International Journal of Engineering & Technology, 7 (4) (2018) 4368-4371

SPC

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 14453
Research paper



Effective classification of diabetes using big data analytics

Nikil P. [1] *, Megha P. Arakeri [1]

[1] *Department of Information Science and Technology, Ramaiah Institute of Technology, Bangalore, Karnataka, India*
**Corresponding author E-mail: nikhilg088@gmail.com*

Abstract

Diabetes Miletus (DM) is a non-communicable disease which has affected more people in India. According to the recent survey, Diabetes Miletus stands at fourth place in the world with India alone accounting to around 50 million. Diabetes Miletus is classified as Type 1 and Type 2 diabetes respectively. This disease may prolong for decades and consequently lead to chronic complications such as foot ulceration, neuropathy, retinopathy and nephropathy. Hospitals produce huge amount of patient data which is stored in the database in a structured or unstructured form. This data must be analyzed using automated tools to extract the knowledge which can be used to classify the diabetic data of the patient and provide appropriate treatment at early stages. Thus, helps in improving the standard of health care in India. The existing systems for analysis of diabetes data takes more time, inaccurate and cannot handle large amount of data. In order to overcome this drawback, automated method is proposed in this paper to handle large amount of diabetes data and to classify it as Type1 and Type2. The proposed method uses Hadoop environment coupled with Map Reduce technique to handle large amount of data. Support Vector Machine (SVM) algorithm is used for classification of diabetes into Type 1, Type 2 and Normal. The experiment is carried out on data ranging from 100 MB to 2 GB. Once the data is classified into Type 1 and Type 2, similar data can be retrieved from the hospital database. Based on this result, effective treatment can be provided to the patient.

Keywords: *Diabetes Miletus; Support Vector Machine; Hadoop; Map Reduce.*

1. Introduction

Diabetes is spreading at an alarming pace across the world. Diabetes Miletus usually occurs because of the immune system incapability to secrete glucose and less insulin production in the pancreas [1]. Diabetes Miletus occurs at early or latter half of human life cycle. Diabetes Miletus is classified as Type1, Type 2 or normal. Type 1 DM usually occurs at the early life cycle of human being [2]. Type 1 is not known until 80% of islet cells are destroyed in the pancreas. Patients affected with Type 1 DM have inability to absorb glucose. Since, the islet cells are completely destroyed, the absorption of glucose is very minimal and needs daily dose of insulin. Patient suffering from Type 1 DM have symptoms such as nausea, vomiting and fatigue problems.

Type 2 DM usually occurs at latter half of human life cycle. Patient associated with Type 2 DM develops incapability to absorb glucose and secrete glucose. Usually, the capability of pancreas to secrete insulin will be less in case of patient associated with Type 2 Diabetes. Usually they develop symptoms such as fatigue, sweating and frequent urination. Over decades patients suffering from Type 1 or Type 2 DM are prone to chronic complications like nephropathy, neuropathy, retinopathy, and foot ulceration.

Data is growing at fast rate especially in the health sector. Data is obtained in health sector is available in sources such as electronic health record, physician notes, accounts etc. Big Data is evolving technology in the current market trend because of its capability to leverage and incorporate huge amount of data [3]. Big Data comprises of [3] V's Velocity, Volume and Variety with Veracity 4 V being optional. Velocity comprises of data in real time usually twitter tweets, stream of data such as demographic information or patient information. Volume comprises huge amount of data in gigabytes or petabytes. Variety comprises of structured or unstructured data [4]. Big Data environment provides better flexibility and memory pooling than traditional database.

In this paper, SVM classification algorithm is used in Map Reduce/Hadoop environment to identify the class of diabetes data. Map Reduce technique consists of two phases: Map phase and Reduce phase, where in map phase constitutes of filtering and sorting technique and reduce phase constitutes of aggregating and combining results. Big data analytics framework is used to handle large amount of data stored in the hospital database.

2. Related works

Various techniques in field of image processing has been deployed for classification of diabetes. Most of the researchers have used Image processing, artificial intelligence and Fuzzy logic techniques to classify the type of diabetes. Big Data is incorporated in most of the health organization due to its ability of scaling out, replication of data and availability factor. Effort has been made by many researchers to analyze the clinical dataset in the Hadoop platform by using different open source components in Big Data.

Rajesh et al. [5] proposed a methodology for classification of diabetes related to mining of dataset using various algorithm techniques. This paper provides techniques and method for efficient classification of diabetes dataset. The limitation in this technique is that the data involved in the classification of diabetes was relatively less. Zolfaghar et al [6] proposed apache Hadoop for storage, retrieval and processing huge volume of data effectively was deployed for analysis of the heart rate and the complications associated with it was highlighted in this technique. They used both logistic regression and Naïve Bayes algorithm for classification of the data. This method was less accurate in predicting the data. Aishwarya et al. [7] came with solutions to diagnose the type of

Copyright © 2018 Nikil P, Megha P. Arakeri. 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.

diabetes by analyzing the patterns in the data found by employing Decision tree and Naïve Bayes algorithm. Timely treatment of patients and efficient way of analyzing the disease is the main focus area in this paper. But, the method developed classification model on small data set and hence inaccurate. Sarvana Kumar et al. [8] proposed predictive algorithm in Hadoop environment to classify the type of diabetes, the type of complications associated and treatment to be followed. This work also provided solutions for type of treatment for efficient care and cure of patient. However, only the framework of the solution was proposed in this paper. Implementation and performance analysis was not presented.

The method proposed by Veena et al. [9] involved computation of missing value and imputation for both numerical and categorical data. A hybrid combination of classification and regression trees (CART) and genetic algorithms was used to impute missing continuous values and self-organizing feature maps (SOFM) to impute categorical values. One limitation with respect to this work was dataset was relatively less. Christy et al. [10] proposed two algorithms namely distance and cluster based algorithm for detecting and removing outliers. The idea of the work was to remove the key attribute rather than the whole dimensional set. Based on this idea, the effective and efficient analysis of dataset was performed. Classification technique used in analysis of this dataset was not efficient. The time factor and dataset involved in the classification of diabetes also played major role.

Abdullah et al. [11] work involved prediction of diabetes using regression based approach with data mining tool. It involved two age groups, where in treatment of old age group had to be carried out immediately and young age groups were delayed. One limitation was the classification was done on two age groups young and old age. Also the analysis was done on particular region which constituted small dataset. Savita et al. [12] proposed the analysis of dataset using hive and R. In this paper, the classification is done using hive and later R for statistical results. The limitation with respect to this paper was more from classification prospective wherein specific machine learning algorithm was not used. Maniruzzaman et al. [13] proposed machine learning technique for classification of diabetes. The Gaussian process classification technique was used for classification of diabetes. This method could not accurately predict the class of unknown data. Nilashi et al. [14] used different algorithm for various process during classification of diabetes. Algorithm such as SOM, PCA and NN were used for clustering, noise removal and classification of diabetes. PIMA Indian Diabetes dataset was used for classification of diabetes which is relatively small in size [15].

From the above survey, it is observed that most of the techniques deployed traditional database with less computing power and more time involved. Hence, in the proposed paper big data technology is adopted, coupled with SVM algorithm for analysis of diabetes data set in an effective and efficient way.

3. Proposed system

The proposed methodology details out the step involved in the classification of diabetes. The proposed methodology discusses about collection of data from various data sources, refining, merging the data and finally classification of diabetes. Once type of diabetes is classified, the most similar patient information can be retrieved from the database. Based on the patient information retrieved, necessary treatment can be provided. The block diagram of the proposed system is shown in Fig.1 and the steps involved in the classification of diabetes are given below.

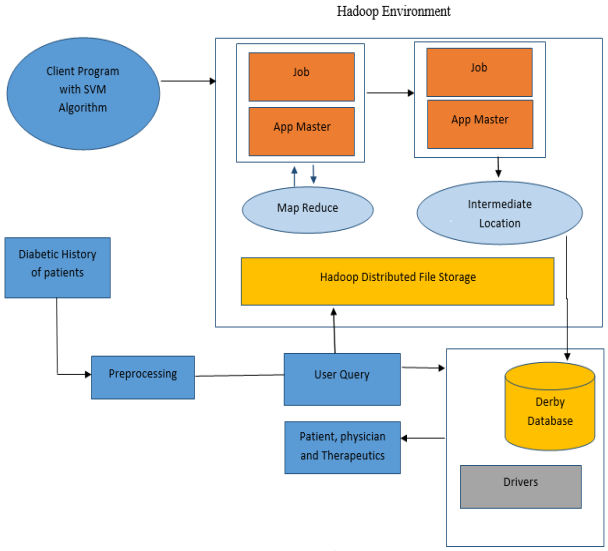


Fig. 1: Block Diagram of Proposed System.

3.1.1. Data acquisition

Diabetes dataset is collected from various sources such as PubMed, UCI machine learning repository and hospitals. The data collected from these sources was in raw format [16]. Thus, the data was converted to a format which is suitable for processing. The diabetes dataset comprises of various parameter such as post prandial sugar level, Fasting blood sugar level, oral intake, and HbA1c, magnesium, potassium and sodium levels [17]. Normally Age is considered as one of the important parameter for the classification of diabetes [18].

3.1.2. Data preparation

Once the data is collected from various sources, it is prepared in a format suitable for processing on Hadoop environment. The raw data collected will be converted into text format. The list of files will be merged using Hadoop copy commands. After data is converted into text, it is stored in local repository. Data in Hadoop can be stored on Hadoop distributed file storage which is same as New Technology file system (NTFS) in Windows [19]. The data from local repository to Hadoop distributed file system can be moved in three ways using flume, sqoop or Hadoop copy commands [20]. In this work, the data from local repository is loaded into Hadoop distributed file system using Hadoop copy commands. Once the data is loaded to the Hadoop distributed file system, the default size will be 256MB in size [21]. For 2 GB file there will be 8 files split across the directory with each file being 256 MB in size [22]. After data is loaded into Hadoop distributed file system, a jar is created with SVM implementation. On the Hadoop client, the jar file is run against the dataset and diabetes is classified into Type 1, Type 2 and normal.

3.1.3. Classification and detection of diabetes

In this phase, the SVM algorithm is used for classification of diabetes into Type 1 DM, Type 2 DM and Normal. SVM algorithm is widely used for classification and pattern solving techniques in the machine learning domain. SVM mainly focuses on two classes either by maximizing or minimizing the hyperplane as shown in Fig.2 [24]. The points close to the hyper plane are called support vectors . The separating hyperplane H is defined as:

$$w \bullet x_i + b \geq +1 \text{ when } y_i = +1 \quad (1)$$

$$w \bullet x_i + b \leq -1 \text{ when } y_i = -1 \quad (2)$$

The two planes H1 and H2 are the planes:

H1: w•x_i +b = +1 (3)

H2: w•x_i +b = –1 (4)

The plane H0 is the median in between, where w•x_i +b =0
The d+ and d- shown in figure 2 indicates the shortest distance to the closest positive point and shortest distance to the closest negative point respectively [25]. The margin (gutter) of a separating hyperplane is d+ + d–.

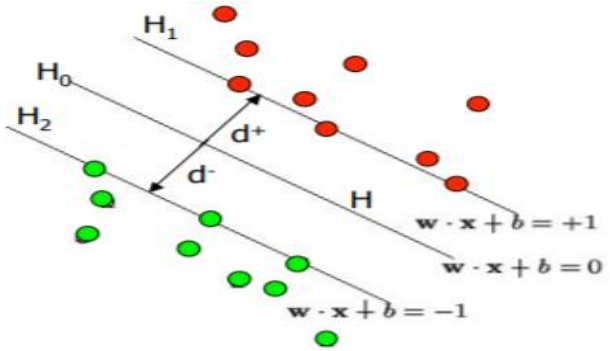


Fig. 2: SVM Technique Used For Classification of Diabetes.

The proposed system uses multiclass classification to classify three different categories during classification of diabetes. Multiclass combination is a technique where in one-many and one-one mapping can be done. For classification purpose, a trained dataset and test dataset are taken in the ratio 70:30.
Also, the proposed methodology uses Map Reduce technique to accommodate huge amount of data for classification purpose. Here the data is spilt, shuffled, aggregated to obtain the suitable result. The data is split based on various parameters such as age, post prandial sugar level, fasting blood sugar level and HBA1C levels [25]. The jar file is created and a command is executed against the jar for processing of data into Type 1 DM, Type 2 DM and Normal. The data will be stored in part files in the Hadoop Distributed File Location. The default size of part files will be 256 MB and will be available in the intermediate location. The algorithm for classification of diabetes data is given below.
Algorithm: Classification of diabetes.
Data: Diabetes Dataset D, Parameters P (p1,p2,p3….pn), Test Data Td and Train Data Tr
Result: Classified Set of Data C1, C2 and C3 denoted with ‘0’,’1’ and ‘2’.
Begin

1) Read the data D. Arrange the required parameters p→p1,p2….pn.
2) After input of data, begin the job for the specified dataset.
3) Tr ← read csv of Train dataset.
4) Td ← read csv of Test dataset.
5) Set the configuration to initiate the map job.
6) For i = 1 to n
7) model← svm(Target~P1+P2+P3,D=Tr,kernel='linear')
8) preds ← predict(model,Td)
9) Shuffle and sort the input to specific combiner
10) After shuffling and sorting, the data is sent to the reducer.In reducer, the data is classified as Type 1 and Type 2 DM with result interpreting as ‘1’ and ‘2’.

End the job.
End

3.1.4. Retrieval of patient information

The last phase uses hive for querying of patients information from the storage. Hive is installed in the user location on the host. After the installation, the configuration related to Hive files is set. In the next step, the hive shell is triggered with Hive command on top of Hadoop shell. Hive works when Hadoop shell daemons are working. After setup of hive, the next step is used for querying purpose. The classified data from the previous step is stored in the intermediate location in Hadoop distributed file system.
Using Hive, an external table is referenced to the intermediate location. The purpose of external table is for storing the information in metadata. Even if table is deleted, the data remains preserved in the storage. This is one of the advantages if a table is referenced externally. After reference, the next step is to query the information from the table. The particular type of diabetes can be queried based on the result obtained from previous step. Normally Type 1, Type 2 and Normal is indicated by ‘1’.’2’ and ‘0’. From the classified information, more details with respect to patient diabetic condition can be retrieved. Also particular patient information can be retrieved using patient id which outlines parameters along with type of diabetes. Also particular patient information can be classified by matching with records residing in the database.

4. Results

The proposed system uses diabetes data, Integrated Development Environment, Hadoop and Hive environment for classification of diabetes and retrieving patient’s information. The information of diabetes can be retrieved by physicians and this form basis for effective treatment to be provided to the patient.
The Hadoop interface uses commands for classification and retrieval purpose of data. Dataset is partitioned in the ratio of 70:30 for training and testing purpose for classification of diabetes. Around 10000 records are used in ratio 70:30 for classification purpose.

a) Performance Analysis

The performance is usually measured in metrics such as accuracy, error rate, precision and recall as given below:

$$\text{Accuracy} = \frac{(T_P+T_N)}{(T_P+F_P+F_N+T_N)} *100 \quad (5)$$

$$\text{Specificity} = \frac{T_N}{(T_N+F_P)} *100 \quad (6)$$

$$\text{Sensitivity} = \frac{T_P}{(T_P+F_N)} *100 \quad (7)$$

$$\text{Precision} = \frac{(\text{Total records identified})}{(\text{Total records identified} + \text{records retrieved postively})} *100 \quad (8)$$

$$\text{Recall} = \frac{(\text{Total records identified})}{(\text{Total records identified} + \text{records retrieved negatively})} *100 \quad (9)$$

Where Tp is true positive, Tn is true negative, Fp is false positive and Fn is false negative. Normally error rate and accuracy are considered as classification performance during classification of diabetes. The error rate and precision is calculated and results are tabulated below for the type of diabetes occurred. Various parameters indicating the performance of classification and retrieval of data are shown in Table.1.
Table 1 indicates the classification performance for different type of diabetes. The error rate, recall, accuracy and precision is provided for different type of diabetes is carried out on 70:30 trained and test data set respectively.

Table 1: Confusion Matrix

	N=3000	Predicted	
		Type 1	Type 2
	Type 1	1024	223
Actual	Type 2	299	1454

Table 2: Performance Quality of Classification Retrieval

Metric	Value in %
Accuracy	82.6
Specificity	82.94
Sensitivity	82.11
Precision	77.39
Recall	82.98

The accuracy obtained in the classification of diabetes is 82.6%. One of the reason for not getting 100% accuracy is because of missing value in the dataset. Moreover good accuracy is obtained during classification of diabetes. This provide basis for the efficient and timely treatment to the patient.

The time taken during classification of diabetes is considered as one of the important parameter for analyzing the efficiency of handling the dataset ranging from small dataset to large dataset. The time efficiency will give an idea of how time plays an important role during classification of diabetes when huge dataset is considered. The time taken for classification of Type 1 and Type 2 diabetes is shown in Fig. 3. The above graph indicates the time with respect to dataset is not growing exponentially. The graph provides clear picture as dataset grows in size, the time taken is relatively less.

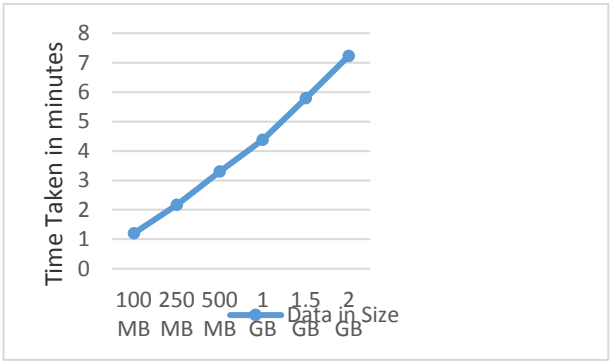


Fig. 3: Time Taken for Classification of Diabetes.

5. Conclusion

This paper provided an automatic method for handling large amount of diabetes data in hospitals. In the proposed system, data from various sources is collected and prepared in a format suitable for processing of dataset on Hadoop environment. The dataset is then loaded into the Hadoop environment. Using SVM algorithm and map reduce techniques, the data is classified into Type 1 DM, Type 2 DM and normal. Finally the classified data can be used to retrieve the most similar patients' data from the database. Based on the information retrieved an efficient and effective treatment can be provided.

The accuracy for classification of diabetes is 82.6 % using SVM algorithm coupled with Map Reduce technique. The time taken during the classification for huge dataset is minimal. For future analysis, spark can be used for classification of diabetes. Here machine learning technique (Mlib) and spark environment can be used for effective analysis.

References

[1] AMERICAN DIABETES ASSOCIATION. "DIAGNOSIS AND CLASSIFICATION OF DIABETES MELLITUS." *DIABETES CARE* 33, NO. SUPPL 1 (2010): S62.

[2] Treece, K. A., R. M. Macfarlane, N. Pound, F. L. Game, and W. J. Jeffcoate. "Validation of a system of foot ulcer classification in diabetes mellitus." *Diabetic medicine* 21, no. 9 (2004): pp. 987-991. https://doi.org/10.1111/j.1464-5491.2004.01275.x.

[3] http://www.intel.com/content/www/us/en/healthcare-it/bigger-data-better-healthcare-idc-insights-white- paper.html

[4] Muni Kumar, N., and R. Manjula. "Role of Big data analytics in rural health care-A

[5] Step towards svasth bharath." *International Journal of Computer Science and Information Technologies* 5, no. 6 (2014): 7172-7178.

[6] Rajesh, K., and V. Sangeetha. "Application of data mining methods and techniques for diabetes diagnosis." *International Journal of Engineering and Innovative Technology (IJEIT)* 2, no. 3 (2012).

[7] Zolfaghar, Kiyana, Nele Verbiest, Jayshree Agarwal, Naren Meadem, Si-Chi Chin, Senjuti Basu Roy, Ankur Teredesai, David Hazel, Paul Amoroso, and Lester Reed. "Predicting risk-of-readmission for congestive heart failure patients: A multi-layer approach." *arXiv preprint arXiv: 1306.2094* (2013).

[8] Iyer, Aiswarya, S. Jeyalatha, and Ronak Sumbaly. "Diagnosis of diabetes using classification mining techniques." *arXiv preprint arXiv: 1502.03774* (2015).

[9] Eswari, T., P. Sampath, and S. Lavanya. "Predictive methodology for diabetic data analysis in big data." *Procedia Computer Science* 50 (2015): 203-208 https://doi.org/10.1016/j.procs.2015.04.069.

[10] Bhat, Veena H., Prasanth G. Rao, S. Krishna, P. Deepa Shenoy, K. R. Venugopal, and Lalit M. Patnaik. "An efficient framework for prediction in healthcare data using soft computing techniques." In *International Conference on Advances in Computing and Communications*. Springer, Berlin, Heidelberg, 2011, pp. 522-532. https://doi.org/10.1007/978-3-642-22720-2_55.

[11] Christy, A., G. Meera Gandhi, and S. Vaithyasubramanian. "Cluster based outlier detection algorithm for healthcare data." *Procedia Computer Science* 50 (2015): 209-215 https://doi.org/10.1016/j.procs.2015.04.058.

[12] Aljumah, Abdullah A., Mohammed Gulam Ahamad, and Mohammad Khubeb Siddiqui. "Application of data mining: Diabetes health care in young and old patients." *Journal of King Saud University-Computer and Information Sciences* 25, no. 2 (2013): 127-136. https://doi.org/10.1016/j.jksuci.2012.10.003.

[13] Sadhana, Savitha Shetty, and S. Shetty. "Analysis of diabetic data set using hive and R." *International Journal of Emerging Technology and Advanced Engineering* 4, no. 7 (2014): 626-9.

[14] Maniruzzaman, Md, Nishith Kumar, Md Menhazul Abedin, Md Shaykhul Islam, Harman S. Suri, Ayman S. El-Baz, and Jasjit S. Suri. "Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm." *Computer methods and programs in biomedicine* 152 (2017): 23-34. https://doi.org/10.1016/j.cmpb.2017.09.004.

[15] Nilashi, Mehrbakhsh, Othman Bin Ibrahim, Abbas Mardani, Ali Ahani, and Ahmad Jusoh. "A soft computing approach for diabetes disease classification." *Health Informatics Journal* (2016):1460458216675500.

[16] Alberti, Kurt George Matthew Mayer, and PZ ft Zimmet. "Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus. Provisional report of a WHO consultation." *Diabetic medicine* 15, no. 7 (1998): 539-553. https://doi.org/10.1002/(SICI)1096-9136(199807)15:7<539::AID-DIA668>3.0.CO;2-S.

[17] American Diabetes Association. "Diagnosis and classification of diabetes mellitus." *Diabetes care* 33, no. Suppl 1 (2010): S62.

[18] Treece, K. A., R. M. Macfarlane, N. Pound, F. L. Game, and W. J. Jeffcoate. "Validation of a system of foot ulcer classification in diabetes mellitus." *Diabetic medicine* 21, no. 9 (2004): pp. 987-991. https://doi.org/10.1111/j.1464-5491.2004.01275.x.

[19] White, Tom. *Hadoop: The Definitive Guide*. " O'reilly Media, Inc.", 2012

[20] Capriolo, Edward, Dean Wampler, and Jason Rutherglen. *Programming Hive: Data Warehouse and Query Language for Hadoop*. "O'reilly Media, Inc.", 2012.

[21] Eadline, Douglas. *Hadoop 2 Quick-Start Guide: Learn the Essentials of Big Data Computing in the Apache Hadoop 2 Ecosystem*. Addison-Wesley Professional, 2015.

[22] Zikopoulos, Paul, and Chris Eaton. *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media, 2011.

[23] Marjanović, Miloš, Miloš Kovačević, Branislav Bajat, and Vít Voženílek. "Landslide susceptibility assessment using SVM machine learning algorithm." *Engineering Geology* 123, no. 3 (2011): 225-234. https://doi.org/10.1016/j.enggeo.2011.09.006.

[24] Provost, Foster, and Tom Fawcett. "Data science and its relationship to big data and data-driven decision making." *Big data* 1, no. 1 (2013): 51-59. https://doi.org/10.1089/big.2013.1508.

[25] Bellamy, Leanne, Juan-Pablo Casas, Aroon D. Hingorani, and David Williams. "Type 2 diabetes mellitus after gestational diabetes: a systematic review and meta-analysis." *The Lancet*373, no. 9677 (2009): 1773-1779. https://doi.org/10.1016/S0140-6736(09)60731-5.