RESEARCHERID
THOMSON REUTERS

[ICRTCET-2018]

ISSN 2348 – 8034
Impact Factor- 5.070

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

THE TAXONOMY OF FEATURE SELECTION TECHNIQUES IN MINING ALGORITHMS

F.Rosita Kamala[*1] & P.Ranjit Jeba Thangaiah[2]

[*1]Doctoral Scholar, Department of Computer Science, Bharathiar University, Coimbatore, India
[2]Associate professor, Department of Computer Applications, Karunya Institute of Technology and Sciences, Coimbatore, India.

ABSTRACT

Selection of attributes or selection of variables is a competent method in dimensionality reduction. It is successful for identifying a subset of optimal significant features in learning. The learning accurateness is improved, as the dimensions of data undergo reduction, to enhance the certainty. The selection of attributes has two key values: a subset of candidate features evaluation and an exploration of the features in the features space. Experiments are conducted on this pragmatic study to observe the merits and demerits of the feature selection methodologies, to provide some strategies on selecting a technique, and to evaluate the accuracy of the classifier prior to and subsequent to feature selection. The recent algorithms approve various procedures to assess the feature subsets effectiveness.

Keywords: *Feature Selection, Filter, Wrapper, Machine Learning, Dimensionality reduction.*

I. INTRODUCTION

The computational cost of irrelevant input features is greater and causes overfitting. Reducing dimensions is the most difficult job, when managing with data of huge dimensions. It is done to decrease the initial features of data and signifies the learning performance. Variable selection techniques are applied for increasing the execution speed of the mining algorithms by significant improvements in the precision of the model's performance. Dimensionality reduction is performed by the techniques, called Feature extraction or variable selection.

Variable selection[1] process chooses a few numbers of related features, that is adequate for the class label's computation. The major reasons for feature selection are computation complexity, reduction in computation cost of dimensions, significant classifier's significance and resulting outcomes for problems.

Feature extraction outputs a few numbers of new dimensions, by the combinations of the initial variables. Using class labels, feature extraction techniques are categorized into supervised or unsupervised.

Selection of features is generally classified into two types: Filter technique and Wrapper technique [1].

The filter methods can be able to select dimension subsets without the usage of any learning algorithm [2]. It selects the weight of features by one or more significant criteria and not depending on the learner. The efficiency of filter methods is very high, faster performance than wrapper method and can scale to large datasets. The cons of filter method is very difficult to identify a criterion filter, which is required for classification procedures.

The wrapper method uses any one data mining algorithm and learning significance is adopted as a criterion for evaluation adjoined with a good classifier, which requires repeated trainings for variant combinations of feature dimensions. In this paper, naive bayes, decision tree, and nearest neighbour classifiers are used to score features for subset evaluation. The optimization algorithms will evaluate the feature subsets in assessing the mining efficiency [3]. The cons is expensive computation cost because of large datasets.

The overall goal of this paper is to analyze the various filter, wrapper and hybrid methods, which affect the feature sets dimension by eliminating the non-performance dimensions and redundant features by increasing the performance of the classifier and achieving better computational time.

Organization of this paper is represented as mentioned below: Section I describes the strategies, and feature selection's pros and cons. Section II describes the associated and relevant work. Section III highlights the different algorithms. Section IV explores the experimental study. Section V concludes the current work and future progression.

II. RELEVANT WORK

Feature selection process

The key steps in a representative method of feature selection are four. They are described as below. 1. Generate subset, 2. Evaluate subset, 3. Stopping Criteria, 4. Result Validation.

***Generate subset*:** The generate subset procedure generates the subsequent feature subset for further processing [4]. It is basically a heuristic search process, specifying a candidate subset with its state for evaluation in the entire search space. It uses an explore strategy to generate subsets of features. The initial point of the search is decided, which influences the direction of search consecutively. The method of search begins with no features or a null set and consecutively adjoins features (forward), or begins with all features or a complete set and consecutively eliminates features (backward), or with an arbitrary subset produced arbitrarily afterward in the final case of the search duration [5]. Features are repeatedly included or eliminated in the initial two cases. The process is worked out by two key issues to initiate in both of the ends to include and eliminate features at the same time (i.e., bi-directional). The search can also initiate with a subset of arbitrarily chosen features to evade being ensnared into neighboring optima [6]. Secondly, one must decide an explore approaches.

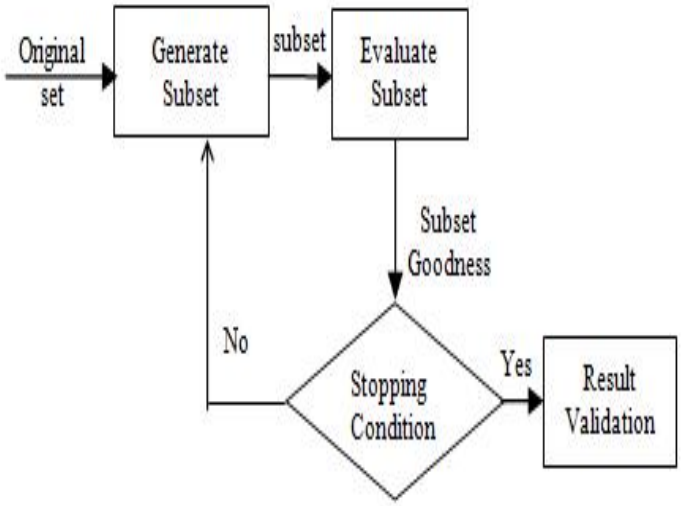

Fig. 1 The four step process of variable selection

ISSN 2348 – 8034
Impact Factor- 5.070

***Evaluate subset*:** An evaluate function evaluates the feature subset. An *evaluate function* computes the subset goodness, got by some subset generate procedure. This result is evaluated for a comparison with the prior best. If the results are considered to be enhanced, it changes the preceding best subset of variables. Always a subset of optimal features is comparative to an assured evaluate function.

***Stopping Condition*:** A *stopping condition* makes a decision, when to stop and computes the time at which stopping event happens in the feature selection process. The stopping c*ondition* rooted in the function which performs evaluation includes: (i) whether an auxiliary adding (or removal) of some dimensions produces an improved subset, and (ii) whether an optimal subset is obtained in line with any evaluate function. The process of feature selection stops by the outcome of the optimal features subset that can be validated later.

***Result Validation*:** It is based on three base learners namely, decision tree, naïve bayes, and nearest neighbour.

III. TAXONOMY OF FEATURE SELECTION TECHNIQUES

Filter Method

Search algorithm and evaluate function are the algorithms of the filter model. The methodology begins with the search method which initiates from the initial subset S_0. Every subset S generated is undergone an evaluation by M (independent measure) which is independent and is found to be in comparison with the best one recognized previously. The search repeats till the specified stopping condition is attained. The algorithm produces the best output subset S_{best} which was the end outcome. Because the filter techniques adopt, an evaluate condition independently with no classifier algorithm's involvement, and is also computationally efficient.

The filter Chi statistic is used to compute the degree of dependence between any two items [7]. This is performed by measuring the observed co-occurrence frequencies with the expected frequencies in a two way contingency table, when they are seemed to be independent. In Chi dependency test, the null and the alternative hypotheses is considered. According to the null hypothesis that any two variables are considered to be independent of each other. According to the alternative hypothesis that there is a kind of some dependence between these two variables. The null hypothesis is tested by comparing the observed frequencies with the expected frequencies based on the assumption that the null hypothesis is true.

The filter correlation is a measure of statistics that shows the degree to which nearly two or more variables change together [8]. A positive correlation measures the degree to which those variables increment or decrement in parallel. A negative correlation measures the degree to which one variable increments as the other decrements. A correlation coefficient is a measure of statistics, that the extent to which it modifies the significance of one variable to the significance of another. When an instability of one variable consistently anticipates a similar instability in another variable, that means that the change in one, causes the change in the other. However, correlation does not involve causation.

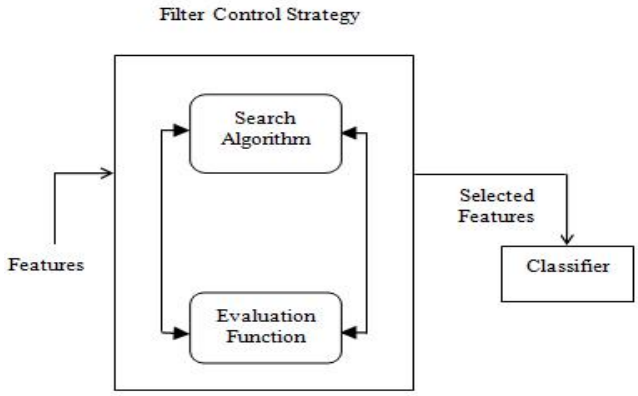

Fig.2 Filter Method

Wrapper Method

The predictor is used as a black box in the wrapper method and the performance of the predictor has variable subsets to be evaluated as the objective function. Since evaluation of 2N subsets has become a NP-hard problem, employing search algorithms found suboptimal subsets are finding a subset heuristically [9]. Different algorithms for search may be employed to identify a variable subset of that optimizes the goal of the function, that is the outcome of the learning significance.

The Sequence Forward Selection (SFS) method begins with a null dataset, which keeps on adding one feature by another feature to give the maximum importance in the subsequent step for the objective function [9,10]. Thus, the latest subset is computed for evaluation. The inclusion of the most important features happens permanently in the subset, if the classification accuracy is maximum. The repetition of the process happens, till the essential features are included. A Sequence Backward Selection (SBS) method [9,10] can also be built, which is more or less similar to SFS. The algorithm begins with the whole set of features and eliminates a single feature at a time, whose elimination provides the minimum reduction in learning significance.

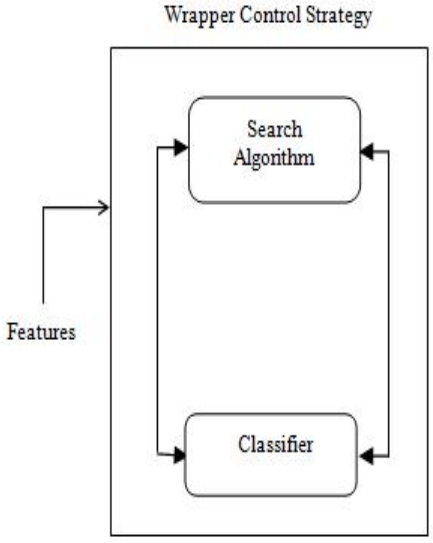

Fig.3. Wrapper Method

The other wrapper methods like Genetic Algorithm(GA), Particle Swarm Optimization(PSO) are also implemented and results are tabulated in Table:5.

IV. EXPERIMENT

Comprehensive experiments are conducted on a variety of UCI (University of California, Irvine) [11] real datasets represented in Table 1 and Table 2 represents the information of the microarray datasets [12] used.

Dataset

In this study, seven datasets are used. Most of them were downloaded from UCI repository, are used to verify the effectiveness of the algorithms. The experimentation is carried out on datasets with more than 30 features. The complete details are described in Table:1 and Table:2.

Table I. Experimental dataset information

Dataset	#Categorical	#Continuous	#Size	#Classes
Sonar	0	60	208	3
Vehicle	0	18	846	5
Iono	0	34	351	3
Chess	36	0	3196	2
Splice	61	0	3190	3

Table II. High dimension microarray datasets

Datasets	#Genes	#Instances	#Classes
Leukemia2C (Leuk2C)	7129	72	2
Central Nervous (CNS)	7129	60	2

The measures for the performance of classification are the computation of learning accuracy, that is calculated by Eq. 1.

$$\text{Accuracy} = (TP+TN) / (TP+FN+FP+TN) \quad (1)$$

Where TP is the number of correctly classified positive occurrences. TN is the number of correctly classified negative occurrences. FN is the number of incorrectly classified positive occurrences as negative. FP is the number of incorrectly classified negative occurrences as positive [4].

Filter and Wrapper Results

In this paper two filter methods and four wrapper methods are adopted for optimal feature subset selection, based on three base learners Decision tree, Naïve Bayes, and Nearest neighbor is presented in Table:3, Table:4 and Table:5. The performance is low for numerical datasets like sonar, vehicle, ionosphere, when compared to categorical datasets in filter method with very less execution time even for microarray datasets. The main disadvantage of filter approach is, when the size of original features set increases, there is a fulmination of the search space dimension.

The wrapper method produces little better performance than filter method and takes more computation time. It is taking nearly or more than 10 minutes for microarray datasets. The primary weakness of wrapper methods is the cost of computations. The hybrid method of filters and wrappers is suggested to overcome these problems [13].

Hybrid Innovation Methods

Hybrid innovation methods consider that the diversification of respective data disparity or procedure discrepancy is not adequate [13]. These two diversity methods are combined. Dittman et al. presents that the closeness amid procedure discrepancy and hybrid innovation is enough higher than their closeness amid data variation [14]. In order to have the maximum closeness, procedure discrepancy and hybrid innovation techniques demonstrate superior learning significance than data diversity. Due to the relative huge number of variables, the procedure discrepancy and hybrid innovation can create more discrepancies for every feature selector. Since the training data volume is small, function and hybrid innovation explain excellent significance. This recommends that the training data size is a vital factor to select the appropriate method for generation of multiple feature selectors.

V. CONCLUSION AND FUTURE PROGRESSION

The different flavours of filter and wrapper methods are explored in this paper, in consideration of the benefits of both feature selection methods. A comprehensive experimental study was performed with seven datasets of UCI repository and microarray datasets with more than 30 features. The evaluations were made for the comparison of the significance of both the methods namely, filter, and wrapper methods. Based on this analysis, the results vindicate that the performance measures like accuracy of the datasets are very low, when compared with the identical datasets of different current methodologies of the confirmed research work. As a future work, this work can be extended to make hybridization of both filter and wrapper methods to make optimal subsets to enhance accuracy and computational time of different types of data.

Table III. Comparison of performance measures precision, recall, accuracy and execution time(sec) of the various filter methods.

Dataset	Correlation				Chi square statistics			
	Recall	Precision	Accuracy	Exec. Time	Recall	Precision	Accuracy	Exec. Time
Sonar	33.33	17.67	53.02	0	33.33	17.67	53.02	0
Vehicle	51.62	51.75	64.21	0	51.62	51.75	64.21	0
Iono	51.15	50.60	70.71	0	51.15	50.60	70.71	0
Chess	98.09	98.05	98.06	0	98.09	98.05	98.06	0
Splice	95.28	94.42	95.30	0	95.28	94.42	95.30	0
Leuk2C	93.33	94	98.57	1	93.33	94	98.57	5
CNS	59.92	61.83	66.67	1	59.92	61.83	66.67	3

Table IV. Comparison of performance measures precision, recall, accuracy and execution time(sec) of the various wrapper methods.

Dataset	Sequence Forward Selection				Sequence Backward Selection			
	Recall	Precision	Accuracy	Exec. Time	Recall	Precision	Accuracy	Exec. Time
Sonar	38.4	39.82	58.3	1	38.4	38.78	54.98	10
Vehicle	52.7	50.95	64.8	2	54	53.40	66.94	4
Iono	55.1	58.68	86.4	0	52.8	51.67	76.12	6
Chess	94.2	94.54	94.3	20	98.3	98.25	98.28	8
Splice	95.7	94.67	95.7	75	95.7	94.93	95.74	18
Leuk2C	95.2	96.05	96.2	602	98.5	97.34	98.21	617
CNS	85.0	86.56	86.4	593	80.6	80.34	81.22	604

Table V. Comparison of performance measures precision, recall, accuracy and execution time(sec) of the various wrapper methods.

Dataset	Genetic Algorithm				PSO			
	Recall	Precision	Accuracy	Exec. Time	Recall	Precision	Accuracy	Exec. Time
Sonar	33.33	17.75	53.24	3	42.1	44.67	64.3	8
Vehicle	53.48	52.86	66.36	15	55.9	55.39	70.4	10
Iono	52.82	51.65	74.18	5	57.9	57.90	85.7	4
Chess	98.36	98.29	98.31	18	97	97.38	97.2	10
Splice	95.64	94.82	95.64	24	77.35	81.42	76.2	10
Leuk2C	100	100	100	28	94.64	94.91	95.1	511
CNS	75.75	73.92	73.33	198	79.83	80.67	81.7	500

REFERENCES

1. *Girish Chandrashekar, and Ferat Sahin, "A Survey on Feature Selection," Computer and Electrical Engineering, vol. 40, pp.16-28, 2014*
2. *V. Kumar, S. Minz, "Feature Selection - A literature Review," Smart Computing Review, vol. 4, no. 3, pp.211-229, June 2014.*
3. *Donghai Guan, Weiwei Yuan, Young-Koo Le, Kamran Najeebulah and Mostofa Kamal Rasel, "A Review of Ensemble Learning Based Feature Selection Method," IETE Technical Review, vol. 31, no. 3, pp.190-198, May June 2014.*
4. *Andreas G. K. Janecek, Wilfried N. Gansterer, Michael A. Demel, Gerhard F. Ecker, "On the Relationship Between Feature Selection and Classification Accuracy," JMLR: Workshop and Conference Proceedings 4, pp.90-105, 2008.*

5. *M.S. Pervez, Farid D. Md., "Literature Review of Feature Selection for Mining Tasks," International Journal of ComputerApplications, vol. 116, no. 21, pp.30-33, April 2015.*
6. *S. Beniwal, J. Arora, "Classification and Feature Selection Techniques in Data Mining," International Journal of Engineering Research & Technology (IJERT), vol. 1, no. 6, 2012, pp.1-6.*
7. *B. Jantawan, C.F. Tsai, "A Comparison of Filter and Wrapper Approaches with Data Mining Techniques for Categorical Variables Selection," International Journal of Innovative Research in Computer and Communication Engineering, 2014.*
8. *M. Naseriparsa, A.M. Bidgoli, T.Varae, "A Hybrid Feature Selection Method to Improve Performance of a Group of Classification Algorithms," International Journal of Computer Applications, vol. 69, no. 17, pp.28-35, 2013.*
9. *P. Pudil, J. Novovicova, J. Kittler, "Floating Search Methods in Feature Selection," Pattern Recognition Letters, vol.15, November 1994, pp.1119–1125.*
10. *J. Reunaanen, "Overfitting in Making Comparisons Between Variable Selection Methods," Journal of Machine Learning Research, vol.3, 2003, pp.1371–1382.*
11. *M. Lichmaan, "UCI Machine Learning Repository" Irvine, CA: University of California, School of Information and Computer Science. http://archive.ics.uci.edu/ml.*
12. *Zhuzx. Microarray Datasets, http :// csse.szu.edu.cn / staff / zhuzx / Datasets.html.*
13. *Rosita Kamala, and Dr. P. Ranjit Jeba Thangaiah, "A Proposed Two Phase Hybrid Feature Selection Method using Backward Elimination and PSO," Int. J. of Appl. Eng. Res, vol. 11, no.1, pp. 77 - 83, 2016.*
14. *J. Dittman, M. Khoshgoftar, R. Wald, and A. Napolitano,"Comparing Two New Gene Selection Ensemble Approaches With the Commonly used Approach," in Proceedings of the11th International Conference on Machine Learning and Applications, Florida, Dec. 12-15, 2012, pp. 184 -191.*