ISSN No. 0976-5697

Volume 7, No. 6(Special Issue), November 2016

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Utilization of Natural Computing Methods for the Classification of Data Streams with Skewed Distribution

Satveer Sharma
(M.Tech Research Scholar)
Computer Engineering
Yadavindra College of Engineering and Technology
Talwandi Sabo
satveersharma07@gmail.com

Er.Meenakshi Bansal
(Assistant Professor)
Computer Engineering
Yadavindra College of Engineering and Technology
Talwandi Sabo
ermeenu10@gmail.com

Abstract- *In last few years there are major changes and evolution has been done on classification of data. Classification of data becomes difficult because of unbounded size and imbalance nature of data. Class imbalance problem become greatest issue in data mining. Imbalance problem occur where one of the two classes having more sample than other classes. This paper has reviewed various natural computing techniques for solving the problem of classification basically based on data streams and skewed method.*

Keywords *- Natural Computing, Swarm Optimization Algorithms, Classifications.*

I. INTRODUCTION

In many real time applications large amount of data is generated with skewed distribution. A data set said to be highly skewed if sample from one class is in higher number than other [1]. In imbalance data set the class having more number of instances is called as major class while the one having relatively less number of instances are called as minor class [2]. Applications such as medical diagnosis prediction of rare but important disease is very important than regular treatment. Similar situations are observed in other areas, such as detecting fraud in banking operations, detecting network intrusions [3], managing risk and predicting failures of technical equipment. In such situation most of the classifier are biased towards the major classes and hence show very poor classification rates on minor classes. It is also possible that classifier predicts everything as major class and ignores the minor class. Various techniques have been proposed to solve the problems associated with class imbalance [4]. But this paper will review various natural computing methods.

II. CLASSIFICATION PROBLEM

Many real applications, such as network traffic monitoring, credit-card fraud detection, and Web click streams, generate continuously arriving data known as data streams. In general, knowledge discovery from stream data is challenging; the data are usually massive and arrive with high speed, making either storing all the historical data or scanning it nearly impossible. Moreover, stream data often evolve considerably over time. In many applications, the response time usually must be short. Data chunks *C*1, *C*2, …,*Ci* arrive one by one, and each chunk contains positive instances *Pi* and negative instances *Qi*. Suppose *C*1, *C*2,…, *Cm* are labeled. At time stamp *m* +1, when an unlabeled data chunk *Cm*+1 arrives, the classification model predicts the labels ofinstances in *Cm*+1 on the basis of previously labeleddata chunks. When experts give the trueclass labels of instances in *Cm*+1, the chunk canjoin the training set, resulting in more andmore labeled data chunks. Because of the storageconstraints, it's critical to wisely select labeledexamples that can represent the currentdistribution well.

III. RELATED WORK

In an imbalance problem the large amount of data is generated that is skewed. A data is said to be imbalanced if one sample is high than other sample. Generally the imbalance methods are divided into three categories: probabilistic method, learning method and classification method.

Seiffert et al. (2010) presented a new hybrid sampling/boosting algorithm, called RUSBoost, for learning from skewed training data. This algorithm provides a simpler and faster alternative to SMOTEBoost, which is another algorithm that combines boosting and data sampling.

Wang, S., Yao, X.,(2012) studied the challenges posed by the multiclass imbalance problems and investigates the generalization ability of some ensemble solutions, including their recently proposed algorithm AdaBoost.NC , with the aim of handling multiclass and imbalance effectively and directly.

Longadge, R., Dongre, S. S., and Malik, L., (2013) described a method to solve the class imbalance problem by multi cluster-based majority under-sampling and random minority oversampling approach. Compared to under-sampling, cluster-based random under-sampling can

978-93-85670-72-5 © 2016 (RTCSIT)

CONFERENCE PAPER
International Conference on
Recent Trends in Computer Science & Information Technology (RTCSIT-2016)
21st August 2016
Guru Nanak College Budhlada, Punjab India

effectively avoid the important information loss of majority class and oversampling will helps to balance data i.e overcomes the drawback of under-sampling that it removes away many useful majority class samples.

Rashu,R.I., Naheena Haq, Rashedur M Rahman (2014) suggested data mining approaches that have been used in business purposes since its inception however, at present it is used successfully in new and emerging areas like education systems. In this paper, use of data mining approaches to predict students' final outcome, i.e., final grade in a particular course by overcoming the problem of imbalanced dataset. Several re-sampling techniques are given to balance the dataset so that to get better performance. Re-sampling techniques include SMOTE, RUS, ROS.

Kwak, J., Lee, T., and Kim, C.O.,(2015) proposed when the class sizes are highly imbalanced, the standard algorithm tend to strongly favor the majority class and provide notably low detection of the minority class as a result. The method proposes an online fault detection algorithm based on incremental clustering. The algorithm accurately finds wafer faults even in severe class distribution skews and efficiently processes massive sensor data in terms of reductions in the required storage.

Sr.No.	Algorithm	Advantages	Disadvantages	Parameters
1.	AdaBoost.NC (2010)	Improve prediction accuracy of minority	Ignore overall performance of classifier	Good accuracy
2.	RUSBoost (2012)	Simple, faster and less complex than SMOTE Boost algorithm	Unable to solve Multiclass imbalance problem	High classification rate
3.	Multi-Cluster Based Approach (2013)	Overcomes the drawback of under-sampling that it removes away many useful majority class samples.	Avoid the important information loss.	Good sampling rate
4.	Data Mining Approaches (2014)	balance the dataset so that to get better performance	Not too good accuracy.	Medium accuracy
5.	Clustering Based Algorithm (2015)	The algorithm accurately finds wafer faults	Time complexity more.	High rate

Table 1: Comparative Study

Many areas are affected by class imbalance problems. The solution provided by many techniques in data mining is helpful but not enough. The consideration of which technique is best for handling a problem of data distribution is highly depends upon the nature of data used for experiment.

IV. NATURAL COMPUTING BASED ALGORITHMS

Natural Computing is the field of research that deals with computational techniques that deal with natural inspiration. It is a highly interdisciplinary field that connects natural science with computing science [5].

A. Artificial Neural Networks (ANN)

The simplest type of Neural Network is ANN (Artificial Neural Network). ANNs are designed closely to the neural structure of the brain. Neural network mainly consists of the layers. Layers are made up of nodes. There are mainly three types of layers in the neural architecture: Input layer, hidden layer and output layer. On the basis of the output layer weights are assigned to get output accordingly to input layer. There are many types of the learning rules in neural network but delta rule is common rule that has been used these days [6].

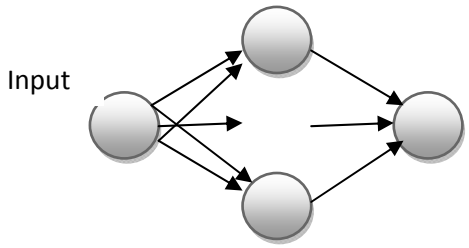

Fig.1: Neural Network Model

B. Evolutionary Algorithms

It comprise a constant or variable size population of persons, a fitness criterion, and hereditarily enthused operators that create the next generationfrom the current one. The first population is characteristically generated randomly or heuristically, and typical operator is mutation and recombination [7]. At each step, the individuals are evaluated according to the given fitness function. The next production is obtained from selected individuals by using genetically stimulated operators. This process of replicated evolution eventually converges towards nearly optimal inhabitants of individuals, from the point of view of the fitness function [8].

978-93-85670-72-5 © 2016 (RTCSIT)

a. Genetic Algorithm (GA)

GAs were first described by John Holland in the 1960s and further developed by Holland and his students and colleagues at the University of Michigan in the 1960s and 1970s. Genetic algorithms (GAs) are computer programs that take off the processes of biological growth in order to explain problems and to make evolutionary systems. The simple GA works as follows [9]:

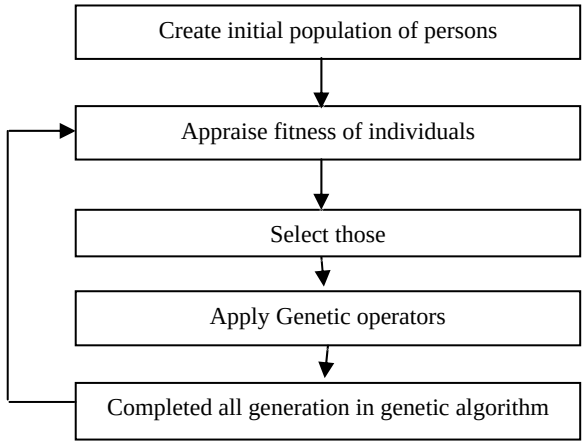

Fig.2: Genetic Algorithm Flowchart

b. Bacterial Foraging Optimization (BFO)

BFO algorithm is first projected by Passino in 2002. It is motivated by the foraging and Chemo tactic behaviors of bacteria, especially the Escherichia coli (E. coli). Locomotion can be achieved during the process of real bacteria forging through the tensile flagella set. Flagella help an E.coli bacterium to fall or swim, that are two essential operations performed by a bacterium at the instance of foraging [10].

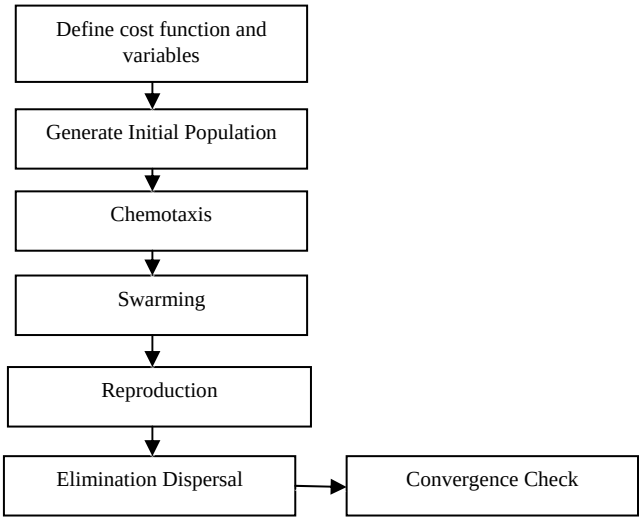

Fig.3: Bacterial Foraging Optimization Flowchart

C. Swarm Optimization

Swarm intelligence, sometimes referred to as collective intelligence, is defined as the difficulty solving performance that emerge from the communication of individual agents. Particle swarm optimization applies this idea to the problem of finding an optimal solution to a given problem by a search through a solution space. The initial set up is a swarm of particles, each representing a possible solution to the problem. [11, 12].

a. Ant Colony Optimization

ACO algorithm is an original intellectual optimization algorithm imply the winged animal swarm practices, which was planned by analyst Kennedy and Dr. Beernaert in 1995. In ACO answer, every individual is called "Burrowing little creature COLONY", which speaks to a potential arrangement. The calculation attains to the best arrangement with the variability of a few particles in the following space. ANT COLONYs seek in the arrangement space captivating after the best ANT COLONY by altering their positions and the wellness oftentimes, the airborne course and speed are controlled by the object function.

V. CONCLUSION AND FUTURE SCOPE

The most of algorithm are more focusing on classification of major sample while ignoring or misclassifying minority sample. The minority samples are those that rarely occur but very important. There are different methods available for classification of imbalance data set which is divided into three main categories, the algorithmic approach, data preprocessing approach and feature selection approach. Each of this technique has their own advantages and disadvantages. In this paper systematic study of each approach is define which gives the right direction for research in class imbalance problem.

REFERENCES

[1] Wang, S., Yao, X., "Multiclass Imbalance Problems: Analysis and Potential Solutions" (2012), IEEE Transactions On Systems, Man, And Cybernetics—Part B: Cybernetics, 42(4).

[2] Seiffert,C., Khoshgoftaar,T.M.,Hulse,J. V., and Napolitano,A., "RUSBoost: A Hybrid Approach to Alleviating Class Imbalance" (2010), IEEE Transactions On Systems, Man, And Cybernetics—Part A: Systems And Humans, 40(1).

[3] Waske,B., Linden,S.D., Benediktsson,J.A., Rabe,A., and Hostert,P., "Sensitivity of Support Vector Machines to Random Feature Selection in Classification of Hyper-

978-93-85670-72-5 © 2016 (RTCSIT)

spectral Data"(2010), IEEE Transactions On Geosciences And Remote Sensing, 48(7).

[4] Guo, X., Yin1,Y., Dong,C., Yang,G., Zhou,G.,. ,"On the Class Imbalance Problem" (2008), Fourth International Conference on Natural Computation,.

[5] Wasikowski, M., and Chen, X. W., "Combating the Small Sample Class Imbalance Problem Using Feature Selection" (2010), IEEE Transactions on Knowledge and Data Engineering, 22(10).

[6] de Castro LN, Von Zuben FJ. Recent developments in biologically inspired computing. Idea Group Publishing; 2004.

[7] Dorigo, M., Blum C., "Ant colony optimization theory: A survey" (2005), Theoretical Computer Science, 344(2–3), 243–78.

[8] Engelbrecht. Fundamentals of computational swarm intelligence. John Wiley & Sons; 2006.

[9] Flake GW. The Computational beauty of nature. MIT Press; 2000.

[10] Freitas AA, Rozenberg G. Data mining and knowledge discovery with evolutionary algorithms. Springer; 2002.

[11] Engelbrecht. (Ed) Fundamentals of Computational Swarm Intelligence. Wiley and Sons, 2005.

[12] Christian Blum · Daniel Merkle (Eds.) Swarm Intelligence, Introduction and Applications. Springer, 2008.

[13] Longadge, R., Dongre, S. S., and Malik, L., "Multi-Cluster Based Approach For Skewed Data In Data Mining" (2013), IOSR journal of computer engineering(IOSR-JCE), 12(6), 66-73.

[14] Rashu,R.I., Naheena Haq, Rashedur M Rahman "Data Mining Approaches To Predict Final Grade By Overcoming Class Imbalance Problem"(2014), 17th International Conference On Computer And Information Technology (Iccit), 215-222.

[15] Kwak, J., Lee, T., and Kim, C.O., "An Incremental Clustering-Based Fault Detection Algorithm for Class-Imbalanced Process Data"(2015), IEEE Transactions On Semiconductor Manufacturing, 28(3), 212-220.

978-93-85670-72-5 © 2016 (RTCSIT)