PERGAMON

Pattern Recognition 32 (1999) 2011–2019

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/pattcog

Fuzzy feature selection

M. Ramze Rezaee, B. Goedhart, B.P.F. Lelieveldt, J.H.C. Reiber*

Division of Image Processing, Department of Radiology, Leiden University Medical Center, P.O. Box 9600, Building 1C2-S, 2300 RC Leiden, Netherlands

Received 26 November 1997; received in revised form 9 December 1998; accepted 9 December 1998

Abstract

In fuzzy classifier systems the classification is obtained by a number of fuzzy If–Then rules including linguistic terms such as *Low* and *High* that fuzzify each feature. This paper presents a method by which a reduced linguistic (fuzzy) set of a labeled multi-dimensional data set can be identified automatically. After the projection of the original data set onto a fuzzy space, the optimal subset of fuzzy features is determined using conventional search techniques. The applicability of this method has been demonstrated by reducing the number of features used for the classification of four real-world data sets. This method can also be used to generate an initial rule set for a fuzzy neural network. © 1999 Pattern Recognition Society. Published by Elsevier Science Ltd. All rights reserved.

Keywords: Feature selection; Fuzzy sets; Multi-dimensional data analysis; Fuzzy neural network; Pattern recognition

1. Introduction

The major objective of pattern classification systems is the automated recognition of data objects of different classes with a minimal rate of misclassification. In most cases, each data object (pattern) is represented numerically by a vector composed of the values of some measurable features. Although all of these features constitute the inputs of a classifier, they have different impacts on the classification performance. Some features may not increase the discriminative power of the classifier among pattern classes. Features have their individual cost components, e.g. related to computational overhead or economical expenses. In addition, some features may be highly correlated and some may even be irrelevant for a specific classification. A reduced feature set requires less training patterns in the training procedure of a pattern classifier, such as a neural network (the curse of dimensionality [1]). In addition, the training procedure would take less time and, due to fewer features (parameters), the classifier would obtain a higher generalization capability. Therefore, one of the crucial steps in the design of a pattern classifier system is the feature selection step by which an appropriate feature subset is selected automatically by the evaluation of candidate feature subsets.

In a typical fuzzy classifier system, the classification is explicitly described by a number of fuzzy If–Then rules. A fuzzy rule may look like IF X is *SMALL* AND Y is *LARGE* then Class1, where X and Y are features and *SMALL* and *LARGE* are fuzzy sets. In each classification rule each feature may be described by different fuzzy sets such as *SMALL*/*LOW*, *MEDIUM*, *LARGE*/*HIGH*, etc. In this paper, we propose a method by which optimal fuzzy sets can be selected automatically by using conventional search techniques and a representative labeled data set. In Section 2, a short survey of feature selection techniques, including conventional methods, genetic algorithms and neural network approaches, is given. The fuzzy feature selection method is motivated in Section 3 and the method is described. In Section 4, a number of data sets are described that was used in the evaluation study. The experimental results from these data sets are

*Corresponding author. Tel.: + 31-71-5263935/2137; fax: + 31-71-5266801.
E-mail address: hreiber@lkeb.azl.nl (J.H.C. Reiber)

PII: S0031-3203(99)00005-9

presented in Section 5. Finally, in Section 6 we discuss the utility and the limitations of our approach.

2. Feature selection techniques

An extensive amount of research has been carried out over the last two decades to obtain reliable methods for feature selection. These methods differ in the evaluation of feature subsets. The evaluation criteria are determined by the characteristics of the features and the specific objectives of the classifier. A number of evaluation criteria such as gain-entropy [2], relevance [3] and contingency table analysis [4] have been developed for the feature values, which lack an intrinsic order such as categorical, symbolic or nominal features (attributes). Features with real values may be evaluated by two approaches. By the first approach the overlap between classes can be measured by a number of interclass distance metrics. These definitions may take the probability density function distribution of different classes into account [1]. A feature subset for which the average class overlap is minimal is considered to be an optimal feature subset. The indices of fuzziness, entropy and π-ness (measures of fuzziness of a set) are also used to define an index of feature evaluation in terms of inter- and intra-class distances [5]. In addition, morphological elements of the data distribution have been used for the evaluation of features [6]. By the second approach the misclassification rate of a classifier is assessed. The error rate of all classifiers (each of which is trained for a specific feature subset) is estimated, so that the optimal feature subset for which the error rate is minimal can be selected.

Since for n number of features 2^n feature subsets exist, the evaluation of all possible feature subsets leads to computational problems for a large value of n, especially when evaluation of each feature subset is costly. To overcome an exhaustive search for finding an optimal feature subset, one of the following three methods may be applied: conventional search techniques, genetic algorithms and neural networks.

By conventional search techniques, feature selection is equivalent to searching a directed graph. An excellent review of methods for feature selection, including search strategies such as sequential forward/backward, bi-directional, beam and branch and bound [7] is provided by Siedlecki and Sklansky [8]. Although the criterion function for the evaluation of feature subsets can be the error rate of a classifier, inter-class distances or any other function, all search techniques may fail when the criterion function does not satisfy the monotonicity condition. This monotonicity condition requires that a criterion function changes monotonically over a sequence of nested feature subsets (features through the layers of the feature subset graph). This condition is, however, not guaranteed if the criterion function is the error rate of a non-Bayesian classifier.

Appropriate features can also be selected by genetic algorithms (GA) [9–12], which do not require the monotonicity condition. In addition, GA in contrast to other feature selection techniques can provide a number of optimal feature subsets. Each feature subset (called a chromosome) is evaluated by a fitness (criterion) function during an optimization cycle. The fitness function includes a number of evaluation criteria reflecting the objective of the classifier. Both the number of features in the subset and the error rate of a classifier have been used by Siedlecki and Sklansky [9] and Bril et al. [10]. Sahiner et al. [12] applied GA to select an optimal feature subset of a data set including 587 features. The fitness function contained two terms: the area under the receiver operating characteristics (ROC), and a penalty term analogous to Brills's utility term [10] that was linearly proportional to the number of selected features. Puch et al. [11] used GA to find an optimal weight value for each feature to warp the feature space.

A different approach in feature selection is based on the examination of the parameters of a trained classifier, e.g. a trained neural network, which is trained by using *all* (input) features of a representative data set. Tarr [13] suggested to use a saliency metric, which uses the sum of the squared weights between the input layer and the first hidden layer of a trained multi-layered feedforward neural network. By the introduction of noise as an extra feature input, Belu and Bauer [14] provided a method, which ranked the features from most significant to least significant. The significance of features was obtained by comparing their saliency with the saliency of the (input) injected noise. Priddy et al. [15] determined the saliency of input features based on the partial derivatives of the output nodes with respect to a given (input) feature. Steppe et al. [16] also used a likelihood-ratio test statistic, by which a trained neural network is pruned in a sequential procedure aimed at selecting the best neural network. A feature was examined by the evaluation of the weights of links between features of interest and the hidden layer and was removed when its associated weights were not statistically different from zero. Setiono and Liu [17] applied a network pruning algorithm that iteratively removed the irrelevant features (inputs) of a trained three-layer feedforward neural network. In each iteration all weights of the connections associated with a specific feature (input) were set to zero and the decrease of the network accuracy was measured. The feature with the smallest accuracy was then removed and the network was retrained in the following iteration.

3. Fuzzy feature selection

In many fuzzy classifier systems the classification is obtained by using a number of fuzzy If–Then rules of the form: if feature f_i is A_j $\dots$ *AND* feature f_k is B_p then class

C_c, where A_j and B_p are fuzzy sets (such as *LOW*, *MEDIUM*, etc.) and "*AND*" is a fuzzy logical operator. In most applications, however, the classification rules are not known in advance and therefore a procedure is needed by which fuzzy rules can be extracted automatically from a representative data set. One current approach is the fusion of fuzzy systems and neural networks [18–23]. For example, Wang and Mendel [19] take all possible fuzzy rules, defined by the number of inputs (features), the number of fuzzy sets for each feature and the number of outputs, into account. The fuzzy neural network proposed by Ishibuchi et al. [20], uses almost all possible fuzzy rules. For an m class problem where p number of features are involved and for each feature s number of fuzzy sets are used, the number of possible fuzzy rules with unspecified consequent (class label) are equal to s^p. Since in most implementations of fuzzy neural networks each fuzzy rule is implemented by a neuron in a hidden rule layer, the number of neurons in the rule layer will increase exponentially, when the number of features p or the fuzzy sets s is increased. This will complicate the training procedure and, in general, the generalization capability of the network will decrease. Furthermore, the examination of weights and neurons, using a pruning algorithm in order to find a reduced fuzzy set, is a difficult task.

From the above, it is clear that a procedure, which identifies a subset of fuzzy sets automatically is required when the number of fuzzy sets or the number of features becomes large. In addition, when an appropriate fuzzy set is found, one can define an initial fuzzy rule base, and train a fuzzy neural network proposed by Horikawa et al. [21] and Jang [22], which allows the training of an initial rule base.

Fuzzy feature selection can be described as follows:

Let a labeled data set $X = \{x_k \mid k = 1,2,\dots,n\}$ include n number of labeled patterns $x_k \in R^p$; p is the number of (real valued) features f_i. If we denote x_k^m as the value of the mth feature f_m of pattern x_k, then each pattern x_k of set X can be represented by a vector

$$x_k = [x_k^1, x_k^2, \dots, x_k^p]. \tag{1}$$

Suppose that all features are represented by a set F:

$$F = \{f_1, f_2, \dots, f_p\}. \tag{2}$$

We can project the original data set to a fuzzy space by using a membership set U defined as

$$U = \{\mu_{11}, \mu_{12}, \dots, \mu_{1q}, \mu_{21}, \mu_{22}, \dots, \mu_{2r}, \dots, \mu_{p1}, \mu_{p2}, \dots, \mu_{ps}\}, \tag{3}$$

where element μ_{ij} is the jth fuzzy set of feature f_i. The indices q, r and s are positive numbers that indicate the cardinality of the fuzzy sets of the first (f_1), the second (f_2) and the pth feature (f_p), respectively. The fuzzy projected set F_X of the original data set X is defined as $F_X = \{(x_k, \mu(x_k)) \mid k = 1,2,\dots,n\}$ where $\mu(x_k)$ is a vector represented as

$$\mu(x_k) = [\mu_{11}(x_k^1), \mu_{12}(x_k^1), \dots, \mu_{1q}(x_k^1), \mu_{21}(x_k^2), \mu_{22}(x_k^2), \dots, \mu_{2r}(x_k^2), \dots, \mu_{p1}(x_k^p), \mu_{p2}(x_k^p), \dots, \mu_{ps}(x_k^p)]. \tag{4}$$

Suppose that the number of fuzzy sets for each feature f_i is denoted as $|f_i|$. The fuzzy sets μ_{ij} are defined as

$$\mu_{ij} : x_k^i \to [0,1]; \forall i \in \{1, 2, \dots, p\} \wedge \forall j \in \{1, 2, \dots, |f_i|\} \wedge k \in \{1, 2, \dots, n\}. \tag{5}$$

The value of $\sum_{i=1}^p |f_i|$ quantifies the resolution of the set U. If the original data is p-dimensional, then its fuzzy projection is represented by a fuzzy space with a dimension equal to $\sum_{i=1}^p |f_i|$.

The fuzzy feature selection determines an optimal combination of fuzzy sets μ_{ij}. If each subset of fuzzy sets can be evaluated by a criterion function $J(.)$ and all possible combinations of fuzzy subsets is denoted by the power set Θ, then fuzzy feature selection becomes one of determining fuzzy set subset U_{optimal} satisfying

$$J(U_{\text{optimal}}) = E(J(U_i)), \quad \forall U_i \subseteq \Theta, \quad \Theta = 2^U, \tag{6}$$

where E may be the minimum or the maximum operator.

The optimal set U_{optimal} can be found by applying one of the three approaches described in Section 2. In case of conventional search or genetic algorithms one can minimize the criterion function $J(.)$. In addition, the third method, neural networks, can also be applied by using the data of the fuzzy projected set F_X as the inputs to a feed-forward neural network.

The fuzzy feature selection can be defined by the following pseudo algorithm:

1. project a labeled data set $X\{x_i \mid i = 1, \dots, n\}$ onto a fuzzy set F_X defined by $U\{\mu_{ij} \mid i = 1, 2, \dots, p \wedge j = 1, 2, \dots, |f_i|\}$. This projection may be defined by linear (e.g. triangular membership function) or nonlinear (e.g. exponential, sigmoid) functions;
2. define a classifier, a criterion function $J(.)$ and $\Theta = 2^U$;
3. use a feature selection method FS;
4. find U_{optimal} by using FS in such a way that $J(U_{\text{optimal}}) = E(J(U_i)); \forall Ui \subseteq \Theta$.

3.1. Dimensionality and sample size consideration

Quite often and certainly in case of the finite learning samples, the performance of the classifiers, based on estimated densities, will improve by involving more features; however, beyond a certain point the classification performance will deteriorate when adding more features. This is common knowledge in pattern recognition applications and is called peaking of classification. The relation between the classification error rate and the ratio of the samples per class to the number of features, was studied, among others, by Kanal [24], Foley [25] and Jain and Chandrasekaran [26]. Foley [25] demonstrated that the estimated error rate is extremely biased if this ratio is less than three. According to Foley, a reasonable engineering rule of thumb appears to be the following: if the ratio of sample to feature size is greater than three, then on average the estimated error rate will be close to the optimum error rate attained by the minimum probability of error classifier. Since in our approach the projection of a data set onto a fuzzy space will increase the dimension of the samples in a finite learning set, it is recommended to choose the number of fuzzy sets per feature in such a way that the ratio of the number of the training samples of each class to the number of generated features be greater than three.

4. Data sets and the parameters of the studies

To demonstrate the applicability of the fuzzy feature selection, four real-world data sets were analyzed. All data sets are public domain.[1] A short description of the data sets and the features are provided in Appendix A. The data sets were the Iris data set of Anderson-Fisher [27,28] (called IRIS in this paper), the Indian diabetes [29] (DIAB), the image segmentation set (IMSEG) and VEHICLES data set. In all studies the criterion function was the minimal error probability of a 5-nearest neighbor estimated by the Bootstrap [30] approach of 100 runs. In each run a data set was generated randomly out of the original data set and the error rate of the classifier was estimated. The default value of $|f_i|$ (the number of fuzzy sets of each feature) was set to two and the ratio of the number of samples per each class to $\sum_{i=1}^{p}|f_i|$ was greater than three in all randomly generated data sets. The fuzzy sets were defined by triangular functions. Fig. 1 shows the fuzzy sets definition of a feature. As this figure illustrates, the maximum and minimum values of data for each feature are used to define the fuzzy sets. For this type of membership functions for a feature f_i the following constraint is valid:

$$\sum_{j=1}^{|f_i|} u_{ij}(x_k) = 1. \tag{7}$$

In order to select the optimal fuzzy sets an exhaustive search is applied, when the number of features was small (data set IRIS and DIAB). For data sets of IMSEG and VEHICLES with a large number of features the sequential backward search method was applied.

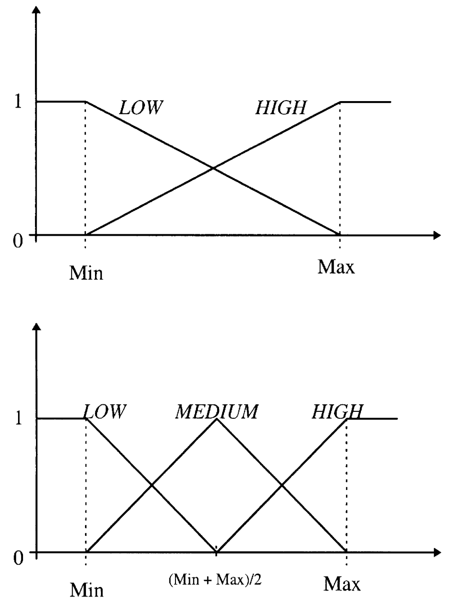

Fig. 1. Two examples of triangular membership functions.

Table 1
The best selected fuzzy features for Iris data set

Fuzzy resolution q	One feature	Two features	Three features
2	*S* PW	*S* PW, *S* PL	*S* PW, *S* PL, *H* PL
3	*M* PL	*M* PL, *S* PW	*M* PL, *S* PW, *M* PW

S: *SMALL*; *M*: *MEDIUM*; *L*:*LARGE*; PW: Petal width; PL: Petal length.

5. Results

5.1. IRIS data set (IRIS)

The average mean of the error rates of the classifier using all four original features was equal to 3.65% with a standard deviation (SD) of 0.02. Table 1 shows the best selected fuzzy sets for the IRIS data set for $|f_i| = 2$ and 3 fuzzy sets. As this table shows if only *SMALL* and *HIGH* membership functionals for each feature are used ($|f_i| = 2$), then the best fuzzy feature is *SMALL* petal width. The classification error rate of a 5-nearest neighbor when the best fuzzy sets are used are shown in Table 2. If

[1] All data, including the description of each data set, are available on ftp.ncc.up.pt/pub/statlog/datasets.

Table 2
Error rates of the best fuzzy features for IRIS data set

Fuzzy resolution (q)	Error rate of fuzzy feature/s (%)					
	One fuzzy set		Two fuzzy sets		Three fuzzy sets	
	Mean	SD	Mean	SD	Mean	SD
2 (*SMALL, HIGH*)	4.0	0.024	3.65	0.022	3.42	0.022
3 (*SMALL, MEDIUM, HIGH*)	15.33	0.035	6.3	0.027	3.64	0.022

SD: Standard deviation.

Table 3
The error rates of the selected features for DIAB set

Selected features	Error rate %	
	Mean	SD
All 8 features	28.07	0.02
L PL	33.27	0.026
L PL and *L* Age	27.6	0.02

SD: Standard deviation; *L*: *LARGE*; PL: Plasma glucose concentration a 2 hours in an oral glucose tolerance test.

Table 4
The error rates of selected features for IMSEG set

No. of features	Error rate (%)		Feature description
	Mean	SD	
All (19 features)	6.98	0.009	See Appendix A
2 fuzzy sets	10.15	0.009	*S* RCR, *S* RRM
3 fuzzy sets	4.22	0.008	*S* RCR, *S* RRM, *L* HM

S: SMALL; *L*: Large; RCR: Region centroid row; RRM: Raw red mean; HM: Hue mean.

only *SMALL* petal width be used, then the average error rate of the classifier is 4% with a SD of 0.024. For $|f_i| = 2$ and by using only three fuzzy sets an average error rate of 3.42% is achieved. The error was caused by the overlap between two classes in the fuzzy space constructed by those three fuzzy sets.

5.2. *Indian diabetes* (*DIAB*)

The average error rate of the classifier was equal to 28.7% when all eight original features were used (Table 3). Again, the average error was estimated by the Bootstrap method of 100 runs. The error rate of the classifier was equal to 27.6% when two fuzzy sets were used.

Table 5
The error rates of selected features for the VEHICLES set

No. of features	Error rate (%)		Feature description
	Mean	SD	
All (18 features)	37.21	0.022	See Appendix A
3 fuzzy sets	33.44	0.020	*L* PrAxis, *L* Max-Length, *L* Elongatedness

SD: Standard deviation; *L*: LARGE.

5.3. *Image segmentation set* (*IMSEG*):

The results of this study are presented in Table 4. The first column shows the number of features used by a 5-nearest neighbor classifier. In the second and third column the average error rate and the SD of 100 bootstrap runs are given, respectively. The fourth column shows which features were used for the classification. An average error rate of 6.98% was obtained when all 19 features of IMSEG were used. By using three fuzzy features the error rate decreased further to 4.22%. This is a decrease by about 3% in the average error rate and also a decrease by about 84% in the number of features used for the classification.

5.4. *VEHICLES data set*

The results of this study are illustrated in Table 5. An error rate of 37.21% was achieved when all 18 original features were used. The error rate of the three optimal fuzzy features was equal to 33.44%. This is a reduction by about 83% for the number of features used for the classification, and a reduction of about 4% of the error rate of the classifier. The three optimal fuzzy features correspond to the fifth, sixth and eighth feature of the VEHICLES data set.

6. Discussion

We have proposed a method by which the optimal fuzzy features of a data set can be selected automatically. In contrast to other feature selection techniques, instead of the original data its fuzzy projection is used to obtain the relevant fuzzy sets. The applicability of this approach has been demonstrated by using four real-world data sets. We have shown that by using simple (triangular) membership functions, a reduced fuzzy feature set may even reduce the classification error rate. The error rate of the classifier for the Iris data using a $|f_i| = 2$ and three fuzzy sets was equal to 3.42%. This value is comparable to the error rates mentioned in various publications [31,32]. The lowest error rate reported for an adaptive fuzzy rule-based classification system by Nozaki et al. [31] was equal to 2%, estimated by leaving-one-out approach and $|f_i| = 4$. The misclassification error rate reported in the same publication was, however, for $|f_i| = 2$ (the same value as in our approach) equal to 8%. The error rate of the Iris data set reported in [32] was equal to 4.7% for a fuzzy min–max classifier model. The estimated error rate for $|f_i| = 2$ and two fuzzy features was equal to 27.6% for the DIAB data set. This error value was even smaller than the error rates for all features used. This error rate is comparable to the average error rate of a neural-network feature selector reported by Setiono and Liu [17], which was equal to 25.71%. Both a decrease in the classification error rate (about 2.76% for three fuzzy features) and the number of features used (about 84%), was also obtained for the IMSEG data set.

In our approach, data is projected onto a fuzzy space using simple triangular membership functions. In addition in most studies the number of fuzzy sets of all features $|f_i|$ was equal to 2. In this context one can argue about the applicability of nonlinear membership functions, e.g. exponential or sigmoid functions or another $|f_i|$ value for a data set. The choice is, however, data dependent and can be specified by expert knowledge. Since membership functions of fuzzy sets are used to partition the feature (input) space, a nonlinear partitioning of the feature space may help to select optimal fuzzy sets. After the identification of optimal fuzzy sets, a higher performance may be obtained by tuning the parameters of the membership functions by a neural network. The results of the Iris data set for two values of $|f_i|$ (2 and 3) suggest that, in case of the triangular membership function, the error rate of the classifier will not decrease by a higher degree of partitioning (a larger $|f_i|$ value), when the error rate is near optimal. In most applications of fuzzy systems, the number of fuzzy features is not larger than 5. Since the computational overhead of fuzzy feature selection is proportional to the resolution of U, one can start the fuzzy features selection procedure by using a $|f_i|$ equal to 2, to avoid the exponential growth of possible fuzzy set combinations. After the identification of the optimal fuzzy sets a second iteration with $|f_i| = 3$ may identify other combinations of fuzzy sets with a lower classification error rate. The iteration may proceed up to a $|f_i|$ equal to 5, or it may stop when the error rate does not further decrease significantly. Due to the dimensionality and sample size consideration, the higher values of $|f_i|$ may affect the accuracy of error estimation. Therefore, the upper value of $|f_i|$ can also be determined by the dimensionality and sample size considerations.

Another possibility for the identification of $|f_i|$ is the partitioning of the feature space by a clustering algorithm such as the fuzzy c-mean [33]. The cluster prototypes can then be used for back projection [34] to obtain the fuzzy membership functionals and the resolution automatically.

We are now investigating the integration of the approach described in this paper in fuzzy neural networks. Since an initial rule set can be generated for a fuzzy neural network after the determination of the optimal fuzzy sets, the parameters of the membership functionals as well as the network connections can be further fine-tuned to obtain a better generalization and classification capability. Further, the automatic determination of the number of fuzzy sets for each feature using fuzzy clustering is under investigation.

7. Summary

In fuzzy classifier systems the classification is obtained by a number of fuzzy If–Then rules including linguistic terms such as *Low* and *High* that fuzzify each feature. In this paper a method is presented by which a reduced linguistic (fuzzy) set of a labeled multi-dimensional data set can be identified automatically. The selected fuzzy sets are optimal in terms of the classification performance. After the projection of the original data set onto a fuzzy space, the optimal subset of fuzzy features is determined using conventional search techniques. The applicability of this method has been demonstrated by reducing the number of features used for the classification of four real-world data sets. An evaluation study showed that even with a reduced number of fuzzy features, a better classification performance could be obtained than the classification based on all available features. For example, in case of the image segmentation data set of outdoor images, a decrease of about 3% in the misclassification rate and a decrease of 84% in the number of features used for classification, was achieved. For further optimization purposes, an optimal set of fuzzy features can also be used to generate an initial rule set for a fuzzy neural network.

Acknowledgements

This research is supported by the Science Foundation, The Netherlands, project LGN.22.2781. We wish to thank the people who provided the public domain data sets used in this paper. Also we would like to thank T. Rauber for using the TOOLDIAG analysis software package. In addition, we wish to thank the referees for their valuable comments and suggestions.

Appendix A

In this appendix the data sets including the features are explained.

A.1. IRIS data set (*IRIS*)

The Iris data set of Anderson–Fisher is a biometric data set consisting of 150 measurements belonging to three flower varieties: Setosa, Versicolor and Virginica. Each class includes 50 observations, in which two variables, length and width of the petal and sepal, are measured. Since the length and the width of each variable are measured, each individual measurement is represented as a point in a four-dimensional measurement space.

A.2. Indian diabetes (*DIAB*)

This is a data set including 768 instances of patients with or without signs of diabetes. Each instance includes eight numeric valued features. Data includes 500 "tested positive for diabetes" and 268 "tested negative" subjects (see Table 6).

A.3. Image segmentation set (*IMSEG*)

All 2310 instances of this data set were drawn randomly from a database of seven outdoor images. The images were hand-segmented to create a classification for every pixel. Each instance is a 3×3 region represented by 19 features describing local pixel information or the output of an image processing algorithm. The instances belong to one of the following classes: brickface, sky, foliage, cement, window, path and grass (see Table 7).

Table 6
Feature description of Indian diabetes (DIAB)

Feature no.	Description
1	Number of times pregnant
2	Plasma glucose concentration a 2 h in an oral glucose tolerance test
3	Diastolic blood pressure (mm Hg)
4	Triceps skin fold thickness (mm)
5	2-h serum insulin (mu U/ml)
6	Body mass index (weight in kg/(height in m)^2)
7	Diabetes pedigree function
8	Age (years)

Table 7
Feature description of image segmentation set (IMSEG)

Feature no.	Description
1	Region-centroid-col: the column of the center pixel of the region.
2	Region-centroid-row: the row of the center pixel of the region.
3	Region-pixel-count: the number of pixels in a region = 9.
4	Short-line-density-5: the results of a line extraction algorithm that counts how many lines of length 5 (any orientation) with low contrast, less than or equal to 5, go through the region.
5	Short-line-density-2: same as short-line-density-5 but counts lines of high contrast, greater than 5
6	Vedge-mean: measure the contrast of horizontally adjacent pixels in the region. There are 6, the mean and standard deviation are given. This attribute is used as a vertical edge detector.
7	Vegde-sd: (see 6)
8	Hedge-mean: measures the contrast of vertically adjacent pixels. Used for horizontal line detection.
9	Hedge-sd: (see 8).
10	Intensity-mean: the average over the region of (R + G + B)/3
11	Rawred-mean: the average over the region of the R value.
12	Rawblue-mean: the average over the region of the B value.
13	Rawgreen-mean: the average over the region of the G value.
14	Exred-mean: measure the excess red: (2R − (G + B))
15	Exblue-mean: measure the excess blue: (2B − (G + R))
16	Exgreen-mean: measure the excess green: (2G − (R + B))
17	Value-mean: 3-D nonlinear transformation of RGB. (Algorithm can be found in Folly and vanadium, Fundamentals of Interactive Computer Graphics)
18	Saturation-mean: (see 17)
19	Hue-mean: (see 17)

A.4. VEHICLES data set

This data set includes 846 samples, each of which includes 18 features extracted from the silhouette of four

Table 8
Feature description of VEHICLES

No.	Name	Description
1	COMPACTNESS	(average perim)**2/area
2	CIRCULARITY	(average radius)**2/area
3	DISTANCE CIRCULARITY	area/(av.distance from border)**2
4	RADIUS RATIO	(max.rad-min.rad)/av.radius
5	PR.AXIS ASPECT RATIO	(minor axis)/(major axis)
6	MAX.LENGTH ASPECT RATIO	(length perp. max length)/(max length)
7	SCATTER RATIO	(inertia about minor axis)/(inertia about major axis)
8	ELONGATEDNESS	area/(shrink width)**2
9	PR.AXIS RECTANGULARITY	area/(pr.axis length*pr.axis width)
10	MAX.LENGTH RECTANGULARITY	area/(max.length*length perp. to this)
11	SCALED VARIANCE ALONG MAJOR AXIS	(second-order moment about minor axis)/area
12	SCALED VARIANCE ALONG MINOR AXIS	(second-order moment about major axis)/area
13	SCALED RADIUS OF GYRATION	(mavar + mivar)/area
14	SKEWNESS ABOUT MAJOR AXIS	(third-order moment about major axis)/sigma_min**3
15	SKEWNESS ABOUT MINOR AXIS	(third-order moment about minor axis)/sigma_maj**3
16	KURTOSIS ABOUT MINOR AXIS	(fourth-order moment about major axis)/sigma_min**4
17	KURTOSIS ABOUT MAJOR AXIS	(fourth-order moment about minor axis)/sigma_maj**4
18	HOLLOWS RATIO	(area of hollows)/(area of bounding polygon)

sigma_maj**2 is the variance along the major axis, sigma_min**2 is the variance along the minor axis, area of hollows is area of bounding poly-area of object.

types of vehicles: a double Decker bus, Chevrolet van, Saab 9000 and Opel Manta 400. The images were acquired by a camera looking downwards at the model vehicles at different angles. All 128 × 128 images were thresholded to produce binary vehicle silhouettes, after which the "salt and pepper" image noise was removed by the application of morphological operations. The features were measured in the resulting images (see Table 8).

References

[1] P.A. Devijver, J. Kittler, Pattern Recognition: A Statistical Approach, Prentice-Hall, Englewood Cliffs, NJ, 1982.

[2] J.R. Quinlan, Introduction of decision trees, in: Machine Learning, vol. 1, Kluwer Acadamic Publishers, Dordrecht, 1986, pp. 81–106.

[3] P.W. Baim, A method for attribute selection in inductive learning systems, IEEE Trans. Pattern Anal. Mach. Intell. 10 (6) (1988) 888–896.

[4] T.W. Rauber, A.S. Steiger-Garcao, Feature selection of categorical attributes based on contingency table analysis, in: Proceedings of the 5th Portuguese Conference on Pattern Recognition, Porto, Portugal, 1993.

[5] S.K. Pal, B. Chakraborty, Fuzzy set theoretic measure for automatic feature evaluation, IEEE Trans. Systems. Man Cybernet. 16 (5) (1986) 754–760.

[6] V. Di Gesu, M.C. Maccarone, Features selection and possibility theory, Pattern Recognition 19 (1) (1986) 63–72.

[7] P.M. Narendra, K. Fukunaga, A branch and bound algorithm for feature subset selection, IEEE Trans. Comput. 26 (1977) 917–922.

[8] W. Siedlecki, J. Sklansky, On automatic feature selection, Int. J. Pattern Recognition Artif. Intell. 2 (2) (1988) 197–220.

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

[10] F.Z. Bril, D.E. Brown, N.M. Worthy, Fast genetic selection of features for neural network classifiers, IEEE Trans. Neural Networks. 3 (2) (1992) 324–328.

[11] W.F. Punch, E.D. Goodman, M. Pei, L. Chia-Shun, P. Hovland, R. Enbody, Further research on feature selection and classification using genetic algorithms, in: Proceedings of the 5th Int. Conf. on Genetic Algorithms, University of Illinois, Urbana-Champaign, IL, vol. 5, 1993, pp. 557–564.

[12] B. Sahiner, H. Chan, D. Wei, N. Petrick, M.A. Helvie, D.D. Adler, M.M. Goodsitt, Image feature selection by a genetic algorithm: application to classification of mass and normal breast tissue, Med. Phys. 23 (10) (1996) 1671–1684.

[13] G. Tarr, Multi-layered feedforward neural networks for image segmentation, Ph.D. Dissertation, School of Engineering, Air Force Institute of Technology, Wright-Patterson AFB OH, 1991.

[14] L.M. Belue, K.W. Bauer, Determining input features for multilayer perceptrons, Neurocomputing 7 (1995) 111–121.

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

[16] J.M. Steppe, K.W. Bauer, S.K. Rogers, Integrated feature and architecture selection, IEEE Trans. Neural Networks. 7 (4) (1996) 1007–1014.

[17] R. Setiono, H. Liu, Neural-network feature selector, IEEE Trans. Neural Networks 8 (3) (1997) 654–662.

[18] H. Takagi, I. Hayashi, NN-driven fuzzy reasoning, Int. J. Approx. Reason 5 (3) (1991) 191–212.

[19] L.X. Wang, J.M. Mendel, Fuzzy basis functions, Universal approximation, and orthogonal least squares learning, IEEE Trans. Neural Networks 3 (5) (1992) 807–814.

[20] H. Ishibuchi, K. Nozaki, H. Tanaka, Pattern classification by distributed representation of fuzzy rules, in: Proceedings of the IEEE International Conference on Fuzzy Systems, San Diego, USA, 1992, pp. 643–650.

[21] S. Horikawa, T. Furuhashi, Y. Uchikawa, On fuzzy modeling using fuzzy neural networks with the backpropagation algorithm, IEEE Trans. Neural Networks 3 (5) (1992) 801–806.

[22] J.S.R. Jang, ANFIS: adaptive-network-based fuzzy inference system, IEEE Trans. Systems Man Cybernet 23 (3) (1993) 665–685.

[23] S.K. Halgamuge, M. Glesner, Neural networks in designing fuzzy systems for real world applications, Int. J. Fuzzy Sets Systems 65 (1) (1994) 1–12.

[24] L. Kanal, On Dimensionality and sample size in statistical pattern classification, Pattern Recognition 3 (1971) 225–234.

[25] D.H. Foley, Consideration of sample and feature size, IEEE Trans. Inform. Theory IT-18 (5) (1972) 618–626.

[26] A.K. Jain, B. Chandrasekaran, Dimensionality and sample size considerations in pattern recognition practice, in: P.R. Krishnaiah, L.N. Kanal (Eds.), Handbook of Statistics, vol. 2, North-Holland, Amsterdam, 1982, pp. 835–855.

[27] E. Anderson, The irises of the Gaspe Peninsula, Bull. Am. Iris Soc. 59 (1935) 2–5.

[28] R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugenics 7 (II) (1936) 179–188.

[29] J.W. Smith, J.E. Everhart, W.C. Dickson, W.C. Knowler, R.S. Johannes, Using the ADAP learning algorithm to forecast the onset of diabetes mellitus, in: Proceedings of the Symp. on Computer Applications and Medical Care, IEEE Computer Society Press, Silverspring, MD, 1988, pp. 261–265.

[30] B. Efron, R. Tibshirani, An Introduction to the Bootstrap. Chapman & Hall, New York, 1993.

[31] K. Nozaki, H. Ishibuchi, H. Tanaka, Adaptive fuzzy rule-based classification systems, IEEE Trans. Fuzzy Systems. 4 (3) (1996) 238–250.

[32] J.F. Hurdle, The synthesis of compact fuzzy Neural circuits, IEEE Trans. Fuzzy Systems. 5 (1) (1997) 44–55.

[33] J.C. Bezdek, Fuzzy mathematics in pattern classification, Ph.D. Dissertation, Applied Mathematics, Cornell University, Ithaca, NY, USA, 1973.

[34] Z. Chi, H. Yan, Image segmentation using fuzzy rules derived from K-means clusters, J. Electron. Imaging 4 (2) (1995) 199–206.

About the Author—M. RAMZE REZAEE was born in Tehran, Iran in 1961. He received the M.Sc. degree from Delft University of Technology, Department of Electrical Engineering, Information Theory Group, in 1993 and the Ph.D. in image processing from the Leiden University in 1998. He is currently a post-doc at the Division of Image Processing, Department of Radiology of Leiden University Medical Center. His research interests are in the areas of fuzzy logic, neural network, pattern recognition, clustering and medical image analysis.

About the Author—B. GOEDHART was born in Delft, The Netherlands, in 1963. He received the M.Sc. degree from Delft University of Technology, Department of Mathematics and Computer Science in 1988 after which he received the Ph.D. degree from the Department of Electrical Engineering of the same university in 1994. He has been a member of LKEB since 1994. Dr. Goedharts research interests are in the areas of artificial intelligence and image processing.

About the Author—B. LELIEVELDT was born in Leiden, The Netherlands in 1969. He received the M.Sc. Degree from the Delft University of Technology, Department of Mechanical Engineering in 1994. He is currently active as Ph.D. student at the Laboratory of Clinical and Experimental Image Processing, Department of Radiology of the Leiden University Medical Center. His research interests are in the field of model based object recognition in medical images.

About the Author—J.H.C. REIBER was born in Haarlem, the Netherlands, in 1946. He received his M.ScEE. degree from the Delft University of Technology in 1971 and the Ph.D. degree in Electrical Engineering in 1975 from Stanford University, California, USA. In 1977 he founded the Laboratory for Clinical and Experimental Image Processing (LKEB) at the Thoraxcenter in Rotterdam, directing the research at the development and validation of objective and automated techniques for the segmentation of cardiovascular images, in particular for quantitative coronary arteriography (QCA), nuclear cardiology and echocardiography. With the move of LKEB in 1990 to the Leiden University Medical Centre (LUMC), the scope broadened to intravascular ultrasound, MRI, CT, etc., also in radiological applications. Since April 1995 he has been a Professor of Medical Image processing, in particular in Cardiovascular Applications at the LUMC and the Interuniversity Cardiology Institute of the Netherlands (ICIN). His research interests include (knowledge guided) image processing and its clinical applications. Since January 1990 he has been co-editor of the International Journal of Cardiac Imaging.