Condensed Representation of Emerging Patterns

Arnaud Soulet, Bruno Crémilleux, and François Rioult

GREYC, CNRS - UMR 6072, Université de Caen
Campus Côte de Nacre
F-14032 Caen Cédex France
{Forename.Surname}@info.unicaen.fr

Abstract. Emerging patterns (EPs) are associations of features whose frequencies increase significantly from one class to another. They have been proven useful to build powerful classifiers and to help establishing diagnosis. Because of the huge search space, mining and representing EPs is a hard task for large datasets. Thanks to the use of recent results on condensed representations of frequent closed patterns, we propose here an exact condensed representation of EPs. We also give a method to provide EPs with the highest growth rates, we call them strong emerging patterns (SEPs). In collaboration with the Philips company, experiments show the interests of SEPs.

1 Introduction

The characterization of classes and classification are significant fields of research in data mining and machine learning. Initially introduced in [5], emerging patterns (EPs) are patterns whose frequency strongly varies between two data sets (i.e., two classes). EPs characterize the classes in a quantitative and qualitative way. Thanks to their capacity to emphasize the distinctions between classes, EPs enable to build classifiers [6] or to propose a help for diagnosis. Nevertheless, mining EPs in large datasets remains a challenge because of the very high number of candidate patterns.

In this paper, we propose two contributions for the efficient extraction of emerging patterns. One originality of our approach is to take advantage of recent progresses on the condensed representations of closed patterns [9]. Firstly, we propose an exact condensed representation of the emerging patterns for a dataset. Contrary to the borders approach (Section 2) which provides the EPs with a lower bound of their growth rate, this condensed representation easily enables to know the *exact* growth rate for each emerging pattern. Secondly, we propose a method to easily provide the emerging patterns having the best growth rates (we call them " strong emerging patterns "). This work is also justified by requests from providers of data. In our collaboration with the Philips company, one notices in practice a high number of EPs and the strong emerging patterns are particularly useful to bring more synthetic and more exploitable results.

The paper is organized in the following way. Section 2 introduces the context and the required notations. Section 3 provides an exact condensed representation of the emerging patterns and proposes the strong emerging patterns. Finally, Section 4 presents the experimental evaluations.

2 Context and related works

Notations and definitions. Let $\mathcal{D}$ be a dataset (Table 1), which is an excerpt of the data used for the search for failures in a production chain (cf. Section 4).

$\mathcal{D}$

Batch	Items						
B_1	C_1		A	B	C	D	
B_2	C_1		A	B	C		
B_3	C_1		A			D	E
B_4		C_2	A	B	C		
B_5		C_2		B	C	D	E
B_6		C_2		B			E

Table 1. Example of a transactional data set

Each line (or *transaction*) of Table 1 represents a batch (noted $B_1, \ldots, B_6$) described by features (or *items*) : $A, \ldots, E$ denote the advance of the batch within the production chain and C_1, C_2 the class values. $\mathcal{D}$ is partitioned here into two datasets $\mathcal{D}_1$ (the right batches) and $\mathcal{D}_2$ (the defective batches). The transactions having item C_1 (resp. C_2) belong to $\mathcal{D}_1$ (resp. $\mathcal{D}_2$). A *pattern* is a set of items (e.g., $\{A, B, C\}$) noted by the string ABC. A transaction t contains the pattern X if and only if $X \subseteq t$.

The concept of emerging patterns is related to the notion of frequency. The frequency of a pattern X in a dataset $\mathcal{D}$ (noted $\mathcal{F}(X, \mathcal{D})$) is the number of transactions of $\mathcal{D}$ which contain X. X is *frequent* if its frequency is at least the frequency threshold fixed by the user. Intuitively, an emerging pattern is a pattern whose frequency increases significantly from one class to another. The capture of contrast between classes brought by a pattern is measured by its growth rate. The *growth rate* of a pattern X from $\mathcal{D}\backslash\mathcal{D}_i$ to $\mathcal{D}_i$ is defined as:

$$GR_i(X) = \frac{|\mathcal{D}| - |\mathcal{D}_i|}{|\mathcal{D}_i|} \times \frac{\mathcal{F}(X, \mathcal{D}_i)}{\mathcal{F}(X, \mathcal{D}) - \mathcal{F}(X, \mathcal{D}_i)} \tag{1}$$

Definition 1 (emerging pattern or EP). *Given threshold $\rho > 1$, a pattern X is said to be an* emerging pattern *from $\mathcal{D}\backslash\mathcal{D}_i$ to $\mathcal{D}_i$ if $GR_i(X) \geq \rho$.*

Let us give some examples from Table 1. With $\rho = 3$, A and $ABCD$ are EPs from $\mathcal{D}_2$ to $\mathcal{D}_1$. Indeed, $GR_1(A) = 3$ and $GR_1(ABCD) = \infty$.

Related Works. Efficient computation of all EPs in high dimensional datasets remains a challenge because the number of candidate patterns is exponential according to the number of items. The naive enumeration of all patterns with their frequencies fails quickly. In addition, the definition of EPs does not provide anti-monotonous constraints to apply a powerful pruning of the search space for methods stemming from the framework of level-wise algorithms [8]. The approach of handling borders, introduced by Dong and al. [5], enables to give a concise description of the emerging patterns. On the other hand, it requires to repeat for all the $\mathcal{D}_i$ the computation of the intervals and it does not provide for each EP its growth rate. Some other approaches exist like [10, 7, 1].

In the following, we focus on the condensed representation based on closed patterns [4]. Such an approach enables the implementation of powerful pruning criteria during the extraction, which improves the efficiency of algorithms [9, 3]. A *closed* pattern in $\mathcal{D}$ is a maximal set of items shared by a set of transactions [2]. The notion of *closure* is linked to the one of closed pattern. The *closure* of a pattern X in $\mathcal{D}$ is $h(X, \mathcal{D}) = \bigcap\{transaction\ t\ in\ \mathcal{D} | X \subseteq t\}$. An important property on the frequency stems from this definition. The closure of X is a closed pattern and $\mathcal{F}(X, \mathcal{D}) = \mathcal{F}(h(X, \mathcal{D}), \mathcal{D})$. In our example, $h(AB, \mathcal{D}) = ABC$ and $\mathcal{F}(AB, \mathcal{D}) = \mathcal{F}(ABC, \mathcal{D})$. Thus, the set of the closed patterns is a condensed representation of all patterns because the frequency of any pattern can be inferred from its closure.

3 Condensed representation and strong emerging patterns

3.1 Exact condensed representation of emerging patterns

Let us move now how to get the growth rate of any pattern X. Equation 1 shows that it is enough to compute $\mathcal{F}(X, \mathcal{D})$ and $\mathcal{F}(X, \mathcal{D}_i)$. These frequencies can be obtained from the condensed representation of frequent closed patterns. Indeed, $\mathcal{F}(X, \mathcal{D}) = \mathcal{F}(h(X, \mathcal{D}), \mathcal{D})$ (closure property) and by the definition of the partial bases $\mathcal{D}_i$, $\mathcal{F}(X, \mathcal{D}_i) = \mathcal{F}(XC_i, \mathcal{D}) = \mathcal{F}(h(XC_i, \mathcal{D}), \mathcal{D})$. Unfortunately, these relations require the computation of two closures ($h(X, \mathcal{D})$ and $h(XC_i, \mathcal{D})$), which it is not efficient. The following properties solve this disadvantage:

Property 1 *Let X be a pattern and $\mathcal{D}_i$ a dataset, $\mathcal{F}(X, \mathcal{D}_i) = \mathcal{F}(h(X, \mathcal{D}), \mathcal{D}_i)$*

Proof. The properties of the closure operator ensure that for any transaction t, $X \subseteq t \Leftrightarrow h(X, \mathcal{D}) \subseteq t$. In particular, the transactions of $\mathcal{D}_i$ containing X are identical to those containing $h(X, \mathcal{D})$ and we have the equality of the frequencies.

It is now simple to show that the growth rate of every pattern X is obtained thanks to the only knowledge of the growth rate of $h(X, \mathcal{D})$:

Property 2 *Let X be a pattern, we have $GR_i(X) = GR_i(h(X, \mathcal{D}))$.*

Proof. Let X be a pattern. By replacing $\mathcal{F}(X, \mathcal{D})$ with $\mathcal{F}(h(X, \mathcal{D}), \mathcal{D})$ and $\mathcal{F}(X, \mathcal{D}_i)$ with $\mathcal{F}(h(X, \mathcal{D}), \mathcal{D}_i)$ in Equation 1, we immediately recognize the growth rate of $h(X, \mathcal{D})$.

For instance, $h(AB) = ABC$ and $GR_1(AB) = GR_1(ABC) = 2$. This property is significant because the number of closed patterns is lower (and, in general, much lower) than that of all patterns [4]. Thus, the frequent closed patterns with their growth rates are enough to synthesize the whole set of frequent EPs with their growth rates. We obtain an *exact* condensed representation of the EPs (i.e. the growth rate of each emerging pattern is exactly known). Let us recall that the borders technique (cf. Section 2) only gives a lower bound of the growth rate.

3.2 Strong emerging patterns

The number of emerging patterns of a dataset can be crippling for their use. In practice, it is judicious to keep only the most frequent EPs having the best growth rates. But thoughtlessly raising these two thresholds may be problematic. On the one hand, if the minimal growth rate threshold is too high, the found EPs tend to be too specific (i.e. too long). On the other hand, if the minimal frequency threshold is too high, EPs have a too low growth rate.

We define here the strong emerging patterns which are the patterns having the best possible growth rates. They are a trade-off between the frequency and the growth rate.

Definition 2 (strong emerging pattern). *A strong emerging pattern X (SEP in summary) for $\mathcal{D}_i$ is the emerging pattern coming from a closed pattern XC_i in $\mathcal{D}_i$ (i.e. C_i does not belong to the SEP).*

SEPs enable to highlight the following important property (due to space limitation the proof is not given here).

Property 3 (SEPs: EPs with maximum growth rate) *Let X be a pattern not containing the item C_i. Then the SEP coming from $h(X, \mathcal{D}_i)$ has a better growth rate than X, i.e. one has $GR_i(X) \leq GR_i(h(X, \mathcal{D}_i)\backslash\{C_i\})$.*

Let us illustrate Property 3 on the elementary example. The pattern BC is not a SEP for class 1 (because $h(BC, \mathcal{D}_1)\backslash\{C_1\} = ABC$), its growth rate is 1 and one has well $GR_1(BC) \leq GR_1(ABC) = 2$ and $\mathcal{F}(BC, \mathcal{D}_1) = \mathcal{F}(ABC, \mathcal{D}_1)$.

Compared to EPs, SEPs have two meaningful advantages: they are easy to discover from the condensed representation of frequent closed patterns (by simply filtering those containing C_i), and they have the best possible growth rates. Let us notice that the emerging patterns based on X and $h(X, \mathcal{D}_i)$ have the same frequency, thus they have the same quality according to this criterion. However, the strong emerging pattern coming from $h(X, \mathcal{D}_i)$ has a stronger (i.e. higher) growth rate and thus offers a better compromise between frequency and growth rate.

4 Experiments

Experiments were carried out on a real dataset within a collaboration with the Philips company. The industrial aim is to identify mistaken tools in a silicon plate production chain. The quality test leads to three quasi-homogeneous classes corresponding to three quality levels. Figure 1 depicts the distributions of EPs according to the length of patterns for a minimal frequency threshold of 1.2%. The number of closed EPs (which stemmed from closed patterns) measures of the size of the condensed representation. Two threshold values of the minimal growth rate (1 and ∞) are used. This figure shows that the number of EPs is very high compared to the number of closed EPs or SEPs. This disproportion does not decrease in spite of the rise of the minimal growth rate. These too large numbers of EPs cannot be presented to an expert for his analysis task. Other experiments have shown that the number of closed EPs and SEPs increases less quickly than the number of EPs when the frequency decreases. So, it is possible to examine longer and less frequent patterns.

After talks with the Philips experts, they have confirmed that the stage suspected by the SEPs was the real cause of the failures (an equipment was badly tuned). This experiment shows the practical contribution of SEPs on real-world data.

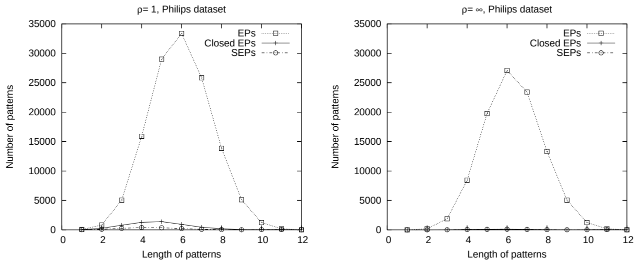

Fig. 1. Comparison between the different kinds of emerging patterns

5 Conclusion

Based on recent results in condensed representations, we have revisited the field of emerging patterns. We have defined an exact condensed representation of emerging patterns for a data base and proposed the strong emerging patterns which are the EPs with the highest growth rates. In addition to the simplicity of their extraction, this approach produces only few SEPs which are particularly useful for helping to diagnosis. So, it is easier to use SEPs than search relevant EPs among a large number of EPs. Dealing with our collaboration with the Philips company, SEPs enable to successfully identify the failures of a production chain of silicon plates. These promising results encourage the use of strong emerging patterns. Further works concern the use of SEPs for classification tasks.

Acknowledgements. The authors wish to thank Philips company. This work has been partially founded by the AS "Discovery Challenge" supported by CNRS.

References

[1] J. Bailey, T. Manoukian, and K. Ramamohanarao. Fast algorithms for mining emerging patterns. In *PKDD'02*, pages 39–50, 2002.
[2] G. Birkhoff. Lattices theory. *American Mathematical Society, vol. 25*, 1967.
[3] J. F. Boulicaut, A. Bykowski, and C. Rigotti. Free-sets: a condensed representation of boolean data for the approximation of frequency queries. *DMKD journal*, 2003.
[4] T. Calders and B. Goethals. Mining all non-derivable frequent itemsets. In *proceedings of PKDD'02*, 2002.
[5] G. Dong and J. Li. Efficient mining of emerging patterns: Discovering trends and differences. In *Knowledge Discovery and Data Mining*, pages 43–52, 1999.
[6] G. Dong, X. Zhang, W. Wong, and J. Li. CAEP: Classification by aggregating emerging patterns. In *Discovery Science*, pages 30–42, 1999.
[7] J. Li and K. Ramamohanarao. The space of jumping emerging patterns and its incremental maintenance algorithms. In *Proc. ICML*, 2000.
[8] H. Mannila and H. Toivonen. Levelwise search and borders of theories in knowledge discovery. *Data Mining and Knowledge Discovery*, 1(3):241–258, 1997.
[9] N. Pasquier, Y. Bastide, T. Taouil, and L. Lakhal. Discovering frequent closed itemsets for association rules. *Lecture Notes in Computer Science*, 1540, 1999.
[10] X. Zhang, G. Dong, and K. Ramamohanarao. Exploring constraints to efficiently mine emerging patterns from large high-dimensional datasets. In *KDD*, 2000.