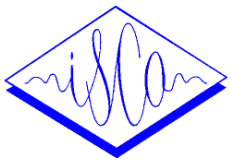

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

Performance Comparison of Recognition Systems Based on the Akaike Information Criterion

Kazuhiko Ozeki
The University of Electro-Communications
1-5-1 Chofugaoka, Chofu-shi, Tokyo, 182 Japan

Abstract

It is widely believed that we should apply a common set of test samples to all the recognition systems under test in order to make a reliable performance comparison. But how much is this true? We discuss this problem based on the Akaike Information Criterion (AIC). It becomes clear that by applying a common set of test samples, more discrimination power can be obtained, as has been believed, as to performance difference than by applying independent sets of samples to each system. The difference between them is, however, not so large as might be expected. The effect of applying a common set of test samples to two systems under test becomes prominent when we measure and utilize the number of samples recognized correctly by both systems in addition to the number of samples recognized correctly by each system.

1. Introduction

In comparing the performance of recognition systems, measuring their sample recognition rates is a common practice. But how many test samples do we need to make a statistically reliable comparison? Also it is widely believed that we should apply a common set of test samples to all the systems under test in order to make a fair comparison. But how much is this true? Furthermore when we apply a common set of test samples to two systems to be compared, we can measure not only the number of samples which are recognized correctly by each system, but also the number of samples which are recognized correctly by both systems. Should we not utilize this information to judge the superiority of one system to the other?

There are statistical theories relevant to these questions such as the theory of *statistical test* and the theory of *confidence interval* [1]. None of the conventional theories, however, is conveniently applicable to these problems. In this paper we try to answer the above questions using the Akaike Information Criterion (AIC), which is a measure of fitness between a statistical model and an observation [2]. When there are two recognition systems to be compared, we set up two statistical models for recognition experiment; in one model it is hypothesized that the two systems have the same recognition rates, and in the other they have different recognition rates. By comparing the AICs for the two models, we can judge whether there exists a significant difference in performance between the two systems. In calculating the AIC, we need to maximize a likelihood function that is given by a sum of multinomial distribution functions. A new mathematical technique is developed for searching the maximum by reducing the number of free parameters in the likelihood function, which makes practical computation feasible.

We exemplify, for various cases, the amount of difference in the sample recognition rate required to assure the significant difference in performance. Through the numerical results, we show the importance of measuring the number of samples that are recognized correctly by both systems to be compared.

2. Statistical Models for Measuring Recognition Performance

2.1 The Akaike Information Criterion (AIC)

Let $f(x_1, x_2, \ldots, x_n \mid \theta)$ $(\theta \in \Theta)$ be a family of probability (density) functions, i.e. a statistical model. The Akaike information criterion AIC of this model for observed data $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n$ is defined as

$$aic = -2l + 2p,$$

where l is the maximum log likelihood

$$l = \max_{\theta \in \Theta} [\log f(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n \mid \theta)],$$

and p is the number of free parameters in θ, i.e. the dimension of the parameter space Θ. The smaller aic means the better fit of the model to the observed data. Therefore if there are two competing statistical models A and B with aic_A and aic_B, respectively, we judge that A is better than B for the observed data if $aic_A < aic_B$.

2.2 Statistical Model for Recognition Experiment

Let X be a pattern space, that is, the set of all the possible objects to be recognized by a recognition system. Each pattern x in X is a pair $x = (y, c)$, where y is the feature of x, and c is its category. We denote the set of all the features by X, and the set of all the categories by C. The category set C is normally a finite set.

A recognition system S is a mapping

$$S: X \to C.$$

The set X_S of patterns recognized correctly by S is given by

$$X_S := \{x \in X \mid x = (y,\ c),\ S(y) = c\}.$$

The recognition rate p_S of S is given by $p_S := \mu(X_S)$, where μ is a probability measure on X. If we randomly choose n samples from X and feed them to the system S, the probability that m samples are recognized correctly is given by

$$B(n, m, ; p_S) = \binom{n}{m} p_S^m (1 - p_S)^{n-m}.$$

2.3 Test on Independent Sample Sets

Suppose we have two recognition systems S and T with recognition rates p_S and p_T, respectively. We choose n samples randomly from X and feed them to S, and choose another n samples randomly and feed them to T. If m_S samples are recognized correctly by S, and m_T samples are recognized correctly by T, then the likelihood of this event is given by $B(n, m_S; p_S)B(n, m_T; p_T)$. We want to know if the two systems have the same recognition rate or not. For that purpose we consider the following two models [3]:

(1) Model with $p_S = p_T$

The AIC for this model is given by

$$\begin{aligned} & aic(1)_e \\ &= -2 \max_{p_S = p_T} [\log B(n, m_S; p_S) B(n, m_T; p_T)] + 2 \times 1 \\ &= -2 \left[(m_S + m_T) \log \frac{m_S + m_T}{2n} \right. \\ &\quad \left. + (k_S + k_T) \log \frac{k_S + k_T}{2n} \right] + 2, \\ &\quad k_S = n - m_S, k_T = n - m_T. \end{aligned}$$

(2) Model with No Constraint

The AIC for this model is given by

$$\begin{aligned} & aic(1)_d \\ &= -2 \max[\log B(n, m_S; p_S) B(n, m_T; p_T)] + 2 \times 2 \\ &= -2 \left[m_S \log \frac{m_S}{n} + k_S \log \frac{k_S}{n} \right. \\ &\quad \left. + m_T \log \frac{m_T}{n} + k_T \log \frac{k_T}{n} \right] + 4. \end{aligned}$$

Note that the number of free parameters is different for each model. In the final expressions for $aic(1)_e$ and $aic(1)_d$, a common term is omitted because only the difference between them is meaningful.

2.4 Test on Common Sample Set

In this case, we choose n samples randomly from X, and feed them to both of S and T. The pattern space X is partitioned into four subspaces: $X_1 := X_S - X_T$, $X_2 := X_T - X_S$, $X_3 := X_S \cap X_T$, and $X_4 := X - (X_S \cup X_T)$, where X_S (X_T) is the set of all the patterns recognized correctly by S (T). Let $p_i := \mu(X_i)$, $(i = 1, \ldots, 4)$, and n_i be the number of samples which fall into X_i. For example, n_1 is the number of samples recognized correctly by S but misrecognized by T. Note that $\sum_i p_i = 1$, $p_i \geq 0$, and $\sum_i n_i = n$. The likelihood of this event is given by

$$\begin{aligned} & P(n_1, n_2, n_3, n_4; p_1, p_2, p_3, p_4) \\ &= \frac{n!}{n_1! n_2! n_3! n_4!} p_1^{n_1} p_2^{n_2} p_3^{n_3} p_4^{n_4}. \end{aligned}$$

We call the above n_1, n_2, n_3, n_4 the *complete information.*

2.4.1 Complete Information Case

We consider here the case where we have the complete information.

(1) Model with $p_S = p_T$

Since $p_S = p_1 + p_3$, and $p_T = p_2 + p_3$, the constraint $p_S = p_T$ is equivalent with $p_1 = p_2$. By this fact, AIC for this model is given by

$$\begin{aligned} aic(2)_e &= -2 \max_{p_1 = p_2} [\log P(n_1, n_2, n_3, n_4; p_1, p_2, p_3, p_4)] \\ &\quad + 2 \times 2 \\ &= -2 \left[(n_1 + n_2) \log \frac{n_1 + n_2}{2n} \right. \\ &\quad \left. + n_3 \log \frac{n_3}{n} + n_4 \log \frac{n_4}{n} \right] + 4. \end{aligned}$$

(2) Model with No Constraint

For this model, AIC is given by

$$\begin{aligned} aic(2)_d &= -2 \max[\log P(n_1, n_2, n_3, n_4; p_1, p_2, p_3, p_4)] \\ &\quad + 2 \times 3 \\ &= -2 \left[n_1 \log \frac{n_1}{n} + n_2 \log \frac{n_2}{n} \right. \\ &\quad \left. + n_3 \log \frac{n_3}{n} + n_4 \log \frac{n_4}{n} \right] + 6. \end{aligned}$$

A common term in $aic(2)_e$ and $aic(2)_d$ is omitted here.

2.4.2 Incomplete Information Case

Customarily we do not measure the above n_1, n_2, n_3, n_4 in a recognition experiment. Instead we only measure the numbers m_S and m_T of samples recognized correctly by the system S and T, respectively. This is to be called *incomplete information*, because there is uncertainty as to n_1, n_2, n_3, n_4. Therefore the likelihood of the event of observing m_S and m_T is the sum of probabilities over all the possible combinations of n_1, n_2, n_3, n_4:

$$\begin{aligned} & Q(n, m_S, m_T; p_1, p_2, p_3, p_4) \\ &= \sum_{n_l \leq n_3 \leq n_u} P(n_1, n_2, n_3, n_4; p_1, p_2, p_3, p_4), \\ &\quad n_1 = m_S - n_3, n_2 = m_T - n_3, \\ &\quad n_4 = n - (m_S + m_T) + n_3, \\ &\quad n_l = \max\{0, m_S + m_T - n\}, \\ &\quad n_u = \min\{m_S, m_T\}. \end{aligned}$$

(1) Model with $p_S = p_T$

Using the likekihood above, we can give AIC for this case as

$$\begin{aligned} aic(3)_e &= -2 \max_{p_1 = p_2} [\log Q(n, m_S, m_T; p_1, p_2, p_3, p_4)] \\ &\quad + 2 \times 2. \end{aligned}$$

No analytic solution for this maximimization problem is known; we have to resort to numerical maximization. Using the principle of EM algorithm [4], we can derive a re-estimation formula for $p_1(=p_2)$, p_3, p_4:

$$\bar{p}_1 = \sum_{n_l \le n_3 \le n_u} V(n_3) \frac{m_S + m_T - 2n_3}{2n},$$

$$\bar{p}_3 = \sum_{n_l \le n_3 \le n_u} V(n_3) \frac{n_3}{n},$$

$$\bar{p}_4 = \sum_{n_l \le n_3 \le n_u} V(n_3) \frac{n - (m_S + m_T) + n_3}{n},$$

where

$$V(n_3) = \frac{P(n_1, n_2, n_3, n_4; p_1, p_1, p_3, p_4)}{\sum_{n_l \le n_3 \le n_u} P(n_1, n_2, n_3, n_4; p_1, p_1, p_3, p_4)},$$
$$n_1 = m_S - n_3, n_2 = m_T - n_3,$$
$$n_4 = n - (m_S + m_T) + n_3.$$

The iterative algorithm based on this formula gives a local maximum. We can not, however, be content with this; we need the global maximum. Since the maximum point $(\hat{p}_1, \hat{p}_3, \hat{p}_4)$ is a fixed point of the re-estimation formula, it follows that

$$\hat{p}_1 + \hat{p}_3 = \sum_{n_l \le n_3 \le n_u} V(n_3) \frac{m_S + m_T}{2n} = \frac{m_S + m_T}{2n}$$

and

$$\hat{p}_1 + \hat{p}_4 = \sum_{n_l \le n_3 \le n_u} V(n_3) \frac{2n - (m_S + m_T)}{2n} = \frac{2n - (m_S + m_T)}{2n}.$$

Because of this constraint among the parameters for the maximum point, we only have to perform a one-dimensional search, which makes practical computation feasible.

(2) Model with No Constraint

In this case AIC is given by

$$aic(3)_d = -2 \max[\log Q(n, m_S, m_T; p_1, p_2, p_3, p_4)] + 2 \times 3.$$

Following a similar reasoning as in the preceeding case, we can show that the parameters for the maximum point $(\hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{p}_4)$ are constrained as

$$\hat{p}_1 - \hat{p}_2 = \frac{m_S - m_T}{n}, \quad \hat{p}_1 + \hat{p}_3 = \frac{m_S}{n},$$

$$\hat{p}_1 + \hat{p}_4 = \frac{n - m_T}{n}.$$

This constraint reduces the dimension of the search space just as in the preceeding case.

3. Performance Comparison in Various Cases

3.1 Independent Sample Sets vs. Common Sample Set

We adopt a criterion recommended in [3] that if $aic_e - aic_d$ is greater than 1, we judge $p_S \neq p_T$, and if $aic_e - aic_d$ is smaller than -1, we judge $p_S = p_T$.

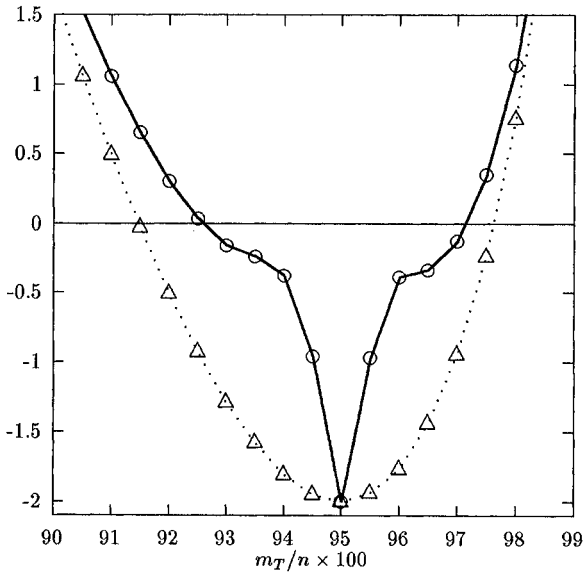

Fig.1. Comparison of Independent Sample Sets Case and Common Sample Set Case.
△: $aic(1)_e - aic(1)_d$, ○: $aic(2)_e - aic(2)_d$,
$n = 200$, $m_S/n \times 100 = 95.0$.

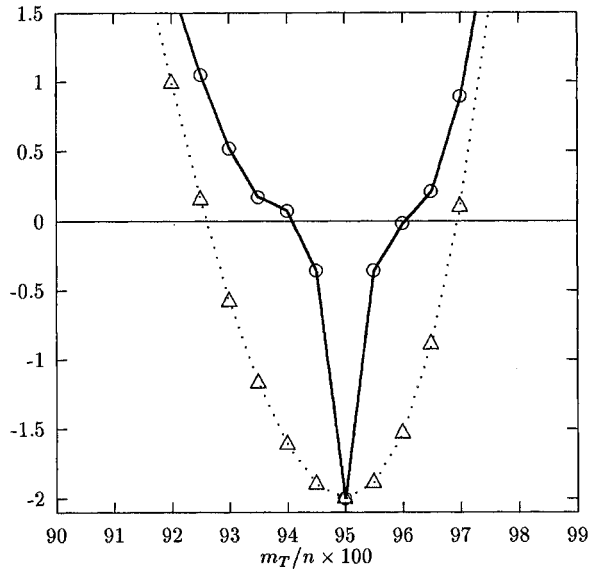

Fig.2. Comparison of Independent Sample Sets Case and Common Sample Set Case.
△: $aic(1)_e - aic(1)_d$, ○: $aic(2)_e - aic(2)_d$,
$n = 400$, $m_S/n \times 100 = 95.0$.

Fig.1 shows that, in the independent sample sets case, if $m_T/n \times 100$ is between 92.5 and 97.0, then $p_S = p_T$, and if $m_T/n \times 100$ is smaller than 90.5 or greater than 98.5, then $p_S \neq p_T$. In the common sample set case, on the other hand, if $m_T/n \times 100$ is between 94.5 and 95.5, then $p_S = p_T$, and if $m_T/T \times 100$ is smaller than 91.0 or greater than 98.0, then $p_S \neq p_T$.

Thus, as has been believed, the discrimination power is larger, that is, a smaller difference between m_S/n and m_T/n supports that $p_S \neq p_T$ in the common sample set case than in the independent sample sets case. The difference of the discrimination power is not, however, so great as might be expected when we want to conclude that $p_S \neq p_T$. Fig.2 shows similar curves as in Fig.1, with only difference of the value of n. From Fig.1 and Fig.2, we see that the discrimination power increases with n, which is a natural consequence. The overall tendency is, however, much the same in both cases.

3.2 Complete Information vs. Incomplete Information

In Fig.3 we see that, in the complete information case, the discrimination power greatly depends on n_3, which is the number of samples recognized correctly by both of S and T. The larger n_3 gives the better discrimination power. We note that $aic(2)_e - aic(2)_d$ in Fig.1 behaves like the best case in Fig.3 while the deviation of m_T/n from m_S/n is small. As the deviation increases, the behavior of $aic(2)_e - aic(2)_d$ in Fig.1 approaches the worst case in Fig.3. Thus, in the incomplete information case, we observe something between the best case and the worst case in the complete information case; we only get a blurred result.

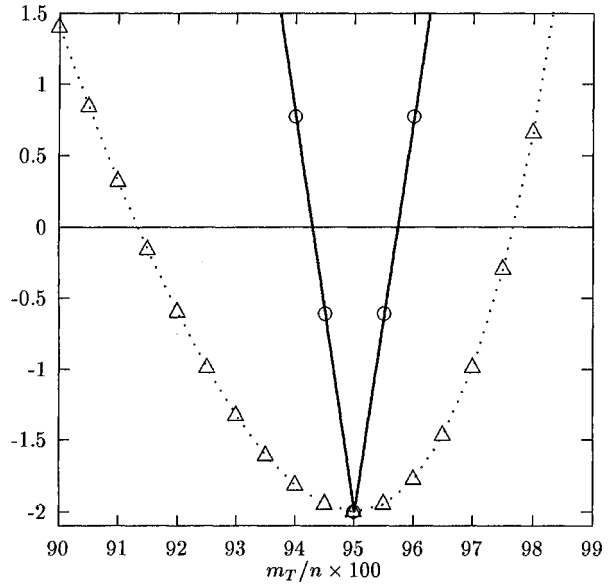

Fig.3. Effect of n_3 in Complete Information Case.
o: $aic(3)_e - aic(3)_d$ (n_3 = maximum, the best case),
△: $aic(3)_e - aic(3)_d$ (n_3 = minimum, the worst case),
$n = 200$, $m_S/n \times 100 = 95.0$.

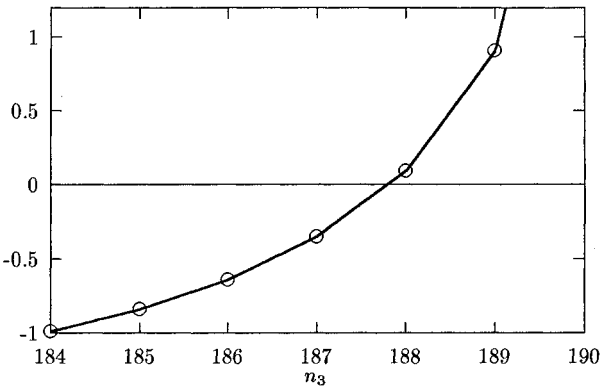

Fig.4. Effect of n_3 in Complete Information Case.
o: $aic(3)_e - aic(3)_d$,
$n = 200$, $m_S/n \times 100 = 95.0$, $m_T/n \times 100 = 97.0$.

Fig.4 clearly shows that we can obtain more detailed information as to the difference between p_S and p_T by measuring, not only m_S and m_T, but also n_3. In this case, we should judge $p_S = p_T$ if $n_3 = 184$, and $p_S \neq p_T$ if $n_3 = 189$. This demonstrates that the sample recognition rates alone are insufficient for the judgement of performance difference.

4. Conclusion

We have discussed methods for comparing performances of recognition systems base on the principle of AIC. It became clear that the discrimination power is greater when a common test sample set is applied than when independent sample sets are applied to two systems to be compared. This fact coincides with our belief. The effect of applying a common sample set is not, however, so large as might be expected as long as we only measure, as is usually practiced, the sample recognition rate of each system. In order to exploit the advantage of using a common sample set, we should additionally measure and utilize the number of samples which are recognized correctly by both systems. By doing so we can obtain, with very little extra cost, far more detailed information as to performance difference between the systems under test.

References

[1] S.S. Wilks: Mathematical Statistics, *John Wiley & Sons, Inc.*, 1962.

[2] H. Akaike: A new look at the statistical model identification, *IEEE Trans.* AC-19 (1974) pp.716-723

[3] : Y. Sakamoto, M. Ishiguro, and G. Kitagawa: Joho-ryo Tokeigaku, *Kyoritsu Shuppan*, 1983.

[4] L.E. Baum: An inequality and associated maximization technique in statistical estimation for probabilistic function of Markov process, *Inequalities*, 3(1972) pp.1-8.