Metric Categorization Relations Based on Support System Analysis

Krassimira Ivanova,
Ilia Mitov,
Krassimir Markov,
Peter Stanchev

Institute of Mathematics and Informatics,
Sofia, Bulgaria
e-mail:kivanova@math.bas.bg

Koen Vanhoof

Hasselt University

Hasselt, Belgium
e-mail:koen.vanhoof@uhasselt.be

Levon Aslanyan,
Hasmik Sahakyan

Institute for Informatics and Automation Problems
Yerevan, Armenia
e-mail:hasmik@ipia.sci.am

ABSTRACT

Theoretical and practical aspects of classification systems and classification learning are considered. Analysis of subject area learning sets and analysis of classification schemes raises a number of nonstandard questions, such as relations between categorization/metadata and logic-combinatorial structuring/clustering of the descriptive part of the input table. Several such questions are are treated in this paper. Software environment of testing and evaluation is the extended system PaGaNe, completed by research methodology of well known logic-combinatorial scheme of pattern recognition.

Keywords

Pattern Recognition, Metadata, Classification, Categorization,

1. INTRODUCTION

The results presented in this work concern the area of classification machine learning systems, for instance [1].

Let us consider a given set of objects described by a set of primary *measurable features* and suppose that this set is *classified* by a number of viewpoints – complementary set of classifiers. All this information is given in a form of a table which consists of two column parts respectively – descriptive or regular part, and metadata part. In majority cases of analysis of such categorization, models follow the simplest case – considering a set of non intersecting classes with one type of classification. Problems arise in this case consist in *classification learning*, and *classification simplification*. In case of complementary classifiers some problems could be mentioned:

(I1) *understanding and evaluating* reasonable requirements for the primary object characterization tools, which are sufficient to validate the characterization given by metadata.

(I2) *finding* logic-combinatorial meaning in feature area that takes informative burden of *describing classes and intersections* given in the metadata area.

Concepts, already known in pattern recognition area, such as support systems, parameterized distances [2] and logic separation [3] will be applied to maintain the novel specific questions of categorization, that is – relations between categorization/metadata and logic-combinatorial structuring/clustering of the descriptive part of the input table.

Section 2 contains basic notions. Possible preprocessing tools are considered in Section 3, containing also the modeling approach. Section 4 describes the programming realization. A small example, which shows the results of the realization of described ideas, is shown in Section 5.

2. BASIC NOTIONS

We assume that objects (descriptive part of input table) are given by *relational forms*, which are tuples of estimated values of groups of properties/attributes on objects. Each attribute has its name and the related domain of possible values; ordinary these are logic, alphanumeric, ranked and numerical attributes. Logic attributes might be assigned by values “true”, ”false”, alphanumeric ones take values from a finite set of possible strings (codes and names) of an alphabet, and numeric domains are machine presented numbers which might particularly act together with an order relation or a metric/distance measure. *Relation* consists of records, which are assignments of attributes by the elements of related domains.

The set of records is partitioned into disjoint (sometime intersecting) classes. A *class* is a subset of the records set consisting of all objects that satisfy the *class condition*. The *classification problem* is to classify new objects, i.e. to construct decision rules that describe objects of each class. The *decision rule* is an operator making a decision about the classification of objects. Classification description might be derived to a *simplified* form which is search effectiveness. The metadata part of description consists of tuples of alphanumeric values that correspond to the concepts in the chosen classifier relevant to the column.

3. PREPROCESSING

Before defining the categorization-metric relation analysis model let us consider several possible first approach/simplification actions to be applied on parts of the descriptive table.

(P1) Identification of groups of features, which give tight/similar descriptions of the objects given in descriptive part, when each feature in the group may play the same role in describing the metadata classes and concepts. It can be done by *cluster analysis by columns* of regular area of the table.

(P2) Similarly *cluster analysis by rows* of regular area is applied. The result concerned sensitively with the considered measure of similarity. The best suitable measure gives clusters highly correlated with classes defined in the metadata area by an individual classifier. To describe the whole complexity of classes by sets of many classifiers and their intersections additional means of similarity are necessary to be developed.

(P3) Metadata part of the descriptive table contains a number of individual classifiers. Application area may apply restrictions on validity of sets and subsets of classifiers. Therefore *consistency structure* of classifiers and classes has to be introduced. Coding the subsets of classifiers by vertices of k dimensional unit cube (where k is the number of features) we form the consistency Boolean function. Any subset of the consistency set of classifiers is also a consistency set (an introduced hypotheses), which implies that the consistency Boolean function is monotone. Other structural properties might appear/applied such as *consistency of intersections***.** Consistency structure is the area where the primary object characterization is to be evaluated *describing classes and intersection* given in the metadata area by (I2).

LOGIC-COMBINATORIAL STRUCTURES OF RECOGNITION

We bring several concepts that were developed in pattern recognition area.

Support system

A set of support systems Ω is defined as a collection of subsets of the set $\{1,2,\ldots,n\}$ where n is the number of features. Let $\omega = \{i_1, i_2, \ldots i_k\}$ be a set from Ω. We call $\tilde{\omega} = (\alpha_1 \ldots \alpha_n)$ the characteristic vector of Ω, when $\alpha_{i_1} = \ldots \alpha_{i_k} = 1$, and the rest of coordinates equals 0. Let's denote by $\tilde{\omega}S$ the ω-part of the description of object S (composed by coordinates $i_1, i_2, \ldots i_k$). Support system is the unit used in comparison of a pair of object descriptions. This is when a set of distances each by a member of Ω is defined. The total distance may be determined as the maximum of these sub-distances, the average of these values or in some other similar way. The application counterpart is that a set of features – not smaller and not larger than a support system is very effective in describing a particular classification.

PARAMETRIZED DISTANCES

Let the descriptive part of initial table consists of the set of objects - $I_0 = \{S_1, S_2, \ldots S_m\}$. Weights for each object S_i are defined, $(\gamma_1(S_i), \ldots \gamma_q(S_i)) = \tilde{\gamma}(S_i)$, and weights $(p_1(\tilde{\omega}), \ldots p_r(\tilde{\omega})) = \tilde{p}(\tilde{\omega})$ for each support system vector $\tilde{\omega}$ are given respectively. Then

$$\Gamma_{\tilde{\omega}}(S, S_i) = f(B_{\tilde{\omega}}(S, S_i), \tilde{\gamma}(S_i), \tilde{p}(\tilde{\omega}))$$

is called the estimate of the object S by S_i and $\tilde{\omega}$. Defining $B_{\tilde{\omega}}(S, S_i)$, the similarity fragment of S and S_i, we also use the distance thresholds τ_j of different features. Distance measures received in this way contain a large number of parameters, and this rich set of distances rise optimization problems in different stages of classification, helping to justify automatically the most effective similarity measures.

LOGIC SEPARATION

The Logic Separation model is based on the implementation of several logically expressed suppositions (constraints, hypothesis) above the elements of the training set. These are some formalisms or additional properties of classification, expressed in terms of Boolean functions and especially – of the Reduced Disjunctive Normal Form (RDNF). Boolean functions appear when considering sets of logical variables (binary properties) $x_1, x_2, \cdots, x_n$, and in case of two classes: K_1 and K_2. Let $\beta \in K_1$, $\gamma \in K_2$ and let α is an unknown object in the sense of classification. We say that γ is separated by the information of β for α if $\beta \oplus \alpha \leq \gamma \oplus \alpha$ where $\oplus$ the bit-vice mod2 summation is (exclusive OR). In simple words this means that the information difference of γ and α is "larger" than of β and α. The first includes directly and completely the second. As a consequence of this assumption we get, that the Reduced Disjunctive Normal Forms of the pair of complementary partially defined Boolean functions describe the complete structure of information enlargement, coming from the training set. Complex categorization given in metadata, transfers consideration from n-cube to the multidimensional multivalue grid, but the principle of class separation and compactness stay effective descriptive elements of classifiers.

THE MODEL OF CATEGORIZATION

We address the problem of understanding and modeling the realistic interconnections between objective categorization given by a set of classifiers and initial object descriptions given by sets of features and native relations of these descriptions.

The overall model is based on several hypotheses which appear in application area analysis. We differentiate three cases – one or several *metrics* (similarity measures) are given as the *application area properties*

(e.g. medical doctor may use in a particular diagnosis several sets of well known comparisons and distances), and they have to be treated as de facto descriptors of the categorization; a large parameterized set of distances is described over the features sets and several *formal optimization* functional are used to estimate the *correspondence of distances to the categorization* (e.g. similarity of categorization tree by metadata and the hierarchical cluster dendrogram, summary distance inside the classes vs. sum of distances between the classes); and finally when several *hypotheses* were found *in terms of* application area *categorization* structures, which helps to find the best distance in the parameterized family of distances. Let outline some example hypotheses.

(H1) *Parameterized smoothness*. Parameterized distances describe in general the classes (their compactness) and interclass separation and boarders (isoperimetry). Those distances are preferable, which give maximal compactness and minimal boarders (smoothness). One of the research goals is *to optimize* the set of all distance parameters being able to give an interpretation to the values found. This is done by modifying the threshold applied, or by involving more suppositions above the set of support systems as follows.

(H2) *Logic join/separation of support systems*. Consider set of all support systems weighted by some compatibility measure to the categorization information. Applying some threshold we get Boolean function that equals 1 on the most attractive support systems. If α and β are two such systems, then all elements, included in interval $I(\alpha,\delta)$ (in terms of unit cube) will be supportive. Given the weights of support systems we may determine the maximal intervals of support systems and apply them to analyze the information with best categorization-metrics relations. It is easy to see that the formalism, describing this structural relation is the reduced disjunctive normal form of Boolean functions. This is the case of a set of nonintersecting intervals.

(H3) *Convexity*. If two intervals considered above are intersecting (intervals correspond to different support systems), then consideration of all pair of vertices – each from one of these intervals and applying (H2) we receive a more general interval, which also will consist of all valid support systems. After a recursion of such steps we will come to (H2) with larger intervals or will finish by the total n-cube as the sign of "equality" of all support systems.

4. PROGRAM REALIZATION

For implementing of presented idea we propose an extension of classical classification methods with the purpose of using of metadata for automatic concept identifying of the founded regularities by the system. We have used as a ground a part of already realized classification algorithm in the experimental intelligent system PaGaNe [4].

The standard classification algorithm in PaGaNe uses feature vectors, which consists of (unique) name of the object, name of the class the given object belongs to, as well as a set of values of attributes that characterize the object and follows the next steps:

- automatically classifies the objects of the training set using hierarchies of information spaces;
- analyses the characteristics of classes with purpose to find combinations of values of characteristics, which are representative for the corresponded classes;
- creates "control" nodes, which vectors contain only the matching values of two or more objects;
- checks the training set for consistency;
- provides a frequency analysis of the values of the objects features in order to reduce the search only in classes, which are real candidates to be possible answer;
- founds representative exceptions in the classes to be used for direct recognition.

The recognition is based on reduced search in the multi-dimensional information space hierarchies.

The enhanced algorithm uses feature vectors with the different structure – the values of preliminary pointed class are equal (all vectors belong to the one general class, which represents the examined area as a whole). Beside of this the vectors contain a second part of values of metadata domains.

At the first stage the learning set is processed by the standard classification algorithm of PaGaNe. As a result we receive a set of control nodes that define specific frequently occurred combinations of attributes. Each of control nodes is connected with the primary objects of the learning set, which were participated in its creation. This way it is known how many and which primary objects contain this combination of attributes. The next step consists of traversing of all metadata positions and finding for each value of these positions one or several control nodes that correspond to this value. The control nodes, connected with examined metadata value, are additionally processed in order to throw out some control nodes that are comprised in other control nodes in the group (each vector, which primary vector includes in the primary vector of another exemplar of examined control nodes, is excluded).

For every metadata value (that define some concept) can be found zero, one or more corresponded control nodes. The reason that corresponded control nodes not exist usually lays in the fact that chosen primary attributes are not enough to correctly define this concept.

If metadata value is connected only with one control node – we can assume that this is the exact name of this control node. The content of this concept is represented as a conjunction of significant values of attributes, contained in corresponded control node. Of course, here also exists risk that primary attributes not represent correctly the examined area (but this is the problem of classification in general).

If the value is connected with more control nodes – it is represented as a disjunction of conjunctions of significant values of attributes, contained in connected control nodes.

5. EXAMPLE

As an example we take a simple database called ZOO from UCI Machine Learning Repository [5], which describes animals using 17 categorical attributes. This database we have expanded manually with three columns of metadata, containing information for the animal respectively:

- in which "Phylum" it belongs to ("Chordata", "Mollusca", "Arthropoda", etc.);
- is "Predator" or "Pray";
- in which "Class" it belongs to ("Mammalia", "Fish", "Aves", "Insecta", etc.).

As a source for additional information we use Encyclopedia of Life [6]. The reason to choice this example was the fact that almost everybody has learned zoology and can easy understand the meaning of the concepts and the attributes.

The system uses this expanded dataset as a learning set and the result is shown on table 1.

Table 1. The result of the working of the enhanced algorithm of PaGaNe (fragment).

```
Phylum : Chordata
    <def>
        backbone: yes
    </def>

Phylum : Mollusca
    <def>
        hair: no
        feathers: no
        eggs: yes
        milk: no
        airborne: no
        backbone: no
        venomous: no
        fins: no
        tail: yes
        domestic: no
    </def>

Phylum : Arthropoda
    <def>
        feathers: no
        milk: no
        toothed: yes
        backbone: no
        fins: no
        catsize: no
    </def>
```

```
Class : Mammalia
    <def>
        feathers: no
        milk: yes
        backbone: yes
        breathes: yes
        venomous: no
    </def>

Class : Fish
    <def>
        hair: no
        feathers: no
        eggs: yes
        milk: no
        airborne: no
        aquatic: yes
        toothed: yes
        backbone: yes
        breathes: no
        fins: yes
        legs: 0
        tail: yes
    </def>

Class : Aves
    <def>
        hair: no
        feathers: yes
        eggs: yes
        milk: no
        backbone: yes
        breathes: yes
        venomous: no
        fins: no
        legs: 2
        tail: yes
    </def>
```

The results of identifying the concepts of "Phylum" are presented in the left column, as well as a part of definition of "Class" are presented in the right column.

It is clearly seen that for instance for definition of Phylum "Chordata" we need only one attribute "backbone" with positive value (which is human definition of this concept) as well as for definitions of "Mollusca" and "Atrhropoda" another attributes become important and the fact that theirs identification not cover human definitions shows that the observed dataset does not contain the attributes, which we use as most specific for these classes (for "Mollusca" – a mantle and nervous system; for "Atrhropoda" – exoskeleton, a segmented body, and appendages). In the same time the system finds regularities to identify these concepts using values of another attributes.

6. CONCLUSION

Application of some concepts, already known in pattern recognition area, such as support systems, parameterized distances and logic separation to solve some novel specific problems of categorization such as discovering the relations between descriptive part and metadata values of the input table in order to use the metadata for automatic concept identifying of the founded regularities by the system, was discussed in the paper.

The concept analysis system PaGaNe is applied to treat the metric-categorization relations It has shown better results in the field of identification scheme in comparison to [1]. Categorization modeling raised questions that brought to the logic combinatorial recognition area.

Acknowledgements

This work is partially financed by Bulgarian National Science Fund under the project **D 002-308 / 19.12.2008** "Automated Metadata Generating for e-Documents Specifications and Standards".

REFERENCES

1. http://www.cs.waikato.ac.nz/ml/weka/, visited on 01.04.2009.

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

3. L. Aslanyan, J. Castellanos, Logic based Pattern Recognition - Ontology content (1), Int. Journal “Information Technologies and Knowledge”, v.1, 2007.

4. I. Mitov, Kr. Ivanova, Kr. Markov, V. Velychko, K. Vanhoof, and P. Stanchev. "PaGaNe" – A Classification Machine Learning System Based on the Multidimensional Numbered Information Spaces. "Intelligent Systems and Knowledge Engineering", 27-28.11.2009, Hasselt, Belgium (in appear).

5. A. Asuncion, D. Newman. UCI Machine Learning Repository. University of California, Irvine, CA, School of Information and Computer Science, http://archive.ics.uci.edu/ml/, visited on 01.04.09.

6. Encyclopedia of Life (EOL), www.eol.org, visited on 05.05.2009.