Investigate the Impact of Dataset Size on the Performance of Data Mining Algorithms

Bondu Venkateswarlu
Research Scholar,
Departmentt of CS&SE,
Andhra University,
Visakhapatnam, India

Dr.GSV Prasad Raju
Professor, School of Distance Education,
Department of CS&SE,
Andhra University,
Visakhapatnam, India

ABSTARCT

KEEL (Knowledge Extraction based on Evolutionary Learning) tool is used to analyze the datasets to access the performance of various existing data mining algorithms. KEEL is an open source tool that can be used for a large no of knowledge data discovery task. It provides a simple GUI based and data flow to design experiments with different datasets to access the behavior of the various existing data mining algorithms. With the combination of classification Algorithms and Bayesian-D preprocessing technique used in KEEL tool, we analyze the performance of classification algorithms by varying the size of dataset records from 500 to 5000. We investigate the impact of dataset size on global classification error, standard deviation global classification error and correctly classified for both training and testing for the classification algorithms such as C4.5-C, AdaBoost-C and C4.5_Binirization-C. From the experimental result reveals the C4.5-C out performed and also found that by varying size from 500 to 5000 the variance of global classification error is 0.001727, standard deviation global classification error is 0.004158 and correctly classified is 0.998267.

Keywords—KEEL, C4.5-C, Adaboost, C4.5_Binirization-C,GUI

1. INTRODUCTION

Knowledge Discovery in Databases (KDD) techniques compose a set of key concepts for understanding the role of Data Mining (DM) and Machine Learning (ML) , providing with a very important tool for their professional training. Most of these techniques are somewhat complex; require a thorough practical analysis in order to understand the most prominent advantages and disadvantages of each one.

This analysis allows a better understanding of each of their components and their behavior over different kind of problems, thus helping to discern which one fits better to a particular case. The underlying problem here is that most of these techniques require a certain programming expertise along with considerable time and effort to write the computer program. In this way, KDD lessons become to mere programming classes instead of focusing into analyzing the different traits that characterize each technique. This problem can only be eased with the use of software tools that relieve students from programming tasks, allowing them to focus into the intrinsic characteristics of the KDD algorithms.

KEEL is an open source Java software tool to assess computational intelligence algorithms for DM problems including regression, classification, clustering and so on.

- It contains a big collection of classical and up-to-date techniques, including preprocessing and post processing approaches, hybrid models, statistical methodologies for contrasting experiments and so forth.
- It allows performing a complete analysis of new computational intelligence proposals in comparison to existing ones.
- KEEL has been designed with for both research and educational.

Data Mining is a detailed process of analyzing the large amounts of data and picking out the relevant information. Data mining can be defined as the extraction of or fetching the relevant information i.e. knowledge discovery from the large repositories of data i.e. the reason it is also called as Knowledge Mining. Data mining is also popularly known as knowledge Discovery in Data Base (KDD) refers to the non-trivial extraction of implicit, previously unknown and potentially useful information from data in data base. Preprocessing includes finding incorrect or missing data. Preprocessing also includes removal of noise or outliers, collecting necessary information to model on account per noise. Data mining is the task being performed to generate the defined results; integration by evaluation is how the data mining results are presented to the user. Different kinds of knowledge require different kinds of representation ex: classification, clustering, association rules etc. in this paper we are using classification and clustering techniques.

The demand on the blood bank sector makes it necessary to exploit the whole potential of stored data efficiently [1]. A fundamental tool to analyze the data gathered by blood bank through their information systems. To classify and predict the number of blood donors based on their age and blood group. J48 algorithm and weka tool has been used to build a data mining model to extract knowledge of blood donor's classification to aid clinical decisions in blood bank center. Different classification algorithms are analyzed for the effective classification of the data. To examine different classification algorithms and to find out a classification techniques with best accuracy rate and least error for the prediction of blood donors. The ability to identify regular blood donor will enable blood banks and voluntary organizations plan systematically for organizing blood donation camps in an effective manner [2]. The classification algorithms are used to identify blood donation behavior. The analysis had been carried out using standard blood transfusion dataset and using the CART decision tree algorithm in Weka. This algorithm provides good classification accuracy based model.

Deigning a model helps to identify the different blood groups with available stock in blood bank [3]. A classification technique is used for analysis of blood bank data set. The blood banks are based on donating blood and used for transfusions are made into medications. The analysis had been carried out using standard blood transfusion dataset. To classify the blood donors based on the sex, blood group, weight and age. The traditional manual data analysis has become in sufficient. Decision tree algorithm is used for blood groups are frequently requested during emergency situations. Two mixed integer linear programming (MILP) models are proposed to determine when a mobile collection must be organized at a location by taking into account human resource capacity, stock level and regulation constraints [4]. The first one is evaluates the amount of blood collected based on demographics, donor generosity and donor availability.

2. DATA MINING MODELS

The Supervised model makes prediction about the unknown data values by using the known values [11]. The Unsupervised model identifies the patterns or relationships in data and explores the properties of the data examined.

The figure below shows the data mining models and tasks that are used in our work.

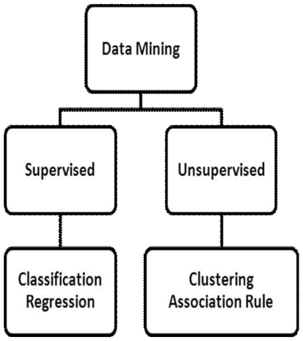


Fig.1 : Data mining models

2.1 Supervised Learning

In supervised learning (often also called directed data mining) the variables under investigation can be split into two groups: explanatory variables and one (or more) dependent variables. The target of the analysis is to specify a relationship between the explanatory variables and the dependent variable as it is done in regression analysis. To apply directed data mining techniques the values of the dependent variable must be known for a sufficiently large part of the data set. Training data includes both the input and the desired results. For some examples the correct results (targets) are known and are given in input to the model during the learning process. The construction of a proper training, validation and test set is crucial. These methods are usually fast and accurate. Have to be able to generalize: give the correct results when new data are given in input without knowing a priori the target.

2.2 Unsupervised Learning

Unsupervised learning is closer to the exploratory spirit of Data Mining. In unsupervised learning situations all variables are treated in the same way, there is no distinction between explanatory and dependent variables. However, in contrast to the name undirected data mining there is still some target to achieve. This target might be as general as data reduction or more specific like clustering. The model is not provided with the correct results during the training. Can be used to cluster the input data in classes on the basis of their statistical properties only Cluster significance and labeling. The labeling can be carried out even if the labels are only available for a small number of objects representative of the desired classes.

2.2.1 Classification

Classification is a data mining function that assigns items in a collection to target categories or classes. Classification is to accurately predict the target class for each case in the data. Classification model could be used to identify blood donors availability based on blood group and location . A classification task begins with a data set in which the class assignments are known. Classifications are discrete and do not imply order. Continuous, floating-point values would indicate a numerical, rather than a categorical, target.

The input data, also called the training set, consists of multiple records each having multiple attributes or features. Each record is tagged with a class label and the objective of classification is to analyze the input data and to develop an accurate description or model for each class using the features present in the data.

2.2.2 Regression

Regression analysis can imply a broader range of techniques that ordinarily appreciated. Statisticians commonly define regression is to understand “as far as possible with the available data how the conditional distribution of some response y varies across subpopulations determined by the possible values of the predictor or predictors”, Regression is a data mining (machine learning) technique used to fit an equation to a dataset. The simplest form of regression, linear regression, uses the formula of a straight line ($y = mx + b$) and determines the appropriate values for m and b to predict the value of y based upon a given value of x. Advanced techniques, such as multiple regression, allow the use of more than one input variable and allow for the fitting of more complex models, such as a quadratic equation.

2.2.3 Clustering

A cluster is a subset of objects which are “similar”. A subset of objects such that the distance between any two objects in the cluster is less than the distance between any object in the cluster and any object not located inside it. A connected region of a multidimensional space containing a relatively high density of objects

Clustering is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters. Help users understand the natural grouping or structure in a data set. Clustering: unsupervised classification: no predefined classes. Used either as a stand-alone tool to get insight into data distribution or as a preprocessing step for other algorithms. Moreover, data compression, outlier’s detection, understands human concept formation.

2.2.4 Association Rules

Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. An association rule has two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent. Association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships. Support is an indication of how frequently the items appear in the database. Confidence indicates the number of times the if/then statements have been found to be true. In data mining, association rules are useful for analyzing and predicting customer behavior. They play an important part in shopping basket data analysis, product clustering, and catalog design and store layout. Programmers use association rules to build programs capable of machine learning. Machine learning is a type of artificial intelligence (AI) that seeks to build programs with the ability to become more efficient without being explicitly programmed.

Real Datasets

The real datasets were all acquired in the UCI Machine Learning Repository (Asuncion and Newman, 2007). These are: ecoli, iris, pima, wine and thyroid.we have to compare with our dataset i.e., blood donor data set.

3 PROPOSED METHODOLOGY

In our proposed methodology, we collection of data from Blood Bank Repositories and apply the KEEL tool to classify the blood donor's information. In this methodology the availability and prediction may be viewed as a type of classification. The problem is usually is to evaluate the work through the training data set and then verify the results by using a test set of data. The following table shows the classification algorithms.

Table 1: Classification Algorithms

Type	Name of the Algorithm
Statistical	Classification, Regression, Bayesian
Distance	K-Means, K-Medoids
Decision Tree	C4.5,Adaboost ,C4.5_Binarization
Neural Network	NN Supervised learning

In this work we use Keel tool as a Data Mining tool. For this research work we have collected 5310 dataset records from various Blood Bank repositories. The dataset have 6 attributes.

Table 2: Attributes of dataset

Name of the Attribute	Description
Name	Name of the donor
Gender	M=male, F=female
Age	Age of the Donor
BG	Blood group of the donor
Place	Place of donor birth
District	District of donor available

This dataset has been implemented in KEEL Data Mining tool used to classify the models based on the dataset. after preprocessing of dataset classification is performed. It will create a model on the test data available and then classify the new data based on the model which is developed using the test data.

3.1 Algorithms

Systems that construct classifiers are one of the commonly used tools in data mining. Such systems take as input a collection of cases, each belonging to one of a small number of classes and described by its values for a fixed set of attributes, and output a classifier [5] that can accurately predict the class to which a new case belongs.

C4.5

C4.5 is a descendant of CLS and ID3. Like CLS and ID3, C4.5 generates classifiers expressed as decision trees, but it can also construct classifiers in more comprehensible rule set form. C4.5 made a number of improvements to ID3. Some of these are:

- Handling both continuous and discrete attributes - In order to handle continuous attributes, C4.5 creates a threshold and then splits the list into those whose attribute value is above the threshold and those that are less than or equal to it.
- Handling training data with missing attribute values - C4.5 allows attribute values to be marked as ? for missing. Missing attribute values are simply not used in gain and entropy calculations.
- Handling attributes with differing costs.
- Pruning trees after creation - C4.5 goes back through the tree once it's been created and attempts to remove branches that do not help by replacing them with leaf nodes.

AdaBoost

Ensemble learning deals with methods which employ multiple learners to solve a problem. The generalization ability of an ensemble is usually significantly better than that of a single learner, so ensemble methods are very attractive. The AdaBoost algorithm proposed by Yoav Freund and Robert Schapire is one of the most important ensemble methods, since it has solid theoretical foundation, very accurate prediction, great simplicity (Schapire said it needs only "just 10 lines of code"), and wide and successful applications.AdaBoost and its variants have been applied to diverse domains with great success. For example, Viola and Jones combined AdaBoost with a cascade process for face detection. They regarded rectangular features as weak learners, and by using AdaBoost to weight the weak learners, they got very intuitive features for face detection.

Boosting refers to the general problem of producing a very accurate prediction rule by combining rough and moderately inaccurate rules-of-thumb.By altering the distribution over the domain in a way that increases the probability of the "harder" parts of the space. Thus forcing the weak learner to generate new classifier that make less mistakes on these parts.

C4.5 Binarization

M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, F. Herrera, Dynamic Classifier Selection for One-vs-One Strategy: Avoiding Non-Competent Classifiers. Pattern Recognition 46:12 (2013) 3412-3424 Multiclassifier learning approach (One-vs-One / One-vs-All) with C4.5 as baseline algorithm To determine a set of decision trees that on the basis of answers to questions about the input attributes predicts correctly the value of the target attribute. Multiclass problems are reduced to binary problems by One-vs-One or One-vs-All strategy. The inference can be done using different aggregations in the OVO case.

Usually, it is easier to construct a classifier to distinguish between two classes than to consider more than two classes in a problem. This is why binarization techniques come up, to deal with multiclass problems by dividing the original problem in more easier to solve binary classification problems which are face up by binary classifiers. These classifiers are usually referred as base learners or base classifiers of the system. Different decomposition strategies can be found in the literature. The most common strategies are called "One-vs-One" (OVO) and "One-vs-All" (OVA). The former consists in dividing the problem in as many binary problems as all the possible combinations between pair of classes, so one classifier is learned to discriminate between each pair and then, the outputs of these base classifiers are combined in order to predict the output class. The latter approach learns a classifier for each class, where the class is distinguished from all other classes, so the base classifier giving a positive answer indicates the output class. In recent years, different methods to combine the outputs of the base classifiers from these strategies have been developed. This approach include the most robust techniques for the OVO scheme and the standard solution (max voting) for OVA one. The base classifier used is the well-known C4.5 decision tree. The decision tree is constructed top-down. In each step a test for the actual node is chosen (starting with the root node), which best separates the given examples by classes. C45 is based on ID3 algorithm. The extensions or improvements of ID3 are that it accounts for unavailable or missing values in data, it handled continuous attribute value ranges, it chooses an appropriate attribute selection measure (maximizing gain) and it prunes the result decision trees

Bayesian Discretizer

X. Wu. A Bayesian Discretizer for Real-Valued Attributes. The. Computer J. 39:8 (1996) 688-691. Discretization of real attributes to transform a set of numerical variables into nominal variables. Input variables may be either real or integer. Bayesian Discretizer is an algorithm that discretizes the non-nominal attributes (real or integer) of a group of instances. The task of a discretization algorithm is to build a set of intervals for each non nominal attribute. The value of the attribute is translated to the interval number to which the value belongs. Bayesian Discretizer is a method for supervised discretization. Initially there are no cut points selected. The process computes a Bayesian measure for each class which will add cut points to the discretization

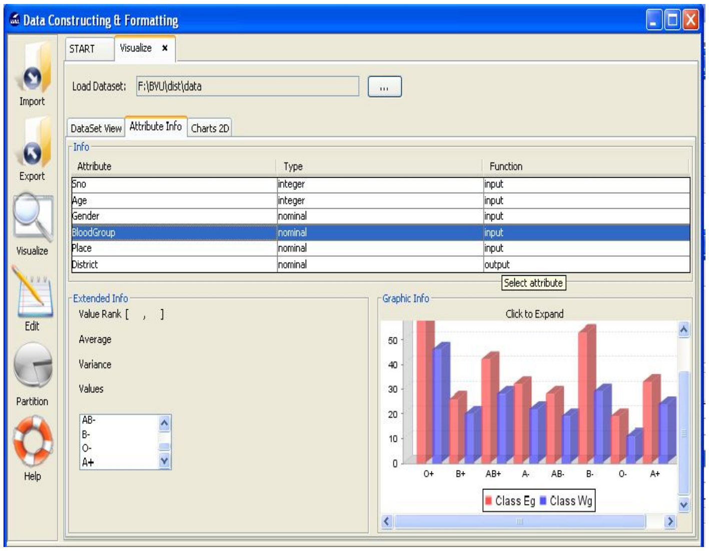
4 EXPERIMENTS AND RESULTS

4.1 Setting Up of Experiment:

In this work is to get the Global Classification Error, Standard Deviation Global Classification Error and correctly classified so that we can classify donor’s samples. We have a data file containing attribute values for 5310 data samples. The data file contains 6 attributes.

Open the KEEL tool interface and select the data management option. Select the source file format of the dataset. The formats admitted are CVS, TXT, PRN, C4.5, Excel, DIF, Property List and Weka. The import option allows a user to transform files in different formats (TXT, Excel, XML, etc.) to the KEEL format.

After specifying the file format used in source file, the path of this file must be specified. Click save button then keel file is created. That file converted from csv to keel format, that file is used in the KEEL experiments. Then we will do the partitions of the keel data file for classification of cross validation, the data file is having training data and testing data. i.e. one partition for training and another one for testing the data.



Load Dataset: F:\BVU\dist\data

Attribute	Type	Function
Sno	integer	input
Age	integer	input
Gender	nominal	input
BloodGroup	nominal	input
Place	nominal	input
District	nominal	output

Fig.2 : Dataset visualization

Fig.2 specifies The visualization options provide graphical information about existing KEEL datasets. There are different options related to this graphical information, where a user can select to view the content of a dataset, specific information about the attributes or to compare two attributes using charts. In the main window of the visualization menu, a user must select the path of source dataset (in KEEL format) that is going to be visualized. When the file is loaded, different information such as Dataset view, Attribute info, Charts 2D, Edit data about the dataset is shown according to the option selected. The Experiments Design section goal is to allow a user to create the desired experiments using a graphical interface.

We can design experiment in two ways, first one is experimental module and second one is Educational module.

In first process, open the experimental module. We import required dataset, when all the necessary dataset are selected, the experiment design process can continue. To do so, the user must click on the white graph panel to set the datasets node of the experiment. We design experiment with different combinations: data-C45- VisClasCheck, data- BayesianD -C45- VisClasCheck, data- AdaBoost.NC-C - VisClasCheck, data- BayesianD - AdaBoost.NC-C - VisClasCheck, data- C45_BinarizationC - VisClasCheck, data- BayesianD - C45_BinarizationC – VisClasCheck. After setting the datasets node we have to design the experiment with necessary options from the following.

First way once a experiment has been designed, the user can generate it through the option Run Experiment of the ’Tools’ menu. Use the tools bar button ZIP. At this point, the software tool will perform several tests about the completeness of the experiment.

We have to select a path for the experiment’s zip file. The generation process generates a ZIP file containing all the elements needed to run the experiment. The experiment generation is completed successfully. First of all, we have to unzip the named ZIP file in the machine that will run the experiment. We will obtain a directory called “experiment Name” (how we named its experiment). Then, we have to place himself into that “experiment Name” folder, and then into the “scripts” subfolder. To run the experiments, we have just to type and run the “java -jar

RunKeel.jar” command. The experiment is thus executed. Once the run of an experiment has finished, the associated result files can be found at the results subdirectory associated to each experiment. The experiment must be run using the RunKeel.jar file located at “experiment/scripts”

In second process, we have to observe the progress of the running of the experiment. For that open the Education module, We import required dataset, when all the necessary dataset are selected, the experiment design process can continue. To do so, we select the datasets node of the experiment. We design experiment with different combinations: data-C45, data-BayesianD -C45-, data- AdaBoost.NC-C, data- BayesianD -AdaBoost.NC-C, data- C45_BinarizationC , data- BayesianD -C45_BinarizationC. When the necessary setup is done in the xperiment, press▶ to run the experiment. Then we have to select start button to start running experiment. We can see the progress of the running experiment in partition area and results in report are of the tool window.

4.2 Result Analysis:

With the combination of classification Algorithms and Bayesian-D preprocessing technique used in KEEL tool, we analyze the performance of classification algorithms by varying the size of dataset records from 500 to 5000.

4.2.1 Test Results: for Different size of Dataset records

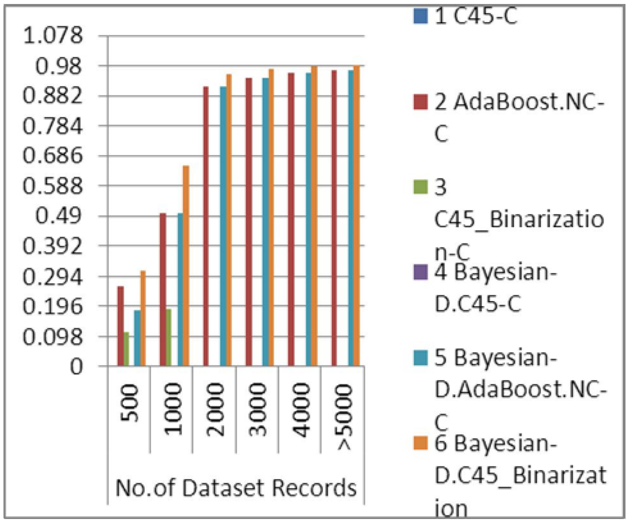


Fig 3: Global Classification Different size of Dataset records

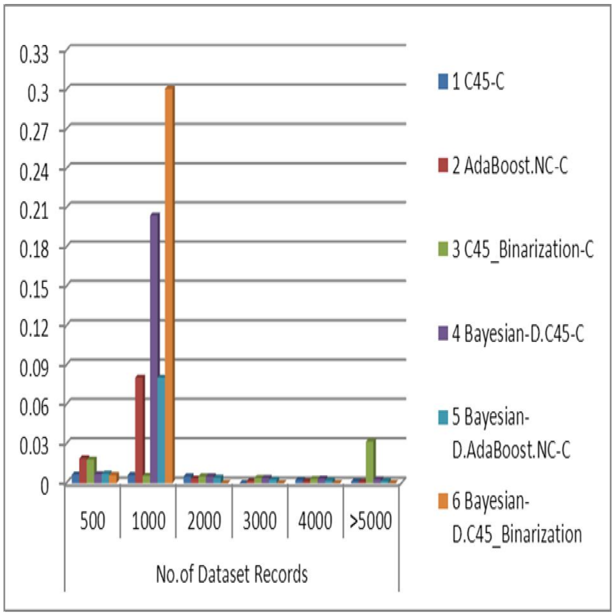


Fig 4: Stddev Global Classification Error Different size of Dataset records

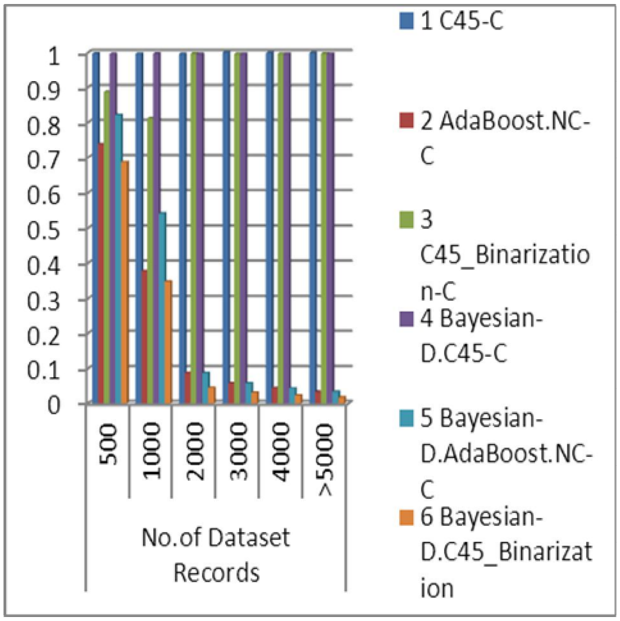


Fig 5: Correctly classified Different size of Dataset records

From the Experimental results, we observe that the impact of data set size is increased the error rates are varies depending on the type of algorithm. The data set size increased the error rate gradually increased and gradually decreased. We observed that when the data set size is large the error rate is minimized. From those results Global Classification Error for Different size of Dataset records is analyzed. Consider the average error rate of different algorithms is observed. The average Global Classification Error for C4.5-C is

0.00181, AdaBoost.NC-C is 0.756345, C45_Binarization-C is 0.051989, Bayesian-D.C45-C is 0.003539, Bayesian-D.AdaBoost.NC-C is 0.743982, Bayesian-D.C45_Binarization is 0.808584. Among all algorithms C4.5-C produced the better result. Fig 3 shows the clear analysis and impact of data set size is observed.

The experimental results of Stddev Global Classification Error. data set size increased the error rates are varies. We observed that when the data set size is large the error rate is minimized. From that results Stddev Global Classification Error for Different size of Dataset records is analyzed. Consider the average error rate of different algorithms are observed. The average Stddev Global Classification Error for C4.5-C is 0.003496, AdaBoost.NC-C is 0.017612, C45_Binarization-C is 0.011172, Bayesian-D.C45-C is 0.037581, Bayesian-D.AdaBoost.NC-C is 0.016215, Bayesian-D.C45_Binarization is 0.061362. out of these C4.5-C has the better result. Fig 4 shows the clear analysis and impact of data set size is observed. When the data set size is 1000 the impact of error rate is increased for all algorithms.

The experimental results of Correctly classified. data set size increased the rate of Correctly classified varies. We observed that when the data set size is large the rate of Correctly classified is increased. From that results rate of Correctly classified for Different size of Dataset records is analyzed. Consider the average rate of Correctly classified for different algorithms are observed. The average rate of Correctly classified for C4.5C is 0.998183, AdaBoost.NC-C is 0.223309, C45_Binarization-C is 0.947908, Bayesian-D.C45-C is 0.996595, Bayesian-D.AdaBoost.NC-C is 0.264143, Bayesian-D.C45_Binarization is 0.191613. out of these C4.5-C has the better performance. Fig 5 shows the clear analysis and impact of data set size is observed.

After observing the results, We investigate the impact of different dataset size on global classification error, standard deviation global classification error and correctly classified for testing the classification algorithms such as C4.5-C, AdaBoost-C and C4.5_Binarization-C and with the combination of preprocessor. From the experimental result reveals that the C4.5-C out performs.

4.2.2 Test Results for Different Dataset

With the combination of classification Algorithms and Bayesian -D preprocessing technique used in KEEL tool, we analyze the performance of classification algorithms on synthetic dataset along with different UCI Machine learning Standard datasets such as ecoli,iris, thyroid, pima,wine.

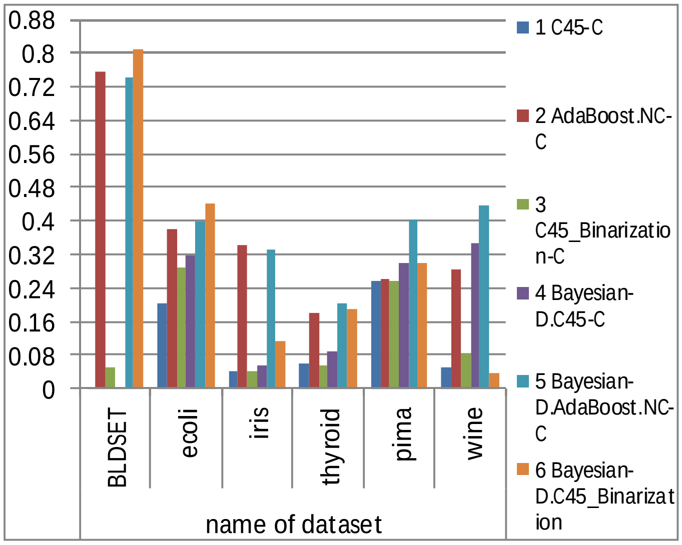


Fig 6: Global Classification Error for Different Dataset

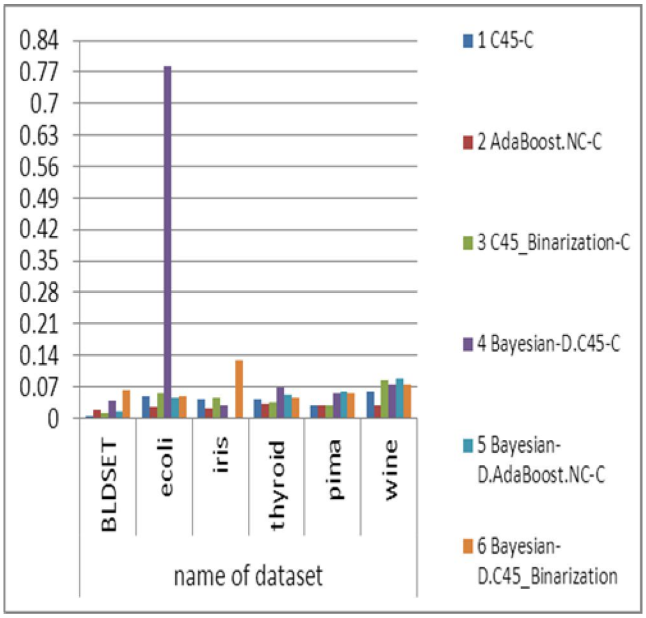


Fig 7: Stddev Global Classification Error for Different Dataset

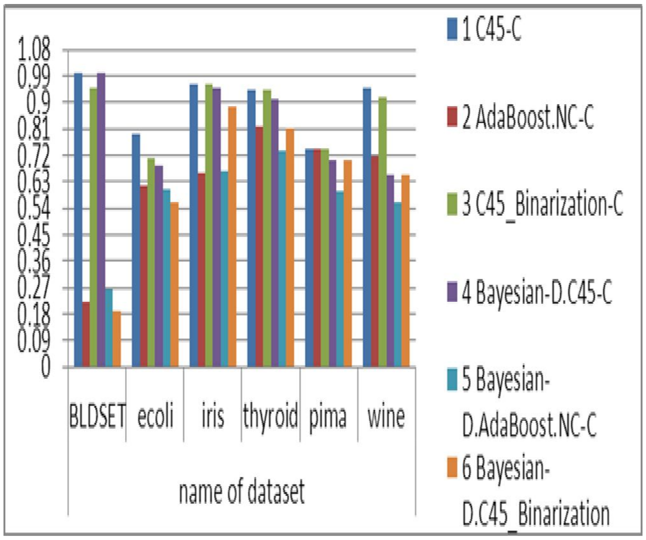


Fig 8: Correctly classified for Different Dataset

From the above fig 6 shows the experimental results of Global Classification Error for Different Datasets is analyzed. The overall performance of our dataset is good when compared to other Standard datasets.

From the above fig 7 shows the experimental results of Stddev Global Classification Error for Different Datasets is analyzed. The overall performance of our dataset is good when compared to other Standard datasets.

From the above fig 8 shows the experimental results of correctly classified for Different Datasets is analyzed. The overall performance of our dataset is good when compared to other Standard datasets.

We investigate the impact of different dataset on global classification error, standard deviation global classification error and correctly classified for both training and testing for the classification algorithms such as C4.5-C, AdaBoost-C and C4.5_Binarization-C. From the experimental result reveals the efficiency of our dataset (BLDSet) out performed

5 CONCLUSIONS AND FUTURE SCOPE

Data mining is the science of extracting the information from large databases to determine knowledge out of data and presenting it in a form that is easily understood to humans. Classification techniques are the main tasks of data mining with broad applications to classify the various kinds of data. it is used to classify the item according to the features of the item with respect to the predefined set of classes. In this research work, the different classification algorithms like C4.5, C4.5_Binarization, AdaBoost and the combination of Bayesion_D preprocessor are discussed and compared. These algorithms are applied on blood donors dataset to find out their accuracy and error rate. We also analyze the performance of these algorithms by varying the dataset size. from simulation results by varying the dataset size C4.5 is outperforms and C4.5_Binarization is moderate

REFERENCES

[1]. “Novel Centroid Selection Approaches For Kmeans-Clustering Based Recommender Systems”, Sobia Zahra a, Mustansar Ali Ghazanfar a, Asra Khalid a, Muhammad Awais Azam a,Usman Naeem b, Adam Prugel-Bennett c, doi 10.1016 2015.03.062, 0020-0255/ 2015 Elsevier .

[2]. “Analysis and Classification of Hardwood Species based on Coiflet DWT Feature Extraction and WEKA Workbench”, Arvind R. Yadav1, R. S. Anand 2, M. L. Dewal3, Sangeeta Gupta4 978-1-4799-2866-8/14/ 2014 IEEE

[3]. “Challenges In Knowledge Discovery And Data Mining In Datasets”, Mykhaylo Lobur1, Yuri Stekh2, Vitalij Artsibasov3 IEEE May 2011

[4]. “Spectral clustering and semi-supervised learning using evolving similarity graphs”, Christina Chrysouli, Anastasios Tefas, doi/10.1016/j. asoc.2015.05.0261568-4946/ 2015 Elsevier

[5]. “Exploration of Soft Computing Models for the Valuation of Residential Premises using the KEEL Tool”, Tadeusz Lasota, Ewa Pronobis, Bogdan Trawiński, Krzysztof Trawiński, 978-0-7695-3580-7/09 2009 IEEE.

[6]. “KEEL: A data mining software tool integrating genetic fuzzy systems”, Jesus Alcala-Fdez, Salvador Garcıa, Francisco Jose Berlanga , Alberto Fernandez, Luciano S ´ anchez, M.J. del Jesus and Francisco Herrera, 978-1-4244-1613-4/08/2008IEEE

[7]. “KEEL: a software tool to assess evolutionary algorithms for data mining problems”, J. Alcalá-Fdez · L. Sánchez · S. García · M. J. del Jesus ·,S. Ventura · J. M. Garrell · J. Otero · C. Romero · J. Bacardit · V. M. Rivas · J. C. Fernández · F. Herrera Published online: 22 May 2008 © Springer-Verlag 2008

[8]. “Investigation of Fuzzy Models for the Valuation of Residential Premises using the KEEL Tool”, Tadeusz Lasota, Jacek Mazurkiewicz, Bogdan Trawiński, and Krzysztof Trawiński, 978-0-7695-3326-1/08 2008 IEEE.

[9]. “Implementation and Integration of Algorithms into the KEEL Data-Mining Software Tool”, Alberto Fernandez, Juli´ an Luengo, Joaquin Derrac, Jes´ us Alcal´ a-Fdez, and Francisco Herrera H. Yin and E. Corchado (Eds.): IDEAL 2009, LNCS 5788, pp. 562–569, 2009. c Springer-Verlag Berlin Heidelberg 2009

[10]. “Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis”, Rashedur M. Rahman, Farhana Afroz Journal of Software engineering and Applications, 2013, 6, 85-97 doi/10.4236/jsea.2013.63013 Published Online March 2013

[11]. “Data Mining Techniques and Their Implementation in Blood Bank Sector –A Review”, Ankit Bhardwaj, Arvind Sharma, V.K. Shrivastava / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 Vol. 2, Issue4, July-August 2012, pp.1303-1309 1303.

[12]. “Application of Knowledge Discovery in Database to Blood Cell Counter Data to Improve Quality Control in Clinical Pathology”, D.Minnie, S.Srinivasan 978-0-7695-4514-1/11 IEEE.

[13]. “Classifying Blood Donors Using Data Mining Techniques” P.Ramachandran, Dr.N.Girija, 3Dr.T.Bhuvaneswari, IJCSET Feb 2011 Vol 1, Issue 1,10-13

[14]. “Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis”, Rashedur M. Rahman, Farhana Afroz Journal of Software engineering and Applications, 2013, 6, 85-97 doi/10.4236/jsea.2013.63013 Published Online March 2013

[15]. “Data Mining to Improve Safety of Blood Donation Process “, Madhav Erraguntla , Peter Tomasulo, Kevin Land, Hany Kamel, Marjorie Bravo, Barbee Whitaker, Richard Mayer, Sarita Khaire 978-1-4799-2504-9/14 2014 IEEE