View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Binus University Repository

A Framework to Mine High-level Emerging Patterns by Attribute-oriented Induction

[1]Maybin K. Muyeba [2]Muhammad S. Khan [3]Spits Warnars [4]John Keane

[1, 3]Sch. of Computing, Maths and Digital Techn., Manchester Metropolitan University, UK
{m.muyeba, s.warnars}@mmu.ac.uk
[2]Department of Computer Science, School of Electrical Engineering and Computer Science,
University of Liverpool, UK
mskhan@liverpool.ac.uk
[4]School of Computer Science, The University of Manchester, UK
john.keane@cs.manchester.ac.uk

Abstract. This paper presents a framework to mine summary emerging patterns in contrast to the familiar low-level patterns. Generally, growth rate based on low-level data and simple supports are used to measure emerging patterns (EP) from one dataset to another. This consequently leads to numerous EPs because of the large numbers of items. We propose an approach that uses high-level data: high-level data captures the data semantics of a collection of attributes values by using taxonomies, and always has larger support than low-level data. We apply a well known algorithm, attribute-oriented induction (AOI), that generalises attributes using taxonomies and investigate properties of the rule sets obtained by generalisation algorithms.

Keywords: attribute-oriented; algorithm; rulesets; high-level, emerging pattern

1 Introduction

Data mining aims to find patterns in data. Recently, emerging patterns [1] have become popular for classification problems [7][11]. Emerging patterns (EP) [9] represent contrasting characteristics between two data sets usually expressed as conjunctions of attribute values in a given class of records. The most familiar approaches use classification [6][10][11]. A pattern is emerging (EP) if its support from one dataset to another increases. A pattern is jumping emerging (JEP) if its support from the previous dataset changes from zero to non-zero.

EPs have been successfully used in classification algorithms with mainly low-level (primitive) data. Low-level data has a tendency to be distinct yet represent semantically similar information e.g. for an attribute “Course”, there may be two different university degree subjects “Chemistry” and “Physics” that are both in category “Science”, a level higher than both subjects. These are two distinct items of data yet they semantically belong to one item “Science”. The problem with low-level EP algorithms is the generation of many EPs because of the combinatorial problem in the number of items and also the use of small supports. As a consequence, most EP classifiers use level-wise border searches to control pattern explosion [8][10]. In contrast using high-level summarised data or attribute taxonomies (is-a hierarchies) that capture significant data features often tends to prune common and irrelevant features usually found from low-level data, leaving only high-level supported items [13]. Attribute taxonomies reveal attribute details at various higher levels in the hierarchy. It is well known that larger supports of an attribute's values occur at higher levels than at lower levels of a given taxonomy [12]. Using this fact, it is imperative that EPs can be used to exploit various taxonomic levels of attributes to express varying support levels of item combinations, and hence more significant EPs.

A well established algorithm for mining *is-a* hierarchies from large data to produce conjunctions of attribute-value pairs is attribute-oriented induction (AOI) [14]. Attribute taxonomies, also known as background knowledge or concept hierarchies, are provided by a domain expert or generated automatically. AOI can generate various types of rule patterns, including discriminant, characteristic and classification rules. In the latter case, there is no need to train the data as AOI searches through the input space using both low-level data and their corresponding taxonomies. The reader is referred to [3] for details on the basic AOI algorithm.

Our motivation is three-fold: firstly, AOI is a versatile algorithm for solving the EP problem using various techniques; secondly, larger supports of attribute values mostly occur higher up the taxonomy than would be for low-level values; thirdly, as there is usually a combinatorial explosion of patterns at a low-level, it makes it more difficult for a user to interpret so many patterns compared with a general pattern i.e. patterns expressed at high taxonomic levels. Thus such general patterns have more expressive potential, and represent global data semantics better, than single primitive patterns.

In this paper we give formal definitions of the problem of mining HEPs and introduce and begin to evaluate an algorithm, AOI-HEP (High-level Emerging Patterns) which mines high-level emerging patterns using an enhanced AOI approach.

The paper is organised as follows: Section 2 presents the background; Section 3 gives the problem definition; section 4, the new AOI-HEP algorithm; section 5 experimentation; and Section 6 presents conclusions.

2 Background

In this section, we introduce formal HEP definitions used in the paper.

Let $D = \{A_1, A_2, .., A_m\}$ be a dataset with m attributes each with a domain $Dom(A_i)$ and N tuples. For each attribute there are a set of values or instances $\{a_i^k\}$, $1 \le k \le m$ and $1 \le i \le N$. We can also label classes to which these instances belong as $C = \{C_1, C_2, .., C_v\}$. An item is an (attribute, instance) pair (A_k, a_i^k). A set of items is an itemset when there is a combination of items, $\bigcup_{i=1}^{n} a_i^k, n < m * N$, for some values of k. The number $m * N$ is the largest possible number of combinations of items for m attributes and N tuples.

Accordingly, in AOI, for each attribute there is a taxonomy $H_i, 1 \le i \le m$ linked to it. To define support, we first represent a characteristic AOI rule pattern as a conjunction of items with instance values of any object x, which can be referred to as a complex pattern [7] of the form:

$$A_1(x) = v_1 \wedge .. \wedge A_k(x) = v_k, \quad [s\%] \tag{1}$$

where $s\%$ is the usual support (% of tuples in the dataset) and $v_i \in \{H_i \cup Dom(A_i)\}$. We see that support is for a more complex expression of item-value pairs than single items. Note that $v_i s$ are not necessarily low-level data or domain values but extracted from attribute hierarchies H_i. In AOI, this pattern forms part of a characteristic rule [5]. The definition of itemset uses the term ruleset as a complex pattern represented by (1). A characteristic rule according to [4] is defined as

$$h \rightarrow e \tag{2}$$

where e is evidence (shown as equation 1) and h is the hypothesis or class description we wish to characterise. Thus we rewrite (1) as a characteristic rule:

$$C_i(x) \rightarrow A_1(x) = v_1 \wedge .. \wedge A_k(x) = v_k \; [s\%]$$

in relation to some hypothetical class C_i. Given two data sets D_1, D_2 with rulesets R_1, R_2, we want to find interesting emerging rule patterns like (3) using supports $s_1 = count_{R_1}(X) / | D_1 |, s_2 = count_{R_2}(X) / | D_2 |$ where X is a given complex pattern (equation 3), namely $X = \{(A_1, v_1),..,(A_k, v_k)\}$, for k attribute-value pairs. Note that we can use many relevant interestingness measures from the literature to compare interestingness of patterns, not only based on support, as highlighted in [4].

3 Problem Definition

Following the definitions of emerging patterns, we formulate the problem as follows: given rule sets R_i, R_j of datasets D_1, D_2, a ruleset is a series of attribute conjunctions. A subset $X \subseteq R_i$ is called a $k - rule\ set$ if $k = | X |$ is the number of attribute-value pairs in X. We need to define subsumption properties for the case where one ruleset subsumes another by one or more values from the taxonomy. A HEP is a ruleset whose support increases from one ruleset of a dataset to another. A ruleset X is a HEP from R_1 of D_1 to R_2 of D_2 if $\sigma = s_2 / s_1 > 0$. Note that $\sigma \in [0, \infty]$ and if $\sigma = \infty$, then the pattern is a jumping high-level emerging pattern (JHEP). If $\sigma = 0$, then there is no HEP, otherwise $0 < \sigma < \infty$ is merely an

$\rho - HEP$ where ρ is a threshold. A HEP pattern r_i can consist of low and high-level values from the taxonomy. There is a subsumption property such that if two HEPs exist and $r_i \subseteq r_j$, then r_i is covered by r_j i.e. r_j has one or more ancestor concept values of some or all values in r_i. There is therefore an order relation $\leq$ defined on child-ancestor pairs (v_i, v'_i) as $v_i \leq v'_i$ for each attribute taxonomy. Below we give subsumption properties that help to find HEPs and JHEPs. All patterns found by the AOI-HEP algorithm with properties as in section 3.1 are HEP patterns.

3.1 Subsumption Properties

P1. Total Subsumption Emerging Pattern (TSEP). We say that ruleset X is totally subsumed by ruleset Y if $x \leq y, \forall x \in X, y \in Y$ and $|X| \leq |Y|$. Note that this property is true for both partial orders i.e. cases where $x = y$ or $x < y$. For example, let $X = \{(a_1, v_1), (a_3, v_3), (a_4, v_4)\}$ and $Y = \{(a_1, v'_1), (a_3, v'_3), (a_4, v'_4)$ where $ancestor(v_k) = v'_k$. Then X is totally subsumed by Y. The TSEP rule condition "=" or "equality-based" may be rare to find as it is equivalent to finding exactly matching rulesets in the two datasets where as "<" or "ancestor-based" means X has some ancestors in Y . The equality condition means the same number of conjunctions and attribute-value pairs exist. We note that the ancestor subsumption property can be equivalent to finding large and frequent itemset in classical frequent itemset mining**.**

P2. Partial Subsumption Emerging Pattern (PSEP). We say that ruleset X is partially subsumed by ruleset Y if there exists some $y \in Y$ which is an ancestor of some $x \in X$ and $|X| < |Y|$. For example, let $X = \{(a_1, v_1), (a_3, v_3), (a_4, v_4)\}$ and $Y = \{(a_1, v_1), (a_3, v'_3), (a_4, v_4)$. There is one partial subsumption value such that $ancestor(v_3) = v'_3$ and all other values satisfy $x = y$. In addition, given $Z = \{(a_1, v_1), (a_3, v_3)\}$, then Z is partially subsumed by Y without (a_4, v_4).

P3. Overlapping Emerging Patterns (OEP). Overlapping emerging patterns occur when there are one or more common patterns between rulesets. . If we have $Y = \{(a_1, v_1), (a_3, v'_3), (a_4, v_4)$ and $Z = \{(a_1, v_1), (a_3, v_3), (a_5, v_5)\}$ then Y overlaps Z except for (a_5, v_5). The pattern (a_5, v_5) absent in Y is a jumping HEP (JHEP) from Y to Z. We call this a partially subsumed and overlapping jumping high-level emerging pattern $p_3 - JHEP$ under property P3. Conversely, the overlapping property can also be used to find diminishing patterns i.e. if suddenly the whole pattern disappears from one dataset to another.

Intuitively, both HEP and $JHEP$ can be obtained from patterns exhibiting properties P1, P2 and P3 by comparing supports using a growth function. The basic emerging pattern problem was highlighted in [1] with a growth rate given in terms of simple support i.e. $Growth-rate(X) = G(X) = \frac{\sup p_2(X)}{\sup p_1(X)}$ where X is an itemset of datasets D_1, D_2 for some threshold ρ. In our case, as we have subsumption properties between rulesets, we represent pattern X from D_1 and pattern Y from D_2 and supports s_1, s_2 respectively as defined earlier (Equation (1)). Given that emerging patterns are a function of supports s_1, s_2, the growth rate can be measured by any function $f(s_1, s_2)$. The subsumption properties hold as follows:

$$G(X,Y) = \begin{Bmatrix} 0 & \mathit{if}\ s_1 = 0, s_2 = 0, \approx P1, P2 \\ \infty & \mathit{if}\ s_1 = 0, s_2 > 0, \approx P1, P2, P3 \\ f(s_1, s_2) & \mathit{otherwise} \approx P1, P2 \end{Bmatrix} \qquad \textbf{(3)}$$

Equation (3) shows that property P1 is synonymous with many classification approaches [1][10] where $f(s_1, s_2) = \frac{s_2}{s_1}$ for measuring emerging patterns when the itemsets match exactly. For rulesets, a coverage function $C(R_i, R_j)$ is used to measure how two rules compare (their similarity) or simply a measure of distance between rules as defined in [2]. This measures the number of attributes in both rules, overlapping and non-overlapping with special conditions. When coverage is determined, rulesets are paired to measure emerging ruleset patterns using the growth function f. This process is equivalent to finding large and frequent itemsets in classical frequent itemset mining and comparing them between datasets as used in emerging patterns. Note that properties P1 and P2 can lead to all three values $(0, \infty, s_2 / s_1)$; P3 is used to help find jumping emerging patterns. Section 4 shows how these patterns are extracted by firstly determining the coverage of rulesets.

4 AOI-HEP Algorithm

The AOI-HEP algorithm uses **a** growth function f and a coverage measure $C(R_i, R_j)$ (similar to the distance metric in [2]) between any rulesets R_i, R_j, given N rulesets**.** The algorithm scans through characteristic rulesets mined from two datasets D_1, D_2 and puts them into relevant pairings according to their coverage using properties P1, P2 and P3 as discussed (Line 3). At Line 4, $C(R_i, R_j)$ checks rule similarity and collects different rulesets e.g. TSEP, OEP etc. After collecting k rulesets, the function f is applied to determine the degree of growth.(Line 8).

```
Input : rulesets $R_{1,} ..., R_N$ from D1, D2; threshold t, using AOI
Output: $EP = \{[ep_1, val_1],..,[ep_m, val_m]\}, Val_i \in [0, \infty], m \le N$
$EP$=emerging Pattern, $i = 1, j = 1$
1. $EP \leftarrow \varnothing$, rulesets$\leftarrow\varnothing$
2. Iterate through the rule sets by comparing $R_i, R_j$ rules
3.  WHILE NOT ($R_i == \varnothing$ and $R_j == \varnothing$) and i, j <= t
4.   if $C(R_i, R_j)$ is satisfied // distance or similarity of rules
5.        $rulesets[k] \leftarrow rulesets[k] + Add(R_i, R_j)$
6. END WHILE
7. FOR $k : 0\, to\, |\, rulesets\, |\, DO$
8.        $EP = EP \cup f(ruleset_k)$ //Apply growth function f
9. OUTPUT $\{EP\}$
```

The result obtained at Line 9 returns a mixture of patterns, OEP, TSEP, PSEP etc. We can apply a ranking function to order the significance of such patterns in terms of their growth-rate. Note that the growth-rate function can be more complex than use of simple support ratios as high-level rule-based emerging patterns have fewer but more complex patterns compared to single itemset patterns.

5 Experimentation

To demonstrate the effectiveness of the proposed AOI-HEP, we have carried out experimentation on breast cancer Wisconsin dataset using 5 attributes that influence cancer diagnosis (699 patterns) [15], and constructed concept hierarchies for each: clump thickness, cellSize, cellShape, bareNuclei and normalNuclei. Dataset D1 had 349 tuples and D2 had 350. Some of the challenges in the experiments and the presented framework was setting an optimal threshold so that the rules and growth rates are meaningful, as the case is in AOI. Evidently, bigger thresholds generate numerous patterns while smaller thresholds generate fewer and meaningless patterns. AOI-HEP was run with thresholds 3, 4 and 5.

Firstly, we assumed that all occurrences of the attribute value “ANY” in the output patterns were meaningless, and so we did not consider threshold 2 in that case. We did not set a growth-rate threshold but ordered all the growth-rates in descending order. Note also that values “-“ in table 1 indicate no patterns found. It is easy to infer from table 1 that TSEP patterns obviously occur when lowest or highest thresholds are used.

Table 1. Patterns from cancer datasets D1, D2 [15]

Threshold	**OEP Growth%**		**PSEP Growth%**		**TSEP Growth%**	
	High	**Low**	**High**	**Low**	**High**	**Low**
3	11.30	0.08	1.50	0.12	0.58	-
4	23.46	1.09	6.00	0.09	-	-
5	3.28	0.46	0.75	0.14	2.49	0.05

The former justifies the need to remove root node "ANY" but in the latter, we can have numerous patterns in which we case we need to pick the strongest ones.

We noted, as per property 1, that TSEP rules (i.e. exactly matching in some cases) were rare to find, but not OEP patterns. Using threshold 3 between D1 and D2, we found one large OEP pattern: Rule 3 (D1): "Clumpthickness=highriskClump AND cellSize=aboutAverage", Rule 1 (D2): "Clumpthickness=lowriskClump AND cellSize=aboutAverage",growth-rate (0.79/0.068)=11.30. In contrast, we found the smallest OEP pattern to be: Rule 1 (D1):"Clumpthickness=lowriskClump AND cellSize=aboutAverage AND cellShape=aboutAverage" overlapping with Rule 3 (D2): "Clumpthickness=highriskClump AND cellSize=aboutAverage", growth-rate (0.04/0.51)=0.08. Technically, the former presents a redundancy in that clump thickness does not discriminate in the growth of the pattern. In the real world, practitioners will find this pattern of concern, noting the growing number of patients with thickening clumps despite average cancer cell sizes. In the latter, practitioners would use the least growing OEP pattern and note the role played by the differentiating attribute "cell shape = about average" between patients with high risk clump thickness and those with low risk clump thickness despite cell sizes being about average.

We observed threshold 5 for patterns. The highest OEP pattern (see table 1) had a growth-rate of 3.28 i.e. rules R2 (D1) and R1 (D2): "clumpthickness=mediumClump AND cellSize=smallSize AND cellShape=smallShape AND bareNuclei=smallNuclei". Basically, higher thresholds could be useful drill-down strategies to check or further validate rules already found using low thresholds. In this case, we could look for high impact patterns, for example the TSEP growth-rate 2.49 for threshold 5 and check whether this pattern is supported sufficiently at higher levels accordingly using some growth-rate threshold.

Further experiments were done on some UCI repository dataset, the adult dataset and similar and interesting patterns were discovered. Generally, the sequence of patterns OEP, PSEP and then TSEP in table 1 denote their order of importance. That is OEPs are quite "frequent" as would be in non-generalisation algorithms, TSEPs are expected in generalisation and merge approaches. They can also be rare but could reflect trends of "similar" subsumed patterns between datasets.

6 Conclusion

The paper has presented a novel framework for mining HEPs using AOI, the AOI-HEP algorithm. This framework has highlighted different aspects of mining emerging patterns by use of more complex rulesets, instead of itemsets, and their subsumption properties that translate into different types of emerging patterns. HEP patterns are particularly representative and informative in relation to large datasets and complex rulesets. Initial evaluation suggests that matching rulesets can use a general function that evaluates coverage or similarity of rules. We intend to apply the algorithm on further diverse real datasets, noting that optimal threshold choices (not too small or too big) could also influence ruleset generation and consequently growth rates. We will also extend the presented framework for mining total subsumption patterns at different hierarchical levels (including root node "ANY") by taking into account features of hierarchical data such as distances, similarity between concepts and appropriate level supports. We also note that the pattern properties presented here are well placed to handle uncertainty or fuzziness in the patterns.

References

1. 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, San Diego, California, United States, August 15 - 18 (1999)
2. Gago, P., Bentos, C. : A metric for selection of the most promising rules. In Zytkow, J. , Quafafou, M. (eds) Proceedings of PKDD'98, pp. 19-27, Nantes, France (1998)
3. Han, J., Cercone, N. , Cai, Y. : Attribute-Oriented Induction in Relational Databases. In: Piatetsky-Shapiro, G. , Frawley, W. J. Frawley (eds) Knowledge Discovery in Databases, pp. 213-228 (1991)
4. Hilderman, R.J. and Hammilton, H.J. : Knowledge Discovery and measures of interest, Kluwer academic, (2001)
5. Kamber, M., Shinghal, R. : Evaluating the interestingness of characteristic rules. In: Proceedings of the Second on Knowledge Discovery and Data Mining (KDD'96). pp 263-266, Portland, Oregon, USA (1996)
6. Ramamohanarao, K. and Fan, H. : Patterns Based Classifiers. World Wide Web, vol. 10(1), pp. 71-83 (2007)
7. García-Borroto, M., Martínez-Trinidad, J. and Carrasco-Ochoa, J.A. : Fuzzy emerging patterns for classifying hard domains, Knowledge and Information Systems (2010)
8. Dong, G. ,Li, J.: Mining border descriptions of emerging patterns from dataset pairs, Knowledge and Information Systems, vol. 8, pp.178-202 (2004)
9. Fan, H. , Ramamohanarao, K.: Fast Discovery and the Generalisation of Strong Jumping Emerging Patterns for building Compact and Accurate Classifiers, IEEE Transactions on Knowledge and Data engineering, vol. 18(6), pp. 721-737 (2006)
10. Dong, G. ,Li, J. : Mining border descriptions of emerging patterns from dataset pairs, Knowledge and Information Systems, vol. 8, pp. 178-202 (2005)
11. Ceci, M., Appice, A. , Malerba, D.: Emerging pattern based classification, LNCS vol. 5181, pp. 283-296 (2008)
12. Agrawal, A . , Srikant, R.: Mining Generalised association rules, VLDB. (1995)
13. Qian, X., Bailey, J. , Leckie, C. : Mining Generalised Emerging Patterns, LNAI, vol. 4304, pp. 295-304 (2006)
14. Chen, Y.L. Wu, Y. Y. , Chang, R.I.: From data to global generalized knowledge, Journal of Knowledge and Information Systems. (2010)
15. UCI Machine Learning Repository http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.names