A dynamic data classification techniques and tools for big data

T Usha Rani¹, CH Sindhu Priyanka² and B S S Monica³

¹,²,³Assistant Professor, Department of CSE, Eluru College of Engineering & Technology, Eluru, India.

Abstract: Big Data is an immense term for working with the large volume and complex data sets. While data set is larger in volume and the conventional processing applications are insufficient subsequently distributed databases are needed. Big data came into reality because the earlier technologies are not able to carry such large amount of data from self-governing sources. To find the important and accurate data from large unstructured data is going to be a difficult task for any user. This is the reason why the classification technique came into picture for big data. With the help of these classification methods unstructured data can be turned into organized form as a result that any user can access the required data easily. These classification techniques can be applied over the big transactional databases to provide data services to users from huge volume of data sets. Classification is one of the aspects of machine learning and there are basically two broad categories one is supervised and other one is unsupervised classification. In this project we worked to study the variants of supervised classification methods upon we discuss some of the tools or frame works to perform classification. For each algorithm we discuss about their advantages and limitations.

1. Introduction
Classification in the way towards find a model which describe and various data classes or concepts. During this model we can derive the class labels which are known for set of training data or data objects by performing analysis. The key objective of this model is to correctly predict the objects class label which is unknown.

The problem given to the Classification [7] should contain some input which is treated as training data having class labels, which is used for determining the class label for unlabeled test data or instance. Classification algorithm consists of two parts. First part is training part and second one is testing part.

- In the first part i.e. training part or learning step, the model was constructed based on the training instances.

- In the second part i.e. testing part or classification step, the unlabeled test instance or given is assigned with a label by using this model.
In some cases like the lazy learning algorithms, will omit the training part entirely therefore the classification on directly performed based on the relationship between the training instances to the test instance. K-nearest neighbour approach is an example for instance-based method. In order to provide efficiency throughout the testing part, nearest neighbour index construction has to perform as a pre-processing part [15]. Either in one of two ways the output of the classification algorithm was given to the test instance:

- When the test instance returns a label then it is known as discrete label.
- By the combination of each and every class label and test instance returns the numerical score. By selecting the class with highest score for the test instance, we can convert numerical score into discrete label.

There is possible to compare relative propensity of test instances which are different and belongs to a particular class of importance and if there is a need we can rank them, these are the main advantage with numerical score.

The classification algorithm is done by defining the unseen test instances into groups like segments. Whereas in clustering is done with the segmentation of examples turned into groups. The main differentiation between these two problems is, in clustering segmentation was done based on the similarities between the feature variables and with not having prior knowledge or understanding the structure of groups. While coming to classification, segmentation was performed based on the training dataset and furthermore knowing about structure on the group where they are in the form of target variable. The result of the classification problem is termed as supervised learning and clustering is termed as unsupervised learning.

The data which is extremely bigger in size and which can't be kept up as far as Mega Byte (MB) and Giga Byte (GB) are named as Big Data. The big data generally in size of Peta Byte (10^15). Different sources of big data are enterprise, public, sensor, social media, and transactions.

In this paper was organized as follows. The next section was discussing about the Data classification techniques. The 3rd Section represents the metrics for evaluating classifier performance. The 4th section describes about the big data issues and tools for big data processing.

2. Data classification techniques
Data preparation is the major issue for classification. This process involves the subsequent activities like selecting, pre-processing and transforming.

2.1. Selecting:
This stage is worried about the selecting the subset of every single accessible data that you will work with. Always there is a strong desire for including the entire data which is accessible, that the “more is better” will hold. You have to consider what data you must essentially need to address the issue or problem you are running on. Make a few assumptions regarding the data that you need and be careful to record those assumptions with the goal that you can be capable of test them later if necessary. Selection of data was determined based on available data, missing data and what data has to be removed [15].

2.2. Pre-processing:
Data pre-processing can sort out the selected data by performing formatting, cleaning, integration, and sampling.

2.3. Data cleaning:
Here missing data was filled, and readdress the invalid data into legitimate one. It identifies the outliers present in the data and inconsistencies are removed from the data source.

2.4. Data integration:
Data from different sources are combined into a single database.

2.5. Data transformation:
Source data was converted into common format which is used for processing.

2.6. Data reduction:
It is the way toward removing unwanted parameters from the data. Hence the data volume will be less and in the meantime it won’t endure on the quality of information.

2.7. Data discretization:
The part of data reduction process. It will substitute the numerical attributes with the nominal attributes.

2.8. Transforming:
Transformed data is ready for performing operations on them. There are generally three basic data transformations which are attribute decompositions, attribute aggregations and scaling. Some of the supervised learning algorithms are as follows:

2.9. Decision Trees:
Decision Trees are the dominant classification on algorithms that are largely used in data mining. A Decision tree is described as flow chart similar to tree structure which involves root node, in non-leaf nodes and leaf nodes. Each non leaf node describes about the test an attribute, each branch describes
about an outcome of that test and each leaf node holds the class label [10]. The fundamental thought of the decision tree is to divide the data recursively into subsets with the end goal that every subset was contains approximately homogenous states of target variable [11]. The selection of splitting criterion is, attribute selection measure has to select such that it “best” separates the given dataset. Some of the famous decision tree algorithms are ID3 [5], C4.5 [2] and CART which utilize Information gain, Gain Ratio, and Gini Index as attribute selection measure correspondingly. C5 is as well as a decision tree based on algorithm which was an enhanced version of C4.5 [2]. When the decision tree was constructed, it is utilized to classify another new instance by performing traversing operation from root node to the leaf node by applying the test criteria at each non leaf node. The class of the leaf node is the class for each instance.

Figure 2. Decision tree

Table1. Attribute selection measures

| Measure     | Formula                                                                 |
|-------------|-------------------------------------------------------------------------|
| Information Gain | \[ \text{info}(D) = - \sum_{i=1}^{m} p_i \log_2 (p_i) \]               |
|             | \[ \text{info}_A(D) = \sum_{i=1}^{p} \frac{|D_j|}{|D|} \times \text{info}(D_j) \] |
|             | \[ \text{Gain}(A) = \text{info}(D) - \text{info}_A(D) \]                |
| Gain Ratio  | \[ \text{splitInfo}_A(D) \]                                           |
|             | \[ = - \sum_{i=1}^{p} \frac{|D_j|}{|D|} \times \log_2 \left( \frac{|D_j|}{|D|} \right) \]                  |
|             | \[ \text{GainRatio}(A) = \frac{\text{Gain}(A)}{\text{splitInfo}_A(D)} \]                                |
| Gini Index  | \[ \text{Gini}_A(D) = 1 - \sum_{i=1}^{p} p_i^2 \]                        |
\[
\begin{align*}
\text{Gini}_A(D) &= \frac{|D_1|}{|D|} \text{Gini}(D) \\
&\quad + \frac{|D_2|}{|D|} \text{Gini}(D_2) \\
\Delta \text{Gini}(A) &= \text{Gini}(D) - \text{Gini}_A(D)
\end{align*}
\]

2.9.1 Advantages.
- Decision trees are easy to understand.
- Non-professional users can easily grasp the decision trees with reasonable number of leaves.
- For better understand ability of the users decision trees are converted into rules.
- Decision trees can hold both numerical and nominal attributes

2.9.2 Disadvantages.
- Most of the algorithms like ID3 [5] and C4.5 [2] necessitate that the target attribute will have single discrete values.
- Decision trees utilize the “divide and conquer” method, which are having a propensity to achieve well if a small number of extremely significant attributes exist, except less so if numerous complex interactions are available.

2.10 Rule based classification
Rule based classification is one of the learned models which representation as a set of IF THEN rules. The rules has been composed of two parts namely rule antecedent which represents the If part, and rule consequent - which represents the else part. IF THEN rule is in the form of

IF condition THEN conclusion

The condition mentioned in the above rule was describing about the antecedent has attribute tests and the consequent contains class prediction [12]. Consider an example as the rule R1

\[
R1: \text{High CHD Risk} \leftarrow \text{gender = female} \times \text{age} > 63 \times \text{body mass index} > \frac{25 \text{kg}}{m^2}
\]

The Above rule describes a group of patients who are female, overweight, and older than 63 years as having the high risk of the disease Coronary Heart Disease (CHD) [13]. The rules are derived from the decision trees or from the data by using some sequential covering algorithm for example, in AQ, CN2 and RIPPER [10]. Coverage and accuracy for rule R is defined as follow

\[
\begin{align*}
\text{coverage}(R) &= \frac{n_{covers}}{|D|} \\
\text{accuracy}(R) &= \frac{n_{correct}}{n_{covers}}
\end{align*}
\]

2.10.1 Advantages.
- As exceptionally expression as decision trees.
• Easy to understand.
• Easy to produce.
• Can classify new instances quickly.
• Performance is similar to decision trees.

2.10.2 Disadvantages.
• It produces the rules with very small coverage.
• Rules may not be exhaustive.

2.11 Bayesian classifier
Bayesian classifiers are known as the statistical classifiers [6]. There are used to predict the class membership probabilities for instance probability to the given tuple which belongs to a particular class. Let X is the data tuple which is called as evidence and H is some hypothesis [10]. Baye’s Theorem is useful for calculating the posterior probability, \( P(H|X) \), from \( P(H) \), \( P(X|H) \) and \( P(X) \).

\[
P(H|X) = \frac{P(X|H)P(H)}{P(X)}
\]

2.11.1 Naive bayesian classifier.
It presumes attributes values which may be conditionally independent from another and they does not exists any dependency relationships between the attributes. This is most accurate classifier when the assumption holds true [10, 15].

Assume that there are m classes which are represented as \( C_1, C_2...C_m \). Predicting whether X is belongs to the class have highest posterior probability which was conditioned on X is done by classifier. Classifier will predict the tuple X whether it belongs to particular class \( C_i \) in case.

\[
P(C_i|X) > P(C_j|X) \text{ for } 1 \leq j \leq m, j \neq i
\]

\[
P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}
\]

2.11.2 Bayesian belief networks (BBN).
The naive bayes classifier will assume that the attributes be independent to one another but in real world scenario [6], the attributes may be correlated with the medical domain where a patient’s symptoms and health state are correlated [6]. Actually dependencies are existing and therefore Bayesian belief networks are used to model the dependencies between attributes using joint conditional probability distributions. Bayesian Belief network consists of a) dependencies among the attributes will represented by using DAG (Directed acyclic graph) and b) Conditional Probability Table (CPT) which is associated with each attribute. The nodes represent attributes and the arcs represent dependencies.

Assume X is a data tuple having \( x_1...x_n \) with some attributes \( Y_1 ... Y_n \). Any node which is present in Bayesian on network was more conditionally independent to their non-descendants if their parents were known. Following equation represents existing joint probability distribution.
\[ P(x_1 \ldots x_n) = \prod_{k=1}^{n} P(x_i|\text{Parents}(y_i)) \]

Where \( P(X_1 \ldots X_n) \) represents the probability of specific combinations of values for \( X \) and represents entries in conditional probability table of \( Y \).

2.11.3 Advantages.
- High speed.
- Handles both numeric and discrete valued attributes.
- Easy to compute.

2.11.4 Disadvantages.
- Optimal Bayes classifier was computationally intractable.
- Naïve Bayes assumptions are violated usually.
- Does not give the accurate results for some cases, where dependency exists among the variables.

2.12 Logistic Regression

Logistic regression [6] is one of the powerful statistical methods used to analyze the datasets which have at least one independent variable for determining the outcome. By using logistic function we can measure the relationship between one or more independent variables and categorical dependent variable through estimating their probabilities. Outcome is two-valued which is in categorical nature. Furthermore it is used for predicting probability of non occurrence or occurrence of an event.

2.12.1 Advantages.
- It can avoid over fitting.
- Feature selection can be done.
- The output can be interpreted as the probability.
- Robust to the noise.

2.12.2 Disadvantages.
- To achieve meaningful as well as stable results it requires more data.
- Identifying the right independent variables.
- Over fitting.

2.13 K-Nearest Neighbour

k - Nearest Neighbour classifier will represent every tuple as one data node in the d-dimensional space, where d holds number of attributes [6]. By this way each and every tuple was stored and remembered during learning stage. When a new tuple whose class is unknown was given the k-NN classifier compares its proximity with k-nearest training tuples and assigns the class of k-nearest neighbours with the majority vote or distance weighted vote to the new tuple. Some of the widely used proximity measures for finding the nearest neighbours include Euclidean distance, Manhattan distance, Simple Matching coefficient, Cosine similarity, and correlation coefficient. K-NN [6] has a wide range of applications which includes cluster analysis, image analysis, pattern recognition, prediction and economic forecasting.

2.13.1 Advantages.
• Fast training.
• Having Complex target functions.
• No loss of information.
• Easy to implement.

2.13.2 Disadvantages:
• Slow to process the query.
• Needs large storage space for remembering all instances.
• Noise sensitive.
• Slow testing.

2.14 Artificial Neural Networks
Artificial neural network is a computational model dependent on the biological neural systems. It comprises of interconnected processing elements which are known as neurons or nodes which cooperate to work together for producing the output function. It is adaptive framework that changes there structure dependent on internal or external information which flows through the network through learning part [16]. An ANN can be a simple perceptron model or a more complex multilayer perceptron model. This model comprises set of output nodes where each input node connected to the output node through weights. The model is trained by adjusting their weights awaiting they fit their input output relationship to the data [9]. The multilayer ANN [9] consists of several intermediate hidden layers connecting the input, output layers and can be used for modelling the complex relationships among the input, output variables. ANN has been successfully applied in clinical medicine domain in these classification and pattern recognition. Neural Networks with Back propagation and the association rule mining which is used for tumour classification in mammograms [16].

2.14.1 Advantages
• Adaptive learning.
• Self organization.
• Fast prediction.
• Easily identify the complex relationships.
• Can handle the noisy data.

2.14.2 Disadvantages.
• All inputs should be converted into numeric.
• Learning will gives local optimum.
• Over fitting.
• Difficult to interpret.
• Processing time for the large neural networks is high.
• Cannot be initialized with the prior knowledge.

2.15 Support Vector Machine
Support Vector Machines (SVM) was one of the supervised learning algorithms which are the most dominant classification algorithms in provisions of analyze data and predictive accuracy which may be used for both classification and regression analysis. It is used for classifying both linear and nonlinear
data. Initially it was designed for two class problems but later it was used for multi class problem also. The essential principle of SVM [3] was to find optimal hyper plane among max distance to the closest point of the two classes. A set of tuples which are closest to optimal hyper plane was called as support vector [1]. SVM uses these support vectors to find optimal hyper plane. Finding the best hyper plane provide the linear classifier, whereas to classify nonlinear data, the original training data was transformed into higher dimension using nonlinear kernel functions such as polynomial, radial, Gaussian, sigmoid etc. SVMs are applied for the numeric prediction plus classification. They have a wide area of application which includes pattern recognition, medicine, bioinformatics, object recognition and prediction. SVM is used to cluster microarray data and extract associated genes for classifying cancer related documents.

2.15.1 Advantages.
- Deterministic algorithm.
- Uses max marginal hyper plane for classifying the linearly separable data.
- Hard to train.
- Can learn very complex functions by using kernels.

2.15.2 Disadvantages.
- Computationally expensive
- Training process may take more time.
- Can solve the binary class problem.

3. Metrics for evaluating classifier performance
Various measures for evaluating how well or how “accurate” of your classifier is on predicting the class label designed for each tuple. While considering the cases of where the class tuples are more or less evenly distributed, in addition to the case where classes are unbalanced. The classifier evaluation measures are summarized in table2 [18].

| Measure                               | formula       |
|---------------------------------------|---------------|
| Accuracy, recognition rate            | \( \frac{TP + TN}{P + N} \) |
| Error rate, misclassification rate    | \( \frac{FP + FN}{P + N} \) |
| Sensitivity ,true positive rate, recall| \( \frac{TP}{P} \) |
| Specificity, true negative rate       | \( \frac{TN}{N} \) |
| precision                             | \( \frac{TP}{TP + FP} \) |
F, F1, F-score, harmonic mean of precision and recall

\[ F_\beta = \frac{(1 + \beta^2) \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}} \]

Training data is utilized to determine a classifier and after the accuracy is assessed dependent on the resulting learning model can result in ambiguous overoptimistic estimation due to over special of the learning algorithm for the data. Instead, it is enhanced to measure classifier’s accuracy on the test set comprising of class labeled tuples which does not used to train the model. Tuples which belong to the main class of interest are called as positive tuples and all previous tuples are called as negative tuples [18].

The performance criteria for evaluating the classifiers are as follows: confusion matrix is shown in Fig3, classification accuracy, error rate, specificity, sensitivity, precision, recall, ROC curve.

3.1 Confusion matrix

\[
\begin{array}{c|cc}
\text{Actual class} & \text{Yes} & \text{No} & \text{Total} \\
\hline
\text{Yes} & TP & FN & P \\
\text{No} & FP & TN & N \\
\hline
\text{Total} & P & N & P+N \\
\end{array}
\]

Figure 3. Confusion matrix shows totals of positive and negative tuples.

In addition of these measures classifiers can also compare with respect to some additional aspects like speed, robustness, scalability and interpretability [7].

Ensemble methods [15] are used to improve the accuracy of classification. An ensemble is a composite model for classification which is made up of combination of classifiers. The aim of ensemble is creating an enhanced composite classification model by combining the series of k-learned models or base classifiers. Class prediction was done by the ensemble methods by considering the votes of the base classifiers. Some of the popular ensemble methods are:

- Bagging: It stands forbootstrap aggregation.
- Boosting.
- Random forests.

4. Tools for Big Data Processing

Issues [17] related to big data are management, storage and processing issues.

4.1 Management issues
Unmanaged data is constantly considered as unnecessary data [8]. Since the big data is formed through collection of numerous heterogeneous sources with various formats, representations etc, to dealing the big data we require high performance and multi dimensional management tools, or else we are probably going to get the unacceptable results. Variety is one of the properties of big data,
consequently to handle the data with various formats and structures; we need more complex data stores through the features of elasticity and scalability also.

### 4.2 Storage Issues

The more information we have to make more accurate decisions. A fine amount of world’s information exists inside the enormous unstructured big data as per big data professionals. From the above statements we need to understand that how much essential this big data for any business organization to grow up.

Yet, tragically we don’t have any devices to accomplish for storing large amount of data for that we are taking example amount of data as a result our decisions, marketing strategies, recommendation systems etc. seem to be exceptionally poor.

### 4.3 Processing Issues

These days, the on-time results are truly matters a great deal especially for business organizations [8]. In the event that the outcomes are not generated accurately and timely, at that point these outcomes will be of least utilize. In the present situation the vast majority of the organizations has transfer their mode of business from ‘brick and mortar’ mode to online mode with the end goal to handle the order to grab the attention of customers and improve the sales globally which the results in tempest of data.

Huge numbers of tools are accessible for processing the big data. Various existing techniques for analyzing big data with accentuation on three essential emerging tools [8] are namely Map Reduce, Apache Spark, and Storm. The most of the available tools are focussed on batch processing, stream processing, and interactive analysis. The majority of batch processing tools depend on the Apache Hadoop infrastructure, for example, Mahout and Dryad. Stream data applications are predominantly used for analysing the real time data. Various open source tools are Apache Hadoop, Apache pig, Cascading Scribe, Apache HBase, Apache Cassandra and Apache S4. There are many open source initiatives in big data mining most popular are Apache Mahout, MOA, R, PEGASUS, and GraphLab. Some of the Hadoop, Apache spark, and Apache Mahout etc. are listed in table3.

#### Table 3. Tools for big data processing.

| Tool         | Description                                                                 | Features                                                                   |
|--------------|------------------------------------------------------------------------------|                                                                            |
| Apache Hadoop| Apache Hadoop is open source software for reliable, distributed, scalable computing. A framework allowing for the distributed processing of large data sets across clusters of computers, the software library uses simple programming models. | Designed to from single servers to thousands of machines. Includes YARN, and Hadoop map reduce modules. |
| Apache HBase | The Hadoop database, Apache HBase is a distributed, scalable, big data store. Data scientists use this open source tool when they need random, real-time read/write access to Big Data. | Linear and modular scalability. Strictly consistent reads and writes. Automatic and configurable shading of tables. |
| Apache Hive  | An Apache software foundation                                                | Command line tool is provided                                              |
project, Apache Hive began as a level project itself. JDBC driver is provided to connect users to Hive.

| Project       | Description                                                                                                                                                                                                 |
|---------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Apache Mahout | An open source Apache foundation project for machine learning and data mining. Simple and extensible program framework for building scalable algorithms. Includes a wide variety of premade algorithms for scala +Apache Spark etc. |
| Apache Pig    | A platform designed for analyzing large datasets, Apache Pig consists of a high-level that is coupled with infrastructure for evaluating such programs. Of pig Latin include ease of programming, optimization opportunities, and extensibility. |
| Apache Spark  | Apache Spark delivers “lightning fast duster computing,” , and this data scientist tool can access diverse data sources such as HDFS, Cassandra, HBase, and s3. More than so high-level operators make it simple to build parallel apps. |
| Apache Storm  | Apache Storm is a tool for data scientist handles distributed and fault-tolerant real-time computation. It also tackles RPC, and more. Free and open source. Reliability process unbounded data streams for real-time processing. Use with any programming language. |

5. Conclusion
In this paper we gave the summary of classification algorithms like Decision Tree, Rule based classification, Bayesian classifier, logistic regression, KNN, ANN and SVM. Big data is extremely large. These methods can be used to organize all kinds of user needs. We have discussed about different performance measures to classify the data along with the different tools to perform classification algorithms on the era of Big Data. We have listed the advantages and limitation of different classification algorithms to provide scope for further research.

6. References
[1] Kotsiantis S B 2007 Supervised Machine Learning A review of classification techniques Informatics p 249-68.
[2] Wei Dai and Wei Ji 2014 A Map Reduce Implementation of C4.5 Decision Tree Algorithm International Journal of Database Theory and Application SERSC.
[3] Lipo Wang 2005 Support Vector Machine: Theory & Applications Springer Berlin.
[4] G Sathyadevi 2011 Application of CART Algorithm in hepatitis Disease Diagnosis IEEE-International Conference on Recent Trends in Information Technology ICRTIT.
[5] Xinyu Shao, Goujan Zhang, Peigen Li and Yubao Chen November 2001 Application of ID3 algorithm in Knowledge acquistion for tolerance design Journal of Materials Processing
Technology 117 Issue 1-2.

[6] Kalpana Kushwaha and Pinkeshwar Mishra September 2016 A Survey on Data Mining using Machine Learning Techniques IJARCCE 5.

[7] Charu C Aggarwal Data Classification Algorithms and Applications IBM T.J.Watson Research Center Yorktown Heights New York USA.

[8] Althaf Rahaman.Sk, Sai Rajesh.K and Girija Rani K Challenging tools on Research Issues in Big Data Analytics International Journal of Engineering Development and Research 6.

[9] Vidushi Sharma, Sachin Rai and Anurag Dev October 2012 Comprehensive Study of Artificial Neural Networks IJARCSSE 2.

[10] Han J, Kamber M and Pei J 2011 Data Mining: Concepts and Techniques 3rd Edition Morgan Kaufmann p 1-703.

[11] Gosain A and Kumar A 2009 Analysis of health care data using different data mining techniques International Conference on Intelligent Agent & Multi-Agent Systems p 1-6.

[12] Dursun D, Walker G and Kadam A 2005 Predicting breast cancer survivability: a comparison of three data mining methods Artificial Intelligence in Medicine.