Volume 42, Number 5
May 2009
ISSN 0031-3203
ELSEVIER
PATTERN
RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
Available online at
ScienceDirect
www.sciencedirect.com

Pattern Recognition 42 (2009) 645 – 654

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

A feature selection technique for generation of classification committees and its application to categorization of laryngeal images

M. Bacauskiene[a], A. Verikas[a,b,*], A. Gelzinis[a], D. Valincius[a]

[a]Department of Applied Electronics, Kaunas University of Technology, Studentu 50, LT-51368 Kaunas, Lithuania
[b]Intelligent Systems Laboratory, Halmstad University, Box 823, S 301 18 Halmstad, Sweden

ARTICLE INFO

Article history:
Received 25 August 2007
Received in revised form 12 August 2008
Accepted 26 August 2008

Keywords:
Feature selection
Variable selection
Classification committee
Genetic search
Support vector machine
Laryngeal image

ABSTRACT

This paper is concerned with a two phase procedure to select salient features (variables) for classification committees. Both filter and wrapper approaches to feature selection are combined in this work. In the first phase, definitely redundant features are eliminated based on the paired *t*-test. The test compares the saliency of the candidate and the noise features. In the second phase, the genetic search is employed. The search integrates the steps of training, aggregation of committee members, selection of hyper-parameters, and selection of salient features into the same learning process. A small number of genetic iterations needed to find a solution is the characteristic feature of the genetic search procedure developed. The experimental tests performed on five real-world problems have shown that significant improvements in classification accuracy can be obtained in a small number of iterations if compared to the case of using all the features available.

1. Introduction

Combining outputs of multiple predictors into a committee (ensemble) output is one of the most important techniques for improving prediction accuracy [1–5]. An efficient committee should consist of predictors that are not only very accurate, but also diverse in the sense that the predictor errors occur in different regions of the input space [6–8]. Manipulating training data set, using different architectures, and employing different subsets of variables are the most popular approaches used to achieve the diversity. To promote diversity of neural networks aggregated into a committee, Liu and Yao [9,10] proposed the so-called *negative correlation learning* approach, according to which, all individual networks in the committee are trained simultaneously, using an error function augmented with a correlation penalty term. In Ref. [11], aiming to find a trade-off between the accuracy and diversity of committee networks, the approach was extended by integrating into the same learning process also the feature selection step. However, to assess and control diversity of predictors and to find the trade-off between the accuracy and diversity is not a trivial task [8,12,13]. For instance, feature selection may influence the quality of a committee in several ways, namely by reducing model complexity, promoting diversity of committee members, and affecting the trade-off between the accuracy and diversity of committee members. Moreover, growing size of data sets, in terms of data points and features, increases the demand for feature selection [14]. Therefore, it seems promising to integrate the steps of training, hyper-parameter and feature selection, and aggregation of members into a committee into the same learning process and to use the prediction accuracy for assessing the quality of the committee.

This paper is concerned with such an approach. The main emphasis of the paper is on feature selection for classification committees. A large variety of feature selection techniques have been proposed for a single predictor [15,16], ranging from the sequential forward selection or backward elimination [17,18], sequential forward floating selection [19] to the genetic [20] or tabu search [21].

Usually feature selection methods are categorized as being based on *filter* or *wrapper* approaches [22]. Sometimes three groups of feature selection methods, namely *filter*, *wrapper*, and *embedded* are distinguished [23,24]. However, *wrapper* and *embedded* approaches are closely related [24]. Filter methods assess the saliency of feature subsets from the data properties without involving the induction algorithm (a classifier in our case). Filter approaches assessing and ranking features individually, on the basis of some feature saliency measure, are sometimes called marginal filters [24]. Marginal filters ignore interaction between features. Such a marginal filter approach is adopted in this work.

[*] Corresponding author at: Intelligent Systems Laboratory, Halmstad University, Box 823, S 301 18 Halmstad, Sweden. Tel.: +46 35 167140; fax: +46 35 216724.
E-mail addresses: marija.bacauskiene@ktu.lt (M. Bacauskiene), antanas.verikas@hh.se (A. Verikas), adas.gelzinis@ktu.lt (A. Gelzinis), donatas.valincius@elinta.lt (D. Valincius).

0031-3203/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2008.08.025

The predictor output sensitivity [25–28] is the most popular measure used to assess the saliency of individual features. Eq. (1) exemplifies such a measure [25,26]

$$\Upsilon_i = \frac{1}{QP}\sum_{j=1}^{Q}\sum_{p=1}^{P}\left|\frac{\partial y_{jp}}{\partial x_{ip}}\right| \tag{1}$$

where y is the predictor output, Q is the number of outputs, P is the number of training samples, and x_{ip} is the i th component of the p th input vector $\mathbf{x}_p$.

However, a saliency measure alone does not indicate how many of the candidate features should be used. The problem is how to find a cut-off point of the ranked list. Torkkola and Tuv suggest using artificial-contrast variables created by randomly permuting values of original N variables across the data points [29]. Some of feature selection procedures are based on making comparisons between the saliency of the candidate and the noise feature [25,26]. One of the main drawbacks of the feature saliency measures is that the measures do not have direct relation to the prediction error. By contrast, in wrapper methods, usually a learner (a classifier in our case) is build and used to evaluate a feature subset. Therefore, the approach is computationally prohibitive for large sets of features.

There are a large variety of techniques to select variables for a single predictor. However, works on feature selection for classification or regression committees are not so numerous [7,30–43]. In the next section, we briefly review the existing techniques. Studying the results obtained by the different authors it seems that the genetic search and wrapper approach based techniques are the most promising techniques for ensemble variable selection. However, genetic search based pure wrapper approaches are computationally prohibitive for large sets of variables.

To mitigate the computational burden problem, we combine both the marginal filter and wrapper approaches in this work. The procedure developed to select salient variables for classification committees consists of two phases. In the first, marginal filter, phase clearly redundant features are eliminated based on the paired t-test comparing the saliency of the candidate feature and the noise feature in a single classifier. The noise feature is assumed to be a Gaussian random variable with zero mean and a given variance. The classifier output sensitivity with respect to the feature is the feature saliency measure utilized. However, any other saliency measure can be applied. Then, in the second phase, the genetic search integrating the steps of training, aggregation of committee members into a committee, search for the optimal hyper-parameter values, and selection amongst the features remaining from the first phase into the same learning process is employed. The committee prediction accuracy is the measure used to assess the committee quality in the genetic search. A small number of genetic iterations needed to find a solution is the characteristic feature of the genetic search procedure developed. The rationale of using the first phase of the procedure is to reduce the computation time needed for the genetic search, since the marginal filter based feature ranking is very fast. We expect that features eliminated in the first phase will not deteriorate the classification accuracy significantly, since eliminated are only those features the statistical saliency of which do not exceed the saliency of the noise feature. If the computation time is not a problem, the first phase of the procedure can be skipped. We use an SVM as a committee member in our tests. However, other types of classifiers can also be utilized.

2. Related work

It has been demonstrated that even simple random sampling in the feature space may be an effective technique for increasing the accuracy of classification committees [30,31]. In Ref. [32], random sampling in the feature space is also used for creating classification committees. First, a relevant variable set size M is found by assessing the classification accuracy of variously sized random subsets of variables. In the next stage, randomly selected M-variable subsets are evaluated. In Ref. [33], random sampling in the feature space is used to create the linear discriminant analysis-based ensemble. In Ref. [44], a slightly modified version of the random sampling technique is utilized. The total number of features used N is equal to $N = N_0 + N_1$, where the first N_0 dimensions are given by the first eigenvectors of the data covariance matrix. The remaining N_1 dimensions are selected randomly. Guo et al. [34] use the random forest algorithm [45] to select the most useful variables. The random forest algorithm ranks the variables in terms of their contribution to the predictive accuracy. To further refine the selected variable subsets, the correlation-based [46] and gain ratio [47] feature selection algorithms were applied.

An early work on genetic algorithms (GA)-based feature selection is by Siedlecki and Sklansky [48]. Vafaie and De Jong applied GAs to select features for rule based classifiers [49]. In Refs. [35,36], GAs have been used for ensemble feature selection by exploring all possible feature subsets. However, only one ensemble was considered in these works. Kim et al. proposed meta-evolutionary ensembles considering multiple ensembles simultaneously [37]. Genetic operators change the size of ensembles and the membership of individual classifiers over time. Each classifier is initialized with randomly selected features and a random ensemble assignment. The representation of a classifier consists of $N + \log_2(G)$ bits, where N is the number of features and G is the maximum number of ensembles. It is believed that feature selection results into diverse ensemble members. Altmcay has also used GAs to explore all possible feature subsets when creating ensembles of evidential k-NN classifiers [38]. As in Ref. [50], each chromosome is designed to represent a different ensemble. In Ref. [51], a GA based technique to create the weighted majority voting-based classification committee is used. The resampled training sets and classifier prototypes included into the committee evolve during training. However, all committee members use the same features.

Several other techniques to select features for committees have been developed. Reduction of the computation burden in one way or another is the common feature of the techniques. Grimaldi et al. apply feature selection to each member of the committee independently [52]. In Ref. [53], a binary decision tree is grown by using a feature with minimal impurity at each node of the tree. Different feature subsets are selected for different resampled learning data sets. Having the ensemble of feature subsets, the features are ranked according to some relevance measure and the most relevant features are used. Cho and Ryu evaluate information content of variables using several measures [54]. Separate committee members are trained independently using the most mutually exclusive variables, as deemed by pairs of the measures. If there are N variables, the most mutually exclusive variables are chosen through the correlation analysis of all possible variable pairs. In a two class classification problem considered in Ref. [55], patterns of the classes are independently partitioned into clusters. Features are then selected independently for each cluster by using those maximizing the distance between the cluster and all patterns of the other class. In Ref. [56], forward sequential selection and backward sequential selection approaches are utilized to select variables for a classification committee. In input decimated ensembles [39], L classifiers are trained, where L is the number of classes. For each classifier (a multi-layer perceptron), a pre-determined number of features having the highest absolute correlation to the targets (presence or absence of the corresponding class) are selected. Each classifier is trained to discriminate amongst all L classes. The technique leads to different variable subsets for different classifiers. It is demonstrated that the technique outperforms ensembles exploiting all the input features, randomly selected features, and features created using PCA. In Ref. [40], a number of training sets are created using a bootstrapping protocol. For each training set a subset of variables is selected based on variable weights in a linear SVM. All the selected variables are used to create a bagged ensemble of nonlinear SVM. In Ref. [41], to create one of L base classifiers in rotation forest, the variable set is randomly split into K (parameter of the technique) subsets and PCA is applied to each subset. All PCs are retained and used as features to train the classifier. It was found that classifiers in the rotation forest are more accurate than those in random forest and AdaBoost. In the technique proposed in Ref. [42], a variable may only be used by one classifier and a selected variable cannot be abandoned. The technique proposed in Ref. [43] may in principle exploit any feature selection algorithm. Feature subsets for committee members are built sequentially trying to prevent the selection algorithm from returning the same feature-subset multiple times. In Ref. [7], a half-&-half bagging and feature selection-based technique for incremental building of classification committees has been proposed.

3. Procedure

By combining the marginal filter and the genetic search based wrapper approaches to feature selection, this work aims at developing a computationally feasible procedure to select features for classification committees as well as to determine suitable hyper-parameter values of the committee members. To assess the feature saliency in the first, marginal filter-based, selection phase, the statistical paired t-test is exploited. K different random splits of the data set into learning, validation, and test subsets are used to estimate the t-statistic. Thus, the t-test allows mitigating the problem of the dependency of the feature saliency assessment results on the way how the data set is split into learning, validation, and test subsets. The genetic search performed in parallel for all committee members aiming to determine the feature sets and the hyper-parameters, provides the flexibility needed to find the trade-off between the accuracy and diversity of the committee members.

The procedure for committee variable selection is summarized in the following steps.

(1) Augment the input vectors with one additional Gaussian noise feature with mean $m=0$ and a given standard deviation s. The value of s depends on the number of initial features. The larger the number, the smaller is the s value. The standard deviation $s=1$ worked well in all our test with the relatively small number of features (up to $\simeq 50$). For the feature sets containing 150–200 features, $s=0.3$ was a good choice.

(2) Randomly assign the available data points into learning S_l, validation S_v, and test S_t data sets, for example 50% for learning, and 25% each for validation and testing.

(3) Train the model. The learning set is used to train the model, the validation set is used to select hyper-parameters of the model. Since we use an SVM with the Gaussian kernel as the model in our tests, the hyper-parameters considered are the kernel width and the regularization constant. We used the GA described in Section 3.3, to find the appropriate hyper-parameter values. The search range for the width parameter σ was limited to the interval given by 0.1 and 0.9 quintile of the $\|\mathbf{x}_i-\mathbf{x}_j\|_2$ statistic, where $\mathbf{x}_i$ denotes the i th data point. The search interval for the regularization constant was limited by the maximum value found from several preliminary experiments. Observe that values for only these two hyper-parameters of a single SVM are determined during the search. To find the appropriate hyper-parameter values, one can also use the Nelder–Mead simplex method [57], for example.

(4) Calculate the saliency score Γ_i,

$$\Gamma_i = \frac{\Upsilon_i}{\max_{l=1,\dots,N}\Upsilon_l}, \quad i=1,\dots,N \tag{2}$$

where Υ_i is given by Eq. (1) and N is the number of features.

(5) Repeat Steps (2)–(4) K times, where the user defined parameter K refers to the number of different random partitions of the data set into learning, validation, and test subsets.

(6) Eliminate the features whose saliency does not exceed the saliency of the noise feature, according to the paired t-test.

(7) Choose the number of committee members L. Construct a chromosome characterizing feature inclusion/noninclusion, regularization, and kernel (in the case of SVM based committee members) parameters of all the committee members. More details on the chromosome definition are given in Section 3.3.

(8) Perform the genetic search.

(9) The committee is given by the parameters encoded in the "best" chromosome found during the genetic search.

3.1. The paired t-test

To assess the equality of the mean saliency of i th feature μ_{Γ_i} and the noise μ_{Γ_n} the paired t-test is defined as suggested in Ref. [26]: Null Hypothesis $\mu_{D_i}=0$, Alternative Hypothesis $\mu_{D_i}>0$, where $\mu_{D_i}=\mu_{\Gamma_i}-\mu_{\Gamma_n}$. To test the null hypothesis, a t^* statistic

$$t^* = \frac{\overline{D}_i}{s_{\overline{D}_i}} \tag{3}$$

is evaluated, where $\overline{D}_i = K^{-1}\sum_{j=1}^{K} D_{ij}$, $D_{ij}=\Gamma_{ij}-\Gamma_{nj}$, Γ_{ij} and Γ_{nj} are the saliency scores computed using Eq. (2) for the i th and the noise feature, respectively, in the j th loop, and

$$s_{\overline{D}_i} = \sqrt{\frac{\sum_{j=1}^{K}(D_{ij}-\overline{D}_i)^2}{K(K-1)}} \tag{4}$$

Under the null hypothesis, the t^* statistic is t-distributed. If $t^* > t_{crit}$, the hypothesis that the difference in the means is zero is rejected, where t_{crit} is the critical value of the t distribution with $\nu = K-1$ degrees of freedom for a significance level of α: $t_{crit}=t_{1-\alpha,\nu}$.

3.2. The SVM output sensitivity, an example

The output of a support vector machine $y(\mathbf{x})$ is given by

$$y(\mathbf{x}) = \sum_{j=1}^{N_s} \alpha_j^* d_j \kappa(\mathbf{x}_j,\mathbf{x}) + b \tag{5}$$

where N_s is the number of support vectors, $\kappa(\mathbf{x}_j,\mathbf{x})$ is a kernel, d_j is a target value ($d_j=\pm 1$), and the threshold b and the parameter α_j^* values are found as a solution to the optimization problem defined by the type of SVM used. In this work, we used the 1-norm soft margin SVM [58]. The parameters α_j satisfy the following constrains:

$$\sum_{j=1}^{N_s}\alpha_j y_j = 0, \quad \sum_{j=1}^{N_s}\alpha_j = 1, \quad 0 \leqslant \alpha_j \leqslant C,\ j=1,\dots,N_s \tag{6}$$

with C being the regularization constant.

For the Gaussian kernel used in this work $\kappa(\mathbf{x}_j,\mathbf{x}_k)=\exp\{-\|\mathbf{x}_j-\mathbf{x}_k\|^2/\sigma\}$, where σ is the standard deviation of the Gaussian, having the j th input vector $\mathbf{x}_j$ presented to the input, the derivative of the

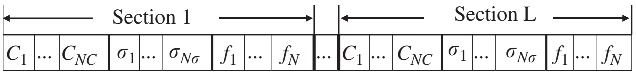
Section 1
Section L
C_1 ... C_{NC} σ_1 ... $\sigma_{N\sigma}$ f_1 ... f_N ... C_1 ... C_{NC} σ_1 ... $\sigma_{N\sigma}$ f_1 ... f_N

Fig. 1. The structure of the chromosome consisting of L sections.

output with respect to the i th feature is given by

$$\frac{\partial y(\mathbf{x}_j)}{\partial x_{ij}} = -\frac{2}{\sigma}\sum_{k=1}^{N_s}\alpha_k^* d_k(x_{ij}-x_{ik})\exp\left\{-\sum_{n=1}^{N}\frac{(x_{nj}-x_{nk})^2}{\sigma}\right\} \qquad (7)$$

The feature saliency measure used in this work is based on the derivative of the classifier output. However, other saliency measures, which do not require differentiable classifier outputs can be applied.

3.3. Genetic search

Information representation in a chromosome, generation of initial population, evaluation of population members, selection, crossover, mutation, and reproduction are the issues to consider when designing a genetic search algorithm.

In this work, a *chromosome* contains all the information needed to build a committee. We consider SVM-based committees and divide the chromosome into sections and each section into parts. The number of sections is equal to the number of committee members L. There are three parts in each section. One part encodes the regularization constant C, one the kernel width parameter σ, and the third one encodes the inclusion/noninclusion of features. The binary encoding scheme has been adopted in this work. Fig. 1 illustrates the chromosome structure, where $C_1 \cdots C_{NC}$ refers to bits used to encode the regularization constant C, $\sigma_1 \cdots \sigma_{N\sigma}$ illustrates bits used to encode the kernel width parameter σ, with NC and $N\sigma$ standing for the number of bits used to encode C and σ, respectively. Features are denoted by $f_1, \ldots, f_N$, where N is the number of features and f_i is 1 or 0 meaning that the i th feature is included or not included into the feature set, respectively.

To generate the *initial population*, information obtained in the first feature selection phase, namely, the maximum number of features, and the initial values of C and σ are exploited. The maximum number of features allowed for one committee member is equal to the number of features determined in the first phase. In the initial population, the features obtained from the first phase are masked randomly and values of the parameters C and σ are chosen randomly from the interval $[C_0 - \Delta C, C_0 + \Delta C]$ and $[\sigma_0 - \Delta\sigma, \sigma_0 + \Delta\sigma]$, respectively. For the width parameter σ, the interval is given by 0.1 and 0.9 quintile of the $\|\mathbf{x}_i - \mathbf{x}_j\|_2$ statistic and the parameter σ_0 is set to the midpoint of the interval. The search interval for the regularization constant C is limited by 0 and the maximum value, C_{max}, found from several preliminary experiments.

The *fitness function* used to evaluate the chromosomes is given by the classification accuracy of the validation set data. The committee output was obtained by averaging the outputs of committee members. Other aggregation approaches can also be applied. For instance, if using the weighted averaging aggregation scheme, the aggregation weights can also be found during the genetic search. To distinguish between more than two classes, the one vs. one pairwise-classification scheme has been used.

The *selection process* of a new population is governed by the fitness values. A chromosome exhibiting a higher fitness value has a higher chance to be included in the new population. The selection probability of the i th chromosome p_i is given by

$$p_i = \frac{r_i}{\sum_{j=1}^{M} r_j} \qquad (8)$$

where r_i is the classification accuracy obtained from the committee encoded in the i th chromosome and M is the population size.

The *crossover operation* for two selected chromosomes is executed with the probability of crossover p_c. If a generated random number from the interval [0,1] is smaller than the crossover probability p_c, the crossover operation is executed. Crossover is performed separately in each section of a chromosome. The crossover point is randomly selected in the "feature mask" part and two parameter parts of each section and the corresponding parts of two chromosomes selected for the crossover operation are exchanged at the selected points.

The *mutation operation* adopted is such that each gene is selected for mutation with the probability p_m. The mutation operation is executed independently in each part of each chromosome section. If the gene selected for mutation is in the feature part of the chromosome, the value of the bit representing the feature in the feature mask (0 or 1) is reversed. To execute mutation in the parameter part of the chromosome, the value of the offspring parameter is mutated by $\pm\Delta\gamma$, where γ stands for C or σ, as the case may be. The mutation sign is determined by the fitness values of the two chromosomes, namely the sign resulting into a higher fitness value is chosen. Thus, to perform the mutation, the computationally expensive fitness evaluation needs to be performed. However, such a choice reduces the total number of genetic iterations required. Furthermore, the mutation probability used is rather low. The way of determining the mutation amplitude $\Delta\gamma$ is somewhat similar to that used in Ref. [59] and is given by

$$\Delta\gamma = w\beta(\max(|\gamma - \gamma_{p1}|, |\gamma - \gamma_{p2}|)) \qquad (9)$$

where γ is the actual parameter value of the offspring, $p1$ and $p2$ stand for parents, $\beta \in [0, 1]$ is a random number, and w is the weight decaying with the iteration number:

$$w = k(1 - t/T) \qquad (10)$$

where t is the iteration number, T is the total number of iterations, and the constant k is chosen experimentally. The constant defines the initial mutation amplitude. The value of $k = 0.4$ worked well in our tests.

In the *reproduction process*, the newly generated offspring replaces the chromosome with the smallest fitness value in the current population, if a generated random number from the interval [0,1] is smaller than the reproduction probability p_r or if the fitness value of the offspring is larger than that of the chromosome with the smallest fitness value.

4. Experimental investigations

4.1. Data used

To test the approach we used five real-world problems. Data characterizing four of the problems are: *US congressional voting records problem*, *The diabetes diagnosis problem*, *Wisconsin breast cancer problem* and *Wisconsin diagnostic breast cancer* (*WDBC*) *problem* are available at: http://www.ics.uci.edu/~mlearn/. The fifth problem concerns classification of laryngeal images [60,61].

Laryngeal images. The task is to automatically categorize color laryngeal images (images of vocal folds) into the *healthy*, *nodular*, and *diffuse* decision classes [60]. Fig. 2 presents characteristic examples from the three decision classes considered.

Due to a large variety of appearance of vocal folds, the categorization task is sometimes difficult even for a trained physician. Fig. 3 provides an example of such a task. The image placed on the right-hand side of the figure comes from the *nodular* class, while the other two are taken from the *healthy* vocal folds. In this case, the only

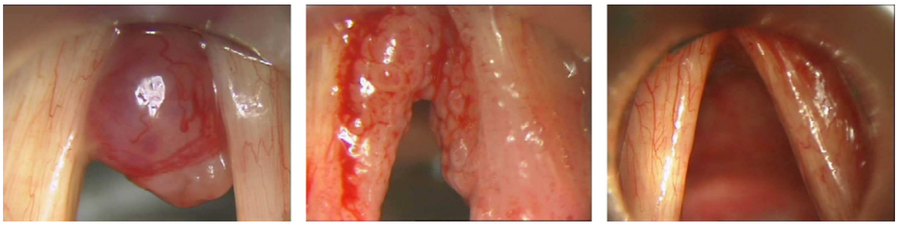

Fig. 2. Images from the *nodular* (left), *diffuse* (middle), and *healthy* (right) classes.

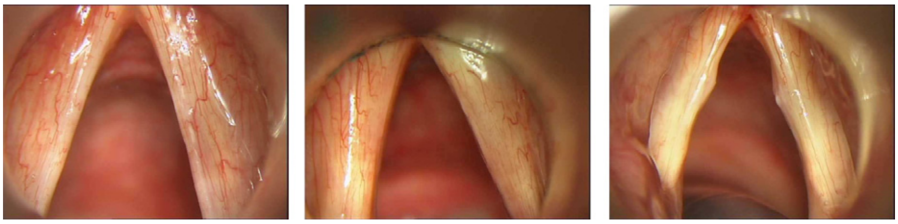

Fig. 3. An example of a laryngeal image coming from the *nodular* class (the right-most) along with two images from the *healthy* class. The slightly convex vocal fold edges in the upper part of the *nodular* class image is the only discriminative feature.

discriminative feature is the slightly convex vocal fold edges in the upper part of the image coming from the *nodular* class.

Aiming to obtain a comprehensive description of laryngeal images, multiple feature sets exploiting information on image color, texture, geometry, image intensity gradient direction, and frequency content are extracted [61]. Image color distribution, distribution of the image intensity gradient direction, parameters characterizing the geometry of edges of vocal folds, distribution of the spectrum of the Fourier transform of the color image complex representation (two types of the frequency content based features), and parameters calculated from multiple co-occurrence matrices are the six sets of features used to describe laryngeal images [61]. Here we present a brief description of the feature sets used.

4.1.1. *Color features*

The approximately uniform $L^*a^*b^*$ color space [62] was employed to represent colors. We characterize the color content of an image by the probability distribution of the color represented by the 3-D color histogram of $N = 4096$ $(16 \times 16 \times 16)$ bins and consider the histogram as an N-vector. Most of bins of the histograms were empty or almost empty. Therefore, to reduce the number of components of the N-vector, the histograms built from a set of training images were summed up and the N-vector components corresponding to the bins containing less than N_α hits in the summed histogram were left aside. Hereby, when using $N_\alpha = 10$ we were left with 733 bins—a vector of 733 components. The color features $\mathbf{x}_C$ are then given by the first 50 principal components computed using the 733-component vector.

4.1.2. *Texture features*

A polynomial $f_k(d)$ of degree n

$$f_k(d) = x_{k0} + x_{k1}d + \cdots + x_{kn-1}d^{n-1} + x_{kn}d^n \tag{11}$$

is fitted to the normalized values of each (k) of the 14 well-known Haralick's coefficients calculated from the co-occurrence matrices evaluated for several distance parameter values [63]. The coefficients are calculated from the average co-occurrence matrix obtained by averaging the matrices calculated for 0°, 45°, 90°, and 135° directions. The distance parameter d values utilized to calculate the co-occurrence matrices were $d = 1,\ 3,\ 5,\ 7,\ 9,\ 11,\ 13,\ 15$. Parameters of the polynomials were then used as components of the measurement vector $\mathbf{x}_T$. Thus, the number of components in the feature vector $\mathbf{x}_T$ is equal to $(n+1) \times 14$. The polynomial degree n is selected experimentally. The set of features defined by the second order polynomial provided the best performance.

4.1.3. *Fourier spectrum based features*

Let $P(u, v)$ be the Fourier spectrum of a color image [61]. The upper part of the frequency plane is divided into M equidistant wedges W_i and the average power

$$\overline{P}_i = \frac{1}{N_{Wi}} \sum_{u,v \in W_i} P(u, v), \quad i = 1, \ldots, M \tag{12}$$

is computed in each of the wedges, where N_{Wi} is the number of distinct frequencies in the wedge W_i. The $\overline{P}_i$ values constitute the feature vector $\mathbf{x}_{F1}$.

To extract the image frequency content based features of the second type, the frequency plane is divided into several rings R_i of different average frequency. The Chi-square χ_i and the entropy M_i of the Fourier power computed in each of the rings constitute the feature vector $\mathbf{x}_{F2}$ [61].

4.1.4. *Image intensity gradient based features*

The features are based on the distribution of the image intensity gradient direction. The direction angle $\alpha(x, y)$ of the image intensity gradient vector $\nabla\mathbf{L}$ is given by

$$\alpha(x, y) = \tan^{-1}(G_x/G_y) \tag{13}$$

where $\nabla\mathbf{L} = [G_x\ G_y]^{\mathrm{T}} = [\frac{\partial L}{\partial x}\ \frac{\partial L}{\partial y}]^{\mathrm{T}}$. The Sobel operator, which consists of a pair of 3×3 convolution masks has been used to estimate the image intensity gradient. We use a histogram to represent the distribution of the angle $\alpha(x, y)$ and the histogram vector $\mathbf{x}_\nabla$ as a feature vector of this type. The number of histogram bins was found experimentally and was equal to 1000.

4.1.5. *Geometrical features*

Geometrical features we use are mainly targeted to characterize the shape of edges of two vocal folds [61]. Two thirds of the lower part of laryngeal images are used in the analysis. Various techniques can be applied to extract color edges. In this work, we exploited the

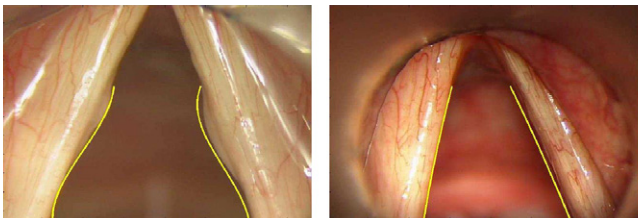

Fig. 4. Laryngeal images coming from the *nodular* (left) and *healthy* (right) classes along with two third order curves used to calculate the geometrical features $\mathbf{x}_G$.

technique based on the difference vector operators [64]. In the next step, all the edge pixels were sorted out into connected components and the small ones were eliminated. Two large connected components, one on the left- and the other on right-hand side of the image being analyzed were then found and used in further analysis.

Having the two connected components, three polynomial curves given by Eq. (14)—one of the first, one of the second, and one of the third order—were fitted to pixels of each of the components

$$p(x) = p_0 + p_1 x + \cdots + p_{n-1}x^{n-1} + p_n x^n, \quad n = 1, 2, 3 \tag{14}$$

Thus, in total, we have 18 parameters p_i characterizing the six curves. The parameters p_i constitute the feature vector $\mathbf{x}_G$. Fig. 4 presents two examples of laryngeal images coming from the healthy and nodular classes along with the third order polynomial curves found.

4.2. Experimental setup

In all the tests, we run an experiment 30 times with different random partitioning of the data set into *learning*, S_l, *validation*, S_v, and *test*, S_t data sets. The learning set is used to train the model, the validation set is used to select hyper-parameters of the model. The mean values and standard deviations of the classification accuracy presented in this paper were calculated from these 30 trials using the test set data. The parameter values used in the genetic search have been found experimentally. The following values worked well in all the tests: $p_c = 0.95$, $p_m = 0.02$, and $p_r = 0.05$.

In the case of laryngeal images, a separate SVM is used to categorize features of each type into the decision classes. The final image categorization is then obtained based on the decisions provided by a committee of support vector machines. In this work, there were 49 images from the *healthy* class, 406 from the *nodular* class, and 330 from the *diffuse* class. Out of the 785 images available, 650 images were assigned to the learning set, 75 to the test set, and 60 to the validation set.

4.3. Results

First, the average test data set classification accuracy obtained from a single SVM without any involvement of the designing procedure proposed was estimated. The optimal values of the regularization constant C and the kernel width σ have been selected experimentally. To select the values, a "qualified guess" was made from several experiments, first. Then, several loops were run to refine the values by keeping one parameter fixed and adjusting the other one, interchangeably.

The upper part of Table 1 presents the average test data set classification accuracy obtained for the first four data sets from a single SVM when using all the original features in the classification process. The number of classes and the number of features available are also given in the table. In the parentheses, the standard deviation of the classification accuracy is provided. The average test data set classification accuracy obtained when using a separate SVM for each type of features extracted from the laryngeal images is presented in Table 2.

Table 1
The average test data set classification accuracy obtained for the different data sets from a single SVM when using: all the original features (upper part) and the selected features (lower part)

Data set	Number of classes	Number of features	Accuracy, % (std. deviation)
Diabetes	2	8	76.9 (1.6)
WBCD	2	9	96.9 (0.8)
Voting	2	16	95.5 (1.0)
WDBC	2	30	97.2 (1.0)
Data set	Average number of iterations	Average number of selected features	Accuracy, % (std. deviation)
Diabetes	8	4	77.6 (1.5)
WBCD	7	6	97.2 (0.6)
Voting	12	3	96.3 (1.0)
WDBC	20	17	98.1 (0.7)

Table 2
The average test data set classification accuracy obtained when using a separate SVM for each type of features extracted from the laryngeal images

Feature type	Number of classes	Number of features	Accuracy, % (std. deviation)
Gradient	3	1000	52.3 (5.8)
Co-occurrence	3	42	83.6 (3.2)
Frequency (*F1*)	3	180	83.4 (3.4)
Frequency (*F2*)	3	40	78.0 (3.0)
Geometrical	3	18	69.2 (3.5)
Color	3	50	91.8 (2.7)

In the next experiment, we studied the effectiveness of the feature selection procedure applied to single SVMs. Both, the feature subsets and the hyper-parameter values were found using the genetic search procedure. The lower part of Table 1 summarizes the results of the test concerning the first four problems. Apart from the average test data set classification accuracy obtained using the selected features, the table also provides the number of selected features and the number of genetic iterations required to achieve the solution. The number of features eliminated in the first selection phase has been equal to 1, 1, 6, and 12 for the *Diabetes*, *WBCD*, *Voting*, and *WDBC* databases, respectively. Observe that the first two problems are characterized by 8 and 9 features, respectively. Thus, there are very few clearly redundant features. The larger number of features eliminated in the first phase for the other two problems significantly speeds up the genetic search executed in the second phase.

As can be seen from Table 1, for all the databases, the average classification accuracy obtained from the single SVMs trained on the selected feature sets is higher than that achieved using all the features available. We remind that the manual and genetic search based tuning of the hyper-parameters (C and σ) was used to obtain results presented in the upper and the lower part of Table 1, respectively. Thus, one may wonder about the source of improvement in classification accuracy. Is the improvement due to the genetic search based parameter tuning or feature selection? Observe that in the case of manual parameter tuning, values of the parameters were selected very carefully. Thus, the improvement is due to feature selection. Results presented in Tables 2 and 3, where the genetic search based parameter tuning is applied in both cases, substantiate this observation. The improvement in classification accuracy obtained for the data presented in Table 1 is rather marginal and is likely to be statistically insignificant for high confidence values. However, as we will see shortly, a considerable improvement in classification accuracy is obtained when classifying laryngeal images represented by a large number of features. The number of genetic iterations needed to achieve the solutions is very small. The population size was equal to 50 in all the tests. Fig. 5 provides two graphs plotting the classification accuracy as a function of the number of genetic iterations for the *WDBC* and *Voting* databases. For each genetic iteration, the performance of the best population member is shown in Fig. 5. The performance achieved by the best member at the end of the search procedure (*max*) is also shown.

The results obtained for the different feature sets characterizing the laryngeal images are summarized in Table 3. The number of features eliminated in the first feature selection phase ranged from 5 to over 700. As can be seen from Tables 2 and 3, a considerable improvement in classification accuracy has been obtained using the proposed SVM designing approach. Observe that even for a single SVM, the genetic search technique was utilized to find both the feature subset and the hyper-parameter values. The number of features chosen is considerably lower than that presented in Table 2, especially for the *Gradient* and *Frequency* (F 1) feature types. On average, a very small number of genetic iterations was required to find the solutions. Fig. 6 provides two graphs plotting the correct classification rate as a function of the number of genetic iterations for the two types of frequency features. For each genetic iteration, the performance of the best population member is shown.

In the last experiment, the effectiveness of the feature selection procedure applied to SVM committees has been studied. Table 4 summarizes the results of the experiment.

All the committees were made of six members, $L = 6$. All six members of the committees built for solving the first four problems used the same initial feature set. Each member of the committee built for solving the Laryngeal problem utilized a different feature set—one of the six available types. For all committee members, features in a chromosome encoding a committee in the initial population were

Table 3
The average test data set correct classification rate obtained for the different types of features extracted from laryngeal images when using a separate SVM for each type of selected features

Feature type	Average number of selected features	Average number of iterations	Accuracy, % (std. deviation)
Gradient	362	17	83.7 (4.4)
Co-occurrence	28	13	85.5 (3.6)
Frequency ($F1$)	78	37	89.7 (2.4)
Frequency ($F2$)	29	13	79.6 (3.5)
Geometrical	10	13	72.1 (3.5)
Color	42	13	92.7 (2.6)

Table 4
The average test data set classification accuracy obtained for the different data sets from a committee when using the selected features

Data set	Average number of selected features	Average number of iterations	Accuracy, % (std. deviation)
Diabetes	5	8	77.7 (1.5)
WBCD	5	14	97.3 (0.6)
Voting	6	37	96.6 (0.8)
WDBC	9	20	98.3 (0.5)
Laryngeal	28	8	95.0 (1.9)

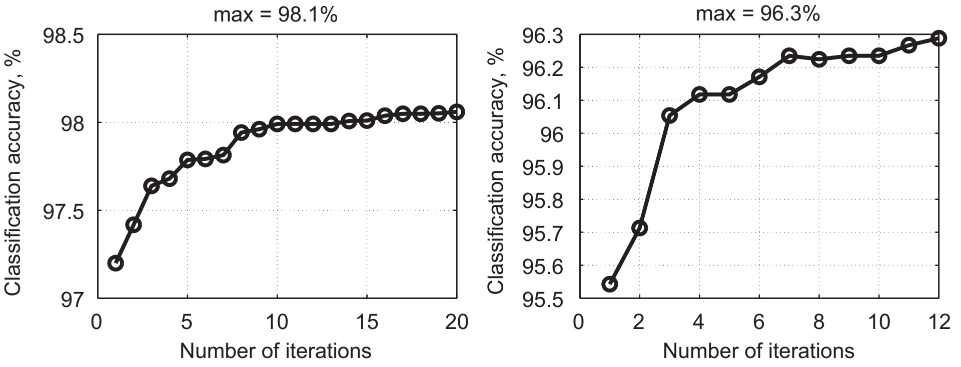

Fig. 5. The test data set classification accuracy obtained from a single SVM as a function of the number of genetic iterations for the Wisconsin diagnostic breast cancer (left) and the US congressional voting records (right) data sets.

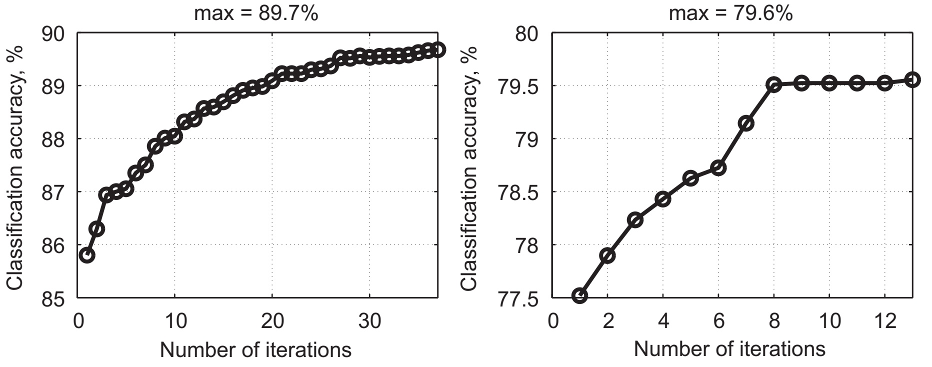

Fig. 6. The test data set classification accuracy obtained from a single SVM as a function of the number of genetic iterations for the two types of frequency features extracted from the laryngeal images.

masked randomly. The average test data set classification accuracy, the average number of features used by one committee member, and the number of iterations needed to obtain the solution are given in Table 4. As can be seen from Table 4, the technique developed is capable of evolving accurate classification committees in a small number of genetic iterations.Table 5 provides the main parameters of members of the committee evolved for categorizing the laryngeal images. Comparing the results presented in Tables 3 and 5 one can find out that the committee members use less features and are less accurate than the separately designed classifiers. However, as it will be demonstrated in the next section, the more accurate individual classifiers when aggregated into a committee provide lower accuracy than that obtained from the committee designed by evolving all its members in parallel.

Fig. 7 provides two graphs plotting the test data set classification accuracy obtained from the committees as a function of the number of genetic iterations for the laryngeal (left) and the WDBC (right) problems.

4.4. Comparisons

Aiming to explore the effectiveness of the proposed procedure for designing classification committees, several comparisons in terms of classification accuracy and computation time have been performed. The set of laryngeal images, represented by the six different feature sets, has been used in the experiments. In the first experiment, the first feature selection phase based on the marginal filter approach was eliminated and the genetic search started from all the available features.

Next, the effectiveness of two other techniques for feature elimination in the first phase of the designing procedure has been investigated. The popularity of the techniques determined the choice. The first technique was based on the Fisher index. The Fisher index for feature k and i,j pair of classes is defined as

$$\mathtt{FI}_k = \frac{(\mu_{k,i} - \mu_{k,j})^2}{(N_i - 1)\sigma^2_{k,i} + (N_j - 1)\sigma^2_{k,j}} \tag{15}$$

where the indices i and j refer to the two classes and $\mu_{k,i}$, $\mu_{k,j}$, $\sigma^2_{k,i}$, and $\sigma^2_{k,j}$ are the class means and variances for variable x_k. The $\mathtt{FI}$ is a measure of the between class spread $(\mu_{k,i} - \mu_{k,j})^2$ in relation to the within class spread $[(N_i - 1)\sigma^2_{k,i} + (N_j - 1)\sigma^2_{k,j}]$, with N_i and N_j being the number of data points in the classes. Since we have three classes, the Fisher index was calculated for the three pairs and then the minimum value of these three was used. Features eliminated in the first phase were those exhibiting the lowest values of the Fisher index. The number of eliminated features was equal to that determined by the t-test used in the proposed technique.

The second technique for the feature elimination in the first phase explored was based on the sequential forward feature selection. The first feature selected was the one providing the highest classification accuracy. The k NN classifier has been used to assess the accuracy, due to no needs for training. The number of k was determined by cross-validation and was equal to 9. The feature added at the j th step was the one providing the highest performance when used together with all the features selected up to the j th step. The inclusion process was terminated when the number of features left aside was equal to the number of features eliminated by the t-test. Thus, the features left aside, were features eliminated in the first phase of the designing procedure. The selected features were then used in the genetic search process.

Fig. 8 plots the test data set classification accuracy as a function of the number of genetic iterations for the four committee designing alternatives. The graph labelled with ’▽’ presents the alternative without using the first feature elimination phase. The results obtained using the proposed technique are denoted by ’□’, while ’o’ and ’×’ denote the Fisher index and k NN based approaches, respectively. Fig. 9 illustrates the effectiveness of the four committee designing alternatives in terms of computation time. The simulations were performed using a PC with a 2.4 GHz clock frequency. The software was written using the C^{++} and Matlab programming languages. As can be seen from the figures, the proposed technique outperformed the other three options in both classification accuracy and computation time. If the number of genetic iterations is increased substantially and parameters of the GA are adjusted accordingly when using all the features in the search, one, of course, would be able to achieve the classification accuracy as high as that obtained by the technique proposed. However, the computation time would increase even more.

Table 5
The main parameters of members of the committee evolved for categorizing the laryngeal images

Feature type	C	σ	Number of selected features	Accuracy, %
Gradient	18	0.02	47	81.2
Co-occurrence	23	0.04	28	84.4
Frequency ($F1$)	37	0.04	28	88.6
Frequency ($F2$)	48	0.04	25	79.5
Geometrical	690	0.02	8	71.5
Color	7	0.04	33	92.2

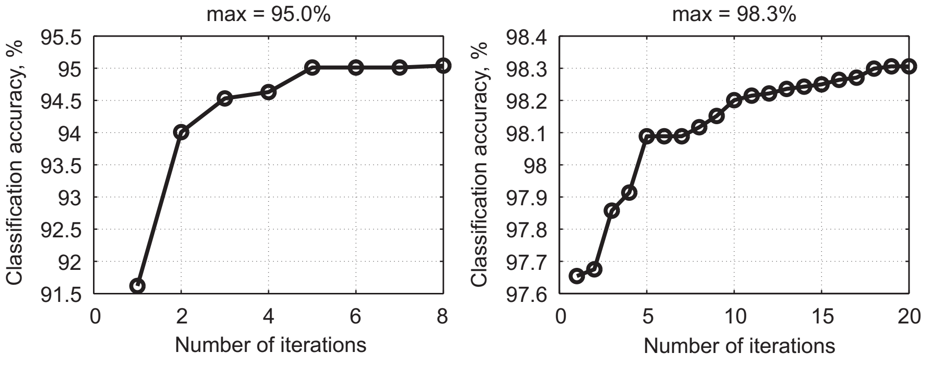

Fig. 7. The test data set classification accuracy obtained from the committee as a function of the number of genetic iterations for the laryngeal (left) and the Wisconsin diagnostic breast cancer (right) data sets.

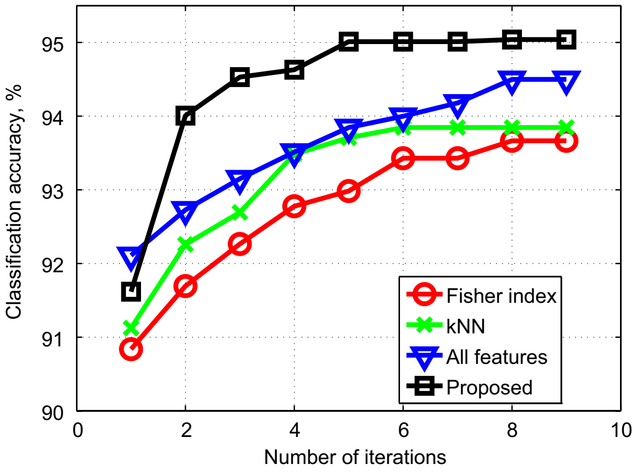

Fig. 8. The test data set classification accuracy as a function of the number of genetic iterations.

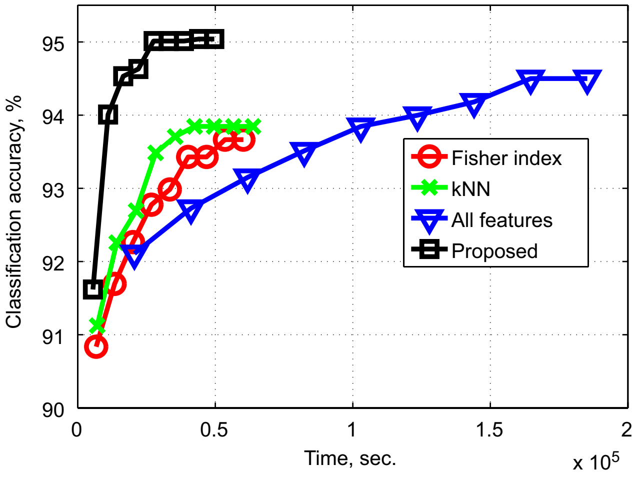

Fig. 9. The test data set classification accuracy as a function of the training time.

In the last experiment, a committee was built using independently designed members, presented in Table 3. Using of independently designed members is a widely adopted approach to creation of classification committees. As mentioned above, the independently designed members are more accurate than those evolved when designing a committee according to the proposed technique, see Table 5. However, the 93.7% classification accuracy achieved from the committee of the independent members was lower than the highest accuracy achieved in this work. Moreover, the number of features used by the independent members is considerably higher.

Thus, the technique proposed outperformed two popular the state-of-the-art approaches to feature selection for classification committees: the genetic search based selection starting the search from all the available features ("All features" curves in Figs. 8 and 9) and selection of features independently for each committee member.

5. Conclusions

A two phase procedure to select salient features for classification committees has been presented in this work. To mitigate the computation burden problem, both the marginal filter and wrapper approaches to feature selection are combined. Elimination of clearly redundant features in the filter approach-based first phase of the procedure speeds up the genetic search executed in the second, wrapper approach-based, phase of the designing process. The paired *t*-test comparing the saliency of the candidate and the noise features is employed for the feature elimination in the first phase.

The experimental tests were performed on: *The US congressional voting records problem*, *The diabetes diagnosis problem*, *Wisconsin breast cancer problem*, *Wisconsin diagnostic breast cancer problem* (http://www.ics.uci.edu/~mlearn/), and the problem of classification of *laryngeal images*. The genetic search integrating the steps of training, aggregation of committee members, selection of hyper-parameters, and selection of salient features into the same learning process allows finding a trade-off between the accuracy and diversity of the members. The committee evolved for classification of laryngeal images substantiates the fact. The committee outperformed its counterpart composed of more accurate members evolved independently. The classification accuracy obtained for these committees was 95.0% and 93.7%, respectively. The experimental investigations have shown that the first phase feature elimination based on the *t*-test is more effective, in terms of the achieved classification accuracy, than the *k* NN and Fisher index based approaches. In the case of feature sets containing a large number of redundant features, the *gradient* and *frequency F1* feature sets characterizing the laryngeal images for example, more than a half of the available features were eliminated in the first phase of the procedure. For these feature sets, the proposed technique allowed to increase the classification accuracy from 52.3% to 83.7% and from 83.4% to 89.7%, respectively. For the *laryngeal images* classification problem characterized by a large number of redundant features, the application of the first feature elimination phase, reduced the genetic search duration almost four times without any increase in the classification error. By contrast, the obtained classification error was slightly (9%) lower. On average, only 10–40 genetic iterations were required to find a solution. However, the genetic iterations can be computationally demanding, since for example the mutation operation entails several fitness evaluations.

Acknowledgments

We gratefully acknowledge the support we have received from the agency for international science and technology development programmes in Lithuania (COST Action IC0602).

References

[1] P.D. Gader, M.A. Mohamed, J.M. Keller, Fusion of handwritten word classifiers, Pattern Recognition Lett. 17 (1996) 577–584.

[2] C.L. Liu, Classifier combination based on confidence transformation, Pattern Recognition 38 (2005) 11–28.

[3] T.G. Dietterich, Ensemble methods in machine learning, in: Lecture Notes in Computer Science, vol. 1857, Springer, Heidelberg, 2000, pp. 1–15.

[4] L.I. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms, Wiley, IEEE, New York, 2004.

[5] A. Verikas, A. Lipnickas, K. Malmqvist, M. Bacauskiene, A. Gelzinis, Soft combination of neural classifiers: a comparative study, Pattern Recognition Lett. 20 (1999) 429–444.

[6] A. Krogh, J. Vedelsby, Neural network ensembles, cross validation, and active learning, in: G. Tesauro, D.S. Touretzky, T.K. Leen (Eds.), Advances in Neural Information Processing Systems, vol. 7, MIT Press, Cambridge, MA, 1995, pp. 231–238.

[7] M. Bacauskiene, A. Verikas, Selecting salient features for classification based on neural network committees, Pattern Recognition Lett. 25 (16) (2004) 1879–1891.

[8] L.I. Kuncheva, C.J. Whitaker, Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, Mach. Learn. 51 (2) (2003) 181–207.

[9] Y. Liu, X. Yao, Ensemble learning via negative correlation, Neural Networks 12 (1999) 1399–1404.

[10] Y. Liu, X. Yao, T. Higuchi, Evolutionary ensembles with negative correlation learning, IEEE Trans. Evol. Comput. 4 (4) (2000) 380–387.

[11] M. Bacauskiene, V. Cibulskis, A. Verikas, Selecting variables for neural network committees, in: J. Wang et al. (Ed.), Lecture Notes in Computer Science, vol. 3971, Springer, Heidelberg, 2006, pp. 837–842.

[12] G. Brown, J. Wyatt, R. Harris, X. Yao, Diversity creation methods: a survey and categorisation, Inf. Fusion 6 (2005) 5–20.

[13] L.I. Kuncheva, R.K. Kountchev, Generating classifier outputs of fixed accuracy and diversity, Pattern Recognition Lett. 23 (5) (2002) 593–600.

[14] G. Forman, Feature selection: we've barely scratched the surface, IEEE Intelligent Syst. 20 (6) (2005) 74–76.

[15] M. Kudo, J. Sklansky, Comparison of algorithms that select features for pattern classifiers, Pattern Recognition 33 (1) (2000) 25–41.

[16] A. Verikas, M. Bacauskiene, Feature selection with neural networks, Pattern Recognition Lett. 23 (11) (2002) 1323–1335.

[17] K.Z. Mao, Orthogonal forward selection and backward elimination algorithms for feature subset selection, IEEE Trans. Syst. Man Cybern. Part B: Cybern. 34 (1) (2004) 629–634.

[18] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (2002) 389–422.

[19] P. Pudil, J. Novovicova, J. Kittler, Floating search methods in feature selection, Pattern Recognition Lett. 15 (1994) 1119–1125.

[20] S. Yu, S.G. Backer, P. Scheunders, Genetic feature selection combined with composite fuzzy nearest neighbor classifiers for hyperspectral satellite imagery, Pattern Recognition Lett. 23 (1–3) (2002) 183–190.

[21] H. Zhang, G. Sun, Feature selection using tabu search method, Pattern Recognition 35 (2002) 701–711.

[22] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1997) 273–324.

[23] A.L. Blum, P. Langley, Selection of relevant features and examples in machine learning, Artif. Intell. 97 (1997) 245–271.

[24] E.K. Tang, P.N. Suganthan, X. Yao, Gene selection algorithms for microarray data based on least squares support vector machine, BMC Bioinformatics 7 (95) (2006).

[25] K.L. Priddy, S.K. Rogers, D.W. Ruck, G.L. Tarr, M. Kabrisky, Bayesian selection of important features for feedforward neural networks, Neurocomputing 5 (1993) 91–103.

[26] J.M. Steppe, K.W. Bauer, Improved feature screening in feedforward neural networks, Neurocomputing 13 (1996) 47–58.

[27] N. Acir, C. Guzelis, Automatic recognition of sleep spindles in EEG via radial basis support vector machine based on a modified feature selection algorithm, Neural Comput. Appl. 14 (2005) 56–65.

[28] T. Evgeniou, M. Pontil, C. Papageorgiou, T. Poggio, Image representations and feature selection for multimedia database search, IEEE Trans. Knowl. Data Eng. 15 (4) (2003) 911–920.

[29] K. Torkkola, E. Tuv, Variable selection using ensemble methods, IEEE Intelligent Syst. 20 (6) (2005) 68–70.

[30] T.K. Ho, The random subspace method for constructing decision forests, IEEE Trans. Pattern Anal. Mach. Intell. 20 (8) (1998) 832–844.

[31] A. Tsymbal, S. Puuronen, D.W. Patterson, Ensemble feature selection with simple Bayesian classification, Inf. Fusion 4 (2003) 87–100.

[32] R. Bryll, R. Gutierrez-Osuna, F. Quek, Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets, Pattern Recognition 36 (6) (2003) 1291–1302.

[33] T. Arodz, D.A. Yuen, A.Z. Dudek, Ensemble of linear models for predicting drug properties, J. Chem. Inf. Modeling 46 (2006) 416–423.

[34] G. Guo, Y. Ma, R. Ward, V. Castranova, X. Shi, Y. Qian, Constructing molecular classifiers for the accurate prognosis of lung adenocarcinoma, Clin. Cancer Res. 12 (11) (2006) 3344–3354.

[35] C. Guerra-Salcedo, D. Whitley, Genetic approach to feature selection for ensemble creation, in: Proceedings of the Genetic and Evolutionary Computation Conference, GECCO-99, Morgan Kaufmann, Los Altos, CA, 1999, pp. 236–243.

[36] D. Opitz, Feature selection for ensembles, in: Proceedings of the 16th National Conference on Artificial Intelligence, Orlando, Florida, 1999, pp. 379–384.

[37] Y. Kim, W.N. Street, F. Menczer, Optimal ensemble construction via meta-evolutionary ensembles, Expert Syst. Appl. 30 (2006) 705–714.

[38] H. Altmcay, Ensembling evidential *k*-nearest neighbor classifiers through multi-modal perturbation, Appl. Soft Comput. 7 (2007) 1072–1083.

[39] K. Tumer, N.C. Oza, Input decimated ensembles, Pattern Anal. Appl. 6 (2003) 65–77.

[40] M.H. Song, C.M. Breneman, J.B. Bi, N. Sukumar, K.P. Bennett, S. Cramer, N. Tugcu, Prediction of protein retention times in anion-exchange chromatography systems using support vector regression, J. Chem. Inf. Comput. Sci. 42 (6) (2002) 1347–1357.

[41] J.L. Rodríguez, L.I. Kuncheva, C.J. Alonso, Rotation forest: a new classifier ensemble method, IEEE Trans. Pattern Anal. Mach. Intell. 28 (10) (2006) 1619–1630.

[42] F. Alkoot, J. Kittler, Feature selection for an ensemble of classifiers, in: Proceedings of the 4th World Multi-conference Systematics, Cybernetics and Informatics, vol. VII, Orlando, Florida, 2000.

[43] S. Gunter, H. Bunke, Feature selection algorithms for the generation of multiple classifier systems and their application to handwritten word recognition, Pattern Recognition Lett. 25 (2004) 1323–1336.

[44] X. Wang, X. Tang, Random sampling for subspace face recognition, Int. J. Comput. Vision 70 (1) (2006) 91–104.

[45] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.

[46] M.A. Hall, G. Holmes, Benchmarking attribute selection techniques for discrete class data mining, IEEE Trans. Knowl. Data Eng. 15 (2003) 1437–1447.

[47] J.R. Quinlan, Induction of decision tree, Mach. Learn. 1 (1986) 81–106.

[48] W. Siedlecki, J. Sklansky, A note on genetic algorithms for large-scale feature selection, Pattern Recognition Lett. 10 (5) (1989) 335–347.

[49] H. Vafaie, K.De Jong, Genetic algorithms as a tool for feature selection in machine learning, in: Proceedings of the 1992 International Conference on Tools with Artificial Intelligence (TAI'92), IEEE Computer Society Press, Arlington, VA, 1992, pp. 200–203.

[50] L.I. Kuncheva, L.C. Jain, Designing classifier fusion systems by genetic algorithms, IEEE Trans. Evol. Comput. 4 (4) (2000) 327–336.

[51] H. Altmcay, Optimal resampling and classifier prototype selection in classifier ensembles using genetic algorithms, Pattern Anal. Appl. 7 (2004) 285–295.

[52] M. Grimaldi, P. Cunningham, A. Kokaram, Discrete wavelet packet transform and ensembles of lazy and eager learners for music genre classification, Multimedia Syst. 11 (5) (2006) 422–437.

[53] X. Li, S. Rao, Y. Wang, B. Gong, Gene mining: a novel and powerful ensemble decision approach to hunting for disease genes using microarray expression profiling, Nucleic Acids Res. 32 (9) (2004) 2685–2694.

[54] S.B. Cho, J. Ryu, Classifying gene expression data of cancer using classifier ensemble with mutually exclusive features, Proc. IEEE 90 (11) (2002) 1744–1753.

[55] L. Nanni, Cluster-based pattern discrimination: a novel technique for feature selection, Pattern Recognition Lett. 27 (2006) 682–687.

[56] A. Tsymbal, P. Cunningham, M. Pechenizkiy, S. Puuronen, Search strategies for ensemble feature selection in medical diagnostics, in: Proceedings of the 16th IEEE Symposium on Computer-Based Medical Systems, IEEE Press, New York, 2003, pp. 124–129.

[57] J.C. Lagarias, J.A. Reeds, M.H. Wright, P.E. Wright, Convergence properties of the Nelder–Mead simplex method in low dimensions, SIAM J. Optim. 9 (1) (1998) 112–147.

[58] J. Shawe-Taylor, N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge University Press, Cambridge, UK, 2004.

[59] K.F. Leung, F.H.F. Leung, H.K. Lam, S.H. Ling, Application of a modified neural fuzzy network and an improved genetic algorithm to speech recognition, Neural Comput. Appl. 16 (4–5) (2007) 419–431.

[60] A. Verikas, A. Gelzinis, M. Bacauskiene, V. Uloza, Integrating global and local analysis of colour, texture and geometrical information for categorizing laryngeal images, Int. J. Pattern Recognition Artif. Intell. 20 (8) (2006) 1187–1205.

[61] A. Verikas, A. Gelzinis, D. Valincius, M. Bacauskiene, V. Uloza, Multiple feature sets based categorization of laryngeal images, Comput. Methods Programs Biomed. 85 (3) (2007) 257–266.

[62] G. Wyszecki, W.S. Stiles, Color Science. Concepts and Methods, Quantitative Data and Formulae, second ed., Wiley, New York, 1982.

[63] A. Gelzinis, A. Verikas, M. Bacauskiene, Increasing the discrimination power of the co-occurrence matrix-based features, Pattern Recognition 40 (9) (2007) 2367–2372.

[64] S.Y. Zhu, K.N. Plataniotis, A.N. Venetsanopoulos, Comprehensive analysis of edge detection in color image processing, Opt. Eng. 38 (4) (1999) 612–625.

About the Author—MARIJA BACAUSKIENE is a senior researcher in the Department of Applied Electronics at Kaunas University of Technology, Lithuania. Her research interests include artificial neural networks, image processing, pattern recognition, and fuzzy logic. She participated in various research projects and published numerous papers in these areas.

About the Author—ANTANAS VERIKAS is currently holding a professor position at both Halmstad University, Sweden and Kaunas University of Technology, Lithuania. His research interests include image processing, pattern recognition, artificial neural networks, fuzzy logic, and visual media technology. He is a member of the International Pattern Recognition Society, European Neural Network Society, International Association of Science and Technology for Development, Swedish Society of Learning Systems, and a member of the IEEE.

About the Author—ADAS GELZINIS received the M.S. degree in Electrical Engineering from Kaunas University of Technology, Lithuania, in 1995. He received the Ph.D. degree in computer science from the same university, in 2000. He is a senior researcher in the Department of Applied Electronics at Kaunas University of Technology. His research interests include artificial neural networks, kernel methods, pattern recognition, signal and image processing, and texture classification.

About the Author—DONATAS VALINCIUS received the M.S. degree in Artificial Neural Networks in 2001 from Kaunas University of Technology, Lithuania. Currently he is a Ph.D. student in the department of Applied Electronics at Kaunas University of Technology. The main fields of his research interests are image analysis, pattern recognition, and learning in artificial neural networks.