International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 25 Issue 7 – JUNE 2018.

Data mining techniques using e-health information for diabetes disease prediction

S.V.Vinodini [#1] and Dr.S.Saravanan[*2]

[#] *Department of CSE, Dhirajlal Gandhi College of Technology, Salem, India*

[*] *Associate Professor, Department of CSE, Dhirajlal Gandhi College of Technology, Salem, India*

***Abstract*— Data mining techniques have been used in many medical applications for predicting and diagnosis a particular disease with better accuracy. These techniques are mainly used to search the hidden patterns present in the medical data. The data selection method is used to select the data from the patient report and analyse those data to find the stage of the disease. This application can classify the disease into three stages as high, medium and low. After analysing the stage level of the patient this application recommends both the medicine and food according to the health condition of the individual patient. This paper mainly focused on diabetes disease and suggests the treatment for the patient.**

***Index Terms*— Deep Learning, Data mining techniques, Prediction Model, Data selection, Cart algorithm.**

I. INTRODUCTION

In recent years the application of Data Mining increases due to its effective way of processing and extracting useful information from a huge database. In general, the professional analyze the patient by the symptoms that are stated by patient. From the symptoms and the reaction of the body according to the symptoms the professional represented as doctor suggest the solution by treatment and their experience according to user body condition. Therefore the symptoms play a major role in analyzing and identifying the disease that user suffering from.

Hence the proposed system concentrates on analyzing the symptoms specifically. This is an initial step to start a diagnosis of a patient if any mistake is done then it leads to wrong suggestion of treatment. The result of effective diagnosis leads to efficient identification of disease therefore these processes are done systematically by means of clustering implementation in medical datasets that concentrates on analyzing the symptoms. Domain relation based analysis is much required for effective analysis and identification of disease.

Diabetes, often referred to by doctors as diabetes mellitus, describes a group of metabolic diseases in which the person has high blood glucose (blood sugar), either because insulin production is inadequate, or because the body's cells do not respond properly to insulin, or both. Patients with high blood sugar will typically experience frequent urination; they will become increasingly thirsty and hungry.

In deep learning, a computer model learns to perform classification tasks directly from images, text, or sound. Deep learning models can achieve state-of-the-art accuracy, sometimes exceeding human-level performance. Models are trained by using a large set of labelled data and neural network architectures that contain many layers.

II. RELATED WORKS

Wrong clinical decision taken by the medical representatives is the problem of the paper. It can lead to serious loss of a life of a patient. To overcome this problem in this paper they used multiclass Naive Bayes algorithm [1] for the prediction a particular disease by analysing the patient's medical report. The proposed system can solve those problems to detect the particular disease and can also help the medical representatives to take a smart clinical decision.

Prediction of heart disease is the issue of this paper which can be resolved by using Data mining techniques. In this paper they used two data mining techniques such as Naive Bayes & Decision tree[2] for the effective prediction of Heart disease. The most accurate predictive system was tested among the two and they found that decision tree was more effective in the prediction of heart disease.

This paper discuss about the prediction of heart disease in diabetic patient [3]. For this prediction purpose they use data mining techniques such as classification, prediction and time series analysis. It shows more accuracy compared to other techniques used. A medical practitioner trying to diagnose a disease based on the medical test results of a patient can be considered as a predictive data mining task.

This paper focused on Data mining techniques on healthcare issue and uses on health care sector. In this paper two major diseases such as heart disease and cancer disease are diagnosed. Knowledge Database Discovery (KDD) [4] process as one of the part in Data mining techniques which is used for diagnosing purpose. By surveying this paper it shows that data mining techniques gives accurate results regarding health issues.

III. MINING HIGH UTILITY ITEMSETS:

A. *Tree Structure*

UP tree is used to maintain the information about the transaction and utility items. In a UP tree, two strategies are applied to reduce the overestimated utilities stored in the node of the tree.The elements which consist in a Up tree are N.name, N.nu, N.parent, N.count, N.hlink and child nodes. the header table is used to facilitate the traversal of Up tree.

The header table consist of the entry records of an each item name and its link.

1) DGU

The global Up Tree is constructed by only two scans of the original database. In the first scan,TU of each item is found and at the same time,of each single items are also found. By TWDC property, the unpromising itemsets are found.the unpromising itemset means which TWU is less than the minimum utility threshold.during the second scan of the database,the transactions are entered into a tree. After retrieved the transaction, the unpromising items should be removed from the transaction and its utilities are also removed from the transaction.NewTU,after pruning unpromising item and sorting the remaining items in any order is known as RTU

2) DGN

By using this strategy DGN, the utilities of the nodes that are closer to the root of a global up tree are reduced.DGN is suitable for the database contains the long transactions. They use the divide and conquer technique in mining processes. The search space are divided into smaller subspaces.

For example,

{b}'s conditional tree

{a} does not contain {b}tree

{d}does not contain {b}and{a}

{c}does not contain {b},{a}and{d}

{e} does not contain {b},{a},{d}and{c}

The searching is starts from bottom of the tree. The nodes does not appear the descendant nodes. the proposed strategies is used for decreasing overestimated utilities is remove the descendant nodes in a tree.

B. DLU and DLN

They are pushing the two more strategies into the FP Growth. By pushing these two strategies overestimated utilities are decreased and the number of PHUIs can be reduced.

1) DLU

The algorithm contains tree steps1. Generate the conditional pattern bases for tracing the trees original path,2.conditional tree are to be constructed is calle local tree. 3.mine the patterns from conditional trees. By using DLU, minimum item utilities are utilized to reduced utilities of local unpromising items in conditional pattern bases. the local unpromising items are subtracted from the path utility of an extracted path.

2) DLN

In DLN, the path are reorganized by pruning unpromising items and resorted in any fixed order. These paths are known as reorganized path.DLU and DLN are can be local version of the DGU. By using , these two strategies, overestimated utilities for itemsets can be locally reduced without losing an actual high utility itemsets.

C. Heuristic rules

Heuristic rules are used for better decision making process. Potential high Utility itemsets are found by four strategies . An association rules for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using different measures of interestingness. A rule states as follows A-> B (The logs A are being visited followed by B) , The support and count is being measured.

An item L with their minimal node utility, path utility and counts, The second step is to find all valid rules with the itemsets . The rules we need to find out in this step can be classified into two groups by verifying whether the item ir is the least frequent item in the whole rule or not: 1. the rules with i_r on the antecedence and ir is the least frequent item in the rule. 2. the rules with i_r on the antecedence, but there exists at least one item with its occurrence count less than i_r on the consequence. We can find out the first kind of rule directly by using the information found in the first step. Our method is to compare any two itemsets IS_A and IS_B found in the first step. If one is the subset of the other, say IS_A is the subset of IS_B. Then let ISC be the

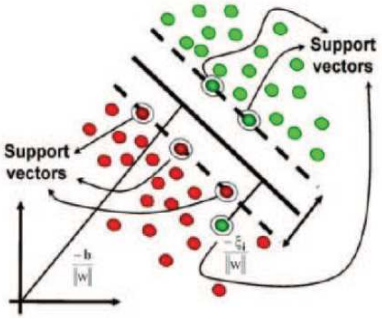

Figure 1:Classification process of SVM

difference of IS_B and IS_A. To verify the validity of the rule IS_A !IS_C , we need to verify the following rule

IS_A-> IS_C =P(IS_A ->IS_c)/P(IS_A)

On solving this equality the subsequent strength of the rule and its validity is being found. Thus on adopting all these the high utility items are evaluated , with strongly saying in terms of rules. Thus Proposed methodology will adopt for many number of transactions/logs.

For Example,

If totally 5 logs are being depicted as high utility logs say (A,B,C,I,K), the possible sets are extracted, which contains 2 combination,3 combination, 4 combination, etc

(A,C) -> 2combination

(C,J)->2 combination

(C,J,K)-> 3 combination

Finally evaluating high utility patterns which means(A,C,K,J).Many user has used A,C,K,J items

Repeatedly, predicted by highest confidence.

D. Identify High Utility Pattern

After finding the PHUIs, now identify actual high utility itemsets and sets of high utilities are produced by scanning the original database one time.

1) Runtime of Classification

The runtime is the time taken to classify sentiments using the existing and proposed methods has shown in figure 3. From the graph, these classifiers are tested on different size of features 10000, 20000, 30000, 40000 and 50000. The proposed method SVM with lexicon based dictionary

approach has taken the shortest time to classify the total online product review dataset, needs 1500ms, 1700ms, 2998ms, 3000 ms in the feature size of 10000, 20000, 30000 and 40000 to preprocess, extract features and classify the sentiments as positive, negative or neutral sentiments on online product review dataset.

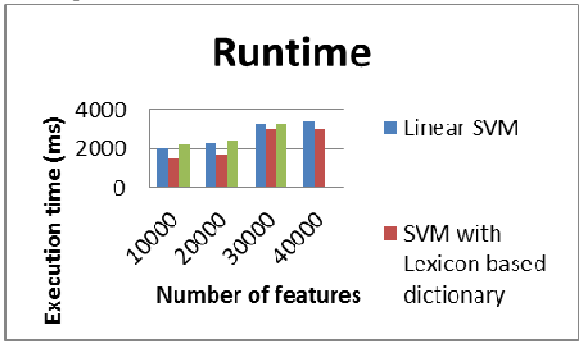

Figure 2: Classification runtime of existing and proposed methods

Existing methods linear SVM takes time 2000ms, 2300, 3255ms, 3400ms and Maximum entropy method has taken 2200, 2450, 3300, 3350 ms to classify the sentiments in the feature size of 10000, 20000, 30000 and 40000. Hence, from the results proposed system has taken less runtime to classify the sentiments compared to existing methods linear SVM and Maximum entropy.

IV. CONCLUSION:

In this paper, CART algorithm has been applied for prediction model to obtain maximum accuracy, for a categorized dataset. Blood Pressure is a remarkable factor in causing diabetes, along with the others such as roadside, heredity and improper diet maintenance. Hence it can be said that one must enjoy every perspective of their life but a little vigilant in one's daily routine does no harm to their health.

REFERENCES

[1] Ajinkya Kunjir, Harshal Sawant and Nuzhat F.Shaikh, "Data Mining and Visualization for Prediction of Multiple Diseases in HealthCare," International Conference on Big Data Analytics and computational Intelligence (ICBDACI), 978-1-5090-6399-4/17/$31.00 © 2017 IEEE.

[2] Priyanka N and Dr.Pushpa RaviKumar, "Usage of Data mining techniques in predicting the Heart diseases – Naïve Bayes & Decision tree," International Conference on circuits Power and Computing Technologies [ICCPCT], 978-1-5090-4967- 7/17/$31.00 © 2017 IEEE.

[3] Charu .V. Verma and Dr. S. M. Ghosh," Review of Cardiovascular Disease in Diabetic Patients using Data Mining Techniques," International Journal of Engineering Science and Computing May 2017 (IJESC), Volume 7 Issue No.5

[4] Gokul Shah and Sangeeta Oswal, "A Study on Data Mining Techniques on Healthcare Issues and its uses and Application on Health Sector," International Journal of Engineering Science and Computing June 2017(IJESC), Volume 7 Issue No.6

[5] Benny Lo, Charence Wong, Daniele Rav`, Fani Deligianni, Guang-Zhong Yang, Javier Andreu-Perez and Melissa Berthelot," Deep Learning for Health Informatics," IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. 21, NO. 1, JANUARY 2017

[6] Ashwini Kamath and Deepashri K.S,"Survey on Techniques of Data Mining and its Applications," Special Issue on International Conference on Emerging Trends in Engineering (ICETE) -2017, International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-6, Issue-2)

[7] Ayush Anand and Divya Shakti," PREDICTION OF DIABETES BASED ON PERSONAL LIFESTYLE INDICATORS," 2015 1st International Conference on Next Generation Computing Technologies (NGCT-2015), 978-1-4673-6809-4/15/$31.00 ©2015 IEEE

[8] Manjiri Harmalkar, Manali Bhoir, Nikita Kamble and Supriya Chaudhary,"Smart Health Prediction System Using Data Mining," International Journal of Scientific Research in Computer Science, Engineering and Information Technology © 2017 IJSRCSEIT | Volume 2 | Issue 2 | ISSN: 2456-3307.

[9] S.Babu and J.Jamila Yasmin Banu," A STUDY ON DIABETES HEALTHCARE PATHWAY PROCESS USING DATA MINING TECHNIQUES," IPASJ International Journal of Computer Science (IIJCS) Volume 5, Issue 9, September 2017 ISSN 2321-5992.

[10] Messan Komi, J un Li, Y ongxin Zhai and Xianguo Zhang," Application of Data Mining Methods in Diabetes Prediction," 2nd International Conference on Image, Vision and Computing 978-1-5090-6238-6/17/$31.00 ©20 17 IEEE

[11] G.Krishnaveni and T.Sudha,"A NOVEL TECHNIQUE TO PREDICT DIABETIC DISEASE USING DATA MINING – CLASSIFICATION TECHNIQUES," International Conference on Innovative Applications in Engineering and Information Technology (ICIAEIT-2017), International Journal of Advanced Scientific Technologies, Engineering and Management Sciences (IJASTEMS-ISSN: 2454-356X) Volume.3, Special Issue.1, March.2017.

[12] Anita Shaikh, Saman Hina and Sohail Abul Sattar," Analyzing Diabetes Datasets using Data Mining," Journal of Basic & Applied Sciences, 2017, 13, 466-471.