FEATURE SELECTION METHODS BASED ON MINIMIZATION OF CPL CRITERION FUNCTIONS

Tomasz Łukaszuk

Faculty of Computer Science, Bialystok University of Technology, Białystok, Poland

Abstract: The feature selection is a method of data analysis commonly used as a preliminary step in the techniques of classification and pattern recognition. It is particularly important in situations when data are represented in high-dimensional feature space. Examples of these are collections of bioinformatics data, particularly data obtained from DNA microarrays. The paper presented two methods of feature selection based on minimizing the CPL criterion function: basic SEKWEM/GENET method, in which the selection of features is done in conjunction with the construction of a linear classifier separating objects from different decision classes, and the RLS method extending the primary method by linear separability relaxation stage in order to obtain a subset of features with better generalization ability. The results of the SEKWEM/GENET and RLS methods were confronted with the results obtained from other common feature selection methods in application to the state of the art microarray data sets.

Keywords: feature selection, CPL criterion function, SEKWEM/GENET algorithm, RLS method

1. Introduction

Nowadays, a lot of companies, administrative and scientific institutions has, and still collects data on various aspects of their bussinesses. Based on the collected data it is possible to carry out the necessary studies and obtain useful information and new knowledge. But often it happens that in the stage of data collection, test objects or phenomena are recorded with as large as possible number of parameters. Also, some types of data, research facilities, by their nature are described in a very large number of attributes. Examples of such data are digitized text and bioinformatics data.

The feature selection is a technique commonly used in data mining. Its aim is the selection from all available features the subset of features relevant to the considered

Zeszyty Naukowe Politechniki Białostockiej. Informatyka, vol. 8, pp. 31-43, 2011.

problem [10]. Best subset should contain the minimum number of features which most affect the quality of the model relating to this problem.

Feature selection is also known as task that consists in removing irrelevant and redundant features from the initial data (features) set [14]. Irrelevant and redundant features means features with no or minimal effect on later decisions.

There are two ways of selecting features set. One consists in making a ranking of features according to some criterion and selecting certain number of the best features. The other is to select a minimum subset of features without learning performance deterioration [14]. In the second way the quality of the whole subset is evaluated.

Important aspects connected with feature selection are models and search strategies. Typical models are filter, wrapper, and embedded. Filter methods use some own internal properties of the data to select features. Examples of the properties are feature dependence, entropy of distances between data points, redundancy. In the wrapper methods the feature selection is connected with the other data analysis technique, such as classification, clustering algorithm, regression. The accompanying technique helps with evaluation of the quality of selected features set. An embedded model of feature selection integrates the selection in model building. An example of such method is the decision tree induction algorithm. At each node a feature has to be selected. Basic search strategies applied in feature selection are forward, backward, floating, branch-and-bound and randomized strategies [14]. Besides there are a lot of modifications and improvements of them.

This paper is engaged in the feature selection by minimization of a special convex and piece-wise linear (CPL) criterion function. The minimization process allows to calculate the parameters of hyperplane separating the learning sets and to find the best set of features ensured the linear separability of them at once.

The remainder of the paper is structured as follows: Section 2 provides a brief description of exploratory analysis techniques based on minimization of CPL criterion function, Sections 3 and 4 contain a more detailed introduction to developed by author feature selection methods SEKWEM/GENET and RLS. Section 5 presents the course and results of experiments involving the comparison of the SEKWEM/GENET and RLS methods with other feature selection methods. Finally, the work is summarized in Section 6.

2. The exploratory analysis techniques based on minimization of CPL criterion function

Let us consider that the test objects O_j $(j = 1,...,m)$ are represented by the feature vectors $\mathbf{x}_j[n] = [x_{j1},...,x_{jn}]^T$ of the same dimensionality n or by points in the

n-dimensional feature space $F[n]$. Feature (attribute) x_i describes a specific numerical value of the i-th parameter, or the result of a specific i-th measurement made on each object O_j. Features can take discrete ($x_i \in \{0,1,...,p\}$) or continuous ($x_i \in \mathbf{R}^1$) values.

Let us take into consideration two disjointed sets C^+ and C^- composed of m feature vectors $\mathbf{x}_j$:

$$C^+ \cap C^- = \emptyset \,. \tag{1}$$

For example vectors from the first set represent patients suffered from certain disease and vectors from the second one represent patients without the disease. The *positive set* C^+ contains m^+ vectors $\mathbf{x}_j$ and the *negative set* C^- contains m^- vectors ($m = m^+ + m^-$).

We are considering the separation of the sets C^+ and C^- by the hyperplane $H(\mathbf{w},\theta)$ in the feature space $F[n]$.

$$H(\mathbf{w},\theta) = \{\mathbf{x} : \langle \mathbf{w},\mathbf{x} \rangle = \theta\} \tag{2}$$

where $\mathbf{w} = [w_1,...,w_n]^T \in \mathbf{R}^n$ is the weight vector, $\theta \in \mathbf{R}^1$ is the threshold, and $\langle \mathbf{w},\mathbf{x} \rangle$ is the inner product.

One way of finding the hyperplane $H(\mathbf{w},\theta)$ (2) is to minimize a properly defined criterion function $\Phi_\lambda(\mathbf{w},\theta)$ [3].

$$\Phi_\lambda(\mathbf{w},\theta) = \sum_{\mathbf{x}_j \in C^+} \alpha_j \varphi_j^+(\mathbf{w},\theta) + \sum_{\mathbf{x}_j \in C^-} \alpha_j \varphi_j^-(\mathbf{w},\theta) + \lambda \sum_{i \in I} \gamma_i \phi_i(\mathbf{w},\theta) \tag{3}$$

where $\alpha_j \geq 0$, $\lambda \geq 0$, $\gamma_i > 0$, $I = \{1,...,n\}$.
The nonnegative parameters α_j determine relative importance (*price*) of particular feature vectors $\mathbf{x}_j$. The parameters γ_i. represent the *costs* of particular features x_i.

The function $\Phi_\lambda(\mathbf{w},\theta)$ is the sum of the penalty functions $\varphi_j^+(\mathbf{w},\theta)$ or $\varphi_j^-(\mathbf{w},\theta)$ and $\phi_i(\mathbf{w},\theta)$. The functions $\varphi_j^+(\mathbf{w},\theta)$ are defined on the feature vectors $\mathbf{x}_j$ from the set C^+. Similarly $\varphi_j^-(\mathbf{w},\theta)$ are based on the elements $\mathbf{x}_j$ of the set C^-.

$$(\forall \mathbf{x}_j \in C^+) \quad \varphi_j^+(\mathbf{w},\theta) = \begin{cases} 1+\theta-\langle \mathbf{w},\mathbf{x}_j \rangle & if \quad \langle \mathbf{w},\mathbf{x}_j \rangle < 1+\theta \\ 0 & if \quad \langle \mathbf{w},\mathbf{x}_j \rangle \geq 1+\theta \end{cases} \tag{4}$$

and

$$(\forall \mathbf{x}_j \in C^-) \quad \varphi_j^-(\mathbf{w},\theta) = \begin{cases} 1+\theta+\langle \mathbf{w},\mathbf{x}_j \rangle & if \quad \langle \mathbf{w},\mathbf{x}_j \rangle > -1+\theta \\ 0 & if \quad \langle \mathbf{w},\mathbf{x}_j \rangle \leq -1+\theta \end{cases} \tag{5}$$

The penalty functions $\phi_i(\mathbf{w},\theta)$ are related to particular features x_i.

$$\phi_i(\mathbf{w},\theta) = \begin{cases} |w_i| & if \quad 1 \leq i \leq n \\ |\theta| & if \quad i = n+1 \end{cases} \tag{6}$$

The criterion function $\Phi_\lambda(\mathbf{w},\theta)$ (3) is the convex and piecewise linear (*CPL*) function as the sum of the *CPL* penalty functions $\varphi_j^+(\mathbf{w},\theta)$ (4), $\varphi_j^-(\mathbf{w},\theta)$ (5) and $\phi_i(\mathbf{w},\theta)$ (6). The basis exchange algorithm allows to find the minimum efficiently, even in the case of large multidimensional data sets C^+ and C^- [2].

$$\Phi_\lambda^* = \Phi_\lambda(\mathbf{w}^*,\theta^*) = min\,\Phi_\lambda(\mathbf{w},\theta) \geq 0 \tag{7}$$

The parameters $\mathbf{w}^*$ and θ^* define the hyperplane $H(\mathbf{w}^*,\theta^*)$ (2), which in an optimal way in terms of linear separability criterion measured by the value of function $\Phi_\lambda(\mathbf{w},\theta)$ (3) separates the data sets C^+ and C^-.

3. SEKWEM/GENET feature selection method

SEKWEM/GENET is the basic algorithm of feature selection based on the minimization of CPL criterion function $\Phi_\lambda(\mathbf{w},\theta)$ (3). Feature selection is done together with the search for the optimal hyperplane $H(\mathbf{w}^*,\theta^*)$ (2) separating the data sets C^+ and C^-. The resulting vector of parameters $\mathbf{w}^*$ may contain a number of factors w_i equal to or close to zero. This condition occurs especially in the case of multidimensional, so called, "long" data. Features x_i corresponding to the coefficients w_i equal or close to zero are rejected, while the features x_i corresponding to other coefficients form a set of selected features.

In order to simplify further considerations let us assume the following augmented form of the feature vectors $\mathbf{y}_j$ and the vector of parameters $\mathbf{v}$:

$$\mathbf{y}_j = [\mathbf{x}_j^T, 1]^T \tag{8}$$

$$\mathbf{v} = [\mathbf{w}^T, -\theta]^T \tag{9}$$

The equation of the separating hyperplane $H(\mathbf{w},\theta)$ (2) will take the form:

$$H(\mathbf{v}) = \{\mathbf{y} : \langle \mathbf{v},\mathbf{y}\rangle = 0\} \tag{10}$$

Determination of parameter $\mathbf{v}^*$ of the optimal hyperplane $H(\mathbf{v}^*)$ (10) is based on the basis exchange algorithm [2]. The algorithm searches in an oriented way vertices $\mathbf{v^k}$ resulting from intersections of hyperplanes h_j^+, h_j^- and h_i (12), respectively associated with the features vectors $\mathbf{y}_j$ belonging to the sets C^+, C^- and the unit vectors $\mathbf{e}_i$ (11).

$$\begin{aligned} &\mathbf{e}_i = [e_{i1}, e_{i2}, ..., e_{ik}, ..., e_{in+1}]^T \\ &(\forall i \in \{1,...,n+1\})(\forall k \in \{1,...,n+1\}) \begin{cases} e_{ik} = 1 \; when \; i = k \\ e_{ik} = 0 \; when \; i \neq k \end{cases} \end{aligned} \tag{11}$$

$$\begin{array}{l}(\forall \mathbf{y}_j \in C^+)\ h_j^+ = \{\mathbf{v} : \langle \mathbf{y}_j, \mathbf{v}\rangle = 1\} \\ (\forall \mathbf{y}_j \in C^-)\ h_j^- = \{\mathbf{v} : \langle \mathbf{y}_j, \mathbf{v}\rangle = -1\} \\ (\forall i\ \in (1,...,n+1))\ h_i = \{\mathbf{v} : \langle \mathbf{e}_i, \mathbf{v}\rangle = 0\}\end{array} \tag{12}$$

Each vertex $\mathbf{v}^k$ is the intersection of at least $(n+1)$ hyperplanes h_j^+, h_j^-, h_i [4]. In the vertex $\mathbf{v}^k$ the following equations are fulfilled:

$$\begin{array}{l}(\forall j \in J_k^+)\ (\mathbf{y}_j)^T \mathbf{v}^k = 1 \\ (\forall j \in J_k^-)\ (\mathbf{y}_j)^T \mathbf{v}^k = -1 \\ (\forall i\ \in I_k)\ (\mathbf{e}_i)^T \mathbf{v}^k = 0\end{array} \tag{13}$$

where J_k^+, J_k^- are the sets of indexes of vectors $\mathbf{y}_j$ belonging respectively to the sets C^+ and C^-, which is compliance with the equation (13), and I_k is the set of indexes of unit vectors $\mathbf{e}_i$ satisfying the last of the equations (13).

Equations (13) can be written in the form of a matrix [4]:

$$\mathbf{B}^k \mathbf{v}^k = \delta \tag{14}$$

$\mathbf{B}^k$ is called the base. The rows of $\mathbf{B}^k$ are formed by features vectors $\mathbf{y}_j$ $(j \in J_k^+ \cup J_k^-)$ or unit vectors $\mathbf{e}_i$ $(i \in I_k)$. δ margins is a vector with components equal to 1, -1 or 0 according to (13).

The coefficients v_i^k of the vector (vertex) $\mathbf{v}^k$ associated with the unit vectors $\mathbf{e}_i\,(i \in I_k)$ in base $\mathbf{B}^k$ are equal to zero ($v_i^k = 0$). This follows from the equality (13). Features x_i corresponding to coefficients v_i^k equal to zero may not be taken into account in the vertex $\mathbf{v}^k$. They do not affect the form of separating hyperplane $H(\mathbf{v}^k)$ (10). Vertex $\mathbf{v}^k$ is completely characterized by the subset of the features F^k:

$$F^k = \{x_i : i\ \in I_k'\} \tag{15}$$

where $I_k' = \{1,...,n\} \setminus I_k$.

Minimizing of the criterion function $\Phi_\lambda(\mathbf{v})$ (3) according to the basis exchange algorithm comes down to appropriate movement between the vertices $\mathbf{v}^k$ until the optimal vertex $\mathbf{v}^*$ is found. Each transition from vertex $\mathbf{v}^k$ to the vertex $\mathbf{v}^{k+1}$ is associated with the replacement of one vector in the base $\mathbf{B}^k$. If a unit vector $\mathbf{e}_l$ exits from the base $\mathbf{B}^k$, it means changing the consideration from the subset of features F^k to the extended subset of features $F^{k+1} = F^k \cup \{x_l\}$. If a unit vector $\mathbf{e}_r$ enters to the base $\mathbf{B}^k$, it means changing the consideration from the subset of features F^k to the reduced subset of features $F^{k+1} = F^k \setminus \{x_r\}$.

Considering the above facts, the process of minimizing the criterion function $\Phi_\lambda(\mathbf{v})$ (3) according to the basis exchange algorithm is connected with a browsing

of subsets of features F^k (15) characterized by the vertices $\mathbf{v}^k$. The optimal vector $\mathbf{v}^*$ corresponds to the optimal (in the sense of linear separability criterion of the sets C^+ and C^- (1)) subset of features F^*.

$$F^1 \rightarrow F^2 \rightarrow ... \rightarrow F^k \rightarrow F^{k+1} \rightarrow ... \rightarrow F^* \tag{16}$$

4. RLS feature selection method

A fact that a model behaves very well in relation to objects from training set does not guarantee that equally well handle with objects inactive in the learning process. It is due to the danger of overfitting. It is worth, by reduction the model quality in relation to training data, to obtain better performance in conjunction with test data [17]. This idea underlies the extended feature selection methods, the relaxed linear separability (RLS) method [4].

The RLS method consists of two calculations stages. In the first stage, in accordance with the previously described basic feature selection scheme, there are determined the optimal subset of features F^* (16) and the optimal parameter vector $\mathbf{v}^*$ (7). In the second stage a linear separability relaxation is performed. The linear separability relaxation consists in the controlled removal from the subset F^* (16) the consecutive least significant features and evaluation so obtained subsets of features [4].

Selecting a feature to remove from the subset F^* (16) (and the next resulting subsets of features) is done through the appropriate increasing the value of cost parameter λ occurring in the expression of the criterion function $\Phi_\lambda(\mathbf{v})$ (3). After increasing the value of parameter λ, optimization of the criterion function $\Phi_\lambda(\mathbf{v})$ is performed. The optimization starts from the previously specified vertex $\mathbf{v}^*$ (7). If the value of λ was increased enough, it will lead to an increase the number of unit vectors $\mathbf{e}_i$ in base $\mathbf{B}^{*1}$ associated with the new optimal vertex $\mathbf{v}^{*1}$, and thus to reduce the number of features of optimal subset of features F^{*1} [4]. Further enhancing the value of parameter λ allows to obtain the next subsets of features F^{*k} with a reduced number of features. It is possible to control the value of λ to obtain a sequence of subsets of features $F^{*1}, F^{*2}, ..., F^{*p}$, where each subset $F^{*(k+1)}$ is equal to the subset F^{*k} minus one least significant feature. The last subset F^{*p} has only one feature.

The measure of quality of feature subsets F^{*k} is the classifier error $e_{LOOCV}(F^{*k})$ estimated by leave-one-out cross validation [7] on the set consisting of all objects from subsets C^+ and C^- (1) with reduced features not belonging to subset of features of F^{*k}. The error $e_{LOOCV}(F^{*k})$ is equal to the fraction of incorrectly classified objects

with one-element test sets created in the validation process.

$$e_{LOOCV}(F^{*k}) = m_{LOOCV}(F^{*k})/m \tag{17}$$

where $m_{LOOCV}(F^{*k})$ is the number of misclassified objects, and m is the total number of objects in sets C^+ and C^- (1).

The RLS method as the best resulting subset of features considers the subset with the smallest error $e_{LOOCV}(F^{*k})$. If there is more than one subset of features with the smallest error $e_{LOOCV}(F^{*k})$, RLS selects the least numerous subset.

5. Empirical studies

5.1 Experimental setup

Three benchmarking feature selection algorithms were selected for an experimental comparison with the SEKWEM/GENET and RLS methods. One of the selected algorithms, ReliefF, is based on feature ranking procedure proposed by Kononenko [13] as an extension of the Relief algorithm [12]. The ReliefF searches for the nearest objects from different classes and weights features according to how well they differentiate these objects. The second one is a subset search algorithm denoted as CFS-SF (Correlation-based Feature Subset Selection - Sequential Forward) [11]. The CFS-SF algorithm is based on a correlation measure which evaluates the goodness of a given feature subset by assessing the predictive ability of each feature in the subset and a low degree of correlation between features in the subset. The third method, Consistency Subset Evaluation - Selection Forward (CSE-SF) also belongs to the subset search selection methods. It searches the space of solutions using the forward selection procedure, and evaluates the found subsets of features using inconsistency measure proposed by Liu and Setiono at work [15] and then developed in the work [5].

The applied algorithms require the determination of certain parameters controlling their work and having an impact on the results returned. The author used, in most cases, the standard parameters recommended by the creators of algorithms.

Studied feature selection methods were compared on the basis of the returned feature space quality. The quality of the feature space was evaluated based on its discriminative power. Four frequently used classification methods and the CPL method were applied to assess the discriminative power of selected feature spaces:

- k Nearest Neighbours (kNN) [6] with $k = 5$ (arbitrary choice)
- Support Vector Machines (SVM) [18] with linear kernel function

– Naive Bayes Classifier (NBC) [6]
– C4.5 Decision Tree Algorithm (C4.5) [16]
– Convex and Piecewise-Linear criterion functions (CPL) [3] with linear relaxation [4]

These five classifiers were designed in the full feature spaces and in the reduced feature subspaces obtained by the feature selection methods. The result characterizing the effectiveness of a classifier for a given data set is the fraction of misclassified objects from the testing set in the process of leave-one-out cross-validation [7]. The effectiveness of a classifier is an assessment of the quality of the feature space and, consequently, a part of assessment of the feature selection method.

The four first classifiers were designed by using Weka's implementation [20]. The Weka's implementation of ReliefF, CSE-SF and CFS-SF was used also for the feature selection and cross validation evaluation of designed classifiers. The CPL classifiers (the fifth type) based on the search for optimal separating hyperplane through minimization of the CPL criterion functions was applied using author's own implementation. Autor's implementation was also used for the SEKWEM/GENET and RLS methods of feature selection. Calculations were performed on a computer with Intel Core2 T5500 processor and 1GB of RAM.

5.2 Data sets

The experiments were carried out on publicly available data sets concerning classification problems related to four different diseases: colon cancer, leukemia, lung cancer and breast cancer.

The Colon cancer [1] contains expression levels of 2000 genes taken in 62 different samples. For each sample it is indicated whether it came from a tumor biopsy or not.

The Leukemia [8] data set contains expression levels of 7129 genes taken over 72 samples. Labels of objects indicate which of two variants of leukemia is present in the sample: acute myeloid (AML, 25 samples), or acute lymphoblastic leukemias (ALL, 47 samples).

The Lung cancer [9] is made of 181 patients with 12533 markers. The samples belong to two lung cancer classes, malignant pleural mesothelioma (MPM, 31 samples) and adenocarcinoma (ADCA, 150 samples).

The Breast cancer [19] data set describes the patients tested for the presence of breast cancer. The data contains 97 patient samples, 46 of which are from patients who had developed distance metastases within 5 years (labelled as "relapse"), the rest 51 samples are from patients who remained healthy from the disease after their initial

diagnosis for interval of at least 5 years (labelled as "non-relapse"). The number of genes is 24481.

Original data sets come with training and test samples that were drawn from different conditions. Here we combine them together for the purpose of cross validation. Data have also been standardized before experiment.

Table 1. The data sets used in testing the feature selection methods

Name	#objects	#features	class sizes	
Colon cancer	62	2000	tumor 40	normal 22
Leukemia	72	7129	ALL 47	AML 25
Lung cancer	181	12533	MPM 31	ADCA 150
Breast cancer	97	24481	relapse 46	non-relapse 51

5.3 Results

Table 2 summarizes the results of examined feature selection methods obtained in the previously described experiment.

System resources are unfortunately insufficient to apply the CFS-SF method to the lung cancer and breast cancer data sets. Available RAM is insufficient compared to the memory complexity of the algorithm.

Comparing the classification errors obtained on the full data sets ("No selection" group of rows), and classification errors on the data sets composed of features selected by each method (next groups of rows) it should be noted that each of the feature selection methods returns a subset of features improving the properties of classifier. It is in line with expectations and the idea of feature selection. However, the improvement of the quality of classifier is different in relation to particular methods. The methods developed by the author (SEKWEM/GENET and RLS) proved to be significantly better compared to other studied methods. This is clearly evident when compared the values of mean errors obtained for all used classification algorithms, listed in Table 2 in the rows "average".

The second criterion of evaluation the feature selection methods is the number of features returned by the procedure. In this aspect, by far the best method is

Table 2. Comparison of feature selection algorithms in terms of number of selected features and classification errors estimated by leave-one-out cross-validation method

		Colon cancer	Leukemia	Lung cancer	Breast cancer
No selection	#features	2000	7129	12533	24481
	kNN	20,97%	15,28%	6,08%	39,18%
	SVM	16,13%	1,39%	1,11%	31,96%
	NBC	16,13%	0,00%	2,21%	47,42%
	C4.5	20,97%	26,39%	3,87%	42,27%
	CPL	9,68%	2,78%	1,11%	25,77%
	average	16,78%	9,17%	2,88%	37,32%
SEKWEM/GENET	#features	39	43	64	78
	kNN	8,06%	0,00%	0,55%	1,03%
	SVM	0,00%	0,00%	0,00%	0,00%
	NBC	6,45%	0,00%	0,55%	8,23%
	C4.5	33,97%	16,67%	2,76%	29,90%
	CPL	0,00%	0,00%	0,00%	0,00%
	average	9,68%	3,33%	0,77%	7,83%
RLS	#features	14	7	3	19
	kNN	8,06%	0,00%	0,00%	5,15%
	SVM	0,00%	0,00%	0,00%	0,00%
	NBC	4,84%	0,00%	2,76%	9,28%
	C4.5	19,35%	12,50%	3,87%	27,84%
	CPL	12,50%	6,67%	3,72%	21,21%
	average	8,95%	3,83%	2,07%	12,70%
ReliefF	#features	15	32	393	43
	kNN	12,90%	4,17%	3,31%	15,46%
	SVM	11,29%	2,78%	0,55%	23,71%
	NBC	9,68%	4,17%	0,55%	20,62%
	C4.5	20,97%	16,68%	3,31%	35,05%
	CPL	12,90%	5,56%	1,11%	19,59%
	average	13,55%	6,67%	1,77%	22,89%
CSE-SF	#features	4	3	2	6
	kNN	17,74%	4,17%	2,21%	27,83%
	SVM	12,90%	5,56%	2,21%	29,90%
	NBC	11,29%	5,56%	2,76%	46,39%
	C4.5	8,06%	5,56%	1,11%	29,90%
	CPL	12,90%	8,33%	2,21%	31,96%
	average	12,58%	5,84%	2,10%	33,20%
CFS-SF	#features	58	81	n/a	n/a
	kNN	8,06%	1,39%	n/a	n/a
	SVM	12,90%	1,39%	n/a	n/a
	NBC	8,06%	0,00%	n/a	n/a
	C4.5	12,90%	20,83%	n/a	n/a
	CPL	12,90%	4,17%	n/a	n/a
	average	10,96%	5,56%	n/a	n/a

the CSE-SF. Among the author's methods much better on this criterion falls RLS algorithm. The SEKWEM/GENET method is characterized by selecting relatively numerous subsets of features.

On the basis of the results it can be quite definitely say that SEKWEM/GENET and RLS methods very well fit for the purpose of feature selection in relation to high dimensional data sets. They choose subsets of features with high quality, using which makes it possible to build much better classification and decision-making rules than on the basis of the starting sets of features.

6. Concluding remarks

The paper presents basic assumptions of the SEKWEM/GENET and RLS methods of feature selection. The basis of both methods is the minimization of the special CPL criterion function. The work also contains the results obtained from applying of described methods with the state of the art high dimensional microarray data. In comparison with other commonly used feature selection methods the author's methods proved to be better considering the quality of the returned sets of features. The measure of the quality of a subset of features is the classification error obtained in the process of leave-one-out cross-validation.

The experiment described in the article is a repetition of the experiment made by the author for his doctoral dissertation. Some of the results obtained this time is a bit different than the results shown in the dissertation. The reason for this is ongoing work on the development and improvement of the SEKWEM/GENET and RLS methods. The results presented in this article have been obtained on the following slightly modified versions of the implementations of the SEKWEM/GENET and RLS algorithms. Recently completed and potential future development activities are aimed at improving the quality of the results, but also take into account emergency situations that occur sometimes after applying the algorithm to a new custom data set.

References

[1] U. Alon, et al., Broad patterns of gene expressions revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, PNAS, 96:6745–6750, 1999.

[2] L. Bobrowski, Design of Piecewise Linear Classifiers from Formal Neurons by Some Basis Exchange Technique, Pattern Recognition, 24(9):863–870, 1991.

[3] L. Bobrowski, Data mining based on convex and piecewise linear (CPL) criterion functions (in Polish), Wyd. Politechniki Białostockiej, Białystok, 2005.

[4] L. Bobrowski, T. Łukaszuk, Feature Selection Based on Relaxed Linear Separability, Biocybernetics and Biomedical Engineering, 29(2):43–59, 2009.

[5] M. Dash, H. Liu, Consistency-based search in feature selection, Artificial Intelligence, 151:155–176, 2003.

[6] O.R. Duda, P.E. Heart, D.G. Stork, Pattern Classification, Second edition, John Wiley & Sons, 2001.

[7] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, San Diego, 1990.

[8] T.R. Golub, et al., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Sciences, 286:531-537, 1999.

[9] J.G. Gordon, R.V. Jensen, L. Hsiao, S.R. Gullans, Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesotheliomar. Cancer Res, 62:4963–4967, 2002.

[10] I. Guyon I., A. Elisseeff, An introduction to variable and feature selection, Journal of Machine Learning Research, 3:1157–1182, 2003.

[11] M.A. Hall, Correlation-based Feature Selection for Machine Learning, PhD thesis, University of Waikato, Dept. of Computer Science, 1998.

[12] K. Kira K., L.A. Rendell, A Practical Approach to Feature Selection, Ninth International Workshop on Machine Learning, 249-256, 1992.

[13] I. Kononenko, Estimating Attributes: Analysis and Extensions of RELIEF, European Conference on Machine Learning, 171-182, 1994.

[14] H. Liu, H. Motoda, Computational methods of feature selection, Chapman & Hall/CRC data mining and knowledge discovery series, Chapman & Hall/CRC, 2008.

[15] H. Liu, R. Setiono, A Probabilistic Approach to Feature Selection - A Filter Solution, 13th International Conference on Machine Learning, Morgan Kaufmann, 319–327, 1996.

[16] J.R. Quinlan, C4.5 - Programs for Machine Learning, Morgan Kaufmann, 1993.

[17] I.V. Tetko, D.J. Livingstone, A.I. Luik, Neural network studies, 1. Comparison of overfitting and overtraining, Journal of Chemical Information and Computer Sciences, 35(5):826–833, 1995.

[18] V.N. Vapnik, Statistical Learning Theory, J. Wiley, New York, 1998.

[19] L.J. van't Veer, et al., Gene expression profiling predicts clinical outcome of breast cancer, Nature, 415(6871):530-536, 2002.

[20] I.H. Witten, E. Frank, Data Mining - Pracitcal Machine Learning Tools and Techniques with JAVA Implementations, Morgan Kaufmann Publishers, 2000.

METODY SELEKCJI CECH BAZUJĄCE NA MINIMALIZACJI FUNKCJI KRYTERIALNYCH TYPU CPL

Streszczenie Selekcja cech jest metodą analizy danych powszechnie stosowaną jako wstępny krok w technikach klasyfikacji czy rozpoznawania wzorców. Ma ona szczególne znaczenie w sytuacji gdy dane reprezentowane są w wysoko wymiarowej przestrzeni cech. Przykładem takich danych są zbiory bioinformatyczne, a w szczególności dane uzyskane na podstawie mikromacierzy DNA. W pracy przedstawione zostały dwie metody selekcji cech bazujące na minimalizacji funkcji kryterialnych typu CPL: podstawowa metoda SEKWEM/GENET, w której selekcja cech dokonywana jest w połączeniu z budową liniowego klasyfikatora separującego obiekty z różnych klas decyzyjnych, oraz metoda RLS rozszerzająca podstawową metodę o etap relaksacji liniowej separowalności w celu uzyskania podzbioru cech o lepszych zdolnościach generalizacji. Wyniki metod SEKWEM/GENET i RLS zostały także skonfrontowane z wynikami uzyskanymi z innych popularnych metod selekcji cech w zastosowaniu do „benchmarkowych" zbiorów danych mikromacierzowych.

Słowa kluczowe: selekcja cech, funkcja kryterialna typu CPL, algorytm SEKWEM/GENET, metoda RLS

Artykuł zrealizowano w ramach pracy badawczej S/WI/2/2008.