ISBN: 978-972-8924-40-9 © 2007 IADIS

FEATURE RANKING BASED ON WEIGHTS ESTIMATED BY MULTIOBJECTIVE OPTIMIZATION

Daniela Zaharie
West University of Timisoara
Bv. V. Parvan, no. 4, 300223 Timisoara, Romania
dzaharie@info.uvt.ro

Diana Lungeanu
University of Medicine and Pharmacy of Timisoara
Pta Eftimie Murgu, no. 2A, 300041 Timisoara, Romania
dilun02@yahoo.com

Stefan Holban
Politehnica University of Timisoara
Bv. V. Parvan, no. 2, 300004 Timisoara, Romania
stefan@cs.utt.ro

ABSTRACT

The aim of this paper is twofold. On the one hand, we analyze a feature ranking technique based on the weights estimated by an evolutionary algorithm for multiobjective optimization. On the other hand, we address the problem of comparing and aggregating different rankings obtained either by applying different methods to the same dataset, or by applying, in the context of distributed data mining tasks, the same method to different datasets.

KEYWORDS

Feature selection, feature ranking, ranking aggregation, multi-objective optimization, evolutionary algorithms, distributed data mining.

1. INTRODUCTION

In the context of medical data mining, feature subset selection (FSS) and feature ranking (FR) are valuable tools in the preliminary steps for identifying risk factors for different pathologies. They allow identifying the relevant features with respect to the data mining task and discarding the redundant or irrelevant ones. In spite of their similar aims, FSS and FR rely on different techniques and generate the results in a different manner. FSS generates a subset of features by applying a searching procedure on the subset space. Each subset is evaluated with respect to its relevance for the envisaged data mining task (e.g. classification, clustering) and the assigned score is a relevance measure for all selected features. This way, no difference is made between the selected features even if they are not equally relevant. In contrast, FR generates a ranking of all features by individually analyzing them, thus the relevance of features' combination is not analyzed and the correlation between features is not fully explored. The limitations of each technique can be addressed by a hybrid approach, as in the method proposed in (Ruiz et al., 2005), where a feature subset is generated by applying a heuristic search over a ranking.

In this paper we present a different approach based on associating weights to the features and on searching for a set of weights which best express the features' relevance to a classification task. The weights can be interpreted as degrees of membership to the subset of relevant features and can be used to generate feature rankings. When the weights are binary, the problem is equivalent to that of FSS, but in the general case the search space is a continuous one. The idea of associating weights to features is a natural one and, in different contexts, has already been explored (Scherf and Brauer, 1997).

To solve the problem of estimating the weights, we had to choose: (i) one or several criteria expressing the quality of a weights set; and (ii) a searching procedure. Usually, one criterion is not powerful enough, so we used several criteria, then formulated the problem as a multi-objective optimization one, and finally solved it using an evolutionary algorithm. The evolutionary algorithm produces a set of reciprocally non-dominated vectors corresponding to a set of weight vectors. Multi-objective optimization based on evolutionary algorithms has previously been applied in feature selection (Pappa et al, 2002), (Handl and Knowles, 2006), but with binary weights instead of continuous ones.

In feature ranking, it is not unusual to obtain different rankings by different methods. Moreover, when analyzing horizontally distributed data which cannot be accumulated in a central server (e.g. for security and privacy requirements, or for organizational reasons) we would obtain several rankings, so natural questions arise: *Which one to choose? How can we compare or aggregate several rankings in order to extract as much as possible information from them?* The problem of aggregating feature rankings has recently been addressed in (Jong et al, 2005) by using ideas from ensemble learning.

2. FEATURE WEIGHTING AS A MULTI-OBJECTIVE OPTIMIZATION PROBLEM

2.1 Problem Formulation

The problem to be solved is very similar to that of feature subset selection but instead of looking for a binary vector, a continuous vector should be estimated. In the case of n features we search for a vector $(w_1,\ldots,w_n)$, with w_i in [0,1], which optimize one or several quality criteria with respect to the classification task. The choice of the quality criteria depends on the technique (wrapper or filter) and on the data mining task (supervised or unsupervised). In the case of filter techniques, one can rely only on the data properties reflected by the training set. For a supervised classification, criteria to be used are: intra-class dissimilarity (C_1), inter-class dissimilarity (C_2), or attribute-class correlation (C_3) (Wang and Fu, 2005). In the case of unsupervised classification, criteria which do not involve the class label, e.g. entropy (C_4) (Handl and Knowles, 2006), should be used instead.

For criteria involving just distances between data (e.g. entropy) or between data and averages (e.g. intra-class and inter-class dissimilarities) the weights can be included by using a weighted distance. For instance, in the case of Euclidean distance this would mean: $d_w(x,y)=\sqrt{\sum_{i=1}^{n} w_i^2 (x^i - y^i)^2}$. When features' averages are involved, the simple average should be replaced with a weighted one. For instance, in the case of attribute-class correlation the weighted version is:

$$C_3 = \left(\sum_{i=1}^{n} w_i c(i)\right) / \left(\sum_{i=1}^{n} w_i\right), \qquad c(i) = \frac{\sum_{j \neq k} |x_j^i - x_k^i| \varphi(x_j, x_k)}{n(n-1)/2} \tag{1}$$

where: x_j^i denotes the component i of data j; $\varphi(x,y)$ is 1 if x and y belong to different classes and it is -0.05 if they belong to the same class (Wang and Fu, 2005). By changing the Euclidean distance with the weighted Euclidean distance, the entropy criterion will be (Handl and Knowles, 2006):

$$C_4 = -\sum_{i,j} (s_{ij} \ln s_{ij} + (1-s_{ij}) \ln(1-s_{ij})); \quad s_{ij} = e^{-\alpha d_w(x_i, x_j)}, \quad \alpha = -\ln 0.5 / \overline{d_w} \tag{2}$$

where $\overline{d}_w$ is the average of weighted distances between all pairs of data. In the general case, in order to take the weights into consideration for the quality criteria computation, it is enough to replace the data components by their product with the corresponding weight. None single criteria is powerful enough to measure the ability of components to correctly discriminate the classes, mainly because they are biased towards extreme cases (e.g. assigning a significant weight to one or very few features or, on the contrary, assigning significant weights to all features). This is why the criteria should be combined such that they

compensate each other's tendency to favor extreme cases. For instance, one usually combines the intra-class dissimilarity (which favors the selection of a small number of features and should be maximized) with the inter-class dissimilarity (which favors the selection of a large number of features and should be minimized).

2.2 Solving the Multi-Objective Optimization Problem

The criteria can be combined into one criterion leading to a single objective optimization problem, as in (Wang and Fu, 2005), or the problem can be treated as a multi-objective optimization one. The technique we propose belongs to the last category and we applied an evolutionary algorithm (e.g. NSGA-II) in order to approximate the Pareto optimal set. By applying a multiobjective evolutionary algorithm one obtains not just one weight vector, but a set of reciprocally non-dominated weight vectors.

As each weight vector leads to a features' ranking, the final ranking can be obtained in different ways. A first variant would be to compute, for each feature, a statistic (average, median or maximum) of the corresponding weights and to further use these statistics in order to establish the ranking. Another variant would be to construct the rankings starting from the weights vectors and to aggregate them as in ensemble feature ranking (Jong, 2004). The aggregation can be based on a voting mechanism which assigns to a feature the most frequent rank. The drawbacks of this approach are related with the fact that, for a given feature, different ranks can have the same maximal frequency while different features can be associated with the same rank. A different approach, which partially avoids these drawbacks, is to compute, for each feature, the averaged rank and to construct the final ranking by increasingly sorting the features based on the averaged ranks. This last variant was used in most of our experiments. Unfortunately, these variants do not necessarily lead to the same final ranking, thus the problem of comparing and aggregating different rankings arises.

2.3 Comparing and Aggregating Rankings

In addition to the situation discussed in the previous subsection, other circumstances with several rankings for the same set of features occur when analyzing data from different locations (distributed data) without transferring them to a central location. Supposing that at each location the data have the same set of features, we could extract a local feature ranking for each dataset and aggregate these local rankings in order to obtain a global one. To deal with the problem of aggregating different rankings, we shall use the concept of *generic ranking*. In order to define the concept, let us denote by F the set of all features. A *generic ranking* is a labeled partition of F in the sense that it consists of disjoint subsets, each one having assigned a rank. For a simple ranking, the subsets are all singletons, suggesting that there is a total order relationship on the features set. In the case of an arbitrary generic ranking, the features belonging to the same subset can be considered to be indiscernible from the point of view of their relevance. Let us suppose that we have a set of simple rankings $R_1,\ldots,R_q$ and we are interested in aggregating them into a generic ranking. If $card(F)=n$, constructing the aggregated generic ranking is equivalent to finding the longest sequence of indices, $1 = k(1) < k(2) < \ldots < k(s) \leq n$, such that for each i in {1,..,s} the sets of features having ranks between $k(i)$ and $k(i+1)-1$ for all analyzed rankings are identical. By convention, $k(s+1)=n+1$. The generic ranking will consist of s subsets of features, $S(1),\ldots S(s)$. The subset $S(i)$ will contain the features having ranks in $\{k(i),k(i)+1,\ldots,k(i+1)-1\}$ with respect to all rankings. In the case of two identical rankings, s will be n and $k(i)=i$. For different rankings s will be less than n. For example, for the rankings $R_1=(4,3,2,1)$, $R_2=(3,4,1,2)$, $R_3=(3,4,2,1)$ the generic ranking will be $(\{3,4\},\{1,2\})$ meaning that, based on these rankings, both features 3 and 4 are more relevant than features 1 and 2. In the case of rankings $R_1=(4,3,1,2)$ and $R_2=(2,1,3,4)$ the ranking obtained by aggregation will be $(\{1,2,3,4\})$ meaning that the features are indiscernible from the point of view of their relevance. The value s can be interpreted as a measure of the similarity between two or several rankings. It is maximal ($s=n$) when the rankings are identical and minimal ($s=1$) when the aggregated generic ranking contains just one set.

3. EXPERIMENTS AND DISCUSSION

This section presents some preliminary results obtained by applying the above described weighting method. In all tests we used the NSGA-II evolutionary algorithm (Deb et al, 2000) to solve the multi-objective optimization problem. We constructed the final ranking based on averaging the rankings generated by the weights corresponding to each Pareto solution. Parameters of the NSGA used in experiments were: 80 (population size), 250 (number of generations), 0.9 (probability of recombination), 1/n (probability of mutation). In order to analyze the behavior of the weighting method and of the different rankings' aggregation, we started with a simple data set consisting of data belonging to two classes and having 10 independent features generated as follows. The first feature was just the class label, thus this should be the most relevant feature. The next 5 features were generated based on two normal distributions having different means for the two classes. Features 7 and 8 were constant values over all data, and features 9 and 10 had random values uniformly generated in [0,1]. Different preliminary experiments were conducted on these data, aimed at identifying an appropriate combination of the criteria. The averaged weights and the corresponding rankings for some individual and different combinations of criteria are presented in Table 1.

Table 1. Rankings (based on averaged weights) obtained by using different simple and combined criteria

Criteria	r1 w1	r2 w2	r3 w3	r4 w4	r5 w5	r6 w6	r7 w7	r8 w8	r9 w9	r10 w10
Intra-class dissimilarity (C_1)	7 0.9	8 0.7	1 0.2	2 0	10 0	5 0	4 0	9 0	6 0	3 0
Inter-class dissimilarity (C_2)	9 1	6 1	5 1	4 1	3 1	2 1	1 1	10 1	8 0.79	7 0.57
Attribute to class correlation (C_3)	1 0.99	5 0	2 0	6 0	3 0	10 0	4 0	9 0	8 0	7 0
Entropy (C_4)	1 0.99	8 0.84	7 0.55	6 0	2 0	9 0	5 0	3 0	10 0	4 0
(C_1,-C_2)	1 0.99	4 0.77	6 0.67	5 0.35	2 0.33	8 0.17	3 0.16	7 0.09	9 0.004	10 0.002
(C_1,-C_2,-C_3)	1 0.99	6 0.67	4 0.58	3 0.23	5 0.21	2 0.20	7 0.08	8 0.08	9 0.003	10 0.002
(-C_2,-C_4)	1 0.99	8 0.91	6 0.68	4 0.45	7 0.33	5 0.26	2 0.18	3 0.13	9 0.006	10 0.0007

The rankings based on just one criterion do not reflect the true relevance of features while the combination (C_1,-C_2,-C_3) is the most appropriate for the analyzed dataset. Table 2 presents comparative results obtained by applying some classical feature selectors implemented in the data mining toolkit Weka (www.cs.waikato.ac.nz/~ml/weka/) and the evolutionary weighting. Although the rankings were different, by aggregating them we obtained a generic ranking which suggested that there were two main subsets of features (the first six were relevant, while the last four were irrelevant), which was in accordance with the way the data had been generated.

Table 2. Rankings obtained by using some classical feature selectors and the evolutionary weighting

Method	r1	r2	r3	r4	r5	r6	r7	r8	r9	r10
Chi-squared, Informational Gain	4	5	1	3	2	6	9	10	7	8
ReliefF	1	6	4	5	3	2	9	8	7	10
Evolutionary weighting	1	6	4	3	2	5	7	8	9	10
Aggregated generic ranking	{1,2,3,4,5,6}			{7,8,9,10)						

In order to analyze the usefulness of the proposed approach in the context of distributed data mining we applied the technique to five independent datasets generated according to the same rules. The results obtained by applying the evolutionary weighting to all of them and the aggregated generic ranking are presented in Table 3. The aggregated generic ranking corresponds to a partition which accurately reflects the particularities of the generated data. Subsets like {7,8,9,10} suggest that the corresponding features should be treated similarly, e.g. if one feature is kept or eliminated, then the same should happen to all the other features in the same subset. This simple example illustrates the possibility of applying the same aggregation technique in the case of feature ranking starting from different data sets.

Table 3. Rankings obtained by applying the evolutionary weighting to five data samples from the same population

Dataset	r1	r2	r3	r4	r5	r6	r7	r8	r9	r10
Dataset1	1	6	4	3	2	5	7	8	9	10
Dataset2	1	5	2	6	4	3	7	8	10	9
Dataset3	1	4	2	5	6	3	8	7	9	10
Dataset4	1	6	4	2	5	3	8	7	10	9
Dataset5	1	5	4	2	3	6	7	8	9	10
Aggregated generic ranking			{1}	{2,3,4,5,6}			{7,8}		{9,10}	

If the aggregated ranking consists of only one set of features, this means that the local rankings are not compatible and suggests that there are statistical differences between data sets.

The final goal of these investigations is to develop a system for risk prediction in obstetrics. Thus we applied the proposed technique for some real data that have been collected in the framework of a collaborative project involving hospitals and clinics of obstetrics and gynaecology. The preliminary dataset we analyzed consisted of 211 instances, each containing 109 binary features related to the presence/absence of some symptoms, and medical procedures concerning mothers and their newly born babies, with the aim of identifying features related to the pre-term birth pathology. By aggregating the rankings obtained through independent runs of the evolutionary algorithm (starting from randomly initialized weights), we obtained a first set of seven relevant features which were plausible from the medical point of view.

4. CONCLUSIONS AND OPEN PROBLEMS

The weighting method based on a multiobjective optimization evolutionary algorithm combines the advantages of feature subset selection and of feature ranking. Moreover, it can be applied both in the case of supervised and in that of unsupervised classification tasks by adequately choosing the optimization criteria. In the context of a distributed data mining task, the concept of generic ranking and the technique allowing the aggregation of several rankings in a generic one can be applied both when different ranking methods are applied to the same dataset and when the same method is applied to different datasets. The main limitation of the proposed approach is related with the computational cost of the evolutionary algorithm. Another open problem is that related to the ability of the proposed weighting method to deal with inconsistencies which are so frequently encountered in medical data. Further work will address the integration of this pre-processing technique into a system for risk prediction in obstetrics.

ACKNOWLEDGEMENT

This work is supported by Romanian grant 99-II CEEX 03 - INFOSOC 4091/31.07.2006.

REFERENCES

Deb, K. et al, 2000. A Fast Elitist Non-dominated Sorting Genetic Algorithm for Multi-objective Optimization: NSGA-II, in M. Schonauer et al (Eds.), *Proceedings of PPSN 2000*, *Lecture Notes in Computer Science*, vol. 1917, pp. 849-858.

Handl, J. and Knowles, J., 2006. Feature Subset Selection in Unsupervised Learning via Multiobjective Optimization, *International Journal of Computational Intelligence Research*, vol.2, no. 3, pp. 217-238.

Jong. K. et al, 2004, Ensemble Learning with Evolutionary Computation: Application to Feature Ranking. In *Proc. PPSN VIII* , Eds. X.Yao et al., Springer, *Lecture Notes in Computer Science,* 3242, pp. 1133-1142.

Ruiz, R. et al, 2005. Heuristic Search over a Ranking for Feature Selection. In J. Cabestany et al (Eds.): *IWANN 2005*, *Lecture Notes in Computer Science,*, vol. 3512, pp. 742-749.

Pappa, G.L. et al, 2002. Attribute selection with a multiobjective genetic algorithm, *LNCS,* vol. 2507, pp. 280-290.

Scherf, M. and Brauer, W., 1997. Feature Selection by Means of a Feature Weighting Approach, *Techn. Report FKI-221-97*, Institut for Informatik, TU Munchen.

Wang, L. and Fu, X., 2005. *Data Mining with Computational Intelligence*, Springer, Berlin, Germany.