International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems
Vol. 4, No. 3 (1996) 245—253
© World Scientific Publishing Company

ON THE EQUIVALENCE BETWEEN FUZZY AND STATISTICAL CLASSIFIERS

LUDMILA I. KUNCHEVA*
Department of Electrical and Electronic Engineering,
Imperial College, Exhibition Road, London SW7 2BT, UK
e-mail: L.Kuncheva@ic.ac.uk

Received (15 July 1995)
Revised (15 March 1996)

In this paper the equivalence between fuzzy systems and two nonparametric techniques for pattern recognition is considered. The conditions under which a fuzzy system coincides with the nearest neighbor rule, and with the Parzen's classifier have been formulated.

Keywords: Pattern recognition, statistical nonparametric classifiers, fuzzy systems

1. Introduction

Pattern classification problem consists in assigning to an object, described as a point in a certain feature space $x \in \mathcal{S}^n$, a class label ω from a predefined set $\Omega = \{\omega_1, \ldots, \omega_M\}$. In the following we will assume that $\mathcal{S}^n$ coincides with $\Re^n$, i.e. each component x_i of x is a real number. The problem of designing a classifier is to find a mapping

$$D : \Re^n \to \Omega \tag{1}$$

optimal in the sense of a certain criterion $\mathcal{J}(D)$, provided we have a finite reference set of labeled samples $\mathbf{Z} = \{Z_1, \ldots, Z_N\}$, $Z_j \equiv [z_{j1} \ldots z_{jn}]^T \in \Re^n$. We will denote by $\theta_j \in \{1, \ldots, M\}$ the index of the class label among $\{\omega_1, \ldots, \omega_M\}$, associated with Z_j.

Usually, the ultimate goal is to design a classifier that assigns class labels with the smallest possible error over the whole feature space. The mapping D will be referred to as a *classifier*.

Let us consider the probabilistic framework where both x and ω are random variables. We denote by $P(\omega_i)$ the prior probability for class ω_i, $i = 1, ..., M$, and

[*] On leave from CLBME, Bulgarian Academy of Sciences,(e-mail:lucy@bgcict.acad.bg)

by $p(x/\omega_i)$ the class-conditional probability density function (p.d.f.). Assuming that $P(\omega_i)$ and $p(x/\omega_i)$ are known, the posterior probability can be calculated by

$$P(\omega_i/x) = \frac{P(\omega_i)\, p(x/\omega_i)}{p(x)}\ , \tag{2}$$

$$p(x) = \sum_k P(\omega_k)\, p(x/\omega_k).$$

Let $\mathcal{J}(D)$ be the overall error rate when using D. $\mathcal{J}(D)$ is called *risk function* and the "best" classifier is the one with minimum $\mathcal{J}$ over $\mathcal{R}^n$. It is a well known result in statistical decision theory that the *optimal* classifier in terms of $\mathcal{J}$ (called also *Bayesian* classifier) is the one that assigns to x the class label ω^* corresponding to the highest posterior probability, i.e.

$$\omega^* = arg \max_{\omega} P(\omega/x)$$

Fuzzy classification techniques are deemed to be a viable extension of classical ones towards handling nonstochastic uncertainty involved in the classification process. Along with "fuzzifying" classical pattern recognition techniques like linear discriminant analysis, k-nearest neighbor, treewise classifiers, etc., some purely fuzzy classification paradigms emerged (see, e.g., Pedrycz[10]). One of these is the fuzzy *if-then rules* based classification. It departures from the trivial look-up table but being its soft version, is much more versatile and powerful. It is just this type of fuzzy systems which has been pointed out by Brown et al.[2] to lack rigorous theory explaining how these systems generalize and also providing insights into the relative merits of differing implementation methodologies.

It has been stated in (Kosko, 1995)[8] that "... fuzzy systems are a class of probabilistic systems". Some studies show the connection between fuzzy systems and Radial Basis Function (RBF) networks[6]. In turn, RBF networks have been shown to be asymptotically Bayes-optimal classifiers by proving their equivalence to the kernel type of nonparametric statistical classifiers[1,3,7,9]. This means that under certain conditions on the basis functions (corresponding to membership functions), and with the number of hidden-layer nodes (corresponding to the rules) tending to infinity, the RBF network approaches the optimal (Bayesian) classification rate. Having these two relations in mind, we can conclude that fuzzy systems are also asymptotically Bayes-optimal classifiers. All this suggests that a more direct link between fuzzy systems and statistical classifiers can be established, excluding RBF networks from the chain.

The paper tries to answer the question: Under what conditions can a fuzzy system be considered as a probabilistic classifier? In Section 2 the notion of fuzzy system for classification is defined. Two nonparametric statistical classification techniques are briefly outlined in Section 3: the k-nearest neighbor rule, and the Parzen's classifier. Section 4 contains the two theorems showing the equivalence between fuzzy and statistical classifiers. Some comments are presented in Section 5.

2. Fuzzy systems for classification

The discussion here will be confined to a particular class of fuzzy systems defined below.

Let $y(x) = [y_1(x), ..., y_M(x)]^T$ denote a vector produced by the classifier D when x is submitted to its input. The coordinates of $y(x)$ correspond to the set of classes Ω. According to (1) we construct $y(x)$ so that

(i) $y_i(x) \in \{0, 1\}$, $i = 1, \ldots, M$, where 1 means "true" and 0 "false" of the respective class label, and

(ii) $\sum_{i=1}^{M} y_i(x) = 1$, $\forall x \in \Re^n$, i.e. one and only one class must be true for any x.

Let $A_i = \{A_{i,1}, \ldots, A_{i,s_i}\}$ be a term set defined over the i-th feature axis, $i = 1, \ldots, n$. Each $A_{i,j}$ is a fuzzy set with membership function

$$\mu_{A_{i,j}} : \Re \to [0, 1].$$

Definition 1 *A fuzzy system for classification is the mapping*

$$\tilde{D} : \Re^n \to \Omega$$

where $\tilde{D}$ is formed as follows:

(a) Using the term sets A_i, $i = 1, ..., n$ we consider K rules of the form:

$$R_k \equiv IF \ (x_1 \ is \ A_{1,k_1}) \ and \ \ldots \ and \ (x_n \ is \ A_{n,k_n})$$

$$THEN \ (y_{i_k}^k \ is \ 1) \ and \ (y_i^k \ is \ 0, \ i \neq i_k, \ i = 1, ..., M), \ k = 1, ..., K,$$

where y_l^k denotes the l-th component of the output vector y^k, associated with the kth rule.

(b) Let $R_k(x)$ denote the strength of activation of the if-part of the rule. $R_k(x)$ is obtained by applying a certain t-norm to the degrees of satisfaction of the clauses $(x_l \ is \ A_{l,k_l})$.

(c) The fuzzy output is a vector $\tilde{y}(x) \in [0, 1]^M$ whose components are the aggregated values

$$\tilde{y}_i(x) = \mathcal{A}_{k=1}^{K}(y_i^k, R_k(x)), \ i = 1, ..., M.$$

where $\mathcal{A}$ is an aggregation operator.

(d) *The final output $y(x)$ is defined by the winner-take-all principle:*

$$y_i(x) = \begin{cases} 1, & \textit{if} \;\; \tilde{y}_i(x) = \max_j \tilde{y}_j(x), \\ 0, & \textit{otherwise} \end{cases}$$

If a tie occurs it is broken at random, i.e. $\tilde{y}_i(x)$ takes value 1 for a randomly chosen one of those classes that have obtained equal maximal support, and 0, for the rest.

In sequel, a *fuzzy system for classification* will be referred to simply as a *fuzzy system.*

3. Two nonparametric statistical classifiers

3.1. *k-nearest neighbor rule*

One of the most theoretically elegant and yet simple classification techniques is the k-nearest neighbor rule (k-NN)[4,5]. According to it, x is assigned the label of the class where belong the majority if its k nearest samples from $\mathbf{Z}$ in terms of a certain metric over the feature space. The rule is asymptotically Bayes-optimal, i.e. the error rate approaches the Bayes one with $N \to \infty$; $k \to \infty$ and $k/N \to 0$.

Another important result is that when k is fixed to 1 (the nearest neighbor rule), and $N \to \infty$, the error rate is bounded from above by twice Bayes error rate.

3.2. *Parzen's classifier*

Parzen's classifier is based on a nonparametric approximation of class-conditional p.d.f.s plugged in the Bayes formula (2). Let $K(x)$ be a *kernel function* (or *Parzen window*) which peaks at zero, is nonnegative, and has integral one over $\Re^n$. The multidimensional kernel function centered around $Z_j \in \Re^n$ is usually expressed in the form $\frac{1}{h^n} K\left(\frac{x - Z_j}{h}\right)$, where h is a smoothing parameter depending on the number of samples N. Then we can approximate the class-conditional p.d.f.s using the sample set $\mathbf{Z}$ by[5,1]

$$\hat{p}_N(x/\omega_i) \;=\; \frac{1}{N_i} \sum_{j,\, \theta_j = i} \frac{1}{h^n} K\left(\frac{x - Z_j}{h}\right), \;\; Z_j \in \mathbf{Z}.$$

where N_i is the number of elements of $\mathbf{Z}$ from class ω_i. The estimate will be asymptotically unbiased if[5]

$$\lim_{N \to \infty} h(N) = 0.$$

Taking as estimates of the prior probabilities:

$$\hat{P}(\omega_i) = \frac{N_i}{N}$$

we arrive at a plug-in estimate of the posterior probability

$$\hat{P}_N(\omega_i/x) = \frac{1}{N\;p(x)} \sum_{j,\;\theta_j=i} \frac{1}{h^n}\,K\left(\frac{x-Z_j}{h}\right). \tag{3}$$

Let $I(i,j)$ be an indicator function such that

$$I(i,j) = \begin{cases} 1, & \text{if } \theta_i = j, \text{ i.e., } Z_j \text{ comes from class } \omega_i; \\ 0, & \text{otherwise} \end{cases}$$

Then (3) can be rewritten as

$$\hat{P}_N(\omega_i/x) = \frac{1}{N}C_1(x) \sum_{j=1}^{N} I(i,j)\;K\left(\frac{x-Z_j}{h}\right). \tag{4}$$

where the term $C_1(x)$ depends on x and N but not on the class label.

By definition, the approximation of the conditional p.d.f.s under the above conditions is asymptotically unbiased. Having plugged the estimates into the Bayes rule (2), it follows that the resultant classifier is also asymptotically Bayes-optimal.

A common choice of the kernel function is the multidimensional Gaussian kernel[5]:

$$\frac{1}{h^n}\,K_G\left(\frac{x-Z_k}{h}\right) = \frac{1}{h^n\,\sqrt{(2\pi)^n}\,\sqrt{det(\Sigma)}}\;\exp^{-\frac{1}{2h^2}\,(x-Z_k)^T\cdot\Sigma^{-1}\cdot(x-Z_k)} \tag{5}$$

For this case the posterior probabilities (4) become

$$\hat{P}_N(\omega_i/x) = \frac{1}{N}C_2(x) \sum_{j=1}^{N} I(i,j)\;K_G\left(\frac{x-Z_j}{h}\right), \quad i = 1, ..., M. \tag{6}$$

Variety of kernel functions are listed in (Fukunaga, 1972)[5]. In fact, a large class of widely used membership functions comply with the definition of a kernel function up to a normalizing constant.

4. Equivalence between fuzzy systems and statistical classifiers

Let $A_i,\ i = 1, ..., n$ be formed in one-to-one correspondence with $\mathbf{Z}$, i.e.

$$A_i = \{A_{i,1}, ..., A_{i,N}\} = \{\text{“}like\_Z_1\_on\_x_i'', \ldots, \text{“}like\_Z_N\_on\_x_i''\} \tag{7}$$

We consider the following set of N rules:

$$\begin{aligned} R_k \;\equiv\; \text{“}like\_Z_k'' \;\equiv\; IF\;\; (x_1\; is\; A_{1,k})\;\; and\; \ldots\; and\;\; (x_n\; is\; A_{n,k}) \\ THEN\;\; (y_{\theta_k}^k = 1)\;\; and\;\; (y_j^k\; is\; 0,\; j \neq \theta_k) \end{aligned} \tag{8}$$

The aggregation operators $\mathcal{A}$ which will be used in the proofs below are

- *max-product*

$$\mathcal{A}_{k=1}^{K}(y_i^k, R_k(x)) \equiv \max_{k=1,\ldots,K} \{y_i^k \cdot R_k(x)\} \;\text{ and}$$

- *COG-product*

$$\mathcal{A}_{k=1}^{K}(y_i^k, R_k(x)) \equiv \frac{\sum_{k=1}^{K} y_i^k \cdot R_k(x)}{\sum_{k=1}^{K} R_k(x)}$$

Theorem 1 *Let a fuzzy system be defined by (7) and (8), with $A_{i,j}$ having membership functions*

$$\mu_{like\_Z_j\_on\_x_i}(x_i) = \exp^{(-(x_i - z_{ji})^2)}$$

Let the AND operation in the if-part be implemented as product, and the aggregation $\mathcal{A}$ be implemented by max-product operator. Then the fuzzy system **is equivalent** *to the nearest neighbor classification rule.*

Proof. The activation strength of the k-th rule is

$$R_k(x) \ = \ \prod_{i=1}^{n} \ \mu_{like\_Z_k\_on\_x_i}(x_i) \ \ = \ \ \exp^{-\sum_{i=1}^{n}(x_i - z_{ki})^2}$$

Therefore, the points equidistant from Z_k will have the same membership value. The closer the point to Z_k, the higher the membership value. The i-th fuzzy output is

$$\tilde{y}_i(x) = \max_{k=1,\ldots,N} \{y_i^k \cdot R_k(x)\}. \qquad (9)$$

Since $y_i^k = 1$ when the class label of Z_k is ω_i, and 0, otherwise, (9) can be rewritten as

$$\tilde{y}_i(x) = \max_{Z_k,\, \theta_k = i} R_k(x).$$

Clearly, $\tilde{y}_i(x)$ is the membership grade corresponding to that vector from $\mathbf{Z}$ which belongs to class ω_i, and which has the smallest Euclidean distance to the input vector x. Therefore, by selecting the class label of the highest output, we assign to x the class of its nearest neighbor. □

Theorem 2 *Let a fuzzy system be defined by (7), (8), with $A_{i,j}$ having membership functions*

$$\mu_{like\_Z_j\_on\_x_i}(x) = \exp^{\left(-\frac{(x_i - z_{ji})^2}{2h^2}\right)}$$

where h is a parameter. Let the AND operation in the if-part be implemented as product, and center-of-gravity (COG) defuzzification operator be used as the aggregation $\mathcal{A}$. Then the fuzzy system **is equivalent** *to the Parzen's classifier.*

Proof. The activation strength of the k-th rule is

$$R_k(x) \ = \ \prod_{i=1}^{n} \ \mu_{like\_Z_k\_on\_x_i}(x_i) \ \ = \exp^{-\frac{1}{2h^2}\sum_{i=1}^{n}(x_i - z_{ki})^2}$$

We can rewrite the above equation as:

$$R_k(x) \ = \exp^{-\frac{1}{2h^2}\ (x-Z_k)^T\cdot\Sigma^{-1}\cdot(x-Z_k)},$$

where the covariance matrix Σ is equal to the identity matrix I_n. Then $R_k(x)$ differs from the Gaussian kernel (5) only by a multiplication constant, so we can equivalently substitute

$$R_k(x) = C_3\ K_G\left(\frac{x-Z_k}{h}\right).$$

The output $\tilde{y}_i(x),\ i=1,\ldots,M$ obtained through COG defuzzification is

$$\tilde{y}_i(x) = \frac{\sum_{k=1}^{N} y_i^k\cdot R_k(x)}{\sum_{k=1}^{N} R_k(x)} \ = \ C_4(x)\cdot\sum_{k=1}^{N} y_i^k\cdot K_G\left(\frac{x-Z_k}{h}\right). \tag{10}$$

Noting that y_i^k plays the role of the indicator function, we observe that equations (6) and (10) differ only by a coefficient that does not depend on the class i. Obviously, the decision obtained by choosing the class label of the output with the maximal value is the same in both cases. Moreover this decision is optimal in Bayesian sense. Therefore the fuzzy system is equivalent to the Parzen's classifier. □

5. Discussion and conclusions

In this paper two theorems have been proven showing the link between fuzzy systems and two nonparametric statistical classifiers both from functional and morphological perspective.

It is worth mentioning that:

- The proved equivalence implies that all the asymptotic properties of the statistical classifiers hold for the fuzzy system under the specified conditions. For the finite-sample case, however, the behavior of both statistical and fuzzy classifiers is difficult to study. Fuzzy systems are believed to possess a richer toolbox for tuning the classification rule, thereby leading to better results.

- The knowledge representation here is not in the form of widely used linguistic categories, e.g., {*small, medium, high*}, or any similar systematic ranking. It is rather in a case-based form. Indeed, if we consider a quantization of x_i into sequential categories, and describe $Z_j \in \mathbf{Z}$ in terms of these, we can arrive at a classical system of fuzzy if-then rules. Then each $A_{i,j}$ will have a linguistic label, like, e.g. *"small"*, *"perfect"*, *"normal"*, etc. In this case, however, the straightforward parallel between the fuzzy and the statistical classifiers considered here will be obscured. Instead, we will have to bring into consideration some classification techniques suitable for discrete variables.

- Generally, the shape of the kernel function is not limited to the Gaussian. Similar equivalence can be searched for replacing Gaussians with some widely used membership functions. It is pointed out by Bishop[1] that instead of increasing enormously the number N (number of the rules in the fuzzy system) we can tune the kernel functions in order to achieve good results in the finite sample case. This suggest using different membership functions for describing the labels of the term set, varying both with respect to the coordinate axes and to the members of the reference set $\mathbf{Z}$.

- The parameter h defines the spanning of the influence of a given sample and also may differ from one to another sample from $\mathbf{Z}$, thus giving another tuning instrument. In theory it is assumed that h is equal for all kernels. The practical guideline for selecting a proper value, however, is very vague: if h is *too small* the approximation will tend to ovetfit the data set and will be too "noisy". If h is *too large*, the estimated densities are oversmoothed, thus obliterating eventual shape features of the p.d.f.s. The compromise can be solved empirically, during the training process of the classifier.

- Classical fuzzy systems use membership functions separately on the feature axes x_i, $i = 1, ..., n$. Then the firing operation is supposed to account for all the dependencies between the features. In contrast, in the considered equivalence, the kernel function can be *nonseparable.* This means that the dependencies between features can be modeled by choosing the proper type of membership function, or better by tuning its free parameters (e.g., the covariance matrix) during training stage.

6. Acknowledgements

This work has been supported by a fellowship kindly provided by the Royal Society and the Foreign & Commonwealth Office. The author wish to express her sincere thanks to the two anonymous referees for their valuable comments and suggestions.

7. References

1. C. M. Bishop, *Neural Networks for Pattern Recognition*, (Clarendon Press, Oxford, 1995).
2. M. Brown, K.M. Bossley, D.J. Mills and C.J. Harris. High dimensional neurofuzzy systems: overcoming the curse of dimensionality, *Proc. FUZZ/IEEE'95*, Yokohama, Japan, 1995, pp. 2139-2146.
3. 5 R.L. Coultrip and R. H. Granger, "Sparse random networks with LTP learning rules approximate Bayes classifiers via Parzen's method", *Neural Networks* **7** (1994) 463-476.
4. B.V. Dasarathy, *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques,* (IEEE Computer Society Press, Los Alamitos, Calofornia, 1990).
5. K. Fukunaga, *Introduction to Statistical Pattern Recognition* (Academic, New York, 1972).

6. J.-S. Roger Jang and C.-T. Sun, "Functional equivalence between radial basis function networks and fuzzy inference systems", *IEEE Transactions on Neural Networks* **1** (1993) 156-159.
7. A Krzyzak, T. Linder and G. Lugosi, "Nonparametric classification using radial basis function nets and empirical risk minimization", *Proc. 12th Int. Conf. on Pattern recognition*, Jerusalem, Israel, 1994, pp. 72-76.
8. B. Kosko, "Combining fuzzy systems", *Proc. FUZZ/IEEE'95*, Yokohama, Japan, 1995, pp. 1855-1863.
9. M. Pawlak and M.F. Yat Fung Ng, "On kernel and radial basis function techniques for classification and function recovering", *Proc. 12th International Conference on Pattern Recognition*, Jerusalem, Israel, 1994, pp. 454-456.
10. W. Pedrycz, "Fuzzy sets in pattern recognition. Methodology and methods", *Pattern Recognition* **23** (1990) 121-146.