Generalization Error of Multinomial Classifier

Sarunas Raudys

Vilnius Gediminas Technical University
Sauletekio 11, Vilnius LT-10223, Lithuania
raudys@ktl.mii.lt

Abstract. Equation for generalization error of Multinomial classifier is derived and tested. Particular attention is paid to imbalanced training sets. It is shown that artificial growth of training vectors of less probable class could be harmful. Use of predictive Bayes approach to estimate cell probabilities of the classifier reduces both the generalization error and effect of unequal training sample sizes.

Keywords: BKS rule, Complexity, Generalization error, Learning, Imbalanced training sets, Multinomial classifier, Sample size.

1 Introduction

In many pattern recognition tasks, the features are discrete: each measurement can assume one of several possible values, such as, the type of an engine, a sex, profession, a presence of certain disease, etc. Then, asymptotically (as the sample size is increasing) optimal classification rule is the Multinomial one [1, 2]. Theoretical analysis of this method is very important for practice since popular *decision tree classifiers* are in fact pruned and tailored versions of the Multinomial classifier. In multiple classifiers system design, we are faced with Multinomial classifier if local (expert) classifiers produce crisp outputs (class labels) and one uses Behavior knowledge space (BKS) method to fuse expert decisions.

Hughes [3] investigated the two class pattern recognition task with categorical valued features. He converted the analysis to investigation of single categorical feature that can assume one of m possible values characterized by m cell probabilities,

$$P_1^{(i)}, P_2^{(i)}, \dots, P_{m-1}^{(i)}, P_m^{(i)} \left(\sum_{s=1}^m P_s^{(i)} = 1, i=1, 2 \right). \quad (1)$$

Hughes used Bayes predictive approach to estimate probabilities (1) assuming that prior distribution of these probabilities is uniform. He derived a mean averaged generalization error, $\bar{P}_N^M$. It is an error averaged over immense variety of potentially possible classification problems defined by uniform prior distribution of probabilities (1). In finite training sample situations, theoretical graphs $\bar{P}_N^M = f(m)$ exhibited clear minima. This results initiated in a number of subsequent small learning sample size investigations (see e.g. [4-8] and references therein).

Characteristic property of many pattern recognition problems is a fact that prior probabilities of pattern classes are different. Available number of training vectors of

the less probable category is very small often. In order to compensate a shortage of training vectors of this pattern class, sometimes designers make significant efforts in order to collect more training vectors of the minority category. In order to investigate usefulness of this strategy we derive equations for mean generalization error of Multinomial classifier for two category case with unequal prior class probabilities, q_1 , q_2 , and unequal training sample sizes, N_1 , N_2 . We show that artificial growth of training vectors of less probable class could be harmful. Use of diverse modifications of predictive Bayes approach to estimate cell probabilities of the Multinomial classifier reduces both the generalization error and effects of imbalanced training sets.

2 Generalization Error of Standard Multinomial Classifier

2.1 The Theory

Let we have E crisp (categorical) valued features, where the e -th feature takes one of m_e states. In two category case, there exist $m = \prod_{e=1}^E m_e$ potential combinations (cells, states) of E features, $x_1, x_2, \dots, x_E$. Denote v_s, u_s , ($s=1, 2, \dots, m$) the first and second class cell probabilities. To create the Multinomial classification rule we have to know prior probabilities of the classes, q_1, q_2 , and $2(m-1)$ probabilities of the cells. The *Bayes decision rule* allocates vector $X = (x_1, x_2, \dots, x_E)^T$, falling into the s -th state, according to a maximum of products $q_1 P_s^{(1)}, q_2 P_s^{(2)}$. If both products are equal among themselves, we make arbitrary decision. The Bayes probability of error is expressed as

$$P_B = \sum_{s=1}^m \min \{ q_1 v_s, q_2 u_s \}. \quad (2)$$

In practice, we estimate probabilities (1) from training data. In standard sample based Multinomial and BKS fusion rules, one utilizes maximum likelihood (ML) estimates:

$$\hat{v}_s = n_s^{(1)} / N_1, \hat{u}_s = n_s^{(2)} / N_2, \quad (3)$$

where $n_s^{(i)}$ is a number of training vectors of i -th pattern class in the s -th cell.

Utilization of Eq. (2) results in the plug-in rule. If $q_2 \neq q_1$, $N_2 \neq N_1$, expected generalization error

$$\bar{P}_N^M = \sum_{s=1}^m [q_1 v_s P \{ q_1 \hat{v}_s < q_2 \hat{u}_s \} + q_2 u_s P \{ q_1 \hat{v}_s > q_2 \hat{u}_s \} + 0.5 (q_1 v_s + q_2 u_s) P \{ q_1 \hat{v}_s = q_2 \hat{u}_s \}]. \quad (4)$$

To analyze finite training sample size behavior, we assume that numbers of training vectors in each single cell, $n_s^{(i)}$, are Multinomial random variables. After some algebraic manipulation utilizing properties of multinomial distribution, following expression for the mean of the generalization error is obtained

$$\bar{P}_N^M = \sum_{s=1}^m \sum_{j=0}^{N_1} \sum_{t=0}^{N_2} \psi_{jt\beta} \frac{N_1!}{j!(N_1-j)!} v_s^j (1-v_s)^{N_1-j} \frac{N_2!}{v!(N_2-v)!} u_s^t (1-u_s)^{N_2-t}, \quad (5)$$

$$\text{where } \psi_{jt\beta} = \begin{cases} q_1 v_s & \text{if } j\beta < t \\ q_2 u_s & \text{if } j\beta > t \\ (q_1 v_s + q_2 u_s)/2 & \text{if } j\beta = t \end{cases} \text{ and } \beta = \frac{q_1}{q_2} \frac{N_2}{N_1}.$$

We see that generalization error depends on $2m$ probabilities, t_s and u_s ($s = 1, \dots, m$), and sample sizes, N_1 and N_2 . If *both* sample sizes, N_1, N_2 , are increasing without bound, expected error, $\bar{P}_N^M$, approaches the Bayes error, i.e.

$$P_B = \lim_{N_1 \rightarrow \infty, N_2 \rightarrow \infty} EP_N^M \quad (6)$$

If *only one of the* sample sizes, e.g. N_2 , is increasing without bound, expected error, $\bar{P}_{N_2 \rightarrow \infty}^M$, becomes

$$EP_{N_2 \rightarrow \infty}^M = \sum_{s=1}^m \sum_{j=0}^{N_1} \psi_{j\beta} \frac{N_1!}{j!(N_1-j)!} v_s^j (1-v_s)^{N_1-j}, \quad (7)$$

$$\text{where } \psi_{j\beta} = \begin{cases} q_1 v_s & \text{if } j\beta < u_s \\ q_2 u_s & \text{if } j\beta > u_s \\ (q_1 v_s + q_2 u_s)/2 & \text{if } j\beta = u_s \end{cases} \text{ and } \beta = \frac{q_1}{q_2} \frac{N_1}{N_2}.$$

2.2 Numerical Example: Case $N_2 = N_1$

Calculation of generalization error (Eq. (5)) requires to know the $2(m-1)$ probabilities, $v_1, v_2, \dots, v_{m-1}, u_1, \dots, u_{m-1}$. If m is very large, it could turn out to be serious difficulty for a practitioner. Therefore, in [8, Section 3.8] for tabulation purposes a simplifying model, **MU**, of distribution of values $v_1, v_2, \dots, v_m, u_1, \dots, u_m$, has been proposed. It was assumed that in two category case, the probabilities of $m/2$ cells of the i -th pattern class are equal among themselves and equal to probabilities of other $m/2$ cells of opposite, the $(3-i)$ -th, class. Thus, in model **MU**, one deals only with two values of the cells' probabilities, $2P_B/m$, and $2(1-P_B)/m$. We present a *few critical comments* below.

Let the training *sample size be very small*. Then in recognition (test) phase, the class label of the s -th cell having probability $P_s^{(i)} = 2P_B/m$ can be easily "confused" with class label of the cell of opposite category having probability $P_s^{(3-i)} = 2(1-P_B)/m$. In model **MU**, the probabilities of each half of the cells are artificially equalized. It means that in very small training sample cases, model **MU** "confuses" class labels more often and overestimates generalization error.

Consider now a situation where *learning sample size is large*. Then in recognition phase, it will be more difficult to confuse the s -th cell of one class having small probability $P_s^{(i)} = 2P_B/m$ with the cell of another class having much higher probability,

Table 1. Expected classification error of Multinomial classifier for “standard” model **MU** (six columns for $m = 26, 38, \dots, 144$) and real world data based model with 128 cells (two very right columns, theory and experimental evaluation). In both data models, the Bayes error $P_B=0.06$.

$N \backslash m$	26	36	46	62	106	144	128RWM	128experim
35	0.1120	0.1507	0.1954	0.2309	0.3103	0.3495	0.1134	0.1138 (73)
50	0.0837	0.1100	0.1374	0.1778	0.2590	0.3037	0.1080	0.1087 (72)
70	0.0688	0.0837	0.1022	0.1336	0.2082	0.2549	0.1026	0.1031 (70)
100	0.0621	0.0603	0.0779	0.0978	0.1571	0.2011	0.0968	0.0976 (65)
200	0.0600	0.0600	0.0613	0.0650	0.0868	0.1121	0.0863	0.0873 (51)
300	0.0600	0.0600	0.0601	0.0607	0.0682	0.0809	0.0808	0.0820 (43)
400	0.0600	0.0600	0.0600	0.0601	0.0627	0.0688	0.0773	0.0887 (37)
500	0.0600	0.0600	0.0600	0.0600	0.0609	0.0638	0.0748	0.0862 (32)
1000	0.0600	0.0600	0.0600	0.0600	0.0600	0.0601	0.0689	0.0705 (21)

$P_s^{(3-i)} = 2(1 - P_B)/m$. In such case, confusions of the class labels will be rare, i.e. model **MU** begins to underestimate generalization error. In six left columns of Table 1 we tabulated generalization errors for model **MU** if $q_2 = q_1 = 0.5$, $N_2 = N_1 = N$.

For comparison of model **MU** with “reality” we have chosen a real-world data obtained in problem of classifying two category eight-dimensional spectral Satellite data by means of multiple classifiers system. Seven multilayer perceptron based classifiers served as seven base experts that produced crisp outputs, 0 (first class) or 1 (the second class). So, $E=7$, $m = 2^7 = 128$. In Fig. 1 we have a scatter diagrams of 126 bi-variate vectors (v_s, u_s) ($s = 2, 3, \dots, 127$), the cell probabilities. Totally, 15,787 vectors from two pattern classes were used to estimate the probabilities. Probabilities of the first (the expert answers are: 0 0 0 0 0 0 0) and the last cells (the expert answers are: 1 1 1 1 1 1 1) differ from remaining ones basically: $v_1 = 0.0247$, $u_1 = 0.7735$ and $v_{128} = 0.7291$, $u_{128} = 0.0114$. All 128 experimentally evaluated values, $v_1, v_2, \dots, v_{128}$, have non-zero probabilities. In majority of the cells, however, the probabilities v_s, u_s are very small. The Bayes error $P_B = 0.06$.

Fig. 1. Scatter diagrams of 126 bi-variate vectors (v_s, u_s) in two scales

We specify this real world data model with diverse 128 cell probabilities as **RWM**. In column “128**RWM**” of Table 1 we present generalization errors calculated according to Equation (5). In the very right column, “128**experim**”, we present generalization error evaluated experimentally (*averages* of 5000 independent experiments where N randomly selected 7D binary vectors of each category were used for training of the Multinomial classifier and the test was performed on remaining 15,787 - $2N$ vectors). In brackets we present *standard deviations* multiplied by 10000. The last two columns show very good agreement between theoretical and experimental evaluations. Comparison of column “128**experim**” with six columns calculated for standardized model MU does not show any god agreement: generalization errors calculated for the RWM data model (marked in bold) coincide with that calculated for different MU models characterized by diverse number off cells, m (marked also in bold).

In *real world problems* usually we have small number of cells with high probabilities v_s , u_s and a large number of ones with small probabilities v_s , u_s . The cells having large probabilities v_s , u_s likely will be classified correctly even in small sample situations. In model MU, however, large probabilities v_s , u_s will be present only in situations where number of cells, m , is small. For that reason, model MU with m comparable with that of real world data overestimates the generalization error in small sample situations (see Table 1).

In large training sample size cases, the cells with moderate differences between probabilities v_s and u_s are not confused any more. For that reason, these cells do not influence an increase in generalization error. There exist, however, large number of cells with small and, therefore, with close probabilities, v_s and u_s . Class labels of these cells can be confused even in large training set situation. For that reason, in large sample case, most important become the cells with small probabilities. When the sample size $N = N_1 = N_2 = 300$, calculation according to Eq. (5) for RWM model gives generalization error $\bar{P}_N^M = 0.0808$. For the standardized data model, MU, this error rate can be obtained if $m=144$ cells ($\bar{P}_N^M=0.0809$). It means that almost empty cells are affecting an increase in the expected generalization error. At this time, one can say that “an effective number of cells” is higher: $m_{\text{effective}}=144$. Thus, in small sample cases (if expected error exceeds the Bayes error 1.5 times and more), standard data model MU overestimates the generalization can, and underestimates it if the sample size is large. Thus, tabulated values for model MU (Table 3.7 in [8] can serve only as a guide of sufficiency of sample size necessary to design BKS fusion rule. In most important pattern recognition problems, one cannot utilize standard data models like MU. One must find a way to evaluate character of distribution of $2(m-1)$ cells’ probabilities and use Equation (5).

2.3 Numerical Example: Case $N_2 \neq N_1$

As a rule, in neural network and pattern recognition literature, however, a sum number of vectors, $n = N_1 + N_2$, has been considered as a single measure of training set size (see e.g. [9 - 11]). Equation (5) points out that for the Multinomial classification rule, the generalization error depends on training set sizes of both pattern classes, N_1 and N_2 . The Multinomial classifier and BKS fusion rules are

heuristically based plug-in classification rules. Here unknown probabilities of the cells are substituted by their maximum likelihood sample estimates. Consequently, *it is not optimal sample based decision rule*. If the number of training vectors of the less probable category is very small, some researchers compensate a shortage of training vectors of this category by additional collection of more training vectors of the minority pattern class.

In Fig. 2 we present theoretical and experimental graphs “the generalization error, $\bar{P}_N^M$, as function of N_2 , the number of training vectors of second pattern class”, in experiments with above mentioned 7D binary data. The number of training vectors of the first class was kept constant: $N_1 = 25$. The number of vectors of the second class, N_2 , varied between 3 and 1800. Prior probability of the first class, q_1 , was set equal to 0.9. Graph 1 in Fig. 1 corresponds to theoretical calculations according to Equation (5). Due to small sample size of the first pattern class, we had large oscillations among different random formations of the training sets. Therefore, 50000 independent experiments with randomly formed learning sets were performed in situations where $N_2 \leq 50$ and 5000 experiments when $N_2 \geq 100$. In Fig. 2 we present mean values.

Both graphs indicate that in situations where prior probabilities of the pattern classes are unequal, artificial increase in training set of smaller category can become harmful: with an increase in N_2 , the generalization error increases from its minimum, 0.107, up to 0.173. Important conclusion follows: *an increase in the number of training vectors of one pattern class not always leads to success. To obtain best generalization results one needs to pay attention to correct balance between numbers of vectors in different pattern classes*. Thus, conventional practice where the designer collects extra training vectors of one pattern class not always is a good policy.

Fig. 2. Generalization error, P_{gen} , as a function of a number of training vectors, N_2 : 1 – theoretical values, 2 – experimental evaluations, 3 – asymptotic error $\bar{P}_{N_2 \rightarrow \infty}^M$; $N_1 = 25$

Specific application of Eq. (5) arises in analysis of decision tree classifier. Decision tree is a useful strategy to tackle small sample problems. In small sample cases, the number of final leaves must be reduced substantially. Formula (5) can be utilized for provisional evaluation of the generalization error of full decision tree as well as for calculation of partial generalization errors corresponding to distinct

branches of the tree. In separate branches, the training sample sizes from opposite categories can differ substantially. It is obvious that formal reduction in a number of the leaves without changing authentic decision rule does not change the classifier's small sample properties. More deep investigation of decision tree classifier suggests that *while simplifying the decision making scheme the number of training errors should increase.*

3 Generalization of the Bayes Predictive Rule

3.1 The Theory: Regularized Multinomial Classifier

To get Bayes predictive rule of Multinomial classifier, one has to know prior distribution of $v_s, u_s, (s = 1, 2, \dots, m)$. Let prior distribution of the probabilities follow a Dirichlet (polynomial) distribution. To be short, we write formulae for one class only:

$$P_{\text{prior}}(v_1, v_2, \dots, v_m) = \beta_{mi} (v_1)^{\gamma_1-1} (v_2)^{\gamma_2-1} \dots (v_2)^{\gamma_{m-1}-1} (v_m)^{\gamma_m-1}, \quad (8)$$

where β_{mi} is a normalizing constant and $\gamma_1, \gamma_2, \dots, \gamma_m$ are parameters of the prior distribution. Then the Bayes estimate of v_j is

$$\hat{v}_j = (n_j^{(1)} + \gamma_j) / (N_1 + \sum_{j=1}^m \gamma_j), \quad (9)$$

If we do not give preference to any particular bin, then $\gamma_1 = \gamma_2 = \dots = \gamma_m = \gamma$. Assume that the prior distribution of probabilities $v_1, v_2, \dots, v_m$ follow uniform distribution, i.e. $\gamma=1$. Then the Bayes predictive estimate of v_j and u_j are

$$\hat{v}_j = (n_j^{(1)} + 1) / (N_1 + m), \quad \hat{u}_j = (n_j^{(2)} + 1) / (N_2 + m). \quad (10)$$

Then, in Equations (5) and (7) one has to use new, index j dependent β value:

$$\beta = \frac{q_1}{q_2} \frac{N_2 + m}{N_1 + m} + \frac{1}{j} \left(\frac{q_1}{q_2} \frac{N_2 + m}{N_1 + m} - 1 \right). \quad (11)$$

Calculations show that Bayes approach with uniform prior distribution of parameters v_j and u_j could reduce generalization error substantially if $N_2 \neq N_1$. Uniform prior distribution contains extremely vague information about the parameters of the model. In order to introduce certain prior information a couple of possibilities will be discussed below. Assume at first that *there exist an independent prior data set* of $N_{pi} = \sum_{j=1}^m r_{ij}$ vectors to be used to determine prior distribution of $v_1, v_2, \dots, v_m$, where r_{is} is a number of prior data set vectors of i -th class in the s -th cell. Then

$$P_{\text{apost}}(v_1, v_2, \dots, v_m) = \beta_{ri} (v_1)^{r_{i1}+1} (v_2)^{r_{i2}+1} \dots (v_m)^{r_{im}+1}, \quad (12)$$

which could be utilized as new prior distribution. Subsequently, new Bayes estimates

$$\hat{v}_j^* = (n_j^{(1)} + r_{1j}) / (N_1 + \sum_{j=1}^m r_{1j} + m), \quad \hat{u}_j^* = (n_j^{(2)} + r_{2j}) / (N_2 + \sum_{j=1}^m r_{2j} + m). \quad (13)$$

In spite of theoretical simplicity utilization of estimate (13) is problematic since researchers do not have “*the independent prior data set*” in typical situations. If such set would be available, it would be merged with training set. At times, one could form such set artificially. One of examples could be a situation where one solves several similar, however, to some extent different pattern recognition tasks. Then specially merged training sets of all tasks could compose “*the independent prior data set*”.

If additional set of vectors is unavailable, one can try using training vectors and some *additional information* to construct “synthetic prior distribution”. For example, one can use k - nearest neighbors directed noise injection [12, 13] originally suggested by R. Duin. A noise injection, in fact, introduces *additional non-formal information*: it declares in an inexplicit way that a space between nearest vectors of one pattern class could be filled with vectors of the same category. One can make an assumption (a guess) that L components, $x_1, x_2, \dots, x_E$ of feature vector $\mathbf{X}$ are statistically independent. Then parameters $\gamma_1, \gamma_2, \dots, \gamma_m$ in Eq. (8) could be evaluated as a scaled by λ ($0 < \lambda < 1$) product of E estimates $\hat{P}_{ie} = n_{ie} / N_i$ ($e=1, \dots, E$). Here n_{ie} stands for a number i -th class training vectors where e -th component takes zero value. Similar to the conventional regularized discriminant analysis (RDA) [2, 8], parameter λ plays a smoothing role. Resembling the RDA, optimal value of λ should decrease with an increase in training sample size and with an increase in complexity of distribution of the cell probabilities $v_1, v_2, \dots, u_m$. Thus, it has to be determined in experimental way.

3.2 Experiments

For illustration of usefulness of “shaky prior information” incorporated into Bayes predictive approach in small learning sample situations, we considered two category 7D binary data set already discussed in Sections 2.2 and 2.3. In Fig. 3a, similarly

Fig. 3 Generalization error, P_{gen} , as a function of λ , the “regularization” constant (a) and of a number of training vectors, N_2 (b): 1 – theoretical values, Equations (5) and (10), 2 – experimental evaluations, 3 – asymptotic error $\bar{P}_{N_2 \rightarrow \infty}^M$; $q_1 = 0.9$, $N_1 = 25$

to the graphs in Fig. 2, we have theoretically and experimentally evaluated dependence of generalization error, P_{gen} , on a number of training vectors, N_2 when $N_1 = 25$.

We assumed that $q_1=0.9$ and used this q_1 value both in theoretical and experimental evaluations of generalization error. Comparison of graphs in Figures 2 and 3a shows that even in case of uniform (almost uninformative) prior distribution Bayes predictive approach allows to reduce generalization error. No increase in generalization error is observed when training set size $N_2 \rightarrow \infty$. Small difference between theoretical experimental graphs observed for large N_2 is caused by the fact that theoretical values are calculated for an entity of situations defined by uniform prior distribution. In the experiment, however, we considered one particular selection of values $v_1, v_2, \dots, u_m$.

In Fig. 3b we have dependence of generalization error on λ , the “regularization” constant when synthetic prior distribution was obtained from training data, and the feature independence assumption was used. The experiments were performed 5000 times with randomly chosen training sets. A presence of minima indicates that, in principle, the performance of standard Multinomial classifier could be improved by introducing techniques similar to regularized discriminant analysis.

4 Concluding Remarks

We derived equations for generalization error of Multinomial classifier for situations when prior probabilities of the pattern classes are different. It was shown that artificial growth of training vectors of less probable class could be harmful if maximal likelihood estimates were used to design the classifier. Use of predictive Bayes approach to estimate cell probabilities of the Multinomial classifier reduces both the generalization error and negative effect of imbalanced training sets. High accuracy of theoretical formulae was confirmed by experiments with real world data set. A few variants of Bayes predictive approach were considered. It was shown that utilization of certain prior assumptions about the data suggests new ways how to regularize the Multinomial classifier and to reduce the generalization error.

References

- [1] Lachenbruch P.A and Goldstein M. Discriminant analysis. *Biometrics* 5(3): 9–85, 1979.
- [2] Duda R.O., Hart P.E. and Stork D.G. *Pattern Classification*. 2nd ed. Wiley, NY, 2000.
- [3] Hughes G.F. On the mean accuracy of statistical pattern recognizers. *IEEE Trans. on Information Theory* IT-14: 55–63, 1965.
- [4] Chandrasekaran B. and Jain A.K. Quantization complexity and independent measurements. *IEEE Trans. Computers*, 23: 102-1-6, 1974.
- [5] Duin R.P.W. *On the Accuracy of Statistical Pattern Recognizers*. Ph.D. dissertation. Delft University of Technology, Delft, 1978.
- [6] Griskevicius D. and Raudys S. On the expected probability of the classification error of the classifier for discrete variables. In S Raudys (editor), *Statistical Problems of Control*, 38:95–112. Institute of Mathematics and Informatics, Vilnius (in Russian), 1979.

- [7] Serych A.P. On use of nonparametric density estimates in pattern recognition. *Proc. of Siberian Physic technical V.D.Kuznetsov Institute*, 63: 13-41, (in Russian), 1973.
- [8] Raudys S. *Statistical and Neural Classifiers: An integrated approach to design*. Springer-Verlag, NY, p. 312, 200.
- [9] Amari S. A universal theorem on learning curves. *Neural Networks*, 6: 161–66, 1993.
- [10] Vidyasagar M. *A Theory of Learning and Generalization*. Springer, London, 1997.
- [11] Vapnik V. *Statistical Learning Theory*. John Wiley and Sons, N.Y., 1998.
- [12] Skurichina M., Raudys S. and Duin R.P.W. (2000). K-NN directed noise injection in multilayer perceptron training, *IEEE Trans. on Neural Networks*, 11(2): 504–511.
- [13] Raudys S. Experts' boasting in trainable fusion rule. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, PAMI 25 (9): 1178-1182, 2003.