International Journal of Computer Applications Technology and Research
Volume 5– Issue 6, 395 - 402, 2016, ISSN:- 2319–8656

A Review on Feature Selection Methods For Classification Tasks

Mary Walowe Mwadulo
Department of Information Technology, Meru University of Science and Technology,
P.O BOX 972-60200 Meru, Kenya.

Abstract: In recent years, application of feature selection methods in medical datasets has greatly increased. The challenging task in feature selection is how to obtain an optimal subset of relevant and non redundant features which will give an optimal solution without increasing the complexity of the modeling task. Thus, there is a need to make practitioners aware of feature selection methods that have been successfully applied in medical data sets and highlight future trends in this area. The findings indicate that most existing feature selection methods depend on univariate ranking that does not take into account interactions between variables, overlook stability of the selection algorithms and the methods that produce good accuracy employ more number of features. However, developing a universal method that achieves the best classification accuracy with fewer features is still an open research area.

Keywords: Feature selection, attribute, dimensionality reduction, optimal subset, classification

1. INTRODUCTION

In recent years, the need to apply feature selection methods in medical datasets has greatly increased. This is because most medical datasets have large number of samples of high dimensional features. This makes it impractical, computationally expensive and causes reduction in classification accuracy when an entire input set is used. Thus, there is a need to reduce the number of features to manageable sizes which can be achieved through feature selection. Choosing an appropriate feature selection method is a non-trivial task, thus the motivation of this review is to make practitioners aware of feature selection methods that have been successfully applied in medical data sets and highlight future trends in this area.

A feature is a distinctive attribute that can be used to measure a process under observation [1]. Feature selection is a dimensionality reduction technique that reduces the number of attributes to a manageable size for processing and analysis [1]. In contrast to other dimensionality reduction techniques, feature selection does not alter the original feature set rather selects a subset by eliminating all the features whose presence in the dataset does not positively affect the learning model [1]. Thus preserves the original semantics of the features which makes it easy to interpret. Using a set of features a machine learning technique can perform classification. Classification is a machine learning task that involves assigning known class labels to training data [2].

The set of features used in model construction in the only source of information for any learning algorithm, thus it is extremely important to select an optimal subset that will be a representative of the original set. Selecting an optimal subset of relevant and non redundant features is a challenging task. Since there is a trend off, if too many features are selected it causes the classifier to have a high workload which can decrease the classification accuracy. On the other hand, if too few features are selected there is a possibility of eliminating features that would have increased the classification accuracy. Thus, there is a need to get an optimal subset of relevant and non redundant features which will give an optimal solution without decreasing the classification accuracy. No known effective method has been devised to select an optimal subset.

Feature selection helps in understanding data, reducing computational requirements, reducing the curse of dimensionality and improving the prediction performance [1]. By combining several feature selection methods, the curse of dimensionality can be reduced and classification accuracy of modeling tasks improved.

The remaining part of this paper is structured as follows: section 2 presents Feature selection techniques, section 3 presents the discussion and section 4 presents conclusion.

2. FEATURE SELECTION TECHNIQUES

Feature selection is a pre processing technique used in machine learning to remove irrelevant and redundant attributes for the purpose of increasing learning accuracy [1]. Feature selection does not only imply to cardinality reduction (imposing an arbitrary or predefined cutoff on the number of attributes that can be considered when building a model) but also the choice of attributes which could be based on presence or lack of interaction among the attributes and the classification algorithm. This means that the modeling tool actively selects or discards attributes based on their usefulness for analysis. Feature selection is necessary because the high dimensionality and vast amount of data poses a challenge to the learning task. In the presence of many irrelevant features some of which do not add much value during the learning process, learning models tend to become computationally complex, over fit, become less comprehensible and decrease learning accuracy. Feature selection is one effective way to identify relevant features for dimensionality reduction. However, the advantages of feature selection come with extra effort of trying to get an optimal subset that will be a true representation of the original dataset. In the context of classification, feature selection techniques can be categorized into Filter methods, wrapper methods, embedded methods and hybrid methods.

2.1 Filter methods

Filter methods are feature ranking techniques that evaluate the relevance of features by looking at the intrinsic properties of the data independent of the classification algorithm [2], [3], [4]. A suitable ranking criterion is used to score the variables and a threshold is used to remove the variable below the threshold [1]. Afterwards this subset of features is used as input to the classification algorithm. Filter methods assess the relevance of features using measures like distance, information, correlation and consistency [5]. Advantages of filter methods are that they are fast, scalable and independent of a learning algorithm. As a result feature selection needs to be performed only once, and then different classifiers can be evaluated [2]. Disadvantages of filter techniques is that they lack interaction with the classifier which makes them generate general results and lower classification accuracy [2], [6]. Filter methods can be categorized into univariate and multivariate. Univariate filter methods ignore feature dependencies which can lead to selection of redundant features and worst classification performance when compared to other feature selection techniques [2]. On the other hand, multivariate filter methods model feature dependencies independent of the classifier. In addition to evaluating class relevance like univariate, they also calculate the dependency between each feature pair [2]. Some univariate filter feature selection methods include:

2.1.1 Information gain (IG)

It is a symmetrical measure of dependency between two variables. The information gained about Y after observing X is equal to the information gained about X after observing Y [6]. Selects candidate features with more information, for each feature a score is obtained based on how much more information about the class is gained when using that feature. The level of features usefulness is determined by how great is the decrease in entropy of the class when considered with the corresponding features individually [3]. Disadvantage of IG is that it favors features with more values even when they may not be more informative [3].

IG is defined as:

$$IG(X; Y) = H(X) - H(X|Y)$$

2.1.2 Gain Ratio (GR)

The Gain Ratio is a non-symmetrical measure that is introduced to compensate for the bias of the IG [6]. GR is given by

$$GR = \frac{IG}{H(X)}$$

When the variable Y has to be predicted, we normalize the IG by dividing by the entropy of X, and vice versa. Due to this normalization, the GR values always fall in the range [0, 1]. A value of GR = 1 indicates that the knowledge of X completely predicts Y, and GR = 0 means that there is no relation between Y and X. In opposition to IG, the GR favors variables with fewer values [6].

2.1.3 Symmetric Uncertainty (SU)

This is a correlation measure between the features and the target class. The Symmetrical Uncertainty criterion compensates for the inherent bias of IG by dividing it by the sum of the entropies of X and Y [6]. Features with a high Symmetric Uncertainty value get a higher value.

SU takes values, which are normalized to the range [0, 1] because of the correction factor 2. A value of SU = 1 means that the knowledge of one feature completely predicts, and the other SU = 0 indicates, that X and Yare uncorrelated [6]. A weakness of SU is that it is biased towards features with fewer values [3]. It is a normalized information theoretic measure which uses entropy and conditional entropy values to calculate dependencies of features [7].

SU is defined as:

$$SU\,(X,Y) = \frac{2\,IG(X;Y)}{H(X)+H(Y)}$$

Multivariate filter techniques which incorporate a degree of feature dependencies that can be used to solve the problem. Some of multivariate filter methods include:

2.1.4 Correlation based Feature Selection (CFS)

Correlation based feature selection method evaluate subsets of features by selecting feature subsets contain features highly correlated with the classification, yet uncorrelated to each other. CFS evaluates a subset by considering the predictive ability of each one of its features individually and also their degree of redundancy (or correlation). This means that given a function, the algorithm can decide on its next moves by selecting the option that maximizes the output of this function [6].

2.1.5 Markov blanket Filter (MBF)

Markov blanket Filter method finds features that are independent of the class label so that removing them will not affect the accuracy [6]. It does not require one to specify a variable ordering, nor to fix an upper bound on the number of parents allowed for each node, and this makes MBF both more general and more appealing for application to domains where no prior knowledge can be used to constrain the learning process [7].

2.1.6 Fast Correlation based Feature Selection (FCBF)

It is a feature selection method which starts with full set of features, uses symmetrical uncertainty to calculate dependences of features and finds best subset using backward selection technique with sequential search strategy. It has an inside stopping criterion that makes it stop when there are no features left to eliminate. It is a correlation based feature subset selection method which runs, in general, significantly faster than other subset selection methods [7].

2.1.7 Minimum Redundancy Maximum Relevance (MRMR)

It is a multivariate feature selection method which maximizes the relevancy of features with the class label while it minimizes the redundancy in each class [6]. It starts with an empty set, uses mutual information (a symmetrical information theoretic measure that measures the amount of information that can be obtained about one random variable by observing another) to weight features and forward selection technique with sequential search strategy to find the best subset of features. It has a parameter k which enables it to stop when there are k features in the selected feature subset [7]. MRMR does not deal with the type of dependency rather the quantity of dependency (it uses mutual information) which can lead to inaccurate ordering of the variables.

Pandey etl. [1], [8], Used information gain for feature selection which showed remarkable result. In [7] Modified Fast Correlation Feature selection method by giving every feature a temporary

predominance in the elimination process and making them start eliminating features from the features which are least correlated with the class. An iteration process that allows one feature to eliminate one feature per iteration which makes elimination process more balanced.

In [2], adopted a two phase feature selection method, where in the first phase, they combined information gain and symmetric uncertainty to generate two subsets of reliable features. In the second phase, the two subsets are merged, weighted and ranked to extract the most important features. Combination of two filtering methods, lead to higher accuracy of intrusion detection. In [5] a four stage Multi Filtration Feature Selection (MFFS) method was introduced. The method adjusts variance coverage and builds the model with the value at which maximum classification accuracy is obtained. In stage one, relevant features are generated using Principal Component Analysis (PCA), stage two, features are ranked using correlation feature selection which is improved by employing symmetric uncertainty in stage three. Finally the system is validated against standard classifier models. The results showed that classification accuracy based on the selected subset by Multi Filtration Feature Selection (MFFS) method was better than that based on the original feature set. Authors in [6] devised a three stage hybrid feature selection approach, that recommended selecting features at the intersection of information gain and Significance analysis of Micro array (SAM). The intersection features are then subjected to mRMR to minimize redundancy in the second stage. Finally, Support Vector Machine Recursive Feature Elimination (SVM-RFE) is applied to choose the most discriminate genes.

Karimi etal. [3], utilized both feature space and sample domain in two phases. The first phase filters and resample the sample domain and the second phase adopted a hybrid procedure by information gain, wrapper subset evaluation and genetic search to find the optimal feature space.

2.2 Wrappers

They use the predictor as a black box and the predictor performance as the objective function to evaluate the variable subset [1], [2],[3]. A search procedure in the space of possible feature subset is defined and various subsets of features are generated and evaluated. The evaluation of a specific subset of features is obtained by training and testing a specific classification model, making this approach tailored to a specific classification algorithm [2].Advantages of this approach is that it includes the interaction between feature subset search and model selection, and the ability to take into account feature dependencies [2].A common drawback is that it has a higher risk of over fitting than filter techniques and are computationally intensive, especially if the building classifier has a high computational cost. Over fitting occurs if the classifier model learns the data too well and provides poor generalization capability.

Wrappers can be categorized into Sequential selection algorithms and Heuristic search algorithms.

2.2.1 Sequential selection algorithms

This algorithm can do forward or backward selection. With Sequential Forward Selection (SFFS) algorithm, you start with an empty set and add one feature for the first step which gives the highest value for the objective function. From the second step onwards the remaining features are added individually to the current subset and the new subset is evaluated. The individual feature is permanently included in the subset if it gives the maximum classification accuracy. The process is repeated until the required number of features is added. This is a naive SFS algorithm since the dependency between the features is not accounted for [1]; this is impractical for feature subset selection from a large number of samples of high dimensionality features [5]. The Sequential Floating Forward Selection (SFFS) [9] algorithm is more flexible than the naive SFS because it introduces an additional backtracking step. The first step of the algorithm is the same as the SFS algorithm which adds one feature at a time based on the objective function. The SFFS algorithm adds another step which excludes one feature at a time from the subset obtained in the first step and evaluates the new subsets. If excluding a feature increases the value of the objective function

then that feature is removed and goes back to the first step with the new reduced subset or else the algorithm is repeated from the top. This process is repeated until the required number of features is added or required performance is reached. The SFS and SFFS methods suffer from producing nested subsets since the forward inclusion was always unconditional which means that two highly correlated variables might be included if it gave the highest performance in the SFS evaluation.

To avoid the nesting effect, adaptive version of the SFFS was developed in [10]. The Adaptive Sequential Forward Floating Selection (ASFFS) algorithm used a parameter r which would specify the number of features to be added in the inclusion phase which was calculated adaptively. The parameter o would be used in the exclusion phase to remove maximum number of features if it increased the performance. The ASFFS attempted to obtain a less redundant subset than the SFFS algorithm.

A different sequential selection approach is Sequential Backward Selection (SBS). It is similar to SFS but the algorithm starts from the complete set of variables and removes one feature at a time whose removal gives the lowest decrease in predictor performance.

2.2.2 Heuristic search algorithms

The heuristic search algorithms evaluate different subsets to optimize the objective function. Different subsets are generated either by searching around in a search space or by generating solutions to the optimization problem.

Genetic algorithms (GA) is a general adaptive optimization search method based Darwinian principle of 'survival of the fittest', GA works with a set of candidate solutions called a population and obtains the optimal solution after a series of iterative computations. GA evaluates each individual's fitness, i.e. quality of the solution, through a fitness function. The fitter chromosomes have higher probability to be kept in the next generation or be selected into the recombination pool using the tournament selection methods. If the fittest individual or chromosome in a population cannot meet the requirement, successive populations will be reproduced to provide more alternate solutions. The crossover and mutation functions are the main operators that randomly transform the chromosomes and finally impact their fitness value. The evolution will not stop until acceptable results are obtained. Associated with the characteristics of exploitation and exploration search, GA can deal with large search spaces efficiently, and hence has less chance to get local optimal solution than other algorithms [11]. GAs offer a particularly attractive approach for problems like feature subset selection since they are generally quite effective for rapid global search of large, non-linear and poorly understood spaces. GAs are based on an imitation of the biological process in which new and better populations among different species are developed during evolution [12]. Thus, unlike most standard heuristics, GA uses information of a population (individuals) of solutions when they search for better solutions.

In [13] combined Symmetric Uncertainty and Genetic Algorithm for feature selection based on the Naïve Bayes classifier. Experimental results conducted over several UCI datasets revealed that higher level of dimensionality reduction was achieved by selecting less number of features than other methods.

In [3] propose a framework based on a genetic algorithm (GA) for feature subset selection that combines various existing feature selection methods. The goal is to effectively utilize useful information from different feature selection methods to select better feature subsets with smaller size and/or higher classification performance in comparison with the existing methods. Multiple selection criteria are combined by a genetic algorithm to improve feature subset selection.

In [12] used a preprocessed statistical parametric mapping software and PCA were used for dimension reduction. Then, independent components of the new data (given by PCA) were estimated using Independent Component Analysis (ICA) method. For feature extraction, LBP histogram extraction technique was used for all estimated components. Genetic Algorithm was used for selection of the most significant histogram bins, in next step. Then, linear discriminant analysis (LDA) is performed to further extract features that maximize the ratio of between-class and within-class variability. Finally, a classifier based on Euclidean distance was used for classification. In [4] adopted an oversampling approach in which the minority class is oversampled by creating synthetic examples rather than by

oversampling with replacement. The synthetic examples are generated in a less application specific manner, by operating in feature space rather than sample domain. Selective Bayesian which uses a forward and backward greedy search method is applied to find a feature subset from the whole space of entire features. It uses the accuracy of Naïve Bayes classifier on the training data to evaluate feature subsets, and considers adding each unselected feature which can improve the accuracy on each iteration. Entropy measure is then calculated and used to measure uncertainty of a class attribute using information gain. Genetic algorithm is applied as a function optimizer.

2.3 Embedded methods

Embedded methods interact with learning algorithm at a lower computational cost than the wrapper approach [1]. It captures feature dependencies and considers not only relations between one input features and the output feature, but also searches locally for features that allow better local discrimination. It uses the independent criteria to decide the optimal subset for a know cardinality. The learning algorithm is used to select the final optimal subset among the optimal subsets across different cardinality [3].This approach has the advantage of including the interaction with the classification model , while at the same time being far less computationally intensive than wrapper methods [2].

2.4 Hybrid methods

Hybrid methods are based on sequential approach where the first step is usually based on filter methods to reduce the number of features used in the second stage. Afterwards a wrapper method is employed to select the desired number of features using this reduced set [4].

3. DISCUSSION

Classification accuracy is very important on medical data sets; however, medical data sets have many features which can be irrelevant and redundant. The irrelevant and redundant features can overload the classifier and lead to decreased classification accuracy. Thus, there is need to reduce the input set to manageable sizes. To solve this problem [2], [3], [6] aspired to reduce the number of features before presenting it for classification.

Selecting an appropriate set of features is extremely important since the feature set selected is the only source of information for any learning algorithm using the data of interest. A goal of feature selection is to avoid selecting too many or too few features than is necessary. If too few features are selected, there is a possibility that the information content in this set of features is low, on the other hand, if too many (irrelevant) features are selected, the effects due to noise may overshadow the information present. Hence this is a trade off that must be considered when applying feature selection methods.

Researchers have tried to address the issue of feature subset selection through filter methods [1], [2], [5], [7] some of which provide a ranking criterion. These methods are fast and scalable; however, they ignore feature dependencies and interaction with the classifier. This makes their results unrealistic since a given feature might provide more information when present with certain other features than when considered by itself. Thus, it is important to consider features not only in relation to the class but also in relation to each other. Again features should be selected as a set, rather than selecting the best features to form the (supposedly) best set. The best individual feature does not necessarily constitute the best set of features. However in most real world situations, it is not know what the best set of features is neither the number of features in such a set.

4. CONCLUSION

After reviewing the work on feature selection, it is observed that obtaining an optimal subset of relevant and non redundant features is a non trivial task. Most of the existing methods in the literature depend on univariate ranking that does not take into account interactions between the variables already included in the selected subsets and the remaining one, overlook stability of the selection algorithm and the methods that produce good accuracy employ more number of features which affects the classification accuracy. This paper attempts to reveal that a holistic and Universal method

that achieves the best classification accuracy with fewer features is still an open research area.

5. REFERENCES

[1] Girish Chandrashekar, Ferat Sahin, (2014). "A survey on feature selection methods". Computers and Electrical Engineering.

[2] Yvan Saeys, Inak Inza, Pedro Larranaga, (2007). "A review of Feature Selection techniques in bioinformatics". Bioinformatics, Oxford University press.

[3] Feng Tan, Xuezheng Fu, Yanqing Zhang, Anu G. Bourgeois, (2008). "A genetic algorithm-based method for feature subset selection". Soft Comput.

[4] Muhammad Shakil Pervez, Dewan Md. Farid ,(2015). "Literature Review of Feature Selection for mining Tasks".International Journal of Computer Application, Vol 116, No. 21.

[5] S. Sasikala, S. Appavu alias Balamurugan, S. Geetha, (2014). "Multi Filtration Feature Selection (MFFS) to improve discriminatory ability in clinical data set". Applied Computing and Informatics.

[6] Hall, M. A. & Smith, L. A. (1998). Practical feature subset selection for machine learning. In C. McDonald (Ed.), Computer Science '98 Proceedings of the 21st Australasian Computer Science Conference ACSC'98, Perth.

[7] Baris Senliol, Gokhan Gulgezen, Lei Yu, Zehra Cataltepe, (2008)." Fast Correlation Based Filter (FCBF) with a Different Search Strategy". Computer and Information Science, 23rd international symponsium.

[8] B.Azhagusundari, Antony Selvadoss Thanamani, (2013). "Feature Selection based on Information Gain". International Journal of Innovative Technology and Exploring Engineering (IJITEE), Vol 2, issue 2.

[9] P. Pudil , J. Novovicova , j. Kittler (1994). "Floating search methods in feature selection". Pattern Recognition Letters.

[10] P. Somol, P. Pudil , J. Novovicova, P. Paclik (1999)." Adaptive Floating search methods in feature selection". Pattern Recognition Letters.

[11] Li Zhuo, Jing Zheng, Fang Wang , Xia Li , Bin Ai , Junping Qian, (2008). "A Genetic Algorithm Based Wrapper Feature Selection Method For Classification of Hyperspectral Images Using Support Vector Machine". The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. 37.

[12] H. Shahamat, A. A. Pouyan, (2014). Feature selection using genetic algorithm for classification of schizophrenia using fMRI data.

[13] Bai-Ning Jiang, Xiang-Qian Ding, Lin-Tao Ma, Ying He, Tao Wang, Wei-Wei Xie, (2008). A Hybrid Feature Selection Algorithm: Combination of Symmetrical Uncertainty and Genetic Algorithms. The Second International Symposium on Optimization and Systems Biology (OSB'08).

[14] Suman Pandey, Anshu Tiwari, Akhilesh Kumar Shirivas ,Vivek Sharma, (2015). "Thyroid Classification using Ensemble Model with feature selection". International Journal of Computer Science and Information Technologies, Vol. 6 (3).

[15] Shailendra Singh, Sanjay Silakari, (2009). "An ensemble approach for feature selection of Cyber Attack Dataset". International Journal of Computer Science and Information Security, Vol 6, No. 2.

[16] Zahra karimi, Mohammad Mansour, Ali Harounabadi, (2013). " Feature Ranking in Intrusion Detection Dataset using Combination of Filtering Methods". International Journal of Computer Applications. Vol. 78. No. 4.

[17] Mehdi Naseriparsa, Amir-Masoud Bidgoli, Touraj Varaee, (2013). "A Hybrid Feature Selection Method to Improve Performance of a Group of Classification Algorithms". International Journal of Computer Applications .Vol. 69 No.17.

[18] Ahmed Soufi Abou-Taleb, Ahmed Ahmed Mohamed, Osama Abdo Mohamed, Amr Hassan Abedelhalim,

(2013). “Hybridizing Filters and Wrapper Approaches for Improving the Classification Accuracy of Microarray Dataset”. International Journal of Soft Computing and Engineering. Vol.3.

[19] Vipin Kumar, Sonajharia Minz, (2014). “Feature Selection: A literature Review”. Smart Computing Review, Vol 4.

[20] Jasmina Novaković, Perica Strbac, Dusan Bulatović, (2011).”Toward Optimal Feature Selection using Ranking Methods and Classification Algorithms”. Yugoslav Journal of Operations Research.

[21] Zena M. Hira, Duncan F. Gillies, (2015) “A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data”. Hindawi Publishing Corporation Advances in Bioinformatics.