This article can be cited as S. Barchinezhad and M. Eftekhari, A New Fuzzy and Correlation Based Feature Selection Method for Multiclass Problems, International Journal of Artificial Intelligence, vol. 12, no. 2, pp. 24-41, 2014.

A New Fuzzy and Correlation Based Feature Selection Method for Multiclass Problems

Soheila Barchinezhad[1], **Mahdi Eftekhari**[2]

[1]Department of Electronic and Computer, Kerman Graduate University of Advanced Technology,
Haft Bagh Blvd, Mahan 7631133131, Kerman, Iran;
Email: s.barchinezhad@kgut.ac.ir

[2]Department of Computer Engineering, Shahid Bahonar University of Kerman,
22 Bahman Blvd, Kerman 7616914111, Iran.
Email: m.eftekhari@mail.uk.ac.ir

ABSTRACT

Feature selection is one of the most important subjects in machine learning and pattern recognition. The main idea in feature selection algorithms is selecting a subset of features which does not include irrelevant and redundant features. In this paper a feature selection algorithm using genetic algorithm and fuzzy sets theory is proposed which we call fuzzy and correlation based feature selection - FCFS. We apply four fuzzy systems to obtain the fitness function in genetic algorithm. This filter method selects a low size feature subset so that the relevancy of each feature with the target is maximized and the redundancy among the selected features is minimized. Relevancy and redundancy are calculated based on Pearson's correlation coefficient criterion. Several experiments are provided to demonstrate the effectiveness of the idea in terms of the classification accuracy and the number of selected features. Some statistical tests are also used to show the significant differences between the proposed method and the other methods.

Keywords: Feature Selection, Fuzzy, Genetic Algorithm, Subset Selection, Filter.

Mathematics Subject Classification: 68T99, 03B52, 68T27, 94D05

1. INTRODUCTION

Feature selection is one of the most important subjects in machine learning, pattern recognition and data mining. In many fields, there are thousands of features which have to be measured and not all of them are relevant to the problem, even some are redundant. In addition, dealing with a large number of features is costly; therefore elimination of irrelevant and redundant features is rather important. What seems to be important is to identify a set of features that are most correlated with the target and and not correlated with the other features. There have been many approaches to feature selection based on a variety of techniques, such as statistical (Zhou and Dillion, 1991), geometrical (Elomaa and Ukkonen, 1994), mathematical programming (Bradley et al., 1998), neural network (Kabir and Islam, 2010), entropy (Luukka, 2011), neurofuzzy (Basak et al., 1998), Genetic Algorithm (GA) (Tsai et al., 2013) and discretization (Liu and Setiono, 1997). Each of these has its own advantages and disadvantages which make them context-specific. Also, there is no universally the best feature selection method.

This article can be cited as S. Barchinezhad and M. Eftekhari, A New Fuzzy and Correlation Based Feature Selection Method for Multiclass Problems, International Journal of Artificial Intelligence, vol. 12, no. 2, pp. 24-41, 2014.

According to (Kabir and Islam, 2010; Luukka, 2011; Lin, 2012; Zhao et al., 2010; Sun et al., 2012; Unler et al., 2011; Sikora and Piramuthu, 2007; Cai et al., 2009; Chen et al., 2011; Covoes and Hruschka, 2011) feature selection techniques, based on their evaluation approach can be classified into the following categories: embedded, filter and wrapper techniques. Feature selection approach is named filter if it is independent from learning algorithm and it is called wrapper if it is related to learning algorithm. By using filter techniques, irrelevant and/or redundant features are filtered before using learning algorithm. In fact these techniques do not use learning algorithms feedback and the features subsets are evaluated according to the other criteria. In this technique the evaluation criteria can be distance based, information based, dependency based or consistency based. In contrast to the filter methods, the wrappers utilize the learning algorithm as a black box to score the subsets of variables according to their predictive power. Therefore they need more time and computations and their results are consequently more accurate compared to the filter approaches. In embedded approach, feature selection is a part of the modelling procedure. The main aim of this approach is to search for the “best” subset of features. It is of great importance when it regards a large number of features but a small sample size. Generally, though the wrapper and embedded methods often outperform the filter methods in terms of accuracy, the filter methods are more usually adopted for feature selection (Peng et al., 2005) due to their simplicity, flexibility and low computational time. According to the analysis above, the filter methods for feature selection which are accurate in multiclass problems are well studies.

In this paper we propose a multiclass filter method for the feature selection using GA and fuzzy sets theory. We also implement it, and evaluate its performance using some benchmark datasets. We choose GA due to its simplicity and its capability as a powerful search mechanism. GA is used for optimizing the fitness function calculated by four fuzzy systems. The inputs of these fuzzy systems are the number of selected features, the feature-target correlation (relevancy) and the feature-feature correlation (redundancy). The results of the experiments reveal that the proposed method is better than the other common methods in terms of classification accuracy and the number of selected features.

The remainder of this paper is organized as follows. Section 2 describes the related works for solving the feature selection problem in recent decades. A preliminary of proposed feature selection algorithm is discussed in section 3, where the proposed methodology is presented. In section 4, the experiments and the results are presented. Some conclusions are drawn in section 5.

2. RELATED WORKS

Feature selection is an optimization problem, therefore in each of the above mentioned approaches various optimization methods such as GA (Tsai et al., 2013), Particle Swarm Optimization (PSO) (Wang et al., 2007), Gravitational Search Algorithm (Purcaru et al., 2013) and Ant Colony Optimization (ACO) (Ahmed, 2006) can be used. In Ref. (Kudo and Sklansky, 2000) a comparative study of optimization

This article can be cited as S. Barchinezhad and M. Eftekhari, A New Fuzzy and Correlation Based Feature Selection Method for Multiclass Problems, International Journal of Artificial Intelligence, vol. 12, no. 2, pp. 24-41, 2014.

algorithms for feature selection, the results of many experiments show that GAs are more suitable than the other heuristic search methods for large and medium sized problems. It can be used in filter and wrapper methods. Ref. (Sikora and Piramuthu, 2007) proposed a genetic based filter method and in Ref. (Huang et al., 2007) we can find a genetic based wrapper method.

In solving feature selection problem, filter approaches are faster to implement, since they estimate the performance of the features without any learning model adapted between the targets and inputs of the data. Selection or removal of a feature is decided on through using some predefined criteria, such as, mutual information, correlation coefficient and feature weighing. Ref. (Peng et al., 2005) presents a theoretical analysis of the minimal-redundancy-maximal-relevance (MRMR) which uses mutual information criterion. Mutual information of two random variables is a quantity that measures the mutual dependence of the two variables. In Ref. (Hall, 1999) a criterion based on redundancy, relevancy and the number of selected features is proposed which uses Pearson's correlation coefficient. This method is based on the hypothesis that a good feature subset should contain the features which are highly correlated with the class, yet uncorrelated with each other. One of the disadvantages of this method is related to its correlation criterion. This criterion is a linear measure and cannot exactly determine the correlation between two variables. In Ref. (Wang et al., 2012), a novel filter framework is presented to select optimal feature subset based on a maximum weight and minimum redundancy (MWMR) criterion. The weight of each feature indicates its importance for some ad hoc tasks (such as clustering and classification) and the redundancy represents the correlations among features. In MWMR, three feature weighting algorithms (Laplacian score, Fisher score and Constraint score) are combined with two redundancy measure (Pearson's correlation coefficient and mutual information).

In some filter algorithms, feature ranking is often used to show which input features are more important, especially when datasets are very large. Feature ranking does not need to create a learning algorithm (Guyon and Elisseeff, 2003). Some of popular feature rankings are Relief (Kira and Rendell, 1992) and its multiclass extension Relief-F (Robnik-Šikonja and Kononenko, 2003), Fisher score (Duda and Hart, 2001), Chi-square score (Liu and Setiono, 1995), Kruskal wallis test (Wei, 1981), Gini index and Information Gain (Yang and Pedersen, 1997). Subset selection is another type of filter methods. MRMR (Peng et al., 2005), Correlation based Feature Selection -CFS (Hall and Smith, 1997), Fast Correlation Based Filter -FCBF (Yu and Liu, 2003), Bayesian logistic regression routine -Blogreg (Cawley and Talbot, 2006) and Sparse multinomial logistic regression algorithm with Bayesian regularization -Sbmlr (Cawley, et al., 2007) are subset selection filter methods. Although current algorithms are widely used for solving the feature selection problem, they also suffer from the following limitations. Firstly some of them evaluate features individually and do not consider feature relevancy. Secondly, most of them are incapable of solving multiclass problems and some are also feature rankings in which the user has to determine the number of the selected features. In this paper we try to solve these limitations.

3. PROPOSED METHOD: FUZZY AND CORRELATION BASED FEATURE SELECTION METHOD (FCFS)

GA (Popov, 2005) which is an evolutionary algorithm (Khan, 2014), is a rapid search for large, nonlinear and poorly understood spaces. A single given dataset may contain more than one optimal feature subset. For larger problems, there may be many more. Typically, only one of these subsets is chosen to perform the necessary dataset reduction. The problem here is how to discern the reductions in order to choose the best one to decrease the data. Unlike other feature selection strategies where one solution is optimized, in genetic based algorithms a population of solutions can result in several optimal (or close-to-optimal) feature subsets as output. In this paper, a fuzzy multi-criteria function, comprising relevancy, redundancy and number of selected features, is used to assess the performance of each features subset obtained through a GA process. Pearson correlation coefficient criterion is used for calculating the redundancy and relevancy. This criterion is a linear measure and cannot determine all correlations between two variables, therefore it is considered as an ambiguous variable in fitness function. The number of selected features is an ambiguous variable too, since we want to find a subset with a low size and the size is not also clearly determined. Therefore fuzzy sets theory sounds appropriate for this multi-criteria function. The proposed algorithm process is shown in Figure 1 which can be described as follows:

```
1) Normalize data based on uniform distribution
2) Use normalized data and select the best feature subset by GA
    a) Initialize population size, chromosome length, and rate of mutation and crossover operations
    b) Initialize population randomly
    c) Calculate $\overline{rcf}$ , $\overline{rff}$ and $k$
    d) Calculate fitness function of each chromosome by fuzzy systems
    e) Do
        - Generate new population by selection, crossover and mutation operations
        - Calculate $\overline{rcf}$ , $\overline{rff}$ and $k$
        - Calculate the fitness function of each chromosome by fuzzy systems
        - Iteration=Iteration+1
    Until satisfaction of stopping criterion
3) Use normalized data and validate the selected subset by 10fold cross validation
```

3.1. Data pre-processing

Some of the learning algorithm such as GAs are often more successful and faster when normal input features are used. Therefor at first, in this paper the data is normalized based on uniform distribution to prevent the computational problems, dispersal of data and data overflows. Then dataset is used as GA's input.

3.2. Chromosome encoding

To use a genetic algorithm, the solution of problem has to be represented as an individual called chromosome. A genetic algorithm uses a population of individuals. It creates a population of individuals and applies genetic operators such as mutation and crossover to evolve the individuals in order to find the best one(s). In this paper, feature subset is represented by a binary string with length equal to the number of original features in the dataset. A one and zero in the j^{th} position in the chromosome denotes the presence or absence of the j^{th} feature in this subset.

3.3. Initial population

In GA process at first an initial population of chromosomes is created. Here the population is created randomly and its size is the same as the determined population size.

3.4. Fitness evaluation

The fitness of an individual solution is its performance measure. This measure is used to favour the selection of successful parents for the pool of new offspring, so that the whole population of solutions incrementally evolves towards a greater fitness. If a filter approach is adopted, the fitness of individuals is calculated using a suitable criterion function. A larger value of criterion function indicates a better feature subset. Such a criterion function could be entropy measure, correlation measure or a mixture of some criteria. To guide the search for the minimal feature subset, the subset size is also incorporated into the fitness function of both filter and wrapper methods. By using fuzzy sets theory we can present a flexible fuzzy multi-criteria system which can appoint trade-off between some antonym criteria and goals. Fuzzy systems have a wide range of applications. Ref. (Joelianto et al., 2013) uses ANFIS (an adaptive neuro-fuzzy inference system) to improve transient response performance of PID controller and in Ref. (Precup et al., 2009) a new stability analysis method for a class of nonlinear time-varying processes based on stabilizing Takagi Sugeno fuzzy logic controller are presented.

3.5. Fuzzy multi-criteria decision making

The algorithm proposed in this paper deals with the feature selection problem as a multi-criteria problem with a single objective function. Therefore, we use multi-criteria decision making to select the features, according to some goals (relevancy, redundancy and the number of selected features), without neglecting the others. The fuzzy set theory enables the representation of multiple criteria with a flexibility that can be exploited to obtain desired trade-offs in order to satisfy contradictory goals, and due to this advantage fuzzy multi-criteria decision making is used in this work. We use four fuzzy systems that calculate the fitness function of individuals (Figure 2). Consequently, the number of selected features (k) and the average of feature-target correlation ($\overline{rcf}$), the number of selected features (k) and the average of feature-feature correlation ($\overline{rff}$), the average of feature-feature correlation ($\overline{rff}$) and the average of feature-target correlation ($\overline{rcf}$) are the inputs of the first, the second and the third fuzzy systems. The fourth fuzzy system aggregates the outputs of three previous systems and the output of this system is the fitness function of the genetic algorithm. The type of membership function which is applied in all systems is Generalized Bell membership function. It can be determined through some experiments. Suppose that

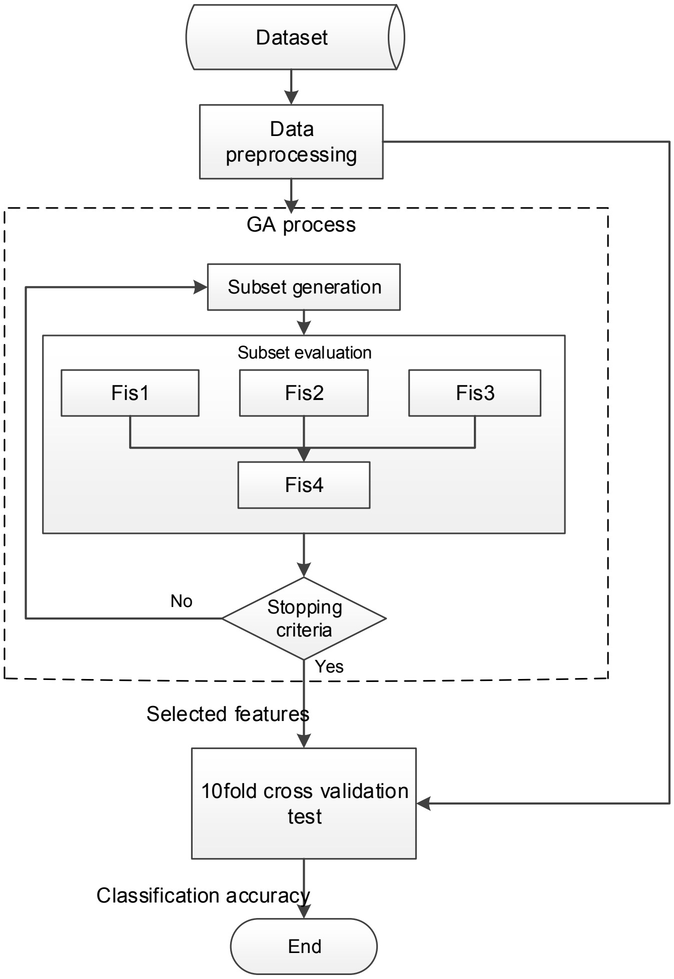

Figure1. The process of fuzzy and correlation based feature selection method (FCFS)

k is the number of the selected features, then we can define the fuzzy sets “low”, “medium” and “high” for the variable k, this is shown in Figure 3. Rules of the fuzzy systems and the weight of each rule are brought in Table 1 – Table 4. The weight of each rule is selected based on the range of k, $\overline{rff}$ and $\overline{rcf}$ which are obtained through some test and trial experiments.

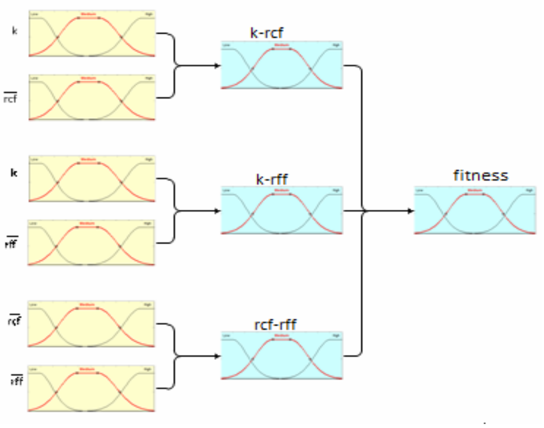

Figure2. The fuzzy systems for evaluating each feature subset

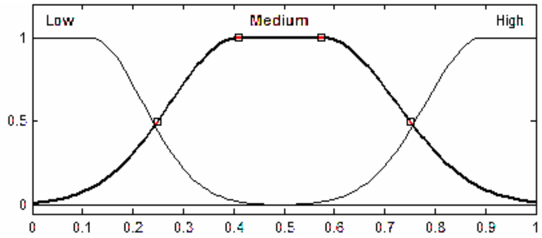

Figure3. Typical membership functions of fuzzy systems

Table 1. Rules of the first fuzzy system

Rule	Weight
1. If (k is Low) and (rcf is High) then (k-rcf is Medium)	1
2. If (k is Low) and (rcf is Medium) then (k-rcf is Medium)	0.7
3. If (k is Low) and (rcf is Low) then (k-rcf is Medium)	0.6
4. If (k is Medium) and (rcf is High) then (k-rcf is High)	1
5. If (k is Medium) and (rcf is Medium) then (k-rcf is High)	1
6. If (k is Medium) and (rcf is Low) then (k-rcf is Low)	1

7. If (k is High) and (rcf is High) then (k-rcf is Medium)	1
8. If (k is High) and (rcf is Medium) then (k-rcf is Low)	1
9. If (k is High) and (rcf is Low) then (k-rcf is Low)	1

Table 2. Rules of the second fuzzy system

Rule	Weight
1. If (k is Low) and (rff is Low) then (k-rff is Medium)	1
2. If (k is Low) and (rff is Medium) then (k-rff is Medium)	0.7
3. If (k is Low) and (rff is High) then (k-rff is Medium)	0.6
4. If (k is Medium) and (rff is Low) then (k-rff is High)	1
5. If (k is Medium) and (rff is Medium) then (k-rff is High)	1
6. If (k is Medium) and (rff is High) then (k-rff is Low)	1
7. If (k is High) and (rff is Low) then (k-rff is Medium)	1
8. If (k is High) and (rff is Medium) then (k-rff is Low)	1
9. If (k is High) and (rff is High) then (k-rff is Low)	1

Table 3. Rules of the third fuzzy system

Rule	Weight
1. If (rcf is Low) and (rff is High) then (rcf-rff is Low)	1
2. If (rcf is Low) and (rff is Medium) then (rcf-rff is Medium)	0.6
3. If (rcf is Low) and (rff is Low) then (rcf-rff is Medium)	0.7
4. If (rcf is Medium) and (rff is High) then (rcf-rff is High)	1
5. If (rcf is Medium) and (rff is Medium) then (rcf-rff is High)	1
6. If (rcf is Medium) and (rff is Low) then (rcf-rff is Low)	1
7. If (rcf is High) and (rff is High) then (rcf-rff is Medium)	1
8. If (rcf is High) and (rff is Medium) then (rcf-rff is Low)	1
9. If (rcf is High) and (rff is Low) then (rcf-rff is Low)	1

Table 4. Rules of the last fuzzy system

Rule	Weight
1. If (k-rcf is High) and (k-rff is High) and (rcf-rff is High) then (fitness is High)	1
2. If (k-rcf is High) and (k-rff is High) and (rcf-rff is Medium) then (fitness is High)	1
3. If (k-rcf is High) and (k-rff is High) and (rcf-rff is Low) then (fitness is Medium)	1
4. If (k-rcf is High) and (k-rff is Medium) and (rcf-rff is High) then (fitness is High)	1

5. If (k-rcf is High) and (k-rff is Medium) and (rcf-rff is Medium) then (fitness is Medium)	1
6. If (k-rcf is High) and (k-rff is Medium) and (rcf-rff is Low) then (fitness is Medium)	0.8
7. If (k-rcf is High) and (k-rff is Low) and (rcf-rff is High) then (fitness is Medium)	1
8. If (k-rcf is High) and (k-rff is Low) and (rcf-rff is Medium) then (fitness is Medium)	0.8
9. If (k-rcf is High) and (k-rff is Low) and (rcf-rff is Low) then (fitness is Low)	1
10. If (k-rcf is Medium) and (k-rff is High) and (rcf-rff is High) then (fitness is High)	1
11. If (k-rcf is Medium) and (k-rff is High) and (rcf-rff is Medium) then (fitness is Medium)	1
12. If (k-rcf is Medium) and (k-rff is High) and (rcf-rff is Low) then (fitness is Medium)	0.8
13. If (k-rcf is Medium) and (k-rff is Medium) and (rcf-rff is High) then (fitness is Medium)	1
14. If (k-rcf is Medium) and (k-rff is Medium) and (rcf-rff is Medium) then (fitness is Medium)	1
15. If (k-rcf is Medium) and (k-rff is Medium) and (rcf-rff is Low) then (fitness is Low)	1
16. If (k-rcf is Medium) and (k-rff is Low) and (rcf-rff is High) then (fitness is Medium)	0.8
17. If (k-rcf is Medium) and (k-rff is Low) and (rcf-rff is Medium) then (fitness is Low)	1
18. If (k-rcf is Medium) and (k-rff is Low) and (rcf-rff is Low) then (fitness is Low)	1
19. If (k-rcf is Low) and (k-rff is High) and (rcf-rff is High) then (fitness is Medium)	1
20. If (k-rcf is Low) and (k-rff is High) and (rcf-rff is Medium) then (fitness is Medium)	0.8
21. If (k-rcf is Low) and (k-rff is High) and (rcf-rff is Low) then (fitness is Low)	1
22. If (k-rcf is Low) and (k-rff is Medium) and (rcf-rff is High) then (fitness is Medium)	0.8
23. If (k-rcf is Low) and (k-rff is Medium) and (rcf-rff is Medium) then (fitness is Medium)	1
24. If (k-rcf is Low) and (k-rff is Medium) and (rcf-rff is Low) then (fitness is Low)	1
25. If (k-rcf is Low) and (k-rff is Low) and (rcf-rff is High) then (fitness is Low)	1
26. If (k-rcf is Low) and (k-rff is Low) and (rcf-rff is Medium) then (fitness is Low)	1
27. If (k-rcf is Low) and (k-rff is Low) and (rcf-rff is Low) then (fitness is Low)	1

3.6. Offspring generation

Offspring solutions are produced from parent solutions by applying selection, crossover and mutation operators. The knowledge about desirable solutions is advantageously stored in the population itself, and implicitly contained in the surviving chromosomes. We take advantage of this principle in selecting the fit solutions to add to mating pool and to use crossover and mutation and producing new generation. Among the selection strategies, roulette wheel selection, Tournament selection and rank-based selection are the most important ones (Srinivas and Patnaik, 1994). In roulette wheel selection, the selection method in this work, the probability of a chromosome being selected is proportional to its fitness. This is a stochastic algorithm and involves the following technique: The individuals are mapped to contiguous segments of a line, so that each individual's segment is equal in size to its fitness. A random number is generated and the individual whose segment spans the random number is selected. The process is repeated until the desired number of individuals is obtained. This technique is analogous to a roulette wheel with each slice proportional in size to the fitness.

3.7. Stopping criterion

A suitable stopping criterion must be chosen. This is typically achieved by limiting the number of generations or by setting some threshold which must be exceeded by the fitness function. If the stopping criterion is not satisfied, then individuals are selected from the current subset pool and the process described above repeats. As soon as satisfaction of stopping criteria GA has finished, the selected feature subset is evaluated in terms of the number of selected features and classification accuracy.

3.8. Correlation coefficient criterion

Pearson's correlation coefficient criterion is used for calculating the correlation between two variables. This criterion is the measure of the strength and direction of the linear relationship between two variables which is defined as the covariance of the variables divided by the product of their standard deviations.

$$p_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \qquad (1)$$

This Criterion is proposed by Karl Pearson (Pearson, 1896). *p* is a measure of the degree of the linear relationship between two variables which is called the correlation coefficient, and may take on any value between +1 and -1. The sign of the correlation coefficient explains the direction of the relationship. A larger value of correlation coefficient indicates which values of two variables tend to change together.

4. IMPLIMENTATION AND RESULTS

4.1. Experimental data

We evaluate the performance of the proposed method on 10 benchmark datasets which are summarized in Table 5. We can find all of these datasets in University of California Irvine Machine Learning Repository-UCI (Newman et al., 1998) and LKC(Kuncheva, 2004). Feature selection procedure is done with all samples in all datasets. In order to show the superiority of our method, eleven common feature selection methods including 9 filter methods and 2 embedded methods are also used to experiment. The summary of properties of all methods is shown in Table 6. All of these algorithms are implemented in a package with ASU feature selection repository (Zhao et al., 2010).

After the optimal feature subsets have been selected, the same independent test is used to estimate the performance for making a fair comparison. To make a reliable comparison among different feature selection methods, the test procedure is done in 10-fold method which is repeated 20 times to generate 20 different classification results. To compare the ranking methods with the proposed method, in the feature ranking methods the number of selected features are the same as the number of returned features by the proposed method. They are chosen from the beginning of the ranked list. In proposed method's experiments, it is important what the parameters of genetic algorithm are. These parameters have been chosen after some trial and error executions. This affects the result of implementation. Therefore these parameters are shown in Table 7.

Table 5. Characteristics of different datasets used for experiments.

Datasets	Features	Samples	Classes	Reference
Parkinson	22	195	2	(Newman et al., 1998)
Soybean	35	307	19	(Newman et al., 1998)
Yeast	8	1484	10	(Newman et al., 1998)
Zoo	16	101	7	(Newman et al., 1998)
Diabetes	8	768	2	(Newman et al., 1998)
Laryngeal	16	353	3	(Kuncheva, 2004)
Page-blocks	10	5473	5	(Newman et al., 1998)
Vowel	10	990	11	(Newman et al., 1998)
Seeds	7	210	3	(Newman et al., 1998)
Dermatology	34	366	6	(Newman et al., 1998)
Libras	90	360	15	(Newman et al., 1998)
Semeion	256	1593	2	(Newman et al., 1998)

Table 6. Properties of different methods used for comparisons.

Algorithm	Learning algorithm	methods type	univariate/ multivariate	Output
Kruskal wallis	supervised	filter	univariate	feature ranking
Gini index	supervised	filter	univariate	feature ranking
Information gain	supervised	filter	univariate	feature ranking
FCBF	supervised	filter	multivariate	feature subset
CFS	supervised	filter	multivariate	feature subset
Blogreg	supervised	embedded	univariate	feature subset
SBMLR	supervised	embedded	multivariate	feature subset
Fisher Score	supervised	filter	univariate	feature ranking
Relief-F	supervised	filter	univariate	feature ranking
Chi-square Score	supervised	filter	univariate	feature ranking

Table 7. Common GA parameters for all datasets

Parameters	Value/Description
Population size	50
Iterations	70
Type of crossover operation	One point crossover
Rate of crossover operation	0.99
Type of mutation operation	Uniform mutation
Rate of mutation operation	0.05
Type of selection operation	Rolette wheel
Stopping criterion of genetic process	Number of generation

4.2. Experimental results

The feature selection process is executed 20 times. The classification experiments with the selected optimal feature subsets are carried out. Classification methods are classical K-Nearest Neighbour (KNN) with K=1, 3, 7 and Binary Tree. It should be noted that the classification accuracy is not probably the same in all 20 rounds. Therefore, it is better to use the average classification accuracy in independent experiments as the index for evaluating the performance of the feature selection algorithms.

The accuracy of the classified samples and the number of the selected features are utilized to evaluate the ability for selecting good features by the proposed algorithm. The performance of each optimal feature set in terms of classification accuracy and the number of selected features for all eleven datasets is shown in Table 8 - Table 11. Meanwhile, the performance of the optimal selected feature subsets by the other ten algorithms are also listed in these Tables. These tables list the average performance in terms of average evaluation function of the 20 runs for all defined algorithms. The evaluation function is based on the number of the selected features and the classification accuracy. It is formulated as follows:

$$EF = \alpha \times (CA) + (1-\alpha) \times (NOF - NSF) / NOF \qquad , \alpha = 0.7 \tag{2}$$

Where *CA*, *NOF* and *NSF* are the classification accuracy, the number of original features and the number of selected features. The results show the superiority of the proposed method. It has the best average evaluation function among result of all datasets in Table 5.

Table 8. Evaluation Function with KNN (K=1).

Dataset	Cfs	FCBF	Sbmlr	Blogreg	MRMR	Fisher	RelifeF	Kruskal	Gini..	Infogain	Chis..	FCFS
Vowel	77.37	77.29	69.23	69.17	74.02	81.99	82.04	81.57	81.57	82.09	82.04	**83.07**
Parkinson	80.19	85.06	67.15	90.89	78.60	81.13	84.15	81.47	81.13	82.64	81.49	86.72
Soybean	78.17	75.63	63.54	64.40	61.58	68.46	75.10	66.74	66.54	76.44	76.25	**79.51**
Laryngeal	51.95	**64.98**	59.83	59.99	64.73	61.69	60.96	58.50	63.18	62.95	61.68	61.84

Seeds	68.40	78.22	69.40	73.70	81.04	73.58	73.58	75.93	82.88	73.53	73.44	**79.06**
Zoo	79.90	**84.44**	72.63	76.44	74.45	80.00	80.15	76.39	70.45	78.50	78.69	80.45
Yeast	40.47	41.97	44.13	36.83	45.51	46.58	47.31	47.23	47.48	47.31	47.25	**48.06**
Diabets	62.93	63.07	67.55	54.87	62.31	64.64	61.53	62.32	64.95	63.03	60.52	64.83
Page blocks	85.97	78.89	83.67	67.18	85.20	84.71	84.38	83.83	84.45	84.87	85.06	**86.38**
Dermatology	83.29	**84.72**	76.72	74.91	74.22	76.48	77.81	80.07	79.95	75.78	75.97	81.50
Libras	82.19	**87.91**	6.823	81.15	74.48	68.15	76.36	76.21	76.26	75.24	74.28	78.06
Semeion	66.66	64.65	59.29	68.58	65.45	67.95	67.97	66.45	67.96	67.88	68.00	**68.74**
Average	71.46	73.90	61.66	68.18	70.13	71.28	72.61	71.39	72.23	72.52	72.06	**74.85**

Table 9. Evaluation Function with KNN (K=3).

Dataset	Cfs	FCBF	Sbmlr	Blogreg	MRMR	Fisher	RelifeF	Kruskal	Gini..	Infogain	Chis..	FCFS
Vowel	75.54	75.58	67.77	67.81	70.91	79.40	79.48	78.55	78.65	79.48	79.33	**80.10**
Parkinson	80.24	84.71	65.27	**88.41**	79.08	82.28	83.36	81.72	81.68	80.59	81.88	85.48
Soybean	77.74	71.57	60.85	63.52	61.94	68.42	75.86	63.91	63.88	76.01	76.02	**77.91**
Laryngeal	55.81	**67.47**	62.29	64.88	65.05	66.05	66.36	64.05	65.31	65.36	64.02	65.37
Seeds	68.47	77.64	69.00	73.84	81.76	75.39	75.41	77.08	**82.74**	75.44	75.24	78.91
Zoo	78.80	**83.77**	69.88	75.93	74.34	79.41	79.67	77.29	71.45	79.69	79.25	80.62
Yeast	41.76	43.47	44.19	37.73	45.84	47.48	48.60	48.63	48.55	48.43	48.37	**48.65**
Diabets	66.15	**66.22**	71.57	56.88	62.31	65.72	63.45	64.60	65.81	64.97	64.55	65.75
Page blocks	**86.72**	78.96	84.03	67.17	85.39	85.08	84.53	84.58	84.57	85.01	85.23	85.13
Dermatology	83.91	**84.37**	77.20	76.05	75.51	76.56	79.13	81.22	81.07	75.81	75.80	81.52
Libras	77.76	**83.29**	62.96	75.81	70.30	60.24	71.79	69.69	69.71	68.33	68.12	76.02
Semeion	66.89	66.08	61.38	**68.70**	65.73	68.16	68.16	66.52	68.20	68.22	68.17	68.46
Average	71.65	73.59	66.37	68.06	69.85	71.18	72.98	71.49	71.80	72.28	72.17	**74.49**

Table 10. Evaluation Function with KNN (K=7).

Dataset	Cfs	FCBF	Sbmlr	Blogreg	MRMR	Fisher	RelifeF	Kruskal	Gini..	Infogain	Chis..	FCFS
Vowel	70.30	70.23	63.24	63.35	68.75	74.21	74.32	73.76	73.93	74.05	74.08	**75.00**
Parkinson	80.75	82.47	64.00	**89.03**	78.73	82.80	82.35	79.21	81.25	81.38	81.93	83.60
Soybean	73.87	69.31	57.25	59.31	59.64	64.61	71.44	61.14	61.02	73.31	72.67	**75.54**
Laryngeal	57.55	**69.93**	65.25	66.13	66.48	67.89	68.01	65.96	67.25	67.17	66.92	66.66
Seeds	69.27	77.97	69.57	73.50	**80.84**	75.78	76.01	78.64	82.71	75.88	75.74	79.81
Zoo	74.84	**80.86**	69.53	74.42	72.77	77.72	78.17	75.99	70.64	76.02	75.62	78.37
Yeast	43.89	46.01	47.17	40.84	49.28	48.07	51.05	51.05	**51.13**	51.04	51.04	50.40

Diabets	67.31	67.42	**73.69**	59.19	65.90	66.80	66.53	65.27	66.81	66.26	65.42	67.05
Page blocks	**87.32**	78.72	83.99	66.82	85.27	84.97	84.39	84.77	84.13	84.97	84.82	85.34
Dermatology	84.43	**85.42**	76.60	76.17	75.05	76.05	79.50	80.32	80.30	74.39	76.02	82.50
Libras	73.13	**79.44**	58.66	71.70	62.27	55.92	65.95	66.60	66.53	64.50	62.92	70.91
Semeion	67.06	66.24	62.42	68.66	65.70	68.16	67.91	66.23	68.17	68.12	68.15	**68.79**
Average	70.81	72.84	65.95	67.43	69.22	70.25	72.14	70.75	71.16	71.42	71.28	**73.66**

Table 11. Evaluation Function with Binary Tree.

Dataset	Cfs	FCBF	Sbmlr	Blogreg	MRMR	Fisher	RelifeF	Kruskal	Gini..	Infogain	Chis..	FCFS
Vowel	62.73	62.19	53.92	53.53	59.48	68.31	68.45	68.04	67.80	68.38	68.55	**69.06**
Parkinson	77.19	**83.09**	60.07	85.85	77.31	80.66	81.10	78.87	81.25	81.79	81.43	81.45
Soybean	74.48	73.12	59.71	61.12	65.35	66.65	73.79	63.55	63.55	74.72	75.11	**75.13**
Laryngeal	50.62	**63.88**	58.57	59.77	62.13	61.31	62.09	59.06	61.98	62.10	61.48	62.75
Seeds	69.09	77.77	69.60	73.27	81.56	74.64	74.84	75.28	**81.74**	75.03	74.61	76.24
Zoo	76.06	79.96	68.54	73.91	76.72	78.58	78.74	79.33	72.69	78.53	78.52	**81.81**
Yeast	41.44	43.44	45.60	37.42	47.90	49.20	50.28	50.33	50.11	49.86	50.20	**50.66**
Diabetes	64.32	64.33	**70.19**	56.69	62.13	64.79	64.10	62.11	64.86	64.71	63.30	64.89
Page blocks	85.85	79.70	85.40	67.55	85.34	85.52	85.28	84.43	85.58	85.69	85.50	**86.00**
Dermatology	82.60	**83.45**	74.81	74.73	72.62	73.75	78.99	78.96	78.64	76.91	74.02	81.25
Libras	67.90	**75.29**	56.26	67.29	60.76	59.73	64.56	64.58	64.76	62.30	61.58	65.49
Semeion	66.21	66.08	63.04	66.57	67.43	65.83	65.34	63.90	65.87	65.72	65.69	**67.39**
Average	68.21	71.03	63.81	64.81	68.23	69.08	70.63	69.04	69.90	70.48	70.00	**71.84**

4.3. Nonparametric tests

Nonparametric tests (NSKI et al., 2012; Conover and Iman, 1981), should be conducted in order to detect whether statistically significant differences occur among the examined algorithms or not. Moreover, these tests rank the algorithms from the best performing one to the worst one. If statistical significance is revealed, then post-hoc procedures should be accomplished to show which pair of algorithms differ significantly. In this paper Friedman test (Friedman, 1937; Garcıa and Herrera, 2008) is used. The Friedman test is a version of the repeated-measures ANOVA which can be performed on ordinal data. The goal of this test is to determine whether there are significant differences among the algorithms considered over given sets of data or not. These experiments were conducted using *KEEL* (Alcalá et al., 2011) (Knowledge Extraction based on Evolutionary Learning), a tool for creating, learning, optimizing and evaluating various models. *KEEL* has been developed in the Java environment by a group of Spanish research centres and is available freely for non-commercial purposes (NSKI et al., 2012).The test determines the ranks of the algorithms for each dataset; the best performing algorithm receives the lowest rank. The average ranking of all algorithms by Friedman test are shown in Table 12.The proposed method – FCFS has the best rank among all algorithms in Freedman test.

Table 12. Average ranking of the algorithms by Friedman procedure

Algorithm	Ranking
FCFS	**2.41**
FCBF	5.25
Relief-F	5.29
Gini-index	5.33
Infogain	5.70
Fisher	6.33
CFS	6.75
Chisquar	6.79
Kruskal	7.45
MRMR	8.08
Blogreg	8.58
Sbmlr	10

An adjusted p-value (probability value) can be directly taken as the p-value of a hypothesis belonging to a comparison of multiple algorithms in this test. If the p-value for an individual null hypothesis of equivalence of rankings by this test is less than the significance level (in our study α =0.05), this hypothesis is rejected (second column in Table 13). In this column the difference between FCFS and all other methods is significant except for Relief-F and FCBF.

After ranking procedure some post-hoc procedures are used on Friedman test results. In Table 13 the result of used post-hoc procedures is presented. Adjusted p-values for the Holm, Hommel, Holland, Rom, Finner, and Li post-hoc procedures for *1 × N* comparisons, where FCFS is the control algorithm, are displayed in Table 13. Thus, our study confirmed some observations made for evaluation function (defined in Eq.(2)) of algorithms. FCFS revealed significantly better performance than other algorithms. In Holm's procedure the p-values less than 0.0125 are bolded in the third column that means the differences between the average raking of FCFS and those of the other methods are significant except for the three last methods in Table 9 (InfoGain, Relief-F and FCBF). In Hommel's procedure similar to the results of Holm's procedure the differences between the FCFS and the last three methods are not significant. In Holland's procedure the results are similar to those of the Holm's and Hommel's procedures. In Finner's procedure the rejected hypothesis is in the last two methods. And as it can be seen in Li's procedure, this procedure rejects more hypotheses compared to the others methods; that means the difference between FCFS and all the others are meaningful except the FCBF.

Table 13. Post Hoc comparison table for *α* = 0.05 (Freedman)

Algorithm	p-value	Holm, Hommel	Holland	Finner	Li
Sbmlr	**0**	**0.004545**	**0.004652**	**0.004652**	**0.049777**
Blogreg	**0.000028**	**0.005**	**0.005116**	**0.009283**	**0.049777**
MRMR	**0.000118**	**0.005556**	**0.05683**	**0.013892**	**0.049777**
Kruskal	**0.000614**	**0.00625**	**0.006391**	**0.018479**	**0.049777**
Chisquar	**0.002956**	**0.007143**	**0.007301**	**0.023045**	**0.049777**
CFS	**0.003241**	**0.008333**	**0.008512**	**0.02759**	**0.049777**
Fisher	**0.007794**	**0.01**	**0.010206**	**0.032114**	**0.049777**
Gini-index	**0.025336**	**0.0125**	**0.012741**	**0.036617**	**0.049777**
InfoGain	**0.047537**	0.016667	0.016952	**0.041099**	**0.049777**
Relief-F	0.050799	0.025	0.025321	0.04556	**0.049777**
FCBF	0.054246	0.05	0.05	0.05	0.05

P-values obtained in by applying post hoc methods over the results of Friedman procedure:

- Holm's procedure rejects those hypotheses that have a p-value ≤ 0.0125.
- Hommel's procedure rejects those hypotheses that have a p-value ≤ 0.0125.
- Holland's procedure rejects those hypotheses that have a p-value ≤0.012741.
- Finner's procedure rejects those hypotheses that have a p-value ≤0.041099.
- Li's procedure rejects those hypotheses that have a p-value ≤ 0.049777.

This article can be cited as S. Barchinezhad and M. Eftekhari, A New Fuzzy and Correlation Based Feature Selection Method for Multiclass Problems, International Journal of Artificial Intelligence, vol. 12, no. 2, pp. 24-41, 2014.

5. CONCLUSION

There have been many approaches to the feature selection based on a variety of techniques. Each of these has its own advantages and disadvantages which make them context-specific, and there is no universally the best feature selection method. In this paper we propose a new filter method - FCFS for the feature selection using genetic algorithm and fuzzy sets. We choose GA due to its simplicity and its capability as a powerful search mechanism. GA is used for optimizing a fitness function calculated by four fuzzy systems. This fitness function considers three criteria e.g. the number of selected features, relevancy and redundancy. Pearson’s correlation coefficient criterion was utilized to compute the relevancy and the redundancy of features. We implement this multiclass filter, and evaluate its performance using some benchmark datasets. It was compared with some common filter and embedded algorithms. The results of comparisons reveal the superiority of the proposed algorithm compared to the others in terms of classification performance and the number of selected features. Some non-parametric statistical tests were utilized in order to reach a scientific comparison of the results. The outcome of these tests confirms the ability of the proposed method in comparison to the others. This multiclass feature selection method is suitable for both discrete and continuous data. Another advantage of the proposed algorithm is that the relevancy and redundancy of the selected features are considered via fuzzy concepts. Therefore, in the proposed fuzzy based measure, the relevancy and redundancy can be defined subjectively according to the expert opinions. Some computational operations are imposed into the algorithm by utilizing the Mamdani FIS (i.e. computations of inference mechanism) which can be considered as disadvantage of the proposed method. Producing a nonlinear relation between the relevancy and redundancy is the added advantage of our proposed method that is achieved by using Mamdani FIS. The aforementioned benefits make the proposed method a useful technique despite of its added load of computation.

REFRENCES

Ahmed, A.A., 2006, Feature subset selection using ant colony optimization, *International Journal of Computational Intelligence, IJCI 2,* 53-58.

Alcalá, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., Herrera, F., 2011, KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic and Soft Computing*, 17, 255-287.

Basak, J., De, R.K., Pal, S.K.,1998, Unsupervised feature selection using a neuro-fuzzy approach. *Pattern Recognition Letters* 19 , 997-1006.

Bradley, P.S., Mangasarian, O.L., Street, W.N.,1998, Feature selection via mathematical programming. *INFORMS Journal on Computing* 10 , 209-217.

Cai, R., Hao, Z., Yang, X., Wen, W., 2009, An efficient gene selection algorithm based on mutual information. *Neurocomputing* 72 , 991-999.

Cawley, G.C., Talbot, N.L., 2006, Gene selection in cancer classification using sparse logistic regression with Bayesian regularization. *Bioinformatics* 22 , 2348-2355.

Cawley, G.C., Talbot, N.L., Girolami, M., 2007, Sparse multinomial logistic regression via bayesian l1 regularisation. *Advances in neural information processing systems* 19, 209-216.

Chen, D., Hu, Q., Yang, Y., 2011, Parameterized attribute reduction with Gaussian kernel based fuzzy rough sets. *Information Sciences* 181, 5169-5179.

Conover, W.J., Iman, R.L., 1981, Rank Transformations as a Bridge between Parametric and Nonparametric Statistics. *The American Statistician* 35 ,124-129.

Covoes, T.F., Hruschka, E.R., 2011, Towards improving cluster-based feature selection with a simplified silhouette filter. *Information Sciences* 181 ,3766-3782.

Duda, R.O., Hart, P.E., Stork, D.G., 2001, *Pattern Classification*. John Wiley & Sons, New York.

Elomaa, T., Ukkonen, E.,1994, A geometric approach to feature selection. *ECML* 94, 351-354.

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.

Garcıa, S., Herrera, F., 2008, An extension on statistical comparisons of classifiers over multiple data sets for all pairwise comparisons. *Journal of Machine Learning Research* 9, 2677-2694.

Guyon, I., Elisseeff, A., 2003, An introduction to variable and feature selection. *Journal of Machine Learning Research* 3, 1157-1182.

Hall, M.A.,1999, *Correlation-based feature selection for machine learning*. PhD Thesis, University of Waikato.

Hall, M.A., Smith, L.A.,1997, Feature subset selection: a correlation based filter approach. *International Conference on Neural Information Processing and Intelligent Information Systems 4, Dunedin, New Zealand.*

Huang, J., Cai, Y., Xu, X., 2007, A hybrid genetic algorithm for feature selection wrapper based on mutual information. *Pattern Recognition Letters* 28 ,13, 1825-1844.

Joelianto, E., Anura, D. C., Priyanto, M. P., 2013, ANFIS hybrid reference control for improving transient response of controlled systems using PID controller. *International Journal of Artificial Intelligence*, 10, 88-111.

Khan, I.H., 2014, A comparative study of evolutionary algorithms, *International Journal of Artificial Intelligence*, 12, 1-17.

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

Kira, K., Rendell, L.A., 1992, A practical approach to feature selection. *International Conference on Machine Learning-ICML 9*2, 249-256.

Kudo, M., Sklansky, J., 2000, Comparison of algorithms that select features for pattern classifiers. *Pattern recognition* 33 ,25-41.

Kuncheva, L.I., 2004, Ludmila Kuncheva Collection. [http://pages.bangor.ac.uk/~mas00a/activities/real_data.htm].

Lin, H.Y., 2013, Feature selection based on cluster and variability analyses for ordinal multi-class classification problems. *Knowledge-Based Systems* 37, 94-104.

Liu, H., Setiono R.,1995, Chi2: Feature selection and discretization of numeric attributes. International Conference on Tools with Artificial Intelligence 24 , 388-391.

Liu, H., Setiono, R., 1997, Feature selection via discretization. *Knowledge and Data Engineering* 9, 642-645.

Luukka, P., 2011, Feature selection using fuzzy entropy measures with similarity classifier. *Expert Systems with Applications* 38 , 4600-4607.

Newman, D., Hettich, S., Blake, C, Merz C, 1998, UCI Repository of machine learning databases. Irvine, CA: University of California. Department of Info. and Comput. Sci. [http://www. ics. uci. edu/~ mlearn/MLRepository. html].

Nski, B.T., Etek, M.S., Telec, Z., Lasota, T., 2012, Nonparametric statistical analysis for multiple comparison of machine learning regression algorithms. *Int. J. Appl. Math. Comput. Sci*. 22 ,867-881.

Pearson, K., 1896, Mathematical Contributions to the Theory of Evolution. *Proceedings of the Royal Society of London* 60 , 489-498.

Peng, H., Long, F., Ding, C., 2005, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 27, 1226-1238.

Popov, A., 2005, *Genetic algorithms for optimization-programs for MATLAB*. User Manual, Hamburg 1, 3-21.

Precup, R.E., Tomescu, M.L., Preitl, S., Petriu, E.M., 2009, Fuzzy logic-based stabilization of nonlinear time-varying systems, *International Journal of Artificial Intelligence*, 3, 24-36.

Purcaru, C., Precup, R.E., Iercan, D., Fedorovici, L.O., David, R.C., Dragan, F., 2013, Optimal robot path planning using gravitational search algorithm, *International Journal of Artificial Intelligence,* 10, 1-20.

Robnik-Šikonja M., Kononenko, I., 2003, Theoretical and empirical analysis of ReliefF and RReliefF. *Machine learning* 53, 23-69.

Sikora, R., Piramuthu, S., 2007, Framework for efficient feature selection in genetic algorithm based data mining. *European Journal of Operational Research* 180, 723-737.

Srinivas, M., Patnaik, L.M., 1994, Genetic algorithms: A survey. *Computer* 27,6 , 17-26.

Sun, X., Liu, Y., Li, J., Zhu, J., Liu, X., Chen, H., 2012, Using cooperative game theory to optimize the feature selection problem. *Neurocomputing* 97, 86-93.

Tsai, C.F., Eberle, W., Chu, C.Y., 2013, Genetic algorithms in feature and instance selection. *Knowledge-Based Systems* 39, 240-247.

Unler, A., Murat, A., Chinnam, R.B., 2011, mr ^{2} PSO: A maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification. *Information Sciences* 181, 4625-4641.

Wang, J., Wu, L., Kong, J., Li, Y., Zhang, B., 2013, Maximum weight and minimum redundancy: A novel framework for feature subset selection. *Pattern Recognition* 46, 1616-1627.

Wang, X., Yang, J., Teng, X., Xia, W., Jensen, R., 2007, Feature selection based on rough sets and particle swarm optimization. *Pattern Recognition Letters* 28, 459-471.

Wei, L., 1981, Asymptotic conservativeness and efficiency of kruskal-wallis test for k dependent samples. *Journal of the American Statistical Association* 76, 1006-1009.

Yang, Y., Pedersen, J.O., 1997, A comparative study on feature selection in text categorizatio*n. International Conference on Machine Learning- ICML 97,* 412-420.

Yu, L., Liu, H., 2003, Feature selection for high-dimensional data: A fast correlation-based filter solution. *International Conference on Machine Learning- ICML* 3, 856-863.

Zhao, Z., Morstatter, F., Sharma, S., Alelyani, S., Anand, A., Liu, H., 2010, Advancing feature selection research. *ASU Feature Selection Repository*, 1-28.

Zhou, X.J., Dillion, T.S.,1991, A statistical-heuristic feature selection criterion for decision tree induction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13 ,834-841.