IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 4, Ver. VI (Jul.-Aug. 2016), PP 31-38
www.iosrjournals.org

A Novel Wrapper-filter Hybrid Method for Candidate SNPs Selection

Farideh Halakou
Department of Computer and Electronics, Kerman Graduate University of Technology, Kerman, Iran.

Abstract: *Genomic studies provide massive amount of data including thousands of Single Nucleotide Polymorphisms (SNPs). The analysis of SNPs helps to identify genetic variants related to complex traits. Therefore, it is essential to provide an efficient method to find a small subset of candidate SNPs as good representatives of the rest of SNPs.In this study, a new feature/SNP selection method based on the relationship between filter and wrapper criteria (i.e. correlation and model accuracy) is proposed. The method is based on the prediction of Mean Square Error (MSE) in terms of the number of features, Mean Feature-Feature Correlation (MFFC) and Mean Feature-Target Correlation (MFTC). It trains a Neural Network (NN) to predict the accuracy in terms of the number of features, MFFC and MFTC.We ran experiments on artificial SNPs datasets, comparing our algorithm with well-known feature selection techniques, and obtained higher accuracy in selecting the candidate SNPs in shorter running time.*

Keywords*:Feature Selection, Filter FS method, Wrapper FS method, Single Nucleotide Polymorphism, candidate SNPs, SNPs selection.*

I. Introduction

Feature Selection (FS) is one of the vital subjects in data mining [1,2], machine learning [3,4] and pattern recognition [5]. It tries to select an optimal subset of original features that are necessary and adequate to explain the target concept. Furthermore, it can improve the learning efficiency, increase the prediction accuracy, and reduce the complexity of models[6,7].

Real-world data sets are often consisted of many irrelevant and/or redundant features due to the lack of prior knowledge about specific problems [8]. If these features are not properly phased out, they may considerably obstruct the model accuracy and learning pace. Optimal feature selection requires an exhaustive search through the whole feature subsets, so it would be intractable and impractical. This problem is known as the 'curse of dimensionality' for large dimensional data sets [9].

FS methods can be divided into two categories: filter and wrapper methods[10]. Filter methods are based on a particular ranking criterion, and independent of any specific learning algorithm. Features are scored and ranked based on certain statistical measures, and the features with the highest ranking values are selected [11]. Due to the computational effectiveness of filter methods, they are mostly used in high-dimension data [12]. However, their accuracies are lower than that of wrappers because of the simplicity of the ranking criterion.

Wrapper FS methods, on the other hand, rely on some learning algorithms to evaluate the performance of different feature subsets according to final criterion [13].They select the features having high prediction accuracy estimated by specific learning algorithms. Considering prediction capability, wrappers can reach higher performance than filter-based methods [8,14]. However, they are often impractical for large scale problems [15].

The hybrid approaches attempt to take advantages of both the filters and wrappers by using their complementary strengths [16,17]. The idea behind a hybrid method is to apply a filter method at first to select a feature subset,and then apply a wrapper method to find the optimal subset of features from the selected feature set. These methods are more feasible in real bioinformatics applications which usually have a large number of features.

In this paper, we have proposed a new wrapper-filter feature selection algorithm. Our method predicts Mean Square Error (MSE) based on the number of features, mean feature-feature correlation and mean feature-target correlation. In addition, it benefits from Genetic Algorithm (GA) to search the problem domain. The methodis applied on Single Nucleotide Polymorphisms (SNPs) data sets. SNPs are primarily responsible for the variation between humans. They promise to significantly advance our ability of understanding and treating diseases [18].

This paper includes six sections. Next section reviews briefly previous works on different feature selection algorithms. Section 3 introduces the concept of the proposed algorithm and explains it in detail. The SNPs data sets and experimental results are explained in section 4. Finally, discussion and conclusions are given in section 5 and 6, respectively.

II. Related Works

So far, many selection methods have been proposed to identify relevant features. This section briefly reviews the recent proposed FS methods. Banerjee and Pal [19] proposed a SVD-entropy based supervised FS algorithm which was a modification of the method proposed by Varshavsky et al. [20] for unsupervised feature selection to improve its performance. The results represented the effectiveness of the developed algorithm in selecting relevant features. Kabiret al.proposed a feature selection algorithm called Constructive Approach for FS (CAFS) based on the idea of the wrapper approach and sequential search strategy [21]. As a learning model, CAFS employed a usual three layered feed-forward neural network. The essential aspect of this algorithm was the automatic determination of neural network architecture through the FS process. The proposed technique combined the FS with the architecture determination of the neural network. They evaluated the performance of CAFS on eight benchmark classification problems. The experimental results showed the essence of CAFS in selecting features with compact neural network architectures. In another work, ElAlami described a feature subset selection algorithm which uses GA to optimize the output nodes of a trained neural network [22]. This algorithm did not depend on the neural network training algorithms or it did not alter the training results. The experimental results demonstrated that the proposed algorithm was very effective in reducing dimensionality and removing irrelevant features.

Some other studies focused on the Mixture of Experts (MoE) techniques [23]which apply 'divide and conquer' approach by jointly training a set of classifiers that are specialized in different regions of the input space. For example, Peralta and Soto [24] proposed a regularized variant of MoE, called RMoE technique, which incorporates an embedded process for local FS using L1 regularization. Experiments with artificial and real-world datasets proved that RMoE improves the classical MoE technique, in terms of accuracy and sparseness of the solution.

Some studies proposed hybrid methods to make use of the advantages of the filter and wrapper methods. For example, a hybrid feature selection method called SU-GA-W was proposed by Jianget al.[13]. This method included two phases. The filter phase removes some features and uses the feature estimation as the heuristic information to guide GA. They adapt symmetrical uncertainty [25] to get feature estimation. The second phase is a GA-based wrapper selector. The effectiveness of this method is demonstrated through empirical study on UCI data sets. Hu and Wu indicated some problems in hybrid FS methods which included filter and wrapper algorithms in first and second stages [14]. In the first stage, selecting suitable evaluation criterion is vital in conventional feature ranking. In the second stage, the wrapper methods may cost too expensive due to the high dimensions. In addition, there is no relationship between the two phases as it is just a combination of the filter and wrapper methods together. Therefore, they proposed a novel Filter-Wrapper Hybrid Method (FWHM) to optimize the efficiency of feature selection. It was designed based on the combination of multi-criterion filters and improved immune wrapper method. The experiments on benchmark model and engineering model proved that FWHM gained better performance both in accuracy and efficiency than the conventional methods.

Several computational methods for SNPs selection have been proposed in the past few years. Mahdevar et al. indicated that molecular haplotyping methods were expensive, difficult, and time consuming; therefore, algorithms for constructing full haplotype patterns from small available data through computational methods are more suitable [26]. They proposed a heuristic method, called GTagger (Genetic Tagger), based on genetic algorithm to find reasonable solution within acceptable time. The algorithm helped to reduce the number of SNPs without losing crucial information. It removed cost of genotyping unnecessary SNPs and reduced the problem size. The GA fitness function was based on the least number of Haplotype Tagging SNPs (htSNPs) and combined with Shannon entropy function. The algorithm was implemented on a variety of simulated and biological data sets. In comparison with the brute force approach, results showed that their method could obtain optimal solutions in almost all cases and ran much faster when the number of SNP sites was large. Long et al. developed a two-step FS method for binary qualities, which included filtering and wrapping [27]. The filter step used information gain and reduced thousands of SNPs to a small number. The wrapper step optimized the performance of the filtered SNPs via a naive Bayesian classifier. An approach based on discretization [28,29] of phenotypic values was developed to allow feature selection in a classification framework. The method was applied to chick death rates (0–14 days of age) on offspringfrom 201 sires in a commercial broiler line, with the goal of identifying SNPs (over 5000) related to offspring death. Experimental results suggested that the FS method, tied with sample partition and subset evaluation procedures, provided a helpful tool for finding important SNPs. For a review of applications of different feature selection techniques in bioinformatics see Saeyset al.[30].

In this paper, a new wrapper-filter method for candidate SNPs selection is presented. Our method works based on a relationship between filter and wrapper criteria. We express the steps of the algorithm in detail in section 3.

III. Proposed Approach

As mentioned earlier, our intention was to find a relationship between filter and wrapper criteria, i.e. correlation and model accuracy, in such a manner that it could be used for feature selection. We called this method as PrMSE_RRL (Predicting MSE based on mean feature-target correlation ($\overline{\text{Rcf}}$), mean feature-feature correlation ($\overline{\text{Rff}}$), and feature Length/number of features (L). The main steps of PrMSE_RRL algorithm are shown in **Error! Reference source not found.**.

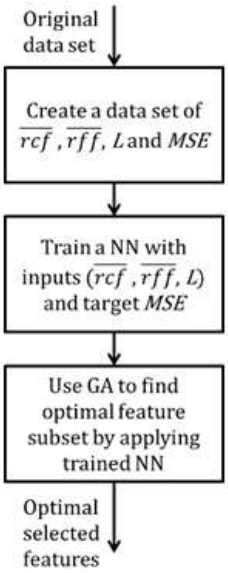

Fig. 1: Proposed PrMSE_RRL feature selection algorithm

Step 1: The process begins with creating a dataset of $\overline{rcf}$, $\overline{rff}$, *L* and *MSE*. For producing one row of the dataset, a random subset of features is selected. Then, *L*, $\overline{rcf}$ and $\overline{rff}$ are calculated for the selected features. We used linear regression to predict the value of target concept via the selected features. The *MSE* value that is calculated by using the linear regression, as well as the tri-tuple ($L, \overline{rff}, \overline{rcf}$) is recorded as one row in the dataset. This procedure is repeated several times and finally a dataset with four columns is generated. The first three columns are $L, \overline{rff}, \overline{rcf}$ and the last one is the values of *MSE*.

Step 2:A feed forward back propagation neural network is trained on the previous data set. Neural network inputs are $L, \overline{rff}, \overline{rcf}$, and its output is *MSE*.

Step 3:In this step, GA is used to find the optimal feature subset. Each individual (chromosome) of GA, codes one selected subset of features that must be evaluated. For each individual solution of GA (a selected subset), $\overline{rff}, \overline{rcf}$ and *L* are calculated. Then, the trained neural network in step (2) is appliedto evaluate this individual solution (i.e. chromosome). The fitness function of GA is set to the predicted *MSE* of neural network as well as the size of the selected features.

One important subject is that in large dimensional domains, the first step of this method is computationally prohibitive and the size of the generated dataset has to be very large. Therefore, to come up with this problem, a filter based method is applied to reduce the dimensionality before applying PrMSE_RRL procedure.

IV. Experimental Results

1.1. Applied Datasets

One of the most important ways to understand the genetic basis of complex diseases such as cancer, drug response or other human phenotypes is genetic association studies. The goal of these studies is to discover relations between DNA variants and such traits by comparing genetic sequence and phenotypes of individuals sampled from a population [27]. SNPs are by far the most rampant of all DNA sequence variations and very useful in genetic association studies. Besides the obvious applications in human disease studies, they are also extremely useful in genetic studies of all organisms, from model organisms to commercially important plants and animals [30]. SNPs most generally refer to single-base differences in DNA among individuals. They occur once in every 300 nucleotides on average, which means there are roughly 10 million SNPs in the human genome. They can act as biological markers, helping scientists to locate genes that are associated with disease.

SNPs are bi-allelic, i.e. the number of different values of SNPs is just two, which are only two nucleotides out of four possible nucleotides may be selected as the values of SNPs [27]. Therefore, each SNP can be represented by a binary variable. Since the number of distinctive combinations of SNP alleles within a block is pretty small, selecting a small subset of SNPs that efficiently represent other SNPs in a given block is a key problem for reducing genotyping costs without losing the ability to detect disease associations [26].

Since SNPs selection is a tough problem in bioinformatics, the proposed feature selection algorithm is evaluated on a set of SNPs data. We used simulated datasets, since in real data the significant SNPs are unknown. One hundred populations each with 500 individuals were artificially produced. The genome of each individual consisted of 9 chromosomes and each chromosome with 101 SNPs leading to a total number of909 SNPs. Seven SNPs out of these 909 SNPs were relevant (SNPs number 31, 71, 132, 172, 253, 334 and 405 which are located on the first five chromosomes). The target concept was a continuous variable (i.e. a phenotype variable such as the weight of individuals with mean of 36.0 and residual error set to 1) and its values were in the range of 32-42.

1.2. Results & Evaluation

In this paper, we compared our proposed method with six FS methods including three filters and three hybrids:

1. CfsSubsetEval is a pure filter FS method which evaluates the worth of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them [32]. Subsets of the features that are highly correlated with the class while having low inter-correlation are preferred.
2. ReliefF algorithm is fairly different to CfsSubsetEval in that it scores individual features rather than feature subsets [4,33,34]. To use ReliefF for feature selection, those features with scores exceeding a user-specified threshold are retained to form the final subset. ReliefF evaluates the worth of an attribute by repeatedly sampling an instance and considering the value of the given attribute for the nearest instance of the same and different class. It can operate on both discrete and continuous data.
3. Decision rule search uses decision rule based heuristic search to eliminate all irrelevant and redundant features, based on domain specific definitions of high, medium, and low correlation [35]. Thresholds to determine the amounts of low, medium, and high are determined by the user which makes the method more flexible.
4. Ck_NNFS, CNNFS, and CRRFS are hybrid methods which use correlation-based feature selection approach as filter method in the first step [36-38]. Then, they use first step's selected features as input of the wrapper method. In this step, Ck_NNFS uses k-Nearest Neighbor (k_NN) algorithm, CNNFS uses a feed-forward back propagation Neural Network, and CRRFS uses Ridge Regression to select the optimal feature subset. In both steps, GA is applied to search the feature space.

As mentioned earlier, our selected data sets were consisted of 909 SNPs which were not manageable by a wrapper method. Therefore, we applied a filter method at first. Correlation-based Feature Selection (CFS) approach was used as a filter method. In this method, feature relevance is measured based on the correlation between a feature and the target concept.Feature redundancy is defined based on the correlation between a feature and other features. This correlation measure is defined by[35]:

$$Merit_S = \frac{L\bar{r}_{cf}}{\sqrt{L + L(L-1)\bar{r}_{ff}}} \quad (1)$$

Where L is the number of features, $\bar{r}_{cf}$ is the mean correlation between each feature and the target function, and $\bar{r}_{ff}$ is the mean correlation between features. The filter approach here uses genetic algorithm as global search to find a subset of relevant features (SNPs) from the original feature set. The population size in GA was 50 individuals within 1000 generations. The crossover and mutation fractions were 0.8 and 0.2, respectively.

The above introduced merit in Equation (1) is used to calculate the fitness of the selected features by GA. The exact formula which we used as fitness function was as follows:

$$f = 10000 \times (1 - Merit) \quad (2)$$

Since *Merit* values are in the range of 0-1, and GA tries to minimize the fitness function, we used*(1-Merit)* as fitness because our intention was to maximize the *Merit* criterion. To enlarge the differences of the *Merit* values in different subsets, we multiplied it by a large number.

After setting these parameters for GA, the number of selected features was in the range 80-130, so they were manageable by the proposed method in the next step.As the first step of PrMSE_RRL, we created adata set of 1000 rows based on the selected SNPs by the CFS method in the previous step. Each row in the data set consisted of$L, \overline{rff}, \overline{rcf}$, and *MSE*. Afterwards, a feed forward neural network was trained on this data set.

Several trial and errors were performed with neural networks containing different number of neurons in the hidden layer to determine the optimal structure which gave the best performance. After the training, genetic algorithm is applied to select the most important SNPs. In this step, the GA population consisted of 100 individuals with 500 generations. The crossover and the mutation fractions were 0.8 and 0.2, respectively. The exact formula used for fitness function was as follows:

$$f = (7 \times \exp(MSE)) + L \quad (3)$$

Where *L* is the number of selected SNPs by the genetic algorithm, and *MSE* is predicted by neural network simulation.Since the *MSE* values were small, we used exponential of the values to enlarge the differences of the *MSE* values for different feature subsets.

The experimental results concerning feature selection performance on the SNPs datasets by using seven different feature selection algorithms are presented in Table I. The terms used in Table I are described as follows (all of them were calculated on 100 data sets):

Precision: this criterion was defined as follow:

$$precision = \frac{number\ of\ relevant\ features\ retrieved}{Total\ number\ of\ retrieved\ features} \quad (4)$$

where *relevant features* are the seven important SNPs and *retrieved features* are the selected SNPs by a FS method. High precision means an algorithm returned more relevant SNPs than irrelevant. Its value is between 0-1.

Recall: this criterion was defined as:

$$recall = \frac{number\ of\ relevant\ features\ retrieved}{Total\ number\ of\ relevant\ features} \quad (5)$$

High recall means an algorithm returned most of the relevant results. Its value is between 0-1.

F-measure: combines recall and precision with equal weights into a single utility function as follows:

$$F = \frac{2 \times precision \times recall}{precision + recall} \quad (6)$$

Its value is between 0-1.

Linked results (%): it determines the selection rate of the five important chromosomes, i.e. chromosomes 1, 2, 3, 4 and 5 in our data sets. It is equal to or greater than precision. This measure is important because the SNPs on the same chromosomes have high correlation with each other. Their correlation depends on their distances reversely, i.e. when two SNPs are near each other, their correlation is high and vice versa. Therefore, FS methods may select the correlated SNPsadjacent tothe important ones.

As mentioned earlier, CfsSubsetEval, ReliefF and Decision rule search were filter FS methods while Ck-NNFS, CNNFS and CRRFS were hybrid ones. In terms of precision, PrMSE_RRL (0.39) was superior overall methods;however, its recall was less than that of filters. This can be justified by considering that the number of selected SNPs in filter methods was greater than hybrids e.g. 190-419 in Decision rule search, so the probability of selecting the candidate SNPs increased significantly in filter methods. Considering F_measure, it can be clearly seen that hybrid methods, except Ck_NNFS, reached better results in comparison to the filter methods.

PrMSE_RRL and CRRFS did not select any irrelevant chromosomes. In addition, PrMSE_RRL selected roughly the correct number of important SNPs (5-7). Moreover, the running time of the method was much less than that of Ck_NNFS and CNNFS as PrMSE_RRL trained the neural network just once, and then applied it as an embedded induction model throughout the GA generations.

Candidate SNPs' selection rate using PrMSE_RRL is listed in Table II. It is clear that there are oscillations in the selection rate foreach important SNP e.g. selection rate of SNPs number 172 and 31 were 6% and 55%, respectively. Among the seven important SNPs, SNP number 31 has the highest selection rate (55%), which means the first important SNP is the most relevant featureto the target. Note that SNPs located on chromosome 2, i.e. SNPs 132,172, have the lowest selection rates compared to the other SNPs.

Table I: A comparison of PrMSE_RRL with six other FS methods on SNPs data

SNPs selection method	No. of selected SNPs[1]	Precision	Recall	F_measure	Linked results (%)	Average time (s) [2]
Hybrid Methods						
PrMSE_RRL	**5-7**	**0.39**	0.34	**0.37**	**100**	775
Ck-NNFS	2-38	0. 12	0.24	0.16	90.88	2506
CNNFS	4-7	0.32	0.26	0.29	99.82	4062
CRRFS	4-7	0.38	0.30	0.34	**100**	35
Filter Methods						
CfsSubsetEval	7-21	0.23	0.36	0.28	74.77	**3**
ReliefF	358-484	0.02	**0.95**	0.03	75.89	30
Decision rule search	190-419	0.02	0.81	0.04	98.80	5

[1] The number of selected SNPs by each method; [2] Average running time of an algorithm in seconds.

Table II: Important SNPs' selection rate using PrMSE_RRL

Important SNPs No.	Selection rate (%) [1]
31	55
71	50
132	11
172	6
253	29
334	42
405	47

[1] Selection rate of each important SNP.

V. Discussion

Based on the results in Table I, the hybrid methods showed better performance although theywere slower than filter methods. Number of selected SNPs and precision of ReliefF and Decision rule search were totally unacceptable in comparison with the others. In addition, they had the lowest F-measures. Among filter methods, CfsSubsetEval had the smallest number of SNPs and highest F_measure. Hybrid methods selected vital chromosomes with a high rate (more than 90%). Furthermore, they achieved higher dimensionality reduction by selecting less number of SNPs than pure filters.PrMSE_RRL, CNNFS, and CRRFS selected the correct number of important SNPs relatively.

In order to obtain the statistical support, the Friedman test [39] between F_measures of each FS method is used. Average ranks obtained by applying the Friedman procedure are given in Table III.

Table III: Average rankings of the algorithms

Algorithm	Ranking
PrMSE_RRL	**2.305**
Ck-NNFS	4.42
CNNFS	3.255
CRRFS	2.745
CfsSubsetEval	3.155
ReliefF	6.4
Decision rule search	5.72

Friedman statistic considering reduction performance distributed according to chi-square with 5 degrees of freedom: 313.112143. P-value computed by Friedman test was 1.5376e-10. Table IIIrepresents that the best performing algorithm was PrMSE_RRL. Hybrid methods had better ranks in comparison to the filters except CfsSubsetEval. To determine whether the differences among the methods were significant or not, the Holm's post-hoc test [40] have been performed. Results achieved on post hoc comparisons for α=0.05 are shown in Table IV.

Holm's procedure rejected those hypotheses with a p-value$\leq$0.01. Therefore, there is no significant difference between algorithms on hypothesis 18 to 21, so PrMSE_RRL has the same result as CRRFS. This conclusion could be resulted by regarding to the ranks of the methods on Table III; there are no significant differences between the rank of PrMSE_RRL and CRRFS.

Finally, to have a pairwise comparison between the methods, the Wilcoxon test [41] was applied. Its results are shown in Table V.Based on this table, PrMSE_RRL and CRRFS are the best performing methods. CNNFS and CfsSubsetEval are equivalent, and they outperform Decision rule search, ReliefF and Ck-NNFS. After them, Ck-NNFS,Decision rule search, and ReliefF got the third to fifth ranks, respectively.These statistical tests confirmed our previous conclusions about the performance of the methods. However, they showed a good quality of the CfsSubsetEval as these tests performed just based on the F-measures.

VI. Conclusions

In this paper, we proposed a new wrapper-filter hybrid method for candidate SNPs selection. We planned to find a relationship between filter and wrapper criteria, so we introduced PrMSE_RRL method. PrMSE_RRL is a novel three-stage filter-wrapper feature selection method which predicts Mean Square Error (MSE) based on the number of features, Mean Feature-Feature Correlation (MFFC) and Mean Feature-Target Correlation (MFTC). Genetic algorithm was used to find optimal feature subset based on the MSE minimization. Our proposed procedure was much faster than a wrapper method because theneural network was trained only once, and then it was used as an embedded induction model throughoutGA generations. Experimental results demonstrated that our method found the candidate SNPs with considerably smaller running time and better accuracy than some other FS methods. Furthermore, it had a considerable dimension reduction ability.

Table IV:P-values for $\alpha = 0.05$

	hypotheses	P	Holm
1	ReliefF vs. PrMSE_RRL	0	0.002381
2	ReliefF vs. CRRFS	0	0.0025
3	DRS[1] vs. PrMSE_RRL	0	0.002632
4	CfsSubsetEval vs. RReliefF	0	0.002778
5	ReliefF vs. CNNFS	0	0.002941
6	DRS vs. CRRFS	0	0.003125
7	CfsSubsetEval vs. DRS	0	0.003333
8	DRS vs. CNNFS	0	0.003571
9	Ck-NNFS vs. PrMSE_RRL	0	0.003846
10	ReliefF vs. Ck-NNFS	0	0.004167
11	Ck-NNFS vs. CRRFS	0	0.004545
12	DRS vs. Ck-NNFS	0.000021	0.005
13	CfsSubsetEval vs. Ck-NNFS	0.000035	0.005556
14	Ck-NNFS vs. CNNFS	0.000137	0.00625
15	CNNFS vs. PrMSE_RRL	0.001873	0.007143
16	CfsSubsetEval vs. PrMSE_RRL	0.005398	0.008333
17	ReliefF vs. DRS	0.026026	0.01
18	**CNNFS vs. CRRFS**	**0.095045**	**0.0125**
19	**CRRFS vs. PrMSE_RRL**	**0.1498**	**0.016667**
20	**CfsSubsetEval vs. CRRFS**	**0.179583**	**0.025**
21	**CfsSubsetEval vs. CNNFS**	**0.743421**	**0.05**

[1] Decision Rule Search

Table V: Summary of the Wilcoxon test.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
CfsSubsetEval (1)	-	•	•	•		∘	∘
ReliefF (2)	∘	-	∘	∘	∘	∘	∘
Decision rule search (3)	∘	•	-	∘	∘	∘	∘
Ck-NNFS (4)	∘	•	•	-	∘	∘	∘
CNNFS (5)		•	•	•	-	∘	∘
CRRFS (6)	•	•	•	•	•	-	
PrMSE_RRL (7)	•	•	•	•	•		-

• = the method in the row improves the method of the column. ∘ = the method in the column improves the method of the row. Upper diagonal of level significance $\alpha = 0.9$. Lower diagonal level of significance $\alpha = 0.95$.

References

[1]. Y. Kim, W. Street, and F. Menczer, "Feature selection in unsupervised learning via evolutionary search,"in*proc. KDD-2000: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 365–369.

[2]. M. Dash, K. Choi, P. Scheuermann, and H. Liu, "Feature Selection for Clustering - A Filter Solution,"in *proc.ICDM 2002: Proceedings ofIEEE International Conference on Data Mining*, pp. 115-122.

[3]. H. Liu, H. Motoda, and L. Yu,"Feature Selection with Selective Sampling,"in*proc.ICML-2002: Proceedings of the 19th International Conference on Machine Learning*, pp. 395-402.

[4]. M. R. Sikonja and I. Kononenko,"Theoretical and empirical analysis of Relief and ReliefF,"*Machine Learning*, vol. 53, pp.23-69, Oct.-Nov. 2003.

[5]. P. Mitra, C.A. Murthy, and S.K. Pal,"Unsupervised feature selection using feature similarity,"*IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, pp.301-312, Mar. 2002.

[6]. A.L. Blum and P. Langley,"Selection of relevant features and examples in machine learning,"*Artificial Intelligence*, vol. 97, pp.245-271, Dec. 1997.

[7]. D. Koller and M. Sahami,"Toward optimal feature selection,"in*proc.ICML-96: Proceedings of the 13th International Conference on Machine Learning*, pp. 284–292.

[8]. M. Dash and H. Liu, "Feature selection for classification,"*Intelligent Data Analysis*, vol. 1, pp.131-156, Mar. 1997.

[9]. S.D. Bhavani, T.S. Rani, and R. Bapi, "Feature selection using correlation fractal dimension: Issues and applicationsin binary classification problems,"*Applied Soft Computing*, vol. 8, pp.555-563, Jan. 2008.

[10]. H. Almuallim and T.G. Dietterich, "Learning boolean concepts in the presence of many irrelevant features,"*Artificial Intelligence*, vol. 69, pp.279-305, Sep. 1994.
[11]. I. A. Gheyas and L. S. Smith, "Feature subset selection in large dimensionality domains," *Pattern Recognition*, vol. 43, pp. 5-13, Jan. 2010.
[12]. R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, pp. 273-324, Dec. 1997.
[13]. B. N. Jiang, X. Q. Ding, and L. T. Ma, "A hybrid feature selection algorithm: combination of symmetrical uncertainty and genetic algorithms," in *proc. OSB'08:Proceedings of the 2nd International Symposium Optimization and Systems Biology*, pp. 152-157.
[14]. M. Hu and F. Wu, "Filter-Wrapper Hybrid Method on Feature Selection," in *proc.GCIS 2010: Proceedings of the 2nd WRI Global Congress on Intelligent Systems*, pp. 98-101.
[15]. H. Liu, J. Sun, L. Liu, and H. Zhang, "Feature selection with dynamic mutual information," *Pattern Recognition*, vol. 42, pp. 1330–1339, Jul. 2009.
[16]. J. Huang, Y. Cai, and X. Xu, "A hybrid genetic algorithm for feature selection wrapper based on mutual information," *Pattern Recognition Letters*, vol. 28, pp. 1825-1844, Oct. 2007.
[17]. R. K. Sivagaminathan and S. Ramakrishnan, "A hybrid approach for feature subset selection using neural networks and ant colony optimization," *Expert Systems with Applications*, vol. 33, pp. 49-60, Jul. 2007.
[18]. S. C. Shah and A. Kusiak, "Data mining and genetic algorithm based gene/SNP selection," *Artificial Intelligence in Medicine*, vol. 31, pp. 183-196, Jul. 2004.
[19]. M. Banerjee and N. R. Pal, "Feature selection with SVD entropy: Some modification and extension,"*Information Sciences*, vol. 264, pp. 118-134, Apr. 2014.
[20]. R.Varshavsky, A.Gottlieb, M.Linial, and D.Horn, "Novel unsupervised feature filtering of biological data,"*Bioinformatics*, vol. 22, pp. 507-513, Jul.2006.
[21]. M. M. Kabir, M. M. Islam, and K. Murase, "A new wrapper feature selection approach using neural network," *Neurocomputing*, vol. 73, pp. 3273-3283, Oct. 2010.
[22]. M. E. ElAlami, "A filter model for feature subset selection based on genetic algorithm," *Knowledge-Based Systems*, vol. 22, pp. 356-362, Jul. 2009.
[23]. R.Jacobs, M.Jordan, S.Nowlan, and G.Hinton,"Adaptive mixtures of local experts,"*Neural Computation*, vol. 3, pp. 79-87, Spring 1991.
[24]. B. Peralta and A Soto,"Embedded local feature selection within mixture of experts,"*Information Sciences*, vol. 296, pp. 176-187, Jun.2014.
[25]. L. X. Zhang, J. X. Wang, Y. N. Zhao, and Z. H. Yang, "A novel hybrid feature selection algorithm: using ReliefF estimation for GA-Wrapper search," in *proc. ICMLC 2003: Proceedings of the International Conference on Machine Learning and Cybernetics*, pp. 380-384.
[26]. Gh. Mahdevar, J. Zahiri, M. Sadeghi, A. N. Dalini, and H. Ahrabian, "Tag SNP selection via a genetic algorithm," *Journal of Biomedical Informatics*, vol. 43, pp. 800-804, Oct. 2010.
[27]. N. Long, D. Gianola, G. M. Rosa, K. A. Weigel, and S. Avendano, "Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers," *Journal of Animal Breeding and Genetics*, vol. 124, pp. 377-389, Sep. 2007.
[28]. S. S. Kannan and N. Ramaraj, "An improved correlation-based algorithm with discretization for attribute reduction in data clustering," *Data Science Journal*, vol. 8, pp. 125-138, Jun. 2009.
[29]. C. J. Tsai, C. I. Lee, and W. P. Yang, "A discretization algorithm based on Class-Attribute Contingency Coefficient," *Information Sciences*, vol. 178, pp.714-731, Feb. 2008.
[30]. Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Gene expression*, vol. 23, pp. 2507-2517, Aug. 2007.
[31]. C. S. Carlson, M. A. Eberle, L. Kruglyak, and D. A. Nickerson, "Mapping complex disease loci in whole genome association studies," *Nature*, vol. 429, pp. 446-452, May 2004.
[32]. M. A. Hall, "Correlation-based feature selection for machine learning," PhD dissertation, Dept. Computer Science, Univ. Waikato, Hamilton, New Zealand, 1999.
[33]. I. Kononenko, "Estimating Attributes: Analysis and Extensions of RELIEF," in *proc. ECML-94: Proceedings of the European Conference on Machine Learning*, pp. 171-182.
[34]. M. R. Sikonja and I. Kononenko, "An adaptation of Relief for attribute estimation in regression," in *proc. ICML-1997: Proceedings of the 14th International Conference on Machine Learning*, pp. 296-304.
[35]. P. E. Lutu and A. P. Engelbrecht , "A decision rule-based method for feature selection in predictive data mining," *Expert Systems with Applications*, vol. 37, pp. 602-609, Jan. 2010.
[36]. F. Halakou, M. Eftekhari, and A. K. Esmailizadeh, "Important SNPs selection via a combination of correlation criterion and k-nearest neighbour algorithm," presented at the 3rdNational Conference on the Applications of Mathematics and Control Theory in Medical Sciences, Bojnourd, Iran, 2011.
[37]. F. Halakou, M. Eftekhari, A. K. Esmailizadeh, and H. Sanatnama, "Using a hybrid of correlation merit, neural network and Genetic Algorithm for selecting important SNPs," presented at the 7th National Congress of Biotechnology, Tehran, Iran, 2011. Available: http://www.ibp.ir/download/new7biotech2011/489.pdf
[38]. F. Halakou, M. Eftekhari, and A. K. Esmailizadeh, "Combination of correlation coefficient and ridge regression for selecting vital features of Single Nucleotide Polymorphisms data set," in *Proc. 2011 the 19th Iranian Conference on Electrical Engineering Conf.*, pp. 3192-3197.
[39]. M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance,"*Journal of the American Statistical Association*, vol. 32, pp. 675-701, Dec. 1937.
[40]. S. Holm, "A simple sequentially rejective multiple test procedure,"*Scandinavian Journal of Statistics*, vol. 6, pp. 65-70, Sep. 1979.
[41]. F. Wilcoxon, "Individual comparisons by ranking methods,"*Biometrics Bulletin*, vol. 1, pp. 80-83, Dec.1945.