Transductive Reliability Estimation for Kernel Based Classifiers

Dimitris Tzikas[1], Matjaz Kukar[2], and Aristidis Likas[1]

[1] Department of Computer Science, University of Ioannina, Greece
[2] Faculty of Computer and Information Science, University of Ljubljana, Slovenia
tzikas@cs.uoi.gr, matjaz.kukar@fri.uni-lj.si, arly@cs.uoi.gr

Abstract. Estimating the reliability of individual classifications is very important in several applications such as medical diagnosis. Recently, the transductive approach to reliability estimation has been proved to be very efficient when used with several machine learning classifiers, such as Naive Bayes and decision trees. However, the efficiency of the transductive approach for state-of-the art kernel-based classifiers was not considered. In this work we deal with this problem and apply the transductive reliability methodology with sparse kernel classifiers, specifically the Support Vector Machine and Relevance Vector Machine. Experiments with medical and bioinformatics datasets demonstrate better performance of the transductive approach for reliability estimation compared to reliability measures obtained directly from the output of the classifiers. Furthermore, we apply the methodology in the problem of reliable diagnostics of the coronary artery disease, outperforming the expert physicians' standard approach.

1 Introduction

Decision-making is usually an uncertain and complicated process, therefore it is often crucial to know the magnitude of diagnosis' (un)reliability in order to minimize risks, for example in the medical domain risks related to the patient's health or even life. One of the reasons why machine learning methods are infrequently used in practice is that they fail to provide an unbiased reliability measure of predictions.

Although there are several methods for estimating the overall performance of a classifier, e.g cross-validation, and quality (reliability and validity) of collected data [1], there is very little work on estimating the reliability of individual classifications. The transductive reliability methodology as introduced in [2] computes the reliability of an individual classification, by studying the stability of the trained model when the training set is perturbed (the newly classified example is added to the training set and the classifier is retrained). For reliable classifications, this process should not lead to significant model changes. The transductive reliability methodology has been applied on traditional classifiers like Naive Bayes and decision trees with interesting results. Here, we examine the effectiveness of this methodology when applied on sparse kernel-based classifiers, such as the Support Vector Machine (SVM) and the Relevance Vector

M.R. Berthold, J. Shawe-Taylor, and N. Lavrač (Eds.): IDA 2007, LNCS 4723, pp. 37–47, 2007.

Machine (RVM), and compare transductive reliability estimations with reliability measures based on the outputs that SVM and RVM provide. Furthermore, we apply the methodology for diagnosis of the coronary artery disease (CAD) using kernel-based classifiers and compare our results to the performance of expert physicians using an established standard methodology.

2 Transduction Reliability Estimations

Transduction is an inference principle that takes a training sample and aims at estimating the values of a discrete or continuous function only at given unlabeled points of interest from input space, as opposed to the whole input space for induction. In the learning process the unlabeled points are suitably labelled and included into the training sample. The usefullness of unlabeled data has also been advocated in the context of co-training. It has been shown [3] that for every better-than-random classifier its performance can be significantly improved by utilizing only additional unlabeled data.

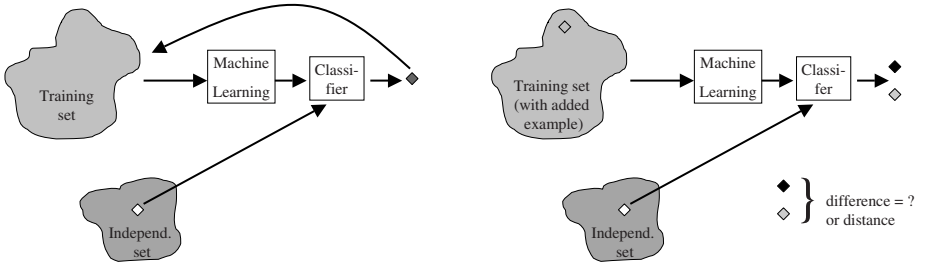
The transductive reliability estimation process and its theoretical foundations originating from Kolmogorov complexity are described in more detail in [2]. In practice, it is performed in a two-step process, featuring an *inductive step* followed by a *transductive step*.

- An *inductive step* is just like an ordinary inductive learning process in Machine Learning. A Machine Learning algorithm is run on the training set, *inducing* a classifier. A selected example is taken from an independent dataset and classified using the induced classifier. An example, labelled with the predicted class is temporarily included into the training set (Figure 1a).
- A *transductive step* is almost a repetition of an inductive step. A Machine Learning algorithm is run on the changed training set, *transducing* a classifier. The same example as before is taken from the independent dataset and is classified using the transduced classifier (Figure 1b). Both classifications of the same example are compared and their difference (distance) is calculated, thus approximating the randomness deficiency.
- After the reliability is calculated, the example in question is removed from the training set.

The machine learning algorithm, whose reliability is being assessed, is assumed to provide a probability distribution p that describes the probability that its input belongs at each possible class. In order to measure how much the model changes, we calculate the distance between the probability distribution p of the initial classifier and the probability distribution q of the augmented classifier, using the Symmetric Kullback-Leibler divergence, or J-divergence, which is defined as

$$J(p,q) = \sum_{i=1}^{n} (p_i - q_i) \log_2 \frac{p_i}{q_i}. \tag{1}$$

$J(p,q)$ is limited to the interval $[0,\infty)$, with $J(P,P) = 0$. For the ease of interpretation, it is desirable for reliability values to be bounded to the $[0,1]$ interval, $J(p,q)$ is normalized in the spirit of Martin-Löf's test for randomness



(a) Inductive step. (b) Transductive step.

Fig. 1. Transductive reliability estimation

[2,4, pp. 129], to obtain the transductive reliability measure (TRE) used in our approach:

$$TRE = 1 - 2^{-J(p,q)}. \tag{2}$$

Due to non-optimal classifiers resulting from learning in noisy and incomplete datasets, it is inappropriate to select *a priori* fixed boundary (say, 0.90) as a threshold above which a classification is considered reliable. To deal with this problem, we split the range $[0, 1]$ of reliability estimation values into two intervals by selecting a threshold T. The lower interval $[0, T)$ contains unreliable classifications, while the higher interval $[T, 1]$ contains reliable classifications. As a splitting point selection criterion we use maximization of the information gain[5]:

$$Gain = H(S) - \frac{|S_1|}{|S|}H(S_1) - \frac{|S_2|}{|S|}H(S_2), \tag{3}$$

where $H(S)$ denotes the entropy of set S, $S_1 = \{x : TRE(x) < T\}$ is the set of unreliable examples and $S_2 = \{x : TRE(x) > T\}$ is the set of reliable results.

Note that our approach is considerably different from that described in [6,7]. Their approach is tailor-made for SVM (it works by manipulating support vectors) while ours requires only that the applied classifier provide a probability distribution. Our approach can also be used in conjunction with probability calibration, e.g. by utilizing the typicalness concept [8,9].

3 Kernel-Based Methods

Kernel methods have been extensively used to solve classification problems, where a training set $\{x_n, t_n\}_{n=1}^{N}$ is given, so that t_n denotes the class label of training example x_n. The class labels t_n are discrete, e.g. $t \in \{0, 1\}$ for binary classification, and they describe the class to which each training example belongs. Kernel methods, are based on a mapping function $\phi(x)$ that maps each training vector to a higher dimensional feature space. Then, inner products between training examples are computed in this new feature space, by evaluating

the corresponding kernel function $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$. This kernel function, provides a measure of similarity between training examples.

Recently, there is much interest in sparse kernel methods, such as the Support Vector Machine (SVM) and the Relevance Vector Machine (RVM). These methods are called sparse because, after training with the full dataset, they make predictions using only a small subset of the available training vectors. In SVM sparsity is achieved through suitable weight regularization, while RVM is a Bayesian model and sparsity is a consequence of the use of a suitable sparse prior distribution on the weights. The remaining training vectors, which are used for predictions are called support vectors (SV) in the case of SVM and relevance vectors (RV) in the case of RVM. The main reason why sparse kernel methods are so interesting and effective, is that during training, they automatically estimate the complexity of the dataset, and thus they have good generalization performance on both simple and complex datasets. In simple datasets only few support/relevance vectors will be used, while in more difficult datasets the number of support/relevance vectors will increase. Furthermore, making predictions using only a small subset of the initial training examples is typically much more computationally efficient.

3.1 Support Vector Machine

The support vector machine (SVM) classifier, is a kernel classifier that aims at finding an optimal hyperplane which separates data points of two classes. This hyperplane is optimal in the sense that it maximizes the margin between the hyperplane and the training examples. The SVM classifier [10] makes decisions for an unknown input vector, based on the sign of the decision function:

$$y_{SVM}(x) = \sum_{n=1}^{N} w_n K(x, x_n) + b \tag{4}$$

After training, most of the weights w are set to exactly zero, thus predictions are made using only few of the training vectors, which are the support vectors. Assuming that the two classes are labeled with '-1' and '1', so that $t_n \in \{-1, 1\}$, the weights $w = (w_1, \ldots, w_N)$ are set by solving the following quadratic programming problem:

$$\begin{aligned} \min_{w,b,\xi} \ & \frac{1}{2} w^T w + C \sum_{n=1}^{N} \xi_n \\ \text{subject to } \ & t_n(w^T \phi(x_n) + b) \geq 1 - \xi_n \\ & \xi_n \geq 0 \end{aligned} \tag{5}$$

where the auxiliary variables $\xi = (\xi_1, \ldots, \xi_N)$ have been introduced to deal with non-separable datasets.

SVM makes predictions based on the decision function of eq. (4). Positive values of the decision function ($y_{SVM}(x) > 0$) correspond to class '1', while negative values ($y_{SVM}(x) < 0$) correspond to class '-1'. Furthermore, the absolute

value of the decision function provides a measure of the certainty of the decision. Values near zero, correspond to points near the decision boundary and therefore may be unreliable, while large values of the decision function should correspond to reliable classifications. In practice, we first obtain probabilistic predictions by applying the sigmoid function $\sigma(x) = 1/(1+\exp(-x))$ to the SVM outputs and then compute the reliability measure as:

$$RE_{SVM} = |2\sigma(y_{SVM}(x)) - 1|. \tag{6}$$

3.2 Relevance Vector Machine

The relevance vector machine (RVM) classifier [11], is a probabilistic extension of the linear regression model, which provides sparse solutions. It is analogous to the SVM, since it computes the decision function using only few of the training examples, which are now called relevance vectors. However training is based on different objectives.

The RVM model $y(x; w)$ is the output of a linear model with parameters $w = (w_1, \ldots, w_N)^T$, with application of a sigmoid function for the case of classification:

$$y_{RVM}(x) = \sigma(\sum_{n=1}^{N} w_n K(x, x_n)), \tag{7}$$

where $\sigma(x) = 1/(1+\exp(-x))$. In the RVM, sparseness is achieved by assuming a suitable prior distribution on the weights, specifically a zero-mean, Gaussian distribution with distinct inverse variance α_n for each weight w_n:

$$p(w|\alpha) = \prod_{n=1}^{N} N(w_n|0, \alpha_n^{-1}). \tag{8}$$

The variance hyperparameters $\alpha = (\alpha_1, \ldots, \alpha_N)$ are assumed to be Gamma distributed random variables:

$$p(\alpha) = \prod_{n=1}^{N} Gamma(\alpha_n|a, b). \tag{9}$$

The parameters a and b are assumed fixed and usually they are set to zero ($a = b = 0$), which provides sparse solutions.

Given a training set $\{x_n, t_n\}_{n=1}^{N}$ with $t_n \in \{0, 1\}$ training in RVM is equivalent to compute the posterior distribution $p(w, \alpha|t)$. However, since this computation is intractable, a quadratic approximation $\log p(w|t, \alpha) \approx (w-\mu)^T \Sigma^{-1}(w-\mu)$ is assumed and we compute matrix Σ and vector μ as:

$$\Sigma = (\Phi^T B \Phi + A)^{-1} \tag{10}$$

$$\mu = \Sigma \Phi^T B \hat{t} \tag{11}$$

with the $N \times N$ matrix Φ defined as $[\Phi]_{ij} = K(x_i, x_j)$, $A = \mathrm{diag}(\alpha_1, \ldots, \alpha_N)$, $B = \mathrm{diag}(\beta_1, \ldots, \beta_N)$, $\beta_n = y_{RVM}(x_n)[1 - y_{RVM}(x_n)]$ and $\hat{t} = \Phi\mu + B^{-1}(t - y)$.

The parameters α are then set to the values α_{MP} that maximize the logarithm of the following marginal likelihood

$$L(\alpha) = \log p(\alpha|t) = -\frac{1}{2}\left[N \log 2\pi + log|C| + t^T C^{-1} t\right], \tag{12}$$

with $C = B^{-1} + \Phi A^{-1} \Phi^T$. This, gives the following update formula:

$$\alpha_n = \frac{1 - \alpha_n \Sigma_{nn}}{\mu_n^2} \tag{13}$$

The RVM learning algorithm iteratively evaluates formulas (10),(11) and (13).

After training, the value of $y_{RVM}(x) = y(x; \mu)$ can be used to estimate the reliability of the classification decision for input x. Values close to 0.5 are near the decision boundary and therefore are unreliable classifications, while values near 0 and near 1 should correspond to reliable classifications. In our experiments, we used the reliability measure

$$RE_{RVM} = |2y_{RVM}(x) - 1|, \tag{14}$$

which takes values near 0 for unreliable classifications and near 1 for reliable classifications.

3.3 Incremental Relevance Vector Machine

An interesting property of the RVM model that can be exploited in the transductive approach, is that it can be trained incrementally, as proposed in [12]. The proposed incremental algorithm, initially assumes an empty model, that does not use any basis functions. Then, it incrementally adds, deletes and re-estimates basis functions, until convergence. It is based on the observation that the marginal likelihood, see eq. (12), can be decomposed as:

$$L(\alpha) = L(\alpha_{-n}) + l(\alpha_n), \tag{15}$$

where $L(\alpha_{-n})$ does not depend on α_n and

$$l(\alpha_n) = \log \alpha_n - \log(\alpha_n + s_n) + \frac{q_n^2}{\alpha_n + s_n}, \tag{16}$$

with $s_n = \phi_n^T C_{-n}^{-1} \phi_n$ and $q_n = \phi_n^T C_{-n}^{-1} \hat{t}$. Here, $C_{-n} = B^{-1} + \sum_{i \neq n} \alpha_n^{-1} \phi_n \phi_n^T$ denotes the matrix C without the contribution of the n-th basis function, so that $C = C_{-n} + \alpha_n^{-1} \phi_n \phi_n^T$, s_n is the "sparseness" factor that measures how sparse the model is and q_n is the "quality" factor that measures how well the model fits the observations. Based on this decomposition, analysis of $l(\alpha_n)$ shows that it is maximized when

$$\alpha_n = \frac{s_n^2}{q_n^2 - s_n} \qquad \text{if } q_n^2 > s_n \tag{17}$$

$$\alpha_n = \infty \qquad \text{if } q_n^2 \leq s_n \tag{18}$$

Based on this result, the following algorithm is proposed in [12]:

1. Initially assume an empty model, set $a_n = \infty$, for all n
2. Select a training point x_n and compute the corresponding basis function ϕ_n as well as s_n and q_n.
   (a) if $q_n^2 > s_n$ and $\alpha_n = \infty$ add the basis function to the model, using eq. (17) to set α_n
   (b) if $q_n^2 > s_n$ and $\alpha_n < \infty$ re-estimate α_n
   (c) if $q_n^2 <= s_n$ remove the basis function from the model, set $\alpha_n = \infty$
3. Compute Σ and μ, using eq. (10) and (11)
4. Repeat from step 2, until convergence.

4 Evaluation of Transductive Reliability Estimations

In this section, we apply the transductive reliability methodology in a series of classification problems and compare the performance of transductive reliability estimations, with respect to the reliability measures that are directly computed based on SVM and RVM outputs. Transductive reliability estimations, are obtained following the procedure described in Section 2. After training the model and computing its output for a new test point x_*, we add this test point to the training set with the predicted label and retrain the model. Transductive reliability estimations are obtained by measuring the distance between the output distributions of the two models.

In the case of RVM we also considered a modification, where we used the incremental training algorithm to obtain fast transductive reliability estimations. Specifically, after adding the new training point x_*, instead of retraining from scratch, we can use the incremental algorithm to continue training the previous model. This is much more computationally efficient, and in the experiments it appears to provide better performance than the standard approach of training from scratch.

In order to evaluate the performance of the reliability estimation methods, we apply the following procedure. We perform leave-one-out cross-validation on the available training dataset and compute a prediction for the class of each training point and a reliability estimation (RE) of this prediction. Afterwards, we can discriminate reliable and unreliable classifications by selecting a threshold (T)

Table 1. Information gain of SVM/RVM reliability estimations and transductive reliability estimations

Method	hepatitis	new-thyroid	wdbc	leukemia
RE_{SVM}	0.106	0.083	0.036	0.054
TRE_{SVM}	0.120	0.092	0.047	0.073
RE_{RVM}	0.109	0.068	0.091	0.089
TRE_{RVM}	0.178	0.062	0.094	0.062
$TRE_{RVM(inc)}$	0.133	0.072	0.106	0.107

for the reliability measure. Using an ideal reliability measure all correct classifications should be labeled reliable ($RE > T$), while all incorrect classifications should be labeled unreliable. Thus, an evaluation of the reliability measure is obtained by computing the percentage of correct and reliable classifications, and the percentage of incorrect reliable classifications. Plotting these percentages, for many values of the threshold, produces an ROC curve, which illustrates the performance of the reliability estimation method.

Although the ROC describes the overall effectiveness of a reliability measure, in practice, a single threshold value has to be used. This is selected by maximizing the information gain, as explained in Section 2. The information gain may also be used to compare the performance of several reliability measures. Table 1 shows the information gain that is achieved by: i) using directly the SVM/RVM reliability estimates RE_{SVM} and RE_{RVM}, ii) using the transduction reliability principle (TRE). Results are shown for three medical datasets from the UCI machine learning repository and the leukemia bioinformatics dataset. It is clear that when SVM is used, transduction provides better information gain for all datasets. The same happens with incremental RVM, while when typical RVM is used, transduction is better in two of the three cases.

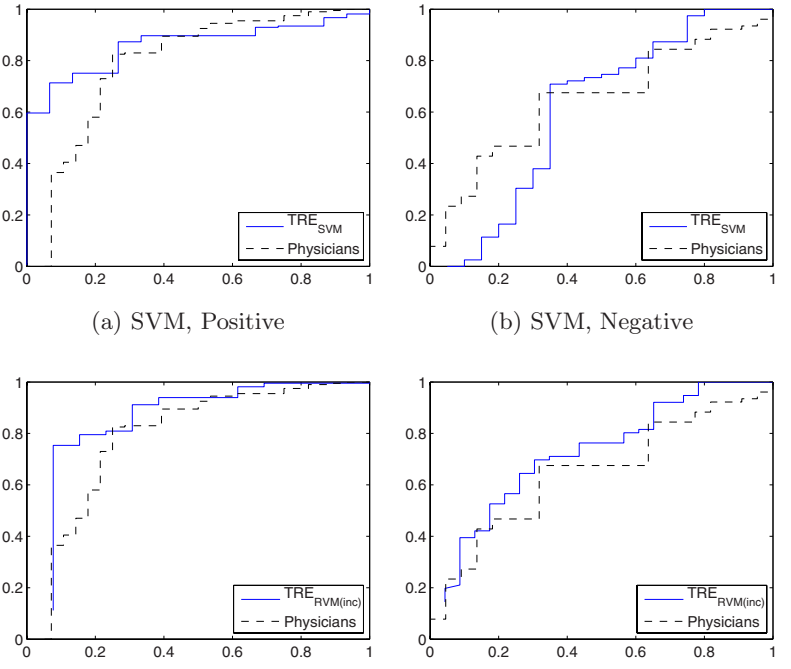
5 Diagnosis of Coronary Artery Disease

Coronary artery disease (CAD) is the most important cause of mortality in all developed countries. It is caused by diminished blood flow through coronary arteries due to stenosis or occlusion. CAD produces impaired function of the heart and finally the necrosis of the myocardium – myocardial infarction.

In our study we used a dataset of 327 patients (250 males, 77 females) with performed clinical and laboratory examinations, exercise ECG, myocardial scintigraphy and coronary angiography because of suspected CAD. The features from the ECG an scintigraphy data were extracted manually by the clinicians. In 228 cases the disease was angiographically confirmed and in 99 cases it was excluded. 162 patients had suffered from recent myocardial infarction. The patients were selected from a population of approximately 4000 patients who were examined at the Nuclear Medicine Department, University Clinical Centre, Ljubljana,

Table 2. Comparison of the performance of expert physicians and machine learning classification methods for the CAD dataset

	Positive			Negative		
Method	Reliable	Errors	AUC	Reliable	Errors	AUC
Physicians	0.72	0.04	0.790	0.45	0.07	0.650
RE_{SVM}	0.65	0.00	0.903	0.30	0.04	0.566
TRE_{SVM}	0.76	0.02	0.861	0.57	0.08	0.672
RE_{RVM}	0.63	0.004	0.842	0.54	0.06	0.729
TRE_{RVM}	0.67	0.013	0.767	0.49	0.05	0.702
$TRE_{RVM(inc)}$	0.69	0.004	0.850	0.54	0.07	0.720



(a) SVM, Positive (b) SVM, Negative

(c) RVM, Positive (d) RVM, Negative

Fig. 2. ROC curves for the transduction reliability measures for SVM and incremental RVM, using the CAD dataset and considering separately the positive and negative examples

Slovenia, between 1991 and 1994. We selected only the patients with complete diagnostic procedures (all four levels) [13].

Physicians apply a stepwise diagnostic process and use Bayes law to compute a posterior probability of disease, based on some diagnostic tests and a prior probability according to the age, gender and type of chest pain for each patient. Reliable diagnoses are assumed to be those whose posterior probability is over 0.90 (positive) or under 0.10 (negative). We considered treating the problem by training an SVM or an RVM classifier and using the transductive reliability principle to estimate the reliability of each classification. For evaluation purposes, we performed leave-one-out cross-validation, and for each example we predicted a class and a reliability of the classification. We then splitted classifications to reliable and unreliable by computing the threshold that maximizes the information gain and measured the percentage of reliable diagnoses (with the reliability measure above some threshold), and errors made in this process (percentage of incorrectly diagnosed patients with seemingly reliable diagnoses).

The results of these experiments are summariazed in Table 2 and furthermore, in Figure 2 ROC curves are plotted separately for the cases of positive and negative examples. The area (AUC) under these curves, which measures the

overall reliability performace, is also shown in Table 2. It can be observed that when the transduction principle is used along with SVM and incremental RVM, better performance is achieved compared to physicians.

Specifically, notice that the transductive SVM, has reliably detected 0.76% of the positive examples and 0.57% of negative examples, which is much better than the percentages of physicians, which are 0.72% and 0.45% respectively More important is the fact that at the same time, transductive SVM made less errors in positive examples, specifically 0.02% when the physicians made 0.04%. In medical diagnosis applications, it is very important that this error rate is kept at very small values. The error rate in negative examples, is 0.08% for physicians and 0.07% for transductive SVM, which is comparable.

When using the RVM model, the error rate of positive examples is dropped even lower to 0.004%. Although, the percentage of reliably detected positive examples (0.63%) is somewhat less than the one of physicians (0.72%), it is improved to 0.69% when using the incremental transduction principle. The RVM percentage of reliably detected negative examples is slightly higher than physicians, while the error rate of negative examples is about the same. Notice, that non-incremental transduction with RVM did not perform as expected, probably because relevant vectors are very sensitive to small changes of the training set.

6 Conclusions

We applied the transduction methodology for reliability estimation on sparse kernel-based classification methods. Experiments on medical datasets from the UCI repository and a bioinformatics gene expression dataset, indicate that, when used with kernel-based classifiers, transductive reliability estimations are more accurate than simple reliability measures based on the outputs of kernel classifiers. Furthermore, we applied the transductive methodology in the problem of CAD diagnosis, achieving better reliability estimation performance compared to the standard physicians procedure.

Acknowledgements

This work was supported in the framework of the "Bilateral S+T cooperation between the Hellenic Republic and the Republic of Slovenia (2004-2006)".

References

1. Hand, D., Mannila, H., Smyth, P.: Principles of Data Mining. MIT Press, Cambridge (2001)
2. Kukar, M., Kononenko, I.: Reliable classifications with Machine Learning. In: Elomaa, T., Mannila, H., Toivonen, H. (eds.) ECML 2002. LNCS (LNAI), vol. 2430, Springer, Heidelberg (2002)
3. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Proceedings of the 11th Annual Conference on Computational Learning Theory, pp. 92–100 (1998)

4. Li, M., Vitányi, P.: An introduction to Kolmogorov complexity and its applications, 2nd edn. Springer, New York (1997)
5. Dougherty, J., Kohavi, R., Sahami, M.: Supervised and unsupervised discretization of continuous features. In: Proc. ICML'95, pp. 194–202. Morgan Kaufmann, San Francisco (1995)
6. Gammerman, A., Vovk, V., Vapnik, V.: Learning by transduction. In: Proceedings of the 14^{th} Conference on Uncertainty in Artificial Intelligence, Madison, Wisconsin, pp. 148–155 (1998)
7. Saunders, C., Gammerman, A., Vovk, V.: Transduction with confidence and credibility. In: Proceedings of the International Joint Conference on Artificial Intelligence, Stockholm, Sweden (1999)
8. Proedrou, K., Nouretdinov, I., Vovk, V., Gammerman, A.: Transductive confidence machines for pattern recognition. In: Elomaa, T., Mannila, H., Toivonen, H. (eds.) ECML 2002. LNCS (LNAI), vol. 2430, pp. 381–390. Springer, Heidelberg (2002)
9. Kukar, M.: Quality assessment of individual classifications in machine learning and data mining. Knowledge and Information Systems 9(3), 364–384 (2006)
10. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(3), 273–297 (1995)
11. Tipping, M.E.: Sparse Bayesian learning and the Relevance Vector Machine. Journal of Machine Learning Research 1, 211–244 (2001)
12. Tipping, M., Faul, A.: Fast marginal likelihood maximisation for sparse Bayesian models. In: Proc. of the Ninth International Workshop on Artificial Intelligence and Statistics (2003)
13. Kukar, M., Kononenko, I., selj, C.G., Kralj, K., Fettich, J.: Analysing and improving the diagnosis of ischaemic heart disease with machine learning. Artificial Intelligence in Medicine: Special Issue on Data Mining Techniques and Applications in Medicine (1999) (in press)