ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

A Review on Prediction of Diabetes Mellitus Disease Using Association Summarization Techniques

Yogita S. Wanjari[1], Prof. Gaurav Y. Kawade[2]

M.Tech Student, Dept. of Computer Science & Engineering, G.H.RC.E, Nagpur, India[1]

Professor, Dept. of Computer Science & Engineering, G.H.R.C.E, Nagpur, India[2]

ABSTRACT: Due to the change in life style of people many diseases can affect on the human body, these diseases are very dangerous for the health care issue. Now a day the most promising and growing disease is the diabetes it is very common and caused to any age group people. But the main fact is that the people have no idea about that they are suffering from such harmful diseases, and the presence of that diabetes some other diseases can also be detected. The main aim of this system is to detect the rapidly developing diabetes mellitus patients. This is very fast growing disease as compare to the other diseases. Early detection of patients with elective risk factor of diabetes is the today's medication need, for that in this system the data mining techniques are used on the diabetic patient's database and then apply the TOPK algorithm and MOPNAR algorithm for identifying that the patients can suffering from such disease or not. For the identification of this disease the association summarization techniques are very effective technique. This system can be used as the expert system in hospital and in the pathologies for the detection of the diabetic patients.

KEYWORDS: Diabetic mellitus, Data Mining, Association Summarization Techniques, TOPK algorithm, MOPNAR algorithm.

I. INTRODUCTION

Today's world is the digital world all the people are surrounded with the electronic or digital gadgets, every individual wants to be digital. In this today's digitalized world every person not just wants to be digital but also fit and fine. Now a day all are gathered with electronic device, all the physical work is replace by the machines it is beneficial to us but it takes major impact on the human body. Some years ago there are not all work done by the machines means theses all are done by the individual, but now it replace by machines, the individual do not take any efforts to do any task because these work is done by the machines. That's why the human life style is going to be change and it affect on the human body. Due to this changed life style many diseases can be caused .but in this 21stcentury everyone wants to be physically fit and they become more health cautious. The people not only used lots of electronic device to reduce this work load but also want to be capable. People know about these changes life style and they also have idea it will impact there lifecycle also, that's why they wants to aware about the disadvantages of this changed life style. Many diseases are caused due to the changes comes in the people life cycle and there is need to be aware about all these diseases .In the list of this disease the first rank disease is the diabetes. Diabetes mellitus is the rapidly growing diseases that affecting 28.8 million people in the US that means 8% people of its total population and the people have no idea that they have such disease. Diabetes is the common and it is caused to anyone it doesn't have some special symptoms that's why it is really essential to early identification of this disease [5] [1]. The presence of diabetes can caused such other co-morbid diseases such as retinopathy, heart failure, mental disability, stroke, neuropathy, hit, skin complication, hearing loss.etc. Appropriate management of patients with risk and changed life cycle and medication can decrease the risk of developing diabetes by 30% to 60% [2]. Some decades ago diabetes is the normal disease but today diabetes is

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

only caused of many co-morbid diseases. So it is the major health care need to detect such kind of diseases early or to detect the risk factor of these diseases.

A WHAT IS DIABETES? [3]

Diabetes can be caused anyone, from any walk of life. Diabetes is a condition in which the body cannot properly process the food which is used as the energy to do the metabolism. The food which enters in the human body is converted into the sugar, or glucose which is used as energy and with the help of this energy all the body work can be done properly. There is an organ which is near to the stomach called pancreas which make a hormone called insulin. Human body is made by the number of cells and the hormone which is called as insulin which serves as key to open the cell and make a way available to enter to the glucose in the cells which are present in the human body. When anyone suffer from the diabetes means the process of hormone secretion of insulin is not done properly or cannot properly use the secreted insulin, an alternatively the presence of the sugar or the glucose in the body will increasingly excess as the required manner because it is not in used to the creation of energy. This is why people refer to it as "sugar" or "high blood sugar", because the sugar level increases in the body. The diabetic is the major and dynamically growing disease in the world it takes more lived than the AIDS and the Cancer. Diabetes is the seventh leading reason of death in the united state.

B. TYPES OF DIABETES:

As the name of the disease it is of two types of diabetes first is the type 1 diabetes and the second one is the type 2 diabetes.

- Type 1:-

  This type 1 diabetes can caused due to the no properly production of insulin in the body it can caused early phase of life means early in the adult age or the teenage or because the 45th year. Type 1 diabetes can also called as insulin dependent diabetes mellitus (IDDM), this type 1 diabetes can directly depend on the production of insulin in the body. The percentage of presence of type 1 diabetes cans approximately 5-10 only. The risk factor of type 1 is less than the type2 diabetes, but the lots of factor are involve in the development of this type of diabetes.

- Type 2:-

  This type 2 diabetes perversely called as non insulin dependent diabetes mellitus (NIDDM) or adult- onset diabetes. Type 2 diabetes having 90-95 percent chances to be diagnosis. This type 2 diabetes can caused when the body does not produce the enough insulin for the proper function or in other hand it can be the insulin resistance. The risk factor of type 2 diabetes is more than the type 1 and it may caused due to the older age, obesity, family history of diabetes, physical inactive, impaired glucose tolerance, prior history of gestational diabetes etc. This gestation diabetes can developed 2-4 percent during the pregnancy and disappear after the pregnancy is over. Obesity and excess of weight are also associated with the higher risk of risk of type 2 diabetes. The type 2 diabetes can be cured due to the control diet, exercise, home blood sugar test etc.

C. NEED FOR EARLY DETECTION OF DISEASE:-

- Diabetes mellitus is the growing epidemic disease.
- According to the US survey approximately 5 million from the 18 million people have the diabetes but they are not aware about it so early identification is very essential.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

- Due to the presence of the diabetes some co-morbid diseases can also affect the human body to save the human life from such life threatening diereses early detection is needed.
- To well known about the factors of diabetes in the particular patients it helps to medicate it and it affect on its fast recovery.
- Economically it is influential to detect diabetes earlier.

D. SYMPTOMS OF DIABETES

- Eyes -Blurred vision
- Central –Polydipsia, Polyphagia, Lethargy
- Systemic- Weight Loss
- Breath- Smell of acetone
- Gastric-Nausea, Vomiting, Abdominal pain

D. PROBLEM DEFINITION OF THE PROPOSED SYSTEM

Traditional approaches have not adequately considered all the disease causing factors and the all attributes of the diabetes and the relevant co-morbid diseases which is caused by the presence of the diabetes. So, the idea is to develop an expert system which enhances the result of the traditional system by a TOPK algorithm and the MOPNAR algorithm. In the traditional approaches they used the rule set summarization techniques which build the lots of rules and from these all rules the detection of diseases is done. But the rules all very large and all of them are not used for the disease detection so there is need to summarizes these rules. This is done by the proposed system and another aim is to increase the accuracy of the system, which is fulfilled by the proposed techniques such as TOPK and the MOPNAR.

E. OBJECTIVES AND SCOPE OF THE PROPOSED SYSTEM

This work aims at early identification of the diabetes and its co-morbid diseases from the patients and introducing the ubiquitous approach to recognize the risk factor of the diabetic patients by using phenomenological Data Mining technique which utilizes TOPK algorithm for building the rules to detect the diabetes and the MOPNAR algorithm for the accurate detection of which patients having the diabetic or which is not. The data mining algorithm TOPK allows us to build the rules based on the all attributes present in the patient's database and the MOPNAR allows to build rule not only on the positive but also in negative approach also. It gives the more effective result of the proposed system. Due to this the recognition of disease will accurate and fast done as compare to the traditional approach.

The developed system will be used in:

1. The system will used in the dialectological hospital.
2. It also used in the pathology lab and in clinical research center.

II. LITERATURE SURVEY

According to Gyorgy J. Simom, Pedro J. Caraballo, Terry M. Therneau, Steven S. Cha, M. Regina Castro, and Peter W. Li (2015) in [1], the term Diabetes is now fastest growing Diseases in today's world it is risky for the human life because it life threading Diseases. So it is today's need to early diagnosis of such diseases according to the author they used four summarization techniques to make the rules for the disease identification and then after that apply the two algorithm from which Bus retained slightly more redundant than the TOPK, But as the result of this paper the second algorithm i.e. TOPK performing slightly better than the Bus.
According to the Hye Soon Kim, A Mi Shin, Mi Kyung Kim, and Yoon Nyun Kim (2011) in [13], Detection of co-morbid diseases relevant with the diabetes is necessary, because if it identify early then it helps in the medication of that diseases. By the investigation it finds the risk factor of the diseases and according to it hypertension plays an

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

important role in between type 2 diabetes and its co-morbid diseases. For the detection of diseases it takes data from medical center in between the period of 1996 to 2007.

Kavitha R. Mohan (2014) in [2], proposed a association rule mining with summarization techniques it apply on the electronic medical records here the author used four association rule mining techniques to identify the set of risk factor on the patients who are at increasing level of diabetes. It build the lot many rules to reduce it used the association rule mining techniques at the end, and from that the large rule set is summarized into the 10-20 rule set which is easy to investigate. It also show the comparison of two algorithm which are used for the diabetic detection from that it is clear which one is better than the other for that they use representative rules.

According to the investigation of Priya R.and Roshma R. (2015) in [5], here in this paper HARS (Hybrid Association Rule Summarization) technique is used for the detection of diabetes and its co-morbid diseases. In this hybrid technique it covers the distributed association mining and summarization techniques, this technique reduce the rules by using generic and SAM (Split and Merge) algorithm. The ARM creates the bulky set of rules which are need to summarized, for the diabetes detection. It also shows the comparative study in between the previous algorithm and HARS algorithm according to this study HARS is slightly more redundant than the top-k and bus. From this it is clear that it is better than the existing system it prediction rate is better than the previous system.

According to N. Mlambo (2016) in [6], this is the survey of data mining techniques, challenges and gives some approaches for improvement, in healthcare, banking, and finance and telecommunication sector. There are many data mining techniques are used to developed the data mining projects such as classification, clustering, association, prediction and sequential pattern. Some key challenges such as private and sensitive data, distributed data and operation, data quality etc. From that survey paper it helps to select the better technique for development of data mining project.

Swati Gupta, A. M. Karandikar (2015) in [7], proposed a automatic detection system for the detection retinopathy. Diabetic retinopathy is the diabetic eye disease and a leading cause of blindness. In this paper the detection of retinopathy is done by using the retinal image and this automatic system reduce the examination time and increase the accuracy rate of detection. This is the review paper which classifies and compares the previously proposed algorithm and techniques to find which one is better.

Meera Walvekar, Geeta Salunke(2015) in [11], In this paper author focus on the detection of retinal disease such as retinopathy it is done by using the image processing, which helps to auto detection of disease by feature extraction of the image only. In this paper the two dataset are used which are STARE and DRIVE.

According to the Swati V. Gupta, Madhuri S. Joshi (2016) in [10], In this paper data mining algorithm MOPNAR is used. This algorithm focuses on the positive as well as negative dependencies with the used of this identification of any object according to their database are very ease.

III. PROPOSED METHODOLOGY

Any diseases can be identified earlier then it is very beneficial to the person who is serving from that disease and physician also. Because if the disease can be identifying early then the treatment according to the disease will early started and it increase the chances to cure the patients from this disease early. Suppose the disease like cancer if it identify early then there is more chances of patient to live or to cure from that diseases, but if it is not identify early then very critical condition for that identifying the risk of that diseases is very necessary. To do this previously some techniques are used which are association rule summarization techniques, there are number of association rules summarization techniques but all are not used only some of them are used which gives the result in there applicability and strength [8]. These techniques are,

A. Survival analysis
B. Association rule mining
C. Distributional association rules
D. Distributional association rules for survival outcomes

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

These four techniques are used to build the rules for identifying risk assessment of the particular diseases. But the rules which are building from these techniques are very large amount and it is difficult to identify the diseases from these large set of rules. In each techniques the different formulation and methodologies is used and from this the rule set is build.[1] But the main problem is large set of rules, from these all rules are not in used means some are frequently occur. By applying the BUS and TOPK algorithm on these rules the diseases can be identify. From the comparison of these two algorithm we conclude that TOPK is effective than the BUS. To overcome the problem in the previous work use the algorithm TOPK and MOPNAR the TOPK which is best from the previously and MOPNAR is the multi objective positive and negative association rules algorithm, it develop the rules based on the positive as well as negative attributes which is very efficient for the detection of any diseases.

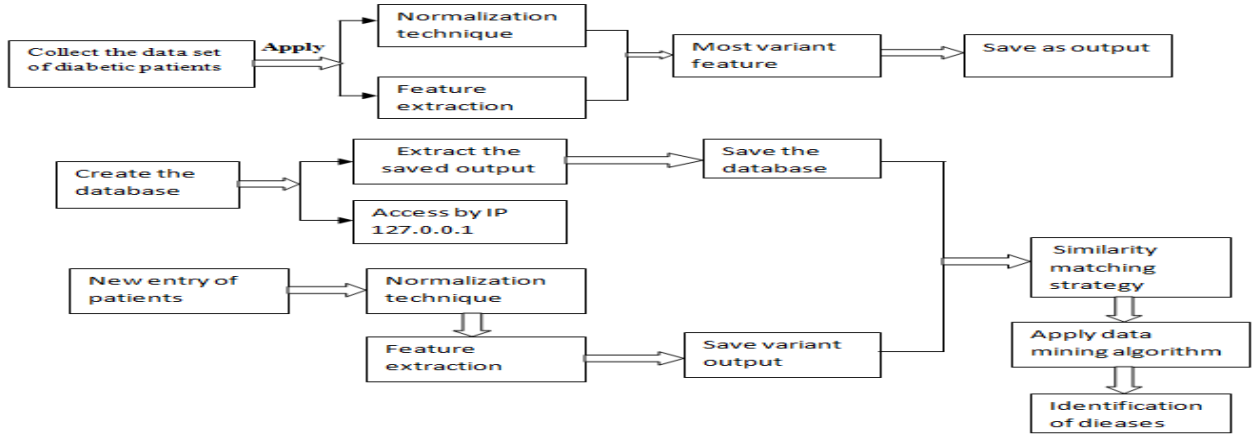

Figure 1:- Flow of proposed system

The proposed system is divided into 6 phases

1) Module 1(dataset collection and feature extraction):- This is the first phase of the proposed system in this phase the data collection is done and it is collected from the UCI repository. In this module first of all perform the normalization by using the min max normalization technique for that first calculate the max value from each column and then divide column entries with that maximum value to get the normalized data which is in between 0 and 1.Then calculate the variance of each column (normalized value) then calculate the mean of variance of each column and if the column variance is greater than the mean then and only then select the column as a most variant column otherwise discard the column. Finally produce the output with all selected column. And collect the dataset of co-morbid diseases i.e. retinopathy and heart diseases

The link from the diabetic patient's dataset is collected.
Diabetes data set: - https://archive.ics.uci.edu/ml/datasets.html
Retinopathy data set: - https://archive.ics.uci.edu/ml/datasets/Diabetic+Retinopathy+Debrecen+Data+Set
Heart disease: - https://archive.ics.uci.edu/ml/datasets/Heart+Disease

AGE	SEX	BMI	BP	S1	S2	S3	S4	S5	S6	Y
59	2	32.1	101	157	93.2	38	4	4.8598	87	151
48	1	21.6	87	183	103.2	70	3	3.8918	69	75
72	2	30.5	93	156	93.6	41	4	4.6728	85	141
24	1	25.3	84	198	131.4	40	5	4.8903	89	206

Fig: - Sample of diabetic patients data set

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

This is the dataset of the diabetic patients which have the attributes such as age of patients, gender of patients, patient's body mass index, patient's blood pressure and the six sugary components (glucose, sucrose, starch, carbohydrates, etc) which are present in the human blood.

2) Module-2(Database creation): In this second Phase the database is created for that the wamp server is used which is best serves for the sql dataset and java development. Create the table in the wamp server as name tbldatabase of attributes id, features and class. All the most variant features are saved in the dataset files extract the class from the dataset file check most variant features are saved already in the database if present then ok otherwise insert the features in the database.

3) Module-3(Access Database): In this third phase, to access the database first open the browser and type the IP address 127.0.0.1 and then click on the phpmyadmin button and then click on the coding folder and finally click on the database table name. After running the code database table is filled with entries. Up to this the database training is over.

4) Module-4(Evaluate the database): In Phase-4, take the new entries of the patients and calculate the normalization and feature extraction then calculate the similarity between the new patients dataset and the previously saved database for that used the JaroWinklerStrategy, which calculate the similarity between the two database based on the features and the class value.

5) Module-5(Identification of Diseases): In this last module apply the data mining algorithm on the final database. First apply TOPK algorithm which build rules for the identification of diseases which include the all the attributes of database and then apply the MOPNAR algorithm which gives exact result which patients are diabetic or which are not. This gives the last and final result of the identification of diabetes.

6) Module -6(Co-morbid Diseases): Apply the same procedure on the co-morbid diseases dataset to identify which patient is suffering from the selected co-morbid diseases.

IV. CONCLUSIONS

This study was conducted to analyze the detection of the diabetes diseases, which is fast growing and need to identify as soon as possible it is the need of today's world. It is notice base on the given literature survey there are many technique to identify that the patients is suffering from the particular disease or not, but these all technique have some laminations which are trying to overcome here. Therefore, the proposed system is defined to increase the accuracy to detection of the particular diabetes disease and its co-morbid disease by using data mining techniques such as TOPK and MOPNAR algorithm. Utilizing the concept of co-morbidity can help to analyzing the co-morbid disease with the diabetes. This appraisal system could be very versatile if it needs to alleviate any changes then it can be able to make in this system. This system comes to fulfill all the goals; this expert system must help physician for detection of disease and to providing to better medication and guideline according to disease. Early identification and then proper medication after detection of disease these all problems are going to solve by the use of this proposed system.

REFERENCES

[1] Gyorgy J. Simom, Pedro J. Caraballo, Terry M. Therneau, Steven S. Cha, M. Regina Castro, and Peter W. Li,"Extending Association Rule Summarization Techniques to Access Risk of Diabetes Mellitus" IEEE Transaction on knowledge and data engineering, Vol. 27, No. 1, pp. 130-141,January 2015.
[2]V. Kavitha, R. Mohan, "ARS: Association Rule Summarization Techniques to Detect Risk Of Diabetes Mellitus," International Journal Of Research In Computer Application And Robotics, Vol. 2, Issue 11, pp. 145-149, November 2014.
[3]https://www.diabetesreasearch.org
[4]https://www.cdc.gov/media/presskits/aahd/diabetes.pdf

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

[5]Priya R., Roshma R.,"Pridiction Of Co-morbid Condition Associated With Diabetes using Split and Merge Algorithm" International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 7, pp. 6716-6722, July 2015.
[6] N. Mlambo, "Data Mining: Techniques, Key Challenges and Approaches for Improvement," International Journal of Advance Research in Computer Science and Software Engineering,Vol. 6 Issue 3, pp. 59-65, March 2016.
[7]Swati Gupta, A.M.Karandikar," A Survey on Method of Automatic Detection of Diabetic Retinopathy", International Journal of Research in IT, Management and Engineering, Vol. 5, Issue 1,pp. 1-7, January 2015.
[8]M. A. Husan,"Summarization in Pattern Mining", 2009.
[9]G. Rahini, P.Dinash Kumar," Electronic Medical Record for Diabetic Mellitus Using Association Rule Mining," International Journal of Advance Research in Biology Engineering Science and Technology, Vol. 2, Issue 15,pp. 247-252, March 2016.
[10]Swati V. Gupta, Madhuri S. Joshi,"Multi-Objective Sentiment Analysis Using Evolutionary Algorithm for Mining Positive &Negative Association Rules," International Journal of Computer Science and information Technologies, Vol. 7, Issue 3,pp. 1362-1368, 2016.
[11] Meera Walvekar, Geeta Salunke,"Detection of Diabetic Retinopathy with Feature Extraction using Image Processing," International Journal of Emerging Technology and Advanced Engineering", Vol. 5, Issue 1, pp. 133-137, January 2015.
[12] Madhavi predhan, Ketki Kohale, Parag Naikade, Ajinkya Pachore Eknath Palwe, "Design of Classifier for Detection of Diabetes using Neural Network and Fuzzy k-Neighbour Algorithm" International Journal of Computational Engineering Research, Vol. 2, Issue 5, pp. 1384-1387, September 2012.
[13] Hye Soon Kim, A Mi Shin, Mi Kyung Kim, and Yoon Nyun Kim,"Co-morbidity Study on Type 2 Diabetes Mellitus Using Data Mining", The Koeran Journal of Internal Medicine , Vol. 27 No.2, pp. 197-202, June 2012.