International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-4 Issue-4, April 2015

A Web-based Parallel Implementation to Classify Multiclass Large Datasets

Rabie Ahmed, Malek Rababah, Mehtab Mehdi, Mohammed Al-Shomrani

***Abstract**: Last few years are witnessed for growing the interest in Web-based Applications. Web applications typically interact with a back-end database to retrieve data to the user as dynamically generated output. In our work, an application is built for classification data sets, especially multi class large data sets, using parallel algorithm PMC-PBC-SVM. Our proposed application presents a general framework for data preprocessing, classification, and prediction. Our application gives an easy and interactive visual interface for classification multi class large data sets which will be useful for both technical and non-technical users.*

Keywords: Web-based Applications, Classification Algorithms, SVM, Parallel processing, Multiclass large Datasets.

I. INTRODUCTION

Classification is an important technique for data analysis that can be used to classify data classes. Data classification is a two-step. In the first step, a model is built using a known set of data classes; which is called training data set. In the second step, the learned derived model is tested using testing data, in which samples are randomly selected and are independent of the training samples, in order to estimate accuracy of this model by comparing the known class label with the class label predicted from learned model. The accuracy of a model on a given test set is the percentage of test set samples that are correctly classified by the model. In general, the learned model derived from the first step is represented in the form of classification rules, decision trees, or mathematical formulae. There are many techniques for data classification such as decision tree induction, Bayesian classification and Bayesian belief networks, Rule-Based classification, Neural Networks. However, Support Vector Machine is the most popular one. Support Vector Machine (SVM) is one of the most effective machine learning algorithms that is based on statistical learning theory. Supervised and non-supervised learning techniques are two types of machine learning techniques for pattern recognition. Support vector machine is one of the supervised machine learning techniques where class label must be known in advance [1].

Manuscript published on 30 April 2015.
* Correspondence Author (s)
Rabie Ahmed, Department of Computer Science, Northern Border University/ College of Computing and Information Technology, Rafha, Saudi Arabia.
Malek Rababah, Department of Computer Science, Northern Border University/ College of Computing and Information Technology, Rafha, Saudi Arabia.
Mehtab Mehdi, Department of Computer Science, Northern Border University/ College of Computing and Information Technology, Rafha, Saudi Arabia.
Mohammed Al-Shomrani, King Abdulaziz University, Faculty of Science, P. O. Box 80203, Jeddah 21589, Saudi Arabia.

The main idea behind the SVM classification algorithm is to separate two point classes of a training data set with a surface that maximizes the margin between them [2]. Therefore, for a training data set D, where

$$D = \{(X_i, Y_i), i = 1,2,\dots,n, X_i \in R^M, Y_i \in \{-1,+1\}\} \qquad (1-1)$$

Which has data samples X_i and labels Y_i, the objective is to find a function $F: R^M \rightarrow R$, for trained parameters W_i and b,

$$F(X_i) = \langle X_i, W_i \rangle + b = \sum_{i=1}^{m} X_i W_i + b \qquad (1-2)$$

Then, for any sample (X,Y), the class assigned by the model is

$$Y = \text{Sign } F(X) = \begin{cases} -1 & \text{if } F(X) \le 0 \\ +1 & \text{if } F(X) > 0 \end{cases} \qquad (1-3)$$

For Multi-class SVM classification, the parallel algorithm that performs the training is the One Versus One (OVO) algorithm which generates $\frac{K(K-1)}{2}$ tasks to be executed concurrently among the available p processors. Then the parallel complexity becomes $\Theta\left(\frac{K M N^2}{P} + T_c\right)$, where T_c is the complexity due to communication for task scheduling and combining the results. The parallel performance is evaluated by the relative speedup (S), which is defined as the ratio of the time taken to solve a problem on a single processing element to the time required to solve the same problem on a parallel computer with p identical processing elements. Also, Efficiency (E) is another way to analyze the parallel implementation, which is defined as the ratio of speedup to the number of processing elements [3]. Two efficient parallel algorithms, SMC-PBC-SVM and PMC-PBC-SVM. SMC-PBC-SVM combines Parallel Binary Class with Serial Multi Class Support Vector Machines for classification while PMC-PBC-SVM combines Parallel Binary Class with Parallel Multi Class Support Vector Machines for classification. The main idea in these algorithms is how to divide a set of processors into two disjoint subsets, one is responsible for multi class case and the other is responsible for binary class case. In SMC-PBC-SVM algorithm, the multi class case group has one processor which is used to solve multi case in serial while the binary class case group has the rest of processors which are used to solve binary case in parallel [4]. On the other hand, in PMC-PBC-SVM algorithm, we divide a set of processors as a grid where each row is used to solve different binary case in parallel which means multi case is solved in parallel [5].

Web based applications are those application which implements on the internet. These types of applications need a web browser. These applications are created in the browser supported programming language, as JSP, Java Script, HTML, CSS etc. WEKA is an example of these types of applications for classification [6]. Web applications commonly use a combination of server-side script (JSP, PHP, etc) and client-side script (HTML, Javascript, etc.) to develop the application. The client-side script deals with the presentation of the information while the server-side script deals with all the hard stuff like storing and retrieving the information.

II. Literature Review

Since last few years are the witnessed of the parallel machine learning algorithms. Many researchers have been shared their thoughts and have given so many algorithms. the parallel support vector machine was introduced in [7]. The graphic processors have been discussed in [8] to describe the solver of the GPUs (Graphic Processing Unit) for SVM. Glenn M. Fung and Mangasarian have constructed a fast algorithm for the multiclass large datasets [9]. This work is very similar to our work but in our work we have implemented it on the web. They used the term PSVM for the Proximal Support Vector Machine while we use the same term as Parallel Support Vector Machine. Elbe Frenk and M. Hall also discussed about the WEKA in the bioinformatics with their classification and clustering problem [6]. Rong-En, FanKai-Wei, ChangCho-Jui, HsiehXiang-Rui, WangChih-Jen and Lin have developed the LIBLINEAR project, which classifies multiclass large dataset using the web [10]. Paul Pavlidis1, Ilan Wapinski and William Stafford Noble have discussed about the SVM classification on the web [11]. The difference between [10] and [11] is that [10] is for the binary data set while [11] is for large data set.

III. PMC-PBC-SVM Algorithm

A new classification algorithm is being introduced because most of the results are just encouraging, so more work is needed to be done to make it more effective. To produce an efficient and effective parallel algorithm for classification which is named PMC-PBC-SVM, a new algorithm is merged parallel binary classification with parallel multi class classification.

The main motive is to divide a set of processors into two subsets. The first one is responsible for the multi class case and the second one is responsible for binary class case. Through PMC-PBC-SVM algorithm, the division of a set of processors as a grid where each row is used to solve different binary case in parallel which means multi case is solved in parallel too, can be done. IT enables us to do two level parallelisms. This new algorithm named PMC-PBC-SVM helps us to combine a parallel Multi Class Support Vector Machine with Parallel Binary Class for classification. Here it certainly becomes more efficient and effective by the following algorithm.

PMC-PBC-SVM Algorithm
Read a data set from input file
Makes sample groups of the same class together
Puts each two classes into one task
Sort $\frac{K(K-1)}{2}$ *tasks based on its size*
Divide processes into two groups (Mulgroup, Bingroup)
q= Sqrt(P);
Key= rank/q;

Mulgroup:
for(i=0; i<q; i++)
*Mulranks[i]= i*q;*
i= 0;

Bingroup:
MPI_Comm_split(MPI_COMM_WORLD,Key,rank,&bincomm)

While(true)
if(i< k(k-1)/2)

if(rank%q== 0)
*Using **Mulgroup***
1- Send Binary Task[key+i] to Bingroup
2- i=q(i+1)*
Otherwise
*Using **Bingroup***
1- Receive Binary Task[key+i] from Mulgroup
2- Solve Task[key+i] in parallel

Otherwise
Break

Build SVM Model
Write SVM Mode into output file

IV. Web-Based Proposed Application

Industries have lots of data but they do not have fast and appropriate knowledge extracting tools in order to benefit from the data. Classification allows industries to make better decisions. Classification aims to create a model from the training data set, which can be used to predict the classification of a new tuple. Classification can be applied in a wide variety of fields such as retail, target marketing, fraud detection and medical diagnosis. Our website is a web application that is able to perform parallel classification in order to minimize the time for data analysis and to maximize the accuracy of prediction

The proposed model consists of three main tasks: Preprocessing, Classification and predication as shown in figure 1. Based on this model a web application is implemented using JSP technology. To use our proposed application, a user must have an account to login the application as shown in figure 2. If a user does not have an account, he can create one via register link in figure 2.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

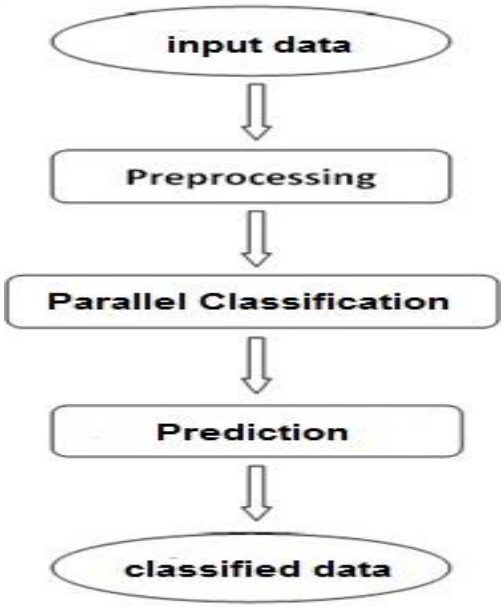

Figure 1: Parallel Classification Model

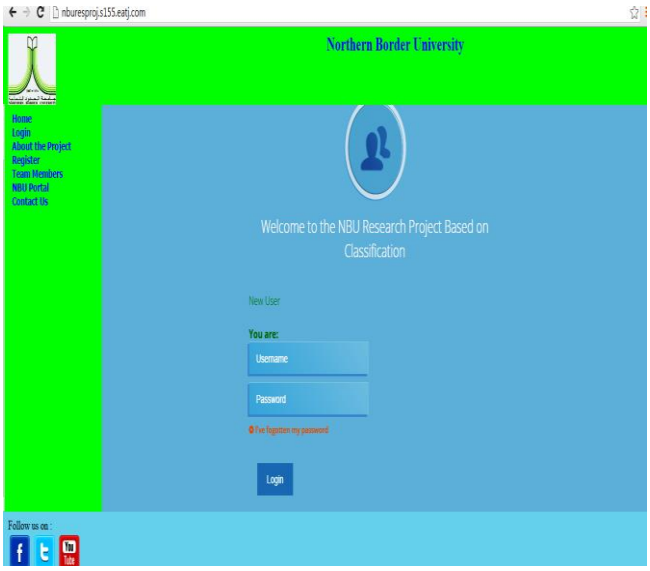

Figure 2: login screen

The main tasks will be explained in details in the following subsections:

4.1 PREPROCESSING OF DATA

Data Preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real world data can be incomplete, inconsistent and /or lack certain behaviors or trends. It is likely to contain many errors. Data preprocessing has been proven to resolve such issues. Data preprocessing prepares raw data for the Parallel Classification process. Real world databases today are susceptible to noisy, missing and inconsistent data. This is because of their huge size (often several gigabytes or more) and because they usually come from multiple, heterogeneous sources. Low quality data will result in low quality data mining results. Therefore, Preprocessing is one of the main steps that affect the accuracy of the results. If we have good preprocessing, then we will achieve correct and accurate results. We will have more control by means of data treatment. This is done by entering the data into an array. When using Preprocessing, you can enter data into an array, which allows you to edit it with ease as you wish.

There are numerous data preprocessing techniques, the most popular ones are used in our proposed application including the following:

i. Data cleaning: this involves handling missing values by ignoring the specific tuples or by filling it with a specific value. In this project, "K" has been applied as the nearest neighboring algorithm in order to handle a missing value.

ii. Data transformation: this involves converting class labels from a "string" format into a "numeric" format. Another data transformation process that is applied to the dataset is "normalization" "Normalization" is where the attribute data is scaled so that it falls within the range -1.0 to 1.0.

iii. Data reduction: data analysis takes a long time when applied to huge amounts of data. This can be performed using data cube aggregation, dimension reduction, data compression, numerosity reduction discretization and concepts hierarchy generation. Because Parallel Classification is used, the dataset will remain as it is. Therefore, we will have more accuracy in our prediction

As showing in Figure 3, preprocessing tab enables the user to select a file, upload it and apply preprocessing techniques to be ready for classification process

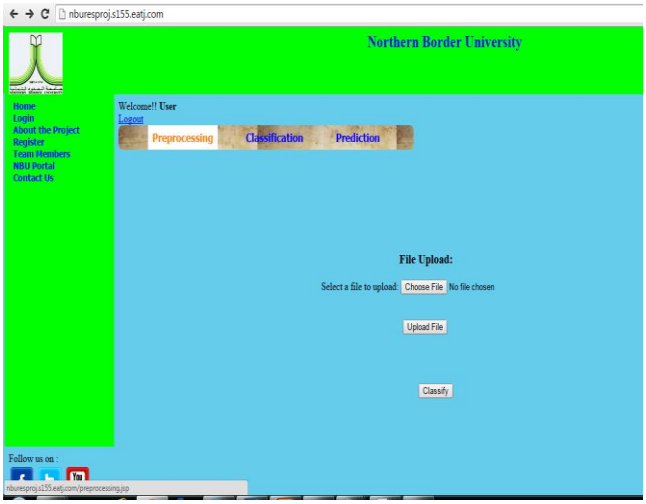

Figure 3 Preprocessing Screen

4.2 PARALLEL CLASSIFICATION

Classification is an important technique of data mining. It plays an important role in prediction, information retrieval, web searches and more. Most present classification techniques are serial, which is not practical for large datasets. Therefore, there is a need for Parallel Classification, which classifies large data set by dividing it among multiple processors in order to increase speed of classification.

We can classify to our data by so many ways, like as Matrix method, Decision Tree, SVM, k-neighbors etc. The main disadvantage of SVM is that to need the large memory for the computation.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Our system can classify the large data online on a very nominal requirement memory. After receiving preprocessed data from the preprocessing step, parallel classification starts to work according to PMC-PBC-SVM algorithm. A registered user can perform the classification step as shown in figure 4, after uploading the data through preprocessing step, and applying classification step, a trained model will created based on the training data set.

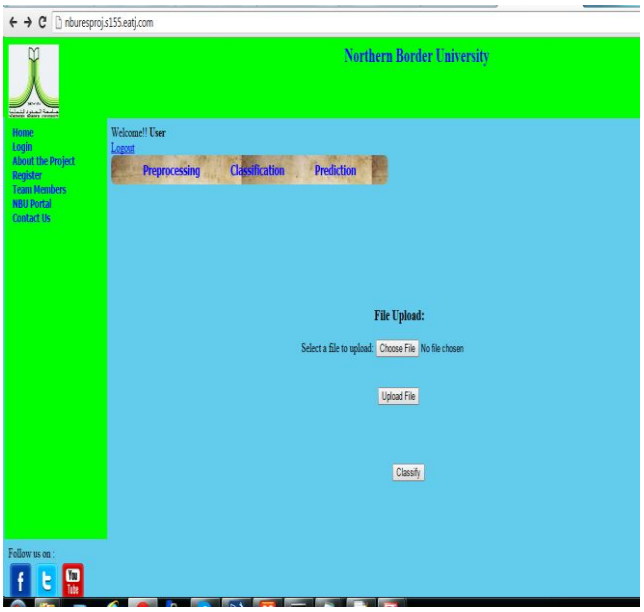

Figure 4 Classification process

4.3 Prediction

Prediction predicts categorical and continuously valued functions. For example: we can build a classification model to categorize bank loan applications, either as “safe” or as “risky”. We can also build a prediction model to predict the expenditure of potential customers (in dollars) on computer equipment, given their income and occupation.

Based on the trained model which created from classification step, unknown data can be predicted using third task of our system and then output the classified data into output file to the user.

V. CONCLUSION AND FUTURE WORK

The results achieved after applying our proposed application with different data sets from different size with both technical and non-technical users were positive through using an easy and interactive interface of our application as shown in the above figures. This application was implemented on the web using JSP and MySQL, so it is the pure online application for classifying the multiclass large datasets.

In future work, the proposed application can be modified by adding visualization of processors performance through computations to get more accurate results.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the Northern Border University for their financially support.

REFERENCES

[1] Stuart Andrews, Ioannis Tsochantaridis and Thomas " Support Vector Machines for Multiple-Instance Learning " ww.robots.ox.ac.uk.
[2] C. Cortes, and V. Vapnik. Support-vector networks. Machine Learning, 1995.
[3] Rajendran. Paralle Support Vector Machines for Mulicategory Classification of Large Scale Data. Dissertation, University of Southern Mississippi, 2007.
[4] Rabie Ahmed, Adel Ali, Chaoyang Zhang. SMC-PBC-SVM: A parallel implementation of Support Vector Machines for data classification. Conference on Parallel and Distributed Processing (PDPTA 2012).
[5] Rabie Ahmed, Mohammed Al-shomrani, “ Two Level Parallelism Implementation to Classify Multiclass Large Datasets”, Oriental Journal of Computers Science & Technology, 2014.
[6] Eibe Frank, Mark Hall , Len Trigg, Geoffrey Holmes , and Ian H. Witten " Data Mining in Bioinformatics using Weka" Frank-etal-bioinformatics Journal.
[7] CHRISTOPHER J.C. BURGES "Data Mining and Knowledge Discovery", 2, 121–167 Kluwer Academic Publishers, Boston. Manufactured in The Netherlands.
[8] Bryan Catanzaro , Narayanan Sundaram Kurt Keutzer, "Fast Support Vector Machine Training and Classification on Graphics Processors" http://parlab.eecs.berkeley.edu/
[9] GLENN M. FUNG , O. L. MANGASARIAN," Multicategory Proximal Support Vector Machine Classifiers " Machine Learning, 59, 77–97, 2005 2005 Springer Science + Business Media, Inc. Manufactured in The Netherlands.
[10] Rong-En, FanKai-Wei ,ChangCho-Jui ,HsiehXiang-Rui, WangChih-Jen Lin " LIBLINEAR: A Library for Large Linear Classification” The Journal of Machine Learning Research
[11] Paul Pavlidis1, Ilan Wapinski and William Stafford "Support Vector Classification for the web" Bioinformatics 2004.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)