Krassimir Markov, Vladimir Ryazanov,
Vitalii Velychko, Levon Aslanyan
(editors)

New Trends in Classification and Data Mining

I T H E A
SOFIA
2010

Krassimir Markov, Vladimir Ryazanov, Vitalii Velychko, Levon Aslanyan (ed.)

New Trends in Classification and Data Mining

ITHEA®

Sofia, Bulgaria, 2010

First edition

Recommended for publication by The Scientific Concil of the Institute of Information Theories and Applications FOI ITHEA

This book maintains articles on actual problems of classification, data mining and forecasting as well as natural language processing:

- new approaches, models, algorithms and methods for classification, forecasting and clusterisation. Classification of non complete and noise data;
- discrete optimization in logic recognition algorithms construction, complexity, asymptotically optimal algorithms, mixed-integer problem of minimization of empirical risk, multi-objective linear integer programming problems;
- questions of complexity of some discrete optimization tasks and corresponding tasks of data analysis and pattern recognition;
- the algebraic approach for pattern recognition - problems of correct classification algorithms construction, logical correctors and resolvability of challenges of classification, construction of optimum algebraic correctors over sets of algorithms of computation of estimations, conditions of correct algorithms existence;
- regressions, restoring of dependences according to training sampling, parametrical approach for piecewise linear dependences restoration, and nonparametric regressions based on collective solution on set of tasks of recognition;
- multi-agent systems in knowledge discovery, collective evolutionary systems, advantages and disadvantages of synthetic data mining methods, intelligent search agent model realizing information extraction on ontological model of data mining methods;
- methods of search of logic regularities sets of classes and extraction of optimal subsets, construction of convex combination of associated predictors that minimizes mean error;
- algorithmic constructions in a model of recognizing the nearest neighbors in binary data sets, discrete isoperimetry problem solutions, logic-combinatorial scheme in high-throughput gene expression data;
- researches in area of neural network classifiers, and applications in finance field;
- text mining, automatic classification of scientific papers, information extraction from natural language texts, semantic text analysis, natural language processing.

It is represented that book articles will be interesting as experts in the field of classifying, data mining and forecasting, and to practical users from medicine, sociology, economy, chemistry, biology, and other areas.

General Sponsor: Consortium FOI Bulgaria (www.foibg.com).

Printed in Bulgaria

ISBN 978-954-16-0042-9

C\o Jusautor, Sofia, 2010

GROWING SUPPORT SET SYSTEMS IN ANALYSIS OF HIGH-THROUGHPUT GENE EXPRESSION DATA

**Arsen Arakelyan, Anna Boyajian, Hasmik Sahakyan,
Levon Aslanyan, Krassimira Ivanova, Iliya Mitov**

Abstract: *Genome wide expression analysis with DNA microarrays has become a mainstay of genomics research. The problem with microarrays is that there is no "gold standard" for microarray-generated data analysis. In most cases, traditional statistical and pattern recognition approaches are not productive in analysis of high dimension low sample size data analysis, and improvements, alternative approaches are currently being developed in order to effectively analyze and interpret microarray data. In this study we design and extend logic-combinatorial scheme that can be used in high-throughput gene expression data. Applied results show that proposed algorithm is able accurately discriminate different biological phenotypes, although some improvements should be further made.*

Keywords: *Pattern Recognition, Functional Pathways, gene expression, post traumatic stress disorder*

ACM Classification Keywords: *I.5 Pattern Recognition, I.5.2 Design Methodology*

Introduction

Genome wide expression analysis with DNA microarrays has become a mainstay of genomics research. It is obvious, that monitoring expression levels for thousands of genes at a time provides insights into cellular processes and responses that cannot be obtained by looking at one or a few genes. Traditional methods for gene expression measurements such as Northern blots can be time-consuming and labor-intensive and are not practical for application on a very large scale. The more global view and increased throughput made possible by the advent of parallel expression measurements with DNA microarrays has therefore opened a new window on cellular activity [Hill et al 2000; Shoemaker et al 2001]. Being introduced in early 1990th, the rapidly increasing popularity of these techniques is evidenced by the number of publications involving microarrays. Gene expression differences between two samples have been applied to disease diagnosis and classification [Bittner et al 2000]. Effects of reagents or drugs on gene expression patterns have been used to test drug efficacies and to determine pharmacological mechanisms [Namba et al 2001; Dan et al 2002]. Correlated mRNA expression profiles under different cellular conditions have been used to predict gene functions [Hughes et al 2000]. DNA and RNA quantifications have also been used for detecting bacterial and viral pathogens, as well as host–pathogen interactions [Grayson et al 2002]. However having undoubted advantages, DNA arrays also has disadvantages that at the moment limit their usage. Currently, both oligonucleotide and spotted cDNA arrays are hybridized and read one at a time and significant time and effort is required to process even a modest number of samples. In many cases, a small number of experiments that cover thousands of genes is not sufficient. It has become increasingly clear that large collections of expression results are much more than the sum of their parts. The analysis of multidimensional expression patterns can reveal new insights that may not be apparent when looking at the results from small numbers of samples [Hughes et al 2000; Ross et al 2000].

Second problem with microarrays are that there is no "gold standard" for microarray-generated data analysis. From the very beginning the primary interest was to identify differentially expressed genes and elucidate related biological processes. The most widely used approach is individual gene analysis which evaluates the significance of individual genes between two groups of samples compared. This type of analysis typically yields a list of

altered genes from a cutoff threshold. Generally the basic approach of analysis is mathematical statistics (MS). Having satisfactory amount of experimental data (statistics) it helps to form conclusions that some properties and postulations take place in some probabilistic level. Simple correlation, regression and hypothesis estimation algorithms are components of the statistical approach. However, strong normality, and independence assumptions make them impractical and not powerful enough. A different situation appears in area of pattern recognition (PR). There is no satisfactory statistics in this case. These heuristics are more responsible and conditional. Learning set is given as a limited number of known classifications but it has to be large enough to describe the class properties in application area. A number of basic approaches are known in PR - Metric Algorithms, Logic Separation (LS), Neural Networks, etc. One of the well-known classes of metric algorithms is the voting (or estimation calculation) model [Zhuravlev 1998]. This is an algorithmic model with a number of additional parameters, requiring optimization during the learning stage. Several improvements and alternative approaches were developed, based on clustering techniques [Divina and Aguilar-Ruiz 2006; Von Borries 2008], support vector machines [Brown et al 2000; Benito et al 2004], "cut-off free" gene ranking [Subramanian et al 2005], dimension reduction [Nicolau et al 2007], and so on, but there is still a room for improvements of existing and development of new algorithms for analysis of high-throughput gene expression data. In this study we design and extend logic-combinatorial scheme for microarray data analysis.

Several assumptions related to the biological nature in terms of classes, features and learning sets were used defining the algorithmic model:

1. Given a dataset of features (gene expression values) and objects characterized by them, and objects are classified in two or more groups (healthy vs. diseased, different treatments, etc.). It is supposed that the biological process exists that is responsible for the differences of classes in terms of phenotype (blue vs. brown eyes, high vs. normal biological activity).
2. If such process exists, features (groups of features) related to that process should be able discriminate given classes of objects.
3. A feature set has higher probability to accurately classify objects, if its subsets, basically, are known to separate these classes.
4. As more features sets representing same biological process is found, as easier and accurate to identify the process itself.

Algorithm

Formal mathematical model considered in this article refers to the natural classification of objects characterized by the sets of features. Objects are characterized by numerical values of features. We suppose a learning set is given which is a list of example objects with their feature assignments, and their properties in terms of membership to some nonintersecting classes. The global problem in area is how to learn classification by limited learning set. Our primary target is in indentifying the features sets of most effective functionality in terms of considered classification. In some point of view we consider a typical problem of applied mathematical statistics or a basic pattern recognition postulation – the supervised learning.

This is true in part and our model gains additional specifics, coming with gene expression data. First of all it is to take into account the extremely large number of features appears here - genes and their expressions. Formally with limited and poor learning set large number of groups of these features may appear that function similarly in object classification. Redundancy appears as a consequence of the learning set limitations vs. very large feature set. If learning set size π is given, then a corresponding number φ of features can effectively participate in process of partitioning the object set into the subsets. $\log \pi$ plays an important role in this. This is minimal power to split the set to separate objects, and this is upper estimate of learning set size because of groups are

supposed not separate objects. Determining such feature φ-sets is the primarily goal of the algorithm we construct. This part is a typical task in pattern recognition area. The technique we apply is the support set systems generation. This technique is introduced first by [Zhuravlev 1998] and now is known in data mining area as frequent subsets [Aslanyan and Sahakyan 2009]. Our algorithm applies a supervised search procedure to construct the set of all support φ-sets of features. Besides this we have to deal with another problem which is more specific, and nonstandard. The applied problem of classification with gene expression data deals with very large sets of objects. The real support sets on these data are large and due to limitations of practical/experimental learning sets we can see only the φ-subsets of them. The same time due to redundancy the quality φ-subsets are mixed with the noisy features and no separation of features is possible due to low learning and large feature sets. Fortunately the application problem helps. The concept of functional pathways is applied. We code a pathway as the corresponding set of features in that pathway. Among several hundreds of pathways considered in a problem only several may correspond to the given classification. We take a hypothesis that such pathways have higher intersections with found support φ-subsets. This is because of real supportive φ-subsets always appear although the noisy ones appear spuriously.

Finally, the situation we are with gene expression data doesn't obey the regular requirement neither statistics nor the pattern recognition. Extremely small learning set and extremely large feature set creates a situation when there is not enough statistical information to treat the hypotheses and when there is a set of features that doesn't allow applying analytic tools to recognize the valid feature subsets. This situation is known in high-throughput gene expression data analysis. Microarray data is known as high dimension low sample size data (HDLSS) [Von Borries 2008] and this specific class of experimental data sets initializes a new algorithmic research area. Current study applies theoretical and algorithmic issue of set systems to solve the problems with HDLSS. In general description we grow the support sets with quality estimations, then apply this estimates to restrict the sets to computability sizes, and after recurrent φ-subsets generation continue through evaluation stage of subsets given by functional pathways.

Consider gene expression data $\Im$ which is a numeric array of n columns representing samples and D rows corresponding to genes. $\Im = \{S_1, S_2, S_3, \ldots, S_n\}$, where S_i is a $1xD$ column containing gene expression data for sample i. HDLSS means $n << D$. Classes which are 2 w.l.o.g. consist of objects: $S_1, S_2 \ldots S_k$ - belonging to the first class (e.g. disease class), and $S_k, S_{k+1} \ldots S_n$ - samples belonging to second class (e.g. healthy class). It is evident that almost any unique row $S_1(i), S_2(i), \ldots, S_n(i)$ of $\Im$ can correctly classify the two classes by linear hyperplane or by some other similar classification mean. The number of such rows might be very large among the D. The same time, it is realistic that different sets of rows are classifying the classes differently. Formally, a collection of subsets of the set $\{1,2,\ldots,n\}$ is known as a set of support systems Ω [Zhuravlev 1998]. In our model support system is the "unit" of comparison of object description pairs. This is when a set of distances, - each by a member of Ω is considered, which collectively describe the differences among the objects of different classes. The application counterpart is that support set is a minimal set of features – so that no any smaller and larger feature set is describing particular classification in a higher approximation. This brings to the problem of determining the proper row subsets (support systems), which provide maximal differences between classes (quality vs. accuracy of classification). In doing this we will eliminate the equivalent (in some sense) rows from one side; and will compose the sets of rows representing different equivalency subsets as approximations to the proper support systems.

1. Classifiers

At first we define **Elementary classifiers**.

These are hyperplane (or neural or any similar) classifiers by small number of rows. **1_classifier** is defined through one single row and its expression values $S_1(i), S_2(i), \ldots, S_n(i)$. Denote by $t(i)$ the classification level/accuracy of this attribute. 1_classifiers $c_1(i)$, $i \in \overline{1,D}$ ranked by classification levels are truncated by a version of sigma's rule. We leave out low classification genes from any further consideration. **2_classifiers** consider the pairs of genes and expression values. Logically 2_classifiers are to be composed by pairs of genes higher ranked by corresponding 1_classifiers. 2_classifiers and in general k **_classifiers** consider any k rows, construct "hyperplanes" and define structures $c_k(i_1,...,i_k)$ and $t_k(i_1,...,i_k)$. They construct convex hulls in areas of two considered classes; consider their partition hyperplane, rank by the classification level and narrows if necessary the set of rows used in future algorithmic steps.

2. Growing Support Systems

Among 2^n elementary classifiers mentioned above, we look for those, which correspond to subsets of genes that are more differently expressed in classes. The simplest way is to start by a 1_classifiers, and grow them step by step to k _classifiers so that the classification level is strictly increasing. Any k _classifier may be considered as a composition of one $k-1$-classifier together with one new row. Concepts $c_k(i_1,...,i_k)$ and $t_k(i_1,...,i_k)$ in this way introduce monotony relation between gene sets, putting them into 1-1 correspondence to the vertices of an n dimensional unit cube. However, practical implementation might be rather hard because of for large k it will become impossible to consider all 2^k sub-classifiers. The search area for these subsets is very large, and appropriate heuristics to combat this complexity is necessary.

We consider several heuristics:

- Sorting 1_classifiers by decreasing forces $t_1(i)$, and eliminating from the further treatment rows with forces lower than the threshold selected. Let the rows in sorted sequence are as $i_1, i_2, ..., i_k, ..., i_D$. An important point of this sequence is the first index j where $t_k(i_1,...,i_k)$ is increasing in area $i_k < j$ and this increase interrupts at the point j. Sorting is applied to the composite multi-feature classifiers because of the comparability of their classification measures.
- Consider an arbitrary elementary classifier $c_k(i_1,...,i_k)$. Compose an n dimensional binary vector, assigning its coordinates $i_1,...,i_k$ as 1. Completing by 0 all reminder coordinates, we create 1-1 correspondence between classifiers and n-cube vertices. Applying hierarchical clustering in n-cube layers we split k-classifiers into the groups by the equivalency relation (after some cut of dendrogram). Similarity measure used in clustering is some correlation between the hyperplanes (their coefficient vectors). We consider the representative sets of clusters. Some of them may give the same force of classifying objects by gene expressions as the whole descriptive table does. In this way we reduce the dimensionality combating the exponential explosion for large n.
- As it was mentioned, 1_classifiers might be directly sorted by their forces. Any k _classifier may be considered as a composition of one $k-1$-classifier $c_k(i_1,...,i_{k-1})$ together with one new row i_k. In terms of class vectors this change means concatenation of a new dimension in direction i_k. Concepts $c_k(i_1,...,i_k)$ and $f_k(i_1,...,i_k)$ in this way introduce monotony relation between gene sets in the same way as the vertices of n dimensional unit cube which are in 1-1 correspondence to elementary classifiers. Considering subsets of different n-cube layers and taking into account monotonity we may apply the chain split technology [Aslanyan and Sahakyan 2009] in finding the best separating gene sets.

It is important to note that chain split (and other known frequent subsets growing algorithms of association rule mining) work on systematized structure of all objects which is a hard computationally. Instead, the representative set mentioned above are a valuable heuristic that may help in reducing the computational complexity in growing.

3. Approximating Support Systems

Mathematical model and algorithms we consider have to implement the application area hypotheses that we defined above. The formal picture that satisfies the hypotheses is a monotone Boolean function over the set of gene expressions. Upper zeros of this function corresponds to the real support sets. Hypotheses say that having very large and satisfactory data the correct growing procedure will achieve these support sets. Having smaller learning sets procedures will deduce some partial of approximate collections of support sets. It is a hard question how to evaluate the quality of these sets. Formal postulation might be an approximation scheme for monotone Boolean functions. Two schemes were considered in this regard:

- Frequency based approach. For an arbitrary binary vector α (a unit cube vertex) vertices of layers of subcube between 0 and that vertex are considered. Frequencies of support sets after the growing stage are computed. Vertex α is determined as a recognized support set if the frequencies are over a given threshold.
- Chain split and interior/exterior sets approach. The chain split recognition of monotone Boolean function is considered [Aslanyan and Sahakyan 2009]. The usual procedure intends to find the exact function and Hansel chains are ideal for this. Consider the extension of chain split with rare chains. Chains are with edges that connect comparable but distance vertex pairs and their relative completions similarly defined, or the same Hansel chains are considered but recognized areas are extended in iterations to the interior and exterior vertices of monotone function by the given threshold. Growing support sets means one way recognition that is recognition through consideration of only (or basically) zero values of function. The same time this is constrained recognition in a sense that hypothetical pathways involve limited number of genes.

Algorithm Realization

Algorithm was realized using Matlab R2008a (Mathworks, Inc). At each iteration selection the list of classifiers and corresponding errors of classification are generated. In order to decrease computational intensity we used very strict cut-off for classifier selection $M(err) - 2 * SD(err)$. Only classifiers with error less than specified threshold were considered for future growing. Search of functional pathways were performed KEGG (Kyoto Encyclopedia of Genes and Genomes, www.genome.jp/kegg/) database using the SOAP/WSDL based web service.

Dataset Description and Preprocessing

To test the functionality of proposed algorithm dataset GDS1020 on gene expression patterns related to post-traumatic stress disorder (PTSD), publicity available in Gene Expression Omnibus repository (http://www.ncbi.nlm.nih.gov/geo/) was used. Description of samples was retrieved from dataset summary. GDS1020 dataset contains gene expression profiles of peripheral blood mononuclear cells (PBMC) isolated from the blood of trauma survivors at admission to emergency room and at 4 month follow-up. Overall, 33 (16 control and 17 PTSD subjects) gene expression profiles were available in dataset. PTSD affected subjects persistently manifested full criteria for PTSD.

From human dataset probes that had at least 50% of present calls were chosen. The missing values were replaced by group mean. Normalization was performed by division of gene expression in individual samples to geometric mean of gene expression of the all samples. The probe identifiers from each platform were converted to the HUGO-approved gene symbols, averaging log transformed expression values of multiple probes targeting the same gene.

Results and Discussion

During data preprocessing 12568 probes were collapsed in 8742 genes that were used in future analyses. Search for the 1-classifier identified 405 genes with classification error less than specified threshold. From selected 405 genes only 149 were included in 77 KEGG functional pathways. These 405 genes were further used for identification of k-classifiers. The iterations stop at 4th step (group of 4 genes) with final classification error 0. During the analysis 605 2-classifiers, 449 3-classifiers and 2270 4-classifiers were formed. Figure 1 shows the increase of classification accuracy in parallel with classifier order increase. When we performed search of KEGG functional pathways using formed classifiers, we have identified only 6 pathways that exactly matched with 2-classifiers and 2 pathways exactly matched with 3-classifiers. None of pathways contain all genes from any of 4-classifiers.

The results showed that the group of 4 genes is enough for precise separation of classes in our case. However, we could not identify any KEGG functional pathways that include these 4-classifiers. Here, several suggestions can be made. First of all, this approach identifies “best genes”, i.e. genes with maximal classification force. If the classification force of other genes from the same pathway is less profound they will not be included in 1-classifiers. Thus, this is very useful tool for the search of biomarker combinations that are able to discriminate classes with maximal accuracy, which is not possible with the use of single biomarker. Second, this can be the result of limited number of pathways annotated in KEGG database. At the moment KEGG database contain information only about 206 pathways. Although all 405 genes were annotated in KEGG, only 149 (36.7 %) of them were associated with one or more functional pathways. In order to check this suggestion, we re-evaluated our data using gene ontology (GO, www.geneontology.org/) database to search terms covered by 4-classifiers. In contrast to KEGG, we identified 841 gene ontology terms that exactly matched 4-classifiers. And finally, it is possible that dataset used in study is not valid input for this approach. The question of validity of initial data are of great importance, however, at the moment we do not have methodology to evaluate their adequacy to the considered problem of functional pathways analysis. In this regard, interpretations given above are conditional and larger and valid datasets needed to be considered.

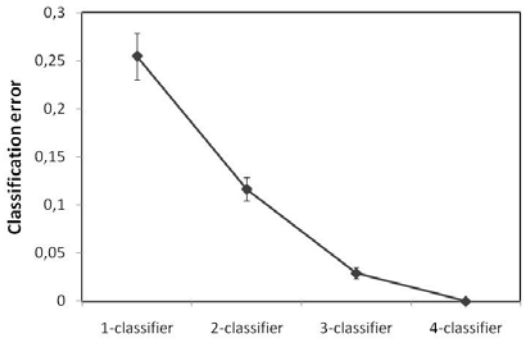

Figure 1. Classification accuracy of growing support sets.

Conclusion

Genome wide expression analysis models and algorithms were considered. It is recognized that in most cases, traditional statistical and pattern recognition approaches suffer to analyze these information due to disbalance between the numbers of features and objects. Traditional in these area high dimension low sample size data (HDLSS) were considered and logic-combinatorial pattern recognition is applied to analysis of this information. The structured algorithm is similar to frequent sets algorithm in data mining. After growing support sets approximation of maximal support sets is considered in a model of monotone Boolean recognition. Applied results show that proposed algorithm is able accurately discriminate different biological phenotypes, although some improvements should be further made.

Bibliography

[Hill et al 2000] A.A. Hill, et al. Genomic analysis of gene expression in C. elegans. Science, 2000, 290: 809–812.

[Shoemaker et al 2001] D.D. Shoemaker, et al. Experimental annotation of the human genome using microarray technology. Nature, 2001, 409: 922–927.

[Bittner et al 2000] M. Bittner, et al. Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature, 2000, 406(6795): 536-540.

[Namba et al 2001] H. Namba, et al. Efficacy of the bystander effect in the herpes simplex virus thymidine kinase-mediated gene therapy is influenced by the expression of connexin43 in the target cells. Cancer Gene Ther, 2001, 8(6): 414-420.

[Dan et al 2002. S. Dan, et al. An integrated database of chemosensitivity to 55 anticancer drugs and gene expression profiles of 39 human cancer cell lines. Cancer Res, 2002, 62(4): 1139-1147.

[Hughes et al 2000] T.R. Hughes, et al. Functional discovery via a compendium of expression profiles. Cell, 2000, 102(1): 109-126.

[Grayson et al 2002] T.H. Grayson, et al. Host responses to Renibacterium salmoninarum and specific components of the pathogen reveal the mechanisms of immune suppression and activation. Immunology, 2002, 106(2): 273-283.

[Ross et al 2000] D.T. Ross, et al. Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet, 2000, 24(3): 227-235.

[Zhuravlev 1998] Yu. Zhuravlev. Selected research publications. Magistr, Moscow, 1998, 420p (in russian).

[Von Borries 2008], G.F. Von Borries, Partition clustering of high dimensional low sampling size data base on p-values, PhD dissertation, Kansas State University, 2008, p. 139.

[Divina and Aguilar-Ruiz 2006] F. Divina and J. S. Aguilar-Ruiz. Biclustering of expression data with evolutionary computation. IEEE Transactions on Knowledge and Data Engineering, 2006, 18: 590–602.

Benito et al 2004] M. Benito, et al. Adjustment of systematic microarray data biases. Bioinformatics, 2004, 20: 105-114.

[Brown et al 2000] M.P.S. Brown, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. PNAS, 2000, 97(1): 262–267.

[Subramanian et al 2005], A. Subramanian, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. PNAS, 2005, 102(43):15545-15550.

[Nicolau et al 2007], M. Nicolau, el at. Disease–Specific Genomic Analysis: Identifying the Signature of Pathologic Biology. Bioinformatics, 2007, 23(8): 957-965.

[Aslanyan and Sahakyan 2009] L. Aslanyan, H. Sahakyan. Chain split and computation in practical rule mining, Information Science and Computing, International book series no. 8., Classification, forecasting, data mining, 2009:132-135.

Authors' Information

***Arsen Arakelyan**, **Anna Boyajian** – "Laboratory of Information Biology" Project of the Institute of Molecular Biology and Institute for Informatics and Automation Problems NAS RA, Yerevan, Armenia, e-mail: aarakelyan@sci.am*

***Hasmik Sahakyan, Levon Aslanyan** –Institute for Informatics and Automation Problems, NAS RA, P.Sevak St. 1, Yerevan 14, Armenia, e-mail: hasmik@ipia.sci.am, lasl@.sci.am*

***Krassimira Ivanova, Ilia Mitov** – Institute of Mathematics and Informatics – Bulgarian Academy of Sciences, Sofia, Bulgaria, e-mail: kivanova@math.bas.bg, mitov@foibg.com*