ISSN: 2321-7782 (Online)

Volume 2, Issue 2, February 2014

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Paper / Case Study

Available online at: www.ijarcsms.com

Classification Technique used to Handle Imbalanced Data

Priyanka U. Kekre[1]
Department of Computer Science & Engg.
G.H. Raisoni College of Engineering
Nagpur – India

Sonali U. Nimbhorkar[2]
Department of Computer Science & Engg.
G.H. Raisoni College of Engineering
Nagpur – India

Abstract: The data in real time applications is expanding at a faster rate. So it has become critical to extract knowledge from such huge and rapidly changing data. The problem arises when the target concept has significantly less number of observations; it is nothing but imbalanced problem. The imbalanced learning focuses on underrepresented data and rare class instances. Classification of data with such underrepresented data is complex and requires an iterative learning module. So accurate classification of imbalanced data requires learning. Thus classification algorithms built to classify such underrepresented data may tend to misclassify the minority class instances. So accurate classification needs refinement of the class boundaries. This paper presents a comparative study of the classifiers and proposed work related to the system which would eliminate the redundant and irrelevant attributes and accurately classify minority instances.

Keywords: classification, data mining, imbalanced learning, minority class, rare class concept.

I. Introduction

Technology is changing at faster rate thus enabling the growth of such huge data. Many real time applications need to handle and analyze such huge data. As the data is growing at a faster rate analyzing such data has become difficult. Many applications produce tons of data every hour; all transactions performed may not be equally important for analysis. So the data needs to be quickly analyzed and classified accordingly.

Many real life applications face imbalance problem and to deal with such fast moving data is a severe problem. Data is said to be imbalanced if there exist an unequal distribution between its classes, to analyze and learn from such data is a challenge. Classifying and learning iteratively is one of the techniques to deal with such data. Thus the training instances acquired during the process forms a model for classifying unseen examples. The dataset is said to be imbalanced if even one class is represented by comparatively less number of instances than the others.

In many real world applications often majority class instances i.e. normal examples are more in number whereas the examples of interest are less in number thus forming the minority class. Another reason for class imbalance problem is its difficulty in collecting instances of some classes. Despite that they are difficult to identify, rare instances generally constitute the target concept in classification tasks. Classification techniques [1],[2], are generally more important to correctly classify such minority class instances.

Some of the real time applications facing imbalance class problem are: identifying fraudulent transactions for e-banking or credit cards, irregularities in accounting data [15], text categorization and network traffic and many more. In real-world applications mis-predicting a rare event can result in serious consequences.

This paper presents a study of various classifications techniques used to handle imbalance data and also presents preliminary work of an effective model which accurately classifies the minority class instances.

II. Related Work

Recent research on Imbalanced class has focused on several groups of techniques. Sampling methods in imbalanced learning applications modify the imbalanced data set by some mechanisms in order to provide a balanced distribution. Sampling methods can be used to balance the dataset either by over-sampling or under-sampling the classes. The over-sampling method appends data to the original data set, whereas under-sampling [2] removes data from the original data set. Over-sampling adds data to the minority class whereas under-sampling removes data from the majority class so as balance the classes. In the case of under-sampling [2], removing examples from the majority class may cause the classifier to miss important concepts present in the majority class. Chavla et al [3] proposed an approach called SMOTE (Synthetic minority oversampling technique) which creates synthetic examples using the minority instances in current training chunk, which in turn are used to balance the training chunk.

Another sampling approach proposed by Gao et al [4] which deals with imbalanced data is SE (Sampling + Ensemble), it differs from the above approach as it processes the stream in batches. SE approach over-samples the positive instances by incorporating the old positive examples along with under-sampling by using disjoint subsets of negative examples.

SERA (Selectively Recursive Approach) an approach proposed by Chen and He [5],in this approach minority examples from previous chunks are selectively absorbed into current training chunk to balance it. Similarity measure used to select minority examples from previous chunks was mahalanobis distance. Thus different from approach used in SE which uses take in all approach.

MuSeRA (Multiple Selectively Recursive Approach) was further work of Chen and He [6] after SERA. In MuSeRA balancing of training chunk is done using mahalanobis distance as similarity measure. It maintains the minority samples accumulated from all the previous training chunks.

REA (Recursive Ensemble Approach) another approach proposed by Chen and He [7] which uses nearest neighbor technique to balance the positive instances. In order to balance the training chunk it adds the positive instances from previous chunks which are nearest neighbors of the positive instances in the current training chunk.

Zhang et al [8] proposed an algorithm to deal with skewed data streams which used clustering+sampling technique. In this approach sampling was carried out by k-means algorithm to form clusters of negative examples in the current training chunk and then the centroid of each of the clusters was calculated. These centroids represented each of the clusters formed. Numbers of clusters formed were equal to the number of positive examples in current training batch. Thus current training batch is updated by taking all positive examples along with centroid of the clusters of negative samples.

Boosting algorithms employs weighting strategy to increase the weights of misclassified instances and decrease the weights of correctly classified ones. Boosting algorithms divide the instances of the imbalanced dataset into two groups: correctly classified instances and misclassified instances. SkewBoost [9] algorithm comprises of two sections, first is a variation of SMOTE and the second is manipulating the weights of the instances in a cost-sensitive approach after every iteration. SMOTE [3] creates synthetic minority class instances. The second part implements boosting algorithm to improve identification of misclassified instances from the earlier iterations.

A decision tree learning algorithm VFDT [10], sub-samples the entire data stream. The sub-sample size is calculated using distribution-free bounds called Hoeffding bounds under the assumption that the data is generated by a stationary distribution.

Nearest Neighbor [12] Classifiers is also called instance-based learners or lazy learners. It provides an accurate way of learning data incrementally. Each processed example is stored and serves as a reference for new data points. Classification is based on the labels of the nearest historical examples.

Streaming Ensemble Algorithm [13] processes the incoming stream in data chunks. The size of those chunks is an important parameter because it is responsible for the trade-off between accuracy and flexibility. Each data chunk is used to train a new classifier. It is then compared with ensemble members. If any ensemble member is “weaker” than that candidate classifier it is dropped and the new classifier takes its place.

The CVFDT [14] algorithm was designed to handle concept drifts. This algorithm is an extension to VFDT; it mines high-speed data streams under the approach of one-pass mining. The one pass mining approach is unable to identify the changes occurred in the model during the data arrival process.

III. Discussion

Sampling methods usually are used for handling imbalanced data sets involve either under-sampling or over-sampling of the original data sets. Every sampling method has its own disadvantages. The major disadvantage of Random undersampling is that it might remove some of prominent instances from the data set. While performing under-sampling process supposes it leaves only the outliers then it would lead to misclassification of the instances which are more closely grouped. Random oversampling creates random instances for minority class. Thus the created attribute values may be drastically different from the values of the original instances. If so it will cause the classifier to get trained in a different way and lead to misclassification of some of the instances. Most of the above mentioned sampling algorithms either under-sample majority or over-sample minority the dataset. Using these methods prediction accuracy of minority classes can be improved. But these methods fail to work efficiently with stream applications because of their infinite data flow. SMOTE [3], SERA [5], MUSERA [6], REA [7] are over-sampling based approaches which in some way over-sample minority class instances to balance the training chunk. SE [4] uses under-sampling as well as over sampling.

Boosting algorithms intentionally increase the weights of examples with higher misclassification. Since misclassified data is assigned higher weights it would cause the classifier to focus on these examples. SkewBoost [9] algorithm creates synthetic values closer to the original instance not like SMOTE which randomly over samples the minority class. It assigns weights to the instances based on the ratio of the class to the entire dataset. The algorithm tries to refine the boundaries for the skewed class dataset, by adding new minority class examples obtained during the learning process. It has not focused on data stream applications and multiple minority class problems.

VFDT [10] is Single Classifier decision tree algorithm. It helps to overcome the long training times issue. It constructs a decision tree by using constant memory and constant time per sample. This algorithm is the combination of decision-tree learning method and sub-sampling of given data stream. This algorithm is unable to cope with concept drifts.

CVFDT [14] Concept Adapting very Fast Decision Tree handles concept drift issue by growing alternate trees and subtrees. This algorithm uses one-pass mining approach for mining high-speed data streams. The algorithm takes more space and the accuracy of this model is not better than the best sliding window model.

IV. Proposed plan of work

From above discussion it has been observed that there is a need of quick and efficient classifier to discover and accurately classify the instances. This section gives a design flow and description of preliminary model for imbalanced data classification.

The proposed model aims to remove the problems related to inefficiency in accurately classifying the data instances. The model aims not only to accurately classify the data but also learn iteratively so as to build predictive model. The Figure1. show the work flow of the proposed system.

The steps that would be followed to obtain accurately classified data are:

1. The system evaluates the dataset properties.
2. The redundant and irrelevant attributes are eliminated.
3. Minority class is discovered and iteratively the boundaries are refined to obtain accurately classified instances.

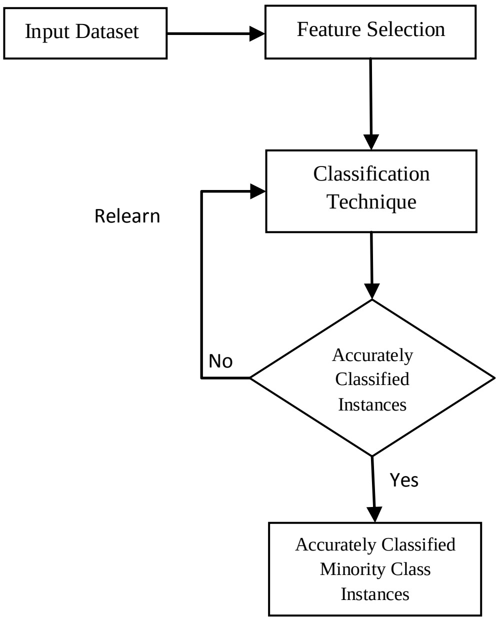
Input Dataset
Feature Selection
Classification Technique
Relearn
No
Accurately Classified Instances
Yes
Accurately Classified Minority Class Instances

Figure 1. Work Flow of proposed plan

Firstly the dataset properties are evaluated. The second step comprises of eliminating the attributes by checking their relevance value. Thus we can say that irrelevant and duplicate columns are eliminated. This is done to improve the accuracy for classifying the instances. The last step identifies new classes and an iterative approach is used to refine the boundaries of already identified classes. So that the instances are accurately classified.

The system will implement a strong classifier switching algorithm. Uncertainty or rare category problem will be addressed by this classifier switching algorithm & will ensure that the best classifier is selected for a given data set. The output obtained is improved every time the dataset is taken as input by relearning.

V. CONCLUSION

Each method discussed above handles imbalanced data in a different manner. Due to imbalance data it is difficult to predict whether the instance belongs to normal class or rare class. For this reason a suitable predicting techniques for classification should be chosen. A strong classification algorithm needs to be chosen to classify the rare instances accurately.

The proposed model introduced above will try to analyze the data and accurately classify the instances. The proposed model is designed to handle and classify high dimensional data. The analysis includes elimination of irrelevant and duplicate data so as to improve the accuracy to further extent. Learning will help to classify the instances more accurately and refine the boundaries of the classes. It will calculate the parameters leading to the correct classification of instances and check the accuracy of the output generated.

References

1. T. M. Hospedales, S. Gong, and Tao Xiang "Finding Rare Classes: Active Learning with Generative and Discriminative Models" IEEE Trans. Data and Knowledge Eng., vol. 25, no. 2, pages 374-386, Feb.2013
2. H. He and E. Garcia, "Learning from Imbalanced Data" IEEE Trans. Data and Knowledge Eng., vol. 21, no. 9, pages 1263-1284, Sept.2009.
3. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. "Smote: synthetic minority over-sampling technique." J. Artif. Int. Res., 16:321-357, June 2002.
4. J. Gao, B. Ding, W. Fan, J. Han, and P. S. Yu. "Classifying data streams with skewed class distributions and concept drifts." IEEE Internet Computing, pages 37-49, Nov. 2008.
5. S. Chen and H. He. "Sera: Selectively recursive approach towards non stationary imbalanced stream data mining" In International Joint Conference on Neural Network (IJCNN), pages 522-529, June 2009.
6. S. Chen, H. He, K. Li, and S. Desai. "Musera: Multiple selectively recursive approach towards imbalanced stream data mining" In Neural Networks (IJCNN), International Joint Conference on, pages1-8, July 2010.

7. S. Chen and H. He. "Towards incremental learning of non stationary imbalanced data stream: a multiple selectively recursive approach" Evolving Systems, pages 1-16, 2011.
8. Y. Wang, Y. Zhang, and Y. Wang "Mining data streams with skewed distribution by static classifier ensemble". In B.-C. Chien and T.-P. Hong, editors, Opportunities and Challenges for Next-Generation Applied Intelligence, volume 214, Springer, pages 65-71, 2009.
9. S. Hukerikar, A. Tumma, A. Nikam, V. Attar "SkewBoost: An Algorithm for Classifying Imbalanced Datasets" IEEE International Conference on Computer & Communication Technology (ICCCT), pages 46-52, 2011.
10. Pedro Domingos, Geoff Hulten, "Mining High Speed Data Streams" KDD-00 in proceeding of sixth ACM SIGKDD international conference on knowledge discovery and data mining, pages 71-80, 2000.
11. A. Godase, V. Attar "Classifier Ensemble for Imbalanced Data Stream Classification", ACM, pages 3-5, Sept.2012.
12. T. Darrell and P. Indyk and G. Shakhnarovich "Nearest Neighbor Methods in Learning and Vision: Theory and Practice." MIT Press, 2006.
13. W .Nick Street, Yong Seog Kim "A streaming ensemble algorithm (sea) for large-scale classification", In Proceedings of the seventh ACM SIGKDD International Conference on Knowledge discovery and data mining, New York, NY, USA, pages 377-382, 2001.
14. Bifet A, Gavald R "Learning from time changing data with adaptive windowing" in SIAM International Conference on Data Mining, pages 443-448,2007.
15. S. Bay, K. Kumaraswamy, M. G. Anderle, R. Kumar, and D. M. Steier. "Large scale detection of irregularities in accounting data." In Proceedings of the Sixth International Conference on Data Mining (ICDM) '06 IEEE Computer Society, pages 75-86, 2006.
16. G. Ditzler, R. Polikar, and N. Chawla. "An incremental learning algorithm for non-stationary environments and class imbalance." 20th International Conference on pattern recognition, pages 2997-3000, Aug. 2010.
17. S. Wang, L. L. Minku, D. Ghezzi, D. Caltabiano, P. Tino and X. Yao "Concept Drift Detection for Online Class Imbalance Learning" IJCNN, pp1-10, Jan. 2013.

AUTHOR(S) PROFILE

Priyanka Kekre received the Bachelors degree in Computer Technology from Priyadarshini College of Engineering, Nagpur in 2007. She has 2 years teaching experience from Priyadarshini College of Engineering and Priyadarshini Indira Gandhi College of Engineering, Nagpur. She also has an industrial experience of 2.3 years from Mahindra Satyam. Her main area of interest includes Data Mining. She is now pursuing Masters in Technology in Computer science and Engineering from Raisoni College of Engineering, Nagpur.

Prof. Sonali U. Nimbhorkar is currently working as professor in G.H. Raisoni College of Engineering. Currently she is pursuing PhD from G.H. Raisoni College of Engineering. She has completed Masters in Technology in Computer science and Engineering from Raisoni College of Engineering, Nagpur. She has many national and international publications to her credit.