Feature Construction and δ-Free Sets in 0/1 Samples

Nazha Selmaoui[1], Claire Leschi[2], Dominique Gay[1], and Jean-François Boulicaut[2]

[1] ERIM, University of New Caledonia
{selmaoui, gay}@univ-nc.nc
[2] INSA Lyon, LIRIS CNRS UMR 5205
{claire.leschi, jean-francois.boulicaut}@insa-lyon.fr

Abstract. Given the recent breakthrough in constraint-based mining of local patterns, we decided to investigate its impact on feature construction for classification tasks. We discuss preliminary results concerning the use of the so-called δ-free sets. Our guess is that their minimality might help to collect important features. Once these sets are computed, we propose to select the essential ones w.r.t. class separation and generalization as new features. Our experiments have given encouraging results.

1 Introduction

We would like to support difficult classification tasks (from, e.g., large noisy data) by designing well-founded processes for building new features and then using available techniques. This is challenging and our thesis is that the recent breakthrough in constraint-based mining of local patterns might provide some results. Considering the case of 0/1 data whose some attributes denote class values[1], many efficient techniques are now available for computing complete collections of patterns which satisfy user-defined constraints (e.g., minimal frequency, freeness, closeness). Our goal is not only to consider such patterns as features but also to be able to predict (part of) the classification behavior based on these pattern properties. In this paper, we discuss preliminary results concerning the so-called frequent δ-free sets in 0/1 samples. When $\delta = 0$, these sets have been studied as minimal generators for the popular (frequent) closed sets. Otherwise ($\delta > 0$), they provide a "near equivalence" perspective and they have been studied as an approximate condensed representation for frequent sets [1]. Furthermore, the minimality of δ-free sets has been exploited for class characterization (see, e.g., [2]) and non redundant association rule mining (see, e.g., [3]). Our guess is that this minimality, in the spirit of the MDL principle, might help to collect relevant features. This is suggested in [4] as a future direction of work, and we provide some results in that direction. Section 2 introduces δ-freeness and our feature construction process. Section 3 reports about classification tasks on both UCI data sets [5] and a real-world medical data set. Section 4 concludes.

[1] It is trivial to derive Boolean data from categorical data and discretization operators can be used to transform continuous attributes into Boolean ones.

N. Lavrač, L. Todorovski, and K.P. Jantke (Eds.): DS 2006, LNAI 4265, pp. 363–367, 2006.

2 Feature Construction by Using δ-Free Sets

Given a potentially large labeled 0/1 data set, our feature construction process consists in three main steps: (1) mining frequent δ-free sets associated to their δ-closure [1, 6] and select those whose δ-closure includes a class attribute; (2) further select the essential patterns w.r.t. some interestingness criteria; (3) encode the original samples in the new representation space defined by these descriptors.

Let $r = (T, I)$ a 0/1 data set where T is a set of objects and I a set of Boolean attributes. An itemset A is subset of I and we recall the definition of useful evaluation functions on itemsets. The frequency of A in r is defined as $freq(A, r) = |support(A, r)|$ where $support(A, r) = \{t \in T / A \in t\}$. Let γ be an integer, A is called γ-frequent if $freq(A, r) \geq \gamma$. The closure of A in r denoted $closure(A, r)$ is the largest superset of A with the same frequency. An itemset A is closed if $closure(A, r) = A$. Since [7], it is useful to formalize this by means of the same-closure equivalence relation. Two itemsets A and B are said equivalent in r ($A \sim_f B$) if $closure(A, r) = closure(B, r)$. Indeed, we have the following properties :

(i) $A \sim_f B \equiv freq(A, r) = freq(B, r)$;
(ii) Each equivalence class contains exactly one maximal (w.r.t. set inclusion) itemset which is a closed set, and it might contain several minimal (w.r.t. set inclusion) sets which are called 0-free sets or free sets for short.
(iii) If A and B are in the same equivalence class and $A \subseteq V \subseteq B$ then V is in the same equivalent class.

Definition 1 (δ-free itemsets and δ-closures). *Let δ be an integer. A is a δ-free itemset if $\forall S \subset A, |freq(S, r) - freq(A, r)| > \delta$. The δ-closure of an itemset A is defined as $closure_\delta(A) = \{X \in I / freq(A, r) - freq(A \cup \{X\}, r) \leq \delta\}$.*

The intuition is that the δ-closure of a set A is the superset X of A such that every added attribute is almost always true for the objects which satisfy the properties from A: at most δ false values are enabled. The computation of every frequent δ-free set (i.e. sets which are both frequent and δ-free) can be performed efficiently [6]. Given threshold values for γ (frequency) and δ (freeness), our implementation outputs each δ-free frequent itemset and its associated δ-closure. Notice that when $\delta = 0$, we collect all the free frequent itemsets and their corresponding closure, i.e., we compute a closed set based on the closure of one minimal generator. Since we are interested in classification, we also assume that some of the attributes denote the class values.

Interestingness measures are needed to select the new features among the δ-frees. Our first measure is based on homogeneity and concentration (HC) [8]. It has been proposed in a clustering framework where formal concepts (i.e., closed sets) were considered as possible bi-clusters and had to be as homogeneous as possible while involving "enough" objects. Our second measure is the well-known information gain ratio (GI). Selecting features among the frequent δ-free itemsets is crucially needed since 0/1 data might contain a huge number patterns which are relevant neither for class discrimination nor for generalization.

At first, the homogeneity and concentration measures were used to respectively maximize the intra-cluster similarity and to limit the overlapping of objects between clusters. Homogeneity is defined as:

$$Homogeneity(A, r) = \frac{|support(A, r)| \times |A|}{divergence(A) + (|support(A, r)| \times |A|)}$$

where $divergence(A) = \sum_{t, A \in t} |t - A|$. If an itemset is pure then its divergence is equal to 0, and its homogeneity is equal to 1. This measure enables to keep the itemsets having many attributes shared by many objects. The concentration is defined as:

$$Concentration(A, r) = \frac{1}{|support(A, r)|} \times \sum_{X \in t, \forall t, A \in t} \frac{1}{|X|}$$

Then, the interestingness measure of an itemset is defined as the average of its homogeneity and its concentration. The more the interestingness is close to 1, the more an itemset is considered as essential for classification purposes.

The filtering of new descriptors is performed in three steps. Once frequent δ-free sets (say A) and their δ-closures (say X) have been extracted, we first retain only the sets X which include a class attribute. The associated minimal generators, i.e., the frequent δ-free sets whose δ-closures involve a class attribute, are selected as potentially interesting new features. We further focus on their supporting sets of objects among the classes. Then, we retain only the patterns that are very frequent in one class and merely infrequent in the other ones. We formulate this condition through the function Gr_Rate defined as follows (C_i is a class attribute):

$$Gr\_Rate(A) = \frac{|support_{C_i}(A)|}{\sum_{j \neq i} |support_{C_j}(A)|}$$

The selected patterns are those for which Gr_Rate is greater than a user-defined threshold. That selection criterion is clearly related to the nice concept of emerging pattern [9]. Finally, we use one of the two interestingness measures cited above to further reduce the number of new descriptors while preserving class separation.

We can now use the selected itemsets as features for encoding a new representation for the original data. In our experiments, we computed the value taken by each new attribute $NewAttr_A$ for a given object t as follows:

$$NewAttr_A(t) = \frac{|A \cap t|}{|A|}$$

It quantifies the number of occurrences of the object in the corresponding itemset. The normalization aims to avoid "unbalanced" values due to occurrences in

large versus small itemsets. As a result, we obtain a numerical database with a number of features derived from the selected frequent (δ-free) itemsets.

3 Experimental Results

We have tested our approach on a real-world medical data set `meningitis`[2] and two data sets from the UCI Repository (`Vote` and `Cars` [5]). We report the best accuracy provided by three classical classification algorithms: Naive Bayes (NB), J48 (i.e., C4.5), and Multilayer Perceptron (NN) available within the popular WEKA platform [10]. The best accuracy results on the original data are 92.36% with J48 and 99.54% with NN for `Cars`; 97.26% for `Vote`; 94% for `meningitis`. We present the results in Table 1 for different values of γ and δ thresholds. For `Cars`, we give the rate for J48 and NN to show that the classifier built from J48 is improved. We improve the NN classifier when $minfreq = 2$ and $\delta = 2$ (99.71%). In general, the results are quite encouraging. Using the Information Gain Ratio instead of homogeneity and concentration has given better results. It should be explained by the fact that the later measures have been targeted towards closed sets and not δ-free ones.

Table 1. Accuracy for `Vote`, `Cars`, and `Meningitis`

Datasets	$minfreq$,δ	#Patterns	Measure	#Selected	Accuracy
Cars	2%,2	450	GI	21	J48=99.25, NN=99.25
			HC	36	J48=97.39, NN=99.71
	4%,4	167	GI	18	J48=98.49, NN=99.31
			HC	16	J48=86.40, NN=86.98
Meningitis	5%,2	60045	GI	7	98.79
			HC	14	93.92
	10%,2	31098	GI	6	97.88
			HC	8	92.71
Vote	5%,3	19859	GI	9	98.40
			HC	15	95.17
	10%,3	10564	GI	10	97.01
			HC	8	96.10

4 Conclusion

We have investigated the use of δ-free sets for feature construction in classification problems. In that context, the main issue was the selection of good patterns w.r.t. class separation and generalization purposes. We specified a filtering process based on both their local and global properties w.r.t. data. Using the new features on several classification tasks with different algorithms has given rise

[2] `meningitis` concerns children hospitalized for acute bacterial or viral meningitis (329 samples described by 60 Boolean attributes.

to quite encouraging results. Our next step will be to use other types of local patterns (e.g., closed or almost-closed ones, emerging patterns) and provide a fair comparison of their interest for feature construction.

Acknowledgments. This research has been carried out while N. Selmaoui and D. Gay were visiting INSA Lyon. This work is partly funded by EU contract IQ FP6-516169 (FET arm of IST). The authors wish to thank P. Francois and B. Crémilleux who provided meningitis, and J. Besson for his technical support.

References

1. Boulicaut, J.F., Bykowski, A., Rigotti, C.: Approximation of frequency queris by means of free-sets. In: Proceedings the 4th European Conference on Principles and practice of Knowledge Discovery in Databases (PKDD). (2000) 75–85
2. Boulicaut, J.F., Crémilleux, B.: Simplest rules characterizing classes generated by delta-free sets. In: 22nd SGAI International Conference on Knowledge Based Systems and Applied Artificial Intelligence, ES'02. (2002) 33–46
3. Becquet, C., Blachon, S., Jeudy, B., Boulicaut, J.F., Gandrillon, O.: Strong association rule mining for large gene expression data analysis: a case study on human SAGE data. Genome Biology **12** (2002)
4. Li, J., Li, H., Wong, L., Pei, J., Dong, G.: Minimum description length principle : generators are preferable to closed patterns. In: Proceedings 21st National Conference on Artificial Intelligence, Menlo Park, California, The AAAI Press (2006) To appear.
5. Newman, D., Hettich, S., Blake, C., Merz, C.: UCI repository of machine learning databases (1998)
6. Boulicaut, J.F., Bykowski, A., Rigotti, C.: Free-sets : A condensed representation of boolean data for the approximation of frequency queries. Data Mining Knowledge Discovery **7** (2003) 5–22
7. Bastide, Y., Taouil, R., Pasquier, N., Stumme, G., Lakhal, L.: Mining frequent patterns with counting inference. SIGKDD Explorations **2** (2000) 66–75
8. Durand, N., Crémilleux, B.: Ecclat : a new approach of clusters discovery in categorical data. In: 22nd SGAI International Conference on Knowledge Based Systems and Applied Artificial Intelligence, ES'02. (2002) 177–190
9. Dong, G., Li, J.: Efficient mining of emerging patterns: discovering trends and differences. In: Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, ACM Press (1999) 43–52
10. Witten, I.H., Frank, E.: Data Mining : Practical machine learning tools and techniques (2nd edition). Morgan Kaufmann Publishers Inc., San Francisco, USA (2005)