Journal of Computing and Security

July 2014, Volume 1, Number 3 (pp. 205–213)

http://www.jcomsec.org

Comparison of Hybrid and Filter Feature Selection Methods to Identify Candidate Single Nucleotide Polymorphisms

Farideh Halakou [a,*]
Mahdi Eftekhari [b]
Ali K. Esmailizadeh [c]

[a] *Department of Information Technology, Kerman Graduate University of Technology, Kerman, Iran.*
[b] *Department of Computer Engineering, Shahid Bahonar University of Kerman, Kerman, Iran.*
[c] *Department of Animal Science, Faculty of Agriculture, Shahid Bahonar University of Kerman, Kerman, Iran.*

ARTICLE INFO.

Article history:
Received: 30 March 2013
Revised: 28 June 2014
Accepted: 1 October 2014
Published Online: 21 February 2015

Keywords:
Feature Selection, Single Nucleotide Polymorphisms, Neural Network, K-Nearest Neighbor, Ridge Regression.

ABSTRACT

During the last decade, applying feature selection methods in bioinformatics has become an essential necessity 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 selection methods named CNNFS, Ck-NNFS, and CRRFS, which are combinations of filter and wrapper techniques. In our methods, filter techniques were applied to remove the irrelevant/redundant features as the first step. Then in the second step, wrapper techniques were exploited to refine the primary feature subset obtained from the first step. Neural Network, k-Nearest Neighbor, and Ridge Regression were injected in the wrapper phase as induction algorithms. Since pure wrapper methods take a long time to run on high dimensional data, we compared our methods with three well-known filter methods, and skipped the wrappers. The results vividly show the performance of hybrid methods in addition to their dimensionality reduction ability in SNPs selection. The CRRFS algorithm brought the most satisfactory results regarding to the precision of recognizing candidate SNPs, and the recall of them in the final SNPs subset.

© 2014 JComSec. All rights reserved.

1 Introduction

Feature Selection (FS) aims to decrease the dimensionality of large scale data sets 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; filter model and wrapper model [1]. The filter model relies on general characteristics of a training data set to select relevant features without involving any learning algorithms while the wrapper model [2–4] requires a predetermined learning model and selects features with the aim of improving the generalization performance of that particular learning model. By taking prediction accuracy into consideration, the wrapper methods can reach better results than the other methods. However, the wrapper methods are less popular and need more computational resources because they use specific learning

* Corresponding author.
Email addresses: fhalakou13@ku.edu.tr (F. Halakou), m.eftekhari@mail.uk.ac.ir (M. Eftekhari), aliesmaili@mail.uk.ac.ir (A. Esmailizadeh)
ISSN: 2322-4460 © 2014 JComSec. All rights reserved.

algorithms.

Filter approaches mainly identify a feature subset from the original feature set by applying certain evaluation criteria, which are independent of learning algorithms. Due to the computational efficiency of filter methods, they are very helpful for high-dimensional data. So far, a lot of filter algorithms, such as Correlation-based Feature Selection (CFS) [5], Markov Blanket Filter (MBF) [6] and Information Gain [7] have been developed.

Hybrid approaches, combining the filters and wrappers take advantage of both methods [8, 9]. Although they are not as fast as pure filters, they can achieve better results. Hybrid methods have less computational cost and 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 features from the selected feature set. 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, CNNFS, Ck-NNFS, and CRRFS, which differ in their wrapper phases. Correlation measure is used as filtering criterion in the proposed methods. Afterwards Neural Network, k-Nearest Neighbor and Ridge Regression are used as induction algorithms in the wrapper phase. In these approaches, in both steps we used Genetic Algorithm (GA) to search the feature set. These mechanisms are applied to solve a recently-emerged bioinformatics problem, named Single Nucleotide Polymorphisms (SNPs) selection. SNPs are primarily responsible for the variation among humans. Their importance revolves around the fact that they significantly increase our ability to understand and treat diseases [10]. We will discuss them in detail in Section 2.1.

2 Materials and Methods

Since hybrid methods take advantage of both the efficiency of filters and accuracy of wrappers, we implemented three hybrid FS methods to find the optimal subsets of SNPs. Figure 1 shows the proposed hybrid feature selection procedure. In all of these hybrid methods, correlation-based FS method was chosen as filter model to remove the most redundant/irrelevant SNPs. Then, a wrapper method is applied to improve the accuracy of the results. We used three different induction algorithms in the wrapper phase namely k-Nearest Neighbor (k-NN), Neural Network (NN), and Ridge Regression (RR). To evaluate the performance of proposed methods, we compare them with

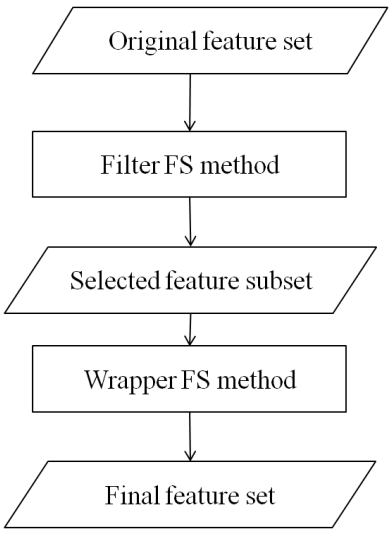

Figure 1. The hybrid feature selection procedure

three pure filters i.e. CFS (Correlation-based Feature Selection) [5], Decision rule search [11], and ReliefF [12–14]. It is important to note that as we were dealing with high dimensional data sets, it would have taken long hours to get the results through the wrapper FS methods so we just skipped them in our evaluation.

2.1 Applied data sets

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 [15]. 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 [16]. SNPs most commonly refer to single-base differences in DNA sequences 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 [17]. 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, selecting a small subset of SNPs that efficiently represent other SNPs in a given block is an important issue for reducing genotyping

costs without losing the ability to detect disease associations. This process is known as Tag SNP selection [18].

Since SNPs selection is a challenging problem in bioinformatics, we evaluated the feature selection algorithms on a set of SNPs data. Since in real data the relevant SNPs are unknown, it would be difficult to precisely compare FS approaches. Therefore, in this research we used simulated data sets. We produced 100 populations in which 500 individuals exist. The genome of each individual was consisted of 9 chromosomes and each chromosome was consisted of 101 SNPs leading to a total number of 909 SNPs. Among these 909 SNPs only 7 of them were relevant i.e. SNPs number 31, 71, 132, 172, 253, 334 and 405 which were located on the first five chromosomes. The target variable (the phenotype) was a continuous quantity with the mean of 36.0, residual error of 1, and values in range of 32-42.

2.2 Evaluation Metrics

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

$$precision = \frac{number\ of\ relevant\ features\ retrieved}{total\ number\ of\ features\ retrieved} \quad (1)$$

where relevant features are the seven important SNPs. Retrieved features are the selected SNPs by a FS method. High precision shows that the algorithm has returned more relevant SNPs than irrelevant. Its values are in the range of 0-1.
Recall: this criterion is defined as:

$$recall = \frac{number\ of\ relevant\ features\ retrieved}{total\ number\ of\ relevant\ features} \quad (2)$$

High recall means that the algorithm has 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:

$$F = \frac{2 * precision * recall}{precision + recall} \quad (3)$$

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

Linked results (%): it defines the selection percentage of first five vital chromosomes. It is equal or greater than precision. Given that the SNPs on the same chromosome have high correlation with each other, this measure is helpful. SNPs correlation has an inverse relationship with their distances i.e. when two SNPs are located near each other, their correlation is high and vice versa.

Selection rate (%): it defines the selection percentage of each important SNP i.e. how many times the vital SNPs were selected by each FS method.
In the following tables the best results stressed in boldface.

3 Results

3.1 Results of filter FS methods

3.1.1 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 among 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 [1].

The results of this method are given in Table 1. As can be clearly seen the precision is pretty low, i.e. it selects a lot of irrelevant/redundant SNPs in most cases. This is also confirmed by its low recall (0.36). However, this method could identify the important chromosomes with a promising rate (74.77%). The selection rate of each candidate SNP through different FS methods is listed in Table 2. It is obvious that CFS did not select important SNPs with the same rate, e.g. selection rate of SNPs number 172 and 405 are 13% and 61%, respectively. Among seven candidate SNPs, SNP number 405 has the highest selection rate (61%). This means that the last important SNP is the most relevant to the target variable in the CFSs̀ point of view. Although this method runs fast it shows inefficient dimension reduction ability.

3.1.2 ReliefF

The ReliefF algorithm is fairly different from 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 data.

As shown in Table 1, ReliefF represents an unacceptable precision (0.02) however it has a high recall (0.95). It means this method selects candidate SNPs and a lot of irrelevant/redundant ones together. Its F-measure is 0.03 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.

Table 1. The results of different FS methods on SNPs data sets

FS method	No. of Selected SNPs [1]	Precision	Recall	F-measure	Linked Results (%)	Average Time [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

Table 2. The selection rate of seven candidate SNPs by different FS methods

Candidate 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

Based on Table 2, ReliefF identified important SNPs with the same rate. Among seven candidate SNPs, SNP number 71 had the highest selection rate (99%).

3.1.3 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 values of low, medium and high are determined by the user which brings flexibility to this method. Based on Table 1, 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 very unstable; i.e. the number of selected SNPs in each run were oscillated significantly (190-419). Nevertheless, linked results of Decision rule search are really high (98.80%). As shown in Table 2, Decision rule search had an extreme power to identify important SNPs on chromosomes number three and four but lack it in cases of the first two chromosomes. It selects SNP number 405 in all data sets.

[1] The number of selected SNPs by a FS method

[2] Average running times of algorithms in seconds. The bold values are the best ones in each column.

3.2 Results of the proposed Hybrid methods

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

$$m = \frac{k * \bar{r}_{cf}}{\sqrt{k + k\,(k-1) * \bar{r}_{ff}}} \tag{4}$$

where k is the number of features, $\bar{r}_{cf}$ is the mean correlation between each feature and the target variable, and $\bar{r}_{ff}$ is the mean correlation between features. In

this step, genetic algorithm was used as global search method to find a subset of relevant SNPs from the original SNPs data set. The population size in GA was set to 50 individuals during 1000 generations. The crossover and the mutation fractions were 0.8 and 0.2, respectively. The above introduced criterion was used to calculate the fitness of the selected SNPs by GA. The exact formula which we used as fitness function is as follows:

$$f = e^{-\alpha m} \tag{5}$$

Since the values of m were very small, we used exponential of its values to enlarge the differences between subsets. As GA needs to minimize the fitness function, we used a negative coefficient to maximize m during the GA generations. Several trials and errors were performed to find the best coefficient which was $\alpha = 3$. After setting these parameters for GA, the size of selected SNPs subsets were in the range of 85-105.

Following the above step, the wrapper method was applied to select the candidate SNPs. We used three different induction algorithms: k-Nearest Neighbor (k-NN), Ridge Regression (RR) and Neural Network (NN). These hybrid methods are named 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).

3.2.1 Ck-NNFS

Instance-based learning methods, such as nearest neighbor are conceptually straightforward 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 instances is retrieved from memory and used to classify the new query instance. K-NN assumes that all instances correspond to points in an n-dimensional space. The nearest neighbors of an instance are defined in terms of the standard Euclidean distance [19].

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

$$f'(x_q) = \frac{\sum_{i=1}^{k} f(x_i)}{k} \tag{6}$$

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 number of nearest neighbors that results the highest accuracy, we performed several trials and errors. Similar to the previous step, genetic algorithm was used as the global search method 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 mutation fractions were 0.7 and 0.3, respectively. Parameters of the fitness function of GA in this step was included the accuracy of k-NN as well as the number of selected SNPs. The precise formula which we used as fitness function is as follows:

$$f = \frac{1000 * L}{(1 + e^{8R^2})} \tag{7}$$

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 the range of 0 to 1, we used exponential of the values to enlarge the differences between the 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 methods, however, its precision is higher than Decision rule search and ReliefF. This can be derived from the fact that Ck-NNFS returns fewer SNPs than Decision rule search and ReliefF (2-38). However Ck-NNFSs F-measure is lower than CFS, it could detect vital chromosomes in most cases (90.88%). Based on Table 2, it is evident that Ck-NNFS has not identified significant SNPs with the same rate, e.g., selection rate of SNPs number 132 and 405 were 0% and 45% respectively. This method weakly identified candidate SNPs located on the second chromosome.

3.2.2 CNNFS

Some learning algorithms such as neural network are often trained more successfully and faster when 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 were the input of the NN. We conducted trial runs with neural networks containing different numbers of hidden nodes to find the optimal number of them. The neural networks accuracy and the size of the SNPs were used as the fitness function to guide the GA in selecting the candidate SNPs. The exact formula which we used as fitness function is described in Eq. (7). The parameters of GA in this model were set just like the previous method.

As shown in Table 1, the precision of CNNFS (0.32) was greater than that of the four previous methods while its recall was lower than filter methods. Just like Ck-NNFS, this happened because CNNFS returned smaller number of SNPs than filter methods (4-7); In addition, CNNFS returned almost the correct number of candidate SNPs. Furthermore, it had a great capability to identify vital chromosomes (99.82). The overall performance of this method calculated by F-

measure was 0.29 (that is the best among all the methods). However, the running time of this method was almost twice as much as that of Ck-NNFSs. As shown in Table 2, CNNFS detected candidate SNPs with the same accuracy, except SNPs located on the second chromosome. This method did not select SNP number 132 in any cases, just like Ck-NNFS.

3.2.3 CRRFS

Ridge Regression (RR) is derived from ordinary Multiple Linear Regression whose goal is to circumvent the problem of predictors 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 show stability and relative importance of the individual predictors [20]. Regression coefficients can be estimated using the following formula:

$$\hat{\beta} = (X^T X + kI)^{-1} X^T y \qquad (8)$$

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. The best value for k was estimated through trial and error. The fitness function of GA in this step consisted of the RR accuracy and the size of SNPs. The exact formula which we used as fitness function is described in Eq. (7). The parameters of GA in this step were just like other two hybrid methods which are presented in Section 3.2.1.

As mentioned in Table 1, the precision of CRRFS (0.38) 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 were not selected in any cases. Similar to CNNFS, the method returned almost the correct number of the candidate SNPs (4-7). Moreover, CRRFS had the highest F-measure as well as the least running time within the hybrid methods (35s). Based on Table 2, CRRFS detected candidate SNPs with the same accuracy, except SNPs located on the second chromosome, like two previous hybrids. Among seven candidate SNPs, SNP number 405 had the highest identification rate (47%), which means the last important SNP is the most relevant with the target variable in the CRRFSs point of view.

4 Discussion

The results in the previous section vividly demonstrated the power of the hybrid methods in SNPs selection. Among filter methods, the precision and number of selected SNPs of ReliefF and Decision rule search were totally improper. In addition, they had the lowest linked results. Conversely, the hybrid methods identified vital chromosomes with high rate (more than 90%). Besides, they achieved higher level of dimensionality reduction by selecting fewer numbers of SNPs than pure filters. It is important to note that CNNFS and CRRFS detected almost the correct number of candidate SNPs.

Based on the results given in Table 1, we can infer that the best performance belonged to CRRFS method, however the results of the CNNFS was also promising. By looking at Table 2, we can say that the SNPs located on the second chromosome were least correlated with the target variable because all of the methods had the least selection rate for them.

In order to obtain the statistical support, the Friedman test [21] on F-measures of FS methods was used. Average ranks obtained by applying the Friedman procedure are given in Table 3. This test also indicated that the best performing algorithm was CRRFS. To determine whether the differences among the methods are significant or not, the Holms post-hoc test [22] was performed. Results obtained on post hoc comparisons for $alpha = 0.05$ are shown in Table 4. Holms procedure rejects those hypotheses with a $p - value \leq 0.0167$. Therefore, there are no significant differences between algorithms on hypothesis 14 and 15 (i.e. CFS vs. CNNFS, and CFS vs. CRRFS). However, we should consider this test was based on just the F-measures and did not take into account the linked results, and number of selected SNPs.

Table 3. Average rankings of the algorithms.

Algorithm	Ranking [1]
CFS	2.425
ReliefF	5.41
Decision rule search	4.73
Ck-NNFS	3.53
CNNFS	2.665
CRRFS	2.24

[1] 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 $\alpha = 0.05$

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

Finally, to have a pairwise comparison between the methods, the Wilcoxon test [23] 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. These statistical tests confirmed our previous conclusions about the performance of the methods.

Table 5. Summary of the Wilcoxon test.

	(1)	(2)	(3)	(4)	(5)	(6)
CFS (1)	-	● [2]	●	●		○ [3]
ReliefF (2)	○	-	○	○	○	○
Decision Rule Search (3)	○	●	-	○	○	○
Ck-NNFS (4)	○	●	●	-	○	○
CNNFS (5)		●	●	●	-	○
CRRFS (6)	●	●	●	●	●	-

[1] Decision Rule Search
[2] ● means the method in the row improves the method of the column.
[3] ○ means the method in the column improves the method of the row. Upper diagonal of level significance $\alpha = 0.9$, lower diagonal level of significance $\alpha = 0.95$.

5 Conclusion

Nowadays, feature selection algorithms play a significant role in data mining and knowledge discovery. In this paper, we proposed three hybrid feature selection methods and compared them with three benchmark filter methods on multiple data sets of SNPs. SNPs provide helpful information on human evolutionary history and lead us to detect genetic variants responsible for human complex diseases. Our proposed hybrid FS methods combine filter and wrapper algorithms to take advantage of both methods i.e. the running time and accuracy. The filter phase removes the irrelevant/redundant features. Then the wrapper phase is 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 method.

Experimental results vividly demonstrated the performance of hybrid methods in SNPs selection in terms of F-measure and the number of selected SNPs. Among three proposed hybrid methods, CNNFS and CRRFS represented higher level of dimensionality reduction. Furthermore, the overall performance of CRRFS, based on several nonparametric statistical tests, was highly encouraging.

References

[1] Ian H. Witten, Eibe Frank, and Mark A. Hall. Chapter 4 - algorithms: The basic methods. In Ian H. WittenEibe FrankMark A. Hall, editor, *Data Mining: Practical Machine Learning Tools and Techniques (Third Edition)*, The Morgan Kaufmann Series in Data Management Systems, pages 85 – 145. Morgan Kaufmann, Boston, third edition edition, 2011. ISBN 978-0-12-374856-0. doi: http://dx.doi.org/10.1016/B978-0-12-374856-0.00004-3. URL http://www.sciencedirect.com/science/article/pii/B9780123748560000043.

[2] Ron Kohavi and George H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97 (12):273 – 324, 1997. ISSN 0004-3702. doi: http://dx.doi.org/10.1016/S0004-3702(97)00043-X. URL http://www.sciencedirect.com/science/article/pii/S000437029700043X. Relevance.

[3] Chun-Nan Hsu, Hung-Ju Huang, and S. Dietrich. The annigma-wrapper approach to fast feature selection for neural nets. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 32(2):207–212, Apr 2002. ISSN 1083-4419. doi: 10.1109/3477.990877.

[4] Md. Monirul Kabir, Md. Monirul Islam, and Kazuyuki Murase. A new wrapper feature selection approach using neural network. *Neurocomputing*, 73(1618):3273 – 3283, 2010. ISSN 0925-2312. doi: http://dx.doi.org/10.1016/j.neucom.2010.04.003. URL http://www.sciencedirect.com/science/article/pii/S0925231210001979. 10th Brazilian Symposium on Neural Networks (SBRN2008).

[5] Mark A. Hall. Correlation-based feature selection for machine learning. Technical report, 1999.

[6] Daphne Koller and Mehran Sahami. Toward optimal feature selection. pages 284–292. Morgan Kaufmann, 1996.

[7] Moshe Ben-Bassat. 35 use of distance measures, information measures and error bounds in feature evaluation. In P.R. Krishnaiah and L.N. Kanal, editors, *Classification Pattern Recognition and Reduction of Dimensionality*, volume 2 of *Handbook of Statistics*, pages 773 – 791. Elsevier, 1982. doi: http://dx.doi.org/10.1016/S0169-7161(82)02038-0. URL http://www.sciencedirect.com/science/article/pii/S0169716182020380.

[8] Rahul Karthik Sivagaminathan and Sreeram Ramakrishnan. A hybrid approach for feature subset selection using neural networks and ant colony optimization. *Expert Systems with Applications*, 33(1):49 – 60, 2007. ISSN 0957-4174. doi: http://dx.doi.org/10.1016/j.eswa.2006.04.010. URL http://www.sciencedirect.com/science/article/pii/S0957417406001187.

[9] Ma L Jiang B, Ding X. A hybrid feature selection algorithm: combination of symmetrical uncertainty and genetic algorithms. In *2nd International Symposium Optimization and Systems Biology*.

[10] Shital C. Shah and Andrew Kusiak. Data mining and genetic algorithm based gene/snp selection. *Artificial Intelligence in Medicine*, 31(3):183 – 196, 2004. ISSN 0933-3657. doi: http://dx.doi.org/10.1016/j.artmed.2004.04.002. URL http://www.sciencedirect.com/science/article/pii/S0933365704000521.

[11] Patricia E.N. Lutu and Andries P. Engelbrecht. A decision rule-based method for feature selection in predictive data mining. *Expert Systems with Applications*, 37(1):602 – 609, 2010. ISSN 0957-4174. doi: http://dx.doi.org/10.1016/j.eswa.2009.06.031. URL http://www.sciencedirect.com/science/article/pii/S0957417409005831.

[12] Kenji Kira and Larry A. Rendell. A practical approach to feature selection. In *Proceedings of the Ninth International Workshop on Machine Learning*, ML92, pages 249–256, San Francisco, CA, USA, 1992. Morgan Kaufmann Publishers Inc. ISBN 1-5586-247-X. URL http://dl.acm.org/citation.cfm?id=141975.142034.

[13] Igor Kononenko. Estimating attributes: Analysis and extensions of relief. In Francesco Bergadano and Luc De Raedt, editors, *Machine Learning: ECML-94*, volume 784 of *Lecture Notes in Computer Science*, pages 171–182. Springer Berlin Heidelberg, 1994. ISBN 978-3-540-57868-0. doi: 10.1007/3-540-57868-4_57. URL http://dx.doi.org/10.1007/3-540-57868-4_57.

[14] Marko Robnik-Sikonja and Igor Kononenko. An adaptation of relief for attribute estimation in regression. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, pages 296–304, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc. ISBN 1-55860-486-3. URL http://dl.acm.org/citation.cfm?id=645526.657141.

[15] Christopher S. Carlson, Michael A. Eberle, Leonid Kruglyak, and Deborah A. Nickerson. Mapping complex disease loci in whole-genome association studies. *Nature*, 429(6990):446–452, May 2004. ISSN 0028-0836. doi: 10.1038/nature02623. URL http://dx.doi.org/10.1038/nature02623.

[16] Yvan Saeys, Iaki Inza, and Pedro Larraaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007. doi: 10.1093/bioinformatics/btm344. URL http://bioinformatics.oxfordjournals.org/content/23/19/2507.abstract.

[17] N. Long, D. Gianola, G.J.M. Rosa, K.A. Weigel, and S. Avendao. Machine learning classification procedure for selecting snps in genomic selection: application to early mortality in broilers. *Journal of Animal Breeding and Genetics*, 124(6):377–389, 2007. ISSN 1439-0388. doi: 10.1111/j.1439-0388.2007.00694.x. URL http://dx.doi.org/10.1111/j.1439-0388.2007.00694.x.

[18] Ghasem Mahdevar, Javad Zahiri, Mehdi Sadeghi, Abbas Nowzari-Dalini, and Hayedeh Ahrabian. Tag {SNP} selection via a genetic algorithm. *Journal of Biomedical Informatics*, 43(5):800 – 804, 2010. ISSN 1532-0464. doi: http://dx.doi.org/10.1016/j.jbi.2010.05.011. URL http://www.sciencedirect.com/science/article/pii/S1532046410000699.

[19] Tom Mitchell. *Machine Learning*.

[20] Bertram Price. Ridge regression: Application to nonexperimental data.

[21] Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701, 1937. doi: 10.1080/01621459.1937.10503522.

URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1937.10503522.

[22] Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. ISSN 03036898. doi: 10.2307/4615733. URL http://dx.doi.org/10.2307/4615733.

[23] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):pp. 80–83, 1945. ISSN 00994987. URL http://www.jstor.org/stable/3001968.

Farideh Halakou received her B.Sc. degree in computer science from Mirdamad University, Gorgan, Iran in 2008, and M.Sc. degree from Kerman Graduate University of Advanced Technology, Kerman, Iran. At present she is a PhD student at Koc University, Istanbul, Turkey. Her research interests involve data mining, feature selection, modeling protein-protein interaction networks, topological and structural analysis of PPI networks.

Mahdi Eftekhari received his B.Sc. in computer engineering from Department of Computer Science and Engineering, Shiraz University, Shiraz, Iran in September 2001. He obtained his M.Sc. and Ph.D. degrees in Artificial Intelligence from the same department in 2004 and 2008 respectively. He has been a faculty member of Computer Engineering Department at Shahid Bahonar University of Kerman, Kerman, Iran since 2008. His research interests include Fuzzy systems and modeling, Evolutionary Algorithms, Data Mining, Machine Learning and Application of intelligent methods in bioinformatics. He is the author and co-author of about 70 papers in cited journals and conferences. Dr. Eftekhari is a member of Iranian Society of Fuzzy Systems.

Ali K. Esmailizadeh did his B.Sc. program in animal production in University of Sistan & Baluchestan, Zahedan, Iran in 1996; M.Sc. in animal science-animal breeding & genetics in University of Tarbiat Modares, Tehran, Iran in 1998, and his PhD in animal science-gene mapping in the University of Adelaide, Australia in 2006. The current position of Ali is associate professor of Shahid Bahonar University of Kerman, teaching animal genetics, advanced statistics in animal science, molecular genetics and genetic engineering, design and analysis of animal science experiments. His research interests include mapping loci underlying complex traits or quantitative trait loci (QTL) in animals and experimental species, applications of genomic technologies to improve animal performance, genome-wide association studies, genomic imprinting and gene mapping.