Chemometrics and Intelligent Laboratory Systems 49 (1999) 105–115

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemometrics

Comparative study of class data analysis with PCA-LDA, SIMCA, PLS, ANNs, and *k*-NN

Yukio Tominaga [*]

Department of Chemistry I, Discovery Research Laboratories, Dainippon Pharmaceutical, Enoki 33-94, Suita, Osaka, 564-0053, Japan

Received 10 November 1998; received in revised form 20 May 1999; accepted 9 June 1999

Abstract

Three types of chemotherapeutic agents, antibacterials, antineoplastics, and antifungals, which are registered in the MDL drug data report (MDDR) database, were used as training data set, and the classification study was performed using the following seven methods: principal component analysis–linear discriminant analysis (PCA-LDA), soft independent modeling by class analogy (SIMCA), partial least-squares2 (PLS2), artificial neural networks (ANNs), nearest neighbor method (NN), combined method of Ward clustering and NN (W-NN), and combined method of genetic algorithms (GAs) and NN (GA-NN). The number of correctly classified samples for each method was decreased by the following order: NN, ANNs, GA-NN, SIMCA, PLS2, W-NN, and PCA-LDA. Using these models, prediction study was then performed for the test set which consists of the drugs registered in the comprehensive medicinal chemistry (CMC) database. The number of correctly predicted samples for each method was decreased by the following order: NN, GA-NN, W-NN, SIMCA, PCA-LDA, ANNs, and PLS2. NN gave the best model from view points of the classification and prediction while overfitting was observed in ANNs and PLS2. Although the fitness and predictiveness of GA-NN and W-NN were inferior to those of NN, the predictiveness of the two methods were superior to PCA-LDA, SIMCA, ANNs, and PLS2. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: PCA-LDA (principal component analysis–linear discriminant analysis); SIMCA (soft independent modeling by class analogy); PLS2 (partial least-squares2); Artificial neural networks; *k*-Nearest neighbor method

1. Introduction

The data structure of quantitative structure–activity relationships (QSAR) data set is classified into symmetric and asymmetric structures [1]. In the symmetric data structure, the samples in the two classes are linearly separable in descriptor space. For such case, linear discriminant analysis (LDA) [2–4] can be used. In the asymmetric (embedded) data structure, the samples in the two classes are not linearly separable in descriptor space, so LDA may fail to provide a model, while soft independent modeling by class analogy (SIMCA) [5,6] or *k*-nearest neighbor method (*k*-NN) gives a successful model [7,8]. The modeling strategies of these methods are substantially different. The former analysis, LDA, focuses on the dissimilarity between classes, whereas the latter methods, SIMCA and *k*-NN, focus on the similarity within a class. In SIMCA modeling, common features of each class are extracted by principle component (PC) model. The SIMCA classification rule determines class membership by the orthogonal projection distance between an unknown sample and the PC model

[*] Corresponding author. Tel.: +81-6-6337-5898; fax: +81-6-6338-7656; E-mail: yukio-tominaga@dainippon-pharm.co.jp

PII: S0169-7439(99)00034-9

of each class. In *k*-NN modeling, the distance between an unknown sample and each sample in the training set is calculated, and the unknown sample is classified into the class that the majority of the *k* nearest neighbors in the training set belongs to.

Typically, the data structure of a symmetric data set may change to asymmetric as the diversity of samples expands. The data structure of an asymmetric data set may also change to complex data structure, e.g., multi-embedded structure, according to further expansion of molecular diversity. The changes in these data structures are schematically illustrated in Fig. 1. The recent advancement in high-throughput biological screening has helped us screen thousands of compounds in a short time. Efficient utilization of the screening results is the key factor to improve the optimization process. The data structure of high-throughput biological screening data is substantially equivalent to the complex data structure mentioned above. To analyze such complex data, it is important to determine an appropriate chemometrics method as well as to perform comparative study with different modeling strategies [9].

In this study, three types of chemotherapeutic agents, antibacterials, antineoplastics, and antifungals, which are registered in the MDL drug data report (MDDR) database, were used as training data set [10]. The training set contains 12 242 samples. The test set consists of 960 samples from comprehensive medicinal chemistry (CMC) database [10]. The modeling of the three categorical data was performed by using principal component analysis–linear discriminant analysis (PCA-LDA), SIMCA, partial least-squares2 (PLS2) [11–13], artificial neural networks (ANNs) [14,15], nearest neighbor method (NN), combined method of Ward clustering and NN (W-NN), and combined method of genetic algorithms (GAs) and NN (GA-NN). The results were compared from view points of the classification and prediction.

2. Method

2.1. Data set

Two-thousand-eighty (2080) antibacterials, 1985 antifungals, and 8177 antineoplastics from MDDR database [10] were used as a training set. Four-hundred-fifty-two (452) antibacterials, 102 antifungals, 406 and antineoplastics from CMC database [10] were used as a test set. These 2D chemical structures were converted into 3D structures by using Converter 95.0 [16] within Insight II. The 3D structures were further optimized by MAXIMIN2 within SYBYL [17].

2.2. Descriptors

Using Tsar v3.1 [18], 156 descriptors were assigned to each molecule. The descriptors are shown in Table 1. All descriptors were regularized to give equivalent variance and means of zero.

2.3. System used for data analysis

All calculations were carried out by using Indigo 2 running version 6.2 of the IRIX operating system.

2.4. Modeling

2.4.1. PCA-LDA

PCA was performed with self-written program which was coded by C. As the matrix (12 242 × 156) of training data set includes an immense number of

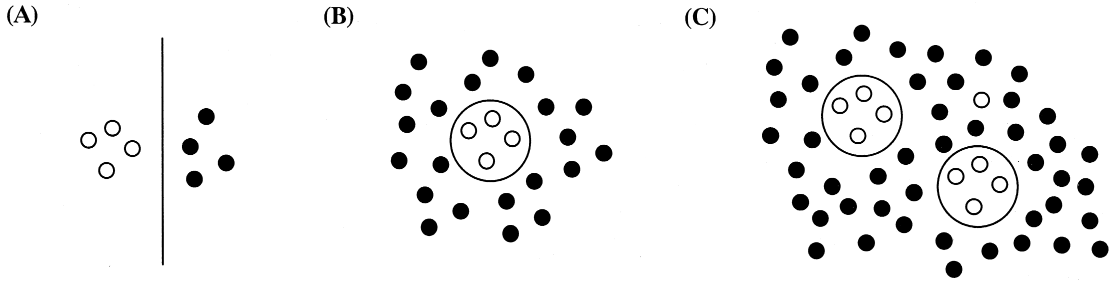

Fig. 1. Schematic representation of data structures. (A) A symmetric data structure. (B) An asymmetric data structure. (C) A complex data structure.

Table 1
Descriptors in this study

Number	Descriptor	Number	Descriptor
1	Molecular mass	39	Kier ChiV4 (ring) index
2	Molecular volume	40	Kier ChiV5 (ring) index
3	Inertia moment 1 size	41	Kier ChiV6 (ring) index
4	Inertia moment 2 size	42	Kappa1 index
5	Inertia moment 3 size	43	Kappa2 index
6	Inertia moment 1 length	44	Kappa3 index
7	Inertia moment 2 length	45	KAlpha1 index
8	Inertia moment 3 length	46	KAlpha2 index
9	Ellipsoidal volume	47	KAlpha3 index
10	Total dipole moment	48	Shape flexibility index
11	Dipole moment X component	49	Randic topological index
12	Dipole moment Y component	50	Balaban topological index
13	Dipole moment Z component	51	Wiener topological index
14	Kier Chi0 (atoms) index	52	Sum of E-state indices
15	Kier ChiV0 (atoms) index	53	Number of atoms
16	Kier Chi1 (bonds) index	54	Number of halogen atoms
17	Kier ChiV1 (bonds) index	55	Number of heteroatoms
18	Kier Chi2 (path) index	56	Number of H-bond donors
19	Kier ChiV2 (path) index	57	Number of H-bond acceptors
20	Kier Chi3 (cluster) index	58	Number of H atoms
21	Kier ChiV3 (cluster) index	59	Number of C atoms
22	Kier Chi4 (cluster) index	60	Number of N atoms
23	Kier ChiV4 (cluster) index	61	Number of O atoms
24	Kier Chi4 (path/cluster) index	62	Number of F atoms
25	Kier ChiV4 (path/cluster) index	63	Number of P atoms
26	Kier Chi3 (path) index	64	Number of S atoms
27	Kier Chi4 (path) index	65	Number of Cl atoms
28	Kier Chi5 (path) index	66	Number of Br atoms
29	Kier Chi6 (path) index	67	Number of I atoms
30	Kier ChiV3 (path) index	68	Number of 3-membered rings
31	Kier ChiV4 (path) index	69	Number of 4-membered rings
32	Kier ChiV5 (path) index	70	Number of 5-membered rings
33	Kier ChiV6 (path) index	71	Number of 6-membered rings
34	Kier Chi3 (ring) index	72	Number of 7-membered rings
35	Kier Chi4 (ring) index	73–93	2D Autocorrelogram (no weighting, separation is from 0 to 20)
36	Kier Chi5 (ring) index	94–114	2D Autocorrelogram (weighted by partial charge, separation is from 0 to 20)
37	Kier Chi6 (ring) index	115–135	3D Autocorrelogram (no weighting, separation is from 0 to 20 A)
38	Kier ChiV3 (ring) index	136–156	3D Autocorrelogram (weighted by partial charge, separation is from 0 to 20 A)

objects, POWER method [19] was used. Using the score vectors of PCA, PCA-LDA was performed with self-written program coded by C.

2.4.2. SIMCA

SIMCA analysis was carried out within SYBYL. Optimum component of PC model was determined for each class by cross-validation. Five cross-validation groups were used.

2.4.3. PLS

The class of each sample was coded by three binary strings. In other words, antibacterials, antineoplastics, and antifungals represent (1,0,0), (0,1,0), and (0,0,1), respectively. As three dependent variables were used, PLS2 analysis was performed within SYBYL. The three predicted scores were used to predict the class of each sample. When the first score was more than or equal to 0.5 and the second and third scores were both less than 0.5, the predicted class of the sample was antibacterial. In the same way, antineoplastics and antifungals were predicted.

2.4.4. ANNs

ANNs analysis was performed within Tsar. Similar to PLS analysis, each class was coded by three binary strings. A three-layer network with input, hidden, and output units was used. The ratio between the number of input variables and the number of variables controlled by the network, ρ, was set to near 2.0, and the configuration of input, hidden, and output units became 156–38–3. Each weight of the bond that connects between hidden and input units or output and hidden units was trained by back-propagation. The prediction of the class of individual samples was performed by using three predicted output scores.

2.4.5. K-NN and subset selection

K-NN analysis was performed by using self-written program coded by FORTRAN 77. K value of one was selected for the three classes.

As the number of samples within training set increased, k-NN analysis requires more distance calculations. To avoid too many distance calculations, subsets were selected from the training set by using two selection methods, Ward clustering [20,21] and genetic algorithms (GAs) [22–25]. *k*-NN with *k* value of 1 was performed based on the selected samples. Ward clustering was carried out within Tsar. One hundred (100), 201, 309, 395, 505, 625, 687, 770, 869, and 969 samples were selected as the representative subsets. GAs were performed using self-written program coded by FORTRAN 77. To classify between the selected samples and non-selected samples, a binary string, either 1 or 0, was assigned to each sample. The sample to which 1 was assigned was the selected sample. The dimension of the string corresponds to the number of samples in the training set. The fitness function was defined as the number of correctly classified samples. When this fitness function is used in the optimization process of GAs, the invaluable final model may be constructed because the number of selected samples in the final model may be identical to the number of the samples within the training set. If all the selected samples were not used to predict the other samples within the training set, the selected sets or strings were omitted during the optimization process of GAs. In other words, when *k*-mean clustering was performed using the selected samples as the seed points, the selected sets or strings were omitted if there were any singletons. A population of 200 randomly selected combinations of strings was generated. Fifty strings with highest fitness scores were then selected as the initial population. Three pairs of strings (parents) were selected using roulette wheel selection method, and crossed-over. Each offspring was subjected to random point mutation with the mutation ratio of less than 1%. The offspring with the highest fitness score and the parent with the closest string to that offspring were selected. The Tanimoto coefficient [26] was used as the similarity measure.

$$\begin{aligned}\text{Tanimoto coefficient}\\ &= \{\Sigma(\text{parent's string})(\text{offspring's string})\}\\ &\quad /\{\Sigma(\text{parent's string})^2 + \Sigma(\text{offspring's string})^2\\ &\quad -\Sigma(\text{parent's string})(\text{offspring's string})\} \qquad (1)\end{aligned}$$

If the fitness score of the offspring was superior to that of the parent, the parent was replaced with the offspring. The optimization process was repeated 500 times to improve the fitness score of the population. When exploration was finished, the model with the highest score in the final generation was selected as the final model.

3. Results and discussion

3.1. Data structure

PCA was performed for the training set (12 242 × 156) to visualize the complex data structure. The first two components can explain 54.4% of the variation of the training set. The scores of the first and second components of each class are shown in Figs. 2–4 for antibacterials, antifungals, and antineoplastics, respectively. The number of subclasses included in each class of agents is significantly large; each plot contains a large number of clusters. That is, the samples in each class may not act by the same mechanism. When comparing the score plots of three classes, it is clear that the total data structure of three classes is complex and multi-embedded.

3.2. PCA-LDA

LDA was performed using the 38 score vectors of PCA, which explain 95.3% of the variation of the training set. The results of PCA-LDA are shown in Tables 2 and 3. In this model, 59.0% of the agents were correctly classified, and 54.2% of the agents were correctly predicted.

3.3. SIMCA

SIMCA analysis was performed by using 156 descriptors. When each PC model used 38 components,

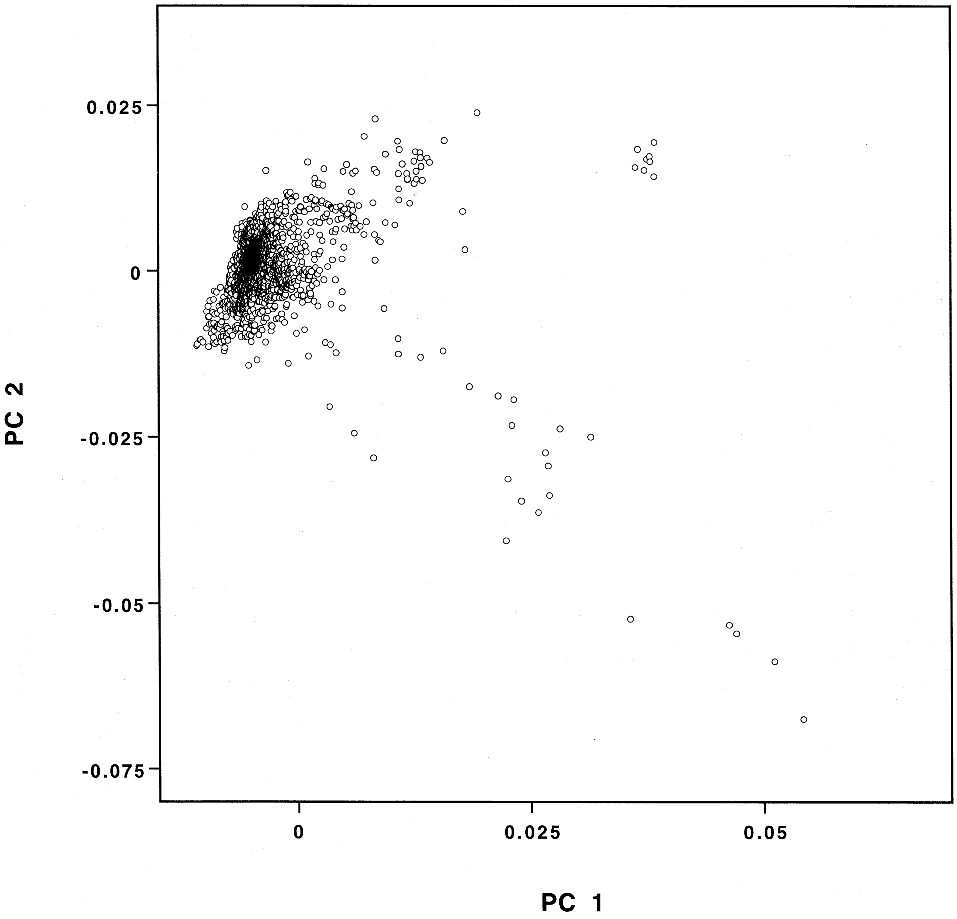

Fig. 2. Score plot of antibacterials.

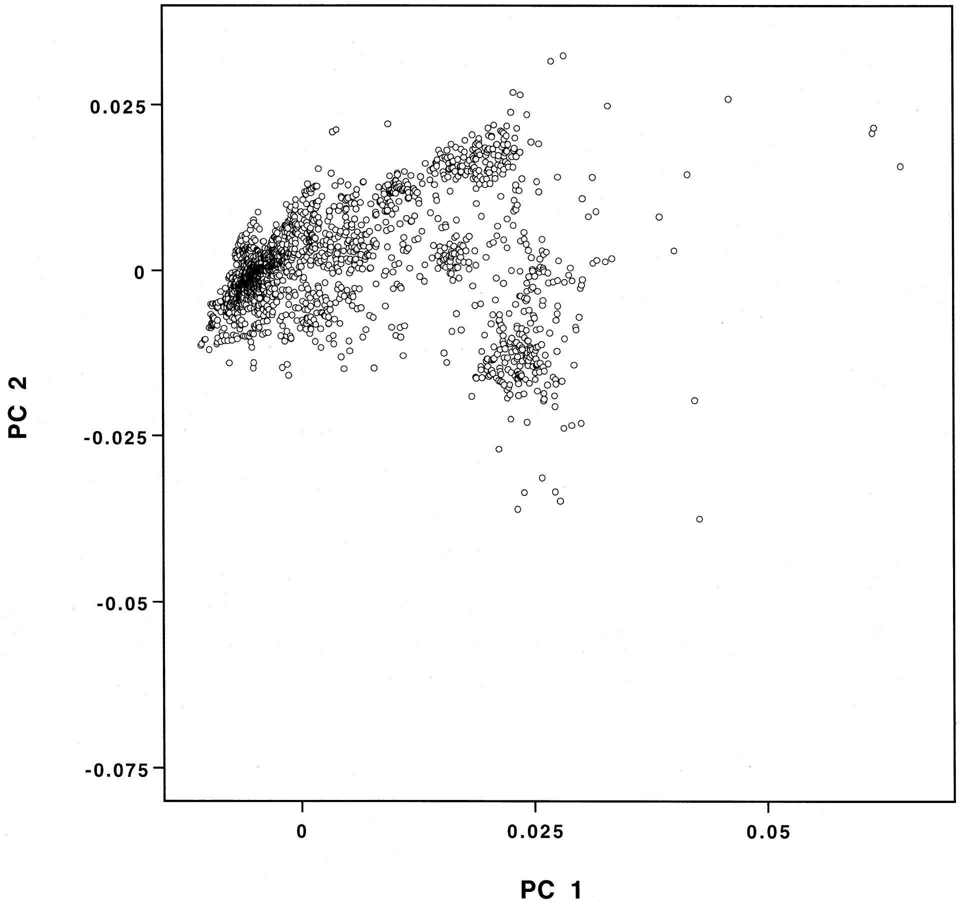

Fig. 3. Score plot of antifungals.

it explained more than 95% of the variation of each class. Thirty-eight or less components of PC model were used for each class to avoid overfitting. An appropriate number of PC was determined for each class by cross-validation. The number of PCs was 38, 38, and 38 for antibacterials, antifungals, and antineoplastics, respectively. The results of SIMCA analysis are shown in Tables 2 and 3. In the SIMCA model, 79.8% of the agents were correctly classified, and 56.0% of the agents were correctly predicted.

3.4. PLS2

PLS2 analysis was performed by using 156 descriptors, and cross-validation study was performed with the cross-validation groups of 10. The results are shown in Fig. 5. The Q^2 of the model increased rapidly at first, then gradually as the number of component increased. The optimum model could not be obtained because the optimum component was not clearly observed. To select the optimum model, the prediction study of the test set was performed according to the classification rule determined in Section 2.4.3. The prediction results are shown in Fig. 6. From the several peaks in the predictiveness, the model with 52 component was selected as optimum model to avoid overfitting. Q^2s of this model were 0.362, 0.335, and 0.400 for antibacterials, antifungals, and antineoplastics, respectively. R^2s of the model were 0.376, 0.354, and 0.412 for antibacterials, antifungals, and antineoplastics, respectively. The

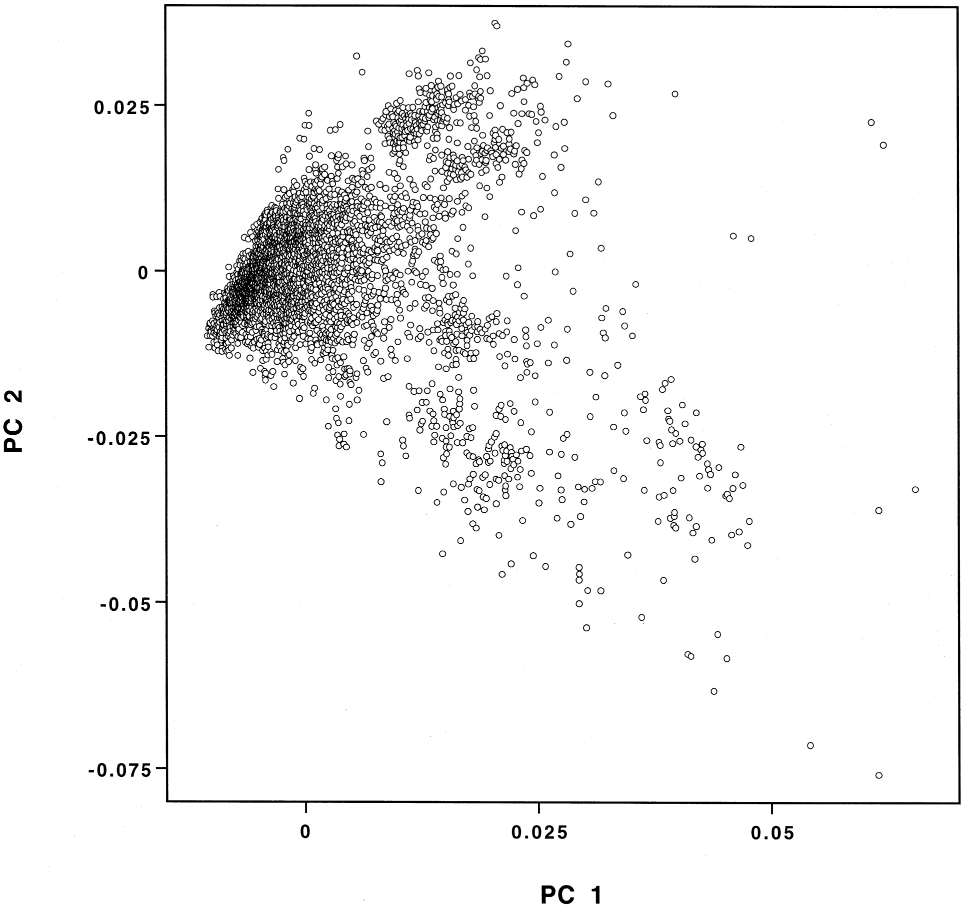

Fig. 4. Score plot of antineoplastics.

results of the classification and prediction are shown in Tables 2 and 3, respectively. In this model, 76.8% of the agents were correctly classified, and 45.5% of the agents were correctly predicted.

3.5. ANNs

ANNs with the configuration of 156-38-3 were trained by back-propagation. The results of the clas-

Table 2
Classification results using seven different approaches
The ratio of those correctly classified is shown in parentheses.

Method	Antibacterials	Antifungals	Antineoplastics	Total
PCA-LDA	414/2080 (19.9%)	924/1985 (46.5%)	5894/8177 (72.0%)	7232/12 242 (59.0%)
SIMCA	1868/2080 (89.8%)	1411/1985 (71.0%)	4713/8177 (57.6%)	9772/12 242 (79.8%)
PLS2	1018/2080 (48.9%)	767/1985 (38.6%)	7617/8177 (93.1%)	9402/12 242 (76.8%)
ANNs	1209/2080 (58.1%)	1143/1985 (57.5%)	7801/8177 (95.4%)	10 153/12 242 (82.9%)
NN	1561/2080 (75.0%)	1761/1985 (88.7%)	7656/8177 (93.6%)	10 978/12 242 (89.6%)
W-NN	1385/2080 (66.5%)	1232/1985 (62.0%)	6751/8177 (82.5%)	9368/12 242 (76.5%)
GA-NN	1336/2080 (64.2%)	1228/1985 (61.8%)	7481/8177 (91.4%)	10 045/12 242 (82.0%)

Table 3
Prediction results using seven different approaches
The ratio of those correctly predicted is shown in parentheses.

Method	Antibacterials	Antifungals	Antineoplastics	Total
PCA-LDA	169/452 (37.3%)	28/102 (27.4%)	324/406 (72.0%)	521/960 (54.2%)
SIMCA	285/452 (63.0%)	47/102 (46.0%)	206/406 (50.7%)	538/960 (56.0%)
PLS2	41/452 (9.07%)	19/102 (18.6%)	377/406 (92.8%)	437/960 (45.5%)
ANNs	82/452 (18.1%)	35/102 (34.3%)	379/406 (93.3%)	496/960 (51.6%)
NN	206/452 (45.5%)	54/102 (52.9%)	368/406 (90.6%)	628/960 (65.4%)
W-NN	189/452 (41.8%)	39/102 (38.2%)	342/406 (84.2%)	570/960 (59.3%)
GA-NN	180/452 (39.8%)	28/102 (27.4%)	368/406 (90.6%)	576/960 (60.0%)

sification and prediction are shown in Tables 2 and 3, respectively. In this model, 82.9% of the agents were correctly classified and 51.6% of the agents were correctly predicted.

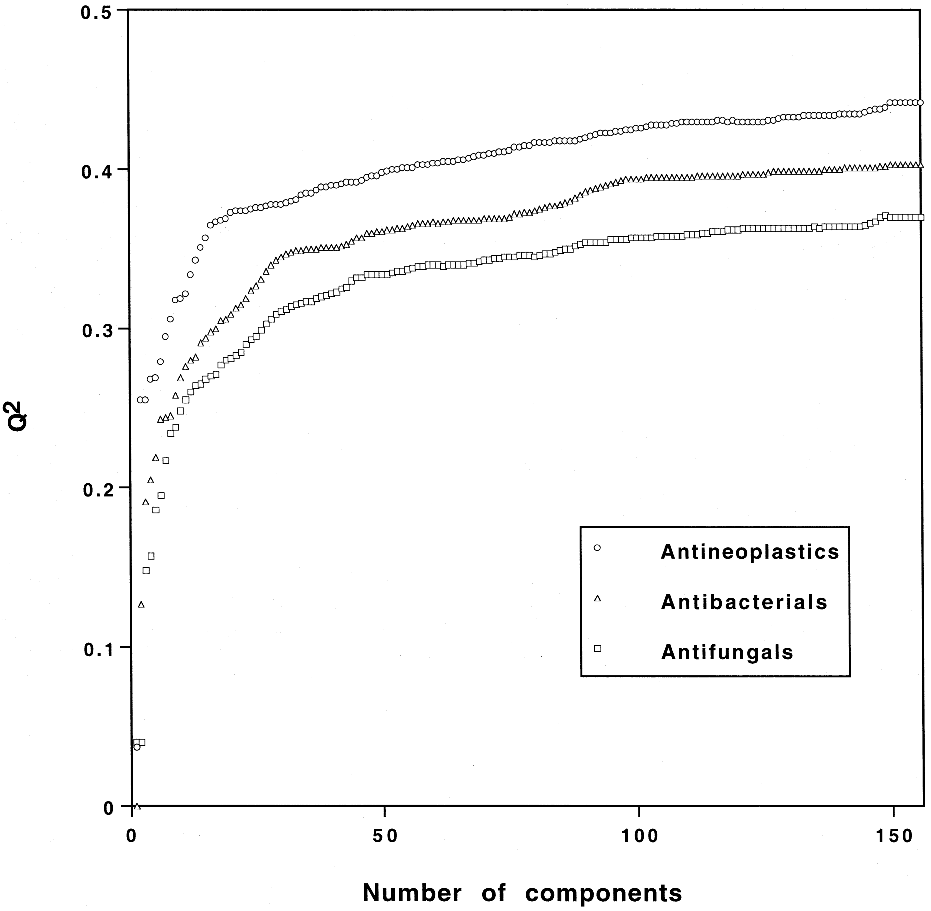

Fig. 5. Plot of Q^2 vs. the number of components.

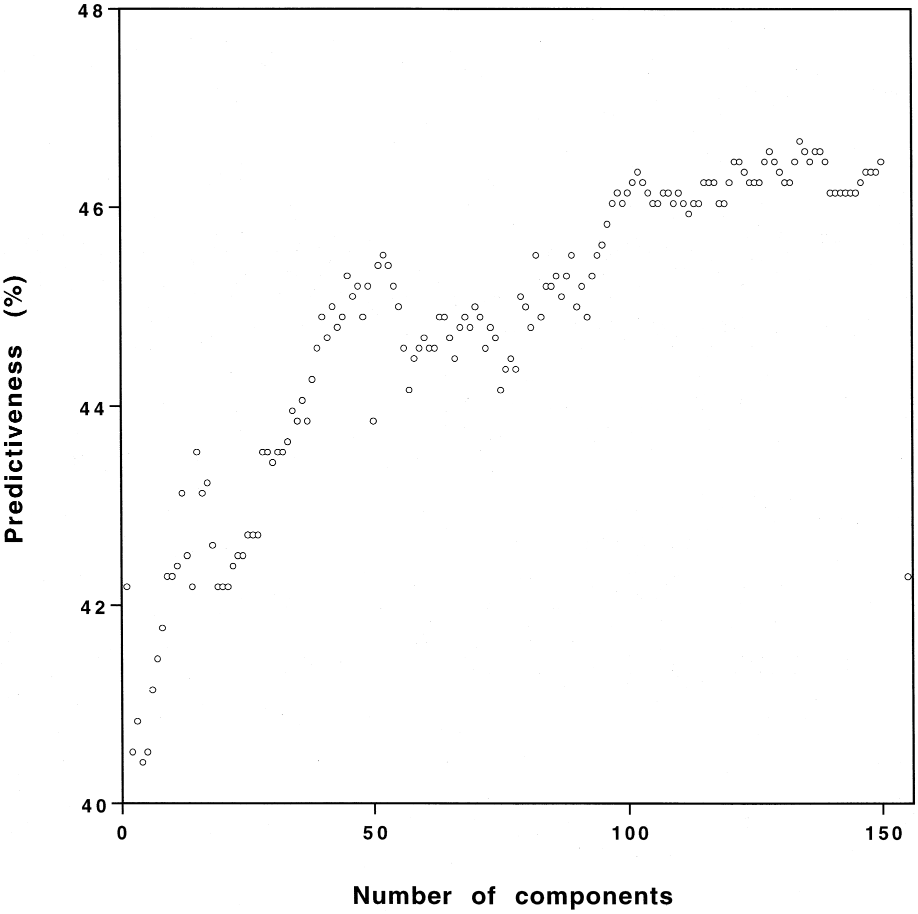

Fig. 6. Plot of predictiveness vs. the number of components. Predictiveness indicates the ratio of correctly classified agents.

3.6. NN

NN analysis was performed by using 156 descriptors. The results of NN analysis are shown in Tables 2 and 3. In this model, 89.6% of the agents were correctly classified, and 65.4% of the agents were correctly predicted.

3.7. Combined method of Ward clustering and NN (W-NN)

Ward clustering was performed, and 100, 201, 309, 395, 505, 625, 687, 770, 869, and 969 samples were selected from the training set as representative subsets. Using these subsets, NN analysis was performed. The results of the classification and prediction by the combined method (W-NN) are shown in Tables 4 and 5, respectively. The number of correctly classified samples in the training set was increased as the number of selected samples increased. On the other hand, the number of correctly predicted samples in the test set increased first, then gradually decreased as the number of selected samples increased. The optimum prediction was achieved when the number of selected samples was 505. In the model with 505 selected samples, 76.5% of the agents were correctly classified, and 59.3% of the agents were correctly predicted.

3.8. Combined method of GAs and NN (GA-NN)

GAs were used to select the subset which can optimize the number of correctly classified samples in

Table 4
The classification results using W-NN approach
N^* indicates the number of selected samples.
The ratio of those correctly classified is shown in parentheses.

N^*	Antibacterials	Antifungals	Antineoplastics	Total
100	789/2080 (37.9%)	1099/1985 (55.3%)	6677/8177 (81.6%)	8565/12 242 (69.9%)
201	1006/2080 (48.3%)	1170/1985 (58.9%)	6221/8177 (76.0%)	8397/12 242 (68.5%)
309	1256/2080 (60.3%)	1230/1985 (61.9%)	6340/8177 (77.5%)	8826/12 242 (72.0%)
395	1306/2080 (62.7%)	1286/1985 (64.7%)	6403/8177 (78.3%)	8995/12 242 (73.4%)
505	1385/2080 (66.5%)	1232/1985 (62.0%)	6751/8177 (82.5%)	9368/12 242 (76.5%)
625	1441/2080 (69.2%)	1269/1985 (63.9%)	6781/8177 (82.9%)	9491/12 242 (77.5%)
687	1456/2080 (70.0%)	1320/1985 (66.4%)	6806/8177 (83.2%)	9582/12 242 (78.2%)
770	1496/2080 (71.9%)	1376/1985 (69.3%)	6827/8177 (83.4%)	9699/12 242 (79.2%)
869	1514/2080 (72.7%)	1387/1985 (69.8%)	6907/8177 (84.4%)	9808/12 242 (80.1%)
969	1517/2080 (72.9%)	1412/1985 (71.1%)	6914/8177 (84.5%)	9843/12 242 (80.4%)

the NN analysis. The optimum model was obtained when the number of selected samples was 629. The results are shown in Tables 2 and 3. In this model, 82.0% of the agents were correctly classified, and 60.0% of the agents were correctly predicted.

3.9. Comparison of the seven methods

For classification, the number of correctly classified samples in the training set decreased by the following order: NN gave the best result (89.6% of samples were correctly classified); ANNs and GA-NN showed similar results (82.9% and 82.0%, respectively); SIMCA, PLS2 and W-NN showed similar results (79.8%, 76.8%, and 76.5%, respectively); and PCA-LDA provided poor result (59.0%). For prediction, the number of correctly predicted samples in the test set decreased by the following order: NN gave the best result (65.4% of samples were correctly predicted); GA-NN and W-NN showed similar results (60.0% and 59.3%, respectively); SIMCA, PCA-LDA, and ANNs (56.0%, 54.2%, and 51.6%, respectively); and PLS2 provided a poor result (45.5%).

The result of predictiveness of each method did not always correlated with that of classification. NN gave the best result in both classification and prediction. Although ANNs were superior to GA-NN, W-NN, SIMCA, and PCA-LDA in classification, they provided the inferior result with overfitting in predic-

Table 5
Prediction results using W-NN approaches
N^* indicates the number of selected samples.
The ratio of those correctly predicted is shown in parentheses.

N^*	Antibacterials	Antifungals	Antineoplastics	Total
100	52/452 (11.5%)	22/102 (21.5%)	368/406 (90.6%)	442/960 (46.0%)
201	102/452 (22.5%)	39/102 (38.2%)	317/406 (78.0%)	458/960 (47.7%)
309	132/452 (29.2%)	36/102 (35.2%)	319/406 (78.5%)	487/960 (50.7%)
395	180/452 (39.8%)	39/102 (38.2%)	321/406 (79.0%)	321/960 (56.2%)
505	342/452 (84.2%)	39/102 (38.2%)	189/406 (41.8%)	570/960 (59.3%)
625	337/452 (83.0%)	39/102 (38.2%)	192/406 (42.4%)	568/960 (59.1%)
687	335/452 (82.5%)	40/102 (39.2%)	192/406 (42.4%)	567/960 (59.0%)
770	327/452 (80.5%)	42/102 (41.1%)	190/406 (42.0%)	559/960 (58.2%)
869	304/452 (74.8%)	41/102 (40.1%)	187/406 (41.3%)	532/960 (55.4%)
969	316/452 (77.8%)	42/102 (41.1%)	184/406 (40.7%)	542/960 (56.4%)

tion. Overfitting was also observed in the PLS2 model. The classification result of PLS2 was superior to that of W-NN and PCA-LDA; however, the prediction result was inferior. NN, GA-NN, and W-NN showed better results in prediction when compared to ANNs, PLS2, SIMCA, and PCA-LDA. The superiority of NN, GA-NN, and W-NN in prediction is mainly due to the combination of the modeling strategy and data structure used in this study. As PCA-LDA put a focus on the dissimilarity between classes, it was not appropriate to model an embedded data set. SIMCA and PLS2 are appropriate to model an embedded data set because they put a focus on the similarity within a class. For the complex data set containing many subclasses, SIMCA and PLS2 provided a high-dimension model, and as a result, overfitting was observed especially in PLS2. In case of ANNs, nonlinear relationships were modeled but the model felled into overfitting. On the other hand, NN, GA-NN, and W-NN showed superiority in prediction because the modeling process properly handled subclasses.

While any explicit model for each class did not constructed by NN, it provided the best results for both classification and prediction. For GA-NN and W-NN, although the fitness and predictiveness were inferior to those of NN, the characteristics of each class were obtained by analyzing the representative samples.

4. Conclusion

Three types of chemotherapeutic agents, antibacterials, antineoplastics, and antifungals which are registered in MDDR database were classified by means of seven methods. Using the seven models, prediction was carried out for the test set from CMC database. NN gave the best model, from view points of both classification and prediction. Overfitting was observed in ANNs and PLS2. Although the fitness and predictiveness of GA-NN and W-NN were inferior to those of NN, these methods showed superiority in prediction to the other methods such as PCA-LDA, SIMCA, ANNs, and PLS2.

Acknowledgements

The author would like to thank the referees for their helpful and constructive comments.

References

[1] W.J. Dunn III, S. Wold, J. Med. Chem. 21 (1978) 1001–1007.

[2] D. Coomans, M. Jonckheer, D.L. Massart, I. Broeckaert, P. Blockx, Anal. Chim. Acta 103 (1978) 409–415.

[3] W. Wu, Y. Mallet, B. Walczak, W. Penninckx, D.L. Massart, S. Heuerding, F. Erni, Anal. Chim. Acta 329 (1996) 257–265.

[4] Y. Mallet, D. Coomans, O. de Vel, Chemom. Intell. Lab. Syst. 35 (1996) 157–173.

[5] S. Wold, Pattern Recognit. 8 (1976) 127–136.

[6] S. Wold, E. Johansson, E. Jellum, I. Bjornson, R. Nesbakken, Anal. Chim. Acta 133 (1981) 251–259.

[7] S. Wold, W.J. Dunn III, J. Chem. Inf. Comput. Sci. 23 (1983) 6–13.

[8] Y. Miyashita, Z. Li, S. Sasaki, Trends Anal. Chem. 12 (1993) 50–60.

[9] W. Werther, H. Lohninger, F. Stancl, K. Varmuga, Chemom. Intell. Lab. Syst. 22 (1994) 63–76.

[10] MDL Information Systems, 14600 Catalina Street, San Leandro, CA 94577.

[11] S. Wold, A. Ruhe, H. Wold, W.J. Dunn III, SIMA J. Sci. Stat. Comput. 5 (1984) 735–743.

[12] P. Geladi, B.R. Kowalski, Anal. Chim. Acta 185 (1986) 1–17.

[13] W.G. Glen, W.J. Dunn III, D.R. Scott, Principal components analysis and partial least squares regression, Tetrahedron Comput. Methodol. 2 (1989) 349–376.

[14] J. Zupan, J. Gasteiger, Anal. Chim. Acta 248 (1991) 1–30.

[15] S.S. So, W.G. Richards, J. Med. Chem. 35 (1992) 3201–3207.

[16] Converter 95.0, MSI 9685 Scranton Road, San Diego, CA 92121-2777, USA.

[17] SYBYL Molecular Modeling Software, version 6.3. Tripos Associates, St. Louis, MO 63144.

[18] Tsar 3.1, Oxford Molecular, Medawar Centre, Oxford Science Park, Oxford, OX4 4GA, England.

[19] W. Wu, D.L. Massart, S. de Jong, Chemom. Intell. Lab. Syst. 36 (1997) 165–172.

[20] F.A. Murtagh, Comput. J. 26 (1983) 354–359.

[21] Y. Tominaga, J. Chem. Inf. Comput. Sci. 38 (1998) 867–875.

[22] C.B. Lucasius, G. Kateman, Chemom. Intell. Lab. Syst. 19 (1993) 1–33.

[23] C.B. Lucasius, G. Kateman, Chemom. Intell. Lab. Syst. 19 (1993) 99–145.

[24] Y. Tominaga, Chemom. Intell. Lab. Syst. 43 (1998) 157–163.

[25] Y. Tominaga, J. Chem. Inf. Comput. Sci. 38 (1998) 1157–1160.

[26] R.D. Brown, Y.C. Martin, J. Chem. Inf. Comput. Sci. 36 (1996) 572–584.