DOI: 10.21276/sjpms.2016.3.3.4

Scholars Journal of Physics, Mathematics and Statistics **ISSN 2393-8056 (Print)**
Sch. J. Phys. Math. Stat. 2016; 3(3):117-122 **ISSN 2393-8064 (Online)**
©Scholars Academic and Scientific Publishers (SAS Publishers)
(An International Publisher for Academic and Scientific Resources)

Research of Imbalanced Data Classification in Data Mining

Xin Hua, Zhou Shao Hua, Hu Jin Yan
Northeast Petroleum University, Daqing City of Heilongjiang Province, 163318, China

***Corresponding Author:**
Xin Hua
Email: xinhuayatou@126.com

Abstract: Classification is one of the most important research contents in data mining and traditional classification methods are relatively mature, when dealing with well-balanced data they can make good performances. But in real world the data is usually imbalanced, that is, most of the data are in majority class and little data are in minority class. Imbalanced data set cause the deduction of the precision of the minority class samples, when it is classified by traditional algorithm, which can tend to favor the more class samples. Making researches on imbalanced datasets are quite important. In order to help readers to have a clear idea of the currently proposed and future work data classification, in view of imbalanced data progress, this paper introduced three developed methods: data level, algorithmic level and developed methods that were the performance evaluation of imbalanced data classification. We are very glad to receive the valuable reference provided by the academics that interested in this field.

Keywords: Data Mining, Imbalanced data, Classification, Performance Evaluation.

INTRODUCTION

The imbalanced datasets are abound in real life and industrial production. In an imbalanced dataset, the majority class have a large percentage for all the samples, while the samples in minority class just occupy a small part of all the samples. For example, the bank needs to build a classifier to predict that whether customers will conduct trust loans in future [1]. While the number of conducting trust loans around only about 1% of total customers. If a classifier predicted that all customers will not conduct trust loans, it has an accuracy of 99%. Obviously, this result does not make any sense. Therefore, we need a classifier that it can effectively predict the minority class to help companies saving money. There is the probability of detection data to diagnose whether the patient is suffering from cancer; the number of cancer patients is very low among healthy people. Imbalanced data classification is also involved in positioning of oil well from a satellite image [2], learning pronunciation of the word [3], distinguishing malicious telephone harassment [4] and risk management, etc.

In recent years, the classification problems of imbalanced datasets have aroused academic attention in data mining and machine learning and it has become one of the hot topics of data mining and machine learning. The American Association for Artificial Intelligence (AAAI) in 2000, which is the first International Conference on the imbalanced data. International Conference on Machine Learning (ICMLZOO3) in 2003 held a special session for imbalanced data problems. Association for Computing Machinery (ACM) in 2004 published a newsletter for the topic. Since the (ICEC) seminars was held on April 27, 2009 in Bangkok, Thailand, more and more conferences for imbalanced data are carried out. With the increasing of sharp and practical problems, many scholars put forward specific methods.

ANALYSIS OF DATA LEVEL

There are three methods for imbalanced data classification: the data level, the algorithm level, and the combination of two levels. On the data level, it is divided into multi-classifier committee and re-sampling approaches.

Multi-classifier committee

Multi-classifier committee [5] approach makes use of all information on a training data set. And it divides the samples with majority class randomly into several subsets, and then takes every subset and all the samples with minority class as training dataset, respectively. After training these datasets separately, several classifiers are available as committees. Multi-classifier committee approach uses all the classifiers to predict a sample and decides the final class to it by the prediction results of the classifiers. Voting is one simple method for making a final class decision to a sample in which a minimum threshold is set up. If the number of classifiers that predict a sample exceeds the minimum threshold, then the final class prediction of this sample can be categorized. Though multi-classifier committee approach does not abandon any sample from MA, it may be inefficient in the training time for all the committees and can-not ensure the quality for every committee. Further selection of the committees will make the predictions more correct and more efficient.

Re-sampling methods are divided into over sampling and under-sampling

(1) Oversampling approach is a common method of dealing with imbalanced data, it changes the distribution of the training data to reduce the imbalanced data. Oversampling is randomly selected small sample, copy and added to the original sample, thereby improving the imbalanced class. However, this method does not add new information, at the same time, it increases the time to build a classifier that may lead an over-fitting problem. SMOTE (Synthetic Minority Over-sampling technique) proposed by Chawla, which aims to improve the oversampling method [6], it artificially synthesizes small sample based on the feature space. The drawback is that when the original data is already too much, which in turn it synthesizes some new small samples, resulting in data burdensome, while neglecting its neighbor samples will increase the chance of overlap between categories. Borderline-SMOTE algorithm that proposed by Han Hui [7], which aims to form a new dataset with small samples of the boundary decision. The disadvantages are its performance depends and selected neighbor numbers. If the number of neighbors is too small that some small samples will be mistaken for noise data. ADASY algorithm [8] that proposed by H. He, it generates different numbers of synthetic data based on characteristics of data distribution itself. Safe-level-SMOTE algorithm [9] that proposed by C. Bunkhumpornpat，its features that synthetic samples are closer to small samples，thus such synthetic samples within the overlapping region of the class becomes more sparse, which will help demarcating borders. These algorithms generate small samples in accordance with some rules and factors in order to reduce the chance of overlap between classes.

(2)Under-sampling approach: under-sampling method is a method that is relatively to re-sampling through reducing majority class to improve the classification performance of the minority class. The most direct way is randomly removing some of the majority class. However，the disadvantage is that it may lose many useful information so people have made many improvements in under-sampling method. One-sided selection algorithm that M. Kubat proposed that devotes to achieve as much as possible without deleting useful sample. Liu Xuying and other people put forward Easy Ensemble and Balance Cascade algorithm [10]. Easy Ensemble combines the under-sampling with Ada Boost，in iterative algorithm，it randomly selects one from a big class and the same number of subclass samples of the sampling training classifier. (One-Sided Selection, OSS) method [11] that proposed by M. Kubat is a representative of under-sampling method. OSS algorithm divides the majority class through determining the distance between samples, which includes noise samples and border samples. NCL method [12] is an improved method of OSS with proposing 3- neighbor classifier to select big categories samples. Based k- nearest neighbor method，J. Zhang and others put forward four under-sampling approaches, according to the experiments, it shows that NearMiss-2 method performs well relatively. (Cluster-Based Under-Sampling, CBUS) [13] put forward by J. Yuan and others, it can maintain the overall structure of the big samples in order to reduce redundant information. (Multiple Random Under-Sampling, MRUS) [14] proposed by H.M. Nguyen and others, it repeats random under-sampling for several times and then averaged them. Multiple random under-sampling method has more stability and high-efficiency compared with random under-sampling.

THE DEVELOPMENT OF ALGORITHM LEVEL

Cost-effective learning algorithm

Cost-effective learning algorithm is based on a hypothesis that the values of minority correct classification are higher than that of majorities, and it supposes that the cost information is known that means the cost matrix is known. Generally, the cost information is provided by experts in this field, which it is a disadvantage of this approach. Currently, there are two aspects in cost-sensitive learning, firstly, to reconstruct the training datasets according to sampling misclassification, which means weighting every samples in training datasets and reconstructing the original samples. The drawback is that it loses some information during the reconstruction process. Secondly, based on the traditional algorithms, it introduced the cost-sensitive factor that small samples with higher costs, big samples with less costs in order to balance the differences between the sample numbers. The document [15] continued to modified cost-effective SVM of Veropoulos，the basic idea is that the cost associated with the slack variables in order to make SVM cost-sensitive. Lee and others have designed SVM of cost-sensitive for multi-class problems and taken the sampling bias [16] into account. The document [17] has proposed a weighted Fisher linear discriminant model (WFLD), which weights scatter matrix within positive and negative classes so that the two kinds of positive and negative samples can balance covariance matrix within the overall scatter class The document [18] proposed a part cost sensitive algorithms, for a new sample, firstly, it selects an appropriate distance metrics, then choosing the k-nearest neighbor samples and training k nearest neighbors in weighting, finally, a classifier is available.

Integrated learning

For imbalanced data**,** the sampling combined with integrated together in most of times and then determined the final result by voting. Ada Boost [19] is an iterative algorithm, at each iteration, the training distributions are given to different weights. It increases misclassification samples' weight after iteration and reduces the weight of the correct classification. Ada Boost has limited ability to improve the recognition rate of positive class samples. However**,** it has made a lot of improvements, such as Ada Cost [20], it proposed that the misclassified minority sample weights higher than that of majorities. Rare Boost [21] alters the update rules of weighting so that misclassified positive class have a higher weight than that of the misclassified negative sample. Document [22, 23] makes a combination between over-sampling and ensemble method, which it not only makes use of oversampling to adjust the class balance but also takes advantage of an integrated approach to improve the overall classified performance in imbalanced datasets. C-SMOTE algorithm in document [24] also adopts the combination of oversampling and integration algorithm.

One-class classifiers

In some cases that the dataset severely imbalanced, classifiers will determine all samples as a big class. One-class classifiers approach is born to solve this problem based on identification method instead of classification. It learns from interested samples through comparing the similarities between new samples and targeted classes to ensure whether the sample belongs to targeted class or not. Hippo [25] is a common one-class learning method, which adapts neural network as classification algorithms. Due to one-class classifiers only needs a class of datasets as the training samples and with less time, therefore, it has been applied in many areas.

Other approaches

The other common approaches including active learning [26], SVM [27], Neural learning [28], random forest to learn imbalanced data [29], subspace method [30], the fuzzy rules [31], feature selection method [32, 33] and the posterior probability SVM model solution method [34], a method for FLDA single positive class are all effective way to learn imbalanced datasets.

COMPREHENSIVE METHOD

Due to data sampling and classification algorithm have their advantages and disadvantages, the current researches combine the two methods. The combination method resamples the original data in order to lower the unbalance and also classifies with unbalanced classification algorithm of compensate data. For example, Rehan Ak-bani [35] used SMOTE+biased-SVM approach to upward sampling minorities to lower the unbalanced degree in algorithm, Veropoulos proposed biased-SVM approach have the different misclassification costs. Rehan Akbani through the experience to verify their approach far exceed one ways. How to achieve the perfect combination of two approaches and how to achieve their advantages have become hot topics.

FOURTH THE EVALUATION CRITERIA AND DEVELOPMENT

In the traditional classification learning, in general, it adapts classification accuracy, that is, the number of correctly classified samples account for the percentage of total number. But for imbalanced data, this method is not scientific. Supposed that the majority class accounted for 99% of total numbers, even if the minority are all considered as wrong, it can be obtained 99% classification accuracy. Therefore, for imbalanced data, scholars have proposed a series of related evaluation criteria, which divided into numerical approach and graphical metric approach. Numerical approach provides thresholds to evaluate classifiers is good or bad, which includes accuracy, precision (precision), recall (recall), F-measure, g mean, AUC [36, 37], Kappa coefficient [38], F1-Measure [39] and so on. The graphical metric approach is a graphical drawing of two-dimensional or three-dimensional image that people can observe easily and intuitively, which includes the ROC curve (receiver operating characteristics curves), precision-recall curves, and cost curve, etc. [40].

Table1: confusion matrix

	ClassificationPositives	ClassificationNegatives
Positives	TP（True Positives）	FN（False Negatives）
Negatives	FP（False Positives）	TN（True Negatives）

Numerical measurement

（1）Accuracy rate： $Acc=\dfrac{TP+TN}{TP+FN+TN+FP}$， error rate： $Err=\dfrac{FP+FN}{TP+FN+TN+FP}$ ；

（2）Precision

$$\text{Precision} = \frac{TP}{TP+FP} \quad ;$$

（3）recall

$$TPrate = \frac{TP}{TP+FN} \quad ;$$

（4）F-value criterion

$$F-value = \frac{\left(1+\beta^2\right)\times recall\times precision}{\beta^2\times recall + precision} \quad ;$$

Here β is a coefficient to adjust Precision and Recall, in most cases it will be equal to 1, F-measure criteria takes the minority class of precision and recall into account, which is one of the commonly used evaluation criteria in imbalanced classification performance;

（5） $G-mean=\sqrt{TPrate\times TNrate}$

where

$$FPrate=\frac{FP}{FP+TN}\ ,\quad TNrate=\frac{TN}{FP+TN}\ ,\quad FNrate=\frac{FN}{TP+FN}\ ,$$

Under the balance of G-mean maintaining the accuracy of positive and negative class. It maximizes the accuracy of two classes. Only when both higher in order to obtain a high G-mean values, it reflects the overall classification performance of the data.

（6）Kappa coefficient

$$kappa=\frac{po-pe}{1-pe}\ ,$$

where

$$po=\frac{TP+TN}{TP+TN+FN+TN}\ ,$$

$$pe=\frac{\left(\frac{A\cdot C}{N}+\frac{B\cdot D}{N}\right)}{N}\ ,\quad A=TP+FN\ ,\quad B=FP+TN\ ,\quad A=TP+FP\ ,\quad D=FN+TN\ ;$$

Kappa coefficient is used to test the degree of consistency of the two predictions, which Po is represented by observe consistency ratio, Pe is the desired consistency ratio. When the two measurements are the same, the Kappa coefficient is 1. When Po is greater than Pe, Kappa coefficient is a positive number, the greater consistency, the better;

(7) F1-Measure

$$F1=\frac{2\times recall\times precison}{recall+precison} \quad ;$$

F1-Measure is a comprehensive evaluation that Precision and recall synthesized, during the experimental process, we hope the higher Precision and the higher Recall, the higher the better, but in fact these two indicators in some cases are negatively correlated. Therefore, the introduction of F1- Measure is a comprehensive combination of these two; it can be seen from the above formula, the higher the F1- Measure, the better classifier.

Graphical measurement

（1）ROC curve（receiver operating characteristics curves ）

ROC curve is the most commonly used evaluation criteria, which the vertical axis represents the TP rate, the horizontal axis represents the FP rate. ROC curve first appeared in signal detection, which aims to balance between the

TP rate and the FP rate. Spackman introduces the ROC analysis into the machine learning that aims for evaluation and comparison algorithms. ROC curve reflects the relationship between the TP rate and FP rate when the classifier parameter changes. (0, 1) for the ideal classification, all samples were classified correctly. Therefore, ROC curve is the closer to the upper left corner, the better the performance of the classifier.

Due to the ROC curve does not give a specific evaluation value, it is inconvenient to compare between the performances of different classifiers, so people often use the area under ROC, AUC as an evaluation index.

$$AUC = \int_0^1 y dx$$

Larger AUC value corresponds to the optimum classifier. AUC is a unique value based on the ROC curve, and it has nothing to do with the mistake classification cost, which is not affected with relevant factors of rules.

(2) Precision-recall curves and the cost curve are also common graphical measurements.

PROSPECTS

Classification of imbalanced data set is one of the tough problem of data classification; the main difficulty is caused by its own characteristics and limitations of traditional algorithm classification. The imbalanced data is important factor that impedes the widely uses of the learning machine in recent years, which it has aroused widespread concerns. Imbalanced problems are prevalent in many practical applications and how to effectively improve its classification performances is the common goal pursued by researchers. For imbalanced data classification, people has raised many solutions and made some progress, but it is necessary to do further research, such as researches on the efficiency of algorithm and time coats and how to adaptive determine the best sampling proportion. So far, the vast majority research of current imbalance problems are considered with the imbalance proportion of data. There is another case of imbalanced data: that is the two classes with a considerable number of data categories, but the differences of the class distribution are quite obvious, a class of relatively concentrated, and the other more dispersed and the current studies on the differences of class distribution are less. In addition, how to mix the feature selection into imbalanced classification algorithm also need further research in future.

Acknowledgement

Youth Foundation of Northeast Petroleum Data mining technique and algorithm research No: NEPUQN2014-24

REFERENCES

1. Ezawa KJ, Singh M, Norton SW; Learning goal oriented Bayesian networks for telecommunications risk management. ICML, 1996; 139-147.
2. Kubat M, Holte RC, Matwin S; Machine learning for the detection of oil spills in satellite radar images. Machine learning, 1998; 30(2-3): 195-215.
3. Van Den Bosch A, Weijters A, Van Den Herik H J, et al. When small disjuncts abound, try lazy learning: A case study[C]//Proceedings of the Seventh Belgian-Dutch Conference on Machine Learning. 1997: 109-118.
4. Fawcett T, Provost FJ; Combining Data Mining and Machine Learning for Effective User Profiling, KDD, 1996; 8-13.
5. Argamon-Engelson S, Dagan I; Committee-based sample selection for probabilistic classifiers. J. Artif. Intell. Res.(JAIR), 1999; 11: 335-360.
6. Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of artificial intelligence research, 2002: 321-357.
7. Han H, Wang WY, Mao BH; Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. Advances in intelligent computing. Springer Berlin Heidelberg, 2005; 878-887.
8. He H, Bai Y, Garcia E, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning[C]//Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on. IEEE, 2008: 1322-1328.
9. Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C; Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2009; 475-482.
10. Liu XY, Wu J, Zhou ZH; Exploratory under sampling for class-imbalance learning. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 2009; 39(2): 539-550.
11. Kubat M, Matwin S; Addressing the curse of imbalanced training sets: one-sided selection. ICML, 1997; 97: 179-186.

12. Mani I, Zhang I; KNN approach to unbalanced data distributions: a case study involving information extraction. Proceedings of Workshop on Learning from Imbalanced Datasets, 2003.
13. Yuan J, Li J, Zhang B; Learning concepts from large scale imbalanced data sets using support cluster machines. Proceedings of the 14th annual ACM international conference on Multimedia. ACM, 2006; 441-450.
14. Nguyen HM, Cooper EW, Kamei K; A comparative study on sampling techniques for handling class imbalance in streaming data. Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS), 2012 Joint 6th International Conference on. IEEE, 2012; 1762-1767.
15. Raskutti B, Kowalczyk A; Extreme re-balancing for SVMs: a case study. ACM Sigkdd Explorations Newsletter, 2004; 6(1): 60-69.
16. Lee Y, Lin Y, Wahba G; Multi-category support vector machines: Theory and application to the classification of microarray data and satellite radiance data. Journal of the American Statistical Association, 2004; 99(465): 67-81.
17. Xie J, Qiu Z; The unbalanced data sets Fisher linear discriminant model. Journal of Beijing jiaotong university: natural science edition, 2006; 30(5): 15-18.
18. Karagiannopoulos MG, Anyfantis DS, Kotsiantis S B, et al. Local cost sensitive learning for handling imbalanced data sets[C]//Control & Automation, 2007. MED'07. Mediterranean Conference on. IEEE, 2007: 1-6.
19. Schapire RE, Singer Y; Improved boosting algorithms using confidence-rated predictions. Machine learning, 1999; 37(3): 297-336.
20. Fan W, Stolfo S J, Zhang J, et al. AdaCost: misclassification cost-sensitive boosting[C]//ICML. 1999: 97-105.
21. Joshi MV, Kumar V, Agarwal RC; Evaluating boosting algorithms to classify rare classes: Comparison and improvements. Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on. IEEE, 2001; 257-264.
22. Chawla NV, Japkowicz N, Kotcz A; Editorial: special issue on learning from imbalanced data sets. ACM Sigkdd Explorations Newsletter, 2004; 6(1): 1-6.
23. Chawla N V, Lazarevic A, Hall L O, et al. SMOTEBoost: Improving prediction of the minority class in boosting[M]//Knowledge Discovery in Databases: PKDD 2003. Springer Berlin Heidelberg, 2003: 107-119.
24. He G, Han H, Wang W; An over-sampling expert system for learning from imbalanced data sets. Neural Networks and Brain, 2005. ICNN&B'05. International Conference on. IEEE, 2005; 1: 537-541.
25. Japkowicz N, Myers C, Gluck M; A novelty detection approach to classification. IJCAI, 19.
26. Constantinopoulos C, Likas A; Semi-supervised and active learning with the probabilistic RBF classifier. Neurocomputing, 2008; 71(13): 2489-2498.
27. Raskutti B, Kowalczyk A; Extreme rebalancing for SVMs: a ease study. SIGKDD Explorations, 2004; 6(1):60-69.
28. Murphey YL, Guo H, Feldkampl A; Neural learnehan Akbaniing from unbalanced data. Applied Intelligence, 2004; 21(2):117-128.
29. Chen C, Liaw A, Breiman L; Using random forest to learn imbalanced data. University of California, Berkeley, 2004.
30. Ahn H, Moon H, Fazzari M J, et al. Classification by ensembles from random partitions of high-dimensional data[J]. Computational Statistics & Data Analysis, 2007, 51(12): 6166-6179.
31. Visa S, Ralescu A; Learning imbalanced and overlapping classes using fuzzy sets. Proceedings of the ICML, 2003; 3.
32. Zheng Z, Wu X, Srihari R; Feature selection for text categorization on imbalanced data. ACM Sigkdd Explorations Newsletter, 2004; 6(1): 80-89.
33. Mladenic D, Grobelnik M; Feature selection for unbalanced class distribution and naive bayes. ICML, 1999; 99: 258-267.
34. Tao Q, Wu G W, Wang F Y, et al. Posterior probability support vector machines for unbalanced data[J]. Neural Networks, IEEE Transactions on, 2005, 16(6): 1561-1573.
35. Akbani R, Kwek S, Japkowicz N; Applying support vector machines to imbalanced datasets. Machine Learning: ECML 2004. Springer Berlin Heidelberg, 2004; 39-50.
36. Daskalaki S, Kopanas I, Avouris N; Evaluation of classifiers for an uneven class distribution problem. Applied artificial intelligence, 2006; 20(5): 381-417.
37. Mozer MC; Optimizing classifier performance via an approximation to the Wilcoxon-Mann-Whitney statistic, 2003.
38. McHugh ML; Interrater reliability: the kappa statistic . Biochemia Medica, 2012; 22(3): 276-282.
39. Maratea A, Petrosino A, Manzo M; Adjusted F-measure and kernel scaling for imbalanced data learning. Information Sciences, 2014; 257: 331-341.
40. Spackman KA; Signal detection theory: Valuable tools for evaluating inductive learning. Proceedings of the sixth international workshop on Machine learning. Morgan Kaufmann Publishers Inc., 1989; 160-163.