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. 4, Issue. 4, April 2015, pg.574 – 579

RESEARCH ARTICLE

PREDICTIVE DATA MINING FRAMEWORK FOR MEDICAL DATA

Mr.Rahul Pahlajani[1], Prof. Mr. Shrikant P. Akarte[2]

[1]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

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

[1] rahulpahlajani@gmail.com, [2] s_akarte25@rediffmail.com

Abstract— In this paper, we focused on developing efficient method for multiclass classification from large of collection predictive data, and present a framework which will work on the cross platform to show data classified manner. In the field of data mining, to find various selective features, classification techniques can be used. This paper presents an innovative and efficient classification technique which includes the processes of feature selection and map reduction, for finding applicable and interesting information also to improve the effectiveness of using data. In proposed system we can take sufficient .txt file as inputs & we multiclass SVM method & generate expected results. Classification is method to perform on poorly & minority class examples when the dataset is extremely imbalanced.

Keywords— Data Mining, Multiclass SVM, imbalanced data, Feature Selection, Framework

I. INTRODUCTION

In today's increasing computational environment, nearly all of our activities from birth until death Stored or leave as digital traces. Health records, schools attended, college record, office record, wages earned, —these and countless other data capturing the details of our daily lives serve as our digital social footprint. Collectively, these digital records—across a group, town, county, state, or nation—form a population's social genome, it act as the footprints of our society in general. If data is properly integrated, analyzed, and interpreted, social genome data could offer crucial insights to best serve our greatest societal priorities like healthcare, economics, education, and employment. The social sciences, health sciences, computer science, and statistics that applies quantitative methods and computational tools for large amount of data[1].

Data mining tasks can be classified to tasks of description and prediction. While description aims at finding human-interpretable patterns and associations, after considering the data as a whole and constructing a model prediction seeks to foretell some response of interest. Although the goals of description and prediction may overlap, the main distinction is that prediction requires the data to include a special response variable .The models generated by some prediction methods may point out some interesting patterns. The goal of predictive data mining in clinical medicine is to construct a predictive model that is sound, makes reliable predictions and helps physicians improve their prognosis, diagnosis or treatment planning procedures.

Working on the open source operating systems are not that much convenient for the user who are familiar with windows, So to overcome this problem we derived a framework which will help to user for easy access [2].

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.

II. LITERATURE REVIEW

Healthcare is a field in which accurate record keeping and communication are critical and yet in which the use of computing and networking technology lags behind other fields. Healthcare professionals and patients are often uncomfortable with computers, and feel that computers are not central to their healthcare mission, even though they agree that accurate record keeping and communication are essential to good healthcare. In current healthcare, information is conveyed from one healthcare professional to another through paper notes or personal communication[3]. For example, in the United States, electronic communication between physicians and pharmacists is not typically employed but, rather, the physician writes a prescription on paper and gives it to the patient. The patient carries the prescription to the pharmacy, waits in line to give it to a pharmacist, and waits for the pharmacist to fill the prescription. To improve this process, the prescriptions could be communicated electronically from the physician to the pharmacist, and the human computer interfaces for the physicians, nurses, pharmacists and other healthcare professionals could be voice enabled.

Current communication mechanisms, based largely on paper records and prescriptions, are old-fashioned, inefficient, and unreliable. In an age of electronic record keeping and communication, the healthcare industry is still tied to paper documents that are easily mislaid, often illegible, and easy to forge. When multiple healthcare professionals and facilities are involved in providing healthcare for a patient, the healthcare services provided aren't often coordinated[4].

A. Classification

Classification divides data samples into target classes. The classification technique predicts the target class for each data points. For example, patient can be classified as "high risk" or "low risk" patient on the basis of their disease pattern using data classification approach. It is a supervised learning approach having known class categories. Binary and multilevel are the two methods of classification. In binary classification, only two possible classes such as, "high" or "low" risk patient may be considered while the multiclass approach has more than two targets for example, "high", "medium" and "low" risk patient. Data set is partitioned as training and testing dataset. Using training dataset we trained the classifier. Correctness of the classifier could be tested using test dataset. Classification is one of the most widely used methods of Data Mining in Healthcare organization. Hu et al. used different classification method such as decision tree, SVM and ensemble approach for analysing microarray data[7].

B. Support Vector Machine (SVM)

The Support Vector Machine (SVM) is a state-of-the-art classification method introduced in 1992 by Boser, Guyon, and Vapnik. The SVM classifier is widely used in bioinformatics (and other disciplines) due to its high accuracy, ability to deal with high-dimensional data such as gene expression, and exibility in modeling diverse sources of data. As existing SVMs were initially developed for binary classification but it could be efficiently extended for multiclass problems. The support vector machine classifier creates a hyper plane or multiple hyper planes in high dimensional space that is useful for classification, regression and other efficient tasks. SVM have many attractive features due to this it is gaining popularity and have promising empirical performance. SVM constructs a hyper plane in original input space to separate the data points. Some time it is difficult to perform separation of data points in original input space, so to make separation easier the original finite dimensional space mapped into new higher dimensional space. Kernel functions are used for non-linear mapping of training samples to high dimensional space[6].

C. Hadoop

Apache Hadoopis an open-source software framework for distributed storage and distributed processing of Big Data on clusters of commodity hardware. In very simple terms, Hadoop is a set of algorithms (frameworks) which allows storing huge amount of data and processing it in a much more efficient and faster manner via distributed processing. So essentially, the core part of Apache Hadoop comprises two things: a storage part (Hadoop Distributed File System or HDFS) and a processing part (MapReduce). Its Hadoop Distributed File System (HDFS) splits files into large blocks (default 64MB or 128MB) and distributes the blocks amongst the nodes in the cluster. For processing the data, the Hadoop Map/Reduce ships code (specifically Jar files) to the nodes that have the required data and the nodes then process the data in parallel. This approach takes advantage of data locality, in contrast to conventional HPC architecture which usually relies on a parallel file system (compute and data separated, but connected with high-speed networking[10].

III. PROPOSED SYSTEM

Figure shows the block diagram of the system. In which user first take input the from the patient database as a text file then read that text file, after reading apply map-reduction technique. In map-reduction values are identified by the system. After that apply multiclass SVM method, in which data is classified. And as a result classified output is displayed in the output window.

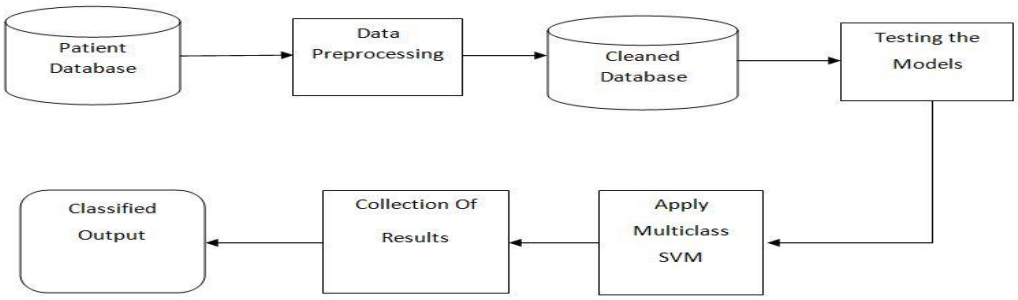

Figure 3.1. Proposed System Architecture

IV. METHOD IMPLEMENTED

In the proposed method the multiclass SVM classifier is applied to the data which is containing the records of the patient. While doing this firstly, it will calculate the actual amount of data present in the input dataset. Then it will perform search operation on the data to clarify that is any duplicate record is present or not, if so, it will avoid it. Then it will classify the data in the desired class to which it belongs to.

Also, as the all process is done with the hadoop which is an open source platform developed by Apache, So it can only be used on open source operating systems like unix. Linux, etc. But for the easy access on the windows platform for the user we have developed an application using cygwin package provider and ecilips. Doing so it gives easy access to Hadoop platform and to reduce the execution time of process on big data[4].

V. EXPERIMENTAL ANALYSIS

A. 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.

B. Hardware and Software Requirements

- **Software Requirements:**

  Operating System: Windows XP/7/8
  Front End : Eclipse IDE
  Back End : Hadoop

- **Hardware Requirement:**

  Processor : Dual Core or Onwords
  Hard Disk : 300 GB
  RAM : 4 GB
  LAN : Enabled

VI. RESULT

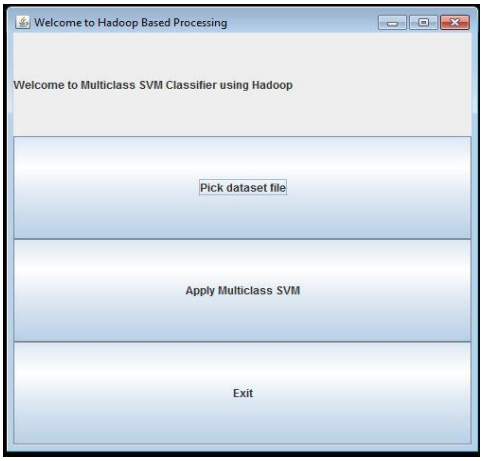

Figure: 6.1. Application Window

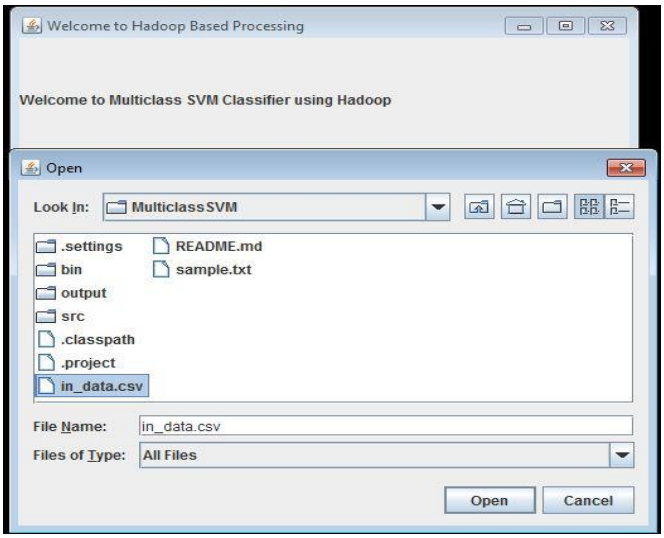

Figure: 6.2. Select Database File From The Storage System

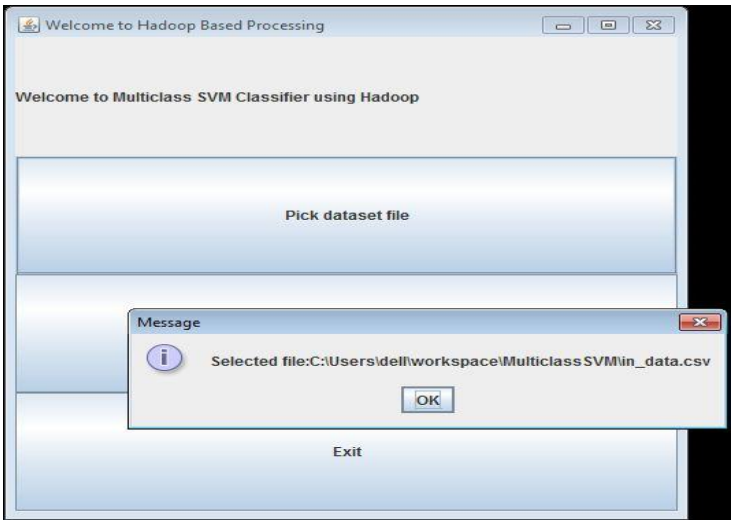

Figure:6.3 Data Set Selected

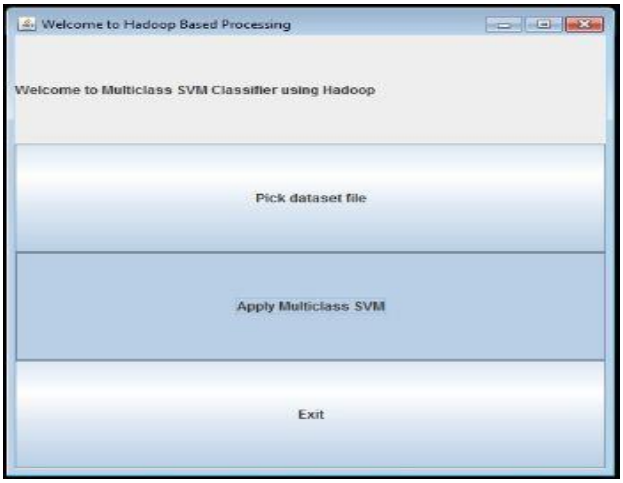

Figure:6.4 Apply Multiclass SVM Classifier Method

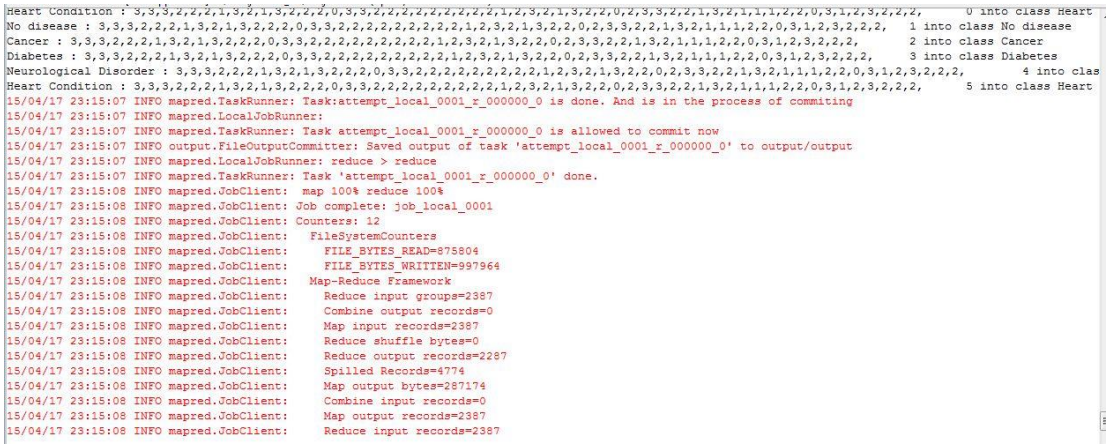

Figure: 6.5 Background Process of MapReduction & SVM

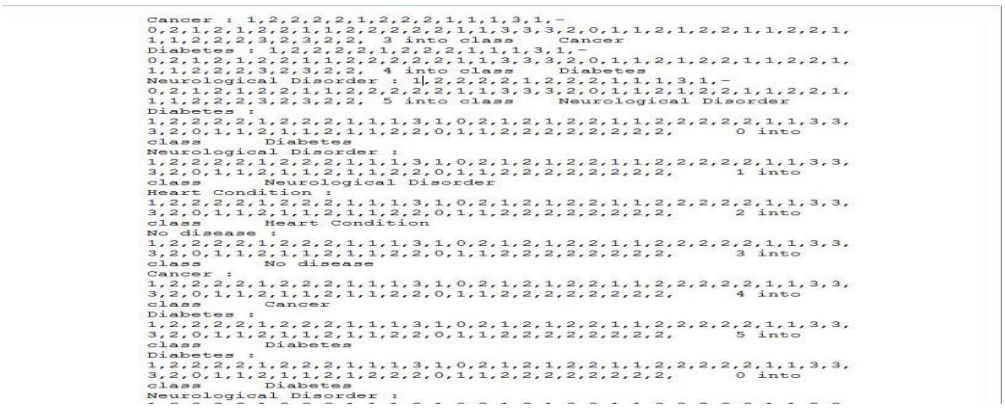

Figure: 6.6 Output Data Set of Classified Data

VII. CONCLUSION

There is no single data mining techniques which give consistent results for all types of healthcare data. The performance of data mining techniques depends on the type of dataset that we have taken for doing experiment. So, we can use hybrid or integrated Data Mining technique such as fusion of different classifiers, fusion of clustering with classification or association with clustering or classification etc.

We developed a cross platform application with the help of cygwin and performed the work on Hadoop Platform. As, a result we got more correct output in less time. Use of the application will definetly reduces the processing time on the big data for biomedical field.

REFERENCES

[1] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009). "The Elements of Statistical Learning: Data Mining, Inference, and Prediction". Retrieved 2012-08-07.

[2]Bellazzi&Zupan, 2008] Riccardo Bellazzi and BlazZupan, "Predictive Data Mining in Clinical Medicine: Current Issues and Guidelines", International Journal of Biological and Medical informatics 77(2008) pg. 81-97.

[3]Dunham & Sridhar, 2006] Dunham M. H. and Sridhar S., "Data Mining: Introductory and Advanced Topics", Pearson Education, New Delhi, ISBN: 81-7758-785-4, 1st Edition, 2006.

[4]C. J. Merz and P. M. Murphy. UCI repository of machine learning databases. Machine-readable data repository http ://www. ics .uci. edu/~mlearn/mlrepository.html, Irvine, CA: University of California, Department of Information and Computer Science, 1998.

[5] Allwein, E. L. Schapire, R. E. and Singer. Y., 2001, Reducing multiclass to binary: a unifying approach for margin classifiers. Journal of Machine Learning Research, 1:113– 141.

[6] Cristianini, N. and Shawe-Taylor, J., 2000, *An Introduction to Support Vector Machines and other Kernel-based Learning Methods*. Cambridge, UK: Cambridge University Press.

[7] Lee, Y., Lin, Y., and Wahba, G., 2001, *Multicategory support vector machines* Tech. Rep. 1043, Department of Statistics, University of Wisconsin, Madison, WI.

[8] Anjulan&Canagarajah, 2009] A. Anjulan and N. Canagarajah, "A Unified Framework for Object Retrieval and Mining", IEEE Transactions on Circuits and Systems for Video Technology. ISSN: 1051 - 8215, Vol. 19, No. 1, January 2009, pg 63-76.

[9] Fayyad et al., 1996] U. Fayyad, G. Piatetsky-Shapiro and P. Smith, "From Data Mining to Knowledge Discovery in Database", American Association for Artificial Intelligence. 1996 August: pp. 37-54.

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

Regards:

Mr.Rahul Pahlajani[1], Prof. Mr. Shrikant P. Akarte[2]

[1]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

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

[1]rahulpahlajani@gmail.com ,[2] s_akarte25@rediffmail.com, +91-8983622584,+91-9226792207.

