Finding Contrast Patterns in Imbalanced Classification based on Sliding Window

Xiangtao Chen & Zhouzhou Liu
College of Information Science and Engineering Hunan University, Changsha 410300, China

ABSTRACT: In the process of the contrast patterns mining, people usually assume that the datasets distribution is basic balance, but in the real world, there are many data sets which class distribution is imbalanced. Considering the problem of contrast patterns mining on the imbalanced data sets, in this paper, we introduce the balance factor, give a new defined contrast patterns called balance emerging patterns(BEPs for short) which suitable for the imbalanced data sets, and propose a new algorithm WBEPM, it construct a sliding window to mine the BEPs on the imbalanced datasets. Experimental results show that the proposed algorithm has a better mining effect than the algorithm for original simple contrast patterns mining, the classification accuracy of the BEPs classifier is higher than that of the previous contrast patterns classifier when deal with the imbalanced data sets.
KEYWORD: Class Imbalance; Balance Factor; Sliding Window; Contrast Pattern Tree

1 INTRODUCTION

Compared to the traditional classification, the classification accuracy of the contrast patterns (Shiwei Zhu et al, 2015) classifier is better than that of the traditional classifier because that the contrast pattern has a strong discriminative capability. At present, there are many contrast patterns mining algorithm. But most of these mining algorithms base on a hypothesis that the class distribution of the data sets is balance, such as Max-Miner algorithm (Bayardo, R.J, 1998), Chen's JEP mining algorithm (CHEN Xiang-tao et al, 2010) and so on. However, in the real life, many data sets are imbalanced, people rarely take the problem that the data sets are imbalanced into account, and the simple definition of contrast pattern is not suitable for the imbalanced data sets because the simple defined contrast patterns are difficult to solve the problem that the classifier will be skewed to majority class, leading to the misclassification of minority class. For example, when dealing with the bank's credit card fraud detection (Sanjeev Jha et al, 2012) information, people concerns about the fraud information recording more, and hope the learned classifier models focused on the identification and classification of fraud information more. In these areas, people hope that more minority class samples could be identified accurately.

When using a general classification algorithm to classify imbalanced data sets, the information of minority class's samples is often covered by that of majority class samples. Meanwhile, the classification error rate come from the minority class samples that is much higher than that come from the majority class samples, and classifier generalization ability is too poor. The main reason for these problems is that most classification algorithms do not consider the sample distribution of the training data sets in each class, and consider the cost of each classification error is the same. When the class distribution is not balanced, the existing algorithms cannot find the suitable patterns for the minority classes, because, for the classify of imbalanced data sets, the False Positive Rates need a low support, but it will result in a large number of contrast patterns be generated in the majority class. So that will consume too much computing resources and get a bad classification result.

In this paper, we propose a novel algorithm WBEPM to mine the contrast patterns on the imbalanced data sets based on a sliding window mechanism. First, redefinition of contrast patterns on imbalanced data sets. We addressed the impact of the relationship between the majority class and the minority class, proposed a new contrast patterns called BEPs, established the basis of the contrast patterns mining on the imbalanced data sets. Then, for the imbalanced data set, we establish a sliding window to split the data sets and reduce the size of data imbalance ratio; improve the generalization ability of

the mining model. In the mining process, we fixed the minority class samples in the window, meanwhile, we let the majority class samples flow across the window, to constitute some sub data sets with the minority class data, and the sub data sets' imbalance ratio is relatively flat. In the window, we adopt based on the sorted frequent pattern tree structure to mining the new contrast pattern. As the majority class samples flowing, it will form many windows, we use the window data to mining the contrast patterns, while building some sub classifiers with the contrast patterns, until the end of window slipping.

2 RELATED CONCEPTS

Assume a database D contains an example of N samples, with attributes I= { i_1 , i_2 , i_3, . . . , i_m} and I is a set of items, an item set X is a subset of I, T is the number of transactions, m is the number of attributes, all continuous attributes are discretized. Item is the duality of attribute name and attribute value. There are two classes denoted as C_p (positive samples or minority class samples) and C_n (negative samples or majority class samples), data sets D_p and D_n correspond to the respective classes C_p and C_n .Using the concept of item sets, each sample is a collection of items, and the training data set is a set of multiple items.

For the given data sets D_p and D_n , contrast patterns describe the significant difference between the data sets of D_p and D_n .Using support expressed as follows:

$$| sup_1(X)- sup_2(X)|>\alpha \quad (1)$$

or

$$sup_1(X)/ sup_2(X)>r \quad (2)$$

Among them, $sup_1(X)$ denote the supports of item set X in D_p ,$sup_2(X)$ denote the supports of item set X in D_n.α and r is a given threshold, when the difference or ratio of $sup_1(X)$ and $sup_2(X)$ is greater than the given threshold, we think X set is contrast patterns, that is to say the patterns in the two data sets with a significant correlation. The (2) is also the definition of emerging patterns (Dong G et al, 1999) which is one of the contrast patterns, if the value of the r is ∞,the patterns is caller jump emerging patterns (CHEN Xiang-tao et al, 2011). The classification effect of the emerging patterns classifier is better than that of C45 (Quinlan, J.R. 1993), CMR (Li, W. et al, 2001) classifier.

However, due to the definition of the above patterns, which based on a basic hypothesis: the class distribution is basic balance. When the class distribution is not balanced, using the method of the pattern definition, that cannot fully reflect the characteristics of the patterns, especially for the minority class that is too sensitive for the value of the support threshold, even the minority class information will be covered fully by the majority class information, and cannot mining the minority class information at all. When considering the unbalanced distribution of data sets, we adopt to reconstruct the concept of contrast patterns.

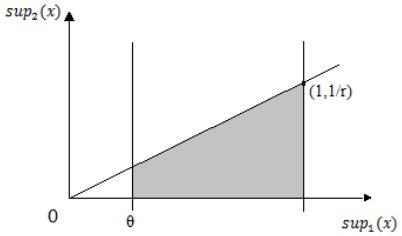

Fig. 1 Pattern Space of Contrast Patterns

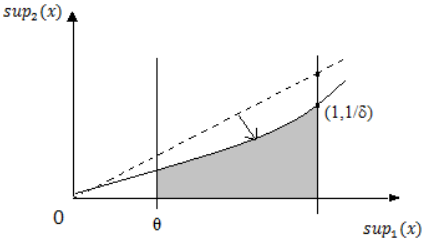

Fig. 2 Pattern Space of BEPs

Figure (1) gives a pattern space (Jinjiu Li et al, 2013) for the contrast patterns. The two-dimensional coordinate space is a plane of the support degree, any point X represents a certain item set X in the plane, the longitudinal coordinates of the X is support degree of the sub data set D_p, the horizontal coordinates of X is support degree of the sub dataset D_n . The slope of X is a derivative of the X's growth rate from D_p to D_n , which is the value of r in the (2). Apparently, according to the definition of contrast patterns ($sup_1(X)/ sup_2(X)>r$), the points in the gray zone is the contrast patterns from D_p to D_n. But in imbalanced data set, due to the presence of minority data, the magnitude of the $sup_1(X)$ and $sup_2(X)$ gap is big, when mining the contrast pattern, is difficult to get a good mining results in the pattern space.

Figure (2) shows us the pattern space of our redefined contrast patterns. The boundary curves of contrast patterns incline to the minority class, and it abandon some contrast patterns near the boundary which affect the result of classification, make the pattern space more incline to the minority class, while ignore the patterns located near by the boundary which have a great possibility to judge a test sample belongs to the majority class.

Here we give a new contrast patterns definition, which have a better adaptability on the imbalanced data sets, we call it Balance Emerging Patterns.

Definition 1. (Balanced Emerging Patterns, BEPs for short)When the itemsets X satisfy the following conditions, we consider it is BEPs:

$$BEPs=\{x|\delta<=a * e^{k*sup_1(X)}/ sup_2(X), sup_1(X)>\theta\} \quad (3)$$

Where k is the balance factor, and δ the minimum contrast coefficient, a<1 is the correlative correction parameter, θ is the minimum support threshold. We use the balance factor to enhance the results of minority classes mining, which makes the pattern boundary offset to the minority class. Balance factor is introduced for the purpose that to weaken the classified tendency of majority class patterns information, that is in order to prevent the boundary offset to the majority class zone.

3 THE ALGORITHM OF PATTERNS MINING

3.1 *Patterns Mining Framework.*

For the characteristics of imbalanced data sets and contrast patterns, we design the contrast pattern mining algorithm WBEPM to dealing with the imbalanced data sets. WBEPM means a contrast patterns mining algorithm under the imbalanced data sets based on a sliding window.

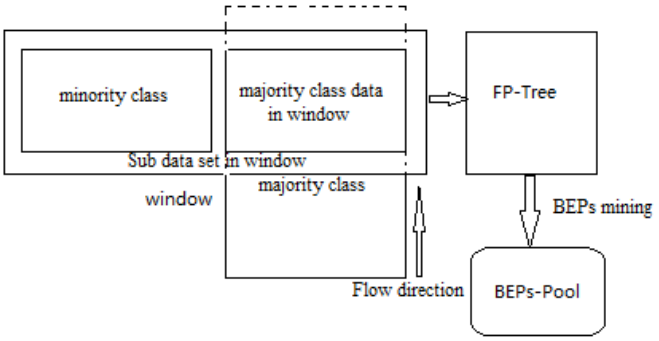

Fig.3. Patterns Mining Framework

The mining algorithm framework shown as Figure (3): Firstly, we determine the size of window according to the numbers of minority class data sets. then we let the minority class data fixed on the window, when the size of window has been determined, let the majority class data flow into the window, until the window is full, the data sets in window is an sub data sets that the imbalance degree is very small, subsequently, the sub data sets was constructed with a FP-tree at the same time mining BEPs, the results will be stored in the BEPs results pool, cycle the above process, until the majority class data is empty.

3.2 *Sliding Window Mechanism.*

James Bailey (James Bailey et al, 2010) give a method of contrast patterns mining in the changing data but it does not adapt the imbalanced data sets. Considering to the diversity distribution of imbalanced data sets, the size of imbalance ratio did not be specific discussed at previous. It very difficult to point out what proportion of the imbalance ratio will influence the results of contrast patterns mining and the classification accuracy of the classifier clearly, because that the contrast patterns mining and the performance of the classifier have a relationship with the data sets distribution and sample size. Therefore, we introduce the sliding window mechanism to divide data sets. Through establish a sliding window to reduce the ratio of data imbalance, we let the minority class data fixed in the window, and the majority data mobile through the window, matching the minority data fixed in the window to constitute a sub data set, and the imbalance ratio of the sub data set is relative gentle. The size of sub data set and the imbalance ratio are controlled by the size of the window. In the process of flowing, we mine the BEPs of the sub dataset in the window, until the end of the window process.

3.3 *Sorted FP-Tree.*

Due to the frequent pattern tree (J.Han et al, 2000) store the main information of database, we use the window mechanism to construct some windows, that request the number of scanning database is the less the better. Because the sorted FP-Tree mining algorithm (Fan H et al, 2002) only needs to scan database two times, and the key information store in the memory as the form of FP Tree, avoid scanning the database many times that brings a large number of I/O time consuming, and it does not need to generate candidate sets, thereby reduce the time of generate and test candidate sets. At the same time the algorithm uses the divide method to mining the patterns, thus in the process of mining, it will greatly reduce the search space. Need to pay attention to is, here, we used two result sets, intermediate result sets Σ_1 and the final result set Σ. Σ_1 kept a suffix generated patterns which generated by an item in the original FP Tree header table, each Σ_1 will be emptied, Σ save the final result.

3.4 *The algorithm WBEPM.*

We show that the algorithm WBEPM mine the contrast patterns on imbalanced data sets based on a sliding window:

Algorithm BEPM

Input: the training data sets D (D_p is the data sets of minority class, and D_n is the data sets of minority class), and the minimum support Sup;
Output: all of the contrast patterns.
Procedure BEPM (D,α){
1 if | D_p |≠0{
2 then W=2*| D_p |*(1+m); //m is a given window correction factor, W is the size of the window
3 while(|D_n|≠0){
4 the first W-|D_p | dates of the negative class data D_n flow into the window

```
5       construction of windows dataset Dwi
6       construction of FP patterns tree   TDwi
7       BEPi =FP-M (TDwi ,α);
8       BEP_Pool+= BEPi ;          //the patterns result set flowinto the pool;
9       negative class data label set to window boundary;
10       |Dn|=|Dn |-(W-|Dp|) and   Dn =new Dn
    }
  }
Procedure FP-M (T, α) {
11  if T contains a single path P then P represents the patternsβif (! CheckIsSuperset (Σ1,β∪α))
    then CheckIsSubSet (Σ1,β∪α)     //Σ1 is the intermediate result set.
12  else {
13      down the T's most left tree, until the support of a node N <sup
    // At this time the number of layers is k
14      for the header item αi in T(i=n, …, k-1)
15      do {produce pattern β=αi∪Sup(β)=Sup(αi); //sort by item
16          construct conditional pattern library of β and conditional FP- tree   Σβ;
17          delete all nodes on the same node of αi
18            if Σβ≠Øthen call FP-M (Σβ,β)
19                    if (T=the initial FP-tree) {
20                    for each patterns βin Σ1:
21                    if (! CheckIsSubSet (Σ, β))
                  add the β to the Σ; emptyΣ1.}
        }
  }
```

Line1-2: Acoording to the minority classes, the training data set to determine the size of the sliding window is more than two times size of the minority class data size. The size of the window marked as W;

Line3-6: when the majority class data set is non-empty, the majority data according to the flow state enter to the window, until the window is full. At this time, a window has constructed, in the window, we use the sorted frequent pattern tree-mining algorithm (FP-M) to mine the BEPs;

Line7-12: the BEPs result set is stored in a classification patterns pool, at the same time, updating the flow data in the window; when the majority class data sets is not empty, return the step 2, otherwise ,using BEPs of the classification patterns pool to build a classifier.

4 EXPERIMENT

4.1 *Experimental Environment.*

In order to verify the effectiveness of the algorithm, we selected eight imbalanced standard data sets for experimental validation, and the continuous attributes of all data sets have been discretized. The experimental environment: Intel (2.50GHz) Core I5-3210 CPU, 4GB Windows8.1, Visual memory, Studio 2013.

The table shows us the 8 different balance standard data sets for our test:

Table.1. The experimental data set

Data sets	Number of Samples	Number of Minority Class Samples	Number of Majority Class Samples	Imbalance Ratio
abalone	502	15	487	1:32
ecoli4	336	20	316	1:15
car-good	1728	69	1659	1:24
car-vgood	1728	65	1663	1:24
flare-F	1066	43	1023	1:23
led7digit	443	37	406	1:10
yeast4	1484	51	1433	1:32
yeast5	1484	44	1440	1:32

4.2 *The Number of Patterns.*

For the number of patterns, on the one hand, considering the algorithm efficiency, it should be control the number of patterns that cannot be too much. On the other hand, minority class pattern mining requires a lower support degree, but low support degrees will produce a huge number of patterns in the majority class, and bring some difficult in complex calculation. Therefore, it is necessary to control the number of patterns. To verify the redefine contrast patterns BEPs close to the patterns boundary that reduce the number of patterns which influence classification results, we use an experiment verify the number of two patterns under the same parameters in the patterns mining.

Table.2. The Number of Patterns

Data sets	EPs	BEPs
abalone	604	525
ecoli4	512	455
car-good	2120	1788
car-vgood	1994	1770
flare-F	890	860
led7digit	510	490
yeast4	1409	1209
yeast5	1440	1190

The results show that 6 data sets are reduced more than 10%, even the yeast5 dataset is reduced by 17.1%, in addition, there have 2 data sets is reduced about 4%.

4.3 *Classification Accuracy*

We mine the BEPs and EPs by the same training data sets, and use the CAEP (Dong G et al, 2001) and the WBEPM to classify our test data sets respectively. We calculate the contribution of the two types of

patterns to test the sample to be classified, and judge the category of the samples. Considering the data characteristics of the minority class and majority class of the imbalanced data sets, we test the classification accuracy of the minority class and the majority class separately. At the same time, we use the classifier evaluation index defined by the confusion matrix to evaluate the classifier of imbalanced data sets. The ROC can show the generalization ability of the classifier. When the test sets and training sets is not identically distributed, ROC curves were used to compare different classifier performance is more appropriate, it will consider the precision and recall at the same time. We give two ROC curves of the data sets Abalone and ecoli4 for comparison.

Table.4. The classification accuracy

Data sets	EPs-Minority	EPs-Majority	BEPs-Minority	BEPs-Majority
abalone	0.64	0.86	0.75	0.88
ecoli4	0.67	0.75	0.72	0.80
car-good	0.70	0.81	0.82	0.85
car-vgood	0.67	0.84	0.76	0.80
flare-F	0.75	0.92	0.79	0.92
led7digit	0.72	0.89	0.83	0.90
yeast4	0.68	0.91	0.75	0.86
yeast5	0.67	0.90	0.75	0.85

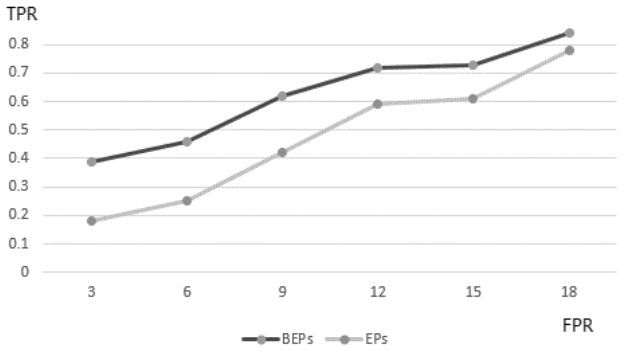

Fig. 5. ROC on abalone

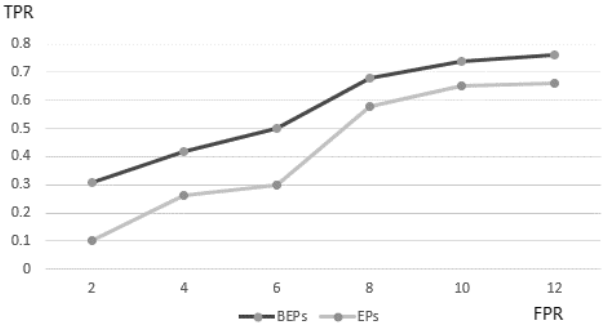

Fig. 6. ROC on ecoli4

For result of classification accuracy from the Table (4) the BEPs classify has a better classification effect than the EPs in the minority class, and classification accuracy is similar between two classes in the majority class. According to the definition of the ROC, when the smaller the FPR and the larger the DR, the classifier is the better. In the Fig (5), the TPR get 50%, the FPR of BEPs and EPs is 60% and 73% respectively, as the FPR is larger, both of the classify get a larger DR, but the BEPs get a larger DR. The Fig (6) shows the same result of both methods. The result reveal that the BEPs classify is more applicable on the imbalanced data sets.

5 CONCLUSIONS AND OUTLOOK

In this paper, we have done two main jobs: At the first, for the influence of the relationship between the majority class and the minority class on imbalanced data sets, we redefine the concept of the contrast patterns. At the Second, we establish a sliding window mechanism to reduce the imbalance ratio of the imbalanced data sets. We introduced a sliding window mechanism, and the imbalance ratio obtained a unified settlement in the window.

There are some problems will be solved in our next work. First, when the algorithm deals with the relative imbalanced data sets, it will encounter the problem that the constructed window is too large. In order to reduce the size of the window, we consider using an appropriate sampling method to let a part of minority class data fixed in the sliding window. Second, we consider the use the window to construct some sub training sets with different imbalance ratio, after trained, using classifier optima selection and the principle of bias towards positive class to integrate all of the sub classifiers.

REFERENCES

Bayardo, R.J: Efficiently mining long patterns from databases. In: SIGMOD 1998, pp. 85–93 (1998).

CHEN Xiang-tao, LU Li-juan: An improved algorithm of mining Strong Jumping Emerging Patterns based on stored SJEP-Tree[C]. In Bio-Inspired Computing: Theory and Applications(BIC-TA), Changsha, China,2010:8-94-898.

CHEN Xiang-tao, LU Li-juan: A New Algorithm Based on Shared Pattern-tree to Mine Shared Emerging in Proc of ICDMW, Vancouver, BC, 2011:1136-1140.

Dong G, Li J. Efficient mining of emerging patterns: discovering trends and differences. In: Proc of 5th Int Conf on Knowledge Discovery and Data Mining. San Diego, 1999, 43-52.

Dong G, Zhang X, Wong L: CAEP: classification by aggregating emerging patterns. Knowledge and information Systems, 2001, 3(10): 131-145.

Fan H, Ramamohanarao K. An efficient single-scan algorithm for mining essential jumping emerging patterns for classification. In: Proc of 6th Int Conf on Knowledge Discovery and Data Mining. Taipei, 2002, 456-462.

James Bailey and Elsa Loekito: Efficient Incremental Mining of Contrast Patterns in Changing Data. In Information Processing Letters 2010 Vol.110 No.3 P88-92.

J.Han, J.Pei, Y.Yin.: Mining frequent patterns without candidate generation. 2000 ACM-SIGMOD (sigod'00) TX, New York: ACM Press, 2000.

Jinjiu Li, Can Wang: Efficient Mining of contrast Patterns on Larger Scale Imbalanced Real-Life Data. In PAKDD 2013, 62-73.

Li, W., Han, J., Pei, J. CMAR: Accurate and efficient classification based on multiple class-association rules. In: ICDM 2001, pp. 369–376 (2001).

Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, Los Altos (1993).

Sanjeev Jha, Montserrat Guillen, J. Christopher Westland: Employing transaction aggregation strategy to detect credit card fraud In Expert Systems with Applications (2012).

Shiwei Zhu, Meilong Ju: A review of contrast pattern based data mining. In Seventh International Conference on Digital Image Processing (ICDIP 2015) Los Angeles, United States 2015.