Comparative study on Single Nucleotide Polymorphisms selection via intelligent methods

Farideh Halakou[*,1], Mahdi Eftekhari[2], Ali K. Esmailizadeh[3]

[1] Department of Information Technology, Kerman Graduate University of Technology, Kerman, Iran.

Tel: +981722772467. Email: farideh.halakou@gmail.com

[2] Department of Computer Engineering, Shahid Bahonar University of Kerman, Kerman, Iran.
Tel and Fax: +983413235901. Email: m.eftekhari@mail.uk.ac.ir

[3] Department of Animal Science, Faculty of Agriculture, Shahid Bahonar University of Kerman, Kerman, Iran.

Tel: +983413202691. Fax: +983413222043. P.O.Box: 76169-133. E-mail: aliesmaili@uk.ac.ir

Abstract
During the last decade, applying feature selection approaches in bioinformatics has become an essential need for model building. This is due to the high dimensional nature of many modeling tasks in bioinformatics of them being Single Nucleotide Polymorphisms (SNPs) selection. In this paper, we propose three hybrid feature/SNP selection methods which combine filters and wrappers. In the methods, at the first step filter techniques were used to remove the irrelevant/redundant features quickly. In the second step, wrapper techniques were utilized to refine the primary feature subset obtained in the first step. We used Neural Network, k-Nearest Neighbor and Ridge Regression as induction algorithms in wrapper phase. We compared our methods with three well-known filters. The results demonstrated that the hybrid methods have much better accuracy and higher level of dimensionality reduction compared to filter methods. They identified important chromosomes with great accuracy. Among three hybrid methods, CNNFS and CRRFS had better dimension reduction ability. The CRRFS algorithm had the most satisfactory results in terms of precision of recognizing vital SNPs and recall of retrieving them in the final subset. This algorithm also showed the best performance comparing to the other hybrids regarding to running time.

Keywords: Feature selection; Single Nucleotide Polymorphisms; Neural Network; K-Nearest Neighbor; Ridge Regression.

Introduction

Feature selection (FS) aims to decrease the dimensionality of large scale datasets without losing useful information. However, searching for an optimal feature subset from a high dimensional feature space is known to be a NP-complete problem. FS algorithms are divided into two categories; the filter model and the wrapper model (Witten and Frank, 2005). The filter model relies on general characteristics of a training dataset to select relevant features without involving any learning algorithms while the wrapper model (Kohavi and John, 1997; Hsu *et al.*, 2002; Kabir *et al.*, 2010) requires one predetermined learning model and selects features with the aim of improving the generalization performance of that particular learning model. Since taking prediction accuracy into consideration, the wrapper methods can reach better results than others. However, the wrapper methods are less general in use and need more computational resources because they use specified learning algorithms.

Filter approaches mainly identify a feature subset from the original feature set by applying given evaluation criteria, which are independent of learning algorithms. Due to the computational efficiency of the filter methods, they are very helpful for high-dimensional data. Nowadays a lot of filter algorithms, such as Correlation-based Feature Selection (CFS) (Hall, 1999), Markov blanket filter (MBF) (Koller and Sahami, 1996) and Information gain (Bassat, 1982) have been developed.

Hybrid approaches, combining the filters and wrappers utilize the advantages of both methods (Sivagaminathan and Ramakrishnan, 2007; Jiang *et al.*, 2008). Although they are not as fast as pure filters, they can achieve better results. Hybrid methods take less computational time and have less complexity than pure wrappers. The idea behind the hybrid method is that a filter method is first applied to select a feature subset and then a wrapper method is applied to find the optimal subset of the features from the selected feature set. In addition, the risk of eliminating relevant features by filter methods is minimized if the filter cut-off point for a ranked list of features is set low.

In this paper, we propose three hybrid FS methods with different wrapper phases, and then compare them to three well-known filters. Correlation measure is used as filtering criterion in the proposed methods, after that Neural Network, k-Nearest Neighbor and Ridge Regression are used as induction algorithms in wrapper phase. In these approaches, in both steps we use Genetic Algorithm (GA) to search the problem domain. These mechanisms are applied to a tough bioinformatics problem, named Single Nucleotide Polymorphisms (SNPs) selection. SNPs are primarily responsible for the variation between humans. Their importance revolves around the fact that they significantly advance our ability to understand and treat diseases (Shah and Kusiak, 2004).

Materials and Methods

As mentioned earlier, wrapper methods use a learning algorithm to measure the goodness of a feature set, so they always achieve better results than pure filters although they perform slowly. The procedure repeats until the result starts to get worse or the number of features reaches a predetermined threshold. Since our selected data set is high dimensional, it would take long hours to get the results through these type of methods.

Unlike wrappers, filter methods can calculate the information of a feature set by various statistical measures. They are often applied to high-dimensional data because they calculate fast. In this paper, we examine three pure filters named CFS (Correlation-based Feature Selection) (Hall, 1999), Decision rule search (Lutu and Engelbrecht, 2010), and ReliefF (Kira and Rendell, 1992; Kononenko, 1994; Sikonja and Kononenko, 1997).

Finally, in hybrid techniques at first a filter method is applied to select a feature subset and then a wrapper one is applied 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 amount of features. The mechanism takes advantage of both the efficiency of filters and the accuracy of wrappers.

In this paper, we implement three hybrid FS methods to find the optimal subsets of SNPs. Fig 1 shows the proposed hybrid feature selection procedure.

Fig. 1.

In all of these hybrid approaches, correlation-based FS method was chosen as filter model to remove the most redundant or irrelevant SNPs. Then, a wrapper model is applied to improve the accuracy of the results. We used three different wrapper models: k-Nearest Neighbor (k-NN), Neural Network (NN), and Ridge Regression (RR).

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 detect relations between genetic variations and such traits, by comparing genetic sequence and phenotypes of individuals sampled from a population (Carlson *et al.*, 2004). Single nucleotide polymorphisms are by far the most prevalent 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 (Saeys *et al.*, 2007). SNPs most commonly 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 diseases. SNPs are bi-allelic, i.e. the number of distinct values of SNPs is just two, which are only two nucleotides out of four possible nucleotides may be selected as the values of SNPs (Long *et al.*, 2007). Therefore each SNP can be represented by a binary variable. Since the number of unique combinations of SNP alleles within a block is pretty small, thus selecting a small subset of SNPs that efficiently represent other SNPs in a given block is an important problem for reducing genotyping costs without losing the ability to detect disease associations. This process is known as Tag SNP selection (Mahdevar *et al.*, 2010).

Since Tag SNP selection is a challenging problem in bioinformatics, we evaluated the feature selection algorithms on a set of SNPs data. In this paper, we used simulated data sets since in real data the relevant SNPs are unknown, so there was no way to precisely compare FS approaches. We produced 100 populations. There were 500 individuals in each population. The genome of each individual was consisted of 9 chromosomes and each chromosome with 101 SNPs leading to a total number of 909 features. Among these 909 SNPs only 7 were relevant (SNPs number 31, 71, 132, 172, 253, 334 and 405 located on the first five chromosomes). The target concept (the phenotype) was a continuous variable with mean of 36.0 and residual error set to 1 and its values were in the range of 32-42.

Results

In all methods coming in the following subsections, we use some expressions which will be described as follows (all of them were calculated on 100 data sets):
Precision: this criterion is defined as follows:

$$precision = \frac{number\ of\ relevant\ features\ retrieved}{Total\ number\ of\ features\ retrieved}$$

Where *relevant features* are the seven important SNPs. *Retrieved features* are the selected SNPs by a FS method. High precision means that the algorithm returned more relevant SNPs than irrelevant. Its values are in the range of 0-1.

Recall: this criterion is defined as:

$$recall = \frac{number\ of\ relevant\ features\ retrieved}{Total\ number\ of\ relevant\ features}$$

High recall means that the algorithm returned most of the relevant SNPs. Its values are in the range of 0-1.

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

$$F = \frac{2 \times precision \times recall}{precision + recall}$$

Its values are also in the range of 0-1.

Linked results (%): it defines selection percentage of first five vital chromosomes. It is equal or greater than precision. Given that the SNPs on a chromosome have high correlation with each other, this measure is helpful. SNPs' correlation relates to their distances reversely, i.e. when two SNPs are located near each other, their correlation is high and vice versa. So FS methods may select the correlated SNPs that are near the important ones.

Selection rate (%): it defines the selection percentage of each important SNP.

In the following tables the best results stressed in boldface.

Results of filter FS methods:

CFS: CFS 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. Subsets of the features that are highly correlated with the class while having low inter-correlation are preferred. The data analysis was conducted using Weka's implementation of this algorithm (Witten and Frank, 2005).

The results of this method are given in Table 1. Precision of the method is pretty low, i.e. it selects a lot of irrelevant/redundant SNPs in most cases. This fact is confirmed by its recall (0.36). The overall combination of precision and recall is calculated by F-measure that its value is 0.28. However, it is noticeable that the method could identify the important chromosomes in most cases (74.77%).

Important SNPs selection rate using different FS methods are listed in Table 2. It is obvious CFS did not select important SNPs with the same rate (power), e.g. selection rate of SNPs number 172 and 405 are 13% and 61%, respectively. Among seven important SNPs, SNP number 405 has the highest selection rate (61%). This means the last important SNP is the most relevant with the target concept in the CFS's point of view. This method is fast enough however it shows inefficient dimension reduction ability.

ReliefF: The ReliefF algorithm is fairly different to CFS in that it scores individual features rather than feature subsets. 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 class data.

AS shown in Table 1, ReliefF shows an unacceptable precision (0.02%) however it has a great recall (95%). It means this method selects important SNPs and a lot of irrelevant/redundant ones together. Its F-measure is 0.02 that indicates the overall poor performance of the method. Nevertheless, linked results of ReliefF are pretty high (75.89%). Finally, the main drawback of this method is its weak dimension reduction ability. In Table 2, it is apparent that ReliefF identifies important SNPs with the same rate. Among seven important SNPs, SNP number 71 has the highest selection rate (99%).

Decision rule search: 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. Thresholds to determine the amounts of low, medium and high are determined by the user and this makes the flexibility of this method.

Based on Table 1, it is obvious that Decision rule search shows an unacceptable precision (0.02%) just like ReliefF even so it has a reasonable recall (0.81). Therefore, its F-measure is 0.04 that indicates its poor performance. Furthermore, this method has weak dimension reduction ability, and it is so unstable; that is, it suffers from the problem of returning very different numbers of SNPs in each run (190-419). Nevertheless, linked results of Decision rule search are really high (98.80). AS shown in Table 2, Decision rule search identifies important SNPs on chromosomes three and four with high power but not in first two chromosomes. It selects SNP number 405 in all datasets.

Results of the proposed Hybrid methods: In all three hybrid methods, we used a correlation-based feature selection approach as filter model. In this method, feature relevance is measured based on the correlation strength between a feature and the class variable. Feature redundancy is defined based on the correlation strength between a feature and other features. This correlation measure is defined as follows:

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

Where k is the number of features, $\bar{r}_{cf}$ is the mean correlation between each feature and the class variable, 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 SNPs from the original SNP data set. The population size in GA is set to 50 individuals and the number of generations is set to 1000. The crossover and the mutation fractions are 0.8 and 0.2, respectively. The above introduced merit is utilized for calculating the fitness of the selected SNPs by GA. The exact formula which we used as fitness function is as follows:

$$(2)\ f = \exp(-\alpha \times merit)$$

Since the merit values are very small, we use exponential of its values to enlarge the differences between subsets. GA needs to minimize the fitness function, thus we use a negative coefficient because our intension is to maximize the merit criterion. Several trials and errors were performed to find the best value which $\alpha = 3$ is chosen. After setting these parameters for GA, the size of selected SNP subsets are 85 to 105.

Following the above step, the wrapper model is applied to select the important SNPs. We used three different induction models: k-Nearest Neighbor (k-NN), Ridge Regression (RR) and Neural Network (NN). The name of these hybrid methods are considered as Ck-NNFS (Correlation-based k-Nearest Neighbor Feature Selection), CNNFS (Correlation-based Neural Network Feature Selection) and CRRFS (Correlation-based Ridge Regression Feature Selection).

Ck-NNFS: Instance-based learning methods, such as nearest neighbor, are conceptually straightforward approaches for approximating real-valued or discrete-valued target functions. Learning in these algorithms consists of simply storing the presented training data. When a new query instance is encountered, a set of similar related instances is retrieved from memory and used to classify the new query instance. K-NN assumes that all instances correspond to points in the n-

dimensional space. The nearest neighbors of an instance are defined in terms of the standard Euclidean distance (Mitchell, 1997).

For approximating continuous valued target functions, the algorithm calculates the mean value of the k nearest training examples using the following formula:

$$(3)\ f'(x_q) = \frac{\sum_{i=1}^{k} f(x_i)}{k}$$

Where x_q is a new instance, x_i is an existing instance and k is the number of nearest neighbors. To reach the optimal numbers of nearest neighbors that resulted in the best performance, we performed several trial and errors. Similar to the previous step, genetic algorithm was used as global search to find the most relevant subset of SNPs. In this step, the population size in GA was set to 100 individuals and the number of generations was set to 100. The crossover and the mutation fractions were 0.7 and 0.3, respectively. The fitness function of GA in this step was set to the accuracy of k-NN as well as the size of selected SNPs. The precise formula which we used as fitness function was as follows:

$$(4)\ f = \frac{1000 \times L}{(1 + \exp(8 \times R^2))}$$

Where L is the number of selected SNPs, and R^2 is the square of correlation coefficient between the predicted output of k-NN and the actual output. The larger value of R^2 is synonymous with higher accuracy of k-NN. Since the R^2 values are in range of 0 to 1, we used exponential of the values to enlarge the differences between subsets.

The results of Ck-NNFS method are shown in Table 1. The recall of the method is lower than that of three previous filter based methods however its precision is higher than Decision rule search and ReliefF. This happens because Ck-NNFS returns very fewer SNPs than Decision rule search and ReliefF (2-38). Its F-measure is lower than CFS but it has a high power to detect vital chromosomes (90.88%). Based on Table 2, it is evident that Ck-NNFS does not identify important SNPs with the same rate e.g. selection rate of SNPs number 132 and 405 are 0% and 45% respectively. This method has less power to identify important SNPs located on the second chromosome.

CNNFS: Some learning algorithms such as neural network are often trained more successfully and faster when the discrete input features, such as those in our data sets are used. We used a feed forward back propagation neural network to evaluate the SNPs. The selected SNPs of the first step are the input subset for the NN. We conducted trial runs with neural networks containing different numbers of hidden nodes to achieve the optimal number of them. The neural network's accuracy and the size of the SNPs are used as fitness function to guide the GA to select the most important SNPs. The exact formula which we used as fitness function is described in Eq. (4). The parameters of GA in this model are set like previous method.

As shown in Table 1, the precision of CNNFS (0.32) is greater than that of the four previous methods while its recall is lower than filter methods. Just like Ck-NNFS, this happens because CNNFS returns very fewer SNPs than filter methods (4-7); In addition, the method relatively returns right number of important SNPs. Furthermore, it has a great power to identify vital chromosomes (99.82). The overall performance of this method calculated by F-measure is 0.29 (that is the best among all methods). However, the running time of the method is almost two times of the Ck-NNFS. AS shown in Table 2, CNNFS identifies important SNPs with the same power, except SNPs located on the second chromosome. This method does not select SNP number 132 in any cases just like Ck-NNFS.

CRRFS: Ridge Regression (RR) is derived from ordinary Multiple Linear Regression whose goal is to circumvent the problem of predictor's collinearity. It uses the Least Squares (LS) as a method for estimating the parameters of the model. Within other regression-related techniques, ridge regression may be viewed as a tool for exploring and extracting information from multifactor data. The ridge trace can give stability and relative importance of the individual predictors (Price, 1977). Regression coefficients can be estimated using the following formula:

$$\hat{\beta} = (X^T X + kI)^{-1} X^T y \quad (5)$$

Where X is the input matrix, k is the ridge parameter and I is the identity matrix. Small positive values of k improve the conditioning of the problem and reduce the variance of the estimates. For defining the best value of k, some trial and error experiments are done. The fitness function of GA in this step consists of the RR accuracy and the size of the SNPs. The exact formula which we used as fitness function was described in Eq. (4). The parameters of GA in this step are just like other two hybrid methods and they are indicated in 5.2.1.

AS mentioned in Table 1, the precision (0.38) of CRRFS is the highest among all of the methods investigated in this study. Moreover its recall (0.30) is higher than those of the previous hybrid methods. The overall performance of this method according to the value of F-measure is 0.34. In addition, irrelevant chromosomes are not selected in any cases. Furthermore, the method relatively returns right number of important SNPs (4-7) just like CNNFS. Moreover, CRRFS has the highest F-measure as well as it has the least running time within the hybrid methods (35s). Based on Table 2, CRRFS identifies important SNPs with the same power, except SNPs located on the second chromosome, like two previous hybrids. Among seven important SNPs, SNP number 405 has the highest identifying rate (47%). This means the last important SNP is the most relevant with the target concept in the CRRFS's point of view.

Discussion

The results indicated for the SNPs selection problem, the hybrid approaches took more running time and returned better results. Number of selected SNPs and precision of ReliefF and Decision rule search were totally unacceptable in compare with other ones. In addition, they had the lowest linked results however hybrid methods identified vital chromosomes with high rate (more than 90%). Running time of CRRFS was much lower than other hybrids, and it was near to filter approaches especially ReliefF. Hybrid methods achieved higher level of dimensionality reduction by selecting less number of SNPs than pure filters. CNNFS and CRRFS relatively detected the right number of important SNPs.

Based on the results given in Table 1, we can say the best performance belonged to CRRFS method with 4-7 selected SNPs, the highest F_measure, no unlinked results, and short running time (35s). After that, CNNFS got the second grade because it selected 4-7 SNPs, very high linked results (99.82), and the second rank of F_measure. It had one drawback, which was its long running time. Comparing Ck-NNFS and CFS methods from Table 1 showed that Ck-NNFS returned better results in identification of vital chromosomes however its F-measure was lower than CFS. The worst performances belonged to Decision rule search and ReliefF with very low F_measure and weak dimension reduction ability.

Based on Table 2, we can say the SNPs located on the second chromosome are the least correlated SNPs with the target concept because all methods have the least selection rate on them. Nevertheless, SNP number 405 (the single important SNP located on the fifth chromosome) is the most correlated SNP with the target concept, and Decisoin rule search identified it in all cases. One drawback of hybrid methods is their disability to identify SNP number 132.

In order to obtain the statistical support, the Friedman test (Friedman, 1937) between F_measures of FS methods was used. Average ranks obtained by applying the Friedman procedure are given in Table 3. It shows the best performing algorithm is CRRFS. To determine whether the differences among the methods are significant or not, the Holm's post-hoc test (Holm, 1979) had been performed. Results achieved on post hoc comparisons for $\alpha = 0.05$ are shown in Table 4.

Holm's procedure rejects those hypotheses that have a p-value ≤ 0.016667. Therefore, there are no significant differences between algorithms on hypothesis 14 and 15 (i.e. CFS vs. CNNFS, and CFS vs. CRRFS). This conclusion could be resulted by regarding to the ranks of methods on Table 3; there are no significant differences between the rank of ReliefF, Decision rule search, and Ck-NNFS.

Finally, to have a pairwise comparison between the methods, the Wilcoxon test (Wilcoxon, 1945) was applied. Its results are shown in Table 5.

Based on this table, CRRFS is the best performing method. CNNFS and CFS are equivalent, and they outperform Decision rule search, ReliefF and Ck-NNFS. After them, Ck-NNFS gets the third rank. Finally, Decision rule search and ReliefF get the fourth and fifth ranks respectively.

These statistical tests confirmed our previous conclusions about the performance of the methods. However they showed a good quality of the CFS method (second grade) because these tests performed based on the F-measure. It must be considered the number of selected SNPs of CFS was much bigger than those of CNNFS (see Table 1).

Conclusion

Nowadays, Feature selection algorithms play an important role in data mining and knowledge discovery. In this paper, we proposed three hybrid feature selection methods and compared them to three benchmark filter methods over multiple data sets of well-known bioinformatics application named SNPs selection. SNPs provide helpful information on human evolutionary history and lead us to detect genetic variants responsible for human complex diseases. Our implemented hybrid feature selection methods combine the filters and the wrappers to take advantage of both methods. The filter phase removed features with weak relevance and then the wrapper phase applied on them to get the final feature subset. We used Neural Network, k-Nearest Neighbor and Ridge Regression as induction algorithms in wrapper phase. In the hybrid methods, the genetic algorithm was used as a global search.

The results demonstrated that:

- The hybrid methods had better performance than pure filter methods.
- Among three hybrid methods, CNNFS and CRRFS had better dimensionality reduction ability.
- The overall performance of CRRFS algorithm was highly encouraging and it confirmed by several nonparametric statistical tests. It showed the best performance over other five methods and had the least running time within hybrids.

References

Bassat M (1982). Pattern recognition and reduction of dimensionality. In: *Handbook of Statistics II*, Krishnaiah and Kanal ed., North-Holland, Amsterdam. PP. 773–791.

Carlson C, Eberle M, Kruglyak L, Nickerson D (2004). Mapping complex disease loci in whole genome association studies. *Nature*. 429: 446–452.

Friedman M (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*. 32: 675–701.

Hall M (1999). *Correlation-based feature selection for machine learning*, PhD Thesis, Department of Computer Science, University of Waikato, Hamilton, New Zealand.

Holm S (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*. 6: 65–70.

Hsu C, Huang H, Schuschel D (2002). The ANNIGMA-wrapper approach to fast feature selection for neural nets, In: *IEEE Transactions on Systems Man, and Cybernetics—Part B: Cybernetics*, Vol. 32, PP. 207–212.

Jiang B, Ding X, Ma L (2008). A hybrid feature selection algorithm: combination of symmetrical uncertainty and genetic algorithms. In: *2nd International Symposium Optimization and Systems Biology*, Lijiang, China. PP. 152-157.

Kabir M, Islam M, Murase K (2010). A new wrapper feature selection approach using neural network. *Neurocomputing*. 73: 3273-3283.

Kira K, Rendell L (1992). A Practical Approach to Feature Selection, In: *9th International Workshop on Machine Learning*, Aberdeen, Scotland, UK. PP. 249-256.

Kohavi R, John G (1997). Wrappers for Feature Subset Selection. *Artificial Intelligence*. 97: 273-324.

Koller D, Sahami, M (1996). Toward optimal feature selection, In: *13th International Conference on Machine Learning*, Bari, Italy. PP. 284–292.

Kononenko I (1994). Estimating Attributes: Analysis and Extensions of RELIEF, In: *European Conference on Machine Learning*, Catania, Italy. PP. 171-182.

Long N, Gianola D, Rosa G, Weigel K, Avendan S (2007). Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. *Journal of Animal Breeding and Genetics*. 124: 377–389.

Lutu P, Engelbrecht A (2010). A decision rule-based method for feature selection in predictive data mining. *Expert Systems with Applications*. 37: 602-609.

Mahdevar Gh, Zahiri J, Sadeghi M, Nowzari A, Ahrabian H (2010). Tag SNP selection via a genetic algorithm. *Journal of Biomedical Informatics.* 43: 800-804.

Mitchell TM (1997). *Machine Learning*. McGraw-Hill, New York.

Price B (1977). Ridge Regression: application to non-experimental data. *Psychological Bulletin*. 84: 759-766.

Saeys Y, Inza I, Larranaga P (2007). A review of feature selection techniques in bioinformatics. *Gene expression*. 23: 2507–2517.

Shah C, Kusiak A (2004). Data mining and genetic algorithm based gene/SNP selection. *Artificial Intelligence in Medicine*. 31: 183-196.

Sikonja M, Kononenko I (1997). An adaptation of Relief for attribute estimation in regression, In: *14th International Conference on Machine Learning*, Nashville, Tennessee, USA. PP. 296-304.

Sivagaminathan R, Ramakrishnan S (2007). A hybrid approach for feature subset selection using neural networks and ant colony optimization. *Expert Systems with Applications*. 33: 49–60.

Wilcoxon F (1945). Individual comparisons by ranking methods. *Biometrics*. 1: 80–83.

Witten I, Frank E (2005). *Data Mining: A Practical Machine Learning Tool with Java Implementation*, Morgan Kaufmann Publishers, San Francisco, California.

Tables

Table 1
The results of different FS methods on SNPs datasets

FS method	No. of selected SNPs[1]	Precision	Recall	F-measure	Linked results (%)	Average time (s)[2]
CFS	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
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

[1] The number of selected SNPs by each method; [2] Average running times of algorithms in seconds. The bold values are the best ones.

Table 2
Important SNPs selection rate using different methods

Important SNPs	Selection rate (%)					
	CFS	ReliefF	Decision rule search	Ck-NNFS	CNNFS	CRRFS
31	36	98	68	27	30	32
71	43	99	67	28	33	38
132	19	92	65	0	0	0
172	13	87	71	14	15	23
253	37	96	99	32	34	36
334	44	97	99	23	32	37
405	61	97	100	45	41	47

Table 3
Average rankings of the algorithms

Algorithm	Ranking
CFS	2.425
ReliefF	5.41
Decision rule search	4.73
Ck-NNFS	3.53
CNNFS	2.665
CRRFS	**2.24**

Friedman statistic considering reduction performance (distributed according to chi-square with 5 degrees of freedom: 245.7814. P-value computed by Friedman Test: 1.4102e-10.)

Table 4
P-values Table for α = 0.05

i	hypotheses	P	Holm
1	ReliefF vs. CRRFS	0	0.003333
2	CFS vs. ReliefF	0	0.003571
3	ReliefF vs. CNNFS	0	0.003846
4	DRS[1] vs. CRRFS	0	0.004167
5	CFS vs. DRS	0	0.004545
6	DRS vs. CNNFS	0	0.005
7	ReliefF vs. Ck-NNFS	0	0.005556
8	Ck-NNFS vs. CRRFS	0.000001	0.00625
9	DRS vs. Ck-NNFS	0.000006	0.007143
10	CFS vs. Ck-NNFS	0.00003	0.008333
11	Ck-NNFS vs. CNNFS	0.001078	0.01
12	ReliefF vs. DRS	0.010165	0.0125
13	CNNFS vs. CRRFS	0.108197	0.016667
14	CFS vs. CNNFS	0.364346	0.025
15	CFS vs. CRRFS	0.484406	0.05

[1] **Decision Rule Search**

Table 5
Summary of the Wilcoxon test.

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

• = 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 α = 0.9, Lower diagonal level of significance α = 0.95.

Figures and Figures Legend

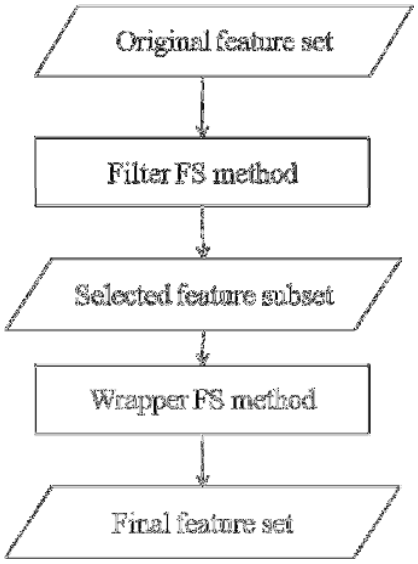

Fig 1. The hybrid feature selection procedure