Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 12, December 2014, pg.181 – 188

REVIEW ARTICLE

A Review on Rapidly Convergence Approach for Handling Class Imbalance Data Set

[1]Ms. Barkha R. Hadke, [2]Prof.Vikrant Chole

[1]Department of Computer Science and Engineering
G. H. Raisoni College of Engineering and Technology, Nagpur, India
barkha.hadke1990@gmail.com

[2]Department of Computer Science and Engineering
G. H. Raisoni College of Engineering and Technology, Nagpur, India
vikrant.chole@raisoni.com

Abstract: In recently class imbalance problem have drawn growing because of their classification difficulties. Class imbalance problem become greatest issue in data mining. Imbalance problem occur where one of the two classes having more sample than other classes. For creating good prediction model, a well balance dataset is very important. As the application area of technology increases the size of data also increases. To handle this major issue of the imbalance class the most existing classification methods and many ensemble methods have been proposed to deal with such imbalance problems. In this paper we examine the different methods of over-sampling and under-sampling techniques to balance data.

Keywords: Imbalanced Classification, Resampling, Over-Sampling, Under-Sampling, Oversampling, SMOTE, Gibbs Sampling

I. Introduction:

Class imbalance problem, is a one of major issue in data mining. This problem referred to the distribution samples in classes in which one class contain large number of samples as compared to another class. Such imbalance distribution of class create a problem during classification as most of the classifier consider only majority class sample for classification, that's why class imbalance is one of the major problem in data

mining. The application of imbalance problem includes fraud detection, medical diagnosis, text classification, oil spills detection, etc. To overcome these problems it uses resampling techniques and methodologies for imbalance problem by either eliminating some data from the majority class or adding some duplicated data to the minority class.

Resampling techniques can be categorized into three groups. Undersampling methods, which create a subset of the original data-set by eliminating instances, oversampling methods, which create a superset of the original data-set by replicating some instances or creating new instances from existing ones; and finally, hybrids methods that combine both sampling methods.

Oversampling is resampling technique which used for balancing the classes by adding samples in class. Undersampling is resampling technique which used for balancing the classes by removing samples from the majority class.For the evaluation of performance of class imbalance methods introduce the concept of True Positive, True Negative, False Positive and False Negative:

- True Positive (TP) – An example that is positive and is classified correctly as positive
- True Negative (TN) – An example that is negative and is classified correctly as negative
- False Positive (FP) – An example that is negative but is classified wrongly as positive
- False Negative (FN) – An example that is positive but is classified wrongly as negative

The methods based on some form of re-sampling methods are listed below:

1. Oversampling the minority class, most commonly implemented as random oversampling, (ROS).
2. Under sampling the majority class, most commonly implemented as random under sampling,(RUS).
3. Use of SMOTE algorithm to artificially synthesize items belonging to the minority class.
4. The implementation of cost sensitive learning.
5. Boosting.

II. RELATED WORK

The class imbalance problem proposed at data level and algorithm levels. At Data level methods consists of resampling the original data set for balancing the classes. At the algorithmic level, it includes adjusting the costs of the class imbalance for adjusting the probabilistic estimate learning. Cost-Sensitive Learning is a type of learning in data mining that takes the misclassification costs. [9] There are many ways to implement cost sensitive learning such that misclassification costs, cost-minimizing techniques and cost Sensitive techniques.

Some strategies are proposed to solve classification problems based on imbalanced data sets. There are many techniques developed to address the class imbalance issue. One of them is resampling technique. Resampling techniques can be used with many base classifiers, such as support vector machine (SVM), C4.5, Naïve Bayes classifier, AdaBoost etc.,

Resampling [12] is divided into two major approaches i.e. under-sampling approach and over-sampling approach. Under-sampling uses only some samples of the majority class to reduce the data size, and removes samples of the majority class to balance a data set. The over-sampling approach is to add more new data instances to the minority class to balance a data set. These new data samples can either be generated by replicating the data samples of the minority class or by applying synthetic methods. However, over-sampling often involves making exact copies of samples which may lead to over-fitting [6].Most existing imbalance learning techniques are only designed for two class problem. Multiclass imbalance problem mostly solve by using class decomposition. In advance sampling techniques it also combines various methods for balancing the classes. There is also alternative methods for class imbalance problem i.e. ensemble- learning, which includes Boosting and Bagging techniques. To overcome on the problem of these imbalance classes it address the existing oversampling approaches which do not consider generating new synthetic samples. As the result, the samples of the minority class represents highly misclassification error.

A. Synthetic Minority Oversampling Technique (SMOTE):

In minority class, it can be balance the class by using non heuristic method through random replication of positive samples, this method replicates existing in minority class and thus their overfitting problem is occurred in minority class.

Chawla proposed Synthetic Minority Over-sampling Technique (SMOTE) [8] an over-sampling approach in which the minority class is over-sampled with replacement. The SMOTE algorithm is used to generate artificial samples of the minority classes in an attempt to rebalance the dataset. It is an oversampling method, whose main idea is to create new minority class samples. SMOTE was tested on a variety of datasets, with varying degrees of imbalance and varying amounts of data in the training set. SMOTE forces focused learning and introduces a bias towards the minority class.

SMOTE creates samples by randomly selecting one of then nearest samples of a minority class. Thus, the overfitting problem is avoided and causes the decision boundaries for the minority class to be spread further into the majority class space.SMOTE Boost algorithm [9] combines SMOTE technique and the standard boosting procedure. It utilizes SMOTE for improving the accuracy over the minority class and utilizes boosting not to sacrifice accuracy over the entire data set.

Chawla also propose SMOTE-NC (Synthetic Minority Over-sampling Technique Nominal Continuous) and SMOTE-N (Synthetic Minority Over-sampling Technique Nominal), the SMOTE can also be extended for nominal features.

As an example, consider the classification of pixels in images as possibly cancerous. A simple dataset might contain 98% normal pixels and 2% abnormal pixels. Thus a majority class gives predictive accuracy of 98% and the purpose of this application is requires a comparatively high rate of correct detection in the minority class and allows for a small error rate in the majority class. In this case, the predictive accuracy is not

appropriate. Therefore Receiver Operating Characteristic (ROC)[10]standard technique is introduced for summarizing classifier performance over a range of tradeoffs the between true positive and false positive error rates.

For the ROC curve, the Area Under the Curve (AUC) is an accepted traditional performance metric.

B. Modified synthetic minority oversampling technique(MSMOTE):

The new modified version of SMOTE is MSMOTE[8]. The samples of minority class can be categories into three different groups i.e *.safe*, *border* and *latent noise* instances by the calculation of the distances among all examples. MSMOTE is also creating new samples as similar to SMOTE, the strategy to select the nearest neighbours is changed with respect to SMOTE that depends on the group previously assigned to the instance. At the first category i.e. safe instances, the algorithm randomly selects a data point from the *k*NN; for border instances, it only selects the nearest neighbor; finally, for latent noise instances, it does nothing.

C. Boosting Method

Boosting [4] is a technique to progress the performance of weak classifiers. AdaBoost [7] is the most known boosting algorithm. In each iteration, the weights are modified with the objective of correctly classifying examples in the subsequently iteration. At the end, all customized models contribute in a weighted vote to classify unlabeled examples. This method is more helpful to deal with class imbalance problem because minority class examples are mainly expected to be misclassified and hence given higher weights in subsequent iterations.

D. Feature selection:

Feature selection is a process of selecting subset relevant features for used in model construction. The techniques for feature selection are commonly classified as filter approach, wrapper approach, and embedded approach. Filter approach and embedded approach are relatively computationally efficient and are commonly applied as a fast feature ranking procedure. In contrast, wrapper approach evaluates features by performing internal classification with a given inductive algorithm. Therefore, they are much more computation intensive. Nevertheless, wrapper approach remains attractive for two reasons. Firstly, wrapper approach evaluates features iteratively with respect to an inductive algorithm. Feature selection is a key step for many machine learning algorithms, especially when the data is high-dimensional

1. Wrapper Method:

Wrapper methods use a predictive model to score feature subsets. Each new subset is used to train a model, which is tested on a hold-out set. Counting the number of mistakes made on that hold-out set gives the score for that subset. As wrapper methods train a new model for each subset, they are very computationally intensive, but usually provide the best performing feature set for that particular type of model.

2. Filter methods

Filter methods use a proxy measure instead of the error rate to score a feature subset. This measure is chosen to be fast to compute, whilst still capturing the usefulness of the feature set. Filters are usually less computationally intensive than wrappers, but they produce a feature set which is not tuned to a specific type of predictive model. Many filters provide a feature ranking rather than an explicit best feature subset, and the cutoff point in the ranking is chosen via cross-validation. Filter methods have also been used as a preprocessing step for wrapper methods, allowing a wrapper to be used on larger problems.

3. Embedded methods

Embedded methods are a catch-all group of techniques which perform feature selection as part of the classification process. Any features which have non-zero regression coefficients are 'selected' by the LASSO algorithm. One other popular approach is the Recursive Feature Elimination algorithm, commonly used with Support Vector Machine. These approaches tend to be between filters and wrappers in terms of computational complexity.

E. Gibbs Sampling

Gibbs sampling is a form of Markov chain Monte Carlo (MCMC) algorithm for approximating joint and marginal distribution by sampling from conditional distributions. This algorithm used for obtaining a sequence of observations which correlated with nearby samples. Gibbs sampling is used to generates the new minority class samples by using the joint probability distribution and interdependencies of data attributes. If the joint distribution is not known explicitly or is difficult to sample from directly, but the conditional distribution is known or easy to sample from. The goal of a Gibbs sampler is to create a Markov chain of random variables that converge to a target probability distribution. Gibbs Sampling algorithm could generate a sequence of samples from conditional individual distributions, which constitutes a Markov chain, to approximate the joint distribution.

Algorithm 1 : Gibbs Sampler

1: $Z^{(0)} = < z_1^{(0)}, \ldots\ldots, z_k^{(0)} >$

2: for t = 1 to T

3: for i = 1 to k

4: $z_i^{(t+1)} \sim P(Z_i \mid z_1^{(t+1)}, .., z_{i-1}^{(t+1)}, z_{i+1}^{(t)} \ldots., z_n^{(t)})$

The Gibbs sampler is a special case of Metropolis-Hastings sampling where the random value is always accepted. The initial value to the Gibbs sampler is considers as univariate conditional distributions in which

the distribution when all of the random variables but one are assigned fixed values. Such conditional distributions are far easier to simulate than complex joint distributions and usually have simple forms. The Gibbs sampler is the most straightforward way to sample from posterior distribution. Gibbs sampling is commonly used in statistical interference, such as Bayesian interference. Bayesian networks are specified as a collection of conditional distributions.

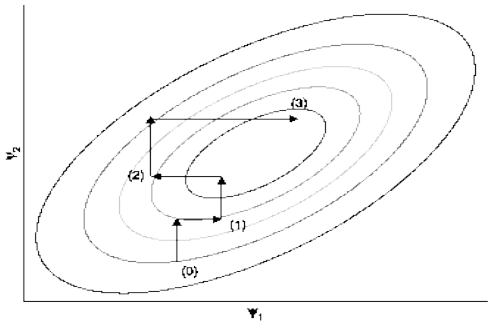

Figure (a): Gibbs sampling algorithm in two dimensions starting from an initial point and then completing three iterations

The approximation in the samples is independent, because they were produced by a process of the previous point in the chain to generate the next point. This is referred to as autocorrelation. The Gibbs sampling can simulate the target distribution by constructing a Gibbs sequence which converges to a stationary distribution that is independent of the starting value. Gibbs Sampling algorithm has been widely used on a broad class of areas, e.g. , Bayesian networks, statistical inference, bioinformatics, econometrics.

III. Conclusion:

This paper reviews all the aspects of handling class imbalance problem in real time dataset.

This paper studied the challenges of class imbalance problem and ensemble the different approaches to deal with class imbalance data.

It works on two specific levels, data level and algorithm level methods. At data level, the most common approach is sampling which is deal with imbalanced data. At the algorithmic level, it adjusts the costs of the various classes imbalance. SMOTE use sampling approach for balancing a data but it has a problem of data overfitting. The probabilistic approach is the new way of sampling method in which samples are generated based on joint probability class distribution, the results of these methods highly sensitive to class imbalance ratio and input dataset. The comparative study suggest that applying Rapidly Convergence on sampling methods will achieve higher balancing ratio on imbalance dataset which will further helps to improve the classification of imbalance dataset.

REFERENCES

[1] Juanli Hu, Jiabin Deng, Mingxiang Sui, "A New Approach For Decision Tree Based On Principal Component Analysis", Proceedings Of Conference On Computational Intelligence And Software Engineering, Page No:1-4, 2009.

[2] K.P.N.V.Satyasree, Dr. J. V. R. Murthy, "An Exhaustive Literature Review On Class Imbalance Problem" Ijettcs,Volume 2, Issue 3, May – June 2013

[3] N. V. Chawla, K.W. Bowyer, L. O. Hall, Andw. P. Kegelmeyer, "Smote synthetic Minority Over-Sampling Technique," J. Artif. Intell. Res., Vol. 16,Pp. 321–357, 2002.

[4] A Review On Ensembles For The Class Imbalance Problem: Bagging-, Boosting-, And Hybrid-Based Approaches, IEEE Transactions On Systems, Man, And Cybernetics

[5] R. LazaEt Al., "Evaluating The Effect Of Unbalanced Data In Biomedical Document Classification," Journal Of Integrative Bioinformatics, Vol. 8, No. 3, Pp. 177, 2011 Sep, 2011.

[6] C. Drummond, R. C. Holte. C "Decision Tree, Class Imbalance, And Cost Sensitivity: Why Under-Sampling Beats Over-Sampling, In: Workshop On Learning From Imbalanced Data Sets" Ii, International Conference On Machine Learning, 2003.

[7] Y. Freund and R. E. Schapire. "A Decision-Theoretic Generalization Of On-Line Learning And An Application To Boosting". Journal Of Computer And System Science, 55(1):119-139, 1997.

[8] N. V. Chawla, A. Lazarevic, L. O. Hall, And K. W. Bowyer. "Smoteboost: Improving Prediction Of The Minority Class In Boosting. In Knowledge Discovery In Databases" Pkdd 2003, Pp. 107–119, 2003.

[9] Dr.D.Ramyachitr, P.Manikandan "Imbalanced Dataset Classification And Solutions: A Review" International Journal Of Computing And Business Research (Ijcbr),Volume 5 Issue 4 July 2014

[10] Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano. "Rusboost: A hybrid approach to alleviating class imbalance", IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 2010, 40(1):185–197.

[11] Dr. Ali Mirza Mahmood "An Overview of Class Imbalance Learning in Knowledge Discovery" Associate Professor, DMS SVH College of Engineering, Machilipatnam. Krishna University, Machilipatnam, Andhra Pradesh, India.

[12] Barnan Das, Narayanan C. Krishnan And Diane J. Cook, Fellow,"RACOG AND WRACOG: Two Probabilistic Oversampling Techniques" IEEE Transaction On Knowledge And Data Engineering - Draft 1

[13] Yun Zhai, Haifeng Sui, Changsheng Zhang "A New Over-sample Method Based on Distribution Density" JOURNAL OF COMPUTERS, VOL. 9, NO. 2, FEBRUARY 2014

[14] Yoav Freund Robert E. Schapire "Experiments with a New Boosting Algorithm" Machine Learning: Proceedings of the Thirteenth International Conference, 1996.

[15] E. Ramentol1, n. Verbiest, r. Bello, y. Caballero1, c. Cornelis and f. Herrera "Smote-Frst: A New Resampling Method Using Fuzzy Rough Set Theory" March 20, 2012

[16] Haibo He, Member, Ieee, And Edwardo A. Garcia "Learning From Imbalanced Data" IEEE Transactions On Knowledge And Data Engineering, Vol. 21, No. 9, September 2009

[17] Sunita Beniwal*, Jitender Arora "Classification and Feature Selection Techniques in Data Mining" International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 6, August – 2012, ISSN: 2278-0181

