Feature Selection Based on Information Theory Filters.

Włodzisław Duch[1], Jacek Biesiada[2], Tomasz Winiarski[1], Karol Grudziński[1], and Krzysztof Grąbczewski[1]

[1] Department of Informatics, Nicholas Copernicus University, Grudziądzka 5, 87-100 Toruń, Poland. http://www.phys.uni.torun.pl/kmk
[2] The Silesian University of Technology, Department of Electrotechnology, Division of Computer Methods, ul. Krasińskiego 8, 40-019 Katowice, Poland.

Abstract. Feature selection is an essential component in all data mining applications. Ranking of futures was made by several inexpensive methods based on information theory. Accuracy of neural, similarity based and decision tree classifiers calculated with reduced number of features. Comparison with computationally more expensive feature elimination methods was made.

1 Introduction

Recent data mining applications in bioinformatics, chemistry and commercial domains are very challenging. In case of bioinformatics a very large ($\sim 10^4 - 10^5$)) number of features are associated with activity of genes (up to 30.000 in humans and even more in some plants), while properties of proteins may be described by more than 100.000 features. All these features may be important for some problems, but for a given task only a small subset of these features is relevant. In commercial applications the situation is similar. Therefore computationally inexpensive methods of filtering features are urgently needed.

Methods of feature selection may be divided into three broad categories. Methods that require evaluation of each potentially useful subset of features by a classifier, called "wrapper methods" [5]. This name is also used for a large class of methods for parameter adaptation. Wrapper methods treat computational intelligence (CI) algorithms as black boxes with some parameters to be determined on the basis of test runs. Finding subsets of features is equivalent to assigning binary weights to inputs. The problem is NP-complete since the number of all subsets 2^N grows exponentially with the number of features N. Second group of feature selection methods is based on specific properties of CI methods. Neural networks allow to compute gradients in respect to internal parameters, allowing for feature selection methods based on regularization techniques that may be more efficient than the blind wrapper approach. The third group of methods is based on evaluation of individual features in respect to the task performed, or filtering features that potentially carry useful information, independently from the final CI method used.

Correlation coefficients and other methods for evaluation of the usefulness of features are usually variants of information-theoretical approaches. Computational

complexity of this approach may be much lower than in the case of wrapper approach, depending on the degree to which interaction among features is taken into account. Several filter methods based on information theory are compared here with k-nearest neighbor (kNN) wrapper method. Two question we attempt to answer are: are filters competitive to computationally more demanding wrapper methods, and are all filters equally good for all methods? CI systems that have been selected include neural network, decision tree and kNN method. First a few information theory filters are described and then results of experimental simulations presented.

2 Filters and wrappers

The simplest wrapper approach based on successive elimination (or addition) of features, leaving those features that lead to highest accuracy, is quite effective [1]. It is equivalent to the best-first search, requiring $N(N-1)/2$ evaluations using CI algorithm on the training set. Obvious variants of the simplest approach include evaluation of results by crossvalidation, and using beam search to avoid local minima.

Ranking methods based on information theory filters evaluate single features, neglecting possible interactions. Setiono [4] has used the concept of normalized information gains $G_i^{'}$ for feature f_i. These information gains can be calculated in the following way. First, information contained in the whole training set is:

$$I(S) = -\sum_{j=1}^{K} p(C_j)\log_2 p(C_j) \tag{1}$$

where $p(C_j) = n_j/n$ is the fraction of samples $\mathbf{X}$ from class C_j, $j = 1..K$. Continuous features are discretized to compute information associated with a single feature. Let n_{ik} be the number of samples for which features f_i takes a value inside the interval $r_k(f_i)$ and n_{ikj} be the number of such samples $\mathbf{X}$ for which $\mathbf{X} \in C_j$. Information contained in the subset S_{ik} of samples with f_i in the interval $r_k(f_i)$ is:

$$I(S_{ik}) = -\sum_{j=1}^{K} p_{ikj}\log_2 p_{ikj}; \quad p_{ikj} = n_{ikj}/n_{ik} \tag{2}$$

Summing (or integrating) $I(S_{ik})$ over all M intervals information E_i contained in all subsets of feature f_i is computed; this information may also be computed directly

$$E_i = \sum_{k=1}^{M} p_{ik} I(S_{ik}); \quad I_i = -\sum_{k=1}^{M} p_{ik}\log_2 p_{ik}; \quad p_{ik} = \frac{n_{ik}}{n} \tag{3}$$

Information gain and normalized information gain is

$$G_i = I(S) - E_i; \quad G_i^{'} = G_i/I_i \tag{4}$$

A feature is more important if its normalized information gain is larger. This method, referred to as “info gain” or IG, treats all features as independent.

Another method, based on mutual information, includes possible interactions between features has been proposed by Battiti [2] and is called BA in the rest of the paper. Let $p(r_k(f)) = p(f \in r_k(f))$ be the probability of finding samples with feature f in the interval $r_k(f)$. Mutual information between two features f,s is defined as:

$$I(f,s) = \sum_{k=1}^{N}\sum_{j=1}^{N} p(r_k(f) \wedge r_j(s)) \cdot \log_2 \frac{p(r_k(f) \wedge r_j(s))}{\cdot p(r_j(s))\, p(r_k(f))} \tag{5}$$

The mutual information between feature f and the set of classes is:

$$I(C,f) = \sum_{i=1}^{K}\sum_{k=1}^{M} p(C_i \wedge r_k(f)) \cdot \log_2 \frac{p(C_i \wedge r_k(f))}{p(C_i) \cdot p(r_k(f))} \tag{6}$$

where $r_1(f), r_2(f), \ldots, r_N(f)$ is a partition of the range of f values into equal intervals and $p(C_i \wedge r_k(f))$ is the probability that vector X from class C_i has feature f in the interval r_k. The sum runs over all these intervals and all the classes. The recommended number of intervals is usually between 16 and 32 [2]. $M = 24$ was used in our experiments, but results may differ significantly for small numbers of intervals. The algorithm for finding the best subset of k features goes as follows:

1. Set F to the whole set of N features and set S to an empty set.
2. Compute the mutual information $I(C,f)$ for every feature $f \in F$ and the set of classes $C = \{C_1, \ldots C_K\}$.
3. Find the feature f that maximizes $I(C,f)$. Move f from the set F to the set in S, $S = S \cap f, F = F - f$.
4. Repeat the following until set S will have k features:
   (a) Compute the mutual information $F(f,s)$ between features $f \in F$ and $s \in S$.
   (b) Choose g as the feature that maximizes $I(C,f) - \beta \sum_{s \in S} I(f,s)$, where β is a parameter in the interval [0.5,1] (as recommended in [2]).
   (c) Move g from F to S.

Small values stress the importance of high mutual information between the feature and set of classes; large values stress more the mutual information with the features already included in the set S.

Pruning decision trees allows to rank features, with the most important features near the root of the tree, and the least important pruned first. In particular the Separability Split Value (SSV) criterion [7] selects features that give the largest gain in separability test, hence it may be used for feature ranking.

3 Numerical experiments

Two datasets were used in numerical experiments: hepatobiliary disorders and hypothyroid problems (these datasets are described in [6]). Feature ranking was determined on the training set using the IG (4), BA for 6 β values (6) and SSV algorithms.

Using these rankings the kNN [1], the IncNet neural network [3], the Feature Space Mapping (FSM) neurofuzzy system [8], and the K* classification method [9] were used to calculate accuracy on the test set, adding consecutively the most important features in a given ranking to the pool of features used. Since results for different β values do not differ significantly accuracy curves have been ploted only for $\beta = 0.5$.

The **hepatobiliary disorders** dataset contains 4 classes, 9 continuous and one binary feature. 373 cases were used for training and feature selection, 163 cases as the test data. This dataset is noisy, has strongly overlapping classes and is rather difficult to classify [6]. The kNN ranking (Table 3) has been done by dropping features [1] on standardized data with k=1, and Minkowski distance function with exp = 0.6, optimized for all features. Some differences in ranking are found, but feature 3 and 4 are in all rankings among the most important while features 8 and 9 are the least significant.

Method	**Most – Least Important**							
IG	3	1	8	4	6	7	5	2	9
SSV best-first	3	7	4	5	1	2	6	8	9
BA $\beta = 0.5 - 0.6$	4	3	7	1	5	2	9	8	6
BA $\beta = 0.7$	4	3	7	1	2	9	5	8	6
BA $\beta = 0.8 - 1.0$	4	3	1	7	2	9	5	8	6
SBL Ranking	1	4	6	7	5	8	9	2	3

Table 1. Results of feature ranking on the heapatobiliary disorders data; see description in text.

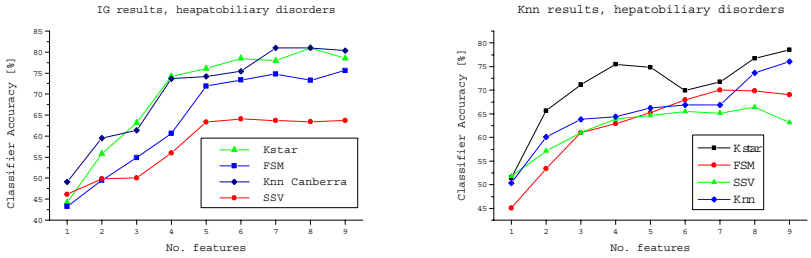

Fig. 1. Left figure: feature selection using normalized information gains; right figure: SBL ranking, Minkowski $exp = 0.6$, $k = 1$.

The same approach was used for the **hypothyroid dataset** (see description in [6]). A total of 3772 cases are used for training (results from one year) and 3428 cases for testing (results from the next year). There are 3 classes but one dominates, with almost 93% of cases. The IG and the SSV with best first search, both compu-

tationally the least expensive methods, have correctly identified the most important features. Interaction of features included in BA has missed important features 18-20, treating them as least important. This has degraded results of all classification methods completely (Fig. 2).

Method	**Most – Least Important**																				
Info gain	17	21	19	18	3	7	13	10	8	15	6	16	5	4	20	12	1	2	11	9	14
SSV best-first	17	21	3	19	18	8	1	20	12	13	15	16	14	11	10	9	7	6	5	3	2
Info interacting $\beta = 0.5$	21	17	13	7	15	12	9	5	8	4	6	16	10	14	2	11	3	18	1	20	19
Info interacting $\beta = 0.7 - 0.8$	21	17	13	15	5	12	9	14	8	4	6	16	7	10	2	11	3	18	1	20	19
Info interacting $\beta = 1.0$	21	17	15	13	5	12	9	14	4	8	6	16	7	10	2	11	3	1	20	18	19
SBL Ranking	17	3	19	8	21	20	18	10	15	13	7	16	9	4	11	5	12	14	6	1	2

Table 2. Results of feature ranking on the hypothyroid dataset; see description in text.

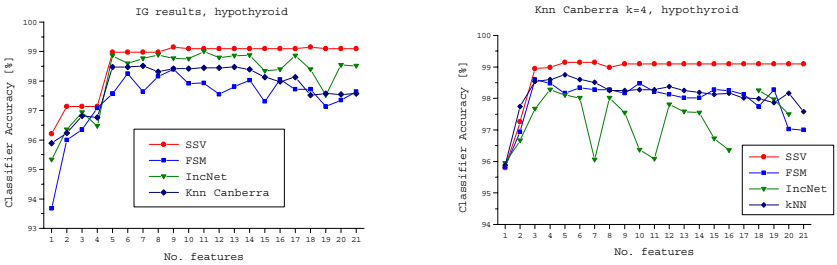

Fig. 2. Hypothyroid data set. Left figure: normalized information gains; right figure: SBL ranking, Canberra, $k = 4$.

4 Conclusions

Several conclusions may be drawn from our study (many results could not be presented in the limited space of this article). Accuracy obtained on wrapper ranking has not been higher than accuracy that may be achieved with filters; in Table 2 features 21 has been placed at the end by the wrapper approach, leading to a serious degradation of performance. Inexpensive filter methods based on information theory may reduce the feature space quite effectively, selecting most important features, but they have several problems. Results may strongly depend on discretization of continuous features and calculation of probabilities. Methods that treat each feature separately are computationally most efficient, but sometimes leave redundant features. Methods that take feature interaction into account are slower, but in principle

should allow to select smaller subsets of important features. Unfortunately (see Fig. 2) such methods may rank low important features since their mutual information with those already selected is high. In multiclass problems, especially with classes that have small percentage of the total number of vectors, feature selection for discrimination of a single class against the rest may give quite different results.

Decision trees allow for independent feature ranking as well as hierarchical ranking including interactions with previous features. This type of methods seem to be competitive with filters based on information theory. After analysis of IG and BA failures we have proposed a new feature selection method that includes feature interaction using a consistency index (in preparation).

Evaluation of the quality of a classifier using overall accuracy only is not sufficient. Much more information is derived from the Receiver Operating Curves [9]. It may also be easier to aggregate (for example by linear combination) several features rather than select them. This is the next step, going beyond feature selection methods. Dependence on the choice of the number of intervals for calculation of information may partially be removed if Gaussian overlapping windows are used instead of intervals.

Acknowledgments: Support by the Polish Committee for Scientific Research, grant 8 T11C 006 19, is gratefully acknowledged.

References

1. Duch W, Grudziński K (1999) The weighted k-NN method with selection of features and its neural realization, 4th Conference on Neural Networks and Their Applications, Zakopane, May 1999, pp. 191-196
2. Battiti R. (1991) Using mutual information for selecting features in supervised neural net learning. IEEE Transaction on Neural Networks **5**, 537-550
3. Jankowski N, Kadirkamanathan V. (1997) Statistical control of RBF-like networks for classification. 7th Int. Conf. on Artificial Neural Networks, Lausanne, Switzerland, Springer Verlag, pp. 385-390
4. Setiono R, Liu H. (1996) Improving Backpropagation learning with feature selection. Applied Intelligence: The International Journal of Artifical Intelligence, *Neural Networks, and Complex Problem-Solving Technologies* **6**, 129-139
5. Kohavi R. (1995) Wrappers for performance enhancement and oblivious decision graphs. PhD thesis, Dept. of Computer Science, Stanford University
6. Duch W, Adamczak R, Grąbczewski K, A new methodology of extraction, optimization and application of crisp and fuzzy logical rules. IEEE Transactions on Neural Networks 12 (2001) 277-306
7. Grąbczewski K. Duch W. (2000) The Separability of Split Value Criterion, *5th Conference on Neural Networks and Soft Computing*, Zakopane, Poland, pp. 201-208
8. Duch W, Diercksen G.H.F. (1995) Feature Space Mapping as a universal adaptive system, Computer Physics Communications **87**, 341–371
9. Witten I.H, Frank E. (2000) Data mining. Morgan Kaufmann Publishers, San Francisco
10. Swets J.A. (1988) Measuring the accuracy of diagnostic systems. Science **240**, 1285-93