ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 3, Issue 3 , March 2016

# Efficient Sequential Pattern Mining Algorithm To Detect Type-2 Diabetes

**Dr.P.Nithya.,B.Uma Maheswari, R.Deepa**

Asistant Professor, Department of Computer Science, PSG College of Arts & Science, Coimbatore, India
Asistant Professor, Department of Computer Science, PSG College of Arts & Science, Coimbatore, India
M.Phil Scholar, Department of Computer Science, PSG College of Arts & Science, Coimbatore, India

**ABSTRACT***:* The specific causes of complex diseases such as Type-2 Diabetes Mellitus (T2DM) have not yet been identified. Nevertheless, many medical science researchers believe that complex diseases are caused by a combination of genetic, environmental, and lifestyle factors. Detection of such diseases becomes an issue because it is not free from false presumptions and is accompanied by unpredictable effects. To solve this problem an existing system multiple classifier approach base type-2 diabetes mellitus detection. In this system dynamic weighted voting scheme called multiple factors weighted combination for classifiers' decision combination. However it does integrate the genetic information and cannot discover complex disease more accurately. To solve this problem a sequential pattern mining approach which is called P-Prefix Span (Percussive-Prefix sequential pattern approach).The main objective of the sequential pattern mining algorithm is used for mining the order of sequence from the specified medical dataset. Based on the gene sequence structure the sequence pattern algorithm discovers the set of frequent subsequences in the dataset. Based on the minimum support count value the frequent patterns are mined and produce interesting patterns which satisfy the conditions.

**KEYWORDS***:* sequential pattern mining, MFWC.

## I. INTRODUCTION

Diabetes mellitus (DM), also called as diabetes which is a set of metabolic diseases, in which there are huge blood sugar levels over a long period. This huge blood sugar makes the symptoms of frequent urination, increased thirst, and increased hunger. Untreated, diabetes can produce a lot of complications. Acute complications include diabetic ketoacidosis and nonketotic hyperosmolar coma. Serious long-term complications include heart disease, stroke, kidney failure, foot ulcers and damage to the eyes.

Diabetes is due to either the pancreas not generating sufficient insulin, or the cells of the body not reacting properly to the insulin produced.

## II. LITERATURE SURVEY

Diabetes is a group of metabolic diseases caused by hyperglycemia this is because of defects in insulin secretion, insulin action and both. Next stage chronic hyperglycemia of diabetes is associated with long term damage, dysfunction, and failure of different organs of body, especially the eyes, kidneys, nerves, heart, and blood vessels. This deficiency leads to destruction of the b-cells of the pancreas with consequent insulin deficiency to abnormalities that result in resistance to insulin action and reaction process. The basis of the abnormalities found in carbohydrate, fat, and protein metabolism in diabetes is deficient action of insulin on target tissues. We propose a method to detect DM initial stage based on three groups of features extracted from tongue images. They include color, texture, and geometry tongue color features with a log gabor filter mechanism. Concerning biological vision criteria, the log-Gabor filters mimic closely analysis the tongue texture features. Relational statistics of natural images have similar shape of trained images which supports the proposed log Gabor filter as an adequate scheme for matching biomedical features.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 3, Issue 3 , March 2016

## III. PROBLEM FORMULATION

The specific causes of complex diseases such as Type-2 Diabetes Mellitus (T2DM) have not yet been identified; nevertheless, many medical science researchers suppose that complex diseases are by a combination of genetic, environmental, and lifestyle factors. Early detection of such diseases can prevent and treat complex diseases when they do not have obvious clinical symptoms. Considering the greatly increased amount of data gathered in medical databases and the availability of historical data on complex diseases, such as patients' blood glucose, traditional manual analysis has become inadequate and naturally leads to the application of data mining techniques to discover interesting patterns so that early detection and successful recommendation for diagnosis becomes possible. Efficient detection of complex diseases such as Type-2 Diabetes Mellitus is important factor.

## IV.PROBLEM SPECIFICATION

### A. EXISTING SYSTEM

In the existing scenario, implement a method named as multiple factors weighted combination (MFWC). This method is used to discover the Type-2 diabetes mellitus (T2DM) with the help of multiple classifier system (MCS). MCS is a set of individual classifiers whose decisions are combined according to certain rules to produce the final output. MCS has many advantages, and studies show that the combination of homogeneous classifiers using heterogeneous features can improve the final result. The dynamic weighting is a better approach which allocates the weights to the output of each individual classifiers and it can change for each input vector in the testing phase.

### B. PROPOSED SYSTEM

In the proposed scenario, we implement a new approach named as sequential pattern mining algorithm for detecting the complex disease more effectively. The main purpose of the sequential pattern mining algorithm is used for mining the order of sequence from the specified medical dataset. Based on the gene sequence structure the sequence pattern algorithm discovers the set of frequent subsequences in the dataset. We have to determine the support threshold value for the given sequences and the algorithm selects the sequence which satisfies the specified threshold value.

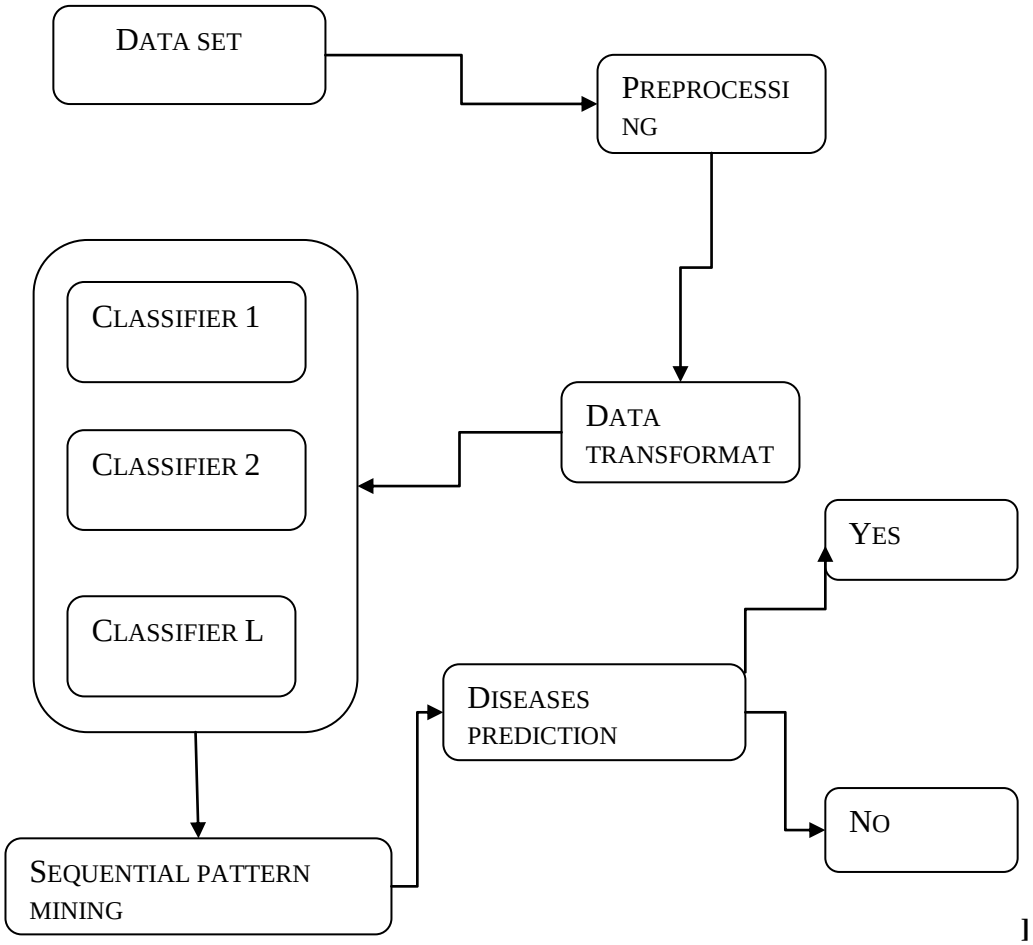
We consider in this scenario the algorithm called as P-PrefixSpan (Precursive-Prefix sequential pattern) algorithm which is based on apriori algorithm. It has to discover the length of sequential pattern and count the support for all gene sequences. It is used to reduce the searching complexity. Based on the minimum support count value the frequent patterns are mined and generate interesting patterns which satisfy the conditions. Hence this algorithm is used to detect the complex disease more accurately.

**Advantages**

1. It is used to integrate the genetic information and handles huge dimensional dataset
2. The GSP P-Prefix Span algorithm is efficiency in finding the complex disease
3. The error rate is reduced in this scenario
4. The accuracy and performance of proposed scenario is improved prominently
5. The time & space consumption of proposed algorithm will be lesser,

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 3, Issue 3 , March 2016

**V ARCHITECTURE DESIGN**




]

## VI.RESULTS AND DISCUSSION

### A. Accuracy

The Accuracy of the system is calculated with the values of the True Negative, True Positive, False Positive, False negative actual class and predicted class outcome it is defined as follows,

$$\text{Accuracy} = \frac{True\ positive + True\ negative}{True\ positive + True\ negative + False\ positve + False\ negative}$$

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 3, Issue 3 , March 2016

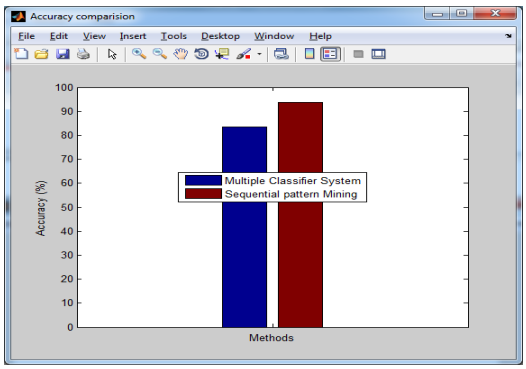



Figure 1. Accuracy comparison

In this graph, x axis will be the two approaches of Type 2 diabetes detection and y axis will be accuracy in %. From the graph see that, accuracy of the proposed system detecting diabetics using sequential pattern mining is better than existing one. From this graph, we can say that the accuracy of proposed system approach is increased, which will be the best one.

B. **Precision**

Precision value is determined based on the retrieval of information at true positive prediction, false positive. In healthcare data precision is determined the percentage of positive outcome returned that are relevant.

Precision =TP/ (TP+FP)

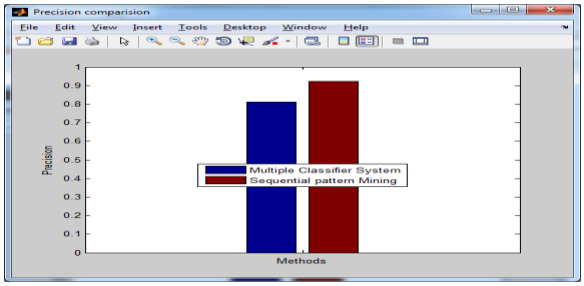



Fig.2. Precision comparison

Compare the methods of detecting type-2 diabetes mellitus using multiple classifier system and detecting type-2 diabetes mellitus using sequential pattern mining. In this graph, x axis will be the two approaches of type-2 diabetics detection and y axis will be precision.The proposed has high precision compare to another one.

**C. Recall**

Recall value is determined based on the retrieval of information at true positive prediction, false negative. Recall in this context is also referred to as the True Positive Rate. In that process the fraction of relevant instances that are retrieved.

Recall =TP / (TP+FN)

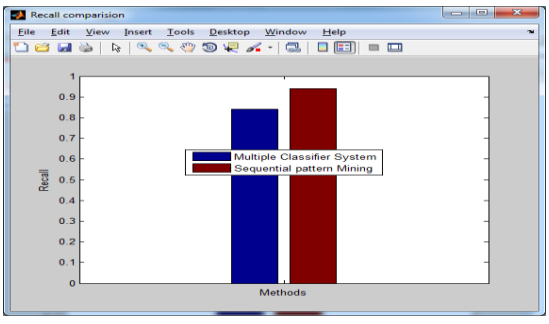



**Fig.3. Recall comparison**

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 3, Issue 3 , March 2016

Compare the methods of detecting type-2 diabetes mellitus using multiple classifier system and detecting type-2 diabetes mellitus using sequential pattern mining. In this graph, x axis will be the two approaches of type-2 diabetics detection and y axis will be recall. From this graph, we can say that the recall of type-2 diabetes mellitus detection is increased, which will be the best one.

**D. F-Measure**

The F-measure of the system is defined as the weighted harmonic mean of its precision and recall, that is, F=1α1P+(1−α)1R, where the weight α ∈ [0,1]. The balanced F-measure, commonly denoted as $F_1$ or just F, equally weighs precision and recall, which means α = 1⁄2

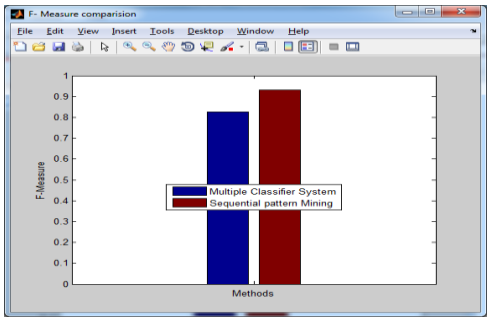



**Fig.4.F measure comparison**

Compare the methods of detecting type-2 diabetes mellitus using multiple classifier system and detecting type-2 diabetes mellitus using sequential pattern mining. In this graph, x axis will be the two approaches of type-2 diabetics detection and y axis will be F-measure .From this graph, we can say that the F measure of diabetics detection is increased, which will be the best one.

## VII. CONCLUSION

The proposed system introduced a sequential pattern mining approach for T2DM by using a P-Prefix Span (Precursive-Prefix sequential pattern approach). Generalized sequential pattern mining is able to substitute all types of conventional sequential pattern mining algorithms with intervals. The proposed method finds out the length of sequential pattern and count the support for all gene sequences. These are used to minimize the searching complexity. By utilizing minimum support count value the frequent sequence are mined. A discovering interesting patterns which satisfy the conditions. We evaluated our method on two T2DM data sets and other complex diseases data from real world with comparisons to multiple classifiers and state-of-the-art fusion methods. The experiments indicated that our proposed method outperforms other methods in terms of accuracy, precision, recall and F-measure.

## REFERENCES

1]. Ramesh Kumar B , Sivapriya V, " Diabetes Mellitus Discovery based on Tongue Texture Features using Log Gabor Filter Mechanism" , International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 3, Issue 9, September 2015

[2] S. Cessie, J.C. Houwelingen, Ridge estimators in logistic regression, Appl. Stat. (1992) 191–201.

[3] P.K. Chan, D.S. Yeung, W.W.Y. Ng, C.M. Lin, N.K. Liu, Dynamic fusion method using localized generalization error model, Inform. Sci. (2012) 1–20.

[4] V. Cheplygina, D.M.J. Tax, M. Loog, Combining instance information to classify bags, in: Multiple Classifier Systems, 2013, pp. 13–24.

[5] C. Cortes, M. Mohri, A. Rastogi, An alternative ranking problem for search engines, Proc. WEA07 (2007) 1–21.

[6] B. Dasarathy, Nearest Neighbor Norms: NN Pattern Classification Techniques, IEEE Computer Society Press, 1991.

[7] G. Dietterich, Machine learning research: four current directions, AI Mag. (1997) 97–136.

[8] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J. Comput. Syst. Sci. (1997) 119–139.

[9] D. Hidalgo, P. Melin, O. Castillo, An optimization method for designing type-2 fuzzy inference systems based on the footprint of uncertainty using genetic algorithms, Expert Syst. Appl. 39 (4) (2012) 4590–4598.

[10] B. Homme, R. KK, J. Valdes, Dynamic pharmacogenetic models in anticoagulation therapy, Clin. Lab. Med. (2008) 539–552.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 3, Issue 3 , March 2016

[11] J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, Biophysics (1982) 2554–2558.
[12] Y. Huang, P. McCullagh, N. Black, R. Harper, Feature selection and classification model construction on type 2 diabetic patient's data, Adv. Data Min. (2005) 153–162.
[13] D.J. Hunter, Gene-environment interactions in human diseases, Nat. Rev. Genet. (2005) 287–298.
[14] G.H. John, P. Langley, Estimating continuous distributions in Bayesian classifiers, in: Eleventh Conference on Uncertainty in Artificial Intelligence, 1995, pp. 338–345.