Analytical Study of Big Data Classification

Neha Khan[1], Mohd Shahid Husain[2], Manish Madhav Tripathi [3]

Department of CSE, ntegral Universiyty, Lucknow, India

***Abstract*— Data over web has been increasing in the past few decades. The main concern for organizations having large datasets is to automatically mine useful information from massive data .Due to the increasing availability of e-documents and the increasing growth of the www, one of the most important and difficult task of automatically categorization of documents has became a key method for organizing the information and discovery of knowledge. The Proper classification of e-documents, online news, blogs, and digital libraries require text mining, machine learning and natural language processing techniques to extract meaningful knowledge. Recently the studies of text mining are gaining more and more importance because of the availability of the increasing number of the e-documents from a variety of data sources. The aim of this paper is to analyse the efficiency of algorithms as the data increases. We will use c4.5 and naïve algorithm to check out the efficiency as the data increases**

***Keywords*- Naive bayes, Precision, FP rate, evolutionary algorithm, evolutionary algorithm.**

I. INTRODUCTION

The main concern of big data is its volume ,complex data sets that are growing day by day.The big data concept comes into existence because the earlier technologies and algorithms are not able to handle huge amount of data. There should be some methods or mechanism that classify unstructured data into oganised form so that the user can be able to access required data easily .The main challenge in big data is data accessing and computing, semantics and domain knowledge about the big data application..Today many of tha applications are suffer from big data problems which includes analysis of traffic risk and business forecasting etc. Therefore to classifiy the data properly and discovery of knowledge from these resources is an important and difficult area for research. In the present information technology world, there is huge availability of data, this huge amount of data or information is of no use until it contains some useful information With the evolution of information technology and computers, huge amount of data can be stored, analyse and calculate to extract useful information for us. Many technologies are used in data mining such as discovery of association rules.

II. C4.5 ALGORITHM

A. Iterative Dichotomiser 3(ID3)

It is used to generate the decision tree from a given dataset. The algorithm is basically used in machine learning and natural learning processes.

Properties

- The algorithm does not guarantee an optimal solution.
- It can overfit to the training data.
- The algorithm is harder to use on continuous data.

Due to the above shortcomings C4.5 came into existence. It was developed by Ross Quinlan to generate the decision tree .It was basically an extension i.e. the improved version of ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and is also known as a statistical classifier. using the concept of information entropy, C4.5 builds decision trees from a set of training data just like ID3 algorithm does where the training data set is a set (S=s1,s2,s3…..sn) of already classified samples.

B. Refinements

Basically c4.5 algorithm is the improved version of ID3 algorithm. Following are the changes made in C4.5

1) It can Handle both continuous and discrete attributes .C4.5 generates a threshold value and splits the list into attribute value is above the threshold and those values which are less than or equal to the threshold value, hence can handle continuous attributes
2) It can handle training data having missing attribute values. C4.5 allows attribute values used the symbol for missing value. These attribute values are not used for gain and entropy calculations.
3) It can handle attributes with differing costs

III. NAÏVE BAYES ALGORITHM

Naïve bayes is basically a data classification technique based on Bayes Theorem. The algorithm is belongs to a family of algorithms having common principles and is not a single algorithm for training classifiers. The algorithm is highly scalable, requires a number of parameters.

A. Application

The naïve bayes algorithm is used in following areas:

- Naïve bayes models are used in a Real time Prediction.
- These models are generally used for Prediction in Multi class.
- Used for Text classification and Sentiment Analysis
- Widely used for Recommendation System

IV. INTRODUCTION TO WEKA

The University of Waikato, New Zealand developed a data mining tool called as WEKA, to implements data mining algorithms for data classifications. Weka is used to implements

algorithms for data pre-processing, classification, regression, clustering, association rules .It contains various visualization tools for displaying graphs also and is an open source software available online. We can directly implement the algorithms on

use. The explorer window have various classifiers like bayes, lazy, meta and tree etc. that are available in weka. Now click on trees, then choose J48 (c4.5 is also termed as J48) which results in following figure

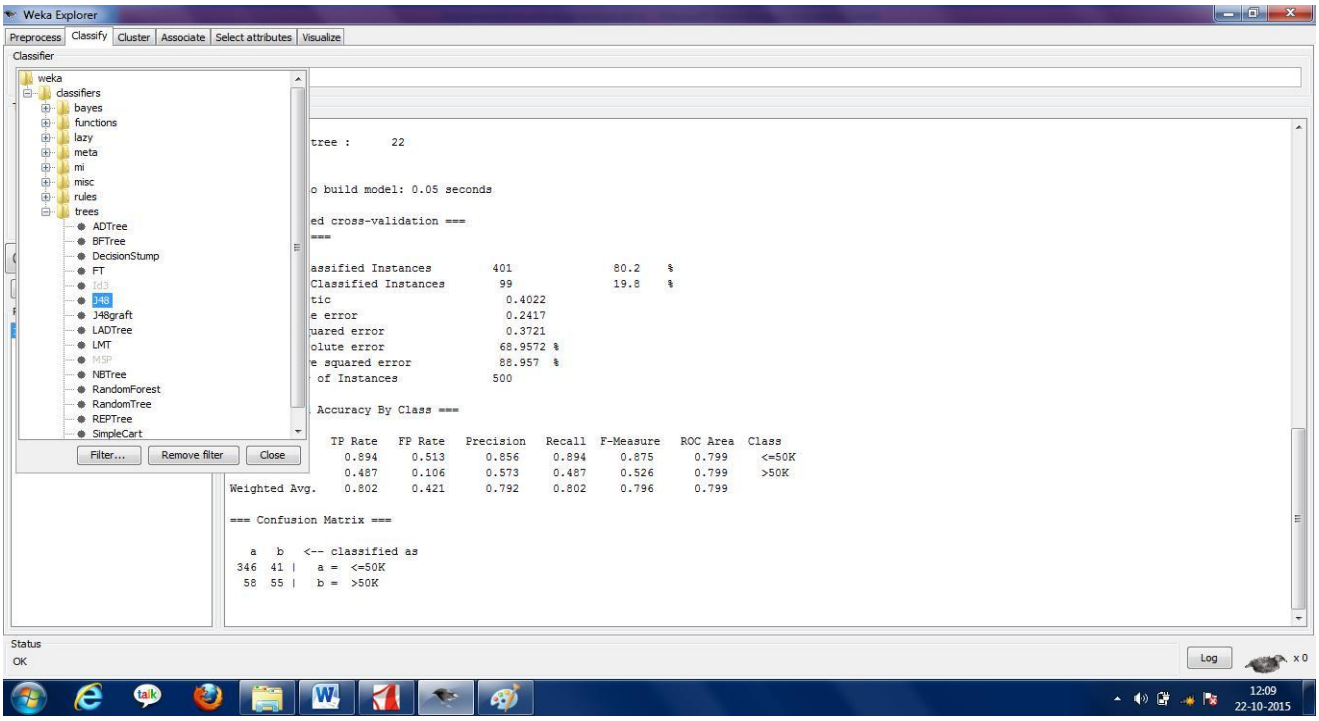


Figure 1: classification panel

a dataset. It can launch from C:\Drive(program files), by clicking on "Start", "Programs" Weka "3-4". The WEKA GUIm(Graphical User Interface) Chooser window appears on the screen, from where we can select one of the four options that appears on the window :

1) **Simple CLI –** CLI is used to provides a simple command- line interface
2) **Explorer -** Explorer is used to provide an environment for exploration of data
3) **Experimenter-** Experimenter is used for performing experiments and conducts statistical tests between various learning schemes.
4) **Knowledge Flow-** Knowledge Flow is based on Java-Beans interface and is also used to set up and running machines

V. Implementation Of C4.5 Algorithm And Naïve Bayes

In order to classify the data in weka, we need to load the dataset on weka first. After opening the weka, weka explorer window appears on the screen, loading of data can be done using window. Here in our implementation we will load dataset named as "Adult" having 500, 2000, 4000, 8000 and16000 instances and 11 attributes. Based on the information contained in the adult dataset, weka will generate a decision tree. Below is the figure showing weka explorer that contains a number of algorithms on the top of the explorer window. We can choose one of the algorithm according to our

After loading the dataset and choosing the algorithm, we click on start button. We classify the data set according to the choosed algorithm, having all the information of number of instances, no. of attributes, class value, time to build the model, FP rate, Precision, F-measure etc.

Weka provides a number of options after classification like decision tree, threshold curve, cost curve etc by clicking on the right on the algorithm. Now we will analyse the efficiency of C4.5 algorithm by increasing the number of instances in the dataset. We will measure the False positive rate (FP rate), Precision and F-measure on increasing the instances and will analyse the efficiency of these parameters.

Similar steps are used to select naïve bayes on weka just like we choose c4.5 algorithm. There is a whole family of baysian is available on weka. Below is the chart and graph of naïve bayes to show efficiency on weka after increasing the data set. (figure follows)

Analysis for C4.5

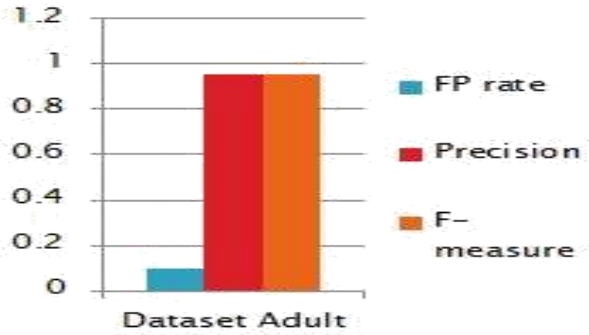
Dataset	FP rate	Precision	F-measure
500	0.421	0.792	0.796
2000	0.057	0.978	0.978
3000	0.013	0.996	0.996
8000	0	1	1
16000	0	1	1
Average	0.098	0.953	0.954

For Dataset Adult

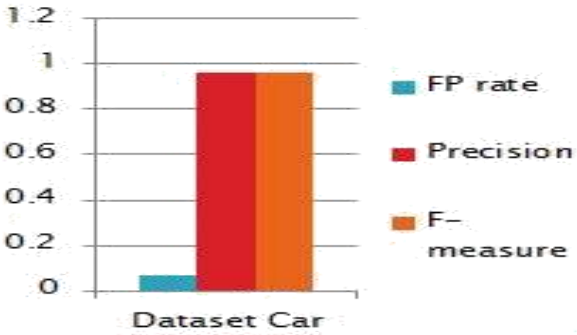
Dataset	FP rate	Precision	F-measure
500	0.098	0.966	0.966
2000	0.001	0.995	0.995
3000	0.237	0.863	0.861
8000	0	1	1
16000	0	1	1
Average	0.067	0.964	0.964

For dataset Car

Chart



For dataset adult



For Dataset Car

Analysis for Naïve bayes

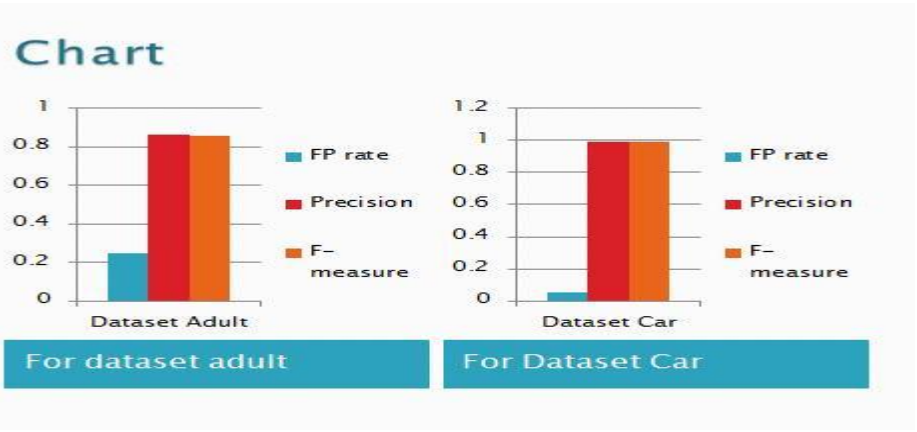
Dataset	FP rate	Precision	F-measure
500	0.29	0.842	0.842
2000	0.244	0.86	0.858
3000	0.237	0.863	0.861
8000	0.229	0.868	0.866
16000	0.228	0.869	0.868
Average	0.245	0.860	0.859

For Dataset Adult

Dataset	FP rate	Precision	F-measure
500	0.131	0.976	0.975
2000	0.042	0.989	0.989
3000	0.056	0.988	0.988
8000	0.024	0.992	0.992
16000	0.024	0.992	0.992
Average	0.055	0.987	0.987

For dataset Car

The graph representation for the naïve bayes algorithm showing the values of parameters as data increases is:



VI. RESULT ANALYSIS

After the implementation of both the algorithms on weka, it is clear from the chart and table that how the FP rate Precision and F-measure are affected as the data increases.

- The False positive rate(FP) rate decreases as the data increases
- The Precision increases as the data increases
- The F-measure increases as the data increases
- The Value of Precison and F-measure increases almost with the same rate as the data increases

VII. CONCLUSION & FUTURE WORK

In this dissertation work we have discuss about the big data in one section, In section 2 we have discuss the evolutionary algorithms that are used for classification of data and in the next section we have discussed about the weka . Our aim is to analyse the efficiency of algorithms on increasing the data. how the parameters changes as the data increases.Finally on implementing C4.5 and Naive Bayes algorithms on Weka ,it is clear from the result that the value of parameters changes as the size of data increases.

- In future we will try to implement the new algorithm on weka which has not implemented yet.
- We will check how effective and efficient this algorithm is, for big data, how the values of parameter changes as the data increases.
- We will also try to compare the algorithms mathematically.

REFERENCES

[1] Liu, Bingwei, et al. "Scalable sentiment classification for Big Data analysis using Naïve Bayes Classifier." Big Data, 2013 IEEE International Conference on. IEEE, 2013.

[2] Suthaharan, Shan. "Big data classification Problems and challenges in network intrusion prediction with machine learning." ACM SIGMETRICS Performance Evaluation Review 41.4, pp. 70-73, 2014.

[3] Singh, Pravesh Kumar, and Mohd Shahid Husain. "Books Reviews using Naıve Bayes and Clustering Classifier." Second International Conference on Emerging Research in Computing, Information, Communication and Applications" (ERCICA-2014), pp. 886-891, 2014.

[4] Ayush Joshi, Jordan Wallwork, khulood Alyahya, Sultanah AlOtaibi, "The Use of Evolutionary Algorithms in Data Mining"

[5] Kumar, Amrender. "ARTIFICIAL NEURAL NETWORKS FOR DATA MINING."

[6] Singh, Pravesh Kumar, and Mohd Shahid Husain. "METHODOLOGICAL STUDY OF OPINION MINING AND SENTIMENT ANALYSIS TECHNIQUES."International Journal on Soft Computing 5(1), (2014).

[7] Singh, Pravesh Kumar, and Mohd Shahid Husain. "ANALYTICAL STUDY OF FEATURE EXTRACTION TECHNIQUES IN OPINION MINING." Computer Science (2013).

[8] Crina Grosan, Ajith Abraham and Monica Chis department of Computer Science,"Swarm Intelligence in Data Mining".