A method for solving the class imbalance Problem in Classification Techniques

Imran Alam[1] Manohar Kumar Kushwaha[2] Vinay Kumar Singh[3]
[1, 2, 3] M.Tech (CS&E)
[1, 2, 3] School of Computing Science and Engineering Department
[1, 2, 3] Galgotias University, Greater Noida, U.P, India.

Abstract— Class imbalance learning refers to learning from imbalanced data sets, in which some classes of examples (minority) are highly under-represented comparing to other classes (majority). The Very skewed class distribution degrades the learning ability of many traditional machine learning methods, especially in the recognition of examples from the minority classes, which are often deemed to be more important and interesting. Although quite a few ensemble learning approaches have been proposed to handle the problem, no in-depth research exists to explain why and when they can be helpful. We investigate mathematical links between single-class performance and ensemble diversity. One method to tackle this problem consists to resample the original training set, either by over-sampling the minority class and/or under-sampling the majority class. we propose two ensemble models (using a modular neural network and the nearest neighbor rule) trained on datasets under-sampled with genetic algorithms. Experiments with real datasets demonstrate the effectiveness of the methodology here proposed.

Keywords: Classification Techniques, Genetic Algorithm, Imbalance, Data Mining.

I. INTRODUCTION

The class imbalance problem has received considerable attention in areas such as Machine Learning and Pattern Recognition. A two-class dataset is said to be imbalanced when one of the classes (the minority one) is heavily under-represented in comparison to the other class (the majority one) .The resulting model (classier) will Enable us to predict the outcome for new unseen examples. We describe the basic classification techniques. Several major kinds of classification method including Decision tree induction, Bayesian networks, K-nearest neighbor classifier, Case-based reasoning. The goal of this report is to provide a comprehensive review of different classification techniques in data mining. Classification Data Mining (DM) Techniques can be a very useful tool in detecting and identifying e-banking phishing websites. Phishing websites is a semantic attack which targets the user rather than the computer. We present a Class imbalance problem has received considerable attention in areas such as Machine Learning and Pattern Recognition.

II. CLASSIFICATION TECHNIQUES

A. *DECISION TREE INDUCTION*

Decision trees are trees that classify instances by sorting them based on feature values. Each node in a decision tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume. Instances are classified starting at the root node and sorted based on their feature values. The same procedure is then repeated on each partition of the divided data, creating sub-trees until the training data is divided into subsets of the same class. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.

B. *Bayesian Networks*

A Bayesian network, Bayesian networks are directed acyclic graphs (DAG) whose nodes represent random variables in the Bayesian sense: they may be observable quantities, latent variables, unknown parameters or hypotheses. Edges represent conditional dependencies; nodes which are not connected represent variables which are conditionally independent of each other. Each node is associated with a probability function that takes as input a particular set of values for the node's parent variables and gives the probability of the variable represented by the node.

C. *K-Nearest Neighbor Classifiers*

Nearest neighbor classifiers are based on learning by analogy. The training samples are described by n dimensional numeric attributes. Each sample represents a point in an n-dimensional space. In this way, all of the training samples are stored in an n-dimensional pattern space.

D. *Neural Networks*

In more practical terms neural networks are non-linear statistical data modeling tools. They can be used to model complex relationships between inputs and outputs or to find patterns in Data. Using neural networks as a tool, data warehousing firms are harvesting information from datasets in the process known as data mining. The difference between these data warehouses and an ordinary database is that there is actual manipulation and cross-fertilization of the data helping users makes more informed decisions.

III. CLASS IMBALANCE PROBLEMS

A. *The class imbalance problem in pattern classification*

In recent years, the class imbalance problem has received considerable attention in areas such as Machine Learning and Pattern Recognition. A two-class data set is said to be imbalanced (or skewed) when one of the classes (the minority one) is heavily under-represented in comparison to the other class (the majority one). This issue is particularly important in real world applications where it is costly to misclassify examples from the minority class, such as diagnosis of rare diseases, detection of fraudulent telephone calls, text categorization, information retrieval and filtering tasks. Traditionally, research on this topic has mainly focused on a number of solutions both at the data and algorithmic levels. However, there have recently appeared

other research lines within the general framework of class imbalance.

B. Resampling techniques

Data level methods for balancing the classes consists of resampling the original data set, either by over-sampling the minority class or by under-sampling and/or under-sampling the majority class, until the classes are approximately equally represented. However, both strategies have shown important drawbacks. Under-sampling may throw out potentially useful data, while over-sampling artificially increases the size of the data set and consequently, worsens the computational burden of the learning algorithm.

C. Two Type of Resampling Techniques

(1) Over-Sampling:- Random Oversampling methods also help to achieve balance class distribution by replication minority class sample.
(2) Under-Sampling:- The most important method in under-sampling is random under-sampling method which trying to balance the distribution of class by randomly removing majority class sample.
(3) Performance Measures:-Most of performance measures for two-class problems are built over a 2 × 2 confusion matrix. From this, four simple measures can be directly obtained:

	Positive prediction	Negative prediction
Positive class	True Positive (TP)	False Negative (FN)
Negative class	False Positive (FP)	True Negative (TN)

Table.1: Confusion matrix N0: -1

A classifier produces four types of examples on testing data,

- TP: the number of correctly classified examples belonging to the positive class.
- TN: the number of correctly classified examples belonging to the negative class.
- FP: the number of misclassified examples belonging to the negative class.
- FN: the number of misclassified examples belonging to the positive class.

A confusion matrix contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. The following table shows the confusion matrix for a two class classifier.

TP and TN denote the number of positive and negative cases correctly classified, while FP and FN refer to the number of misclassified positive and negative examples, respectively. The most widely used metrics for measuring the performance of learning systems are the error rate and the accuracy, defined as

Error = (FP +FN)/ (TP +FN+TN+FP) and

Accuracy = (TP + TN)/ (TP + FN + TN + FP), respectively.

Geometric Mean: - $G = \sqrt{a+ \cdot a-}$

Where a+ is the accuracy on cases from the minority class

$$a\wedge + = TP/ TP+FN$$

And a- is the accuracy on cases from the majority class

$$a\wedge - = TN/TN+FP$$

This measure tries to maximize the accuracy on each of the two classes while keeping these accuracies balanced.

IV. Challenges and Potential Solutions for Multi-Class Imbalance Problems

As an important extension to two-class cases, multi-class imbalance problems pose new challenges that have not drawn much attention. The infectiveness of two-class imbalance learning techniques caused by multi-class has been reported. In chapter 6, we aimed to nod out what problems multi-class can cause and how it acts the classification performance in the presence of imbalanced data. Two types of multi-class imbalance problems, i.e. the multi-minority and multi-majority cases, were studied by applying random oversampling and under sampling techniques. Both types showed strong negative correlations with the performance measures, which implied that the performance decreases as the number of imbalanced classes increases. The multi-majority case was shown to be more harmful than the multi-minority case, because the imbalance rate became more severe. Regarding the class imbalance learning techniques, oversampling did not help the classification and caused overstating; under sampling was sensitive to the number of minority classes and surged from great performance loss on majority classes. This is the rest systematic study of multi-class for class imbalance learning by providing separate and in-depth discussions of multi-minority and multi-majority cases. The results reveal possible issues that a class imbalance learning technique could confront when dealing with multi-class tasks, and provide guidance for designing better solutions.

V. Genetic Algorithms

The most basic structure of the GA proposed by Holland, begins with a set of possible solutions (population) codified as a chain of bits (called chromosome), later with the use of a method to evaluate the behavior (fitness) of each chromosome, the parents of the next population are determined.

In this work we modify the GA proposed by Diaz et al., to reduce the processing time of the GA, in addition to the 0′s, some chromosomes are reduced in 20%, that is to say, during the evaluative process, several genes marked with a different value of 0 or 1 were ignored. The leaving-one-out method was used as fitness method and, an elitist method select the best solutions in each step and uses these chromosomes to apply the genetic operators: crossover and mutation. The former, consists of the uniform crossover and, next, randomly change 10% of the genes in each chromosome.

VI. MIXTURES OF EXPERTS

A Mixture of Experts (ME) or modular network solves a complex computational task by dividing it into a number of simpler subtasks and then combining their individual solutions. Thus, a ME consists of several expert neural networks (modules), where each expert is optimized to perform a particular task of an overall complex operation. An integrating unit, called gating network, is used to select or combine the outputs of the modules (expert networks) in order to form the final output of the modular network. In the more basic implementation of these networks, all the modules are of a same type, but different schemes could be also used.

VII. EXPERIMENTAL RESULTS

This section expose the experimental results obtained with two ensemble models: using mixture of experts and using the NN rule, both of them trained on under sampled subsamples by a GA. The section was dividing in three parts. The first one, describe the method used for transform the datasets in a problem with two classes. The second part exposes the evaluation criterion for the imbalance problem here used. Finally, the experimental results are shown in the third part.

Dataset	Positive samples	Negative samples	Majority class
Cancer	191	355	1
Pima	268	500	1
Glass	17	197	1,2,3,4, 5,6,8, 9
German	300	700	1
Phoneme	1586	3818	1
Vehicle	212	634	2,3,4
Sat image	629	5809	1,2,3,5,6

Table. 2: Description of the data sets

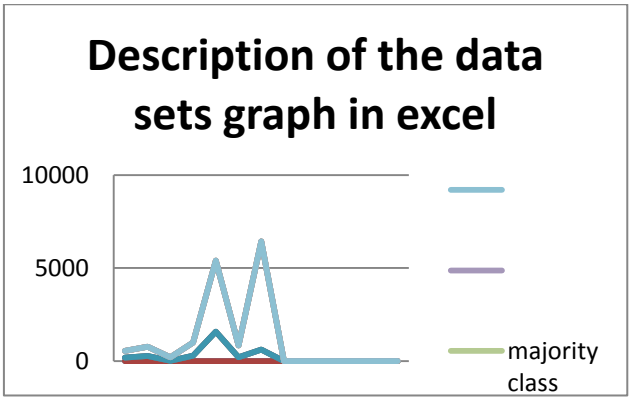

Fig. 1: Description of the data sets graph in excel

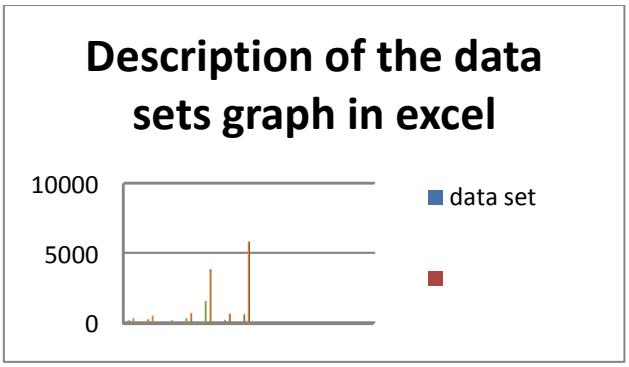

Fig. 2: Description of the data sets graph in excel

VIII. CONCLUSION

There are we proposed the sampling method for classification problems. We can generate a method for remove the classification problems. Then we can get a good performance in future. In many real-world applications, supervised pattern recognition methods have to cope with imbalanced TSs. we propose a new methodology focused on the solution methods approach, which combines an under-sampling method using a GA and an ensemble trained with the solutions given by the GA.

REFERENCES

[1] G.Kesavaraj,Dr.S.Sukumaran,"A Study on Classification Techniques in Data Mining" july 4-6,2013.
[2]]Thair Nu Phyu" Survey of classification Techniques in Data Mining". 2009.
[3] Shweta Joshi " Categorizing the Document using Multi class classification in Data Mining". 2011.
[4] V .Garacia and J. S Sancher" The class imbalance problem in pattern classification and learing.
[5] Mr. Rushilongade, Ms. Snehlata s. Dongre" Class imbalance problem in data mining:Review"(2013).
[6] Laura cleofas , Rasa Valdovinos " use of Ensemble Based on GA For Imbalance problem".(2009).
[7] Sundar.C, M.Chitradevi and Dr.G.Geetharamani ―Classification of Cardiotocogram Data using Neural Network based Machine Learning Technique‖ International Journal of Computer Applications (0975 – 888) Volume 47– No.14, June 2012.
[8] K Priya, R. GeethaRamani and ShomonaGracia Jacob ―Data Mining Techniques for Automatic recognition of Carnatic Raga Swaram notes‖ International Journal of Computer Applications (0975 – 8887) Volume 52– No.10, August 2012.
[9] Smitha .T, V. Sundaram ―Comparative Study Of Data Mining Algorithms For High Dimensional Data Analysis‖ International Journal Of Advances In Engineering &Technology, Sept 2012.IJAET ISSN: 2231-1963.
[10] Neural Networks for Data Mining. http://www.google.com./(16-042014)(21,24,&29May2014).
[11] Barandela, R., Sánchez, J.S., García, V., Rangel, E.: Strategies for Learning in Class Imbalance Problems. Pattern Recognition 36, 849–851 (2003).

[12] Sundar.C, M.Chitradevi and Dr.G.Geetharamani ―Classification of CardiotocogramData using Neural Network based Machine Learning Technique‖ International Journal of Computer Applications (0975 – 888) Volume 47–No.14, June 2012.
[13] Barandela, R., Sánchez, J.S., García, V., Rangel, E.: Strategies for Learning in Class Imbalance Problems. Pattern Recognition 36, 849–851 (2003)
[14] Woods, K., Doss, C., Bowyer, K.W., Solk, J., Priebe, C., Kegelmeyer, W.P.: ComparativeEvaluation of Pattern Recognition Techniques for Detection of Microcalcifications in Mammography. International Journal of Pattern Recognition and Artificial Intelligence 7, 1417–1436 (1993)
[15] Fawcett, T., Provost, F.: Adaptive Fraud Detection. Data Mining and Knowledge Discovery 1, 291–316 (1996)

