An Evolutionary Based Multi-Objective Filter Approach for Feature Selection

Mahdieh Labani[1], Parham Moradi[1], Mahdi Jalili[2], Xinghuo Yu[2]

[1] Department of Computer Engineering, University of Kurdistan, Sanandaj, Iran
[2] School of Engineering, RMIT, Melbourne, Australia

***Abstract*- Feature selection is one of the important research areas in pattern recognition. The aim of feature selection is to select those of informative features to improve the classifier's performance. In this paper, we propose a novel multi-objective algorithm based on mutual information for feature selection, called multi-objective mutual information (MOMI). The proposed method identifies a set of features with minimal redundancy and maximum relevancy with the target class. Several experiments are performed to evaluate the performance of MOMI compared to that of well-known and state-of-the-art feature selection methods over five benchmark datasets. The results show that in most cases MOMI achieves better classification performance than others.**

***Keywords* - multi-objective optimization; mutual information; feature selection, dimensionality reduction.**

I. Introduction

Data mining is a useful tool to analyze data and extract useful information. A major issue associated with data mining is large numbers of redundant features of data which might lead to reduce the accuracy of classifiers. An effective solution to this issue is removing redundant features is select the effective ones; this is known as feature selection in the literature. Feature selection has been successfully applied to many research areas such as web page classification [1], text categorization [2], and Gene selection [3].

Generally, feature selection methods can be classified into four categories including filter, wrapper, embedded and hybrid methods. The filter approaches evaluate the usefulness of each feature using statistical properties of the data. The wrapper approach employs a learning algorithms to search the solution space for a good subset of features[3, 4]. In the embedded model, a given learning algorithm is trained in such a way that a reduced feature set and also the learning model is simultaneously constructed. Finally, the hybrid approach has advantages of both the filter and wrapper approaches [2]. The wrapper, hybrid and embedded methods can often select better subsets with higher accuracy compared to filter methods due to using a learning algorithm in their processes. On the other hands, filter methods are much faster than others, and thus are widely used, especially for huge feature spaces [5-15].

The filter-based feature selection methods can be classified into univariate and multivariate methods. In univariate methods, features are assessed based on a given relevance criterion [7]. On the other hand, multivariate methods consider both irrelevant and redundant features in their processes, and thus provide better efficiency compared to univariate methods. Examples include MIFS [16], NMIFS [12], and UFSACO [6]. Although, multivariate methods are effective, most of them are based on mutual information which has two major limitations. First, they use a greedy search mechanism which often generates local optimal solutions. The second problem is that they need user-defined parameters to establish a balance between relevancy and redundancy metrics which should be determined precisely for different datasets.

In this paper, we propose a novel multi-objective filter-based feature selection algorithm – called Multi Objective Mutual Information (MOMI). The proposed method considers both the relevancy and redundancy concepts in its evaluation process. MOMI uses NSGA II [17] – a multi-objective genetic algorithm – in its search process. In the proposed method, the mutual information between selected features and also the mutual information between features and classes are used as its objectives. We apply the proposed feature selection on a number of benchmark datasets and show its effectiveness over state-of-the-art methods.

II. Multi-Objective Optimization

Multi-objective problem is a process of optimizing two or more competing objectives subject to certain constraints. An optimization problem with M conflicting objective functions can be formulated as follows:

$$\min[f_1(x), f_2(x), ..., f_M(x)] \quad , s.t. x \in X \tag{1}$$

where X denotes a set of feasible solutions. Generally, a solution x is said to be non-dominated (i.e., Pareto optimal) if there is no other feasible solution which dominates x. In other words, we say that solution x_1 dominate x_2 if the following conditions are met:

$$\{f_i(x_1) \le f_i(x_2)\} \wedge \{\exists i = 1, 2, ..., M : f_i(x_1) < f_i(x_2)\} \quad \forall i = 1, 2, ..., M \tag{2}$$

978-1-5090-5573-9/17 $31.00 © 2017 IEEE
DOI 10.1109/WCCCT.2016.44

III. Proposed multi-objective feature selection method

In this section, we describe the proposed multi-objective feature selection method (MOMI). The proposed method consists of four steps including: (1) Initialization, (2) Fitness assignment, (3) Population ranking and (4) Mating selection. In the first step, an initial population is randomly generated. Then, in the second step for each solution, the value of each objective is calculated. In the third step, an iterative non-dominated sorting process is employed to identify Pareto optimal solutions. In the fourth step, crossover and mutation operators are applied on the Pareto optimal set to produce a new population. Finally, if the stop conditions are met (e.g., predefined number of iterations or generations) the algorithm stops.

The proposed method starts with a random population. To initialize the solutions, a vector with the length of the number of selected features is used. The index of each feature (i.e., feature number) in this vector indicates that the corresponding feature is selected.

The relevancy and redundancy in the multi-objective feature selection methods are often considered as their objective functions. To compute the relevancy value of each feature, the mutual information between the feature and the target class variable is computed (Eq. (3)). Similarity, summation of mutual information between each two features is considered as the redundancy value of a feature set (Eq. (4)). These objectives are defined as follows:

$$Rel(C,S) = \sum_{f_i \in S} MI(f_i, S) \tag{3}$$

$$Red(C,S) = \sum_{f_i, f_j \in S} MI(f_i, f_j) \tag{4}$$

where $MI(X,Y)$ shows mutual information of two variables X and Y that are defined as follows:

$$MI(X;Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x).p(y)} \tag{5}$$

where $p(x,y)$p(x,y) is joint probability of x and y which occurs together and $p(x)$ and $p(y)$ are probability distribution of variables x and y, respectively.

The proposed method attempts to identify a set of features with maximum relevancy with the target class and minimum redundancy between the set. To consider these objectives in the search process, one should minimize both $-Rel(C,S)$ and $Red(C,S)$.

The next step is to rank the population. The aim of population ranking is to assign a rank value to each solution, and thus identifying the non-dominated solutions. For this purpose, in each iteration the population is divided into dominated and non-dominated solutions, which are denoted by n_p and S_p, respectively. The first front consists of non-dominated solutions. The second non-dominated front is also recognized among S_p. This process is continued until all solutions are placed in their proper fronts. Then, a metric called crowding distance is used to rank each solution, which estimates the density of solutions by computing the mean distance of two points on either side of these points along each objectives. The crowding distance for a solution i is defined as follows:

$$d_i = \sum_{j=1}^{2} d_i^j$$

where d_i^j is the crowding distance of solution i to j-th objective and obtained as follows:

$$d_i^j = \frac{\left| f_j^{i+1} - f_j^{i+1} \right|}{f_j^{\max} - f_j^{\min}}$$

where f_j^{max} and f_j^{min} are maximum and minimum values of j-th objective, respectively.

The proposed method, searches the solution space using the crossover and mutation operators. In the crossover operation, each two parents generate two new offspring. In the mutation operator, feature indices in each offspring chromosomes mutate a new offspring with a mutation probability. Finally, the current and new populations are combined to form a new population for further processes.

IV. Simulation results

To evaluate the performance of the proposed method, a set of experiments were performed to compare the proposed method (MOMI) with a number of well-known and state-of-the-art filter-based feature selection methods including IG [8], GI [9], GR [15], UFSACO [6], RRFS [18], and mRMR [11]. All experiments were conducted in windows 7 environment on a PC having core i5 processor and 8 GB RAM. We used Weka framework [19] for implementing Naïve Bayes (NB) and Support Vector Machine (SVM) classifiers. The proposed method includes a number of adjustable parameters. The values of the parameters were chosen after a number of predefined runs, and they may not be optimal values. The population size is set to 100. Maximum number of cycles was also set to 100.

The experiments were performed on a number of datasets including Hepatitis, Dermatology and Spam Base taken from the UCI machine learning repository [20]. Additional details about these datasets are given in Table 1.

Table 1. Datasets used in the experiments

Datasets	Features	Patterns	Classes
Hepatitis	19	155	2
Spam Base	57	4601	2
Dermatology	34	366	2

The efficiency of the proposed method was evaluated in terms of F-measure and the number of selected features. F-measure combines both precision and recall and is defines as follows:

$$F = 2.\frac{P.R}{P+R}$$

where precision P is the fraction of retrieved instances that are relevant and recall R is the fraction of relevant instances that are retrieved.

Tables 2 reports the F-measure results achieved by the proposed method and when NB classifier is used. The best results for each dataset are indicated in bold face and the second best is identified in bold face and underlined. From Table 2 it can be seen that in most cases the proposed method outperformed other methods and achieved higher F-measure values. IG, GR, GI are uni-variate methods. The proposed method is also compared with two well-known multivariate methods including RRFS and UFSACO and the obtained results are shown in Table 3. This table reports the mean of F-measure results over five independent runs when SVM classifier is used. The standard deviation values are shown inside parentheses. It is clear from the results that the proposed method achieves the best result in most cases, while having the second-best performance for other cases. For example, for Glass dataset the proposed method obtained the highest F-measure when different numbers of features were selected.

Table 2. Comparison of F-measure of algorithms when using NB classifier.

Dataset	# features	Feature selection methods				
		MOMI	*IG*	*GR*	*GI*	*mRMR*
Hepatitis	4	0.82	**0.90**	0.81	0.79	0.85
	10	**0.89**	0.86	0.85	0.80	**0.89**
Spam Base	11	**0.91**	0.75	0.80	0.73	0.48
	29	0.86	**0.87**	0.77	**0.88**	0.72
Dermatology	7	**0.81**	0.75	0.57	0.68	**0.81**
	17	0.92	**0.95**	0.89	0.89	**0.97**

Table 3. Average F-measure values using SVM classifier over 5 independent runs.

Dataset	#features	Feature selection methods		
		MOMI	*UFSACO*	*RRFS*
Hepatitis	4	**0.97(0.042)**	**0.95(0.165)**	0.81(0.21)
	10	**0.98(0.014)**	0.70(0.105)	**0.82(0.29)**
Spam Base	11	**0.91(0.038)**	0.78(0.076)	**0.92(0.065)**
	29	**0.99(0.012)**	0.87(0.065)	**0.95(0.034)**
Dermatology	7	**0.88(0.013)**	**0.85(0.148)**	0.83(0.198)
	17	**0.94(0.006)**	0.89(0.021)	**0.97(0.16)**

Table 4 compares the execution time of the proposed method with multivariate feature selection method when 20 percent of Features on each dataset was selected. These show that the method considered a trade-off between the execution time and the quality of the solution. It can be concluded from the Tables 2 and 3 that the proposed method achieved better qualitative results by spending a little more time compared to other feature selection methods.

Table 4. Execution time (in seconds) comparison of feature selection methods.

Dataset	Feature selection methods			
	MOMI	*mRMR*	*UFSACO*	*RRFS*
Hepatitis	48.77	0.02	**0.01**	**0.01**
Spam Base	74.45	150.16	0.10	**0.01**
Dermatology	93.92	0.42	0.02	**0.01**

Finally the following conclusions can made from the conducted experiments:

- The achieved results reveal that the proposed method performs better than the traditional univariate feature selection methods such as IG, GI and GR. This is because of the fact that these methods do not take into account the similarity between features.
- Although mRMR, UFSACO and RRFS are multivariate methods, the results show that the proposed method significantly performed better than these method. This is due to the fact that MOMI takes an advantage of multi-objective algorithms to consider various objectives simultaneously in its process.

V. Conclusion

In this paper, a novel multi-objective feature selection method (MOMI) was proposed. In each iteration, the MOMI produces a set of non-dominant solutions by analyzing both the relevancy and redundancy of the selected features. We compared the performance of the proposed method with a number of well-known and state-of-the-art methods by applying them on five datasets (Glass, Wine, Hepatitis, SpamBase and Dermatology) and using NB and SVM as classifiers. The experimental results show that the proposed method can effectively remove irrelevant and redundant features which leads to achieve the best performance.

References

[1] A. Selamat and S. Omatu, "Web page feature selection and classification using neural networks," *Information Sciences,* vol. 158, pp. 69-88, 2004.

[2] H. Uğuz, "A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm," *Knowledge-Based Systems,* vol. 24, pp. 1024-1032, 2011.
[3] S. Tabakhi, A. Najafi, R. Ranjbar, and P. Moradi, "Gene selection for microarray data classification using a novel ant colony optimization," *Neurocomputing,* vol. 168, pp. 1024-1036, 2015.
[4] P. Moradi and M. Gholampour, "A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy," *Applied Soft Computing,* vol. 43, pp. 117-130, 2016.
[5] S. Tabakhi and P. Moradi, "Relevance–redundancy feature selection based on ant colony optimization," *Pattern Recognition,* vol. 48, pp. 2798-2811, 2015.
[6] S. Tabakhi, P. Moradi, and F. Akhlaghian, "An unsupervised feature selection algorithm based on ant colony optimization," *Engineering Applications of Artificial Intelligence,* vol. 32, pp. 112-123, 2014.
[7] P. Moradi and M. Rostami, "A graph theoretic approach for unsupervised feature selection," *Engineering Applications of Artificial Intelligence,* vol. 44, pp. 33-45, 2015.
[8] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *ICML*, 2003, pp. 856-863.
[9] L. E. Raileanu and K. Stoffel, "Theoretical comparison between the gini index and information gain criteria," *Annals of Mathematics and Artificial Intelligence,* vol. 41, pp. 77-93, 2004.
[10] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Advances in neural information processing systems*, 2005, pp. 507-514.
[11] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 27, pp. 1226-1238, 2005.
[12] P. A. Estévez, M. Tesmer, C. A. Perez, and J. M. Zurada, "Normalized mutual information feature selection," *Neural Networks, IEEE Transactions on,* vol. 20, pp. 189-201, 2009.
[13] H. Zare and M. Niazi, "Relevant based structure learning for feature selection," *Engineering Applications of Artificial Intelligence,* vol. 55, pp. 93-102, 10// 2016.
[14] Q. Gu, Z. Li, and J. Han, "Correlated multi-label feature selection," in *Proceedings of the 20th ACM international conference on Information and knowledge management*, 2011, pp. 1087-1096.
[15] T. M. Mitchell, "Machine learning. WCB," ed: McGraw-Hill Boston, MA:, 1997.
[16] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *Neural Networks, IEEE Transactions on,* vol. 5, pp. 537-550, 1994.
[17] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *Evolutionary Computation, IEEE Transactions on,* vol. 6, pp. 182-197, 2002.
[18] A. J. Ferreira and M. A. Figueiredo, "An unsupervised approach to feature discretization and selection," *Pattern Recognition,* vol. 45, pp. 3048-3060, 2012.
[19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter,* vol. 11, pp. 10-18, 2009.
[20] C. Blake and C. J. Merz, "{UCI} Repository of machine learning databases," 1998.