Statistical Analyses of Various Error Functions for Pattern Classifiers

Sang-Hoon Oh

Mokwon University, Doan-dong, Seo-gu, Daejon, Korea
shoh@mokwon.ac.kr

Abstract. There are various error functions for pattern classifiers. This paper analyzes the error functions such as MSE(mean-squared error), CE(cross-entropy) error, AN(additive noise) in MSE, MLS(mean log square) error, and nCE(*n*th order extension of CE) error functions in a statistical perspective. Also, the analyses include CFM(classification figure of merit). The results of analyses provide considerable insights into the properties of different error functions.

Keywords: Classifier, error functions, statistical analysis, optimal solution.

1 Introduction

Pattern classifiers can be implemented using a discriminant function. The functional value corresponds to the degree of confidence that an input pattern belongs to a certain class and the decision of classification is done by selecting the class of maximal discriminant value [1]. Alternatively, the classifiers can be implemented based on the posteriori probabilities and this provides the Bayes classifier [2]. However, it is difficult to estimate the p.d.f.(probability density function) or the probability distribution of samples. The Parzen's window can estimate the p.d.f. of samples by locating the window function at each sample [3]. Still this method requires enough number of samples for the accurate estimation of the p.d.f.

In many cases, the discriminant function approach attains better performance than the posteriori probability approach and this supports the popularity of discriminant functions. Conventionally, MSE(mean-squared error) function is used to train the classifier whose outputs become discriminant values [4]. As a variant of MSE, Wang and Principe proposed the additive noise method in the desired signal of output [5]. In order to deal with outliers of samples, Liano proposed MLS(mean-log square) error function [6]. CE(cross-entropy) error is another error function for performance improvement [7] and nCE(*n*th order extension of CE) error is a more advanced formulation of CE [8][9]. All these error functions are tried to be minimized when training classifiers. CFM(classification figure of merit) function is another approach to be maximized during training of classifiers [10].

In this paper, the various error functions are analyzed in a statistical way in order to provide insights into the properties of them. Section 2 describes the mathematical analyses and the comparisons among them. Finally, Section 3 concludes this paper.

G. Lee, D. Howard, and D. Ślęzak (Eds.): ICHIT 2011, CCIS 206, pp. 129–133, 2011.

2 Statistical Analyses of Various Error Functions

Let $\mathbf{x}=[x_1,x_2,\cdots,x_N]^T$ be an input sample and $\mathbf{y}=[y_1,y_2,\cdots,y_M]^T$ be an output vector of a classifier, of which desired vector $\mathbf{t}=[t_1,t_2,\cdots,t_M]^T$ is coded as follows:

$$t_k=\begin{cases}+1, & \text{if } \mathbf{x} \text{ originates from class } k\\ -1, & \text{otherwise.}\end{cases} \tag{1}$$

Classifiers are trained to minimize the distance between **t** and **y**. MSE [4] defined by

$$E_{MSE}(\mathbf{x})=\frac{1}{2}\sum_{k=1}^{M}(t_k-y_k(\mathbf{x}))^2 \tag{2}$$

is usually used as a distance measure. In the limit $P\to\infty$, the minimizer of E_{MSE} for all P training patterns converges (under certain regularity conditions, Theorem 1 in [11]) towards the minimizer of the function

$$E\{E_{MSE}(\mathbf{X})\}=E\left\{\frac{1}{2}\sum_{k=1}^{M}(T_k-y_k(\mathbf{X}))^2\right\}, \tag{3}$$

where $E\{\cdot\}$ is the expectation operator, T_k is the random variable denoting the desired value, and $\mathbf{X}$ is the random vector denoting the input sample. Since targets are coded as in (1), the square term in (3) can be written as

$$E\{(T_k-y_k(\mathbf{X}))^2\}=\int\left[(1-y_k(\mathbf{x}))^2Q_k(\mathbf{x})+(-1-y_k(\mathbf{x}))^2(1-Q_k(\mathbf{x}))\right]f(\mathbf{x})d\mathbf{x} \tag{4}$$

where $Q_k(\mathbf{x})=\Pr[\mathbf{X}\text{ originates from class }k\mid\mathbf{X}=\mathbf{x}]$ is the posterior probability. Let us seek the function $\mathbf{b}=[b_1,b_2,\cdots,b_M]^T$ minimizing the criterion (3) (in the space of all functions taking values in (-1,+1)). For fixed $Q_k(\mathbf{x})$, $0<Q_k(\mathbf{x})<1$, the optimal solution is given by $\mathbf{b}(\mathbf{X})$, where the components of **b** are given by [8][11]

$$b_k(\mathbf{x})=E\{T_k\mid\mathbf{x}\}=2Q_k(\mathbf{x})-1,\ k=1,2,...,M. \tag{5}$$

For performance improvement of MSE, additive noise in the desired signal was adopted as

$$E_{AN}(\mathbf{x})=\frac{1}{2}\sum_{k=1}^{M}(t_k+n_k-y_k(\mathbf{x}))^2 \tag{6}$$

where n_k is the zero-mean white noise with variance σ^2 [5]. Then, since the random noise is independent of the desired and real output values,

$$\begin{aligned}E\{E_{AN}(\mathbf{X})\}&=\frac{1}{2}\sum_{k=1}^{M}\int\left[\int(T_k+N_k-y_k(\mathbf{x}))^2f(N_k)dN_k\right]f(\mathbf{x})d\mathbf{x}\\&=\frac{1}{2}\sum_{k=1}^{M}\left[\int(T_k-y_k(\mathbf{x}))^2f(\mathbf{x})d\mathbf{x}+\sigma^2\right].\end{aligned} \tag{7}$$

Here, N_k is the random variable denoting the noise. By applying (4), the minimizer of (7) has the optimal solution vector whose components are the same with (5).

CE [7] error is defined by

$$E_{CE}(\mathbf{x}) = -\sum_{k=1}^{M} [(1+t_k)\ln(1+y_k(\mathbf{x})) + (1-t_k)\ln(1-y_k(\mathbf{x}))]. \tag{8}$$

When we derive the optimal solution for minimizing

$$E\{E_{CE}(\mathbf{X})\} = -\sum_{k=1}^{M} \int [2Q_k(\mathbf{x})\ln(1+y_k(\mathbf{x})) + 2(1-Q_k(\mathbf{x}))\ln(1-y_k(\mathbf{x}))] f(\mathbf{x}) d\mathbf{x}, \tag{9}$$

the result is the same with the MSE case.

As an extension of CE, nCE error [8] is defined by

$$E_{nCE}(\mathbf{x}) = -\sum_{k=1}^{M} \int \frac{t_k^{n+1}(t_k - y_k(\mathbf{x}))^n}{2^{n-2}(1-y(\mathbf{x})_k^2)} dy_k \ . \tag{10}$$

The optimal solution for minimizing $E\{E_{nCE}(\mathbf{X})\}$ is derived and the kth component is given by [8]

$$b_k(\mathbf{x}) = g(h_n(Q_k(\mathbf{x}))), \tag{11}$$

where

$$h_n(q) = \left(\frac{1-q}{q}\right)^{\frac{1}{n}} \text{ and } g(u) = \frac{1-u}{1+u}. \tag{12}$$

Also, $g \circ h_n$ is strictly increasing.

In order to suppress the huge amount of weight updating by outliers, MLS error function [6] was proposed by

$$E_{MLS}(\mathbf{x}) = \sum_{k=1}^{M} \log\left(1 + \frac{1}{2}(t_k - y_k(\mathbf{x}))^2\right). \tag{13}$$

Then,

$$\begin{aligned} E\{E_{MLS}(\mathbf{X})\} = \sum_{k=1}^{M} \int [Q_k(\mathbf{x}) \log\left(1 + \frac{1}{2}(1-y_k(\mathbf{x}))^2\right) \\ + (1-Q_k(\mathbf{x}))\log\left(1 + \frac{1}{2}(1+y_k(\mathbf{x}))^2\right)] f(\mathbf{x}) d\mathbf{x}. \end{aligned} \tag{14}$$

Using the same procedure, the kth component of optimal solution vector for minimizing $E\{E_{MLS}(\mathbf{X})\}$ can be derived by

$$b_k(\mathbf{x}) = h^{-1}(Q_k(\mathbf{x})) \tag{15}$$

where

$$h(y) = \frac{y^3 - y^2 + y + 3}{-2y^2 + 6}. \tag{16}$$

It is easy to show that $h^{-1}(q)$ is a strictly increasing function of $q \in (0,1)$.

Contrary to the error functions which are minimized during training of classifiers, CFM is a criterion function to be maximized. CFM is defined by

$$CFM(\mathbf{x}) = \sum_{k \neq c} \frac{1}{1 + \exp(-\beta(y_c(\mathbf{x}) - y_k(\mathbf{x})))} \tag{17}$$

where y_c denotes the correct node and y_k denotes the incorrect node [10]. Therefore, the expectation is

$$E\{CFM(\mathbf{X})\} = \sum_{k=1}^{M} E\left\{ \int \frac{1 - Q_k(\mathbf{x})}{1 + \exp(-\beta(y_c(\mathbf{x}) - y_k(\mathbf{x})))} f(\mathbf{x}) d\mathbf{x} \right\}. \tag{18}$$

Since the fraction term in the integral is a monotonic increasing function of $y_c(\mathbf{x}) - y_k(\mathbf{x})$, $E\{CFM(\mathbf{X})\}$ is maximized when $y_c(\mathbf{x}) - y_k(\mathbf{x})$ is maximized. For a specific $\mathbf{x}$ in the class c, y_c is trained to be 1 and y_k $(k \neq c)$ is trained to -1. That is,

$$E\{y_k(\mathbf{x})\} = (+1) \times Q_k(\mathbf{x}) + (-1) \times (1 - Q_k(\mathbf{x})). \tag{19}$$

As a result, the *k*th component of optimal solution is given by

$$b_k(\mathbf{x}) = 2Q_k(\mathbf{x}) - 1. \tag{20}$$

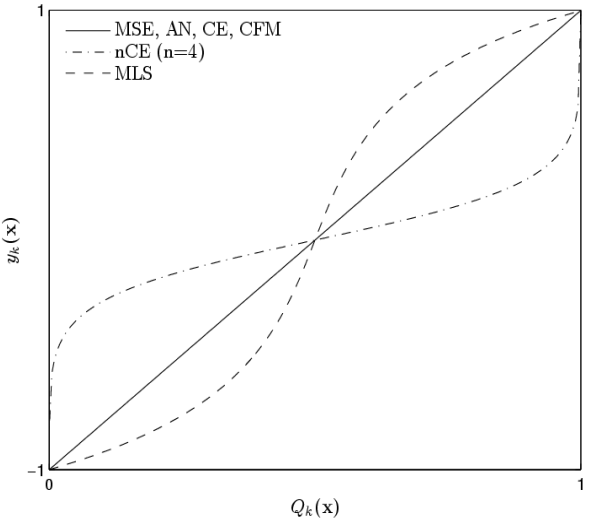

Fig. 1. The optimal solutions of $y_k(\mathbf{x})$ for minimizing the expectations of MSE, AN, CE, nCE, and MLS error functions and for maximizing the expectation of CFM. $y_k(\mathbf{x})$ denotes the *k*th output of classifier and $Q_k(\mathbf{x})$ denotes the posterior probability $\Pr[\mathbf{X}$ originates from class $k \mid \mathbf{X} = \mathbf{x}]$ when a random vector $\mathbf{X}$ is presented to the classifier.

Fig. 1 shows the optimal solutions of the various error or criterion functions. In the MSE, AN, CE, and CFM cases, the optimal solution is proportional to $Q_k(\mathbf{x})$. The optimal solution of nCE has a rapid slope when $Q_k(\mathbf{x})$ is near to 0 or 1. On the contrary, MLS has an optimal solution which is gentle when $Q_k(\mathbf{x})$ is near to 0 or 1. Although the curves of optimal solutions are different in some cases, all the error functions have optimal solutions which are strictly increasing functions of $Q_k(\mathbf{x})$. Therefore, the Bayes classifier can be defined with the decision rule "decide *k,* if k = max $y_k(\mathbf{x})$".

3 Conclusions

This paper analyzed various error or criterion functions for classifiers in a statistical perspective. MSE, AN, CE, and CFM have the same optimal solution of classifier output, which is proportional to the posterior class probability. nCE and MLS have some rapid or gentle slopes when the posterior class probability is near to 0 or 1. For all error or criterion functions, the Bayes classifier can be defined with the "max" rule.

References

1. Park, W.J., Kil, R.M.: Pattern Classification with Class Probability Output Network. IEEE Trans. Neural Network 20, 1659–1673 (2009)
2. Fukunaga, K., Kessel, D.: Nonparametric Bayes error estimation using unclassified samples. IEEE Trans. Inf. Theory 19, 434–439 (1973)
3. Parzen, E.: On the estimation of a probability density function and mode. Ann. Math. Statist. 33, 1065–1076 (1962)
4. Rumelhart, D.E., McClelland, J.L.: Parallel Distributed Processing. MIT Press, Cambridge (1986)
5. Wang, C., Principe, J.C.: Training Neural Networks with Additive Noise in the Desired Signal. IEEE Trans. Neural Networks 10, 1511–1517 (1999)
6. Liano, K.: Robust Error Measure for Supervised Neural Network Learning with Outliers. IEEE Trans. Neural Networks 7, 246–250 (1996)
7. van Ooyen, A., Nienhuis, B.: Improving the Convergence of the Backpropagation Algorithm. Neural Networks 4, 465–471 (1992)
8. Oh, S.-H.: Improving the error back-propagation algorithm with a modified error function. IEEE Trans. Neural Networks 8, 799–803 (1997)
9. Oh, S.-H.: Error Back-Propagation Algorithm for Classification of Imbalanced Data. Neurocomputing 74, 1058–1061 (2011)
10. Hampshire II, J.B., Waibel, A.H.: A Novel Objective Function for Improved Phoneme Recognition Using Time-Delay Networks. IEEE Trans. Neural Networks 1, 216–218 (1990)
11. White, H.: Learning in Artificial Neural Networks: A Statistical Perspective. Neural computation 1, 425–464 (1989)