An Aggregate Model for Prognosticate Diabetic Disease using Dissimilar Feature Selections with Upright Classification Techniques

P.Anitha, P.R.Tamilselvi

Abstract— Our aims are to find the accuracy of classification with the normalisation in different types and the features in the techniques of selection on Diabetic Mellitus and the Pima Indian Diabetic dataset. Data Mining is the process of extraction. It extracts the previous unknown, valid and important information from the large amount of the data bases and can make the crucial decisions using the information. The classification methods are K-Nearest Neighbour and J48 decision tree can be applied to the data set of original and as well as the dataset with the pre-processed dataset. All the process of pre-processing can be applied to Pima Indian Diabetic Dataset to analyse the classification performance in terms of accuracy rate. The performance metrics is used to identify the accuracy classification is Recall, F-measure, Sensitivity and specificity, Precision, and Accuracy. The simulation is done by R tool.

Keywords: Data Mining, Health Informatics, J48, KNN, Dataset.

I. INTRODUCTION

Data Mining [8] is the field of co-operative and bringing together the techniques from various field The fields like learning the machines behaviour about feature, recognizing the learning of patterns, huge diabetic databases, the analysis about statistical data and many of the visualization techniques to extract the processed data from huge amount of databases. Knowledge Discovery in Databases (KDD) is used to mine the data from the irrelevant detection and from the previously not known and potential helpful processed data [15] in Databases. The steps involved in the mining process consists interactively beginning from the collections of unprocessed data to some other form of the entirely new processed knowledge.

The feature selection and classification technique provide better accuracy and reduces irrelevant data. The intentionally reduced dataset improves the result in the ancient methods which drops the noise of the data. The ensemble model gives a greatest advantage by furnishing good results of predicting diabetic disease.

II. LITERATURE SURVEY

Amit kumar Dewangan and Pragati Agrawaldesigned a model using multilayer perceptron and Bayesian classification. They used various techniques and software tools. In this work they focus on sensitive and accurate data.

V. AnujaKumari and R.Chitra proposed a classification model using Support vector machine to classify the diabetic diagnosis. This is a machine learning method. This is used to classify a high dimensional huge data set. They use the Pima Indian database for diabetic.

MadhuriPanwar, Amit charyya, Rishad A. Shafik and Dwaipayan Biswas proposed a method for diabetic diagnosis using the classifier called K-nearest neighbor algorithm. They have done various comparative analysis with the existing reports. The work has shown many results by feature reduction.

III. METHODOLOGY

Step 01: For the mining of data, initially the data require to be acquired from the information of obtainable source. The data origin can be distributed in the heterogeneous diabetic databases and aggregation techniques are applied in diabetic databases.

Step 02: when the aggregated information is collected in the warehouse of data for particular data mining activities. The activities having a mind of particular purpose, the information can be selected from the data which warehoused. These activities for a specified task of data mining are functional process of discovery is applied.

Step 03: The duties of in data mining is the procedure of finding the fascinating models in the data. That can be normally interpreted as data mining in the sense of restricted and such as rule mining with association.

Step 04: The particular patterns are identified, and the patterns are developed further. The development of additional patterns can be implied to the research performance analysis of "interestingness" for the particular issues analysed and the measures of deployed patterns in discovery.

Diabetes: Now-a-days Diabetes [1][4] is the most common disease in all age of group. It is a problem in the human body, so it doesn't produce any insulin. Insulin is essential in our body for the glucose production.

Manuscript published on November 30, 2019.
* Correspondence Author

P. Anitha*, Research Scholar & Assistant Professor, Vellalar College for Women (Autonomous), Erode–638012, Tamilnadu, India. (Email: anithavcw@gmail.com)

Dr.P.R.Tamilselvi, Assistant Professor, Govt Arts and Science College, Komarapalayam – 638183, Tamilnadu, India. (Email: selvipr2003@gmail.com)

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

When our human body affect by diabetes the insulin doesn't secrete or secrete only little. In this, glucose is available only plenty and the blood stream unable to that. They are various types of diabetes namely, Type-1 Diabetes, Type-2 Diabetes and Gestational Diabetes. Diabetes can cause severe complication in health including blindness, kidney failure, heart disease and lower-extremity amputations.

Type-1: Diabetes occurs only when our body's immune system is attacked and pancreas beta cells are destroyed. It causes insulin deficiency. It may recover by Type-1 diabetes insulin.

Type-2: It is caused by deficiency of relative insulin. In this type it produces the insulin but it is not enough to control the blood glucose. This type is common in the age of above 40.

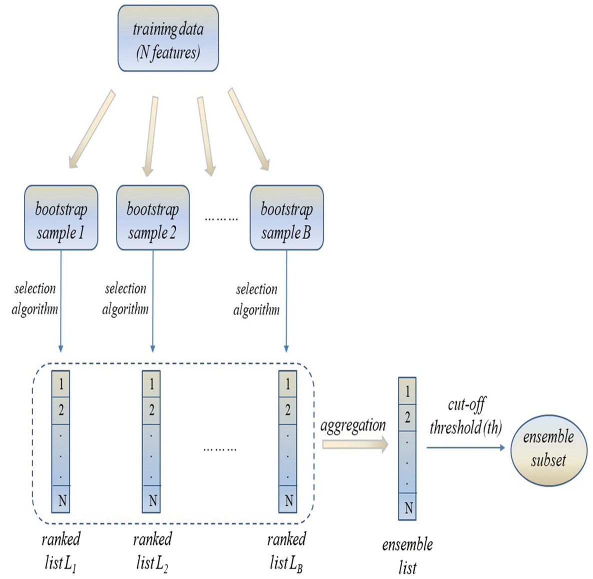

Figure 2.1 Group Ranking Approach

Methodology: Now-a-days many people are affected by the diabetic mellitus disease. The patient's counts are getting increased in day by day and it is most dangerous leading disease. The hospitals are maintaining the medical records of patients in electronic medium who are affected by various diseases. It is more difficult to maintain the information and collect the other information for the manual records. Data mining techniques are used for collecting the knowledgeable information from the databases in medical field. Supervised learning can design models for describing the data classes, where the class attributes are involved in construction. There are many steps to divide the disease of diabetic of mellitus or non-diabetic mellitus an example, and the Simulation of R tool is utilized for uncomplicated implementation.

1. Normalized data for attributing class.
2. The more redundant feature [11] is selected to utilize the patterns, with linear correlation of the selection features of Rank 13][14] and Recursive and Feature by the Elimination methods.
3. Apply transformation of data techniques such as Max-Min and Z-Score normalization technique is applied as transformation of data.
4. For previously processed data or raw data the KNN algorithm is applied.
5. For the unprocessed information or data to be processed the Decision Tree is applied.
6. The performance metrics such as accuracy, precision and recall are calculated.
7. Examination of the results for recognizing the effect of sorting of parameter and data change over on separation [12] for Diabetic Mellitus Disease Data Set.

IV. IMPLEMENTATION OF CLASSIFICATION ALGORITHMS

K-Nearest Neighbor (KNN)

KNN is the supervised learning algorithm that classifies the new data using them minimum distance from the data of new to the K-Nearest Neighbor[5].

Euclidean distance is used to define the closeness.

$$\mathrm{dist}(X,Y) = \sqrt{\sum^{n}_{i=0}} (X_i-Y_i)^2\text{------} \qquad (1)$$

In each numeric attribute, the difference between the values of corresponding data and there attribute in tuple X and in tuple Y, square the difference and accumulate it.

Min-max normalization, for example, it is used to change over a value X_{mm} of the characteristics of numeric A to X_{mm} in the range [0, 1] by calculating

$$X_{mm} = \frac{X-\min(X)}{\max(X)-\min(X)} \quad \text{-----} \qquad \text{-(2)}$$

Rules, using a single rule with highest confidence, predicate the class label of a new tuples. Z-score Normalization transforms the data by converting an average of zero and standard deviation method of the values to a common scale with,

$$v' = \frac{v - \dot{F} \quad \text{----}}{\sigma_F} \qquad (3)$$

Decision Tree (J48): Decision tree is the way of representing the rules which will be class or value. In this algorithm it takes dataset as input. The entropy and the gained information is used to calculate the attributes of dataset training. The attributes are ranked by using the gained information. The attributes which has the maximum gained information it refers as the root attribute.

The process will be performed continuously until no attributes are left. Node refers as a selection attribute for a tree. And at-last the tree with conclusion and leaf nodes are obtained[13]. Leaf motes show the class label. When the parent tree is developed and it is kept for future structure. And this algorithm for developing the verdict tree for mining the data is called as ID3. It assign a top-down fashion, greedy look over through the space of feasible branches with no retreating.

ID3 utilize the entropy and take the processed data to make a decision tree. C4.5 is an efficient algorithm by Ross Quinlan to be utilized to produce a decision tree for deploying. C4.5 is an expansion of the decision tree algorithm of Quinlan's earlier ID3 algorithm.

V. RESULTS AND DISCUSSION

In the proposed performed research work done, the Pima Indian Diabetic Dataset is first assigned to the data of previously processed and uses the data.

The examined data are assigned to normalization by couple of transformation highlighted techniques namely Max-Min and Z-Score Normalization. The examined data are then assigned to feature sorted process using parameter selection. The tasks are then divided to couple ofdistinct classification methods. During previously processing the data, it is detected that there is no missing of values in the analysed datasets.

In the fore coming process, the dataset is repeatedly assigned to feature selection which is utilized by three distinct methods of feature selection. In "Redundant Remove Feature" method, the parameters of highly correlated are detected and then the features are removed.

Table 1. KNN Normalization

	Diabetic MellitusDisease		Recall	Precision	Accuracy	F-measure
	No. of patterns	No. of parameters				
J48	867	7	0.8082	0.6484	0.7195	0.8160
Normalizationprocess of Z-Score with KNN	867	7	0.9195	0.8791	0.8988	0.9280
Normalization process of Max-Min With KNN	867	7	0.8000	0.8889	0.8421	0.8929

In the Pima Indian Diabetic Dataset, age is largely equivalent parameter and it is destroyed for the division process. The fore coming feature selection method is the important rank features, by utilizing the obtained vector [3]model, the focused features can be analysed from a information. It gives ins, bmi, glu, npreg and skin are the six peculiar features for classification.

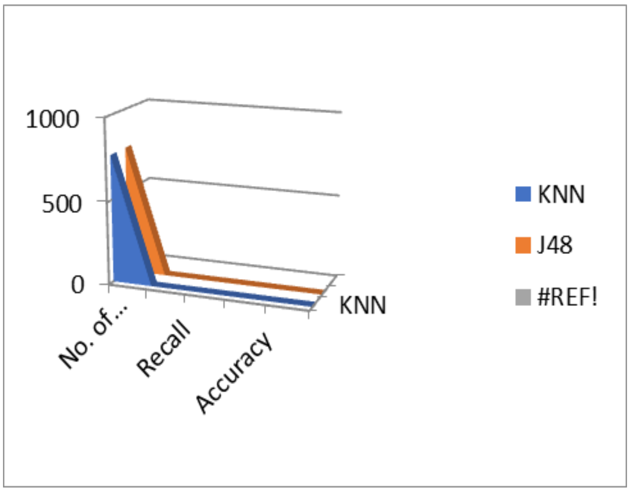

Figure 4.1 Analysis Measures of Classification performance with Pre-processing and Recursive Feature Elimination Approach

In this proposed research work the different method used is called "Feature Elimination by Recursion" method. And the method provides the exact results. Random forest algorithm is utilized for examining the model in each iteration.

And the above algorithm expands all the subsets of the data attributes. It gives ped, ins, bmi, glu, skin and npreg are the subset characteristics for the classification.

Table 2. Classification performance before preprocessing

	Diabetic Mellitus Disease		Recall	Precision	Accuracy	F-measure
	No. of patterns	No of parameters				
KNN	768	9 - class Label including	0.4111	0.7400	0.5286	0.7360
J48	768	9 - class Label includes	0.5000	0.7500	0.6000	0.7600

The datasets divided by utilizing J48 and KNN decision tree by with and without feature selection and previous processing.

The dataset of each division is calculated and contrasted to verify the variation in accuracy due to the data pre-processing. The proposed research utilized65% of the data is examined. During examination it is concluded that it is a training data and 35% of data is analysed as it is used as testing data from the Pima Indian diabetic disease dataset which is taken and processed from UCI Machine Learning Repository [16].

The dataset is taken and processed using the database. Obtained data is used in the parameters in Pima Indian Diabetic Dataset. While analysing, the parameters, 1.65% is processed by glu parameter, 5.5% processed by bp parameter, 13% is processed by npreg parameter,28.5% is processed by skin parameter, 2.4% processed by bmi parameter and 47.6% processed by ins parameter. After analysing all the resultant parameters which are detected and change over with its normalized data. The classifier examined with raw data.

It is observed from the table. The J48 decision tree has the more Accuracy rate. By this research the following measures can be evaluated.

Recall: Recall is calculated using the fraction of correct instances among all the instances that belong to the relevant subset i.e (Recall is known as the Actual True Positive rate).

Recall = TP / (TP + FN) ---- (5)

Precision determines the fraction of records that actually are the positive in the group. The classifier has been declared as the positive class.

Precision = TP / (FP + TP) ---- (4)

F-parameter is a feature that joins recall and precision. The harmonic mean is said to be as recall and precision parameter, in the existing F-parameter.

F-parameter = 2 /((Recall.Precision) / Recall + Precision)) ---- (6)

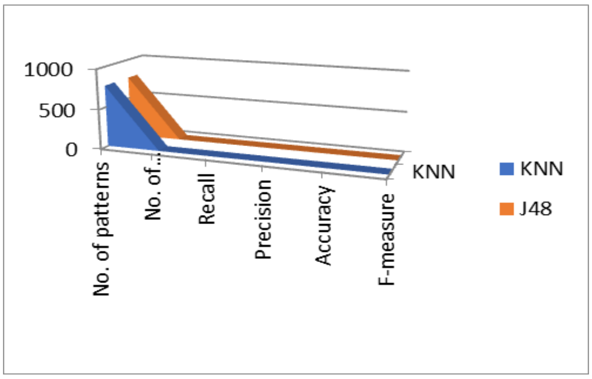

Fig.4.2 Examining the Measures of Classification Performance before Pre-processing

According to the properly predicted cases the accuracy is the calculated in fraction. [6].

Accuracy = ((FN*TP) / (FP*TP)) ----- (7)

VI.CONCLUSION

In medical diagnosis the mining of data played a vital role. The data division techniques are largely utilized in medical diagnosis field for the predicting the disease [9] and diagnosis of peculiar diseases. After examining the pima Indian dataset, the results are detected and then the results are normalized .Further, three distinct parameter used in electing the access are deployed to form the database subset to obtain a vital characteristics for the division algorithm. Couple of distinct characteristics of change over techniques are proposed on diabetic dataset and examined with the famousJ48 and KNN classifier.

Recall and accuracy are taken as the most vital parameters in the medical diagnosis field of a critical patient of diabetic disease. Examination results predict that the patient's reports conclusion will process normalised parameters and parameter selection methods are largely expands the accuracy of feature classification. The attribute performance of whole three parameter selection methods with J48 classifier and KNN decision classifier are analysed. KNN with the normalized Z-Score data change over methods and Feature Elimination using Recursion method produce good results for accuracy in Pima Indian Diabetic Dataset.

In the future expansion of our work, it is planned to process numerical processed data which can be expanded to feature categorical data. Hence forth, a professional analyzation approaches for parameter election that can be detected. It is concluded that it acquires minimum cost effective. Distinct hybrid optimization [7]can be utilized for effective analysis of some peculiar pre-processing methods.

REFERENCES

1. Amit kumarDewangan, Pragati Agrawal, "Classification of Diabetes Mellitus Using Machine Learning Techniques", in May 2015.
2. V.AnujaKumari, R.Chitra, "Classification of Diabetes Disease Using Support Vector Machine", in March -April 2013.
3. Jiawei Han, MichelineKamber, Jian Pei, "Data Mining Concept and Techniques", Elsevier Ins, 2012.
4. KanakSaxena, Richa Sharma, 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions) "Diabetes Mellitus Prediction System Evaluation Using C4.5 Rules and Partial Tree", 2015.
5. Madhuri Panwar, Amit Acharyya, Rishad A. Shafik, Dwaipayan Biswas,6th International Symposium on Embedded Computing and System Design,"K-Nearest Neighbor Based Methodology for Accurate Diagnosis of Diabetes Mellitus", IEEE Digital Xplore, 2016.
6. NirmalaDevi,Balamurugan and Swathi,IEEE International Conference ON Emerging Trends in Computing, Communication and Nanotechnology (ICECCN),"An amalgam KNN to predict Diabetes Mellitus",2013.
7. Omar S.Soliman, EmanAboElhamd, "Classification of Diabetes Mellitus is using Modified Particle in Swarm Optimization and Least Squares Support Vector Machine", Research Gate, Feb 2014.
8. M.Panda, S.Dehuri, M.R.Patra, "Modern Approaches of Data Mining Theory and Practice", Alpha Science International Limited, 2015.
9. Priyanka Shetty, Sujata Joshi, "A Tool for Diabetes Monitoring and Prediction Using Data Mining Technique",I.J. Information Technology and Computer Science, 2016, 11, 26-32,2016.
10. SushmitaMitra, TinkuArchaya, "Data Mining Multimedia, Soft Computing And Bioinformatics", John Wiley & Sons,.Ins, 2003.
11. Seijo-Pardo B, Porto-Díaz I, Bolón-Canedo V, Alonso-Betanzos A (2017) Ensemble feature selection: homogeneous and heterogeneous approaches. Knowledge-Based System 118:124–139.
12. Rojas-Thomas JC, Mora M, Santos M (2017) Neural networks ensemble for automatic DNA microarray spot classification. Neural ComputerApplications. https://doi.org/10.1007 /s00521-017-3190-6.
13. Golay J, Leuenberger M, Kanevski M (2017) Feature selection for regression problems based on the Morisita estimator of intrinsic dimension. Pattern Recognition 70:126–138.
14. Neumann U, Heider D (2018) Ensemble feature selection for regression problems. In: European conference on data analysis (ECDA 2018), book of abstracts, pp 19.
15. Brahim AB, Limam M (2017) Ensemble feature selection for high dimensional data: a new method and a comparative study Adv Data Analysis Classification 1:2– https://doi.org/10.1007/s11634-017-0285-y.
16. Witten, Frank, Hall, Pal, "DATA MINING: practical machine learning tools and techniques", Morgan Kaufmann, Burlington,2016.
17. https://archive.ics.uci.edu/ ml/machine-learning-databases/pima-indians-diabetes/pima-indians-diabetes.data

AUTHORS PROFILE

Ms.P.Anitha.,received the M.C.A. from Bharathiar University and M.Phil Computer Science from Periyar University. She is working as an Assistant Professor, Department of Computer Applications , Vellalar College for Women, Erode, Tamil Nadu, India . She has been involving herself in teaching for the past 13 years. 10 M.Phil Scholars have completed their research under her guidance.She is pursuing her Ph.D in Periyar University. Her area of research interest is Data Mining. She has published 5 publications. She had attended and presented manypapers in national and international seminars and conferences. She has guided P.G and U.G students for doing their semester project, mini and major projects.

Dr.P.R.Tamilselvi received the M.Sc and M.Phil Computer Science from Bharathiar University. She has completed Ph.D from Anna University, Chennai. She has guided 20 M.Phil scholars from various universities. She has presented and published many papers in national and international conferences and journals. She has organized many seminars, workshops and conferences. Currently she is working as an Assistant Professor at Government Arts and Science College, Komarapalayam. She has guided P.G and U.G students for doing their semester project, mini and major projects. Her area of research interest is Medical Imaging.