A Procedure to Compute Prototypes for Data Mining in Non-structured Domains

J. Méndez, M. Hernández, and J. Lorenzo

Dpto. de Informática y Sistemas, Universidad de Las Palmas de Gran Canaria, 35017 Las Palmas, Spain. [jmendez, mhernandez,jlorenzo]@dis.ulpgc.es

Key words: learning, data mining, knowledge discovery, symbolic clustering.

Abstract. This paper describes a technique for associating a set of symbols with an event in the context of knowledge discovery in database or data mining. The set of symbols is related to the keywords in a database which is used as an implicit knowledge source. The aim of this approach is to discover the significant keyword groups which best represent the event. A significant contribution of this work is a procedure which obtains the representative prototype of a group of symbolic data. It can be used for both, unsupervised learning to describe classes, and supervised learning to compute prototypes. The procedure involves defining an objective function and the subsequent hypothesis-exploring system and obtaining an advantageous procedure regarding computational costs.

1 Introduction

Knowledge Discovery in Database and Data Mining involve a set of techniques used in an automated approach to exhaustively explore and bring to the surface complex relationships in very large datasets [4]. Data mining aims at finding useful regularities in large datasets. The interest in the field is motivated by the growth of computerized data collections and by the high potential value of patterns discovered in those collections [7]. Clustering is a technique used in a large number of data mining applications in different domains [9] [6] because it deals with two important issues: machine learning to generate abstractions and statistics to deals with data noise. Both, clustering and supervised learning, can need a procedure to compute class prototypes. This paper describes a technique for finding some symbolic description of an event. The symbolic description is established through labeled or attributed groups present in keywords of the database. The technique may be applied to the interpretation of any characteristic or event related with a database as long as an appropriate search engine exists. A similar approach is described by Guigó and Temple[10] applied to protein databases.

Biosequences Databases is a field where Data Mining tools play important and enabling roles [14] [13]. In most protein databases each entry contains, addicionality to its name and amino acid sequence, a set of annotated data which may be keywords associated with the functionality of protein characteristics. Likewise it may be accompanied by other data of diverse scientific interest. The information considered for each protein entry is a 3-tuple [*name,keywords,sequence*]. The

sequence is useful for locating the patterns in it. Different tools may be used for this task. Some of them are general purpose and have relative low performance, others include specific search engines for matching in protein sequences and DNA, as, for example BLAST [12].

Most dataset used in Machine Learning are attribute based. Each attribute can be a numeric value or a symbol take from a discrete set. Attribute based datasets are structured domains in which an attribute can be considered as a dimension in a continuous or discrete space. In this paper non-structured domain of samples are considered, that is only as set of symbols without other organization or algebraic class. Samples in a non-structured domain have not fixed cardinality or length. In the work of Guigó and Temple[10] a logical combination of keywords is given as a result. Let k be a boolean expression of keywords in the database and $F(k)$ the set of proteins that are matched by a query in the database. Let p be a protein pattern or motif, and $H(p)$ the set of proteins that are matched by a search engine in the database. Guigó and Temple consider k to be a possible explanation of the p pattern if the similarity or correlation between $F(k)$ and $H(p)$ is maximum. That is, $s(F(k), H(p))$ must be maximized, where $s(A, B)$ is a similarity measure between A and B sets, expressed using the set cardinalities $|A|$ and $|B|$ as:

$$s(A, B) = \frac{|A \cap B|}{\sqrt{|A||B|}} \tag{1}$$

The approach used in this paper is very different. Rather than determine a boolean expression of keywords, *it involves inducing which types or classes, and its prototypes M, are derived from $H(p)$ by means of a process of symbolic clustering appropriate to the nature of the data considered.* Rather than obtain a *or-of-and* expression of keywords, it obtains a *set-of-set* of keywords. Prior to the process of symbolic clustering, the keywords for each protein in $H(p)$ is obtained based on the information contained in the database and selecting some keyword types. Annotated databases as SWISS-PROT [1] includes a lot of attributes which must be selected. The resulting data from this process include a protein dataset with its selected keywords. It is referred to as the items I set. The proposed procedure can be used in all applications which mach with:[*item-name, keywords, other-data*].

Clustering is a useful tool for analyzing data because it is an inductive method which can discover certain regularities in the experimental data. It is possible to discover natural types or data clusters which are derived from the use of distance or similarity measures from the data themselves [11][3]. The set of samples is made up of the set of items $I = \{I_i; i = 1, \ldots, n\}$ associated with the set of proteins obtained by the search engine. Each item I_i includes the set of symbols associated with the database keywords in $U = \{s_j; j = 1, \ldots, m\}$. Each item may be expressed as: $I_i = \{s_j, u_{ij}; j = 1, \ldots, m\}$ where $s_j \in U$. The value $u_{ij} \in \{0, 1\}$ is the membership degree of s_j in the item I_i. This value may also be defined as: $u_{ij} = |I_i \cap \{s_j\}|$. The addressed problem involves determining the groups of items

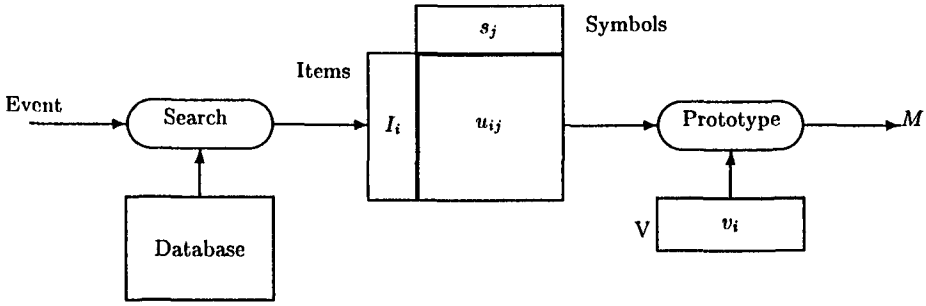

Fig. 1. Prototype Computation from the data obtained after a search in a Database

that best represent the natural classes of the total I set. The characterization of these groups is determined through their respective prototypes or means.

2 Computing Prototypes

This section is concerning with the problem of how to determine the symbol set contained in M so that it is most similar to a non null set of items I. Specifically, it establishes a measure of similarity as an objective function expressed as:

$$f(M) = \frac{1}{n_v} \sum_{i=1}^{n} v_i s(I_i, M) = \frac{1}{n_v} \sum_{i=1}^{n} v_i \frac{|M \cap I_i|}{\sqrt{|M||I_i|}} \tag{2}$$

where n_v it is given for:

$$n_v = \sum_{i=1}^{n} v_i > 0 \tag{3}$$

The term $v_i \in \{0, 1\}$ is the degree with which item I_i participates in the group which is used for computing the prototype. To carry out the computation different hypothetical solutions will be considered. Let X be a hypothesis about M that does not include a symbol s_k and X' a new hypothesis obtained by including this symbol, that is $X \xrightarrow{s_k} X'$.

$$X' = X \cup \{s_k\} \qquad X \cap \{s_k\} = \emptyset \tag{4}$$

The objective function value of X' can be obtained recursively from that of X as follows:

$$f(X') = \frac{\sqrt{|X|} f(X) + \lambda_k}{\sqrt{|X| + 1}} \tag{5}$$

where λ_k is the relevance factor of the symbol s_k given by the following expression:

$$\lambda_k = \frac{1}{n_v} \sum_{i=1}^{n} v_i \frac{u_{ik}}{\sqrt{|I_i|}} \tag{6}$$

This factor is the weighted average of the presence of a symbol in the different items. Previous results are deduced from: $|\{s_k\}| = 1$ and $|A \cup B| = |A| + |B| - |A \cap B|$.

Let S and P be the union and intersecction closures respectively:

$$S = \bigcup_{\forall i, v_i > 0} I_i \qquad P = \bigcap_{\forall i, v_i > 0} I_i \tag{7}$$

Let R be the complement of P into S, that is: $P \cup R = S$ and $P \cap R = \emptyset$, it is also expresed as : $R = S - P$. Let $h = |R|$ be its cardinality.

Theorem 1 *If a symbol s_k does not belong to the S union, then it does not belong to the prototype M.*

Proof: If $s_k \notin S$, in such case $s_k \notin I_i$, and for it $u_{ik} = 0$, verifying itself as $\lambda_k = 0$. For the hypotheses X and X' it stands that:

$$f(X') - f(X) = \frac{\sqrt{|X|} f(X)}{\sqrt{|X| + 1}} - f(X) < 0 \tag{8}$$

It generates a decreasing of the objective function value, and therefore such symbols are never found in the prototype M for which this value should be maximum.

Theorem 2 *If a symbol s_k belongs to the intersection P, then it belongs to the prototype M.*

Proof: If $s_k \in P$, in such case $s_k \in I_i$, and for it $u_{ik} = 1$, verifying itself as:

$$\lambda_k = \sigma = \frac{1}{n_v} \sum_{i=1}^{n} \frac{v_i}{\sqrt{|I_i|}} \tag{9}$$

The relevance of a symbol that participates in P is the greatest possible designated σ. For the hypotheses X and X' it is given that:

$$f(X') - f(X) =$$
$$\frac{1}{n_v} \sum_{i=1}^{n} \frac{v_i}{\sqrt{|I_i|}} \left(\frac{|X \cap I_i| + 1}{\sqrt{|X| + 1}} - \frac{|X \cap I_i|}{\sqrt{|X|}} \right) > 0 \tag{10}$$

This result is positive due to: $|X \cap I_i| \leq |X|$. It generates a increasing of the objective function value, and therefore such symbols must be found in the prototype M for which this value should be maximum. From all the previous results, we may deduce the following general conclusion:

Corollary 1 *M does not include any symbol that is not contained in S and therefore $M \subseteq S$, and in the same way, M includes all the symbols that are contained in P and therefore $P \subseteq M$. Thus, the M prototype contains the intersection P and is contained in the union S, that is: $P \subseteq M \subseteq S$.*

Given the graph $G =< N, L >$ compound of a set of nodes N and links L, each node is made up of the symbols contained in P and a combination of the contained in R, so the items as well as the closure sets are included in the N. The number of graph nodes is given by: $n_G = 2^h$. The set of nodes $N = \{N_0, \cdots, N_{2^h-1}\}$ is made up of all the combinations of possible symbols. It is verified that: $N_0 = P$ and $N_{2^h-1} = S$. Each link is determined by the inclusion or union with a symbol belonging to R. In the set of links, $L = \{l_{ij}\}$, l_{ij} represents a directed link between the node N_i and N_j, so if $l_{ij} = s_k$ then $N_j = N_i \cup \{s_k\}$. Each hypothesis of the solution corresponds to a N_i node, having associated a value of objetive function $f(N_i)$. The process of finding the solution to the problem becomes a search for the graph node which has the maximum value in the objective function. Given that the number of nodes in the graph is 2^h, this problem seems **NP** class.

The node N_b has a distance $\Delta(N_b, N_a)$ from another node N_a, if it is verified that: $N_a \subset N_b$ and $|N_b| = |N_a| + \Delta(N_b, N_a)$. The objective function value of N_b can be expressed recursively from the value of that of N_a in the following way, where: $\sum_a^b \lambda$ is the sum of the relevance factor of all the symbols of difference between both nodes.

$$f(N_b) = \frac{\sqrt{|N_a|} f(N_a) + \sum_a^b \lambda}{\sqrt{|N_a| + \Delta(N_b, N_a)}} \tag{11}$$

Given that all the nodes can be derived from the one associated with P, the objective function of a N_i node can be expressed as:

$$f(N_i) = \frac{\sqrt{|P|} f(P) + \sum_P^{N_i} \lambda}{\sqrt{|P| + \Delta(P, N_i)}} = \frac{|P|\sigma + \sum_P^{N_i} \lambda}{\sqrt{|P| + \Delta(P, N_i)}} \tag{12}$$

Let $\{\xi_1, \ldots, \xi_h\}$ the set of maximal relevance factors obtained by sorting the relevance factor, λ_i, in such way that: $\xi_1 = \max(\lambda_i)$ and $\xi_h = \min(\lambda_i)$. The relevant symbols $\{w_1, \ldots, w_h\}$ are the results of sorting in R the symbols based on the relevance factor in decreasing order. Let, η_i, be the factor of maximum cumulative relevance:

$$\eta_0 = 0 \tag{13}$$

$$\eta_i = \eta_{i-1} + \xi_i = \xi_1 + \cdots + \xi_i \tag{14}$$

Theorem 3 *The highest value of the objective function from among all the nodes that have a distance $i \geq 0$ from the node N_0 is given as:*

$$f_i = \frac{|P|\sigma + \eta_i}{\sqrt{|P| + i}} \tag{15}$$

Table 1. False Positive and True Negative samples. All errors are contained in `class3` and `class7` which are the most spread classes.

Sample	Own Class	Most Similar Class
seasnake	class3	class4
slug	class7	class6
tortoise	class3	class2
worm	class7	class6

Proof: The number of nodes found at a distance i of N_0 is $\binom{h}{i}$. For all of them the denominator of the objective function is the same, obtaining the highest possible value choosing that which have the highest numerator value, that is, the sum of values of maximum relevance. A recursive expresion can be used also to compute f_i:

$$f_0 = \sigma\sqrt{|P|} \qquad C_0 = |P| \tag{16}$$

$$f_{i+1} = \frac{\sqrt{C_i} f_i + \xi_{i+1}}{\sqrt{C_{i+1}}} \qquad C_{i+1} = C_i + 1 \tag{17}$$

Theorem 4 *The highest value of the objective function of the nodes contained in G is contained in the set $F = \{f_0, f_1, \cdots, f_h\}$.*

Proof: From the fact that the highest value of the graph found among the distances $i = 0, 1, 2, \ldots, h$. In such a way that $f_0 = f(P)$ and $f_h = f(S)$.

If f_T is the highest value of the contents in F, then the solution of the M prototype is given as:

$$M = \begin{cases} P & \text{if } T = 0 \\ P \cup \{w_1, \cdots, w_T\} & \text{otherwise} \end{cases} \tag{18}$$

It must be emphasized that the set of operations carried out for obtaining this solution is polinomially expressable on h,m and n, so the problem seems to be in the **P** class.

3 Computing Prototypes with dependent Symbols

In the previous section non-structured applications where all symbols are independent have been considered. Many real applications are well structured attribute based, in this case different values in an atttribute are dependent symbols which can be considered as mutually exclusive. Previous procedure can be extended to deal with dependent symbols. Let s_j and s_k be two boolean related symbols, that is: $s_j = \bar{s_k}$, eg. in UCI Zoo dataset [5] `[aquatic, no-aquatic]`. In this case is verified that:

$$u_{ij} + u_{ik} = 1 \tag{19}$$

So is verified that:

$$\lambda_j + \lambda_k = \sigma \tag{20}$$

In a general case, given a set of exclusive symbols $B = \{b_1, \ldots, b_l\}$ (eg. in UCI Soybean dataset attribute **`area-damage =[scatterd, low-areas, upper-areas, whole-field]`**) it is verified that:

$$\sum_{i=1}^{l} \lambda(b_i) = \sigma \tag{21}$$

To compute prototypes it must be included an additional exclusion rule of a symbol if any other in the exclusive set is already included. Let B a set of mutually exclusive symbols, if b_t is the symbol with maximal relevance in B, then b_t symbol *can* be include in the prototype M and the others in B *must* be excluded.

Some experimental attribute based datasets can include unknow samples. Some approaches can be used to deal with unknow samples [8]. One of them is based in to assign unknow data to all possible symbols in this attribute. In this case is verified that:

$$\sum_{i=1}^{l} \lambda(b_i) \geq \sigma \tag{22}$$

Other approach is based on not assigning unknow data to any possible symbols in this attribute. In this case is verified that:

$$\sum_{i=1}^{l} \lambda(b_i) \leq \sigma \tag{23}$$

4 An Example

A practical case of prototype computation is presented using the UCI Zoo dataset [5]. This is an attribute based dataset with a small number of attributes, and so it is not the natural domain to apply the described procedure. However it can be applied to this type of domains including the concept of dependent symbols. In this case a dataset with 16 attributes is used to code 101 samples grouped in 7 predefined classes. Each sample corresponds to an animal which is classified into a zoological class. All attributes, except `legs` attribute, are boolean. In this paper this last attribute is considered as boolean with `[legs, no-legs]` values. The goal in this problem is to find a prototype for each class which minimizes missclassifications. A simple approach is to compute the prototype which maximizes the average similarity with all samples contained in that class. Four

missclassifications in a dataset of 101 samples are obtained which are included in Table 1. Table 2 shows the average similarity $f(M)$ for each class, the relative relevance λ_k/σ and the attribute values of missclassified samples. Relative relevance can play an important role to introduce fuzzy related concept [2].

References

1. Bairoch A. and Apweiler R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL. *Nucleic Acids Res.*, (25):31–36, 1997.
2. Zadeh L. A. Fuzzy sets. *Information and Control*, 8:338–352, 1965.
3. Jain A.K. and Dubes R.C. *Algorithms for Clustering Data.* Printice Hall, 1988.
4. Moxon B. Defining data mining. *DBMS online*, August 1996. http://www.dbmsmag.com/9608d53.html.
5. Merz C.J. and Murphy P. UCI repository of machibe learning databases. Technical report, Departament of Information and Computer Science, University of California, Irvine, CA, 1996. http://www.ics.uci.edu/mlearn/MLRepository.html.
6. Mannila H. Methods and problems in data mining. In *Proc. Int. Conf. on Database Theory.* Springer-Verlag, January 1997.
7. Toivonen H. Discovery of frecuent patterns in large data collections. Technical Report Report A-1996-5, Dept. of Computer Science, University of Helssinki, Finlad, 1996.
8. Quinlan J.R. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.
9. Decker K.M. and Focardi S. Technology overview: A report on data mining. Technical Report CSCS TR-95-02, Swiss Scientific Computer Center, May 1995.
10. Guigó R. and Temple F.S. Inferring correlation between database queries: Analysis of protein sequence patterns. *IEEE PAMI*, 25(10):1030–1041, 1988.
11. Duda R.O., , and Hart P. *Pattern Classification and Scene Analysis.* Wiley and Sons, 1973.
12. Altschul S.F., Gish W., Miller W., Myers E.W., and Lipman D.J. Basic local alignment search tool. *J. Mol. Biol.*, (215):403–410, 1990.
13. Fayyad U.M., Haussler D., and Stolorz P. KDD for science data analysis; issues and examples. In *Proc. Second Int. Conf. on Knowledge Discovery and Data Minig.* AAAI Press, August 1996.
14. Fayyad U.M., Piatetsjy-Shapiro G., and Smyth P. *Advances in Knowledge Discovery and Data Mining.* AAAI/MIT Press, 1996.

Table 2. Mean similarity of each class, relative relevance λ_k/σ of symbols included in each class and missclassified items [`seasnake, slug, tortoise, worm`]. For all classes is verified that $\sigma = 0.25$. An contradictory case is realted to `tortoise` sample, which being included in the `class3`, is most similar with `class2` having the `airbone` attribute. This can not be considered as a error because `tortoise` have 12 matched attributes with `class3` and 13 with `class2`, and not all animals in `class2` have `airbone` attribute as relative relevance shows.

attribute	class1	class2	class3	class4	class5	class6	class7	seas.	slug	tort.	worm
$f(M)$	0.907	0.912	0.875	0.951	0.953	0.922	0.887				
airborne		0.80				0.75					
aquatic				1.00	1.00		0.60	1			
backbone	1.00	1.00	1.00	1.00	1.00			1		1	
breathes	1.00	1.00	0.80		1.00	1.00			1	1	1
catsize	0.78									1	
domestic											
eggs		1.00	0.80	1.00	1.00	1.00	0.90		1	1	1
feathers		1.00									
fins				1.00							
hair	0.95					0.50					
legs	0.92	1.00			1.00	1.00	0.60			1	
milk	1.00										
predator	0.54		0.80	0.69	0.75		0.80	1			
tail	0.85	1.00	1.00	1.00				1		1	
toothed	0.98		0.80	1.00	1.00			1			
venomous								1			
no-airborne	0.95		1.00	1.00	1.00		1.00	1	1	1	1
no-aquatic	0.85	0.70	0.80			1.00			1	1	1
no-backbone						1.00	1.00		1		1
no-breathes				1.00			0.70	1			
no-catsize		0.70	0.80	0.69	1.00	1.00	0.90	1	1		1
no-domestic	0.80	0.84	1.00	0.92	1.00	0.88	1.00	1	1	1	1
no-eggs	0.98							1			
no-feathers	1.00		1.00	1.00	1.00	1.00	1.00	1	1	1	1
no-fins	0.90	1.00	1.00		1.00	1.00	1.00	1	1	1	1
no-hair		1.00	1.00	1.00	1.00		1.00	1	1	1	1
no-legs			0.60	1.00				1	1		1
no-milk		1.00	1.00	1.00	1.00	1.00	1.00	1	1	1	1
no-predator		0.55				0.88			1	1	1
no-tail					0.75	1.00	0.90		1		1
no-toothed		1.00				1.00	1.00		1	1	1
no-venomous	1.00	1.00	0.60	0.92	0.75	0.75	0.80		1	1	1