

Classification Algorithm based on NB for Class Overlapping Problem

Haitao Xiong¹, Ming Li², Tongqiang Jiang¹ and Shouxiang Zhao¹

¹School of Computer and Information Engineering, Beijing Technology and Business University, 100048 Beijing, China

²School of Business Administration, China University of Petroleum, 102249 Beijing, China

Received: 7 Oct. 2012, Revised: 21 Dec. 2012, Accepted: 23 Dec. 2012

Published online: 1 Jun. 2013

Abstract: Class overlapping is thought as one of the toughest problems in data mining because the complex structure of data. The current classification algorithms show little consideration of this problem. So when using this traditional classification algorithms to resolve this problem, classification performance is not good for samples in overlapping region. To meet this critical challenge, in this paper, we pay a systematic study on the class overlapping problem and propose a new classification algorithm based on NB for class overlapping problem (CANB). CANB uses NB to find class overlapping region and use this region and non-overlapping region in NB classification model learning separately. Experimental results on bench mark and real-world data sets demonstrate that CANB can improve the classification performances for class overlapping problem stably and effectively.

Keywords: Class Overlapping, Classification, Overlapping Region, Naive Bayes

1 Introduction

For some useful classification applications such as fraud detection, network intrusion detection, and character recognition, there is often occurs that some samples from different classes have very similar characteristics which are called overlapping samples for they usually reside in overlapping regions in the feature space. The problem is recognized as the so-called class overlapping problem. It has become one of the toughest problems in data mining and business intelligence communities. As we know, lots of researchers have found that class overlapping problem is one of the important factor of bad classification performance [1,2,3]. So lots of researches have studied the feasible solution for this problem [4,5].

For its complicated resolving methods, in the literature, great methods have been proposed to improve the classification performance in class overlapping problem. Prati et al. [6] developed a systematic study using a set of artificially generated datasets. Results showed that the degree of class overlapping had a strong correlation with class imbalance. Moreover, experiments with fuzzy classifiers showed that the overlapping affected more than the imbalance [7]. These researches, however, mainly work on artificially generated data sets

and focus on the effectiveness of basic classifiers in the presence of class overlapping. Therefore, from a practice point of view, there is still a critical need in conducting a systematic study on the schemes that can find and handle overlapping regions for the real-world data sets.

Our work is in order to resolve this hard work based on our before research in [8]. A new Classification Algorithm based on NB for class overlapping problem (CANB) for class overlapping problem was proposed in this paper. Specifically, we have two major contributions in CANB as follows. One is that we use the classification algorithm Naive Bayes (NB) [9] creatively to capture the overlapping regions in class overlapping problem resolving application. The other is that we bring this overlapping regions into classification model learning by NB and model the data sets with the presence of class overlapping by overlapping region and non-overlapping region separately.

The remainder of the paper is organized as follows. Section 2 describes the preliminaries and define the problem studied in this paper. Section 3 describes classification algorithm based on NB for class overlapping problem. Then in Section 4, we show the systematic experimental results of bench mark data sets and real-world data set. Then the analysis of the

* Corresponding author e-mail: xionghaitao@th.btbu.edu.cn

experimental results is shown. Finally, we conclude our work and the future work in Section 5.

2 Preliminaries

In this section, we first review the main issues related to class overlapping problems which have been mentioned in our previous work. Then, we present the problem definition.

2.1 Class overlapping problems

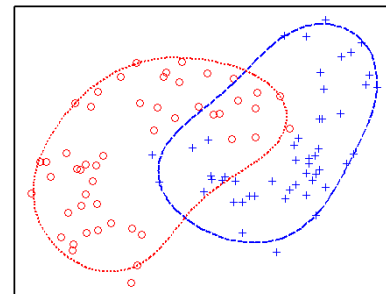
Class overlapping problems are often of special interests in data mining scope. Real-world data sets usually involve class overlap. A class overlap region in $\Omega \in R^n$ an n -dimensional feature space R^n can be described as a region where at least two different classes C_a and C_b simultaneously exhibit probability densities greater than zero, i.e., $p(x|C_a) > 0$ and $p(x|C_b) > 0$, $\forall x \in \Omega$ and $a \neq b$ [11]. It is well known that if the two class conditional distributions have a large overlapping volume, the classification accuracy is poor [12]. Therefore, people often want to uncover subtle patterns that may be overlap with data from different class. In the literature, the class overlapping problems has been addressed from the following two perspectives [1,4,6,13] and bring two kinds of overlapping which were call "concept overlapping" and "sample overlapping" by this paper.

Concept overlapping perspective. This is a macro viewpoint. From this viewpoint, the concepts of one class and the other class has overlapping region in the feature space.

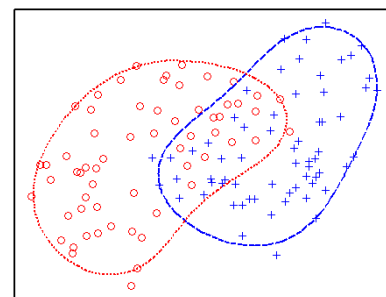
Sample overlapping perspective. This is a micro viewpoint. From this viewpoint, some samples from one class have similar characteristics or overlap with some samples from the other class in feature space.

Let us illustrate class overlapping by some examples shown in Fig. 1. In the figure, the dotted dashed lines denoted the concept or boundaries of the class. Fig. 1a is the example of concept overlapping, while Fig. 1b is the example of sample overlapping. As can be seen, concept overlapping is easily happen while two classes close. But in this time, sample overlapping is not need to happen. If sample overlapping happens, it means concept overlapping is already happen. Previous research have shown that classification errors are always lying on the edge of data [1,14]. And this is the region where concept overlapping often occurs. However, samples can be sparsely distributed so that non-overlapping samples of them are far away from the boundary in overlapping region, and thus have little impact during the boundary learning process in overlapping region. As a result, overlapping samples tend to have more power to decide the boundary. Using non-overlapping samples for learning may shift the boundary of overlapping region,

and lead to a large misclassification rate for the normal class. So learning on class overlapping region separately can improve the performance of classification.



(a) Concept overlapping example.



(b) Sample overlapping example.

Fig. 1 Class overlapping examples.

2.2 Problem definition

For traditional classifiers, all the samples from a class are used to represent a specific concept and are used as a whole for building the classification model. However, it is quite normal to see that the concept of samples from a class may have overlapping region with the concept from other class. Because in real-world, sample overlapping rarely happen and it is hardly to be handle through schemes on learning process. The most common overlapping is concept overlapping. When it is combined with the well-known class imbalance problem, the situation becomes even more complicated.

In the other side, C2C e-commerce fraud in China is also very severe. There were a large number of buyers who had ever confronted fraud in the C2C e-commerce and the fraud behaviors of seller include: misrepresentation of items, shill bidding, seller collusion, fee stacking, non-delivery of items and so . Fraud has become an important factor which prevents C2C e-commerce from further developing. Taobao is the leading online C2C e-commerce company in China. Currently, the reputation systems chosen by most C2C

e-commerce sites to prevent fraud mainly use simple summation or average of ratings. Summation of ratings is simply to sum the number of positive ratings and negative ratings separately, and to keep a total score as the positive score minus the negative score. The reputation systems adopted by eBay and Taobao use summation of ratings. Average of ratings is to compute the reputation score as the average of all ratings which is used by Amazon. Reputation systems can hardly depict the traders true reputation, and often be attacked by fraudsters. As C2C e-commerce develops, more and more buyers and sellers participate in it. Meanwhile, the number of fraud transaction also runs up remarkably. Non-fraud transaction is represented by a large number of transactions while fraud transaction is represented by only a few. So it is extremely difficult to extract the fraud patterns in C2C e-commerce and cause the class imbalance problem. So there is need to design an classification algorithm for C2C e-commerce fraud detection.

3 Classification algorithm based on NB for Class Overlapping Problem

In this section, we proposes a new Classification Algorithm based on NB for class overlapping problem (CANB) according to the features of class overlapping. The aim of CANB is to improve the classification accuracy of NB in class overlapping problem resolving.

Naive Bayes (NB) classifier is a simple probabilistic classification algorithm based on applying Bayes' theorem with strong independence assumptions [9,15]. NB assumes that the presence of a particular feature of a class is unrelated to the presence of any other feature, given the class variable. Depending on the precise nature of the probability model, NB can be trained very efficiently in a supervised learning setting.

Theoretically, bayes optimal classifier minimizes the error rate exists using the conditional probability of a positive class outcome given a particular attribute vector. Then applying the bayes conditional probability rule using the conditional probability of a particular attribute vector given that the class outcome is positive and the prior probability of the event that the class outcome is positive.

So, NB use the conditional probability and the prior probability to decide the label of unknown samples, and the sample x is assigned to class C_i when the blow condition is ture:

$$P(x|C_i)P(C_i) > P(x|C_j)P(C_j), 1 \leq i, j \leq n, i \neq j, \quad (1)$$

where C is classes, and n is the number of classes.

In spite of their naive design and apparently over simplified assumptions, NB often work much better in many complex real-world situations than one might expect. NB can give better predictive accuracy than other

algorithms such as C4.5 and BP when attributes are normally distributed and not redundant [9,10]. While attributes are not normally distributed and redundant, it will show lower predictive accuracy.

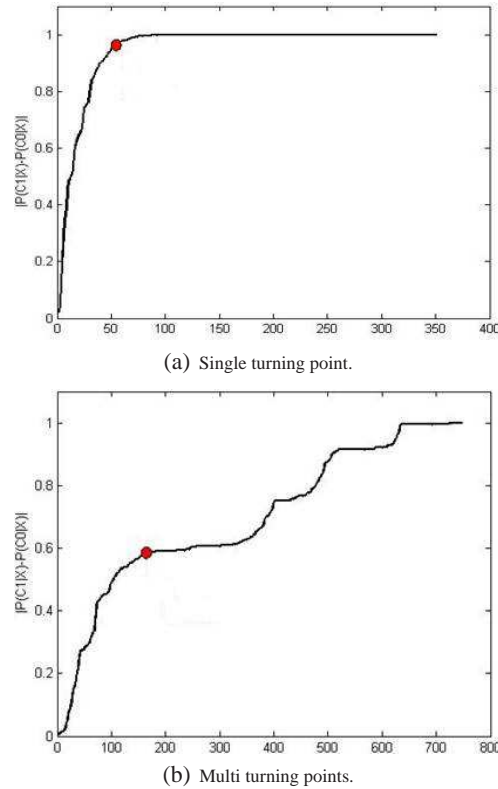


Fig. 2 The graph of complementary cumulative distribution for NB.

As we known, NB does not consider the class overlapping problem, so it is difficult to substantively solve this problem which is often occur in real data[10]. But the advantage of NB is also obvious. The running time of NB is low and the classification performance of NB is well for normal data sets. For these reason, CANB firstly use NB to find class overlapping region and non-overlapping region. This is through the threshold of dividing r . if that value of sample s is less than r , then we thought this sample is belong to overlapping region.

Firstly, $P(C_i|x)$ of each samples will be calculated by NB. Then we can get the value of $|P(C_i|x) - P(C_j|x)|$, in which i and j are different classes. These values are the important things to decide which samples are belong to overlapping region. So they can be used to generate the graph of complementary cumulative distribution of $|P(C_i|x) - P(C_j|x)|$. After finding out the turning point of $|P(C_i|x) - P(C_j|x)|$ for all samples, we can get the threshold of dividing r . Fig. 2 shows how the graph of

complementary cumulative distribution and the threshold of dividing in NB works. There are two kinds of turning points in this Fig. Fig. 2(a) is single turning point and Fig. 2(b) is multi turning points. CANB can deal with this two kinds of turning point. If there exists no turning point in the graph of complementary cumulative distribution, then CANB will use the threshold of dividing that user have input.

Then the overlapping region can be define by the condition which is described blow:

$$O = \{x \in D \mid |P(C_i|x) - P(C_j|x)| < r\}, \quad (2)$$

where D is the samples of training data set.

In the end, we can divide data set into two regions: overlapping region and non-overlapping region. Divided data set with the two regions is generated on this approach. Through this way, the original data structure does not change, this is because NB algorithm is designed based on the features of data, and the date sets are divided into several more sophisticated regions which are good for classification model learning.

After finding overlapping region, CANB learns on the data in overlapping and non-overlapping regions separately. It can get two models. One is the model Mo of the data in overlapping region, and the other is the model Mn of the data in non-overlapping region. Each test data will be tested on the model which is determined by that if the test data is lying on non-overlapping regions or overlapping regions to predict the class of the test data.

Lastly, classification model Mo and Mn are used to predict the label of test samples, and the assigned pseudo sub-class label for each test samples is converted into the label of its parent-class. Figure 3 shows the pseudo-code of CANB.

4 Experiment and Analysis

4.1 Experimental Data Sets and Metrics

The main goal of this research is to gain some insights on NB of solving class overlapping problem and its interrelationship with classification performance. We will investigate how CANB affect the performance of learning in such conditions.

To make this comparison, five real-world binary data sets from UCI [16] are selected for comparing experiments. Table 1 summarizes the data sets employed in this study. For each data set, it shows the data set names, the number of examples (#Examples) and the number of attributes (#Attributes).

Experiments are divided into two scenarios. The first one is in benchmark data sets. It uses NB to find out overlapping ratios of the five data sets. Then it compares the CANB with other classic classification algorithms and see how overlapping ratios affect the performance of CANB. The second one is in real-world data set. It uses

Table 1 Characteristics of Data Sets.

Data Set	#Examples	#Attributes
Inosphere	351	34
Wine	178	12
Transfusion	748	5
Diabetes	768	8
Glass	214	10

CANB and NB in C2C e-commerce data sets to see the performance of these two classification algorithms.

The experiments in this paper adopt a ten-fold cross-validation method. Each data set will be divided into ten equal parts, using nine folds as the training set and the remaining block as an independent test set. F-measure metrics of the both classes are used to compare the performance of CANB and other four classification algorithms with distinct natures, which are NB, k -NN ($k=5$), SVMs (with linear kernel) and C4.5. F-measure is defined as below [17]:

$$F\text{-measure} = \frac{2 \times TP}{2 \times TP + FP + FN}, \quad (3)$$

where TP, TN, FP and FN are the numbers of True Positives, True Negatives, False Positives and False Negatives, respectively.

4.2 Experimental Results of Benchmark Data Sets

The experiment is over a collection of above five binary data sets with decreasing class overlapping ratios. Experiment are divided into three scenarios. The first one uses three different schemes and four distinct classifiers to deal with class overlapping problem for data sets with different overlapping ratios. The second one compares the original scheme with the best scheme in first experiment and finds what may affect classification performance.

Firstly, the overlapping ratio of five data sets will be calculated in CANB which is described in previous section. Table 2 shows the overlapping ratios of five data sets. Using NB to find out overlapping region and non-overlapping region of each data sets, this data sets can be divided into two parts: non-overlapping and overlapping. After that, CANB and four distinct classifiers will be used. F-measure of both classes will be calculated for comparison.

Then we investigate how F-measure value changes as the increase of the overlapping ratio of data sets. Table 3 shows the performances of CANB and other four traditional classifiers on five binary data sets. In table, "class 1" and "class 1" means the two classes in binary data sets. The results in bold indicate the best values. As can be seen, F-measures of CANB for the both classes in binary data sets are higher than other four traditional

```

CANB(Classification Algorithm based on NB for class overlapping problem)
Input:    D: a training data set.
            T: a test data set.
            C: class of data set.
            n: number of class.
            r: threshold of dividing.
            NB: Naive Bayes classification algorithm.
Output:  L: label of classification results.
Procedure:
Phase 1: finding class overlapping region by NB
1.   for class i = 1 to n
2.       Mi = train(NB,D(i));
3.   end for
4.   for class i = 1 to n
5.       for each sample s in D(i)
6.           for class j = i to n
7.               if ( $|P(C_i|x) - P(C_j|x)| < r$ )
8.                   Do = Do ∩ s;
9.               end if
10.          end for
11.      end for
12.  end for
Phase 2: training
13.  for class i = 1 to n
14.      Moi = train(Do , NB);
15.      Mni = train(D-Do , NB);
16.  end for
Phase 3: predicting
17.  L' = predictLabel(T , Mo, Mn);
18.  L = convertLabel(L');
    
```

Fig. 3 The pseudo-code of CANB algorithm

Table 2 Overlapping Ratio of Data Sets.

Data Set	Overlapping ratio
Wine	7.30%
Ionosphere	19.37%
Glass	23.80%
Transfusion	24.79%
Diabetes	24.87%

classifiers, or have almost the same results. While seeing into the overlapping ratios, for two data sets, *Ionosphere* , *Glass*, *Transfusion* and *Diabetes* with overlapping ratio more than 10%, CANB performs almost better than the other four classification algorithm. While for the remaining three data sets with overlapping ratio less than 10%, performance of CANB is close to the NB. This is because that data sets are divided into small parts will be good for precisely learning and time costing. If the overlapping ratio is high, the merit of this dividing is more useful. These observations support that using

CANB can improve the performance of classification in class overlapping problem.

Another observation is that as the overlapping ratio increases, the gaps between the CANB and other four classification algorithm are more and more obvious except for those high F-measure values. This is reasonable for that high F-measure can difficultly be improved. And because low overlapping ratio means that there are little data lying on overlapping region, CANB can hardly improve the overall performance.

4.3 Experimental Results of Real-world Date Set

In this subsection, we demonstrate an application of CANB for e-commerce fraud detection. The fast and wide development of Internet has made C2C e-commerce become more and more popular because of low cost and high efficiency. During the high development of C2C e-commerce, hidden problems have been exposed. The

Table 3 Classification Results by AdaBoost with Re-sampling Methods on Imbalance Data Sets

Data sets	F-measure(class 1)					F-measure(class 2)				
	CANB	NB	k-NN	SVMs	C4.5	CANB	NB	k-NN	SVMs	C4.5
Wine	0.979	0.958	0.967	0.250	0.932	0.992	0.998	0.979	0.579	0.928
Ionosphere	0.901	0.781	0.783	0.861	0.874	0.951	0.935	0.825	0.912	0.892
Glass	0.780	0.619	0.717	0.522	0.721	0.826	0.766	0.782	0.662	0.818
Transfusion	0.591	0.269	0.317	0.142	0.358	0.805	0.695	0.574	0.518	0.691
Diabetes	0.879	0.643	0.554	0.015	0.614	0.917	0.819	0.650	0.535	0.751

virtual internet transaction will make it not easy to check the identification of both sides in a transaction and customers have difficulty in buying products because of asymmetric information of product quality.

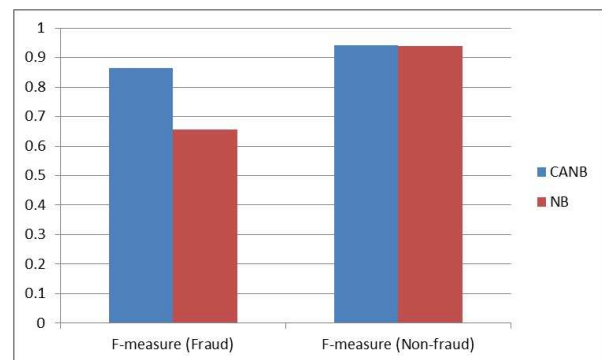
To ensure the safety of transactions and enhance the trust between users, Taobao adopts real-name authentication and differentiates authentication processes of individual users and business users. An individual user only needs to provide proof of identity authentication, but a business user needs to provide business license certification, and one user can not apply for these two kinds of authentication simultaneously. Taobao also builds its own reputation systems, and traders can see trust ratings of each other. In addition, it launches a third-party payment platform for reducing transaction risk and providing the function of fraud reporting. When a user finds that a fraud transaction occurs, he can report the suspected fraud transaction and the fraudsters account to Taobao. When Taobao identifies there is a fraud transaction, it will prohibit the fraudster from conducting fraud behavior through the ways of warning, investigation and closure. But due to the huge number of transactions, Taobao can not respond quickly to fraud transactions to prevent fraud, which leads to low customer satisfaction. So we can continuously collect data from taobao and update taobao data set. ways of warning, investigation and closure. But due to the huge number of transactions, Taobao can not respond quickly to fraud transactions to prevent fraud, which leads to low customer satisfaction. So we can continuously collect data from taobao and update taobao data set.

Firstly, we choose cell phone as the study object because fraud in cell phone is more serious than other commodities in china and collect cell phone data from Taobao from December 2011 to March 2012. The collected data includes account information of sellers, their reputation information, and transaction information. The fraud behaviors considered in this study are mainly misrepresentation of items, fee stacking and non-delivery of items. After data filtering and cleaning, blank and incorrect data was deleted. The final data is composed by 24250 sellers in which 816 are fraud sellers, and the number of non-fraud sellers is much larger than that of fraud cases. Table 4 shows the detail characteristics of collected sellers in taobao. The data set is transformed to a simple data set which is more suitable for learning.

Table 4 Characteristics of Collected Sellers in Taobao.

Type	Count
All Sellers	24250
Fraud Sellers	816
Non-fraud Sellers	23434

In Fig. 3, we can observe that CANB shows the better performances on both classes. If we take a closer look on the F-measures produced by pure NB and CANB, we can notice that CANB greatly improved the ability of NB in detecting more fraud e-commerce selles. The above indicates that CANB can help to divide a complex concept of class overlapping into simple concepts which are easy to be classify separated and take less training time for model learning.

**Fig. 4** The performance of CANB and NB on C2C e-commerce data set.

5 Conclusion

In this paper, we took a systematic study on the classification algorithm that were proposed specifically for handling the class overlapping problem. The classic classification algorithm NB was first employed to find the overlapping regions in the data. Then by extensive experiments on various real-world data sets, we found

CANB that modeling the overlapping and non-overlapping regions separately was the best algorithm for solving the class overlapping problem. By further comparative studies on four widely used classifiers, we found that NB enjoyed a much better performance using CANB. Finally, when combined with the class overlapping ratio, we found that as the increase of the degree of class overlapping, CANB showed higher improvements to the classification performance. And the reason for the best performance of CANB in experiments is that it can weaken the affection of class brought by class overlapping problem in class overlapping region and samples from non-overlapping region. Our study also opens up several directions for future research in fraud detection. This study conducts experiments only for the selected algorithms in the C2C e-commerce fraud detection model. In the future, more studies are needed to use different classification algorithm to detect fraud in C2C e-commerce.

Acknowledgements

This research was partially supported by the National Natural Science Foundation of China (No. 71201004, 71101153), the Scientific Research Common Program of Beijing Municipal Commission of Education (No. KM201310011009), and the Research Foundation for Youth Scholars of Beijing Technology and Business University (No. QNJJ2011-39).

References

- [1] C.L. Liu, *International Journal on Document Analysis and Recognition* **11**, 53-56 (2008).
- [2] A. Huo, H. Zhang, L. Zhang, H. Li and M. Hou, *Applied Mathematics & Information Sciences* **6**, 19-23 (2012).
- [3] Q. Zhu, Z. Zhang and Y. Xie, *Applied Mathematics & Information Sciences* **6**, 363-369 (2012).
- [4] V. García, R.A. Mollineda and J.S. Sánchez, *Pattern Analysis & Applications* **11**, 269-280 (2008).
- [5] A.M. Breda and A.F. Santos, *Applied Mathematics & Information Sciences* **6**, 123-134 (2009).
- [6] R.C. Prati, G.E. Batista and M.C. Monard, *Proc. of the Mexican International Conference on Artificial Intelligence*, 312-321 (2008).
- [7] S. Visa and A. Ralescu, *Proc. International Conference on Machine Learning 2003 workshop: learning with imbalanced data sets II*, 97-104 (2003).
- [8] H.T. Xiong, J.J. Wu and L. Liu, *Proc. of the 2010 International Conference on Electronic-Business Intelligence*, 491-497 (2010).
- [9] D. Soria, J.M. Garibaldi, F. Ambrogi, E.M. Biganzoli and I.O. Ellis, *Knowledge-Based Systems* **24**, 775-784 (2011).
- [10] J. Woo, H.J. Choi and H.K. Kim, *Applied Mathematics & Information Sciences* **6**, 291S-302S (2012).
- [11] R. Kretschmar, N.B. Karayiannis, and F. Eggimann, *Proc. of the International Joint Conference on Neural Networks*, 517-522 (2003).
- [12] A.K. Jain, R.P.W. Duin and J. Mao, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**, 4-37 (2000).
- [13] A. Gangemi, D.M. Pisanelli and G. Steve, *Data and Knowledge Engineering* **31**, 183-220 (1999).
- [14] J. Wu, H. Xiong and J. Chen, *Data Mining and Knowledge Discovery* **20**, 191-220 (2010).
- [15] Y. Yun, *Applied Mathematics & Information Sciences* **6**, 737-743 (2012).
- [16] C. L. Blake, C. J. Merz, UCI repository of machine learning databases, <http://kdd.ics.uci.edu>.
- [17] H. He and E.A. Garcia, *IEEE Transactions on Knowledge and Data Engineering* **20**, 1263-1284 (2010).



Haitao Xiong

obtained his PhD at Beihang University in 2011 and since then he is a lecturer at School of Computer and Information Engineering, Beijing Technology and Business University. He has published over 10 papers in peer reviewed journals. His

research interests include machine learning, data mining, business intelligence and management information system.