A Comprehensive Review of Filter and Wrapper Methods for Feature Selection

M. Anjali[*1] & Dr. P. Verma[2]
[*1]Assistant Professor, Department of Computer Science, Vellalar College for Women, India
[*]Associate Professor, Department of Computer Science, Gobi Arts and Science College, India

ABSTRACT

In health care, automatic disease diagnosis is a precious tool because of limited observation of the expert and uncertainties in medical knowledge. Progresses in medical information technology have enabled healthcare industries to automatically collect huge quantity of data through clinical laboratory examinations. To explore these data, the past few years have envisaged the use of Computer Aided Diagnosis (CAD) systems in many hospitals and screening sites. Machine learning techniques are gradually introduced to construct the CAD systems owing to its well-built capability of extracting complex relationships in the biomedical data. Data mining is a pioneering and attractive research area due to its vast application areas and task primitives. Data classification is one of the most important tasks in data mining. Feature Selection is also known as Attribute selection which selects subset of features from original set by removing the irrelevant and redundant features. This paper focus on the literature review of two feature selection techniques namely, filter approach and wrapper approach.

KEYWORDS: Data mining, Classification, Feature selection, Filter approach, Wrapper approach.

I. INTRODUCTION

Data mining plays a vital role in medical field for disease diagnosis. It offers lot of classification techniques to predict the disease accuracy [1]. The computer based analysis system indicates the mechanized medical diagnosis system. This mechanized diagnosis system support the medical practitioner to make good decision in treatment and disease [2].Classification maps data into predefined groups or classes. It is frequently referred to as supervised learning because the classes are determined before examining the data [3]. Usage of CAD systems for diagnosis provides multiple advantages [4]. Feature Selection plays an essential role in the process of data mining. It is necessary to deal with the excessive number of features, which can become computational burden on the learning algorithms as well as various feature extraction techniques. It is also essential, even when computational resources are not limited, since it improves the accuracy of the machine learning tasks. In general, features are characterized as relevant, irrelevant and redundant. The advantages of feature selection are

- Improving the data quality.
- Increasing the accuracy of the resulting model.
- It reduces the dimensionality of the feature space, to limit storage requirements and increase algorithm speed.
- It removes the redundant, irrelevant or noisy data.
- Performance improvement, to gain in predictive accuracy.
- Feature set reduction, to save resources in the next round of data collection or during utilization.
- Data understanding, to gain knowledge about the process that generated the data or simply visualizes the data.

II. FEATURE SELECTION TECHNIQUES

The Filter Approach

Filter approach or Filter method shown in Fig 1. This method selects the feature without depending upon the type of classifier used. The advantage of this method is that, it is simple and independent of the type of classifier used so feature selection need to be done only once and drawback of this method is that it ignores the interaction with the classifier, ignores the feature dependencies, and lastly each feature considered separately[5][6]

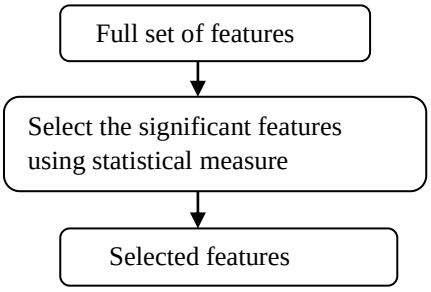

Fig 1 Filter Approach

The Wrapper Approach

Wrapper approach or Wrapper method is shown in Fig 2. In this method the feature is dependent upon the classifier used, i.e. it uses the result of the classifier to determine the goodness of the given feature or attribute. The advantage of this method is that it removes the drawback of the filter method, i.e. it includes the interaction with the classifier and also takes the feature dependencies and drawback of this method is that it is slower than the filter method because it takes the dependencies also. The quality of the feature is directly measured by the performance of the classifier [5][6].

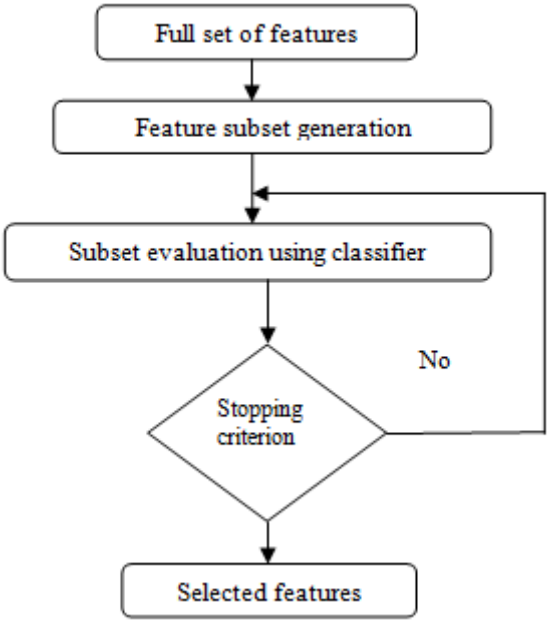

Fig 2.Wrapper Approach

III. LITERATURE REVIEW

The feature selection is to select a subset of variables from the input data which can efficiently describe the input data while reducing effects from noise or relevant variables and still provide good prediction results[7][8]. The filter-based approaches are independent of the supervised learning algorithm therefore offer more generality and they are computationally cheaper than the wrapper and embedded approaches. For processing the high-dimensional data, the filter methods are suitable rather than the wrapper and embedded methods. Wrapper-based approach generates the feature subsets using any one of the searching techniques and evaluates these subsets using the supervised learning algorithm in terms of classification error or accuracy [9]. The wrapper method seems to be a “brute force” method. Wrapper methods perform a search in the space of feature subsets such as classification performances on a cross-validation of the training set which provided better results than filter methods. But wrapper approaches increase the computational cost [10].

Donghai Guan, et al. [11] reviewed and compared two techniques of integrating feature selection and ensemble learning, (1) Feature selection for ensemble learning (ENfs) and (2) Ensemble learning for feature selection (FSen). This approach obtained predictive accuracy superior to conventional feature selection methods for supervised learning. Moreover, its most prominent advantage is the ability to handle stability issue that is usually poor in existing feature selection methods.

S´anchez-Maro˜no, et al. [12] proposed a new wrapper method, called Incremental ANOVA and Functional Networks-Feature Selection (IAFN-FS) for dealing with multiclass problems based in classical algorithms, such as C4.5 and Naïve Bayes. The multiple binary classifiers approach obtained better results in accuracy, although it has the drawback of selecting a higher number of features.

Akin Ozcift and Arif Gulten [13] used a rotation forest ensemble decision tree algorithm wrapped with best first search strategy. The wrapper uses forward selection to choose the optimum subset on the Erythemato-Squamous diseases dataset. The discrimination ability of selected features is evaluated using several machine learning algorithms and the diversity of the training data using the bagging algorithm.

Yvan Saeys, et al. [10] proposed the method of ensemble feature selection techniques for high dimension data which can be used to yield more robust feature selection techniques. As well Sangkyun Lee, et.al [14] presented a method of an extension to RapidMiner which delivers implementations of algorithms which is well suited for very high-dimensional data. These experiments were conducted on a microRNA-expression dataset.

Yu , et al.[15] propose a new framework of feature selection which avoids implicitly handling feature redundancy and turns to efficient elimination of redundant features via explicitly handling feature redundancy. Relevance definitions divide features into irrelevant features, weakly relevant features and strongly relevant features; redundancy definition divides weakly relevant features into redundant and non redundant ones. Thus produces the final subset. Its advantage is decoupling relevance and redundancy analysis and allows a both efficient and effective way in finding a subset that approximates an efficient subset. It uses C & F correlations for redundancy analysis and only C-correlation for relevance analysis.

Yu, et al. correlation based filter approach is making use of symmetric uncertainty method. Symmetric uncertainty measure how much a feature is related to another feature. This involves two steps. The first step shows how to decide whether a feature is relevant to the class or not; and second how to decide whether such a relevant feature is redundant or not when considering it with other relevant features. The solution to the first question can be using a user- defined threshold SU value, as the method used by many other feature weighting algorithms (e.g., Relief). The answer to the second question is more complicated because it may involve analysis of pair wise correlations between all features (named F correlation), which results in a time complexity of O(N2) associated with the number of features N for most existing algorithms. To solve this problem, FCBF algorithm is proposed. FCBF means Fast Correlation-Based Filter Solution [16]. This algorithm involves two steps. First step is select relevant features and arrange them in descending order according to the correlation value. Second step is remove redundant features and only keeps predominant ones. The information theory and linear correlation is used to propose a new algorithm for best feature selection with less time complexity.

Butterworth, et al.[17] introduces an algorithm for feature selection that clusters attributes using a special metric of Barthelemy-Montjardet distance and then uses a hierarchical clustering for feature selection. Hierarchical algorithm generates clusters that are placed in cluster tree which is commonly known as dendrogram. The dendrogram of resulting cluster hierarchy to choose the most representative attributes. Clustering's are obtained by extracting those clusters that are situated at given height in this tree.

Biesiada, et al.[18] introduces an algorithm for filtering information based on the Pearson $\chi 2$ test approach has been implemented and tested on feature selection. This is useful for high dimensional data where no sample set is large. This test is frequently used in biomedical data analysis and used only for nominal (discretized) features. This algorithm has only one parameter, statistical confidence level that two distributions are identical. Empirical comparisons with four other features selection algorithms (FCBF, CorrSF, ReliefF and ConnSF) are done to find quality of feature selected. This algorithm works fine with the linear SVM classifier. Similar to other correlation-based filters, PRBF (Pearson's Redundancy Based Filter) algorithm is essential and much lower than ReliefF.

Das [19], describes the advantages and disadvantages of filter and wrapper methods for feature selection and proposes a new hybrid algorithm that uses boosting and incorporates into a fast filter method with some of the features of wrapper methods for feature selection. Empirical results are reported on six real-world datasets from the UCI repository, showing that hybrid algorithm is competitive with wrapper methods while being much faster, and scales well to datasets with thousands of features.

Arauzo-Azofra, et al.[20] presents Relief is a well known and good feature set estimator. Feature selection methods try to find a subset of the available features to improve the application of a learning algorithm. Feature set estimators evaluate features individually. On artificial datasets, the proposed feature set measure based on relief can be better than the wrapper approach to guide a common feature selection search process. In this paper, the method is compared with a consistency measure and the highly reputed wrapper approach. The main disadvantage of this system is, it measure low accuracy of the search process.

Roslina, et al. makes use of Support Vector Machines to predict hepatitis and applied wrapper based feature selection method to identify relevant features before classification. Combining wrapper based methods and Support vector machines produced good classification results [21].

Sartakhti, et al. also presented a novel machine learning method using hybridized Support Vector machine and simulated annealing to predict hepatitis. They obtained high classification accuracy rates [22].

Harb, et al. proposed the filter and wrapper approaches with Particle Swarm Optimization (PSO) as a feature selection method for medical data. They applied different classifiers to the datasets and compared the performance of the proposed methods with another feature selection algorithm based on genetic approach. Their results illustrated that the proposed model shows the best classification accuracy among the others [23].

Huang, et al. relates a filter-based feature selection method using inconsistency rate measure and discretization, to a medical claims database to predict the adequacy of duration of antidepressant medication utilization. They used logistic regression and decision tree algorithms. Their results suggest it may be feasible and efficient to apply the filter-based feature selection method to reduce the dimensionality of healthcare databases [24].

Inza, et al. examined the crucial task of accurate gene selection in class prediction problems over DNA microarray datasets. They used two well-known datasets involved in the diagnosis of cancer such as Colon and Leukemia. The results showed up that the filter and wrapper based gene selection approaches lead to considerably improved accuracy results in comparison to the non-gene selection procedure, coupled with interesting and notable dimensionality reductions [25].Sathyadevi, et al.used CART, C4.5 and ID3 algorithms to diagnose hepatitis disease effectively. According their results, CART algorithm performed best results to identify to disease [26].

In wrapper approach, the searching is an overhead since the searching technique does not have the domain knowledge. In order to overcome the searching time overhead, Inza, et al used estimation of Bayesian network algorithm for feature subset selection using naive Bayes and ID3 (Iterative Dichotomiser 3) [27]. In general, the searching method may lead to increase in computational complexity, since the training data is split for evaluation. In order to overcome this issue, Grimaldi, et al used an aggregation principle with sequential search [28]. Dy & Brodley developed a wrapper-based approach for unsupervised learning using order identification (recognizing the number of clusters in the data) with the expectation maximization (EM) clustering algorithm using maximum likelihood (ML) criterion [29].

The Maldonado & Weber developed a wrapper approach-based feature selection by combining support vector machine (SVM) with kernel functions. This method uses the sequential backward selection for feature subset generation and these subsets are validated in terms of classification error to identify the best subset [30].In order to minimize the searching overhead, Gütlein, et al used the search algorithm namely ORDERED-FS that orders the features in terms of resubstitution error to identify their irrelevancy [31]. Kabir, et al developed a wrapper-based constructive approach for feature selection (CAFS) using neural network (NN). In this method, the correlation measure is used to remove the redundancy in the searching strategy for improving the performance of NN [32].

Stein et al proposed an ant colony optimization-based feature selection with wrapper model. In this approach, the ant colony optimization is used as a searching method in order to reduce the searching overhead such as blind search or forward selection or backward elimination searching methods [33]. Furthermore, to minimize the searching overhead, Zhuo, et al presented a wrapper-based feature selection using genetic algorithm with support vector machine for classifying the hyper-spectral images [34].

Meyer & Bontempi proposed a filter-based approach that uses double input symmetrical relevance (DISR) metric for feature selection. This approach returns the selected features that contain more information about the target-class than the information about other features [35]. Lin & Tang introduced an information theory-based conditional infomax feature extraction (CIFE) algorithm to measure the class-relevancy and redundancy for feature selection [36]. Brown et al used the conditional redundancy (CondRed) metric for selecting the significant features from the dataset [37].

In the recent past, the clustering technique is also adopted in feature selection. Song et al developed a feature selection framework and adopted the graph-based clustering technique to identify the similarity among the features for removing the redundant features [38]. Li et al incorporated the clustering algorithm with the chi-square statistical measure to select the features from statistical data [39]. Cai et al developed a spectral clustering-based feature selection (MCFS) for selecting the significant features from the datasets [40]. Sotoca & Pla developed a feature selection method for classification based on feature similarity with hierarchical clustering [41].

Further, it is observed that the filter-based methods are computationally better than the wrapper [42] and embedded [43] methods. Therefore, the filter-based methods can be a suitable choice for high-dimensional space. The filter-based methods achieve high generality since they do not use the supervised learning algorithm.

IV. CONCLUSION

The task of a feature selection algorithm is to provide with a computational solution to the feature selection problem. In this study, a review of literature was carried on the basis of two feature selection techniques namely, filter approach and wrapper approach. This study asserted that feature selection methods are capable to improve the performance of learning algorithms and also helps to analyze the methodology behind each algorithm in selecting the more relevant features and removing irrelevant features.

V. REFERENCES

[1] Srinivasan B and Pavya K, “Diagnosis of Thyroid Disease Using Data Mining Techniques: A Study” in International Research Journal of Engineering and Technology, Vol.3, no.11, 2016, PP. 1191-1194.
[2] Srinivasan B and Pavya K, “A study on data mining prediction techniques in healthcare sector” in International Research Journal of Engineering and Technology, Vol.3, no.3, 2016, PP. 552-556.
[3] Srinivasan B and Pavya K, “ A Comparative Study on Classification Algorithms in Data Mining” in International Journal of Innovative Science, Engineering & Technology, Vol. 3,no.3, 2016, PP. 415-418.
[4] Pavya k, and Srinivasan B. " Enhancing Filter Based Algorithms for Selecting Optimal Features from Thyroid Disease Dataset” International Journal of Advanced Research in Computer Science, Vol.8, no.9,2017, PP. 184-188.
[5] Srinivasan B and Pavya K, “Feature Selection Techniques in Data Mining: A Study” in International Journal of Scientific Development and Research (IJSDR), Vol.2, no.6, 2017, PP. 594-598.
[6] Asir Antony Gnana Singh D, Appavu S and Jebamalar Leavline E, “Literature review on Feature Selection methods for high dimentional data” in International Journal of Computer Applications,Vol.136,no.1,2017,pp.9-17.
[7] Chandrashekar G and Sahin F, “A survey on feature selection methods” in Computer and Electrical Engineering, 2014, pp.16-28.
[8] Kumar V and Minz S,” Feature Selection: A literature Review” in Smart Computing Review, 2014.
[9] Dash M and Liu H, “Feature selection for classification” in Intelligent data analysis, vol. 1, no.1, 1997, pp.131-156.
[10] Saeys Y, Abeel T and Peer YV, “Robust Feature Selection Using Ensemble Feature Selection Techniques” in W. Daelemans (Eds.): ECML PKDD, 2008, pp. 313–325.
[11] Guana D, Yuana W, Leea YK, Najeebullaha K and Rasela MK, “ A Review of Ensemble Learning Based Feature Selection” in IETE Technical Review,2014.
[12] Sanchez-Marono N, Alonso-Betanzos A and Calvo-Estevez RM., “A Wrapper Method for Feature Selection in Multiple Classes Datasets” in J. Cabestany (Eds.): IWANN, 2009, pp. 456–463.
[13] Ozcift A and Gulten A, “ A Robust Multi-Class Feature Selection Strategy Based on Rotation Forest Ensemble Algorithm for Diagnosis” in J Med Syst, 2012, pp.941–949.
[14] Lee S, Schowe B and Sivakumar V, “Feature Selection for High-Dimensional Data with RapidMiner” in Technical Report of TU Dortmund University,2011.
[15] Yu L and Liu H, “Efficient Feature Selection via Analysis of Relevance and Redundancy” in J.Machine Learning Research, vol.10,no. 5, 2004, pp. 1205-1224.
[16] Yu L and Liu H, “Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution” in Proc.20th Intl Conf.Machine Leaning, vol. 20, no. 2, 2003, pp. 856-863.
[17] Butterworth R, Piatetsky-Shapiro G, and Simovici D.A, “On Feature Selection through Clustering” in Proc. IEEE Fifth Intl Conf.Data Mining, 2005, pp. 581-584.
[18] Biesiada J and Duch W, “Features Election for High-Dimensional data a Pearson Redundancy Based Filter” in Advances in Soft Computing, vol. 45, 2008, pp. 242-249.
[19] Das S, “Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection” in Proc. 18thIntl Conf. Machine Learning, 2001, pp. 74- 81.
[20] Arauzo-Azofra A, Benitez J.M, and Castro JL, “A Feature Set Measure Based on Relief” in Proc. Fifth Intl Conf. Recent Advances in Soft Computing, 2004, pp. 104-109.
[21] Roslina A.H and Noraziah A, “Prediction of hepatitis prognosis using support vector machine and wrapper method,” in Proc. the Fuzzy Systems and Knowledge Discovery, Yantai Shandong, 2010, pp. 2209-2211.
[22] Sartakhti J.S, “Hepatitis disease diagnosis using a novel hybrid method” in Computer Methods and Programs in Biomedicine, vol. 108,issue 2, 2011, pp. 570-579.
[23] Harb H and Desuky A.S, “Feature selection on classification of medical datasets based on particle swarm optimization” in International Journal of Computer Applications, vol. 104, no. 5, 2014, pp. 14-17.

[24] Huang S, Wulsin L.R, Li H, and Guo J, "Dimensionality reduction for knowledge discovery in medical claims database: Application to antidepressant medication utilization study" in Computer Methods and Programs in Biomedicine, vol. 93, 2009, pp. 115-123.
[25] Inza I, Larranaga P, Blanco R and Cerrolaza A.J, "Filter versus wrapper gene selection approaches in DNA microarray domain" in Artificial Intelligence in Medicine, vol. 31, 2004, pp. 91-103.
[26] Sathyadevi G, "Application of cart algorithm in hepatitis disease diagnosis," in Proc. the Recent Trends in Information Technology, Chennai, Tamil Nadu, 2011, pp. 1283-1287.
[27] Inza I, Larrañaga P, Etxeberria R and Sierra B, "Feature subset selection by Bayesian network-based optimization" in Artificial intelligence, vol. 123, no. 1, 2000, pp.157-184.
[28] Grimaldi M, Cunningham P and Kokaram A, "An evaluation of alternative feature selection strategies and ensemble techniques for classifying music "in Proceedings of Fourteenth European Conference on Machine Learning and the Seventh European Conference on Principles and Practice of Knowledge Discovery in Databases, Dubrovnik, Croatia, 2003.
[29] Dy J.G and Brodley CE, "Feature subset selection and order identification for unsupervised learning"in proceedings In Proceedings of the Seventeenth International Conference on Machine Learning, 2000,pp. 247–254.
[30] Maldonado S and Weber R, "A wrapper method for feature selection using support vector machines" in Information Sciences, 2009,pp.2208-2217.
[31] Gütlein M, Frank E, Hall M and Karwath A, "March. Large-scale attribute selection using wrappers" in Proceeding of IEEE Symposium on Computational Intelligence and Data Mining, Nashville, TN, USA, 2009,pp. 332-339.
[32] Kabir M.M, Islam M.M and Murase K, "A new wrapper feature selection approach using neural network" in Neurocomputing, vol. 73, no. 16, 2010,pp.3273-3283.
[33] Stein G, Chen B, Wu A.S and Hua K.A, "March. Decision tree classifier for network intrusion detection with GA-based feature selection" in Proceedings of the forty-third ACM Annual Southeast regional conference, Kennesaw, GA, USA, vol. 2, 2005, pp. 136-141.
[34] Zhuo L, Zheng J, Li X, Wang F, Ai B and Qian J, "A genetic algorithm based wrapper feature selection method for classification of hyperspectral images using support vector machine "in Proceedings of Geoinformatics and Joint Conference on GIS and Built Environment: Classification of Remote Sensing Images, 2008,pp. 71471J-71471J.
[35] Meyer P.E and Bontempi G, "On the use of variable complementarity for feature selection in cancer classification"in Applications of Evolutionary Computing, 2006, pp. 91-102.
[36] Lin D and Tang X, "Conditional infomax learning: an integrated framework for feature extraction and fusion" in Proceeding of ninth European Conference on Computer Vision, Graz, 2006,pp. 68-82.
[37] Brown G, Pocock A, Zhao M.J and Luján M, "Conditional likelihood maximisation: a unifying framework for information theoretic feature selection" in The Journal of Machine Learning Research, vol.13, no.1, 2012, pp.27-66.
[38] Song Q, Ni J and Wang G, "A fast clustering-based feature subset selection algorithm for high-dimensional data "in IEEE Transactions on Knowledge and Data Engineering, vol. 25, no.1, 2013, pp.1-14.
[39] Li Y, Luo C, and Chung SM, "Text clustering with feature selection by using statistical data" in IEEE Transactions on Knowledge and Data Engineering, vol. 20, no.5, 2008, pp.641-652.
[40] Cai D, Zhang C and He X, "Unsupervised feature selection for multi-cluster data" in Proceedings of the sixteenth ACM SIGKDD international conference on Knowledge discovery and data mining, Washington, 2010, pp. 333-342.
[41] Sotoca J.M and Pla F, "Supervised feature selection by clustering using conditional mutual information-based distances" in Pattern Recognition, vol. 43, no.6, 2010, pp.2068-2081.
[42] Freeman C, Kulić D and Basir O, "An evaluation of classifier-specific filter measure performance for feature selection" in Pattern Recognition, vol.48, no.5, 2015, pp.1812-1826.
[43] Frénay B, Doquire G and Verleysen M, "Estimating mutual information for feature selection in the presence of label noise" in Computational Statistics & Data Analysis, vol. 71, 2014, pp.832-848