Asian Journal of Applied Research (AJAR) 2017

E-ISSN: 2395-1702 P-ISSN: 2395-0382
Volume 03- Issue 01-, pp-01-04

Research Paper Open Access

SURVEY ON K-NEAREST NEIGHBOR APPROACH FOR BIG DATA CLASSIFICATION BASED ON MAP REDUCE

Vigneshwaran.R[1], Balaji.V[2], Manikandan.A[3], Dr. Danapaquiame.N[3]

PG Scholar [1], PG Scholar[2], PG Scholar [3], Associate Professor[4]
Department of Computer Science & Engineering,
Sri Manakula Vinayagar Engineering College- Puducherry
Vigneshwaran260593@gmail.com[1], vbbalajir@gmail.com[2] , n.danapaquiame@gmail.com[3]

Abstract

In the data mining one of the most well known methods is K-Nearest Neighbor classifier because of its simple and effectiveness. Due to its way of working, the applications of this classifier may be restricted the problems with a creation number of examples, especially, when the runtime matters. However, the classification of large amounts of data is becoming a necessary task in a great number of real world applications. This paper uses a variety of datasets, and analyzes the impact of data volume, data dimension and the value of k from many perspectives like time and space complexity, and accuracy. We then analyze each step from load balancing, accuracy and complexity aspects. We identify three generic steps for KNN computations on map reduce: data preprocessing, data partitioning and computation. Overall, this paper can be used as to tackle KNN-based practical problems in the context of big data.

Keywords: Classification algorithm, Mapreduce, Paralelism algorithm, algorithm based on Mapreduce

1) 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 algorithms 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 Map Reduce. The Map reduce model was developed by Google to run data-intensive applications on a distributed infrastructure like commodity cluster [3].

2) Related Work

Due to the huge increase in the size of the data and the great amount of information that are added every second the word big data became one of the most important words nowadays. It becomes troublesome to perform efficient analysis using the current traditional techniques on the big data. So, big data put forward a lot of challenges. The Map Reduce paradigm has become one of the most challengeable areas in this era due to its infrastructure and importance in dealing with big data. Also the Hadoop have gained a lot of attention because it's an open source and can deal with big data. According to the importance of the big data classification there are a lot of related works and I will list some of them below. The k-Nearest Neighbor method is placed in the top ten data mining techniques. rajesh P, Anchalia, and Kaushik Roy used this well-known classification technique to classify big data. They run this method on an apache Hadoop environment that of course uses the MapReduce paradigm to process on big data. They faced a lot of problems and the most important one is balancing between the friendly user interface and performance. They implemented the k nearest neighbor on the Hadoop using multiple computers to delete the limitations of computational capability and speeding up the processing time. This is done by having groups of systems working together and connected over a network. They also compared their results using a MapReduce K Nearest Neighbor with sequential K Nearest Neighbor and concluded that the MapReduce k nearest neighbor gives better performance than the sequential K Nearest Neighbor with big data [7]. Nasullah Khalid Alham, Maozhen Li, Yang Liu, and Suhel Hammoud used a MapReduce support vector machine technique. They named this technique MRSMO technique (MapReduce based distributed SVM algorithm for automatic image annotation). Their technique depends on partitioning the training data into subsets and sent these subsets across groups of computers. They evaluated their technique in an experimental environment and it result in a significant reduction in the training time and a high accuracy in both binary and multiclass classification [8]. Ke Xu , Cui Wen , Qiong Yuan, Xiangzhu He , and Jun Tie used the parallel Support Vector Machine based on MapReduce method for classification of emails which is a big data set. They implement an experiment on this data set and used many techniques in evaluation but the support vector machine based on MapReduces show a significant reduction in the training time. Big data sets are very complex to be analyzed using classical Support Vector Machine but the parallel Support Vector Machine depending on MapReduce and can deal easily with big data. The MapReduce distribute the subsets of the training data among many nodes to improve the computation speed and improve the accuracy [9]. Now days, the mobile data set became one of the most challenging big data set due to its continuous production. Classification on this data requires high specifications because of its nature. This data have three main challenges; which are the difficulty in keeping both the accuracy and the efficiency, the huge increase in the load of the system, and the data noise. Zhiqiang Liu, Hongyan Li, and Gaoshan Miao used the Back Propagation Neural Network MapReduce technique to classify big data of mobiles. They implemented a technique called MBNN (MapReduce-based Back propagation Neural Network) to be used in classification of data. A lot of experiments are performed using the cloud computing platform and concluded that the MBNN have the characteristics of superior efficiency, good scalability and anti-noise [10]. Changlong Li, Xuehai Zhou, and Kun Lu Implemented an Artificial Neural

Networks in Map Reduce paradigm. They represented it to accomplish the parameter configuration automatically for Map Reduce. Their technique can adjust its software configuration and hardware configurations to the system automatically giving the cluster, Map Reduce job, and frameworks. Their technique also can determine the ideal configuration of the system in suitable time with the help of ANN. They experiment their technique and show that it result in a great influence in optimizing the system performance and speeding the system up [11].

3) Proposed Work

We first introduce the reference algorithms that compute KNN over MapReduce. They are divided into categories:a)Exact solutions b) Approximate solutions. Although based on different methods, all of these solutions follow a common workflow which consists in three ordered steps: 1) data preprocessing 2) data partitioning 3) KNN computation.

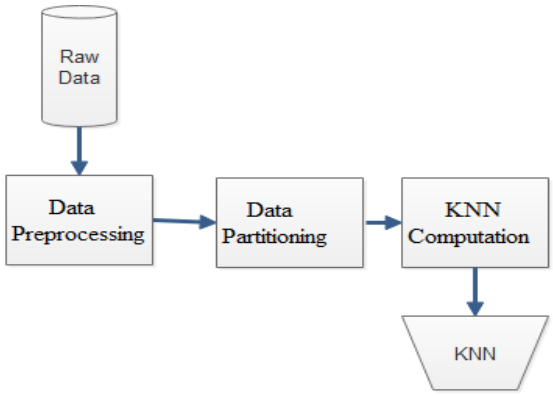

Figure 1: KNN MAP REDUCE

Data Preprocessing

The idea of data preprocessing is to transform the original data to benefit from particular properties. This step is done before the partitioning of data to pursue two different goals: 1) either to reduce the dimension of data 2) to select central points of data clusters.

3.2) Data Partitioning and selection

Map reduce is a shared-nothing platform, so in order to process data on map reduce, we need to divide the data set into independent pieces, called partitions.

Conclusion

In this paper, we have studied existing solutions to perform the KNN operations in the context of Map reduce. We have first approached this problem from a workflow point of view. We have pointed out that all solutions follow three main steps to compute KNN over map reduce, namely preprocessing of data, partitioning and actual computation. Finally, Map Reduce, especially through its hadoop implementation, is well suited for data Processing of static data. The efficiency of these methods on data stream.

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 Map Reduce and Hadoop",Journal of Computational Information Systems 7: 16 (2011) 5956-5963.
[7] P Anchalia, Prajesh, and Kaushik Roy. The K-Nearest Neighbor Algorithm Using MapReduce Paradigm. Fifth International Conference on Intelligent Systems, Modelling And Simulation. 2014. Web. 15 Oct. 2015.
[8] Nasullah Khalid Alham, Maozhen Li, Yang Liu, and Suhel Hammoud, (2011). a MapReduce-based distributed SVM algorithm for automatic image annotation
[9]. Xu, K., Wen, C., Yuan, Q., He, X., & Tie, J. (2014). A MapReduce based Parallel SVM for Email Classification. Journal of Networks JNW. [10]. Zhiqiang Liu; Hongyan Li ; Gaoshan Miao.MapReduce-based Backpropagation Neural Network over large scale mobile data [11]. Changlong Li1, Xuehai Zhou1, Kun Lu1. Implementation of Artificial Neural Networks in MapReduce Optimization.