International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-12, October 2019

Density Based Feature Selection Method for Medical Datasets

Manonmani.M, Sarojini Balakrishnan

Abstract*: High dimensional data are found in the medical domain that needs to be processed for improved data analysis. In order to deal with the curse of dimensionality, feature selection process is employed in almost all data mining applications. In this research work, Density based Feature Selection (DFS) method that ranks the features by finding the Probability Density Function (PDF) of each feature is applied to medical datasets that suffer from the curse of dimensionality. The DFS method is a filter based approach that selects the most discriminatory features from the given feature set. The feature selection method evaluates the importance of the feature with regard to the target class using density function. The DFS method has major advantages over other methods, since it is based on the ranking method to select the most discriminatory features from the whole feature set. This research work finds the best feature subset that can be used in prediction and classification of medical datasets imbibed with high dimensionality. The DFS method based on PDF is applied on the three medical datasets namely Chronic Kidney Disease (CKD) dataset, Breast Cancer Wisconsin Dataset and Parkinsons Dataset. The proposed feature selection method evaluates the merit of each feature, assign weights to the feature and rank the features based on their feature density. The reduced feature subset is then validated by the application three classification algorithms namely Support Vector Machine (SVM), Gradient Boosting, and Convolutional Neural Network (CNN). The performance of the classification algorithms are evaluated based on the performance metrics Accuracy, Sensitivity and Specificity. Experimental results indicate that the performance of the classification algorithms SVM, Gradient Boosting, and CNN is improved after the feature selection process.

Keywords*: Curse of dimensionality, Filter method, Density based Feature Selection, Probability Density Function, SVM, Gradient Boosting, CNN.

I. INTRODUCTION

Medical data is growing in an unprecedented speed with the growth of internet and telecommunications. The data derived from the medical sector is very huge and varies in its interpretation and storage structure. The data stored in medical database consists of many parameters or features that describe the human body condition in addition to analysis of future health condition of the patients.

Revised Manuscript Received on October 30, 2019.
***** Correspondence Author
Manonmani M*, Research Scholar, Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India.
Dr. Sarojini Balakrishnan, Assistant Professor (SS), Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India.

This results in the formation of clinical datasets with high-dimensionality. These clinical data can be utilized to develop computational models which in turn help in carrying out classification. High-dimensional medical data increases the complexity of computational models and thus reduces the efficiency. Hence it becomes necessary to design analytical models that provide accurate results of with less number of features while retaining important information of the original data. Reduction of dimensionality leads to simplified form of the original data while the integrity of the data is also maintained [1]. There is a possibility of unconnected data entities resulting from large multi-dimensional data which may even contain noisy and redundant data. [2]. Hence there is a necessity to reduce the dimensions of the data to make viable use of huge medical data.

Dimensionality reduction is the process of obtaining a reduced representation of the actual data without any loss in information. High dimensional clinical data is difficult to process because there might be more features than the number of observations made [3]. Also, when there is high dimension in the dataset, every observation in the dataset will appear equidistant from all other observation and thus classification and clustering becomes harder. Other challenges in dimensionality reduction is finding a reduced feature subset without any loss of information in the original feature set and increasing the accuracy of classification with the reduced feature subset. Dimensionality reduction helps to reduce the time and space required for storing the huge amount of medical data since data in the medical sector accumulates at an unprecedented speed. Moreover, reducing the dimensionality helps in enhanced visualization of the data. In fact, less dimensional data results in low computational time and effort in contrast to high dimensional medical data. More importantly, reducing the dimensionality in the medical domain helps in identifying the features that are important in predicting and classifying the instances as prone to illness or not. Dimensionality reduction techniques helps in finding the most discriminatory features from the original dataset that helps in early prediction of chronic diseases that might otherwise be difficult with data with high dimension [4].

The dimensions of the data can be reduced using feature extraction and feature selection. Feature extraction method transforms high-dimensional data into a dataset of few dimensions by combining the features using feature extraction techniques to form a reduced feature subset. Feature selection techniques finds a reduced set of features to create a data model of low dimension by considering the entire feature set and selecting only the most discriminator feature subset that contribute required information for the task under consideration [5].

Methods of feature selection are preferred to methods of feature extraction for medical data as the original features and values are maintained and a reduced feature subset is selected without any loss of information [6]. Interpretability of data becomes easier with low computational cost, effort and time. The reduced feature subset produces a class distribution which is more correct to the class distribution of the original dataset. This ensures that the feature selection technique is more reliable and provides results that are easier to interpret and analyze.

Filter, wrapper, embedded and hybrid methods are the four types of feature selection methods [7]. Among the feature selection techniques, this research work aims to select the features based on a filter approach known as Density based Feature Selection (DFS). The DFS method finds the most discriminatory features by a ranking method that is based on the estimating the Probability Density Function (PDF) for each feature. This method finds the most important features from the original dataset by calculating the PDF for each feature in the original feature set.

The PDF method is advantageous when compared to other methods, as ranking of the features is based the densities of the features. This implies that, the feature with high probability is considered and the other feature with low probability and the final feature subset is derived based on the weight of the feature that contributes more information to the accuracy of classification results [1].

To systematically exploit the data represented in a high-dimensional medical datasets and to improve the performance of classification algorithms, we propose a filter method for feature selection technique using DFS. The following sections of the paper are organized as follows. Section 2 deals with literature survey, section 3 deals with methodology, section 4 deals with experimental results and discussion. The conclusion is given in section 5.

II. LITERATURE REVIEW

Mina Alibeigi et al [8] have proposed a novel unsupervised feature selection algorithm which finds the reduced feature subset based on the density of the original features. The feature values are scaled to an interval of [0,1]. After that, the probability density functions for each feature is evaluated using Kernel Density Estimation (KDE) method. Among similar features that exhibit similar probability density function value, one among the similar feature is which has the highest density is removed till the discriminatory feature subset is derived. The proposed method is evaluated against three bench mark datasets taken from the UCI repository. The performance of the proposed method is compared with CfsSubsetEval, ConsistencySubsetEval, SF with Entropy and the results reveal that this method better classification accuracy for the reduced feature subset.

V´ıtor Santos et al [9] have applied FS techniques to large datasets using Feature Ranking (FR) algorithms. The feature subset is derived using three measures for each class which are statistical between-class distance, interclass overlapping measure and an estimate of class impurity. The authors have evaluated their results against breast cancer X-ray images that is available in KDD Cup 2008 website. The algorithms for ranking the features used in this work include Information Gain, Gain Ratio, Symmetrical Uncertainty, and Chi-square. The proposed ensemble feature ranking algorithm has achieved better results in terms of Area Under Curve (AUC) for evaluating classifier performance. Experimental results indicated that for large datasets, Naïve Bayes achieved higher AUC and lower FPR.Maciej Kusy [10] has analyzed the difficulties encountered in feature selection and feature extraction for classifying medical datasets based on Probabilistic Neural Network (PNN). The author has analyzed three types of PNN models. Results reveal that feature selection applied with variable importance procedure has shown increased prediction ability for PNN1, PNN2 and PNN3 models for classification task. Each classification case has revealed decreased computational time needed to complete the classification task. Min Zhu et al [11] have introduced an Improved Niche Genetic Algorithm (INGA) which uses a self-adaptive niche-culling operation for dimensionality reduction. The population diversity is improved in the niche environment by the use of self-adaptive niche-culling operation in predicting 28-day death in sepsis patients. Experimental results reveal that the INGA algorithm has produced a feature subset consisting of 10 features that are important in predicting death in sepsis patients from the original feature set of 77 features. The model was evaluated against BP, SVM and RF classifiers and the results reveal that RF-INGA has achieved the highest accuracy of 92% for the reduced feature subset.

I. Beheshti et al [12] have presented a novel statistical approach for feature selection based on probability density function (PDF) for dealing with high-dimensional data. In order to develop an automatic computer-aided diagnosis (CAD) technique, the authors have explored statistical patterns extracted from structural MRI (sMRI) data on four systematic level. The proposed feature selection method was applied to MRI data for early identification of Alzheimer's disease (AD). The performance of SVM classifiers namely, SVM-linear and SVM-RBF were evaluated for the feature selection process based on the proposed PDF method and to the standard PLS-method. Experimental results reveal that PDF-based feature selection method has achieved better accuracy that the existing PLS-based feature selection method for early identification of AD.

Rattanawadee Panthong and Anongnart Srivihok [13] have presented a wrapper based feature selection process for reducing the dimensionality in medical datasets. The methods that have been used in this work include SFS, SBS and ensemble algorithms namely Bagging and AdaBoost based on the evaluation of the subset. The feature selection algorithms are evaluated against thirteen datasets taken from the UCI repository containing different attributes and dimension. Decision Tree and Naïve Bayes algorithm are applied to evaluate the results of the feature selection process. Experimental results reveal that SFS algorithm using Decision Tree has obtained an accuracy of 89.60% compared to the other methods.

S. Sasikala et al [14] have introduced a four-stage procedure known as Multi-Filtration Feature Selection (MFFS) for deriving optimal feature subset in medical data mining. In this method, a parameter called 'variance coverage' is adjusted in this method to achieve maximum classification accuracy.

The proposed method was evaluated against 22 medical datasets and the performance of four different classifiers, Naïve Bayes, SVM MLP and J48 was also evaluated. Experimental results reveal that the proposed MFFS method yields promising results on feature selection and classification accuracy for medical data mining.

Zhang B and Cao P [15] have proposed a feature selection algorithm based on redundant removal (FSBRR) to classify high dimensional medical data. In the proposed method, two redundant features are determined by finding the relationship between feature and class attribute and also finding the relationship between feature and feature. Furthermore, to quantify the redundancy condition, redundancy feature framework based on Mutual Information (MI) is defined and the redundant features are removed. The effectiveness of the feature selection algorithm is evaluated against eight high dimensional medical datasets and three classification algorithms namely RF, kNN and SVM. Experimental results indicate that FSBRR method can remove the redundant features effectively thereby improving the accuracy of classification algorithm.

III. METHODOLOGY

The proposed feature selection method is based on ranking the features based on Probability Density Function. Parametric and non-parametric approaches are the two methods for finding the density function based on the probability [1]. In the parametric approach, the mean and variance of the features are determined and are similar to KNN method of feature selection. On the other hand, non-parametric approach does not consider any previous knowledge about the densities of the features. It finds the density of the features straight from the instances hence they are used in most of the research work. The non-parametric method to estimate the PDF of a feature is represented in Eq. (1).

$$p(x) \cong k/ N * V \qquad (1)$$

where, p(x) is the value of estimated probability density function for instance x, V is the volume surrounding x, N is the total number of instances and k is the number of instances in V. Non-parametric method of calculating the density of each feature can be done in two ways. In the first approach, the value of k can be fixed and the value V can be determined. In the next approach, the value of V can be fixed and the value of k can be determined. The first approach is similar to the KNN method which suffers from local noise. The second approach is known as Kernel Density Estimation (KDE) in which the volume V is fixed and the value of k is determined. In this research work, the probability density functions for each of the features are computed according to KDE method, i.e., according to the formula represented in Eq. (1).In the proposed method, the feature weight of each feature is estimated and based on the feature weight, the features are ranked. The features that exhibit more feature weight have high probability of inclusion in the feature subset compared to features with low weightage. The methodology of the proposed feature selection methods is described in "Fig.1".

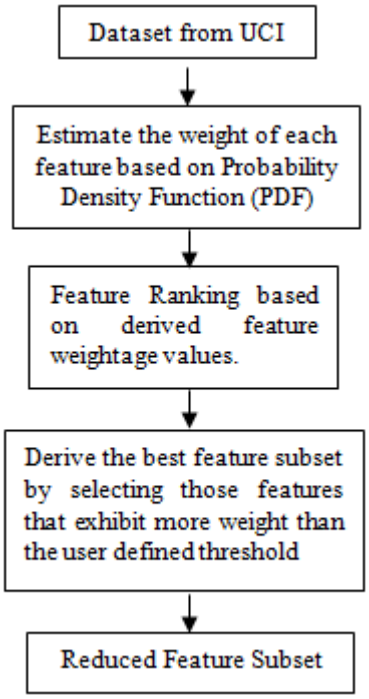

Fig.1. Proposed feature selection Process

As shown in "Fig. 1", the first step in our feature selection approach is estimating the weight of each feature based on the probability density function of each feature. Having estimated the weight of each feature based on PDF, the features are ranked based on high probability of inclusion in the feature subset because of more weight to low probability of inclusion in the feature subset. Those features which show less weight than the user defined threshold are removed and the other features form the final feature subset that contributes to better prediction and classification.

The derived feature subset is then evaluated based on the performance metrics of three classification algorithms namely SVM, Gradient Boosting and CNN. The performance metrics viz., Accuracy, Sensitivity and Specificity are derived from the confusion matrix.

IV. EXPERIMENTS AND RESULTS

Experiments were carried out on three chronic illness datasets namely, Chronic Kidney Disease (CKD) dataset, Breast Cancer Wisconsin dataset, and Parkinsons dataset that are taken from the UCI Machine Learning repository. The description of the three datasets is provided in Table I.

Table- I: Description of the datasets used for experimental analysis of the proposed research work.

S.No.	Dataset	No. of Attributes	No. of Instances
1.	CKD	25	400
2.	Breast Cancer Wisconsin	32	569
3.	Parkinson	23	197

The missing values in the datasets are filled using kNN method. After preprocessing the data using kNN method, the preprocessed datasets are given as input to the proposed feature selection method.

After feature selection, the number of features selected from the CKD dataset is 19 features from the original feature set, excluding the class attribute. Similarly, number of features selected for Breast Cancer Wisconsin Dataset and Parkinsons Dataset are 25 and 16 respectively excluding the class. The feature selection results and the corresponding ranking of the features for the three datasets are given in Table II.

Table- II. Feature selection results for the three datasets

S.No.	Dataset	No. of features selected	List of features selected based on ranking
1.	CKD	19	2, 10, 3 17, 18, 1, 4, 5, 6, 19, 7, 16, 9, 12, 13, 11, 8, 14, 15
2.	Breast Cancer Wisconsin	25	12, 13, 11, 27, 28, 26, 8, 7, 29, 6, 25, 9, 5, 18, 30, 17, 16, 19, 10, 15, 2, 14, 1, 22, 21
3.	Parkinsons	16	10, 21, 22, 17, 20, 18, 14, 9, 11, 12, 13, 5, 7, 4, 15, 8

The feature reduction achieved for CKD, Breast Cancer Wisconsin Dataset, and Parkinsons Dataset are 24%, 21.8%, and 30.4% respectively. The feature subset derived after the proposed feature selection algorithm consists of the most discriminatory features that enhances the accuracy of prediction of chronic diseases without any loss in the original information.The reduced feature subset is evaluated against three classification algorithms SVM, CNN, and Gradient Boosting. The performance metrics namely accuracy, sensitivity and specificity are analyzed for the three classifier for the whole feature set and reduced feature subset based on 10-fold cross- validation. The results of evaluation of the performance metrics of the three classification algorithms for CKD, Breast Cancer Wisconsin Dataset and Parkinsons Dataset are depicted in Table III, IV and V respectively.

Table- III. Comparison of the accuracy of three classifiers before and after feature selection for CKD dataset.

Classification Algorithm	Before Feature Selection (%)			After Feature Selection (%)		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
SVM	90.25	92.8	86	92	94.4	88
Gradient Boosting	92.25	91.6	93	95	96	94.6
CNN	92.75	92	94	95.5	94.4	97.3

Table- IV. Comparison of the accuracy of three classifiers before and after feature selection for Breast Cancer Wisconsin Dataset.

Classification Algorithm	Before Feature Selection (%)			After Feature Selection (%)		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
SVM	87.3	84.4	89	90.5	90.6	90.4
Gradient Boosting	91.04	89.6	91.8	94.2	97	92.4
CNN	92.7	92.4	93	95.9	96.7	95.5

Table- V. Comparison of the accuracy of three classifiers before and after feature selection for Parkinsons Dataset.

Classification Algorithm	Before Feature Selection (%)			After Feature Selection (%)		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
SVM	90.2	89.5	90.4	92.3	95.8	91.1
Gradient Boosting	92.3	89.5	93.2	95	91.6	97.2
CNN	94.3	95.8	93.8	95.9	100	94.5

It is evident from table 3,4 and 5 that the accuracy of the three classifiers has improved after feature selection for the undertaken three medical datasets. For the CKD dataset, CNN has achieved the highest accuracy of 95.5 % after feature selection as against an accuracy of 92.75% before feature selection. From table 4, it is evident that the CNN algorithm has shown the highest accuracy of 95.9% after feature selection for both Breast Cancer Wisconsin Dataset and Parkinsons Dataset compared to the accuracy of SVM and Gradient Boosting classification algorithms. Overall, the accuracy, sensitivity and specificity of all the three classification algorithm has increased after feature selection when comparing the performance metrics of the classifiers before feature selection. Experimental results indicate that the proposed feature selection algorithm can efficiently select the most discriminatory features that are important in prediction and classification. The sensitivity and specificity values are significant in any medical diagnosis. High sensitivity helps in accurately diagnosing a diseased individual as having the disease and high specificity helps in accurately diagnosing a non-diseased individual as not having the disease. Experimental results indicate that the classification algorithms have attained high sensitivity and high specificity which in turn helps in accurate diagnosis of patients suffering from chronic illness.

V. CONCLUSION

This research work aims to select the best feature subset with informative and significant features of medical dataset that aid in early identification of chronic illness. The main objective of the proposed feature selection approach is to apply DFS using PDF which ranks the features based on their weight. This ranking method helps to identify the most important features from most significant to least significant that contribute to prediction and classification of the medical datasets. The feature reduction achieved for CKD, Breast Cancer Wisconsin Dataset, and Parkinsons Dataset after applying the proposed method are 24%, 21.8%, and 30.4% respectively. Thus there is a reduction in the dimensionality of the original datasets that in turn help in increasing the accuracy of prediction and classification. The performance metrics of the three classification algorithms, SVM, Gradient Boosting and CNN has shown improved results in accuracy, sensitivity and specificity when comparing the results before and after feature selection. CNN algorithm has shown the highest accuracy for CKD, Breast Cancer Wisconsin Dataset and Parkinsons Dataset. Thus, experimental results indicate that the proposed DFS method based on filter approach addresses the problem of high dimensionality by reducing the number of features from the original dataset. The reduced feature subset is the most discriminatory feature subset that that help in early diagnosis of diseases for patients who are suffering from chronic illness for a long term with high predictive accuracy.

REFERENCES

1. Mina Alibeigi et al. (2009). “Unsupervised Feature Selection Using Feature Density Functions.” World Academy of Science, Engineering and Technology, International Journal of Computer and Information Engineering, 3(3): 847- 852.

2. Houari er al. (2016). “Dimensionality Reduction in Data Mining: A Copula Approach.” Expert Systems with Applications. 64. 247-260. 10.1016/j.eswa.2016.07.041.
3. Priyanka Jindal and Dharmender Kumar. (2017). “A Review on Dimensionality Reduction Techniques.” International Journal of Computer Applications. 173(2): 42-46.
4. Divya Jain and Vijendra Singh. (2018). “An Efficient Hybrid Selection model for Dimensionality Reduction.” Procedia Computer Science. 132: 333-341.
5. Robert E. Colgan et al. (2013). “Analysis of medical data using dimensionality reduction techniques.” 10.13140/2.1.2270.1762.
6. Mary Walowe Mwadulo. (2016). “A Review on Feature Selection Methods For Classification Tasks.” International Journal of Computer Applications Technology and Research. 5 (6): 395 – 402.
7. Elhoseny, M. et al. (2019). “Intelligent Diagnostic Prediction and Classification System for Chronic Kidney Disease.” Scientific reports, 9(1): 9583.
8. Mina Alibeigi et al. (2011). “Unsupervised Feature Selection Based on the Distribution of Features Attributed to Imbalanced Data Sets.” International Journal of Artificial Intelligence and Expert Systems, 2(1): 14-22.
9. V´ıtor Santos et al. (2014). “Ensemble feature ranking applied to medical data.” Procedia Technology, 17. 223 – 230.
10. Maciej Kusy. (2015). “Dimensionality Reduction for Probabilistic Neural Network in Medical Data Classification Problems.” International Journal of Electronics and Telecommunications, 61(3): 289-300.
11. Min Zhu et al. (2015). “Dimensionality Reduction in Complex Medical Data: Improved Self-Adaptive Niche Genetic Algorithm.” Computational and Mathematical Methods in Medicine. Hindawi Publishing Corporation, Volume 2015, Article ID 794586, 12 pages.
12. I. Beheshti et al. (2015). “Probability distribution function-based classification of structural MRI for the detection of Alzheimer’s disease.” Computers in Biology and Medicine, 64. 208-216.
13. Rattanawadee Panthong and Anongnart Srivihok. (2015). “Wrapper Feature Subset Selection for Dimension Reduction Based on Ensemble Learning Algorithm.” Procedia Computer Science, 72. 162-169.
14. S. Sasikala et al. (2016). “Multi Filtration Feature Selection (MFFS) to improve discriminatory ability in clinical data set.” Applied Computing and Informatics, 12. 117-127.
15. Zhang B and Cao P. (2019). “Classification of high dimensional biomedical data based on feature selection using redundant removal.” PLoS ONE, 14 (4): 1-19.

AUTHORS PROFILE

Manonmani M has completed her Master degree in M.Sc. Software Engineering (Applied Science) at Amrita Institute of Technology and Science, Coimbatore in 2005 after which she has done her B.Ed in Computer Science in 2013. She had been working as a Computer Science teacher in Chinmaya Vidhyalaya CBSE School from 2013 to 2016 and now she is undertaking doctoral research at Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore. Her research area in computer science includes data mining and her current research work is focused on Feature Selection Methods for Medical Datasets. She has published papers that analyze the various feature selection methods during her course of Ph.D in Computer Science from 2017 till date.

Dr.Sarojini Balakrishnan, received her Post-Graduate Degree, M.C.A from Alagappa University, Karaikudi 1993 and research degrees, M.Phil and Ph.D in Computer Science from Mother Teresa Women's University, Kodaikanal in 2001 and 2010 respectively. She has been teaching Post Graduate Courses in Computer Science, Applications and Engineering for more than 23 years. She has published research papers in reputed journals and International Conferences. She has completed a research project funded by UGC in the domain of medical informatics. Her research areas include Medical Informatics, Web mining, and Information Security.