Correlation based Ensemble Feature Selection Algorithm for Diagnosis of Diabetics

R. Kuppuchamy, T. Kamalavalli, S. Vinothini, N. Jayalakshmi, N. Vallileka

Abstract: ***One of the preprocessing steps is data cleaning and feature selection in data mining. Feature selection has more efficiency regarding dimensionality reduction, eliminating irrelevant data, improving the accuracy and enhancing the output comprehensibility. This paper utilizes wrapper / hybrid-filter based feature selection method for feature selection and extraction from medical dataset. From the extracted information, the individual features are evaluated by calculating a rank value where it helps to choose highly correlated data from the entire dataset. Selected features are classified using the popular C4.5 classifier. To experiment the proposed method, the benchmark dataset is obtained from the UCI repository. It is a famous machine learning repository used by several earlier research works to evaluate the performance of their proposed methods. Finally, the accuracy of the classification method shows that our proposed method outperforms than the existing methods.***

Keywords: C 4.5, Correlation, Information gain, Feature selection, Classification

I. INTRODUCTION

One of the major researches works and involving in computing industries is data mining, is used to extract the expecting information from the voluminous data [1]. Data mining process has more ability to identify and detect hidden patterns available in the entire dataset. Nowadays, it is mainly used in medical industry for diagnosing medical data. Data mining involves various processes like data learning, preprocessing, normalization, dimensionality reduction, clustering and classification. In other words, it can able to predict the unknown patterns in the large volume of data. These above said processes are highly essential in data mining methods [2]. However, feature selection is one of the major tasks where it reduces the dimensional space, includes selecting discriminating features, mapping function and reduce the complexity [3].

One of the diseases which cannot be cured easily or suddenly is Diabetes and called as life-long disease. Diabetes is characterized by the sugar-level occurred in the blood, where it is because of the deficiency of insulin in the blood. It may be due to the human body not able to produce the amount of insulin. There two types of diabetes are called as Type-1 and Type-2 diabetes, Type-1 is the former level and the next level is Type-2 diabetes. Basically, most of the people affected by Type-2 diabetes [4], especially adults.Obesity, unhealthy diet, sedentary lifestyle and inheritance of the family are the main factors of diabetes. Author in [5] stated that diabetes occurs mostly due to genetic behavior of the family and environmental issues. One of the other common reasons for diabetes is sedentary life-behavior and obesity, where it is due to the emerging technologies, people have less active and physical exercise make obesity and leads to diabetes. To identify the level or type of diabetic, the medical industry uses data mining models for analyzing the medical data (diabetes data). Hence this paper aimed to use data mining model for learning and analyzing the diabetes data.

Main process of the feature selection method is labelling the unlabeled data [6]. Instead of selecting too many features, only the best features need to be selected for predicting the required data pattern. Choosing a greater number of features overshadow the patient information and degrades the classification/prediction accuracy. So, the author in [7] discussed only about feature selection methods.

In order to provide better results, this paper used the integration of information-gain-joint model with correlation-based feature selection model to select the relevant and most appropriate features from the dataset. The best features chosen from the dataset is feed as input to decision tree classifier. Section-2 presents a related work about the feature selection. Section-3 discusses wrapper & filter feature selection methods. Further section-3 describes information gain; Correlation based feature selection as subset selection and Decision tree as classification. Section 4 briefs about the proposed study. Section-5 and 6 discussed about the experimental results and conclusion of the proposed method.

II. RELATED WORK

It is well known that the data size/volume can be reduced by feature selection method is one of the ways. While choosing the features, it links a highly amount of data by considering the feature as the label. One label can connect more data; hence it reduces the data. Some cases, by eliminating redundant features can also use for dimensionality reduction. Author in [8] proposed a hybrid feature-selection model for high-dimensional data where it split the works and do the process. Hence it is suitable for any size of dataset. Author in [9] used correlation-based features selection method by investigating the independency among the variables. It also has applied clustering method for finding the correlation. Author in [10] proposed gain-ratio method for obtaining correlated features.

Revised Manuscript Received on January 2, 2020.

R. Kuppuchamy, Department of MCA, PSNA College of Engineering and Technology, Dindigul, Tamilnadu, India.

T. Kamalavalli, Department of MCA, PSNA College of Engineering and Technology, Dindigul, Tamilnadu, India.

S. Vinothini, Department of MCA, PSNA College of Engineering and Technology, Dindigul, Tamilnadu, India.

N. Jayalakshmi, Department of MCA, PSNA College of Engineering and Technology, Dindigul, Tamilnadu, India.

N. Vallileka, Department of MCA, PSNA College of Engineering and Technology, Dindigul, Tamilnadu, India.

Author in [11] used Information-gain method for fetching direct features available on the dataset. This method used the main features from genetic algorithm, KNN and Naive Bayes method obtaining the rank-based features. Author in [12] proposed Filters integrated with information-gain method for dimensionality reduction and support vector machine used as classifier. This method can eliminate irrelevant information and noise from the hyperspectral images. P. Jaganthan et al. [13] proposed a method called threshold fuzzy entropy which is used to identify the relevance features. Selected features are measured by Radial Basis Function classifier for medical data classification. JaganathanPalanichamy et al. [14] calculated appropriate features using context level information based on conditions.

The main aim of the paper is to maximum relevancy and minimum redundancy. JaganathanPalanichamy and KuppuchamyRamasamy [15] implemented the filter-based feature-selection procedure based on mutual information. From the above discussion and their experimental results, it is found that the feature selection by combined correlation-based filter approach with information gain is marginal improvement for classification accuracy is better only by using information gain approach, than the other methods.

III. FEATURE SELECTION

One of the main pre-processing tasks is feature selection. It selects all the available features from the input dataset and map the whole data, which automatically reduces the number of features. It increases the speed of the data mining process and accuracy of the prediction. Some of the irrelevant features, redundant features may produce negative results in prediction and increases the computational complexity. It can be reduced only by an efficient feature-selection and classification algorithm. Most of the earlier research works have used wrapper and filter method. Wrapper method uses significant feature extraction and it provides better performance than the filer method. Comparing with filter methods the wrapper methods extracts the optimized features for feature selection to predict the data pattern. Wrapper methods are too costly regarding the dimensionality reduction in high dimensional data [16].

The next method is filter method where it was the only method used very first as a classification method. Filter methods are not dependent with another methods. Comparing with the wrapper methods filter methods are simple, scalable and fast regarding computation methodology. Filter based selected features can be used more times for classification [8].

Information gain

One of the classification algorithms is ID3, where it uses the information gain, which is used to calculate the value of the features w.r.t feature label. ID3 is used in decision classifier algorithm. The calculated values are used to recognize the important features and mapped with the data. It automatically reduces the data size. All the features obtain the information-gain value utilized for deciding about which features need to be selected. Hence Decision Tree (DT) method is used for classification. DT method utilizes entropy value for recognizing the best feature to be selected for classification. Sametime features can also be selected using a threshold value. The condition behind the threshold value calculation is, the feature information should be bigger than the threshold value. Mathematically expressing the classification process is given here:

The number of occurrences (n) be represented as a set S, the k number of classes as C. Fraction value P (Ci, S) from S has the class Ci, the predictable class can be written as:

$$Info(S) = -\sum_{i=1}^{k} P(C_i, S) \times log(P(C_i, S)) \qquad -1$$

Feature “A” has the different value “v”, where “A” is the root of the tree, and the weighted sum value of the tree and the sub-tree is used for predicting the information in accordance to the different values. The set of occurrence Si and Ai is the value of the feature A is written as:

$$Info_A(S) = -\sum_{i=1}^{v} \frac{|S_i|}{|S|} \times Info(S_i) \qquad -2$$

Then the different information among information of (Si) and information (A, S) provides the information gained by partitioning S in accordance the gain value of A is:

$$Gain(A) = Info(S) - Info_A(S) \qquad -3$$

The highest amount of information-gain is used for obtaining the accurate classes from the available classes as the target class.

Correlation based Feature Selection (CFS)

One of the main feature selection method is Correlation based Feature Selection (CFS) method assesses the subset features used to measure the individual prognostic skill of the features. The coefficient of correlation is used for correlating the association among the features of the subset and the classes of the variable including with the inter and intra-correlation among the features. Relevant features group increases correlation among the features and classes, eliminates the inter-correlation. CFS method is mainly used in DT method for selecting the best features. It also combined with various searching methods like forwards selection, bi-directional search and etc. The following equation represents the CFS calculation.

$$r_{zc} = \frac{k\overline{r_{zi}}}{\sqrt{k+k(k-1)\overline{r_{ii}}}} \qquad -4$$

Where, r_{ZC} represents the correlation among the feature sum and feature subset sum. The number of features is represented as k, r_{Zi} represents the average value of the correlation and r_{ii} represents the average value of the inter-correlation among the feature subsets.

Information gain

Decision tree is a well-known technique which was first proposed by J R Quinlan [12]. The basic idea of decision tree is a greedy algorithm that constructs decision trees in a top-down recursive divide and conquer manner. First, select a feature to place at the root node and make one branch for each possible value. This splits up the example set into subsets, one for every value of the feature. Now the procedure can be repeated recursively for each branch, using only those instances that actually reach the branch. Whenever all instances at a node have the same classification, stop developing that part of the tree.

The information gain measure is used to select the test attribute at each node in the tree. The first measure is called as entropy is defined as

$$Entropy(S) = -\sum_{i=1}^{c} p_i log_2(p_i) \qquad - 5$$

Where pi is the proportion of S belongs to class i. The information gain, Gain(S,A) of an attribute A, the expected reduction in entropy caused by knowing the value of attribute A is defined as

$$Gain(S,A) = Entropy(S) - \sum_{value\ (A)} \frac{S_v}{S} \times Entropy(S_v) \qquad -6$$

Where Sv is the subset of S for which attribute A has value v. The attribute with the highest information gain is chosen as the test attribute for the given set S [3].

IV. PROPOSED METHOD

Feature selection develops computation efficiency and classification accuracy in classification problems with multiple features, as not all features essentially influence classification accuracy. Selecting suitable features improves the classification accuracy.

In this study, the feature selection method is used to select features and classification algorithms to assess the performance of the proposed method. Figure 1 shows the process of the ensemble feature selection method. In the first phase, Information gain (IG) was used to measure the importance of each feature with respect to the class variable. Weka [15] is a collection of machine learning algorithms for data mining tasks, which is used to calculate the information gain value of each feature and arrange the features in accordance with their information gain value. A feature with a larger information gain value specify higher discrimination of features compared to other category features and that features can be used to calculate classification results. After determining the information gain values for all features, a threshold for the results is recognized. The results of the information gain values are zero after the computation process are irrelevant for classification. If the information gain value of the feature is larger than the threshold, the features are selected for classification and other are not selected.

In the second phase, Correlation based feature selection method is used. This method is to select the most relevant features from selected features which are information gain value of the feature is larger than the threshold value. Exhaustive search is used as search method with Correlation based feature selection as subset estimating method. C4.5 is a comparatively talented decision tree algorithm and the Weka version of J48 is used to classify the selected features.

V. EXPERIMENT AND RESULTS

A. Data Set

In this study, the Pima Indians Diabetes Database was used and analyzed. The data was collected by the US National Institute of Diabetes and Digestive and Kidney Diseases and available at UCI repository [13]. The diagnostic, binary-valued variable investigated is whether the patient shows signs of diabetes according to World Health Organization criteria. The database contains details of 768 females all of which are older than 21 and 8 features detailed in Table I and the 9 being the class variable.

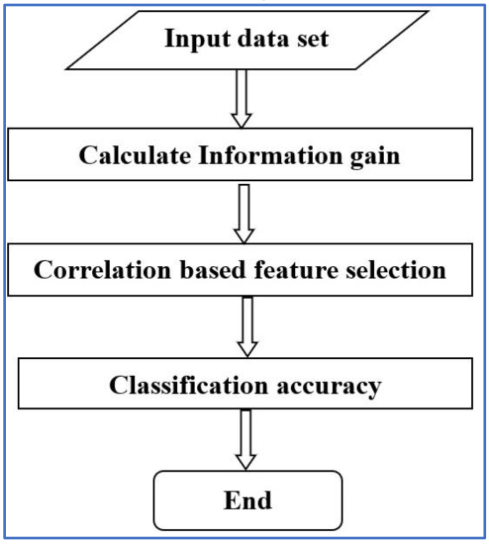

Input data set
Calculate Information gain
Correlation based feature selection
Classification accuracy
End

Fig. 1: Architecture of proposed method

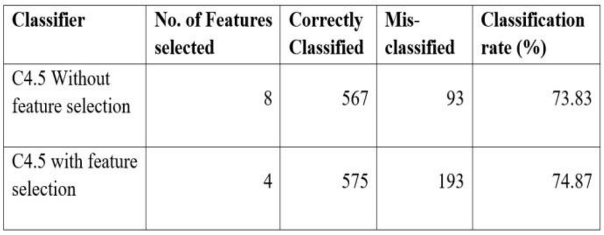
Table I: Pima Indians Diabetics Data set Description

Attribute Number	Attribute Description	Mean	Standard deviation
1	Number of times pregnant	3.8	3.4
2	Plasma glucose concentration	120.9	32.0
3	Diastolic blood pressure	69.1	19.4
4	Triceps skin fold thickness	20.5	16.0
5	2-Hour serum insulin	79.8	115.2
6	Body mass index	32.0	7.9
7	Diabetes pedigree function	0.5	0.3
8	Age	33.2	11.8

B. Discussion

The performance of the proposed method is used to the decision tree as a classifier using ten-fold cross validation is analyzed. Each dataset is divided into ten partitions, and each method is run ten times, using a different partition as test set each time, with the other nine as training set[9]. The results demonstrate that the proposed method is 74.87% (Four inputs and one output class attribute) accurate as compare to 73.83% (eight inputs and one output class attribute) in Pima Indiana Diabetes data set as shown in table II. The proposed method produces best classification accuracy for the chosen medical datasets compared to the without feature selection.

Table II: Classification performance comparison



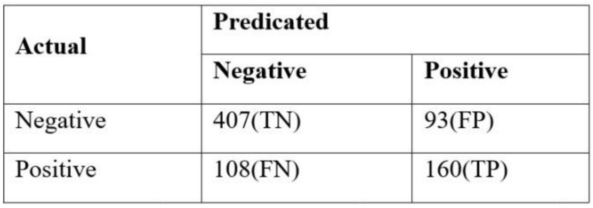
Classifier	No. of Features selected	Correctly Classified	Mis-classified	Classification rate (%)
C4.5 Without feature selection	8	567	93	73.83
C4.5 with feature selection	4	575	193	74.87

Confusion Matrix Analysis

Confusion matrix analysis was conducted to see the success rate of the study. The results of the confusion matrix without feature selection and with feature selection are shown in table III and table IV respectively.

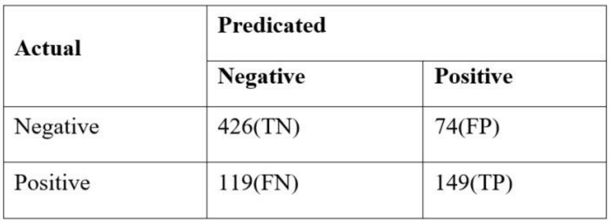
The entries of the confusion matrix have the following meaning:

Table III: Confusion matrix for without feature selection



Actual	Predicated	
	Negative	Positive
Negative	407(TN)	93(FP)
Positive	108(FN)	160(TP)

Table IV: Confusion matrix for with feature selection



Actual	Predicated	
	Negative	Positive
Negative	426(TN)	74(FP)
Positive	119(FN)	149(TP)

The true positive (TP) and true negatives (TN) are correct classifications. A false positive (FP) and true negative (FN) are incorrect classifications [5].

D. Classification Accuracy

In this study, the accuracy for the dataset is measured as the ratio of the number of correctly classified instances to the total number of examined instances. For the given confusion matrix, mathematically stated, this translates to:

$$Accuracy = \frac{TN+TP}{TN+FP+FN+TP} \qquad (7)$$

The result of the accuracy was shown in table 5.

E. Sensitivity and Specificity

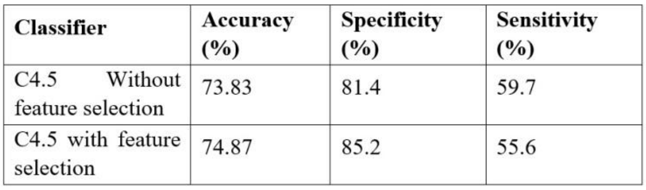
Sensitivity is a measure of accuracy of diagnosis of true cases of diabetic patient. Specificity is a measure of accuracy of diagnosis of false cases of diabetic patient.

$$Sensitivity = \frac{TP}{FN+TP} \qquad (7)$$

$$Specificity = \frac{TN}{TN+FP} \qquad (7)$$

The result of the sensitivity and specificity was shown in table V. From the result sensitivity is lower than specificity.

Table V: Comparison of performance



Classifier	Accuracy (%)	Specificity (%)	Sensitivity (%)
C4.5 Without feature selection	73.83	81.4	59.7
C4.5 with feature selection	74.87	85.2	55.6

VI. CONCLUSTION

In this proposed study, we used a ensemble two stage method to perform feature selection. The decision tree served as classifier of the given medical data set. The experimental result shows that the feature selection is used to very effective in reducing dimensionality, removing irrelevant data and to increasing learning accuracy, improving result comprehensibility. The information gain is used to evaluate the ranking of the individual features and the correlation based feature selection is used for the actual feature selection. C4.5 classification is used to classify the selected features and validate the method into confusion matrix analysis and obtain the best result. This has been applied to UCI data set namely Pima Indians diabetes data set. We could achieve successful feature dimensionality reduction and increased classification accuracy. The proposed ensemble feature selection method might possibly be applied in other areas in the future.

REFERENCES

1. J. Han and M. Kamber , "Data Mining: Concepts and Techniques" San Francisco, Morgan Kauffm ANN Publishers, 2001.
2. Liu. H, and Motoda. H, “Feature Selection For Knowledge Discovery And Data Mining”, Boston: Kluwer Academic Publishers, 1998.
3. RahmanMukras , NirmalieWiratunga , Robert Lothian , SutanuChakraborti and David Harper,“Information Gain Feature Selection for Ordinal Text Classification using Probability Re-distribution,” Proc. of IJCAI Text link Workshop, 2007
4. Screening for type-2 diabetes: Report of World Health Organization & International Diabetes Federation meeting. http://www.who.in/diabetes/publications/en/screeningmnc03.pdf.
5. An article on current status of Diabetes Mellitus in India by National Institute of Health US. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3920109
6. Huan Liu and Lei Yu, “Toward Integrating Feature Selection Algorithms For Classification And Clustering”, IEEE Transactions On Knowledge and Data Engineering, Vol. 17, No. 4, PP. 491 – 502, 2005.
7. Martín-Valdivia M R, Díaz-Galiano M.C, Montejo-Raez A, and Ureña-López L A, “Using information gain to improve multi-modal information retrieval systems,” Information Processing & Management, Vol. 44, No. 3, PP 1146-1158, May 2008.
8. Vidyavathi, B. M, and Ravikumar. C. N, “A Novel Hybrid Filter Feature Selection Method for Data Mining”, Ubiquitous Computing and Communication Journal, Volume 3, No. 3, 2008
9. K. Michalak and H. Kwasnicka, “Correlation Based Feature Selection Strategy in Classification Problems”, International Jornal of Applied Mathematics and Computer Science, Vo 16, No. 4, PP. 503–511, 2006.
10. AshaGowdaKaregowda , A. S. Manjunath&M.A.Jayaram, "Comparative Study of Attribute Selection Using Gain Ratio and Correlation based Feature Selection" International Journal Of Information Technology And Knowledge Management Vol. 2, No. 2, Pp. 271-277, 2010.
11. Swati Jadhav, Hongmei He, Karl Jenkins “ Information gain directed genetic algorithm wrapper feature selection for credit rating“ Applied Soft Computing, Volume 69, pp 541-553, 2018.

12. AsmaElmaizi, HasnaNhaila, ElkebirSarhrouni, Ahmed Hammouch, and ChafikNacir "A novel information gain based approach for classification and dimensionality reduction of hyperspectral images" Procedia Computer Science, Vol.148, PP 126-134, 2019.
13. P.Jaganathan and R.Kuppuchamy, "A threshold fuzzy entropy based feature selection for medical database classification" International Journal of Computers in Biology and Medicine, Vol. 43, No. 12, PP. 2222-2229, 2013.
14. JaganathanPalanichamy and KuppuchamyRamasamy, "An Improved Feature Selection Algorithm with Conditional Mutual Information for Classification Problems" IEEE International Conference on Human Computer Interactions (ICHCI), Chennai, India, 2013.
15. JaganathanPalanichamy and KuppuchamyRamasamy, "A Novel Feature Selection Algorithm with Supervised Mutual Information for Classification" International Journal of Artificial Intelligence Tools, Vol. 22, No. 4, PP. 1350027(14Pages), 2013.
16. Kohavi and G. H. John, "Wrapper for Feature Subset Selection," Artificial Intelligence, Vol. 97, No. 1-2, PP. 273-324, 1997.

AUTHORS PROFILE

R. Kuppuchamy, Department of MCA, PSNA College of Engineering and Technology, Dindigul, Tamilnadu, India.

T. Kamalavalli, Department of MCA, PSNA College of Engineering and Technology, Dindigul, Tamilnadu, India.

S. Vinothini, Department of MCA, PSNA College of Engineering and Technology, Dindigul, Tamilnadu, India.

N. Jayalakshmi, Department of MCA, PSNA College of Engineering and Technology, Dindigul, Tamilnadu, India.

N. Vallileka, Department of MCA, PSNA College of Engineering and Technology, Dindigul, Tamilnadu, India.