Advances in Engineering Research (AER), volume 148
3rd Workshop on Advanced Research and Technology in Industry Applications (WARTIA 2017)

Research on Building Bank Anti-fraud Model Based on Tri-training Semi-Supervised Learning and Fuzzy SVM Active Learning

Xiaoguo Wang[1,a], Luxi Liu[1,b,*]

[1]School of Electronics Engineering, Tongji University, Cao'an Road, Shanghai, China

[a]xiaoguowang@tongji.edu.cn, [b]1531705@tongji.edu.cn, [c] email

*corresponding author

Keywords:semi-supervised learning, active learning, tri-training, fuzzy SVM, anti-fraud.

Abstract: With the rapid development of Internet finance and its applications, online banking fraud is becoming increasingly frequent. How to accurately identify the fraud data among the huge amount of transactions, is the urgent needs of Third Party Payment Center, Channel Department and other departments of banks. As to this problem, this paper proposes a recognition method based on tri-training semi-supervised learning and fuzzy SVM active learningaiming at researching thathow to build aneffectiveanti-fraud model for banks. Experimental results show that the method has encouraging recognition accuracy, which provides an effective scheme for banks' anti-fraud model training, building and fraud recognition.

1. Introduction

With the rapid development of Internet finance, the traditional anti-fraud risk control model has been difficult to cope with the evolution of complex Internet fraud. The Central Bank has made it clear that it is necessary to "build the anti-fraud information management system of banking". Constructing the intelligent anti-fraud model and accurately identifying suspicious transactions are becoming the urgent needs of the bank. Semi-supervised learning and active learning as the research hotspot in the field of machine learning, provide theoretical and technical support for constructing anti-fraud model. In this paper, the construction of bank anti-fraud model will be discussed and studied by combining the characteristics of semi-supervised learning and active learning.

2. Semi-supervised learning and active learning

2.1. Semi-supervised learning and tri-training

Cooperative banks have massive trading data with only a small amount of labeled data. It takes lots of manpower and time to access more labeled data. Traditional supervised learning requires sufficient labeled data as training data, while unsupervised learning only uses unlabeled data and it's difficult to ensure high accuracy. Semi-supervised learning can take full advantage of unlabeled data and labeled data, not only ensuring a sufficient quantity of the training set, but also being able to build a model with higher accuracy [1].

From the semi-supervised learning is proposed, there are a variety of new learning methods, collaborative learning as a classic semi-supervised learning algorithm is mainly used for image retrieval, natural language processing. Tri-training is based on the collaborative learning, abandoning the full redundant view and using three classifiers. Different classifiers choose different training subsets by bootstrap sampling to ensure the difference of different classifiers. After training, the three classifiers are integrated by voting to obtain the final classifier.

2.2. Active learning and fuzzy SVM

Active learning can filter out unlabeled data that has more information and is more useful for model's training to be labeled by professors, raising the number of the labeled data. Then the new

labeled data will be trained again, the process of "label-retrain" will be iterated for many times to get the final classifier with higher performance. Active learning has been widely used[2] in the text classification[3], image retrieval[4], intrusion detection and other aspects. In this paper, there is a small amount of noise data, that is, there may be fraud data in the labeled non-fraud data. Fuzzy SVM can reduce the impact of noise data on the model.

SVM is a general classifier based on statistical learning theory, which can solve the high dimension problem, and has good generalization ability and robustness. Theoretically, SVM can get the global optimal solution, but in fact, due to the existence of noise, SVM will have a greater impact on the optimal classification surface. On the basis of SVM, fuzzy SVM evaluates the data and objective function by fuzzy membership and penalty. The noise point will have smaller fuzzy membership and smaller weight, which can greatly reduce the impact to the optimal classification surface.

3. Design

Combined with semi-supervised learning and active learning, insufficient number of labeled data can be improved more targeted, and hidden information that unlabeled data carries can be fully excavated, improving the accuracy of anti-fraud model from the whole.

3.1. Selection of tri-training algorithm

The process of building anti-fraud model is the process of data mining and learning, the process of judging whether the transaction data is fraud is the process of classifying data. The methods of data mining commonly used are decision tree, k-means, SVM, Apriori, EM, AdaBoost, kNN, naive Bayesian and so on. Among them, Apriori is mainly used for mining Boolean association rules; k-means, EM is mainly used for clustering; AdaBoost is easy to be effected by noise, and the training time will be too long when the amount of data is too large; kNN based on Continental distance, but the relationship between transaction data can not be accurately expressed in the Euclidean distance; Naive Bayesian requires that the conditions are independent, but many fraud transactions often occur in a short time, and transaction hour, amount, quantity are not independent of each other.

In contrast, decision tree shows good performance[5]. In the bank data, there are some attributes missing values, the use of decision tree can reduce the impact of missing values. In addition, decision tree can train a large number of transaction data in a relatively short time, and get relatively good results. The results of the decision tree training are more explanatory, easy to understand, and can be derived from the model to fraud rules.

3.2. Data processing

Reduced sampling: There is billions of bank data, it takes lots of time to train such a large amount of data. In order to reduce the training time, unlabeled data will be reduced sampled. This process assumes that the fraud data is evenly distributed in the transaction data, and the proportion of fraud data to the transaction data remains unchanged.

Data expansion: Because the amount of fraud data (labeled data) is very small, all data features can not be obtained from a little data, so fraud data expansion is necessary. This process assumes that data belonging to the same category is adjacent in the feature space and that the data between any two of the same categories also belong to that category. Through the data expansion, the number of labeled data sets has been improved, and more feature information has been provided for the model training.

3.3. Algorithm description

Figure 1 is a pseudo-code description of the algorithm in this paper. Assuming that the initial labeled data set is L, the unlabeled data set is U, the random sampling algorithm is bootstrap, the number of data that needs to be labeled by experts for each iteration is m, and the number of sampling from the labeled data set and the unlabeled data set is respectively Ml and Mu.

```
// Initial classifier
for i = [0,1,2]
    TrainingSet[i] = bootstrap(L,Ml)
classifier = tritraining(TrainingSet)
// The process of "label-retrain"
while(U is not empty and some condition) {
    UnlabelTrainingSet = bootstrap(U,Mu)
    FSVM(UnlabelTrainingSet)
    ToLabelDataSet = getMinFuzzy(UnlabelTrainingSet, m)
    NewLabelDataSet = Label(ToLabelDataSet)
    L = L + NewLabelDataSet
    U = U - NewLabelDataSet
    for i = [0,1,2]
    TrainingSet[i] = bootstrap(L)
    classifier = tritraining(TrainingSet)
}
```

Figure 1 This caption has one line so it is centred.

Firstly, get three different train data sets from L by random sampling, and use tri-training to train to get initial classifier. If U is not empty and some certain iterate conditions are satisfied, go to next step. The certain iterate conditions are like, the number of iterations does not reach the specified maximum times, the accuracy of classification does not reach the specified minimum accuracy, the training time dost not reach the specified maximum time, and so on. When looping, random sampling M(M > m) data from U, get suspicious data by FSVM, and calculate the fuzzy membership of all suspicious data. m data with the smallest fuzzy membership will be labeled by experts. Add new labeled data to L and deleted from U. Use tri-training to train new L again to get new classifier until the end of loop meets the conditions. The classifier get in the last time is the final classifier.

4. Experiments

4.1. Experiment data

The experiment uses cooperative bank third party transaction data, as Table 1 shows. T represents normal data, and F represents fraud data.

Table 1Experiment data.

Total number		Labeled data		Unlabeled data	Test data	
T	F	T	F		T	F
115107	61	22460	31	61765	30882	30

4.2. Design

For fuzzy SVM，use 10 cross-validation，the penalty parameter varies {2^0,2^1,…,2^15 }，kernel function is K(xi, xj)=exp(-γ ‖ xi-xj ‖ ^2)， γ varies {2^(-15),2^(-14),…,2^0 }。

The experiment use tri-training, fuzzy SVM and combination of tri-training and fuzzy SVM respectively to build model and compare the performance of different algorithm in bank anti-fraud.

Experiment environment: Linux, 2GHz CPU, python 3.7.

4.3. Experiment data

In this paper, 100 experiments have been tested on data set and the results are the average with two decimal. The algorithm of this paper is marked as TFSVM and the results show as Table 2. Among them, F represents the number of fraud data which model classifies, and TF represents the number of true fraud data which model classifies.

Table 2Result.

Algorithm	Average		Maximum
	F	TF	TF
Tri-training	47.36	4.07	9
FSVM	0	0	0
TFSVM	67.54	6.77	10

The experiment uses recall and precision as a measure of model effect. Because FSVM does not recognize any fraud data, Figure 2 only compares the recall and precision of tri-training and TFSVM.

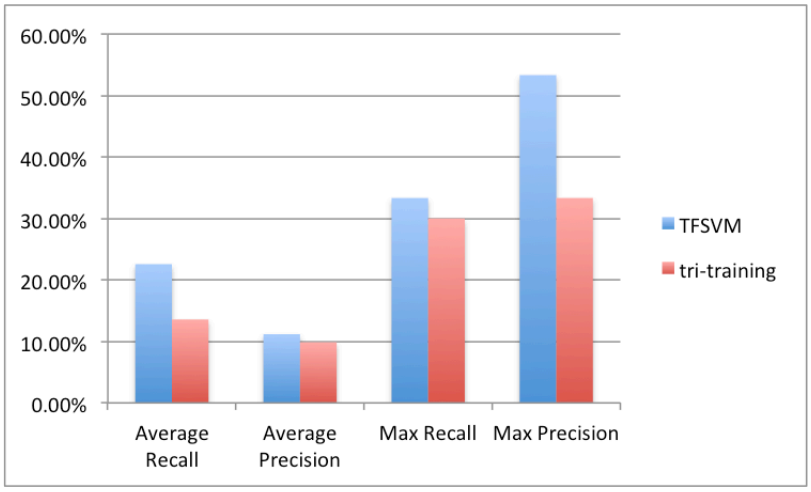

Figure 2 Recall and precision.

Because the number of positive samples (normal data) in the experiment is much larger than the negative sample (fraud data), the imbalance of the sample ratio leads to the separation of the FSVM from the positive sample, and the test data are judged as positive samples. Models created by tri-training can identify a certain amount of fraud data.

The algorithm proposed in this paper, on the basis of Tri-training, uses FSVM to increase the number of labeled data, and expands the negative data, so that the sample ratio is corrected in a certain degree. It can be found from the experiment that TFSVM both improve the recall, but also improve the precision.

5. Conclusion

For the demand of anti-fraud modeling and transaction recognition in bank transaction, based on semi-supervised learning and active learning, this paper studies the modeling and application of tri-training and fuzzy SVM. Experiments show that this method can recognize user transaction accurately, and provides an effective way to solve the problem of bank fraud recognition.

References

[1] Qian, T., Liu, B., Chen, L., &Peng, Z. (2014). Tri-Training for Authorship Attribution with Limited Training Data. In ACL (2), 345-351.

[2] Settles, B. (2010). Active learning literature survey. University of Wisconsin, Madison, 52(55-66), 11.

[3] Youngdoo Son, Jaewook Lee. (2016). Active learning using transductive sparse Bayesian regression.Information Sciences, 240-254.

[4] IoannisSarafis, ChristosDiou, AnastasiosDelopoulos. (2015). Building effective SVM concept detectors from clickthrough data for large-scale image retrieval. International Journal of Multimedia Information Retrieval, 2015, Vol.4 (2), pp.129-142.

[5] Delen, D., Kuzey, C., &Uyar, A. (2013). Measuring firm performance using financial ratios: A decision tree approach. Expert Systems with Applications, 40(10), 3970-3983.