www.ijiarec.com

ISSN:2348-2079

Volume-5 Issue-1

International Journal of Intellectual Advancements and Research in Engineering Computations

Efficient sequential pattern mining algorithm and improved MFWC scheme for discovering complex disease and type-2 diabetes mellitus

S.Ranjithkumar [1] P.Haritha, S.jayaselvamathi, J.Kiruthika and S.Logeshwari [2,3,4,5]

[1]Assistant Professor, Department of CSE, SNSCE, Coimbatore -107, Tamil Nadu, India.

[2,3,4,5] UG Scholar, Department of CSE, SNSCE, Coimbatore -107, Tamil Nadu, India.

[1]sranjith54@gmail.com [2]harithapremkumar23@gmail.com [3]madhumadhie02@gmail.com

[4]kiruthigajayagopal16@gmail.com [5] logeshwari5333@gmail.com

ABSTRACT

Diabetes mellitus is one among the complex diseases for which specific causes have not yet been identified. Nevertheless, many medical science researchers believe that complex diseases are caused by environmental, genetic and abnormal cholesterol and triglyceride levels. 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 introduced multiple classifier approach base type-2 diabetes mellitus detection. In this system, we introduced a voting scheme which is dynamic 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 the proposed system is introduced a sequential pattern mining approach which is called Frequent Pattern growth approach. The main objective of the sequential pattern algorithm is to check and mine data sets based on the sequential order. Based on the gene sequence structure the sequence pattern algorithm discovers the set of frequent sub sequences in the dataset. The minimum support count value is identified to produce interesting patterns which satisfy the conditions. Hence this algorithm is used to detect the complex disease more accurately. The experimental results show that the proposed system achieves high performance compared with the existing system in terms of accuracy, precision, recall and f-measure.

INTRODUCTION

Data mining is the process of collecting, searching through, and analyzing a large amount of data in a database, as to discover patterns or relationships. Generally, data mining is the search for hidden patterns that could be present in huge databases. Data mining discovers patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. Data mining is becoming a gradually more important tool to make over this data into information. Data mining requires the use of data analysis tool to determine previously unknown, valid patterns and relationships in huge volume data.

Diabetes mellitus (DM), also called as diabetes is a disease in which the body's ability to respond to the hormone insulin is reduced readily, which results in abnormal metabolism of carbohydrates and elevated levels of glucose in the blood.

Type 1 DM results from the body's failure to generate sufficient insulin. This form was earlier referred to as childhood-onset diabetes, and insulin-dependent diabetes mellitus (IDDM). The reason is unknown.

SIGNS AND SYMPTOMS

This huge blood sugar makes the symptoms of frequent urination(Polyurea), increased thirst(Polydipsia), and increased hunger(Polyphagia). Untreated, diabetes can produce a lot of complications. Acute complications Hyperglycemia hyperosmolar state, Diabetic coma, Respiratory infections

Serious long-term complications include heart disease, stroke, kidney failure, foot ulcers and damage to the eyes.

2. Diabetic emergencies

Diabetics may experience life-threatening **emergencies** from too much or too little insulin in their bodies. Too much insulin can cause a low sugar level (hypoglycaemia), which can lead to insulin shock.

3. Complications

Diabetes is a disease which has the risk of long-term complications. The major long-term complications relate to damage to blood vessels. Diabetes doubles the risk of cardiovascular disease and in 2010; diabetes was mentioned as a cause of death in a total of 234,051 certificates. Other "macro vascular" diseases are the stroke and peripheral vascular disease.

The primary micro vascular complications of diabetes include Dyslipidemia, Blindness and Eye Problems, hypertension diabetic foot ulcers, proximal diabetic neuropathy muscle wasting and failing.

LITERATURE SURVEY

1.Bayu Adhi Tama , Rodiyatul F. S. , Hermansyah Faculty of Computer Science, University of Sriwijaya Faculty of Medicine, University of Sriwijaya. **“An Early Detection Method of Type-2 Diabetes Mellitus in Public Hospital”**

2. Ramesh Kumar B, Sivapriya V, “Diabetes **Mellitus Discovery based on Tongue Texture Features using Log Gabor Filter Mechanism**”

PROBLEM SPECIFICATION:

In the existing scenario, we introduced 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 a special classifier system called multiple classifier systems (MCS). Multiple classifier combination methods can be considered some of the most robust and accurate learning approaches. In MCS a set of individual classifiers are combined and the final report is published. MCS has many advantages, and studies show that the combination of homogeneous classifiers using heterogeneous features can improve the final result. The existing system of 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.

Disadvantage:

The existing method not only considers the local accuracy factor for each classifier and uses a validation set to estimate the classification accuracy at the global level but also concerns the relationship between testing and training samples with generalization error because the generalization error of a classifier is a key function to measure the performance of a classifier generalized to unseen samples. We take physiological data, e.g., blood pressure, as input data and convert them to input vectors through a data transformation process. All values are normalized between 0 to 1. The input vectors will then be used by a classifier for training to generate a classification model than can identify whether a person is a T2DM patient or not. It is used to overcome the issues of static weighting approaches. In this scenario, for specified patient dataset, the model or label is created based on the training classifier. Then the classification model is used to predict the input data is belongs to complex diseases or not.

METHODOLOGY

4.1 Proposed system

In the proposed scenario, we introduced a new approach named as sequential pattern mining algorithm for detecting the complex disease more effectively. The main purpose of this (pattern mining) algorithm is used for mining the order of the sequence from any medical dataset. Based on the gene sequence structure this algorithm used discovers the set of frequent subsequences in the dataset. We have to determine the STV (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 FP-growth (Frequent Pattern growth) algorithm which uses a special internal structure called as FP-tree (Frequent Pattern tree. The length of sequential patterns and the count of genes is identified and calculated. It is used to reduce the searching complexity and is most memory efficient. The frequent data patterns are mined based on the minimum support count value. Hence this algorithm is used to detect the complex disease more accurately.

Advantages

It is used to integrate the genetic information and handles huge dimensional dataset

The FP-growth algorithm is efficiency in finding the complex disease

The error rate is reduced in this scenario

The accuracy and performance of proposed scenario is improved prominently

The time & space consumption of proposed algorithm will be lesser

4.2 proposed algorithm

Output: The complete set of sequential patterns

Step 1: Start Step

Step 2: Construct FP-tree

Step 2a: Scan database D once. Collect F, the set of frequent items (genes), and their support counts. Sort F in support count descending order as L, the list of frequent items (genes)

Step 2b: Create the root of an FP-tree, and label it as "null." For each data in D do the following. Select and sort the frequent items in D according to the order of L. Let the sorted frequent item list in Trans be [p|P], where p is the first element and P is the remaining list. Call insert tree ([p|P], T), which is performed as follows. If T has a child N such that N.item-name=p.item-name, then increment N's count by 1; else create a new node N, and let its count be 1, its parent link be linked to T, and its node-link to the nodes with the same item-name via the node-link structure. If P is nonempty, call insert tree (P, N) recursively.

Step 3: The FP-tree is mined by calling FP growth (FP tree, null), which is implemented as follows. Procedure FP growth (Tree, α)

Step 3a: if Tree contains a single path P then

Step 3b: for each combination (denoted as β) of the nodes in the path P

Step 3c: generate pattern β∪α with support count = minimum support count of nodes in β;

Step 3d: else for each ai in the header of Tree {

Step 3e: generate pattern β = ai ∪α with support count = ai.support count;

Step 3f: construct β's conditional pattern base and then β's conditional FP tree Treeβ;

Step 3g: if Treeβ 6= / 0 then

Step 3h: call FP growth (Treeβ, β) ;}

Step6: End

IMPLEMENTATION

5.1 Module identification

1. Input data

We take input data as physiological data, e.g., blood pressure and convert them to input vectors through a data transformation process. And then noise can be removed by using pre-processing process.

2. Ensemble of classifiers

Assuming there are symbolic L base classifiers and a variable X is the sample to be classified, we calculate the final decision in the oracle for X, as denoted in Equation.

Z= _(c_(j∈c))^argmax ∑_(l=1)^L▒w_(c_j)^l (X) f_(c_j)^l (x) ....(1)

Where Z is the predicted class for X, C is the set of all possible classes, w_cj^l(X) is the weight assigned to the lth classifier based on the oracle calculation and f_ (c_j) ^l (x) is the decision value of the lth classifier for X. If the lth classifier predicts that X belongs to a class c _j, we give X a value 1; otherwise, the value will be set to 0. From Eq. (1), we know that the function w_cj^l(X) is the critical part for the entire system.

3. Multiple factors weighted combination

Our proposed dynamic weighted voting scheme is called MFWC, in which means there are multiple factors are used together to calculate the weight that will be used later for classification.

1. Local k-NN accuracy

From training data, the nearest K samples were founded corresponding to the sample to be classified using the Euclidean distance to calculate the distance between two samples. Assuming that two samples exist, a variable X is the sample to be classified, and a variable Y is one of the training samples. Each sample has n attributes, represented as $(\mathrm{x}_1, \mathrm{x}_2, \ldots . \mathrm{x}_\mathrm{n})$and $(y_1, y_2 \ldots . y_\mathrm{n}$, then the distance can be calculated as:

$$D(X, Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

We then calculate the LKA by averaging classification accuracy as follows

$$LKA= \frac{\sum_{i=1}^{K} f_i}{K}$$

Where f (i) is the classification result from the classification model generated from the training data for the ith sample. The value is 1 if the classification is right; otherwise, the value is set to 0.

2. Global k-NN accuracy

As like LKA, global k-NN accuracy (GKA) is calculated using mean classification accuracy, but the K samples are selected from a validation data set using the k-NN rule. Using a validation data set has been proven to be effective on estimating a model's classification accuracy. By this means, we can determine the real performance of the classifier in a global perspective model generated from the training data for the ith sample. The value is 1 if the classification is right; otherwise, the value is set to 0.

3. Diversity k-NN accuracy

The degree of difference among classifiers is measured by diversity factor. We should assign more weight to a base classifier if it has higher accuracy in diversity. The uncertain samples were founded from training samples and validation data set that only a certain percentage of base classifiers can correctly classify. Given that we will use only five classifiers in our multiple classifier systems, we set the percentage from 20% to 80%. We then calculate the diversity k-NN accuracy (DKA) similar to LKA and GKA, but the K samples are selected from the uncertain samples. The upper bound (UB) and lower bound (LB) of classification accuracy are computed, as shown in Eqs. (4) and (5):

$UB = \mu\times (1-d) +d$

$LB = \mu\times d\ (1-d)/N$

Where l is the mean individual classification accuracy, a variable d is the proportion of selected uncertain samples and N is the number of the classes. If a base classifier's DKA is not in the range of UB and LB, then the DKA of this classifier will be ignored in the voting to improve the ensemble.

4. Localized generalization error bound

The function of generalization error of a classifier which measures the performance of a classifier generalized to unseen samples. We should assign more weight to the classifier if the classifier has smaller generalization error bound. We consider training error and also the sensitivity of each classifier; we applied the L-GEM method [41] to calculate the localized generalization error bound (LGEB).

First, we need to find the greatest distance DMAX between the sample X to be classified and its K neighbourhoods ($Y_1, Y_2, \dots\dots\dots\dots . Y_N$) from training samples according to Eq. (2) :

D^{Max}= max (d(X,Y_i))

We then calculate the LGEB:

$$\text{LGEB} = \sqrt{\frac{1}{k}} \sum_i^k err(f_i y_i) + \sqrt{(D^{\max)\ 2}} + \sum_i^k (\frac{\partial f}{\partial y_i})^{\ T} + (\frac{\partial f}{\partial y_{\ i}})$$

Where

Err (f,y_i) = f (y_i)-F (y_i)

And

$$(\frac{\partial f}{\partial y_{\ i}}) = [\ \frac{df}{dy_{11}},\ \frac{df}{dy_{12}} \dots . \frac{df}{dy_{1n}}\ ]^{\ T}$$

Here, f (y_i) is the function to calculate the confidence for each decision to Yi between 0 to 1; and ($\frac{\partial f}{\partial y_{\ i}}$) is the final decision for Yi which is 0 or 1. and ($y_{i1}, y_{i2} \dots . y_{in}$) are the features of Y $_i$.

In our problem, the function of LR can be defined as follow:

$f^{\ LR}(y_i\ ) = \beta . \sum_{m=1}^n Y_{im}$

Where b is the set of regression coefficients $\beta_0, \beta_1, , \beta_{n.}$

Similarly, for NB, the core function is defined as:

$F^{\ NB}(y_i\ )$ =argmaxp (C_j)$\prod_{m=1}^n p(y_{im}|c_j)$

Where p () is the probability model of NB, and the sensitivity term can be defined as:

$f^{\ NB}(y_i\ )$

$\frac{\partial f^{\ NB} Y_i}{\partial\ y_{it}}$ =argmaxp (C_j)$\prod_{m=1}^t p(y_{im}|c_j)$

5. Weighted voting combination

The weighted voting system is our combination method. This method employs the idea that not all voters are equal, and this approach is a good way to combine decisions. Some voters might carry more weight than others. Assume a voter V _1 has ʎ _1 votes, voter V _2 has ʎ 2 votes, . . . , and voter V _n has ʎ _n votes, and a quota q for voting should be passed, then we have{q:: ʎ_1,ʎ_2,,,ʎ_n}. We first need to rank all four factors among all base classifiers and assign a score to each factor. We then calculate the weight w_ij^l (X) for each classifier l,

w_ij^l (X)= (∑_(l=1)^L▒ 〖l×s_l 〗)/q

Where Sl is the number of times the classifier gets an l score. Therefore, for a final decision, each factor is combined and contributes to the weight assigned to each base classifier. Regarding the value of q, considering that we apply a weighted voting system to the MCS and each base classifier of the system has four factors with an associated score to contribute, the maximum value a classifier can achieve is 4L.

We can be further converted to:

$$w_{ij}^{l}(X) = \frac{\sum_{l=1}^{L} l \times s_l}{1 + 4\sum_{l=2}^{L} l}$$

6. Time complexity

We need to consider for dynamic fusion of the time complexity methods because it is critical. Similar to other methods, MFWC has two stages, i.e., training and testing, which we are going to analyze the two stages. Let M denote the number of training samples, n is the number of attributes for each training sample and L is the number of classifiers. For a typical fusion method in the training stage, majority voting [22], the time complexity of an MCS is O (Mnl).For MFWC, no additional training time is required for the calculation of LKA, GKA, and DKA, but to calculate the generalization error the same amount of training time is needed depending on classification methods. Therefore, the MFWC's time complexity in the training stage is O(2Mnl). The same as other fusion methods of testing stage, the distance calculation between K training or validation samples and testing samples time need to be counted. Hence, the time complexity is O (Knl). In practice, we can add as many classifiers as we like the system as long as the testing time is not longer than the time of the prescript diagnosis.

7. Prefix Span (Precursive-Prefix sequential pattern) algorithm

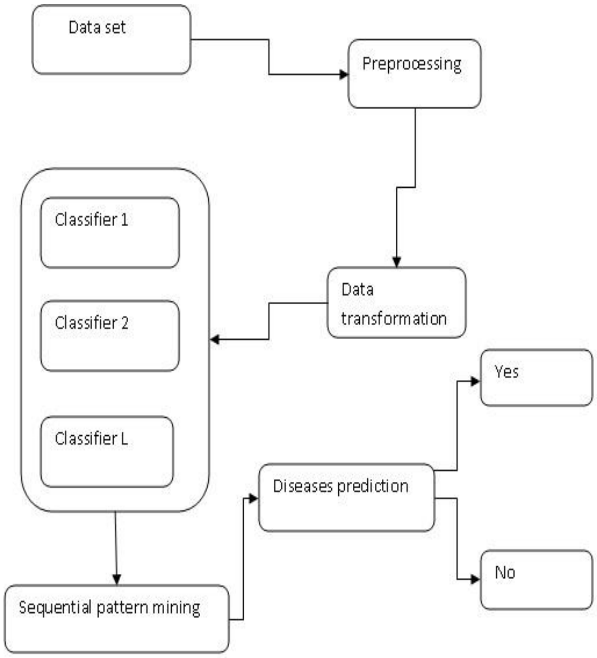
We extend the sequential database projection operation in FP-growth to handle both interval extended sequence and item interval constraints. The FP-growth has been proposed for finding the relevant frequent patterns from sequences. Its major idea is that, instead of projecting sequence databases by considering all the possible occurrences of frequent subsequence's, the frequent prefix of projection is done on which results in processing time of higher efficiency of the algorithm. A method that mines the complete set of frequent item sets without frequent-pattern growth is called candidate generation, or simply FP-growth, which adopts a divide-and-conquer strategy.

Frequent items are represented into a frequent-pattern tree by compressing a database; the item set association information is retained. It then divides the compressed database into a set of conditional databases; each associated with one frequent item or pattern fragment and mines each such database separately. There exist these approaches to counting support value; the genes constraint approach and extended sequence approach.

The genes constraint approach involves extraction of sequences satisfying not only a user-specified minimum support constraint but also user-specified constraints, such as maximum/minimum intervals. The extended sequence approach extends sequences by inserting pseudo genes which represent gene intervals. After extending the original sequences, it extracts frequent sequential patterns from them.

Finally, it is used for complex diseases prediction. The algorithm introduced for mining sequential patterns. The added constraints could filter out less important patterns and reduce the memory space required for storing projected databases.

8. Architecture design



CONCLUSION AND FUTURE WORK:

A. Conclusion

The proposed system introduced a sequential pattern mining approach for T2DM by using an FP-growth algorithm. All types of conventional sequential pattern mining algorithms with intervals are substituted by a generalized sequential pattern. The proposed method finds out the length of sequential pattern and counts the support for all gene sequences. These are used to minimize the searching complexity. The frequent sequence is mined by utilising the minimum support count value. A discovering interesting patterns which satisfy the conditions. Two T2DM data sets and other complex diseases data are evaluated from the 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.

B. Future work

In future various sequential pattern mining approaches are used for detecting the complex diseases.

REFERENCE:

[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. Computer. 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.

[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.

[15] J. Kittler, M. Hatef, R. Duin, J. Matas, On combining classifiers, IEEE Trans. Pattern Anal. Mach. Intell. (1998) 226–239.

[16] N.Shyamambika, N.Thillaiarasu "A Survey on Acquiring Integrity of Shared Data with Effective user Termination in the Cloud"International Conference on Intelligent Systems and Control (ISCO16), DOI: 10.1109/ISCO.2016.7726893, IEEE Explore, 2016.

[17] N.Thillaiarasu, Dr.S.Chenthur Pandian "Enforcing Security and Privacy over Multi– loud Framework Using Assessment Techniques", International Conference on Intelligent Systems and Control (ISCO16), DOI: 10.1109/ISCO.2016.7727001, IEEE Explore, 03 N2016.

[18] N. Shyamambika and N. Thillaiarasu,"Attaining Integrity, Secured Data Sharing and Removal of Misbehaving Client in the Public Cloud using an External Agent and Secure Encryption Technique" Advances in Natural and Applied Sciences. 10(9); Pages: 421-431.