Feature Ranking, Selection and Discretization.

Włodzisław Duch[1,2], Tomasz Winiarski[1], Jacek Biesiada[3], and Adam Kachel[3]

[1] Dept. of Informatics, Nicholaus Copernicus University, Toruń, Poland
http://www.phys.uni.torun.pl/kmk
[2] School of Computer Engineering, Nanyang Technological University, Singapore
[3] Division of Computer Methods, Dept. of Electrotechnology, The Silesian University of Technology, Katowice, Poland

Abstract. Many indices for evaluation of features have been considered. Applied to single features they allow for filtering irrelevant attributes. Algorithms for selection of subsets of features also remove redundant features. Hashing techniques enable efficient application of feature relevance indices to selection of feature subsets. A number of such methods have been applied to artificial and real-world data. Strong influence of continuous feature discretization and very good performance of separability-based discretization has been noted.

1 Introduction

Attention is the basic mechanism that the brain is using to select relevant information for further processing. Without initial selection of information the brain would be overflooded with information and could not function. In many applications of computational intelligence methods the situation is similar: large number of features irrelevant to a given task is provided, making the analysis of data very difficult. Many algorithms for filtering information have been devised in the past, either ranking the relative importance of features, or selecting subsets of features [1, 2]. Information theory is most often use as a basis for such methods, but feature relevance indices based on correlation, purity of classes or consistency are also used. Unfortunately relative advantages and weaknesses of these methods are not know.

Feature ranking methods evaluate the relevance of each feature independently, thus leaving potentially redundant features. Feature selection methods search for best subsets of features, offering larger dimensionality reduction. Exhaustive search with performance evaluation on all possible subsets of features provides the golden standard. Although for larger number of features n it is not realistic (the number of all subsets is 2^n), sometimes it may be performed using only the subset of high-ranking features. Finding useful subsets of features is equivalent to assigning binary weights to inputs, so feature and subset selection is a special case of feature transformation. Some classification and approximation methods may benefit from using feature relevance indices as scaling factors. Splitting criteria in decision trees rank features and may be used to define feature relevance indices.

We will consider here only general feature selection methods, independent of any specific properties of classification methods, such as using regularization techniques to trained neural networks [3]) or pruning in decision trees [4]. This paper attempts to elucidate the following questions: is there any significant difference between performance of feature relevance indices; how to convert the feature ranking (filtering) methods into feature selection methods; is greedy search sufficient for finding optimal subsets (comparing to selection based on evaluation of all possible subsets); how strongly different methods of discretization influence ranking and selection. To answer some of these questions a number of feature relevance indices have been programmed, and hashing techniques were implemented to use them for feature selection, evaluating whole subsets of features and thus including interaction of features. Artificial data has been created using 4 Gaussians (one per class) in 8 dimensions, with the last four features being noisy copies of the first four. Results are also compared on real data with the "golden standard", or selection made by evaluation of all subsets.

Feature relevance indices used to estimate the importance of a given feature are presented in the next section. Several feature ranking and feature selection methods based on information theory and other approaches are presented in the third section. In the fourth section empirical tests are made on artificial data for which correct ranking of features is known. Calculations of classification accuracy on the well known hypothyroid dataset are presented in section five. The paper is finished with a number of conclusions.

2 Feature relevance indices

In the simplest case we have two classes $K = 2$ and n binary features $X_i = 0, 1, i = 1 \ldots n$. Ranking of these features is always done independently. For feature X_i the joint probability $p(C_j, X_i)$ is a 2 by 2 matrix that carries full information that may be derived from this feature. The “informed majority classifier" (i.e. knowing the X_i value) makes in this case optimal decisions: if $p(C_0, X_i = 0) > p(C_1, X_i = 0)$ then class C_0 should be chosen for $X_i = 0$, giving a fraction of $p(C_0, X_i = 0)$ correct and $p(C_1, X_i = 0)$ erroneous predictions. The same procedure is used for all other X_i values (for binary feature only $X_i = 1$), leading to the following accuracy of informed majority classifier:

$$IMC(X) = \sum_i \max_j \left(p(C_j, X = x_i) \right) \tag{1}$$

with $x_0 = 0$ and $x_1 = 1$. To account for the effects of class distribution the expected accuracy of the uninformed majority classifier (i.e. the base rate =$\max_i p(C_i)$) should be subtracted from this index but since this is a constant for a given dataset it will not change the ranking. Since no more information is available two features with the same accuracy $IMC(X_a) = IMC(X_b)$ should be ranked as equal.

This reasoning may be extended to a multiclass cases and multivalued discrete features, and since continuous features are usually discretized it may cover all cases. In particular, $IMC(X_i)$ may be used for local feature ranking for each vector $\mathbf{X}$ subject to classification. In this case there is no global ranking or selection of features, and no problem with averaging the importance of features over all data.

The accuracy of the majority classifier is one way of measuring how “pure" are the bins to which $X = x$ feature value falls to (cf. [10] fro more on purity or consistency-based indices). Other feature relevance indices estimate how concentrated the whole distribution is. The *Gini* impurity index used in CART decision trees [5] sums the squares of the class probability distribution for a tree node. Summing squares of all joint probabilities $P(C,X)^2 = \sum_{i,x} p(C_i, X=x)^2$ (where x is a set of values or intervals) gives a measure of probability concentration. For N_x values (or intervals) of X variable $P(C,X)^2 \in [1/KN_x, 1]$ and should be useful for feature ranking.

Association between classes and feature values may also be measured using χ^2 values, as in the CHAID decision tree algorithm [6]. Information theory indices are most frequently used for feature evaluation. Information contained in the joint distribution of classes and features, summed over all classes, gives an estimation of the importance of the feature:

$$\begin{aligned} I(C,X) &= -\sum_{i=1}^{K} \int p(C_i, X=x) \lg_2 p(C_i, X=x) dx \qquad (2)\\ &\approx -\sum_{x} p(X=x) \sum_{i=1}^{K} p(C_i, X=x) \lg_2 p(C_i, X=x) \end{aligned}$$

where $p(C_i, X=x), i=1\dots K$ is the joint probability of finding the feature value $X=x$ for vectors $\mathbf{X}$ that belong to some class C_k (for discretized continuous features $X=x$ means that the value of feature X is in the interval x), and $p(X=x)$ is the probability of finding vectors with feature value $X=x$, or within the interval $X \in x$. Low values of $I(C,X)$ indicate that vectors from single class dominate in some intervals, making the feature more valuable for prediction. Joint information may also be calculated for each discrete value of X or each interval, and weighted by $p(X=x)$ is probability:

$$WI(C,X) = -\sum_{x} p(X=x) I(C, X=x) \qquad (3)$$

To find out how much information is gained by considering feature X information contained in the $p(X=x)$ probability distribution and in the $p(C)$ class distribution should be taken into account. The resulting combination $M_I(C,X) = I(C) + I(X) - I(C,X)$, called “mutual information" or “information gain", is computed using the formula:

$$\begin{aligned} M_I(C,X) = \; & -I(C,X) - \sum_{i=1}^{K} p(C_i) \lg_2 p(C_i) \\ & - \sum_{x} p(X=x) \lg_2 p(X=x) \qquad (4) \end{aligned}$$

Mutual information is equal to the Kullback-Leibler divergence between the joint and the product probability distribution, i.e. $M_I(C,X) = D_{KL}(p(C,X)|p(C)p(X))$. A feature is more important if its mutual information is larger.

Various modifications of the information gain have been considered in the literature on decision trees (cf. [4]), such as the gain ratio $IGR(C,X) = M_I(C,X)/I(X)$, or the Mantaras distance $1 - MI(C,X)/I(C,X)$ (cf. [7]). Another ratio $IGn(C,X) = M_I(C,X)/I(C)$, called also “an asymmetric dependency coefficient", is advocated in [8], but it does not change the ranking since $I(C)$ is a constant for a given database.

Correlation between distributions of classes and feature values is also measured by the entropy distance $D_I(C,X) = 2I(C,X) - I(C) - I(X)$, or by the symmetrical uncertainty coefficient $U(C,X) = 1 - D_I(C,X)/(I(C) + I(X)) \in [0,1]$.

As noted in [5] the splitting criteria do not seem to have much influence on the quality of decision trees. The same phenomenon may be observed in feature selection: that actual choice of feature relevancy index has little influence on feature ranking. There is another, perhaps more important issue here, related to the accuracy of calculation of feature relevancy indices. Features with continuous values are discretized to estimate $p(X = x)$ and $p(C,X = x)$ probabilities. Alternatively, the data may be fitted to a combination of some continuous one-dimensional kernel functions, for example Gaussian functions, and integration may be used instead of summation. Effects of discretization are investigated in the next section.

3 Artificial data: ranking and discretization

To test various ranking and selection methods, and to test accuracy of discretization methods, we have created several sets of artificial data. 4 Gaussian functions with unit dispersions, each one representing a separate class, have been placed in 4-dimensional space. The first Gaussian has center at $(0,0,0,0)$, the second is at $a(1,1/2,1/3,1/4)$, the third at $2a(1,1/2,1/3,1/4)$ and the fourth at $3a(1,1/2,1/3,1/4)$, so in higher dimensions the overlap is high. For $a = 2$ the overlap in the first dimension is already strong, but it is clear that feature ranking should go from X_1 to X_4. To test the ability of different algorithms for dealing with redundant information additional 4 features have been created by taking $X_{i+4} = 2X_i + \varepsilon$, where ε is uniform noise with unit variance. Ideal feature ranking should give the following order: $X_1 \succ X_5 \succ X_2 \succ X_6 \succ X_3 \succ X_7 \succ X_4 \succ X_8$, while ideal feature selection should recognize linear dependences and select new members in the following order: $X_1 \succ X_2 \succ \ldots \succ X_8$. In addition the number of generated points per Gaussian was varied from 50 to 4000 to check statistical effects of smaller number of samples. 1000 points appeared to be sufficient to remove all artifacts due to the statistical sampling.

First naive discretization of the attributes was used, partitioning each into 4, 8, 16, 24, 32 parts with equal width (results for equal number of samples in each bin are very similar). A number of feature ranking algorithms based on different relevance indices has been compared, including the weighted joint information $WI(C,X)$, mutual information (information gain) $MI(C,X)$, information gain ratio $IGR(C,X)$, transinformation matrix with Mahalanobis distance $GD(C,X)$ [11], methods based on purity indices, and some correlation measures.

Ideal ranking was found by $MI(C;X)$, $IGR(C;X)$, $WI(C,X)$, and $GD(C,X)$ feature relevance indices. The $IMC(X)$ index gave correct ordering in all cases except for partitioning into 8 parts, where features X_2 and X_6 reversed (6 is the noisy version of 2). A few new indices were checked on this data and rejected since some errors have been made. Selection for Gaussian distributions is rather easy using any evaluation measure, and this example has been used as a test of our programs.

Feature ranking algorithms may be used for feature selection if multidimensional partitions, instead of one-dimensional intervals, are used to calculate feature relevancy indices. A subset of m features $\mathcal{F} = \{X_i\}, i = 1 \ldots m$ defines m-dimensional subspace and all probabilities are now calculates summing over cuboids obtained as Cartesian products of one-dimensional intervals. The difficulty is that partitioning each feature into k bins k^m multidimensional partitions are created. The number of data vectors is usually much smaller than the number of m-dimensional bins, therefore hashing techniques may be used to evaluate all indices for subsets of features $\mathcal{F}$. Greedy search algorithm has been used, adding always only a single feature that leads to the largest increase of the relevance index of the expanded set.

Feature selection on the 4 Gaussian data in 8 dimensions has proved to be a more challenging task. The order of first four features X_1 to X_4 is much more important than the order of the last four that do not contribute new information, and therefore may be added in random order. Discretization seems to be much more important now; partitions into 4 and 32 parts always led to errors, and partition into 24 parts gave the most accurate result. Information gain ratio $IGR(C;\mathcal{F})$ gave correct ordering for 8, 16, and 24 bin partitions, finding noisy versions of X_3 and X_4 more important than original versions for 32 bins. The $IMC(\mathcal{F})$ index works well for 24 bins, replacing correct features with their noisy versions for other discretizations. Performance of Battiti's algorithm based on pairwise feature interactions ($MI(C;X_i) - \beta MI(X_i;X_j)$, with $\beta = 0.5$) was very similar to $IMC(\mathcal{F})$.

SSV univariate decision tree applied to this data selected X_1, X_2, X_6, X_3 and X_7 as the top-most features, removing all others. Univariate trees are biased against slanted data distributions, and therefore may not provide the best feature selections for this case. Since the relevancy indices computed with different discretization differ on more than a factor of two from each other and results seem to depend strongly on discretization, the use of new discretization scheme based on the Separability Split Value (SSV) criterion [13] has been investigated. In our previous study [12] very good results were obtained for feature ranking using SSV discretization. Results are presented only for the real data case.

4 Experiments on real data

The well-know **hypothyroid dataset** has been used for these experiments, containing results from real medical screening tests for hypothyroid problems [14]. The class distribution is about 93% normal, 5% of primary hypothyroid and 2% of compensated hypothyroid type. The data offers a good mixture of nominal (15) and numerical (6) features. A total of 3772 cases are given for training (results from one year) and 3428 cases for testing (results from the next year). Comparison with results of other classifiers has been provided elsewhere [3], the data is used here only for evaluation of feature selection.

First, the information gain ratio $IGR(C,X)$ index has been computed for all features using various naive discretizations, such as the equi-width partitioning into 4 to 32 bins. This index (and all others) grows with the number of bins, showing the importance of local correlations, but still for the most important feature it has never reached 0.4 (Fig. 1). Using the SSV separability criterion to discretize features into the same number of bins, mutual information values that were twice as large have been obtained, with high information content even for 4 or 6 bins. In general this increase leads to much more reliable results, therefore only this discretization has been used further.

Strong correlations between features makes this dataset rather difficult to study. Subsets of features have been generated, analyzing the training set using the mutual information (or normalized information gain [8]), Battiti's information gain with pairwise feature interaction [9], and using other indices presented here for evaluation of selected subsets. Selection of feature subsets has also been provided by SSV decision tree using both best first search and the beam search mode to create the tree [13]. Features that are at the top of the tree are considered to be most important, and SSV performs its own discretization. An additional feature selection has been made with the k nearest neighbor (kNN) method as a wrapper, consecutively dropping the least important feature to determine smaller feature subset, until a single feature is left. This is of course computationally very expensive.

A number of classifiers implemented in the GhostMiner software, developed in our laboratory[4], have been applied to the hypothyroid data, starting from a single feature and adding more in the order indicated by different methods. Here only the Support Vector Machine (SVM) results are reported (as implemented by N. Jankowski, in preparation), but other results lead to similar conclusions. The best discretization should lead to the feature selection methods that achieve higher accuracy with smaller number of features, but the final results depend also on the relevancy index used, and on the search procedure for subset selection.

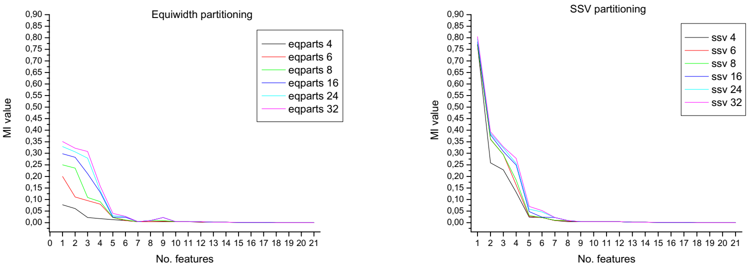

Fig. 1. Values of mutual information for the hypothyroid data. Left figure: with equiwidth partitioning; right figure: with SSV partitioning.

kNN wrapper approach created very good feature subsets, finding 3 to 5 element subsets that led to the highest accuracy. Please note that there is no discretization involved in this calculation. Overall results from kNN selection were slightly worse only for 6-element feature subset, where SSV found a better subset of features {17, 21, 3, 19, 18, 8}. Mutual information (and other indices of relevance, not shown here) with SSV discretization was not able to find 3 and 4-element subsets. kNN results suggest that instead of adding features perhaps dropping them would be a better approach. This can be

[4] http://www.fqspl.com.pl/ghostminer/

Method	Most Important – Least Important																				
kNN	17	3	8	19	21	5	15	7	13	20	12	4	6	9	10	18	16	14	11	1	2
BA $\beta = 0.5 - 06$	17	21	15	13	7	9	5	12	8	6	4	16	14	10	11	2	3	1	18	20	19
IGR	17	21	19	18	3	7	13	10	8	15	6	16	5	20	4	1	12	2	11	9	14
SSV BFS	17	21	3	19	18	8	1	20	12	13	15	16	14	11	10	9	7	6	5	4	2
SSV beam	17	8	21	3	19	7	9	11	10	12	14	15	16	18	20	13	6	5	4	2	1

Table 1. Results of feature selection for the hypothyroid dataset; mutual information has been calculated using SSV discretization.

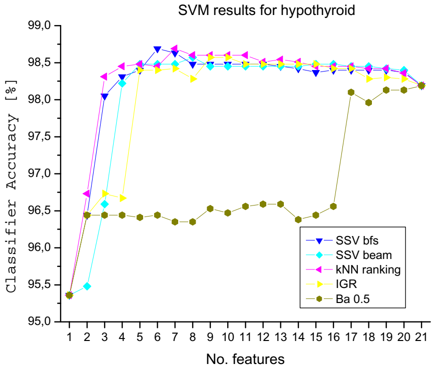

Fig. 2. Accuracy of SVM calculations for hypothyroid data, results obtained on subsets of features created by five methods.

done with our implementation of selection algorithms and is now currently investigated. Finally pairwise feature interaction included in the Battiti method leaves out important features, even for small values of β, indicating that interaction among many features should be taken into account, as it is done with other selection method presented here.

5 Conclusions

Over 20 feature ranking methods based on information theory, correlation and purity indices have been implemented and using hashing techniques applied also to feature selection with greedy search. Only a few results for artificial and real data have been presented here due to the lack of space. Several conclusions may be drawn from this and our more extensive studies:

- The actual feature evaluation index (information, purity or correlation) may not be so important, therefore the least expensive indices (purity indices) should be used.
- Discretization is very important; naive equi-width or equi-distance discretization may give unpredictable results; entropy-based discretization is better but more costly, with the separability-based discretization offering less expensive.
- Selection requires calculation of multidimensional evaluation indices, done effectively using hashing techniques.

- Continuous kernel-based approximations to calculation of feature relevance indices are a useful, although little explored, alternative.
- Ranking is easy if global evaluation of feature relevance is sufficient, but different sets of features may be important for separation of different classes, and some are important in small regions only (cf. decision trees), therefore results on Gaussian distributions may not reflect real life problems.
- Local selection and ranking is the most promising technique.

Many open questions remain: discretization method deserve more extensive comparison, fuzzy partitioning may be quite useful, ranking indices may be used for feature weighting, not only selection, selection methods should be used to find combination of features that contain more information.

References

1. Liu H, Motoda H. (1998) *Feature Extraction, Construction and Selection: A Data Mining Perspective.* Kluwer Academic Publishers.
2. Liu H, Motoda H. (1998) *Feature Selection for Knowledge Discovery and Data Mining.* Kluwer Academic Publishers.
3. Duch W, Adamczak R. and Grąbczewski K. (2001) Methodology of extraction, optimization and application of crisp and fuzzy logical rules. *IEEE Transactions on Neural Networks* **12**: 277-306
4. Quinlan J.R. (1993) *C4.5: Programs for machine learning*. San Mateo, Morgan Kaufman
5. Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA.
6. Kass, G.V. (1980). An exploratory technique for investigating large quantities of categorical data. Applied Statistics **29**:119 – 127.
7. de Mantaras L.R. (1991) A distance-based attribute selection measure for decision tree induction. Machine Learning **6**, 81-92.
8. Sridhar D.V, Bartlett E.B, Seagrave R.C. (1998) Information theoretic subset selection. Computers in Chemical Engineering **22**, 613-626.
9. Battiti R. (1991) Using mutual information for selecting features in supervised neural net learning. IEEE Transaction on Neural Networks **5**, 537-550.
10. Hall M.A. (1998) Correlation based feature selection for machine learning. PhD thesis, Dept. of Comp. Science, Univ. of Waikato, Hamilton, New Zealand.
11. Lorenzo J, Hernández M. and Méndez J. (1998) GD: A Measure Based on Information Theory for Attribute Selection. Lecture Notes in Artificial Intelligence 1484:124-135.
12. Duch W, Winiarski T, Grąbczewski K, Biesiada J, Kachel, A. (2002) Feature selection based on information theory, consistency and separability indices. Int. Conf. on Neural Information Processing (ICONIP), Singapore, Vol. IV, pp. 1951-1955.
13. Grąbczewski K, Duch W (2000) The Separability of Split Value Criterion, *5th Conf. on Neural Networks and Soft Computing*, Zakopane, Poland, pp. 201-208.
14. C.L. Blake, C.J. Merz, UCI Repository of machine learning databases (2001)
http://www.ics.uci.edu/ mlearn/MLRepository.html.
15. Duch W, Grudziński K. (1999) The weighted k-NN method with selection of features and its neural realization. 4th Conf. on Neural Networks and Their Applications, Zakopane, May 1999, pp. 191-196.
16. Duch W, Diercksen G.H.F, Feature Space Mapping as a universal adaptive system, Computer Physics Communications **87**, 341–371 (1995)