Listy Biometryczne - Biometrical Letters
Vol. 40 (2003), No. 1, 15-22

On results obtained by discriminant analysis and classification trees for medical data with considerably different group sizes

Piotr Jurkowski[1], **Małgorzata Ćwiklińska-Jurkowska**[2], **Piotr Korbal**[3]

[1]Laboratory of Informatics and Research Methodology, [2]Department of Medical Informatics, [3]Clinic of Obstetrics and Gynecology, The Ludwik Rydygier Medical University, ul. Skłodowskiej 9, 85-094 Bydgoszcz, Poland
[1]jurkomal@mail.atr.bydgoszcz.pl

SUMMARY

The paper investigates the results of discriminant analysis and classification trees, when sizes of groups are considerably different, on the basis of a real medical dataset. Probabilities a priori proportional to sizes of groups and cross-validation assessment of classification errors for Bayesian parametric and nonparametric discrimination and for classification trees are used. For considered methods very high specificity but low sensitivity was obtained. To be useful for aiding diagnosis, the classification procedure on the basis of the given database should be able to provide not only the satisfactory global classification error, but also good sensitivity and (or) specificity, depending on the character of medical problem or doctor's preferences. To achieve this goal one can try different actions and then check the chosen procedure by a testing sample.

KEY WORDS: discriminant analysis, classification trees, cross-validation, specificity, sensitivity, medical protection, low-birth weight.

1. Introduction

Some medical databases available in Internet to examine the effectiveness of different new methods of pattern recognition have such a structure that they consist of considerably smaller set of individuals from one disease than from another one.

Additionally, these databases have sometimes the disadvantage that one or more predicting variable is nearly the same as the grouping variable (recognition). Observation of such variable can require a very precise, advanced technical examination, often invasive or expensive, so rarely recommended or available for the physician.

The aim of this work is to study the results of discriminant analysis and classification trees, when sizes of groups are considerably different, on the basis of a real medical dataset.

2. Material

Our data set forms a part of data collected in Clinic of Obstetrics and Gynecology in Bydgoszcz. We used 35 mixed variables (where variables numbered 2, 22, 23 are quantitative, the rest are qualitative):

1. Accident: car accident
2. Age: mother's age
3. Amniocen: Amniocentesis
4. Anaphyla: anaphylactic shock
5. Angina: angina and similar infections
6. Asphyxia: danger of asphyxia
7. Bleeding: Bleeding during pregnancy
8. CardDis: cardiac diseases (without coronary heart disease)
9. CathInfl: common catharal infections in the second and third trimester of pregnancy
10. CathInfl: common catharal infections in the first trimester of pregnancy
11. CNSabnor: central nervous system abnormalities
12. Diabetes: diabetes
13. DiabPreg: diabetes of pregnant women
14. Dystroph: dystrophy
15. Epilepsy: epilepsy
16. Gestosis: gestosis
17. HarvSut: harvett suture
18. Hyperten: hypertension
19. InfLabor: infection proved in laboratory tests
20. InflamUp: inflammation of upper breathing tract
21. Insulin: insulin therapy
22. NoLabor: labor number
23. NoPregna: pregnancy number
24. Occupat: mother's occupation
25. RenFail: renal failure
26. SerolCon: serological conflict
27. Soiling: soiling
28. TORCH: TORCH infection (Toxoplasma gondi, Other viruses -e.g., varicella and parvovirus- and Rubella, Cytomegalo (CMV) and Herpes viruses)
29. Transfus: fetal intrauterine transfusion
30. Twins: Twins
31. UrinDeas: urinary tract diseases
32. USGabnor: abnormalities in USG examine
33. UtCervIn: uterus cervix insufficiency
34. Vomits: vomits
35. WithoutDr: pregnancy without doctors care

According to doctors' knowledge, these variables are supposed to predict premature newborn (when number of weeks of pregnancy at the time of labor HBD is smaller than 27) or a low birth weight baby (with birth weight $<$ 2500 g). So we considered two ways of grouping 508 babies (Tab. 1, Tab. 2).

Table 1. Number of babies in predicting premature newborn

Group	No of babies
0 Premature newborn	440
1 Control	68
Total	508

Table 2. Number of babies in predicting a low birth weight newborn

Group	No of babies
0 Low birth weight baby	442
1 Control	66
Total	508

3. Methods

First we performed the classical parametric (linear and quadratic) discriminant analysis and next the nonparametric Bayes discrimination (nearest neighbor and kernel method) with the parameters (respectively, the number of neighbors k and radius r) chosen to give the smallest leave-one-out classification errors. For all discriminant procedures (Krzanowski 1994, Mc Lachlan 1992, Webb 2002) we used the set of all variables or a subset of selected variables.

We selected a subset of variables that should produce good discrimination, using stepwise mixed forward-backward selection. The criterion was the Wilks' λ. The Wilks' statistic for the overall discrimination is computed as the ratio of the determinant of the within-groups covariance matrix over the determinant of the total covariance matrix:

$$\lambda = \det(W)/\det(T).$$

The F approximation to Wilks' λ is computed following Rao (1951).

We used also the classification trees such as the classical CART method (Classification and Regression Trees, Breiman et al. 1989) and the novel QUEST one (Quick, Unbiased, Efficient Statistical Tree, Loh and Shih 1997).

Both discriminant methods and classification trees divide the multivariate measurement space into disconnected regions. Classification trees may be considered as a generalized discriminant analysis, because many classification tree methods are forms of the recursive discriminant analysis. This is visible especially when in each node the linear discrimination based on more than one variable is performed. QUEST has a quadratic character of splitting one variable in each node.

Classification trees are nonparametric, because they do not assume any underlying family of probability distribution. We haven't got a big dataset, so we cannot divide it into learning and testing sample. Thus, as the measure of goodness-of-fit of classification we use the cross-validation error for all studied methods.

4. Results and discussion

Because the groups differ much in sizes, we used probabilities a priori that are proportional to sizes of groups (both for discriminant analysis and for classification trees). Using equal probabilities a priori gave worse results – as we expected.

The method selected by the stepwise procedure has not necessarily to give the best possible model and the Wilks' λ can be not the best measure of discriminatory power for the application. Though the method is the most appropriate for the multivariate normal distribution with the common covariance matrix, it is also often used for variables not fullfiling this assumption. When we connect the selection of the model with the (medical) knowledge of the data and the leave-one-out procedure for assessing the error, it can be a valuable help.

The results of selection of variables using Wilks' λ statistic are presented in Tables 3 and 4. These tables contain the partial R^2, Wilks' λ, probability for Wilks' λ of smaller values based on the F approximation to Wilks' lambda statistic and the average squared canonical correlation. For two considered discriminant problems, respectively, 8 and 6 variables were chosen. For comparison we now present the values of Wilks' λ for all 35 variables corresponding to discrimination into groups studied in Table 1 ($\lambda = 0.799$, $p < 0.0024$) and corresponding to Table 2 ($\lambda = 0.809$; $p < 0.0054$). So we can see that the subsets of chosen variables have the values of Wilks' λ (equal to 0.0 and 0.866 – see Tables 3 and 4) similar to Wilks'λ statistic for all 35 variables.

Table 3. Summary of the stepwise variables' selection of the most discriminating variables for predicting premature newborn, on the basis of Wilks' λ statistic

Step	Variable entered	Partial R^2	Wilks' Lambda	P value	Average squared canonical correlation
1	Insulin 1	0.0543	0.946	0.0001	0.054
2	Amnicen 2	0.0290	0.918	0.0001	0.082
3	NoLabor 3	0.0167	0.029	0.0001	0.097
4	Vomits 4	0.0161	0.888	0.0001	0.112
5	DiabPreg 5	0.0146	0.875	0.0001	0.125
6	Angina 6	0.0107	0.866	0.0001	0.134
7	Twins 7	0.0107	0.857	0.0001	0.143
8	WithoutDr 8	0.0078	0.850	0.0001	0.150

Table 4. Summary of the stepwise variables' selection of the most discriminating variables for predicting low birth weight babies on the basis of Wilks' λ statistic

Step	Variable entered	Partial R^2	Wilks' Lambda	P value	Average squared canonical correlation
1	Insulin 1	0.0543	0.946	0.0001	0.054
2	NoPregna 2	0.0290	0.918	0.0001	0.082
3	Vomits 3	0.0167	0.029	0.0001	0.097
4	Angina 4	0.0161	0.888	0.0001	0.112
5	Accident 5	0.0146	0.875	0.0001	0.125
6	Diabetes 6	0.0107	0.866	0.0001	0.134

Some variables selected in Tables 3 and 4 are risk factors of considered diseases (pathologies). We used discriminant analysis for all variables and also for the subsets of chosen, the most discriminating variables.

The results of discriminant analysis for two considered problems are summarized in Tables 5 and 6.

We studied also sensitivity (patients properly recognized as healthy – Parmigiani, 2002) and specificity (patients properly recognized as ill). Often increasing one of them may cause decreasing the other. When deciding to choose a method with bigger specificity or sensitivity, we can obtain bigger global classification error, but the aid is then often more reasonable and useful for doctors.

When we chose the classification tree method on the basis of only the global classification error, we got results similar to the discriminant methods: specificity was often very low. So, we used classification trees with different parameters and in different combinations to select the ones that give a small global classification error (cross-validation) and which provide the most sensible aid for doctors. We did not

Table 5. Global cross-validation classification error for predicting premature newborn with specificity and sensitivity for all variables and for selected variables

Method	All variables (35)			Selected variables (8)		
	Specificity	Sensitivity	Global Error	Specificity	Sensitivity	Global Error
Linear	98,0%	60%	0.14	99%	5%	0.13
Quadratic	62,0%	75%	0.36	97%	13%	0.10
Kernel[1]	97,0%	30%	0.14	100%	11%	0.12
Kernel[2]	66.9%	100%	0.28	100%	8%	0.11
Nearest neighbour ($k = 6$)	94,0%	21%	0.16	100%	0%	0.13

k – number of neighbours

[1] normal, radius $r = 0.8$, pooled covariance matrix

[2] normal, radius $r = 0.8$, not pooled covariance matrix

Table 6. Global cross-validation classification error for predicting a low birth weight baby with specificity and sensitivity for all variables and for selected variables

Method	All variables (35)			Selected variables (6)		
	Specificity	Sensitivity	Global Error	Specificity	Sensitivity	Global Error
Linear	97.8%	6.3%	0.14	99.8%	4.9%	0.12
Quadratic	99.7%	4.9%	0.31	45.0%	95.1%	0.48
Kernel[1]	96.7%	34.4%	0.11	99.8%	4.9%	0.13
Kernel[2]	68.5%	100.0%	0.27	33.5%	97.6%	0.58
Nearest neighbour ($k = 6$)	98.2%	25.0%	0.11	100.0%	0%	0.13

k – number of neighbours
[1] normal, radius $r = 0.8$, pooled covariance matrix
[2] normal, radius $r = 0.8$, not pooled covariance matrix

obtain any improvement. The same situation was when for the classification tree procedure we used only variables chosen by Wilks' λ (Tab. 3, 4).

The structure of studied data is such that even one variable can give good percent of global correct classifications. For example, we obtain global classification error equal to 0.11 for the classification tree presented in Figure 1, using only Insulin variable. Specificity is highest (100%), but sensitivity is very unsatisfactory (0.15%).

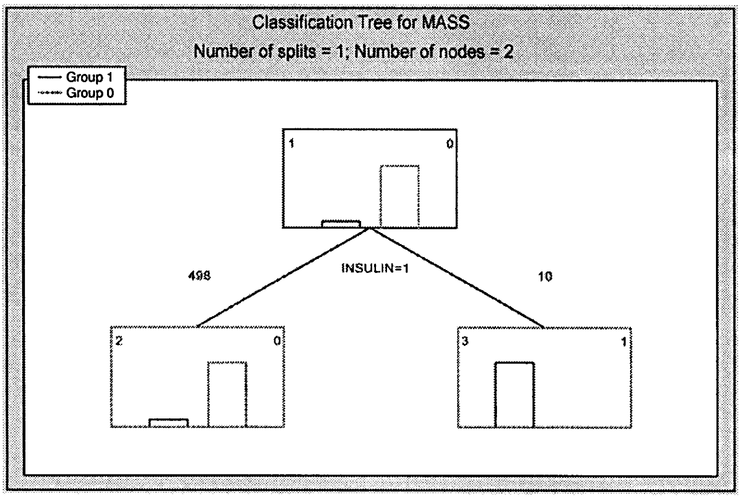
Classification Tree for MASS
Number of splits = 1; Number of nodes = 2
Group 1
Group 0
1
0
498
INSULIN=1
10
2
0
3
1

Figure 1. An example of a classification tree for predicting a low birth weight baby with small global cross-validation error (0.11) and highest specificity(100%) but not sufficient, very low sensitivity (15%).

Therefore, the set of data is not very interesting from the physician point of view, who wants to have a tool to aid the diagnosis. A doctor is interested not only in the global error of classification. For medical applications the specificity and sensitivity are also very important. Such datasets can be rather a material for theoretical study, than the basis of aiding diagnosis after the results of discriminant analysis or classification trees.

5. Concluding remarks

Using many various classification tree methods (and with different parameters) has not clearly improved sensitivity, and results were similar to discriminant analysis.

In aiding medical diagnosis, investigating not only global error of classification, but also sensitivity and specificity is very important. So the percent of correct classifications to each of the discriminated groups should be also considered to choose the best method. In the situation when the global classification error is small, but sensitivity or specificity is nor satisfying, we can consider different actions.

The researcher may define different costs of misclassification to interesting groups (to obtain better classifications for a smaller group and a little worse for a bigger one) or perhaps increasing much smaller groups. However, obtaining bigger groups can be sometimes very difficult for a medical domain of research. One can also consider choosing another set or subset of variables.

Another way could be trying methods from different areas of (statistical) pattern recognition (Webb 2002) or artificial intelligence, eg. neural networks with specially chosen parameters or weights or modified minimized criterion function.

For choosing the procedure it would be recommended to check the correctness of the classification on the basis of the testing sample. However, sometimes difficulties in obtaining satisfying samples for all interesting classification levels can come from the character of data – then finding a good method of classification for aiding diagnosis can be very difficult.

References

Breiman L., Friedman J. H., Olshen R. A., Stone C. H. (1989). *Classification and regression trees.* Wadsworth, Belmont.

Krzanowski W. J. (1994). Quadratic location discriminant functions for mixed categorical and continuous data. *Statistics and Probability Letters* **19**, 91-95.

Mc Lachlan G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition.* J. Wiley & Sons.

Loh W. Y. Shih Y. S. (1997). Split selection methods for classification trees. *Statistica Sinica* **7**, 815-840.

Parmigiani G. (2002). *Modeling in Medical Decision Making: A Bayesian Approach.* Wiley & Sons, LTD.

Rao C.R. (1951). *Linear statistical inference and its applications.* Wiley, New York.

Webb A. R. (2002). *Statistical Pattern Recognition*, 2nd Edition. Wiley & Sons, LTD.

Received 15 January 2003

Wyniki analizy dyskryminacji i drzew klasyfikacyjnych dla danych medycznych o znacznie różniących się liczebnościach grup

STRESZCZENIE

Badano wyniki analizy dyskryminacji i drzew klasyfikacyjnych na podstawie zbioru rzeczywistych danych medycznych, gdy wielkości grup znacznie się różnią. Przyjęliśmy prawdopodobieństwa a priori proporcjonalne do rozmiarów grup oraz oszacowanie błędu klasyfikacji metodą krzyżową (cross-validation) – dla bayesowskich metod parametrycznych i nieparametrycznych dyskryminacji, jak i drzew klasyfikacyjnych. Dla badanych metod otrzymano wysoką swoistość, lecz niską czułość. Aby procedura klasyfikacyjna, otrzymana na podstawie zbioru informacji medycznych, była użyteczna do wspomagania diagnozy, powinna dostarczać nam nie tylko zadowalająco mały łączny błąd klasyfikacji do wszystkich grup, lecz także inne błędy, np. dawać satysfakcjonującą lekarza czułość oraz (lub) specyficzność, w zależności od jego preferencji i charakteru zagadnienia medycznej klasyfikacji. Proponujemy kilka możliwych procedur, aby to osiągnąć.

SŁOWA KLUCZOWE: analiza dyskryminacji, drzewa klasyfikacyjne, krzyżowy błąd klasyfikacji, czułość, swoistość, klasyfikacje medyczne, wcześniactwo, niska waga urodzeniowa.