A REVIEW ON CLASSIFICATION METHODS USING BIG DATA ANALYTICS

Bhavna Rawal[1], Dr. Ruchi Agarwal[2]

[1]Department of Computer Science & Engineering, Sharda University, (India)

[2]Department of Computer Science & Engineering, Sharda University, (India)

ABSTRACT

Big data is a source by which we can extract the knowledge and useful information from large amount of data, for analyzing this type of data we require the machine learning techniques. This technology is able to handle the storage of huge data. There are some mechanisms which classify the data into organized form and access useful data for user. So, Classification methods provide required data to user from huge amount of data. And there are two most popular techniques of classification, supervised and unsupervised. But we focused on supervised classification technique and also explain the basics of machine learning to analyze the data. Further this paper compares the methods in terms of advantages and disadvantages.

Keywords: Big Data, Decision Tree, Machine learning, Supervised classification, Support Vector Machine (SVM)

I. INTRODUCTION

Dealing with huge amount of data of different type like structured, unstructured and semi-structured is big data. Big data is required for modeling the data and transforming it into useful information and knowledge with the help of machine learning techniques. There are some challenges to analyze the big data because of its complexity (Li et al., 2014) (Suthaharan, 2014). Big data analytics is process of understanding the features of massive data by extracting the hidden patterns also called knowledge discovery [1]. Big data can be characterized with the help of four V's: Volume, Velocity, Variety and Veracity [IBM]. Where volume means scale of data, variety means different type of data as structured, semi-structured and unstructured data, velocity that is use for analysis of data and veracity use for checking the uncertainty of data. The structured data is a data in which data can be stored in a database in form of row and columns, when data can be stored partially that means semi-structured and unorganized data is unstructured data [2]. Big data is very essential for increasing the storage capacities and processing powers of systems. Previously Wesley Becari and Luana Ruiz experimented on different classifiers and trained and tested those with iCub tactile for sensor grasping data. With the help of this they created an intelligent system which is capable to provide robust solution of sensor problems and textures [8].

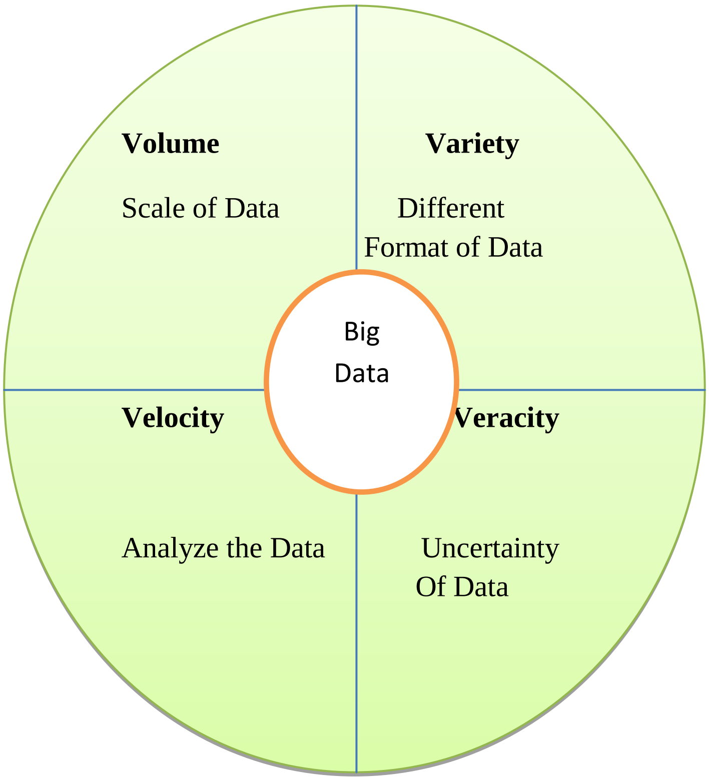

Fig. 1 Features of Big Data

1.1 Machine Learning

Machine learning is very helpful to analyze the big data problems [1]. Machine learning algorithms are used to create the models for better understanding of the proposed system. For this the detailed knowledge is required to analyze the data properly. So, Machine learning provides the number of algorithms and methods which helps in the observations of the system. Learning describes the requirements of the classification model in term of inputs and outputs.

1.2 Classification

Classification is one of the most widely used data mining technique which helps in knowledge discovery and future plans of the classifiers [3]. Classification is an essential task of data mining for solving the problems related to the huge amount of and hidden patterns [1]. It can also provide intelligent decision making. For classification, there are some steps: first learn the model with the help of classification methods then evaluate this to find out the rate of accuracy finally use it. Classification is categorized into supervised and unsupervised classification. But the main focus of this paper is supervised classification.

II. SUPERVISED CLASSIFICATION

Supervised classification is a machine learning task to discover the attribute relationship by creating the learning models [4]. It is required to accurately classify the data. There are the ‘training sets’ to predict the unseen values for correctly classified the data and then categorize according to the similarity. Such technique of data mining is use for specific target values for prediction [5]. To do so it requires the subsets of training data like action and measurement. In this way classification become easier to correctly classify the data. Using these features of training data, the goal of the supervised learning is analyze the large datasets by creating and evaluating the models. There are some methods of supervised classification like Decision Tree and Support Vector Machine and two models a Classification and Regression model [7].

2.1 Decision Tree

Decision tree is used to handle large amount data and for checking the efficiency and accuracy of datasets [6]. Many problems can be solved by creating the tree that uses the meta-learning. DT is beneficial for handling with continuous data, dealing missing values and pruning the tree. It helps to extract knowledge and useful information from large amount of data for accurate classification. DT is a learning process on large datasets for fast and accurate classification. Although decision tree is very time consuming because of huge amount of data. To overcome this there are some DT algorithms like ID3 and C4.5. They are very popular decision tree methods because of their efficient features and provide fast and robust solution for specific problems.

2.2 Support Vector Machine

This supervised learning method also called the support vector network that is used to analyze the data for classification purpose [7]. It is a machine learning technique which performs many task and activities to provide efficient learning and also using nonlinear mapping for classification. It helps to create a classification model for new instances and very helpful in data categorization. SVM can be used for recognize the characters also used for higher accuracy of classification. It is required for handling the complex data and over fitting problems.

III. CONCLUSION

This paper discussed about the supervised classification techniques and their requirements. Decision tree and SVM both are the popular classification techniques where DT is easy to understand and SVM have ability to learn of features of data space, it is fast and robust for classification purpose. They are organizing user requirement by providing the speed, rate of accuracy and many efficient features. There are some limitations of both methods, DT works on separate training sets and SVM required Kernel tricks. In some cases DT is better than SVM and sometime SVM is better than DT. So, both are the essential Classification techniques.

REFERENCES

[1] Jiawei Han and Micheline Kamber-Data Mining: Concepts and Techniques, 3rd edition, first volume, 2011.

[2] M. Saravanan, A.M. Toufeeq, S. Akshaya and V. L. Jayasre Manchari,”Exploring new privacy approaches in a scalable classification framework.” 2014 International Conference on Data Science and Advanced Analytics (DSAA), Oct 2014.

[3] E. Osaba, E. Onieva, A. Moreno, P. Lopez-Garcia, A. Perallos and P.G. Bringas, "Decentrallised intelligent transport system with distributed intelligence based on classification techniques," IET Intelligent Transport System, vol. 10, no. 10, pp. 674-682, Dec. 2016.

[4] M. Rahim, P. Cluciu and S. Bougacha, "Impact of perceptual learning on resting-state fMRI connectivity: A supervised classification study," 2016 24th European Signal Processing Conference (EUSIPCO), Aug. 2016.

[5] Y. Vitthal, Prof. B. Mahip, "Review on Data Mining with Big Data," International Journal of Computer Science and Mobile Computing, Vol.3 Issue. 4,pp. 97-102, April-2014.

[6] S. Sardari and M. Eftekhari,"A fuzzy decision approach for imblanced data classification," 2016 6th International Conference on Computer and Knowledge Engineering (ICCKE), Oct, 2016.

[7] S. Nan, L. Sun, B. Chen, Z. Lin and K-A, Toh, "Density-Dependent Quantized least Squares Support Vector Machine for Large Data Sets," IEEE Transactions on Neural Networks and Learning System, vol. 28, no. 1, pp. 94-106, Jan. 2017.

[8] W. Becari, L. Ruiz, B.G.P.Evaristo and F.J. Ramirez-Fernandez,"Comparative analysis of classification algoriyhms on tactile sensors."2016 IEEE International Symposium on Consumer Electronics (ISCE),sep. 2016.