Active Learning for Mining Big Data

Sadia Jahan (ID: 012162006)

Department of Computer Science and Engineering

United International University

A thesis submitted for the degree of

MSc in Computer Science & Engineering

September 2019

Abstract

Active learning also known as an optimal experimental design, is a process for building a classifier or learning model with less number of training instances in the semi-supervised setting. It's a well-known approach that is used in many real-life machine learning and data mining applications. Active learning uses a query function and an oracle or expert (e.g., a human or information source) for labeling unlabeled data instances to boost up the performance of a classifier. Labeling the unlabeled data instances is difficult, time-consuming, and expensive. In this paper, we have proposed an approach based on cluster analysis for selecting informative training instances from large number of unlabeled data instances or big data that helps us to select less number of training instances to build a classifier suitable for active learning. The proposed method clusters the unlabeled big data into several clusters and find the informative instances from each cluster based on the center of the cluster, nearest neighbors of the center of the cluster, and also selecting random instances from each cluster. The objective is to find the informative unlabeled instances and label them by the oracle for scaling up the classification results of the machine learning algorithms to be applied on big data. We have tested the performance of the proposed method on seven benchmark datasets from UC Irvine Machine Learning Repository employing following five well-known machine learning algorithms: C4.5 (decision tree induction), SVM (support vector machines), Random Forest, Bagging, and Boosting (AdaBoost). The experimental analysis proved that proposed method improves the performance of classifiers in active learning with less number of training instances.

Dedication

"The dissertation of my work is dedicated to my parents, my husband and my Supervisor for their support and inspiration. Special gratitude to my Supervisor, without his help and direction it would not have been possible to complete my work."

Published Papers

Work relating to the research presented in this thesis has been published by the author in the following peer-reviewed conference:

1. Sadia Jahan, Swakkhar Shatabda, and Dewan Md. Farid, "Active Learning for Mining Big Data," 21stInternational Conference on Computer and Information Technology (ICCIT 2018), 21-23 December 2018, Dhaka, Bangladesh, pp. 1-6, and IEEE Xplore Digital Archive, DOI: 1109/ICCITECHN.2018.8631973.

Acknowledgements

I would like to express my gratitude to all those who have helped me to complete my MSc in Computer Science and Engineering Thesis.

I am especially grateful to my supervisor Dr. Dewan Md. Farid, Associate Professor, Department of Computer Science and Engineering at United International University, for his continuous guidance, advice, effort and veritable suggestion throughout the research.

I am also grateful to Head Examiner Dr. Swakkhar Shatabda, Associate Professor, Department of Computer Science and Engineering at United International University, for his suggestion and valuable review for both conference and thesis book.

I would like to thank Dr. Mohammad Nurul Huda, Professor and MSCSE Director, Department of Computer Science and Engineering at United International University for reviewing my thesis book and his valuable comments.

I also like to thank Rubaiya Rahtin Khan, Assistant Professor, Department of Computer Science and Engineering at United International University for her valuable time to review my thesis book.

Lastly, I would like to express my sincere appreciation to all of them who have willingly helped me out with their abilities.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

This chapter will present an overview of the depiction of the problem with problem statements. We also discuss about the different research challenges what we are going to face in the whole scenario. The objectives and contributions of this thesis have been presented at the last part of this chapter. The chapter ends with the organization of the thesis.

1.1 Motivation

In the era of big data, data are always generated and already reached in a volume of more than 1000 Exabyte and is expected in next ten years it will increase 20-fold. Big data mining employing machine learning tools and techniques is the process of organising and analysing an immense volume, variety, and velocity of data for decision making [8, 10]. Everyday huge amounts of data are being generated from different sources like cloud, business management and various machines & devices. So this era is for data torrent where in many situations the term *Big Data* appears. But, most of the time essential data is inaccessible to user due to its incomplete and unstructured types [11, 12]. It is a tiring task to manage and arrange out large amount of data, which has to be automated as much as possible [13]. Machine learning algorithms take advantage of simplifying the task of automated digital objects classification [14, 15]. For training, the machine learning algorithms take a large amount of labeled data as acceptable classification accuracy is needed [16, 17]. Data labelling is time consuming and expensive task as it has to be done manually. In recent years, there has been

prime interest in improving accuracy of classification methods, which tries to exploit unlabelled data [18–21]. The semi-supervised learning methods have been applying to improve the accuracy of unlabelled data classification for last couple of decades. Unlabelled data are often enormous and it is a tedious work to obtain label data from unlabelled data, which is also costly and time consuming process. In terms of labelling unlabelled data, active learning is the process of classifying semi-supervised data in machine learning [7].

In supervised learning, for achieving accurate predictive model a bunch of labeled training instances are needed [22]. But, the performance of the learning models can be declined due to wrongly labeled/contradictory instances [23]. To label the unlabelled training data an exhaustive and rigorous analysis is needed. Active learning in Reinforcement learning for data mining is a good solution for addressing this problem [24]. An expert/oracle is used in the active learning process to label the unlabelled data that improve the classification accuracy by posing minimum queries to user/expert. The number of training instances to learn a concept in active learning process can often be much lower than the number required in normal supervised learning process. Though manually labelling of unlabelled big data is very challenging and expensive.

1.2 Objectives and Research Challenges

In this study, the main purpose is to collect informative instances as less as possible by maintaining the same accuracy of the total dataset through active learning process. Though many methods are established for collecting informative instances by active learning but most of them are for binary or only for one class dataset. Active learning works properly by using informative instances from unlabeled data for supervised big data which is more effective and informative for labeling those unlabeled instances. In addition, finding the most informative instances for semi-supervised big data is too much complex and time consuming. So, our main objectives are to establish a new method in active learning process for collecting informative instances from any types of datasets and for multiple classes. The main objectives of this thesis are as follows:

- Cluster unlabeled instances from big data (semi-supervised) in N number where N will be defined by user.

- Collect informative instances from each cluster based on random collection of the center of the cluster, nearest neighbor data point of the center of the cluster.
- Label those informative instances by the Oracle/ User.
- Scaling up the classification results.
- Boosting up the performance of existing machine learning algorithm.

So, considering the key objectives of this study, the following are the subsequent research challenges are addressed.

- How to work our framework effectively and efficiently?
- How to find the informative unlabeled instances and label from big data?
- How to apply active learning to use multi-class datasets to predict newly instance accuracy?
- What is the main advantage of applying active learning by using our framework for classification?
- Which machine learning techniques perform better in active learning algorithm?

1.3 Thesis Contributions

The main contributions of this research work are summarized as follows:

1. We synthesized the literature on various active learning techniques for big data analysis and proposed a method using cluster analysis to find informative training instances in active learning for mining semi-supervised big data.
2. We have used different supervised and semi-supervised benchmark data sets from UCI machine learning repository and KEEL (Knowledge Extraction based on Evolutionary Learning) dataset repository with multiple classes.
3. To collect informative instances from unstructured big data, we follow our two major objectives like random collection and collect the Nearest Neighbor point of the center of the cluster.

4. Proper training method for the classification is done using KNN (K-nearest neighbors), C4.5 (decision tree inductions), SVM (support vector machine), Random forest, Bagging and Boosting (AdaBoost) on our proposed framework.

5. The performance is evaluated with different machine learning techniques and our experimental outcomes served as an alternative source of knowledge to fill the gaps of the traditional big data reports and surveys.

1.4 Organization of the Thesis

The organization of this thesis is as follows:

Chapter 2 describes literature review of active learning and big data with challenges of active learning and big data analysis also with analytic of active learning and big data management.

Chapter 3 will present proposed active learning method for mining big data. Data collection, processing and selection from clustering. And different machine learning approaches like KNN, C4.5, SVM, Random Forest, Bagging, Boosting, Scaling, Data reducing on our proposed framework.

Chapter 4 will focus on experimental analysis with dataset description, experimental setup and experimental result.

Chapter 5 will discuss about our thesis and conclusion with future work.

Chapter 2

Related Work

We are living in an epoch where data are being generated at a torrent rate which is called big data. Those data are in a different format, structure, size, value, etc. For converting them into a correct format and classifying them we need machine learning. Active learning is a part of machine learning which can classify that huge data set. This chapter is a discussion of the analysis of big data with active learning, its challenges, big data analysis and resolution of active learning for big data management.

2.1 Big Data Analysis with Active Learning

2.1.1 Big Data

In our modern information technology industry data are generated through different sources like various devices, machines, web technologies, social media, mobile, genomics, meteorology, biological, complex physical simulation and environmental research, finance and business to health care, etc. Figure 2.1 shows that different size and speed of data with numerous values, structures, and qualities from several sources are generated big data.

As a result, the rate of data is blowing up at an unprecedented speed. For example, within 2020, more than 30 billion devices will be connected, as estimated by ABI Research. Twitter processes almost 8TB tweets daily and over 70M each day. The behavior of users, their feedback of online transaction these types of information are collecting and selling by most popular sites and social media like Amazon, Google, and

Figure 2.1: Sources of Big Data [1].

Facebook. Even different types of apps also collect and sell user information [8, 9, 25–29]. IDC (International Data Corporation) published a report in Figure 2.2, where they showed that from 2018 to 2025, 175 Zettabytes (ZB) digital data will be generated.

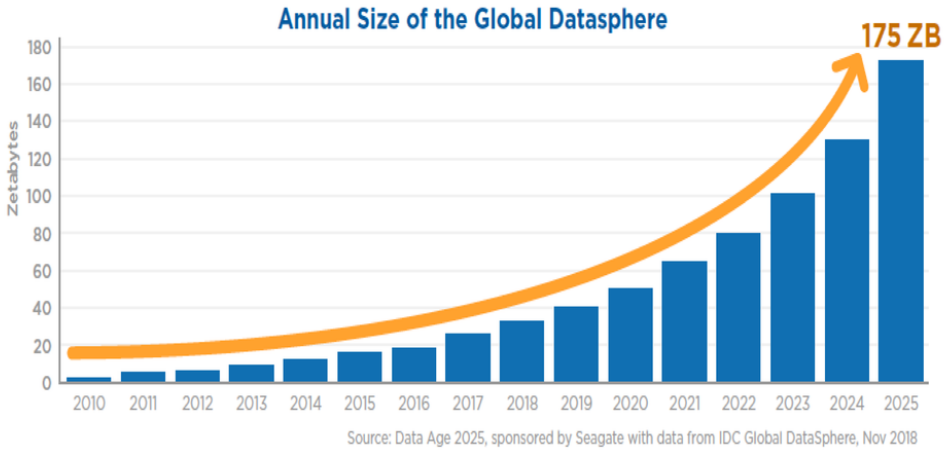


Figure 2.2: Data Generation Trend [2].

World economic forum showed their result in Figure 2.3 of market growth of big data And Worldwide Big Data market revenues on software and services are projected to increases from $\$42B$ in 2018 to $\$103B$ in 2027 which is shown in Figure 2.4

Gartner in 2012 [30] defined big data in more detailed: Big Data are high-volume, high-velocity, and/or high-variety information assets that require new forms of process-

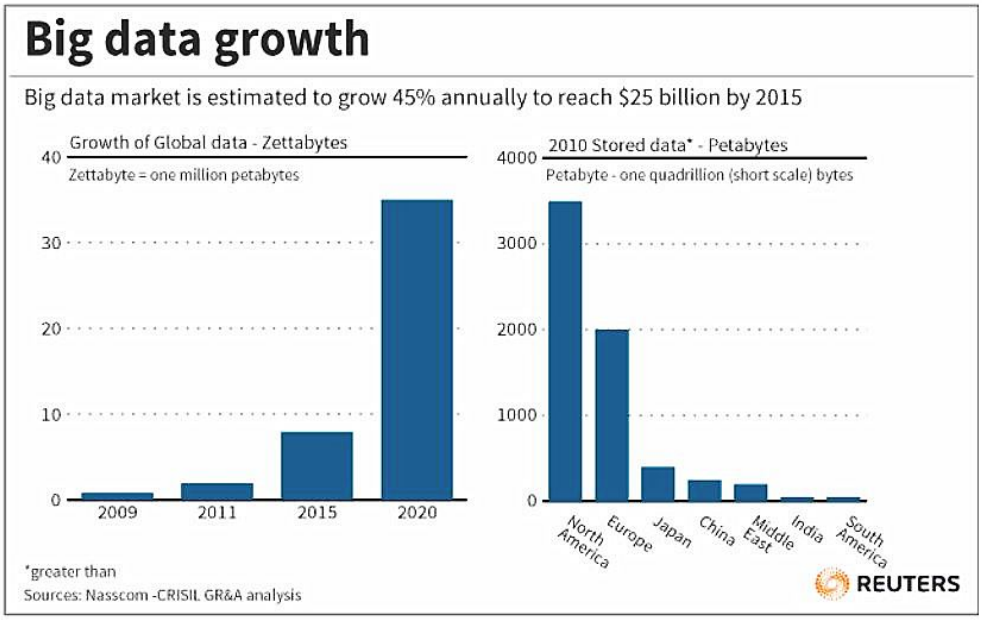


Figure 2.3: Growth of Big Data [3].

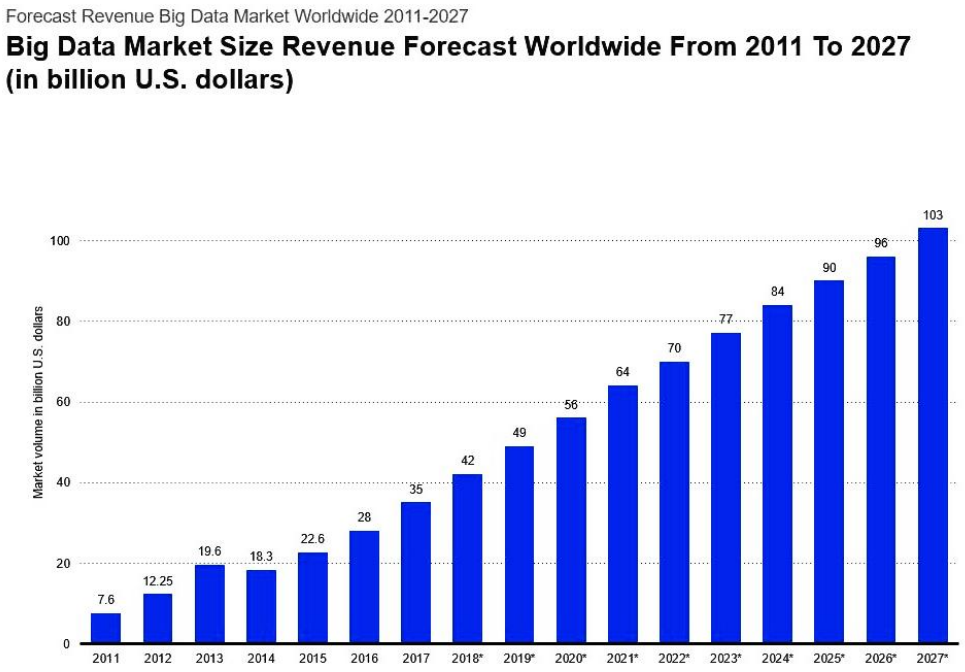


Figure 2.4: Worldwide Big Data Market Size Revenue [4].

ing to enable enhanced decision making, insight discovery and process optimization [31] .

In addition, HACE theorem is proposed by Xindong Wu et al. [8] where features of big data were characterized and from the data mining perspective big data processing model also proposed. Various explanations exist in the definition of big data from 3Vs to 4Vs. Doug Laney and Wu et al. used 3Vs as fundamental characteristics for defining Big data. Those 3Vs are i) volume ii) velocity and iii) variety which indicates the types and sources of data ranges [8, 32]. In 2013 both Berman and Katal et al. [33, 34] argued that for a special requirement and better description more Vs with other characteristics also exist in Figure 2.5.

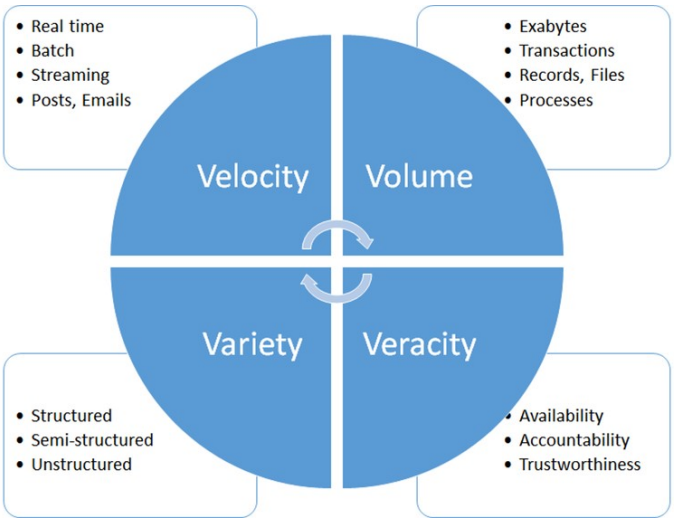


Figure 2.5: Characteristics of Big Data [5].

By using big data revolution many industries, government, scientific field, economic and social sector become able to extract valuable insight and optimize their plan with strategies by taking decision from insight. Different models, opportunities, examples, and projects in many sectors such as politics, commerce, healthcare, tourism etc. are proposed by Fatima-Zahra Benjelloun et al. [35] in 2015. By using Big data tools and approaches, recent research for Health Informatics is presented by Matthew Herland et al. [36] in 2014, where she showed that data gathered at multiple levels, including molecular, tissue, patient and population levels and medical science finds a way to analyze in best way and answer as many medical questions as possible. In distributed

computing system e-science which is computationally intensive science, is implemented usually and by using e-science many big data application issue can be resolved [37],[38]

In Big Data analytics and knowledge discovery machine learning techniques plays a vital role with advances in available computational power [39–42]. Machine learning takes the advantage of simplifying the task of automated classification of big data [14, 15]. In data mining model for lightening the processing model feature selection has been used popularly. For performance evaluation a collection of Big Data with exceptionally large degree of dimensionality are put under test of new feature selection algorithm was proposed by Simon Fong et al. [43] in 2015. At the same year in 2015, Yanfeng Zhang et al. [44] designed and implemented incremental process which is the promising approach to refreshing the mining results and to avoid the expense of re-computation from scratch it utilizes previously saved states. It was the extension to Map-Reduce which is used most widely as a framework for mining Big Data.

2.1.2 Active Learning

Todays world is technology world where every moment data is generated in an infinite rate that we can say that we are living in data deluge. Most of the time those data are in different format, structure and due to their incompleteness and unstructured types it is inaccessible to user [9]. To organize and conduct those large amounts of data which need to be automated is an exhausting work.

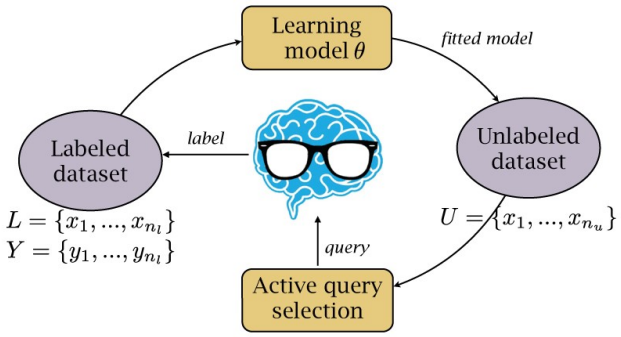


Figure 2.6: Active Learning for Labeling Unlabeled Data [6].

The algorithm of machine learning takes this tedious work to classify the digital object and for doing this work we need to train our machine by giving large amount of labeled data with classification accuracy. Though data labeling is the most costly and swallowing work but in recent years, exploiting unlabeled data is the prime interest [45]. Unlabeled data is often enormous and not difficult to obtain but it is a tedious work to obtain label data which is costly and time consuming. In terms of labeling unlabeled data, active learning uses oracle or human experts for labeling the unlabeled data by applying query function. In supervised learning or classification, for achieving accurate predictive model, a bunch of correctly labeled training instances are needed [22]. The performance of the learning models can be declined due to wrongly labeled or contradictory instances [23]. To label the unlabeled training data, an exhaustive and rigorous analysis is needed. Active learning in Reinforcement learning for data mining is a good solution for addressing this problem [24].

An expert or oracle is used in the active learning process to label the unlabeled data that improve the classification accuracy by posing minimum queries to user or expert. The number of training instances required to learn a concept in active learning can often be much lower than the number required in traditional supervised learning process. Though manual labeling of unlabeled big data is very challenging and expensive. Active learning in machine learning is the process of achieving high classification results using less number of training instances to learn a concept that can often be much lower than the number required in typical supervised learning. It interactively queries an expert to label the unlabeled instances.

The objective is to train a classifier with the help of the expert knowledge to improve the performance of the classifier [46]. Suppose, we have a semi-supervised data set, D, which contains both labeled instances, D_L and unlabeled instances, D_U. Initially, D is bifurcated into D_L and D_U. Then, a classifier or the learning model, M^* is trained using labeled data, D_L. In contrast, a subset of unlabeled instances, $X_U \in D_U$ is chosen from the unlabeled data D_U and the expert or the user is requested to label $X_U \rightarrow X_L$. Finally, X_L is added to D_L and re-trained the ensemble model, M^* again. This process continues until the user is satisfied. The main challenge in active learning is to select the subset of unlabeled instances from the original unlabeled data. The process of active learning is illustrated in Figure 2.7.

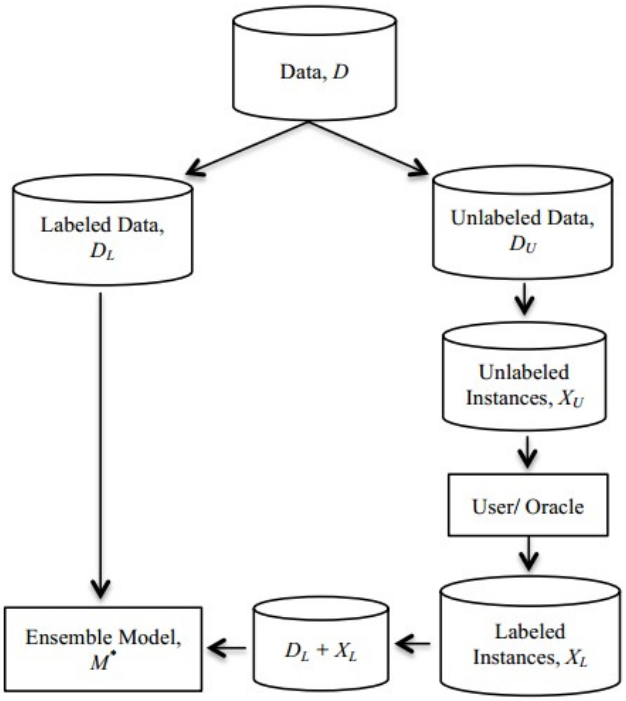


Figure 2.7: Existing Active Learning Process [7].

2.2 Challenges of Active Learning and Big Data Analysis

We have faced some important challenges which already exists when we analyse active learning and big data. In this part we are going to discuss about those challenges.

2.2.1 Challenges of Big Data

Large size, heterogeneous, autonomous sources of data known as big data with decentralized and distributed control and need to pursue complicated and evolving relationship among data. For discovering useful knowledge from Big Data those characteristics fall it an extreme challenge. P. Ahrens [47] and N Gudivada et al.[48] discussed the challenges of big data, that is how we can acquire, store, transfer, search, share, analyze and visualize those data. The main problem of big data is storing and retrieving. Leskovec et al. [49] in 2011 stated in their book that explore the huge volume of data and for future action finding the knowledge and useful information is the most fundamental challenge for Big Data applications. RDBMS (relational database management system) are able to handle traditional data management and analysis systems and these

only applicable for structured, semi-structured and unstructured data but not able to handle huge volume and heterogeneity of big data. The barrier of the development of big data applications are discussed in some literature [50–52] and the main challenges of big data are: Data representation, Redundancy reduction and Data compression, Data life cycle management, Analytical mechanism, Data confidentiality, Energy management, Expendability and scalability, Cooperation, Data with a larger scale, higher diversity, and more complex structures. In particular, big data safety is confronted with the security related challenges like Big data privacy, Big data safety mechanism, Big data application in information security, Data quality [53]. Data inconsistency and incompleteness, scalability, timeliness and data security [54, 55] are also the challenges of big data. Many methods already applied for security purposes and secure sum protocol is one of them. In 2015, Jahan et al. [56] designed a secure sum protocol by using trusted third party where they proved that their security is stronger and its complexity is also better than others. In 2013, Katal et al. [34] discussed the main challenges for the IT professionals and analytical challenges of big data. The first challenge is to design of such systems which would be able to handle such large amount of data efficiently and effectively and the other one is to filter the most important data from all the data collected by the organization.

In 2015, for confronting software architects, Ian Gorton, John Klein [57] described the challenges of Big Data systems. Both data and deployment architectures are tightly linked to the distributed software architecture quality attributes this also showed up. To design the solution based on quality strategy such as scalability, availability and performance based on specific big data technologies, helps the researchers to think of a particular technology to a specific set of architecture tactics.

In literature [58] and [8] analyze big data challenges with data mining. In 2013, though Fan and Bifet [58] discussed the challenges of Big data with data mining but no possible solution and classification of those challenges is not provided. Wu et al. [8], in 2014 discussed challenges according to the V′s dimensions into three tires categorization, Tier I, Tier II and Tier III and consider that data mining deals with machine learning, shows in Figure 2.8. Big Data mining platforms is in Tier I which is the center of the research challenges of three tier structure. Accessing and computing of low-level data is focused in this tier. User privacy issue, application domain knowledge and high-level semantics are concentrated by Tier II. Data sharing with privacy and big data

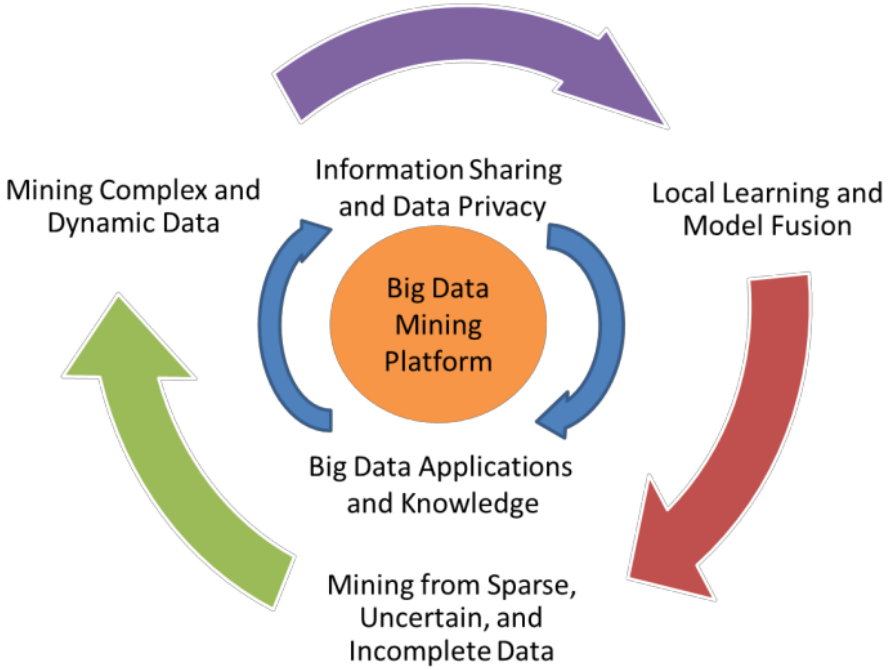


Figure 2.8: Big Data Processing Framework [8].

application domain with knowledge are the most two important issues of this tier. And last tier, Tier III can handle challenges on Big Data Mining algorithms.

2.2.2 Challenges of Active Learning

As the goal of active learning is to achieve high accuracy by using a smaller number of instances, in most active learning the common challenge is to construct a good classifier as initial classifier contains the performance of active learning. Another common challenge is cost, as it′ll be too much costly for manual labeling. An active learning algorithm may waste labeling budget and label on meaningless example by designing poor classifier [59]

In 2012, Lughofer [60] concerned another new challenge which could be faced in many industrial systems. Based on the sample selection, after trained an initial classifier is needed to update and evolve the classifiers during on-line operation phase. In 2012, Zliobaite et al. [61] showed some additional challenges for active learning. For streaming data, at that time data distribution may changes and models need to be adopted with those changes. In order to acquire effective achievement for labelling strategies they

defined three requirements. Their first strategy was, the machine needs to be able to label data at any time so that the budget will not be exceeded over time. The second strategy was, to be able to adapt to changes over all the instance space a strategy should query. The last one was, for any kind of unbiased changes, monitoring and detection strategy need to preserve in the input data distribution.

2.3 Analytics of Active Learning and Big Data Management

The processing capacity of traditional systems already exceeds the management of storage system due to increasing amount of big data [62]. By experts and practitioners, distributed computing has been widely used, to boost up sequential solutions in medium-size data before the advent of Big Data [8].

Nowadays, Map-Reduce is the parallel programming able to handle big data processing for public cloud computing platform where Big Data services are provided. In machine learning and data mining algorithm Map-Reduce parallel programming is applied. For enhancing the real-time nature, large-scale data processing have been received a significant amount of attention for improving the performance of Map-Reduce [8]. To handle different types of structured and unstructured data in various Map-Reduce framework like: Apache Hadoop, Skynet, Sailfish and File-Map have been developed [35]. Based on the simple Map-Reduce programming model on multi-core processors, Chu et al. [41] in 2006 proposed a method which is applicable to a large number of machine learning algorithms. In this framework, k-Means, Naïve Bayes, Linear SVM, Gaussian discriminant analysis, expectation maximization, the independent variable analysis, back-propagation neural networks and locally weighted linear regression algorithm these 10 data mining algorithms are introduced.

In 2017, Sowmya and Suneeth [9] proposed a research work in the development for real time application in Big Data of single platform. Their work was form of 3 abstraction levels which is shown in Figure 2.9.

The three abstraction levels were, i) Data Acquisition which is consist of data collection, transmission, and pre-processing units, ii) Data Processing which is the part of storage and integration of data by using different transformation method to obtain process data, and the last part was iii) Data Services/Retrieval which provide

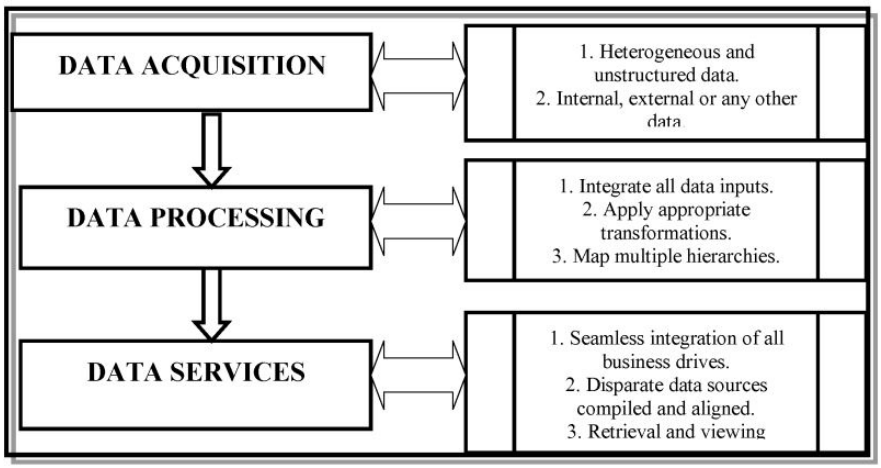


Figure 2.9: Framework of Data Mining using Big Data [9].

an interface and give the services for the user to access and collect data from different sources and use them.

For predicting the classification accuracy from unlabeled big data, active learning is needed to select a minimum number of unlabeled instances which is very demanding work. In each iteration, the process of querying the expert for labeling the unlabeled data makes the process slow and costly but active learning needs to increase the accuracy by using a minimum number of instances. Pool based active learning is used to select instances from unlabeled data. In 2014-15, Demir and Guo [63, 64] mentioned different methods for selecting instances like representative sampling, uncertainty sampling, random sampling, local uncertainty sampling, committee based active learning, etc. A common strategy of active learning is the classifier-based sampling where uncertainty builds in a classification [65, 66]. To evaluate the uncertainty of the data from the active pool for visual concept recognition, a semi-supervised batch mode multi-class active learning algorithm presented by Yang et al. in 2015 [67]. Though depending on the pool set, classification performance is varied but this method can evaluate the informativeness and exploit uncertainty into multiple classes.

In 2016, Dimitriadou et al.[68] proposed an Active Learning-based approach for Interactive Data Exploration (AIDE) where several exploration techniques are presented and a number of samples are presented in a minimizes way. To explore both labeled and unlabeled data-set in a two-way process (forward and Backward), a bi-directional active learning algorithm is proposed by Zhang et al. [67] in 2015. In 2016, Park and Kang

[69] proposed an active learning method which can compute the selective probability by measuring the method of concept drift on the data stream. A query-by-committee framework is introduced by Dou et al. [70] in 2017, where no previous labeled data is needed. A new strategy for data-driven classifier was proposed by Edwin Lughofer [60] in 2012, where initial learner or labeling was not required. By using this method from off-line training phase samples are selected and most informative instances exist near the center of the cluster and the border of the cluster. Following a certainty-based criterion called on-line training, the samples are updated incrementally. This method is a combination of both unsupervised and supervised active learning thats why this method is called hybrid active learning. Another method which is almost similar to the hybrid method was established in 2009 by Hu et al. [71], where based on random selection pre-labeled examples from initial collection are needed. In this method incrementally update classifier is not included for any types of on-line active learning

2.4 Summary

Active learning is a part of machine learning that can classify a huge set of unlabeled big data. The algorithm of AL(Active Learning) performs the tremendous work of classifying and labeling digital object using less number of training instances with the help of oracle/expert knowledge. There are many challenges exist regarding big data management and including data security. Though AL works efficiently although there is a high chance of misclassification if it cannot collect informative instances properly.

Chapter 3

Proposed Method

This chapter will be a discuss about different types of approaches which is used in our proposed framework of active learning for mining big data.

3.1 Approaches of Our Proposed Framework

In our framework, we have proposed two objectives for collecting informative training instances in the active learning process for supervised and semi-supervised big data analysis. These are:

- To collect Nearest Neighbor data point of the center of the cluster
- To collect informative instances randomly from each cluster.

After collecting data by using these objectives, we have used five classification methods to predict the accuracy of providing new test data and for clustering, scaling and data reducing method is used. These are:

3.1.1 C4.5 (Decision Tree Induction)

C4.5 is the classifier used in Data Mining which can generate a decision by using a certain sample of data and a decision tree. The most common conventional decision tree was ID3. But ID3 had some problem and faced over-fitting. To tackle those problem new decision tree C4.5 comes, which also able to handle missing value and by using continuous data, C4.5 can generate a generalized model.

3.1.2 SVM (Support Vector Machine)

SVM is used in both classification and regression problem. The main objective of SVM is to find N-dimensional hyper-plane where N is the number of features. Each data item is placed in the value of a particular coordinate that distinctly classifies the data points

3.1.3 Random Forest

Random Forest is an ensemble learning method for classification and regression. It is called ensemble learning because it combines more than one algorithm for classifying objects. It can build multiple decision trees and to get more stable and accurate prediction, then it merge those decision trees and take a vote for final consideration.

3.1.4 Bagging

Bagging is a bootstrap aggregation which is a simple and powerful ensemble method used in statistical classification and regression. Though this method is also ensemble like the random forest, but it can avoid over-fitting and can reduce the variance which had high variance.

3.1.5 Boosting

Boosting is an ensemble method that combines weak learners to form strong classifier. Three boosting algorithms exist they are: AdaBoost (Adaptive Boosting), Gradient Tree Boosting and XGBoost. In our framework, we have used Adaboost algorithm which is more popular for boosting the performance of binary class classification problem of the decision tree. It is also sensitive to noisy data and outliers. By giving weights it can generate rules for finding weak learner and strong classifier.

3.2 Proposed Framework

Active learning is a technique to build a classifier with a minimum number of training instances. In this section, we have discussed our proposed framework of active learning of big data analysis. The idea of active learning is that if a classifier can be trained

using the selected smaller number of training instances, it can perform better than traditional machine learning methods.

3.2.1 Proposed Active Learning Framework for Mining Big Data

We have proposed our framework for both supervised and semi-supervised data set, although we have focused our framework for providing perfect accuracy of semi-supervised dataset. In our proposed framework we have worked on a collection of semi-supervised big data, D. We divide these semi-supervised big data D into two subsets, labeled data D_L and unlabeled data D_U. By using labeled data D_L we generate a model/classifier M. On the other hand, we divide D_U into several cluster $\{C_1, C_2, \cdots, C_k\}$ where each cluster have sub-datasets $\{D_1, D_2, \cdots, D_n\}$ and each sub dataset contain instances $\{x_1, x_2, \cdots, x_N\}$.

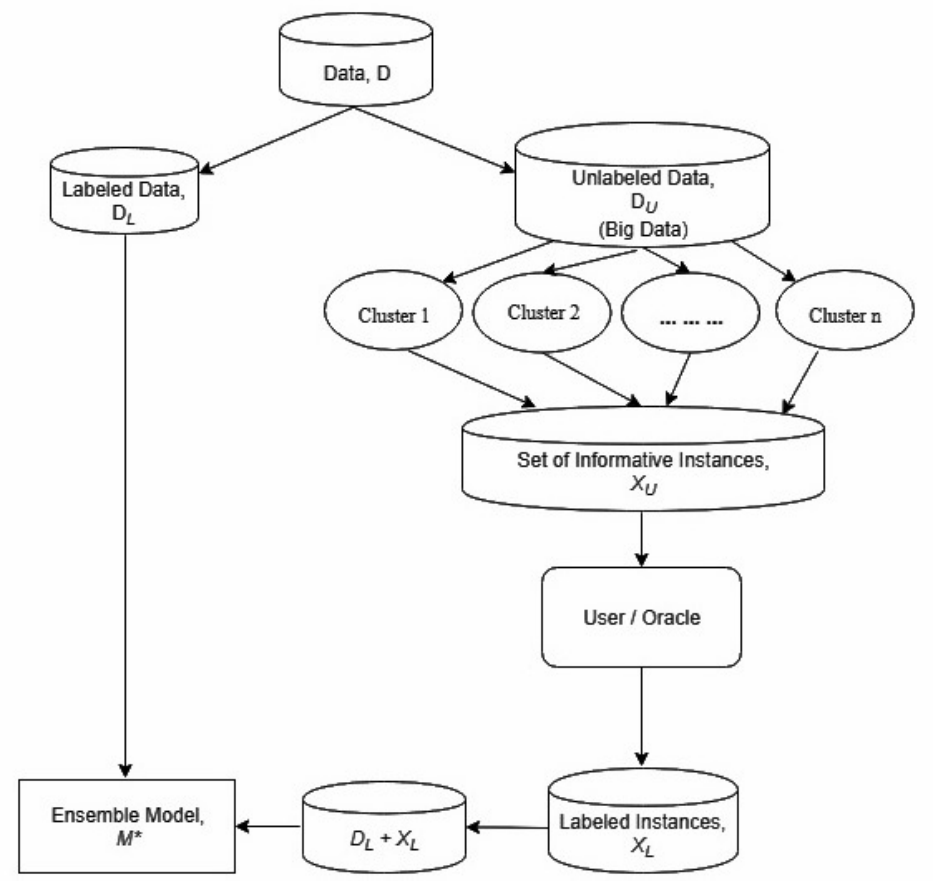


Figure 3.1: Proposed Active Learning Framework.

According to the way of active learning we need to collect informative instances from each cluster by using our proposed method. From each cluster C_i we collect informative instances in two ways. Firstly, X_U is collected from the nearest neighbor data point of the center of the cluster and secondly, X_U is collected from randomly taking instances from each C_i. After collecting unlabelled training instances, we labeled those instances with the help of user or oracle by using generated model M and update D_L by adding those labeled instances. Finally, we retrained the classifier M again by updated D_L. When new test data comes, updated generated model M, predict the accuracy. Figure 3.1 shows the whole procedure of our proposed framework. Algorithm 1 illustrates the proposed framework of active learning of big data analysis.

Algorithm 1 Proposed Active Learning

Input: Semi-supervised data, D;

Output: An active learning model, M;

Method:

1: divide D into D_L, and D_U;
2: create a model, M using D_L;
3: divide D_U into several sub-datasets $\{D_1, D_2, \cdots, D_n\}$;
4: **for** each $D_i = \{x_1, x_2, \cdots, x_N\}$ **do**
5: cluster D_i into $\{C_1, C_2, \cdots, C_k\}$;
6: **for** each C_i **do**
7: $X_U \leftarrow$ the cluster centre, $x_m \in C_i$;
8: $X_U \leftarrow$ nearest neighbors of the centre, $x_m \in C_i$;
9: $X_U \leftarrow$ random instances from C_i;
10: **end for**
11: $X_L \leftarrow X_U$ by expert/ oracle;
12: $D_L \leftarrow X_L$;
13: **end for**
14: retrain, M using D_L;
15: return M;

3.3 Summary

In this research, we have proposed a new technique with an algorithm for collecting informative instances in active learning process. For collecting informative instances in

our proposed method we need to collect and process our dataset like handling missing value and categorical instances and scaling those data before clustering. For calculate accuracy by using our technique we have applied some popular classification method like Decision tree (C4.5), Support Vector Machine (SVM), Random Forest, Bagging and Boosting.

Chapter 4

Experimental Analysis

This chapter are presented the detailed information about dataset which is used in our experiment. Also, we have discussed about the experimental setup and experimental results on our proposed method.

4.1 Data Collection and Processing Methods

As we have worked on semi-supervised datasets and collected on various types of data. We have processed differently formatted data to decorate them in the same format.

4.1.1 Data Collection

For our experiment, we have collected different types and formatted dataset from UCI machine learning repository and KEEL(Knowledge Extraction based on Evolutionary Learning) dataset repository. Both supervised and semi-supervised dataset have been collected. The supervised dataset is used for ensuring that our framework performs correctly and by using semi-supervised we have showon the result of our experiment. As we collected both train and test data and if sometimes test data is not available in that case from training data, we split those data into 70/30 that means 70% dataset for training and 30% dataset for testing. As our framework will work more efficiently for huge dataset, that's why for gathering huge amount of data in one dataset, we collect the same dataset in different range labeling, merge them and keep unique data. We also separated both label and unlabeled data.

4.1.2 Data Processing

As our framework will be able to the evaluate accuracy of any types and formatted dataset so, we have collected different types of dataset. After collecting data, we need to process those dataset. Moreover, we used jupyter notebook which is an open-source web application and python for coding. For a categorical dataset, we separated both integer and object type data. For changing the format from object to integer we use label encoder which can encode labels with a value between 0 to n. And if any row contains ? or null data we also encode those data using label encoder. After encoding, we merge those data. Here, when we separated integer and object data, we have faced that after separating integer data though data is an integer, but its attribute format is object type. In that case, we needed to change its attribute type from object to integer. For processing dataset we take the following steps.

4.1.2.1 Missing Value Handling

Sometimes our dataset contains missing/null value, value with ? and , operator. As null/missing value is not possible to handle or not possible to convert into other values so we just remove those instances. And if instance contains ?, or , value, then we convert this value into numeric value by encoding method.

4.1.2.2 Categorical Instances

For categorical feature and instances, we convert those instances into numeric value by using encoding process using Scikit-learn library. Scikit learn is a free software machine learning library which is used for python programming. Though there exist two encoding process, 'Label encoder' and 'One-Hot' encoder but we use Label encoder in our prepossessing for converting our object type data into integer.

4.1.2.3 Label Encoder

Label encoder is an encoding process which converts any object(text) type data into an integer(numeric) by labeling the value between 0 to n_classes. Label encoding is a property of Scikit-learn library. For using the label encoding process we need to import library 'sklearn import prepossessing' and 'le = prepossessing.Label Encoder()'. Though another encoding process, 'One-Hot' encoder also exists but it just encodes

object value into a binary value. So, 'One-Hot' encoder will be applicable for solving any binary class classification problem.

4.1.2.4 Data Scaling

For applying clustering technique we need to apply normalization or standardization. Normalization is the technique of structuring a relational database in normal forms with reducing the data redundancy and improving the data integrity. On the other hand, standardization is the technique of bringing data into a common format for collaborating and large-scale analytics. It mainly structures disparate datasets into a Common Data Format. There are two methods exists for data standardization. First one is min-max scaling which scale value into different means and standard deviations but in equal ranges. There is at least one observed value at 0 and 1 endpoints. The second one is Z score scaling, where all variables in the dataset have equal means (0) and standard deviations (1) but different ranges. As we have used k-means cluster so we have applied min-max scaling method before clustering.

4.1.3 Clustering for Data Selection

We have already identified that the main purpose of active learning is to collect less number of informative training instances to build a classification model. As our framework is to apply clustering technique in active learning for collecting those informative instances more easily, we clustered the unlabeled dataset. Now, the question is how many cluster need to do? The answer is it depends on the user. But in our framework, we have proposed two objectives for collecting informative instances. In the experiment section, we have shown the detail experimental result and the discuss on everything.

4.2 Datasets Information

We have applied the following seven real benchmark datasets from UCI machine learning repository, where 4 datasets are supervised and 3 are semi-supervised. Table 4.1 presents the details of the datasets.

1. Poker Hand Dataset (Poker)

2. Thyroid Disease Dataset (Thyroid)

3. Yeast Dataset (Yeast)
4. KDD Cup 1999 Dataset (KDD Cup)
5. Abalone Dataset (Abalone)
6. Contraceptive Method Choice Dataset (Contraceptive)
7. MAGIC Gamma Telescope Dataset (Magic)

Table 4.1: Datasets Description

No.	Datasets	No. of Features	Types of Features	Working Instances	Training Instances	Testing Instances	Classes	Category
1	Poker	11	Categorical, Integer	1025010	25010	1000000	10	Supervised
2	Thyroid	29	Categorical, Real	9172	6421	2751	6	Supervised
3	Yeast	9	Real	1484	1039	445	10	Supervised
4	KDD Cup	42	Categorical Integer	4000000	2800000	1200000	23	Supervised
5	Abalone	9	Categorical, Integer	7508	1506 (Labeled) 2250 (Unlabeled)	3751	29	Semi Supervised
6	Contraceptive	9	Categorical, Integer	2650	530 (Labeled) 795 (Unlabeled)	1325	3	Semi Supervised
7	Magic	11	Real	53720	17705 (Labeled) 18897 (Unlabeled)	17117	2	Semi Supervised

4.3 Experimental Setup

We have used accuracy, precision, recall, f1-score, and 10-fold cross-validation to evaluate our proposed method by using 5 different machine learning algorithms like C4.5 (Decision Tree), SVM (Support Vector Machine), Random Forest, Bagging and Boosting. Then we have compared the main accuracy with our proposed clustering method where we apply 2 techniques (Random Sampling, Instances near to the center of the cluster) to evaluate our proposed method. So, our result is based on four possible classification outcomes. And for any types of classification measurement confusion matrix is needed.

Confusion Matrix: A Confusion matrix is a tool for analyzing different classes that classifier can recognize tuples. It is a table which is used by a classification/classifier to mark out the achievement of known true value on a set of test data. Measuring accuracy, precision, recall confusion matrix needed immensely.

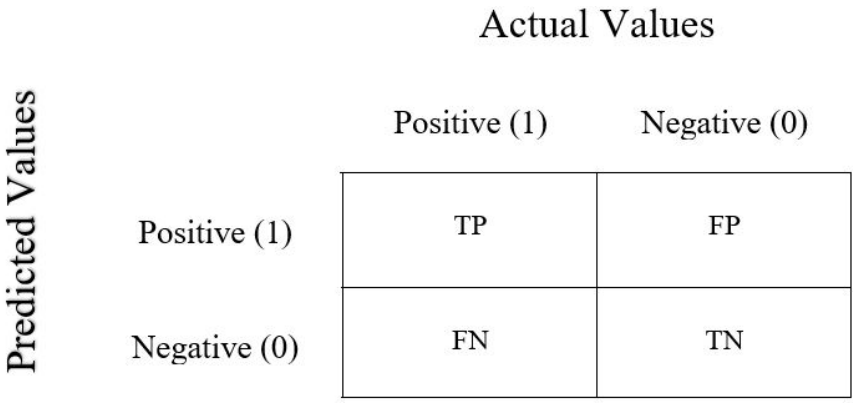


Figure 4.1: Confusion Matrix.

True Positive (TP): When the machine predicts the positive value, that means both predicted and actual class is positive/yes (1) or class are labeled correctly by classifier, it is called true positive.

True Negative (TN): When the machine predicts the negative value, that means both predicted and actual class are negative/no (0) or classifier can label the class correctly where value is negative is called true negative.

False Positive (FP): Machine predicts the value is positive/yes (1) but its actual value is negative/no (0) that is classifier label the class incorrectly as positive is called false positive

False Negative (FN): Machine predicts the value is negative/no (0) but its actual value is positive/yes (1) that is classifier label the class incorrectly as positive is called false positive.

Accuracy: When we train a model, we measure its accuracy either trained accuracy or prediction accuracy. Trained accuracy represents that our trained model is being trained correctly by measuring its ratio. Predicted accuracy represents the predicted observation of the total observation or known value. On the other hand, accuracy is

the percentage of the test set which is classified correctly.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{4.1}$$

Precision: Precision is the measurement of exactness of actual class. That is predicted result is being consistent when measurement is repeated or how well a result can be determined when it is measured.

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

Recall: Recall is the measurement of completeness of the actual class. When we predict on test set then the ratio of positive observation on all the observation of actual positive is called recall.

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

F Score: F1 score / F measurement is the combination of both precision and recall that it is taken both false positives and false negatives into accounts.

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

4.4 Experimental Results

The classification results are summarized in Table 4.2 to Table 4.6 for C4.5 (decision tree induction), SVM (support vector machines), Random Forest, Bagging, and Boosting (AdaBoost) respectively on each dataset. In these table we have shown the actual accuracy of each dataset where all instances are used. As we have applied five different classification method for calculating accuracy that's why each Table 4.2 to Table 4.6 represent the actual accuracy with their precision, recall and F-score results.

Table 4.2: Classification Results of C4.5 with 10-Fold Cross Validation

Datasets	Accuracy (%)	Precision (weighted average)	Recall (weighted average)	F-Score (weighted average)
Poker	47.85	48.62	47.85	48.23
Thyroid	61.19	61.15	61.19	61.16
Yeast	51.01	50.96	51.01	50.91
KDD Cup	99.99	99.99	99.99	99.99
Abalone	51.72	51.94	51.72	51.68
Contraceptive	67.22	67.27	67.22	67.24
Magic	64.85	77.21	64.85	51.05

Table 4.3: Classification results of SVM with 10-Fold Cross Validation

Datasets	Accuracy (%)	Precision (weighted average)	Recall (weighted average)	F-Score (weighted average)
Poker	50.12	25.12	50.12	33.47
Thyroid	70.71	70.37	70.71	68.68
Yeast	57.98	60.02	57.98	55.88
KDD Cup	99.99	99.99	99.99	99.99
Abalone	24.10	13.59	24.10	16.83
Contraceptive	51.66	51.73	51.66	51.51
Magic	64.83	42.03	64.83	50.99

Table 4.4: Classification results of Random Forest with 10-Fold Cross Validation

Datasets	Accuracy (%)	Precision (weighted average)	Recall (weighted average)	F-Score (weighted average)
Poker	57.43	54.61	57.43	54.61
Thyroid	69.18	67.48	69.18	68.29
Yeast	61.57	60.67	61.57	60.35
KDD Cup	99.99	99.99	99.99	99.99
Abalone	53.11	54.29	53.11	52.89
Contraceptive	68.66	68.47	68.47	68.53
Magic	78.95	84.11	78.95	75.92

Table 4.5: Classification results of Bagging with 10-Fold Cross Validation

Datasets	Accuracy (%)	Precision (weighted average)	Recall (weighted average)	F-Score (weighted average)
Poker	58.70	55.96	58.70	56.14
Thyroid	69.04	68.30	69.04	68.41
Yeast	59.78	59.35	59.78	59.08
KDD Cup	99.99	99.99	99.99	99.99
Abalone	52.73	53.68	52.73	52.47
Contraceptive	69.79	69.62	69.79	69.65
Magic	78.90	84.08	78.90	75.86

Table 4.6: Classification results of Adaboost with 10-Fold Cross Validation

Datasets	Accuracy (%)	Precision (weighted average)	Recall (weighted average)	F-Score (weighted average)
Poker	48.00	48.72	48.00	48.36
Thyroid	61.52	61.38	61.52	61.43
Yeast	50.34	50.64	50.34	50.39
KDD Cup	99.99	99.99	99.99	99.99
Abalone	51.64	52.15	51.64	51.59
Contraceptive	68.50	68.56	68.50	68.52
Magic	82.54	86.24	82.54	80.69

Classification accuracy (%) of different classifiers in the proposed active learning method with randomly selected instances from each cluster are tabulated in Table 4.7. From the result, we can achieve almost the same and sometimes better classification results using 50% of the training instances which is shown in Figure 4.2 to Figure 4.4. In this experiment, our main aim was to keep the same accuracy by using less number of instances as less as possible. At first, we divide the whole unlabeled dataset into 2 cluster, then from each cluster, we select data randomly for calculating accuracy. In this case, firstly we took 100(%) data from each cluster, calculate accuracy by using five classification methods and compare their accuracy with their actual accuracy where all instances were present. Then at the same process we took 90(%), 80(%),70(%),60(%) and lastly 50(%). We ended up our experiment at 50(%) data from whole data because we can summarize that for random selection if we take 50(%) data from each cluster of the total data then we can achieve the almost same and sometimes better classification results. So, 50 (%) is the standard measurement for random selection.

Table 4.7: Accuracy (%) of Different Classifiers in Active Learning with Randomly Selected Instances from Each Cluster (Cluster = 2).

Datasets	Randomly Taking Data	C4.5	SVM	Random Forest	Bagging	Boosting
Poker	100(%)	47.90	50.12	56.07	56.42	48.47
	90(%)	48.96	50.12	55.06	56.89	50.09
	80(%)	50.46	50.12	54.78	57.04	48.72
	70(%)	49.24	50.12	54.98	57.53	50.16
	60(%)	48.24	50.12	54.40	57.28	47.07
	50(%)	45.61	50.12	54.21	57.41	47.71
Thyroid	100(%)	60.79	70.71	69.18	68.57	60.36
	90(%)	61.30	70.71	69.13	67.84	60.90
	80(%)	60.79	70.71	68.55	68.35	60.79
	70(%)	61.66	70.71	68.95	67.62	60.97
	60(%)	60.86	70.71	69.24	68.24	60.54
	50(%)	60.90	70.71	69.42	67.84	60.83
Yeast	100(%)	73.93	55.56	77.98	77.52	73.03
	90(%)	74.06	57.10	77.46	77.31	72.32
	80(%)	77.52	54.78	74.89	71.63	72.47
	70(%)	72.76	56.74	73.39	77.88	68.27
	60(%)	67.42	51.68	71.31	73.78	71.16
	50(%)	66.82	52.47	69.96	68.61	66.82
KDD Cup	100(%)	99.99	99.99	99.99	99.99	99.99
	90(%)	99.99	99.99	99.99	99.99	99.99
	80(%)	99.99	99.99	99.99	99.99	99.99
	70(%)	99.99	99.99	99.99	99.99	99.99
	60(%)	99.99	99.99	99.99	99.99	99.99
	50(%)	99.99	99.99	99.99	99.99	99.99
Abalone	100(%)	51.74	24.18	52.76	52.76	51.99
	90(%)	51.96	24.29	52.72	53.11	51.96
	80(%)	52.39	24.18	53.45	53.79	52.12

Table 4.7: Accuracy (%) of Different Classifiers in Active Learning with Randomly Selected Instances from Each Cluster (cluster = 2)(*Continue*)

Datasets	Randomly Taking Data	C4.5	SVM	Random Forest	Bagging	Boosting
Abalone	70(%)	52.19	24.26	52.95	53.32	51.77
	60(%)	52.47	24.21	53.22	53.24	51.32
	50(%)	52.04	24.18	53.29	52.81	51.51
Contraceptive	100(%)	67.37	51.96	69.49	68.58	68.05
	90(%)	66.09	51.44	68.54	69.34	67.75
	80(%)	67.07	51.28	69.76	68.05	67.52
	70(%)	66.84	51.74	69.68	69.64	67.90
	60(%)	68.28	51.89	68.36	70.17	68.73
	50(%)	67.29	51.36	69.46	70.69	68.20
Magic	100(%)	98.93	79.02	98.35	98.29	98.79
	90(%)	98.76	79.09	98.28	98.29	98.92
	80(%)	98.83	78.93	98.32	98.26	98.83
	70(%)	98.85	79.06	98.13	98.24	98.86
	60(%)	98.88	79.02	98.32	98.20	98.91
	50(%)	98.88	78.99	98.14	98.05	98.69

Table 4.8 shows the classification accuracy (%) of different classifiers that were trained with the nearest neighbor instances of the center of the cluster. In this experiment, at first, we cluster dataset into a different cluster. Here we take 100 to 500 number for cluster number and from each cluster we just take the nearest neighbor data point from the center of the cluster. As from each cluster, only one data point is picked up, that's why we need to increase the number of the cluster. We also followed the same procedure as the previous table for calculating the accuracy, just instances collection was different. In this experiment, we can see that the center of the cluster from each cluster also increase the accuracy using less number of training instances. And if cluster number increases the accuracy also increases. The result effect is shown in Figure 4.5 to Figure 4.7

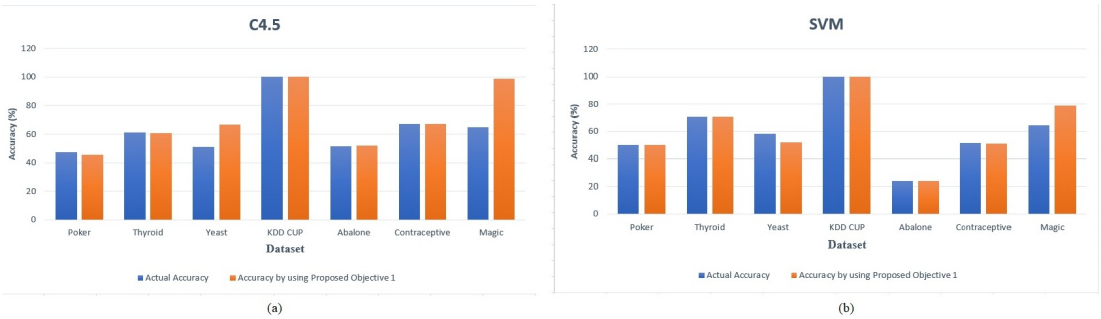


Figure 4.2: C4.5(a) and SVM(b): Comparison of Actual Accuracy and the Accuracy of Random Collection (50% instances) from Each Cluster.

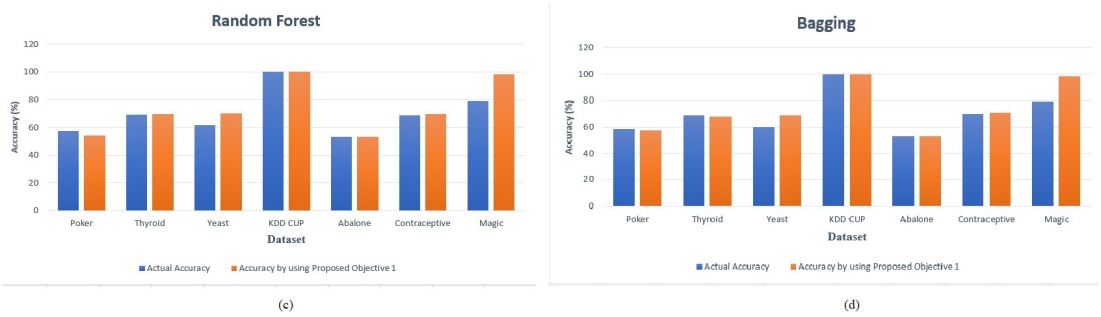


Figure 4.3: Random Forest(c) and Bagging(d): Comparison of Actual Accuracy and the Accuracy of Random Collection (50% instances) from Each Cluster.

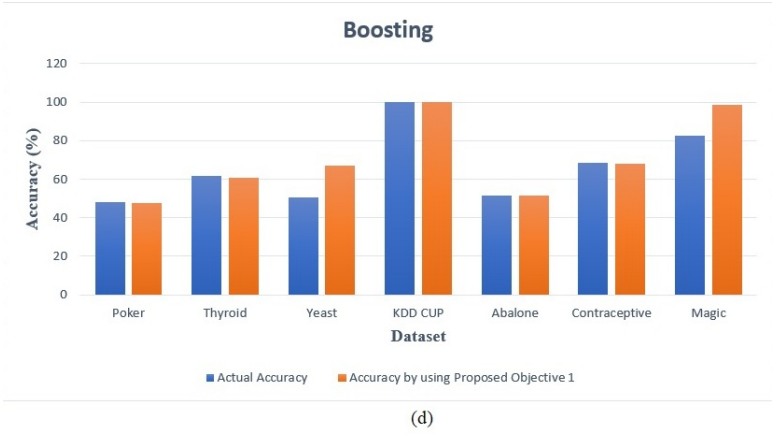
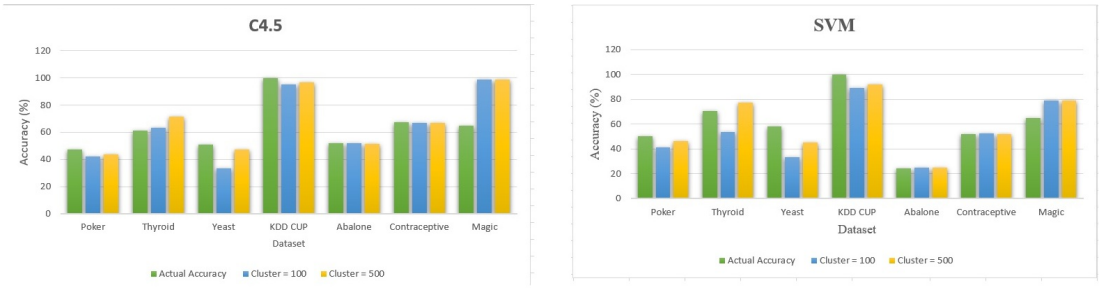


Figure 4.4: Boosting(e):Comparison of Actual Accuracy and the Accuracy of Random Collection (50% instances) from Each Cluster.

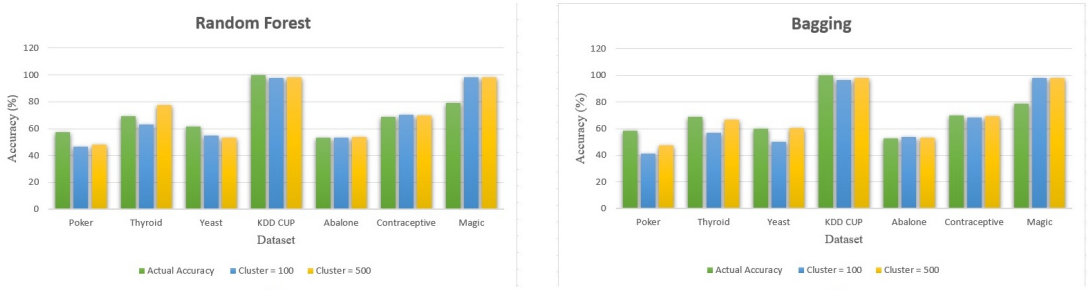
Table 4.8: Accuracy (%) of Different Classifiers in Active Learning with Cluster Centre and Nearest Neighbors Instances from Each Cluster

Datasets	Number of Cluster	C4.5	SVM	Random Forest	Bagging	Boosting
Poker	100	40.42	41.24	46.45	41.16	40.81
	200	42.74	44.32	47.03	46.59	41.73
	300	43.69	45.35	48.38	46.87	41.91
	400	44.05	46.13	47.31	48.22	43.15
	500	43.56	46.11	48.06	47.57	44.87
Thyroid	100	63.33	53.33	63.33	56.67	63.33
	200	70.00	75.00	64.33	66.67	66.67
	300	71.11	76.67	74.22	77.78	65.56
	400	72.50	74.17	76.33	72.50	68.33
	500	71.33	77.33	77.33	79.33	74.00
Yeast	100	33.33	33.33	54.67	50.00	40.00
	200	50.00	36.67	53.33	53.33	46.67
	300	47.78	31.11	52.22	55.56	45.56
	400	50.00	44.17	55.00	52.50	47.50
	500	47.33	45.33	53.33	60.67	55.33
KDD Cup	100	95.01	89.00	97.56	96.34	95.21
	200	95.43	90.34	97.89	96.45	95.89
	300	95.69	89.23	97.93	97.45	96.49
	400	96.59	91.42	98.54	97.69	97.50
	500	96.79	91.96	98.44	97.99	97.69
Abalone	100	51.99	24.69	53.36	53.51	52.31
	200	51.24	24.78	53.98	53.43	52.17
	300	52.19	24.63	53.48	53.39	52.12
	400	51.45	25.08	53.59	53.35	52.31
	500	51.64	25.00	53.85	53.16	51.91
Contraceptive	100	66.77	52.19	70.39	68.28	67.82
	200	67.29	52.27	69.11	67.98	68.28
	300	65.41	52.27	70.77	70.24	68.58



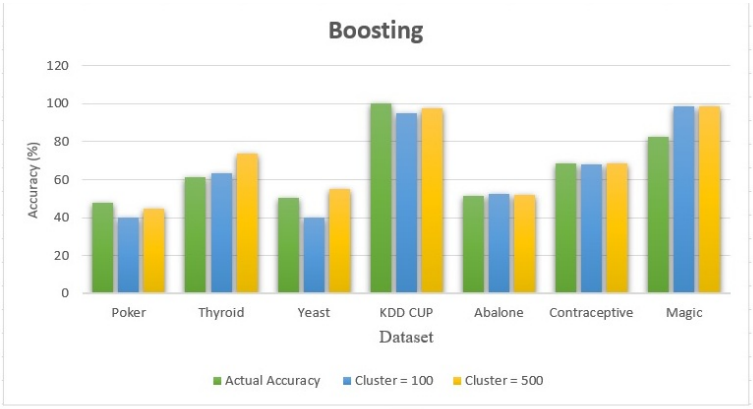
(a) (b)

Figure 4.5: C4.5(a) and SVM(b): Comparison of Actual Accuracy and the Accuracy of Collected Informative Instances which belongs to the Near of Cluster Center



(c) (d)

Figure 4.6: Random Forest(c) and Bagging(d): Comparison of Actual Accuracy and the Accuracy of Collected Informative Instances which belongs to the Near of Cluster Center



(e)

Figure 4.7: Boosting(e): Comparison of Actual Accuracy and the Accuracy of Collected Informative Instances which belongs to the Near of Cluster Center

Table 4.8: Accuracy (%) of different classifiers in active learning with cluster centre and nearest neighbors instances from each cluster.(*Continue*)

Datasets	Number of Cluster	C4.5	SVM	Random Forest	Bagging	Boosting
Contraceptive	400	66.69	51.81	70.32	68.43	68.20
	500	66.54	51.81	69.64	69.56	8.50
Magic	100	98.81	78.89	98.07	97.98	98.89
	200	98.87	78.89	98.14	98.10	98.81
	300	98.86	78.89	98.14	98.05	98.85
	400	98.85	78.89	98.34	98.16	98.88
	500	98.87	78.89	98.17	98.05	98.88

4.5 Summary

In this thesis, we have proposed an approach to find the informative unlabeled instances from a large number of unlabeled data/big data, which represent the total dataset and ask the expert to label only the selected informative unlabeled instances. We have used clustering technique to find the informative unlabeled instances and labeling them by experts/users to improve the process in active learning which showed in Table 4.7 and Table 4.8. We have considered the nearest neighbor data instances of the center of the cluster including the center of the cluster for selecting informative instances. Also, we have taken some instances randomly from each cluster. In a cluster, instances are similar to each other, so we are taking a few instances from each cluster. The experimental analysis proved that we could achieve almost the same and sometimes better classification results using 50% of the training instances. Also, considering only nearest neighbor instances of the center of the cluster from each cluster increase the classification results.

Chapter 5

Conclusion and Future Work

This chapter contains the discuss of the total thesis work with its limitations and conclude with future work.

5.1 Discussion

The main purpose of this thesis work was to collect informative instances as less as possible by maintaining the same accuracy of the total instance of a dataset through the active learning process. Though many methods are established for collecting informative instances by active learning but most of them are for binary or only for a single class dataset. So, we have proposed some new methods in the active learning process to collect informative instances from any types of the datasets with multiple classes. The main action for collecting informative instances from big data (semi-supervised) was to cluster them into N number, where N will be defined by the user. As in experiment section we have shown that, more accuracy comes when cluster number is excessive. After clustering, we have collected informative instances either randomly or the nearest neighbor point of the center of each cluster. In our experiment section, we have shown these two methods on different types of dataset. In our first method, we have been trying to collect informative data randomly from each cluster. Then with the help of user/oracle, we have labeled them and update the trained model newly and predict the accuracy of upcoming test data. We also use five different methods (C4.5, SVM, Random forest, Bagging and Boosting) for measuring the accuracy of this method and compare those accuracies with its actual accuracy. As our main goal was to establish

the same or better accuracy by taking as fewer instances as possible. Thats why we have been trying to increase the percentage rate of randomly taking data from each cluster. Then the result showed that, by taking at least 50% data from each cluster we can get almost same and sometimes better accuracy. So, we can easily avoid other 50% data from each cluster of any dataset which will not be too much effective for generating a model. In the second method, we have been trying to take data which is located at the center of the cluster and nearest neighbor point of the center of the cluster. As we are taking just one data point from each cluster as an informative instance so we need to increase the number of the cluster. We can see that for increasing the number of the cluster the accuracy rate also increases. For getting the actual or better accuracy sometimes 200 cluster was enough but sometimes we needed 500 cluster for keeping the actual accuracy of that dataset. As we applied five different classification methods (C4.5, SVM, Random Forest, Bagging, Boosting and KNN) in our two objectives and these methods also used for acquiring the actual classification accuracy. After applying our two objectives by using these methods we can see that in every step Random forest and Bagging methods are given better accuracy than others. Sometimes Boosting algorithm is given better result but C4.5 and SVM were not good than other methods. We also notice that though two objectives are proposed by ourselves but our second objective which collect instances at the nearest neighbor of the center of the cluster was the easiest way. The reason is, we need to cluster the unlabeled data as much as possible then just we need to collect informative data which was closer of the cluster center. As we mentioned that our proposed method will work for semi-supervised big data and we also proved it in our experiment. For our experiment, we have tried to select big data though our method will work for less amount of semi-supervised data.

In our experiment, we have followed our two objectives, random collection from each cluster and collection of the nearest neighbor data of the center of the cluster. But we didn't give any relationship about data and cluster number with random selection and nearest neighbor data of the center of the cluster. For first objective we already said that we can easily avoid 50% data from each cluster but sometimes better accuracy achieved from more or less than 50% of the random collection. Similarly, by increasing the number of the cluster, accuracy also increase but at which point we need to end up we didn't fix it.

5.2 Conclusion

In this paper, we have addressed a research problem that is labeling unlabeled big data is very costly, expensive, time consuming. To solve this problem our main objective was to find the informative unlabeled instances and label them with the help of oracle and scaling up the classification results using the machine learning algorithms. We have proposed a framework that selectively labels instance from each cluster based on the nearest neighbor of the center of the cluster and by selecting random instances from each cluster. We have used clustering technique to find the informative unlabeled instances and labeling them by experts/users to improve the process in active learning. The experimental analysis have proved that our proposed framework address the above mentioned research challenges using a 50% of the training instances to guide the instance selection process. In addition, our framework achieved the better classification results than other methods by using a fewer number of informative instances.

5.3 Future Work

In our work we have applied clustering techniques in active learning process for collecting most informative instances. In future, at first we will try to build a relationship with cluster number and number of informative instances. We will apply instance-weighting approach to find the informative data instances and grey box model in active learning to improve the classification results in semi-supervised learning.

Bibliography

[1] "10 vs of big data," https://www.xenonstack.com/images/blog/10-vs-of-big-data. viii, 6

[2] "Global datasphere to hit 175 zettabytes by 2025, idc," https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper, November 27, 2018. viii, 6

[3] "Big data growth," https://blogs.thomsonreuters.com/answerson/big-data-business-opportunities-infographics/, October, 2012. viii, 7

[4] "Big data market size revenue forecast worldwide from 2011 to 2027 (in billion u.s. dollars)," https://www.statista.com/statistics/254266/global-big-data-market-forecast/. viii, 7

[5] C. Küçükkeçeci and A. Yazıcı, "Big data model simulation on a graph database for surveillance in wireless multimedia sensor networks," *Big data research*, vol. 11, pp. 33–43, 2018. viii, 8

[6] "Active learning," https://www.epfl.ch/labs/cvlab/research/machine-learning/page-123199/. viii, 9

[7] 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. viii, 2, 11

[8] 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. viii, 1, 6, 8, 12, 13, 14

[9] R. Sowmya and K. Suneetha, "Data mining with big data," in *Intelligent Systems and Control (ISCO), 2017 11th International Conference on.* IEEE, 2017, pp. 246–250. viii, 6, 9, 14, 15

[10] 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

[11] 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

[12] 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

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

[14] 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, no. 15, pp. 5895–5906, November 2013. 1, 9

[15] D. M. Farid, L. Zhang, C. M. Rahman, M. Hossain, and R. Strachan, "Hybrid decision tree and naïve bayes classifiers for multi-class classification tasks," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1937–1946, March 2014. 1, 9

[16] D. M. Farid and C. M. Rahman, "Mining complex data streams: discretization, attribute selection and classification," *Journal of Advances in Information Technology*, vol. 4, no. 3, pp. 129–135, August 2013. 1

[17] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*, 3rd ed. Morgan Kaufmann, July 2011. 1

[18] S. Xiong, J. Azimi, and X. Z. Fern, "Active learning of constraints for semi-supervised clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 43–54, 2014. 2

[19] D. M. Farid, A. Nowé, and B. Manderick, "A feature grouping method for ensemble clustering of high-dimensional genomic big data," in *Future Technologies Conference*, San Francisco, United States, December 2016, pp. 260–268.

[20] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, June 2010.

[21] C. C. Aggarwal and C. K. Reddy, Eds., *Data Clustering: Algorithms and Applications*, ser. Chapman and Hall/CRC Data Mining and Knowledge Discovery Series. Chapman and Hall/CRC, August 2013. 2

[22] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. Morgan Kaufmann, January 2011. 2, 10

[23] D. M. Farid and C. M. Rahman, "Assigning weights to training instances increases classification accuracy," *International Journal of Data Mining & Knowledge Management Process*, vol. 3, no. 1, pp. 129–135, January 2013. 2, 10

[24] A. Krishnakumar, "Active learning literature survey," Technical Report, University of California, Santa Cruz, Tech. Rep., 2007. 2, 10

[25] A. Lheureux, K. Grolinger, H. F. Elyamany, and M. A. Capretz, "Machine learning with big data: Challenges and approaches," *IEEE Access*, vol. 5, no. 5, pp. 777–797, 2017. 6

[26] R. Krikorian, "Twitter by the numbers, twitter," *On-line]. Available: http://www. slideshare. net/raffikrikorian/twitter-by-thenumbers*, 2010.

[27] M. Than, "Billion devices will wirelessly connect to the internet of everything in 2020," *ABI Research 30*, 2013.

[28] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," *Health information science and systems*, vol. 2, no. 1, p. 3, 2014.

[29] O. Y. Al-Jarrah, P. D. Yoo, S. Muhaidat, G. K. Karagiannidis, and K. Taha, "Efficient machine learning for big data: A review," *Big Data Research*, vol. 2, no. 3, pp. 87–93, 2015. 6

[30] E. Savitz, "Gartner: 10 critical tech trends for the next five years," *October2012¡ http://www. forbes. com/sites/ericsavitz/2012/10/22/gartner-10-critical-tech-trends-for-the-next-five years*, 2012. 6

[31] C. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on big data," *Information Sciences*, vol. 275, pp. 314–347, 2014. 8

[32] D. Laney, "3d data management: Controlling data volume, velocity and variety," *META group research note*, vol. 6, no. 70, p. 1, 2001. 8

[33] J. J. Berman, *Principles of big data: preparing, sharing, and analyzing complex information*. Newnes, 2013. 8

[34] A. Katal, M. Wazid, and R. Goudar, "Big data: issues, challenges, tools and good practices," in *Contemporary Computing (IC3), 2013 Sixth International Conference on*. IEEE, 2013, pp. 404–409. 8, 12

[35] F.-Z. Benjelloun, A. A. Lahcen, and S. Belfkih, "An overview of big data opportunities, applications and tools," in *Intelligent Systems and Computer Vision (ISCV), 2015*. IEEE, 2015, pp. 1–6. 8, 14

[36] M. Herland, T. M. Khoshgoftaar, and R. Wald, "A review of data mining using big data in health informatics," *Journal of Big data*, vol. 1, no. 1, p. 2, 2014. 8

[37] T. Hey and A. E. Trefethen, "The uk e-science core programme and the grid," in *International Conference on Computational Science*. Springer, 2002, pp. 3–21. 9

[38] B. Jacob, M. Brown, K. Fukui, N. Trivedi *et al.*, "Introduction to grid computing," *IBM redbooks*, 2005. 9

[39] J. Lin and A. Kolcz, "Large-scale machine learning at twitter," in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. ACM, 2012, pp. 793–804. 9

[40] A. Smola and S. Narayanamurthy, "An architecture for parallel topic models," *Proceedings of the VLDB Endowment*, vol. 3, no. 1-2, pp. 703–710, 2010.

[41] C.-T. Chu, S. K. Kim, Y.-A. Lin, Y. Yu, G. Bradski, K. Olukotun, and A. Y. Ng, "Map-reduce for machine learning on multicore," in *Advances in neural information processing systems*, 2007, pp. 281–288. 14

[42] B. Panda, J. S. Herbach, S. Basu, and R. J. Bayardo, "of decision tree ensembles," *Scaling Up Machine Learning: Parallel and Distributed Approaches*, 2011. 9

[43] S. Fong, R. Wong, and A. Vasilakos, "Accelerated pso swarm search feature selection for data stream mining big data," *IEEE transactions on services computing*, no. 1, pp. 1–1, 2016. 9

[44] Y. Zhang, S. Chen, Q. Wang, and G. Yu, "i ̂2 mapreduce: Incremental mapreduce for mining evolving big data," *IEEE transactions on knowledge and data engineering*, vol. 27, no. 7, pp. 1906–1919, 2015. 9

[45] A. Ghasemi, H. R. Rabiee, M. Fadaee, M. T. Manzuri, and M. H. Rohban, "Active learning from positive and unlabeled data," in *2011 IEEE 11th International Conference on Data Mining Workshops*. IEEE, 2011, pp. 244–250. 10

[46] M. Zuluaga, A. Krause, and M. Püschel, "ε-pal: an active learning approach to the multi-objective optimization problem," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 3619–3650, 2016. 10

[47] J. Ahrens, B. Hendrickson, G. Long, S. Miller, R. Ross, and D. Williams, "Data-intensive science in the us doe: case studies and future challenges," *Computing in Science & Engineering*, vol. 13, no. 6, pp. 14–24, 2011. 11

[48] V. N. Gudivada, R. Baeza-Yates, and V. V. Raghavan, "Big data: Promises and problems," *Computer*, no. 3, pp. 20–23, 2015. 11

[49] J. Leskovec, A. Rajaraman, and J. D. Ullman, *Mining of massive datasets*. Cambridge university press, 2014. 11

[50] A. Labrinidis and H. V. Jagadish, "Challenges and opportunities with big data," *Proceedings of the VLDB Endowment*, vol. 5, no. 12, pp. 2032–2033, 2012. 12

[51] S. Chaudhuri, U. Dayal, and V. Narasayya, "An overview of business intelligence technology," *Communications of the ACM*, vol. 54, no. 8, pp. 88–98, 2011.

[52] D. Agrawal, P. Bernstein, E. Bertino, S. Davidson, U. Dayal, M. Franklin, and J. Widom, "Challenges and opportunities with big data, a community white paper," 2012. 12

[53] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mobile networks and applications*, vol. 19, no. 2, pp. 171–209, 2014. 12

[54] D. Agrawal, P. Bernstein, E. Bertino, S. Davidson, U. Dayal, M. Franklin, J. Gehrke, L. Haas, A. Halevy, J. Han *et al.*, "Challenges and opportunities with big data 2011-1," 2011. 12

[55] R. T. Kouzes, G. A. Anderson, S. T. Elbert, I. Gorton, and D. K. Gracio, "The changing paradigm of data-intensive computing," *Computer*, vol. 42, no. 1, pp. 26–34, 2009. 12

[56] I. Jahan, N. N. Sharmy, S. Jahan, F. A. Ebha, and N. J. Lisa, "Design of a secure sum protocol using trusted third party system for secure multi-party computations," in *Information and Communication Systems (ICICS), 2015 6th International Conference on.* IEEE, 2015, pp. 136–141. 12

[57] I. Gorton and J. Klein, "Distribution, data, deployment: software architecture convergence in big data systems," *IEEE Software*, vol. 32, no. 3, pp. 78–85, 2015. 12

[58] 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. 12

[59] C. Dou, D. Sun, G. Li, and R. K. Wong, "Active learning with density-initialized decision tree for record matching," in *Proceedings of the 29th International Conference on Scientific and Statistical Database Management.* ACM, 2017, p. 14. 13

[60] E. Lughofer, "Hybrid active learning for reducing the annotation effort of operators in classification systems," *Pattern Recognition*, vol. 45, no. 2, pp. 884–896, 2012. 13, 16

[61] I. Zliobaite, A. Bifet, B. Pfahringer, and G. Holmes, "Active learning with drifting streaming data," *IEEE transactions on neural networks and learning systems*, vol. 25, no. 1, pp. 27–39, 2014. 13

[62] S. Ramírez-Gallego, A. Fernández, S. García, M. Chen, and F. Herrera, "Big data: tutorial and guidelines on information and process fusion for analytics algorithms with mapreduce," *Information Fusion*, vol. 42, pp. 51–61, 2018. 14

[63] B. Demir and L. Bruzzone, "A multiple criteria active learning method for support vector regression," *Pattern recognition*, vol. 47, no. 7, pp. 2558–2567, 2014. 15

[64] H. Guo and W. Wang, "An active learning-based svm multi-class classification model," *Pattern recognition*, vol. 48, no. 5, pp. 1577–1597, 2015. 15

[65] P. Jain and A. Kapoor, "Active learning for large multi-class problems," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on.* IEEE, 2009, pp. 762–769. 15

[66] C. Campbell, N. Cristianini, A. Smola *et al.*, "Query learning with large margin classifiers," in *ICML*, 2000, pp. 111–118. 15

[67] X. Zhang, S. Wang, and X.-c. Yun, "Bidirectional active learning: A two-way exploration into unlabeled and labeled data set." *IEEE Trans. Neural Netw. Learning Syst.*, vol. 26, no. 12, pp. 3034–3044, 2015. 15

[68] K. Dimitriadou, O. Papaemmanouil, and Y. Diao, "Aide: an active learning-based approach for interactive data exploration," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 11, pp. 2842–2856, 2016. 15

[69] C. H. Park and Y. Kang, "An active learning method for data streams with concept drift," in *Big Data (Big Data), 2016 IEEE International Conference on.* IEEE, 2016, pp. 746–752. 16

[70] M.-R. Bouguelia, Y. Belaïd, and A. Belaïd, "An adaptive streaming active learning strategy based on instance weighting," *Pattern Recognition Letters*, vol. 70, pp. 38–44, 2016. 16

[71] W. Hu, W. Hu, N. Xie, and S. Maybank, "Unsupervised active learning based on hierarchical graph-theoretic clustering," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 5, pp. 1147–1161, 2009. 16