Filter methods

Włodzisław Duch

Department of Informatics, Nicolaus Copernicus University,
Grudziądzka 5, 87-100 Toruń, Poland, and
Department of Computer Science, School of Computer Engineering,
Nanyang Technological University, Singapore 639798
Google: Duch

1 Introduction to filter methods for feature selection

Feature ranking and feature selection algorithms may roughly be divided into three types. The first type encompasses algorithms that are built into adaptive systems for data analysis (predictors), for example feature selection that is a part of embedded methods (such as neural training algorithms). Algorithms of the second type are wrapped around predictors providing them subsets of features and receiving their feedback (usually accuracy). These wrapper approaches are aimed at improving results of the specific predictors they work with. The third type includes feature selection algorithms that are independent of any predictors, filtering out features that have little chance to be useful in analysis of data. These filter methods are based on performance evaluation metric calculated directly from the data, without direct feedback from predictors that will finally be used on data with reduced number of features. Such algorithms are usually computationally less expensive than those from the first or the second group. This chapter is devoted to filter methods.

The feature filter is a function returning a relevance index $J(\mathcal{S}|\mathcal{D})$ that estimates, given the data $\mathcal{D}$, how relevant a given feature subset $\mathcal{S}$ is for the task Y (usually classification or approximation of the data). Since the data and the task are usually fixed and only the subsets $\mathcal{S}$ vary the relevance index may be written as $J(\mathcal{S})$. In text classification these indices are frequently called “feature selection metrics” [19], although they may not have formal properties required to call them a distance metric. Instead of a simple function (such as a correlation or information content) some algorithmic procedure may be used to estimate the relevance index (such as building of a decision tree or finding nearest neighbors of vectors). This means that also a wrapper or an embedded algorithm may be used to provide relevance estimation to a filter used with another predictor.

Relevance indices may be computed for individual features $X_i, i = 1 \dots N$, providing indices that establish a ranking order $J(X_{i_1}) \leq J(X_{i_2}) \cdots \leq$

$J(X_{i_N})$. Those features which have the lowest ranks are filtered out. For independent features this may be sufficient, but if features are correlated many of important features may be redundant. Moreover, the best pair of features do not have to include a single best one [48, 8]. Ranking does not guarantee that the largest subset of important features will be found. Methods that search for the best subset of features may use filters, wrappers or embedded feature selection algorithms. Search methods are independent of the evaluation of feature subsets by filters, and are a topic of Chapter 5. The focus here is on filters for ranking, with only a few remarks on calculation of relevance indices for subsets of features presented in Sec. 8.

The value of the relevance index should be positively correlated with accuracy of any reasonable predictor trained for a given task Y on the data $\mathcal{D}$ using the feature subset $\mathcal{S}$. This may not always be true for all models, and on theoretical grounds it may be difficult to argue which filter methods are appropriate for a given data analysis model. There is little empirical experience in matching filters with classification or approximation models. Perhaps different types of filters could be matched with different types of predictors but so far no theoretical arguments or strong empirical evidence has been given to support such claim.

Although in the case of filter methods there is no direct dependence of the relevance index on the predictors obviously the thresholds for feature rejection may be set either for relevance indices, or by evaluation of the feature contributions by the final system. Features are ranked by the filter, but how many are finally taken may be determined using the predictor in a "wrapper setting". This "filtrapper" approach is computationally less expensive than the original wrapper approach because the evaluation of the predictor's performance (for example by a cross-validation test) is done only for a few pre-selected feature sets. There are also theoretical arguments showing that this technique is less prone to overfitting than pure wrapper methods [40]. In some data mining applications (for example, analysis of large text corpora with noun phrases as features) even relatively inexpensive filter methods, with costs linear in the number of features, may be prohibitively slow.

Filters, as all other feature selection methods, may be divided into local and global types. Global evaluation of features takes into account all data in a context-free way. Context dependence may include different relevance for different tasks (classes), and different relevance in different areas of the feature space. Local classification methods, for example nearest neighbor methods based on similarity, may benefit more from local feature selection, or from filters that are constructed on demand using only data from the neighborhood of a given vector. Obviously taking too few data samples may lead to large errors in estimations of any feature relevance index and the optimal tradeoff between introduction of context and the reliability of feature evaluation may be difficult to achieve. In any case the use of filter methods for feature selection depends on the actual predictors used for data analysis.

In the next section general issues related to the filter methods are discussed. Section 3 is focused on the correlation based filtering, Sec. 4 on relevance indices based on distances between distributions and Sec. 5 on the information theory. In Section 6 the use of decision trees for ranking as well as feature selection is discussed. Reliability of calculation of different indices and bias in respect to the number of classes and feature values is very important and is treated in Section 7. This is followed by some remarks in Sec. 8 on filters for evaluation of feature redundancy. The last section contains some conclusions.

2 General issues related to filters

What does it mean that the feature is relevant to the given task? *Artificial Intelligence* journal devoted in 1996 a special issue to the notion of relevance (Vol. 97, no. 1–2). The common-sense notion of relevance has been rigorously defined in an axiomatic way (see the review in [4]). Although such definitions may be useful for the design of filter algorithms a more practical approach is followed here. [31] give a simple and intuitive definition of relevance that is sufficient for the purpose of feature selection: a feature X is relevant in the process of distinguishing class $Y = y$ from others if and only if for some values $X = x$ for which $\mathcal{P}(X = x) > 0$ the conditional probability $\mathcal{P}(Y = y|X = x)$ is different than the unconditional probability $\mathcal{P}(Y = y)$. Moreover, a good feature should not be redundant, i.e. it should not be correlated with other features already selected. These ideas may be traced back to the test theory [20] developed for psychological measurements.

The main problem is how to calculate the strength of correlations between features and classes (or more generally, between features and target, or output, values), and between features themselves. The Bayesian point of view is introduced below for the classification problems, and many other approaches to estimation of relevance indices are described in subsequent sections. Some of these approaches may be used directly for regression problems, others may require quantization of continuous outputs into a set of pseudo-classes.

Consider the simplest situation: a binary feature X with values $x = \{0, 1\}$ for a two class $y = \{+, -\}$ problem. For feature X the joint probability $\mathcal{P}(y, x)$ that carries full information about the relevance of this feature is a 2 by 2 matrix. Summing this matrix over classes ("marginalizing", as statisticians say) the values of $\mathcal{P}(x)$ probabilities are obtained, and summing over all feature values x gives *a priori* class probabilities $\mathcal{P}(y)$. Because class probabilities are fixed for a given dataset and they sum to $\mathcal{P}(y = +) + \mathcal{P}(y = -) = 1$ only two elements of the joint probability matrix are independent, for example $\mathcal{P}(y = -, x = 0)$ and $\mathcal{P}(y = +, x = 1)$. For convenience notation $\mathcal{P}(y_i, x_j) = \mathcal{P}(y = i, x = j)$ is used below.

The expected accuracy of the majority classifier (MC) $A_{\text{MC}} = \max_y \mathcal{P}(y)$ is independent of the feature X because MC completely ignores information about feature values. The Bayesian Classifier (BC) makes optimal decisions based on the maximum *a posteriori* probability: if $x = x_0$ then for $\mathcal{P}(y_-, x_0) > \mathcal{P}(y_+, x_0)$ class y_- should always be selected, giving a larger fraction $\mathcal{P}(y_-, x_0)$ of correct predictions, and smaller fraction $\mathcal{P}(y_+, x_0)$ of errors. This is equivalent to the *Maximum-a-Posteriori* (MAP) rule: given $X = x$ select class that has greater posterior probability $\mathcal{P}(y|x) = \mathcal{P}(y, x)/\mathcal{P}(x)$. The Bayes error is given by the average accuracy of the MAP Bayesian Classifier (BC). For a single feature, the Bayes error is given by:

$$A_{\mathrm{BC}}(X) = \sum_{j=0,1} \max_i \mathcal{P}(y_i, x_j) = \sum_{j=0,1} \max_i \mathcal{P}(x_j|y_i)\mathcal{P}(y_i). \qquad (1)$$

Precise calculation of “real” joint probabilities $\mathcal{P}(y_i, x_j)$ or the conditional probabilities $\mathcal{P}(x_j|y_i)$ using observed frequencies require an infinite amount of the training data, therefore such Bayesian formulas are strictly true only in the asymptotic sense. The training set should be a large, random sample that represents the distribution of data in the whole feature space.

Because $A_{\mathrm{MC}}(X) \leq A_{\mathrm{BC}}(X) \leq 1$, a Bayesian relevance index scaled for convenience to the $[0, 1]$ interval may be taken as:

$$J_{\mathrm{BC}}(X) = (A_{\mathrm{BC}}(X) - A_{\mathrm{MC}}(X))/(1 - A_{\mathrm{MC}}(X)) \in [0, 1]. \qquad (2)$$

The $J_{\mathrm{BC}}(X)$ may also be called “a purity index”, because it indicates how pure are the discretization bins for different feature values (intervals). This index is also called “the misclassifications impurity” index, and is sometimes used to evaluate nodes in decision trees [17].

Two features with the same relevance index $J_{\mathrm{BC}}(X) = J_{\mathrm{BC}}(X')$ may be ranked as equal, although their joint probability distributions $\mathcal{P}(y_i, x_j)$ may significantly differ. Suppose that $\mathcal{P}(y_-) > \mathcal{P}(y_+)$ for some feature X, therefore $A_{\mathrm{MC}}(X) = \mathcal{P}(y_-)$. For all distributions with $\mathcal{P}(y_-, x_0) > \mathcal{P}(y_+, x_0)$ and $\mathcal{P}(y_+, x_1) > \mathcal{P}(y_-, x_1)$ the accuracy of the Bayesian classifier is $A_{\mathrm{BC}}(X) = \mathcal{P}(y_-, x_0)+\mathcal{P}(y_+, x_1)$, and the error is $\mathcal{P}(y_+, x_0)+\mathcal{P}(y_-, x_1) = 1-A_{\mathrm{BC}}(X)$. As long as these equalities and inequalities between joint probabilities hold (and $\mathcal{P}(y_i, x_j) \geq 0$) two of the probabilities may change, for example $\mathcal{P}(y_+, x_1)$ and $\mathcal{P}(y_+, x_0)$, without influencing $A_{\mathrm{BC}}(X)$ and $J_{\mathrm{BC}}(X)$ values. Thus the Bayesian relevance index is not sufficient to uniquely rank features even in the simplest, binary case. In fact most relevance indices cannot do that without additional conditions (see also Sec. 7).

This reasoning may be extended to multi-valued features (or continuous features after discretization [36]), and multi-class problems, leading to probability distributions that give identical J_{BC} values. The expected accuracy of a Bayesian classification rule is only one of several aspects that could be taken into account in assessment of such indices. In the statistical and pattern recognition literature various measures of inaccuracy (error rates, discriminability), imprecision (validity, reliability), inseparability and resemblance (resolution, refinement) are used (see [24, 15] for extended discussion). Knowing the joint

$\mathcal{P}(y, x)$ probabilities and using the MAP Bayesian Classifier rule confusion matrices $\mathcal{F}_{ij} = N(y_i, y_j)/m = M_{ij}/m$ may easily be constructed for each feature, representing the joint probability of predicting sample from class y_i when the true class was y_j:

$$\mathcal{F}\left(\text{true}, \text{predicted}\right) = \frac{1}{m}\begin{bmatrix} M_{++} & M_{+-} \\ M_{-+} & M_{--} \end{bmatrix} = \frac{1}{m}\begin{bmatrix} \text{TP} & \text{FN} \\ \text{FP} & \text{TN} \end{bmatrix} \tag{3}$$

where M_{++} is the number of hits or true positives (TP); M_{--} is the number of hits in the y_- class, or true negatives (TN); M_{-+} is the number of false alarms, or false positives (FP) (for example, healthy people predicted as sick), and M_{+-} is the number of misses, or false negatives (FN) (sick people predicted as healthy), and the number of samples m is the sum of all M_{ij}.

Confusion matrices have only two independent entries because each row has to sum to $\mathcal{F}_{+j} + \mathcal{F}_{-j} = \mathcal{P}(y_j)$, the *a priori* true class probability (estimated as the fraction of all samples that belong to the class y_j). Class accuracies, or conditional probabilities that given a sample from class y it will be really classified as class y are usually taken as the two independent variables. In medical informatics $S_+ = \mathcal{F}_{++}/\mathcal{P}(y_+) = \mathcal{F}(y_+|y_+)$ is called sensitivity or true positive rate (in information retrieval the name recall or detection rate is used), and $S_- = \mathcal{F}_{--}/\mathcal{P}(y_-) = \mathcal{F}(y_-|y_-)$ is called specificity. These diagonal elements of the conditional confusion matrix $\mathcal{F}(y_i|y_i)$ reflect the type of errors that the predictor makes. For example, sensitivity shows how well sick people (class $y = +$) are correctly recognized by classification rule based on some feature (results of a medical test), and specificity shows how well healthy people (class $y = -$) are recognized as healthy by the same test. Generalization to the K-class case is obvious. Standard classifier accuracy is obtained as a trace of the $\mathcal{F}(y_i, y_j)$ matrix, or Acc $= \sum_i \mathcal{F}(y_i|y_i)\mathcal{P}(y_i)$. The arithmetic average of class accuracies $\mathcal{F}(y_i|y_i)$ is called a balanced accuracy

$$\text{Acc}_2 = \frac{1}{K}\sum_{i=1}^{K} \mathcal{F}(y_i|y_i). \tag{4}$$

The Balanced Error Rate BER=1 − Acc$_2$ is particularly useful evaluation measure for unbalanced datasets. For feature ranking, using accuracy-based relevance indices, such as the $A_{\text{BC}}, J_{\text{BC}}$ indices, is equivalent to comparing $\mathcal{F}(y_+, y_+) - \mathcal{F}(y_+, y_-)$ (true positives minus false positives), while using balanced accuracy is equivalent to $\mathcal{F}(y_+|y_+) - \mathcal{F}(y_+|y_-)$ (true positives ratio minus false positives ratio), because terms that are constant for a given data will cancel during comparison. This difference may be rescaled, for example by using [19]:

$$BNS = G^{-1}\left(\mathcal{F}(y_+|y_+)\right) - G^{-1}\left(\mathcal{F}(y_+|y_-)\right) \tag{5}$$

where $G^{-1}(\cdot)$ is the z-score, or the standard inverse cumulative probability function of a normal distribution. This index, called bi-normal separation index, worked particularly well in information retrieval (IR) [19]. Another simple criterion used in this field is called the Odds Ratio:

$$\mathrm{Odds} = \frac{\mathcal{F}(y_+|y_+)\mathcal{F}(y_-|y_-)}{\mathcal{F}(y_+|y_-)\mathcal{F}(y_-|y_+)} = \frac{\mathcal{F}(y_+|y_+)(1-\mathcal{F}(y_-|y_+)}{(1-\mathcal{F}(y_+|y_+))\mathcal{F}(y_-|y_+)} \tag{6}$$

where zero probabilities are replaced by small positive numbers.

Ranking of features may be based on some combination of sensitivity and specificity. The cost of not recognizing a sick person (low sensitivity) may be much higher than the cost of temporary hospitalization (low specificity). Costs of misclassification may also be introduced by giving a factor to specify that $\mathcal{F}_{+-}$ type of errors (false positive) are α times more important than $\mathcal{F}_{-+}$ type of errors (false negative). Thus instead of just summing the number of errors the total misclassification cost is $E(\alpha) = \alpha\mathcal{F}_{-+} + \mathcal{F}_{+-}$. For binary feature values the BC decision rule has no parameters, and costs $E(\alpha)$ are fixed for a given dataset. However, if the $\mathcal{P}(y, x)$ probabilities are calculated by discretization of some continuous variable z so that the binary value $x = \Theta(z - \theta)$ is calculated using a step function Θ, the values of sensitivity $\mathcal{F}(y_+|y_+;\theta)$ and specificity $\mathcal{F}(y_-|y_-;\theta)$ depend on the threshold θ, and the total misclassification cost $E(\alpha, \theta)$ can be optimized with respect to θ.

A popular way to optimize such thresholds (called also "operating points" of classifiers) is to use the receiver operator characteristic (ROC) curves [24, 45]. These curves show points $R(\theta) = (\mathcal{F}(y_+|y_-;\theta),\ \mathcal{F}(y_+|y_+;\theta))$ that represent a tradeoff between the false alarm rate $\mathcal{F}(y_+|y_-;\theta)$ and sensitivity $\mathcal{F}(y_+|y_+;\theta)$ (true positives rate). The Area Under the ROC curve (called AUC) is frequently used as a single parameter characterizing the quality of the classifier [25], and may be used as a relevance index for BC or other classification rules. For a single threshold (binary features) only one point $R = (\mathcal{F}(y_+|y_-), \mathcal{F}(y_+|y_+))$ is defined, and the ROC curve has a line segment connecting it with points $(0, 0)$ and $(1, 1)$. In this case AUC= $\frac{1}{2}(\mathcal{F}(y_+|y_+) + \mathcal{F}(y_-|y_-))$ is simply equal to the balanced accuracy Acc_2, ranking as identical all features that have the same difference between true positive and false positive ratios. In general this will not be the case and comparison of AUCs may give a unique ranking of features. In some applications (for example, in information retrieval) classifiers may have to work at different operating points, depending on the resources that may change with time. Optimization of ROC curves from the point of view of feature selection leads to filtering methods that may be appropriate for different operating conditions [7].

A number of relevance indices based on modified Bayesian rules may be constructed, facilitating feature selection not only from the accuracy, but also from the cost or confidence point of view. The confusion matrix $\mathcal{F}(y_1, y_2)$ for the two-class problems may be used to derive various combinations of accuracy and error terms, such as the harmonic mean of recall and precision called the $F1$-measure,

$$J_F(X) = 2\mathcal{F}_{++}/(1 + \mathcal{F}_{++} - \mathcal{F}_{--}), \tag{7}$$

well-justified in information retrieval [50]. Selection of the AUC or balanced accuracy instead of the standard accuracy corresponds to a selection of the relative cost factor $\alpha = \mathcal{P}(y_-)/\mathcal{P}(y_+)$ [15]. An index combining the accuracy and the error term $J(\gamma) = \mathcal{F}_{--}+\mathcal{F}_{++}-\gamma(\mathcal{F}_{-+}+\mathcal{F}_{+-}) = A-\gamma E$ does not favor one type of errors over another, but it may be used to optimize confidence and rejection rates of logical rules [14]. For $\gamma = 0$ this leads to the A_{BC} Bayesian accuracy index, but for large γ a classification rule that maximizes $J(\gamma)$ may reduce errors increasing confidence in the rule at the expense of leaving some samples unclassified. Non-zero rejection rates are introduced if only significant differences between the $\mathcal{P}(y, x)$ values for different classes are kept, for example the feature is may be rejected if $|\mathcal{P}(y_+, x) - \mathcal{P}(y_-, x)| < \theta$ for all values of x.

From the Bayesian perspective one cannot improve the result of the maximum *a posteriori rule*, so why is the $J_{\mathrm{BC}}(X)$ index rarely (if ever) used, and why are other relevance indices used instead? There are numerous theoretical results [12, 2] showing that for any method of probability density estimations from finite samples convergence may be very slow and no Bayes error estimate can be trusted. The reliability of $\mathcal{P}(y, x)$ estimates rapidly decreases with a growing number of distinct feature values (or continuous values), growing number of classes, and decreasing number of training samples per class or per feature value. Two features with the same $J_{\mathrm{BC}}(X)$ index may have rather different distributions, but the one with lower entropy may be preferred. Therefore methods that compare distributions of feature and class values may have some advantages [47]. An empirical study of simple relevance indices for text classification shows [19] that accuracy is rather a poor choice, with balanced accuracy (equivalent to comparison of AUCs for the two-class problems) giving much higher recall at similar precision. This is not surprising remembering that in the applications to text classification the number of classes is high and the data are usually very unbalanced ($\mathcal{P}(y_+)$ is very small).

Distribution similarity may be estimated using various distance measures, information theory, correlation (dependency) coefficients and consistency measures, discussed in the sections below. Some theoretical results relating various measures to the expected errors of the Bayesian Classifier have been derived [51, 49] but theoretical approaches have met only with limited success and empirical comparisons are still missing. Features with continuous values should be discretized to estimate probabilities needed to compute the relevance indices [37, 36]. Alternatively, the data may be fitted to a combination of some continuous one-dimensional kernel functions (Gaussian functions are frequently used), and integration may be used instead of summation.

The relevance indices $J(X)$ introduced above are global or context-free, evaluating the average usefulness of a single feature X. This may be sufficient in many applications, but for some data distributions and for complex domains features may be highly relevant in one area of the feature space and not relevant at all in some other area. Some feature selection algorithms (such as Relief described below) use local information to calculate global, averaged indices. Decision trees and other classification algorithms that use the "divide and conquer" approach hierarchically partitioning the whole feature space, need different subsets of features at different stages. Restricting calculations to the neighborhood $O(\mathbf{x})$ of some input vector $\mathbf{x}$, local or context-dependent, relevance indices $J(X, O(\mathbf{x}))$ are computed.

In multiclass problems or in regression problems features that are important for specific target values ("local" in the output space) should be recognized. For example, if the data is strongly unbalanced, features that are important for discrimination of the classes with small number of samples may be missed. In this case the simplest solution is to apply filters to multiple two-class problems. In case of regression problems filters may be applied to samples that give target values in a specific range.

3 Correlation-based filters

Correlation coefficients are perhaps the simplest approach to feature relevance measurements. In contrast with information theoretic and decision tree approaches they avoid problems with probability density estimation and discretization of continuous features and therefore are treated first.

In statistics "contingency tables" defined for pairs of nominal features X, Y are frequently analyzed to determine correlations between variables. They contain the numbers of times $M_{ij} = N(y_i, x_j)$ objects with feature values $Y = y_j, X = x_i$ appear in a database. In feature selection m training samples may be divided into subsets of M_{ij} samples that belong to class $y_i, i = 1 \dots K$ and have a specific feature value x_j; summing over rows of the M_{ij} matrix marginal distribution $M_{i\cdot}$ of samples over classes is obtained, and summing over columns distribution $M_{\cdot j}$ of samples over distinct feature values x_j is obtained. The strength of association between variables X, Y is usually measured using χ^2 statistics:

$$\chi^2 = \sum_{ij} (M_{ij} - m_{ij})^2 / m_{ij}, \text{ where } m_{ij} = M_{i\cdot} M_{\cdot j} / m, \tag{8}$$

Here m_{ij} represent the expected number of observations assuming X, Y independence. Terms with $m_{ij} = 0$ should obviously be avoided (using sufficient data to have non-zero counts for the number of samples in each class and each feature value), or replaced by a small number. If feature and target values were completely independent $m_{ij} = M_{ij}$ would be expected, thus large differences show strong dependence. To estimate the significance of the χ^2 test an incomplete gamma function $Q(\chi^2|\nu)$ is used [41]. The number of degrees of freedom ν is set to $K - 1$. This approach is justified from the statistical point of view only if the number of classes or the number of feature values are large. In contrast to the Bayesian indices the χ^2 results depend not only on the joint probabilities $\mathcal{P}(x_i, y_j) = N(x_i, y_j)/m$, but also on the number of samples m, implicitly including the intuition that estimation of probabilities from small samples is not accurate and thus the significance of small correlations is rather low. χ^2 statistics have been used in several discretization methods combined with feature selection [37, 36].

The linear correlation coefficient of Pearson is very popular in statistics [41]. For feature X with values x and classes Y with values y treated as random variables it is defined as:

$$\varrho(X,Y) = \frac{E(XY) - E(X)E(Y)}{\sqrt{\sigma^2(X)\sigma^2(Y)}} = \frac{\sum_i (x_i - \bar{x}_i)(y_i - \bar{y}_i)}{\sqrt{\sum_i (x_i - \bar{x}_i)^2 \sum_j (y_i - \bar{y}_i)^2}}. \tag{9}$$

$\varrho(X,Y)$ is equal to ± 1 if X and Y are linearly dependent and zero if they are completely uncorrelated. Some features may be correlated positively, and some negatively. Linear coefficient works well as long as the relation between feature values and target values is monotonic. Separation of the means of the class distributions leads to an even simpler criterion:

$$\mu(X,Y) = \frac{\mu(y_+) - \mu(y_-)}{(\sigma(y_+) + \sigma(y_-))}, \tag{10}$$

where $\mu(y_+)$ is the mean value for class y_+ vectors and $\sigma(y_+)$ is the variance for this class. For continuous targets a threshold $y < \theta$ divides vectors into y_+ and y_- groups. The square of this coefficient is similar to the ratio of between-class to within-class variances, known as the Fisher criterion [17]. The T-test uses slightly different denominator [41]:

$$T(X,Y) = \frac{\mu(y_+) - \mu(y_-)}{\sqrt{\sigma(y_+)^2/m_+ + \sigma(y_-)^2/m_-}}, \tag{11}$$

where $m_\pm$ is the number of samples in class $y_\pm$. For ranking absolute values $|\varrho(X,Y)|$, $|\mu(X,Y)|$ and $|T(X,Y)|$ are taken. How significant are differences in $\varrho(X,Y)$ and other index values? The simplest test estimating the probability that the two variables are correlated is:

$$\mathcal{P}(X \sim Y) = \mathrm{erf}\left(|\varrho(X,Y)|\sqrt{m/2}\right), \tag{12}$$

where erf is the error function. Thus for $m = 1000$ samples linear correlations coefficients as small as 0.02 lead to probabilities of correlation around 0.5. This estimation may be improved if the joint probability of X, Y variables is binormal. The feature list ordered by decreasing values (descending order) of the $\mathcal{P}(X \sim Y)$ may serve as feature ranking. A similar approach is also taken with χ^2, but the problem in both cases is that for larger values of χ^2 or correlation coefficient, probability $\mathcal{P}(X \sim Y)$ is so close to 1 that ranking becomes impossible due to the finite numerical accuracy of computations. Therefore an initial threshold for $\mathcal{P}(X \sim Y)$ may be used in ranking only to determine how many features are worth keeping, although more reliable estimations may be done using cross-validation or wrapper approaches. An alternative is to use a permutation test, computationally expensive but improving accuracy for small number of samples [9] (see also Neal and Zhang, this volume).

If a group of k features has already been selected, correlation coefficients may be used to estimate correlation between this group and the class, including inter-correlations between the features. Relevance of a group of features grows with the correlation between features and classes, and decreases with growing inter-correlation. These ideas have been discussed in theory of psychological measurements [20] and in the literature on decision making and aggregating opinions [26]. Denoting the average correlation coefficient between these features and the output variables as $r_{ky} = \bar{\varrho}(\mathbf{X}_k, Y)$ and the average between different features as $r_{kk} = \bar{\varrho}(\mathbf{X}_k, \mathbf{X}_k)$ the group correlation coefficient measuring the relevance of the feature subset may be defined as:

$$J(\mathbf{X}_k, Y) = \frac{k r_{ky}}{\sqrt{k + (k-1) r_{kk}}}. \tag{13}$$

This formula is obtained from Pearson's correlation coefficient with all variables standardized. It has been used in the Correlation-based Feature Selection (CFS) algorithm [23] adding (forward selection) or deleting (backward selection) one feature at a time.

Non-parametric, or Spearman's rank correlation coefficients may be useful for ordinal data types. Other statistical tests of independence that could be used to define relevance indices, such as the Kolmogorov-Smirnov test based on cumulative distributions and G-statistics [41].

A family of algorithms called Relief [43] are based on the feature weighting, estimating how well the value of a given feature helps to distinguish between instances that are near to each other. For a randomly selected sample $\mathbf{x}$ two nearest neighbors, $\mathbf{x}_s$ from the same class, and $\mathbf{x}_d$ from a different class, are found. The feature weight, or the Relief relevance index $J_R(X)$ for the feature X, is increased by a small amount proportional to the difference $|X(\mathbf{x}) - X(\mathbf{x}_d)|$ because relevance should grow for features that separate vectors from different classes, and is decreased by a small amount proportional to $|X(\mathbf{x}) - X(\mathbf{x}_s)|$ because relevance should decrease for feature values that are different from features of nearby vectors from the same class. Thus $J_R(X) \leftarrow J_R(X) + \eta(|X(\mathbf{x}) - X(\mathbf{x}_d)| - |X(\mathbf{x}) - X(\mathbf{x}_s)|)$, where η is of the order of $1/m$. After a large number of iterations this index captures local correlations between feature values and their ability to help in discrimination of vectors from different classes. Variants include ratio of the average over all examples of the distance to the nearest miss and the average distance to the nearest hit, that self-normalizes the results [22]:

$$J_R(X) = \frac{E_x(|X(\mathbf{x}) - X(\mathbf{x}_d)|)}{E_x(|X(\mathbf{x}) - X(\mathbf{x}_s)|)}. \tag{14}$$

The ReliefF algorithm has been designed for multiclass problems and is based on the k nearest neighbors from the same class, and the same number of vectors from different classes. It is more robust in the presence of noise in the data, and includes an interesting approach to the estimation of the missing values. Relief algorithms represent quite original approach to feature selection, that is not based on evaluation of one-dimensional probability distributions [43]. Finding nearest neighbors assures that the feature weights are context sensitive, but are still global indices (see also [29] for another algorithm of the same type). Removing context sensitivity (which is equivalent to assuming feature independence) makes it possible to provide a rather complex formula for ReliefX:

$$
\begin{aligned}
J_{\mathrm{RX}}(Y,X) &= \frac{GSx}{(1-Sy)Sy}; \quad \text{where} \\
Sx &= \sum_{i=1}^{K} \mathcal{P}(x_i)^2; \quad Sy = \sum_{j=1}^{M_Y} \mathcal{P}(y_j)^2 \qquad (15) \\
G = \sum_j \mathcal{P}(y_j)(1-\mathcal{P}(y_j)) - &\sum_{i=1}^{K} \left(\frac{\mathcal{P}(x_i)^2}{Sx} \sum_j \mathcal{P}(y_j|x_i)(1-\mathcal{P}(y_j|x_i)) \right).
\end{aligned}
$$

The last term is a modified Gini index (Sec. 6). Hall [23] has used a symmetrized version of $J_{\mathrm{RX}}(Y,X)$ index (exchanging x and y and averaging) for evaluation of correlation between pairs of features. Relief has also been combined with a useful technique based on the successive Gram-Schmidt orthogonalization of features to the subset of features already created [22]. Connection to the Modified Value Difference Metric (MVDM) is mentioned in the next section.

4 Relevance indices based on distances between distributions

There are many ways to measure dependence between the features and classes based on evaluating differences between probability distributions. A simple measure – a difference between the joint and the product distributions – has been proposed by Kolmogorov:

$$
D_{\mathrm{K}}(Y,X) = \sum_i \sum_{j=1}^{K} |\mathcal{P}(y_j,x_i) - \mathcal{P}(x_i)\mathcal{P}(y_j)|. \qquad (16)
$$

This is very similar to the χ^2 statistics except that the results do not depend on the number of samples. After replacing summation by integration this formula may be easily applied to continuous features, if probability densities are known or some kernel functions have been fitted to the data. It may reach zero for completely irrelevant features, and it is bounded from above:

$$0 \leq D_{\mathrm{K}}(Y,X) \leq 1 - \sum_i \mathcal{P}(x_i)^2, \tag{17}$$

if the correlation between classes and feature values is perfect. Therefore this index is easily rescaled to the $[0,1]$ interval. For two classes with the same *a priori* probabilities Kolmogorov measure reduces to:

$$D_{\mathrm{K}}(Y,X) = \frac{1}{2}\sum_i |\mathcal{P}(x_i|y=0) - \mathcal{P}(x_i|y=1)|. \tag{18}$$

The expectation value of squared *a posteriori* probabilities is known as the average Euclidean norm of the conditional distribution, called also the Bayesian measure [49]:

$$J_{\mathrm{BM}}(Y,X) = \sum_i \mathcal{P}(x_i) \sum_{j=1}^{K} \mathcal{P}(y_j|x_i)^2, \tag{19}$$

It measures concentration of the conditional probability distribution for different x_i values in the same way as the Gini index (Eq. 39) used in decision trees (Sec. 6).

The Kullback-Leibler divergence:

$$D_{\mathrm{KL}}((\mathcal{P}(X)||(\mathcal{P}(Y)) = \sum_i \mathcal{P}_Y(y_i) \log \frac{\mathcal{P}_Y(y_i)}{P_X(x_i)} \geq 0, \tag{20}$$

is used very frequently, although it is not a distance (it is not symmetric). The KL divergence may be applied to relevance estimation in the same way as the χ^2 statistics:

$$D_{\mathrm{KL}}(\mathcal{P}(X,Y)||\mathcal{P}(X)\mathcal{P}(Y)) = \sum_i \sum_{j=1}^{K} \mathcal{P}(y_j,x_i) \log \frac{\mathcal{P}(y_j,x_i)}{\mathcal{P}(x_i)\mathcal{P}(y_j)}. \tag{21}$$

This quantity is also known as “mutual information” $MI(Y,X)$. The Kullback-Liebler measure is additive for statistically independent features. It is sensitive to the small differences in distribution tails, which may lead to problems, especially in multiclass applications where the relevance index is taken as the average value of KL divergences between all pairs of classes.

The Jeffreys-Matusita distance (JM-distance) provides a more robust criterion:

$$D_{\mathrm{JM}}(Y,X) = \sum_i \sum_{j=1}^{K} \left[\sqrt{\mathcal{P}(y_j,x_i)} - \sqrt{\mathcal{P}(x_i)\mathcal{P}(y_j)}\right]^2. \tag{22}$$

For Gaussian distributions D_{JM} is related to the Bhattacharya distance. Because $D_{JM} \leq 2(1-\exp(-D_{KL}/8))$ an exponential transformation $J_{KL} = 1-\exp(-D_{KL}/8)$ is sometimes defined, reaching zero for irrelevant features and growing to 1 for a very large divergences, or highly relevant features.

There is some evidence that these distances are quite effective in remote sensing applications [6].

The Vajda entropy is defined as [49]:

$$J_{\mathrm{V}}(Y,X) = \sum_i \mathcal{P}(x_i) \sum_{j=1}^{K} \mathcal{P}(y_j|x_i)(1-\mathcal{P}(y_j|x_i)), \tag{23}$$

and is simply equal to the $J_{\mathrm{V}}(Y,X) = 1 - J_{\mathrm{BM}}(Y,X)$. The error rate of the Bayesian Classifier is bounded by the Vajda entropy, $A_{\mathrm{BC}}(X) \leq J_{\mathrm{V}}(Y,X)$. Although many other ways to compare distributions may be devised they may serve as better relevance indicators only if tighter error bounds could be established.

In the memory-based reasoning the distance between two vectors X, X' with discrete elements (nominal or discretized), in a K class problem, is computed using conditional probabilities [52]:

$$VDM(X,X';Y)^2 = \sum_i \sum_{j=1}^{K} \left|\mathcal{P}(y_j|x_i) - \mathcal{P}(y_j|x'_i)\right|^2 \tag{24}$$

This formula may be used to evaluate feature similarity when redundant features are searched for.

5 Relevance measures based on information theory

Information theory indices are most frequently used for feature evaluation. Information (negative of entropy) contained in the class distribution is:

$$H(Y) = -\sum_{i=1}^{K} \mathcal{P}(y_i)\log_2 \mathcal{P}(y_i), \tag{25}$$

where $\mathcal{P}(y_i) = m_i/m$ is the fraction of samples $\mathbf{x}$ from class $y_i, i = 1..K$. The same formula is used to calculate information contained in the discrete distribution of feature X values:

$$H(X) = -\sum_i \mathcal{P}(x_i)\log_2 \mathcal{P}(x_i). \tag{26}$$

Continuous features are discretized (binned) to compute information associated with a single feature or some kernel functions are fitted to approximate the density of X values and integration performed instead of summation. Information contained in the joint distribution of classes and features, summed over all classes, gives an estimation of the importance of the feature. Information contained in the joint distribution is:

$$H(Y,X) = -\sum_i \sum_{j=1}^{K} \mathcal{P}(y_j, x_i) \log_2 \mathcal{P}(y_j, x_i), \tag{27}$$

or for continuous features:

$$H(Y,X) = -\sum_{j=1}^{K} \int \mathcal{P}(y_j, x) \log_2 \mathcal{P}(y_j, x) dx, \tag{28}$$

where $\mathcal{P}(y_j, x_i), j = 1 \ldots K$ is the joint probability (density for continuous features) of finding the feature value $X = x_i$ for vectors $\mathbf{x}$ that belong to some class y_j and $\mathcal{P}(x_i)$ is the probability (density) of finding vectors with feature value $X = x_i$. Low values of $H(Y,X)$ indicate that vectors from a single class dominate in some intervals, making the feature more valuable for prediction.

Information is additive for the independent random variables. The difference $MI(Y,X) = H(Y) + H(X) - H(Y,X)$ may therefore be taken as "mutual information" or "information gain". Mutual information is equal to the expected value of the ratio of the joint to the product probability distribution, that is to the Kullback-Leibler divergence:

$$MI(Y,X) = -\sum_{i,j} \mathcal{P}(y_j, x_i) \log_2 \frac{\mathcal{P}(y_j, x_i)}{\mathcal{P}(y_j)\mathcal{P}(x_i)} = D_{KL}(\mathcal{P}(y_j, x_i) | \mathcal{P}(y_j)\mathcal{P}(x_i)). \tag{29}$$

A feature is more important if the mutual information $MI(Y,X)$ between the target and the feature distributions is larger. Decision trees use closely related quantity called "information gain" $IG(Y,X)$. In the context of feature selection this gain is simply the difference $IG(Y,X) = H(Y) - H(Y|X)$ between information contained in the class distribution $H(Y)$, and information after the distribution of feature values is taken into account, that is the conditional information $H(Y|X)$. This is equal to $MI(Y,X)$ because $H(Y|X) = H(Y,X) - H(X)$. A standard formula for the information gain is easily obtained from the definition of conditional information:

$$\begin{aligned} IG(Y,X) = H(Y) - H(Y|X) &= H(Y) + \sum_{ij} \mathcal{P}(y_j, x_i) \log_2 \mathcal{P}(y_j|x_i) \qquad (30) \\ &= H(Y) - \sum_{ij} \mathcal{P}(x_i) \left[-\mathcal{P}(y_j|x_i) \log_2 \mathcal{P}(y_j|x_i) \right], \end{aligned}$$

where the last term is the total information in class distributions for subsets induced by the feature values x_i, weighted by the fractions $\mathcal{P}(x_i)$ of the number of samples that have the feature value $X = x_i$. Splits induced by tests in nodes of decision trees are usually not based directly on all attribute values and thus information gain in general is different from mutual information, but for the feature selection purposes these two quantities are identical.

It is not difficult to prove that the Bayes error A_{BC} is bounded from above by half of the value of the conditional information and from below by the Fano inequality,

$$\frac{H(Y|X)-1}{\log_2 K} \leq A_{\mathrm{BC}} \leq \frac{1}{2}H(Y|X), \tag{31}$$

although the left side is usually negative and thus not useful. Minimizing $H(Y|X) = H(Y) - MI(Y,X)$, or maximizing mutual information, leads to an approximation of Bayes errors and optimal predictions. Error bounds are also known for the Renyi entropy that is somehow easier to estimate in on-line learning than the Shannon entropy [18].

Various modifications of the information gain have been considered in the literature on decision trees (cf. [42]), aimed at avoiding bias towards the multivalued features. These modifications include:

$$\begin{aligned}
IGR(Y,X) &= MI(Y,X)/H(X), && (32)\\
D_H(Y,X) &= 2H(Y,X) - H(Y) - H(X), && (33)\\
D_M(Y,X) &= 1 - MI(Y,X)/H(Y,X), && (34)\\
J_{\mathrm{SU}}(Y,X) &= 1 - \frac{D_H(Y,X)}{H(Y)+H(X)} = 2\frac{MI(Y,X)}{H(Y)+H(X)} \in [0,1]. && (35)
\end{aligned}$$

where IGR is the information gain ratio, D_H is the entropy distance, D_M is the Mantaras distance [11] and J_{SU} is the symmetrical uncertainty coefficient. The J_{SU} coefficient seems to be particularly useful due to its simplicity and low bias for multi-valued features [23].

The J-measure:

$$J_J(X) = \sum_i \mathcal{P}(x_i) \sum_j \mathcal{P}(y_j|x_i) \log \frac{\mathcal{P}(y_j|x_i)}{\mathcal{P}(y_j)}, \tag{36}$$

has been initially introduced to measure information content of logical rules [44], but it is applicable also to the feature selection [32].

[38] has defined an index called “average weight of evidence”, based on plausibility, an alternative to entropy in information:

$$J_{\mathrm{WE}}(X) = \sum_{j=1}^{K} \sum_i \mathcal{P}(x_i) \left| \log \frac{\mathcal{P}(y_j|x_i)(1-\mathcal{P}(y_j))}{(1-\mathcal{P}(y_j|x_i))\mathcal{P}(y_j)} \right|. \tag{37}$$

Minimum Description Length (MDL) is a general idea based on the Occkam’s razor principle and Kolmogorov’s algorithmic complexity [35]. The joint complexity of the theory inferred from the data and the length of the data encoded using this theory should be minimal. MDL has been applied to the construction of decision trees and the selection of features [32]. As in the description of χ^2 test, m training samples are divided into subsets of M_{ij} samples that belong to class $y_j, j = 1 \dots K$ and have a specific feature value

$x_i, i = 1 \dots M_x$. The number of bits needed for optimal encoding of the information about the class distribution for m training samples is estimated (this number is fixed for a given dataset), and the same estimation is repeated for each partitioning created by a feature value (or interval) x. Combinatorics applied to the information coding leads to the MDL formula expressed using binomial and multinomial coefficients $m!/m_1! \dots m_K!$ in the following way [32, 23]:

$$MDL(Y,X) = \log_2 \frac{m!}{M_{1\cdot}! \dots M_{K\cdot}!} + \log_2 \binom{m+K-1}{K-1} \qquad (38)$$
$$- \sum_{j=1}^{M_x} \log_2 \binom{M_{\cdot j}+K-1}{K-1} - \sum_{j=1}^{M_x} \log_2 \frac{M_{\cdot j}!}{M_{1j}! \dots M_{Kj}!},$$

where $M_{i\cdot}$ and $M_{\cdot j}$ are marginal distributions calculated from the M_{ij} matrix. The final relevance index $J_{\mathrm{MDL}}(Y,X) \in [0,1]$ is obtained by dividing this value by the first two terms representing the length of the class distribution description. A symmetrized version of MDL relevance index is used in [23], calculated by exchanging features and classes and averaging over the two values.

6 Decision trees for filtering

Decision trees select relevant features using top-down, hierarchical partitioning schemes. In the deeper branches of a tree only a small portion of all data is used and only local information is preserved. In feature selection global relevance is of greater importance. One way to achieve it is to create a single-level tree (for algorithms that allow for multiple splits), or a tree based on a single feature (for algorithms that use binary splits only) and evaluate their accuracy. An additional benefit of using decision trees for continuous features is that they provide optimized split points, dividing feature values into relatively pure bins. Calculation of probabilities $\mathcal{P}(x_j)$ and $\mathcal{P}(y_i|x_j)$ needed for the estimation of mutual information and other relevance indices becomes more accurate than with the naïve discretization based on the bins of equal width or bins with equal number of samples. Mutual information calculated after discretization based on a decision tree may be a few times larger than using naive discretization [16].

The 1R decision tree algorithm [28] is most appropriate for feature filtering because it creates only single level trees. Features are analyzed searching for a subset of values or a range of values for which vectors from a single class dominate. The algorithm has one parameter (called the "bucket size"), an acceptable level of impurity for each range of the feature values, allowing for reduction of the number of created intervals. Performance may be estimated using the $J_{\mathrm{BC}}(Y,X)$ index, and the optimal bucket size may be evaluated using cross-validation or bootstrap sampling that can help to avoid the bias for large number of intervals but will also increase computational costs.

The C4.5 tree [42] uses information gain to determine the splits and to select the most important features, therefore it always ranks as the most important features that are close to the root node. The CHAID decision tree algorithm [30] measures association between classes and feature values using χ^2 values, as in Eq. 8. Although the information gain and the χ^2 have already been mentioned as relevance indices the advantage of using decision trees is that automatic discretization of continuous features is performed.

The Gini impurity index used in the CART decision trees [5] sums the squares of the class probability distribution for a tree node, $J_{\mathrm{Gini}}(Y) = 1 - \sum_i \mathcal{P}(y_i)^2$. Given a feature X a split into subsets with discrete feature values x_j (or values in some interval) may be generated and Gini indices in such subsets calculated. The gain is proportional to the average of the sum of squares of all conditional probabilities:

$$J_{\mathrm{Gini}}(Y,X) = \sum_j \mathcal{P}(x_j) \sum_i \mathcal{P}(y_i|x_j)^2 \in [0,1], \tag{39}$$

giving a measure of the probability concentration useful for feature ranking. This index is similar to the entropy of class distributions and identical with the Bayesian measure Eq. 19.

The Separability Split Value (SSV) criterion is used to determine splits in decision tree [21] and to discretize continuous features [15, 13], creating a small number of intervals (or subsets) with high information content. It may also be used as feature relevance index. The best "split value" should separate the maximum number of pairs of vectors from different classes. Among all split values that satisfy this condition, the one that separates the smallest number of pairs of vectors belonging to the same class is selected. The *split value* for a continuous feature X is a real number s, while for a discrete feature it is a subset of all possible values of the feature. In all cases, the *left side* (LS) and the *right side* (RS) of a split value s is defined by a test $f(X,s)$ for a given dataset $\mathcal{D}$:

$$\begin{aligned} LS(s,f,\mathcal{D}) &= \{\mathbf{x} \in \mathcal{D} : f(\mathbf{x},s) = T\}\} \\ RS(s,f,\mathcal{D}) &= \mathcal{D} - LS(s,f,\mathcal{D}), \end{aligned} \tag{40}$$

where the typical test $f(\mathbf{x},s)$ is true if the selected feature $x_i < s$ or (for discrete feature) $x_i \in \{s\}$. The *separability of a split value s* is defined for a given test f as:

$$\begin{aligned} \mathrm{SSV}(s,f) = 2 &\sum_{i=1}^{K} |\mathrm{LS}(s,f,\mathcal{D}_i)| \cdot |\mathrm{RS}(s,f,\mathcal{D}-\mathcal{D}_i)| \\ &- \sum_i \min\left(|\mathrm{LS}(s,f,\mathcal{D}_i)|, |\mathrm{RS}(s,f,\mathcal{D}_i)|\right), \end{aligned} \tag{41}$$

where $\mathcal{D}_k$ is the subset of $\mathcal{D}$ vectors that belong to the class k. If several features separate the same number of pairs of training vectors the second term ranks higher the one that separates a lower number of pairs from the same class. This index has smilar properties to Gini and is easily calculated for both continuous and discrete features. For 10 or less feature values all subsets are checked to determine the simplest groupings, for a larger number of unique values the feature is treated as ordered and the best split intervals are searched for. In the feature selection applications of the SSV, splits are calculated and applied recursively to the data subsets $\mathcal{D}_k$, creating a single-feature tree. When pure nodes are obtained the algorithm stops and prunes the tree. The Bayesian Classifier rule is applied in each interval or for each subset created by this algorithm to calculate the $J_{\mathrm{SSV}}(Y,X)$ relevance index. More complex tree-based approaches to determine feature relevance use pruning techniques [15].

7 Reliability and bias of relevance indices

How good are different relevance indices? Empirical comparisons of the influence of various indices are difficult because results depend on the data and the classifier. What works well for document categorization [19] (large number of classes, features and samples), may not be the best for bioinformatics data (small number of classes, large number of features and a few samples), or analysis of images. One way to characterize relevance indices is to see which features they rank as identical. If a monotonic function could transform one relevance index into another the two indices would always rank features in the same way. Indeed such relations may be established between some indices (see Sec. 4), allowing for clustering of indices into highly similar or even equivalent groups, but perhaps many more relations may be established.

The ranking order predicted by the mutual information and other information theoretic measures, and by the accuracy of the optimal Bayesian Classifier using information contained in a single feature, is not identical. It is easy to find examples of binary-valued features where BC and MI predictions are reversed. Consider three binary features with the following class distributions: $\mathcal{P}(Y,X) = \begin{pmatrix} 0.50 \; 0.00 \\ 0.25 \; 0.25 \end{pmatrix}$, $\mathcal{P}(Y,X') = \begin{pmatrix} 0.45 \; 0.05 \\ 0.20 \; 0.30 \end{pmatrix}$, $\mathcal{P}(Y,X'') = \begin{pmatrix} 0.41 \; 0.09 \\ 0.10 \; 0.40 \end{pmatrix}$. The J_{BC} relevance indices for the three distributions are 0.50, 0.50, 0.62, the MI values are 0.31, 0.21, 0.30, and the J_{Gini} indices are 0.97, 0.98, and 0.99. Therefore the ranking in descending order according of the Bayesian relevance is $X'', X = X'$, mutual information gives X, X'', X', and the Gini index predicts X, X', X''.

The differences between relevance indices are apparent if the contour plots showing lines of constant values of these three indices are created for probability distributions $\mathcal{P}(y,x) = \begin{pmatrix} a \; 0.5-a \\ b \; 0.5-b \end{pmatrix}$. These contour plots are shown in Fig. 1 in the (a,b) coordinates. The $J_{\mathrm{BC}}(Y,X)$ index is linear, the $MI(Y,X)$ has logarithmic nonlinearity and the Gini index has stronger quadratic nonlinearity. For many distributions each index must give identical values. Unique ranking is obtained asking for “the second opinion”, that is using pairs of indices if the first one gives identical values. In the example given above the Bayesian relevance index could not distinguish between X and X', but using mutual information for such cases will give a unique ranking X'', X, X'.

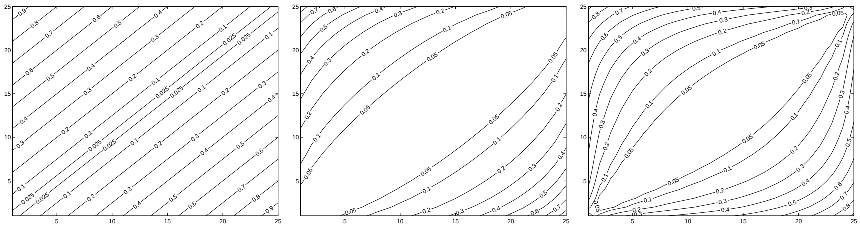

Fig. 1. Contours of constant values for BC relevance index (left), MI index (middle) and Gini index (right), in a, b coordinates.

Calculation of indices based on information theory for discrete features is straightforward, but for the continuous features the accuracy of entropy calculations based on simple discretization algorithms or histogram smoothing may be low. The literature on entropy estimation is quite extensive, especially in physics journals, where the concept of entropy has very wide applications (cf. [27]). The variance of the histogram-based mutual information estimators has been analyzed in [39]. A simple and effective way to calculate mutual information is based on Parzen windows [33]. Calculation of mutual information between pairs of features and the class distribution is more difficult, but interesting approximations based on the conditional mutual information have been proposed recently to calculate it [34].

Filters based on ranking using many relevance indices may give similar results. The main differences between relevance indices of the same type is in their bias in relation to the number of distinct feature values, and in their variance in respect to the accuracy of their estimation for small number of samples. The issue of bias in estimating multi-valued features has initially been discussed in the decision tree literature [42]. Gain-ratio and Mantaras distance have been introduced precisely to avoid favoring attributes with larger number of values (or intervals). Biases of 11 relevance indices, including information-based indices, Gini, J-measure, weight of evidence, MDL, and Relief, have been experimentally examined for informative and non-informative features [32]. For the two-class problems biases for a large number of feature values are relatively small, but for many classes they become significant. For mutual information, Gini and J-measure approximately linear increase (as a function of the number of feature values) is observed, with steepness proportional to the number of classes. In this comparison indices based on the Relief (Sec. 3) and MDL (Sec. 5) came as the least biased. Symmetrical uncertainty coefficient

J_{SU} has a similar low bias [23]. Biases in evaluation of feature correlations have been examined by Hall [23].

Significant differences are observed in the accuracy and stability of calculation of different indices when discretization is performed. Fig. 2 shows convergence plots of 4 indices created for overlapping Gaussian distributions (variance=1, means shifted by 3 units), as a function of the number of bins of a constant width that partition the whole range of the feature values. Analytical values of probabilities in each bin were used to simulate infinite amount of data, renormalized to sum to 1. For small (4-16) number of bins errors as high as 8% are observed in the accuracy of J_{BC} Bayesian relevance index. Convergence of this index is quite slow and oscillatory. Mutual information (Eq. 21) converges faster, and the information gain ratio (Eq. 32) shows similar behavior as the Gini index (Eq. 39) and the symmetrical uncertainty coefficient J_{SU} (Eq. 35) that converge quickly, reaching correct values already for 8 bins (Fig. 2). Good convergence and low bias make this coefficient a very good candidate for the best relevance index.

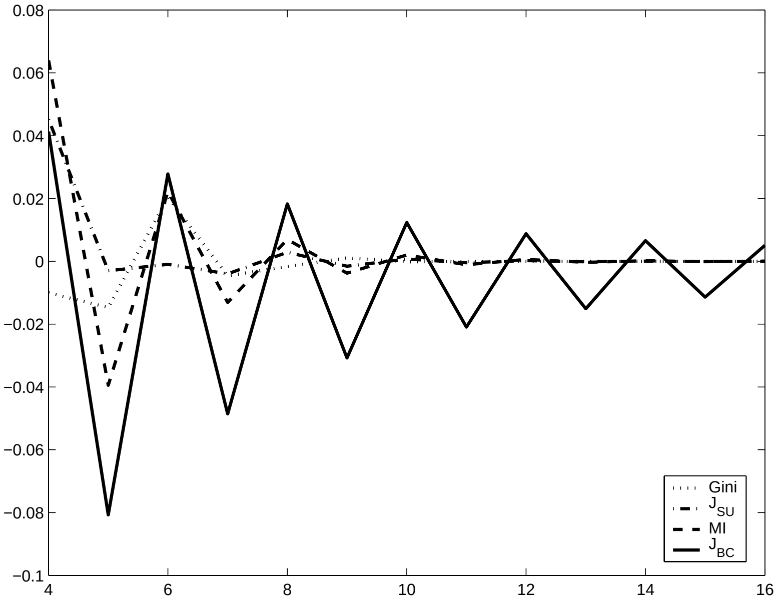

Fig. 2. Differences between the Gini, J_{SU}, MI, and J_{BC} indices and their exact value (vertical axis), as a function of the number of discretization bins (horizontal axis).

8 Filters for feature selection

Relevance indices discussed in the previous sections treat each feature as independent (with the exception of Relief family of algorithms Sec. 3 and the group correlation coefficient Eq. 13), allowing for feature ranking. Those features that have relevance index below some threshold are filtered out as not useful. Some feature selection algorithms may try to include interdependence between features. Given a subset of features $\mathbf{X}$ and a new candidate feature X with relevance index $J(X)$ an index $J(\{\mathbf{X}, X\})$ for the whole extended set is needed. In theory a rigorous Bayesian approach may be used to evaluate the gain in accuracy of the Bayesian classifier after adding a single feature. For k features the rule is:

$$A_{\mathrm{BC}}(\mathbf{X}) = \sum_{x_1, x_2, ..x_k} \max_i \mathcal{P}(y_i, x_1, x_2, \ldots x_k) \tag{42}$$

where the sum is replaced by integral for continuous features.

This formula converges slowly even in one dimension (Fig. 2), so the main problem is how to reliably estimate the joint probabilities $\mathcal{P}(y_j, x_1, x_2 \ldots x_k)$. The density of training data $\propto \mathcal{P}(x)^k$ goes rapidly to zero with the growing dimensionality k of the feature space. Already for 10 binary features and less than 100 data samples less than 10% of 2^{10} bins are non-empty. Although various histogram smoothing algorithms may regularize probabilities, and hashing techniques may help avoiding high computational costs [16], a reliable estimation of $A_{\mathrm{BC}}(\mathbf{X})$ is possible only if the underlying distributions are fully known. This may be useful as a "golden standard" to calculate error bounds, as it is done for one-dimensional distributions, but it is not a practical method.

Calculating relevance indices for subsets selected from a large number of features it is not possible to include full interactions between all the features. Note however that most wrappers may evaluate full feature interactions, depending on the classification algorithm used. Approximations based on summing pair-wise interactions offer a computationally less expensive alternative. The CFS algorithm described in Sec. 3 is based on Eq. 13, calculating average correlation coefficients between features and classes and between different features. Instead of a ratio for some relevance indices that may measure correlation or dependency between features one may use a linear combination of the two terms: $J(Y, X; \mathcal{S}) = J(Y, X) - \beta \sum_{s \in \mathcal{S}} J(X, X_s)$, where the user-defined constant β is introduced to balance the importance of the relevance $J(Y, X)$ and the redundancy estimated by the sum of feature-feature relevancies. Such algorithm has been used with mutual information as relevance measure by [3]. In this way redundancy of features is (at least partially) taken into account and search for good subsets of features may proceed at the filter level. A variant of this method may use a maximum of the pair relevance $J(X, X_s)$ instead of the sum over all features $s \in \mathcal{S}$; in this case β is not needed and fewer features will be recognized as redundant.

The idea of inconsistency or conflict – a situation in which two or more vectors with the same subset of feature values are associated with different classes – leads to a search for subsets of features that are consistent [10, 1]. This is very similar to the indiscernability relations and the search for reducts in rough set theory [46]. The inconsistency count is equal to the number of samples with identical features, minus the number of such samples from the class to which the largest number of samples belong (thus if there is only one class the index is zero). Summing over all inconsistency counts and dividing by the number of samples m the inconsistency rate for a given subset is obtained. This rate is an interesting measure of feature subset quality, for example it is monotonic (in contrast to most other relevance indices), decreasing with the increasing feature subsets. Features may be ranked according to their inconsistency rates, but the main application of this index is in feature selection.

9 Summary and comparison

There are various restrictions on applications of the relevance indices discussed in the previous sections. For example, some correlation coefficients (such as the χ^2 or Pearson's linear correlation) require numerical features and cannot be applied to features with nominal values. Most indices require probabilities that are not so easy to estimate for continuous features, especially when the number of samples is small. This is usually achieved using discretization methods [36]. Relevance indices based on decision trees may automatically provide such discretization, other methods have to rely on external algorithms.

In Table 1, information about the most popular filters is collected, including the formulas, the types of inputs X (binary, multivalued integer or symbolic, or continuous values), and outputs Y (binary for 2-class, multivalued integer for multiclass problems and continuous for regression).

The first method, Bayesian accuracy A_{BC}, is based on observed probabilities $\mathcal{P}(y_j, x_i)$ and provides a "golden standard" for other methods. Relations between the Bayesian accuracy and mutual information are known 31, and such relations may be inferred for other information-based indices, but in general theoretical results of this sort are difficult to find and many indices are simply based on heuristics. New methods are almost never tested against Bayesian accuracy for simple binary features and binary classes. Differences in ranking of features between major relevance indices presented in Sec. 7 are probably amplified in more general situations, but this issue has not been systematically investigated so far.

Other methods that belong to the first group of methods in Tab. 1 are somehow special. They are based on evaluation of confusion matrix elements and thus are only indirectly dependent on probabilities $\mathcal{P}(y_j, x_i)$. Confusion matrix may be obtained by any classifier, but using Bayesian approach for classification balanced accuracy, area-under-curve (AUC), F-measure, Bi-normal separation and odds ratio are still the best possible approaches, assuming specific costs of different type of errors.

Many variants of a simple statistical index based on separation of the class means exist. Although these indices are commonly applied to problems with binary targets extension to multiple target values is straightforward. In practice pair-wise evaluation (single target value against the rest) may work better, finding features that are important for discrimination of small classes. Feature values for statistical relevance indices must be numerical, but target values may be symbolic. Pearson's linear correlation coefficient can be applied only for numerical feature and target values, and its averaged (or maximum) version is used for evaluation of correlations with a subset of features. Decision-tree based indices are applicable also to symbolic values and may be computed quite rapidly. Some trees may capture the importance of a feature for a local subset of data handled by the tree nodes that lie several levels below the root. The Relief family of methods are especially attractive because they may be applied in all situations, have low bias, include interaction among features and may capture local dependencies that other methods miss.

Continuous target values are especially difficult to handle directly, but distance-based measures of similarity between distributions may handle them without problems. Kolmogorov distance and other relevance indices from this group may be expressed either by a sum of discrete probabilities or an integral over probability density functions. Bayesian measure, identical with the Gini index for discrete distributions, generalizes it to continuous features and continuous targets. The only exception in this group is the Value Difference Metric that has been specifically designed for symbolic data.

Indices based on information theory may also be used for continuous features and targets if probability densities are defined. Information gain ratio and symmetrical uncertainty coefficient are especially worth recommending, sharing low bias with the MDL approach (Sec. 5), and converging in a stable and quick way to their correct values.

10 Discussion and conclusions

Filters provide the cheapest approach to the evaluation of feature relevance. For a very large number of features they are indispensable, and only after filtering out most of the features other, more expansive feature selection methods become feasible.

Many approaches to filters discussed in the preceding sections show that there is no simple answer to the question: which relevance index is the best to construct a good filter? If there is sufficient data and joint probabilities may be estimated in a reliable way there is no reason why Bayesian relevance J_{BC} should not be used. After all other relevance indices, and in particular indices based on the theory of information, are only approximations to the Bayesian relevance. Unfortunately this index seems to be the most difficult

Method		**X**			*Y*			Comments
Name	Formula	B	M	C	B	M	C	
Bayesian accuracy	Eq. 1	+	s		+	s		Theoretically the golden standard, rescaled Bayesian relevance Eq. 2.
Balanced accuracy	Eq. 4	+	s		+	s		Average of sensitivity and specificity; used for unbalanced dataset, same as AUC for binary targets.
Bi-normal separation	Eq. 5	+	s		+	s		Used in information retrieval.
F-measure	Eq. 7	+	s		+	s		Harmonic of recall and precision, popular in information retrieval.
Odds ratio	Eq. 6	+	s		+	s		Popular in information retrieval.
Means separation	Eq. 10	+	i	+	+			Based on two class means, related to Fisher's criterion.
T-statistics	Eq. 11	+	i	+	+			Based also on the means separation.
Pearson correlation	Eq. 9	+	i	+	+	i	+	Linear correlation, significance test Eq. 12, or a permutation test.
Group correlation	Eq. 13	+	i	+	+	i	+	Pearson's coefficient for subset of features.
χ^2	Eq. 8	+	s		+	s		Results depend on the number of samples m.
Relief	Eq. 15	+	s	+	+	s	+	Family of methods, the formula is for a simplified version ReliefX, captures local correlations and feature interactions.
Separability Split Value	Eq. 41	+	s	+	+	s		Decision tree index.
Kolmogorov distance	Eq. 16	+	s	+	+	s	+	Difference between joint and product probabilities.
Bayesian measure	Eq. 16	+	s	+	+	s	+	Same as Vajda entropy Eq. 23 and Gini Eq. 39.
Kullback-Leibler divergence	Eq. 20	+	s	+	+	s	+	Equivalent to mutual information.
Jeffreys-Matusita distance	Eq. 22	+	s	+	+	s	+	Rarely used but worth trying.
Value Difference Metric	Eq. 22	+	s		+	s		Used for symbolic data in similarity-based methods, and symbolic feature-feature correlations.
Mutual Information	Eq. 29	+	s	+	+	s	+	Equivalent to information gain Eq. 30.
Information Gain Ratio	Eq. 32	+	s	+	+	s	+	Information gain divided by feature entropy, stable evaluation.
Symmetrical Uncertainty	Eq. 35	+	s	+	+	s	+	Low bias for multivalued features.
J-measure	Eq. 36	+	s	+	+	s	+	Measures information provided by a logical rule.
Weight of evidence	Eq. 37	+	s	+	+	s	+	So far rarely used.
MDL	Eq. 38	+	s		+	s		Low bias for multivalued features.

Table 1. Summary of the relevance measures suitable for filters. Features X and targets Y may be of the B = binary type (+), M = mutivalued, s (symbolic), or i (integer) only (symbolic implies integer), or C = continuous, real numbers (+). Methods that cannot work directly with continuous values need discretization.

to estimate reliably (see Fig. 2), leaving room for other approaches. In some applications including costs of different types of misclassifications (Sec. 2) is a better choice of relevance index, leading to the balanced accuracy (Eq. 4), F-measure or optimization of ROC curves. Evaluation of all such quantities will suffer from the same problem as evaluation of the Bayesian relevance J_{BC}, and therefore other, approximate but more reliable methods should be studied.

Different approaches to relevance evaluation lead to a large number of indices for ranking and selection. Certainly more papers with new versions of relevance indices for information filters will be published, but would they be more useful? As noted in the book on CART [5] the splitting criteria do not seem to have much influence on the quality of decision trees, so in the CART tree an old index known as Bayesian measure J_{BM} (Eq. 19) or Vajda Entropy (Eq. 23) has been employed, under the new name "Gini". Perhaps the actual choice of feature relevance indices also has little influence on performance of filters. For many applications a simple approach, for example using a correlation coefficient, may be sufficient.

Not all options have been explored so far and many open questions remain. Similarities, and perhaps equivalence up to monotonic transformation of relevance indices, should be established. The reliability of estimation of relevance indices – with the exception of entropy estimations – is not known. Biases towards multi-valued features of several indices have been identified but their influence on ranking is not yet clear. Little effort has been devoted so far towards cost-sensitive feature selection. In this respect the accuracy of Bayesian classification rules and other performance metrics related to logical rules are worth investigating.

Not much attention has been paid towards specific class-oriented and local, context-dependent filters. Some problems (especially in bioinformatics) require the simultaneous identification of several features that may individually have quite poor relevance. The paradigmatic benchmark problems of this sort are the parity problems, starting from the XOR. Only context-dependent local feature selection methods (like Relief, or filter methods applied to vectors in a localized feature space region) seem to be able to deal with such cases. Although our knowledge of filter-based feature selection has significantly grown in recent years still much remains to be done in this field.

Acknowledgment: I am grateful to Jacek Biesiada for drawing my attention to many papers in this field, to Norbert Jankowski and Krzysztof Grąbczewski for critically reading this paper, and to Isabelle Guyon for many helpful remarks. This work has been supported by the grant from the Polish Committee of Scientific Research 2005-2007.

References

1. H. Almuallim and T.G. Dietterich. Learning with many irrelevant features. In *Proceedings of the 9th National Conference on Artificial Intelligence (AAAI-91)*, pages 547–552, 1991.
2. A. Antos, L. Devroye, and L. Gyorfi. An extensive empirical study of feature selection metrics for text classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(7):643–645, 1999.
3. R. Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Trans. on Neural Networks*, 5:537–550, 1994.
4. D.A. Bell and H. Wang. A formalism for relevance and its application in feature subset selection. *Machine Learning*, 41:175–195, 2000.
5. L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.
6. L. Bruzzone, F. Roli, and S.B. Serpico. An extension of the jeffreys-matusita distance to multiclass cases for feature selection. *IEEE Transactions on Geoscience and Remote Sensing*, 33(6):1318–1321, 1995.
7. F.M. Coetzee, E. Glover, L. Lawrence, and C.L Giles. Feature selection in web applications by roc inflections and powerset pruning. In *Proceedings of 2001 Symp. on Applications and the Internet (SAINT 2001)*, pages 5–14, Los Alamitos, CA, 2001. IEEE Computer Society.
8. T.M. Cover. The best two independent measurements are not the two best. *IEEE Transactions on Systems, Man, and Cybernetics*, 4:116–117, 1974.
9. D.R. Cox and D.V. Hinkley. *Theoretical Statistics*. Chapman and Hall/CRC Press, Berlin, Heidelberg, New York, 1974.
10. M. Dash and H. Liu. Consistency-based search in feature selection. *Artificial Intelligence*, 151:155–176, 2003.
11. R.L. de Mantaras. A distance-based attribute selection measure for decision tree induction. *Machine Learning Journal*, 6:81–92, 1991.
12. L. Devroye, L. Gyrfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, Berlin, Heidelberg, New York, 1996.
13. W. Duch, R. Adamczak, and K. Grąbczewski. A new methodology of extraction, optimization and application of crisp and fuzzy logical rules. *IEEE Transactions on Neural Networks*, 12:277–306, 2001.
14. W. Duch and L. Itert. A posteriori corrections to classification methods. In L. Rutkowski and J. Kacprzyk, editors, *Neural Networks and Soft Computing*, pages 406–411. Physica Verlag, Springer, Berlin, Heidelberg, New York, 2002.
15. W. Duch, R. Setiono, and J. Zurada. Computational intelligence methods for understanding of data. *Proceedings of the IEEE*, 92(5):771–805, 2004.
16. W. Duch, T. Winiarski, J. Biesiada, and A. Kachel. Feature ranking, selection and discretization. In *Proceedings of Int. Conf. on Artificial Neural Networks (ICANN)*, pages 251–254, Istanbul, 2003. Bogazici University Press.
17. R.O. Duda, P.E. Hart, and D.G. Stork. *Patter Classification*. John Wiley & Sons, New York, 2001.
18. D. Erdogmus and J.C. Principe. Lower and upper bounds for misclassification probability based on renyis information. *Journal of VLSI Signal Processing Systems*, 37(2-3):305–317, 2004.
19. G. Forman. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3:1289–1305, 2003.

20. E.E. Ghiselli. *Theory of psychological measurement.* McGrawHill, New York, 1964.
21. K. Grąbczewski and W. Duch. The separability of split value criterion. In *Proceedings of the 5th Conf. on Neural Networks and Soft Computing*, pages 201–208, Zakopane, Poland, 2000. Polish Neural Network Society.
22. I. Guyon, H.-M. Bitter, Z. Ahmed, M. Brown, and J. Heller. Multivariate nonlinear feature selection with kernel multiplicative updates and gram-schmidt relief. In *BISC FLINT-CIBI 2003 workshop, Berkeley, Dec. 2003*, 2003.
23. M.A. Hall. *Correlation-based Feature Subset Selection for Machine Learning.* PhD thesis, Department of Computer Science, University of Waikato, Waikato, N.Z., 1999.
24. D.J. Hand. *Construction and assessment of classification rules.* J. Wiley and Sons, Chichester, 1997.
25. J.A. Hanley and B.J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143:29–36, 1982.
26. R.M. Hogarth. Methods for aggregating opinions. In H. Jungermann and G. de Zeeuw, editors, *Decision Making and Change in Human Affairs.* D. Reidel Publishing, Dordrecht, Holland, 1977.
27. D. Holste, I. Grosse, and H. Herzel. Bayes' estimators of generalized entropies. *J. Physics A: Math. General*, 31:2551–2566, 1998.
28. R.C. Holte. Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 11:63–91, 1993.
29. S.J. Hong. Use of contextual information for feature ranking and discretization. *IEEE Transactions on Knowledge and Data Engineering*, 9:718–730, 1997.
30. G.V. Kass. An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29:119–127, 1980.
31. R. Kohavi and G. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1996.
32. I. Kononenko. On biases in estimating the multivalued attributes. In *Proceedings of IJCAI-95, Montreal*, pages 1034–1040, San Mateo, CA, 1995. Morgan Kaufmann.
33. N. Kwak and C-H. Choi. Input feature selection by mutual information based on parzen window. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:1667–1671, 2002.
34. N. Kwak and C-H. Choi. Input feature selection for classification problems. *IEEE Transactions on Neural Networks*, 13:143–159, 2002.
35. M. Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications.* Text and Monographs in Computer Science. Springer, Berlin, Heidelberg, New York, 1993.
36. H. Liu, F. Hussain, C.L. Tan, and M. Dash. Discretization: An enabling technique. *Journal of Data Mining and Knowledge Discovery*, 6(4):393–423, 2002.
37. H. Liu and R. Setiono. Feature selection and discretization. *IEEE Transactions on Knowledge and Data Engineering*, 9:1–4, 1997.
38. D. Michie. Personal models of rationality. *J. Statistical Planning and Inference*, 21:381–399, 1990.
39. R. Moddemeijer. A statistic to estimate the variance of the histogram based mutual information estimator based on dependent pairs of observations. *Signal Processing*, 75:51–63, 1999.

40. A.Y. Ng. On feature selection: learning with exponentially many irrelevant features as training examples. In *Proceedings of the 15th International Conference on Machine Learning*, pages 404–412, San Francisco, CA, 1998. Morgan Kaufmann.
41. W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical recipes in C. The art of scientific computing.* Cambridge University Press, Cambridge, UK, 1988.
42. J.R. Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufman, San Mateo, CA, 1993.
43. M. Robnik-Sikonja and I. Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Machine Learning*, 53:23–69, 2003.
44. P. Smyth and R.M. Goodman. An information theoretic approach to rule induction from databases. *IEEE Transactions on Knowledge and Data Engineering*, 4:301–316, 1992.
45. J.A. Swets. Measuring the accuracy of diagnostic systems. *Proceedings of the IEEE*, 240(5):1285–1293, 1988.
46. R.W. Swiniarski and A. Skowron. Rough set methods in feature selection and recognition. *Pattern Recognition Letters*, 24:833–849, 2003.
47. K. Torkkola. Feature extraction by non parametric mutual information maximization. *Journal of Machine Learning Research*, 3:1415–1438, 2003.
48. G.T. Toussaint. Note on optimal selection of independent binary-valued features for pattern recognition. *IEEE Transactions on Information Theory*, 17:618–618, 1971.
49. I. Vajda. *Theory of statistical inference and information.* Kluwer Academic Press, London, 1979.
50. C.J. van Rijsbergen. *Information Retrieval.* Butterworths, London, 1979.
51. T.R. Vilmansen. Feature evaluation with measures of probabilistic dependence. *IEEE Transactions on Computers*, 22:381–388, 1973.
52. D.R. Wilson and T.R. Martinez. Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research*, 6:1–34, 1997.