Active Learning with Clustering for Mining Big Data

Md. Ibrahim (ID: 011133061)
Salman Masud (ID: 011141051)
Reza E Rabby (ID: 011141131)

Department of Computer Science & Engineering

United International University
United City, Madani Avenue, Badda, Dhaka 1212, Bangladesh

A thesis submitted for the degree of

BSc in Computer Science & Engineering

May, 2019

Abstract

Big data mining is become a key research issue nowadays. It's costly and also time-consuming to extract knowledge from big data. Big data is so big, it contains millions of data points that's why it's very difficult to build a learning model using machine learning and data mining algorithms. The main problem is to fit the hole data into the computer memory, which is quite impossible. Therefore, we need more scalable, robust, and adaptive learning algorithms. The exiting mining algorithms are design to handle relatively small datasets with fix number of class labels. In this paper, we have proposed a new method to select a few/ less number of training instances that we consider them as informative instances from a set of large data/ big data using clustering techniques. We have applied our proposed method in active leaning process for classifying big data. Active learning is a machine learning process in supervised learning where an oracle is ask to label the unlabelled training instances. It's very challenging and difficult task for connoisseur to label a large number of unlabelled data. Therefore, finding informative unlabelled training instances is necessary for learning from big semi-supervised data. We have collected six benchmark datasets from UCI machine learning repository and tested our proposed method using following machine learning algorithms: naïve Bayes (NB) Classifier, decision tree (DT) classifier (i.e. C4.5 and CART), Support Vector Machines (SVM), Random Forest, Bagging, and Boosting (AdaBoost).

This work is devoted to our mother and father.

Acknowledgements

Our time in United International University has been deeply motivated and guided by the people here. We are truly obliged towards them. Without their support and motivation, this thesis would have been never have surfaced and even if it did, it would have ended up in blunder. We would like to thank Dr. Dewan Md. Farid - Associate Professor,Department of Computer Science & Engineering, United International University for his insight and knowledge he shared with us. Without his help and guidance, We would have been most unfortunate. Again, we can not stop but appreciate what our advisor Dr. Dewan Md. Farid has done to us. He taught us about researches and how they are conducted. We also learned a great deal from our faculties. Without their help and teachings, We would not be here today.

Our deepest gratitude goes out for our family for without their support and motivation, we could not even had the opportunity to study in United International University let alone do our thesis. It is true sometimes it was frustrating to hear their bickering about our studies but now we realize how impactful those bickering words were to us because those were the words that drove us to where we are today.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Knowledge mining from big data is really hard and costly process. Big data is an expression used to mean a huge volume of both organised and unstructured data/ information [1, 2]. Big data is so big and normally inexactly organised that is very difficult to process using conventional database management techniques or relational database management techniques. Generally, the volume of data is too large in big data and it grows excessively quickly over the time [3]. Mining big data can possibly help organisations to enhance activities and decision making. In the present digital era, every day petabytes/ Exabytes of information comprising of billions to trillions of data records are generating from every sector of our life like web, medical science, online business/ e-business, finance & banking, bioscience etc. Data in big data that garner from various sources afterward organised and analysed to gain knowledge. Big data mining is the process of extracting knowledge to uncover large hidden information from the massive amount of complex data or databases [4, 5]. The data in big data comes in different forms including two-dimensional tables, images, documents and complex records from multiple sources. It must support search, retrieval and analysis. The 3 V's define big data: Volume (the quantity of data), Variety (the category of data) and Velocity (the speed of data in and out) [6]. It might suggest throwing a few more V's into the mix: Vision (having a purpose/ plan), Variation (ensuring that the data conforms to a set of specifications) and Validation (checking that its purpose is fulfilled) [7, 8].

Collecting and managing big data then extracting knowledge and informative information from big data is quite impossible using existing relational database management systems, so machine learning for mining big data become very popular in the recent time [9, 10]. Machine learning (ML) is an advance form of artificial intelligence (AI) that gives the ability to the machine to learn from historical data and grow from experiences without being explicitly programmed [11, 12].Machine learning can be divided into 2 different types: supervised and unattended. Further these can be divided into: semi-supervised Learning and semi-unsupervised learning. Supervised learning creates a basis of knowledge from the previously classified models that helps to classify new models. The objective of this learning is to map the input to a class. This model then can be used to correctly classify unseen instances. Unsupervised learning is used to draw inductions from datasets comprising of information without labeled responses. The most widely recognised unsupervised learning technique is cluster analysis, which is utilised for exploratory data analysis to discover hidden patterns or grouping data. Semi-supervised learning on the other hand is half-way between supervised and unsupervised learning. The primary aim of SSL is to overcome both supervised and unattended learning disadvantages. Supervised learning requires an enormous volume of training data for the classification of the test data. Unsupervised learning, on the contrary, requires no data which clusters the information on a similar basis. [13, 14].

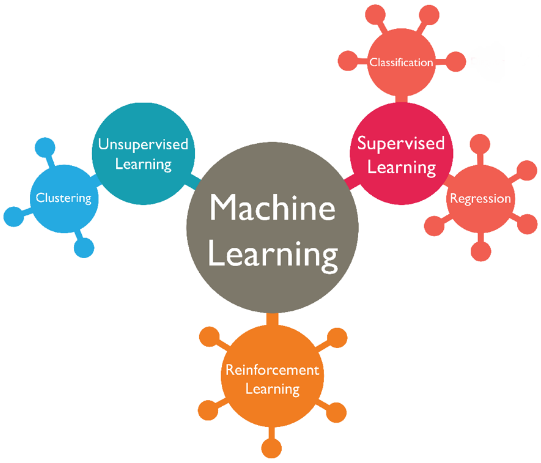

Figure 1.1: Branches of Machine Learning.

In this paper, we have proposed a cluster-based approach for building a classifier in active learning process with less number of training instances. The idea is to select a small chunk of data from the big data that is informative enough to build a learning model. These small chunks of data are known as the informative instances, and using these informative instances we can build classifier as these small chunks of data represent the big data. The proposed approach uses active learning for mining big data. Active learning is a special case of semi-supervised machine learning in which a learning algorithm is able to interactively query the oracle/ user to obtain the desired outputs of unlabelled data. We have collected six benchmark datasets from UCI machine learning repository [15] and tested our proposed method using following machine learning algorithms: naïve Bayes (NB) Classifier, decision tree (DT) classifier (i.e. C4.5 and CART), Support Vector Machines (SVM), Random Forest, Bagging, and Boosting (AdaBoost) [16–18].

1.2 Objectives of the Thesis

To achieve our goal, we use our general approach to active learning to develop theoretical foundations, supported by experiment results, for scenarios in each of the three previously mentioned learning tasks: classification, parameter estimation, and structure discovery. We overcome each of these three tasks by focusing our work on two frequent models of machine learning: C4.5(Optimized method for Decision tree) and Simple Kmeans. For the process of classification, C4.5(Optimized method for Decision tree) have laid strong theoretical foundations and overwhelming and experimental successes. Decision trees, which can be used for the purposes of classifications and predictions, are a tool to support decision making. As a decision tree can accurately classify data and make effective predictions, it has already been employed for data analyses in many application domains. Real life applications of decision tree are in business management, engineering, and health-care management etc. We develop a model for active learning with decision tree and demonstrate that active learning can significantly improve the performance of this already strong classifier.

One of the sorts of unsupervised learning is K-means clustering. This sort is explicitly utilized when we are given a circumstance where we need to manage unlabeled information. The manner in which this algorithm works is it makes k number of clusters iteratively and allocates every data point dependent on the highlights given. Data points are assembled together dependent on their similarity to one another. After the k-means algorithm is executed, the accompanying outcomes are created:

- The centers or Centroids of the k number of cluster which helps in identifying the new data.
- The identification or the labels for the training data.

By combining these two algorithms, we show that with active learning, we can actually reduce the number of experiments needed to determine the result.

1.3 Organization of the Thesis

This thesis is divided into five parts. Following this introductory part, which presents background information on the algorithms we use and why we used them. It also describes why we have choosen active learning and how we are going to implement it using our chosen algorithms.

Chapter 2 In this Chapter contains information regarding the concept of Active Learning and all the Related Works of the thesis.

Chapter 3 Describes Ensemble Models in details that is, discusses about Random Forest and AdaBoost and also discusses on clustering. It also overviews our proposed method and how we have implemented it which is broadly covered in chapter 4.

Chapter 4 Contains all the information about the experiments we have carried out for proposing our model. It contains the datasets we have used. It contains details about our experimental setup and also experimental results.

Chapter 5 Present conclusions and future work.

Chapter 2

Related Work

2.1 Active Learning

In semi-supervised learning, labelling unlabelled data points is time consuming and costly process. Active learning is a process of labelling unlabelled data by querying the oracle/ expert and then builds a learning model using these labelled data [19]. Usually, a set of unlabelled instances is selected randomly from unlabelled big data and ask human expert to label them [20]. Fig. 2.1 shows the active learning process. In the last decades, several process of active learning has been introduced like: uncertainty sampling, query-by-committee, expected model modification, expected error or variance reduction and information gain.Incidents of maximum uncertainty were preferred for sampling. Based on the uncertainty measurement, sample uncertainty may be divided into two categories: maximum label entropy and minimum distance from decision limit. Query-by-committee is a process of training model based on the available labelled data.

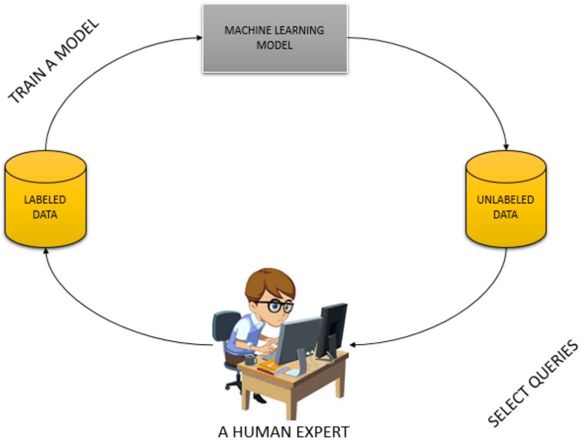

Figure 2.1: Active learning process in machine learning.

The committee is constructed either in one of these two ways: (1) sampling different models, and (2) applying ensemble learning (RandomForest, Bagging, and Boosting). All the models vote they're predictions on the unlabelled pool of data. The examples with maximum disagreements are chosen for labelling. Then the committee is then again retrained after including the new labelled instances. Expected model change selects the instances whose inclusion brings the maximum change in the learning model. Expected error reduction selects instances that reduce the expected generalisation error the most. Variance reduction selects training instances that reduce the model variance by most.

2.2 Semi-Supervised Learning

To understand the conception of Semi-Supervised Learning, we should initial perceive what's supervised and unsupervised learning. Supervised machine learning algorithms will apply what it's learned from the past to new data using instances that are tagged to predict the long run. By analyzing the training data to find out from it, the training algorithm comes up with a hard and fast function which is able to predict a result concerning the output values. the training algorithm can also—also will—can even—may also—may compare its output worth with the right output worth and if it finds a mismatch it can correct its error. And one in all the simplest things concerning this can be that the algorithm becomes terribly correct with adequate training. The process of learning a collection of rules from instances, that means the examples in an exceedingly training is what we call artificial machine learning. To be a lot of general, we've got to form a classifier that may be recycled to alter from new instances. Here, below we have a tendency to are showing the progression of applying supervised mil to a true world drawback,

The first and vital stage is gathering the dataset. If a vital expert is offered, then she/he will recommend that attributes or options are the foremost informative. If insufferable, then the best technique is that of "brute-force" that solely means that determinative everything offered so the informative and pertinent options are often isolated. However, a dataset gathered by "brute-force" technique isn't best or acceptable for induction. On a contrary it contains in most cases noise and missing feature values, and thus needs vital pre-processing [21]. The second stage is that the information readiness and information pre-handling. looking on issues, numerous analysts have assortment of ways to deal with settle on from to deal with missing data [22]. [23] have presented an investigation of ongoing strategies for exception (commotion) location. These scientists have perceived the methods' remunerations and inadequacies. Occasion decision isn't exclusively acclimated handle commotion anyway to adapt to the impracticableness of gaining from appallingly colossal datasets. there's an assorted variety of techniques for inspecting occasions from an outsized dataset [24]. The process of characteristic and removing as several extraneous and redundant options as attainable is understood as Feature set choice [25]. This lessens the spatial property of the information and empowers data processing algorithms to control quicker and a lot of expeditiously. The circumstance that a lot of options depend upon each other usually to a fault influences the accuracy of supervised mil classification models. this method is named feature construction/conversion. These new generated options could result in the development of a lot of summarizing and correct classifiers. what is more, the detection of silver options contribute to higher unambiguousness of the created classifier, and a more robust understanding of the learned conception.

Now Supervised learning can be split into 2 categories:

- Classification.
- Regression.

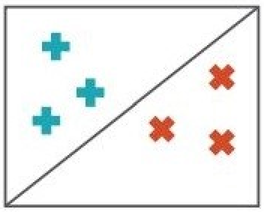

Figure 2.2: Classification.

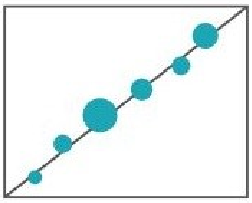

Figure 2.3: Regression.

Classification is a function of data mining to narrate and distinguish data classes or designs.The objective is to provide exact labels of classes in instances with known attribute values but unknown class values. The purpose of classification It is a form of data analysis that extracts (called classification) models that describe major data classes. It is a two-step process:

- Learning step (or training phase) where a classification model, classifier is constructed. By analysis of a training data set consisting of instances and their associated class labels a classification algorithm builds a classifier.
- Classification step where the classification.

Regression could be a technique of modeling a target worth supported independent predictors. This technique is wide used for statement and searching for cause and impact relationship between the two variables. Regression techniques principally dissent supported the amount of independent variables and therefore the sort of relationship between the freelance and dependent variables.Linear regression is probably one in all the foremost well-known and well understood algorithms in statistics and machine learning. The other foremost space of machine learning is that the unsupervised learning. The difference between supervised and unsupervised learning isn't entirely shrill, but the core a part of unsupervised learning is that we aren't given any solid data on however well we are performing arts. this can be in divergence to, say, classification wherever we are given manually labelled training data set. unsupervised learning incorporates clump (where we try and notice teams of data instances that are kind of like every other) and model building (where we have a tendency to try and build a model of our domain from our data).Unsupervised learning is wherever we don't have any output variable however solely have one input variable. The goal for unsupervised learning is to model the underlying structure or distribution within the data so as to find out a lot of concerning the data. there's no sure or specific answers and there's no teacher. it's left fully on the algorithm on the way to present the data.

One of the ways of Unsupervised learning is Clustering. Clustering: A clustering issue is wherever you might want to get the innate groupings inside the information, such as gathering clients by completing behavior.It is along these lines by separating the populace or information focuses into assortment of groups such information focuses inside similar gatherings are a ton of sort of like elective information focuses inside a similar group and not at all like the data focuses in elective groups. it's basically a lot of articles on the possibility of likeness and distinction between them. So by joining these two methodologies we tend to get through a cross breed approach as what's called Semi-administered Learning.

Semi-supervised learning algorithms are prepared on a blend of named and unlabelled information. this can be useful for a few reasons. To start with, the strategy for naming gigantic measures of information for regulated learning is normally restrictively long and overrated. What's a great deal of, an inordinate measure of marking will force human predispositions on the model.which means as well as plenty of unlabeled knowledge throughout the coaching method really tends to enhance the accuracy of the ultimate model whereas reducing the time and price spent building it.

For that reason, semi-supervised learning could be a win-win to be used cases like webpage classification, speech recognition, or maybe for genetic sequencing.In semi-supervised learning, labeled data is employed to assist determine that there are specific groups of webpage varieties gift within the knowledge and what they could be. The algorithm is then trained on unlabelled knowledge to outline the boundaries of these webpage varieties and will even determine new kinds of webpages that were any old within the existing human-inputted labels.

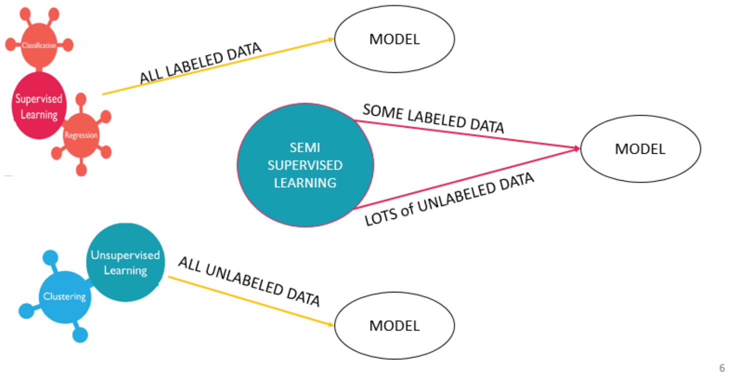

Figure 2.4: Semi-Supervised Learning.

2.3 Big Data

Big Data is a term used to describe a huge but exponentially expanded data collection with time. Big data is so large and complex that neither traditional data management tools can store it or efficiently process it [13]. Big data are some examples. Business figures from the New York Stock Exchange represent approximately one terabyte of new business information a day, and the statistics of social media show that every day more than 500 terabytes of new data are entered into Facebook's social media databases. Mainly photo and video uploads, message exchanges, commentary, etc. are generated in this data. There are three different types of large data: structured, unstructured and semi-structured. Any data stored, accessed and processed in a fixed format is known as *structured data* information. Over the course of time, talented computers have been able to develop techniques to work with and derive value from these data.(where the format is well known in advance). Today, however, we are anticipating problems when the size of such data is increasing to an enormous degree, typical sizes of several zettabytes. The *Employee* table in a database is an example of structured data. In contrast, unstructured data are any unknown format data or structure is unstructured. Data are not structured. Besides the large size of the un structured data, its treatment for deriving value from it also presents several challenges. The heterogeneous source of information consisting of simple text files, photos, videos, etc. is one typical example of unstructuring data. Nowadays, companies have a great deal of information available, but unfortunately, because this is raw or unstructured, they don't know how to derive value. Both forms of data can be found in semi-structured data. Semi-structured data may be considered as structured data, but is not defined in relation DBMS, for instance, by table definition. The data represented in an XML file is an example of semi-structured data. Big data are a transparency infrastructure for the manufacturing industry that is able to detect uncertainties such as incoherent performance and availability of components.The conceptual framework for predictive production begins

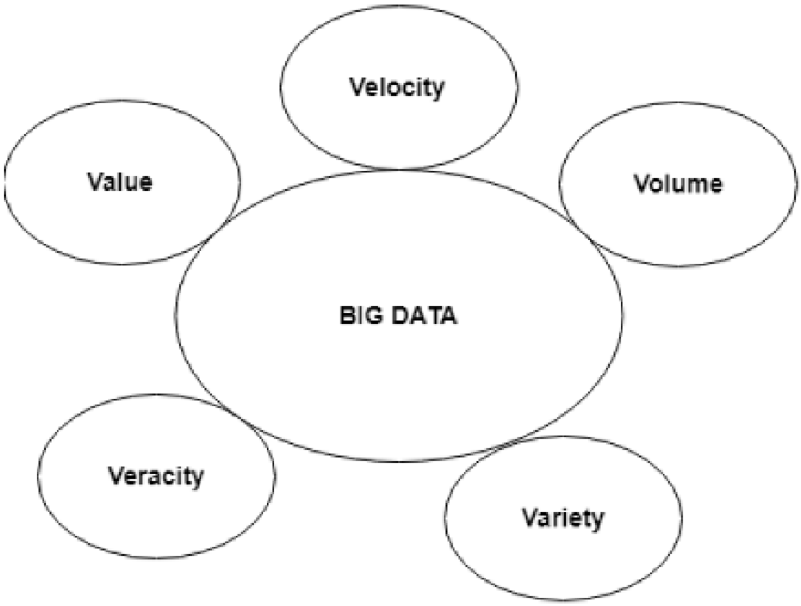

Figure 2.5: Big Data.

with data acquisition in big data applications, where various sensor data, like pressure, vibration, acoustics, voltage, current and control can be obtained.The combination of sensory and historical information forms the basis for production Big data. Many companies use Big Data, but may not have the basic assets from a security perspective, especially for marketing and research. In the event of a breach of safety of big data, the legal impact and reputational damage would be even more serious than at present. Many companies use the technology in this new era to store and analyze data petabytes concerning their company, business and customers. This makes it even more critical to classify the information.

Chapter 3

Proposed Method

In active learning, the classifier investigates the raw unlabelled data and requests labels from an oracle/ expert. The oracle then labels these raw data, which the classifier can use to improve its performance. But, this process is cost effective if the size of the data is small but becomes very time consuming when the size of the data becomes relatively large/ big unlabelled data. So the method that we have proposed deals with the problem associated with big unlabelled data. In our method, with the help of ensemble clustering such as component clustering, we divide the big data into a number of clusters/ groups. Then from the clusters, we choose the most informative instances. We pick the data that are closest to the centroids and the data instances, which are on the boundary of the clusters. Now that we have reduced the amount of data instances for processing, we then make the oracle manually handle this data instances and label them. Then we run ensemble classifiers with the newly labelled data and produce an output. This method of ours not only reduces the heavy workload that comes with handling large chunks of unlabelled data, but also it makes it much more efficient, faster and reliable. For the research purpose, we have proposed two methods and both of them are given below:

- Identifying less number of unlabelled instances using clustering approach.
- Applying active learning with ensemble classifiers.

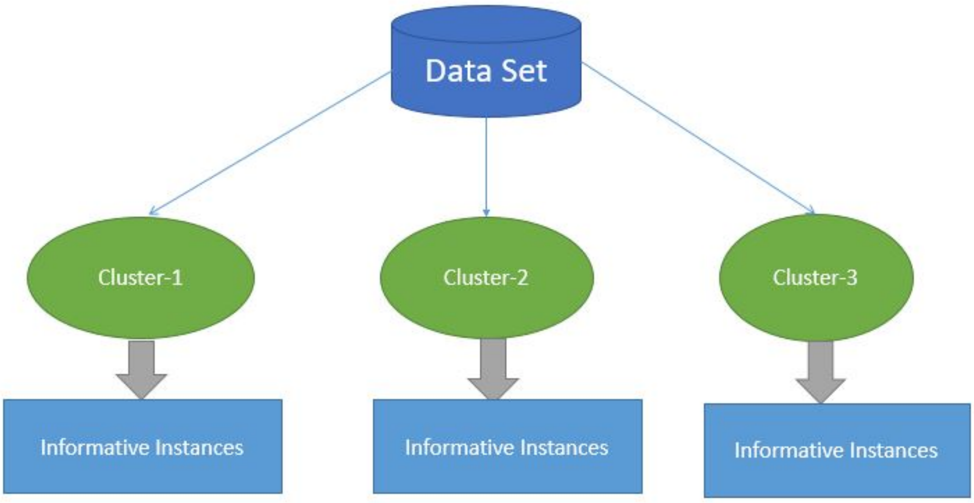

Figure 3.1: Clustering.

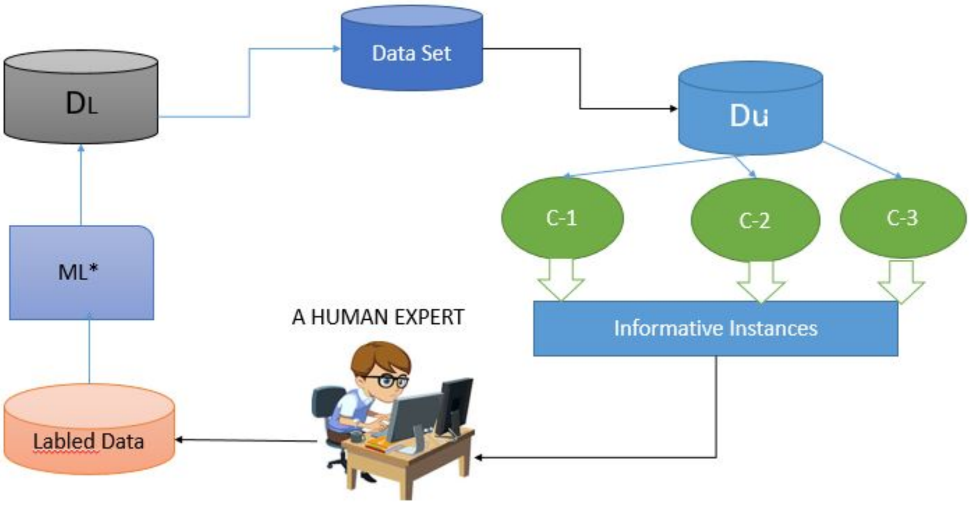

Figure 3.2: Clustering in active learning.

3.1 Algorithms

We have been used eight algorithms for research and they are :

- Naive Bayes Algorithm.
- Sequential Minimal Optimization (SMO).
- J48 Decision Tree.
- Random Forest.
- Bagging.
- CART.
- AdaBoost.

A brief description of the Algorithms that we used:

3.1.1 Naive Bayes Algorithm

The Bayesian Classification constitutes both a supervised learning method and a statistical classification method.Assumes a probabilistic model underlying and enables us to understand uncertainties about the model in a principled manner through the probability of results. This classification is named after Thomas Bayes (1702-1761), who proposed the Bayes theorem It can solve diagnostic and predictive issues.

Usage of NB :
There are several usages of this algorithm like,
i. Probabilistic learning method (Naive Bayes text classification).Classification devices from Naive Bayes are one of the best known classifying algorithms for text documents.
ii. The most famous use of the Bavarian Naive text classification is spam filtering. It uses a naive Bayes classifier for spam e-mail identification.

3.1.2 Sequential Minimal Optimization(SMO)

John C. Platt suggested the SMO algorithm in 1998 and has become the fastest quadratic optimization algorithm for programming, in particular for linear SVM and low data performance. The optimization problem is usually solved by[26] SMO. It is used for training SVM and the tool Iibsvm is used. It categories and handles a problem accordingly into many subproblems. Two multipliers ai, anda2 can be represented as,
0:Sal,a2:Sc
YI al + Y2 a2=k

Procedure:

- This requirement is the minimum number of multipliers that can be optimised at each stage is two: when one multiplier is updated, there should be a minimum of one additional multiplier to keep the condition true. .
- At each step, SMO selects two elements, αi and αj to optimize jointly, find the perfect values for those parameters, as all other elements are fixed, and updates the α vector accordingly.
- A heuristic thing is the choice of both points while analytically optimizing the two multipliers..
- Although more iterations are needed to converge, the algorithm uses as few operations as possible so that some magnitude orders can accelerate overall.

3.1.3 J48 Decision Tree

The C4.5 algorithm is used in Weka as the classification known as J48 in building decision trees. Classifiers, like filters, are hierarchically organized: J48 has weka.classifiers.trees.J48 full name. In the text box next to the button Choose: Read J48–C 0.25 –M 2. Classifier is presented. The default settings for this classification are set in this text.

Procedure:

In view of the univariate decision tree sample data, three types of approaches are available:

- Construction:First, check if all cases are classified, then the tree is a blade that is tagged with that class. Calculate the information and data gain for each attribute.
- In this process, the information gain is counted: "Entropy" is used. Entropy is a data disorder measurement. Bits, nats or bans are measured for the entropy. The uncertainty measurement of any random variable is also referred to. Suppose there is a fair coin, if there is a single jerk on that coin than the entropy. A number of two fair coins will have two-bit entropy. Now if the coin is unfair and this results in a lower rate of entropy.
- Tree Pruning: Pruning is a technique which is very important for the creation of tree because of the outliers. The overfitting also applies. Datasets can include small, unspecified subsets of instances. Tailing may be used to classify them correctly. Two types of cutting are available:

  1. Post pruning (carried out after tree creation) .
  2. Online pruning (Conducted during tree creation).

3.1.4 Random Forest

Random Forest is a flexible machine study algorithm that is easy to use and produces great results most of the time, even without hyperparameters tuning.It is also one of the most frequently used algorithms, because it is simple and can be used for classification and regression tasks.In this article, you will learn about the function of the random wood algorithm and a number of other things.

Procedure:
Random Forest is a supervised learning algorithm. It creates a forest and makes it random somehow, as is seen from its name. The forest it constructs is a set of decision trees, usually formed by the ”bagging” method.The general idea of the bagging method is that the overall result is increased by a combination of learning models.One major advantage of random forests is the ability to make the most updated machine learning systems, both for classification and regression problems.I will speak about random forests as classification is occasionally considered to be the machine learning building block.

3.1.5 Bagging

Bagging meaning Bootstrap Aggregation. Bagging is an ensembling process – where a model is trained on each of the bootstrap samples and the final model is an aggregated model of the all sample models. For a numeric target variable /regression problem the predicted outcome is an average of all the models and in the classification problems, the predicted class is defined based on plurality.

3.1.6 CART

For the development of both classification and regression trees, both CART algorithms can be used. The impurity (or purity) measure used in building decision tree in CART is Gini Index. The CART algorithm decision tree is always a binary decision tree (each node will have only two child nodes). where i and j are target variables.

3.1.7 AdaBoost

AdaBoost, Adaptive Boosting short form. ML meta-algorithm found by Yoav Freund and Robert Schapire is AdaBoost. They were awarded the Gödel Prize for their work in 2003. AdaBoost, the number one algorithm for binary classification, was actually successful. AdaBoost is used for improving the performance of decision books concerning binary classification problems. AdaBoost can be cast off to increase the performance of all machine learning algorithms. It’s best used for weak students. These models are precision above random possibilities for a classification problem. The most suitable and therefore the most popular AdaBoost algorithm are one-level decision trees. Because they are so short and contain only one classification decision, they are often referred to as decision-making stumps.

Each example is weighted in the training dataset. The weight of the initial is: Weight(xi) = 1/n When xi is the i-th instance of training , n is the number of training

Chapter 4

Experimental Analysis

4.1 Experiment

This section presents datasets, experimental setup and experimental results.

4.1.1 Dataset Descriptions

We have collected following six benchmark datasets from UCI machine learning repository [27] that are shown in Table 4.1.

Table 4.1: Dataset details.

Dataset	Instance	Attribute	Characteristics	Area
Breast-Cancer	286	10	Multivariated	Life
Soyabean	683	36	Multivariated	Life
Glass	214	10	Multivariated	Physical
German-Credit	1000	21	Multivariated	Financial
Vote	435	17	Multivariated	Social
HypoThyroid	3772	30	Multivariated	Life

4.1.2 Experimental Setup

We have used accuracy, precision, recall, and F-score that are shown in Eqs. 4.1 to 4.4 where TP, TN, FP and FN are true positive, true negative, false positive, and false negative respectively with 10-fold cross validation to evaluate the learning algorithms.

$$accuracy = \frac{\sum_{i=1}^{|X|} assess(x_i)}{|X|}, x_i \in X \tag{4.1}$$

$$precision = \frac{TP}{TP + FP} \tag{4.2}$$

$$recall = \frac{TP}{TP + FN} \tag{4.3}$$

$$F - score = \frac{2 \times precision \times recall}{precision + recall} \tag{4.4}$$

4.1.3 Experimental Results

For experimental results, we have considered the weighted average for precision, recall and F- score analysis for each dataset. The detailed results are presented in Tables 4.2 to 4.9. In Table 4.2, for the breast cancer dataset we found the performance of SMO is maximum among all the classifiers. In SMO, previous accuracy was 69.58% and using our proposed method the new accuracy is 75.29%, old precision was 0.75 and new is 0.78 , Recall was 0.85 and new is 0.92 and F-score was 0.79 and new is 0.84. In Table 4.3, for Glass dataset the performance of SMO is also maximum, the accuracy increases 40% for SMO. In SMO previous accuracy was 56.07% and new accuracy is 96.62%, precision was 0.59 and new is 1.0, recall was 0.56 and f-score was 0.54 and new is 0.84.

Table 4.2: Results on Breast-Cancer dataset using 10-fold cross validation on 286 instances and 83 instances.

Algorithm	Accuracy (%) old	Accuracy (%) new	Precision old	Precision new	Recall old	Recall new	F-Score old	F-Score new
Naïve Bayes	71.67	71.76	0.77	0.76	0.83	0.9	0.8	0.82
J48	75.52	74.11	0.75	0.74	0.96	0.88	0.84	0.81
CART	69.23	73	0.71	0.73	0.93	0.98	0.81	0.84
SMO	69.58	75.29	0.75	0.78	0.85	0.92	0.79	0.84
Random Forest	69.8	69	0.74	0.87	0.9	0.9	0.8	0.81
Bagging	69.23	70.58	0.72	0.73	0.91	0.95	0.8	0.82
AdaBoost	70.27	70.41	0.77	0.71	0.82	0.88	0.79	0.84

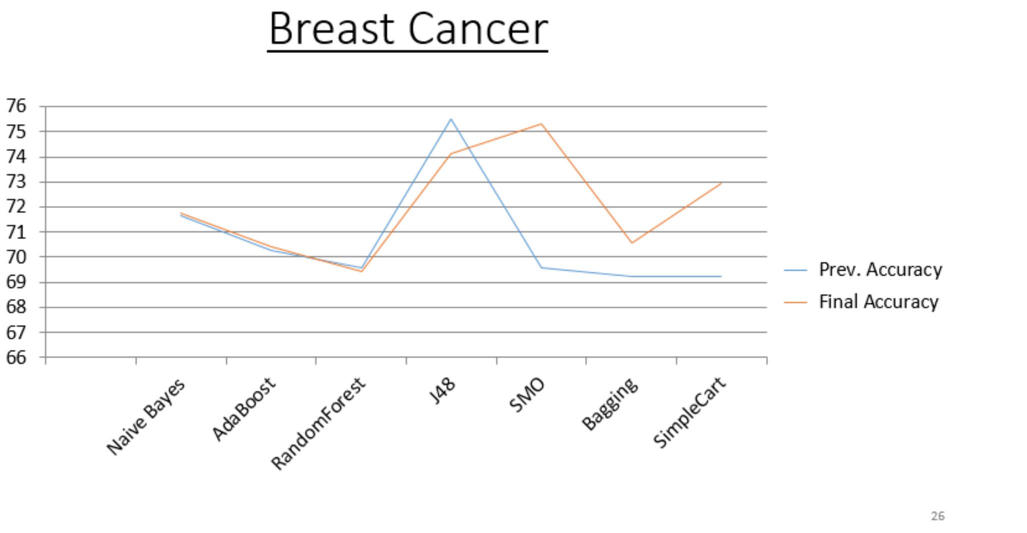

Figure 4.1: Result of Breast Cancer Dataset.

Table 4.3: Results on Glass dataset using 10-fold cross validation on 214 instances and 32 instances.

Algorithm	Accuracy (%) old	Accuracy (%) new	Precision old	Precision new	Recall old	Recall new	F-Score old	F-Score new
Naïve Bayes	48.59	81.25	0.45	0.66	0.72	0.66	0.56	0.66
J48	72.91	87.55	0.73	0.75	0.98	0.66	0.82	0.81
CART	70.56	71.81	0.00	0.00	0.71	0.00	0.74	0.00
SMO	56.07	96.62	0.59	1.00	0.56	0.66	0.54	0.84
Random Forest	79.90	80.25	0.78	0.57	0.87	0.33	0.82	0.44
Bagging	72.48	71.87	0.72	0.82	0.82	0.83	0.77	0.87
AdaBoost	44.88	75.31	0.45	0.01	1.00	0.00	0.62	0.00

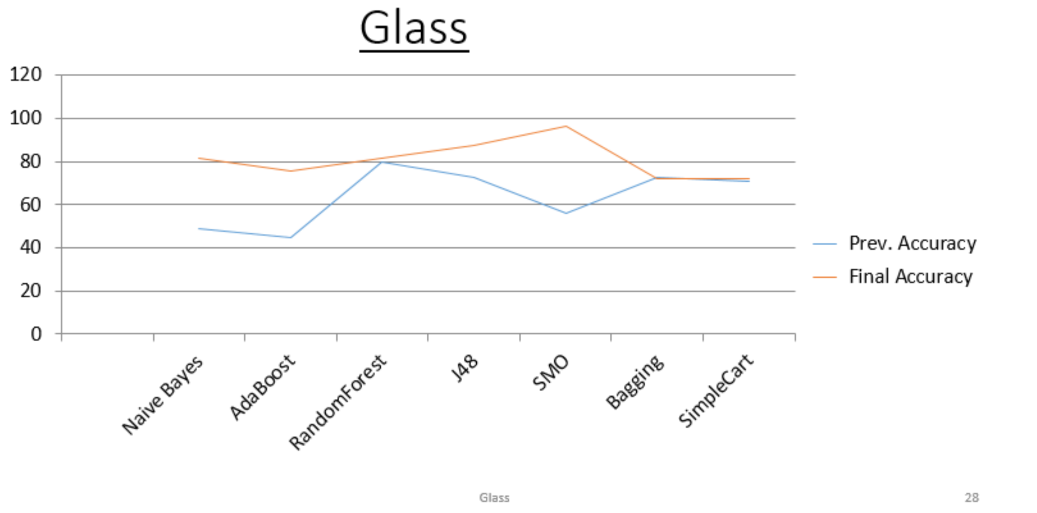

Figure 4.2: Result of Glass Dataset

In Table 4.4, for German credit Dataset we choose 257 instances after clustering. But, the performance of precision, recall and f-score are not increased from the old results. Then we increased instances in Table 4.5 from 257 to 410. By using minimum number of instances we found the maximum accuracy for RandomForest, the accuracy was 76.41% and new accuracy is 80.12%. In Table 4.6, Soybean datasets initially we found on 23 instances but again performance of classifier was not acceptable. So, after increasing instances 225 in Table 4.7, the accuracy for the both J48 and Bagging is increased. For J48 accuracy: old 91.50% and new 95.26% and for Bagging: old 85.65% and new 88.86%. But Precision , recall and f-score increased maximum for j48. In Table 4.8, vote dataset the performance of Bagging is maximum, accuracy was 95.63%, but new accuracy 96.11%. In Table 4.9, Hypothyroid dataset the maximum performance is shown by J48, CART, RandomForest, and Bagging.

Table 4.4: Results on German Credit dataset using 10-fold cross validation on 1000 instances and 257 instances.

Algorithm	Accuracy (%) old	Accuracy (%) new	Precision old	Precision new	Recall old	Recall new	F-Score old	F-Score new
Naive Bayes	75.44	76.72	0.80	0.77	0.86	0.84	0.83	0.81
J48	70.51	72.41	0.78	0.74	0.97	0.94	0.84	0.81
CART	73.92	65.75	0.78	0.72	0.88	0.87	0.83	0.79
SMO	75.11	75.92	0.79	0.78	0.87	0.86	0.83	0.82
Random Forest	76.40	70.45	0.78	0.74	0.97	0.94	0.84	0.81
Bagging	74.70	67.70	0.78	0.72	0.88	0.87	0.83	0.79
AdaBoost	69.50	71.59	0.73	0.74	0.87	0.89	0.80	0.81

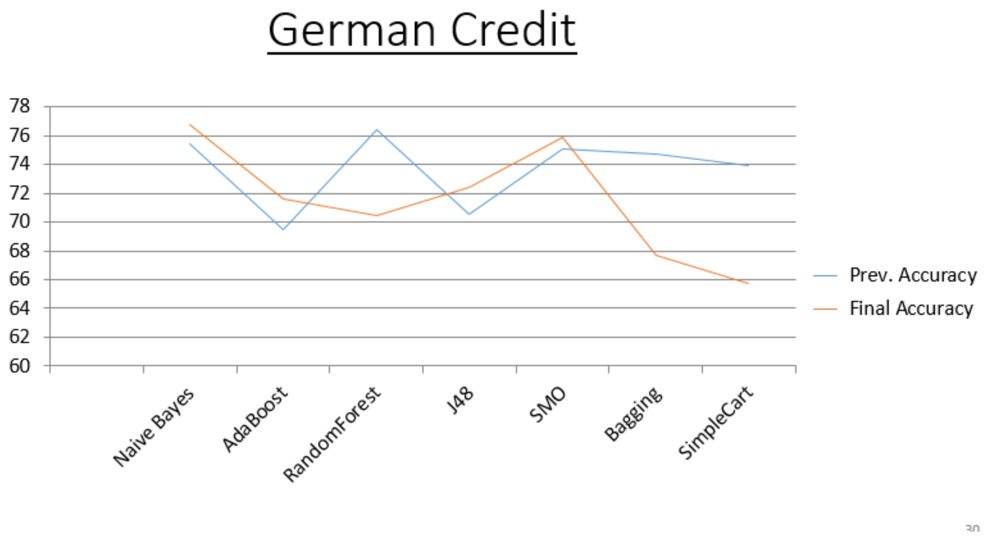

Figure 4.3: Result of German credit Dataset.

Table 4.5: Results on German Credit dataset using 10-fold cross validation on 1000 instances and 410 instances.

Algorithm	Accuracy (%) old	Accuracy (%) new	Precision old	Precision new	Recall old	Recall new	F-Score old	F-Score new
CART	73.92	74.28	0.78	0.75	0.88	0.90	0.83	0.88
RANDOM FOREST	76.41	80.12	0.78	0.79	0.97	0.98	0.84	0.81
Bagging	74.70	76.20	0.78	0.78	0.88	0.88	0.83	0.89

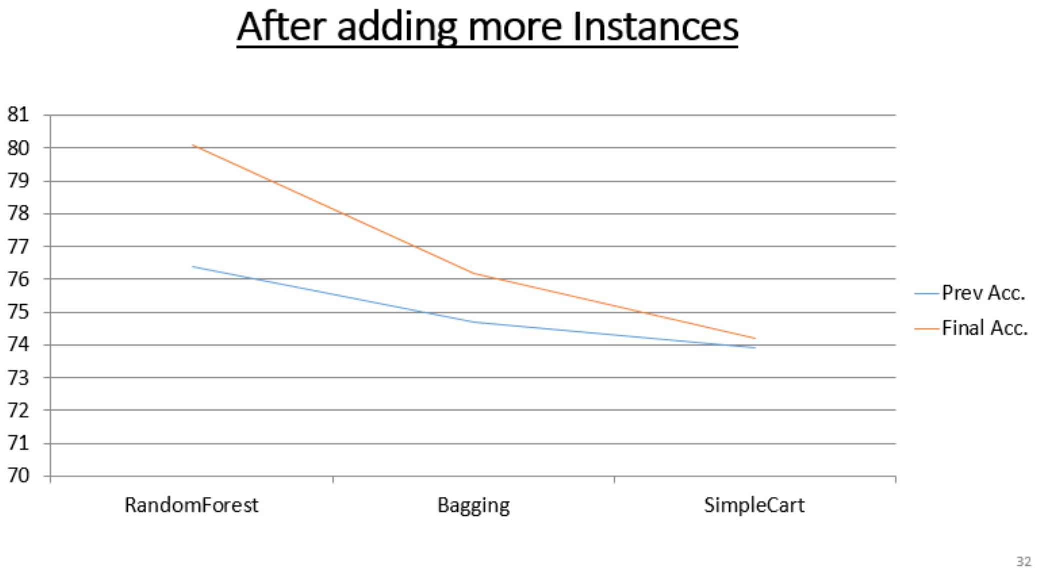

Figure 4.4: Results of German Credit dataset.

Table 4.6: Results on Soyabean dataset using 10-fold cross validation on 683 instances and 23 instances.

Algorithm	Accuracy (%) old	Accuracy (%) new	Precision old	Precision new	Recall old	Recall new	F-Score old	F-Score new
Naive Bayes	98.26	78.26	1.00	0.00	1.00	0.00	1.00	0.00
J48	91.50	78.26	0.95	0.00	0.95	0.00	0.95	0.00
CART	91.06	65.21	1.00	0.00	1.00	0.00	1.00	0.00
SMO	93.85	94.30	1.00	0.00	1.00	0.00	1.00	0.00
Random Forest	92.99	78.26	1.00	0.00	1.00	0.00	1.00	0.00
Bagging	85.65	60.86	0.95	0.00	1.00	0.00	0.94	0.00
AdaBoost	27.96	47.86	1.00	0.00	1.00	0.00	1.00	0.00

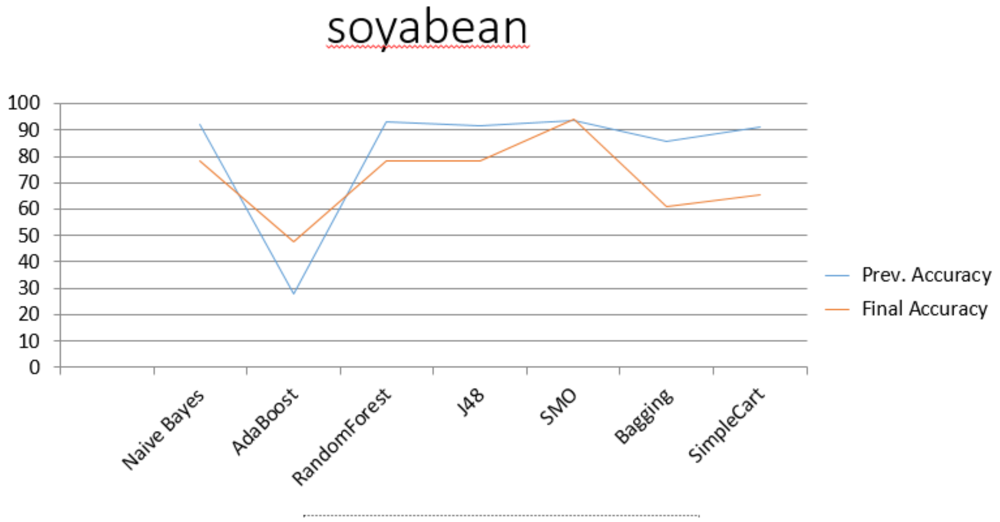

Figure 4.5: Results of Soyabean Dataset.

Table 4.7: Results on Soyabean dataset using 10-fold cross validation on 683 instances and 225 instances.

Algorithm	Accuracy (%) old	Accuracy (%) new	Precision old	Precision new	Recall old	Recall new	F-Score old	F-Score new
Naive Bayes	92.26	92.62	1.00	0.00	1.00	0.00	1.00	0.00
J48	91.50	95.26	0.95	0.00	0.95	1.00	0.95	1.00
CART	91.06	91.21	1.00	1.00	1.00	0.00	1.00	0.00
Random Forest	92.99	93.26	1.00	1.00	1.00	0.00	1.00	1.00
Bagging	85.65	88.86	0.95	0.00	1.00	0.00	0.94	0.00

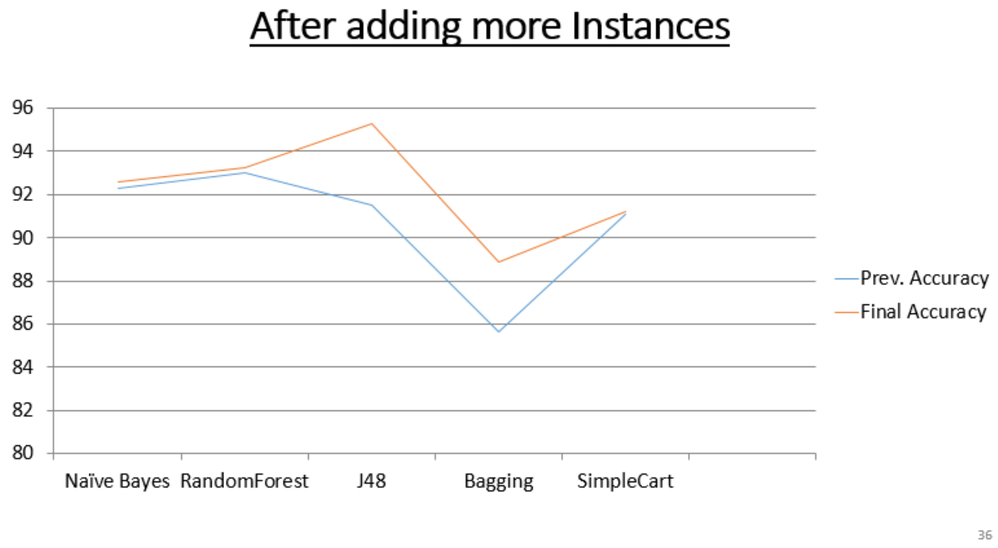

Figure 4.6: Result of Soyabean Dataset.

Table 4.8: Results on VOTE dataset using 10-fold cross validation on 435 instances and 80 instances.

Algorithm	Accuracy (%) old	Accuracy (%) new	Precision old	Precision new	Recall old	Recall new	F-Score old	F-Score new
Naive Bayes	90.11	90.29	0.94	0.95	0.89	0.83	0.91	0.89
J48	96.32	96.11	0.97	0.91	0.97	1.00	0.97	0.95
CART	95.40	96.11	0.97	1.00	0.95	0.91	0.96	0.95
SMO	96.09	96.11	0.97	1.00	0.96	0.91	0.96	0.95
Random Forest	96.92	97.08	0.96	1.00	0.97	0.93	0.96	0.96
Bagging	95.63	96.11	0.97	1.00	0.95	0.91	0.96	0.95
AdaBoost	95.40	95.14	0.97	0.97	0.95	0.91	0.96	0.94

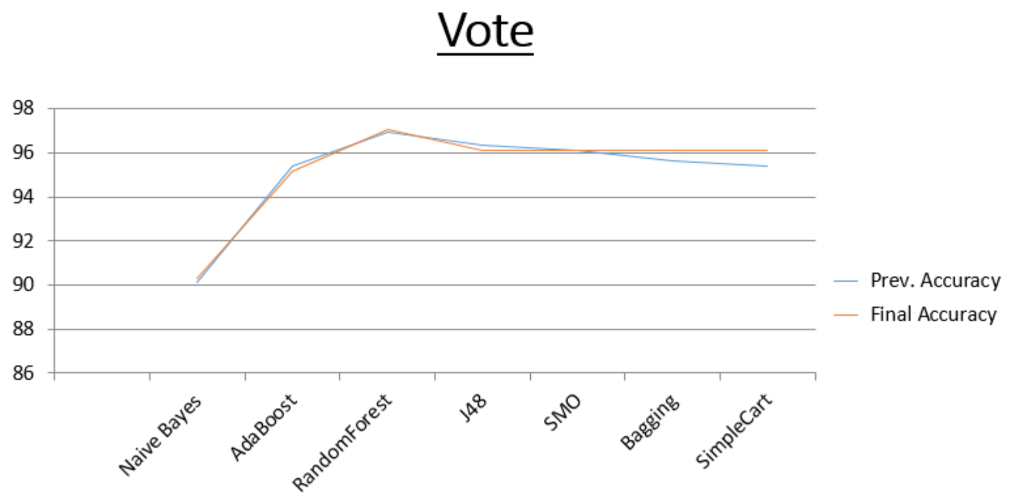

Figure 4.7: Results of VOTE Dataset.

Table 4.9: Results on Hypothyroid dataset using 10-fold cross validation on 3772 instances and 770 instances.

Algorithm	Accuracy (%) old	Accuracy (%) new	Precision old	Precision new	Recall old	Recall new	F-Score old	F-Score new
Naive Bayes	94.96	96.64	0.95	0.97	0.99	0.99	0.97	0.98
J48	99.35	99.64	0.99	0.99	0.99	0.99	0.99	0.99
CART	99.39	99.64	1.00	0.99	0.97	0.99	0.99	0.99
SMO	92.95	94.78	0.93	0.94	0.99	1.00	0.96	0.97
Random Forest	99.28	99.64	0.99	0.99	0.99	0.99	0.99	0.99
Bagging	99.46	99.64	1.00	0.99	0.99	0.99	0.99	0.99
AdaBoost	94.24	94.34	0.97	0.99	0.98	0.97	0.98	0.98

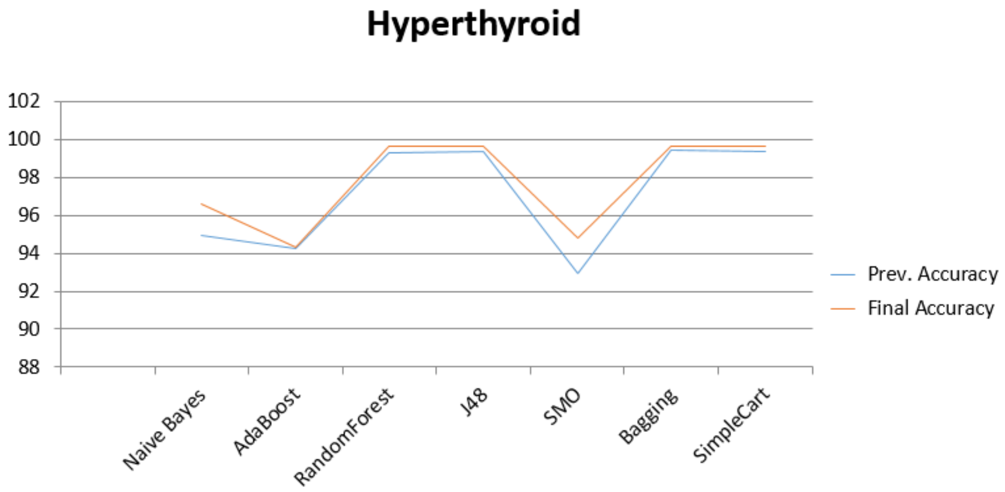

Figure 4.8: Results of Hyperthyroid Dataset.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

In this paper, we have presented a cluster-based method for selecting informative unlabelled instances in the process of active learning. We have applied seven popular machine learning classifiers (e.g. naïve Bayes classifier, J48, CART, SMO, RandomForest, Bagging and AdaBoost) and increased their performance employing our proposed method. The proposed method generates learning models with better accuracy even with less number of instances that are small chunks of data as we consider them as informative instances. Since unlabelled big data is undoubtedly tough to deal with even for a human expert, we have selected informative instances from the centroids and from the boundary of the clusters after clustering the data. So, the expert can label the less number of unlabelled data. In future, we will apply the proposed method for mining real-life semi-supervised traffic big data.

5.2 Future Work

In future,we will work with real life Big Datasets.This time we choose only center oriented data but in future we will work with border oriented data also.

Bibliography

[1] A. L'heureux, K. Grolinger, H. F. Elyamany, and M. A. M. Capretz, "Machine learning with big data: Challenges and approaches," *IEEE Access*, vol. 5, pp. 7776–7797, 2017. 1

[2] N. Garg, S. Singla, and S. Jangra, "Challenges and techniques for testing of big data," *Procedia Computer Science*, vol. 85, pp. 940–948, 2016. 1

[3] H. Özköse, E. S. Arı, and C. Gencer, "Yesterday, today and tomorrow of big data," *Procedia-Social and Behavioral Sciences*, vol. 195, pp. 1042–1050, 2015. 1

[4] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE Transactions on Knowledge & Data Engineering*, vol. 26, no. 1, pp. 97–107, 2014. 1

[5] D. M. Farid, A. Nowé, and B. Manderick, "An ensemble clustering for mining high-dimensional biological big data," *International Journal of Design & Nature and Ecodynamics*, vol. 11, pp. 328–337, 2016. 1

[6] D. M. Farid, M. A. Al-Mamun, B. Manderick, and A. Nowé, "An adaptive rule-based classifier for mining big biological data," *Expert Systems with Applications*, vol. 64, pp. 305–316, December 2016. 1

[7] V. López, S. del Río, J. M. Benítez, and F. Herrera, "Cost-sensitive linguistic fuzzy rule based classification systems under the mapreduce framework for imbalanced big data," *Fuzzy Sets and Systems*, vol. 258, pp. 5–38, 2015. 1

[8] A. Elragal, "Erp and big data: The inept couple," *Procedia Technology 16 (2014) 242 – 249 Procedia Technology 16 (2014) 242 – 249 Procedia Technology 16 (2014) 242 – 249 Procedia Technology*, vol. 16, pp. 242–249, 2014. 1

[9] W. Fan and A. Bifet, "Mining big data: Current status, and forecast to the future," *ACM SIGKDD Explorations Newsletter*, vol. 14, no. 2, pp. 1–5, 2013. 1

[10] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. U. Khan, "The rise of "big data" on cloud computing: Review and open research issues," *Information Systems*, vol. 47, pp. 98–115, 2015. 1

[11] F. Rayhan, S. Ahmed, S. Shatabda, D. M. Farid, Z. Mousavian, A. Dehzangi, and M. S. Rhaman, "iDTI-ESBoost: identification of drug target interaction using evolutionary and structural features with boosting," *Scientific Reports*, vol. 7, 2017. 1

[12] F. Rayhan, S. Ahmed, D. M. Farid, A. Dehzangi, and S. Shatabda, "CFSBoost: cumulative feature subspace boosting for drug-target interaction prediction," *Journal of Theoretical Biology, Elsevier*, vol. 464, no. 2-3, pp. 1–8, 2019. 1

[13] D. M. Farid, A. Nowé, and B. Manderick, "An ensemble clustering for mining high-dimensional biological big data," in *International Conference on Big Data*, Alicante, Spain, May 2016, pp. 131–142. 2, 9

[14] ——, "A feature grouping method for ensemble clustering of high-dimensional genomic big data," in *(FTC)*, San Francisco, United States, December 2016, pp. 260–268. 2

[15] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml 2

[16] D. M. Farid, A. Nowé, and B. Manderick, *Ensemble of Trees for Classifying High-Dimensional Imbalanced Genomic Data*, ser. Lecture Notes in Networks and Systems, R. B. e. In: Y. Bi, S. Kapoor, Ed. Springer, Cham, 2017, vol. 15. 2

[17] D. M. Farid, L. Zhang, C. M. Rahman, and M. H. R. Strachan, "Hybrid decision tree and naïve bayes classifiers for multi-class classification tasks," *Expert Systems with Applications*, vol. 41, pp. 1937–1946, March 2014.

[18] D. M. Farid, L. Zhang, A. Hossain, C. M. Rahman, R. Strachan, G. Sexton, and K. Dahal, "An adaptive ensemble classifier for mining concept drifting data streams," *Expert Systems with Applications*, vol. 40, pp. 5895–5906, November 2013. 2

[19] D. M. Farid, A. Nowé, and B. Manderick, "Combining boosting and active learning for mining multi-class genomic data," in *25th Belgian-Dutch Conference on Machine Learning (Benelearn)*, Kortrijk, Belgium, September 2016, pp. 1–2. 5

[20] S. Jahan, S. Shatabda, and D. M. Farid, "Active learning for mining big data," in *21st International Conference on Computer and Information Technology (ICCIT)*, Dhaka, Bangladesh, December 2018, pp. 1–6. 5

[21] Z. Zhang, L. Zhang, S. Croll, and M. Chopp, "Angiopoietin-1 reduces cerebral blood vessel leakage and ischemic lesion volume after focal cerebral embolic ischemia in mice," *Neuroscience*, vol. 113, no. 3, pp. 683–687, 2002. 6

[22] G. E. Batista and M. C. Monard, "An analysis of four missing data treatment methods for supervised learning," *Applied artificial intelligence*, vol. 17, no. 5-6, pp. 519–533, 2003. 6

[23] V. Hodge and J. Austin, “A survey of outlier detection methodologies,” *Artificial intelligence review*, vol. 22, no. 2, pp. 85–126, 2004. 6

[24] W. Reinartz and V. Kumar, “The mismanagement of customer loyalty,” *Harvard business review*, vol. 80, no. 7, pp. 86–95, 2002. 6

[25] L. Yu and H. Liu, “Efficient feature selection via analysis of relevance and redundancy,” *Journal of machine learning research*, vol. 5, no. Oct, pp. 1205–1224, 2004. 6

[26] G. Nadiammai and M. Hemalatha, “Perspective analysis of machine learning algorithms for detecting network intrusions,” in *Computing Communication & Networking Technologies (ICCCNT), 2012 Third International Conference on.* IEEE, 2012, pp. 1–7. 13

[27] D. Dheeru and E. K. Taniskidou, “Uci machine learning repository (2017),” *URL http://archive. ics. uci. edu/ml*, 2017. 17