International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163
Volume 1 Issue 7 (August 2014) http://ijirae.com

Comparative Study of Classification Algorithms Based On MapReduce Model

Seyed Reza Pakize *
Department of Computer, Islamic Azad University, Yazd Branch, Yazd, Iran.

Abolfazl Gandomi
Department of Computer, Islamic Azad University, Yazd Branch, Yazd, Iran.

Abstract— Nowadays Classification is an important data mining problem. Classification Algorithms can be used for classifying the interested users. Today, all focus of the researchers and companies toward to big data. thus, many classification algorithms have been proposed in the past decades. Many of them have very limitation and weakness. One of the most important limitations is the time when we want to use this classification algorithm on the very large datasets which have poor run-time performance and high Computation time and cost. to covers this limitations, many researchers using this classification algorithm based on MapReduce. In this paper, we have studies these classification algorithms. then, we comparison with the traditional models. finally, highlighting the advantages of Mapreduce Models into traditional models.

Keywords— Classification algorithm, Mapreduce, Parallelism algorithm, algorithm based on Mapreduce.

I. INTRODUCTION

Nowadays, many researchers and companies toward to big data. Big Data is a loosely defined term used to describe data sets so large and complex that they become awkward to work with using standard statistical software it has to deal with large and complex datasets which usually includes datasets with sizes[1]. There are many algorithm in different field of big data. Data mining algorithms are divided in four classes, including association rule learning, Clustering, Classification and Regression. classification algorithm deals with associating an unknown structure to a well known structure which is an important data mining problem[2]. it now, many classification algorithms have been proposed for big data. Many of them have limitation and weakness. Such as: low performance in large dataset, poor run-time performance when the training set is large, high Computation cost. to covers these limitations, many researchers using classification algorithm based on MapReduce. The Mapreduce model was developed by Google to run data-intensive applications on a distributed infrastructure like commodity cluster [3].Rest of the paper is organized as follows. Section 2 presents the Mapreduce. Section 3 will describe the classification algorithm based on Mapreduce. Comparison and discussion are reported in section 4. Finally, the overall conclusions of this study are presented in section 5.

II. MAPREDUCE OVERVIEW

MapReduce [4] is a new distributed programming paradigm which widely used to run parallel applications for large scale datasets processing. MapReduce uses key/value pair data type in map and reduce functions and it was inspired by the map and reduce primitives present in Lisp and many other functional languages [3]. The computations in MapReduce expressed as two functions: Map and Reduce function. Map function requires the user to handle the input of a pair of key value and produces a group of intermediate key and value pairs. Reduce function is also provided by the user, which handles the intermediate key pairs and the value set relevant to the intermediate key value[5]. Reduce function mergers these values, to get a small set of values. in the MapReduce Programming Model [6], Map algorithm include three steps: first, Hadoop and MapReduce framework produce a map task for each Input Split, and each Input Split is generated by the Input Format of job. Each <Key,Value> corresponds to a map task. in second step, Execute Map task, process the input <key,value> to form a new <key,value>. and in last step, Mapper's output is sorted to be allocated to each Reducer. Reducer algorithm also included three steps[6]: first, MapReduce will assign related block for each Reducer (Shuffle). next step, the input of reducer is grouped according to the key (sorting step). and finally step is Secondary Sort. If the key grouping rule in the intermediate process is different from its rule before reduce.

III. CLASSIFICATION ALGORITHMS BASED ON MAPREDUCE

In this section we briefly introduce some of the Classification Algorithm.

A. *K Nearest Neighbor*

K Nearest Neighbor (KNN) is one of the most widely used classification Algorithm in data mining, which based on learning by analogy, that is by comparing a given test tuple with training tuples which are similar to it[7]. kNN classification determines the decision boundary locally that was developed from the need to perform discriminate analysis when reliable parametric estimates of probability densities are unknown or difficult to determine. If k = 1, then the object is simply assigned to the class of its nearest neighbor. In practical applications, k is in units or tens rather than in hundreds or thousands [7]. K is a user-defined constant, and an unlabeled vector or test point is classified by assigning the label which is most frequent among the k training samples nearest to that query point. In k-NN classification [8], the output is a class membership. Many researches [10, 11] is proposed based on the centralized paradigm where the kNN join is performed on a single, centralized server. parallelization KNN algorithm improves the classification efficiency[11].

B. Support Vector Machine

Title Support Vector Machine (SVM) is one of the most successful classification algorithms in the data mining area, but its long training time limits its use. the aim of SVM is to find optimal separating hyper plane by maximizing the margin between the two classes, which offer the best generalization ability for future data[12]. Their computation and storage requirements increase rapidly with the number of training vectors. SVM uses statistical learning theory to maximize generalization property of generated classifier model. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a on probabilistic binary linear classifier. Support vector machines(SVM) is a learning technique which has been successfully applied in many application areas. Support Vector Machines (SVM) are the classifiers which were originally designed for binary classification[13]. A good separation is achieved by the hyper plane that has the largest distance to the nearest training data point of any class, since in general the larger the margin the lower the generalization error of the classifier. The computation cost of SVM is square proportion to the number of training data[15,16]. Classical SVM model is difficult to analyse large scale practical problems. Parallel SVM can improve the computation speed greatly. Z. Sun and G. Fox et al. [14] proposed the parallel SVM model based on iterative MapReduce, which showed that the parallel SVM based on iterative MapReduce is efficient in data intensive problems. To improve scalability, a parallel SVM Algorithm is developed, which reduces memory use through parallel computation[15].

C. Naive Bayes

Naıve Bayes is an important supervised classification method which is based on applying Bayes' theorem with strong independence assumptions. It can predict class membership probabilities. The basic idea in Naive Bayes approaches is to use the joint probabilities of words and categories to estimate the probabilities of categories given a document [17]. Bayesian classifiers are statistical classifiers. They can predict class membership probabilities. Na¨ıve Bayes classifiers often work much better in many complex real-world situations than one might expect [18]. Naïve Bayes is appropriate for Map reduce architecture. Zhou et al. [19] implementation parallel Naïve Bayes algorithm based on Mapreduce model. This implementation is divided into training and prediction stages. the training stages including three stages. First, the Input Format which is belonged to the Hadoop framework loads the input data into small data blocks known as data fragmentation, and the size of each data fragmentation is 5M, and the length of all of them is equal, and each split is divided into records. second, the map function statistics the categories and properties of the input data, including the values of categories and properties. and third stage, the reduce function aggregates the number of each attribute and category value, which results in the form of (category, Index1:count1, Index2:count2, Index3:count3, … , Indexn:countn), and then output the training model. Prediction Stage, Predicate the data record with the output of the training model. Implementation of Naïve Bayes based on Map Reduce has very good performance and reduced the training time[20].

D. C4.5

C4.5 [21] is a standard algorithm for inducing classification rules in the form of decision tree. it uses a divide-and-conquer approach to growing decision tree. the default splitting criterion used by C4.5 is gain ratio, an information-based measure that into account different numbers of test outcomes. [22] There are two limitation in traditional decision tree algorithms. First, when the volume of dataset is extremely big, building a decision tree can be very time consuming. and Second, although parallel computing in clusters can be leveraged in decision tree based classification algorithms [23]. to address this limitations, Wei die and Wei ji [23] suggested a parallel version of C4.5 based on MapReduce. Also Gong-Qing Wu [24], proposes a new method MReC4.5 for parallel and dis-tributed ensemble classification.

IV. COMPARISON OF CLASSIFICATION ALGORITHMS

So far we have described the classification algorithm. In this section, we evaluated and compared these classification algorithm with mapreduce model. table 1 shows the limitation of traditional model of these classification algorithms. also shows the advantages of these algorithm when using Mapreduce model. as you can see, all of this algorithm have same limitation a time when the dataset is very large. e.g., traditional SVM algorithm required large memory and continues high computation time in large dataset. C4.5 Decision boundaries are rectilinear and low performance in large datasets. and also, Naïve Bays and KNN have low performance and high computation times in large data sets. Also as you can see, in the Mapreduce column (Based on Mapreduce model) all of these weaknesses are address

.

TABLE I
COMPARISON OF TRADITIONAL OF CLASSIFICATION ALGORITHM WITH MAPREDUCE MODEL

Classification	Traditional (limitation in old model)	Based on MapReduce
SVM	1. requirement large memory in very large dataset 2. high computation time in large dataset	1. Reduce training time 2. Reduce computation time 3. Increases the performance
C4.5	1. Decision boundaries are rectilinear 2. A sub-tree can be replicated several times	1. Minimize the communications cost 2. time efficiency and scalability 3. reduce the execution time
Naïve Bays	1. it has strong feature independence assumptions 2. low performance in large dataset	1. Improves the performance 2. reduce the training time 3. able to process large database
KNN	1. Performance depends on the number of Dimensions. 2. have poor run-time performance when the raining set is large. 3. high Computation cost.	1. reduce the communications cost 2. increases the performance

V. CONCLUSION

In this paper, we explained four popular classification algorithm based on Mapreduce model. Then, compared this algorithm with traditional models highlighting advantages and dis-advantages of each. all the limitations and dis-advantages of traditional models, covers with Mapreduce model, which is increases the performance and reduce computational and executorial time. we can conclude, one of the best advantage of these classification algorithm based on Mapreduce model is able to use and gets good result in very large datasets.

REFERENCES

[1] C. Snijders, U. Matzat, U. Reips," 'Big Data': Big gaps of knowledge in the field of Internet ", International Journal of Internet Science 2012, 7 (1), 1–5.
[2] A. N. Nandakumar and N. Yambem, "A Survey on Data Mining Algorithms on Apache Hadoop Platform", International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 1, January 2014.
[3] J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. In Proceedings of the 6th conference on Symposium on Operating Systems Design & Implementation -Volume 6, pages 10–10, Berkeley, CA, USA, 2004. USENIX Association.
[4] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," Commun. ACM, vol. 51, no. 1,pp. 107–113, 2008.
[5] Dean J, Ghemawat S. Simplified data processing on large clusters. In: 6th conference on Symposium on Operating Systems Design & Implementation(OSDI); 6-8 December 2004; Berkeley, USA: ACM. pp. 107-113.
[6] Ping ZHOU, Jingsheng LEI and Wenjun YE "Large-Scale Data Sets Clustering Based on MapReduce and Hadoop", Journal of Computational Information Systems 7: 16 (2011) 5956-5963.
[7] H. Jiawei and M. Kamber. Data mining: concepts and techniques. San Francisco, CA, itd: Morgan Kaufmann, 2001.
[8] Tomasev, Nenad, Milos Radovanovic, Dunja Mladenic, and Mirjana Ivanovc, "Hubness-based fuzzy measures for high-dimensional k-nearest neighbor classification." International Journal of Machine Learning and Cybernetics 5, Vol.3, 2014.
[9] A. Stupar, S. Michel, and R. Schenkel. RankReduce processing k-nearest neighbor queries on top of MapReduce. In LSDS-IR, pages 13–18, 2010.
[10] C. Zhang, F. Li, and J. Jestes. Efficient parallel knn joins for large data in MapReduce. In EDBT, 2012.
[11] Wei Lu et al., "Efficient Processing of k Nearest Neighbor Joins using MapReduce", Journal Proceedings of the VLDB Endowment VLDB Endowment Hompage archive, Volume 5 Issue 10, June 2012.
[12] C. Cortes and V. Vapnik. Support-vector networks. Machine learning, 20(3):273– 297, 1995.
[13] C. Chang, C. Lin. "LIBSVM: a library for support vector machines." ACM Transactions on Intelligent Systems and Technology, 2011, 27(2): 1-27.
[14] Zhanquan Sun and Geoffrey Fox, " Study on Parallel SVM Based on MapReduce".
[15] Kiran M et al.,"Verification and Validation of MapReduce Program Model for Parallel Support Vector Machine Algorithm on Hadoop Cluster", IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 3, No 1, May 2013.
[16] Zhanquan Sun ―Study on Parallel SVM Based on MapReduce‖ in conference on worldcomp2012.
[17] S. Dhillon and K. Kaur, " Comparative Study of Classification Algorithms for Web Usage Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 7, July 2014.

[18] A. K. Santra and S. Jayasudha, "Classification of Web Log Data to Identify Interested Users Using Naïve Bayesian Classification" , International Journal of Computer Science Issues, Vol.9, Issue 1, January 2012.
[19] L. Zhou, H. Wang and W. Wang, " Parallel Implementation of Classification Algorithms Based on Cloud Computing Environment", TELKOMNIKA, Vol.10, No.5, September 2012.
[20] C.T. Chu, S.K. Kim, Y.A. Lin, Y.Y. Yu, G. Bradski, A.Y. Ng, and K. Olukotun. Map-reduce for machine learning on multicore. In Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference, page 281. The MIT Press, 2007.
[21] J.R. Quinlan,"C4.5: programs for machine learning", Morgan Kaufmann, (1993).
[22] J. R. Quinlan," Improved Use of Continuous Attributes in C4.5", Journal of Arti_cial Intelligence Research 4 (1996) 77-90.
[23] W. Dai, W. Ji," A MapReduce Implementation of C4.5 Decision Tree Algorithm", International Journal of Database Theory and Application Vol.7, No.1 (2014), pp.49-60.

