ALGORITHMS FOR BIG DATA AND DATA SCIENCE

K.S. Praveenkumar, R. Gunasundari*

Abstract

The major challenge of big data in research, industry and society is the availability of suitable technology and its proper usage. This is one of the important questions in Information Management. Another problem related to big data is its storage, security and its applications. In many applications, the amount of data does not make sense, but way it is used matters. Organizations must be capable to utilize the huge data for the advantage of their organization's growth. There are some algorithmic methods that provide better results for analysing a large amount of data. In many of the application areas, efficient algorithms will be essential in order to obtain the required efficiency. Security and storage remain critical issues for IT departments to manage the large quantity of data. Organizations must find an adaptable mechanism to validate and use the data by keeping its confidential nature. When the data become too large, the organization has to face more difficulty in processing it. But it is known that this is where data science comes in. Many organizations have faced this problem and identified that they can generate more accurate results or predictions, if data science is used in a better way. Generally, statistical algorithms are used to sort, classify and process data. The paper discusses data science and some of the algorithms used in.

Keywords: Data Science, Classification, Regression, Similarity Matching, Data Science Algorithms.

I INTRODUCTION

Data are everywhere and the generating sources are limitless. The growth rate of data almost doubles every two years. Such data can be unstructured, structured and semi-structured. Data Science is the area that manages all items related to data such as data cleaning, preparation, and analysis. Data Science combines the best of statistics, mathematics, programming and problem-solving. The traditional data base applications are not capable of handling big data. Big Data processing begins with raw data which are not summarized and so are difficult to in the memory of a traditional computer. In a majority of situations Big Data can be used for making better decisions and strategic plans for an organization. Data Analytics is the science that analyses all the available data and to reach a conclusion. The conclusion can be in terms of a prediction, information, or a strategic plan. The purpose of data analysis depends on the nature of the, organization and its requirements.

II Big Data and Data Science

The name Big Data points to the large amounts of data produced from different data sources and in different formats. Big data is very popular because it can manage structured, unstructured and semi-structured data from a number of different sources and in different formats [1]. This is achieved through the use of advanced data analysis tools.

Data Science is a mixture of a number of tools, algorithms, and machine learning principles. Its goal can be vary depending on the objective of the organization. Data Science analyses a problem in different ways such as a new processes for data modelling, new prototype model, new algorithm, predictive model, etc. to arrive at a conclusion. The person who performs a deep analysis of data is called

Department of Computer Science,
Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India
*Corresponding Author

data scientist. He can use different new machine learning algorithms to identify the presence of a specified event or a particular item in the future. This involves identifying unknown patterns, correlations, and market trends.

In data analysis the roles and responsibilities of a Data Analyst and a Data Scientist are almost similar. Even though they are performing similar duties, they differ in the implementation part. Data Science handles data basically with slicing and dicing of a large quantity of data. Beside these operations they use different techniques to obtain hidden patterns and useful information from the data. The responsibilities of Data Scientists are to open up the hidden facts in the complex web of unstructured data. This type of data or information helps business people or organizations to set their market goal. The Data Science applications and the technologies can improve the concept of Machine Learning models of visualized data.Data scientists need to be able to combine complex data which are available in different forms by using proper analysis and the scientific method. There are three methodologies in applied data science, namely classification, regression and similarity matching.

III CLASSIFICATION

It is an important data mining method and it is a process of assigning items in a collection to target categories or classes. The main objective of classification is to predict the target class for each case in the data. In the case of classification, if the class assignments are known, then the classification task begins with a data set.To determine the target class the training dataset can be used. This training dataset will set boundary conditions which are used to determine each target class. When the boundary conditions are determined, the next step is to predict the target class. The entire process is known as classification.

There are a number of classification algorithms such as Naive Bayes classifier, k-nearest neighbour, Decision trees etc.

IV REGRESSION

The most-commonly used forecasting method is the Regression method. Regression can be confused with classification methods because the process of using known values to predict an outcome is the same. Regression is actually a form of machine learning approach. It is mainly used to predict a continuous value based on some variables. It is a form of supervised learning strategy. In supervised learning, a model is used to identify some features from the existing data. Based on the existing data the regression model creates its own knowledge base [2].Based on this existing data model or the knowledge base the model can make predictions for outcomes on new data in future.In other words, regression analysis investigates the relationship between target (dependent) and predictor (independent variable). Different kinds of regression techniques are used to make predictions. All the techniques are based on three metrics, namely shape of regression line, type of dependent variables and type of independent variables. Among different techniques the most commonly used regressions are,stepwise regression,polynomial regressionlinear regression, logistic regression, ridge regression, elasticnet regression,and lasso regression.

V SIMILARITY MATCHING

The technique of similarity matching is to identify similar data or individual items based on the information known about the item or data. If two entities, -it can be products, services, or companies-are similar in some way they share other characteristics as well [3]. This can be used to find customers for the targeted marketing campaigns or for managing the company's image with targeted online ads. Data scientists can use these principles in algorithms, to analyze massive amounts of data. Algorithms can be developed from statistical models, which are helpful for interpreting graphical models, where there are multiple

unknowns and some special dependency or relationship exists between the unknowns. Some very influencing algorithms mostly depend on unsupervised machine-learning so they can refine their effectiveness as they are used, despite the depth of the data and the number of unknowns.

VI ALGORITHMS USED IN DATA SCIENCE

Nowadays decision makers strongly believe in the analysis of their data to predict future or to predict decisions. How to analyse such a massive amount of data-, is a - problem. In data science, different number of algorithms are built on statistical models for data scientists to satisfy these needs. Which algorithm is chosen is based on the goal of the organization. There are many algorithms, based on Classification, Regression, and Similarity.

6.1 Clustering Algorithm – K-Means

Among the sets of unsupervised learning,K-means clustering algorithm is an important one. This algorithm can be used for known clustering issues.It requires two inputs, namely-'k' (number of clusters) and the training set (m) = {x1, x2, x3,……..., xm}.

The algorithm focuses to divide 'n' observations into 'k' clusters. Each observed data set belongs to the cluster with the nearest mean, and it serves as a prototype of the cluster. This algorithm tries to create an inter-cluster similarity and also retaining difference between intra-clusters as much as possible.

6.2 Association Rule Mining Algorithm

Association rule mining[4]is one of the data mining methods. It is used to identify associations among items or itemsets. In today's big data environment, association rule mining can be used with big data. Association rules are conditional statements. They use "if –then" rules to discover relationships between unrelated data in a relational database repository [5]. Association rule is based on two values:- namely support value and confidence values of data set. The algorithm has to satisfy the predefined minimum support and confidence as they are set by the user.

6.3 Linear Regression Algorithms

As mentioned earlier, regression is a technique used for predicting continuous values based on certain inputs. Linear regression is used to fit data that can fit into a straight linear line. It includes the category of machine learning regression algorithm. This algorithm helps to point out patterns from a set of data. The entire process of this algorithm discovers the mean value out of a given data set. The algorithm uses least squares function for calculating the mean, and then maps it out onto the rest of the data points. The remaining functions can be used to connect data points and to smooth out the differences between the points. [6].

6.4 Logistic Regression Algorithms

Logistic regression is a statistical analysis tool-is mainly used for predictive analysis. Its application can be extended to machine learning also. Like any other algorithm, logistic regression algorithm requires input data. Since it is used for predictive analysis, the input data may be a historical data. Like other algorithms, it is one of the important tools in the filed ML. Depending on the quality of input data or historical data the algorithm can do its prediction well. This type of prediction method can play an important role in data preparation activities. In short, the regression model algorithm predicts by identifying the relationship between the input data and the existing new data. For example, the logistic regression can be used to predict whether a high school student will be admitted to a particular college.

6.5 C4.5

C4.5algorithm is a standard algorithm which is used for implementing classification rules in the form of decision tree. It is an expansion of ID3 [7] algorithm. The C4.5

algorithm is based on splitting of its attribute set, and the splitting is based on information gain ratio. As already said, it is an extension of ID3 algorithm, but in ID3 the attribute splitting is based on the information gain of a number of objects. The use of information gain ration of attributes helps the C4.5 algorithm to avoid attributes with many values.[8]. The algorithm generates a tree and the development or growth of the tree is based on depth-first strategy. In a large data set, there can be a number of noisy data, numeric attributes, and missing values. The algorithm can handle all these types of data. This is one of the advantages of C4.5 algorithm. Sometimes the algorithm needs to handle continuous valued data. In this situation the algorithm creates a threshold value for its variables and splits the variables based on the threshold value. The two sets are, attributes whose value is above the threshold value and attributes whose value is below the threshold value. After creating a tree-like structure the algorithm improves the branches of the tree by applying some pruning function. Finally the algorithm creates a decision tree.

6.6 Support vector machine (SVM)

In the category of Supervised Learning algorithms, SVM (Support Vector Machine) is an important machine learning algorithm. The SVM is considered as the tool used for Regression problems and Classification problems. It is known that the main objectives of Machine learning are classification and prediction. There are a number of machine learning algorithms available to do so according to the dataset. SVM is considered as the linear model for classification and regression problems. At the same time this algorithm can solve linear and non-linear problems too good results for a number of practical problems. The basic idea of SVM is to classify the data when the algorithm creates a line. By using this line the algorithm separates the data into classes. It is also an algorithm that splits the data into different sets based on the input data set and the generated line. The quality of the algorithm depends on the creation of the line. This line splits the data set into different classes.Some studies say that SVM is not suitable for large datasets. If the data contain more noise, SVM cannot provide good result. That means if the data contains more noise , the target classes will overlap the line threshold. At the same time the algorithm works better if there is a clear separation between the classes. Also, the algorithm supports well, when the number of dimensions is larger than the number of samples taken.

6.7 Apriori

The most significant problem of data mining is the frequent item set mining on big datasets. The bestknown basic algorithm for frequent mining item set is Apriori. Apriori algorithm is one of the old algorithms which was developed by R Agrawal and R.Srikant in 1994. It is mainly for the data set where frequent itemsets and relevant association rules are used in data mining. The algorithm is fully based on generation of frequent item set. The algorithm ends when the frequent itemsets cannot be extended next. The main advantage of this algorithm is it's easy for implementation. But apriori algorithm is very slow compared to other algorithms of same range. As the name implies the algorithm uses previous knowledge of frequent itemsets properties to reach conclusions. To find 'k' number of frequent itemsets, this algorithm uses 'k+1' itemsets.

6.8 Naive Bayesian

Naive Bayes is one of the probabilistic algorithms mainly used for classification problems and is based on Bayes' Theorem. It is also one of the classification techniques. The theory behind the algorithm is very simple; Naive Bayes classifier assumes the presence of a special feature in a class and it is unrelated to the presence of any other feature in the class. Like Apriori algorithm it is also easy to implement. But this algorithm is much faster than Apriori algorithm. Another advantage of this algorithm is that it requires fewer training datasets. The Naive Bayseian

algorithm can make probabilistic prediction and it can be used for both continuous and discrete data.

VII CONCLUSION

All the above methodologies or algorithms are data science tools for doing statistical analysis to large data sets. All the algorithms are basically doing three things: classifying data, identifying similarities, and predicting trends. Data Science is one of the areas where Big Data is analyzed in an effective way. With the effective analysis using suitable tools, analyse can be done on large data stored in a complex system. Without the help of a proper tool, it will be very difficult to interpret huge data into meaningful information. The rapid increase of data leads to the invention of new technologies. Organizations can choose right algorithms which provide good result depending on their needs. The insights gained from Big Data enable organizations to be driven by business intelligence and insight, which themselves are driven by complex metrics and analyses. Data Science is here to stay as a necessary part of the Big Data toolset.

REFERENCES

[1] A survey on big data management and job scheduling. Sreedhar C, Kasiviswanath N, Reddy PC Int J Comput Appl. 2015;130(13):41–49

[2] Book Big Data Analytics with Java by Rajat Mehta, chapter 5

[3] Book: Data Classification: Algorithms and Applications, by CharuAggarwal, CRC Press, 2014, chapter 2

[4] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. In: ACM SIGMOD conference. New York City: ACM Press; 1993. p. 207–16

[5] Dr. R NedunchezhianDirector of Research KIT – **Kalaignar Karunanidhi Institute of Technology Coimbatore**, K.Geethanandhini PG Scholar Department of CSE KIT – **Kalaignar Karunanidhi Institute of Technology** Coimbatore. Association Rule Mining on Big Data –A Survey. International Journal of Engineering Research & Technology (IJERT)ISSN: 2278-0181 Vol. 5 Issue 05, May-2016

[6] Book: Data Classification: Algorithms and Applications, by CharuAggarwal, CRC Press, 2014, chapter 2

[7] R. Quinlan,"Improved use of continuous attributes in C4.5", arXiv preprint cs/9603103, (1996).

[8] Wei Dai and Wei Ji, School of Economics and Management, Hubei Polytechnic University, Huangshi 435003, Hubei, P.R.China A Map Reduce Implementation of C4.5 Decision Tree Algorithm, International Journal of Database Theory and ApplicationVol.7, No.1(2014), pp.49-60 http://dx.doi.org/ 10.14257/ijdta.2014.7.1.05.