MRK-SVM: An Effective Technique for Big Data In Health Care Sector

K. Sharmila1, Dr. S. A.Vethamanickam2
1Asst Prof. & Research Scholar, Dept. of Computer Science, Vels University, Chennai, India.
2Research Advisor, Chennai, India.First Author, Second Author, Third Author

Abstract— In this modern world, Big Data plays an imperative role in health care segment. Big data is a promising phrase that describes any ample amount of structured, semi-structured and unstructured data that has likely to be mined for information. Among any other dangerous Non Communicable Diseases (NCD), Diabetes Mellitus is a foremost health vulnerability in developing countries such as India. The sensitive nature of DM is coupled with long term complications and several health disorders. So this paper presents the progress of a hybrid model for the big dataset using the combination of clustering and classification techniques in Hadoop . The Apache Hadoop has become a world-wide adoption for Big Data. The above hybrid model consists of two phase. In the foremost phase, the K-means clustering is used to discover and stamp out the incorrectly classified instance. In the succeeding phase a fine tuned classification is performed using Support Vector Machine (SVM) by taking the right clustered occurrence of earliest phase. The result shows that our approach to this hybrid model will be efficient to predict the patients with diabetic who are all having the risk of Cardio Vascular Disease, Nephropathy, and Retinopathy and at the same time guarantees the timely treatment of the patients at the precise time.

Index Terms— Keyword: Big Data, Diabetes, Apache Hadoop, K-means, SVM.

1 INTRODUCTION

As the technology is mounting the size of data is also growing accordingly. So people are living in the world of big data. The term big data refers to the dataset of huge size which are incapable to store in typical database. People are deluged with this continuous increasing amount of data processing which is a storm of data flowing in almost all science research areas like web data, biomedical, Bio-Informatics and other disciplines.

Diabetes mellitus has become a global hazard. Diabetes mellitus is a group of metabolic diseases characterized by raised blood glucose levels (hyperglycemia). There are two major types of diabetes. In type 1 diabetes (insulin-dependent),the body completely stops producing insulin and this form of diabetes generally seen in children or young adults, but can occur at any age. Type 2 (non insulin-dependent) diabetes results when the body doesn't construct enough insulin and/or is unable to make use of insulin properly. This form of diabetes generally occurs in people who are over 40, overweight, sedentary and have a family history of diabetes. Chronic hyperglycemia can leads to visual impairment, blindness, kidney disease, nerve damage, amputations, heart disease, and stroke etc.

The world health organization has estimated the number of diabetics may reach up to 60 million in the world by 2025 and India's contribution to it would be 30 million.. Hence this is a foremost issue and attentiveness towards this disease is essential.

A massive amount of data gets accumulated in the hospitals, most of them just get stored in some form of files which are never turned back, if we analyze this data properly they help in detecting the diabetic related diseases which are to treated as early as possible to avoid the death rate due to this disease. An approach to this hybrid model will help in generating interesting facts which remained unrepealed otherwise; hence taking into consideration this technique used to analyze the diabetic dataset.

Hadoop is an open source software framework for large scale data storage and processing of data sets on a Map Reduce programming model for large scale datasets. MapReduce is a distributed programming model which works on large scale dataset by dividing the huge datasets in smaller chunks. Users specify the computation in terms of a map and a reduce function, and the process automatically parallelizes the computation across large-scale clusters of machines and schedules inter-machine communication to make efficient use of the technique.

2 OVERVIEW OF METHOLOGIES

2.1 K-Means:

Clustering is an important data mining task employed in dataset exploration and in other settings where one wishes to group the objects into clusters . Among the various algorithms, K-means algorithm is the most well-known and commonly used clustering method. It takes the input parameter, k, and partitions a set of n objects into k clusters. The algorithm proceeds as follows: Firstly, it randomly selects k objects from the whole objects which represent initial cluster centers. Each remaining object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster center. The new mean for each cluster is then calculated. This process iterates until the criterion function converges.
Steps of K-means:

- At the initial stage we randomly choose K points from the dataset to choose
the cluster center.
- repeat
- Each object is assigned to the most similar cluster based on the mean value of the objects in a cluster.
- Update the mean value of a cluster

- Until the mean values of clusters remain unchanged.

2.2 SVM

Support Vector Machine (SVM) is one of the most popular and effective algorithms in machine learning. Support Vector Machines (SVMs, also support vector networks) are supervised learning models that used to analyze data by classification and regression analysis. The goal of SVM is to find the best possible separating hyper plane where the margin separates. It works well in both linear and non-linear conditions.

In Linear SVM,We provide a training dataset of n points of the form

$(\vec{x}_1, y_1),\, \ldots ,\, (\vec{x}_n, y_n)$

* y_i are either 1 or −1, each indicating the class to which the point \vec{x}_i belongs.

*vec{x}_i is a p-dimensional real vector.

Any hyperplane can be written as the set of points \vec{x} satisfying

$$\vec{w}\cdot\vec{x} - b=0,\,$$

where {\vec{w}} is the (not necessarily normalized) normal vector to the hyperplane. The parameter $\tfrac{b}{\|\vec{w}\|}$ determines the offset of the hyperplane from the origin along the normal vector {\vec{w}}.

3 Methodologies

This section presents experiments on big data sets to verity the effectiveness of the MRK-SVM algorithm. Testing accuracy and response time are used to measure the performance of hybrid algorithm. In this section we shall discuss about how the hybrid model algorithm works to get the expected results using the combination of K-means and SVM.

4 Proposed Data Model

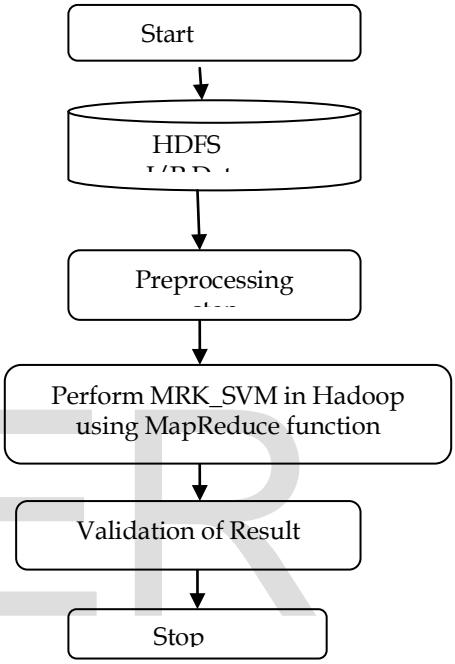
In this model, the big dataset is loaded into HDFS which would then be available for the mapper automatically upon Hadoop streaming process. The Mapper then implements the K Means algorithm to cluster the input data. This clustered data is then used on the SVM algorithm to classify and they are split into number of small datasets. The Mapper finally sends out these data as key-value pairs. Then, Hadoop shuffles and the respective reducer process receive them. The reducer then aggregates these values and saves the results back on to HDFS.

Steps to implement MRK-SVM algorithm:

- Read the input data from the HDFS.
- Input data are received by the Mapper.
- Initialize the center of the clusters and attribute the closest cluster to each data point.
- Set the position of each cluster to the mean of all data points, otherwise repeat steps above two steps until convergence.
- Now the clustered outputs data are used for building classifiers by executing the training dataset and create a model of our data.
- Prediction is done on test dataset and on the model file created by trained dataset.
- End the process with the predictions output file.

The above algorithm is used to predict patients with high risks of getting diseases like Nephropathy, Retinopathy and Cardio Vascular Diseases. This result will help the people who are at high risk due to diabetics which is a major health hazard in developing and developed countries.

Flow diagram of MRK-SVM:



5 Results:

The above algorithm is used to predict the patients with diabetic who are all having the risk of Cardio Vascular Disease, Nephropathy, and Retinopathy. This result will help the people having the risk due to diabetics which is a major health hazard in developing and developed countries.

6 Conclusions

The diagnosis of diabetes is an important real-world problem in day-to-day life. Detection of diabetes related diseases in its early stages is the key for treatment of dreadful diseases . This paper shows how the hybrid model helps in detecting the diabetic's related diseases efficiently and accurately .In future it is planned to assemble the information from different locales or climatic changes over the south India and make a more precise note and make this model as general perceptive for diabetes conclusion.

REFERENCES

[1] "Rule based Classification for Diabetic Patients using Cascaded K-Means and Decision Tree C4.5", Asha Gowda Karegowda e.tal., International Journal of Computer Applications (0975 – 8887) Volume 45– No.12, May 2012.

[2] "Parallel K-Means Clustering Based on MapReduce", Weizhong Zhao, Huifang Ma and Qing He, M.G. Jaatun, G. Zhao, and C. Rong (Eds.): Cloud-Com 2009, LNCS 5931, pp. 674–679, 2009.Springer-Verlag Berlin Heidelberg 2009.

[3] "K-SVM: An Effective SVM Algorithm Based on K-means Clustering", Yukai Yao, Yang Liu, Yongqing Yu, Hong Xu, Weiming Lv, Zhao Li, Xiaoyun Chen, JOURNAL OF COMPUTERS, VOL. 8, NO. 10, OCTOBER 2013.

[4] "Survey On Data Mining Algorithm And Its Application In Healthcare Sector Using Hadoop Platform", K.Sharmila& S.A.Vethamanickam, International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 5, Issue 1, January 2015.

[5] "Support vector machines based on K-means clustering for real-time business intelligence systems", Jiaqi Wang etal., Int. J. Business Intelligence and Data Mining, Vol. 1, No. 1, 2005.

[6] "Big Data Analytics Predicting Risk of Readmissions of Diabetic Patients", Saumya Salian, Dr. G. Harisekaran, International Journal of Science and Research (IJSR), ISSN (Online): 2319-7064 Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438.

[7] "The Prediction, Diagnosis and Treatment of Diabetes Mellitus Using an Intelligent Decision Support System Framework", Adekunle Y.A, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 3, March 2015.

[8] "Benchmarking of Data Mining Techniques as Applied to Power System Analysis", Can ANIL, Department of Information Technology, Uppsala University.

[9] "Diagnosis Of Diabetes Using classification Mining Techniques", Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015.

[10] "Analysis of Diabetic Data Set Using Hive and R", Sadhana, Savitha Shetty, International Journal of Emerging Technology and Advanced Engineering, Volume 4, Issue 7, July 2014.

[11] "K-means Clustering Optimization Algorithm Based on MapReduce", Weizhong Zhao, Huifang Ma and Qing He, M.G. Jaatun, G. Zhao, and C. Rong (Eds.): CloudCom 2009, LNCS 5931, pp. 674–679, 2009.c_Springer-Verlag Berlin Heidelberg 2009.

[12] "Analysis of a Population of Diabetic Patients Databases with Classifiers", Murat Koklu and Yavuz Unal, International Journal of Medical, Health, Pharmaceutical and Biomedical Engineering Vol:7, No:8, 2013.