International Journal of Engineering & Technology, 7 (3.4) (2018) 13-16

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



Migrating From Data Mining to Big Data Mining

Gourav Bathla[1]*, Himanshu Aggarwal[1] , Rinkle Rani[2]

[1]*Punjabi University Patiala, India*
[2]*Thapar University Patiala, India*
**Corresponding author E-mail:gouravbathla@gmail.com*

Abstract

Data mining is one of the most researched fields in computer science. Several researches have been carried out to extract and analyse important information from raw data. Traditional data mining algorithms like classification, clustering and statistical analysis can process small scale of data with great efficiency and accuracy. Social networking interactions, business transactions and other communications result in Big data. It is large scale of data which is not in competency for traditional data mining techniques. It is observed that traditional data mining algorithms are not capable for storage and processing of large scale of data. If some algorithms are capable, then response time is very high. Big data have hidden information, if that is analysed in intelligent manner can be highly beneficial for business organizations. In this paper, we have analysed the advancement from traditional data mining algorithms to Big data mining algorithms. Applications of traditional data mining algorithms can be straight forward incorporated in Big data mining algorithm. Several studies have analysed traditional data mining with Big data mining, but very few have analysed most important algortihsm within one research work, which is the core motive of our paper. Readers can easily observe the difference between these algorthithms with pros and cons. Mathemtics concepts are applied in data mining algorithms. Means and Euclidean distance calculation in Kmeans, Vectors application and margin in SVM and Bayes therorem, conditional probability in Naïve Bayes algorithm are real examples. Classification and clustering are the most important applications of data mining. In this paper, Kmeans, SVM and Naïve Bayes algorithms are analysed in detail to observe the accuracy and response time both on concept and empirical perspective. Hadoop, Mapreduce etc. Big data technologies are used for implementing Big data mining algorithms. Performace evaluation metrics like speedup, scaleup and response time are used to compare traditional mining with Big data mining.

Keywords: *Clustering, Classification, Big Data, Hadoop, MapReduce*

1. Introduction

Large scale of data is generated due to social networking sites, business transactions, sensor networks and digital traces. Important information extracted from this data is very crucial for organizations. Traditional data mining algorithms were designed for limited scale of data and for data which were in structured format. 10 algorithms are described in IEEE ICDM which are - C4.5, K-means, SVM, Apriori, EM, PageRank, AdaBoost, kNN, Naïve Bayes, and CART [1]. Kmeans, Naïve Bayes and SVM algorithms are analysed and observed with the advantages of Big data technologies implemention in traditional data mining algorithms. Big data technologies incorporated in data mining algorithm can be used for handling large scale of data which is semi-structured and unstructured.

Data mining algorithms are implemented using traditional technologies and it is migrated to Big data technologies. In Big data, classification and clustering are very important algorithms and most frequently used also [2]. Classification is used in web searching and CRM [3] .Hadoop and Mapreduce are used in several research works to prove improvement in mining algorithms as compared to traditional approaches.

Mapreduce is a framework that split job using map phase for running job on distributed nodes and combines the results in reduce phase [4]. Google File Systems is used by MapReduce to store dataset in distributed format [5]. Map divides the dataset to distributed it on nodes in key/value pair and reduce phase output from different nodes is updated. Mappers and Reducers are developed to deploy any algorithm on Big data. Runtime system in Mapreduce can handle machine failures and schedule communication amongst nodes [6]. Hadoop is open source implementation of Mapreduce. Mahout is also used for improving clustering and classification algorithms [7].Experiment analysis proves that Big data mining algorithms provide better results in context of speedup, scaleup and CPU time.

The rest of the paper is organized as follows, in Section 2 traditional Kmeans is explained and improvement is shown with the use of Mapreduce. In Section 3, Naïve Bayes algorithm is elaborated and comparison is explained in detail with Big data algorithms. In Section 4, SVM algorithm is described with its Big data algorithm. Finally, Section 5 concludes the paper.

2. KMeans

Kmeans is the most commonly used clustering algorithm. It is used to split dataset into specified K clusters. In a set of n integers, K points are determined which are called centers, where mean squared distance is minimized [8]. K value is fixed which is used to cluster objects. K centers are selected and objects are assigned

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

in these clusters based on similarity score. Several distance measures like cosine similarity, Euclidean distance is used to find distance between data objects and centers. It is used to cluster similar objects with proven accuracy. Intra-cluster objects have high similarity and inter-cluster objects have low similarity. It is algorithm which partition dataset in Voronoi diagram [9].
Calculation of distance is time consuming step which takes maximum time amongst several steps. Distance calculation between center with object is independent of other object distance calculation with other center [6]. Distance computation can be distributed on nodes. In each iteration, centers are updated to be used by nodes for next iteration. Kmeans can not provide good response time if implemented using centralized approach. In this paper, Kmeans calculation of distance is distributed on nodes using big data technologies like hadoop and mapreduce. The advantage is that Kmeans can be deployed for large scale of data.
Euclidean distance is mostly used to find similarity between data objects and cluster centroids.

$$d = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \quad (1)$$

2.1. Kmeans Algorithm:

(1) Initialize data points (d1, d2..... dn), and assign it to one of k clusters (C1, C2.....Ck) based on similarity score.
(2) Repeat step 3 and 4 until there is no change in centroid data points with previous centroid data points.
(3) Calculate mean of all data objects in a particular cluster and assume mean as new center points.
(4) Calculate similarity score with new centroid points with data points.
(5) Exit

2.2. Kmeans for Big Data:

Distance between data objects and centers can be calculated individually without updation from other objects distance calculation from centers. The main objective to deploy Kmeans on Big data is reduction in I/O and network cost [10]. Distance between objects and centers are calculated on clusters using map, reduce steps combines the updated centers.
Map: Data objects are globally shared amongst all nodes. Centers dimensions are already set at each node. Distance computation on every node is done in distributed manner. Similar objects based on distance calculated are selected as member of that particular cluster.
Reduce: Mean of objects selected on clusters are calculated and new centers dimensions are set for every cluster. This information is sharable to clusters and next iteration proceeds to calculate new centers. Iteration continues until there are centers where previous values of new values are same.

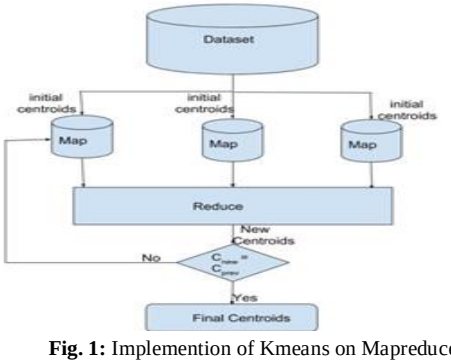


Fig. 1: Implemention of Kmeans on Mapreduce

Speedup, scaleup and CPU response time to prove that implemention using Big data technologies is much better than existing approaches.

$$\text{speedup} = \frac{S1}{Sm} \quad (2)$$

where S1 is speed on single node and Sm is speed on m nodes. Speedup is not linear as there is communication cost amongst nodes. Figure 2 shows the experiment results of [6] where hadoop version 0.17.0 and clusters with two 2.8 GHz cores are used.
Dataset is constant and numbers of nodes are increased to analyse speedup. It is clear from Figure 2 that sppedup is not linear when Kmeans is deployed on mapreduce. Dataset is split into 1GB, 2 GB, 4GB and 8 GB and deployed on 2, 3 and 4 nodes. The key observation from this experiment analysis is that Kmeans can be improved by using big data technologies.

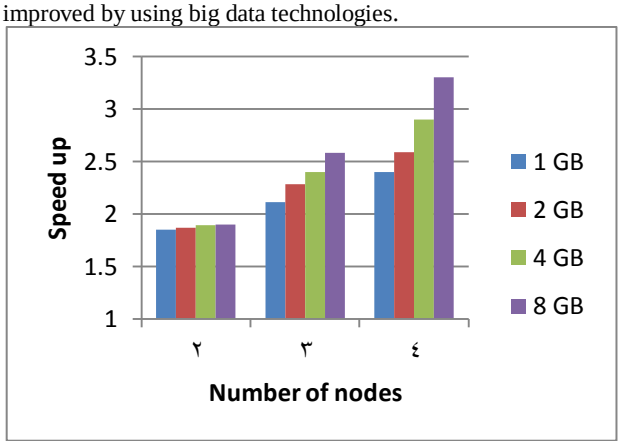


Fig. 2: Speedup of Kmeans implemented on Big data

3. Naïve Bayes Algorithm

This algorithm uses supervised approach i.e. given a set of predefined classes, any new object is assigned to class based on similarity with a particular class. This algorithm is simple and there is no need of iterative approach [X Wu]. This algorithm assumes each feature of object vector as independent . It is used in text classification and spam filtering [1]. It is proved to be a effective machine learning technique [5] . P(c|d) is the probability that object with vector (d1, d2……dn) has highest similarity with class c. It is calculated with the use of conditional probability.

$$P(c|d) = \frac{P(d|c)\,P(c)}{P(d)} \quad (3)$$

where, P(d|c) is calculated with univariate vector calculation between object d and class c. P(c) is calculated as the probability of class c in complete training set.

3.1. Naïve Bayes for Big data

The advantage of Naïve Bayes algorithm is that objects features are independent of each other. Calculation of likelihood of an object with in any particular class can be executed in parallel. Classification accuracy and CPU time are improved significantly using hadoop and mapreduce. In Equation 4, P(d) is constant value as this is calculated using dataset. The motive in Naïve bayes is to optimize P (d|c) P(c).

$$\text{P (c|d)} = \frac{P(c)}{P(d)} \prod_{i=1}^{n} P(d_i\,|c) \quad (4)$$

This algorithm is implemented on Hadoop and Mapreduce in parallel. Training of data is first submitted to map, paramenters rules are defined and model is generated in reduce phase. This model along with testing data is submitted to next mapper and reducer. Data is checked along with trained model to finalize the class.
Map : In this step, dataset is split and object parameters are analyzed and trained to classify.
Reduce : Object parameters results are combined to assign object in a particular class. The class for this object is written on HDFS.

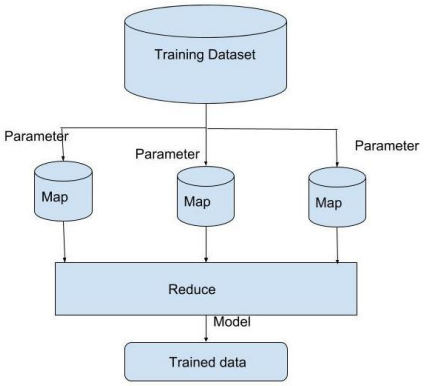


Fig. 3: Training for generating model in Naïve Bayes

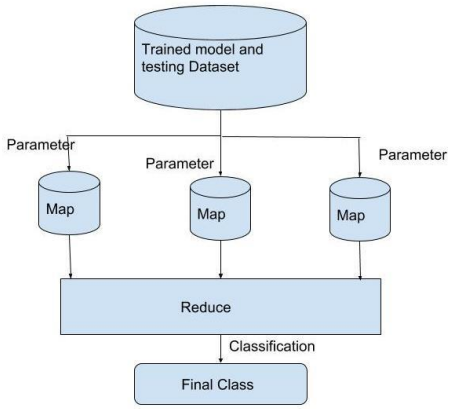


Fig. 4: Classification using Naïve Bayes

Accuracy is used as evaluation parameter to analyse the scaling of this algorithm. In [5], It is observed that when dataset is small, it provides accuracy which is not predictable due to not proper training of data. When dataset size is increased, accuracy is 80%. The following table shows the accuracy with the increase in size of dataset in [5].

Dataset size	2	20	200	400	600	800
Accuracy	0.84	0.72	0.79	0.80	0.81	0.81

Fig. 5: Accuracy for Big data mining using dataset variation

This table proves that accuracy of Naïve Bayes Classifier for large dataset is better as compared to traditional classification algorithm.

4. SVM

Support Vector Machine is classification technique which is very popular to classify different varieties of objects. It is used in image classification, text classification and machine learning techniques. It was introduced by VN Vapnik [11]. It is supervised learning technique which provides the largest distance between vectors. It is like hyperplanes that divide the training data by maximum margin [12]. SVM is also called maximum margin classifier [13]. Support vector classifier uses hyperplane which maximizes the margin between classes [14]. Vectors which are near the hyperplane are called Support Vectors. SVM is used in many applications like handwritten character recognition, text categorization and information extraction [2]. Several optimization techniques can be used to improve SVM like Ant Colony and genetic algorithm [13]. In [15], enclosing ball clustering technique is used to deploy SVM on Big data. Training complexity depends upon the size of large dataset [15]. In the following subsection, SVM is improved by using Big data technologies.

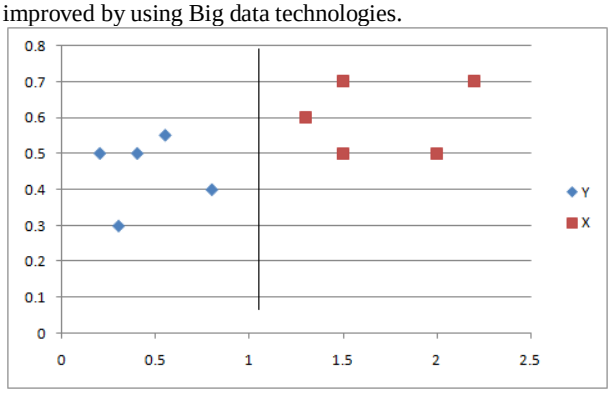


Fig. 6: Traditional margin Classifier

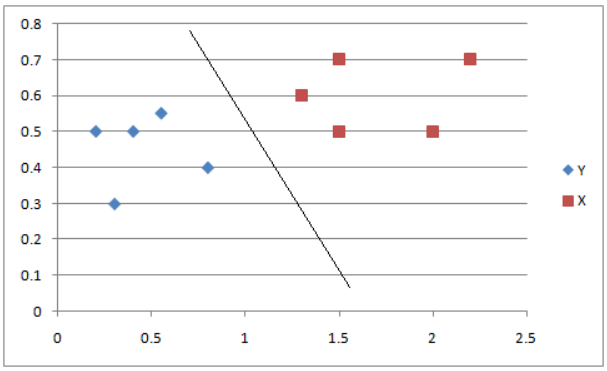


Fig. 7: Maximum margin Classifier

It is clear from Figure 6 that data points are divided by simple margin techniques, whereas in Figure 7, maximum margin technique is applied to maximize the distance between data points. It is the reason that SVM is also called maximum margin algorithm.

4.1. Svm for Big Data

Large scale of data is splitted into small sets and processed on distributed nodes. It is clear from Figure 8 that hierarchical structure is followed to classify objects. Objects distance calculated in sub parts are combined and shared with next sub part. In this manner, final SVM classify objects with less CPU time.

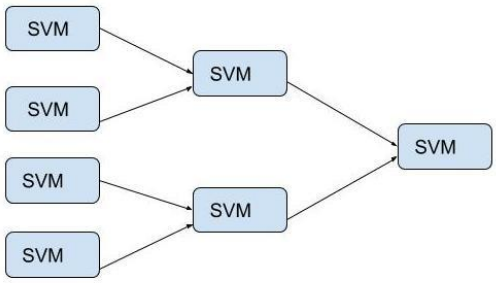


Fig. 8: Distributed SVM classification

Map: In this phase, local SVM is computed from splitted dataset.
Reduce: Global SVM is computed after combining local SVM. Final support vector is the output of this phase.

In [13], SVM is distributed on hadoop and mapreduce with varying number of support vectors. Computation time is less when SVM is implemented on multiple nodes as compared to single node. It is shown in this work that when dataset size is 128 MB, execution time is 96.6 secs on 3 nodes and when dataset size is increased to 1024 MB, execution time is 376.8 secs on 3 nodes. When dataset size is 1024 MB, execution time is 289.45 secs on 4 nodes. This analysis clearly shows the advantage of multi- node cluster use in SVM implemention.

5. Conclusion

Data mining algorithms are very popular amongst researchers as it provides a lot of important information from dataset. In this paper, we have mentioned that there are 10 data mining algorithms which are defined to be most effective in ICDM. Kmeans, Naïve Bayes and SVM algorithms are analysed in detail. These algorithms are proved for accuracy but when migrated to Big data, it can not perform better due to limited storage and processing. We have shown that these algorithms can work well for large scale of data using Big data technologies. Hadoop and Mapreduce are used for implementing these algorithms for Big data. Mapper and Reducer for algorithms are explained to make it clear to reader that CPU time and accuracy is better as compared to traditional approaches. Mathematical concepts like mean, vectors, bayes theorem and conditional probability are used in traditional data mining algorithms and also in big data mining algorithms.

References

[1] Wu X, Kumar V, Quinian JR, Ghosh J,Yang Q, Motoda H, McLachlan GJ Liu B, Yu PS , Zhou Z, Steinbach M, Hand DJ and Steinberg D(2007) , Top 10 algorithms in data mining, Knowledge and Information Systems, vol. 14, no.1, pp. 1-37.

[2] Demidova L, Nikulchev E and Sokolova Y(2016), Big Data Classification Using The SVM Classifiers With The Modified Particle Swarm Optimization And The SVM Ensembles, International Journal of Advanced Computer Science and Applications (IJACSA), vol.7, no. 5, pp.294-312.

[3] He Q, Zhuang F, Li J and Sh Z (2010), Parallel implementation of classification algorithms based on MapReduce, In International Conference on Rough Sets and Knowledge Technology, pp. 655-662, Springer.

[4] Dean J and Ghemawat S (2008), MapReduce: Simplified Data Processing on Large Clusters, Communications of the ACM, vol. 51, no.1, pp. 107-113.

[5] Liu B, Blasch E, Chen Y, Shen D and Chen G (2013), Scalable sentiment classification for Big Data analysis using Naïve Bayes Classifier ,IEEE Conference on Big Data.

[6] Zhao W, Ma H and He Q (2009), Parallel K-Means Clustering Based on MapReduce, in Cloud Com LNCS 5931, pp. 674-679.

[7] Owen S and Owen S (2012), Mahout in action.

[8] Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R and Wu AY (2002), An efficient k-means clustering algorithm: Analysis and implementation, IEEE transactions on pattern analysis and machine intelligence, vol. 24, no. 7, pp.881-892.

[9] Cai X, Nie F and Huang H (2013), Multi-View K-Means Clustering on Big Data, IJCAI, pp. 2598-2604.

[10] Cui X, Zhu P, Yang X, Li K and Ji C (2014), Optimized big data K-means clustering using MapReduce, The Journal of Supercomputing, vol. 70, no. 3, pp.1249-1259.

[11] Vapnik VN (1995), Editor,The Nature of Statistical Learning Theory, Springer- Verlag.

[12] Tong S and Koller D (2001), Support Vector Machine Active Learning with Applications to Text Classification, Journal of Machine Learning Research, pp. 45-66.

[13] Priyadarshini A and Agarwal S (2015), A Map Reduce based Support Vector Machine for Big Data Classification, IJDTA, vol. 8 no. 5, pp. 77-98.

[14] Hearst MA, Dumais ST, Osuna E, Platt J and Scholkopf B (1998), Support vector machines, IEEE Intelligent Systems and their applications, vol. 13, no. 4, pp.18-28.

[15] Cervantes J, Li X, Yu W and Li K (2008), Support vector machine classification for large data sets via minimum enclosing ball clustering, Neurocomputing, vol. 71, no. 4-6, pp.611-619.