Pattern Teams

Arno J. Knobbe[1,2], Eric K. Y. Ho[1]

[1]Kiminkii, Postbus 171, NL-3990 DD, Houten, The Netherlands
a.knobbe@kiminkii.com, e.ho@kiminkii.com
[2]Utrecht University, P.O. box 80 089, NL-3508 TB Utrecht, the Netherlands

Abstract Pattern discovery algorithms typically produce many interesting patterns. In most cases, patterns are reported based on their individual merits, and little attention is given to the interestingness of a pattern in the context of other patterns reported. In this paper, we propose filtering the returned set of patterns based on a number of quality measures for pattern sets. We refer to a small subset of patterns that optimises such a measure as a *pattern team*. A number of quality measures, both supervised and unsupervised, is proposed. We analyse to what extent each of the measures captures a number of 'intuitions' users may have concerning effective and informative pattern teams. Such intuitions involve qualities such as independence of patterns, low overlap, and combined predictiveness.

1 Introduction

Over the last few years, there has been an increasing interest in pattern discovery algorithms. In this branch of Data Mining, the emphasis lies on discovering a collection of local patterns that satisfy a number of inductive constraints provided by the user, rather than on the induction of a single global model of the data. Typically, a pattern represents some subgroup of the data, and patterns are selected on the basis of the support of the subgroup and one or more constraints on interestingness measures, for example based on the correlation with a target concept. Common examples of such Data Mining settings are frequent pattern discovery (itemsets, trees, graphs, etc.) [11], association rule mining [12] and subgroup discovery [1, 3, 9, 10, 12]. In most cases, patterns are reported based on their individual merits, and little attention is given to the interestingness of a pattern in the context of other patterns reported. As a result, the outcome of a pattern discovery exercise is often a large collection of patterns, with high levels of redundancy, that is hard to inspect manually. It is our aim in this paper to improve the effectiveness of pattern discovery algorithms by considering the quality of patterns in the context of other patterns reported.

Let us consider a busy end-user, who has only very limited time to inspect the outcome of a pattern discovery exercise. If only a few patterns can be considered, a small yet effective set of patterns needs to be selected. Having already seen one or more patterns, the next pattern presented needs to both perform well, and be substantially different from the first patterns. If the next pattern effectively covers almost the same set of individuals as any combination of the previous ones (even though syntactically it might be completely different), it provides little new information. We present a method of selecting a small subset of patterns – a *pattern*

team – that optimises some given quality measure for sets of patterns. In this paper we consider four candidate quality measures that promote different desirable properties of sets of patterns. Depending on the (implicit) expectations of the end-user, different methods of pattern selection can thus be applied, resulting in different pattern teams.

The quality measures presented are inspired by a number of intuitions end-users typically have about pattern mining results. We use the following list of intuitions, and test to what extent the four measures satisfy these intuitions:

I1 No two patterns should cover (approximately) the same set of examples.

I2 No pattern should cover (approximately) the complement of another pattern.

I3 No pattern should cover (approximately) a logical combination of two or more other patterns.

I4 Patterns should be (approximately) mutually exclusive.

I5 When using patterns as input to a classifier, the pattern set should lead to the best performing classifier.

I6 Patterns should lie on the convex hull of all patterns in ROC-space.

Clearly, these six intuitions cannot all hold at the same time. In fact, some intuitions are to some extent competing (e.g. I2 and I4). One should think of these intuitions as descriptions of the kind of expectations an end-user may have about the set of patterns returned. Typically, we will only be interested in satisfying one, or a few of these intuitions.

An important characteristic of the four quality measures presented is that they are syntax-independent (although one can envisage more syntax-oriented measures). This means that pattern sets are solely judged on the subgroups of individuals covered by each pattern, and the potential overlap or independence of these subgroups. As a result, our methodology applies to any mining paradigm where patterns represent subgroups. This includes complex domains such as structured and multi-relational domains, where rich pattern languages make purely syntactical comparisons of patterns difficult or expensive.

An extended version of this paper can be found in [5]. This version contains more examples, experiments and extended descriptions of related work.

2 Pattern Team Discovery

The process of filtering patterns will generally be preceded by a pattern discovery phase. The *pattern discovery* task can be defined as follows: given a pattern collection $\boldsymbol{P}$, a set of interestingness measures $\phi_1, \ldots, \phi_l$, $\phi_i : \boldsymbol{P} \rightarrow [0, 1]$, and a set of threshold values $\sigma_1, \ldots, \sigma_l$, $\sigma_i \in [0, 1]$, find all patterns $p \in \boldsymbol{P}$ such that $\forall i : \phi_i(p) \geq \sigma_i$. In an alternative setting, the top k patterns with respect to one of the interestingness measures is returned.

In both settings, the outcome will typically be a large set of interesting patterns P_ϕ with considerable levels of redundancy. In this paper, we therefore propose a second phase, consisting of a *pattern team discovery* task: given a set of interesting patterns P_ϕ, and a quality measure for pattern sets $\Phi : 2^{\boldsymbol{P}} \rightarrow \mathbf{R}$, find a pattern set $P \subseteq P_\phi$ of size k such that $\Phi(P) \geq \Phi(Q)$ for all $Q \subseteq P_\phi$ of size k.

In this case, we are interested in a single pattern team of specified size k, but other variants of this task can be imagined. For example, one could query for *all* pattern teams of size k, or for an optimal pattern team regardless of its size. Alternatively, one

can imagine a discovery task that returns all (or the top k) pattern sets P such that their quality exceeds some threshold: $\Phi(P) \geq \Sigma$. This setting however defies the purpose of pattern filtering, as potentially many pattern sets will be returned.

Finding a pattern team of size k for any given pattern set quality measure Φ potentially involves the consideration of $\binom{n}{k}$ subsets of P_ϕ, where $n = |P_\phi|$. In fact, Mielikäinen et al. [7] show that the general pattern team discovery problem is NP-hard by relating it to the set-covering problem, making it infeasible for all but small values of k. Fortunately, for specific quality measures, it is possible to find optimal pattern sets efficiently, or to find approximations that can be shown to perform reasonably well [7]. In [4], we provide some example algorithms for *joint entropy*, one of the quality measures considered in this paper. The thorough treatment of efficient implementations of pattern team discovery is outside the scope of this paper.

3 Quality Measures

In this section, we present four quality measures for pattern sets. Two of the four measures work in an unsupervised fashion: they consider properties of the subgroups covered (specifically independence and mutual exclusiveness), and ignore the potential predictive qualities of a pattern. Note that these two measures work on pattern sets discovered by algorithms working in either supervised or unsupervised fashion. Conversely, we could create a pattern team using one of the two supervised quality measures on patterns discovered in unsupervised mode (e.g. frequent patterns). We thus have a choice for supervised or unsupervised both for the initial pattern discovery phase, as well as the subsequent pattern filtering phase.

The following basic definitions will allow us to define the four quality measures for pattern sets formally. We assume that our database d is a bag of labelled objects $i \in D$, referred to as *individuals*, taken from a domain D. Furthermore there is a function l: $d \rightarrow \mathbf{R}$ that specifies the label of an individual. If the labels just have values 0 and 1, we interpret the individuals as belonging to the negative (F) and positive (T) classes, respectively. Alternatively, we treat the mining task as a regression problem. We refer to the size of the database as $N = |d|$.

We assume nothing about the syntax of the pattern language, and treat a pattern simply as a function p: $D \rightarrow \{0, 1\}$. We will say that a pattern p *covers* an individual i *iff* $p(i) = 1$. A *subgroup* $S(d, p)$ implied by a pattern is now simply the set of individuals $i \in d$ that are covered by p: $S(d, p) = \{i \in d \mid p(i) = 1\}$. For brevity we will omit the d from now on. $s(p) = |S(p)|$ refers to the size of the subgroup implied by p. Furthermore, we will use expressions like $l(i) = 1$ to denote patterns related to the label of individuals, such that $S(l(i) = 1)$ for example denotes the set of positive cases.

When talking about sets of patterns $P = \{p_1, \ldots, p_k\}$ of size k, an individual may be covered by some patterns in P and not by others. In order to represent such *contingencies*, we introduce *codes* $c \in \{0, 1\}^k$. The subgroup implied by a given pattern set P and a code c is defined by $S(P, c) = \{i \in d \mid p_1(i) = c_1, \ldots, p_k(i) = c_k\}$. $s(P, c) = |S(P, c)|$ is the size of the subgroup implied by P and c.

Joint Entropy The first quality measure for pattern teams is based on our work on maximally informative k-itemsets (miki's) [4]. In essence, we treat each pattern in P_ϕ

as a binary feature (an item), and for each pattern set $P \subseteq P_\phi$ of size k, compute the joint entropy [4] of the features in P. A miki is then simply the itemset (pattern set) that optimises this joint entropy. Our first quality measure Joint Entropy $H(P)$ is hence defined as:

$$H(P) = -\sum_{c \in \{0,1\}^k} \frac{s(P,c)}{N} \lg \frac{s(P,c)}{N}$$

The entropy is a measure for the uniformity of the distribution of individuals over the different contingencies. A uniform distribution is achieved if for all patterns $s(p) = N / 2$, and all patterns are independent. A pattern team that optimises the joint entropy will hence optimise the power to distinguish between individuals. Note that H is unsupervised, as well as insensitive to replacing one or more patterns with their complement [4]. Patterns are merely used to distinguish two complementary sets of individuals.

Exclusive Coverage The second quality measure is inspired by intuition I4: pattern sets that reduce the amount of overlap between patterns are favoured. Because overlap is less likely with patterns of low support, we will also have to promote the support of individual patterns. The Exclusive Coverage $EC(P)$ quality measure counts the coverage that is exclusive for each pattern, and is defined as follows:

$$EC(P) = \sum_i \left(s(p_i) - \left| S(p_i) \cap \bigcup_{j \neq i} S(p_j) \right| \right)$$

Note that this measure counts the coverage of subgroups that correspond to the codes that contain only a single 1.

DTM Accuracy Unlike the first two measures, the third quality measure is supervised: it determines the quality of a pattern team on the basis of how well a simple classifier is able to predict the label of individuals, given the patterns as feature set. Finding the optimal pattern team hence amounts to selecting a pattern set by means of a *wrapper approach* [2].

The classifier of choice is the *Decision Table Majority* classifier [6, 8], also known as a *simple decision table*. The idea behind this classifier is to build from the pattern set a contingency table for each possible code, and compute the relative frequency of positive cases for each contingency. For contingencies that do not appear in the database, the relative frequency of positive cases is based on that of the whole database (i.e. the prior). An individual is now classified by computing its code, and returning the majority class within the associated subgroup. This simple approach works surprisingly well, under two conditions: the features (i.e. patterns) have a low cardinality, and the decision table should be based on a relatively small number of features selected from a larger set by means of a wrapper [6]. These conditions clearly hold for our application. The following definition captures the workings of a DTM classifier. The function f computes a conditional probability estimate for $l(i) = 1$ for a code c, given a set of patterns P:

$$f(P,c) = \frac{|\, S(P,c) \cap S(l(i)=1) \,|}{s(P,c)}, \text{ if } s(P, c) > 0,$$

$$f(P,c) = \frac{s(l(i)=1)}{N}, \text{ if } s(P, c) = 0.$$

The DTM Accuracy $Acc(P)$ quality measure uses the DTM classifier to determine how predictive a pattern set is by computing the accuracy of the classifier by means of cross-validation. This will reduce the risk of choosing a pattern set that over-fits. As a more efficient alternative, one might consider the purity (the accuracy based on in-sample testing) of the DTM classifier as a quality measure. Informal experimentation has shown that results thus obtained are very close to the cross-validated accuracy.

It is important to note that instead of a DTM classifier, any classifier can be applied. Furthermore, obtaining a good classifier is not our primary goal: we are merely using a classifier in order to obtain a well-performing pattern team.

Area Under Curve The Area Under Curve $AUC(P)$ quality measure computes the area of the convex hull of the patterns in P in ROC-space [1]. The quality measure is computed by plotting the patterns in P in ROC-space, along with the points (0, 0), (1, 0), (1, 1), and computing the area of the convex hull of these $k + 3$ points.

4 Intuitions and Quality Measures

The different quality measures introduced in the previous section capture different aspects of pattern sets. In this section, we examine how these measures fit the intuitions introduced in Section 1. Furthermore, we analyse the correlation between measures, and hence to what extent measures capture similar qualities of pattern sets. As the quality measures are chosen such that they capture at least one intuition perfectly, we can use the correlations between measures to understand how they map to intuitions. If a certain measure is uncorrelated with another measure that is designed to fit a particular intuition, then this first measure cannot be useful for said intuition. The following experiment will support our discussion of quality measures.

Experiment The database under consideration is the multi-relational database Mutagenesis [5]. It contains structured descriptions of 188 molecules that fall in two classes: mutagenic (66.5%) and non-mutagenic. Although multiple versions of the database exist, with various amounts of information about the molecular structure and properties of the molecules and atoms, we will use a version that contains the basic molecular structure, as well as two numeric attributes on the molecule level (Lumo and LogP). Additionally we have added two aggregated attributes on the molecule level, describing the number of atoms and the number of distinct elements. Hence there is a certain level of redundancy in the database, which may lead to different patterns capturing more or less similar properties of the molecules. Furthermore, the availability of multiple numeric attributes allows for a large range of decision boundaries, which should lead to redundancy in the patterns, as well as moderate variations of patterns.

Predictive patterns were discovered using Safarii in supervised mode. Subgroups were discovered using the absolute value of the novelty (a.k.a. weighted relative accuracy) interestingness measure, and a minimum support threshold of 5%. Relatively moderate search conditions were used in order to arrive at a manageable result set. The outcome is a collection of 51 multi-relational patterns describing subgroups that show a substantial deviation in mutagenicity, either positive or negative. The original database, as well as the propositionalised version of the 51 binary features can be obtained from the authors.

These results lead to the following conclusions. The results are also summarised in Figure 1. Each cell describes how useful a particular measure is for a given intuition. If a particular quality measure was defined with a specific intuition in mind, the corresponding cell is coloured grey. For supporting data, see [5].

Joint Entropy The quality measure *H* clearly captures Intuitions I1 to I3 [4]. If two patterns cover almost the same subgroup, having both of them in the pattern team will not give a significant improvement in uniformity of the distribution over just having one of the two (I1), and hence *H* will favour pattern sets with more diversity among the patterns. The same holds for patterns that follow directly from multiple other patterns (I3). As *H* is insensitive to replacement with complementary patterns, I2 applies. *H* and EC turn out to be uncorrelated, and hence a pattern team optimised with respect to *H* can not be expected to satisfy I4.

Surprisingly, patterns teams optimised for *H* perform reasonably well for classification, despite the unsupervised nature of this measure. Therefore, we can say that *H* captures I5 to a reasonable degree. *H* and *AUC* are uncorrelated, and hence *H* is not a good measure for finding pattern sets on the convex hull in ROC space (I6).

Exclusive Coverage The measure *EC* penalises overlap, and thus having two similar patterns is unlikely. *EC* therefore satisfies to some extent I1, with the exception of patterns with low support (and thus low penalty) that sometimes appear in copies. I2 and I3 however are not satisfied, because complements and logical redundancy are promoted. Clearly I4 is satisfied. The performance of a DTM classifier is unrelated to the patterns being mutually exclusive: *EC* does not satisfy I5. The same holds for I6.

DTM Accuracy The measure *Acc* correlates quite well with *H*, and hence captures I1 to I3 moderately well. This should be no surprise, as redundancy in the pattern set cannot benefit the classification score. It turns out that *Acc* in general does not provide mutually exclusive pattern sets (I4). Clearly I5 is satisfied. The experiment shows a very slight correlation between the classification score and the area under curve (I6). At least, poorly performing pattern sets consist of patterns below the convex hull in ROC space. A higher correlation might be expected if only positive patterns would have been produced in the initial discovery phase, as many predictive patterns now appear below the diagonal.

Area Under Curve As follows from the discussion above, the *AUC* measure is really only useful for I6. A pattern team consisting of patterns on the convex hull apparently is not very useful as input to a classifier. The only purpose of such a pattern team would therefore be to provide patterns that are optimal individually, rather than as a team.

	Joint Entropy	Exclusive Coverage	DTM Accuracy	AUC
Intuition 1	very high	moderate	high	
Intuition 2	very high		high	
Intuition 3	very high		high	
Intuition 4		very high		
Intuition 5	high		very high	low
Intuition 6			low	very high

Figure 1. Informal analysis of how well the different quality measures fit the six intuitions.

5 Related Work

The domain of feature selection [2, 6, 8] provides good inspiration for pattern filtering techniques, since every pattern in our view can be interpreted as a virtual binary feature. When selecting a feature selection technique, one has to make sure that one or more of our intuitions is satisfied. Many feature selection methods consider the quality of individual features, for example based on correlation with the target concept, and thus potentially produce redundant feature sets. Selection techniques that do consider the value of features in the context of others are more precisely referred to as *feature subset selection* techniques. Wrapper methods are good examples of such techniques [6, 8]. For an overview, see [2].

A domain concerned with the production of a concise set of interesting patterns is known as Subgroup Discovery [1, 3, 9, 10, 12]. The typical approach is to define an interestingness measure, often related to correlation with the target, and then find the top k patterns with respect to this measure. Unfortunately, such techniques often do not consider potential redundancy, and therefore suffer from the same limitations as many feature selection methods. Zimmermann et al. [12] describe a method called CorClass for finding the top k predictive association rules, based on interestingness measures such as novelty, information gain or X^2. The convexity of such measures can be used to find the best rules efficiently. However, due to the redundancy among these rules, relatively high values of k ([12] proposes $k = 1000$) are needed to at least include the essential dependencies required for obtaining good predictive scores. A range of well-known rule combination strategies is used.

For more related work, see [5].

6 Conclusions and Further Work

We have presented a method for reducing the number of patterns returned to the user by a pattern discovery algorithm. The method works by selecting from the (potentially large) collection of patterns deemed interesting by the discovery algorithm a small set of patterns that optimises some quality function for pattern sets. We refer to such an optimal set of patterns of specific size as a *pattern team*. By only allowing a small number of patterns in the pattern set, and selecting the right quality measure, the resulting pattern team reduces the amount of redundancy between

patterns, while retaining as much of the information captured by the patterns as possible. We have presented four measures that capture different qualities of pattern sets. Two unsupervised measures, *Joint Entropy* and *Exclusive Coverage*, promote independence or reduce overlap, respectively. The remaining two supervised measures, *DTM Accuracy* and *Area Under Curve*, are based and how well the pattern set performs as input to a simple classifier, and how well it performs as a collection of points in ROC-space, respectively. Initial experimentation shows that Joint Entropy and DTM Accuracy produce the most useful results, and satisfy a number of intuitive expectations of end-users concerning non-redundancy and predictive quality of the patterns returned.

We have implemented the proposed pattern team discovery scheme in the Safarii system. Space limitations unfortunately prevent a detailed description of algorithmic aspects of the efficient computation of pattern teams. Interesting optimisations over naïve implementations can however be obtained for the quality measures presented [4], and we intend to extend our work in this direction. Furthermore, alternative quality measures can be thought of. Apart from quality measures on the level of pattern sets, one could envisage selecting pattern sets on the basis of inductive queries based on relationships between its member patterns, for example by requiring a certain amount of dissimilarity between every pair of patterns. A pattern team would thus optimise a given quality measure, as well as satisfy a number of inductive constraints.

References

1. Fürnkranz, J., Flach, P., *ROC 'n' Rule Learning – Towards a Better Understanding of Covering Algorithms*, Machine Learning, 58, 39–77, Springer, 2005
2. Guyon, I., Elisseeff, A., *An Introduction to Variable and Feature Selection*, Journal of Machine Learning Research 3, 1157-1182, 2003
3. Knobbe, A.J., *Multi-Relational Data Mining*, Ph.D. dissertation, 2004, http://www.kiminkii.com/thesis.pdf
4. Knobbe, A.J., Ho, E.K.Y, *Maximally Informative* k-*Itemsets and their Efficient Discovery*, in Proceedings of KDD 2006, 2006
5. Knobbe, A.J., Ho, E.K.Y, *Pattern Teams*, long version, 2006, http://www.kiminkii.com/publications.html
6. Kohavi, R., *The Power of Decision Tables*, In Proceedings of ECML '95, 1995
7. Mielikäinen, T., Mannila, H., *The Pattern Ordering Problem*, In Proceedings PKDD 2003, LNAI 2838, 2003
8. Pfahringer, B., *Compression-Based Feature Subset Selection*, In Proceedings of IJCAI '95, 1995
9. *Safarii Multi-Relational Data Mining Environment*, http://www.kiminkii.com/safarii.html, 2006
10. Scheffer, T., Wrobel, S., *Finding the Most Interesting Patterns in a Database Quickly by Using Sequential Sampling*, Machine Learning Research 3, 2002
11. Yan, X., Cheng, H., Han, J, Xin, D., *Summarizing Itemset Patterns: A Profile-Based Approach*, In Proceedings KDD'05, 2005
12. Zimmermann, A., De Raedt, L., *CorClass: Correlated Association Rule Mining for Classification*, In Proceedings DS 2004, 2004