Metric Generalization and Modification of Classification Algorithms Based on Formal Concept Analysis

Evgeny Kolmakov

Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Russia

Abstract. FCA-based classifiers can deal with nonbinary data representation in different ways: use it directly or binarize it. Those algorithms that binarize data use metric information from the initial feature space only as a result of scaling (feature binarization procedure). Metric approach in this area allows one significantly reducing classification refusals number and provides additional information which can be used for classifier training. In this paper we propose an approach which generalizes some of existing FCA classification methods and allows one to modify them. Unlike other algorithms, the proposed classifier model uses initial metric information together with order object-attribute dependencies.

Keywords: classification, pattern recognition, formal concept analysis

1 Introduction

Formal concept analysis (FCA) is a branch of applied lattice theory allowing one to formalize some machine learning models. It provides tools to solve various tasks in many domains of computer science, such as knowledge representation and management, data mining, including classification and clustering. There are many FCA-based classification algorithms known [6]. One of the particular features of FCA methods is that object $x \in \mathbb{X}$ is being described using binary attributes. However, in many cases attributes can be, e.g., real numbers, graphs, etc. There are classification methods using nonbinary representation directly, e.g., see these works on pattern structures [4], [5], but many classifiers use it only after scaling procedure. The scaling procedure is the transformation of the initial feature space $\mathcal{F}$ into the Boolean cube B^n. It leads to the significant loss of the metric information provided by $\mathcal{F}$ space. In this paper we propose generalizations and modifications of several FCA-based classifiers, which use scaling procedure, by introducing new classifier model on the basis of class estimates. It generalizes straight hypotheses-based algorithm [1] and both of GALOIS classification procedures [3]. We also define the pseudometric on arbitrary finite lattice, which is based on the ideas from Rulearner rules induction algorithm [2] and so has intelligible interpretation in terms of formal concepts and concept lattice.

In what follows we keep to standard lattice theory and FCA definitions. Therefore here we briefly describe some basic definitions, classifiers and introduce

the notation which is used further. Let G and M be an arbitrary sets called the *set of objects* and the *set of attributes* respectively and $I \subseteq G \times M$ be a binary relation. The triple $\mathbb{K} = (G, M, I)$ is called a *formal context*. The following $(\cdot)'$ mappings define a *Galois connection* between 2^G and 2^M sets partially ordered by set-theoretic inclusion:

$A' = \{m \in M \mid gIm \text{ for all } g \in A\}$, $B' = \{g \in G \mid gIm \text{ for all } m \in B\}$.

A pair (A, B), such that $A \subseteq G, B \subseteq M$ and $A' = B, B' = A$ is called a *formal concept* of $\mathbb{K}$ with *formal extent* A and *formal intent* B. For object $g \in G$ we write g' instead of $\{g\}'$. Define the "projection" mappings $ext : (A, B) \mapsto A$ and $int : (A, B) \mapsto B$. Formal concepts of a given context $\mathbb{K}$ form a complete lattice denoted by $\mathfrak{B}(\mathbb{K})$. It is called the *concept lattice* of a context $\mathbb{K}$. Let $\langle L, \wedge, \vee \rangle$ be a lattice and $x \in L$. By x^∇ (x^Δ) we denote the order ideal (filter) generated by x. By $At(L), J(L)$ and $M(L)$ we denote the set of all atoms, join-irreducible and meet-irreducible elements of L respectively. The function $f\colon L \to \mathbb{R}$ is called *supermodular* if $f(x) + f(y) \leqslant f(x \vee y) + f(x \wedge y)$ for all $x, y \in L$.

A concept C is called *consistent* if all objects in $ext(C)$ belong to the same class. Both GALOIS classification procedures are described in [3]. GALOIS(1) calculates the similarity $\Gamma_C(x)$ between an object x and each consistent concept C, then x is assigned to the class corresponding to C with the highest value of $\Gamma_C(x)$. GALOIS(2) finds all consistent concepts C satisfying $int(C) \subseteq x'$, then x is assigned to the most numerous class in the previous set.

Let $\mathbb{K} = (G, M, I)$ be a context and $w \notin M$ be a target attribute. The input data for classification may be described by three contexts w.r.t. w: the *positive context* $\mathbb{K}_+ = (G_+, M, I_+)$, the *negative context* $\mathbb{K}_- = (G_-, M, I_-)$ and the *undefined context* $K_\tau = (G_\tau, M, I_\tau)$. G_-, G_+ and G_τ are sets of positive, negative and undefined objects respectively. $I_\epsilon \subseteq G_\epsilon \times M$, where $\epsilon \in \{-, +, \tau\}$ are binary relations that define structural attributes. Galois operators in these contexts are denoted by $(\cdot)^+, (\cdot)^-$, and $(\cdot)^\tau$ respectively. A formal concept of a positive context is called a *positive concept*. Negative and undefined concepts are defined similarly. If the intent B_+ of a positive concept (A_+, B_+) is not contained in the intent g^- of any negative example $g \in G_-$, then it is called a *positive hypothesis* with respect to the property w. A positive intent B_+ is called *falsified* if $B_+ \subseteq g^-$ for some negative example $g \in G_-$. Negative hypotheses are defined similarly. By "hypotheses-based classifier" we mean the classification procedure from [1], which can be described as follows. If unclassified object $g \in G_\tau$ contains a positive but no negative hypotheses, it is classified positively, similar for negative. If g does not contain any positive or negative hypothesis (insufficient data) or contains both positive and negative hypotheses (inconsistent data), then no classification happens.

2 Generalization and Modification of Algorithms

The common drawback of the FCA-based classifiers using binary features after scaling is that they forget the initial feature space metric structure. The main idea of this paper is to use this metric information together with order-

theoretic relations between objects and attributes provided by a concept lattice. It is important that $\mathcal{F}$ and B^n spaces with additional structures (metric and formal context) are being used at the same time, providing more possibilities for classifier training methods.

2.1 Metric estimates

Denote by $\mathcal{H}_+$ and $\mathcal{H}_-$ the sets of concepts constructed from a training set, which intents are positive and negative hypotheses respectively. We assume that $(\mathcal{F}, \rho)$ is a metric space, and let $S(x, A)$ be the similarity measure (based on the metric from $\mathcal{F}$, see examples in Section 3) between an object x and a set of objects A. Let us define the estimates for positive and negative classes:

$$\Gamma_+(x) = \sum_{C \in \mathcal{H}_+} I(x, C) S(x, ext(C)), \quad \Gamma_-(x) = \sum_{C \in \mathcal{H}_-} I(x, C) S(x, ext(C)),$$

where $I(x, C) = [int(C) \subseteq x']$ and $[\cdot]$ is the indicator function. Then the classifier will have the following form: $a(x) = \text{sign}\, \Gamma(x) = \text{sign}(\Gamma_+(x) - \Gamma_-(x))$.

Proposition 1. *If hypotheses-based classifier correctly predicts class label for an object then* $a(x) = \text{sign}\, \Gamma(x)$ *does the same.*

In comparison with hypotheses-based classifier the number of classification refusals is reduced, but the total error rate can increase.

2.2 Analogy with algorithms based on estimate calculations

To calculate the estimates in the method above we use positive and negative hypotheses sets, i.e. special subsets of concept lattice. Such calculation of estimates can be generalized to an arbitrary concept lattice subsets somehow characterizing individual classes $y \in Y$.

Let $\mathcal{C}$ be the set of concepts which we call *the support concepts system.* Suppose that each concept from $\mathcal{C}$ characterizes only one class $y \in Y$, that is $\mathcal{C} = \bigsqcup_{y \in Y} \mathcal{C}_y$, where Y is the set of classes. Then define the estimate of object x for class y as follows:

$$\Gamma_y(x) = \sum_{C \in \mathcal{C}_y} S(x, C).$$

The classifier will have the following form: $a(x) = \arg\max_{y \in Y} \Gamma_y(x)$. The estimates of this type are similar to the estimates used in estimate calculations methods [7] and the sets $\mathcal{C}_y$ are the support sets analogues.

Consider specific examples of support concepts system $\mathcal{C}$, similarity measure $S(x, C)$ and analyze corresponding classifiers:

1. $\mathcal{C} = \mathcal{H}_+ \bigsqcup \mathcal{H}_-$ are positive and negative hypotheses sets,
   $S(x, C) = [int(C) \subseteq x'] \hat{S}(x, ext(C))$, where $\hat{S}(x, ext(C))$ is the given similarity measure. The corresponding classifier was described above.
2. $\mathcal{C} = \bigsqcup_{y \in Y} \mathcal{C}_y$ is the consistent concepts set.
   If $S(x, C) = |(M \setminus int(C)) \cup x'|$ we get modified GALOIS(1) algorithm.
   If $S(x, C) = [int(C) \subseteq x']$ we get GALOIS(2) algorithm.

2.3 Analogy with metric classifiers

Let $\mathcal{C} = \{C_1, \ldots, C_n\}$ be the support concepts system. Suppose that there is the distance measure ρ in $\mathcal{F}$ space. Sort $\mathcal{C}$ in increasing order w.r.t. the values of the distance $\rho(x, C_i)$ between object x and concepts C_i:

$$\rho(x, C_x^{(1)}) \leqslant \rho(x, C_x^{(2)}) \leqslant \cdots \leqslant \rho(x, C_x^{(n)}),$$

where $C_x^{(i)}$ is the i-th neighbour of x among $\mathcal{C}$, $y_x^{(i)}$ is the class, characterized by $C_x^{(i)}$ concept. Define the estimate of object x for class y:

$$\Gamma_y(x) = \sum_{i=1}^{n} w_i(x)[y_x^{(i)} = y],$$

$w_i(x)$ is x i-th neighbour weight (positive function non-increasing w.r.t. i).

The defined estimates are completely analogous to the metric classifiers estimates, except that the neighbours here are not objects but support concepts.

Thus choosing the suitable weights $w_i(x)$ we get analogs of all known metric classifiers (kNN, Parzen window, potential functions and others), but in terms of concepts. For example:

- $w_i(x) = [i \leqslant k]$ is k nearest neighbours method;
- $w_i(x) = [i \leqslant k]w_i$ is k weighted nearest neighbour method (w_i depends only on the neighbour number);
- $w_i(x) = K(\frac{\rho(x, C_x^{(i)})}{h(x)})$ is Parzen window method ($K(z)$ is non-increasing positive-valued function defined on $[0, 1]$, $h(x)$ is the window width).

All the proposed methods are the generalizations of the existing methods and can be used for their modifications. They use both metric information from $\mathcal{F}$ and object-attribute dependencies provided by concept lattice. This allows to reduce the number of classification refusals and error rate.

2.4 Pseudometric on the set of concepts

Another approach which uses the notion of similarity in FCA algorithms is to define a distance function on the set of concepts. In Rulearner algorithm ([2]) the most important characteristics of concept lattice element u were the value of the function $\text{cover}(u) = |J(L) \cap u^{\nabla}|$ and $M(L) \cap u^{\Delta}$ set. The comparison of lattice elements is performed on the basis of these characteristics. In the case of reduced context, this ties up with a fact, that every concept is characterized by its extent (distinct objects correspond to join-irreducible elements) and intent (distinct features correspond to meet-irreducible elements). Thus, $\text{cover}(u)$ corresponds to the number of objects from training set covered by the concept u, and $M(L) \cap u^{\Delta}$ corresponds to the attributes characterizing u. We use these observations to define the distance function on an arbitrary finite lattice. Due to the propositions dual to theorems 3.1 and 3.3 from [8], the following theorem holds.

Theorem 1. *Let $\langle L, \wedge, \vee\rangle$ be a lattice and $f\colon L \to \mathbb{R}$ is isotone and supermodular function, then $d_f(x,y) = f(x) + f(y) - 2f(x \wedge y)$ defines a pseudometric on this lattice.*

Consider arbitrary finite lattice $\langle L, \wedge, \vee\rangle$, non-empty subset $D \subseteq L$ and a function $f\colon L \to \mathbb{Z}_+$, defined as follows: $f(x) = |D(x)|$, where $D(x) = D \cap x^{\nabla}$.

Proposition 2. *The function $f(x)$ is isotone and supermodular.*

Proof. The isotone property of f follows from the following chain of implications:

$$x \leqslant y \Rightarrow x^{\nabla} \subseteq y^{\nabla} \Rightarrow D(x) \subseteq D(y) \Rightarrow f(x) = |D(x)| \leqslant |D(y)| = f(y).$$

To prove supermodularity consider the following:

$$f(x)+f(y) = |D(x)|+|D(y)| = |D(x)\cup D(y)|+|D(x)\cap D(y)| \leqslant f(x\vee y)+f(x\wedge y).$$

To prove the last inequality observe that $D(x) \cup D(y) \subseteq D(x \vee y)$ follows from the following inclusions:

$$x \leqslant x \vee y \Rightarrow D(x) \subseteq D(x \vee y), \quad y \leqslant x \vee y \Rightarrow D(y) \subseteq D(x \vee y).$$

The equality $D(x) \cap D(y) = D(x \wedge y)$ follows from $x^{\nabla} \cap y^{\nabla} = (x \wedge y)^{\nabla}$. □

Thus, according to the theorem above, the function $f(x)$ induces the pseudometric $d_f(x,y)$ on the lattice, defined by the following equality:

$$d_f(x,y) = f(x) + f(y) - 2f(x \wedge y).$$

The value of $d_f(x,y)$ has simple interpretation.

Proposition 3. *$d_f(x,y) = |D(x) \oplus D(y)|$, where $A \oplus B = (A \setminus B) \cup (B \setminus A)$.*

Proof. From the proof of the proposition above we conclude that the equality $D(x) \cap D(y) = D(x \wedge y)$ holds. Consider the chain of equalities:

$$\begin{aligned} f(x) + f(y) - 2f(x \wedge y) &= |D(x)| + |D(y)| - 2|D(x \wedge y)| = \\ &= |D(x)| + |D(y)| - 2|D(x) \cap D(y)| = \\ &= |D(x) \cup D(y)| + |D(x) \cap D(y)| - 2|D(x) \cap D(y)| = \\ &= |D(x) \cup D(y)| - |D(x) \cap D(y)| = |D(x) \oplus D(y)|. \end{aligned}$$

□

Corollary 1. *If $\langle L, \wedge, \vee\rangle$ is a finite Boolean algebra and D is the set of all atoms of L, then $d_f(x,y)$ is exactly the Hamming distance.*

In order to compare formal concepts it is reasonable to choose $D = J(L)$ or $D = At(L)$. In terms of this pseudometric two concepts are the closer, the less object concepts are covered by only one of them. Moreover, the cover(u) and $d_f(x,y)$ functions are tied: cover$(u) = d_f(u, \bigwedge L)$. One of the drawbacks of the defined distance measure is that the number of elements from D covered by $x \wedge y$ is not taken into account. In some cases it may lead to inadequate distance estimates.

Possible modifications:

1. Various normalizations to take the number of elements into account, e.g.:

$$d(x,y) = \frac{|D(x) \oplus D(y)|}{|D(x) \cup D(y)|}.$$

2. Weighting elements of D, e.g. let $D = J(L)$ and w_e be the proportion of the hypotheses covering $e = (g'', g')$. Then $d(x,y)$ will have the following form:

$$d(x,y) = \sum_{e \in D(x) \oplus D(y)} w_e.$$

The distance between concepts can be applied to modify the classification algorithms mentioned above. For example, let an object x be classified with hypotheses-based algorithm. Suppose there are two positive hypotheses H_1^+, H_2^+ and two negative hypotheses H_1^-, H_2^- for the classification of x. In this case the algorithm refuses to classify x. Suppose we know the concept distances $d(H_1^+, H_2^+), d(H_1^-, H_2^-)$ and also $d(H_1^+, H_2^+) \gg d(H_1^-, H_2^-)$. Then it is natural to classify x as positive, because the distant concepts (in terms of the proposed measure) are less "correlated" (since they cover many distinct object concepts), and hence their answers are more significant. Distance between concepts can also be used for reducing the size of concepts system (used by classifier, e.g. consistent concepts) in order to improve generalization ability of classifier, reduce the overfitting and remove concepts based on noisy data.

3 Experiments

In this section the experimental results are presented. The algorithms have been tested on two data sets taken from UCI Machine Learning Repository [9]: SPECT and SPECTF Heart Data Set (training set consists of 80 objects, testing set consists of 187 objects, 22 binary attributes in SPECT, 44 real-valued attributes in SPECTF) and Liver Disorders Data Set (training set consists of 150 objects, testing set consists of 195 objects, 6 real-valued attributes, 30 binary attributes (after scaling)). Tested algorithms: GALOIS(1, 2), Rulearner, straight hypotheses-based algorithm, modified GALOIS(1) (described by the second example in Section 2.2), modified hypotheses-based algorithm with metric estimates (described in Section 2.1 with different similarity functions). Euclidian metric $\rho(x,y)$ was used in $\mathcal{F}$ space in both experiments. Similarity function: $S(x,C) = K(\rho(x,C), a)$, where $K(r,a)$ and $\rho(x,C)$ are one of the following functions:

$$K_1(r,a) = \frac{1}{1+\exp(ar)}, \qquad K_2(r,a) = \frac{1}{r+a}.$$

$$\rho_1(x,C) = \inf_{c \in C} \rho(x,c), \quad \rho_2(x,C) = \frac{1}{|C|} \sum_{c \in C} \rho(x,c), \quad \rho_3(x,C) = \sup_{c \in C} \rho(x,c).$$

We introduce the following notation: ν_c is the proportion of classified objects, $\nu_r = 1 - \nu_c$ is the proportion of refused classifications, e_t is total error rate (including refusals), e_r is the error rate among classified objects.

Algorithm	ν_c	ν_r	e_t	e_r
GALOIS(1)	1	0	0.1604	0.1604
Modified GALOIS(1)	1	0	0.0856	0.0856
GALOIS(2)	1	0	0.0802	0.0802
Rulearner	0.7487	0.2513	0.2727	0.0286
Hypotheses-based	0.5936	0.4064	0.6150	0.1842
$K = K_1, a = 0.0125, \rho = \rho_1$	0.8021	0.1979	0.3155	0.1467
$K = K_1, a = 0.0125, \rho = \rho_2$	0.8021	0.1979	0.2888	0.1133
$K = K_1, a = 0.0125, \rho = \rho_3$	0.8021	0.1979	0.2834	0.1067
$K = K_1, a = 1, \rho = \rho_2$	0.7273	0.2727	0.3422	0.0956
$K = K_2, a = 1, \rho = \rho_1$	0.8021	0.1979	0.2941	0.1200
$K = K_2, a = 1, \rho = \rho_2$	0.8021	0.1979	0.3209	0.1533

Table 1. SPECT and SPECTF Heart Data Set. Experimental results.

Algorithm	ν_c	ν_r	e_t	e_r
GALOIS(1)	1	0	0.4605	0.4605
Modified GALOIS(1)	1	0	0.5590	0.5590
GALOIS(2)	1	0	0.4359	0.4359
Rulearner	0.9795	0.0205	0.4564	0.4450
Hypotheses-based	0.2923	0.7077	0.8256	0.4035
$K = K_1, a = 1, \rho = \rho_1$	0.8821	0.1179	0.5231	0.4593
$K = K_1, a = 0.01, \rho = \rho_2$	0.8974	0.1026	0.5436	0.4914
$K = K_1, a = 0.25, \rho = \rho_3$	0.8872	0.1128	0.5385	0.4798
$K = K_2, a = 200, \rho = \rho_1$	0.8974	0.1026	0.4769	0.4171
$K = K_2, a = 150, \rho = \rho_2$	0.8974	0.1026	0.4564	0.3943
$K = K_2, a = 150, \rho = \rho_3$	0.8974	0.1026	0.4667	0.4057

Table 2. Livers Disorder Data Set. Experimental results.

The aim of the experiments was to compare FCA classification methods and not to achieve low error rate in solving particular tasks. Hence we used simple scaling procedure: normalizing all attributes to $[0, 1]$ interval and then applying interval-based nominal scaling (the number of intervals was chosen to be 5). It explains high error rate of all classifiers in the second task. Individual scaling (e.g. scaling with floating-size intervals) for each task may significantly reduce error rate, but this work is not focused on this problem. From the results above we may conclude that for hypotheses-based algorithm modifications the number of refusals is substantially reduced together with total error rate e_t. Modified GALOIS(1) classifier improved GALOIS(1) on the first data set and disimproved it on the second. This may be due to the different nature of binary data description: in the first case 22 binary attributes were obtained from 44

real-valued using complex binarization procedure, while in the second one this procedure was very simple. The choice of $K(r, a)$ and $\rho(x, C)$ affects only e_t but not ν_r, hence their accurate selection may improve classification quality.

4 Conclusions

In this paper we have formally described and experimentally studied a new approach to classification which encompasses the usage both of metric information provided by the initial feature space and the order object-attribute dependencies. Also we have defined the pseudometric on arbitrary finite lattice, which has intelligeble interpretation in terms of concepts and hence can be used for comparing concepts in order to improve FCA classification methods. Further developments can be focused on studying of classifiers obtained from the proposed model by fixing the support concepts system $\mathcal{C}$ and the similarity measure $S(x, C)$ and on the possibilities of choosing such support concepts system that allows to construct only a part of concept lattice.

References

1. S.O. Kuznetsov: Complexity of Learning in Concept Lattices from Positive and Negative Examples. Discrete Applied Mathematics, 2004, No. 142(13), pp. 111-125.
2. M. Sahami: Learning classification Rules Using Lattices. N. Lavrac and S. Wrobel eds., pp. 343346, Proc ECML, Heraclion, Crete, Greece (April 1995).
3. C. Caprineto, G. Romano: GALOIS An order-theoretic approach to conceptual clustering. In proceedings of ICML93, pp. 3340, Amherst, USA (July 1993).
4. M. Kaytoue, S.O. Kuznetsov, A. Napoli, S. Duplessis, Mining gene expression data with pattern structures in formal concept analysis. Information Sciences, Volume 181, Issue 10, 15 May 2011, pp. 1989-2001, Information Science, 2011.
5. S.O. Kuznetsov, Scalable Knowledge Discovery in Complex Data with Pattern Structures. In: P. Maji, A. Ghosh, M.N. Murty, K. Ghosh, S.K. Pal, Eds., Proc. 5th International Conference Pattern Recognition and Machine Intelligence (PReMI'2013), Lecture Notes in Computer Science (Springer), Vol. 8251, pp. 30-41, 2013.
6. O. Prokasheva, A. Onishchenko, S. Gurov. Classification methods based on Formal Concept Analysis. FCAIR 2013 Formal Concept Analysis Meets Information Retrieval. Workshop co-located with the 35th European Conference on Information Retrieval (ECIR 2013). March 24, 2013, Moscow, Russia. National Research University Higher School of Economics, pp. 95-104. ISSN 1613-0073
7. Yu.I. Zhuravlev: An Algebraic Approach to Recognition and Classification Problems, in Problems of Cybernetics, Issue 33 (Nauka, Moscow, 1978), pp. 568 [In Russian].
8. Dan A. Simovici: Betweenness, Metrics and Entropies in Lattices. Proceedings of the 38th International Symposium on Multiple Valued Logic. 22-24 May, 2008, Dallas, TX, USA. IEEE Computer Society Washington, pp. 26-31. ISSN 0195-623X
9. Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.