Hybrid wrapper-filter approaches for input feature selection using Maximum relevance-Minimum redundancy and Artificial Neural Network Input Gain Measurement Approximation (ANNIGMA)

Shamsul Huda, John Yearwood, Andrew Stranieri
CIAO, GSITMS, University of Ballarat, Victoria, Australia
s.huda@ballarat.edu.au, j.yearwood@ballarat.edu.au, a.stranieri@ballarat.edu.au

Abstract

Feature selection processes improve the accuracy, computational efficiency and scalability of classification process in data mining applications. This paper proposes two filter and wrapper hybrid approaches for feature selection techniques by combining the filter's feature ranking score in the wrapper stage. The first approach hybridizes a Mutual Information (MI) based Maximum Relevance (MR) filter ranking heuristic with an Artificial Neural Network (ANN) based wrapper approach where Artificial Neural Network Input Gain Measurement Approximation (ANNIGMA) has been combined with MR (MR-ANNIGMA) to guide the search process in the wrapper. The second hybrid combines an improved version of MI based (Maximum Relevance and Minimum Redundancy; MaxRel-MinRed) filter ranking heuristic with the wrapper heuristic ANNIGMA (MaxRel-MinRed-ANNIGMA). The novelty of our approach is that we integrate the capability of wrapper approach to find better feature subset by combining filter's ranking score with the wrapper-heuristic's score that take advantages of both filter and wrapper heuristics. The performances of the hybrid approaches have been verified using synthetic, bench mark data sets and real life data set and compared to both independent filter and wrapper based approaches. Experimental results show that hybrid approaches (MR-ANNIGMA and MaxRel-MinRed-ANNIGMA) achieve more compact feature sets and higher accuracies than filter and wrapper approaches alone.

Keywords: Hybrid Feature Selection, Wrapper, Filter Maximum-Relevance, Maximum-Relevance and Minimum Redundancy, ANNIGMA wrapper,

1 Introduction

Feature selection is an important and frequently used data pre-processing technique in machine learning (Blum, Langely 1997), (John, Kohavi 1994), data mining (Dash, Liu 1997), medical data processing (Puronnen et al 2000) and statistical pattern recognition areas (Bne-Bassat 1982), (Mitra et al. 2002). Due to rapid advances of computational technologies and internet, datasets are getting larger and larger. To use datasets with thousands of features for decision making, prediction or classification purposes by using data mining techniques is a challenge for researchers and practitioners because the performance of data mining methodologies degrades with huge volumes of training data (Blum, Langely 1997), (John, Kohavi 1994), (Dash, Liu 1997). Therefore, feature selection from datasets by removing irrelevant, redundant or noisy features is a primary task for machine learning researchers. Given an m-dimensional dataset, a feature selection algorithm needs to find optimal feature subset from the 2^m subsets of the feature space. Therefore finding an optimal feature subset is computationally expensive (Kohavi et al. 1997). The performance of a feature selection algorithm depends on its evaluation criterion and search strategies.

Significant research works have appeared in the literature on feature selection. These can be grouped broadly into three main categories based on the evaluation criteria: I) the filter model (Dash, Liu 1997), (Kwak et al. 2002), (Wang et al. 1999), (Hall 2000) II) the wrapper model (Kwak et al. 2002) (J.G. Dy et al. 2000) (Hsu et al. 2002) (Dash, Liu 1997)) and III) hybrid models (Zhu et al. 2002). The filter models are based on the intrinsic characteristics of the data and do not involve the application of an induction algorithm. Filter models are computationally cheap due to its evaluation criteria. However, feature subsets selected by filter may result in poor prediction accuracies, since they are independent from the induction algorithm. In contrast, the wrapper model (Kwak et al. 2002) (J.G. Dy et al. 2000) (Hsu et al. 2002) uses a predetermined induction algorithm and uses predictive accuracy as the evaluation criteria for the feature selection. However, wrapper models face huge computational overhead due to the use of the induction algorithm's performance criteria as its evaluation criteria. In (Hsu et al. 2002), (Kohavi et al. 1997), a hybrid of genetic algorithm and filter heuristic were proposed where GA framework works as subset generation process and filter heuristic improves local search. Despite significant researches on evaluation criteria and search strategies, current generation feature selection literature lacks the work that can combine the merit of wrapper and filter approaches.

In this paper, we propose a hybrid wrapper and filter approach by using the filter's feature ranking score with the wrapper heuristics in the wrapper stage to speed up the search process and find optimal feature subset in the wrapper stage. In our approach, we hybridize two novel filter heuristics with Artificial Neural Network Input Gain Measurement Approximation (ANNIGMA) wrapper heuristic. The first proposed hybrid feature selection approach uses Mutual Information (MI) based Maximum Relevance (MR) filter ranking heuristics with

ANNIGMA. The second hybrid approach uses an improved version of MI-based filter heuristic Maximum Relevance and Minimum Redundancy; MaxRel-MinRed) with the ANNIGMA (MaxRel-MinRed-ANNIGMA). The novelty of our approach is that we use a wrapper and filter hybrid that combines the filter's ranking score with the wrapper-heuristic's score to guide the search process in the wrapper stage. The proposed approaches avoid the computational overhead of hybrid GA-based approaches (Hsu et al. 2002), (Kohavi et al. 1997) and takes advantage of both filter and wrapper heuristics which are absent in the traditional GA-based hybrid approaches (Hsu et al. 2002), (Kohavi et al. 1997). This type of hybrid approach is a new concept and has not been explored yet in the literature.

The rest of the paper is organized as follows. The next section introduces some related literature. The proposed hybrid of wrapper-filter feature selection algorithm using the combination of filter heuristic Maximum-Relevance-Minimum-Redundancy (MaxRel-MinRed) and Artificial Neural Network Input Gain Measurement Approximation (ANNIGMA) is described in Section 3. It also discusses the hybrid approach using Maximum-Relevance (MR) and ANNIGMA. Section 4 presents experimental results and discussion. Conclusions of this study are presented in the last section.

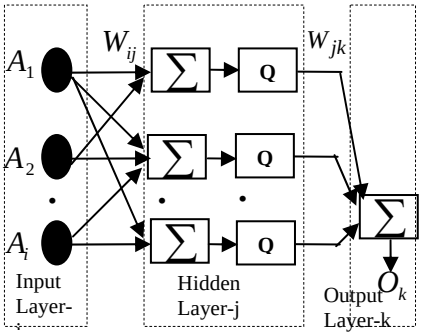

Figure 1. A single hidden layer Multi layer Perceptron (MLP) neural network in wrapper approach

2 Related Work.

2.1 Generalized Filter and wrapper approaches

Filter approaches start from an initial subset either empty or full and generate a new subset each step by following a search strategy (either forward or backward or bi-directional) through the feature space. Each generated subset is evaluated using filter heuristics. If a current subset has a higher evaluation score than previous, it is assigned as the current best subset. The search process stops on a user defined stopping criteria based on the score and number of optimal feature set. The final subset can be justified further using an induction algorithm. In the wrapper approach (Kwak et al. 2002) (J.G. Dy et al. 2000) (Hsu et al. 2002) generated subsets are evaluated using a predetermined induction algorithm. However, subsets in the wrapper approach are evaluated by the predictive accuracies of a trained classifier, therefore are more significant than those in the filter approach. In the Neural network based wrapper literature, search process in the wrapper can also guided by a wrapper heuristic such as Artificial Neural Network Input Gain Measurement Approximation (ANNIGMA) (Hsu et al. 2002) which shows significant improvement over the wrapper alone.

2.2 Artificial Neural Network Input Gain Measurement Approximation (ANNIGMA)

Artificial Neural Network Input Gain Measurement Approximation (ANNIGMA)(Hsu et al. 2002) is a weight analysis based wrapper heuristic that ranks features by relevance based on the weight associated with feature in a Neural Network based wrapper approach. Features that are irrelevant or redundant will produce more error than relevant features. Therefore, during training, weights of noisy features are controlled in such a way that they contribute to the output as least as possible. ANNIGMA (Hsu et al. 2002) is based on the above strategy of the training algorithm. For a two layer Neural Network, (Figure 1) if i, j, k are the input, hidden and output layer and Q is a logistic activation function (1) of the first layer and second layer has a linear function, then output of the network is as (2).

$$Q(x) = (1/(1+\exp(-x))) \quad (1)$$

$$O_k = \sum_j Q\left(\sum_i A_i \times W_{ij}\right) \times W_{jk} \quad (2)$$

Then local gain is defined as (3)

$$LG_{ik} = \frac{\Delta O_k}{\Delta A_i} \quad (3)$$

According to C.N. Hsu and H.J. Huang et.al. (Hsu et al. 2002), the local gain can be written in terms of network weight as (4):

$$LG_{ik} = \sum_j |W_{ij} \times W_{jk}| \quad (4)$$

Then ANNINGMA score for feature-i (F_i) is the local gain (LG) normalized based on a unity scale as (5)

$$ANNIGMA(F_i) = \frac{LG_{ik}}{\max imum\ (i)\, LG_{ik}} \quad (5)$$

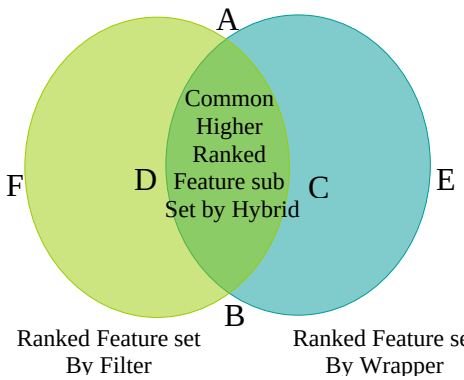

Figure 1.1 Venn diagram for combined heuristics

3 Hybrid feature selection algorithms using Maximum Relevance and Minimum Redundancy Filter Heuristic and Artificial Neural Network Input Gain Measurement Approximation Wrapper Heuristics

Standard filter approaches can extract knowledge of the intrinsic characteristics from real data. However filter

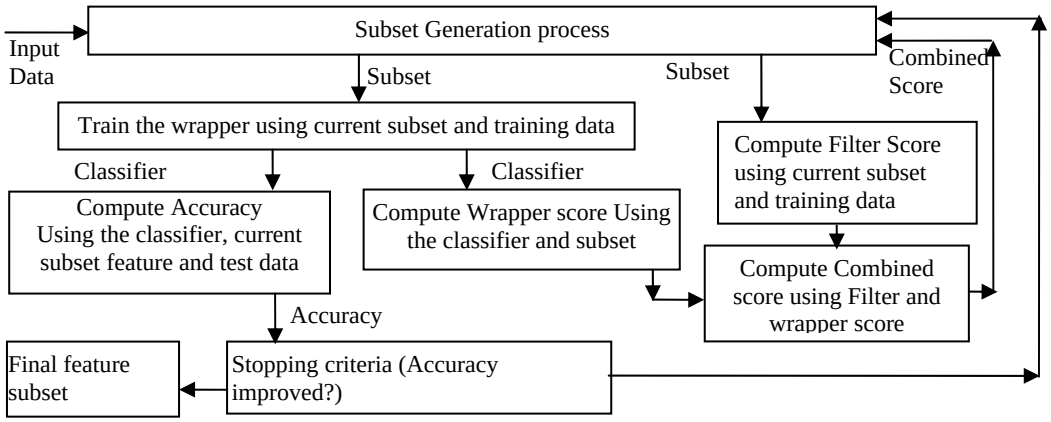

Figure 2. General Framework for proposed hybrid of wrapper and Filter feature selection approaches

approaches do not use any performance criteria based on predictive accuracies. This does not guarantee that final feature subset will do better in classification/prediction tasks. In contrast, the wrapper approaches (Kwak et al. 2002), (J.G. Dy et al. 2000), (Hsu et al. 2002) use a predetermined induction algorithm and different search strategies (Jain et al. 1997), (Ferri et al. 1994) to find the best feature subset. Use of predictive-accuracy based evaluation criteria in the wrapper ensures good performance from the selected feature subset. However repeated execution of the induction algorithm (in the worst case exponential search space) in the search process incurs a high computational cost in the wrapper approach.

In this paper, proposed hybrid approaches introduce the filter heuristic in the wrapper stage and take advantages of both approaches which is able to find more significant features than either wrapper and filter alone. The idea behind this approach can be explained by the Venn-diagram (in figure 1.1.). If the two feature subsets (ACBF and ADBE figure 1.1) are separately ordered/ranked according to their score, then common higher ranked feature subset (ACBD) is the strongly recommended most significant feature subset by the both feature selection algorithms. If the scores of both algorithms are normalized on the same scale and combined (summed), then feature subsets with higher combined scores provide the common higher ranked feature subset from both algorithms. A Backward Elimination (BE) search strategies based on the combined score along with the wrapper evaluation criteria can find the most significant features. Performance of the combined score may be affected due to performance of the incorporated filter for a particular wrapper approach in the hybrid. However, different filter approaches can be combined to find a suitable hybrid for a particular wrapper heuristic and vice-versa. In this paper, we have combined two filter heuristics: mutual information based Maximum Relevance (MR), Maximum Relevance-Minimum Redundancy (MaxRel-MinRed) with Artificial Neural Network Input Gain Measurement Approximation (ANNIGMA) based wrapper. Here, we have focused on a Neural network based wrapper and different filter heuristics. We will use other wrapper approaches in a future work. The following sub-sections describe different heuristics and steps of the proposed hybrid algorithms.

3.1 Maximum Relevance (MR)

Relevant features provide more information about the class variable than irrelevant features. Therefore mutual information based maximum relevance (Wang et al. 1999) is a good heuristic to select salient features in data mining area. If S is a set of features F_i and class variable is c, the maximum relevance (Wang et al. 1999) can be defined as (6).

$$\max imum\ \mathrm{Re}levance(S,c) = \frac{1}{|S|}\sum_{f_i \in S} I(F_i;c) \quad (6)$$

$I(F_i;c)$ is the mutual information between F_i and class c which is defined as (7).

$$I(F_i;c) = H(F_i) - H(F_i \mid c) \quad (7)$$

$H(F_i)$ is the entropy of F_i with the probability density function $p(f_i)$ where F_i takes discrete values from the set $F = \{f_1, f_2 \ldots f_i\}$, then $H(F_i)$ is defined as (8)

$$H(F_i) = -\sum_{f_i \in F} p(f_i)\log p(f_i) \quad (8)$$

$H(F_i \mid c)$ in (2) is the conditional entropy between F_i and c and is defined as (9)

$$H(F_i \mid c) = -\sum_{f_i \in F}\sum_{c_i \in C} p(f_i, c_i)\log p(c_i \mid f_i) \quad (9)$$

where class variable c takes the discrete values from the set $C = \{c_1, c_2, \ldots c_i\}$.

3.2 Maximum Relevance and Minimum Redundancy

Maximum relevance (MR) (Wang et al. 1999) can select features that are highly relevant to class. However MR may contribute to redundancy. When two features are highly dependent on each other, the corresponding class discriminative ability of the two features would not be affected much if one of them were removed. Therefore, to avoid the redundancy in MR, a redundancy function is incorporated with maximum relevance as (10)

$$MR - Minimum\ \mathrm{Re}dundancy = \max s\frac{1}{|S|}\sum_{f_i \in S} I(F_i;c) - Minimum\ \mathrm{Re}dundancy \quad (10)$$

Minimum Redundancy is defined as (11)

$$(\mathrm{Re}d, c) = \frac{1}{|S|^2}\sum_{i,j \in S} I(F_i;F_j) \quad (11)$$

Where $I(F_i;F_j)$ is the mutual information between the features F_i *and* F_j.

3.3 Computation of combined score in proposed hybrid algorithm-1: Hybrid of Maximum Relevance and ANNIGMA (MR-ANNIGMA)

The proposed MR-ANNIGMA uses Artificial Neural Network as the classification algorithm in the wrapper stage. An n-fold cross-validation approach has been used in MR-ANNIGMA to train the wrapper. In each fold we compute the ANNIGMA score for every feature. Then after training of all folds, the ANNIGMA score is averaged as (12):

$$ANNIGMA(F_i)_{average} = \left(\frac{1}{n}\right)(ANNIGMA(F_i)_1 + .. + ANNIGMA(F_i)_n) \quad (12)$$

While computing the combined score in the proposed ANNIGMA, the relevance of a feature in the current subset is computed from the individual score which is scaled to the maximum individual relevance of the subset. Thus relevance of a feature in a subset in the hybrid approach is as (13)

$$\text{Re}levance(F_i) = \frac{I(F_i;c)}{\max imum\ (f_i \in S)\ \ I(F_i;c)} \quad (13)$$

The combined score of filter's heuristic and wrapper's heuristic in the proposed MR-ANNIGMA is computed as (14).

$$Combined\ \ Score(MR-ANNIGMA) = \frac{I(F_i;c)}{\max imum\ (f_i \in S)\ \ I(F_i;c)} + ANNIGMA(F_i)_{average} \quad (14)$$

3.4 Computation of combined score in proposed hybrid algorithm-2: Hybrid of Maximum Relevance-Minimum-Redundancy and ANNIGMA (MaxRel-MinRed-ANNIGMA).

As the MR-ANNIGMA, the proposed (MaxRel-MinRed –ANNIGMA) uses ANN as the wrapper. The ANNIGMA score is computed as (12). An incremental search method (Peng et al. 2005) is used to compute the Maximum Relevance and Minimum Redundancy score as (15). Maximum Relevance and Minimum Redundancy (MaxRel-MinRed) score is the difference of maximum relevance score of a candidate features in the candidate set and redundancy score between the corresponding feature with a feature in the goal set.

$$\text{MaxRel\_MinRed Score} = \max_{F_i \in F-F_{l-1}} \left(\frac{1}{|S|} \sum_{f_i \in S} I(F_i;c) - \frac{1}{l-1} \sum_{F_j \in F_{l-1}} I(F_i;F_j) \right) \quad (15)$$

Since the MaxRel-MinRed score is a difference of feature score which is relative to the search iteration, while computing the combined score in the hybrid, an equivalent weighted score of MaxRel-MinRed score is computed for each feature. First the features are ordered according to their ranks in the MaxRel-MinRed incremental search method (Peng et al. 2005). Then equivalent weighted score is computed form their ranking on a unity scale. Orders of the feature ranking are incremental integers starting from one to total number of features in the data set where top-ranked has a maximum score of one. Therefore, equivalent weighted MaxRel-MinRed score is as (16)

$$\text{Weighted MaxRel\_Min Red Score}(F_i) = 1 - (Rank\ \ feature\ (F_i)\ \ in\ \ \text{MaxRel\_Min Red} / |F|) \quad (16)$$

The combined score of filter's heuristic and wrapper's heuristic in the proposed (MaxRel-MinRed –ANNIGMA) is computed as (17).

$$Combined\ \ Score(\text{MaxRel\_MinRed ANNIGMA}; F_i) = \text{Weighted MaxRel\_MinRed Score}(F_i) + ANNIGMA(F_i)_{average} \quad (17)$$

Algorithm-1 and 2: Procedure (Hybrid Wrapper-Filter approach)

Input: $D(F_1, F_2, ...F_m)$ // Training data with m features

Output: S_{BEST} //an optimal subset of features

Begin

1. Let S=whole set of m features $F_1, F_2, ...F_m$

2. S_0 =Initial set of feature which records all generated subsets with accuracy

// Apply a BE search strategy

3. for N = 1 to m-1

4. Current set of feature $S_{current}$ =S

5. Compute Filter score by (6) and (10)

6. for fold=1 to n

7. Train the network with $S_{current}$

8. Compute ANNGMA of all features

9. Compute Accuracy

10. endfor

11. Compute average accuracy of all folds for $S_{current}$

12. Compute average ANNIGMA of $S_{current}$ by (12)

13. Compute combined score for every feature in $S_{current}$ by (14 to 19) for hybrids

14. Rank the features in $S_{current}$ using the combined score in descending order

15. $S_0 = S_0 \cup S_{current}$

16. Update the current feature set $S_{current}$ by removing the feature with lowest score

17. endfor

18. S_{BEST} = Find the subset form S_0 with the highest accuracy.

19. return S_{BEST}

End

3.5 Detail steps of Hybrid algorithms (MaxRel-MinRed-ANNIGMA and MR-ANNIGMA)

The detail algorithm of hybrid approaches is described in algorithm-1 and 2 and Figure 2.

3.5.1 Search strategies and subset generation in MaxRel-MinRed-ANNIGMA and MR-ANNIGMA

Both of the hybrid approaches use a Backward Elimination (BE) search strategy to generate a subset of features. Initially hybrid starts with the full feature set. Subset generation in BE is guided by the wrapper-filter hybrid heuristic score. The combined score computation follows the steps of sub-sections (3.1 to 3.4). When the number of features in BE process is significantly reduced compared to total feature, the filter score component is weighted less than the wrapper score as (18) and (19)

$$Combined\ \ Score(\text{MaxRel - MinRed ANNIGMA}; F_i) = (u * \text{Weighted MaxRel\_MinRed Score}(F_i)) + (v * ANNIGMA(F_i)_{average}) \quad \text{----------(18)}$$

$$Combined\ \ Score(MR-ANNIGMA) = (u * \frac{I(F_i;c)}{\max imum\ (f_i \in S)\ I(F_i;c)}) + (v * ANNIGMA(F_i)_{average}) \quad \text{----------(19)}$$

where $1 \le u, v \ge 0$

3.5.2 Wrapper step in MR-ANNIGMA/ MaxRel-MinRed-ANNIGMA

Both the proposed MR-ANNIGMA and MaxRel-MinRed-ANNIGMA, use a single hidden layer Multi Layer Perceptron (MLP) Network (Figure-1) in the wrapper stage. An n-fold cross validation approach has been applied in the training of the network. The evaluation criterion of feature subset is based on the average prediction accuracy over n-fold of the wrapper (MLP network). In Algorithm-1 and 2, steps-1 to 11 computes the average accuracy over n-folds for the current subset of features. Step-12 to step-14 computes the hybrid scores and ranks the features based on their combined score. Step-15 to step-16 generates new subset based on the feature ranking and keep records of evaluated feature subsets with their accuracy. The BE processes in MR-ANNIGMA and MaxRel-MinRed-ANNIGMA update MR, MaxRel-MinRed and ANNIGMA and the combined score in every iteration. The combined score guides the subset generation. The BE continues until a single feature is remaining in the current subset. The subset with highest accuracies or close to the highest accuracies with fewer features than it is chosen as the final feature subset.

4 Experimental Results and discussion

The proposed hybrids (MR-ANNIGMA and MaxRel-MinRed-ANNIGMA) have been tested on both synthetic and UCI Machine learning repository data sets (Asuncion et al.) and Tobacco control policy evaluation dataset (Thompson et al. 2006) in Table-1. For each data set in Table-1, the data is normalized in the range [-1, 1]. CAIM (Kurgan et al. 2004) discretization technique has been used for continuous attributes while computing filter score. A single hidden layer neural network with the different network configuration for each data set (Table-1) is used.

The results of the hybrids (MR-ANNIGMA and MaxRel-MinRed-ANNIGMA) have been compared to filter approaches MR, MaxRel-MinRed and the wrapper ANNIGMA. Each of the above five algorithms were tested using 10-fold cross validation and executed for 10-trials. In the BE process 2/3 iterations use (u=v=1) and last 1/3 iterations uses (u=0.3, v=0.7). The average accuracies from 10 trials were considered for final accuracies and described in Table 2 to Table 8.

Data set	Hidden nodes	Hidden Layer Transfer function	Output Transfer function	Max. epoc
Synthetic	5	tansig	purelin	250
Wine	6	tansig	purelin	150
Ionosphere	22	tansig	purelin	300
Cancer (Diagnostic	12	tansig	logsig	400
Sonar	24	tansig	purelin	200
Pima	6	tansig	purelin	200
Tobacco	22	tansig	purelin	350

Table 1: Network construction data for different data sets.

4.1 Evaluation and experimental analysis of the search process in the hybrids using combined score on a Synthetic Data set

This synthetic data set has been constructed with 6 features as Table-II where W, X, Y and Z are uniformly distributed on [-0.5, +0.5].

Features	Description
Featrure-1 (F_1):	X
Featrure-2 (F_2):	3X+1
Featrure-3 (F_3):	W
Featrure-4 (F_4):	W-Y
Featrure-5 (F_5):	Z
Featrure-6 (F_6):	2Z + 1

Table 2 : Description of Input Features of synthetic data

Index of BE Iteration	Total Features in BE iterations	MR	MaxRel-MinRed	ANNI-GMA	MR-ANNIGMA	MaxRel-MinRed-ANNIGMA
i)	6	6.917	77.167	7.167	76.883	77.267
ii)	5	7.467	77.683	7.117	77.700	77.717
iii)	4	**7.783**	**77.700**	**7.900**	**77.850**	77.900
iv)	3	6.500	75.650	6.767	76.583	**77.967**
v)	2	6.217	76.100	6.217	76.550	76.250
vi)	1	7.250	67.283	67.600	67.617	67.317

Table 3: Accuracies for five algorithms at different iterations of BE for synthetic data set

$$C = \begin{cases} 0 & if\ \ (2X - W) < 0 \\ 1 & if\ \ (2X - W) \ge 0 \end{cases} \quad (20)$$

$$C = \begin{cases} 0 & if\ \ (Y + W) > 0 \\ 1 & if\ \ (Y + W) \le 0 \end{cases} \quad (21)$$

The class variable C is binary type and has been generated by using two rules (20) and (21). 600 samples have been generated in which 300 samples' class values have been computed using (20) and rest 300 samples' class values have been computed using (21). It is seen from Table-2 that class C can be computed using Features-(F_1 or F_2) and (F_3 and F_4). Features (F_5 and F_6) are irrelevant to class variable C and F_2 is redundant with F_1. Therefore, the actual salient feature sets are

(F_1, F_3, F_4) or (F_2, F_3, F_4). All five algorithms have been executed on this synthetic data set. The different iterative accuracies of BE process in five algorithms have been given in Table-3.

First Iteration (Total 6 Features)	Accuracies (77.167%)	Second Iteration (5 Features)	Accuracies (77.117%)	Third Iteration (Total 4 Features)	Accuracies (77.900%)
Feature Index	ANNIGMA Score	Feature Index	ANNIGMA Score	Feature Index	ANNIGMA Score
3	0.974419	3	0.999538	3	0.997381
4	0.548960	1	0.799970	1	0.531413
1	0.547774	2	0.783565	2	0.473748
2	0.516855	4	0.283010	4	0.421396
5	0.105015	5	0.127403	5	Removed
6	0.099459	6	Removed	6	Removed

Table 4: Accuracies and features' score at different iterations of BE in ANNIGMA for synthetic data set. Removed means corresponding feature has been removed in this iteration

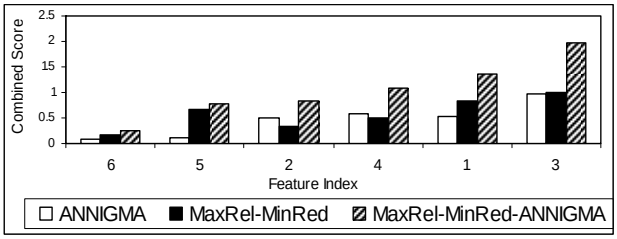

Figure 3. Synthetic Data: The combined score in the first iteration of BE process in the hybrid (MaxRel-MinRed -ANNIGMA) when total features is 6. Y-axis gives the combined score and X-axis gives the feature's serial no.

The ranking order of features in MR is (3, 1, 2, 4, 5, 6) in synthetic data. The BE process for MR finds best accuracy with four features (3, 1, 2, 4) as in Table-3 (iteration-iii) where MR incorporates the redundant feature-2. The ANNIGMA score of features in different iterations of BE for ANNIGMA heuristic is given in Table-4. It is seen in Table-4 that order of features changes (according to the score) in different iterations of BE in ANNIGMA process and corresponding accuracies also changes. In 3rd iteration (iteration-iii, Table-3) of BE, ANNIGMA finds best accuracies 77.9% with total four features (3, 1, 2, 4) which includes redundant feature-2. The ranking order of features in MaxRel-MinRed is (3, 1, 5, 4, 2, 6). The BE process accuracies is given in Table-III for MaxRel-MinRed which finds best accuracy with four features (3, 1, 5, 4) as in Table-3 (iteration-iii) but this final feature set is different from both ANNIGMA and MR.

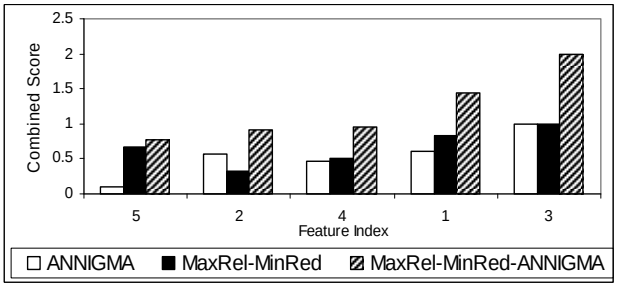

Figure .4. Synthetic Data: The combined score in the 2nd iteration of BE in the hybrid (MaxRel-MinRed -ANNIGMA) when total features is 5. Y-axis gives the combined score and X-axis gives the feature's serial no.

The ranking of features and their score in BE process for (MaxRel-MinRed-ANNIGMA) is given in (Figure 3, 4, 5). In Figure 3, feature-6 has the lowest combined score at first iteration and has been removed after first iteration. In Figure-4, in second iteration, ANNIGMA finds feature-5 as the lowest score, MaxRel-MinRed finds feature-2 as the lowest, however (MaxRel-MinRed – ANNIGMA) finds feature-5 as the lowest. Therefore, feature-5 has been removed after second iteration. In Figure-5, feature-2 is removed after 3rd iteration since the lowest combined score. (MaxRel-MinRed –ANNIGMA) finds the highest accuracy 77.967% (Table-3) with only three features (3, 1, 4) and this final feature set has no (irrelevant or redundant) component and is the correct salient features of the synthetic data set. This shows the significance of hybridization of wrapper and filter approaches in MaxRel-MinRed-ANNIGMA.

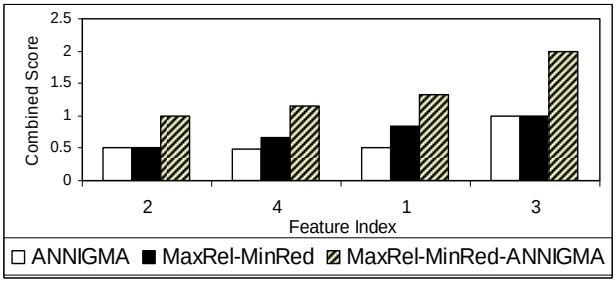

Figure 5. Synthetic Data: The combined score in the third iteration of BE process in the hybrid (MaxRel-MinRed -ANNIGMA) when total features is 4. Y-axis gives the combined score and X-axis gives the feature's serial no.

Total Features	ANNI-GMA (%)	MR (%)	MR-ANNIGMA (%)	MaxRel-MinRed (%)	MaxRel-MinRed -ANNIGMA(%)
13	97.528	97.697	97.416	97.578	97.472
12	96.910	96.966	97.921	97.079	97.809
11	97.753	96.517	96.966	96.124	97.079
10	97.022	96.798	96.854	97.360	97.640
9	97.079	97.022	97.472	96.966	97.921
8	97.191	**97.640**	97.360	**97.865**	96.517
7	97.697	96.742	96.966	97.360	97.697
6	**97.921**	97.416	**98.034**	96.798	97.865
5	96.124	96.348	96.124	97.360	**98.315**
4	95.225	95.337	95.618	96.517	96.124
3	94.438	94.719	94.888	93.876	93.539
2	90.000	90.618	90.169	91.910	90.393
1	78.933	78.989	78.944	79.270	79.270

Table 5: Accuracies for five algorithms at different iterations of BE process for wine data set.

4.2 Evaluation and experimental analysis of the search process in the hybrids using combined score on Wine data set (Asuncion et al.)

This data set has total 13 real/integer valued attributes with no missing values. The detailed accuracies in different iterations of BE process for wine data set is given in Table-5. The wrapper approach (ANNIGMA) achieves an accuracy of (97.921%) for 6 attributes (7,10,12,13,2,1). The filter-MR achieves accuracy (97.640%) for 8 attributes (7,10,13,12,1,11,6,2) and the filter-MaxRel-MinRed achieves accuracy 97.865% for 8 attributes (7,1, 10,13,11,12,6,5) which is different from final set of MR.

The hybrid process (MR-ANNIGMA) starts with 13 attributes where attribute-8 has the lowest score for ANNIGMA, attribute-3 has the lowest MR score and the hybrid finds attribute-8 as the lowest (Figure 6). Therefore the hybrid eliminates attribute-8 after the first cycle. In the next cycle of BE (Figure 7), the hybrid re-computes all feature's score resulting in attribute-3 attaining the lowest combined score. Therefore it eliminates attribute-3. The hybrid (MR-ANNIGMA) continues BE process and achieves the highest accuracy (98.034%) for six attributes (10,7,1,13,12,11) which is different from final feature set of ANNIGMA. The hybrid MaxRel-MinRed-ANNIGMA achieves the highest accuracy 98.315% for 5 attributes (7,10,1,13,11). Therefore MaxRel-MinRed-ANNIGMA obtains the smallest feature set in all algorithms with the highest accuracy.

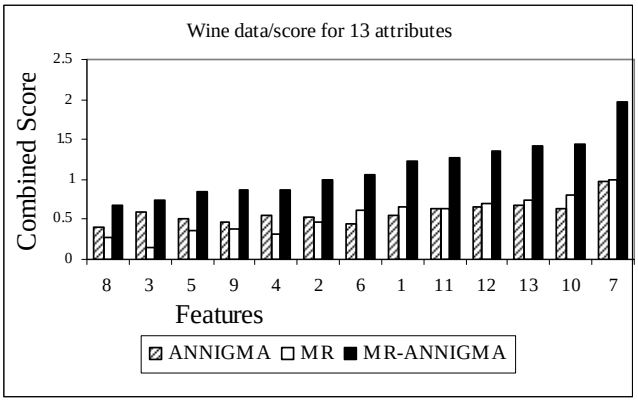

Figure 6. The combined score in the BE search process in the hybrid (MR-ANNIGMA) when total features is 13. Y-axis gives the combined score and X-axis gives the feature's serial no.

4.3 Evaluation and experimental analysis of the search process in the hybrids using combined score on a real life data set: Tobacco Control Policy Evaluation data set (Thompson et al. 2006)

Tobacco Control Policy Evaluation data set: The proposed algorithm has also been tested on a real life data set - "Tobacco Control Policy Evaluation data set". This data set is constructed through International Tobacco Control Policy Evaluation Project (ITC Project) of World Health Organization (WHO) (ITCEP). ITC completed a four country survey (ITC-4 tobacco data) (ITCEP), (Fong et al. 2005) with a target of estimating the impact of psychological and behavioural impact of the key policies of Framework Convention on Tobacco Control (FCTC)) (ITCEP), (Fong et al. 2005), (Thompson et al. 2006), (Heyland et al. 2006) organized by the World Health Organization (WHO). The Four-Country Survey was made among randomly selected smokers in four English-speaking countries: Canada, the United States, the United Kingdom, and Australia. ITC-4 participant smokers are adult who have smoked more than 100 cigarettes in their lifetimes and have smoked at least once in the past 30 days.

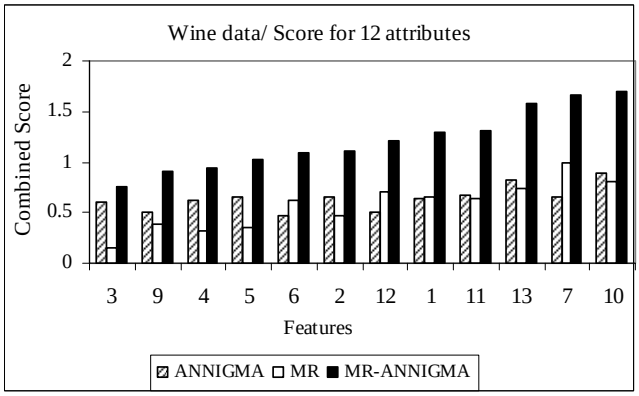

Figure 7. The combined score in the BE search process in the hybrid (MR-ANNIGMA) when total features is 12. Y-axis gives the combined score and X-axis gives the feature's serial no.

The survey consists of four waves. More than seventy five questions have been considered to evaluate the impact of tobacco control policy measures among smoking population. Survey question are mainly based on psychosocial – beliefs about smoking, beliefs about quitting, psychosocial questions such as perceived risk and health worry, smoking behaviour such as total minutes to first cigarette, addictedness to cigarettes), knowledge of health effects/tobacco constituents, socio-demographic questions such income, smokers' reaction and outcome on cessation advice and services, smokers' reactions on warning labels, advertising, monitoring of anti-tobacco campaigns, price/taxation and sources of tobacco, smokers' reactions and effect on smoking restrictions.

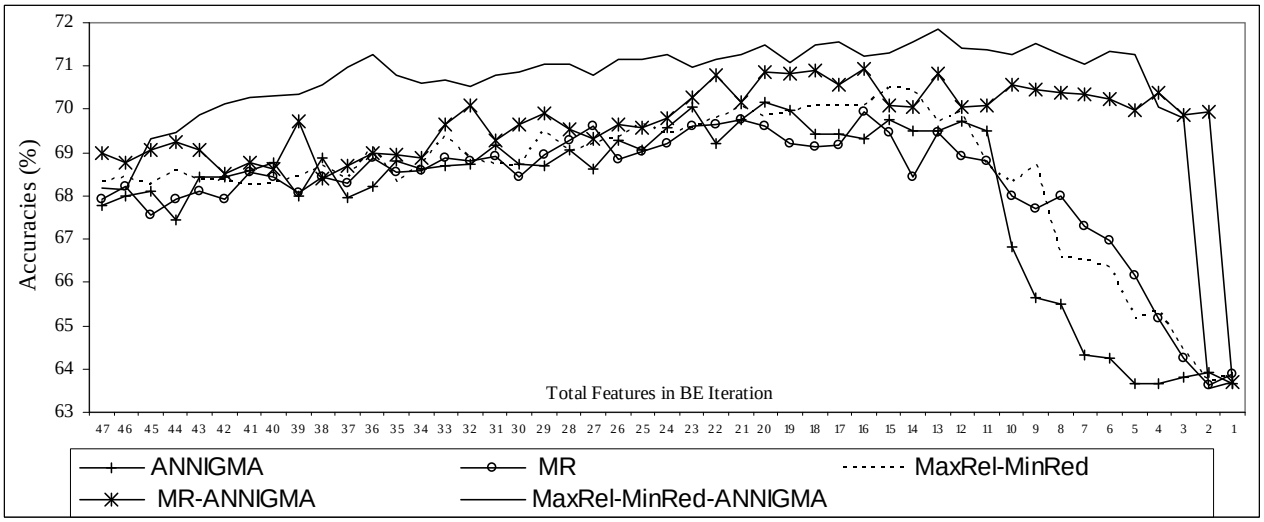

Figure 8. Accuracies in different iterations of BE process in five algorithms for Tobacco Control Policy Evaluation data. Y-axis gives the accuracies and X-axis gives the total number of feature in different iterations of BE process for five algorithms.

The main outcome questions is whether the smokers' have made any attempt to stop smoking since they were interviewed last or they have stopped smoking for about 6 months. We have used here Wave-1 data set of tobacco data. There are 77 attributes in the Wave-1 data set (Integer and Real) with 6682 examples. All five algorithms have been executed on tobacco data. Initially for all algorithms, tobacco data has been pre-processed and the attributes are ordered based on their corresponding heuristic score in descending. Then last 30 attributes with lowest rank are discarded to reduce computational overhead and top-ranked 47 attributes have been used for evaluation of each algorithm. 10-fold cross validation with 10 trials is applied for each of the five algorithms (MR, MaxRel-MinRed, ANNIGMA, MR-ANNIGMA and MaxRel-MinRed-ANNIGMA). The average accuracies over 10 trials for five algorithms for different iterations of the BE process is presented in Figure-8. ANNIGMA achieves 67.793% at 47 attributes. Accuracies over iterations continue to increase up to 70.048% at 23 attributes which is the highest in all iterations. Therefore 23 attributes is the final feature set from ANNIGMA. The filter MaxRel-MinRed achieves accuracy 70.485% for 15 features. It also achieves an acceptable accuracy 70.443% for 14 features. (MR) achieves highest accuracies 69.94% for 16 features. It also achieves an acceptable accuracy 69.45% for 15 features. The hybrid MR-ANNIGMA approach obtains the highest accuracy 70.922% with 16 features. The MR-ANNIGMA also finds a second highest 70.835% for 13 features which is also acceptable. MaxRel-MinRed-ANNIGMA approach obtains the highest accuracy 71.869% with 13 features. The MaxRel-MinRed-ANNIGMA also finds 71.536% accuracy for 9 features which is also acceptable. The results show that hybrid approaches (MR-ANNIGMA, MaxRel-MinRed-ANNIGMA) achieve better accuracies using fewer features than filter or wrapper alone. This demonstrates the significance of the hybridization in the proposed approaches in searching for the most important feature set.

4.4 Evaluation and experimental analysis of the search process in the hybrids using combined score on Ionosphere data set (Asuncion et al.)

This data set has total 34 real valued attributes with no missing values. The detailed accuracies in BE process for the ionosphere data set for all algorithms is described in Table-6. The filter approach (MR) achieves highest accuracy of (91.311%) with 15 attributes and (91.083%) with 11 attributes which is closer to the highest and considered as final feature set of it. The ANNIGMA achieves (90.057%) with four attributes. The filter MaxRel-MinRed achieves the highest accuracy of (91.339%) with 13 attributes. Hybrid MR-ANNIGMA achieves (92.137%) with four attributes. However these final four attributes (5,21,3,6) are different from final feature set (6,24,15,14) of ANNIGMA. The hybrid MaxRel-MinRed-ANNIGMA achieves the highest accuracy of (92.792%) with three attributes (5,6,3) (Table 6 and 7). It is seen that our hybrid approaches achieve the highest accuracy with very compact feature set (less than five features). However, MaxRel-MinRed-ANNIGMA performs best among all algorithms with highest accuracy (92.792%) and fewest features (3). This proves the significance of the hybrid approaches to select the most salient feature set.

Total Features in BE Iterations	MR	ANNI-GMA	MaxRel-MinRed	MR-ANNI-GMA	MaxRel-MinRed-ANNI-GMA
34	84.188	86.439	84.672	86.011	84.615
33	85.299	87.322	87.094	82.222	83.618
32	84.131	84.615	85.499	84.929	82.279
31	84.986	86.268	86.467	87.607	86.382
30	87.066	85.328	90.256	86.467	89.060
29	88.234	85.613	89.886	86.724	88.091
28	87.521	84.843	88.775	84.501	88.917
27	84.615	86.752	87.464	85.157	85.755
26	87.179	83.789	88.519	85.271	85.726
25	87.350	88.120	88.575	86.724	87.920
24	85.271	83.504	87.578	84.046	86.923
23	86.524	85.556	87.009	83.875	88.661
22	88.946	85.527	87.749	87.436	88.262
21	86.752	86.724	87.578	85.755	88.034
20	85.442	86.524	87.607	84.387	87.151
19	89.687	88.547	88.262	89.829	87.892
18	89.487	89.687	90.228	88.946	88.860
17	90.456	88.063	89.145	87.464	90.969
16	88.405	87.208	89.430	88.718	89.886
15	**91.311**	87.464	90.798	88.063	91.567
14	89.373	89.744	89.459	89.630	89.772
13	90.684	86.325	**91.339**	88.917	91.368
12	90.513	88.262	90.085	89.715	90.997
11	**91.083**	86.524	89.886	89.402	91.823
10	90.627	88.376	89.060	91.197	92.023
9	89.829	88.177	88.860	88.433	91.994
8	89.174	86.980	88.632	89.402	93.020
7	87.037	88.632	90.855	90.712	90.969
6	86.040	**90.057**	90.256	91.738	93.048
5	88.405	89.829	88.860	91.595	94.330
4	87.407	**90.057**	90.883	**92.137**	91.880
3	88.746	89.886	89.715	86.752	**92.792**
2	87.236	87.977	89.630	87.037	90.197

Table 6: Accuracies for five algorithms at different iterations of BE process for Ionosphere data set (Asuncion et al.).

Table-7 summarizes the final accuracies and number of optimal feature selected for all data sets (described in Table-1) for all algorithms. It shows that the proposed hybrid approaches achieves very compact feature sets in all data sets trialled with higher accuracies than both filter and wrapper alone. This demonstrates that the hybridization of filter and wrapper in the MR-ANNIGMA and MaxRel-MinRed-ANNIGMA lead to improved predictive accuracy with fewer features. However MaxRel-MinRed-ANNIGMA performs better than MR-ANNIGMA and finds smallest feature set with highest accuracies.

Data set		MR (%)	ANNIGMA (%)	MaxRel-MinRed (%)	MR-ANNIGMA (%)	MaxRel-MinRed ANNIGMA (%)	Other (%)
Wine	Accuracy	97.640	97.921	97.865	98.034	98.315	98.2 (Huang et al. 2008)
	Total Features	8	6	8	6	5	6
Ionosphere	Accuracy	91.311	90.057	91.339	92.137	92.792	92.51 (Huang et al. 2008)
	Total Features	15	6	13	4	3	4
Cancer (Diagnostic)	Accuracy	96.148	96.287	96.92	97.065	97.71	94.9 (II-Seok et al. 2004)
	Total Features	21	14	17	16	15	
Sonar	Accuracy	83.506	83.606	83.073	84.236	84.594	83.4 (Optiz et al. 1999)
	Total Features	17	40	12	16	15	
Pima	Accuracy	76.95	76.71	77.04	77.17	77.173	77.0 (Hsu et al. 2002)
	Total Features	{8,2,6}=3	{2,6,7,1,5}=5	6	{7,6,2}=3	{7,6,2}=3	
Tobacco	Accuracy	69.45	70.048	70.443	70.835	71.536	
	Total Features	15	23	14	13	9	

Table 7 Detailed accuracies for five algorithms and number of features in final feature sets for all data sets.

4.5 Computational Performance

The hybrid algorithms runs a backward elimination (BE) process where each iteration involves computational time in training the network, the computation of MR, MaxRel-MinRed score, ANNIGMA and hybrid score. Computation of MR score and ANNIGMA has linear time complexity in terms of feature dimensionality. At the beginning when all features are used, the time for training and computing scores (MR, ANNIGMA, and hybrid) would be the highest. Subsequent computation will take less time. Computational performance has been described in Table-8. The experimental platform was 3.2-GHz Pentium-4 CPU with 1GB of RAM. Table-8 shows that MaxRel-MinRed-ANNIGMA takes more time than MR-ANNIGMA.

Data Set	MR (Hrs)	ANNIGMA (Hrs)	MaxRel-MinRed (Hrs)	MR-ANNI-GMA (Hrs)	MaxRel-MinRed-ANNI-GMA (Hrs)
Synthetic	0.1219	0.1208	0.1225	0.1239	0.1246
Wine	0.1021	0.1304	0.1114	0.1428	0.1479
Ionosphere	0.8223	0.840	0.8261	0.8545	0.9081
Cancer	0.8543	0.8012	0.8745	0.8731	0.8834
Sonar	0.9014	0.9415	0.9124	0.9512	1.0972
Pima	0.2214	0.214	0.2573	0.2421	0.2588
Tobacco	4.241	4.520	4.631	5.342	5.459

Table 8: Computational time (Hours) for five algorithms for all data sets.

5 Conclusions

This paper proposes two novel hybrids of wrapper and filter approaches for input feature selection problem. The novelty of our approaches is that these integrate knowledge (from the intrinsic characteristics of data) obtained by the filter approach into the wrapper approach and combines the wrapper's heuristic score with the filter's ranking score in the wrapper stage of the hybrid. To the best of our knowledge, the idea of our approach is new and has not been explored yet in the literature. The first proposed hybrid combines a mutual information (MI) based Maximum Relevance (MR) filter ranking heuristic with an Artificial Neural Network (ANN) based wrapper approach where Artificial Neural Network Input Gain Measurement Approximation (ANNIGMA) has been combined with MR (MR-ANNIGMA). The proposed second hybrid algorithm combines an improved version of MR (Maximum Relevance and Minimum Redundancy; MaxRel-MinRed) filter ranking heuristic with the ANNIGMA (MaxRel-MinRed-ANNIGMA). The combined heuristics in the hybrids: (MR-ANNIGMA and MaxRel-MinRed-ANNIGMA) take the advantages of the complementary properties of the both filter and wrapper heuristics and guide the wrapper to find optimal and compact feature subsets in the wrapper. The approaches have been tested using synthetic data, bench mark machine learning data sets and real life-Tobacco Control Policy Evaluation data sets with varying number of features and sample size. Our experiments show that hybrid algorithms (MR-ANNIGMA and MaxRel-MinRed-ANNIGMA) ranks the features in such a way that the internal BE process of the wrapper step generates better subsets of features than both filter and wrapper approaches in terms of wrapper evaluation criteria and achieves higher accuracies and smaller feature sets than both filter and wrapper approaches. However, MaxRel-MinRed-ANNIGMA outperforms all other algorithms. In the future we will use other search strategies such as bidirectional search with the proposed approaches and then will evaluate the approaches on the rest of the Waves' data of tobacco control data as well as other bench mark data sets.

6 References

Blum, A.L, and Langely,P, "Selection of relevant features and examples in Machine Learning", Artificial Intelligence, Vol 69, pp 245-271,1997

John, G.H., Kohavi, R. and Pfleger, K, " Irrelevant Feature and the subset selection problem", Proc. Of 11th Int. conference on Machine Learning, pp 121-129, 1994

Dash, M. and H. Liu, "Feature selection for classification", Intelligent data analysis: An International Journal, vol-1, no-3, pp 131-156, 1997

Puronnen,S., Tsymbal,A. and I. Skrypnik, "Advanced local feature selection in Medical Diagnostics", Proc. 13th IEEE symp. computer-based medical diagnostics, 2000.

Bne-Bassat,M., "Pattern recognition and Reduction of dimensionality", Handbook of Statistics-II, P.R. Krisnaiah and L.N. Kanal eds. Pp 773-791, 1982

Mitra, P., Murthy, C.A., and S.K. Pal, "Unsupervised Feature selection using Feature similarity", IEEE Transaction on Pattern Analysis and Machine Intelligence, vol 24 pp 301-312, March 2002.

Kwak,N., and C. Choi, Member, "Input Feature Selection by Mutual Information Based on Parzen Window", IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 24, No. 12, December 2002

Wang,H., Bell,D., and F. Murtagh, "Axiomatic Approach to Feature Subset Selection Based on Relevance", IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 21, No. 3, March 1999

Hall,M.A., "Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning," Proc. 17th Int'l Conf. Machine Learning, pp. 359-366, 2000.

Kwak,N., and C. Choi, "Input Feature Selection for Classification Problems", IEEE Transaction on Neural Networks, Vol. 13, No. 1, January 2002

J.G. Dy and C.E. Brodley, "Feature Subset Selection and Order Identification for Unsupervised Learning," Proc. 17th Int'l Conf. Machine Learning, pp. 247-254, 2000.

Hsu,C.N, H.J. Huang, and D. Schuschel, "The ANNIGMA-Wrapper Approach to Fast Feature Selection for Neural Nets", IEEE Transactions on Systems, Man and Cybernetics-Part B: Cybernetics, Vol. 32, No. 2, April 2002

Zhu,Z., Y. S Ong, M. Dash, "Wrapper-Filter feature selection algorithm using a memetic framework", IEEE Transactions on Systems, Man and Cybernetics-Part B: Cybernetics, Vol. 32, No. 2, April 2002

Kohavi,R., and G.H. John, "Wrappers for Feature Subset Selection," Artificial Intelligence, vol. 97, nos. 1-2, pp. 273-324, 1997

II-Seok Oh, Ji-Seon Lee and Byung-Ro Moon, "Hybrid genetic algorithms for Feature selection", IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 26,2004.

Jain,A., and D. Zongker, "Feature selection: Evaluation, Application and small sample performance", IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 19, No. 2, pp 153-158, Feb 1997.

Ferri,F.J, P. Pudil, M. hatef and J.Kittler, "Comparative study of techniques for large-scale feature selection", Pattern recognition in practice IV, E.S. Gelsema et.al. eds pp 403-413, 1994

Asuncion, A. and Newman, D. J.. UCI Machine Learning Repository Irvine, CA: University of California, School of Information and Computer Science. [http://www.ics.uci.edu/~mlearn/MLRepository.html].

Optiz,D., and R. Maclin, "Popular Ensemble methods: An Empirical study," Journal of Artificial Intelligence Research, vol 11, pp 169-198,199

Huang,J.J., Y.Z Cai and X.M. Xu, "A parameter less feature ranking algorithm based on MI", Neurocomputing, (Vol-71), 2008,pp 1656-1668.

Kurgan,LA., et al., "CAIM discretization algorithm", IEEE Trans. on Knowledge and Data Engineering, 26, 2004, pp 145-153.

Peng,H., C. Ding and F. Long, "Minimum Redundancy-Maximum Relevance Feature selection", IEEE intelligent Systems, Nov 2005, pp 70-71.

ITCEP,International Tobacco Control policy Evaluation Project (ITCEP), http://www.itcproject.org/. WHO, 2008, WHO REPORT on the global TOBA CCO epidemic, 2008, The MPOWER package

Fong,G.T, K. M. Cummings, R. Borland, G. B. Hastings, P. Hyland, G. A. Giovino, D. Hammond, and M. E. Thompson, "The conceptual framework of the International Tobacco Control (ITC), Policy Evaluation Project," Tobacco Control, vol. 15, Suppl. 3, pp. 3-11, 2005.

Thompson,M.E., G. T. Fong, D. Hammond, C. Boudreau, P. Driezen, P. Hyland, R. Borland, K. M. Cummings, G. B. Hastings, M. Siahpush, A. M. Machintosh, and F. L. Laux, "Methods of the International Tobacco Control (ITC) Four Country Survey", Tobacco Control, vol. 15, Suppl. 3, pp. 12-18, 2006.

Hyland,A., R. Borland, Q. Li, H-H. Yong, A. McNeill, G. T. Fong, R. J. O'Connor, and K. M. Cummings, "Individual- level predictors of cessation behaviours among participants in the International Tobacco Control (ITC) Four Country Survey", Tobacco Control, vol. 15, Suppl. 3, pp. 83-94, 2006.