2010 Fourth International Conference on Network and System Security

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

Shamsul Huda[1], John Yearwood[2], Andrew Strainieri[3]
Centre for Informatics and Applied Optimisation
GSITMS, University of Ballarat, Victoria, Australia
1. s.huda@ballarat.edu.au, 2. j.yearwood@ ballarat.edu.au, 3. a.stranieri@ballarat.edu.au

Abstract—**Feature selection is an important research problem in machine learning and data mining applications. This paper proposes a hybrid wrapper and filter feature selection algorithm by introducing the filter's feature ranking score in the wrapper stage to speed up the search process for wrapper and thereby finding a more compact feature subset. The 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 novelty of our approach is that we use hybrid of wrapper and filter methods that combines filter's ranking score with the wrapper-heuristic's score to take advantages of both filter and wrapper heuristics. Performance of the proposed MR-ANNIGMA has been verified using bench mark data sets and compared to both independent filter and wrapper based approaches. Experimental results show that MR-ANNIGMA achieves more compact feature sets and higher accuracies than both filter and wrapper approaches alone.**

I. INTRODUCTION.

Feature selection is an important and frequently used data pre-processing techniques in machine learning [1] [2], data mining [3], medical data processing [4] and statistical pattern recognition areas [5] [6]. Due to rapid advances of computational, data transfer and storage technologies, large volumes of data with thousands of features is very common. To use huge datasets for decision making, prediction or classification purposes by using data mining techniques is a challenging task for researchers and practitioners because the performance of data mining methodologies degrades with huge volumes of training data [1], [2], [3]. Therefore, dimensionality reduction of the training data by removing irrelevant, redundant or noisy features is a primary task for machine learning researchers.

The identification of an optimal feature subset speeds up data mining algorithms and improves their performance measures such as predictive accuracies. Given an m-dimensional dataset, a feature selection algorithm needs to find optimal feature subset from the 2^m subsets of feature space. Therefore finding an optimal feature subset is computationally expensive [14]. The performance of a feature selection algorithm depends on its evaluation criterion and search strategies.

Feature selection algorithms developed with different evaluation criteria can be grouped broadly into three main categories: 1) the filter model [3], [7], [8], [9] 2) the wrapper model [10] [11] [12] [3]) and hybrid models [13]. The filter model involves the application of an algorithm to a dataset prior to its use for data mining. This makes it independent of any induction algorithm. Diverse filter models have been advanced including ones that use a relevance measure[8] and others that deploy a distance measure [5] to estimate the goodness of the feature subset. Filter models are computationally cheap because they are applied to the dataset as a pre-processing step and, unlike wrapper models, do not use a data mining algorithm. However, because filter models are independent from the induction algorithm used in the mining phase, feature subsets selected may result in poor prediction accuracies.

In contrast, the wrapper model [10] [11] [12] uses a predetermined induction algorithm and uses its performance as the evaluation criteria for the feature selection. The selection algorithm directly uses the performance of the induction algorithm to guide the search path in the feature space for optimal feature subset. However, wrapper models face huge computational overhead due to the use of the induction algorithm's performance criteria as its evaluation criteria.

Hybrid models [12], [14] take advantage of the complementary properties of the both approaches. In [12], Z. Zhu et al. proposed a hybrid of wrapper-filter approach (hybrid of genetic algorithm and filter heuristic) where a filter ranking method was used in the genetic algorithm (GA) framework to speed up the genetic search process by the improvement of local search using a filter heuristic. In [14], a hybrid of GA approach has been proposed for feature selection. GA-based approaches face huge computational overheads due to the evaluation of the induction algorithm embedded in the GA fitness function.

In this paper, we propose a hybrid of wrapper and filter approach by using the filter's feature ranking score in the wrapper approach to speed up the search process in the wrapper stage for optimal feature subset selection using an induction algorithm. In the approach, we hybridize novel Mutual Information (MI) based Maximum Relevance (MR) filter ranking heuristics with Artificial Neural Network (ANN) based wrapper approach where the Artificial Neural Network Input Gain Measurement Approximation

978-0-7695-4159-4/10 $26.00 © 2010 IEEE
DOI 10.1109/NSS.2010.7

(ANNIGMA) wrapper heuristic has been combined with MR (MR- ANNIGMA) to accelerate the wrapper search process.

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 approach avoids the computational overhead of hybrid GA-based approaches [12], [14] and takes advantage of both filter and wrapper heuristics which are absent in the traditional GA-based hybrid approaches [12], [14]. 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 Maximum Relevance (MR) filter heuristics and Artificial Neural Network Input Gain Measurement Approximation (MR-ANNIGMA) is described in Section III. Section IV presents experimental results and discussion. Conclusions of this study are presented in the last section.

II. Related Work.

A. Standard Filter approach

Figure-1 presents a standard filter approach for feature selection.

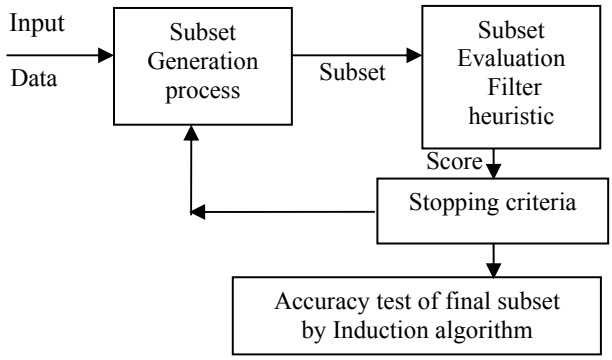

Figure 1. A standard filter approach

Standard filter approaches use a subset generation process which may start from an empty set or start with full feature set and then either forward or backward or bi-directional search strategies are followed. Generated subsets are evaluated using filter heuristics such as information gain, co-relation measure [9], mutual information [7], and maximum-relevance [8]. 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.

B. Wrapper approach

Figure-2 presents a standard wrapper approach for feature selection. In the wrapper approach [10] [11] [12], generated subsets are evaluated using a predetermined induction algorithm. This means the induction algorithm is trained repeatedly with the training data for each subset. This is computationally very expensive. Different search strategies such as sequential backward elimination (SBE) [12], sequential forward elimination (SFE) [12] or bidirectional search approaches are used in the subset generation process. GA based search approaches have also been used [13], [15] where subsets are generated by the GA population. Some comparisons of different search strategies have been made in [16], [17]. However, subsets in wrapper approach are evaluated by the predictive accuracies of trained classifier, therefore are more significant than those in the filter approach which only depends on feature redundancy or relevance.

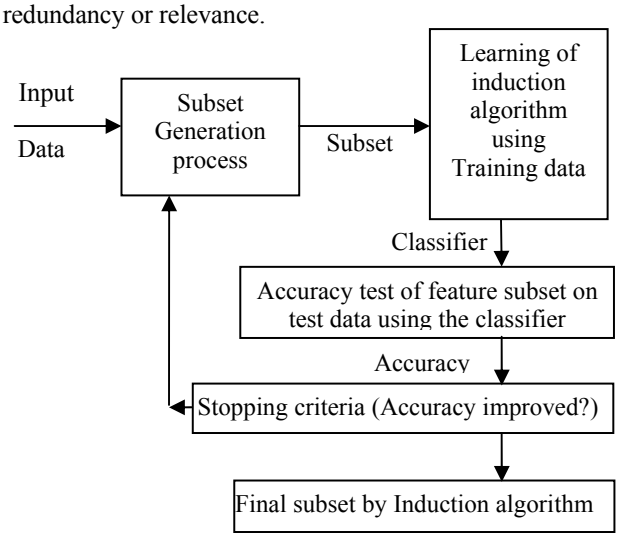

Figure 2. A standard wrapper approach

III. Hybrid feature selection algorithm using Maximum Relevance (MR) and Artificial Neural Network Input Gain Measurement Approximation (MR-ANNIGMA).

In the proposed approach, we hybridize wrapper and filter approach to take advantage of both approaches. Standard filter approaches can extract knowledge of the intrinsic characteristics from real data. However filter approaches do not use any performance criteria based on predictive accuracies. This does not guarantee that selected feature subset will do better in classification/prediction tasks. Usually the wrapper approach [10], [11], [12] uses a predetermined induction algorithm and different search strategies [16], [17] 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 search process incurs a high computational cost in the wrapper approach.

In this paper, we propose a hybrid approach that introduces the filter heuristic in the wrapper stage to speed up the search process in the wrapper. We also employ a wrapper heuristic and combine it with the filter heuristic. We have combined mutual information based Maximum Relevance (MR) filter heuristics and Artificial Neural Network Input Gain Measurement Approximation (MR-ANNIGMA) based wrapper heuristic (MR-ANNIGMA).

The following sub-sections describe different heuristics and steps of MR-ANNIGMA.

A. Mutual information based Maximum Relevance (MR)

Mutual information provides statistics that summarize the degree of relevance between the features and class variable. Relevant features provide more information about the class variable than irrelevant features. Usually, features selection process finds those features that provide as much information as possible about the class variable. Therefore maximum relevance [8] 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 [8] can be defined as:

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

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

$$I(F_i;c) = H(F_i) - H(F_i \mid c) \quad (2)$$

$H(F_i)$ is the entropy of F_i with the probability density function $p(f_i)$ where F_i takes discrete values from the set $F = \{f_1, f_2 ... f_i\}$, then $H(F_i)$ is defined as (3)

$$H(F_i) = -\sum_{f_i \in F} p(f_i)\log p(f_i) \quad (3)$$

$H(F_i \mid c)$ in (2) is the conditional entropy between F_i and c and is defined as (4)

$$H(F_i \mid c) = -\sum_{f_i \in F}\sum_{c_i \in C} p(f_i, c_i)\log p(c_i \mid f_i) \quad (4)$$

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

B. Artificial Neural Network Input Gain Measurement Approximation (ANNIGMA) wrapper heuristic

Artificial Neural Network Input Gain Measurement Approximation (ANNIGMA)[12] is a weight analysis based heuristic that ranks features by relevance based on weight associated with feature in a Neural Network based wrapper approach. Features that are irrelevant or redundant will produce more error than relevant features. In a standard neural network, if noisy features have high associated weight, they will produce high error rates. Therefore, during training, weights of noisy feature are controlled in such a way that they contribute to the output as least as possible.

ANNIGMA [12] is based on the above strategy of the training algorithm. For a two layer Neural Network, (Fig.3) if i, j, k are the input, hidden and output layer and Q is a logistic activation function (5) of the first layer and second layer has a linear function, then output of the network is as (6).

$$Q(x) = (1/(1+\exp(-x))) \quad (5)$$

$$O_k = \sum_j Q\left(\sum_i A_i \times W_{ij}\right) \times W_{jk} \quad (6)$$

Then local gain is defined as (7)

$$LG_{ik} = \frac{\Delta O_k}{\Delta A_i} \quad (7)$$

According to C.N. Hsu and H.J. Huang et.al. [12], the local gain can be written in terms of network weight as (8):

$$\text{Local Gain } LG_{ik} = \sum_j |W_{ij} \times W_{jk}| \quad (8)$$

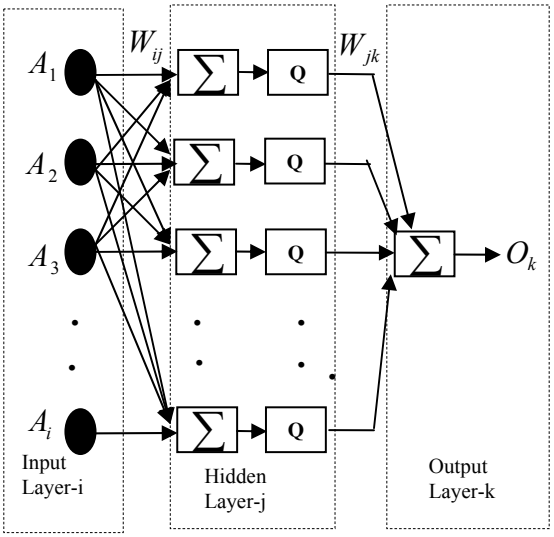

Figure 3. A single hidden layer neural network in the MR-ANNIGMA hybrid approach

Then ANNINGMA score for feature-i (F_i) is the local gain (LG) normalized based on a unity scale as (9)

$$ANNIGMA(F_i) = \frac{LG_{ik}}{\max imum_{(i)} LG_{ik}} \quad (9)$$

C. Computation of combined score using the filter's score MR and wrapper's score ANNIGMA in the proposed MR-ANNIGMA

The proposed MR-ANNIGMA uses Artificial Neural Network as the induction algorithm in the wrapper. 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 (10)

$$ANNIGMA(F_i)_{average} = \left(\frac{1}{n}\right)(ANNIGMA(F_i)_1 + .. + ANNIGMA(F_i)_n) \quad (10)$$

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 (11)

$$\text{Re}levance(F_i) = \frac{I(F_i;c)}{\max imum_{(f_i \in S)} I(F_i;c)} \quad (11)$$

The combined score of filter's heuristic and wrapper's heuristic in the proposed MR-ANNIGMA is computed as (12).

$$Combined\ Score = \frac{I(F_i;c)}{\max imum\ (f_i \in S)\ I(F_i;c)} + ANNIGMA(F_i)_{average} \quad (12)$$

D. Detail steps of MR-ANNIGMA

The detail algorithm of MR-ANNIGMA is described in algorithm-1.

1) Search strategies in MR-ANNIGMA and subset generation:

MR-ANNIGMA uses a Backward Elimination (BE) search strategy to generate a subset of features. Initially it starts with the full feature set. To guide the BE for subset generation, we have used a wrapper-filter hybrid heuristic score that combines mutual information based Maximum Relevance (MR) and Artificial Neural Network Input Gain Measurement Approximation (ANNIGMA)[12]. The combined score computation follows the steps of sub-sections (III-A, III-B and III-C).

2) Wrapper step in MR-ANNIGMA and subset evaluation:

In the proposed MR-ANNIGMA, we use a single hidden layer Multi Layer Perceptron (MLP) Network (Fig.3) 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, 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 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 process in MR-ANNIGMA updates MR, ANNIGMA and the combined score in every iteration. The combined score guides the subset generation. The BE search continues until a single feature is remaining in the current subset.

Algorithm-1: Procedure (MR-ANNIGMA)

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 subsets which records all generated subsets with their accuracy

// Apply a backward elimination search strategy

3 for N = 1 to m-1

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

5 Compute MR score by (1) and (11)

6 for fold=1 to n

7 Train the network with $S_{current}$

8 Compute ANNGMA of all features by (9)

9 Compute Accuracy

10 endfor

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

12 Compute average ANNIGMA of $S_{current}$ by (10)

13 Compute the combined score for every feature in $S_{current}$ by (10), (11) and (12)

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 eliminating the feature with lowest combined score

17 endfor

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

19 return S_{BEST}

End

IV. RESULTS AND DISCUSSION

The proposed hybrid (MR-ANNIGMA) has been tested on UCI Machine learning repository data set [18]. For each data set, the data is normalized in the range [-1, 1]. A single hidden layer neural network is used. The network for the wrapper step for each of the four data sets was constructed according to the description in Table-1.

TABLE 1. NETWORK CONSTRUCTION DATA FOR DIFFERENT DATA SETS.

UCI Data set	Hidden nodes	Hidden Layer Transfer function	Output Layer Transfer function	Max. epoc
Ionosphere	22	tansig	purelin	300
Wisconsin Cancer (Diagnostic)	12	tansig	logsig	400
Sonar	24	tansig	purelin	200
Pima	6	tansig	purelin	200

The results of the hybrid (MR-ANNIGMA) have been compared to filter approach (MR) and wrapper approach (ANNIGMA). Each of the above three algorithms were tested using 10-fold cross validation and executed for 10-trials. The average accuracies from 10 trials were considered for final accuracies and described in Table 2 to Table 6.

A. Ionosphere data set [18]

This data set has total 34 real valued attributes with no missing values and was selected because it is a commonly used dataset that is an appropriate size and complexity for this trial. The hybrid process starts with 34 attributes where attribute-10 has the lowest score for ANNIGMA, attribute-2 has the lowest MR score and the hybrid finds attribute-2 as the lowest (Fig. 4). Therefore the hybrid eliminates attribute-2 after the first cycle of the BE search process. In the next cycle of BE (Fig. 5), the hybrid re-computes all feature's score resulting in attribute-3 attaining the lowest score for

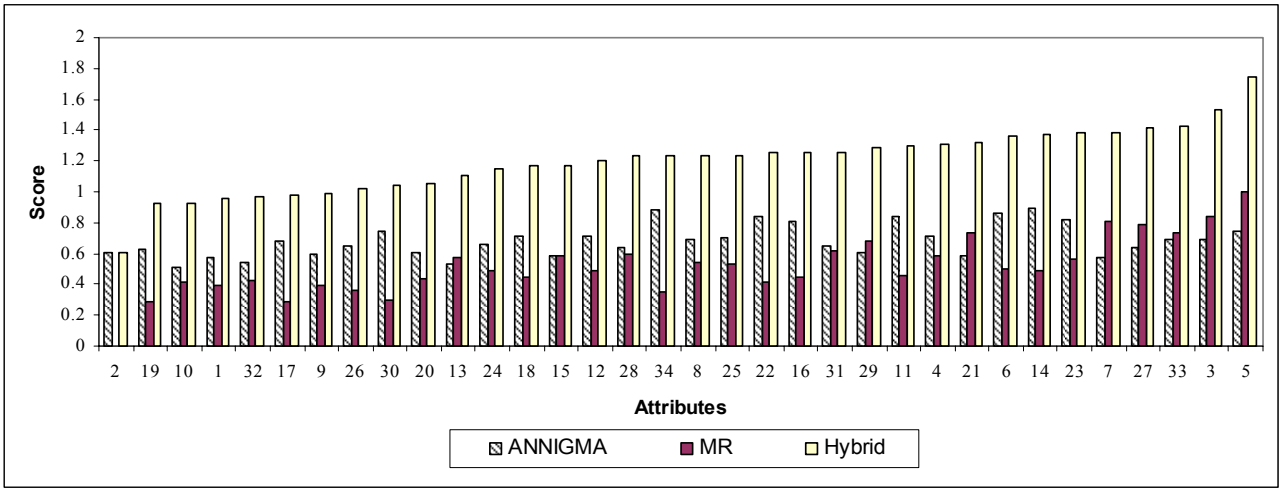

Figure 4. The computed combined score in the BE search process in the hybrid (MR-ANNIGMA) when total features is 34. Y-axis gives the combined score and X-axis gives the feature's serial no.

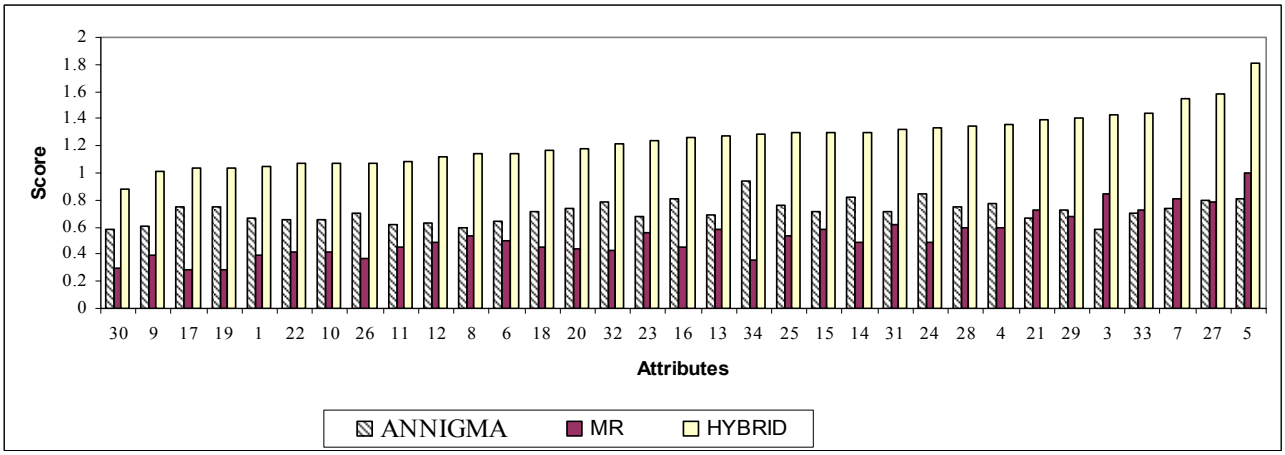

Figure 5. The computed combined score in the BE search process in the hybrid (MR-ANNIGMA) when total features is 33.

TABLE 2. FINAL ACCURACIES (%) OF MR, ANNIGMA AND HYBRID (MR-ANNIGMA) FOR DIFFERENT DATA SETS AND FINAL FEATURE SET ACHIEVED BY DIFFERENT ALGORITHMS.

Data set		Filter- MR(%)	Wrapper-ANNIGMA(%)	Hybrid (Proposed) (%)	Other (%)
Iono-sphere	Accuracy	90.14	90.057	92.137	89.8 [12]
	Features	{5,3,7,27,33,21,29,31,28,4,15, 13,23,8} =14 Features	{6, 24, 15, 14 } = 4 Features	{5,21,3,6} =4 Features	
Cancer (Diag-nostic)	Accuracy	96.148	96.499	96.626	94.9 [15]
	Features	{27,17,29,22,18,23,1,15,2,9,3 0, 7,24,20,25,8,19,3,4,21} =21 features	{22,28,2,9,25,5,1,21,10,26,3 , 29,18,12,27,8,23,25,7} =19 Features	{28,21,23,24,8,1,3, 4,7,27,14,2,22} =13 Features	
Sonar	Accuracy	83.506	83.606	84.236	83.4 [19]
	Features	{33,5,43,19,41,2,57,55,36,28, 17,10,14,27,16,20 }=17 Features	40 features	{11,12,10,9,13,21, 49,17,48,28,31,47, 36,44,37,43} =16 Features	
Pima	Accuracy	76.953	76.710	77.174	77.0 [12]
	Features	{8,2,6}=3 Features	{2,6,7,1,5}=5 Features	{7,6,2}=3 Features	

ANNIGMA, attribute-17 has the lowest MR score and the hybrid finds attribute-30 as the lowest. Therefore it eliminates attribute-30 after second cycle of BE. When the total attributes is eight (Fig. 6), the lowest score for ANNIGMA is attribute-27, attribute-14 has the lowest MR score and hybrid finds attribute-4 as the lowest. Therefore BE eliminates attribute-4. The BE process continues in MR-ANNIGMA and the highest accuracy (92.137%) is obtained with only four attributes (5, 21, 3, 6) in Table-2 and Table-3.

TABLE 3. ACCURACIES (%) OF MR, ANNIGMA AND HYBRID (MR-ANNIGMA) FOR DIFFERENT SET OF ATTRIBUTES DURING THE BE PROCESS FOR IONOSPHERE DATA SET.

Total Attributes	MR %	ANNI-GMA %	Hybrid %	Total Attributes	MR %	ANNI-GMA %	Hybrid %
34	85.897	86.439	86.011	18	**90.741**	89.687	88.946
33	83.561	87.322	82.222	17	88.177	88.063	87.464
32	86.553	84.615	84.929	16	88.746	87.208	88.718
31	87.550	86.268	87.607	15	88.519	87.464	88.063
30	86.923	85.328	86.467	14	**90.142**	89.744	89.630
29	87.664	85.613	86.724	13	88.006	86.325	88.917
28	85.385	84.843	84.501	12	88.917	88.262	89.715
27	85.527	86.752	85.157	11	88.034	86.524	89.402
26	83.390	83.789	85.271	10	88.547	88.376	91.197
25	87.179	88.120	86.724	9	88.234	88.177	88.433
24	84.359	83.504	84.046	8	86.895	86.980	89.402
23	85.413	85.556	83.875	7	87.550	88.632	90.712
22	87.920	85.527	87.436	6	87.892	**90.057**	91.738
21	85.185	86.724	85.755	5	88.063	89.829	91.595
20	85.442	86.524	84.387	4	88.462	**90.057**	**92.137**
19	88.917	88.547	89.829	3	87.550	89.886	86.752
				2	87.208	87.977	87.037

The detailed accuracies in BE process for the ionosphere data set is described in Table-2 and Table 3. The filter approach (MR) achieves an accuracy of (90.741%) with 18 attributes and (90.142%) with 14 attributes. The ANNIGMA achieves (90.057%) with four attributes. However these 4 attributes (6, 24, 15, 14) are different from final feature set (5,21,3,6) of the hybrid which has been described in Table 2. It is seen that our hybrid approach achieves the highest accuracy with very compact feature set (only 4 features in Table 2). This proves the significance of the hybrid approach in searching for the most important feature set. Detailed results are given in Table 2.

B. Wisconsin cancer (Diagnostic) data set [18]

This dataset has 30 real valued attributes with no missing values. Class variable is binary. Detailed accuracies for three algorithms (MR, ANNIGMA and hybrid) at different iterations of BE process is given in Table 4 and Table 2. In Table-4, the filter approach with 30 attributes resulted in an average generalisation across the ten trials with ten cross validation sets in each trial of 95.9%. The wrapper (ANNIGMA) alone resulted in a predictive accuracy of 96.4% and the hybrid (MR-ANNIGMA) was also similar. The removal of one feature at time according to the BE procedure led to little change in predictive accuracy across the three experimental conditions for three algorithms to the point where 22 features achieve almost identical predictive accuracy as 30.

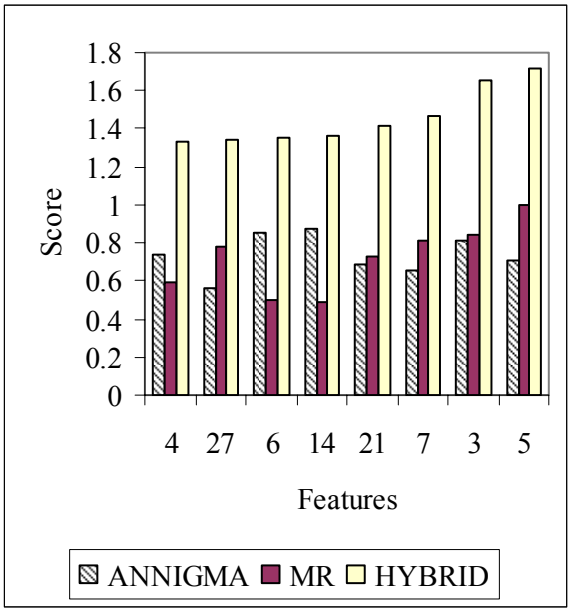

Figure 6. The computed combined score in the BE search process in the hybrid (MR-ANNIGMA) when total features is 8. Y-axis gives the combined score and X-axis gives the feature's serial no.

TABLE 4. ACCURACIES (%) OF MR, ANNIGMA AND HYBRID (MR-ANNIGMA) FOR DIFFERENT SET OF ATTRIBUTES DURING THE BE PROCESS FOR WISCONSIN CANCER (DIAGNOSTIC) DATA SET.

Total Attributes	MR %	ANNI-GMA %	Hybrid %	Total Attributes	MR %	ANNI-GMA %	Hybrid %
30	95.937	96.446	96.268	17	95.57	96.112	96.534
29	95.955	96.004	96.235	16	95.639	96.27	96.409
28	96.27	96.428	96.217	15	95.516	96.147	96.61
27	96.095	96.237	96.201	14	95.657	96.287	96.219
26	95.987	96.465	96.357	13	95.462	95.652	**96.626**
25	96.025	96.464	96.129	12	95.218	95.531	96.274
24	95.831	96.27	95.971	11	95.357	95.9	**96.467**
23	95.885	96.007	96.514	10	94.427	95.635	94.66
22	96.114	96.2	96.008	9	91.743	95.723	94.38
21	**96.148**	96.216	96.023	8	91.2	95.599	95.135
20	95.918	96.323	96.078	7	91.884	94.407	94.522
19	95.412	**96.499**	96.253	6	90.936	88.857	94.434
18	95.586	96.268	96.326	5	91.501	89.263	94.381
				4	90.464	89.455	93.171

In Table-4, the filter (MR) achieves the highest accuracy 96.14% for 21 attributes. After this the trend changes below 21 attributes where accuracy degrades down to 90% for MR. The wrapper achieves the highest accuracy 96.499% for 19 attributes. After 19 attributes accuracy degrades for wrapper to 89%. The hybrid approach achieves the highest accuracy 96.626% for 13 attributes and 96.467% for 11 attributes (Ref. Table 4). After this the hybrid's accuracy degrades to 93%. However the hybrid approach achieves the highest accuracy 96.626% in three algorithms using only 13 attributes (Ref. Table 2 and Table 4).

C. *Sonar data set [18]*

This data set has total 60 real valued attributes with binary class variable. MR, ANNIGMA and hybrid (MR-ANNIGMA) algorithms have been tested on the data set using 10 trials and 10-fold cross validation for each trial. The average accuracies over 10 trials for different iterations of BE process for three algorithms is described in Table 5 and Table 2. In table-5, it is seen that MR achieves a highest accuracy of 83.506% with 17 features. ANNIGMA achieves the highest accuracies 84.327% with 53 features, 83.606% with 40 features. After that its accuracy decreases to 75% with 4 features. The hybrid approach achieves the highest accuracy 84.717% with 34 features, 84.674% with 25 features. The hybrid also achieves an acceptable accuracy 84.236% with 16 features (Table 2 and Table 5) which is closer to the highest. Therefore, the hybrid approach obtains higher accuracy (84.236%) using fewer features than both filter and wrapper approaches.

TABLE 5. ACCURACIES (%) OF MR, ANNIGMA AND HYBRID (MR-ANNIGMA) FOR DIFFERENT SET OF ATTRIBUTES DURING THE BE PROCESS FOR SONAR DATA SET.

Total Attributes	MR %	ANNI-GMA %	Hybrid %	Total Attributes	MR %	ANNI-GMA %	Hybrid %
60	82.436	81.827	82.336	31	82.283	80.865	81.659
59	81.063	80.192	81.967	30	81.935	82.885	80.692
58	81.714	83.942	81.802	29	81.491	82.452	80.213
57	82.281	82.067	83.130	28	82.431	80.048	83.544
56	82.439	82.404	79.860	27	81.564	79.087	81.742
55	82.278	81.827	81.025	26	81.835	81.587	82.992
54	81.749	81.010	83.226	25	82.406	79.183	**84.674**
53	82.762	**84.327**	81.226	24	82.251	78.846	82.426
52	81.952	81.875	81.744	23	81.659	77.837	82.767
51	82.792	82.740	82.659	22	83.211	79.231	83.902
50	82.368	81.635	83.830	21	83.376	80.048	81.311
49	81.511	82.692	82.624	20	83.173	79.135	82.534
48	81.328	82.404	84.479	19	83.358	78.269	84.288
47	82.990	82.788	82.544	18	82.952	78.990	83.729
46	82.719	82.163	82.035	17	**83.506**	81.827	81.221
45	81.216	83.221	82.847	16	83.053	78.750	**84.236**
44	82.584	80.240	82.962	15	82.810	77.596	81.301
43	81.907	82.019	83.474	14	83.035	79.375	82.326
42	81.967	79.904	83.231	13	83.231	77.981	81.459
41	81.965	81.538	85.531	12	83.073	78.365	83.030
40	81.664	83.606	83.148	11	82.138	76.154	82.639
39	82.123	82.692	82.754	10	80.336	74.904	76.792
38	82.368	81.202	83.278	9	82.258	73.173	78.276
37	81.995	81.587	83.812	8	81.454	71.731	77.754
36	83.143	81.058	80.509	7	79.534	71.250	78.135
35	82.148	82.644	81.241	6	80.195	72.885	73.306
34	80.972	81.827	**84.717**	5	80.714	75.144	73.709
33	82.940	82.356	82.634	4	77.569	75.385	76.313
32	82.559	83.365	82.569				

D. *Pima diabetes data set [18]*

Pima Indians Diabetes data set has eight real valued attributes and binary class variable. There is no missing value. 10-fold cross validation with 10 trials is applied for each of the three algorithms (MR, ANNIGMA and hybrid). The average accuracies over 10 trials for three algorithms for different iterations of the BE process is described in Table 6. The filter (MR) achieves highest accuracies 77.018% for 7 features. It also achieves an acceptable accuracy 76.953% for 3 features. The wrapper (ANNIGMA) shows accuracy 76.99% for 7 features and 76.710% for 5 features. The hybrid approach obtains the highest accuracy 77.253% with 6 features. The hybrid also finds a second highest 77.174% for 3 features which is also acceptable. The results show that hybrid approach achieves 77.174% accuracy using fewer features than wrapper and equal to the number for filter's corresponding accuracy (76.953%). The selected features are described in Table 2.

TABLE 6. ACCURACIES (%) OF MR, ANNIGMA AND HYBRID (MR-ANNIGMA) FOR DIFFERENT SET OF ATTRIBUTES IN THE ITERATIONS OF THE BE PROCESS FOR PIMA DIABETES DATA SET.

Total Attributes	MR %	ANNI-GMA %	Hybrid %
8	76.315	76.486	76.250
7	**77.018**	**76.999**	76.888
6	76.523	76.159	**77.253**
5	75.807	**76.710**	77.109
4	75.521	76.559	75.703
3	**76.953**	76.502	**77.174**
2	76.589	76.049	76.445

Table 2 summarizes the final accuracies for all data sets for algorithms (filter-MR, wrapper-ANNIGMA and the hybrid-MR-ANNIGMA). Table 2 provides the selected final feature set and corresponding accuracies. It is seen in Table 2 that the proposed hybrid approach achieves very compact feature sets in all data sets trialled with higher accuracies than both filter and wrapper. This demonstrates that the hybridization of filter and wrapper in the MR-ANNIGMA leads to improved predictive accuracy with fewer features.

The hybrid algorithms runs a backward elimination (BE) process where each iteration involves computational time in training the network, the computation of MR score, ANNIGMA score 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. Therefore considering all features' time as constant for subsequent iterations, BE process in the hybrid generates a computational complexity of O(m) where m is the total number features in a data set.

V. CONCLUSIONS

This paper proposes a novel hybrid (MR-ANNIGMA) of wrapper and filter approaches for feature selection problem in data mining /machine learning applications. The novelty of our approach is that this introduces 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 takes

advantages of the complementary properties of the both filter heuristic and wrapper heuristics. In the (MR-ANNIGMA) approach, we combine a Mutual Information (MI) based Maximum Relevance (MR) filter ranking score with Artificial Neural Network Input Gain Measurement Approximation (ANNIGMA) based wrapper ranking score to generate the feature subset and thereby guide the search process in the wrapper.

The approach has been tested using bench mark machine learning data sets with varying number of features and sample size. Our experiments show that combined heuristic score in the MR-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. Thereby, the hybrid algorithm (MR-ANNIGMA) achieves higher accuracy and smaller feature set than both filter and wrapper approach. In the future we will use other search strategies such as bidirectional search and filter approaches with the proposed approaches.

REFERENCES:

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

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

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

[4] S. Puronnen, A, Tsymbal and I. Skrypnik, "Advanced local feature selection in Medical Diagnostics", Proc. 13th IEEE symposium computer-based medical diagnostics, pp 25-30, 2000.

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

[6] P. Mitra, C.A. Murthy 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.

[7] N. Kwak 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

[8] H Wang, D Bell, 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

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

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

[11] 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.

[12] C.N. Hsu, 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

[13] Z Zhu, 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

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

[15] 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, No. 11, Nov 2004.

[16] A. Jain 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.

[17] F. J. Ferri, 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

[18] 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]. 1997.

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