Big Data Algorithms – Literature Review

Dr. R.K.Nandhini,
Assistant Professor,
Department of Computer Science,
Chikkanna Government Arts college,
Tirupur, India.
krishnandhini@yahoo.com

Ramesh Balasubramaniam,
Research Scholar,
Department of Computer Science,
Chikkanna Government Arts college,
Tirupur, India.
Rameshbala50@gmail.com

Abstract

Collection and storage of data for processing and analysis has been around for a very long time, since the inception of File or Database Systems. Big Data describes the large volume of data that bombards businesses today. Today's data has grown exponentially which in itself is a big topic of discussion – Big Data: Boon or Ban. The significance of data is not in its raw form but in the way the data is being used by organizations. Analyzing Big Data, businesses can derive strategic industry decisions and receive insights to new product or service innovations. This paper aims to analyze some of the different algorithms and methods scientists can use to evaluate Big Data and touch upon other authors views on this topic in our literature review section.

Keyword: Big data, algorithms, data mining, analytics, decision making

INTRODUCTION

A 2016 poll conducted by KDnuggets [1] showsthat average usage of algorithms and methods has increased by 8.1% since a similar poll was conducted in 2011. What algorithms do data scientists use? Figure 1 below shows the top 10 algorithms/methods and their share of voters.The top methods are Regression, Clustering, Decision Trees/Rules, and Visualization. Most popular new options in 2016, K-nearest neighbors, PCA and Random forests.

Not all data is Big Data requiring the tools and advanced capabilities that Big Data applications typically require. Still advanced analytics can play an important role in enhancing customer experience, reducing cost, better targeted marketing and making existing processes more efficient. Knowing what to do with the data gives a key competitive advantage for businesses over their competitors. There are several algorithms built on statistical models that data scientists use to create analytical platforms. Below we will address some of the most commonly used big data algorithms then other authors views on algorithms round it off with our views.

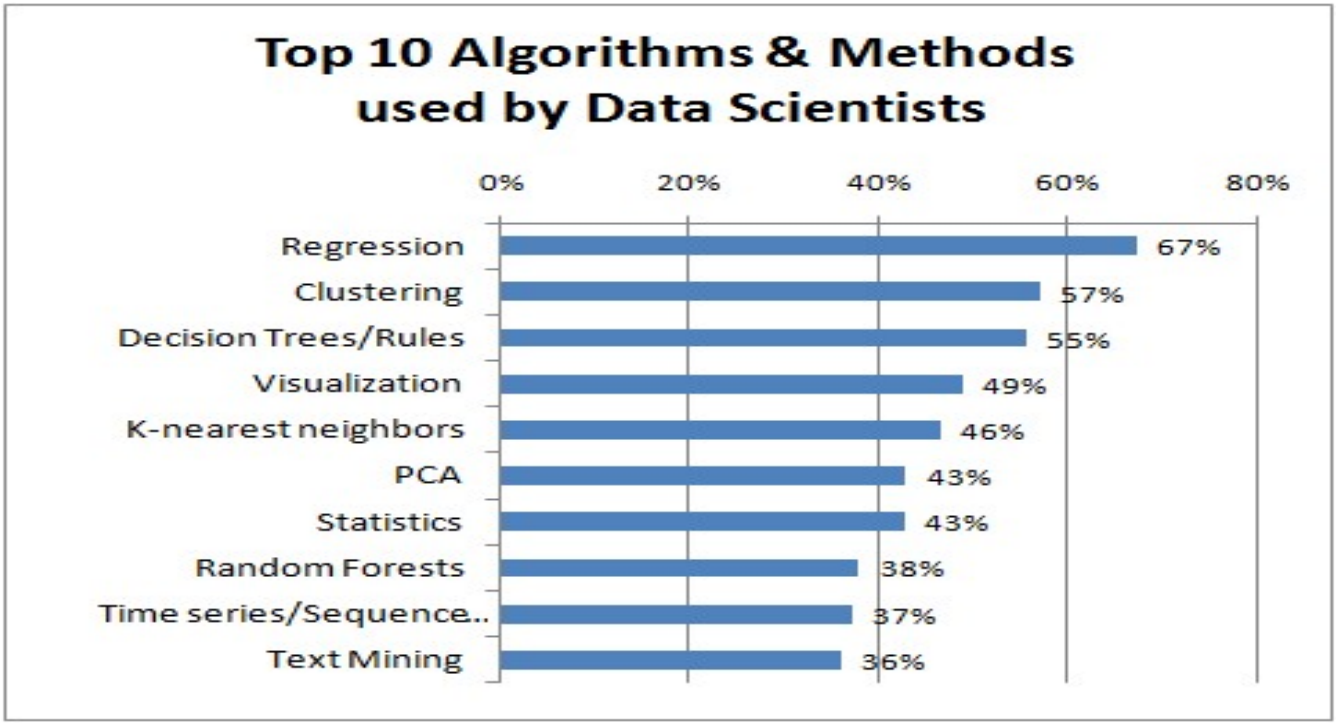

Figure 1. Top 10 Algorithms & Methods

MOST COMMONLYUSED ALGORITHMS

Regression Algorithms

There are many types of regression algorithms - Linear, Logistic, Ridge, Lasso, Logic and more. The oldest and most commonly known is linear regression.

This algorithm is used to find approximate real values based on predictors or independent variables. It is used to estimate the strength and direction of the relationship between variables that are *linearly* related to each other [2]. A relationship between linear variables are established by plotting a best line given by the equation

$Y = aX + b$

where

a is the *slope*, or the change in *Y* due to a given change in *X*

b is the *intercept*, or the value of *Y* when $X = 0$

But is linear regression a technique worthy of using on very large data? There are different opinions. In general, it is used on computations (on small data) that can easily be carried out by a human being by design. Ridge regression is a more robust version of linear regression, putting constraints on regression coefficients to make them much more natural, less subject to over-fitting, and easier to interpret. [3]Bayesian Regression assumes prior knowledge about the regression coefficients. This method is designed for rational incorporation of prior information (information external to the data) into the process of analysis. By doing so, it offers solutions on how to analyze multiple exposures [4]

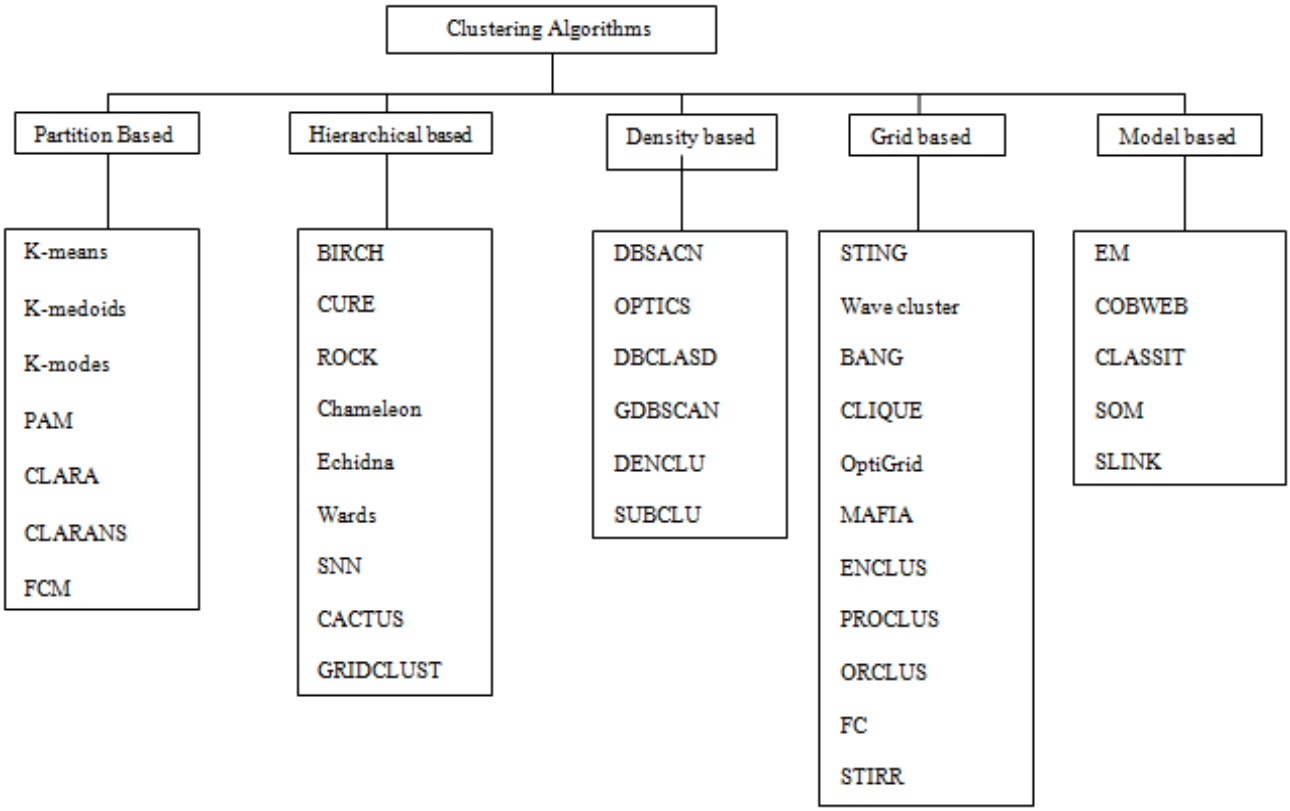

Figure 2. Various Clustering Algorithms

Clustering

Clustering is the task of grouping a set of objects in such a way that objects in thesame group are more similar to each other than to those in other groups. Below Figure 2 shows various Clustering Algorithms classifiedaccording to different categories. It is a main task of exploratory mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics [5]. One of the most popular clustering methods is k-means

K-Means Algorithm

K-means algorithm is an effective and powerful method in exploring the structure in a data set. K-means clustering algorithm partitions the dataset into " k " number of subsets [7]. In K-Means each cluster is represented by itscentroid. Centroids can be considered as the average point, also can be called as mean of points within the cluster. K-means provide effective result on dataset with numeric attributes, whereas noise and outliers affects the effectiveness and efficiency of the algorithm. K-mean classifies n instance in a dataset to k clusters, it finds an optimal solution which minimizes the objective function value. [20] K-means clustering algorithm is based on assignment step which finds the nearest cluster for each point using distance metrics between the point and the cluster center. In updating step the computing of cluster centers will be performed based on the current clustermember points.

The basic algorithm is very simple:
1. Select K points as initial centroids.
2. Repeat
3. Form K clusters by assigning each point to its closest centroid.
4. Recomputed the centroid of each cluster until centroid does not change.

For example, if a hospital wants to open emergency units in a state, the accident-prone regions should be identified and located. It should be ensured that these units are at a minimum distance from the accident zone and at a significant distance from each other. For this, clusters can be identified in such a way that their centroids define the placement of the emergency units. [6]

Over the past years, various extensions of the classicalk-means algorithm have been developed, for example,kernel k-means [9], spherical k-means [10], Minkowskimetric weighted k-means [11], fuzzy c-means [12] etc. Themajority of them is modified to speed up calculations orfor specific tasks. Due to its low computational cost andeasily parallelized process, the classical k-means algorithmis well known for its efficiency in clustering large data sets,but some modifications of k-means are introduced as veryspecific tools for big data analysis.

Decision Tree

It is a supervised learning algorithm that is used in classification problems. A very popular algorithm among data scientists for data mining, decision tree works by splitting a population in as many distinct ways as possible. The construction of a decision tree does not require any domain knowledge or parameter setting, and therefore appropriate for exploratory knowledge discovery. As it is in tree form it is easy to understand and assimilate by humans. [6]

Classification is done by using the most significant attributes (independent variables) at each level to form homogenous groups. An added advantage is that it will work for categorical as well as continuous dependent variables.[8] Given below Figure 3 illustrates a typical decision tree application. Here a given population is observed and classified to understand routine customer patterns better and improve the existing business model. For heterogeneous groups, techniques such as Gini impurity, information gain, variable reduction, etc. are also used.

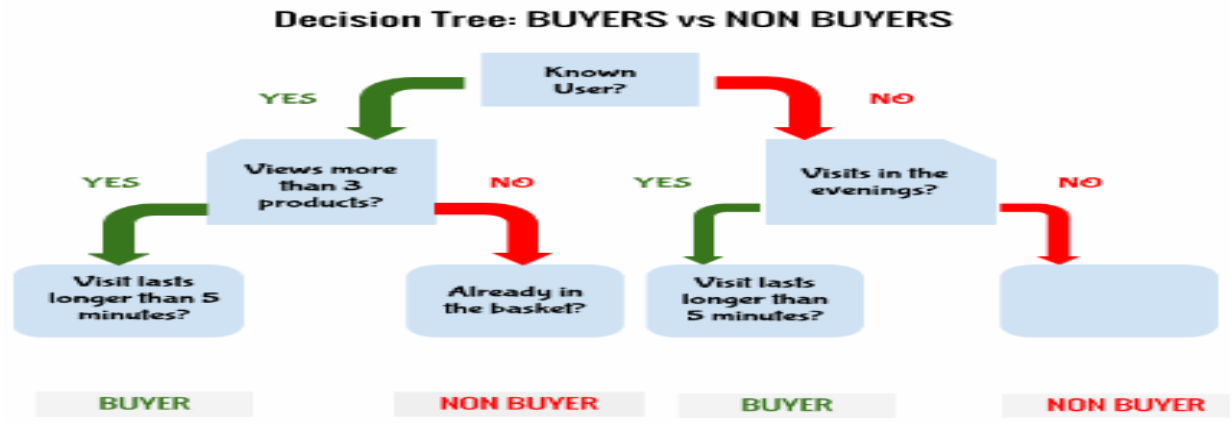

Figure 3. Decision Tree Applications

Which algorithm is to beused? This is a primary skill of a data scientist. Choosing the right algorithm for an organization is a combination of science and art. The "artistic" part is based on data mining experience, combined with knowledge of the business and its customer base. Algorithms play a crucial role in delivering business queries accurately. It's always a good idea to blend multiple techniques together to improve your regression, clustering or segmentation algorithms. An example of such blending is hidden decision trees.

LITERARTURE REVIEW

Dongyu Lin & Dean P. Fosterpresent a solution to address the problem of how to improve the speed of variable selection algorithms for large-scale data. They use the classical linear model as their basis to propose a fast and accurate algorithm, Variance Inflation Factor (VIF) regression, for doing feature selection. Using one-pass search over the predictors and an efficient computation method to test each potential predictor VIF regression avoids the pitfalls of linear regression such as model over-fitting and marginal False Discovery Rate (mFDR). The authors compare VIF with classic stepwise regression, Lasso algorithm and two other algorithms GPS and FoBa and conclude that VIF is much faster than other regression algorithms with feature selection and is as accurate as the best of the slower algorithms.

Steven L. Scott, Alexander W. Blocker et.al state that there is a need for algorithms that perform distributed approximate Bayesian analyses on very large data sets with minimal communication. The Consensus Monte Carlo breaks the data into groups (called "shards"), gives each shard to a worker machine which does a full Monte Carlo simulation from a posterior distribution given its own data, and then combine the posterior simulations from each worker to produce a set ofglobal draws representing the consensus belief among all the workers. They say that depending on themodel, the resulting draws can be nearly indistinguishable from the draws that would have beenobtained by running a single machine algorithm for a very long time. They quote examples of consensusMonte Carlo for simple models where single-machine solutions are available, for largesingle-layer hierarchical models, and for Bayesian additive regression trees (BART).

Shirin Abbasi and Babak Vazirihavestudied the main clustering algorithm K-Means and suggested some ways that consider main parameters in cloud environments and how to improve massive clustering. They state that K-means is one of the main data mining algorithms available in different software packages as open codes. Though they say the algorithm is simple and fast they highlight 2 problems. One, dependency on initial cluster centers which arises for first clusters and two, the number of clusters that is specified with K needs to be determined at the beginning of the algorithm. They offer improvisations to K-means by implementing it with map-reduce function this way improving the initial cluster centers and its operational distance between sample points. They also suggest for greater reliability using indexing methods in combination to clustering algorithms.

Xiao Cai, Feiping Nie, et al. propose a new robust large-scale multi-view K-means clustering method tointegrate heterogeneous representations of large-scale data. They

address two main computational challenges in large-scale data clustering (1) How to integratethe heterogeneous data features to improve the performanceof data categorizations? (2) How to reduce the computationalcost of clustering algorithm for large-scale applications. Using six benchmark data sets they have demonstrated in their paper that their proposed method consistently achieves better clustering performances. Using a common cluster indicator, they search a consensus pattern and do clustering across multiple visual feature views. By imposing the structured sparsitynorm on the objective function, their method is robust to the outliers in input data.

Sudipto Guha, Rajeev Rastogi et al.propose a new clustering algorithm called CURE that is more robust to outliers, and identifies clusters having non-spherical shapes and wide variances in size. By representing each cluster by a certain fixed number of pointsthat are generated by selecting well scattered points from the cluster and then shrinking them toward the center of the cluster by a specified fraction the authors demonstrate that CURE adjusts well to non-spherical shapes and dampens the effects of outliers. The authors say that by employing a combination of random sampling and partitioning, CURE not only outperforms existing algorithms but also scales well for large databases without sacrificing clustering quality.

Jooyeol Yun1, Jun won Seo et al present a method to construct a fuzzy decision tree using Iris flower data set. As the real world contains a lot of fuzzy and uncertain data the authors stress upon the need for fuzzy algorithms which combine fuzzy set theory and entropy. Using a method of fuzzydecision tree that uses the distance from an average for the index of uncertainty the authors propose a fuzzy decision tree induction method for fuzzy data of which data is obtained by a membership function. They conduct an experiment inorderto provethevalidity, and to compare it with former ID3 algorithm.

CONCLUSION

A survey of more than 400 executives in many sectors revealed that companies with better analytics capabilities were twice as likely to be in the top quartile of financial performance in their industry, five times more likely to make decisions faster than their peers and three times more likely to excel [20]. Business is really a contract between people. The more information the provider of products or services has about a customer, the better it can help. When used to enhance the customer experience, big data doesn’t just offer benefits to businesses; it can help all of us.

REFERENCES

[1] Piatetsky , Gregory . "Top algorithms and methods used by data scientists." KDNuggests, Sept. 2016. Web. May 2017.

[2] S. Y. Hwang, H. Wang, J. Tang, and J. Srivastava, “A probabilistic approach to modeling and estimating the QoS of web-servicesbased workflows,” Info. Sci., vol. 177, pp. 5484–5503, 2007.

[3] Granville,Vincent. "10 types of regressions. Which one to use?" Data Science Central, 21 July 2014. Web. 20 May 2017.

[4] Bayesian Data Analysis by A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin ISBN 0-412-03991-5, Chapman and Hall, New York, New York (Telephone: 800-842-3636), 1995, 526 pp., $60.50 hardback

[5] Verma, Amit, Iqbaldeep Kaur, & Amandeep Kaur. "Algorithmic Approach to Data Mining and Classification Techniques." *Indian Journal of Science and Technology* [Online], 9.28 (2016): n. pag. Web. 20 May. 2017

[6]Rachel, Anju. "Top 5 Algorithms to Make Your Machine Intelligent." Travancore Analytics, 06 Feb. 2017. Web. 20 May 2017.

[7] A. Avizienis, J.-C. Laprie, B. Randell, and C. Landwehr, ‘‘Basic Concepts and Taxonomy of Dependable and Secure Computing,’’ IEEETrans. Dependable Secure Comput., vol. 1, no. 1, pp. 11 /33, Jan.-Mar. 2004.

[8] Changala, Ravindra, Annapurna Gummadi, Yedukondalu, and UNPG Raju.] *Classification by Decision Tree Induction Algorithm to Learn Decision Trees from the class-Labeled Training Tuples.* International Journal of Advanced Research in Computer Science and Software Engineering, Apr. 2012.

[9] I. Dhillon, Y. Guan, B. Kulis,” Kernel k-means: spectral clustering and normalized cuts”, in Proceeding KDD’04 Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp.551–556, 2004.

[10] I. Dhillon and D. Modha,” Concept decompositions for large sparse text data using clustering”, Machine Learning, vol. 42, no. 1–2, pp. 143–175, 2001.

[11] R.C. de Amorim and B. Mirin, ”Minkowski metric, feature weighting and anomalous cluster initializing in k-means clustering”, Pattern Recognition, vol. 45, no. 3, pp. 1061–1075, 2012.

[12] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Kluwer Academic Publishers, 1981.

[13] Lin, Dongyu, and Dean P. Foster. "VIF Regression: A Fast Regression Algorithm for Large Data." *2009 Ninth IEEE International Conference on Data Mining* (2009): n. pag. Web.

[14]Scott, Steven L., Fernando V. Bonassi, Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. *Bayes and Big Data: The Consensus Monte Carlo Algorithm.* N.p., 13 Oct. 2013. Web. <http://www.rob-mcculloch.org/some_papers_and_talks/papers/working/consensus-mc.pdf>.

[15] Abbasi, Shirin, and Babak Vaziri. "Clustering Algorithms in Big data." *International Academic Journal of Science and Engineering* 2.9 (2015): 26-36. *IAIEST*. Web. 6 May 2017.

[16] Cai, Xiao, Feiping Nie, and Heng Huang. "Multi-View K -Means Clustering on Big Data." *Twenty-Third International Joint Conference on Artificial Intelligence* (n.d.): n. pag. Web. doi=10.1.1.415.8610

[17] Guha, Sudipto, Rajeev Rastogi, and Kyuseok Shim. "Cure: an efficient clustering algorithm for large databases." *Information Systems* 26.1 (2001): 35-58. Web

[18] Yun, Jooyeol, Jun Won Seo, and Taeseon Yoon. "The New Approach on Fuzzy Decision Forest." *Lecture Notes on Software Engineering* 4.2 (2016): 99-102. Web.

[19] Mahieu, Hristophe De, and Gregory Garnier. "New possibilities in big data analytics in oil and gas." *Gulf Times* 26 Mar. 2015: n. pag. Print.

[20] Aggarwal, Charu C., and Chandan K. (1980-) Reddy. *Data clustering: algorithms and applications*. Boca Raton: CRC Press/Taylor & Francis Group, 2014. Print.