parameters of the formula $P_i^1(\overline{y}_i^1)$ in the exponent (the number of variables or the number of atomic formulas with the same predicate). These parameters are less then the corresponding parameters of the formula $A(y_1, y_2, \ldots, y_l, d_1, \ldots, d_r)$.

These estimates have the form

— $O(\sum_{i=1}^{n_1}(l+r)^{\|\overline{y}_i^1\|})$, $\|\overline{y}_i^1\|$ be the number of arguments in $P_i^1(\overline{y}_i^1)$, for an exhaustive algorithm;

— $O(\sum_{i=1}^{n_1}\sum_{j:p_j\ is\ in\ P_i^1(\overline{y}_i^1)} a_j^{\alpha_j}))$, a_j and α_j be the number of atomic formulas with the predicate p_j in $A(y_1, y_2, \ldots, y_l, d_1, \ldots, d_r)$ and $P_i^1(\overline{y}_i^1)$ respectively, for a logical algorithm.

2. Substitute into $A(y_1, y_2, \ldots, y_l, d_1, \ldots, d_r)$ instead of every sub-formula,which is isomorphic to $P_i^1(\overline{y}_i^1)$, atomic formula $p_i^1(y_i^{1,j})$, where $y_i^{1,j}$ is a variable for a list of variables and constants of $\overline{y}_i^1$ (index j changes from 1 to the number of occurrences sub-formulas isomorphic to $P_i^1(\overline{y}_i^1)$). It is possible because we have unifiers of every such sub-formulas with $P_i^1(\overline{y}_i^1)$). An elementary conjunction $A^1(\overline{x}^1)$ is received. Here $\overline{x}^1$ is a list of the initial variables, 1-level variables and constants that remains explicitly in the formula as arguments.

This item is fulfilled in linear under the notation length of $A(y_1, y_2, \ldots, y_l, d_1, \ldots, d_r)$ number of steps. Note, that the notation length of $A^1(\overline{x}^1)$ is not greater than the notation length of $A(y_1, y_2, \ldots, y_l, d_1, \ldots, d_r)$ and is strictly less if at least one sub-formula has been changed by an atomic one of the first level.

3. While checking $S^1(D) \Rightarrow \exists\overline{x}^1 A^1(\overline{x}^1)$ first of all check atomic formulas of $A^1(\overline{x}^1)$ with first-level predicates and find possible values for first-level variables. After that check atomic formulas of $A^1(\overline{x}^1)$ with initial predicates taking into account the values of initial variables that have occurrences into lists defining first-level variables.

The number steps estimates for checking atomic formulas with first-level predicates and finding values for first-level variables are

— $O(t_1^{l_1})$ (t_1 be the number of first-level constants in $S^1(D)$, l_1 be the number of first-level variables in $A^1(\overline{x}^1)$) for an exhaustive algorithm;

— $O(\sum_{i:p_i^1\ is\ in\ A^1(\overline{x}^1)} {s_i^1}^{a_i^1})$ (s_i^1 and a_i^1 be the numbers of occurrences of the predicate p_i^1 in the description $S^1(\omega)$ and in the formula $A^1(\overline{x}^1)$ respectively) for a logical algorithm.

Note that if at least one sub-formula has been changed by an atomic first-level one then the number of initial variables in $\overline{x}^1$ and the number of atomic formulas with initial predicates in $A^1(\overline{x}^1)$ decreases in comparison with the Boolean query $A(y_1, y_2, \ldots, y_l, d_1, \ldots, d_r)$.

The number steps estimates for checking checking atomic formulas with initial predicates and finding values for initial variables are

— $O(t^{l-l_1})$ (t_1 be the number of first-level constants in $S^1(D)$, l_1 be the number of first-level variables in $A^1(\overline{x}^1)$) for an exhaustive algorithm;

— $O(\sum_{i:p_i\ is\ in\ A^1(\overline{x}^1)} {s_i}^{a_i-a_i^1})$ (s_i^1 and a_i^1 be the numbers of occurrences of the predicate p_i^1 in the description $S^1(\omega)$ and in the formula $A^1(\overline{x}^1)$ respectively) for a logical algorithm.

If we sum the estimates for items 1 – 3 we obtain complete estimates

— $O(\sum_{i=1}^{n_1}(l+r)^{\|\overline{y}_i^1\|}) + O(t_1^{l_1}) + O(t^{l-l_1}) = O(t_1^{l_1} + t^{l-l_1})$ for an exhaustive algorithm;

— $O(\sum_{i=1}^{n_1}\sum_{j:p_j\ is\ in\ P_i^1(\overline{y}_i^1)} a_j^{\alpha_j}) + O(\sum_{i:p_i^1\ is\ in\ A^1(\overline{x}^1)} {s_i^1}^{a_i^1}) + O(\sum_{i:p_i\ is\ in\ A^1(\overline{x}^1)} {s_i}^{a_i-a_i^1}) = O(\sum_{i:p_i^1\ is\ in\ A^1(\overline{x}^1)} {s_i^1}^{a_i^1}) + O(\sum_{i:p_i\ is\ in\ A^1(\overline{x}^1)} {s_i}^{a_i-a_i^1})$ for a logical algorithm.

It is obvious that if at least one sub-formula has been changed by an atomic first-level one then the number steps estimates for checking the Boolean query sequent from the two-level data base is less than its sequent from the initial data base estimates.

Conclusion

Essential differences of the well-known NP-complete problem Conjunctive Boolean Query while its implementation as a problem of pattern recognition in the frameworks of logic-objective approach and as a problem of the data base use are shown.

In the both implementations it is possible to construct a level description of a fixed input data which increases the time complexity of multiple implementation. The estimates of number of steps while using a two-level data base are proved in the paper.

Algorithms of level description of classes are yet developed in the previous papers of the author, but there is only an approach to developing of algorithms of level data base construction.

Acknowledgements

The paper is published with partial support by the ITHEA ISS (www.ithea.org) and the ADUIS (www.aduis.com.ua)

Bibliography

Garey M.R., Johnson D.S., "Computers and Intractability: A Guide to the Theory of NP-Completeness", Freeman, New York, 1979.

Kosovskaya T. Discrete Artificial Intelligence Problems and Number of Steps of their Solution // International Journal on Information Theories and Applications, Vol. 18, Number 1, 2011. P. 93 ï£¡ 99.

Kosovskaya T. Construction of Class Level Description for Efficient Recognition of a Complex Object // International Journal "Information Content and Processing", Vol. 1,No 1. 2014. P. 92 – 99.

Kosovskaya T.M. Level system of formulas for decreasing the number of proof steps of formulas simulating some Artificial Intelligence problems // CLMPS. 15-th Congress of Logic, Methodology and Philosopfy of Science. Book of abstracts. 3 - 8 August 2015, University of Helsinki. P. 283.

Petrov D.A. Algorithms of extraction of a maximal common up to the names of variables predicate formulas and their implementation // Proc. of the 9-th Conference "Information Technology in Management". St.Petersburg, 2016. P. 97 –102. (In Russian)

Authors' Information

Tatiana Kosovskaya *- Dr., Professor of St.Petersburg State University, University av., 28, Stary Petergof, St.Petersburg, 198504, Russia; e-mail: kosovtm@gmail.com*
Major Fields of Scientific Research: Logical approach to artificial intelligence problems, Theory of Computational Complexity of Algorithms

Examples of NP-complete essential restrictions of the SUBSET SUM PROBLEM

Nikolay K. Kosovskii, Tatiana Kosovskaya, Michail Starchak

Abstract: *The problem SUBSET SUM Garey, Johnson, [1979] may be interpreted as the problem of solvability in $\{0,1\}$ numbers checking of a linear Diophantine equation with positive coefficients and constant term. This problem is the one of the most simply formulated mathematical problems for which it is proved that it is NP-complete. Its essential restriction under which it remains to be NP-complete are offered in the paper.*

Keywords: *NP-completeness, SUBSET SUM, linear Diophantine equation*

ACM Classification Keywords*: F.1.3 Complexity Measures and Classes, Reducibility and completeness; F.2.1 Analysis of Algorithms and Problem Complexity, Numerical Algorithms and Problems, Number-theoretic computations*

Introduction

Algorithms for the solving of a system of linear equations have broad implementation during computer simulation Krivyi [2016]. The use of a computer while solving mathematical and discrete problems involves to take into account the effectiveness of an algorithm to be programmed. At present an algorithm without proved polynomial-time complexity is regarded as not effective.

The notion of an NP-complete problem was introduced by Cook S.A. in 1971 Cook [1971]. At present for such a problem a polynomial-time algorithm is not found and there is a hypothesis that it does not exist. That's why the proof of NP-completeness for different mathematical problems are very actual (see fo example Kosovskii, Starchak, [2016]).

To prove NP-completeness of a problem the notion of polynomial reducibility is often used.

A very representative list of NP-complete problems is in Garey, Johnson, [1979]. The problem SUBSET SUM may be pointed as one of the most simply formulated mathematical problems.

SUBSET SUM Garey, Johnson, [1979]

Instance: Finite set A, a size $s(a) \in \mathbb{Z}^+$ for each $a \in A$ and a positive number B.

Question: Is there exists a subset $A' \subseteq A$ such that the sum of sizes of the elements in A is exactly B?

This problem may be formulated as the problem of solvability in $\{0,1\}$-numbers checking of a linear Diophantine equation with positive coefficients and constant term.

SUBSET SUM

Instance: A set of positive integers $\{s_1,\ldots,s_n\}$, a set of variables $\{x_1,\ldots,x_n\}$ and a positive integer B.

Question: Is there exists a $\{0,1\}$-solution of the equation $s_1x_1+\cdots+s_nx_n=B$?

The analysis of an NP-complete problem allows to extract such its subproblems that some of them remain to be NP-complete and the other turn into polynomial-time ones. The extracting of such subproblems is an important step to the defining the domain of input data which allows effective computer implementation.

Essential restriction of the problem SUBSET SUM, under which it remains to be NP-complete are offered in the paper. The essence of the restriction consists in the fact that the constant term of the equation is written in positional number system with the help of the only one non-zero figure.

To prove the result presented in the paper NP-completeness of the problem ONE-IN-THREE 3SAT Garey, Johnson, [1979] is used.

ONE-IN-THREE 3SAT

Instance: Set U of variables, collection C of clauses over U such that each clause $c \in C$ has $|c| = 3$.

Question: Is there a truth assignment for U such that each clause in C has exactly one true literal?

Its restriction when each clause does not contain the symbol of negation is used in the paper. Note, that NP-completeness of such a restriction is announced in Garey, Johnson, [1979] but the reference to the proof is absent.

Main results

Let's prove the polynomial-time reduction of the problem ONE-IN-THREE 3SAT when each clause does not contain the symbol of negation to the main problem ONE-IN-THREE 3SAT.

Lemma 1. *The problem ONE-IN-THREE 3SAT is polynomially reducible to its subproblem ONE-IN-THREE 3SAT when each clause does not contain the symbol of negation.*

Proof. First of all show that $z = 1$ is equivalent to the consistency in $\{0, 1\}$-numbers of the system of equations

$$\begin{cases} y_1 + y_2 + z &= 1 \\ y_2 + y_4 + z &= 1 \\ y_3 + y_4 + z &= 1 \\ y_1 + y_3 + z &= 1 \end{cases}.$$

If $z = 1$ then the consistency is evident and $y_1 = y_2 = y_3 = y_4 = 0$.

If $z = 0$ then we have the system

$$\begin{cases} y_1 + y_2 &= 1 \\ y_2 + y_4 &= 1 \\ y_3 + y_4 &= 1 \\ y_1 + y_3 &= 1 \end{cases},$$

which is not consistent because the rank of its matrix is 3 and the rank of its augmented matrix is 4.

If the constant *true* is interpreted as the number 1 and the constant *false* is interpreted as the number 0, then the set of clauses $\{y_1 \vee y_2 \vee z,\ y_2 \vee y_4 \vee z,\ y_3 \vee y_4 \vee z,\ y_1 \vee y_3 \vee z\}$ in the problem ONE-IN-THREE 3SAT gives the condition $z = true, y_1 = y_2 = y_3 = y_4 = false$.

This statement allow to introduce in the problem ONE-IN-THREE 3SAT 4 new variables. One of which (the variable z) is identical *true* and the others y_1, y_2, y_3, y_4 are identical *false*. For every variable x in the problem ONE-IN-THREE 3SAT we can introduce a new variable $\overline{x}$ and the clause $x \vee \overline{x} \vee y_1$. The received in such a manner set of clauses does not contain the symbol of negation.

It is evident that the new set of variables and the new set of clauses may be obtained by a polynomial-time under the length of ONE-IN-THREE 3SAT input data algorithm. □

The proof of the following lemma uses NP-completeness of the following problem Schrijver [1986].

Lemma 2. *The problem of consistency in non-negative integers of the system of linear Diophantine equations of the form $a_1x_1 + ... + a_nx_n = 1$, with coefficients $a_i \in \{0, 1\}$ and exactly three non-zero coefficients in every equation, is NP-complete.*

Proof. The problem is from the class **NP** as the consequence of the NP-completeness of the general problem of consistency in non-negative integers of the system of linear Diophantine equations of the form $a_1x_1 + ... + a_nx_n = 1$ (corollary 18.1a in Schrijver [1986]).

The problem ONE-IN-THREE 3-SAT is polynomial-time reducible to the problem under consideration. The constants *true* and *false* are encoded with 1 and 0 respectively. As ONE-IN-THREE 3-SAT remains NP-complete even with no negated literals in disjunctions the representation of every disjunction of the form $x \vee y \vee z$ by $x + y + z = 1$ completes this polynomial reduction. □

Let SUBSET SUM-1f be the restriction of the problem SUBSET SUM when the constant term of the equation is written in positional number system with the help of the only one non-zero figure.

SUBSET SUM-1f

Instance: A set of positive integers $\{s_1, \ldots, s_n\}$ and a positive integer B written in a positional number system with the help of the only one non-zero figure.

Question: Is there exists a $\{0, 1\}$ solution of the equation $s_1x_1 + \cdots + s_nx_n = B$?

Theorem 1. *The problem SUBSET SUM-1f is NP-complete.*

Proof. SUBSET SUM-1f belongs to the class **NP** as it is a subproblem of the NP-complete problem SUBSET SUM.

Let $u_1, \ldots, u_n, c_1, \ldots, c_m$ be input data of the problem ONE-IN-THREE 3SAT when each clause does not contain the symbol of negation.

If the constant *true* is interpreted as the number 1 and the constant *false* is interpreted as the number 0, then the truth with exactly one true literal of each clause c in the form $w_1 \vee w_2 \vee w_3$ (where w_1, w_2, w_3 are variables) in C may be interpreted as $w'_1 + w'_2 + w'_3 = 1$ (where w'_1, w'_2, w'_3 are variables with values from $\{0, 1\}$.

The jth equation multiply by 2^{j-1} ($j = 1, \ldots, m$) and sum the results. The equation in the form

$$c'_1 + 2c'_2 + \cdots + 2^{m-1}c'_m = 1 + 2 + \cdots + 2^{m-1}$$

is received. It is solvable in $\{0, 1\}$-numbers if and only if the problem ONE-IN-THREE 3SAT when each clause does not contain the symbol of negation is solvable.

This is a is polynomial-time reduction because the lengths of the binary notation of the numbers 2^{j-1} ($j = 1, \ldots, m$) are not greater then m, where m is the number of clauses in the input data.

The figure 1 may be changed by any non-zero figure f in the positional number system with the radix of number system p. In such a case the cofactor 2^{j-1} must be changed by $p^{j-1}f$ ($j = 1, \ldots, m$). □

Conclusion

NP-completeness of the restriction of the problem SUBSET SUM when the constant term of the equation is written in positional number system with the help of the only one non-zero figure is proved in the paper.

Acknowledgements

The paper is published with partial support by the ITHEA ISS (www.ithea.org) and the ADUIS (www.aduis.com.ua)

Bibliography

Cook S.A., "The complexity of theorem-proving procedure", Proc. 3rd Ann. ACM Symp. on Theory of Computing, Association for Computing Machinery, New York, pp. 151 – 158.

Garey M.R., Johnson D.S., "Computers and Intractability: A Guide to the Theory of NP-Completeness", Freeman, New York, 1979.

N. K. Kosovskii, T. M. Kosovskaya, and N. N. Kosovskii, "NP completeness conditions for verifying the consistency of several kinds of systems of linear diophantine discongruences," Vestn. St. Petersburg Univ.: Math. 49, 18 – 22 (2016). ©Allerton Press, Inc., 2016.

Kosovskii N.K., Starchak M.R., "NP-complete problems for greatest common divisor of values of linear polynomials", Proceedings of the 9th conference ITU-2016, St. Petersburg, 2016, pp. 71-72. (in Russian)

Kosovskii N.K., Kosovskaya T.M., Kosovskii N.N., Starchak M.R., "NP-complete problems for systems of divisibilities of values of linear polynomials", Vestn. St. Petersburg Univ.: Math., 2017, to be published. (in Russian)

Krivyi S.L. "Linear constrains and their solving methods", International Journal "Information theories and applications, 2016, Vol. 23, Number 2, pp. 103 – 200. (in Russian)

Schrijver A. "Theory of Linear and Integer Programming" // John Wiley and Sons, New York, 1986.

Schaefer T.J. "The complexity of satisfiability problems" // Proceedings 10th Symposium on Theory of Computing, ACM Press, 216-226 (1978).

Authors' Information

Nikolay K. Kosovskii *- Dr., Professor of Computer Science Chair of St. Petersburg State University, University av., 28, Stary Petergof, St. Petersburg, 198504, Russia; e-mail: kosov@NK1022.spb.edu*

Major Fields of Scientific Research: Mathematical Logic, Theory of Computational Complexity of Algorithms

Taniana Kosovskaya *- Dr., Professor of Computer Science Chair of St. Petersburg State University, University av., 28, Stary Petergof, St. Petersburg, 198504, Russia; e-mail: kosovtm@gmail.com*

Major Fields of Scientific Research:Logical Approach to Artificial Intelligence Problems, Theory of Computational Complexity of Algorithms

Mikhail Starchak *- PhD student of Computer Science Chair of St. Petersburg State University, University av., 28, Stary Petergof, St. Petersburg, 198504, Russia; e-mail: mikhstark@gmail.com*

Major Fields of Scientific Research: Theory of Computational Complexity of Algorithms

COMPUTER-BASED BUSINESS GAMES' RESULT ANALYSIS

O. Vikenteva, A. Deriabin, N. Krasilich, L. Shestakova

***Abstract**: Given research considers the Business Intelligence analysis of computer based business games. A tool environment, called Competence-based Business Game Studio (CBGS), is applied for business games' design and development. An approach is proposed that allows designing and conducting business games based on enterprises business processes. Consequently, CBGS may be considered as a universal product with respect to domain. Competence-based Business Game Studio consists of several subsystems. The Analysis Subsystem makes possible to exclude human factor from the process of player skills and knowledge assessment, the latter are scored employing an automated approach based on formal parameters. This paper defines the source data for Analysis Subsystem as well. Data warehouse containing multidimensional data marts was designed for the evaluation of player's competency. Two info-cubes were developed: the first info-cube is proposed to assess players' actions, the second one - to identify bottlenecks within business processes using efficiency assessment of Decision Making Points. In order to collect information about players Complex Analysis methods are proposed for implementation: such as aggregation, navigation and filtering. To evaluate business game quality three types of Decision Making Points should be distinguished. Decision Making Points completed by players are allocated to the aforementioned types using cluster analysis (PAM-algorithm) and supervised classification.*

***Keywords**: business intelligence methods, data warehouse, competencies, active learning methods, business-game.*

***ACM Classification Keywords:** K.3 Computers and Education: K.3.2 Computer and Information Science Education – Information systems education. I. Computing Methodologies: I.2 Artificial Intelligence: I.2.1 Applications and Expert Systems – Games.*

Introduction

The implementation of game mechanics implies an increase of a player's involvement into the learning process by simulation of real-life conditions. Moreover, player's actions are evaluated in accordance with the set of competencies and criteria. There are a lot of researches in the business games area. For instance, one of the most popular and complex business games products are SimulTrain, Innov8, BrandPro. However, most of such systems focus on a certain domain.

The proposed approach to the creation of competence-based educational environment consists of the development of design, technical, organizational and methodological tools for implementing one of the active methods of forming competencies that is named competence-based business games [Vikentyeva, 2013]. The approach is multi-model and is based on the development of domain-specific models applied in design and execution stages [Vikentyeva, 2015].

Competence-based educational system (CBGS – Competence-based Business Games Studio) should consist of several subsystems. The CBGS structure is presented in Figure 1. [Vikentyeva, 2013].

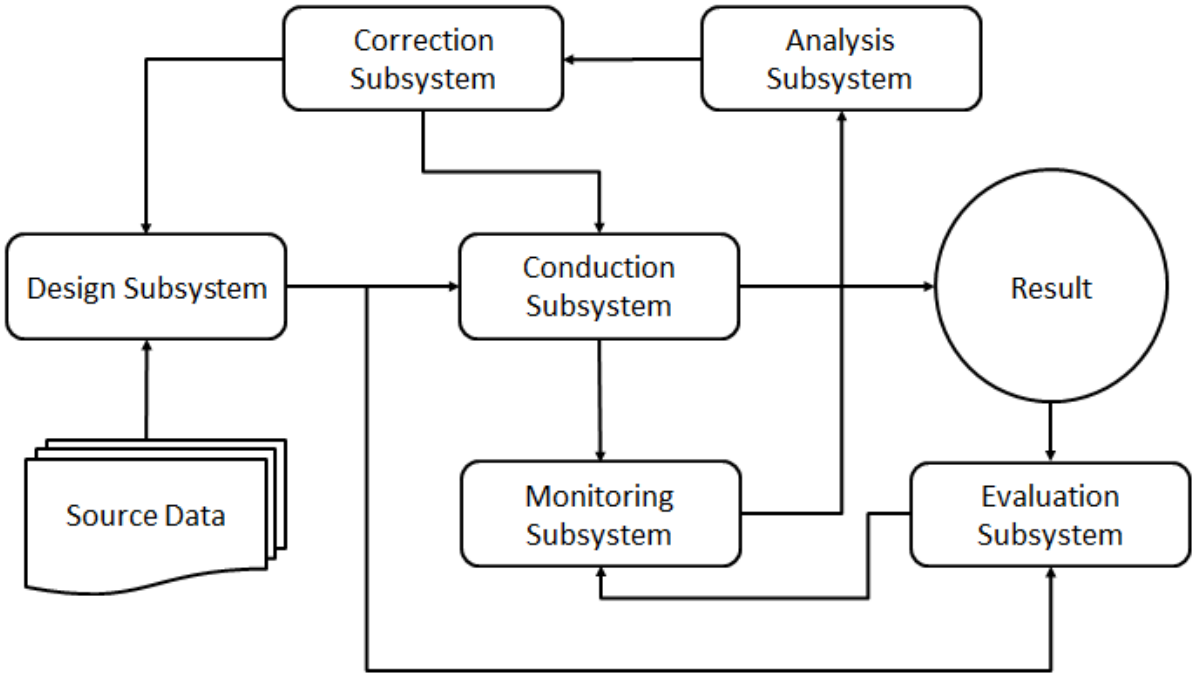

Figure 1. Structure of CBGS

Nowadays prototypes of following subsystems are developed:

- Design Subsystem. Business Processes Models are building within the Design Subsystem. These models are transformed from weakly formalized format based on real business processes models into formalized form with the use of graphical models editor.
- Conduction Subsystem. Source data for the subsystem are game plan and information about resources used during the game. The mechanism testing users is named Decision Making Point (DMP). DMP determines the course of game when a user has chosen resources.
- Evaluation Subsystem. It allows evaluating player's actions based on tests.

– Monitoring Subsystem. The subsystem implements two modules for working with databases: one is design to work with the database of operational data obtained during the game, the second works with a database of the results of players' testing.

This research issues related to business games results analysis using Business Intelligence methods are considered.

The process of human resources knowledge evaluation is subjective since it implies the influence of human factor. The CBGS's Analysis Subsystem allows excluding human factor due to automated approach to assess the trainee competency based on formal parameters. Nowadays there is a lot of research in the field of Educational Data Mining (EDM). EDM aims to apply Data Mining methods to extract information related to the learning process [Hung, 2012], [Jeong, 2013], [Sahedani, 2013].

The Analysis Subsystem should assess player's competences (knowledge, skills, experiences) based on his choice of resources within DMPs. DMPs allow the player to choose the sequence of operations of a business process. Data of passed games have to be compared with the reference model developed within the Design Subsystem.

It is important to take into account that the reason of a trainee inability to complete the game with 100% success might be the Game bottlenecks. Some algorithms may be not trivial even for experts of a corresponding business process as model of unified educational business process (UEBP) including DMP is automatically generated. Business game scenario is being built based on UEBP.

Analysis Subsystem should perform two major analysis procedures [Vikentyeva, 2016]:

– Player's actions analysis that allows providing player's characterization based on all business games, which the player participated.
– Game analysis to its correction in the case of bottlenecks identification. Such analysis has to be conducted for all DMPs.

Data Sources for Analysis Subsystem

The reference model of business process is created within the Design Subsystem. Business Process Design database stores the correct sequence of operations for each business process as well as a set a set of resources for every operation. Data for tables «Business Process», «Operation», «Resources» have to be loaded from this database [Vikentyeva et al., 2015].

Competence is a set of knowledge, skills, experience and personal characteristics, that are needed for successful performance of tasks [Kozodaev, 2015].

The concept of competence is considered in the learning process. It is important to understand that personal characteristics and experience of players are not considered within the project, because it is

extremely difficult to evaluate experience level in a short time. Therefore, competence will be defined as a set of knowledge, skills necessary for successful passage of a business game.

Process of competences planning implies creation of matrix defining the dependence between operations of business processes and competences [Vikentyeva et al., 2013]. Within different business processes the same operations can be characterized by different competences.

It is possible to identify the relationship between operations and set of competences. In order to determine to what extend is competence formed and what knowledge and skills a player has, the resources that the player chooses to perform operations should also be included in the multidimensional array of competencies, since they are the ones that determine whether the player possesses the necessary set of knowledge and skills to perform the operation (the player knows which resources to choose and can apply them).

This structure can easily be formed in a multidimensional data warehouse, developed within the framework of the analysis subsystem [Vikentyeva, 2016]. The schema of database storing the results of passing games is also considered in the research [Vikentyeva, 2016].

Based on the data that can be extracted from the Design Subsystem and the Conduction Subsystem, it can be determined that the evaluation of the players' actions should be carried out according to three criteria:

- Correspondence of the sequence of operations performed by the learner during the game to the reference model.
- Competence of the player (within a single game).
- Satisfactory time of passing the game.

Data Warehouse Info-objects

Data warehouse info-objects are divided into two types [Kolb, 2012]:

- A characteristic is a sequence of values of one of analyzed parameters. Characteristics may include master data, texts and hierarchies;
- A key figure is a data quantitatively characterizing the set of characteristics.

Table 1 presents characteristics that are created within the designing data warehouse.

Table 1. Characteristics Developed in the Data Warehouse

Characteristic Name	Type	Amount of Symbols
Time of a Game	Time	–
Game Number for the Player	Integer	–
Player	String	5
Business Process	String	255
Operation	String	255
Resource Type	String	255
Resource	String	255
Competence	String	255
Competence Type (Knowledge/Skill)	String	6
Knowledge/Skill Name	String	255
Operation Number in the Reference Model	Integer	–
Actual Operation Number	Integer	–

Table 2 presents key figures that are created within the designing data warehouse.

Table 2. Key Figures Developed in the Data Warehouse

Key Figure Name	Type	Unit of Measurement
The Deviation in Operations Sequence	Integer	–
Operation Performance Indicator	Integer (0 or 1)	–
Resource Selection Indicator	Integer (0 or 1)	–
Formed Percentage of Knowledge/Skill	Number	Percentage
Maximum Percentage of Knowledge/Skill	Number	Percentage

Data Warehouse Info-Providers

Within the developed data warehouse multidimensional data marts (info-cubes) are used.

In accordance with the functional requirements for the Analysis Subsystem it is necessary to design two info-cubes:

- Evaluation of Players' Actions.
- Search of Business Game Bottlenecks.

Table 3 represents the set of info-objects that are included into the info-cube designed for evaluation of players' actions [Vikentyeva, 2016].

Table 3. Structure of Info-cube Designed for Evaluation of Players' Actions

Dimension	Characteristics
Time	Time of a Game
Game	Player
	Game Number for the Player
	Business Process
	Operation
	Resource
Competence	Competence
	Competence Type (Knowledge/Skill)
	Knowledge/Skill Name
Key Figures	
	The Deviation in Operations Sequence
	Operation Performance Indicator
	Formed Percentage of Knowledge/Skill

With the use of this set of data, the following reports can be obtained:

- The percentage of each competence formation for the player. The report will show aggregated data on competences.
- List of knowledge and skills that a player possesses or does not possess.
- Correspondence of actual operation sequence of a game to the reference model.

Table 4 represents the set of info-objects that are included into the info-cube designed for searching business game bottlenecks [Vikentyeva, 2016].

Table 4. Structure of Info-cube Designed for Searching Business Game Bottlenecks

Dimension	Characteristic
Game	Player
	Business Process
	Game Number for the Player
Decision Making Point	Operation
	Resource Type
	Resource
Key Figures	
	Resource Selection Indicator
	Maximum Percentage of Knowledge/Skill

By applying clustering to the data bottlenecks in decision making point (DMP) can be detected.

The Process of Loading Data into Data Warehouse Info-providers

Into the info-cubes data is loaded from the following databases:

- Database for business processes' modeling.
- Database for competence planning.
- Database of actual results of game.

The algorithms for loading data into the info-cube designed for evaluation of players' actions are presented in Table 5.

Table 5. The Algorithms for Loading Data into the Info-cube Designed for Evaluation of Players' Actions

Dimension	Characteristics	Algorithm of Data Loading	Source Database
Time	Time of a Game	Formula: End Time-Start Time of a Game	Database of actual results of game
Game	Player	Direct assignment	Database of actual results of game
	Game Number for the Player	Count distinct Business Process ID with the actual Business Process ID, Player ID and Start Time less or equal the Game Start Time	Database of actual results of game
	Business Process	Direct assignment	Database for business processes' modeling
	Operation	Direct assignment	Database for business processes' modeling
	Resource	Direct assignment	Database for business processes' modeling
Competence	Competence	Direct assignment	Database for competence planning
	Competence Type (Knowledge/Skill)	Defined by table type	Database for competence planning
	Knowledge/Skill Name	Direct assignment	Database for competence planning
Key Figures			
	The Deviation in Operations Sequence	Formula: Operation Number within the Reference Model for the Game – Actual Operation Number	Database for business processes' modeling Database of actual results of game
	Operation	If the operation is present in	Database of actual results of

Dimension	Characteristics	Algorithm of Data Loading	Source Database
	Performance Indicator	the database of the actual results of games for a particular game and for a specific player, then 1, otherwise 0	game
	Formed Percentage of Knowledge/Skill	If the resource characterizing knowledge/skill is selected within the specified business process and operation, then the percentage of knowledge/skill within the competence is assigned, otherwise 0	Database for competence planning Database of actual results of game

Table 6. The Algorithms for Loading Data into the Info-cube Designed for Searching Business Game Bottlenecks

Dimension	Characteristics	Algorithm of Data Loading	Source Database
Game	Player	Direct assignment	Database of actual results of game
	Game Number for the Player	Count distinct Business Process ID with the actual Business Process ID, Player ID and Start Time less or equal the Game Start Time	Database of actual results of game
	Business Process	Direct assignment	Database for business processes' modeling
Decision Making Point	Operation	Direct assignment	Database for business processes' modeling
	Resource	Direct assignment	Database for business processes' modeling

	Resource Type	Direct assignment	Database for business processes' modeling
Key Figure			
	Resource Selection Indicator	Direct assignment (if resource was selected, then 1, otherwise 0)	Database of actual results of game
	Maximum Percentage of Knowledge/Skill	Direct assignment	Database for competence planning

Data Analysis Algorithms for the Info-cube Designed for Evaluation of Players' Actions

Complex Analysis method is applied for Evaluation of Players' Actions. The player's competence within a single business process may be defined by several ways:

- The total competence of player based on actual results of game. Aggregation on Business Process and calculation of average percentage of competence are necessary for this analysis. Other characteristics are not considered. Sample of data includes Game Number, Player, Business Process, Competence, Formed Percentage of Knowledge/Skill.
- The percentage of competence obtained by a player within a single game. Such a sample will determine the degree of competence obtained by the player within the operation. Sample of data includes Game Number, Player, Business Process, Operation, Competence, Formed Percentage of Knowledge/Skill.
- Possession of certain knowledge and skills. For this type of analysis, the data should be fully detailed. Sample of data includes Game Number, Player, Business Process, Operation, Resource, Competence, Competence Type (Knowledge or Skill), Knowledge/Skill Name, Formed Percentage of Knowledge/Skill (the key figure is restricted by condition «>0»).
- Unformed knowledge and skills of the player. For this type of analysis, all the data within a single game must be aggregated by Operations and Knowledge/Skills. Sample of data includes Game Number, Player, Business Process, Operation, Competence, Competence Type (Knowledge or Skill), Knowledge/Skill Name, Formed Percentage of Knowledge/Skill (the key figure is restricted by condition «==0»).

- In addition to the degree of the player's competence, the data set of the Info-cube also allows to determine the deviation of actual operations' sequence from the reference model. Sample of data includes Game Number, Player, Business Process, Operation, The Deviation in Operations Sequence (the key figure is restricted by condition «<>0»).
- In addition, it is possible to identify which operations from the reference model were not performed. Sample of data includes Game Number, Player, Business Process, Operation, Operation Performance Indicator (the key figure is restricted by condition «==0»).
- A player progress. This type of analysis is performed by comparison of all results of passing a particular game if the player participates in the game not for the first time.

Data Analysis Algorithms for the Info-cube Designed for Searching Business Game Bottlenecks

Models of real business processes performed at enterprises can't be used in the design of business games, therefore the concept of a model of a unified educational business process (UEBP) is introduced [Vikentyeva, 2015]. UEBP reflects the essential invariant characteristics of business processes of enterprises. UEBP can be quite complex and include not only consistent actions, but also various business conditions, repetitive operations. UEBP must contain operations that simulate the learning situation in the Business Game. The learning situation is understood as the situation in which decisions are made in the process of selecting resources for performing operations and/or the next operation of business process, etc. The learning situation allows to form or verify the player's competencies.

The Business Game is an interactive test for each player, and, as it is known, the tests should include questions, the correctness of the answers to which has a normal distribution. Therefore, there are two types of Decision Making Points taking a role of bottlenecks in Business Game or UEBP. Types of such points are the following:

- Simple DMPs are DMPs in which almost nobody makes mistakes even passing a game for the first time.
- DMPs of increased complexity are DMPs in which even the most competent players make the same mistakes.

The data analysis for the search for "bottlenecks" in Business Game should be implemented using one of the Data Mining methods - clustering. At this stage of the design, we are looking for clusters of three types of DMPs:

- Simple DMPs.

- DMPs of normal complexity.
- DMPs of increased complexity.

Set of characteristics of the same type is used for each combination of Business Process and Decision Making Point. Training sample is formed for all Business Processes.

The following characteristics must be used for DMPs' clustering:

- Amount of mistakes made when selecting mandatory resources.
- Amount of mistakes made when selecting optional resources.
- Formed percentage of the player's competence within each game.

During factor analysis it was revealed that the characteristics must be normalized in order to reduce the amount of data. Normalization of data is performed by calculating the following values for each DMP:

- Average rate of mistakes made when selecting mandatory resources.
- Average rate of mistakes made when selecting optional resources.
- Average rate of the player's competence within each game.

These average values represent three dimensions in the characteristic set for clustering.

After that it was necessary to identify the most appropriate clustering algorithm for finding bottlenecks in Business Game.

It is important to understand that the search for problem Decision Making Points needs to be done in two stages, that is, clustering is performed two times. Simple Decision Making Points need to be identified in a sample that includes the results of absolutely all games, including games of players with low level of competencies. Decision Making Points of increased complexity should be identified only among those games for which users have received high assessment, that is, the average player's competence within a business process is at least 75%. The second sample allows clearing the data from the unsuccessful traineeship due to a lack of knowledge of business processes.

The paper [Barsegian, 2004] provides a description of the clustering algorithms that is later is used for algorithms' comparison.

Comparison of clustering algorithms will be performed according to the following criteria:

- The total number of clusters is known (three clusters: simple DMPs, DMPs of normal complexity and DMPs of increased complexity).
- The volume of data sets may vary.
- The form of the clusters is arbitrary.
- Ease of work with multidimensional objects.

– The distance between clusters is small.

These criteria were singled out on the basis of the initial data (data in info-cube for searching bottlenecks in Business Game), requirements for the result of data analysis and analysis of clustering methods.

The comparison is made by the method "from the inverse", that is, it is determined which algorithms do not satisfy the criteria in question. A comparison of clustering algorithms is presented in Table 7.

Table 7. Clustering Algorithms Comparison

Algorithm Name	Known Total Number of Clusters	Variable Volume of Data Sets	Arbitrary Form of Clusters	Ease of Work With Multidimensional Objects	Small Distance Between Clusters
AGNES (Agglomerative Nesting)	No	Yes	Yes	Yes	Yes
CURE	Yes	No	Yes	Yes	Yes
DIANA (Divisive Analysis)	No	Yes	Yes	Yes	Yes
BIRCH	No	Yes	No	Yes	Yes
MST	No	Yes	Yes	No	Yes
K-means	Yes	Yes	Yes	Yes	No
Maximin	No	Yes	Yes	Yes	Yes
PAM	Yes	No	Yes	Yes	Yes
CLOPE	No	No	Yes	Yes	Yes
Self-organizing Map	Yes	No	Yes	Yes	Yes
HCM	Yes	No	Yes	Yes	Yes
Fuzzy C-means	Yes	No	Yes	Yes	Yes

The results of the comparison show that no algorithm fully meets all the criteria. However, it is important to take into account that under large volumes of data, databases with a multimillion-number transactions and a large set of characteristics are understood. As factor analysis allowed reducing characteristic set to three dimensions and OLAP technology allows getting aggregated data, it is assumed that actually the volume of data is not large in the general sense. Thus, evaluating the characteristics of the algorithms, it was decided to use the PAM algorithm for DMPs' clustering, since the training sample will not include a huge number of objects, the number of clusters is set and equal to three, the algorithm is less sensitive to emissions, the occurrence of which cannot be predicted in advance. Clusters will be classified in remoteness from the reference model.

In order to exclude of possible clustering mistakes related to fixed numbers of clusters two technical DMPs should be added into training sample. For instance all DMPs might be of normalized complexity, but PAM algorithm will distribute them into three sets anyway as this condition is set initially. Technical DMPs with following parameters {0; 0; 100} and {100; 100; 0} representing simple DMP and DMP of increased complexity properly allows getting rid of this problem. Here the first parameter is average rate of mistakes made when selecting mandatory resources, the second - average rate of mistakes made when selecting optional resources and the third - average rate of the player's competence gained during the DMP performance. Such technical DMPs should not be displayed to the player as an output, but allow avoiding errors associated with a fixed set of clusters.

In addition to clustering, supervised classification should also be applied to evaluate the quality of DMPs' design. Since DMPs distributed in clusters «Simple DMP» and «DMP of increased complexity» might be simple or complex but have not worthless id their parameters are not equal to {0; 0; 100} or {100; 100; 0}. The decision about such points redesign has to be made by the developer of UEBP, however DMPs having parameters equal to {0; 0; 100} or {100; 100; 0} should be highlighted singularly since they require redesign doubtlessly.

Conclusion

Since each resource is related with knowledge or skill analysis of player's competency is possible.

To conduct analysis of player's competency corresponding info-cube was designed. Applying Complex Analysis methods such as aggregation, navigation and filtering following reports regarding player's competency can be obtained:

– Player's competency within a business process.

- The percentage of different player's competences.
- Bottlenecks in player's knowledge and skills.
- The reasons of the lack of player's competency.
- The list of the most qualified participants.
- Weaknesses of players.
- Knowledge and skills not acquired by players previously.
- The average time taken to complete a single iteration of the business process.
- The progress of players in time.

To conduct analysis of Business Game another info-cube was designed. The analysis of the Business Game includes an assessment of the degree of successful DMPs performance in order to determine if DMPs were designed correctly. To assess the quality of the business game, it is proposed to distinguish three types of Decision Making Points:

- Simple DMPs.
- DMPs of normal complexity.
- DMPs of increased complexity.

Clustering is used to distribute all DMPs to these types. To determine the most appropriate clustering algorithm, a characteristic set was determined:

- Amount of mistakes made when selecting mandatory resources.
- Amount of mistakes made when selecting optional resources.
- Formed percentage of the player's competence within each game.

In order to increase operating speed of clustering algorithm, it was necessary to reduce the number of analyzed transactions. Therefore, it was decided to normalize the analyzed indicators and characteristic set was reformulated as follows:

- Average rate of mistakes made when selecting mandatory resources.
- Average rate of mistakes made when selecting optional resources.
- Average rate of the player's competence within each game.

Since the number of clusters is knows and the volume of training sample is not large due to normalization of analytic set it was decided to use the PAM in order to assess DMPs. In addition to clustering supervised classification should be applied.

Based on implementation of clustering and supervised classification algorithms the UEBP developer is able to identify DMPs that are recommended to revision as well as DMPs that must be revised as they distort the results of player's competence assessment.

Bibliography

[Barsegian, 2004] Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Методы и модели анализа данных: OLAP и Data Mining. СПб.: БХВ-Петербург, 2004, С. 67-75.

[Hung, 2012] Hung J.L., Rice K., Saba A. An Educational Data Mining Model for Online Teaching and Learning. Journal of Educational Technology Development and Exchange, 2012, pp. 77-94.

[Jeong, 2013] Jeong H. Educational Data Mining. How Students' Self-motivation and Learning Strategies Affect Actual Achievement. Department of Computer Science, Indiana University-Purdue University Fort Wayne, 2013.

[Kolb, 2012] Kolb E. BW310: BW - Enterprise Data Warehousing – Germany: SAP AG, 2012.

[Kozodaev, 2015] Козодаев М.А. Оценка проектного персонала: не забыть бы, для чего это делается (часть 1). Управление проектами и программами, 2015.

[Sahedani, 2013] Sahedani K.S., Supriya Reddy B. A Review: Mining Educational Data to Forecast Failure of Engineering Students. International Journal of Advanced Research in Computer Science and Software Engineering, 2013, pp. 628-635.

[Vikentyeva, 2013] Викентьева О.Л., Дерябин А.И., Шестакова Л.В. Концепция студии компетентностных деловых игр. Современные проблемы науки и образования, № 2, 2013, http://www.science-education.ru/108-8746.

[Vikentyeva et al., 2013] Викентьева О.Л., Дерябин А.И., Шестакова Л.В. Функциональные требования к студии компетентностных деловых игр. Вестник Пермского национального исследовательского политехнического университета. Электротехника, информационные технологии, системы управления. № 8, 2013, С. 31-40.

[Vikentyeva, 2015] Викентьева О.Л., Дерябин А.И., Шестакова Л.В., Лебедев В.В. Многомодельный подход к формализации предметной области. Информатизация и связь. №3, 2015, С.51-56.

[Vikentyeva et al., 2015] Викентьева О.Л., Дерябин А.И., Шестакова Л.В., Красилич Н.В. Проектирование редактора ресурсов информационной системы проведения деловых игр. Вестник Пермского национального исследовательского политехнического университета. Электротехника, информационные технологии, системы управления. №16, 2015, С. 68-87.

[Vikentyeva, 2016] Vikentyeva O., Deryabin A., Krasilich N., Shestakova L. Employment of Business Intelligence Methods for Competences Evaluation in Business Games. International Journal "Information Technologies & Knowledge", V. 10, №3, 2016, pp. 286-299.

Authors' Information

Olga Vikentyeva – *National Research University Higher School of Economics, City of Perm, Perm, Russia, e-mail: oleovic@rambler.ru.*

Major Fields of Scientific Research: General theoretical information research, Multi-dimensional information systems

Alexandr Deryabin – *National Research University Higher School of Economics, City of Perm, Perm, Russia, e-mail: paid2@yandex.ru.*

Major Fields of Scientific Research: General theoretical information research, Multi-dimensional information systems

Nadezhda Krasilich – *National Research University Higher School of Economics, City of Perm, Perm, Russia, e-mail: mefaze@yandex.ru.*

Major Fields of Scientific Research: General theoretical information research, Multi-dimensional information systems

Lidiia Shestakova – *National Research University Higher School of Economics, City of Perm, Perm, Russia, e-mail: L.V.Shestakova@gmail.com.*

Major Fields of Scientific Research: General theoretical information research, Multi-dimensional information systems

TABLE OF CONTENTS