Using Highly Expressive Contrast Patterns For Classification - Is It Worthwhile?

Elsa Loekito and James Bailey

NICTA Victoria Laboratory
Department of Computer Science and Software Engineering
University of Melbourne, Australia
{eloekito,jbailey}@csse.unimelb.edu.au

Abstract. Classification is an important task in data mining. Contrast patterns, such as emerging patterns, have been shown to be powerful for building classifiers, but they rarely exist in sparse data. Recently proposed disjunctive emerging patterns are highly expressive, and can potentially overcome this limitation. Simple contrast patterns only allow simple conjunctions, whereas disjunctive patterns additionally allow expressions of disjunctions. This paper investigates whether expressive contrasts are beneficial for classification. We adopt a statistical methodology for eliminating noisy patterns. Our experiments identify circumstances where expressive patterns can improve over previous contrast pattern based classifiers. We also present some guidelines for i) using expressive patterns based on the nature of the given data, ii) how to choose between the different types of contrast patterns for building a classifier.

Key words: Expressive contrasts, emerging patterns, disjunctive emerging patterns, classification, quantitative association rules

1 Introduction

Classification is a well studied area in data mining. Contrast patterns [1, 2] capture strong differences between classes, and useful for building accurate classifiers. Existing pattern-based classifiers consider simple contrasts, such as emerging patterns [3], which are conjunctions of attribute values. A highly expressive class of contrast, namely disjunctive emerging patterns [4], allows disjunctions as well as conjunctions of attribute values. Their use for classification is an open question though, which we aim to answer in this paper.

Expressive contrasts can potentially overcome some of the limitations of simple contrasts. E.g. the following disjunctive pattern from the *income* [5] data set: $[age \in [30..39] \wedge (industry =$ 'manufacturing' $\vee$ 'transportation')] differentiates males from females, being true for more than 10% of the males but not true for any female. If the two industries were considered individually, the non-disjunctive combination $[age \in [30..39] \wedge industry =$ 'manufacturing'], would be true for far fewer males, thus, a weaker contrast. This issue often arises when the data is

sparse, or lacking in data instances. Despite their low frequency, rare contrasts can be useful for classification, but they are often not identified.

Since emerging patterns assume discrete data, the rarity of contrasts can also result from the data discretisation used, when the input data set has a continuous-valued domain. We call this problem the *resolution problem*. In coarsely discretised data, patterns may be lacking class-distinguishing ability, but in a finely discretised data, patterns may be lacking frequencies (or support). Expressive patterns provide a solution to this problem, by allowing several discrete attribute-values to be combined into a disjunction.

Expressive contrasts may help remedy the above-mentioned situations, but they may also have limitations: i) an increased number of patterns become available, ii) more patterns may be noisy. E.g., $[(age \in [20..24] \vee [40..44]) \wedge industry =$ 'manufacturing'] is a valid disjunctive pattern, but the two age groups, [20..24] and [40..44], may be irrelevant. Such irrelevance within a pattern may, in turn, cause misclassification. To address this issue, we propose a method for statistically testing the significance of disjunctive patterns.

This paper investigates the advantages and disadvantages of using highly expressive contrasts, instead of simple contrasts, for classification. We aim to answer the following questions: i) When should disjunctions be allowed in contrast patterns for building a classifier? ii) Which types of contrast patterns are most suitable for various data characteristics? Our contributions are three-fold:

- We propose a classifier model based on disjunctive emerging patterns [4]. To eliminate noise, we use a statistical significance method, similar to that used in [6], which is based on the Fisher's Exact Test. To test the significance of each element in a pattern, we extend the testing methodology by using the *negative representation* of the pattern, which is a conjunction of the negated attribute values. The use of statistical tests on negative conjunctions has not been previously studied.
- We present experimental results using several real [5] data sets, to study the accuracy of our classifier. We use an existing contrast pattern based classifier [1] as a baseline. It shows that the disjunctive classifier is superior for sparse data, and as good as the baseline for dense data. Moreover, data discretisation or data sparsity has low influence on the classification accuracy when expressive contrasts are used.
- Based on our findings, we present a series of recommendations for practitioners, which answer the two questions posed earlier, regarding when disjunctions should be allowed in contrast patterns, and which types of contrasts are most suitable for classifying data with particular characteristics.

2 Contrast Pattern Definitions

A dataset D is defined upon a set of k attributes (also referred as dimensions) $\{A_1, A_2, \ldots, A_k\}$. For every attribute A_i, the domain of its values (or items) is denoted by $dom(A_i)$. Let I be the aggregate of the domains across all the

attributes, i.e. $I = \bigcup_{i=1}^{k} dom(A_i)$. An *itemset* is a subset of I. Let P and Q be two itemsets. We say P *contains* Q if Q is a subset of P, $Q \subseteq P$, and P is a superset of Q. A *dataset* is a collection of transactions, each transaction T is a set of attribute-values, i.e. $T \subset I$. The number of transactions in D is denoted by $|D|$. The *support* of an itemset P in dataset D, denoted by $support(P, D)$, is the transactions in D which contain P, divided by $|D|$ ($0 \leq support(P, D) \leq 1$).

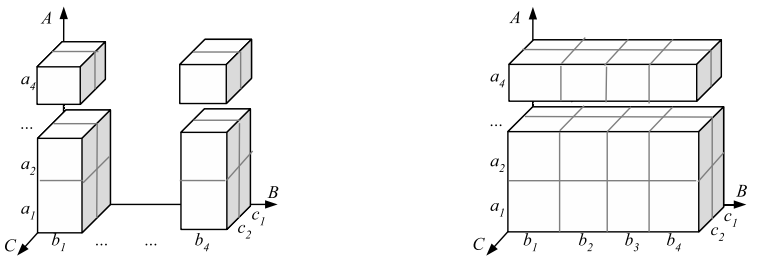
Assume two classes in dataset D, namely D_p (the positive class) and D_n (the negative class). The support ratio of an itemset between two classes, termed as *growth rate (gr)*: $gr(P, D_p, D_n) = \frac{support(P,D_p)}{support(P,D_n)}$. Each itemset is associated with a discriminating power (or *contrast strength*): $strength(P, D_p, D_n) = support(P, D_p) * \frac{gr(P,D_p,D_n)}{1+gr(P,D_p,D_n)}$. Given support thresholds α and β, an **Emerging Pattern (EP)** [3] is a simple contrast pattern, defined as an itemset P, s.t. $support(P, D_n) \leq \beta$ (i.e. infrequent in D_n), and $support(P, D_p) \geq \alpha$ (i.e. frequent in D_p). Moreover, P is a **minimal emerging pattern** if it does not contain other emerging patterns. A **Jumping Emerging Pattern (JEP)** is an EP which has an infinite growth rate. In the remainder of this paper we use the term *pattern* to refer to an emerging pattern. The *support* of a pattern refers to its support in the positive class.

A **Disjunctive Emerging Pattern (DEP)** is an itemset P which contains one or more items from the domain of every attribute, and satifies two support constraints: i) $support(P, D_p) \geq \alpha$, and ii) $support(P, D_n) \leq \beta$. E.g. Given a dataset with three attribute domains $\{a_1, a_2, a_3, a_4\}$, $\{b_1, b_2, b_3, b_4\}$, $\{c_1, c_2, c_3, c_4\}$, and $x = \{a_1, a_2, a_4, b_1, b_4, c_1, c_2\}$ is a DEP. DEPs express contrasts as conjunctions of disjunctions (CNF), where disjunctions are only allowed between items within attributes. The boolean function that x represents, denoted $f(x)$, is $(a_1 \vee a_2 \vee a_4) \wedge (b_1 \vee b_4) \wedge (c_1 \vee c_2)$. The dataset projection into multi-dimensional space considers x as a subspace (see Fig. 1a). Thus, we can calculate *support* by counting the transactions which are subsets of x.

For an attribute with an ordered domain, a set of adjacent items within the same dimension (or attribute) is called a *contiguous* itemset. We call a set of non-occuring items between two adjacent items (within the same dimension) as a *gap*. If the gap in each dimension of x is no larger than a given minimum threshold g, then we say that x is a *g-contiguous* itemset, where $0 \leq g \leq k-2$, k is the number of domain items for that attribute. Moreover, a disjunctive pattern is a **g-contiguous pattern** if it does not contain non g-contiguous itemsets in any of its dimensions. Consider Fig. 1, x is g-contiguous for $g \geq 2$, and y is g-contiguous for $g \geq 1$.

3 Classification by Significant Expressive Contrast Patterns

For the purpose of our study, we use the existing JEP-classifier framework [2] as a baseline, which is highly accurate for dense and large datasets. It is based on minimal JEPs, which are considered the most powerful JEPs for classifica-



(a) $x = \{a_1, a_2, a_4, b_1, b_4, c_1, c_2\}$ (b) $y = \{a_1, a_2, a_4, b_1, b_2, b_3, b_4, c_1, c_2\}$

Fig. 1. Geometric representations of disjunctive patterns

tion, since their supports are largest. To adapt the framework, our classifier uses maximal disjunctive patterns which have infinite growth rate. Given a test instance T, all patterns which contain T can be found from each class. Based on its distinguishing class frequencies, a JEP favors D_p over D_n. Each pattern which occurs in T makes a *contribution* to classify T as an instance of D_p, based on its *support*. The JEP classifier then chooses the class which has the highest total contribution to be the winner.

Since disjunctive patterns are relatively longer (i.e. contain more items) than the simple patterns, intuitively not every item makes an equally-high contribution to the contrast strength of a pattern. Thus, we propose two levels of significance testing: i) external significance: tests whether the pattern is highly associated with the class, ii) internal significance: tests whether each element in a pattern makes a significant contribution in the pattern's strength.

3.1 Statistical Fisher Exact Test and Externally Significant Patterns

Work in [6] showed that the Fisher Exact Test (FET) is useful for finding statistically significant association rules, which makes it potentially useful for contrast patterns as well. To test the significance of a pattern P, FET uses a 2x2 contingency table containing the support of P and its complemented support in each class (shown in Table 1). The test returns a p-value, which is a probability that the *null-hypothesis* should be accepted, i.e. there is no significant association between the pattern and the class. If the p-value is below the *significanceLevel* (typically 0.05), we reject the hypothesis and say P is **externally significant**. Given a contingency table $[a, b; c, d]$, and $n = a + b + c + d$. The p-value is computed by:

$$p([a, b; c, d]) = \sum_{i=0}^{min(b,c)} \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!(a+i)!(b-i)!(c-i)!(d+i)!} \quad (1)$$

3.2 Internally Significant Disjunctive Emerging Patterns

The testing methodology for significant association rules [6] tests whether the inclusion of each condition significantly contributes to the rule's association.

Table 1. The contingency table for testing the significance of association between pattern P and class C

D	P	$\neg P$
C	$a = support(P, C)$	$b = support(\neg P, C)$
$\neg C$	$c = support(P, \neg C)$	$d = support(\neg P, \neg C)$

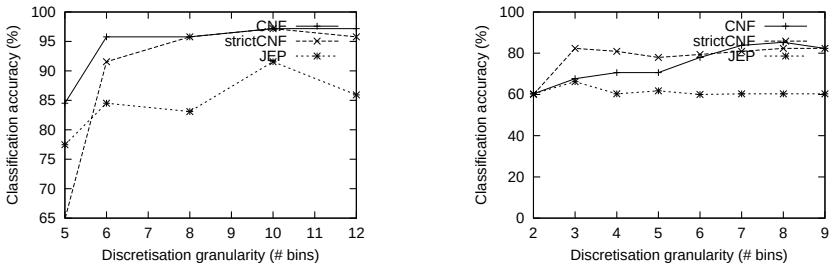
However, it was originally fashioned for purely conjunctive rules. To adapt the method for our needs, we use a *negative representation* of a disjunctive pattern, which is a pure conjunction of negated items. A pattern is significant if each of the negated items makes a significant contribution. This differs from previous work on significant association rules, which consider conjunctions of positive items, instead of negative items. E.g. The NNF (Negative Normal Form) representation of a disjunctive pattern x in Fig. 1a, denoted $f_N(x)$, is the conjunction of the non-occurring items: $f_N(x) = (\neg a_3) \wedge (\neg b_2 \wedge \neg b_3) \wedge (\neg c_3 \wedge \neg c_4)$.

Given ordered attribute domains, a disjunctive pattern can be projected to a subspace, possibly with some holes in it (correspond to *gaps*). Small holes may not be worth retaining if they contain very few data instances from the positive class. On the other hand, big holes may be necessary if they contain many data instances from the negative class. A gap is a *significant gap* if it passes the internal significance test. We call the generalisation of a pattern that is obtained by filling-in a gap as the *gap-filled generalisation*. A gap is maximal if it is not a subset of another gap. If all maximal gaps in a pattern are significant, then we say that the pattern is **internally significant**.

E.g. Reconsider pattern $x = \{a_1, a_2, a_4, b_1, b_4, c_1, c_2\}$. It contains three maximal gaps: $\neg\{a_3\}$, $\neg\{b_2, b_3\}$, $\neg\{c_3, c_4\}$. These correspond to the negative representation of x. The significance of a gap $\neg z$ is calculated between x and its generalisation (by inverting $\neg z$ to z). Let $z = \{b_2, b_3\}$. Let y be the gap-filled generalisation of x s.t. $y = x \cup \{b_2, b_3\} = \{a_1, a_2, a_4, b_1, b_2, b_3, b_4, c_1, c_2\}$. We can calculate the p-value using Eq. 1 and the contingency table in Table 1, by letting $P = \neg z = \neg\{b_2, b_3\}$, $C = D_p|y$, and $\neg C = D_n|y$, where $D_p|y$ (resp. $D_n|y$) refers to transactions in D_p (resp. D_n) which support y. A low p-value indicates the significance of gap $\neg\{b_2, b_3\}$ in x.

3.3 Classification by Significant Disjunctive Emerging Patterns

Our classifier is built based on the maximal disjunctive patterns which have an infinite growth rate. Using only those patterns, however, may overfit the training data. In real situations, there may be training instances which have significant association with the class, but are overlooked, due to the strict *infinite growth rate* constraint. To eliminate this problem, our classifier allows some limited constraint violation by filling-in the insignificant gaps, based on two criteria: i) the gap is not significant in the original pattern, and ii) the resulting gap-filled pattern is externally significant. Thus, all patterns which are used by the classifier are externally and internally significant. We refer to such patterns as **significant disjunctive patterns**.



(a) breast-cancer-w (dense) (b) horse-colic (dense)

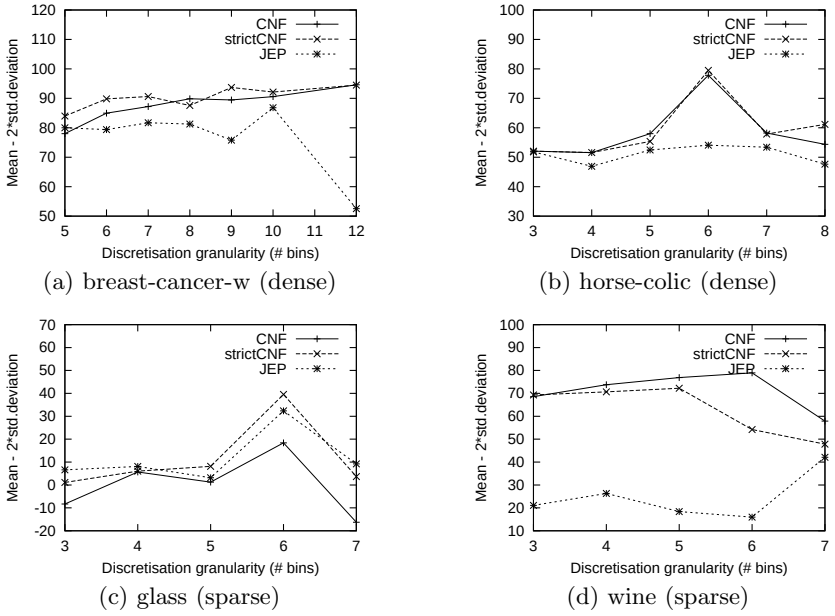
Fig. 2. Comparison of classification accuracy w.r.t. data discretisation's granularity

4 Experimental Results and Discussion

In this section, we study the performance of our classifier described in Section 3.3, based on significant disjunctive patterns, which we call *CNF-Classifier*. We will compare its classification performance against *strictCNF-Classifier*, which is also based on significant disjunctive patterns, but does not employ the significance testing (strictly imposing the support constraints on the patterns). As a baseline, we also use the Jumping Emerging Pattern Classifier (*JEPC*) [2]. The accuracy is based on 10-fold stratified cross validation. We use four data sets [5], which contain continuous attributes, and categorise them by their sparsity/density. The first two data sets are dense, namely *breast-cancer-w* and *horse-colic*, which contain two classes. The other data sets, *wine* and *glass*, contain multiple classes and are considered sparser. The *glass* data set is greatly imbalanced and extremely sparse, having 7 classes with only a few instances in each class.

Performance comparison with respect to discretisation granularity: In this experiment, we vary the number of bins (or discretised intervals) when discretising each data set using equal-density discretisation. Fig. 2 shows the classification accuracies from two data sets. In the *breast-cancer-w* data set, it is shown that the CNF-Classifier has the highest accuracy for all scenarios. Given finer granularties (i.e. more bins), strictCNF-Classifier and CNF-Classifier are able to outperform JEPC by 12% accuracy. In the *horse-colic* data set, the CNF-Classifier is more accurate than the strictCNF-Classifier when 6 or more bins are used, but it is less accurate otherwise. It shows that the significance test is useful when the data is finely discretised. The JEPC has the lowest accuracy in this data set.

Sensitivity of classification with respect to the support constraint: We now compare the sensitivity of the classifier w.r.t. to the minimum support of the contrast patterns. Fig. 3 shows the lower bound of the accuracy for various support thresholds, which is computed as *(mean - 2 st.dev)*, for each discretisation granularity. In the dense data sets, JEPC has the lowest lower bound and its accuracy greatly varies across the discretisation granularities. When a 12-bin discretisation was used for the *breast-cancer-w* data set, the JEPC has a mean accuracy of 79%, which indicates its large deviation or sensitivity w.r.t.



(a) breast-cancer-w (dense) (b) horse-colic (dense)

(c) glass (sparse) (d) wine (sparse)

Fig. 3. Comparison of mean - 2 st.dev of the classification accuracy (over various minimum support of the patterns) w.r.t. the discretisation granularities

the support constraint. (the relevant figure is not included in this paper, due to space limitation). The other classifiers, on the other hand, have mean accuracies of 96% and 99%, showing their low sensitivity w.r.t. the support constraint. In the sparse *glass* data set, the strictCNF-Classifier has a similar performance to JEPC, whereas the CNF-Classifier has a high sensitivity w.r.t the minimum support threshold and the data discretisation. In the less sparse *wine* data set, CNF-Classifier has the highest lower bound accuracy, and JEPC has the lowest lower bound.

Practical recommendations for users: Answering the questions posed at the beginning of this paper, we now present our recommendations:

When should disjunctions be allowed in contrast patterns for building a classifier? Disjunctions should be allowed in contrast patterns when the data is sparse, that is when the classes are imbalanced, or when the data is finely discretised, e.g. 8 bins or finer.

Which types of contrast patterns are most suitable for various data characteristics? When the data is sparse, expressive contrasts are more appropriate than simple contrasts. The significance test should be performed, except when the data is greatly imbalanced. Simple contrasts are useful for dense and coarsely discretised data sets.

5 Related Work

A contrast pattern is similar to a highly confident class association rule [7]. More expressive association rules have been studied [8], but they allow DNF (disjunction of conjunctions) rules, instead of CNF, which is the kind of rules considered in this paper. Our significance testing methodology could be extended for disjunctive association rules. Previous work on significant association rules [9, 6] only considers conjunctive rules. In an ordered domain, contiguous disjunctive patterns correspond to quantitative association rules [10], which are conjunctions of intervals of ordered values, however gaps are disallowed in a quantitative association rule. The negative representation of a disjunctive pattern in this paper is similar to a negative association rule [11], but the rule's antecedent contains only negative items, and the consequent contains a class label.

6 Conclusion and Future Work

In this paper, we investigated the advantages and disadvantages of using expressive (in the form of CNF combinations) contrast patterns in classification. We proposed a statistical testing for finding significant CNF patterns, which can also be adopted for disjunctive association rules or negative association rules. As our results suggest, expressive forms of patterns can be beneficial for classification, being less sensitive to the data sparsity. For future research, we would like to investigate their use in other types of classifiers and other data mining tasks.

References

1. Dong, G., Zhang, X., Wong, L., Li, J.: CAEP: Classification by aggregating emerging patterns. Discovery Science 1999 (1999) 30–42
2. Dong, G., Li, J., Ramamohanarao, K.: Making use of the most expressive jumping emerging patterns for classification. KAIS **3** (2001) 131–145
3. G.Dong, Li, J.: Efficient mining of emerging patterns: Discovering trends and differences. In: Proc. of KDD'99. (1999) 43–52
4. Loekito, E., Bailey, J.: Fast mining of high dimensional expressive contrast patterns using ZBDDs. In: Proc. of KDD'06. (2006) 307–316
5. Murphy, P.M., Aha, D.W.: UCI repository of machine learning databases. University of California, Dept. of Information and Computer Science, Irvine, CA
6. Verhein, F., Chawla, S.: Using significant, positively associated and relatively class correlated rules for associative classification of imbalanced datasets. In: Proc. of ICDM'07. (2007) 679–684
7. Liu, B., Hsu, W., Ma, Y.: Integrating classification and association rule mining. In: Proc. of KDD'98. (1998) 80–86
8. Navati, A.A., Chitrapura, K.I., Joshi, S., Krishnapuram, R.: Mining generalised disjunctive association rules. In: Proc. of CIKM'01. (2001) 482–489
9. Webb, G.I.: Discovering significant rules. In: Proc. of KDD'06. (2006) 434–443
10. Srikant, R., Agrawal, R.: Mining quantitative association rules in large relational tables. In: Proc. of SIGMOD'96. (1996)
11. Wang, H., Zhang, X., Chen, G.: Mining a complete set of both positive and negative association rules from large databases. In: Proc. of PAKDD'08. (2008) 27–38