www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 4 Issue 5 May 2015, Page No. 11783-11787

Efficient Multiclass Classification Model For Imbalanced Data.

Mr. Roshan M.Pote [1], Prof. Mr. Shrikant P. Akarte [2]

[1] ME (CSE) ,Second Year,Department of CSE, Prof. Ram Meghe Institute Of Technology and Research, Badnera,Amravai
Sant Gadgebaba Amravati University, Amravati, Maharashtra, India – 444701.
roshan4892@gmail.com

[2] Assistant Professor, Department of CSE, Prof. Ram Meghe Institute Of Technology and Research, Badnera Amravati.
SantGadgebabaAmravatiUniversity, Amravati, Maharashtra, India – 444701.
s_akarte25@rediffmail.com

Abstract: ***In this paper, we focused on developing efficient mining algorithm for multiclass classification from large of collection imbalanced data. And gives well sorted data. In the field of data mining, classification techniques can be used to find various selective features. This paper presents an innovative and efficient classification technique which includes the processes of feature selection and map reduction, to improve the effectiveness of using data for finding relevant and interesting information. In proposed system we can take sufficient .txt file as inputs & we apply variance algorithm & generate expected results. Classification is method to perform on poorly & minority class examples when the dataset is extremely imbalanced.***

Keywords: Data Mining, Multiclass Classification, imbalanced data, Feature Selection.

1. INTRODUCTION

One of the important data mining methods in biomedical research is classification. in the classification task, training examples are required to presage a target class of an unseen example. Nevertheless, training data sometimes have imbalanced class distribution. Inadequacy of absolute amount of examples of some classes for training a classifier is one of major reasons for the problem [1]. A well balanced dataset is very important for creating a good prediction model. Medical datasets are often not balanced in their class labels. Most existing classification methods tend to perform poorly on minority class examples when the dataset is extremely imbalanced. This is because they aim to optimize the overall accuracy without considering the relative distribution of each class [2]. In the field of biomedical, the issue of learning from these imbalanced data is highly important because it can invent useful knowledge to make important decision on the other hand it can also be extremely costly to misclassify these data. In machine learning, multiclass or multinomial classification is the problem of classifying instances into more than two classes [5]. While some classification algorithms naturally permit the use of more than two classes, others are by nature binary algorithms; these can, however, be turned into multinomial classifiers by a variety of strategies.

The remainder of this paper is organized as follows. Section III provides overview of Proposed System, which Section IV describes Methods Implemented. Section V describes Experimental Analysis. Section VI describes Result of Experiment, Finally section VII concludes this paper.

2. LITERATURE REVIEW

2.1 Imbalanced Data

Many researchers have studied the problem of imbalanced data classification in order to improve the performance of classification models. However, most of them only concentrate their works on binary classification. General ideas such as feature selection, sampling-based approach, and cost sensitive learning can easily be extended to multiclass problem, but it is rather difficult for algorithm-specific techniques [2]. There are many research works that try to improve traditional techniques or develop new algorithms to solve the class imbalance problem. However, most of those studies are focused only on binary case or two classes. Only a few researches have been done for multiclass imbalance problem that is much more common and complex in the real-world application.

We will study the multiclass imbalanced data problem, and developed new classification algorithms that can effectively handle the imbalance problem in many biomedical domains. Crammer K, Singer Y, Cristianini N, Shawe-taylor J,Williamson B. Implemented the algorithm of multiclass kernel-based vector machines. Wasikowski M, Chen X W. worked on the small sample class imbalance problem using feature selection. Similarly, Chen X, Gerlach B, Casasent D. introduced support vectors for imbalanced data classification [4].

2.2 Data Mining

The most commonly accepted definition of "data mining" is the discovery of "models" for data. A "model," however, can be one of several things. Statisticians were the first to use the term "data mining." Originally, "data mining" or "data dredging" was a derogatory term referring to attempts to extract information that was not supported by the data. For ex.: Suppose our data is a set of numbers. This data is much simpler than data that would be data-mined, but it will serve as an example. A statistician might decide that the data comes from a Gaussian distribution and use a formula to

compute the most likely parameters of this Gaussian. The mean and standard deviation of this Gaussian distribution completely characterize the distribution and would become the model of the data. More recently, computer scientists have looked at data mining as an algorithmic problem. In this case, the model of the data is simply the answer to a complex query about it. There are many different approaches to modeling data [6].

2.3 Hadoop Overview

Hadoop is a distributed computing framework released by Apache Foundation, it is Google's open source implementation of the cloud computing model, and it can be efficient, reliable, scalable way to process data. Its core idea is to build on a large number of cheap and efficient cluster hardware devices, in the form of software processing to provide storage and computing environment for the huge amounts of data, and provide a unified standard interface, is a highly scalable distributed computing systems.

Hadoop is a MapReduce programming model and mass data. It has made a lot of simulation system in the cloud computing, such a calculation based on the concept of cloud modeling and simulation platform of COSIM-CSP system, a new mode of the networked manufacturing, private cloud framework for visual simulation, and the military training system [7].

2.4 Classification

Classification, which is the task of assigning objects to one of several predefined categories, is a pervasive problem that encompasses many diverse applications. Examples include, detecting spam email messages based upon the message header and content, categorizing cells as malignant or benign based upon the results of MRI scans, and classifying galaxies based upon their shapes. The input data for a classification task is a collection of records. Each record, also known as an instance or example, is categorized by a tuple (x, y), where x is the attribute set and y is a special attribute, designated as the class label (also known as category or the target attribute). The attributes set in a dataset for classification can be either discrete or continuous but the class label must be a discrete attribute. This is the key characteristic that distinguishes classification from regression, a predictive modeling task in which y is a continuous attribute [10].

2.5 Multiclass Classification

In multiclass classification, given a set of labeled examples with labels selected from a finite set, an inductive procedure builds a function that (hopefully) is able to map unseen instances to their appropriate classes.

3. PROPOSED SYSTEM

Figure shows the block diagram of the system. In which user first input the text file then read that text file, after reading apply feature selection. In feature selection process values are identified by the system. After that apply variance calculation algorithm, in which variant form of valued are reduced. And result is displayed in the output window.

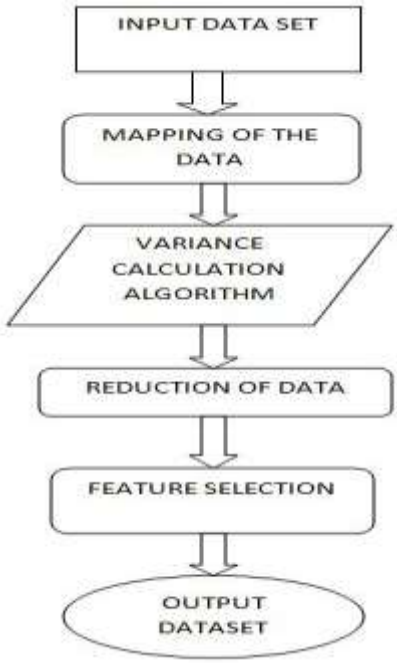

Figure 3.1: Proposed System Architecture

4. METHOD IMPLEMENTED

To remove redundant attributes from the data set & to choose the features that are most informative in classification task. We use the *variance* that is calculated in Naïve Bayesian. The features are filtered by variance corresponding to different class value. The reasons of why we choose variance of filter criterion are:

- Calculating variance is a step in fitting data to Gaussian distribution. It is time saving compared to other methods.
- The larger the variance is, the wider these data distribute. That means, even the feature have some little change or big change, it won't influence the result of learning much compared to data with small variance. Hence, output is more sensitive to features with small variance. The way we filter the features is by removing the largest variance one by one.

Here first normalize all the numerical features. This is because we need to make sure all the numerical features are in the same domain of [0, 1]. In the later part, when we measure the variance, same domain of features would be comparable. The second step is a step of fitting numerical data into Gaussian distribution. It is a step which is accomplished by the Gaussian naïve Bayesian approach. Assume $G = \{G_1, \ldots . G_m\}\ (m \leq n)$ is the numerical feature variable set $(G \subseteq F)$, and *V* is the class variable with class values $\{v_1 \ldots \ldots v_k\}$ (k >1), *k* **m* Gaussian distributions would be estimated and correspondingly, *k* **m* variances are computed. For each class variable, there are *m* variances corresponding to different features. For each feature vector $G_i(i \in (1, \ldots . , m))$, there are *k* variances corresponding to different class value. The third step is to calculate the average variance value for each feature. We simply sum the *k* variance together and divided by *k* . In this way, we can obtain the number of *m* average variance. According to the value of the average variance, we filter the features one by one from the largest value of variance.

5. EXPERIMENTAL ANALYSIS

5.1 Requirement Analysis

For the implementation of this system we used Ecillips IDE with Hadoop. In computer programming, Eclipse is an integrated development environment (IDE). It contains a base workspace and an extensible plug-in system for customizing the environment. Written mostly in Java, Eclipse can be used to develop applications. The Eclipse software development kit (SDK), which includes the Java development tools, is meant for Java developers. Users can extend its abilities by installing plug-ins written for the Eclipse Platform, such as development toolkits for other programming languages, and can write and contribute their own plug-in modules. Eclipse uses plug-ins to provide all the functionality within and on top of the runtime system. Also we used Haoop, Apache Hadoop is a framework that allows for the distributed processing of large data sets across clusters of commodity computers using a simple programming model. It is designed to scale up from single servers to thousands of machines, each providing computation and storage.

5.2 Hardware and Software Requirements

A. Hardware Requirement:

Processor : Dual Core or Onwards
RAM : 4 GB RAM
HDD : 40 GB
LAN : Enabled

B. Software Requirement:

Operating Platform : WINDOWS XP/ Windows 7
Front End : Eclipse-SDK
Back End : Hadoop

6. RESULT

Figure: 6.1. Application Window

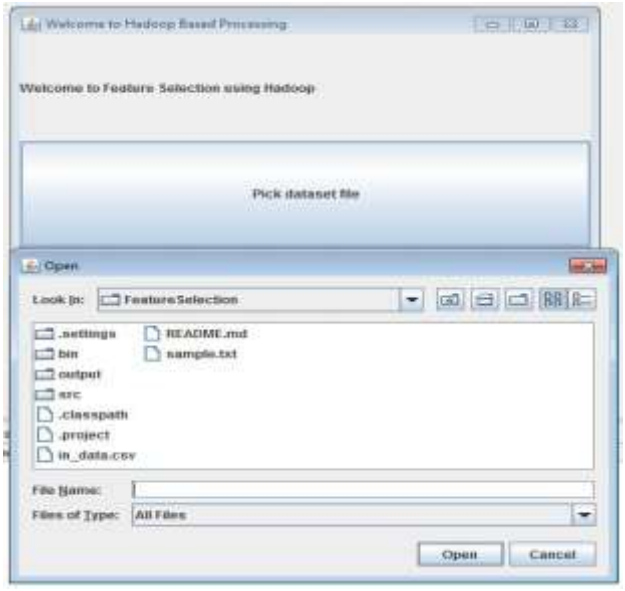

Figure:6.2 Select database file from the storage system

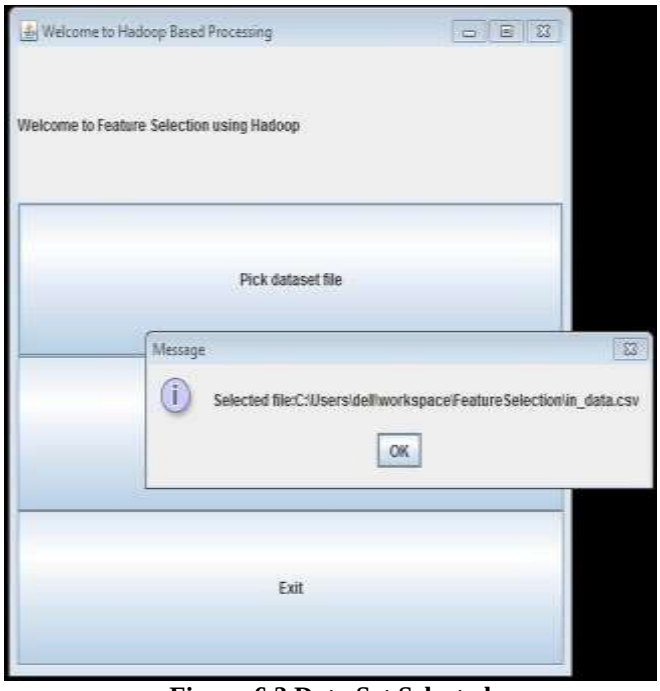

Figure:6.3 Data Set Selected

Figure:6.4 Apply Feature Selection Method

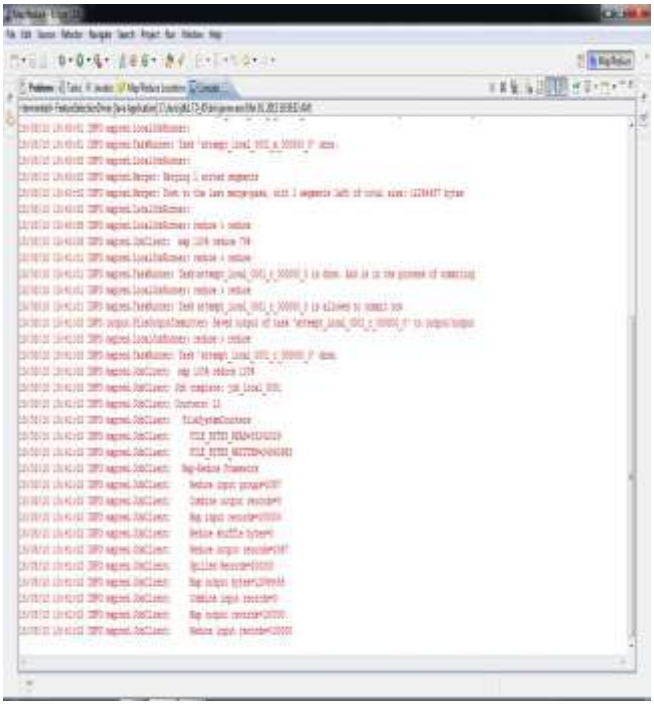

Figure: 6.5 Background Process of MapReduction & Feature Selection

```
3,3,3,2,2,2,1,3,2,1,2,2,2,2,0,2,3,2,2,2,2,2,2,1,2,1,1,3,2,1,3,2,2,0,2,3
,3,2,2,1,3,2,1,2,1,2,2,0,3,1,1,2,2,2,2, No disease
3,3,3,2,2,2,1,3,2,1,3,2,2,2,0,2,3,2,2,2,2,2,2,1,2,1,1,3,2,1,3,2,2,0,2,3
,3,2,2,1,3,2,1,2,1,2,2,0,3,1,1,2,2,2,2, No disease
3,3,3,2,2,2,1,3,2,1,3,2,2,2,0,2,3,2,2,2,2,2,2,1,2,1,2,3,2,1,3,2,2,0,2,3
,3,2,2,1,3,2,1,2,1,2,2,0,3,1,1,3,2,2,2, No disease
3,3,3,2,2,2,1,3,2,1,3,2,2,2,0,3,3,2,2,2,2,2,2,2,2,1,2,3,2,1,3,2,2,0,2,3
,3,2,2,1,3,2,1,1,1,2,2,0,3,1,2,3,2,2,2, No disease
1,2,2,1,2,1,3,2,3,1,3,2,1,1,-
0,3,1,3,1,3,0,2,1,2,1,2,2,1,2,1,2,3,2,3,0,1,2,2,1,2,1,2,1,1,1,3,1,2,-
0,2,3,1,2,2,3,1, Cancer
1,2,2,1,2,1,3,2,3,1,3,2,1,1,0,2,1,3,1,3,-
0,2,1,2,1,2,2,1,2,1,2,3,2,3,0,1,2,2,1,2,1,2,1,1,1,3,1,2,0,2,3,1,2,2,3,1,
    Cancer
1,2,2,1,2,1,3,2,3,1,3,2,1,1,0,3,1,3,1,3,-
0,2,1,2,1,2,2,1,2,1,2,3,2,3,0,1,2,2,1,2,1,2,1,1,1,3,1,2,0,2,3,1,2,2,3,1,
    Cancer
1,2,2,1,2,1,2,2,2,3,1,2,1,3,1,-
0,2,1,3,1,2,2,1,1,2,1,2,2,2,2,2,2,3,3,2,0,1,2,2,2,2,2,2,1,1,1,1,1,2,2,3
,3,3,3,2,1, Cancer
1,2,2,2,1,2,2,2,1,1,2,2,-0,2,0,1,3,1,3,2,0,2,2,1,-
0,2,1,3,2,1,2,1,2,2,0,2,2,2,2,2,1,2,2,1,2,2,3,1,1,2,3,1,3,2,2,2,
    Cancer
1,2,2,2,1,1,2,2,2,1,1,2,2,0,2,0,1,3,1,3,2,0,2,2,1,0,2,1,2,2,1,2,1,2,2,0,2,2
,2,2,2,1,2,2,1,2,2,3,1,1,2,3,1,3,2,2,2, Cancer
1,2,2,2,1,2,2,2,2,1,1,2,2,1,2,0,1,3,1,3,2,0,1,2,1,0,2,1,2,2,1,2,1,2,2,0,2,2
,2,2,2,1,2,2,1,2,2,2,2,1,1,2,3,1,3,1,2,2, Cancer
1,2,2,2,1,2,2,2,1,1,2,2,1,2,0,1,3,1,3,2,0,1,2,1,0,2,1,2,2,1,2,1,2,2,0,2,2
,2,2,2,1,2,2,1,2,2,2,2,1,1,2,3,1,3,2,2,2, Cancer
```

Figure: 6.6 Output Data Set of Selected Features Amongst The Redundant Data

7. CONCLUSION

A model for multiclass classification of imbalanced data is presented here. A large number of models are currently working in many locations. Which are working on the single class classification of the imbalanced data which is very costly as well as time consuming process. Now, it's a need to think apart from single class classification. The proposed system uses the feature selection variance algorithm to works & processes the multiclass data sets. And doing so it is found that the algorithm works correctly and this multiclass class classification is beneficial for saving time required for classification process on large scale, also for saving the storage space on storage space by avoiding repetition of data.

8. FUTURE SCOPE

The system can be upgraded with classification methods using Support Vector Machine (SVM). Support Vector Machine can be helpful for the researchers to achieve more accuracy in classification as well as it will also helpful for the prediction analysis to combine the result from both the systems.

References

[1] Piyaphol Phoungphol__, Yanqing Zhang, Yichuan Zhao. Robust Multiclass Classification for Learning from Imbalanced Biomedical Data, pp619-628 Volume 17, Number 6, December 2012

[2] Chawla N V, Japkowicz N. Editorial: Special issue on learning from imbalanced datasets. SIGKDD Explorations, 2004, 6: 1-6.

[3] M. S. Kim, "An Effective Under-Sampling Method for Class. Imbalance Data Problem," in *Proc. 8th International Symposium on Advance intelligent System (ISIS 2007)*, 2007.

[4] Y. Liu *et al.*, "Combining integrated sampling with SVM ensembles for learning from imbalanced datasets," *Information Processing & Management*, vol. 47, no. 4, pp. 617-631, Jul, 2011.

[5] R. Laza *et 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] J.Han an K.C.-C.chang,"Data Mining for Web Intelligence," Computer,vol.35,no.11,pp.64-70, Nov.2002

[7] hadoop.apache.org, Apache Foundation.

[8] UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science.

[9] Mike Wasikowski, Member and Xue-wen Chen, "Combating the Small Sample Class Imbalance Problem Using Feature Selection", IEEE Transactions on Knowledge and Data Engineering, Vol. 22, No. 10, October 2010.

[10] S¸ eyda Ertekin1, Jian Huang, L´eon Bottou, C. Lee Giles "Active Learning in Imbalanced Data Classification."

Author Profile

Mr. Roshan M. Pote,

ME (CSE), Second Year,Department of CSE,Prof. Ram Meghe Institute Of Technology and Research, Badnera, Amravati. Sant Gadgebaba Amravati University, Amravati, Maharashtra, India – 444701.

Prof. Mr. Shrikant P. Akarte,
[2] Assistant Professor, Department of CSE, Prof. Ram Meghe Institute Of Technology and Research, Badnera,Amravati. Sant Gadgebaba Amravati University, Amarvati, Maharashtra, India – 444701.