Algorithms for Classification and Clustering

Anamika Chaudhary[1]
Department of CSE
JIET Group of Institutions
Jodhpur, Rajasthan, India

Rajeev K R Singh[2]
Department of CSE
JIET Group of Institutions
Jodhpur, Rajasthan, India

***Abstract*—Feature selection is an important topic in data mining, especially for high dimensional dataset. Feature selection is a process commonly used in machine learning, wherein subsets of the features available from the data are selected for application of learning algorithm. The best subset contains the least number of dimensions that most contribute to accuracy. Feature selection methods can be decomposed into three main classes, one is filter method, another one is wrapper method and third one is embedded method. This paper presents an empirical comparison of feature selection methods and its algorithm. In view of the substantial number of existing feature selection algorithms, the need arises to count on criteria that enable to adequately decide which algorithm to use in certain situation. This paper reviews several fundamental algorithms found in the literature and assess their performance in a controlled scenario.**

Keywords— ***Feature selection, Filter method, Wrapper method, Information gain, Feature Selection Algorithms.***

I. INTRODUCTION

The feature selection problem is inescapable in inductive machine learning or data mining setting and its significance is beyond doubt. The main benefit of a correct selection is the terms of learning speed, speculation capacity or simplicity of the induced model. On the other hand there are the straight benefits related with a smaller number of features: a reduced measurement cost and hopefully a better understanding of the domain. A feature selection algorithm (FSA) is a computational solution that should be guided by a certain definition of subset relevance although in many cases this definition is implicit or followed in a loose sense. This is so because, from the inductive learning perspective, the relevance of a feature may have several definitions depending on precise objective (Caruana and Freitag, 1994). Thus the need arises to count on common sense criteria that enable to adequately decide which algorithm to use or not to use in certain situation (Belanche and González, 2011). The feature selection algorithm can be classified according to the kind of output one are giving a (weighed) linear order of features and second are giving a subset of the original features. In this research, several fundamental algorithms found in the literature are studied to assess their performance in a controlled scenario. This measure computes the degree of matching between the output given by a FSA and the known optimal solution. Sample size effect also studied. The result illustrates the strong dependence on the particular conditions of the FSA used and on the amount of irrelevance and redundancies in the data set description, relative to the total number of feature. This should prevent the use of single algorithm even when there is poor knowledge available about the structure of the solution. The basic idea in feature selection is to detect irrelevant and/or redundant features as they harm the learning algorithm performance (Lee and Moore, 2014). There is no unique definition of relevance, however it has to do with the discriminating ability of a feature or a subset to distinguish the different class labels (Dash and Liu, 1997). However, as pointed out in the paper (Guymon and Elisseeff, 2003a), an irrelevant variable may be useful when taken with others and even two irrelevant variables that are useless by themselves can be useful when taken together.

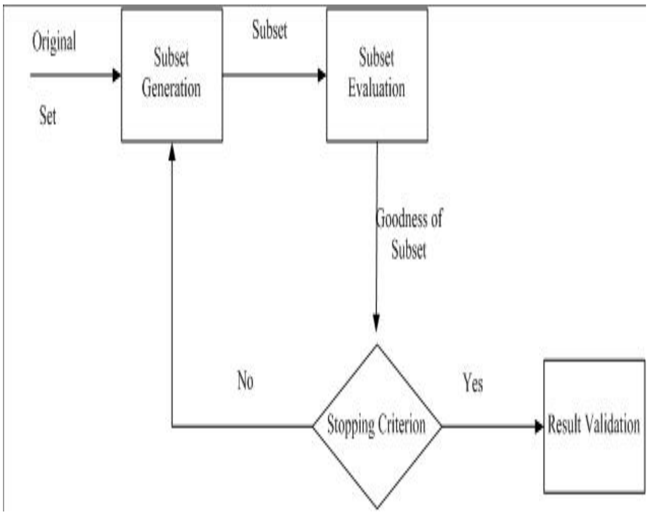

Figure 1. Feature Selection Criteria

II. THE FEATURE SELECTION PROBLEM

Let X be the original set of features which cardinality |X| = n. The continuous feature selection problem (also called feature weighing) refers to the assignment of weights w_i to each feature $x_i \in X$ in such a way that the order corresponding to its theoretical relevance is preserved. The binary feature selection problem (also called feature subset selection) refers to the choice of a subset of feature that jointly maximizes a certain measure related to subset relevance. This can be carried out directly as many FSA (Almuallim and Dietterich, 1991: Caruana and Freitag, 1994) or setting a cut point in the output of this continuous problem solution. Although both types can be seen in a unified way (the latter case corresponds to the assignment of weights in {0, 1}), these are quite different problems that reflect different design objectives. In the continuous case, one is interested in keeping all the features but in using them differentially in the learning process. On the contrary in the binary case one is interested in keeping just a subset of the features and (most likely) using them equally in the learning process.

A common instance of the feature selection problem can be formally stated as follows. Let J be a performance evaluation measure to be optimized (say to maximize) defined as $J: P(X) \rightarrow R^{+} \cup \{0\}$. This function accounts for a general evaluation measure that may or may not be inspired in a precise and previous definition of relevance. Let $C(x) \geq 0$ represent the cost of variable x and call $C(X')$ $= \sum_{x \in X'} c(x)$ for $X' \in p(X)$. Let $C_{x=} C(X)$ be the cost of the whole feature set. It is assumed here that c is additive, that is, $C(X' \cup X'') = C(X') + C(X'')$ (Belanche and González ,2011).

A. *Relevance of a Feature*

The purpose of a FSA is to identify relevant feature according to a definition of relevance. However, the notion of relevance in machine learning has not yet been rigorously defined on a common agreement (Bell and Wang, 2000). Let E_i, with $1 \leq i \leq n$, be domains of feature $X= \{x_{1 \ldots\ldots,} x_n\}$: an instance space is defined as $E= E_1 \times \ldots\ldots \times E_n$. where an instance is a point in this space. Consider P a probability distribution on E and T a space of labels (classes). It is desired to model or identify an objective function c: $E \longrightarrow T$ according to its relevant feature. A data set S composed by $|S|$ instances can be seen as the result of sampling E under p a total of $|S|$ times and labeling its element using c.

A Primary definition of relevance(Blum and Langley, 1997) is the notion of being " relevant with respect to an objective". It is assumed here to be classification objective.

Definition 1 (Relevance with respect to an objective)

A feature $x_i \in X$ is relevant to an objective c if there exist two examples A, B in the instance space E such that A and B differ only in their assignment to x_i and $c(A) \neq c(B)$. In other words, if there exist two instances that can only be classified thanks to x_i . This definition has the inconvenience that the learning algorithm can not necessarily determine if a feature x_i is relevant or not, using only a sample S of E. Moreover, if a problem representation is redundant (e.g. some features are replicated), it will never be the case that two instance differ only in one feature. A proposal oriented to solve this problem (John et al., 1994) include two notions of relevance, one with respect to a sample and another with respect to distribution.

Definition 2 (Strong relevance with respect to S)

A feature $x_i \in X$ is strongly relevant to the sample S if there exist two examples $A, B \in S$ that only differ in their assignment to x_i and $c(A) \neq c(B)$. That is to say, it is the same definition 1, but now $A, B \in S$ and the definition is respect to S.

Definition 3 (Strong relevance with respect to P)

A feature $x_i \in X$ is strongly relevant to an objective c in the distribution p if their exist two examples $A,B \in E$ with $p(A) \neq 0$ and $p(B) \neq 0$ that only differ in their assignment to x_i and $c(A) \neq c(B)$.

This definition is natural extension of definition 2 and contrary to it, the distribution p is assumed to be known.

Definition 4 (Weak relevance respect to S)

A feature $x_i \in X$ is weakly relevant to the sample S if there exist a proper $X' \supset X$ ($x_i \in X'$) where x_i is strongly relevant with respect to S. A weakly relevant feature can appear when a subset containing at least one strongly relevant feature is removed.

Definition 5 (Relevance as a complexity measure) (John et al., 1994)

Given a data sample S and an objective c, define $r(S,c)$ as the smallest number of relevant feature to c using Definition 1 only in S, and such that the error in S is the least possible for the inducer.In other words, it refers to the smallest number of features required by a specific inducer to reach optimum performance in the task of modeling c using S.

Definition 6 (Incremental Usefulness) (Caruana and Freitag, 1994)

Given a data sample S , a learning algorithm L, and subset of feature X', the feature x_i is incrementally useful to L with respect to X' if the accuracy of the hypothesis that L produces using the group of features $\{ x_i \} \cup X'$ is better than the accuracy reached using only the subset of features X'. This definition is especially natural n FSAs that search in the feature subset space in an incremental way, adding or removing features to a current solution. It is also related to a traditional understanding of relevance in the philosophy literature.

Definition 7 (Entropic relevance) (Wang, Bell, and Murtagh, 1998)

Denoting the Shannon entropy by $H(x)$ and the mutual information by $I(x;y) = H(x)- H(x|y)$ (the difference of entropy in x generated by the knowledge of y), the entropic relevance of x to y is defined as $r(x:y) = I(x:y)|H(y)$. Let X be the original set of feature and let C be the objective seen as a feature, a set $X' \supset X$ is sufficient if $I(X':C) = I(X,C)$. For a sufficient set X' it turns out that $r(X';C) = r(X;C)$. The most favorable set is that sufficient set $X' \supset X$ for which $H(X')$ is smaller. (Molina, Belanche, and Nebot, 2002)

B. *Feature Selection*

The main objective of feature selection are that it reduces the dimensionality of feature space, speedup and reduce the cost of learning algorithms, improve the predictive accuracy of classification algorithm, and also improve the visualization and the comprehensibility of the induced concepts. The feature selection algorithm may be based on three major criterions such as based on some evaluation measure, based on search organization and based on the generation of successors (Guyon and Elisseeff, 2003a).

TABLE1.

FEATURE SELECTION METHODS CHARACTERIZATION BASED ON DIFFERENT CRITERION AND THEIR TYPES

Characterizatio	**Types**
Evaluation measure	Distance
	Divergence
	Information
	Dependenc
	Accuracy
Search Organization	Exponential
	Sequential
	Random
Generation of successors	Forward
	Backward
	Compound
	Random
	Weight

C. *General Methods for Feature Selection*

The relationship between a FSA and the inducer chosen to evaluate the usefulness of the feature selection process can take three main forms such as Filter, Wrapper and Embedded.

1) *Filter Methods*

These methods select features based on discriminating criteria that are relatively independent of classification. Several methods use simple correlation coefficients similar to Fisher's discriminant criterion. Others adopt mutual information or statistical tests (t-test, F-test). Earlier filter-based methods evaluated features in isolation and did not consider correlation between features. Recently, methods have been proposed to select features with minimum redundancy. The methods proposed use a minimum redundancy-maximum relevance (MRMR) feature selection framework. They supplement the maximum relevance criteria along with minimum redundancy criteria to choose additional features that are maximally dissimilar to already identified ones. By doing this, MRMR expands the representative power of the feature set and improves their generalization properties (Guyon and Elisseeff,2003a).

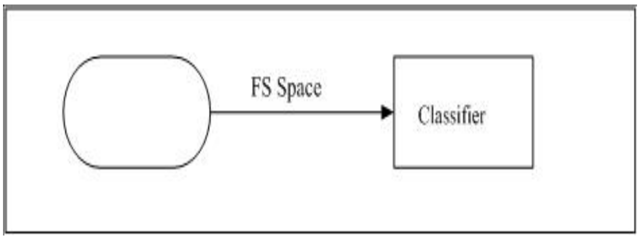

Figure 2. Filter Methods

2) *Wrapper Methods*

Wrapper methods utilize the classifier as a black box to score the subsets of features based on their predictive power. Wrapper methods based on SVM have been widely studied in machine-learning community. SVM-RFE (Support Vector Machine Recursive Feature Elimination), in each recursive step, it ranks the features based on the amount of reduction in the objective function. It then eliminates the bottom ranked feature from the results. A number of variants also use the same backward feature elimination scheme and linear kernel (Kohavi and John, 1997a).

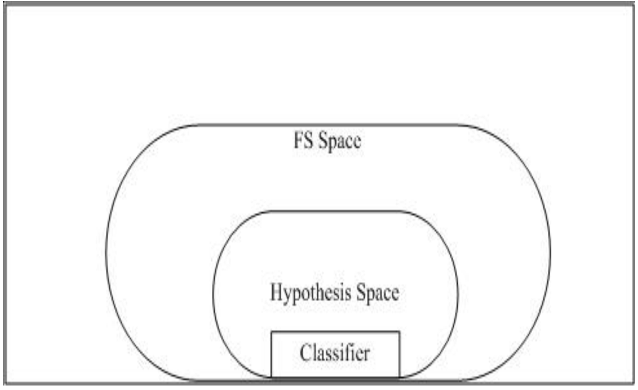

Figure 3. Wrapper Methods

3) *Embedded Method*

The inducer has its own FSA (either explicit or implicit). The methods to induce logical conjunctions provide an example of this embedding. Other traditional machine learning tools like decision trees or artificial neural networks are included in this scheme (Guyon and Elisseeff, 2003a)

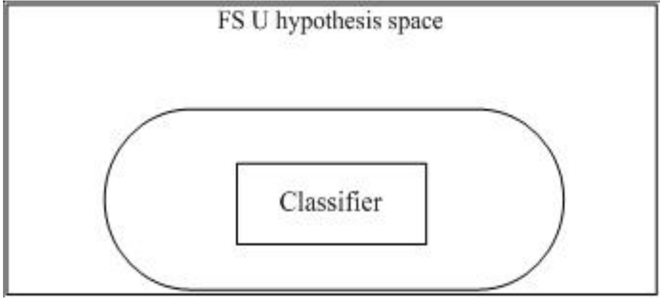

Figure 4. Embedded Methods

D. *Filter Based Feature Selection Method*

Filter based feature selection methods may be broadly categorized into two categories-:

1) Supervised

In supervised learning, the data is assigned to be known before computation and are used in order to learn the parameters that are really significant for clusters. Each object in supervised learning comes with a pre assigned class label.

2) Unsupervised

In unsupervised learning the datasets are assigned to segments without the cluster being known. Supervised and Unsupervised learning approaches further classified in univariate and Multivariate data. In univariate data analysis it is assumed that the response variable is influenced only by one other factor whereas in multivariate data analysis it is assumed that the response variable is influenced by multiple factors and even combination of factors (Guyon and Elisseeff, 2003a). Classification of filter based feature selection methods on the basis of supervised and unsupervised learning is shown below in Table-2, evaluation function used by filter based feature selection method is shown in Table-3 and brief description of filter based feature selection methods is shown in Table-4.

TABLE 2

CLASSIFICATION OF FILTER BASED FEATURE SELECTION METHODS ON THE BASIS OF SUPERVISED AND UNSUPERVISED LEARNING

Filter based feature selection methods	**Supervised**		**Unsupervised**	
	Univariate	**Multivariate**	**Univariate**	**Multivariate**
Relief F(Robnik-Šikonja and Kononenko, 2003)	No	Yes	No	No
mRmR(Peng, Long, and Ding, 2005)	No	Yes	No	No
FCBF(Yu and Liu, 2003)	No	Yes	No	No
Fisher score(Duda, Hart, and Stork, 2001)	Yes	No	No	No
SVM-RFE(Furey et al., 2000)	No	Yes	No	No
t-test(Duda, Hart, and Stork, 2001)	No	No	Yes	No
Entropy based(Duda, Hart, and Stork, 2001)	No	No	Yes	No
Laplacian Score(He, Cai, and Niyogi, 2005)	No	No	Yes	No
PCA(Duda, Hart, and Stork, 2001)	No	No	No	Yes

TABLE 3
EVALUATION FUNCTION USED BY FILTER BASED FEATURE SELECTION METHOD

Basic Criterion/ Evaluation function used	**Examples**
Distance based measures	Euclidean distance
Information theory based measure	Entropy, Information gain, mutual information
Data dependency measure	Correlation coefficient
Consistency based measure	Minimum feature bias

TABLE 4

BRIEF DESCRIPTION OF FILTER BASED FEATURE SELECTION METHODS

Filter based feature selection method	**Basic criterion**	
Supervised feature selection method	Fisher Score (Duda, Hart, and Stork, 2001)	Distance based, univariate filter method evaluating each feature individually
	Relief F (Robnik-Šikonja and Kononenko, 2003)	A multivariate filter taking into account dependencies between features
	mRmR (Peng, Long, and Ding, 2005)	Information theory based uses mutual information
	FCBF (Yu and Liu, 2003)	Based on information gain, Fast correlation based filter
	SVM-RFE (Furey et al., 2000)	Ranks features based on their coefficients in the SVM classifier.
Unsupervised rank based feature selection methods	t-test score (Duda, Hart, and Stork, 2001)	Statistical, rank based feature selection approach
	Bhattacharya distance	
	Entropy rank feature	
	Principal component analysis (PCA)	PCA finds a linear projection of high dimensional data into lower dimensional subspace
	Laplacian score based	It is unsupervised feature selection algorithm. It is based on Laplacian Eigen maps.

E. WRAPPER BASED FEATURE SELECTION METHOD

Wrapper methods (Hall, 1999) are feedback methods which merge the machine learning algorithm in the feature selection process. Wrapper method search through the space of feature subset using a learning algorithm to guide the search. A search algorithm “wrapped” around the classification model. In search procedure the space of possible feature subset is defined and generated various subsets of features. Wrapper method can be divided in two groups these are deterministic and wrapper methods.

Deterministic wrapper method

This method search through the space of available feature either forward or backward. In forward selection single attribute are added to initially an empty set of attributes.

Randomized wrapper method

Randomized wrapper algorithms search the next feature subset partly at random. Single feature or several features at once can be added, removed or replaced from various feature set. The brief descriptions of wrapper based feature selection methods is shown in Table-5.

TABLE 5

BRIEF DESCRIPTION OF WRAPPER BASED FEATURE SELECTION METHODS

Wrapper	**Deterministic**	Simple, Interact with the classifier, Models feature dependencies, Less computationally intensive than randomized method	Risk of over fitting, More prone than Randomized algorithm to getting stuck in a local optimum(Greedy Search), Classifier dependent Selection	Sequential forward selection(SFS), Sequential backward elimination(SBE), Plus L Minus R, Beam Search
	Randomized	Less prone to local optima, Interact with the classifier, Model feature dependencies	Computationally Intensive, classifier dependent selection, High risk of over fitting than deterministic algorithms	Simulated annealing, Randomized hill climbing, Genetic algorithms

F. EMBEDDED FEATURE SELECTION METHOD

Embedded method (Saeys, Inza, and Larrañaga, 2007) sometime also referred as nested subset method. It acts as an integral part of machine learning algorithm itself. During the operation of classification process, the algorithm itself decides which attribute to use and which to ignore. Embedded approach depends on a specific learning algorithm. Embedded methods are faster than wrapper methods. Decision trees are the best example of embedded method.

Feature Selection Algorithms

1) CHI (X^2Statistics)

This method measure the lack of independence between a term and category. CHI-Squared is the common statistical test that measures divergence from the distribution expected if one assumes the feature occurrence is actually independent of class value. The X^2 test is applied to test the independence of two events, where two events A and B are defined to be independent if $P(AB) = P(A)P(B)$ or equivalently $P(A|\ B) = P(A)$ and $P(\ B|A) = P(B)$.(Liu and Setiono 1995) Feature selection using the X^2 statistics is analogous to performing a hypothesis test on the distribution of the class as it relates to the values of the feature in question. Under the null hypothesis, if p of the instance have a given value and q of the instances are in a specific class, $(p.q)/n$ instances have a given value and are in a specific class(n is the total number of instances in the data set)(Liu and Motoda, 2007). This is because p/n instances have the value and q/n instances are in the class, and if the probabilities are independent their joint probability is their product. Given the null hypothesis, the X^2 static measure how far away the actual value is from the expected value.

2) Euclidian Distance

Euclidian Distance is the most common use of distance. In most cases when we talk about distance refer to Euclidian Distance. It examines the root of square differences between coordinates of a pair of object. For each feature X_i calculate Euclidian distance from it to all other features in sample. Euclidian distance $d(X_i\ ;\ Y_i)$ between features X_i and Y_i is calculated using the formula $distance(x,y) = \{\Sigma i\ (x_i - y_i)2\ \}½$ (Dash and Liu, 1997)**.**

This distance generally computed from raw data and not from standardized data.

3) t-Test

The t-test assesses whether the means of two groups are statistically different from each other. This analysis is appropriate whenever you want to compare the means of two groups, and especially appropriate as the analysis for the posttest-only two-group randomized experimental design. The formula for the t-test is a ratio. The top part of the ratio is just the difference between the two means or averages. The bottom part is a measure of the variability or dispersion of the scores(Guyon and Elisseeff, 2003b).

Information Gain

Information gain, of a term measures the number of bits of information obtained for category prediction by the presence or absence of the term in a document. Information Gain measures the decrease in entropy when the feature is given vs absent. This is the application of a more general technique, the measurement of informational entropy, to the problem of deciding how important a given feature is(Kira and Rendell, 1992). Informational entropy, when measured using Shannon entropy, is notionally the number of bits of data it would take to encode a given piece of information. The more space a piece of information takes to encode, the more entropy it has. Intuitively, this makes sense because a random string has maximum entropy and cannot be compressed, while a highly ordered string can be written with a brief description of the string's information. In the context of classification, the distribution of instances among classes is the information in question. If the instances are randomly assigned among the classes, the number of bits necessary to encode this class distribution is high, because each instance would need to be enumerated. On the other hand, if all the instances are in a single class, the entropy would be lower, because the bit-string would simply say "All instances save for these few are in the first class."(Joachims 1998) Therefore function measuring entropy must increase when the class distribution gets more spread out and be able to applied recursively to permit finding the entropy of subsets of the data. The following formula satisfies both of these requirements:

$H(D) = - \Sigma\, (n_i/n) \log(n_i/n)\ i = 1, \ldots l$

where dataset D has $n = | D |$ instances and n_i members in class c_i, $i = 1, \ldots, l$.

The entropy of any subset is calculated as:

$H(D|X) = - \Sigma\, (|Xj|/n) H(D|X\text{-}Xj)$

where $H(D|X = Xj)$ is the entropy calculated relative to the subset of instances that have a value of Xj for attribute X. If X is a good description of the class, each value of that feature will have little entropy in its class distribution; for each value most of the instances should be primarily in one class. The information gain of an attribute is measured by the reduction in entropy (Kira and Rendell, 1992) defined as

$IG(X) = H(D) - H(D|X)$

The greater the decrease in entropy when considering attribute X individually, the more significant feature X is for prediction.

Correlation based feature selection (CFS)

Correlation based feature selection (CFS) searches feature subsets according to the degree of redundancy among the features. The evaluator aims to find the subsets of features that are individually highly correlated with the class but have low inter-correlation(Hall, 1999). The subset evaluators use a numeric measure, such as conditional entropy, to guide the search iteratively and add features that have the highest correlation with the class(Saeys, Inza, and Larrañaga, 2007). The downside of univariate filters for example information gain is, it does not account for interactions between features, which is overcome by multivariate filters for example CFS. CFS evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Correlation coefficients are used to estimate correlation between subset of attributes and class, as well as inter-correlations between the features. Relevance of a group of features grows with the correlation between features and classes, and decreases with growing inter-correlation. CFS is used to determine the best feature subset and is usually combined with search strategies such as forward selection, backward elimination, bi-directional search, best-first search and genetic search(Yu and Liu, 2004).

Equation for CFS is given:

$$r_{zc} = \frac{k\overline{r_{zi}}}{\sqrt{k + k(k-1)\overline{r_{ii}}}}$$

Where r_{zc} is the correlation between the summed feature subsets and the class variable, k is the number of subset features, r_{zi} is the average of the correlations between the subset features and the class variable, and r_{ii} is the average inter-correlation between subset features.

1) Fast Correlation based Feature Selection (FCBF)

Fast Correlation based Feature Selection (FCBF) (Yu and Liu, 2003) uses also the symmetrical uncertainty measure. But the search algorithm is very different. It is based on the "predominance" idea. The correlation between an attribute X^* and the target Y is predominant if and only if $\rho_{y,x}{}^{*} \geq \delta$et for all $X(X \neq X^*)$, $\rho_{X,X}{}^{*} < \rho_{Y},X^*$

Concretely, a predictor is interesting if its correlation with the target attribute is significant (delta is the parameter which allows to assess this one); there is no other predictor which is more strongly correlated to it.

Algorithm for FCBF

1. S is the set of candidate predictors, $M = \emptyset$ is the set of selected predictors
2. Searching X^* (among S) which maximizes its correlation with $Y \rightarrow \rho y,x^*$
3. If $\rho y,x^* \geq \delta$ add X^* into M and remove X^* from S
4. Remove also from S all the variables X such $\rho x,x^* \geq \rho y,x^*$
5. If $S \neq \emptyset$ then GOTO (2), else END of the algorithm

This approach is very useful when we deal with a dataset containing a very large number of candidate predictors. About the ability to detect the "best" subset of predictors and it is similar to CFS.

2) Sequential forward selection (SFS)

Sequential Forward Selection(Jain and Zongker, 1997) is the simplest greedy search algorithm. Starting from the empty set, sequentially add the feature x+ that results in the highest objective function $J(Y_k\text{+x+})$ when combined with the features Y_k that have already been selected.

Algorithm

1.Start with the empty set $Y_0=\{\phi\}$

2.Select the next best feature X^+ = argmax $[J(Y_k\text{-}X)]$; $x ¢ Y_k$

3.Update $Y_{k+1} = Y_k + X^+; K = K+1$

4.Goto 2

SFS performs best when the optimal subset has a small number of features. When the search is near the empty set, a large number of states can be potentially evaluated. Towards the full set, the region examined by SFS is narrower since most of the features have already been selected. The search space is drawn like an ellipse to emphasize the fact that there are fewer states towards the full or empty sets. As an example, the state space for 4 features is shown. Notice that the number of states is larger in the middle of the search tree. The main disadvantage of SFS is that it is unable to remove features that become obsolete after the addition of other features.

3) *Sequential Backward Elimination (SBE)*

Sequential Backward Elimination(Mao, 2004) works in the opposite direction of SFS. Also referred to as SBS (Sequential Backward Selection). Starting from the full set, sequentially remove the feature *x*− that results in the smallest decrease in the value of the objective function *J(Y-x*−). Notice that removal of a feature may actually lead to an increase in the objective function *J(Yk-x*−)>J(*Yk*). Such functions are said to be non-monotonic.

Algorithm

```
1. Start with the full set $Y_0 = X$
2. Remove the worst feature $X^- =$ argmax $[J(Y_k-X)]; x\ Y_k$
3. Update $Y_{k+1}=Y_k- X^-$ ; $k=k+1$
4. Goto 2
```

SBS works best when the optimal feature subset has a large number of features, since SBS spends most of its time visiting large subsets. The main limitation of SBS is its inability to reevaluate the usefulness of a feature after it has been discarded.

III. CONCLUSIONS

This paper presented an empirical comparison of feature selection methods and its algorithm. In view of the substantial number of existing feature selection algorithms, the need arises to count on criteria that enable to adequately decide which algorithm to use in certain situation. This paper also reviewed several fundamental algorithms found in the literature and assess their performance in a controlled scenario.

REFERENCES

[1] Almuallim, H., & Dietterich, T. G. (1991, July). Learning with Many Irrelevant Features. In AAAI (Vol. 91, pp. 547-552).

[2] Athanasakis, D., Shawe-Taylor, J., & Fernandez-Reyes, D. (2013). Principled Non-Linear Feature Selection. arXiv preprint arXiv:1312.5869.

[3] Belanche, L. A., & González, F. F. (2011). Review and evaluation of feature selection algorithms in synthetic problems. arXiv preprint arXiv:1101.2320.

[4] Bell, D. A., & Wang, H. (2000). A formalism for relevance and its application in feature subset selection. Machine learning, 41(2), 175-195.

[5] Caruana, R., & Freitag, D. (1994, July). Greedy Attribute Selection. In ICML(pp. 28-36).

[6] Craven, M., DiPasquo, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K., & Slattery, S. (2000). Learning to construct knowledge bases from the World Wide Web. Artificial intelligence, 118(1), 69-113.

[7] Dash, M., & Liu, H. (1997). Feature selection for classification. Intelligent data analysis, 1(3), 131-156.

[8] Ding, C., & Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. Journal of bioinformatics and computational biology, 3(02), 185-205.

[9] Duda, R. O., Hart, P. E., & Stork, D. G. (2001). Pattern classification. 2nd.Edition. New York.

[10] Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., & Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics,16(10), 906-914.

[11] Gu, Q., Li, Z., & Han, J. (2012). Generalized fisher score for feature selection.arXiv preprint arXiv:1202.3725.

[12] Gupta, P., Doermann, D., & DeMenthon, D. (2002). Beam search for feature selection in automatic SVM defect classification. In Pattern Recognition, 2002. Proceedings. 16th International Conference on (Vol. 2, pp. 212-215). IEEE.

[13] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. The Journal of Machine Learning Research, 3, 1157-1182.

[14] Hall, M. A. (1999). Correlation-based feature selection for machine learning(Doctoral dissertation, The University of Waikato).

[15] Jain, A., & Zongker, D. (1997). Feature selection: Evaluation, application, and small sample performance. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 19(2), 153-158.

[16] Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features (pp. 137-142). Springer Berlin Heidelberg.

[17] Kohavi, R., Sommerfield, D., & Dougherty, J. (1996, November). Data mining using &mscr; &lscr; &cscr;++ a machine learning library in c++. In Tools with Artificial Intelligence, 1996., Proceedings Eighth IEEE International Conference on (pp. 234-245). IEEE.

[18] Ladha, L., & Deepa, T. (2011). FEATURE SELECTION METHODS AND ALGORITHMS. International Journal on Computer Science & Engineering, 3(5).

[19] Lee, K., Joo, J., Yang, J., & Honavar, V. (2006). Experimental comparison of feature subset selection using GA and ACO algorithm. In Advanced Data Mining and Applications (pp. 465-472). Springer Berlin Heidelberg.

[20] Lee, M. S., & Moore, A. W. (2014, June). Efficient algorithms for minimizing cross validation error. In Machine Learning Proceedings 1994: Proceedings of the Eighth International Conference (p. 190). Morgan Kaufmann.

[21] Liu, H., & Motoda, H. (Eds.). (2007). Computational methods of feature selection. CRC Press.

[22] Liu, H., & Setiono, R. (1996, July). A probabilistic approach to feature selection-a filter solution. In ICML (Vol. 96, pp. 319-327).

[23] Mao, K. Z. (2004). Orthogonal forward selection and backward elimination algorithms for feature subset selection. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 34(1), 629-634.

[24] Masaeli, M., Dy, J. G., & Fung, G. M. (2010). From transformation-based dimensionality reduction to feature selection. In Proceedings of the 27th International Conference on Machine Learning (ICML-10) (pp. 751-758).

[25] Molina, L. C., Belanche, L., & Nebot, À. (2002). Feature selection algorithms: A survey and experimental evaluation. In Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on (pp. 306-313). IEEE.

[26] Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy.Pattern Analysis and Machine Intelligence, IEEE Transactions on, 27(8), 1226-1238.

[27] Reinartz, T. (1999). Focusing solutions for data mining: analytical studies and experimental results in real-world domains. Springer-Verlag.