Human Machine Interaction in the Digital Era – Prof. J. Dhilipan et al. (eds)

An Effective Analysis of Feature Selection Methods for High Dimensional Data—A Survey

29

M. Kalaivani[1]
Research Scholar, Department of Computer Science, Vels Institute of Science, Technology & Advanced Studies (VISTAS) Chennai, Tamil Nadu

K. Abirami[2]
Assistant Professor, Vels Institute of Science, Technology & Advanced Studies.

K. Dharmarajan[3]
Associate Professor, Vels Institute of Science, Technology & Advanced Studies

Abstract In the field of genomics, High dimensional data is primarily utilized to detect the essential genes that play a vital role in determining the disease diagnosis using expression levels. The number of features in the High dimensional dataset is extremely very high when compared to the samples present in the dataset. The features in the dataset are usually given as input to a learning algorithm for classification of diseases. However, in the High dimensional data most features are redundant and irrelevant or noisy which will decrease the learning accuracy. To solve these problems, Feature selection technique is employed a significant role. Feature selection is one of the important preprocessing step for prediction and classification of disease. It aims to find informative features, selecting a small subset of relevant features from the original set of features by removing the redundant and irrelevant features from the dataset which can reduce the computational time and improving the classification accuracy. Due to increase in dimensionality of High dimensional data imposes a significant challenge to many existing feature selection methods in terms of prediction and accuracy of the model. This research work analyses about the use of various Feature selection methods that can select prominent attributes from the High dimensional dataset for classification of diseases.

Keywords Classification, Feature selection, High dimensional data, Prediction, Preprocessing

1. Introduction

High dimensional data occur when the feature count (P) higher than the number of samples (N), represented as P>N. and contains a large number of redundant and irrelevant attributes. To select the significant features from the data is a major challenging task. The main application of High dimensional data is in the field of Genomics such as Clinical decision support system for classification of the type of disease and facilitates the clinical analysis of those genes that are responsible for a particular disease especially Cancer Classification. In particular Deoxyribonucleic Acid (DNA) Microarray is an effective tool that supports to monitor the level of gene expression in an organism. The microarray dataset are represented as images that are converted into two dimensional matrices where the rows represent genes, and the columns represent samples (Brazma

[1]mkvanimca@gmail.com, [2]abiramidharmarajan@gmail.com, [3]dharmak07@gmail.com

DOI: 10.1201/9781003428466-29

and Vilo, 2001). The total number of samples in the dataset is significantly lesser when compared to the number of features (genes) commonly known as "curse of dimensionality" (Hamim et al., 2021). Genes are now taken into consideration with the development of biomedical research for the classification of a disease, particularly cancer, for prognosis or diagnosis of disease at an early stage. It is necessary to reduce the dimensions prior to classification because a high number of features may lead to increase in computational time and memory consumption. In an attempt to overcome these issues, Feature selection algorithms have been applied prior to classification. This research work analyses the different Feature Selection methods applied in High dimensional data used to remove redundant and irrelevant features for better classification of the model.

The rest of this paper is organized as follows. Section II discusses about the Role of Feature selection methods. Section III explores the different Feature Selection methods. Analysis of Feature Selection methods applied for High dimensional data by different researchers is discussed in section IV. Finally, Section V concludes the suitability of Feature Selection methods in High dimensional data.

2. Role of Feature Selection Methods

Dimensionality reduction is one of the primary pre-processing method in machine learning to determine the significant attributes in the data and can be broadly classified into two ways (Sahu et al., 2018). Feature selection and Feature extraction. Feature selection, selects the subset of features from the original set whereas Feature extraction is to generate new features from the existing set. Feature selection techniques can be broadly classified into three important methods. Supervised, Unsupervised and Semi Supervised method. A Framework for feature selection methods are shown in Fig. 29.1. In Supervised Feature Method the availability of class label information allows feature selection algorithms to efficiently select discriminative and relevant features to distinguish features from different classes.

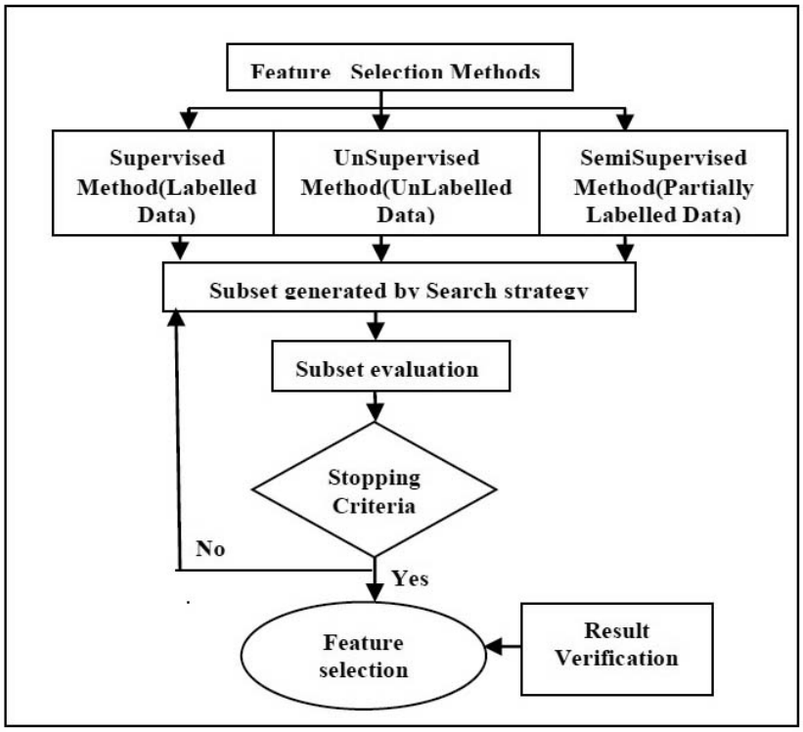

Fig. 29.1 Framework for feature selection methods

Source: Subset evaluation—Made by Author

In Unsupervised Feature Selection method, new features are created by combining the original features. The absence of class label information allows selecting discriminative features and produces high performance without prior knowledge is required and the limitation is to neglect the feasible correlation between features. Semi supervised method is the combination of Supervised and Unsupervised methods. It learns the data from the combination of both labeled and unlabeled data to evaluate

the features (Sechidis & Brown 2018). It selects features by utilizing unlabeled data when there is limited number of labeled data. According to the subset generated by search strategy, the commonly used search methods are Sequential, Exponential and Random Search. In sequential search, feature selection methods can be used Sequential Backward Elimination, Sequential Forward Selection and Bidirectional search (Cai et al., 2018). Feature evaluation criteria is mainly categorized into distance, consistency, dependency, correlation and information. Stopping Criteria indicates the end of the process. Finally, evaluate the accuracy of the method.

3. Classification of Feature Selection Methods

In machine learning, Feature selection method plays an important role in the field of biomedical data analysis particularly cancer classification with an increase in the number of data dimensions. Feature selection methods are classified into the following categories: Filter ,Wrapper ,Embedded, Hybrid and Ensemble method. Filter method consists of selecting attributes based on inherent characteristics of the data without using classifier and generally involves non-iterative computation. Each attribute is analyzed individually by employing its basic statistical properties and can be classified as Univariate and Multivariate filters. Univariate filters evaluate the individual feature whereas multivariate filters estimate the whole feature subset. Wrapper approaches are using a specific learning algorithm to select subsets of features (Panthong & Srivihok,2015) which is more accurate than filter methods. The number of operations required to obtain the feature subset and run each time a new learning method is applied. Computationally difficult method when compared to filter method. In Embedded methods, the feature selection algorithm is incorporated into the learning process (Jovic et al., 2015) and takes the advantages of its own attribute selection process and performs Feature selection and Classification at the same time. The hybrid method integrates the benefits of Filter, Wrapper and Embedded methods. Filter method is used as an initial step to reduce the attribute dimensions particularly when dimensions in the feature space are high and remove irrelevant and similar features with less computation cost. In the next step of process, Wrapper or Embedded method is applied to the selected features and evaluated to determine the best feature subset. An ensemble method is based on two steps. Initially, two or more component learners are trained either sequential method or parallel method and finally aggregating their predictions based on algorithms.

4. Analyzing Feature Selection Methods

High dimensional dataset normally consists of a huge number of features, but all the features are not contributing to the goodness of classification. Due to redundant and noisy features of High-Dimensional dataset, Feature selection is one of the essential pre-processing step which is used to identify the significant attributes present in the dataset and produce higher performance accuracy with less processing time. Various Feature Selection methods are used to select the significant attributes in the High dimensional data. This section discusses about different Feature selection methods are applied by many researchers in High dimensional dataset. D.M. Deepak Raj et al. (2020) proposed a new feature weighting algorithm called boundary margin relief (BMR) to predict the feature weights through the metric of local hyper plane to determine the set of the closest hits and misses and calculate the features weight by increasing the hyper plane margin. The proposed method identified non redundant features and produced higher accuracy value.

Dewi Pramudi Ismi et al. (2016) developed a model using k-means partitional cluster method. The model solves the curse of dimensionality by partitioning the data into clusters and removed similar and irrelevant features. Abdulrauf Garba Sharifai et al. (2020) have designed the multiple filter method with correlation based redundancy CBRMFA. In this method, Mutual Information, Symmetric Uncertainty and Euclidean distance was applied to select the relevant features from each filter method and aggregating the features using the union operator. The top N ranking features are selected to form a new set of features based on the threshold value. Correlation between features was computed and finally sequential forward search was applied to select the optimal set of features. Manikandan et al., (2017) proposed a wrapper based feature selection approach using symmetrical uncertainty method to calculate the feature weight based on Entropy and information gain values. The selected features are arranged in descending order and given as input to the classifier and the accuracy of each attribute is compared with the previous attribute and finally selected the higher accuracy features. Liuzhi Yin et al. (2013) discussed class imbalanced data using feature selection and developed two approaches. In the first method, large classes are partitioned into pseudo-subclasses with equal sizes and find the efficiency of features with the partitioned data to reduce the biased value of the imbalanced class. In the next step, Hellinger distance was applied to select the essential features. Compared the results with Pearson Correlation, Mutual information gain and Fisher Score methods. Yongbin Zhu et al. (2022) have developed a method referred as HFIA based on AI Optimization technique. Fisher Score and clonal section algorithm are applied to explore the feature subset. The result showed that the proposed method produced 91.67% of accuracy rate.

Aiguo Wang et al. (2017) have described a novel method to combine the MB method into WBFS. The developed method, eliminated the redundant and noisy attributes using a classifier which helps to speed up the process and select the significant features. Jamshid pirgazi et al. (2019) have designed a new metaheuristic attribute selection technique and implemented in two phases. In the first phase, reliefF was employed to select the attributes based on the rank and SFL,IW Subset Selection relevance algorithm was applied in the second stage and selected effective features in the dataset to increase the rate of accuracy. Amirreza Rouhi et al. (2017) have presented an algorithm using hybrid and ensemble method. FCBF filter method was applied to reduce the features space and different metaheuristic optimization algorithms are applied independently to the selected attributes and finally combined the important attributes. Chaonan Shen et al. (2021)proposed a model that combined MLP and IGWO. Network was trained with Lasso method to evaluate the hidden layer neurons using the weights. In the next step, IGWO was applied to reduce the size of the features space and also selected an optimal number of features. The results showed that 93.05% of accuracy for eleven tumor dataset. Rania Saidi et al. (2019) have proposed a hybrid selection technique using Pearson Correlation Coefficient and Genetic algorithm. In the initial stage, GA was applied for the Ionosphere dataset using Random Forest as evaluation function. Pearson Correlation Coefficient was applied in the next stage and combined the resulting attribute subsets to increase the classification accuracy. Mohammed Loey et al., (2020)have designed an algorithm consisting of three processing stages to reduce the feature space. At the initial stage, Information Gain (IG) was applied in the dataset and selected the most relevant features based on the weight value. To optimize the features using GWO ptimization and SVM classifier was employed to evaluate the features. Abdulrauf Garba Sharifai et al. (2021) have presented multiple filter approach with hybrid Grasshopper optimization and Simulated Annealing (HGOASA) technique. According to the threshold value, the top ranked attributes from each filter are selected. In the next stage, optimization approach was applied based on global search method to select the prominent set of features. Annavarapu & Dara ,2021 have described a cluster based feature selection method. Partitional kmeans cluster analysis with snr rank method was applied to reduce the feature dimension space of the dataset. CLA combined with ant colony optimization (CLACO) was applied on the selected attributes to obtain the significant feature subset and analysed using several classifiers.

5. Conclusion

Feature selection is an essential pre-processing step for the discriminate analysis of very High-Dimensional data because it contains a large number of attributes and limited sample sizes. Due to curse of dimensionality it is vital to implement Feature selection algorithm to reduce the high dimensional feature space to low dimensional feature space and select the most significant features. The primary goal of the method is to enhance the high accuracy of the model by the elimination of redundant, irrelevant and noisy features. In this process, Filter methods are very fast and simple, whereas each feature is measured separately and thus it does not take into account of the dependencies among the features. The Wrapper methods using exhaustive search to generate optimal solutions, whereas its limitation is that a risk of over fitting. Embedded methods are lower risk of over fitting but faster running time as compared to a wrapper method. Hybrid feature selection method can combine the benefits of Filter, Wrapper and Embedded methods which can significantly improve the performance by increasing the rate of accuracy and decreasing the execution time for classification of diseases. Ensemble methods are less prone to over fitting and certain regularization are more suitable to specific types of learners. The future work determines to develop an effective feature selection algorithm for identifying significant features as well as remove redundant and irrelevant attributes present in the High dimensional data, analysing the merits and demerits of the surveyed methodology.

References

1. Brazma, A., & Vilo, J. (2001). Gene expression data analysis. *Microbes and Infection*, *3*(10), 823–829.
2. Hamim, M., El Mouden, I., Ouzir, M., Moutachaouik, H., & Hain, M. (2021). A novel dimensionality reduction approach to improve microarray data classification. *IIUM Engineering Journal*, *22*(1), 1–22.
3. Sahu, B., Dehuri, S., & Jagadev, A. (2018). A study on the relevance of feature selection methods in microarray data. *The Open Bioinformatics Journal*, *11*(1).
4. Sechidis, K., & Brown, G. (2018). Simple strategies for semi-supervised feature selection. *Machine Learning*, *107*(2), 357–395.
5. Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, *300*, 70–79.
6. Panthong, R., & Srivihok, A. (2015). Wrapper feature subset selection for dimension reduction based on ensemble learning algorithm. *Procedia Computer Science*, *72*, 162–169.
7. Jovic, A., Brkic, K., & Bogunovic, N. (2015). A review of feature selection methods with applications. In 2015 38th international convention on information and communication technology, electronics and microelectronics (pp. 1200–1205). IEEE.

8. Raj, D. D., & Mohanasundaram, R. (2020). An efficient filter-based feature selection model to identify significant features from high-dimensional microarray data. *Arabian Journal for Science and Engineering*, *45*, 2619–2630.
9. Ismi, D. P., Panchoo, S., & Murinto, M. (2016). K-means clustering based filter feature selection on high dimensional data. *International Journal of Advances in Intelligent Informatics*, *2*(1), 38–45.
10. Sharifai, A. G., & Zainol, Z. (2020). The correlation-based redundancy multiple-filter approach for gene selection. *International Journal of Data Mining and Bioinformatics*, *23*(1), 62–78.
11. Manikandan, G., Susi, E., & Abirami, S. (2017). Feature selection on high dimensional data using wrapper based subset selection. In *2017 Second International Conference on Recent Trends and Challenges in Computational Models* (pp. 320–325). IEEE.
12. Yin, L., Ge, Y., Xiao, K., Wang, X., & Quan, X. (2013). Feature selection for high-dimensional imbalanced data. *Neurocomputing*, *105*, 3–11.
13. Zhu, Y., Li, T., & Li, W. (2022). An Efficient Hybrid Feature Selection Method Using the Artificial Immune Algorithm for High-Dimensional Data. *Computational Intelligence and Neuroscience*, *2022*.
14. Wang, A., An, N., Yang, J., Chen, G., Li, L., & Alterovitz, G. (2017). Wrapper-based gene selection with Markov blanket. *Computers in biology and medicine*, *81*, 11–23.
15. Pirgazi, J., Alimoradi, M., Esmaeili Abharian, T., & Olyaee, M. H. (2019). An Efficient hybrid filter-wrapper metaheuristic-based gene selection method for high dimensional datasets. Scientific reports, 9(1), 18580.
16. Rouhi, A., & Nezamabadi-pour, H. (2017). A hybrid feature selection approach based on ensemble method for high-dimensional data. In *2017 2nd conference on swarm intelligence and evolutionary computation (CSIEC)* (pp. 16–20). IEEE.
17. Shen, C., & Zhang, K. (2021). Two-stage improved Grey Wolf optimization algorithm for feature selection on high-dimensional data. *Complex & Intelligent Systems*, 1–21.
18. Saidi, R., Bouaguel, W., & Essoussi, N. (2019). Hybrid feature selection method based on the genetic algorithm. *Machine learning paradigms: theory and application*, 3–24.
19. Loey, M., Wajeeh Jasim, M., El-Bakry, H. M., Hamed N. Taha, M., & Khalifa, N. E. M. (2020). Breast and colon cancer classification from gene expression profiles using data mining techniques. *Symmetry*, *12*(3), 408.
20. Sharifai, A. G., & Zainol, Z. B. (2021). Multiple filter-based rankers to guide hybrid grasshopper optimization algorithm and simulated annealing for feature selection with high dimensional multi-class imbalanced datasets. *IEEE Access*, *9*, 74127–74142.
21. Annavarapu, C. S. R., & Dara, S. (2021). Clustering-based hybrid feature selection approach for high dimensional microarray data. *Chemometrics and Intelligent Laboratory Systems*, *213*, 104305.