Fuzzy vs Statistical Classifiers

Bayesian Classifier, Singleton Fuzzy Model

Olaf Wolkenhauer

Control Systems Centre

UMIST

o.wolkenhauer@umist.ac.uk

www.csc.umist.ac.uk/people/wolkenhauer.htm

Contents

1. Learning Objectives

- Fuzzy clustering groups unlabelled data into a fixed number of classes and hence can be used to design classifiers.
- Specific fuzzy classifiers can be shown to be formally equivalent to optimal statistical classifiers.
- If-then rule-based fuzzy classifiers provide an intuitive framework to interpret data.

2. Multivariate Analysis

▷ Separation:

- Discriminant Analysis (exploratory...discriminants)
- Classification (rules...classifiers)

▷ Grouping:

- Clustering

3. Classification, Discrimination

Assumptions:

- ▷ The data, $\mathbf{m}_j \in \mathbb{R}^r$, are assumed to comprise c clusters.
- ▷ The number of clusters c is assumed to be known.
- ▷ A training sample of data is available from each cluster.

✘ Formulate rules for assigning new unclassified (unlabelled) observations to one of the clusters.

▷ Assign to an object (described as a point $\mathbf{x}$ in the feature space $X_1 \times \cdots \times X_r$) a class label C from the set $\mathcal{C} = \{C_1, \ldots, C_c\}$.

▷ Assume that $X_1 \times \cdots \times X_r$ coincides with $\mathbb{R}^r$ and that have available a set of (labelled) training data $\mathbf{M} = \{\mathbf{m}_1, \ldots, \mathbf{m}_d\}$, $\mathbf{m}_j = [m_{1j}, \ldots, m_{rj}]^T \in \mathbb{R}^r$.

▷ Denote by $b_i \in \{1, 2, \ldots, r\}$ the index of the class label among $\{C_1, \ldots, C_c\}$, associated with $\mathbf{m}_j$.

✘ The problem is to design a classifier, i.e to specify a mapping ψ such that each object $\mathbf{x}$ is associated with one class C_i :

$$\psi \, : \, \mathbb{R}^r \to \mathcal{C} \, .$$

4. Probabilistic Classifier

▷ Let $\mathbf{x}$ and C are random variables.

▷ Let $Pr(C_i)$ be the prior probability for class C_i, $i = 1, \ldots, c$.

▷ Denote by $p(\mathbf{x}|C_i)$ the class-conditional probability density function.

✘ Bayesian decision theory: design optimal classifier with a small error, that is, assign to $\mathbf{x}$ a class label C^* corresponding to the highest posterior probability, i.e

$$C^* = \arg\max_{C} \; Pr(C|\mathbf{x}) \; .$$

Where the posterior probability is calculated by

$$Pr(C_i|\mathbf{x}) = \frac{Pr(C_i)\; p(\mathbf{x}|C_i)}{p(\mathbf{x})} \; , \tag{1}$$

$$p(\mathbf{x}) = \sum_{k} Pr(C_k)\; p(\mathbf{x}|C_k) \; .$$

4.1. Kernel Density Estimation

- ▷ Parzen's kernel estimator: nonparametric approximation of a probability density function.
- ▷ Let $K(\mathbf{x})$ be a *kernel function* (also referred to as a *Parzen window*) which peaks at zero, is nonnegative, and whose integral equals one over $\mathbb{R}^r$.
- ▷ The multidimensional kernel function centered around $\mathbf{m}_j \in \mathbb{R}^r$ can be expressed in the form

$$\frac{1}{h^r} K\left(\frac{\mathbf{x} - \mathbf{m}_j}{h}\right)$$

where h determines the window with and hence is a *smoothing parameter*.

▷ We can approximate the class-conditional probability density using the sample set $\mathbf{M}$ by

$$\hat{p}(\mathbf{x}|C_i) = \frac{1}{d_{C_i} h^r} \sum_{j:\ b_j = i} K\left(\frac{\mathbf{x} - \mathbf{m}_j}{h}\right) , \qquad \mathbf{m}_j \in \mathbf{M} ,$$

where d_{C_i} is the number of elements of $\mathbf{M}$ from class C_i.

▷ Finally, we estimate the prior probabilities in (1) by

$$\widehat{Pr}(C_i) = \frac{d_{C_i}}{d} .$$

▷ Inserting both approximations into (1), we obtain the following estimate of the posterior probability :

$$\widehat{Pr}(C_i|\mathbf{x}) = \frac{1}{d \cdot h^r \cdot p(\mathbf{x})} \cdot \sum_{j:\ b_j = i} K\left(\frac{\mathbf{x} - \mathbf{m}_j}{h}\right) . \tag{2}$$

- ▷ Introducing an indicator function $\zeta_{C_i}(\mathbf{m}_j)$,

$$\zeta_{C_i}(\mathbf{m}_j) = \begin{cases} 1\ , & \text{if } b_j = i, \text{ i.e., } \mathbf{m}_j \text{ comes from class } C_i; \\ 0\ , & \text{otherwise.} \end{cases}$$

- ▷ We can rewrite (2) as

$$\widehat{Pr}(C_i|\mathbf{x}) = \frac{1}{d} \cdot a_1(\mathbf{x}) \cdot \sum_{j=1}^{d} \zeta_{C_i}(\mathbf{m}_j) K\left(\frac{\mathbf{x} - \mathbf{m}_j}{h}\right)\ , \qquad (3)$$

where factor $a_1(\mathbf{x})$ depends on $\mathbf{x}$ but not on the class label.

▷ Using the multidimensional Gaussian kernel

$$\frac{1}{h^r} K_G\left(\frac{\mathbf{x}-\mathbf{m}_j}{h}\right) = \frac{1}{h^r\sqrt{(2\pi)^r}\sqrt{|\mathbf{\Sigma}|}}\exp\left(-\frac{1}{2h^2}(\mathbf{x}-\mathbf{m}_j)^T\mathbf{A}^{-1}(\mathbf{x}-\mathbf{m}_j)\right) \ , \tag{4}$$

where $\mathbf{\Sigma}$ is the covariance matrix.

▷ Using the Gaussian kernel we have for the posterior probabilities (3)

$$\widehat{Pr}(C_i|\mathbf{x}) = \frac{1}{d}\cdot a_1(\mathbf{x})\cdot\sum_{j=1}^{d}\zeta_{C_i}(\mathbf{m}_j)K_G\left(\frac{\mathbf{x}-\mathbf{m}_j}{h}\right) \ . \tag{5}$$

5. Fuzzy Classifier

- Product inference engine,
- Singleton input data,
- Centre average defuzzifier : ...nonlinear mapping

$$\begin{aligned} f \,:\, X &\rightarrow Y \\ \mathbf{x} &\mapsto f(\mathbf{x}) \end{aligned}$$

where

$$f(\mathbf{x}) = \frac{\sum_{i=1}^{n_R} y_0^{(i)} \cdot \prod_{k=1}^{r} \mu_{A_{ik}}(x_k)}{\sum_{i=1}^{n_R} \prod_{k=1}^{r} \mu_{A_{ik}}(x_k)} \,. \tag{6}$$

5.1. Equivalence of Fuzzy and Statistical Classifiers

▷ From (6), for any class C_i,

R_j : IF x_1 is A_{j1} AND ... AND x_r is A_{jr},
THEN $y^j_{i'} = 1$ and $y^j_i = 0$, $\forall i \neq i'$, $i = 1, \ldots, c$, $j = 1, \ldots, d$.

where y^j_i denotes the i^{th} component of the output vector $\mathbf{y}_j$, associated with the j^{th} rule.

▷ Each A_{jk} is a fuzzy set with membership function

$$\mu_{A_{jk}} : \mathbb{R} \to [0, 1] \; .$$

▷ We define

$$\mu_{A_{jk}}(\mathbf{x}) = \exp\left(-\frac{(x_k - m_{kj})^2}{2h^2}\right)$$

where h is a parameter and the membership functions evaluate the similarity of any given $\mathbf{x}$ with $\mathbf{m}_j$.

▷ Let the activation strength ('firing level') of the j-th rule is

$$\begin{aligned}\beta_j(\mathbf{x}) &= \prod_{k=1}^{r} \mu_{A_{jk}}(x_k) \\ &= \exp\left(-\frac{1}{2h^2}\sum_{k=1}^{r}(x_k - m_{kj})^2\right) \\ &= \exp\left((\mathbf{x}-\mathbf{m}_j)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{m}_j)\right) \ .\end{aligned}$$

Where $\mathbf{A}$ is an identity matrix.

▷ We notice that $\beta_j(\mathbf{x})$ differs from the Gaussian kernel (4) only by a constant. We therefore write

$$\beta_j(\mathbf{x}) = a_2 \cdot K\left(\frac{\mathbf{x}-\mathbf{m}_j}{h}\right) \ .$$

▷ The output of the fuzzy classifier, with respect to class C_i, is obtained as

$$y^i = \frac{\sum\limits_{j=1}^{d} y_i^j \cdot \beta_j(\mathbf{x})}{\sum\limits_{j=1}^{d} \beta_j(\mathbf{x})} \qquad \text{.. equivalent to (6)!}$$

$$= a_3(\mathbf{x}) \cdot \sum_{j=1}^{d} y_i^j \cdot K_G\left(\frac{\mathbf{x}-\mathbf{m}_j}{h}\right) . \tag{7}$$

5.2. Conclusions

- ▷ Since y_i^j functions as an indicator function for $\mathbf{m}_j$ with respect to C_i, we find that equations (7) and the posterior probability of the statistical classifier (5) differ only by a factor which does not depend on the class i.
- ▷ In both cases, for the fuzzy classifier and the statistical classifier a decision is obtained by choosing the class label for which (7) and (5) is largest.
- ▷ We conclude that a fuzzy system can be shown to be equivalent to a probabilistic classifier (which is known to be asymptotically optimal in the Bayesian sense).

References

[1] Bishop, C.M. : *Neural Networks for Pattern Recognition.* Clarendon Press, 1996.

[2] Bezdek, J.C. : *Pattern Recognition with Fuzzy Objective Function Algorithms.* Plenum Press, 1981.

[3] Bezdek, J.C and Pal, S.K. (eds.) : *Fuzzy Models for Pattern Recognition.* IEEE Press, 1992.

[4] Fukunaga, K. : *Introduction to Statistical Pattern Recognition.* 2nd. ed., Academic Press, 1990.

[5] Höppner, F. et.al. : *Fuzzy Cluster Analysis.* Wiley, 1999.

[6] Johnson, R.A. and Wichern, D.W. : *Applied Multivariate Statistical Analysis.* 4th ed., Prentice Hall, 1998.

[7] Kuncheva, L. : *On the equivalence between fuzzy and statistical*

classifiers. Int. J. of Uncertainty, Fuzziness and Knowledge-Based Systems. Vol. 4, No. 3 (1996), pp. 245–253.

[8] Pal, S.K. and Mitra, S. : *Neuro-Fuzzy Pattern Recognition.* Wiley, 1999.

[9] Ripley, B.D. : *Pattern Recognition and Neural Networks.* Cambridge University Press, 1996.