A Novel Approach For Mining Emerging Patterns in Rare-class Datasets

Hamad Alhammady
Etisalat University College, UAE
Email: hamad@euc.ac.ae

Abstract - Mining emerging patterns (EPs) in rare-class databases is one of the new and difficult problems in knowledge discovery in databases (KDD). The main challenge in this task is the limited number of rare-class instances. This scarcity limits the number of emerging patterns that can be mined for the rare class. In this paper, we propose a novel approach for mining emerging patterns in rare-class datasets. We experimentally prove that our method is capable of gaining enough knowledge from the rare class; hence, it increases the performance of EP-based classifiers.

I. INTRODUCTION

The rare-class problem is faced in many real life applications. These applications include direct marketing, web log analysis, and intrusion detection in security networking systems. The main challenge in this problem is the scarcity of the rare cases. This challenge prevents classification methods from gaining enough knowledge from the rare class.

In this paper, we introduce a new method for mining emerging patterns (EPs) in rare-class datasets. EPs are a new kind of patterns introduced recently [2]. They have been proved to have a great impact in many applications [1] [5] [6] [8] [9]. EPs can capture significant changes between datasets. They are defined as itemsets whose supports increase significantly from one class to another. The discriminating power of EPs can be measured by their growth rates. The growth rate of an EP is the ratio of its support in a certain class over that in another class. Usually the discriminating power of an EP is proportional to its growth rate.

For example, the Mushroom dataset, from the UCI Machine Learning Repository [7], contains a large number of EPs between the poisonous and the edible mushroom classes. Table 1 shows two examples of these EPs. These two EPs consist of 3 items. e1 is an EP from the poisonous mushroom class to the edible mushroom class. It never exists in the poisonous mushroom class, and exists in 63.9% of the instances in the edible mushroom class; hence, its growth rate is ∞ (63.9 / 0). It has a very high predictive power to contrast edible mushrooms against poisonous mushrooms. On the other hand, e2 is an EP from the edible mushroom class to the poisonous mushroom class. It exists in 3.8% of the instances in the edible mushroom class, and in 81.4% of the instances in the poisonous mushroom class; hence, its growth rate is 21.4 (81.4 / 3.8). It has a high predictive power to contrast poisonous mushrooms against edible mushrooms.

TABLE I
EXAMPLES OF EMERGING PATTERNS

EP	Support in poisonous mushrooms	Support in edible mushrooms	Growth rate
e1	0%	63.9%	∞
e2	81.4%	3.8%	21.4

e1 = {(ODOR = none), (GILL_SIZE = broad), (RING_NUMBER = one)}
e2 = {(BRUISES = no), (GILL_SPACING = close), (VEIL_COLOR = white)}

II. RELATED WORK

Traditional classification accuracy (percentage of correctly classified instances in all classes) is not a suitable metric to measure the performance of classifiers in the rare-class problem. For example, suppose we have an imbalanced dataset consisting of two classes, and the major class contributes 95% of the instances. Then, the traditional accuracy can be increased to at least 95% by assuming all data instances belong to the major class.

TABLE II
CONFUSION MATRIX

	Classified as major class	Classified as rare class
Actual Major class	*TM*	*FR*
Actual Rare class	*FM*	*TR*

TM = number of major-class instances classified as major-class instances
FR = number of major-class instances classified as rare-class instances
FM = number of rare-class instances classified as major-class instances
TR = number of rare-class instances classified as rare-class instances

According to the confusion matrix in table 2, the traditional accuracy is defined as follows.

$$Accuracy = \frac{TM + TR}{TM + FR + FM + TR} \quad (1)$$

This accuracy is dominated mainly by the performance of classifiers on the major class. This is because of the large ratio between the number of major-class instances and the number of rare-class instances in the training set. The *F-measure* (F) [10] is a suitable alternative metric to evaluate

T. Sobh (ed.), Innovations and Advanced Techniques in Computer and Information Sciences and Engineering, 207–211.

classifiers in rare-class classification. This metric evaluates a classifier based on both *precision* (P) and *recall* (R) as follows.

$$P = \frac{TR}{TR + FR} \tag{2}$$

$$R = \frac{TR}{TR + FM} \tag{3}$$

$$F = \frac{2PR}{P + R} \tag{4}$$

The F-measure has been used to evaluate a number of classification methods designed specifically for rare-class problems.

There is a number of techniques proposed for rare-class problems. On of these techniques is EPRC [3]. This approach is based on applying some improving stages to maximize the discriminating power of rare-class EPs. These stages include generating new undiscovered rare-class EPs, pruning low-support EPs, and increasing the support of rare-class EPs.

EPDT [4] aims at supporting decision trees in rare-class problems. It consists of two steps. First, new non-existing rare-class instances are generated. Second, the most important rare-class instances are over sampled. These two steps increase the performance of decision trees as they work together toward balancing the rare class with the major class.

Two-phase rule induction (PNrule) [11] tries to find the best tradeoff between recall and precision to achieve the highest possible f-measure. It consists of two phases. In the first phase it seeks high recall objective using P-rules. These P-rules detect the presence of the target class. In the second phase the technique seeks high precision objective using N-rules. These N-rules detect the absence of the target class. The P-rules, mined in the first phase, are not accurate. The reason is that they cover many major-class instances beside the rare-class instances. This is because the high interference between the major and rare classes due the scarcity of the rare-class. This problem affects the f-measure negatively.

III. EMERGING PATTERNS AND CLASSIFICATION

Let $obj = \{a_1, a_2, a_3, \ldots a_n\}$ is a data object following the schema $\{A_1, A_2, A_3, \ldots A_n\}$. $A_1, A_2, A_3 \ldots A_n$ are called attributes, and $a_1, a_2, a_3, \ldots a_n$ are values related to these attributes. We call each pair (attribute, value) an item.

Let I denote the set of all items in an encoding dataset D. *Itemsets* are subsets of I. We say an instance Y contains an itemset X, if $X \subseteq Y$.

Definition 1. Given a dataset D, and an itemset X, the support of X in D, $s_D(X)$, is defined as

$$s_D(X) = \frac{count_D(X)}{|D|} \tag{5}$$

where $count_D(X)$ is the number of instances in D containing X.

Definition 2. Given two different classes of datasets D_1 and D_2. Let $s_i(X)$ denote the support of the itemset X in the dataset D_i. The growth rate of an itemset X from D_1 to D_2, $gr_{D_1 \to D_2}(X)$, is defined as

$$gr_{D_1 \to D_2}(X) = \begin{cases} 0, & \text{if } s_1(X) = 0 \text{ and } s_2(X) = 0 \\ \infty, & \text{if } s_1(X) = 0 \text{ and } s_2(X) \neq 0 \\ \frac{s_2(X)}{s_1(X)}, & \text{otherwise} \end{cases} \tag{6}$$

Definition 3. Given a growth rate threshold $\rho > 1$, an itemset X is said to be a ρ-emerging pattern (ρ-EP or simply EP) from D_1 to D_2 if $gr_{D_1 \to D_2}(X) \geq \rho$.

Let $C = \{c_1, \ldots c_k\}$ is a set of *class labels*. A *training dataset* is a set of data objects such that, for each object *obj*, there exists a class label $c_{obj} \in C$ associated with it. A *classifier* is a function from attributes $\{A_1, A_2, A_3, \ldots A_n\}$ to class labels $\{c_1, \ldots c_k\}$, that assigns class labels to unseen examples.

IV. MINING EPS IN RARE-CLASS DATASETS

The major problem in mining EPs in rare-class datasets is that the number of the rare-class EPs is very small compared to the major-class EPs. Work in [3] aims at solving this problem by generating additional EPs. In this paper, we propose a novel method for mining a large number of rare-class EPs to fill the gap between the rare class and the major class.

First, let us investigate the main reason behind the shortage in rare-class EPs. Mining EPs involves some sort of comparison between the small population in the rare class and the large population in the major class. That is, the mining process aims at finding patterns that exist frequently in the small number of rare-class instances and that do not exist very frequently in the large number of major-class instances. This difficult restriction limits the number of rare-class EPs because the rare-class instances are compared at the same time with all the major-class instances.

Our proposed approach is based on mining rare-class EPs by comparing the rare-class instances with subsets of the major-class instances instead of the whole range of data.

The details of our approach are as follows. Suppose that the rare class (RC) and the major class (MC) consist of R and M instances, respectively. The major class is divided into a number of subsets, MS_j such as $j = \{1, \ldots, M/R\}$ and the number of instances in each subset is R. Rare-class EPs are mined from RC against each subset of MC. That is the mining process is divided into M/R sub processes rather than one as in the normal case. The results of each sub process are a reasonable number of rare-class EPs because the number of

instances in *RC* and each subset MS_j is identical. These rare-class EPs are distributed in *M/R* sets ($REPS_j$) each of which is related to one of the mining sub processes.

The rare-class EPs in the *M/R* sets are combined in one set, the EPs are then ranked in a descendent order according to their strength. The strength of an EP *e*, *strg(e)*, is defined as follows.

$$strg(e) = \frac{gr(e)}{1+gr(e)} * s(e) \quad (7)$$

The strength of an EP is proportional to both its growth rate (discriminating power) and support. Notice that if an EP has a high growth rate and a low support its strength might be low. In addition, if it has a low growth rate and a high support its strength might also be low.

Suppose that the number of major-class EPs is *MEP*. The set of ranked rare-class EPs is divided into two subsets. The first subset (called the final subset) contains *MEP* strongest rare-class EPs. That is, the number of final rare-class EPs equals the number of major-class EPs. The second subset (called the pending subset) contains the remaining rare-class EPs that are not yet included in the final subset.

The final stage of our approach involves comparing the final subset of rare-class EPs with the major-class EPs. If an EP exists in both the final subset of rare-class EPs and the major-class EPs, then it is eliminated from the final subset and the strongest EP in the pending subset is added to the final subset. This process ensures that noisy rare-class EPs are eliminated from the final subset and the strongest EPs are added to this subset.

- After completing our approach, we end up with two sets of EPs. On of them is related to the major class and the other is related to the rare class. These two sets have almost the same number of EPs. That is, rare-class EPs are not rare compared to the major-class EPs. The two sets can be used with any EP-based classifier (such as BCEP [5]) to classify unlabeled instances in both classes. Figure 1 sketches our proposed method.

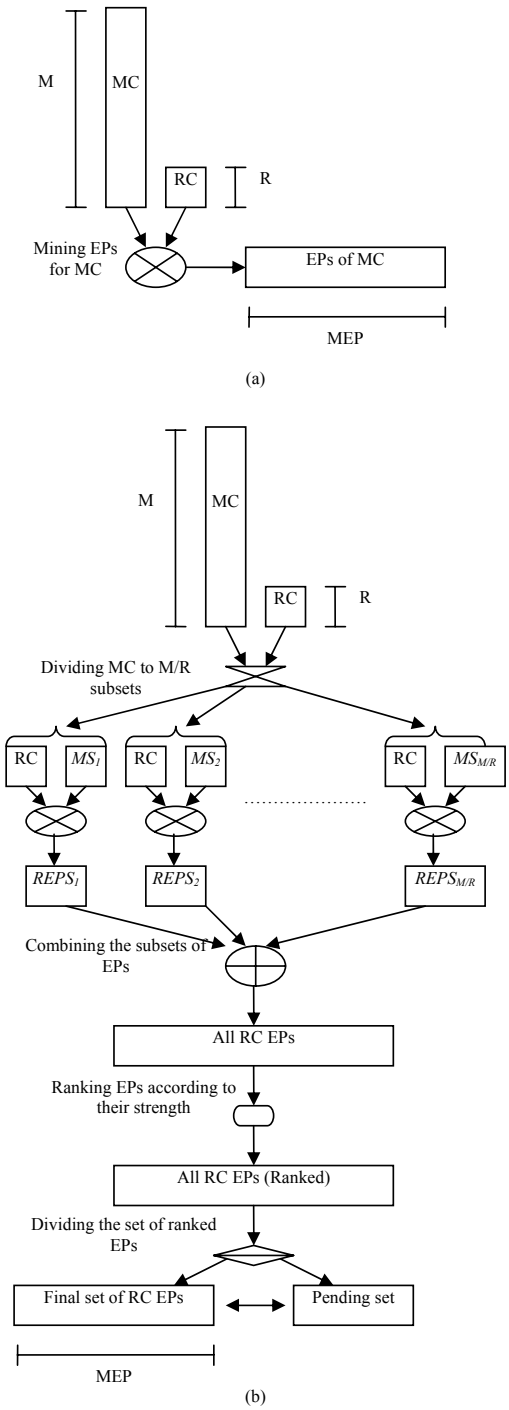

Figure1. (a) Mining major-class EPs. (b) Mining rare-class EPs using DEP.

Rare-class EPs mining using our method (we call it division for mining EPs - DEP) are motivated by the following points:

- Dividing the major class into subsets enables the discovery of unseen rare-class EPs. These unseen EPs are covered by the overwhelming amount of data in the major class.
- Using the strength function to evaluate the rare-class EPs ensures that noisy EPs have minimum effect.

V. EXPERIMENTAL EVALUATION

We conduct experiments on 12 datasets from UCI repository of machine learning databases [7]. These datasets are disease, hypothyroid, sick-euthyroid, and nine binary datasets formed from the king-rook-king dataset[1]. Table 3 lists these datasets and the percentage of the rare-class for each one of them. We compare our proposed method (DEP) with Pnrule [11], boosted PNrule, EPRC [3], and EPDT [4]. The results are shown in table 4[2].

The following points summarize the results:

- Our proposed method, DEP, outperforms all the other methods on all datasets.
- DEP has the highest average F-measure.

VI. CONCLUSIONS

Mining emerging patterns (EPs) in rare-class datasets is one of the challenging problems in data mining. This problem is considered as the main reason behind the failure of EP-based classifiers in the rare-class classification. In this paper, we propose a new technique for mining EPs in rare-class datasets. Our proposal is based on dividing the mining process into a number of sub processes and then combining the resulted EPs according to their strength. We experimentally prove that our method is effective in rare-class classification.

TABLE III
RARE-CLASS DATABASES

Dataset	Percentage of the rare class
Disease	1.6
Hypothyroid	4.7
Sick-euthyroid	9.2
KRK-5	1.7
KRK-8	5.1
KRK-9	6.1
KRK-10	7.1
KRK-11	10.2
KRK-13	14.9
KRK-14	16.2
KRK-15	7.7
KRK-16	1.4

[1] A binary rare-class problem is formed by considering a class as a rare class and union of the other classes as one major class.

[2] The first three datasets are not included in the average due to the unavailable results for PNrule. We could not find more results for this technique from published research neither from their authors.

TABLE IV
F-MEASURE COMPARATIVE RESULTS

Datasets	PNrule	BPNrule	EPRC	EPDT	DEP
Disease	-	-	73.5	74.9	**76.8**
H-thyroid	-	-	93.7	94.3	**95.6**
S-thyroid	-	-	88.3	88.7	**90.2**
KRK-5	63.5	65.8	65.1	65.5	**68.4**
KRK-8	52.7	61.8	66.1	66.9	**69.8**
KRK-9	43.4	59.1	66	66.7	**70.1**
KRK-10	42.1	54.6	58.2	55.9	**63.3**
KRK-11	49	58.6	58.9	58.3	**64.9**
KRK-13	58.5	61.6	64.5	65.3	**69.5**
KRK-14	61.7	72.9	74.3	74	**78.8**
KRK-15	66.1	72.1	74.9	74.8	**77.4**
KRK-16	56.4	70.2	78.5	78.2	**83.3**
Average	54.8	64.1	67.4	67.3	**71.7**

REFERENCES

[1] H. Alhammady, and K. Ramamohanarao. Expanding the Training Data Space Using Emerging Patterns and Genetic Methods. In Proceeding of the 2005 SIAM International Conference on Data Mining, New Port Beach, CA, USA.

[2] G. Dong, and J. Li. Efficient Mining of Emerging Patterns: Discovering Trends and Differences. In Proceedings of the 1999 International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA.

[3] H. Alhammady, and K. Ramamohanarao. The Application of Emerging Patterns for Improving the Quality of Rare-class Classification. In Proceedings of the 2004 Pacific-Asia Conference on Knowledge Discovery and Data Mining, Sydney, Australia.

[4] H. Alhammady, and K. Ramamohanarao. Using Emerging Patterns and Decision Trees in Rare-class Classification. In Proceedings of the 2004 IEEE International Conference on Data Mining, Brighton, UK.

[5] H. Fan, and K. Ramamohanarao. A Bayesian Approach to Use Emerging Patterns for Classification. In Proceedings of the 14th Australasian Database Conference (ADC'03), Adelaide, Australia.

[6] Guozhu D., Xiuzhen Z., Limsoon W., and Jinyan L.. CAEP: Classification by Aggregating Emerging Patterns. In Proceedings of the 2nd International Conference on Discovery Science (DS'99), Tokyo, Japan.

[7] C. Blake, E. Keogh, and C. J. Merz. UCI repository of machine learning databases. Department of Information and Computer Science, University of California at Irvine, CA, 1999.
http://www.ics.uci.edu/~mlearn/MLRepository.html.

[8] H. Alhammady & K. Ramamohanarao. Mining Emerging Patterns and Classification in Data Streams. In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI), Compiegne, France (Sep 2005), pp. 272-275.

[9] H. Alhammady & K. Ramamohanarao. Using Emerging Patterns to Construct Weighted Decision Trees. In IEEE

Transactions on Knowledge and Data Engineering. Volume 18, Issue 7 (July 2006), pp. 865-876.

[10] Van Rijsbergan, C. J. (1979). Information retrieval. London, UK: Butterworths.

[11] Joshi, M. V., Agarwal, R. C., & Kumar, V. (2001). Mining needle in a haystack: classifying rare classes via two-phase rule induction. In Proceedings of the ACM-SIGMOD International Conference on Management of Data (ACM SIGMOD), Santa Barbara, CA, USA, pp. 91-102.