International Journal of Engineering & Technology, 7 (4) (2018) 3103-3108

SPC

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i4.18066
Research paper



Efficient fuzzy frequent pattern tree mining technique to predict chronic kidney diseases

S. Dilli Arasu [1] *, Dr. R. Thirumalaiselvi [2]

[1] *Research Scholar, Department of Computer Applications (Ph.D.),Bharath Institute of Higher Education and Research (BIHER), (Declared as Deemed-to-be-University under Section 3 of UGC Act 1956),# 173 Agharam Road, Selaiyur, Chennai - 600 073, Tamil Nadu, India*
[2] *Research Supervisor / Assistant Professor, Department of Computer Science, Government Arts College for Men - (Autonomous), Nandanam, Chennai - 600 035, Tamil Nadu, India.*
**Corresponding author E-mail: dilliarasu@yahoo.com*

Abstract

Data mining is an important role in huge number of applications, which are manufacturing, aerospace, business organizations, government sectors, and medical industry. In the medical field, the data mining is mostly used for detecting disease and diagnoses. The medical analysis is performed in patients to explore the disease are enormous which results in huge data collection. However, in the large number of records some important data are missing. In a dataset, the presence of missing data is a common problem in statistical analysis and it degrades the performance of the classifier model while using the dataset as a training sample. The Weighted Average Ensemble Learning Imputation (WAELI) was proposed for filling the missing values in the clinical dataset by using single value imputation and multiple value imputation models. Furthermore, the priority-assigning algorithm was assigned along with WAELI to each feature (WAELI-FPA) for selecting the high priority features. In this paper, the performance of WAELI-FPA is further improved by modifying the classifier model such as mining fuzzy frequent pattern tree (MFFPT) and orthogonal partial least square (O-PLS) to achieve the optimal result. Initially, the O-PLS is efficiently performed in the reduction of missing valued after that allocating the priority to each feature in the clinical database. Second, the MFFPT is performed which is initially determined the frequent data and assigning fuzzy membership values then utilizing priority mechanism to predict the kidney disease. The performance of proposed model is tested in terms of accuracy, precision, recall and F-Measure.

Keywords: *Missing Values; Dataset; Statistical Analysis and Imputation.*

1. Introduction

Data mining (Rogeith, V., & Magesh, S. 2017) is a used for the healthcare industry to enable health systems systematically. It uses data for analytics to identify incompetence and best practices that increase the care and reduce costs. Medical treatment is facing a challenge of knowledge discovery from the growing volume of data. Nowadays huge data are collected continuously through health examination and medical treatment.

Chronic kidney disease (CKD) (Couser, W. G., et.al.2011) is a global public health problem, affecting approximately 10% of the population worldwide and is increasing in prevalence. CKD is often associated with an increased risk of hospital admission, morbidity and death due to cardiovascular disease and the progressive loss of kidney function. Patients with CKD have a high potential for developing atherosclerosis and other types of syndromes. These syndromes have significant effects on their quality of life and survival. The objective of this research work is to predict kidney diseases by using efficient techniques of data mining.

In existing research, Weighted Average Ensemble Learning Imputation (WAELI) was proposed to process the pre-processing for filling the missing values in the clinical dataset. The single value imputation model was presented to predict the missing value in small dataset and the multiple value imputation models was presented to predict the missing value in the huge dataset. For single value imputation model, the expectation maximization (EM) and random forest (RF) models were utilized and for multiple value imputation model, the CART and C4.5 were predicted the missing values. Furthermore, the priority assigning algorithm was assigned to each feature, instead of giving all attributes to classifiers the priority assigning algorithm is used to select the high priority features then those features are given to the classifiers such as support vector machine (SVM), artificial neural network (ANN), Partial Least Squares Discriminant Analysis (PLS-DA), mining frequent pattern tree (MFPT) it predict the kidney disease effectively.

In this paper, the performance of WAELI model is further improved by enhancing priority assigning through modifying classifier models which are mining fuzzy frequent pattern tree (MFFPT) and Orthogonal partial least square (O-PLS). The performance of proposed model is tested with accuracy, precision, recall and F-Measure.

2. Literature survey

In this paper analyze the effect of class imbalance (Yildirim, P. 2017) in training data when developing neural network classifier

Copyright © 2018 S. Dilli Arasu, Dr. R. Thirumalaiselvi. 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.

for medical decision making on kidney disease. The neural networks were widely used in various kinds of applications including data mining and decision systems. Back propagation neural networks were popular neural network methods which can be trained to recognise various patterns. The importance of these networks was employed based on multilayer perceptron with various learning rate values for prediction of chronic kidney disease. This research efficiently used in medical data mining and reveals that sampling algorithm improves the performance of multilayer perceptron with optimum learning rate parameter for learning process.

In this paper presented two classifiers (Chen, Z., et.al.2016) namely fuzzy rule building expert system (FuRES) and fuzzy optimal associative memory (FOAM) for diagnosing of patients with chronic kidney disease (CKD). The aforementioned both classifiers are provided better robustness and high yielded prediction rates while FuRES had better robustness than FOAM especially when the training and prediction sets each contained similar values of noise. The FuRES classifier was more robust than FOAM with a better tolerance to deviations in measured data. The use of fused data from original and composite datasets for building classification models would be a promising approach for the diagnoses of other diseases as well as CKD diagnostic classifications.

This research work was proposed an association rule (Ilayaraja, M., & Meyyappan, T. 2013) based apriori data mining technique that finds the frequency of diseases affecting patients. The research was made on patients from various geographical locations and at various time periods. The experimental result shows the proposed research was achieve optimal accuracy to predict disease in patients.

In this paper, the decision supporting method was proposed (Jung, H., et.al.2015) for chronic disease patients based on mining frequent pattern tree. The proposed method was measured for pain-related decision making by chronic disease-suffering patient using a frequent pattern tree for data pre-pressing, extraction and data mining of conventional medical data. By utilizing the information of patients which were the pain-related decision making, normalization can be applied to the frequent pattern tree of data mining. The mining frequent pattern tree is processed which identified potential but important information in large amounts of data so that it enables extraction of similar information patterns in a certain patient by using conventional chronic disease information for the support of pain-related decision making. The proposed method was reduced time and expenses for searching information for pain decision making of chronic disease patients who were frequently exposed to pain, unlike acute disease patients, and enables standardized and specific decision making.

In this paper an association rule mining method (Lee, D. G., et.al.2013) to discover interesting patterns which was included medical knowledge. The proposed method was removed imprecise patterns and discovers target patterns which contained associations between blood factors and disease history. The association that blood factors affect to disease history was defined as target pattern. The confidence threshold was ignored the support of itemset in consequence of patterns. However, the reliability and the reasonability of a target pattern were subjective in the medical domain.

In this paper, a hybrid model was presented (Seera, M., et.al.2015) which was consisted of the fuzzy ARTMAP (FAM) neural network, classification and regression tree (CART). FAM was useful for tracking the stability-plasticity dilemma per-training to data-based learning systems which CART was useful for depicting its learned knowledge explicitly in a tree structure. The combination of both models, FAM-CART had capable of learning data samples stably and at the same time, explaining its predictions with a set of decision rules. From the result, FAM-CART was outperformed compared to other conventional model.

In this paper presented the various classification techniques (Kunwar, V., et.al.2016) such as Naïve Bayes and Artificial Neural Network (ANN) for predicting Chronic Kidney disease. This research was considered following factors which were age, diabetes, blood pressure, RBC count etc. Furthermore, this work can be extend by considering other parameters such as food type, working environment, living conditions, availability of clean water, environmental factors etc for detection of kidney disease. Further research can be conducted by using other classifiers which are Fuzzy logic, KNN.

In this paper a novel technique was proposed (Moustafa, A., et.al.2015) namely fuzzy frequent pattern ubiquitous streams (FFP_USTREAM). The novel technique was integrated fuzzy concepts with ubiquitous data streams and performing sliding window approach to mine fuzzy association rules. Furthermore, the proposed model was helped in many practical situations to make more significant and flexible decision which was included determining stock required in retail applications, determining methods of treatment in medical applications and determining methods of precaution in road safety applications. The proposed model was handled distributed homogenous data streams. Thus, there was a need to analyze the cases where the distributed flow of data stream was heterogeneous and the data sets were incompatible.

In this paper, a new mechanism was proposed (Joy, R., & Sherly, K. K. 2016) for predicting disease based on faster-IAPI algorithm. This algorithm was supported incremental mining and which had capable of generating frequent patterns from a massive data store in two database scan. The proposed method was assured the advantages of in-memory computation over other models using spart RDD framework. The proposed approach was applied to determine correlation between various symptoms of patients in faster and efficient manner and provides the support for the prediction of occurrence of disease based on the symptoms.

In this paper, the classification model was proposed (Radha, N., & Ramya, S. 2015) to predict the chronic kidney disease using various machine learning algorithms. The following model were utilized for this research which were decision tree (DT), naïve bayes (NB), support vector machine (SVM) and k-nearest neighbors (k-NN). The experimental result shows the k-nearest neighbour was provided better diagnosis of chronic kidney disease.

3. Proposed methodology

In this section, kidney disease is predicted from clinical database by assigning priority for each feature in the dataset which utilize priority assigning algorithm to feed the most important features to the classifiers. The performance of WAELI-FPA is extended by enhancing priority assigning through mining fuzzy frequent pattern tree (MFFPT) and orthogonal partial least square (O-PLS).

3.1. Orthogonal partial least square (O-PLS)

In this section, the orthogonal partial least square (O-PLS) is allocated the priority to each feature in the prediction of chronic kidney diseases. The O-PLS is the supervised classification tool which is the extension version of PLS. The O-PLS is the very powerful dimension-reduction and visualization tool. It affords better interpretability and transparency compared to PLS. It is more practical and robust than traditional least squares regression methods, because it effectively handles collinearity of variables, noise in both blocks of variables and missing data. The OPLS is introduced by Trygg, possesses built in orthogonal signal correction (OSC) that filters out some variance in the X-matrix unrelated

to Y. The OPLS model separates systematic variation in the X-block into two parts named as predictive and orthogonal respectively. This model has been successfully applied to acquire new information of specific combination of metabolite, protein and transcript correlation. Hence, OPLS can integrate multiple data blocks to improve interpretation and identification of relevant information.

An OPLS model can be written as follows

$$X = TP^T + T_{orth}P_{orth}^T + E \quad (3.1)$$

$$Y = TP^T b + f = Tq + f \quad (3.2)$$

$$b = Pq \quad (3.3)$$

Where the orthogonal nXn score matrix for X and Y was represented by X, P is the orthogonal pXn loading matrix representing the regression coefficients of X on T, T_{orth} is the orthogonal nXn score matrix for X and Y, P_{orth} is its corresponding orthogonal pXn loading matrix and E is the nXp residual matrix of X. Values of y and b are calculated from equation (3.2) and (3.3) respectively. The regression coefficient b (Equ 3.3) is used to evaluate the contribution of the original variables to the final model.

Algorithm 1: Orthogonal partial least square (O-PLS)

Step 1: Calculate a dingle O-PLS model to discriminate between extraction and treatment.
Step 2: Maximizes the covariance between variables.
Step 3: The OPLS model is defined Equation (3.1) and (3.2)
Step 4: The scaling association of blocks predicts underlying structure that is present in all data blocks.
Step 5: Performed to compute a weighted sum of the respective association matrix from each block to determine a consensus space optimising the prediction accuracy.
Step 6: The weight optimisation step improves the prediction ability.
Step 7: Terminate Process.

3.2. Mining fuzzy frequent pattern tree (MFFPT)

In this section, the MFFPT is assigned the priority to each feature for achieving the optimal results in prediction of chronic kidney disease. Assume the clinical dataset; perform the frequent pattern mining to identify the set of frequent data. However, the frequent of data is defined according to the fuzzy membership functions. If the information of the fuzzy association at the appearance of a set of frequent data is captured with a compact data structure, restrict the repetitive scanning of the original database. When multiple data share to some degree a set of frequent data's, merge the shared fuzzy sets by aggregating their co-occurrences with a count value.
The global frequency list is determined by processing all the process of the database. The first pre-processing step is the thresholding of the fuzzy membership functions with a thresholdθ. Values smaller than θ are ignored since it is implied it has the condition possesses vary marginally the gene at the corresponding state. The second step is to sum over all the processes.
After constructing the global frequent data's, construct the frequent data for every condition.

Algorithm 2: Construction of Fuzzy frequent Pattern Tree

Input: A clinical database for which with some data can associate a degree of appearance. Also a minimum support thresholdθ.

Output: The fuzzy frequent pattern tree (FFPT) of the database.

Method: The steps for the construction of the fuzzy frequency pattern tree are as follows

1. Computation of the set of global frequent data with an aggregation of their fuzzy counts of appearance. Let Flist be the list of these data's. Each data is considered as feature of chronic disease.

2. Creation of the root of the FFPT. For the data of each process do the following process. First, select the frequent data in process and sort them in accordance with Flist. This step creates the condition data frequent list described previously. Then, denote the sorted condition data list of process by where head is the data at the head of the list tail the remaining frequent data. Finally, denote by μ_{head} (Process) the membership degree of the data head at the process.

The insertion is performed with a function that implements the following algorithm for each process.

```
FFPTreeNode insertFFPT (head, Tail, TreeNode)

If TreeNode has a child node Child such that
        Child.ItemID=head then
        Child.count=Child.count+ μ_head(Process)
Else {
        Create a new node Child;
        Child.parent=TreeNode;
        Child.count =μ_head(Process);
        Connect Child to the Node-Links structure
    } // else
Return child;
```

For every process, insert the whole conditions item list frequent item list by using the following pseudocode

```
InsertionPoint=root;

While the process data list is not empty do
        Get head and tail element
        InsertionPoint=insertFFPT (head, Tail, InsertionPoint)
End;
```

In order to facilitate the subsequent data mining, construct a header table structure that has one entry for each frequent 1-dataset (i.e., the data appearing at the global Flist). This entry keeps the name of the data, its total frequency count and the node links pointers that connect with a linked list of all the information concerning the data at the fuzzy FPT.

Mining the frequent patterns for fuzzy frequent pattern tree to achieve the optimal solution, which are described as following ways.

```
For all frequent patterns of the HeaderTable do
// path detection for the current frequent data fdata
        NodeListPointer=head link pointer for the data fdata
        from the header table
        Conditional Pattern Base (CPB) =null;
 While NodeListPoniter not null do
         CurrentPath=path from root to the current frequent data
        pointed by NodeListPoniter
        CPB=CPB+AdjustFrequencyOfData (CurrentPath,
        NOdeLitPointer, fdata)
        NodeListPointer=NodeListPointer.next;
End while

// mining of CPB

If the CPB is a single path
        Output all the combinations as frequent data's
Else
        RecursiveMine (CPB)
```

Disconect currently examined frequent data fdata from the FFPT since all frequent data concerning it were examined.

The function AdjustFrequencyOfData (Path,Node,Data) adjusts the frequency of the data data over the whole path path as the frequency of the last node of the path.

4. Result and discussion

In this section, evaluate the result of the proposed WAELI- FPA–MFFPT and WAELI- FPA -OPLS-DA and existing WAELI- FPA in terms of Root Mean Square (RMSE), Mean Absolute Error (MAE), Accuracy, Precision, Recall and F-Measure. The dataset relation Chronic Kidney Disease is used for evaluation. It contains attributes are age, blood pressure, specific gravity, albumin, sugar, red blood cells, pus cell, pus cell clumps, bacteria, blood glucose random, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema, anemia. The missing values are represented in the dataset as ?.

4.1. RMSE and MAE

RMSE and MAE value are used to evaluate the models the RMSE is not a good indicator of average model performance and might be a misleading indicator of average error, and thus the MAE would be a better metric for that purpose.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|e_i|^2}$$

Where n is the n samples of model errors ϵ calculated as ($\boldsymbol{e_i}$, i = 1, 2,..... , n).

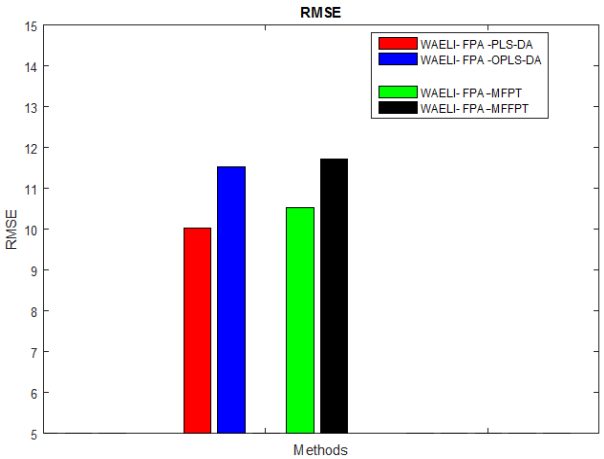


Fig. 4.1: Comparison of RMSE.

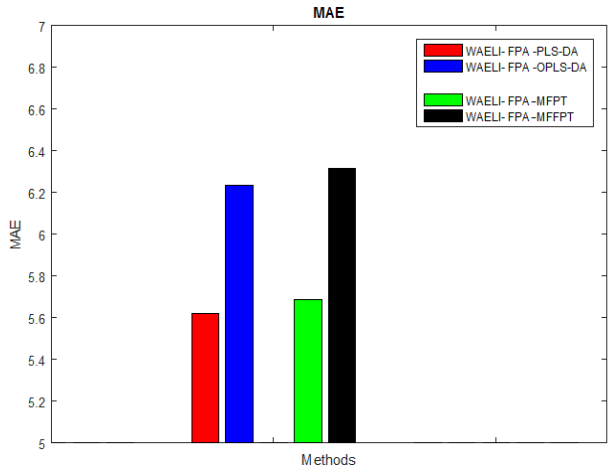


Fig. 4.2: Comparison of MAE.

Figure 4.1, 4.2 shows the proposed WAELI- FPA-OPLS- DA, WAELI- FPA–MFFPT has high value than the existing WAELI- FPA-PLS- DA and WAELI- FPA–MFPT in term of RMSE and MAE values.

4.2. Accuracy

Social Accuracy is defined as the proportion of true positives and true negatives among the total number of results obtained. Accuracy is evaluated as,

$$Accuracy = \frac{(True\ positive + True\ negative)}{(True\ positive + True\ negative + False\ positive + False\ negative)}$$

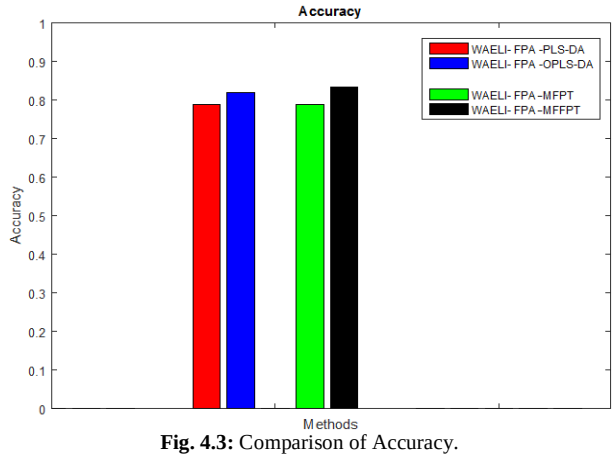


Fig. 4.3: Comparison of Accuracy.

Figure 4.3 shows the comparison of accuracy between proposed WAELI- FPA-OPLS- DA, WAELI- FPA–MFFPT and the existing WAELI- FPA-PLS- DA, WAELI- FPA–MFPT in terms of accuracy values. The result shows the proposed methods provides better value compare to existing methods.

4.3. Precision

Precision value is evaluated according to the relevant information at true positive prediction, false positive.

$$Precision = \frac{True\ positive}{(True\ positive + False\ positive)}$$

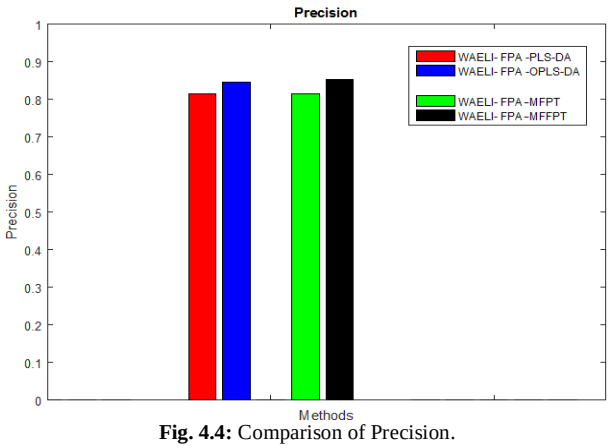


Fig. 4.4: Comparison of Precision.

Figure 4.4 shows the comparison of Precision between proposed WAELI- FPA-OPLS- DA, WAELI- FPA–MFFPT and the existing WAELI- FPA-PLS- DA, WAELI- FPA–MFPT in terms of Precision values. The experimental result describes the proposed methods are provided better performance than existing methods.

4.4. Recall

The Recall value is evaluated according to the retrieval of information at true positive prediction, false negative.

$$Recall = \frac{True\ positive}{(True\ positive + False\ negative)}$$

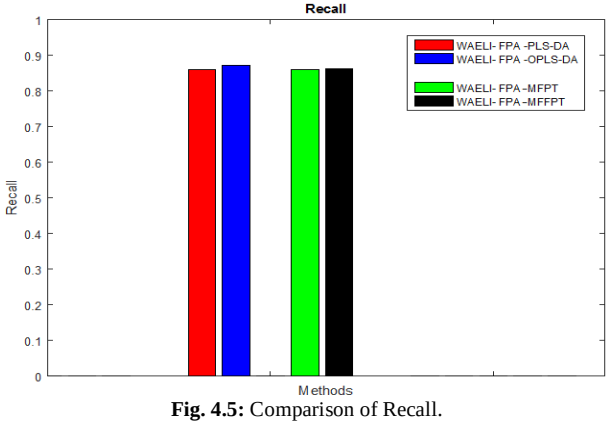


Fig. 4.5: Comparison of Recall.

Figure 4.5 shows the comparison of Recall between existing WAELI- FPA-PLS- DA, WAELI- FPA–MFPT and proposed WAELI- FPA-OPLS- DA, WAELI- FPA–MFFPT through recall values. The result assures the proposed method provides better performance than existing methods.

4.5. F-measure

F-measure is calculated from the precision and recall value. It is calculated as:

$$f - measure = 2 \times \left(\frac{precision \times recall}{precision + recall}\right)$$

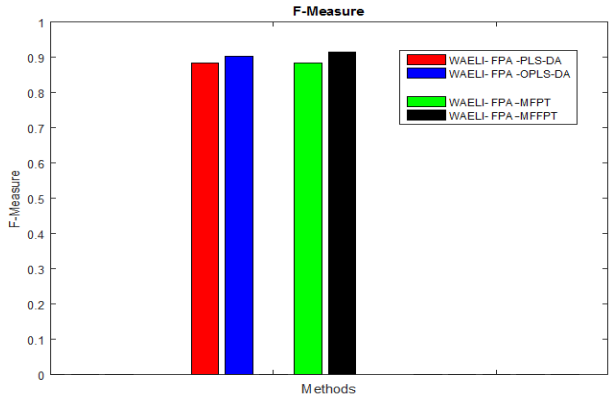


Fig. 4.6: Comparison of F-Measure.

Figure 4.6 shows the comparison of F-Measure with existing WAELI- FPA-PLS- DA, WAELI- FPA–MFPT and proposed WAELI- FPA-OPLS- DA, WAELI- FPA–MFFPT with F-Measure values. The outcome result shows the proposed methods are provided better result compared to existing methods.

Table 1: Comparison between WAELI- FPA -PLS-DA, WAELI- FPA -OPLS-DA, WAELI- FPA–MFPT and WAELI- FPA–MFFPT

	WAELI- FPA -PLS-DA	WAELI- FPA -OPLS-DA	WAELI- FPA–MFPT	WAELI- FPA–MFFPT
RMSE	10.0254	11.5223	10.5261	11.7216
MAE	5.6210	6.2314	5.6845	6.3124
Accuracy	0.7885	0.8201	0.7883	0.8326
Precision	0.8137	0.8459	0.8140	0.8523
Recall	0.8600	0.8721	0.8596	0.8623
F-Measure	0.8841	0.9018	0.8844	0.9145

5. Conclusion

In this paper, the modifying the two classifier model is performed such as mining fuzzy frequent pattern tree (MFFPT) and orthogonal partial least square (O-PLS) to efficiently predict the kidney disease in patients. The both classifier models are improved the classification accuracy of predicting kidney disease by assigning priority on each feature in the clinical dataset. Initially, the O-PLS is performed efficiently in the reduction of missing values and assigning priority to each feature. Second, the MFFPT is performed which is initially determines the frequent data then performed priority to predict the kidney disease. The performance of the proposed approach is evaluated in terms of RMSE, MSE, Accuracy, Precision, Recall and F-Measure.

References

[1] Rogeith, V., & Magesh, S (2017). A SURVEY ON HEALTH CARE DATA USING DATA MINING TECHNIQUES. International Journal of Pure and Applied Mathematica, 117(16).

[2] Couser, W. G., Remuzzi, G., Mendis, S., & Tonelli, M. (2011). The contribution of chronic kidney disease to the global burden of major noncommunicable diseases. Kidney international, 80(12), 1258-1270 https://doi.org/10.1038/ki.2011.368.

[3] Yildirim, P. (2017, July). Chronic Kidney Disease Prediction on Imbalanced Data by Multilayer Perceptron: Chronic Kidney Disease Prediction. In Computer Software and Applications Conference (COMPSAC), 2017 IEEE 41st Annual (Vol. 2, pp. 193-198). IEEE.

[4] Chen, Z., Zhang, Z., Zhu, R., Xiang, Y., & Harrington, P. B. (2016). Diagnosis of patients with chronic kidney disease by using two fuzzy classifiers. Chemometrics and Intelligent Laboratory Systems, 153, 140-145. https://doi.org/10.1016/j.chemolab.2016.03.004.

[5] Ilayaraja, M., & Meyyappan, T. (2013, February). Mining medical data to identify frequent diseases using Apriori algorithm.

In Pattern Recognition, Informatics and Mobile Engineering (PRIME), 2013 International Conference on (pp. 194-199). IEEE.

[6] Jung, H., Chung, K. Y., & Lee, Y. H. (2015). Decision supporting method for chronic disease patients based on mining frequent pattern tree. Multimedia Tools and Applications, 74(20), 8979-8991.

[7] Lee, D. G., Ryu, K. S., Bashir, M., Bae, J. W., & Ryu, K. H. (2013). Discovering medical knowledge using association rule mining in young adults with acute myocardial infarction. Journal of medical systems, 37(2), 9896. https://doi.org/10.1007/s10916-012-9896-1.

[8] Seera, M., Lim, C. P., Tan, S. C., & Loo, C. K. (2015). A hybrid FAM–CART model and its application to medical data classification. Neural Computing and Applications, 26(8), 1799-1811. https://doi.org/10.1007/s00521-015-1852-9.

[9] Kunwar, V., Chandel, K., Sabitha, A. S., & Bansal, A. (2016, January). Chronic Kidney Disease analysis using data mining classification techniques. In Cloud System and Big Data Engineering (Confluence), 2016 sixth International Conference (pp. 300-305). IEEE.

[10] Moustafa, A., Abuelnasr, B., & Abougabal, M. S. (2015). Efficient mining fuzzy association rules from ubiquitous data streams. Alexandria Engineering Journal, 54(2), 163-174. https://doi.org/10.1016/j.aej.2015.03.015.

[11] Joy, R., & Sherly, K. K. (2016, March). Parallel frequent itemset mining with spark RDD framework for disease prediction. In Circuit, Power and Computing Technologies (ICCPCT), 2016 International Conference on (pp. 1-5). IEEE.

[12] Radha, N., & Ramya, S. (2015). Performance Analysis of Machine Learning Algorithms for Predicting Chronic Kidney Disease.