Extraction of a Subset of Concepts from the Frequent Closed Itemset Lattice: A New Approach of Meaningful Clusters Discovery

Nicolas Durand[1] and **Bruno Crémilleux**[2]

Abstract. In this paper we present a new idea for the discovery of meaningful clusters from categorical data (which is an usual situation, e.g. web data analysis). Our method extracts a subset of concepts from the frequent closed itemsets lattice, using an evaluation measure. This method is promising, and first experiments give attractive results.

1 INTRODUCTION

For some years, there has been a very high interest for the web data analysis in order to create new applications and services. In such a context, it is important to have efficient knowledge discovery methods on categorical data. There are works on clustering based on association rules. A family of methods [7] consists in grouping transactions into clusters in order to minimize an intra-cluster and an inter-cluster costs, but the characterization of clusters is not obtained. [4] presents a method of association rules hypergraph k-partitioning. A clustering of items is obtained, but transactions are not straightforwardly ranked into clusters. To cope with these problems, we can use some conceptual classification methods [3]. These methods create a concept hierarchy, generally represented by a lattice. Every concept can be seen as a cluster with its properties (i.e. items) and transactions. Unfortunately, the number of concepts is very high in real-world applications. KDD's results (the frequency [1] and the condensed representations like the frequent closed itemsets [5][2][6]) can be used to return a preliminary selection of clusters. Nevertheless, the number of concepts remains high and the hierarchy can not be used nor presented to an expert. Starting from the frequent closed itemsets (seen as clusters), we propose here a method, based on the definition of a cluster evaluation measure, to select the most interesting concepts gathering similar transactions. From the frequent closed itemsets side, we think that it is an original use.

2 IDEAS AND METHOD

2.1 Frequent closed itemsets

Let $\mathcal{D} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ be a data mining context, $\mathcal{O}$ a set of transactions, $\mathcal{I}$ a set of items, and $\mathcal{R}$ a binary relation between transactions and items. For $O \subseteq \mathcal{O}$ and $I \subseteq \mathcal{I}$, we define :

$$f(O) = \{i \in \mathcal{I} | \forall o \in O, (o,i) \in \mathcal{R}\}$$
$$g(I) = \{o \in \mathcal{O} | \forall i \in I, (o,i) \in \mathcal{R}\}$$

$f(O)$ associates with O, items common to all transactions $o \in O$, and $g(I)$ associates with I, transactions related to all items $i \in I$. The operators $h = f \circ g$ and $h\prime = g \circ f$ are the Galois closure operators.

Let X be an itemset, $X \in \mathcal{I}$. X is a closed itemset iff $h(X) = X$. In other terms, a closed itemset is a maximal set of items shared by a set of transactions. This notion is a key point to highlight clusters. Indeed, closed itemsets capture all the similarities among a set of transactions. Then, a cluster is a concept of the closed itemsets lattice and is composed to an itemset X and its transactions $g(X)$. The frequency of X is $\mathcal{F}(X) = \frac{|g(X)|}{|\mathcal{O}|}$. An itemset X is frequent if its frequency is at least the frequency threshold $minfr$. For the following, X denotes a frequent closed itemset, and L the set of the frequent closed itemsets.

2.2 Cluster evaluation measure

Intuitively, a relevant cluster has to be as homogeneous as possible and should gather "enough" transactions. Translated into the usual clustering framework, it means that we have to maximize the intra-cluster similarity (called here $homogeneity$) and minimize the inter-clusters similarity. We use a $concentration$ measure to limit the overlapping of transactions between clusters. We will see that a slight overlapping may be understandable and useful in some applications. The $score$ of a cluster is the average of its homogeneity and concentration and we will select clusters with high score.

$$score(X) = \frac{1}{2}(homogeneity(X) + concentration(X))$$

Let us define homogeneity and concentration. For the homogeneity, we want to favour clusters having many items shared by many transactions. Homogeneity of a cluster X is computed from its size, $g(X)$ (i.e. its number of transactions) and a divergence measure. The $divergence$ is the number of items not in X, for each transaction of $g(X)$.

$$homogeneity(X) = \frac{\mathcal{F}(X) \times |\mathcal{O}| \times |X|}{divergence(X) + (\mathcal{F}(X) \times |\mathcal{O}| \times |X|)}$$

where $divergence(X) = \sum_{t \in g(X)} |f(t) - X|$.

We have $0 \leq homogeneity(X) \leq 1$. If a cluster is pure (i.e. $\forall t \in g(X) \quad f(t) = X$), its divergence is equal to 0, and its homogeneity equals 1. The more a cluster supports transactions with items not belonging to X, the more its homogeneity leads to 0.

For the concentration, we want to favour clusters having transactions appearing the least in the whole set of the clusters. Concentration limits the overlapping of transactions between selected clusters.

[1] France Telecom R&D, 42 rue des coutures, F-14066 Caen Cédex 4, France, email: nicola.durand@francetelecom.com

[2] Université de Caen, F-14032 Caen Cédex, France, email: cremilleux@info.unicaen.fr

Concentration of a cluster X is defined by taking into account the number of clusters where each transaction appears.

$$concentration(X) = \frac{1}{|g(X)|} \times \sum_{t \in g(X)} \frac{1}{\mathcal{F}'(t)}$$

where $\mathcal{F}'(t)$ is the number of clusters where t occurs (frequency of t in L).

We have $0 \leq concentration(X) \leq 1$. If all transactions $g(X)$ occur only in X, then concentration(X) = 1. The more the transactions of $g(X)$ are frequent in the whole set of clusters, the more concentration(X) leads to 0. Finally, we have $0 \leq score(X) \leq 1$.

2.3 Selection algorithm

Let M an integer corresponding to a number of transactions that a selected cluster must classify. Selected clusters from the frequent closed itemsets lattice are provided by the following way. At first, the score of each cluster of L is computed. The cluster having the highest score is selected. Then as long as there are transactions to classify (i.e. which do not belong to any selected cluster) and clusters remain, select the cluster having the highest score and containing at least M transactions not yet classified. The value of M reduces the overlapping between clusters and transactions.

3 EXPERIMENTS

We tested our method on the well-known database Mushroom composed of 8124 transactions and 116 items. Values of the class (`edible` and `poisonous`) have been deleted, but used for the assessment of the results. We set M to $minfr \times |\mathcal{O}|$. As $minfr \times |\mathcal{O}|$ is the minimum number of transactions in a cluster, if a cluster has exactly $minfr \times |\mathcal{O}|$ transactions, then all its transactions have not been classified yet. If a cluster has more than $minfr \times |\mathcal{O}|$ transactions, some transactions may be in common with other selected clusters. We noted experimentally that $M = minfr \times |\mathcal{O}|$ is a good trade-off of overlapping.

For $minfr = 20\%$ (888 clusters), we get 4 clusters and a cluster of remaining transactions (see table 1). Here, no transaction is shared by several clusters. For $minfr = 5\%$ (9738 clusters), we obtain 16 clusters and a cluster of remaining transactions[3] (see table 2). Slight overlapping involves only clusters 14 and 16.

Table 1. Mushroom, $minfr$=20%

cluster	#p	#e
1	0	1728
2	1728	0
3	528	1120
4	1296	768
remainder	364	592

Table 3 shows the results with the method of Wang et al. [7]. 14 clusters are obtained hierarchically (by splitting trash clusters) and using several $minfr$.

Results given by our approach seem more understandable: 13 clusters among 17 are pure and we do not use an hierarchical decomposition requiring several values for $minfr$. We use a lower value of $minfr$, which is possible with efficient algorithms for mining frequent closed itemsets [2][6].

[3] The fact that some clusters have the same number of transactions is a pure coincidence.

Table 2. Mushroom, $minfr$=5%

cluster	#p	#e
1	0	432
2	0	432
3	0	432
4	0	432
5	648	0
6	648	0
7	432	0
8	432	0
9	432	0
10	432	0
11	0	768
12	0	512
13	352	96
14	288	896
15	0	416
16	72	560
remainder	180	160

Table 3. Mushroom, Wang et al.

cluster	#p	#e
1	0	94
2	0	13
3	0	6
4	26	682
5	30	2631
6	37	121
7	61	69
8	287	0
9	3388	61
10	77	372
11	0	9
12	10	19
13	0	21
14	0	110

We work currently on web data analysis coming from France Telecom R&D (clusters of user accesses and clusters of web pages). In such situations, it is useful that a transaction (e.g. a web page) appears in few clusters. It allows to retrieve pages from taxonomies corresponding to several points of view. For instance, if we have two clusters of pages described by their keywords (noted $A, B, \ldots$), $C_1 = (\{A, B, C\}\{1, 2, 4\})$ and $C_2 = (\{D, E\}\{1, 3, 5\})$, it is understandable and useful to obtain the page 1 in these two clusters, in order to retrieve it from several kind of queries. Let us note that a clustering algorithm producing a partition of transactions gives page 1 in one cluster, that means a single point of view.

4 CONCLUSION

We propose a new method to discover meaningful clusters from categorical data. Such clusters correspond to concepts selected from the frequent closed itemsets lattice. Unlike usual techniques, our approach does not use a global measure of similarity between transactions but is based on an evaluation measure of a cluster. This method does not require to fix the number of clusters beforehand. Slight overlapping between clusters may appear, which is useful in situations where a set of clusters (and not necessarily a partition) is required such in web mining. As exceptions are present in real-world data, further work is to relax the constraint of closure to allow for few exceptions in a cluster.

REFERENCES

[1] R. Agrawal and R. Srikant, 'Fast Algorithms for Mining Association Rules', in *the 20th VLDB Conference*, Santiago, Chile, (1994).

[2] J. F. Boulicaut and A. Bykowski, 'Frequent Closures as Concise Representation for Binary Data Mining', in *the 4th PAKDD*, Japan, (2000).

[3] C. Carpineto and G. Romano, 'Galois: An Order-Theoretic Approach to Conceptual Clustering', in *the Machine Learning Conf.*, (1993).

[4] E. H. Han, G. Karypis, V. Kumar, and B. Mobasher, 'Hypergraph Based Clustering in High-Dimensional Data Sets : a Summary of Results', *Bulletin of the Technical Commitee on Data Engineering*, **21**(1), (1998).

[5] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, 'Efficient Mining of Association Rules Using Closed Itemset Lattices', *Information Systems*, **24**(1), 25–46, Elsevier, (1999).

[6] G. Stumme, R. Taouil, Y. Bastide, N. Pasquier, and L. Lakhal, 'Computing Iceberg Concept Lattices with TITANIC', *Journal on Knowledge and Data Engineering*, (2002).

[7] K. Wang, X. Chu, and B. Liu, 'Clustering Transactions Using Large Items', in *ACM CIKM*, USA, (1999).