LOGIC BASED PATTERN RECOGNITION - ONTOLOGY CONTENT (2)[1]

Levon Aslanyan, Vladimir Ryazanov

Abstract: Logic based Pattern Recognition extends the well known similarity models, where the distance measure is the base instrument for recognition. Initial part (1) of current publication in iTECH-06 reduces the logic based recognition models to the reduced disjunctive normal forms of partially defined Boolean functions. This step appears as a way to alternative pattern recognition instruments through combining metric and logic hypotheses and features, leading to studies of logic forms, hypotheses, hierarchies of hypotheses and effective algorithmic solutions. Current part (2) provides probabilistic conclusions on effective recognition by logic means in a model environment of binary attributes.

1. Introduction

Pattern Recognition consists in reasonable formalization (ontology) of informal relations between object's visible/measurable properties and of object classification by an automatic or a learnable procedure [1]. Similarity measure [1] is the basic instrument for many recognition formalisms but additional means are available such as logical terms discussed in part (1) of current research [2]. Huge number of recognition models follow the direct goal of increasing recognition speed and accuracy. Several models use control sets above ordinary learning sets, others use optimization and other direct forces. Besides, more alternative notions are available to describe algorithmic properties. In existing studies the role of these notions is underestimated and less attention is paid to these components. In part (1) the attention is paid to implementing the learning set through its pairs of elements rather than the elements separately. The following framework is considered: given a set of logical variables (properties) $x_1, x_2, ..., x_n$ to code the studied objects, and let we have two types/classes for classification of objects: K_1 and K_2. Let $\beta \in K_1$, and $\gamma \in K_2$, and α is an unknown object in sense of classification. We say, that γ is separated by the information of β for α if $\beta \oplus \gamma \le \beta \oplus \alpha$, where $\oplus$ is $mod\, 2$ summation. Formally, after this assumption, the reduced disjunctive normal forms of two complementary partially defined Boolean functions appear to describe the structure of information enlargement of the learning sets. The idea used is in knowledge comparison. α is an object of interest. Relation $\beta \oplus \gamma \le \beta \oplus \alpha$ informs that the descriptive knowledge difference of β and α is larger than the same difference of β and γ. This approach we call logic separation. While notion of similarity gives the measure of descriptive knowledge differences, the logic separation describes areas which are preferable for classes and learning set elements. In general the question is in better use of learning set. The learning set based knowledge, which is used by recognition procedure, at least is supposed to reconstruct the learning set itself. It is indeed negative when these information is not able to reconstruct the learning set. It is easy to check that the similarity knowledge can't reconstruct an arbitrary learning set, and only special sets allow reconstructing of objects by their distances [3]. Restructuring power is high when comparison is used for the set of all attribute subsets. Theoretically such structures are studied in discrete tomography problems [4], but practically even the use of pairs draws to known hard computational area of disjunctive normal forms.

[1] The research is supported partly by INTAS: 04-77-7173 project, http://www.intas.be

Consider pairs of elements of the learning set, where each pair contains elements of different classes (the case of 2 learning classes is supposed). It was shown [2] that the logical separators divide the object space into three areas, where only one of these areas needs to be treated afterward by AEA (algorithms of estimation analogies – voting algorithms) [1]. This set is large enough for almost all weakly defined Boolean functions, but for the functions with compactness property it is small. Let, for $0 \le k_0 < k_1 \le n$, F_{n,k_0,k_1} be the set of all Boolean functions defined as follows: each of them has zero (false) value on the vertices of k_0-sphere centered at $\tilde{0}$, and has one (true) value on $(n-k_1)$-sphere centered at $\tilde{1}$. On the remainder vertices of n-cube the assignment/evaluation is arbitrary. These functions (for appropriate choice of k_0 and k_1) satisfy the compactness assumptions [8], and their quantity is not less than $2^{\varepsilon(n)2^n}$ for an appropriate $\varepsilon(n) \to 0$ with $n \to 0$. For these functions we have also, that for recovering the full classification by means of logical separators procedure, it is enough to consider a learning set which consists of any $n2^{n-\varepsilon(n)\sqrt{n}}$ or more arbitrary points. This is an example of postulations which will be considered below. It is relating the metric and logic structures and suppositions, although separately studies of these structures are also important. The follow up articles will describe the mixed hierarchy of recognition metric-logic interpretable hypotheses, which helps to allocate classification algorithms to the application problems.

2. Structuring by Logic Separation

Let f be a Boolean function (it might be partially or completely defined). Let N_f denotes the reduced disjunctive normal form of f and sets $N_0^f,\dots,N_3^f$ [2] define areas, in which N_f and $N_{\bar{f}}$ take values {0,1}, {1,0}, {0,0} and {1,1} correspondingly. Identical to $N_0^f,\dots,N_3^f$, similar areas are defined by logic separation - $M_0^f,\dots,M_3^f$.

Let $f_0 \in P_2(n)$ (a completely defined Boolean function of n variables) and $f_0(\tilde{\alpha}) = 1$. Denote by $t(f_0,\tilde{\alpha})$ the number of k-subcubes included in N_{f_0} and covering the vertex $\tilde{\alpha}$. Let m_k is the average number of $t(f_0,\tilde{\alpha})$ calculated for all $f_0 \in P_2(n)$ and $\tilde{\alpha} \in N_{f_0}$. It is easy to check that

$$m_k = \frac{2^n C_n^k 2^{2^n-2^k}}{2^n 2^{2^n-1}} = \frac{C_n^k}{2^{2^k-1}}.$$

Dispersion d_k of the same value $t(f_0,\tilde{\alpha})$ is expressed as

$$d_k = C_n^k \sum_{j=0}^{k} C_k^j C_{n-k}^{k-j} 2^{-2^{k+1}+2^j+1} - \left(\frac{C_n^k}{2^{2^k-1}}\right)^2.$$

Applying the Chebishev inequality to above measures $t(f_0,\tilde{\alpha})$, m_k, d_k leads to the conclusion:

Proposition 1(8). $t(f_0,\tilde{\alpha}) \sim \frac{C_n^k}{2^{2^k-1}}$ for almost all pairs $f_0 \in P_2(n)$ and $\tilde{\alpha} \in N_{f_0}$, when $n \to \infty$ and $\frac{C_n^k}{2^{2^k}} \to \infty$.

Taking into account that for almost all Boolean functions the number of 1-vertices is equivalent to $2^{n-1}, n \to \infty$, we obtain that for almost all functions $f_0 \in P_2(n)$, almost all 1-vertices are covered by the number of k-intervals from N_{f_0}, which is equivalent to $\frac{C_n^k}{2^{2^k-1}}$, when $n \to \infty$ and $\frac{C_n^k}{2^{2^k}} \to \infty$. Particularly, this fact might be used to adjust the postulation in Proposition 7, [2]. Indeed, the $\frac{C_n^k}{2^{2^k-1}}$ intervals, coming from a common fixed vertex, cover not less than $\frac{C_n^k}{2^{2^k-1}}$ vertices of an n-cube.

Now consider arbitrary placement of any l points into the vertices of an n-cube M. Estimate for almost all functions $f_0 \in P_2(n)$ (see Proposition 1(8)) the main value of the number of vertices $\tilde{\alpha} \in N_{f_0}$, which are not covered by any of the k-intervals included in N_{f_0} which is pricked by our l vertices:

$$\mu(n,k,l) \prec (1+\varepsilon_1(n))2^{n-1}\frac{C^l_{2^n-(1+\varepsilon_2(n))C_n^k/2^{2^k-1}}}{C^l_{2_n}}, n \to \infty,\ \varepsilon_1(n) \to 0,\ \varepsilon_2(n) \to 0, \text{ and } \frac{C_n^k}{2^{2^k}} \to \infty .$$

Proposition 2(9). If $\frac{C_n^k}{2^{2^k}} \to \infty$ and $l \geq \varphi(n)\frac{2^n 2^{2^k}}{C_n^k}$, where $\varphi(n) \to \infty$ as $n \to \infty$, then random l vertices for almost all functions $f_0 \in P_2(n)$ prick such sets of k-subcubes included in N_{f_0}, which cover almost all N_{f_0}.

In case of $k = [\log\log n]$ we conclude that the minimal number l satisfying the above proposition, is not greater than $2^n n^2 / C_n^{[\log\log n]}$.

Notice, that in conditions of Proposition 7 [2] and Proposition 2(9) only the usability of condition F_0 (logic separation) is mentioned, so that these are the conditions, when usage of F_0, as a rule, doesn't imply to significant errors. Also, it is important, that we applied the condition F_0 to the whole class $P_2(n)$, although it was supposed for problems, satisfying compactness suppositions. So, it is interesting to know how completely the class $P_2(n)$ satisfies to these suppositions.

Let us bring now a particular justification of compactness conception [8]. Let $f_0 \in P_2(n)$. We call the vertex $\tilde{\alpha} \in M$ boundary vertex for function f_0, if the sphere $S(\tilde{\alpha},1)$ of radius 1 centered at $\tilde{\alpha}$, contains a vertex for which f_0 has the opposite value to $f_0(\tilde{\alpha})$. Denote by $\Gamma(f_0)$ the set of all boundary vertices of f_0. We will say that the function hipping (completion) procedure obeys the compactness conditions, if $|\Gamma(f_0)| = o(2^n), n \to \infty.$

It is easy to calculate that the average number of boundary vertices of functions $f_0 \in P_2(n)$ is almost 2^n. This shows that $P_2(n)$ contradicts the compactness conditions. The same time we proved that the use of the F_0 rule in a very wide area $P_2(n)$ doesn't move to a sensitive error. Below we consider an example problem, which obeys the compactness assumptions, and will follow the action of the rule F_0 on that class. Before that we justify some estimates for the set M_3^f.

Consider the class $\Phi_2(n, k(n), l(n))$ of all of partial Boolean functions, for which $|M_0| = l(n)$ and $|M_1| = k(n)$. We'll deal with the case $l(n) = o(2^n)$ and $k(n) = o(2^n)$. Estimate now the quantitative characteristics of sets M_0^f, M_1^f and M_3^f.

First estimate the average number of vertices of the cube, which are achievable from set M_0:

$$C_{03} \geq 2^n \frac{C^{k(n)}_{2^n - l(n) - 2^j}}{C^{k(n)}_{2^n - l(n)}} \left(1 - \frac{C^{l(n)}_{2^n - \sum_0^j C_n^j}}{C^{l(n)}_{2^n}} \right), \ j = 0,1,\cdots$$

Proposition 3(10). If $k(n)$ and $l(n)$ are $o(2^n)$, $n \to \infty$ and there exists a j_0, that $k(n) 2^{j_0 - n} \to 0$ and $2^{-n} \sum_0^{j_0} C_n^i l(n) \to \infty$, then for almost all functions of class $\Phi_2(n, k(n), l(n))$, $|M_1^f| \approx o(2^n), n \to \infty$.

To except the trivial cases in the pattern recognition problems we have to suppose, that $k(n) \cong l(n), n \to \infty$. Then it is clear that choosing appropriate values for j_0 we get $|M_1^f| = o(2^n)$ and $|M_0^f| = o(2^n)$ for almost all functions of class $\Phi_2(n, k(n), l(n)), n \to \infty$.

Let us give an other estimation of c_{03}:

$$c_{03} \geq \sum_{j=0}^{n} C_n^j \frac{C^{l(n)}_{2^n - 2^j}}{C^{l(n)}_{2^n - 1}}.$$

If $\lambda(n)$ is the minimal value for which

$\sum_{i = n/2 - \lambda(n)}^{n/2 + \lambda(n)} C_n^i \sim 2^n, n \to \infty$, then $\lambda(n) \approx \sqrt{n}$.

From here we conclude:

Proposition 4(11). If $l(n) \geq 0$ and $2^{-n} k^2(n) 2^{C(n)\sqrt{n}} \to 0$ as $n \to \infty$ for $\forall c(n)$ - restricted, then almost ever $M_1^f \sim o(2^n)$.

So, for the small values of $k(n)$ and $l(n)$ from the each vertex of set $M_0 \cup M_1$, almost all vertices of the n-unite-cube almost ever are achievable. Comparing this, for example with [5] we find that for these classes F_0 works ineffectively.

3. Logic Separation on Compact Classes

Consider problems, satisfying the compactness assumptions. First of all it is evident, that for $M_0 \cup M_1 \supseteq \Gamma(f_0)$ the continuation of function f made on base of F_0, exactly correspond to the final result f_0. Taking into account that by the given description of the compactness assumptions $\Gamma(f_0) = o(2^n), n \to \infty$, we receive that in problems, satisfying the compactness assumptions we can point out learning sets of size $o(2^n), n \to \infty$, which allow to complete and exact continuation of function f_0 on base of condition F_0 only.

Let $\tilde{\alpha} \in M$ and $0 \le k_1 \le k_2 \le n$. Consider functions $f_0 \in P_2(n)$, for which $M_0(f_0) \supseteq S(\tilde{\tilde{\alpha}}, n-k_2)$, $M_1(f_0) \supseteq S(\tilde{\alpha}, n-k_1)$ and which receive arbitrary values on vertices of sets $S(\tilde{\alpha}, k_2 - 1) \supseteq S(\tilde{\alpha}, k_1)$.

Denote the class of these functions by $K(n)$. It is evident, that for $\left|S(\tilde{\alpha}, k_2 - 1) \setminus S(\tilde{\alpha}, k_1)\right| = o(2^n)$ all the constructed functions satisfy the given formalisms for the compactness assumptions, and that the quantity of these functions is not less than $2^{\varepsilon_1(n)2^n}$, where $\varepsilon_1(n)$ is an arbitrary function of $n, \varepsilon_1(n) \to 0$ with the $n \to \infty$.

Take a point $\tilde{\beta} \in M, \rho(\tilde{\alpha}, \tilde{\beta}) = k, k < k_1$. It is evident that no more than $C_n^{[n/2]} \cong 2^n \sqrt{\frac{2}{\pi n}}, n \to \infty$ subsets of any fixed size are coming out from any point of n-cube. From the other hand it is evident, that it is enough to take $k_1 - k = o(\sqrt{n})$ as $n \to \infty$ to get the $\left|S(\tilde{\alpha}, k_1) \setminus S(\tilde{\alpha}, k)\right| = o(2^n)$. Suppose, that $\left|M_0(f_0) \cup M_1(f_0)\right| = l$, and that l points appear as the result of their appropriate placement on the vertices of the n-cube M, when all of these placements are equally probable. Estimate the probability of reaching of this point $\tilde{\beta}$ from zeros of function f_0.

$\tau_l \le 2^n \sqrt{\frac{2}{\pi n}} \frac{C_{2^n - 2^{\varepsilon_2(n)\sqrt{n}}}}{C_{2^n}^l}, \varepsilon_{2(n)} \to 0$ with $n \to \infty$.

From here we conclude the

Proposition 5(12). Let $f_0 \in K(n, k_1, k_2)$ and f_1 -- is the continuation of function f on base of condition F_0. If $l \ge n2^{n - \varepsilon_2(n)\sqrt{n}} = o(2^n)$, $n \to \infty$ and the set $M_0(f) \cup M_1(f)$ is formed as a random collection of points of size l from the set M, then almost ever the function f_1 is the continuation for f, which converges to the f_0 by the accuracy, tending to 1 with the $n \to \infty$.

Conclusion

Logic Separation is an alternative approach to pattern recognition hypotheses and formalisms, while the base concept uses the similarity approach. Structures appearing in this relation are based on terms of Reduced Disjunctive Normal Forms of Boolean Functions. Propositions 1-5(8-12) provide additional knowledge on quantitative properties of areas appearing in extending classification by means of compactness and logic separation principles.

Bibliography

1. Zhuravlev Yu. I. On an algorithmic approach to the problems of recognition and classification. Problemi Kibernetiki, 33, (1978) 5--68.
2. Aslanyan L. and Castellanos J., Logic based pattern recognition – ontology content (1), iTECH-06, 20-25 June 2006, Varna, Bulgaria, Proceedings, pp. 61-66.
3. Gavrilov G. and Sapojenko A., Collection of exersises of discrete mathematics, NAUKA, Moscow, 1973, 368 p.
4. Herman G.T. and Kuba A., editors. Discrete Tomography: Foundations, Algorithms and Applications. Birkhauser, 1999.
5. Zhuravlev Yu. I. Selected Scientific Works, Publishing House Magister, Moscow, (1998) 417p.
6. Aslanyan L. H. On a pattern recognition method based on the separation by the disjunctive normal forms. Kibernetika, 5, (1975), 103--110.
7. Vapnik V. and Chervonenkis A. Theory of Pattern Recognition. "Nauka", 1974.
8. Aslanyan L. H. The Discrete Isoperimetric Problem and Related Extremal Problems for Discrete Spaces. Problemy Kibernetiki, 36, (1979), 85--128.
9. Nechiporuk E. I. On topological principles of self-correcting. Problemy Kibernetiki, 21, (1969), 5--102.
10. Graham N., Harary F., Livingston M. and Stout Q. Subcube Fault-Tolerance in Hypercubes. Information and Computation 102 (1993), pp. 280{314.
11. Glagolev V. V. Some Estimations of D.N.F. for functions of Algebra of Logic. Problemy Kibernetiki, 19, (1967), 75--94.

Authors' Information

Levon Aslanyan – Institute for Informatics and Automation Problems, NAS Armenia, P.Sevak St. 1, Yerevan-14, Armenia; e-mail: lasl@sci.am

Vladimir Ryazanov - Computing Centre of the Russian Academy of Sciences, 40 Vavilova St., Moscow, GSP-1, 119991, Russian Federation, rvvccas@mail.ru