Available at www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

Pattern Recognition Letters 24 (2003) 833–849

Pattern Recognition Letters

www.elsevier.com/locate/patrec

Rough set methods in feature selection and recognition

Roman W. Swiniarski [a,*], Andrzej Skowron [b]

[a] Department of Mathematical and Computer Sciences, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182, USA
[b] Institute of Mathematics, Warsaw University, Banacha 2, 02-097 Warsaw, Poland

Abstract

We present applications of rough set methods for feature selection in pattern recognition. We emphasize the role of the basic constructs of rough set approach in feature selection, namely reducts and their approximations, including dynamic reducts. In the overview of methods for feature selection we discuss feature selection criteria, including the rough set based methods. Our algorithm for feature selection is based on an application of a rough set method to the result of principal components analysis (PCA) used for feature projection and reduction. Finally, the paper presents numerical results of face and mammogram recognition experiments using neural network, with feature selection based on proposed PCA and rough set methods.

Keywords: Pattern recognition; Rough sets; Feature selection

1. Introduction

Reduction of pattern dimensionality via feature extraction and feature selection (see, e.g., Kittler, 1986; Liu and Motoda, 1998a,b) belongs to the most fundamental steps in data preprocessing. Feature selection is often isolated as a separate step in processing sequence. Features constituting the object's pattern may be irrelevant (having no effect on processing performance) or relevant (having an impact on processing performance). Features can be redundant (dependent), and may have a different discriminatory or predictive power. We present rough set methods and principal components analysis (PCA) in context of feature selection in pattern recognition.

The paper begins with a some preliminaries of rough set approach (Pawlak, 1991). We emphasize the special role of reducts in feature selection, including dynamic reducts (Bazan et al., 1994, 1998). Then, we present a short overview of feature selection problem including open-loop and closed-loop feature selection methods. This section focuses on the discussion on feature selection criteria including rough set based methods. The next section presents a short description of the PCA as a method of feature projection and reduction. It also contains description of rough set based methods, proposed jointly with PCA, for feature projection and reduction. The following section describes results of numerical experiments on face and

* Corresponding author.
E-mail addresses: rswiniar@sciences.sdsu.edu (R.W. Swiniarski), skowron@mimuw.edu.pl (A. Skowron).

0167-8655/03/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0167-8655(02)00196-4

mammogram recognition using the proposed rough set based method for feature selection and neural networks. This section also contains short description of feature extraction from facial images using singular value decomposition (SVD) and feature extraction from mammograms which is based on histograms.

2. Rough set preliminaries

Rough set theory has been introduced by Zdzisław Pawlak (Pawlak, 1991) to deal with imprecise or vague concepts. In recent years we witnessed a rapid growth of interest in rough set theory and its applications, worldwide (see, e.g., Skowron, 2000). Here, we introduce only the basic notation from rough set approach used in the paper.

Suppose we are given two finite, non-empty sets U and A, where U is the *universe* of *objects*, *cases*, and A—a set of *attributes*, *features*. The pair $\mathrm{IS} = (U, A)$ is called an *information table*. With every attribute $a \in A$ we associate a set V_a, of its *values*, called the *domain* of a. By $\boldsymbol{a}(x)$ we denote a data pattern $(a_1(x), \ldots, a_n(x))$ defined by the object x and attributes from $A = \{a_1, \ldots, a_n\}$. A data pattern of IS is any feature value vector $\boldsymbol{v} = (v_1, \ldots, v_n)$ where $v_i \in V_{a_i}$ for $i = 1, \ldots, n$ such that $\boldsymbol{v} = \boldsymbol{a}(x)$ for some $x \in U$.

Any subset B of A determines a binary relation $I(B)$ on U, called an *indiscernibility relation*, defined as follows:

$$xI(B)y \quad \text{iff } a(x) = a(y) \text{ for every } a \in B$$

where $a(x)$ denotes the value of attribute a for object x.

The family of all equivalence classes of $I(B)$, i.e., the partition determined by B, will be denoted by $U/I(B)$, or simply U/B; an equivalence class of $I(B)$, i.e., the block of the partition U/B, containing x will be denoted by $B(x)$.

If $(x, y) \in I(B)$ we will say that x and y are B-indiscernible. Equivalence classes of the relation $I(B)$ (or blocks of the partition U/B) are referred to as B-elementary sets. In the rough set approach the elementary sets are the basic building blocks (concepts) of our knowledge about reality. The unions of B-elementary sets are called B-definable sets.

The indiscernibility relation will be further used to define basic concepts of rough set theory. Let us define now the following two operations on sets

$$B_*(X) = \{x \in U : B(x) \subseteq X\},$$
$$B^*(X) = \{x \in U : B(x) \cap X \neq \emptyset\}$$

assigning to every subset X of the universe U two sets $B_*(X)$ and $B^*(X)$ called the B-lower and the B-upper approximation of X, respectively. The set

$$BN_B(X) = B^*(X) - B_*(X)$$

will be referred to as the B-boundary region of X.

If the boundary region of X is the empty set, i.e., $BN_B(X) = \emptyset$, then the set X is *crisp* (*exact*) with respect to B; in the opposite case, i.e., if $BN_B(X) \neq \emptyset$, the set X is referred to as *rough* (*inexact*) with respect to B.

Several generalizations of the classical rough set approach based on approximation spaces defined by (U, R), where R is an equivalence relation (called indiscernibility relation) in U, have been reported in the literature (for references see the papers and bibliography in (e.g., Polkowski and Skowron, 1998; Skowron, 2000).

Sometimes we distinguish in an information table (U, A) a partition of A into two classes C, $D \subseteq A$ of attributes, called *condition* and *decision* (*action*) attributes, respectively. The tuple $\mathrm{DT} = (U, C, D)$ is called a *decision table*. Any such decision table where $U = \{x_1, \ldots, x_N\}$, $C = \{a_1, \ldots, a_n\}$ and $D = \{d_1, \ldots, d_k\}$ can be represented by means of a data sequence (called also data set) of data patterns $((\boldsymbol{v}_1, \mathrm{target}_1), \ldots, (\boldsymbol{v}_N, \mathrm{target}_N))$ where $\boldsymbol{v}_i = \boldsymbol{C}(x_i)$, $\mathrm{target}_i = \boldsymbol{D}(x_i)$, and $\boldsymbol{C}_i = (a_1(x_i), \ldots, a_n(x_i))$, $\boldsymbol{D}_i = (d_1(x_i), \ldots, d_k(x_i))$, for $i = 1, \ldots, N$. It is obvious that also any data sequence defines a decision table. The equivalence classes of $I(D)$ are called decision classes.

Let $V = \bigcup\{V_a | a \in C\} \cup V_d$. Atomic formulae over $B \subseteq C \cup D$ and V are expressions $a = v$ called *descriptors* (*selectors*) over B and V, where $a \in B$ and $v \in V_a$. The set $\mathscr{F}(B, V)$ of formulae over B and V is the least set containing all atomic formulae over B and V and closed with respect to the

propositional connectives $\wedge$ (conjunction), $\vee$ (disjunction) and $\neg$ (negation).

By $\|\varphi\|_{\mathrm{DT}}$ we denote the meaning of $\varphi \in \mathscr{F}(B,V)$ in the decision table DT which is the set of all objects in U with the property φ. These sets are defined by $\|a=v\|_{\mathrm{DT}} = \{x \in U | a(x) = v\}$, $\|\varphi \wedge \varphi'\|_{\mathrm{DT}} = \|\varphi\|_{\mathrm{DT}} \cap \|\varphi'\|_{\mathrm{DT}}$; $\|\varphi \vee \varphi'\|_{\mathrm{DT}} = \|\varphi\|_{\mathrm{DT}} \cup \|\varphi'\|_{\mathrm{DT}}$; $\|\neg\varphi\|_{\mathrm{DT}} = U - \|\varphi\|_{\mathrm{DT}}$. The formulae from $\mathscr{F}(C,V)$, $\mathscr{F}(D,V)$ are called condition formulae of DT and decision formulae of DT, respectively.

Any object $x \in U$ belongs to a *decision class* $\|\bigwedge_{a\in D} a = a(x)\|_{\mathrm{DT}}$ of DT. All decision classes of DT create a partition of the universe U.

A *decision rule* for DT is any expression of the form $\varphi \Rightarrow \psi$, where $\varphi \in \mathscr{F}(C,V)$, $\psi \in \mathscr{F}(D,V)$, and $\|\varphi\|_{\mathrm{DT}} \neq \emptyset$. Formulae φ and ψ are referred to as the *predecessor* and the *successor* of decision rule $\varphi \Rightarrow \psi$. Decision rules are often called "IF ... THEN ..." rules.

Decision rule $\varphi \Rightarrow \psi$ is *true* in DT if and only if $\|\varphi\|_{\mathrm{DT}} \subseteq \|\psi\|_{\mathrm{DT}}$. Otherwise one can measure its *truth degree* by introducing some inclusion measure of $\|\varphi\|_{\mathrm{DT}}$ in $\|\psi\|_{\mathrm{DT}}$.

Each object x of a decision table determines a *decision rule* $\bigwedge_{a\in C} a = a(x) \Rightarrow \bigwedge_{a\in D} a = a(x)$.

Decision rules corresponding to some objects can have the same condition parts but different decision parts. Such rules are called *inconsistent* (non-deterministic, conflicting, possible); otherwise the rules are referred to as *consistent* (certain, sure, deterministic, non-conflicting) rules. Decision tables containing inconsistent decision rules are called inconsistent (non-deterministic, conflicting); otherwise the table is consistent (deterministic, non-conflicting).

Numerous methods have been developed for different decision rule generation (see, e.g., Polkowski and Skowron, 1998; Skowron, 2000).

Another important issue in data analysis is discovering dependencies between attributes. Let D and C be subsets of A. We will say that D depends on C in a degree k $(0 \leqslant k \leqslant 1)$, denoted $C \Rightarrow_k D$, if

$$k = \gamma(C,D) = \frac{|\mathrm{POS}_C(D)|}{|U|}$$

where

$$\mathrm{POS}_C(D) = \bigcup_{X\in U/D} C_*(X)$$

called a *positive region* of the partition U/D with respect to C, is the set of all elements of U that can be uniquely classified to blocks of the partition U/D, by means of C.

If $k = 1$ we say that D depends totally on C, and if $k < 1$, we say that D depends partially (in a degree k) on C.

The coefficient k expresses the ratio of all elements of the universe, which can be properly classified to blocks of the partition U/D, employing attributes C and will be called the *degree of the dependency*.

The coefficient $1 - \gamma(C,D)$ can be called the inconsistency degree of DT (see Liu and Setiono, 1996).

The ability to discern between perceived objects is important for constructing many entities like reducts, decision rules or decision algorithms. In the classical rough set approach the discernibility relation $\mathrm{DIS}(B) \subseteq U \times U$ is defined by $x\mathrm{DIS}(B)y$ if and only if $\mathrm{non}(xI(B)y)$. However, this is in general not the case for the generalized approximation spaces (one can define indiscernibility by $x \in I(y)$ and discernibility by $I(x) \cap I(y) = \emptyset$ for any objects x, y).

The idea of Boolean reasoning (Brown, 1990) is based on construction for a given problem P a corresponding Boolean function f_P with the following property: the solutions for the problem P can be decoded from prime implicants of the Boolean function f_P. Let us mention that to solve real-life problems it is necessary to deal with Boolean functions having huge size and large number of variables.

It is important to note that the methodology allows to construct heuristics having a very important *approximation property* which can be formulated as follows: expressions generated by heuristics (i.e., implicants) *close* to prime implicants define approximate solutions for the problem.

Given an information system IS a *reduct* is a minimal set of attributes $B \subseteq A$ such that $I(A) = I(B)$. Finding a minimal reduct is NP-hard; one can also show that for any m (sufficiently

large) there exists an information system with m attributes having an exponential number of reducts. There exist fortunately good heuristics that compute sufficiently many reducts in an acceptable time.

Let IS be an information system with n objects. The discernibility matrix of IS is a symmetric $n \times n$ matrix with entries c_{ij} as given below. Each entry consists of the set of attributes upon which objects x_i and x_j differ.

$$c_{ij} = \{a \in A | a(x_i) \neq a(x_j)\} \quad \text{for } i, j = 1, \ldots, n$$

A discernibility function f_{IS} for an information system IS is a Boolean function of m Boolean variables $a_1^*, \ldots, a_m^*$ (corresponding to the attributes $a_1, \ldots, a_m$) defined by

$$f_{\text{IS}}(a_1^*, \ldots, a_m^*) = \wedge \left\{ \vee c_{ij}^* | 1 \leqslant j \leqslant i \leqslant n, c_{ij} \neq \emptyset \right\}$$

where $c_{ij}^* = \{a^* | a \in c_{ij}\}$. In the sequel we will write a_i instead of a_i^*.

The discernibility function f_{IS} describes constraints which should be preserved if one would like to preserve discernibility between all pairs of discernible objects from IS. It requires to keep at least one attribute from each non-empty entry of the discernibility matrix, i.e., corresponding to any pair of discernible objects. One can show (Skowron and Rauszer, 1992) that the sets of all minimal sets of attributes preserving discernibility between objects, i.e., reducts correspond to prime implicants of the discernibility function f_{IS}.

The intersection of all reducts is the so-called *core*. It is well known that choosing random reduct as a relevant set of features in information system will give rather poor results. Hence, several techniques have been developed to select relevant reducts or their approximations. Among them is one base on so-called *dynamic reducts* (Bazan et al., 1994, 1998). The attributes are considered relevant if they belong to dynamic reducts with a sufficiently high stability coefficient, i.e., they appear with sufficiently high frequency in random samples extracted from a given information system.

There are several kinds of reducts considered for decision tables. We will discuss one of them. Let $\mathscr{A} = (U, A, d)$ be a decision system (i.e., we assume, for simplicity of notation, the set D of decision attributes consists of one element d only, $D = \{d\}$ and $C = A$). The generalized decision in $\mathscr{A}$ is the function $\partial_A : U \to \mathscr{P}(V_d)$ defined by

$$\partial_A(x) = \{i | \exists x' \in U x' \text{ IND}(A) x \text{ and } d(x') = i\}$$

A decision system $\mathscr{A}$ is called consistent (deterministic), if $|\partial_A(x)| = 1$ for any $x \in U$, otherwise $\mathscr{A}$ is inconsistent (non-deterministic). Any set consisting of all objects with the same generalized decision value is called the generalized decision class. The decision classes are denoted by C_i where the subscript denotes the decision value.

It is easy to see that a decision system $\mathscr{A}$ is consistent if, and only if, $\text{POS}_A(d) = U$. Moreover, if $\partial_B = \partial_{B'}$, then $\text{POS}_B(d) = \text{POS}_{B'}(d)$ for any pair of non-empty sets B, $B' \subseteq A$. Hence, the definition of a decision–relative reduct: a subset $B \subseteq A$ is a relative reduct if it is a minimal set such that $\text{POS}_A(d) = \text{POS}_B(d)$. Decision–relative reducts may be found from a discernibility matrix: $M^d(\mathscr{A}) = (c_{ij}^d)$ assuming $c_{ij}^d = c_{ij} - \{d\}$ if $(|\partial_A(x_i)| = 1$ or $|\partial_A(x_j)| = 1)$ and $\partial_A(x_i) \neq \partial_A(x_j)$; $c_{ij}^d = \emptyset$, otherwise. Matrix $M^d(\mathscr{A})$ is called the decision–relative discernibility matrix of $\mathscr{A}$. Construction of the decision–relative discernibility function from this matrix follows the construction of the discernibility function from the discernibility matrix. One can show that the set of *prime implicants* of $f_M^d(\mathscr{A})$ defines the set of all decision–relative reducts of $\mathscr{A}$. Because the core is the intersection of all reducts, it is included in every reduct, i.e., each element of the core belongs to some reduct. Thus, in a sense, the core is the most important subset of attributes, since none of its elements can be removed without affecting of the classification power of attributes. Yet another kind of reducts, called reducts relative to objects can be used for generation of minimal decision rules from decision tables (Skowron, 2000). In some applications, instead of reducts we prefer to use their approximations called α-reducts, where $\alpha \in [0, 1]$ is a real parameter. For a given information system $\mathscr{A} = (U, A)$ the set of attributes $B \subseteq A$ is called α-reduct if B has non-empty intersection with at least $\alpha \cdot 100\%$ of non-empty sets $c_{i,j}$ of the discernibility matrix of $\mathscr{A}$. Different kinds of reducts and their approximations are discussed in literature as a basic constructs for reasoning about data repre-

sented in information systems or decision tables. It turns out that they can be efficiently computed using heuristics based on Boolean reasoning approach.

3. Feature selection

Feature selection is a process of finding a subset of features, from the original set of features forming patterns in a given data set, optimal according to the given goal of processing and criterion. An optimal feature selection is a process of finding a subset $A_{\text{opt}} = \{a_{1,\text{opt}}, a_{2,\text{opt}}, \ldots, a_{m,\text{opt}}\}$ of A, which guarantees accomplishment of a processing goal by minimizing a defined feature selection criterion $J_{\text{feature}}(A_{\text{feature\_subset}})$. A solution of an optimal feature selection does not need to be unique.

One can distinguish two paradigms in data model building, and potentially in an optimal feature selection (minimum construction paradigms): the Occam's razor and minimum description length principle (Rissanen, 1978).

In the virtue of the minimum construction idea, one of the techniques for the best feature selection could be based on choosing a minimal feature subset that fully describes all concepts (for example classes in prediction-classification) in a given data set (Almuallim and Dietterich, 1991; Pawlak, 1991). Let us call this paradigm a minimum concept description. However, this approach, good for a given (possibly limited) data set, may not be appropriate for processing of unseen patterns. A robust processing algorithm, with the associated set of features (reflecting complexity), is a trade-off between the ability of processing a given data set, versus generalization ability.

The second general paradigm of optimal feature selection, mainly used in classifier design, relates to selecting a feature subset which guarantees the maximal between-class separability for the reduced data sets. This relates to the discriminatory power of features.

Feature selection methods consists of two main streams (Duda and Hart, 1973; Fukunaga, 1990; Bishop, 1995; John et al., 1994): *open-loop methods* and *closed-loop methods*. The open-loop methods (filter method) are based mostly on selecting features using between-class separability criterion (Duda and Hart, 1973). They do not use a feedback from predictor quality for the feature selection process. The closed-loop methods (John et al., 1994) called also the wrapper methods, are based on feature selection using a predictor performance (and thus forming a feedback in processing) as a criterion of feature subset selection. A selected feature subset is evaluated using as a criterion $J_{\text{feature}} = J_{\text{predictor}}$ a performance evaluation $J_{\text{predictor}}$ of a whole prediction algorithm for the reduced data set containing patterns with the selected features as pattern's elements.

Let us consider a problem of defining a feature selection criterion for a prediction task based on an original data set T containing N cases $(\boldsymbol{a}, \text{target})$ constituted with n-dimensional input patterns $\boldsymbol{a}$ and *target* pattern of output. Assume that the m-feature subset $A_{\text{feature}} \subseteq A$ ought to be evaluated based on the closed-loop type criterion. A reduced data set T_{feature}, with patterns containing only m-features from the subset A_{feature}, should be constructed. Then a type of predictor $\text{PR}_{\text{feature}}$ (for example k-nearest neighbors, or neural network), used for feature quality evaluation, should be decided. This predictor ideally should be the same as a final predictor PR for a whole design; however, in simplified sub-optimal solution, a computationally less expensive predictor can be used only for feature selection purpose. Let us assume that, for the considered feature set A, a reduced feature data set A_{feature} has been selected and a predictor algorithm $\text{PR}_{\text{feature}}$ based on A_{feature}, used for feature evaluation, decided. Then, evaluation of feature quality can be provided using one of methods used for the final predictor evaluation. This will require defining a performance criterion $J_{\text{PR}_{\text{feature}}}$, of a predictor $\text{PR}_{\text{feature}}$, and an error counting method which will show how to estimate a performance through averaging of results. Consider as an example a hold-out error counting method for predictor performance evaluation. In order to evaluate performance of a predictor $\text{PR}_{\text{feature}}$, an extracted feature data set T_{feature} is split into a N_{tra} case training set $T_{\text{feature,tra}}$, and a N_{test} case test set $T_{\text{feature,test}}$ (hold out for testing). Each case $(\boldsymbol{a}^i_f, \text{target}^i)$ of both sets contains a feature pattern

$\boldsymbol{a}_f^i$ labeled by a target^i. The evaluation criteria can be defined separately for prediction-classification and prediction-regression.

We will consider defining feature selection criterion for a prediction-classification task, when a feature subset T_{feature} case contains pairs $(\boldsymbol{a}_f, c_{\text{target}})$ of a feature input pattern $\boldsymbol{a}_f$ and a categorical type target c_{target} taking value corresponding to one of possible r decision classes C_i. The quality of classifier $\text{PR}_{\text{feature}}$, computed basing on the limited size test set $T_{\text{feature,test}}$ with N_{test} patterns, can be measured using the following performance criterion $J_{\text{PR}_{\text{feature}}}$ (here equal to a feature selection criteria J_{feature})

$$J_{\text{PR}_{\text{feature}}} = \hat{J}_{\text{all miscl}} = \frac{n_{\text{all miscl}}}{N_{\text{test}}} \times 100\% \tag{1}$$

where $n_{\text{all miscl}}$ is the number of all misclassified patterns, and N_{test} is the number of all tested patterns. This criterion estimates the probability of error (expressed in percent) by the relative frequency of error. Usually some statistical methods (e.g., cross-validation techniques) are used to obtain better estimation of the predictor quality.

An overview of feature selection methods can be find in (Liu and Motoda, 1998a,b). Let us only mention that several methods of feature selection are inherently built in a predictor design procedure (Quinlan, 1993) and some methods of feature selection merge feature extraction with feature selection. A feature reduction (pruning) method for a self-organizing neural network map, based on concept description, is suggested in (Lobo et al., 1997).

We will concentrate in the following sections on rough set approach to feature selection and on some relationships of rough set methods with the existing ones.

3.1. Feature selection based on rough sets

Rough set approach to feature selection can be based on the minimal description length principle (Rissanen, 1978) and tuning methods of parameters of the approximation spaces to obtain high quality classifiers based on selected features. We have mentioned before an example of such parameter with possible values in the powerset of the feature set, i.e., related to feature selection. Other parameters can be used e.g., to measure the closeness of concepts (Skowron, 2000).

One can distinguish two main steps in this approach.

In the first step, by using Boolean reasoning relevant kinds of reducts from given data tables are extracted. These reducts are preserving exactly the discrenibility (and also some other) constraints (e.g., reducts relative to objects in process for minimal decision rules generation).

In the second step, by means of parameter tuning reduct approximations are extracted. These reduct approximations allow for shorter concept description than the exact reducts and they are still preserving the constraints to a sufficient degree to guarantee, e.g., a sufficient approximation quality of the described (induced) concept (Skowron, 2000).

In using rough sets for feature selection two cases can be distinguished, namely global and local feature selection scheme. In the former case the relevant attributes for the whole data table are selected while in the latter case the descriptors of the form (a, v) where $a \in A$ and $v \in V_a$ are selected for a given object. In both cases we are searching for relevant features for the object classification. In the global case we are searching for features defining a partition (or covering) of the object universe. This partition should be relevant for describing together with some other features the approximation of partition (or part of it) defined by the decision attribute. In the local case we are extracting descriptors defining together with some other descriptors a relevant neighborhood for a given object with respect to a decision class.

Using of rough sets (Pawlak, 1991; Bazan et al., 1994, 1998) to feature selection was proposed in several contributions (see, e.g., Swiniarski and Nguyen, 1996; Swiniarski et al., 1995). The simplest approach is based on calculation of a core for discrete attribute data set, containing strongly relevant features, and reducts, containing a core plus additional weakly relevant features, such that each reduct is satisfactory to determine concepts in the data set. Based on a set of reducts for a data set some criteria for feature selection can be formed, for example selecting features from a minimal

reduct, i.e., a reduct containing minimal set of attributes. In order to find a robust (well-generalizing) feature subset, dynamic reducts were proposed (Bazan et al., 1994, 1998). The selection of dynamic reduct is based on the cross-validation method. The methods of dynamic reducts generation have been applied for relevant feature extraction, e.g., for dynamic selection of features represented in discretization as well as in the process of relevant decision rules inducing. In order to find a robust feature subset, reflecting generalization, dynamic reducts were proposed in (Bazan et al., 1994, 1998), construction of which is based on cross-validation in feature subset selection. Some other methods based on non-invasive data analysis and rough sets are reported in (Duentsch and Gediga, 1997).

Let us now summarize the applications of rough set methods for feature selection in closed loop. The method is based on searching first for short (dynamic) reducts or reduct approximations. This step can be realized using for example software systems like ROSETTA (see www page: http://www.idi.ntnu.no/~aleks/rosetta/rosetta.html or RSES: see www page: www.roughsets.org). The next step is based on genetic algorithm with the fitness function measuring the quality of the selected reduct approximation B dependent, among others, on (i) the quality of the reduct approximation by the set B, (ii) the cardinality of the feature set B, (iii) the discernibility power of the feature set B with respect to the discernibility between decision classes measured, e.g., by means of the approximation quality of a D-reduct by B, (iv) the number of equivalence classes created by a feature set on a given data set and/or the number of rules generated by this set (Wróblewski, 2001), (v) the closeness of concepts (Skowron, 2000), (vi) the conflict resolution strategy (Szczuka et al., 2001). The parameters used to specify and compose the above components into a fitness function are tuned in evolutionary process to obtain the classifier of the highest quality using the feature set B. The classifier quality is measured by means of the quality of new object classification. Let us finally mention recently reported results based on ensambles of classifiers constructed on the basis of different reducts (see, e.g., Wróblewski, 2001).

In the following subsections we point out some relationships of rough set approach with the existing methods for feature selection. The conclusion is that these methods are strongly related to extraction of different kinds of reducts.

3.2. Relevance of features

There have been both deterministic and probabilistic attempts to define feature relevancy (Almuallim and Dietterich, 1991; John et al., 1994; Pawlak, 1991).

Rough set theory (Pawlak, 1991; Skowron, 2000) defines deterministic strong and weak relevance for discrete features and discrete targets. For a given data set a set of all strongly relevant features forms a *core*. A minimal set of features satisfactory to describe concepts in a given data set, including a core and possibly some weakly relevant features, form a *reduct*. A core is an intersection of reducts.

We will show that different kinds of definitions of relevant features correspond to different kinds of reducts.

Let us denote by $\boldsymbol{a}_i$ a vector of features (attributes) $(a_1, a_2, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n)$ obtained from the original feature vector $\boldsymbol{a}$ by removing a_i. By $\boldsymbol{v}_i$ is denoted a value of $\boldsymbol{a}_i$ (John et al., 1994).

A feature a_i is *relevant* if there exists some value v_i of that feature, a decision value (predictor output) v, and value $\boldsymbol{v}_i$ (generally a vector) for which $P(a_i = v_i) > 0$ such that

$$P(d = v, \boldsymbol{a}_i = \boldsymbol{v}_i | a_i = v_i) \neq P(d = v, \boldsymbol{a}_i = \boldsymbol{v}_i) \qquad (2)$$

In the light of this definition a feature a_i is relevant if probability of a *target* (given all features) can change if we remove knowledge about a value of that feature.

In (John et al., 1994) other definitions of *strong* and *weak* relevance were introduced.

A feature a_i is strongly relevant if there exists some value of that feature v_i, a value v (predictor output) of decision d and a value $\boldsymbol{v}_i$ of a $\boldsymbol{a}_i$ for which $P(a_i = v_i, \boldsymbol{a}_i = \boldsymbol{v}_i) > 0$ such that

$$P(d = v | \boldsymbol{a}_i = \boldsymbol{v}_i, a_i = v_i) \neq P(d = v | \boldsymbol{a}_i = \boldsymbol{v}_i) \qquad (3)$$

Strong relevance implies that a feature is indispensable, i.e., that its removal from a feature vector will change prediction accuracy.

Let us assume $\mathrm{DT} = (U, A, d)$ be a decision table where $V_d = \{1, \ldots, r\}$. The decision d defined the (target) classes $C_s = \{x \in U | d(x) = s\}$ for $s = 1, \ldots, r$. We define a new decision table $\mathrm{DT}_d = (U, A, d_A)$ assuming $d_A(x) = (\mu^A_{C_1}(x), \ldots, \mu^A_{C_s}(x))$ for $x \in U$. It means that the new decision is equal to the probability distribution defined by the case x in decision table DT. Now one can show that the reducts relative to such decision, called frequency related reducts (Ślęzak, 2001), consists of relevant features in the above defined sense, any maximal set of relevant features is a reduct of this kind.

One can also define reducts corresponding to the relevant features specified by means of the definition of relevant feature as well as the following one.

A feature a_i is weakly relevant if it is not strongly relevant, and there exists a subsequence $\boldsymbol{b}_i$ of $\boldsymbol{a}_i$, for which there exist: some value of that feature v_i, a decision value (predictor output) v of d, and a value $\boldsymbol{v}_i$ of vector $\boldsymbol{b}_i$, for which $P(a_i = v_i, \boldsymbol{b}_i = \boldsymbol{v}_i) > 0$ such that

$$P(d = v | \boldsymbol{b}_i = \boldsymbol{v}_i, a_i = v_i) \neq P(d = v | \boldsymbol{b}_i = \boldsymbol{v}_i) \tag{4}$$

One can observe that a weak relevance indicates that a feature might be dispensable (i.e., not relevant), however, sometimes (in companion with some other features) it may improve prediction accuracy.

A feature is relevant if it is either strongly relevant or weakly relevant, otherwise it is irrelevant. We can see that irrelevant features will never contribute to prediction accuracy, thus can be removed.

It has been shown in (John et al., 1994) that for some predictor designs feature relevancy (even strong relevancy) does not imply that a feature must be in an optimal feature subset.

3.3. Criteria based on mutual information

Entropy can be used as a mutual information measure of data set for feature selection. Let us consider a decision table (data set) $\mathrm{DT} = (U, A, d)$. Assume $A = \{a_1, \ldots, a_n\}$. Then any *n-dimensional* pattern vector $\boldsymbol{a}(x) = (a_1(x), \ldots, a_n(x))$ where $x \in U$ is labelled by a decision class from $C = (C_1, \ldots, C_r)$. The value of mutual information measure for a given feature set $B \subseteq A$ can be understood as the suitability of feature subset B for classification.

If initially only probabilistic knowledge about classes is given, then the uncertainty associated with the data can be measured by entropy

$$E(C) = -\sum_{i=1}^{r} P(C_i) \log_2 P(C_i) \tag{5}$$

where $P(C_i)$ is the a priori probability of a class C_i occurrence. It is known that entropy $E(C)$ is an expected amount of information needed for class prediction.

As a measure of uncertainty, the conditional entropy $E(C|B)$, upon subset of features B, can be defined for discrete features as

$$E(C|B) = -\sum_{\text{all } \boldsymbol{v}} P(\boldsymbol{v}) \left(\sum_{i=1}^{r} P(C_i|\boldsymbol{v}) \log_2 P(C_i|\boldsymbol{v}) \right) \tag{6}$$

More generally for continuous features we have

$$E(C|B) = -\int_{\text{all } \boldsymbol{v}} p(\boldsymbol{v}) \left(\sum_{i=1}^{r} P(C_i|\boldsymbol{v}) \log_2 P(C_i|\boldsymbol{v}) \right) \tag{7}$$

where $p(\boldsymbol{v})$ is a probability density function. The mutual information $\mathrm{MI}(C, B)$ between the classification and feature subset B is measured by a decrease of uncertainty about prediction of classes given knowledge about patterns $\boldsymbol{v}$ formed from features B

$$J_{\text{feature}}(B) = \mathrm{MI}(C, B) = E(C) - E(C|B) \tag{8}$$

One can consider entropy related reducts (Ślęzak, 2001) and Boolean reasoning to extract relevant feature sets with respect to the entropy measure. Moreover, using Boolean reasoning one can search for frequency related reducts preserving probability distributions to a satisfactory degree.

3.4. Criteria based on inconsistency count

An example of criteria for feature subset evaluation can be the inconsistency measure (Pawlak, 1991; Liu and Setiono, 1996).

The idea of attribute reduction can be generalized by introducing a concept of significance of attributes which enables to evaluate attributes not only in the two-valued scale dispensable—relevant (indispensable) but also in the multi-value case by assigning to an attribute a real number from the interval $[0,1]$ that expresses the importance of an attribute in the information table.

Significance of an attribute can be evaluated by measuring the effect of removing the attribute from an information table.

It was shown previously that the number $\gamma(C,D)$ expresses the degree of dependency between attributes C and D, or the accuracy of the approximation of U/D by C. It may be now checked how the coefficient $\gamma(C,D)$ changes when attribute a is removed. In other words, what is the difference between $\gamma(C,D)$ and $\gamma(C-\{a\},D)$. The difference is normalized and the significance of attribute a is defined by

$$\sigma_{(C,D)}(a) = \frac{(\gamma(C,D)-\gamma(C-\{a\},D))}{\gamma(C,D)} = 1-\frac{\gamma(C-\{a\},D)}{\gamma(C,D)}$$

Coefficient $\sigma_{C,D}(a)$ can be understood as a classification error which occurs when attribute a is dropped. The significance coefficient can be extended to sets of attributes as follows:

$$\sigma_{(C,D)}(B) = \frac{(\gamma(C,D)-\gamma(C-B,D))}{\gamma(C,D)} = 1-\frac{\gamma(C-B,D)}{\gamma(C,D)}$$

The inconsistency rate used in (Liu and Setiono, 1996) for a reduced data set can be expressed by $J_{\text{inc}}(B) = \sigma_{(C,D)}(B)$.

Another possibility is to consider as relevant the features that come from approximate reducts of sufficiently high quality.

Any subset B of C is called an approximate reduct of C and the number

$$\varepsilon_{(C,D)}(B) = \frac{(\gamma(C,D)-\gamma(B,D))}{\gamma(C,D)} = 1-\frac{\gamma(B,D)}{\gamma(C,D)}$$

is called an it error of reduct approximation. It expresses how exactly the set of attributes B approximates the set of condition attributes C with respect to determining D.

Several other methods of reduct approximation based on measures different from positive region have been developed. All experiments confirm the hypothesis that by tuning the level of approximation the quality of the classification of new objects may be increased in most cases. It is important to note that it is once again possible to use Boolean reasoning to compute the different types of reducts and to extract from them relevant approximations.

3.5. Criteria based on interclass separability

Some of the criteria for feature selection which are based on interclass separability are based on an idea of Fisher's linear transformation: a good feature (with a high discernibility power) should cause a small within-class scatter and a large between-class scatter (Duda and Hart, 1973; Fukunaga, 1990).

Rough set approach also offers methods to deal with interclass separability. In (Skowron, 1995) so-called D-reducts have been investigated. These reducts preserve not only discernibility between required pairs of cases (objects) but also they allow to keep the distance between objects from different decision classes above a given threshold (if this is possible).

3.6. Criteria based on minimum concept description

Open loop type criteria of feature selection based on minimum construction paradigm were studied (Almuallim and Dietterich, 1991) in machine learning and in statistics for discrete features noise free data sets. The straightforward techniques of best feature selection could be choosing a minimal feature subset that fully describes all concepts (for example classes in classification) in a given data set, (see, e.g., Almuallim and Dietterich, 1991; Pawlak, 1991). Here a criterion of feature selection could be defined as Boolean function $J_{\text{feature}}(B)$ with value 1 if a feature subset B is satisfactory to describe all concepts in a data set, otherwise having a value 0. The final selection would based on choosing a minimal subset for which a criterion gives value 1.

An idea of feature selection, with the minimum concept description criterion, can be extended by using concept of reduct defined in theory of rough sets (Pawlak, 1991; Skowron, 2000). A reduct is a minimal set of attributes that describes all concepts. However, a data set may have many reducts. If we use definition of the above open-loop feature selection criterion, we can see that for each reduct B we have maximum value of the criterion $J_{\text{feature}}(B)$. Based on a paradigm of the minimum concept description, we can select a minimum length reduct as the best feature subset. However, the minimal reduct is good for ideal situations where a given data set fully represents a domain of interest. For real life situations, and limited size data sets, other reduct (generally other feature subset) might be better for generalizing prediction. A selection of robust (generalizing) reduct, as a best open-loop feature subset, can be supported by introducing an idea of dynamic reduct (Bazan et al., 1994, 1998).

3.7. Feature selection with individual feature ranking

One of straightforward feature selection procedures is based on an evaluation of predictive power of individual features, then ranking such evaluated features, and eventually choosing the first best m features (Kudo and Sklansky, 2000). A criterion applied to an individual feature could be of either of the open-loop or closed-loop type. This algorithm has limitations and assumes independency of features, also relies on an assumption that the final selection criterion can be expressed as a sum or products of the criteria evaluated for each feature independently. It can be expected that a single feature alone may have a very low predictive power, whereas this feature when put together with others, may demonstrate significant predictive power.

One can attempt to select a minimal number $\hat{m}$ of the best ranked features that guarantees a performance better or equal to a defined level according to a certain criterion $J_{\text{feature,ranked}}$.

One of criteria evaluating predictive power of a feature could be defined by the rough set measure of significance of the feature (attribute) discussed before.

4. PCA and rough sets for feature projection, reduction and selection

Orthonormal projection and reduction of pattern dimensionality may improve the recognition process by considering only the most important data representation, possibly with uncorrelated elements retaining maximum information about the original data and with possible better generalization abilities.

We will discuss PCA for feature projection and reduction, followed by the joint method of feature selection using PCA and rough sets method.

4.1. PCA for feature projection and reduction

We generally assume that our knowledge about a domain is represented as a limited size sample of N random n-dimensional patterns $\boldsymbol{x} \in \boldsymbol{R}^n$ representing extracted object's features. We assume that an unlabeled training data set $T = \{\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^N\}$ can be represented as a $N \times n$ data pattern matrix $\boldsymbol{X} = [\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^N]^{\mathrm{T}}$. Let the eigenvalues of the covariance matrix $\boldsymbol{R}_x$ of $\boldsymbol{X}$ are arranged in the decreasing order $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \lambda_n \geqslant 0$, with the corresponding orthonormal eigenvectors $\boldsymbol{e}^1, \boldsymbol{e}^2, \ldots, \boldsymbol{e}^n$. Then the optimal linear transformation

$$\boldsymbol{y} = \widehat{\boldsymbol{W}}_{\mathrm{KLT}}\boldsymbol{x} \tag{9}$$

is provided using the $m \times n$ optimal Karhunen–Loéve transformation (KLT) matrix $\widehat{\boldsymbol{W}}_{\mathrm{KLT}} = [\boldsymbol{e}^1, \boldsymbol{e}^2, \ldots, \boldsymbol{e}^m]^{\mathrm{T}}$ composed with m rows being the first m orthonormal eigenvectors of the original data covariance matrix $\boldsymbol{R}_x$. The optimal matrix $\widehat{\boldsymbol{W}}_{\mathrm{KLT}}$ transforms the original n-dimensional patterns $\boldsymbol{x}$ into m-dimensional ($m \leqslant n$) feature patterns $\boldsymbol{y}$

$$\boldsymbol{Y} = (\widehat{\boldsymbol{W}}_{\mathrm{KLT}}\boldsymbol{X}^{\mathrm{T}})^{\mathrm{T}} - \boldsymbol{X}\widehat{\boldsymbol{W}}_{\mathrm{KLT}}^{\mathrm{T}} \tag{10}$$

minimizing the mean least square reconstruction error.

The open question remains, which principal components to select as the best for a given processing goal.

We have applied PCA, with the resulting KLT (Duda and Hart, 1973; Bishop, 1995), for the orthonormal projection (and reduction) of reduced SVD patterns $\boldsymbol{x}_{\mathrm{svd},r}$ representing recognized face images.

The selection of the best principal components for classification purpose is yet another feature selection problem. In the next section we will discuss an application of rough sets for feature selection/reduction.

4.2. Application of rough set based reducts for selection of discriminatory features from principal components

The PCA, with resulting linear KL projection, provides feature extraction and reduction optimal from the point of view of minimizing the reconstruction error. However, PCA does not guarantee that selected first principal components, as a feature vector, will be adequate for classification. Nevertheless, the projection of high dimensional patterns into lower dimensional orthogonal principal components feature vectors might help for some data types to provide better classification.

In many applications of PCA an arbitrary number of the first dominant principal components is selected as a feature vector. However, these methods do not cope with the selection of the most discriminative features well suitable for classification task. Even assuming that the KL projection can help in classification, and can be used as a first step in the feature extraction/selection procedure, still an open question remains: "Which principal components to choose for classification"?

One of possibilities for selecting features from principal components is to apply rough set theory (Pawlak, 1991; Skowron, 2000). Specifically, defined in rough set computation of a reduct can be used for selection some of principal components being a reduct. Thus these principal component will describe all concepts in a data set. For a suboptimal solution one can choose the minimal length reduct or dynamic reduct as selected set of principal components forming a selected, final feature vector.

The following steps can be proposed for PCA and rough set based procedure for feature selection. Rough sets assume that a processed data set contains patterns labeled by associated classes, with the discrete values of its elements (attributes, features). We know that PCA is predisposed to transform optimally patterns with real-valued features (elements). Thus after realizing the KLT, resulting projected patterns features must be discretized by some adequate procedure. The resulting discrete attribute valued data set (an information system) can be processed using rough set methods.

Let us assume that we are given a limited size data set T, containing N cases labeled by associated classes

$$T = \{(\boldsymbol{x}^1, c^1_{\mathrm{target}})(\boldsymbol{x}^2, c^2_{\mathrm{target}}), \ldots, (\boldsymbol{x}^N, c^N_{\mathrm{target}})\} \tag{11}$$

Each case $(\boldsymbol{x}^i, c^i_{\mathrm{target}})$ $(i = 1, 2, \ldots, N)$ is constituted with a n-dimensional real-valued pattern $\boldsymbol{x}^i \in \boldsymbol{R}^n$ with corresponding categorical target class c^i_{target}. We assume that a data set T contains N_i $(\sum_i^l N_i = N)$ cases from each categorical class c_i, with the total number of classes denoted by l.

Since PCA is an unsupervised method, first, from the original, class labeled data set T, a pattern part is isolated as $N \times n$ data pattern matrix

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}^1 \\ \boldsymbol{x}^2 \\ \vdots \\ \boldsymbol{x}^N \end{bmatrix} \tag{12}$$

where each row contains one pattern. The PCA procedure is applied for extracted pattern matrix $\boldsymbol{X}$, with resulting full size $n \times n$ optimal KL matrix $\boldsymbol{W}_{\mathrm{KLT}}$ (where n is a length of an original pattern $\boldsymbol{x}$). Now, according to the designer decision, the number $m \leqslant n$ of first dominant principal components has to be selected. Then the reduced $m \times n$ KL matrix $\boldsymbol{W}_{\mathrm{KLT}}$, containing only first m rows of a full size matrix $\boldsymbol{W}$, is constructed. Applying the matrix $\boldsymbol{W}_{\mathrm{KLT}}$ the original n-dimensional pattern $\boldsymbol{x}$ can be projected using transformation $\boldsymbol{y} = \boldsymbol{W}_{\mathrm{KLT}}\boldsymbol{x}$, into the reduced *m-dimensional* patterns $\boldsymbol{y}$ in the

principal components space. The entire projected $N \times m$ matrix $\boldsymbol{Y}$ of patterns can be obtained by the formula $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}_{\mathrm{KLT}}^{\mathrm{T}}$.

At this stage, the reduced, projected data set, represented by $\boldsymbol{Y}$ (with real-valued attributes), has to be discretized. As a result, the discrete attribute data set represented by the $N \times m$ matrix $\boldsymbol{Y}_d$ is computed. Then, the patterns from $\boldsymbol{Y}_d$ are labeled by the corresponding target classes from the original data set T. It forms a decision table DT_m with *m-dimensional* principal component related patterns. From the decision table DT_m one can compute the selected reduct $A_{\mathrm{feature,reduct}}$ of size l (for example minimal length or dynamic reduct) as a final selected attribute set. Here a reduct computation is pure feature selection procedure. Selected attributes (being a reduct) are some of elements of projected principal components vector $\boldsymbol{y}$.

Once the selected attribute set has been found (as a selected reduct), the final discrete attribute decision table $\mathrm{DT}_{f,d}$ is composed. It consist of these columns from the discrete matrix $\boldsymbol{Y}_d$ which are included in the selected feature set $A_{\mathrm{feature,reduct}}$. Each pattern in $\mathrm{DT}_{f,d}$ is labeled by the corresponding target class. Similarly one can obtain a real-valued resulting reduced decision table $\mathrm{DT}_{f,l}$ extracting (and adequately labeling by classes) these columns from the real-valued projected matrix $\boldsymbol{Y}$ which are included in the selected feature set $A_{\mathrm{feature,reduct}}$. Both resulting reduced decision tables can be used for a classifier design.

Algorithm. Feature extraction/selection using PCA and rough sets

Given: A N-case data set T containing n-dimensional patterns, with real-valued attributes, labeled by l associated classes $\{(\boldsymbol{x}^1, c^1_{\mathrm{target}}), (\boldsymbol{x}^2, c^2_{\mathrm{target}}), \ldots, (\boldsymbol{x}^N, c^N_{\mathrm{target}})\}$.

1. Isolate from the original class labeled data set T, a pattern part as $N \times n$ data pattern matrix $\boldsymbol{X}$.
2. Compute for the matrix $\boldsymbol{X}$ the covariance matrix $\boldsymbol{R}_x$.
3. Compute for the matrix $\boldsymbol{R}_x$ the eigenvalues and corresponding eigenvectors, and arrange them in descending order.
4. Select the reduced dimension $m \leqslant n$ of a feature vector in principal components space using defined selection method, which may base on judgement of the ordered values of computed eigenvalues.
5. Compute the optimal $m \times n$ KLT matrix $\boldsymbol{W}_{\mathrm{KLT}}$ based on eigenvectors of $\boldsymbol{R}_x$.
6. Transform original patterns from X into m-dimensional feature vectors in the principal component space by formula $\boldsymbol{y} = \boldsymbol{W}_{\mathrm{KLT}}\boldsymbol{x}$ for a single pattern, or formula $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}_{\mathrm{KLT}}$ for a whole set of patterns (where $\boldsymbol{Y}$ is $N \times m$ matrix).
7. Discretize the patterns in $\boldsymbol{Y}$ with resulting matrix Y_d.
8. Compose the decision table DT_m constituted with the patterns from the matrix $\boldsymbol{Y}_d$ with the corresponding classes from the original data set T.
9. Compute a selected reduct from the decision table DT_m treated as a selected set of features $A_{\mathrm{feature,reduct}}$ describing all concepts in DT_m.
10. Compose the final (reduced) discrete attribute decision table $\mathrm{DT}_{f,d}$ containing these columns from the projected discrete matrix $\boldsymbol{Y}_d$ which are correspond to the selected feature set $A_{\mathrm{feature,reduct}}$. Label patterns by corresponding classes from the original data set T.
11. Compose the final (reduced) real-valued attribute decision table $\mathrm{DT}_{f,r}$ containing these columns from the projected discrete matrix $\boldsymbol{Y}_d$ which are correspond to the selected feature set $A_{\mathrm{feature,reduct}}$. Label patterns by corresponding classes from the original data set T.

The results of discussed method of feature extraction/selection depend on a data set type and three designer decisions:

1. Selection of dimension $m \leqslant n$ of projected pattern in the principal component space.
2. Discretization method (and resulting quantization) of projected data.
3. Selection of a reduct.

First, for the selected dimension m, the applied quantization method may lead to the decision table DT_m for which no reduct exists. Then, a designer should return to the discretization step

and select other discretization. Even, if for all possible discretization attempts a reduct cannot be found, then a return is realized to the stage of selecting a dimension m of reduced feature vector y in principal component space. It means that possibly the projected vector does not contain satisfactory set of features. In this situation a design procedure should provide the next iteration with selected larger value of m. If for $m = n$ a reduct cannot be found, a data set is not classifiable in precise deterministic sense. Lastly, selection of reduct will impact an ability of a designed classifier to generalize prediction for unseen objects.

5. Numerical experiments

5.1. Face recognition

As a demonstration of a role of rough set methods for feature selection/reduction we have carried on numerical experiments of face recognition. We have considered ORL (see web page: www.cam-orl.co.uk/facedatabase.html) gray scale face image data sets. We have provided separately recognition experiments for 10 category data sets, and 40 category data sets of face images. Each category was represented by 10 instances of face images. Each gray scale face image was of the dimension 112×92 pixels. Feature extraction from face images has been provided by SVD.

Classification of face images was performed with a single hidden layer error back propagation neural network, learning vector quantization (LVQ) neural network, and rule-based rough set classifier.

5.1.1. SVD as a feature extraction from face images

SVD can be used to extract features from images. A rectangular $n \times m$ real image represented by $n \times m$ matrix $\boldsymbol{A}$, where $m \leqslant n$, can be transformed into a diagonal matrix by means of SVD. Assume the rank of matrix $\boldsymbol{A}$ is $r \leqslant m$. The matrices $\boldsymbol{A}\mathrm{A}^{\mathrm{T}}$ and $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ are non-negative, symmetric and have the identical eigenvalues λ_i. For $m \leqslant n$ there are at most $r \leqslant m$ non-zero eigenvalues. The SVD transform decomposes matrix $\boldsymbol{A}$ into the product of two orthogonal matrices $\boldsymbol{\Psi}$ of dimension $n \times r$, and $\boldsymbol{\Phi}$ of dimension $m \times r$ and a diagonal matrix $\boldsymbol{\Lambda}^{1/2}$ of dimension $r \times r$. The SVD of a matrix (image) $\boldsymbol{A}$ is given by

$$\boldsymbol{A} = \boldsymbol{\Psi}\boldsymbol{\Lambda}^{1/2}\boldsymbol{\Phi}^{\mathrm{T}} = \sum_{i=1}^{r} \sqrt{\lambda_i}\boldsymbol{\psi}_i\boldsymbol{\phi}_i^{\mathrm{T}} \tag{13}$$

where the matrix $\boldsymbol{\Psi}$, and $\boldsymbol{\Phi}$ have r orthogonal columns $\boldsymbol{\psi}_i \in \boldsymbol{R}^n$, $\boldsymbol{\phi}_i \in \boldsymbol{R}^m$ $(i = 1, \ldots, r)$, respectively (representing orthogonal eigenvectors of $\boldsymbol{A}\mathrm{A}^{\mathrm{T}}$ and $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$). The square matrix $\boldsymbol{\Lambda}^{1/2}$ has diagonal entries defined by

$$\boldsymbol{\Lambda}^{1/2} = \operatorname{diag}\left(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_r}\right) \tag{14}$$

where $\sigma_i = \sqrt{\lambda_i}$ $(i = 1, 2, \ldots, r)$ are the singular values of the matrix $\boldsymbol{A}$. Each λ_i, $(i = 1, 2, \ldots, r)$ is the non-zero eigenvalue of $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ (as well as $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$). Given a matrix $\boldsymbol{A}$ (an image) decomposed $\boldsymbol{A} = \boldsymbol{\Psi}\boldsymbol{\Lambda}^{1/2}\boldsymbol{\Phi}^{\mathrm{T}}$, and since $\boldsymbol{\Psi}$ and $\boldsymbol{\Phi}$ have orthogonal columns, thus the SVD transform of the image $\boldsymbol{A}$ is defined as

$$\boldsymbol{\Lambda}^{1/2} = \boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{\Phi} \tag{15}$$

If the matrix A represents an $n \times m$ image, then r singular values $\sqrt{\lambda_i}$ $(i = 1, 2, \ldots, r)$ from the main diagonal of the matrix $\boldsymbol{\Lambda}^{1/2}$, can be considered as extracted features of the image. These r singular values can be arranged as an image feature vector (SVD pattern) $\boldsymbol{x}_{\mathrm{svd}} = [\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_r}]^{\mathrm{T}}$ of an image.

Contrary to PCA, SVD is a purely matrix processing technique and not a direct statistical technique. The SVD decomposition is applied to each face image separately as a face feature extraction, whereas eigenfaces (Turk and Pentland, 1991) are obtained by projection of face vectors into principal component space derived statistically from the covariance matrix of the set of images.

Despite of the expressive power of the SVD transformation it is difficult to say arbitrarily how powerful the SVD features could be for a classification of face images.

The r-element SVD patterns can be heuristically reduced by removing its r_r trailing element which values are below heuristically selected threshold ϵ_{svd}. This can result in $n_{\mathrm{svd},r} = r - r_r$ element

reduced SVD patterns $x_{svd,r}$. In the next sections we discuss techniques of finding reduced set of face image features.

5.1.2. ORL data sets

The entire image data set was divided into training and test sets: 70% of these sub-images were used for the training set. Given original face images set, we have applied feature extraction using SVD of matrices representing image pixels. As a result, we have obtained for each image a 92 element x_{svd} SVD pattern with features being the singular values of an object matrix ordered in the descending order. In the next step we have carried out several simple classification experiments using SVD patterns of different length in order to estimate the sub-optimal reduction of those patterns. These patterns are obtained by cutting of trailing elements from the original 92-element SVD pattern.

These experiments have helped to select 60-element reduced SVD patterns $x_{svd,r}$. Then, according to the proposed method, we have applied PCA for feature projection/reduction based on reduced SVD patterns from the training set. Similarly as for the reduction for SVD pattern, we have provided several classification experiments for different length of reduced PCA patterns. These patterns are obtained by considering only a selected number of the first principal components. Finally, the projected 60-element PCA patterns have been in this way heuristically reduced to 20-element reduced PCA patterns $x_{svd,r,pca,r}$. In the last preprocessing step the rough set method has been used for the final feature selection/reduction of the reduced PCA continuous-valued patterns. For discretization of the continuous reduced PCA features we have applied the method of dividing each attribute value range into 10 evenly spaced bins. The discretized training set was used to find relevant reducts, e.g., the minimal reduct. This reduct was used to form the final pattern. The training, and the test sets (decision tables) with real-value pattern attributes were reduced according to the selected reduct.

In the paper we describe the simplest approach for relevant reduct selection. Existing rough set methods can be used to search for other forms of relevant reducts. Some of such methods we have mentioned in Section 3.1. Among them are those based on ensambles of classifiers (Dietterich, 1997). In our approach first a set of reducts of high quality is induced. Such set is used to construct a set of predictors and next from such predictors the global predictor is constructed using evolutionary approach (for details see, e.g., Wróblewski, 2001). Predictors based on these, more advanced methods make possible to achieve predictors of better quality. Certainly, the whole process of inducing such classifiers needs more time.

In all these cases statistical methods, e.g., cross validation techniques, are used to estimate the robustness of the constructed predictors.

5.1.3. Neural network classifier

The designed error backpropagation neural network classifier was composed with input layer, one hidden layer and output layer followed by the class choosing module. The network learning algorithm had momentum and adaptive learning techniques built into it. First, we have studied 10 category data set with 90% cases in the training set and 10% cases in the test set. We have selected the 5-element reduct basing on the reduced 20-element PCA pattern of the training set. The neural network with 50 neurons in the hidden layer has been designed. The number of hidden neurons was chosen on the basis of performed experiments. The neural network has provided 99% correct classification of the test set. The rough set rule based classifier for the discretized data set restricted to the attributes from the 5-element reduct has exhibited 100% accuracy.

We have also studied 40 category data set with total number of 400 cases. For this data sets we have selected 7-element reduct of the 320-case training set as a base for the final feature selection of reduced PCA patterns. The error backpropagation neural network with 300 neurons has been designed. The number of neurons in the hidden layer has been chosen experimentally. The neural network has provided 96.25% correct classification of the 320 case training set, and 75.5% accuracy for the 80 case test set. We have applied the resilient backpropagation algorithm as a network training function that updates weight and bias val-

ues, with the performance criterion goal 0.000299. The rough set rule based classifier for the discretized data set restricted to the attributes from the 7-element reduct has exhibited 94.5% accuracy for the 80 case test set.

The LVQ neural network, trained for the training set with reduced final patterns has provided 95.8% accuracy for the test set with 28 cases. The network has been trained for 200 code-book vectors and $k = 4$ neighbors.

The SVD has demonstrated a potential as a feature extraction method for face images. The processing sequence: SVD, PCA with KLT, and rough set approach created possibilities for a significant reduction of pattern dimensionality with increase of classification accuracy and generalization. The considered classifiers have demonstrated ability to recognize face images after such substantial reduction of pattern length.

5.2. Recognition of mammographic images

We have provided numerical experiments of mammographic images recognition. The MIAS MiniMammographic Database with 1024×1024 pixels images has been used in numerical experiments (Suckling et al., 1994). The database contains three types of class-labeled images: normal, benign (abnormal), and malignant (abnormal). For each abnormal image the coordinates of centre of abnormality and proximate radius (in pixels) of a circle enclosing the abnormality, have been given. For classifications the centre locations and radii apply to clusters rather than to the individual classifications. We have provided numerical experiments of recognition of normal and abnormal images (two category classification). We have selected randomly 144 images for recognition experiments. This set was divided into 128 case training set and 16 case test set.

From the original 1024 pixel gray scale mammographic image we have extracted a 64×64 pixels sub-image around the center of abnormality (or at the average coordinate for normal cases). For the selected 64×64 pixels sub-image we have applied a histogram method as a feature extraction method.

5.2.1. Feature extraction based on histogram

We have applied a histogram method for feature extraction from the mammographic images. First, for a given $n_r \times n_c$ image with n_r rows and n_c columns, the histogram $n_{hb} \times n_c$ matrix H for a given number of bins n_{hb} is extracted. Second, the $n_h = n_{hb} \times n_c$ element histogram pattern x_{h} is formed from the histogram matrix H by concatenating of its subsequent columns of H: $\boldsymbol{x}_h = [\boldsymbol{h}'_1 \cdots \boldsymbol{h}'_{nc}]'$. The 10 bin histogram has been used to form the $10 \times 64 = 640$ element histogram pattern.

5.2.2. Classification

For the extracted histogram patterns in the next phase the PCA method has been applied for pattern projection into principal component space followed by heuristic reduction of PCA pattern length. In the final preprocessing step the rough set methods have been applied for the final feature selection and data sets reduction.

Classification of mammographic images have been performed by the single hidden layer, error back propagation neural network. The PCA technique has allowed heuristic reduction of projected patterns to the length of 60 elements. Finally the rough set technique has resulted with 8 element reduced final patterns. For the reduced by rough set histogram patterns the error backpropagation network has provided 75.0% accuracy for the test set.

The sequence of data mining steps, including application of histogram for feature extraction, PCA, and rough set for projection and feature selection, has showed a potential for designing of neural network classifiers for mammographic images.

6. Conclusion

We have presented a rough set method and its role in feature selection for pattern recognition. We have proposed the sequence of data mining steps, including application of SVD, histograms, PCA, and rough sets for feature selection. This processing sequence has shown a potential for feasible feature extraction and feature selection in

designing of neural network classifiers for face images and mammographic images. The discussed method provides substantial reduction of pattern dimensionality. Rough set methods have shown ability to reduce significantly the pattern dimensionality and have proven to be viable data mining techniques as a front end of neural network classifiers.

Acknowledgements

The research has been partially supported by the COBASE project from NSF-National Research Council USA, National Academy of Sciences USA and Poland 2000–2001. Moreover, the research of Andrzej Skowron has been partially supported by the State Committee for Scientific Research of the Republic of Poland (KBN) research grant 8 T11C 025 19 and by the Wallenberg Foundation grant.

References

Almuallim, H., Dietterich, T.G., 1991. Learning with many irrelevant features. In: Proceedings of the Ninth National Conference on Artificial Intelligence. AAAI Press, Menlo Park, CA, pp. 552–574.

Bazan, J., 1998. A comparison of dynamic and non-dynamic rough set methods for extracting laws from decision system. In: Polkowski, L., Skowron, A. (Eds.), Rough Sets in Knowledge Discovery, Vol. 1. Physica-Verlag, Heidelberg, pp. 321–365.

Bazan, J., Skowron, A., Synak, P., 1994. Dynamic reducts as a tool for extracting laws from decision tables. In: Proceedings of the Symposium on Methodologies for Intelligent Systems, Charlotte, NC, LNAI 869. Springer-Verlag, Berlin, pp. 346–355.

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford University Press, Oxford.

Brown, F.M., 1990. Boolean Reasoning. Kluwer Academic, Dordrecht.

Dietterich, T., 1997. Machine learning research: Four current directions. AI Magazine 18 (4), 97–136.

Duda, R.O., Hart, P.E., 1973. Pattern Recognition and Scene Analysis. Wiley, New York.

Duentsch, I., Gediga, G., 1997. Statistical evaluation of rough set dependency analysis. Int. J. Human Comput. Stud. 46, 589–604.

Fukunaga, K., 1990. Introduction to Statistical Pattern Recognition. Academic Press, New York.

John, G., Kohavi, R., Pfleger, K., 1994. Irrelevant features and the subset selection problem. In: Machine Learning: Proceedings of the Eleventh International Conference (ICML-94). Morgan Kaufmann, Los-Altos, CA, pp. 121–129.

Kittler, J., 1986. Feature selection and extraction. In: Young, T.Y., Fu, K.S. (Eds.), Handbook of Pattern Recognition and Image Processing. Academic Press, New York, pp. 59–83.

Kudo, M., Sklansky, J., 2000. Comparison of Algorithms that Select Features for Pattern Classifiers. Pattern Recognit. 33 (1), 25–41.

Liu, H., Motoda, H., 1998a. Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic, Boston.

Liu, H., Motoda, H. (Eds.), 1998b. Feature Extraction, Construction and Selection: A Data Mining Approach. Kluwer Academic, Boston.

Liu, H., Setiono, R., 1996. A probabilistic approach to feature selection—A filter solution. In: Kumar, S. (Ed.), 13th International Conference on Machine Learning (ICML'96), Bari, Italy, pp. 319–327.

Lobo, V., Moura-Pires, F., Swiniarski, R., 1997. Minimizing the number of neurons for a SOM-based classification, using Boolean function formalization. Internal report San Diego State University, Department of Mathematical and Computer Sciences, 08/4/97.

Pawlak, Z., 1991. Rough Sets—Theoretical Aspects of Reasoning about Data. Kluwer Academic, Dordrecht.

Polkowski, L., Skowron, A. (Eds.), 1998. Rough Sets in Knowledge Discovery, vols. 1 and 2. Physica-Verlag, Heidelberg.

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufman, Los Altos, CA.

Rissanen, J., 1978. Modeling by shortest data description. Automatica 14, 465–471.

Skowron, A., Rauszer, C., 1992. The discernibility matrices and functions in information systems. In: Słowiński (Ed.), Intelligent Decision Support—Handbook of Applications and Advances of the Rough Sets Theory. Kluwer Academic, Dordrecht, pp. 331–362.

Skowron, A., 1995. Extracting laws from decision tables. Computat. Intell. 11 (2), 371–388.

Skowron, A., 2000. Rough sets in KDD (plenary talk). In: Shi, Z., Faltings, B., Musem, M. (Eds.), Proceedings of Conference on Intelligent Information Processing (IIP2000), 16-th World Computer Congress (IFIP'2000), Beijing, 19–25 August, 2000. Publishing House of Electronic Industry, Beijing, pp. 1–17.

Ślęzak, D., 2001. Approximate decision reducts. Ph.D. Thesis, Warsaw University.

Swiniarski, R. et al., 1995. Feature selection using rough sets and hidden layer expansion for rupture prediction in a highly automated production system. In: Proceedings of the 12th International Conference on Systems Science, Wroclaw, Poland.

Swiniarski, R., Nguyen, J., 1996. Rough set expert system for exture classification based on 2D spectral features. In: Proceedings of the Third Biennial European Joint Confer-

ence on Engineering Systems Design and Analysis ESDA'96. Montpellier, France, pp. 3–8.

Suckling, J. et al., 1994. The mammographic image analysis society digital mammogram database. Exerpta Medica: Int. Cong. Ser. 1069, 375–378.

Turk, M.A., Pentland, A.P., 1991. Face Recognition Using Eigenspaces. In: Proceedings CVPR'91, pp. 586–591.

Wróblewski, J., 2001. Adaptive methods for object classification. Ph.D. Thesis, Warsaw University.