INTRODUCTION

Millions of users around the world are using social network sites daily. The analysis of social network data has several potential in real world applications such as friend recommendation system, product recommendation for e-commerce and suspect identification in anti-terrorism. It has been observed that social network data sizes have been growing rapidly. It leads to a need of machine learning algorithms which will scale rapidly with the number of examples in the data set. The computational complexity of such algorithms increases when they are applied on large data sets. There is a need to study such machine algorithms whose training time remains same though the data size increases. Most learning algorithms are medium scale they assume that data can be stored on memory and can be scanned repeatedly. So there is a need of executing machine learning algorithms on distributed environment which will classify data accurately. The social network data is of unstructured type. There is a need of technique which will classify this data accurately within less time.

With a massive datasets in social Networks (e.g. web logs), it would take days to train a machine learning algorithm. Developing, testing, and deploying such a system which would clearly be prohibitive. As the data size is very large, for training this large dataset distributed cloud computing environment can be used.

1.1 Big Data

While Big Data is perhaps not an entirely new concept, it is certainly a hot topic of the modern, digital era. However, it is an evolving concept and any standards created must take into account that what is classed as ‘Big Data’ today is likely to change rapidly over the next few years. Standards must therefore look beyond the ‘here and now’ of how Big Data is currently being used and instead seek to establish frameworks for dealing with data sets that represent a significant logistical challenge[11]. Taking decisions based on large amount of past and current data is one of emerging topic today[4,5]. For example, predicting exchange rates or forecasting stock prices based on past records. This data is enormous in volume. Data is changing with respect to time. Also data format is changing enormously. Decision tree algorithms are more suitable for this kind of data as it gives visual representation of data which is beneficial for analysis. Traditional decision tree algorithms like C4.5, ID3, and Random forest cannot handle
large volume of data due to their large memory requirement [16]. These methods are based on global optima which need to copy whole data onto memory to design a model. So the new approach is required which will take care three dimensions of Big data (Volume, Variety and Velocity). Our approach is based on distributed implementation of Hoeffding trees. Hoeffding trees are based on Hoeffding bound which is used to reduce tree height and increase accuracy. MapReduce architecture is used to implement proposed distributed system.

As the usage and popularity of online social networks is growing rapidly, there is a need to do analysis of social network data. Different web companies like yahoo, Google, Facebook need to do large scale data analysis to obtain information from Big data. This analysis will help different business analytics to get comments or feedback about the products or services. This will help to analyse relationship that exist between different customers or relationship between customer behaviour. Using this data business companies can improve their existing product and service and can also expand their market place.

By processing this large amount data user can take a decision. It is useful for business analyst for doing accurate analysis and for taking correct decisions. Better decisions will help to reduce risks and to reduce cost also. Many machine learning algorithms are available in literature for doing data analysis but these algorithms are not suitable for doing big data analysis.

Handling Big data need to handle data in three different dimensions:
   a) Volume-As data is growing rapidly
   b) Variety-As type of data is diverse
   c) Velocity- As data is arriving continuously fast

Most of the existing systems given in literature handle at most two dimensions at a time like Mahout. It is a distributed machine learning environment it handles volume and variety dimension at a time. Other systems handle variety and velocity at a time.

1.2 Machine Learning Algorithms

In 1959, Arthur Samual has made a comment about machine learning that “Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed.”

In 1997, Tom Mitchell gave a well definition of machine learning. It is “A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.”
Machine Learning is a type of Artificial Intelligence which makes computers able to learn without programmer support. Machine Learning (ML) plays vital role in wide range of critical applications like data mining, image processing, robotics etc.

1.3 Types of Machine Learning Algorithms

1) Supervised Learning

2) Unsupervised Learning

3) Reinforcement Learning

1) Supervised Learning

Target variable is predicted from the set of independent variables. Using set of independent variables, function is generated to map input to desired output. All data is labelled and the algorithms learn to predict output from the input data. Output datasets are provided to train the machine. Examples of supervised learning algorithms are regression, decision trees, random forest, KNN etc.

2) Unsupervised Learning

In Unsupervised learning there is no target or outcome variable to predict. It is used to model structure or pattern of data to learn more about data. In this learning, input variable is provided without output variable. Unsupervised algorithms are further divided into two types

a) Clustering

   Clustering is a technique in which inherent groups of data are discovered. K-means is a clustering algorithm

b) Association

   It is used to discover rules which describe large portion of data. Apriori is an association mining algorithm.

3) Reinforcement Learning

Reinforcement Learning is used to train the machine for specific decisions. Machine is allowed to train itself continually using trial and error. It learn from past experiences and tries to take accurate decision. Example of this is reinforcement learning.
1.4 List of Common Machine Learning Algorithms

Commonly used machine learning algorithms are as given below

1. Linear Regression
2. Logistic Regression
3. Decision Tree
4. SVM
5. Naive Bayes
6. KNN
7. K-Means
8. Random Forest
9. Dimensionality Reduction Algorithms
10. Gradient Boost & Adaboost

1. **Linear Regression**

   It is used to predict value of variable based set of continuous variables. Relationship between dependent and independent variables is established by fitting best line. This best line is called as regression line and its equation is given by $Ax+b=y$

2. **Logistic Regression**

   It is a classification algorithm. It is used to estimate value of discrete variable based on a set of independent variables. It predicts probability of occurrence of event by fitting data to logit function. As it predicts probability, its output lies between 0 and 1.

3. **Decision Tree**

   It is a supervised algorithm used for classification. It is used for classification of continuous as well as categorical data. It is used to divide population into two or more sets based on most significant attribute.

4. **Support Vector Machine**

   It is used for classification of data. In this algorithm, each data item is plotted as a point in n-dimensional space. Data item value will denote coordinate of the point. Line will be drawn to divide data items into two linearly separable groups.
5. **Naïve Bayes**

   It is a classification algorithm. It is based on Bayes theorem. It is easy to build and useful large set of data. It assume that value of particular variable is independent of any other variable of feature set. It assigns class labels to problem instances which are represented as vector of feature values and class labels are drawn from some finite set. Bayes theorem provides way to find posterior probability $P(c \mid x)$ from $P(c), P(x)$ and $P(x|c)$.

   $$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$  \hspace{1cm} (1.1)

   $P(c|x)$ is posterior probability of class $c$ given attribute $x$.

   $P(c)$ is posterior probability of class $c$

   $P(x|c)$ is likelihood that is probability of predictor $x$ given class $c$

   $P(x)$ is the prior probability of predictor $x$

6. **KNN (K Nearest Neighbour)**

   It is used for classification as well as regression purpose. It stores all available cases and classify new cases by taking votes from $k$ neighbours. Class being assigned to new case is calculated by using distance function. Distance function can be Euclidean and hamming distance. Euclidean function can be used for continuous variables while hamming can be used categorical values.

7. **K-Means**

   It is a type of unsupervised learning. It solves clustering problem. It is used to classify given data through different clusters.

8. **Random Forest**

   It is a classification and regression method. In random forest, collections of trees are used. To classify new object based on attributes, each tree gives a classification. It is implemented by creating multiple decision trees at training time and output will be class which is mode of classes individual trees.

9. **Dimensionality Reduction Algorithms**

   Day by day data is increasing due to voluminous online transactions, ecommerce usage etc.
More details of data are available. Data contains large amount of features. So, it becomes difficult to create robust model. In such cases, dimensionality reduction algorithms are used. These algorithms can be combined with other classification algorithms like decision trees, random forest etc. These algorithms are based on correlation matrix, missing values ratio etc.

10. Gradient Boosting and AdaBoost

Gradient Boosting and AdaBoost is used to make accurate prediction from plenty of data. It combines multiple weak classifiers to build one strong classifier.

1.5 Machine Learning Algorithms used for Classification and Prediction

Classification and prediction algorithms are used to decide to which set of categories a new example will belong. It is done based on training set of observations whose category membership is known. Classification is an example of pattern recognition. It is a supervised type machine learning. An algorithm that implements classification is called classifier.

Classification is of two type’s binary and multiclass classification. Binary classification is classification into two classes while multiclass classification is classification into multiple classes. Classification and prediction algorithm is demonstrated in below figure 1.1.

![Figure 1.1: Standard Classifier](image)

Classifier performance depends on characteristics of data. There is no single algorithm which works best for all types of problems.

1.6 Machine Learning Benefits

- Machine learning algorithms are often accurate than human drafted rules/models
- Don’t need human expert r programmer
- Automatic method to search for hypothesis explaining data
It is cheap and flexible

It can be applied to any training data.

1.7 Distributed Machine Learning

In this Big data era distributed machine learning has become more important. As data is growing in terms of volume, variety and velocity, sequential machine learning algorithms outperforms. Technology is growing rapidly, so, large capacity processors, RAMs are available with machines. Sequential machine learning algorithms does not have capacity to utilize computational power of modern computers. So, there is need to design distributed machine learning algorithms which will make use of whole computational power as well as time required to train the machine learning algorithms will get reduced.

1.8 Distributed Machine Learning Framework

Mahut [76] is one of the distributed Machine Learning Framework. It run on the top of HADOOP. Different machine learning algorithms libraries are available in Mahut. Classification, clustering, mining algorithms are implemented in Mahut. It is used to process large amount of data. Spark[77] is another distributed Machine Learning Framework. It is mainly used for real time big data analysis. MLlib is an apache sparks library which contains different classification, clustering, pattern mining algorithms. It provides high-level APIs in Java, Scala and Python, and an optimized engine that supports general execution graphs. It also supports a rich set of higher-level tools including Spark SQL for SQL and structured data processing, MLlib for machine learning, GraphX for graph processing, and Spark Streaming.

1.9 Streaming Machine Learning

To process large amount of data, streaming machine learning algorithms are used. Streaming machine learning algorithm reads real time data stream and apply machine learning algorithms onto it. It is characterized by high data rate and volume like social media data, ATM logs etc. This stream is infinite in nature.
So, analysis of this data is not possible in one pass due to inability memory or disk to hold large number of records. Streaming machine learning algorithms are also called as online algorithms. Different frameworks are available for streaming machine learning. MOA (Massive Online Analysis) is one of the popular streaming machine learning API. It contains different classification, clustering machine learning algorithms. MOA has their own stream generator classes.

1.10 Rationale and Significance of Work

This problem statement proposes a system which will take care of all three dimensions (Volume, Variety & velocity) of Big data by using distributed Machine Language framework. When big data is processed and stored, additional dimensions come into play, such as governance, security, and policies. Choosing an architecture and building an appropriate big data solution is challenging because so many factors have to be considered.

1.11 Motivation

Extracting Knowledge efficiently from massive and online data has been a popular research topic. This data is generated and changed very rapidly. Classification models are required which will accommodate data on the fly. Models are required to read one example only once. These models need to presents knowledge in tree/rules like format that can be understood by both human and machine. So decisions can be taken very rapidly and easily. Traditional machine learning algorithms are static in nature which builds a static model by loading full data. When new data arrive, the whole data (old + new) is reloaded to update algorithm. So, this model is not suitable for voluminous data. There is a need of dynamic induction methods which will be based on incremental learning.
1.12 Problem Definition

Investigating and Designing a Big Data Induction Model (BIM) based on Decision Tree for classification and prediction of data without compromising accuracy with good visualization along with less memory consumption and training time by reading every sample only once.

1.12.1 Problem Scope

Proposed Model is designed for multi label classification of structured as well as unstructured data. It can be used to classify balanced and unbalanced data. It is also designed to do sentiment analysis of real time streams which will classify data in two classes positive and negative. Model can be applied on data in the form of graph only after pre-processing.

Performance of proposed model is checked by comparing its performance with existing ML algorithms by using different data sets [6][7]. Distributed model performance is analysed by comparing it with distributed C4.5 [39].

1.13 Research Objectives

This research work in thesis pertains to the following specific research objectives:

1) To Collect and Pre-process Big data
   To collect data set from various sources like Twitter, MOA, Weka, MEB, UCI Repository, NODEXL and apply pre-processing techniques onto it.

2) To propose algorithm for data cleaning
   To investigate and propose algorithm for cleaning Social Media Data to remove unused, noisy, missing values.

3) To propose algorithm for pattern classification of big data and compare outcome with existing distributed algorithms
   To investigate and propose Big Data Induction Model (BIM) algorithm for pattern classification of big data and compare outcome with existing distributed algorithms.

4) To propose prediction algorithm for big data
   To investigate and propose Big Data Induction Model (BIM) algorithm for prediction of big data and compare outcome with existing distributed algorithms.
5) To propose variety, volume and velocity parameters under big data environment for classification and prediction.

To evaluate variety, volume and velocity parameters of proposed Big Data Induction Model (BIM) algorithm using different parameters like accuracy, time and memory.

1.14 Summary

In chapter 1, introduction of the research topic is explained. Then motivation behind the research is elaborated. Machine learning application for classification and prediction of data along with different types of machine learning is explained. Introduction of different machine learning algorithms is explained in detail.