Information-Theory Analysis of Cell Characteristics in Breast Cancer Patients

David Blokh

C. D. Technologies Ltd., Israel

david_blokh@012.net.il

Abstract

A problem of selecting a subset of parameters containing a maximum amount of information on all parameters of a given set is considered. The proposed method of selection is based on the information-theory analysis and rank statistics. The uncertainty coefficient (normalized mutual information) is used as a measure of information about one parameter contained in another parameter. The most informative characteristics are selected from the set of cytological characteristics of breast cancer patients.

Keywords

Normalized mutual information, Newman-Keuls test, Cytological characteristics, Breast cancer

1. Introduction

The selection of a subgroup of parameters containing a maximum amount of information on all parameters of a given group of parameters is an important problem of medical informatics. This problem has applications in analyzing oncology patients data [1], mathematical modeling of tumor growth [2] and developing intellectual medical systems [3, 4].

In the present article, a subgroup of parameters describing cytological characteristics of breast cancer patients and containing the maximum amount of information on all parameters of this group is selected.

This selection is required for constructing models of cancer diagnosis and prognosis.

The problem under consideration is rather difficult. First, this group includes both discrete and continuous parameters; second, the distributions of continuous parameters are non-Gaussian and, third, the interrelations between parameters are nonlinear.

Any method of selecting information-intensive parameters is based on the use of a measure of parameters correlation. In the majority of methods, a correlation coefficient is used as such a measure. However, the application of the correlation coefficient suggests that parameters distributions are Gaussian, and the correlations of parameters are linear. Therefore, in the present article, the uncertainty coefficient (normalized mutual information) is used as a measure of parameters correlation. This coefficient evaluates nonlinear correlations between parameters with arbitrary distributions.

Thus, this article presents a method of selecting the most informative parameters, which has no restrictions on the distribution of the parameters and on the correlations between the parameters.

Application of the uncertainty coefficient has made it possible to obtain interesting results in medicine [5] and, in particular, in oncology [6, 7, 8]. The approach proposed in [6] is presented in monograph [9].

DOI : 10.5121/ijbb.2013.3101

2. Materials and Methods

2.1. Materials

To illustrate the method, we used data from the Wisconsin Breast Cancer Database [10]. These data cover 663 patients, each patient being presented by 10 cytological parameters, 9 of which are continuous (clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitoses) and one is discrete (Class).

2.2. Data Analysis

2.2.1. Statement of the problem

Assume that the initial data on ***n*** objects are presented in the form of a $n \times m$ array $[a_{kj}]$, where each row ***k*** is an object described by ***m*** parameters. It is needed to find a parameter or a subgroup of parameters containing the greatest amount of information about all ***m*** parameters.

2.2.2. Description of the algorithm

The algorithm of selecting a subgroup of the most informative parameters from the entire group of parameters includes four procedures. A short description of each procedure is as follows. A more complete description of the application of information-theory analysis to the selection problem is presented in [11].

1. Discretization.

This procedure transforms parameters having continuous values into parameters having discrete values.

If an acceptable method of a continuous parameter discretization is unavailable, use a formal approach to the discretization [12].

2. Construction of the uncertainty coefficient matrix.

For *i*-th and *j*-th parameters $1 \le i, j \le m$, calculate the uncertainty coefficient c_{ij} [5] and construct $m \times m$ uncertainty coefficient matrix $[c_{ij}]$.

3. Construction of the rank matrix.

For each column of the matrix $[c_{ij}]$, we rank its elements and assign rank 1 to the smallest element of the column. We obtain the matrix $m \times m$ of ranks $[r_{ij}]$, where each column of the matrix contains ranks from 1 to m.

We estimate the amount of information about all m parameters contained in the i-th parameter by the sum of all the entries of the i-th row of the matrix $[r_{ij}]$.

4. Application of the multiple comparison method.

Apply the multiple comparison method to the sums of $[r_{ij}]$ matrix rows [13]. This gives a clustering of parameters that contains the desired subgroup of parameters.

Remark. Programs for the algorithm of uncertainty coefficient calculation and the Friedman test are implemented in the package SPSS [14].

3. Results

We consecutively perform the four procedures of the selection algorithm.

1. Categories of parameters contained in the Wisconsin Breast Cancer Database are given below.

Parameter name	Category
1. Clump Thickness:	1,2,3,4,5,6,7,8,9,10.
2. Uniformity of Cell Size:	1,2,3,4,5,6,7,8,9,10.
3. Uniformity of Cell Shape:	1,2,3,4,5,6,7,8,9,10.
4. Marginal Adhesion:	1,2,3,4,5,6,7,8,9,10.
5. Single epithelial cell size	1,2,3,4,5,6,7,8,9,10.
6. Bare Nuclei:	1,2,3,4,5,6,7,8,9,10.
7. Bland Chromatin:	1,2,3,4,5,6,7,8,9,10.
8. Normal Nucleoli:	1,2,3,4,5,6,7,8,9,10.
9. Mitose:	1,2,3,4,5,6,7,8,9,10.
10. Class:	2-benign, 4-malignant

2. Calculate the uncertainty coefficients c_{ij} $1 \le i, j \le 10$ for the parameters of the Wisconsin Breast Cancer Database and construct the $[c_{ij}]$ matrix of uncertainty coefficients (Table 1).

Table 1. Uncertainty coefficient matrix $[c_{ij}]$.

No	Parameter name	1	2	3	4	5	6	7	8	9	10
1	Clump thickness	1	.222	.194	.144	.170	.205	.136	.180	.171	.439
2	Uniformity of cell size	.181	1	.406	.263	.305	.319	.243	.301	.251	.676
3	Uniformity of cell shape	.165	.419	1	.238	.250	.282	.211	.266	.209	.556
4	Marginal adhesion	.105	.232	.205	1	.211	.239	.198	.196	.175	.382
5	Single epithelial cell size	.135	.295	.235	.231	1	.240	.191	.256	.255	.503
6	Bare nuclei	.137	.259	.223	.219	.202	1	.197	.222	.154	.491
7	Bland chromatin	.123	.267	.226	.247	.218	.268	1	.245	.155	.499
8	Normal nucleoli	.121	.245	.211	.181	.216	.223	.181	1	.180	.386
9	Mitoses	.071	.125	.103	.100	.133	.096	.071	.111	1	.226
10	Class	.180	.339	.269	.215	.259	.301	.225	.236	.223	1

3. Rank the $[c_{ij}]$ matrix (Table 1) columns and obtain a rank matrix $[r_{ij}]$ (Table 2).

Table 2. Rank matrix $[r_{ij}]$.

No	Parameter name	1	2	3	4	5	6	7	8	9	10	Sum of ranks
1	Clump thickness	10	2	2	2	2	2	2	2	3	4	31
2	Uniformity of cell size	9	10	9	9	9	9	9	9	8	9	90
3	Uniformity of cell shape	7	9	10	7	7	7	7	8	6	8	76
4	Marginal adhesion	2	3	3	10	4	4	6	3	4	2	41
5	Single epithelial cell size	5	7	7	6	10	5	4	7	9	7	67
6	Bare nuclei	6	5	5	5	3	10	5	4	1	5	49
7	Bland chromatin	4	6	6	8	6	6	10	6	2	6	60
8	Normal nucleoli	3	4	4	3	5	3	3	10	5	3	43
9	Mitoses	1	1	1	1	1	1	1	1	10	1	19
10	Class	8	8	8	4	8	8	8	5	7	10	74

Consider Table 2 as the Friedman statistical model [15] and examine the row effect of this table.

Hypotheses

H_0: There is no row effect ("null hypothesis").

H_1: The null hypothesis is invalid.

Critical range. The sample is "large"; therefore, the critical range is the upper 1%-range of χ^2_9 distribution.

Calculation of the χ^2-criterion [15] gives $\chi^2 = 48.48$.

The critical range is $\chi^2_9 > 21.67$. Since $48.48 > 21.67$, the null hypothesis with respect to Table 2 is rejected. Thus, according to the Friedman test, the row effect exists. Hence, there is a difference between the rows under consideration.

4. For multiple comparisons, we use the Newman-Keuls test [13]. We obtain $\left|R_j - R_{j+1}\right| > 8.15$,

where R_j and R_{j+1} are the j-th and $(j+1)$-th elements of "Sum of ranks" column of Table 2. Using the multiple comparisons method, we construct the parameter clustering shown in Table 3.

Table 3. Parameter Clustering.

No	Cluster	Parameter name	Sum of ranks
1	Cluster 1	Uniformity of cell size	90
2	Cluster 2	Uniformity of cell shape	76
3		Class "benign" or "malignant"	74
4		Single epithelial cell size	67
5		Bland chromatin	60
6	Cluster 3	Bare nuclei	49
7		Normal nucleoli	43
8		Marginal adhesion	41
9	Cluster 4	Clump thickness	31
10	Cluster 5	Mitoses	19

The obtained clustering has the following properties:

For two neighboring clusters of Table 3, the smallest element of one cluster and the greatest element of another cluster located nearby are significantly different (α_T=0.01);

Elements belonging to the same cluster do not differ from each other (α_T=0.01);

The differences between Cluster 1 (Uniformity of cell size) and all the characteristics are statistically significant (α_T=0.01).

The characteristic "Uniformity of cell size" contains the greatest amount of information about all the characteristics under study. It corresponds to the thesis that the uniformity of cell size in unicellular organisms and within tissues in multicellular organisms could be necessary to maintain the homeostasis of transcription and proliferation [16]. Thus, an adequate result has been obtained using the proposed algorithm.
We believe that the processing other biomedical data using the proposed algorithm will also give adequate results.

4. Conclusions

In the present paper, a method of selecting parameters based on the assessment of correlations between single parameters is considered. However, problems arising in medical practice often require the selection of parameters taking into account correlations between groups of parameters. The development of methods of selecting informative parameters taking into account not only correlations between single parameters, but also correlations between groups of parameters, should be a subject of further research.

References

1. G. S. Atwal, R. Rabadan, G. Lozano et al., (2008) "An Information-Theoretic Analysis of Genetics, Gender and Age in Cancer Patients", *PLoS ONE*, vol. 3, no. 4, 7 p.
2. R. Molina-Pena, M. M. Alvarez, (2012) "A Simple Mathematical Model Based on the Cancer Stem Cell Hypothesis Suggests Kinetic Commonalities in Solid Tumor Growth", *PLoS ONE*, vol. 7, no. 2, 11 p.
3. S. M. Anzar, P. S. Sathidevi, (2012) "An Efficient PSO Optimized Integration Weight Estimation Using D-Prime Statistics for a Multibiometric System", *International Journal on Bioinformatics & Biosciences*, vol. 2, no. 3, pp. 31-42.
4. P. Preckova, J. Zvarova and K. Zvara, (2012) "Measuring diversity in medical reports based on categorized attributes and international classification systems", *BMC Medical Informatics and Decision Making*, vol. 12, 11 p.
5. J. Zvarova, M. Studeny, (1997) "Information theoretical approach to constitution and reduction of medical data", *International Journal of Medical Informatics*, vol. 45, no. 1-2, pp. 65-74.
6. D. Blokh, I. Stambler, E. Afrimzon, et al., (2007) "The information-theory analysis of Michaelis-Menten constants for detection of breast cancer", *Cancer Detection and Prevention*, vol. 31, no. 6, pp. 489-498.
7. D. Blokh, N. Zurgil, I. Stambler, et al., (2008) "An information-theoretical model for breast cancer detection", *Methods of Information in Medicine*, vol. 47, no. 4, pp. 322-327.
8. D. Blokh, I. Stambler, E. Afrimzon, et al., (2009) "Comparative analysis of cell parameter groups for breast cancer detection", *Computer Methods and Programs in Biomedicine*, vol. 94, no. 3, pp. 239-249.
9. P. J. Gutierrez Diez, I.H. Russo, J. Russo, (2012) *The Evolution of the Use of Mathematics in Cancer Research*, Springer, New York.
10. W.H. Wolberg, O.L. Mangasarian, (1990) "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", *Proceedings of the National Academy of Sciences*, Vol. 87, No. 23, pp. 9193-9196.
11. D. Blokh, (2012) "Clustering financial time series via information-theory analysis and rank statistics", *Journal of Pattern Recognition Research*, vol. 7, no. 1, pp. 106-115.
12. G.V. Glass, J.C. Stanley, (1970) *Statistical Methods in Education and Psychology*, Prentice-Hall, New Jersey.
13. S.A. Glantz, (1994) *Primer of Biostatistics*, 4th ed., McGraw-Hill, New York.
14. A. Buhl, P. Zofel, (2001) *SPSS Version 10*, Addison-Wesley, Munchen.
15. W.J. Conover, (1999) *Practical Nonparametric Statistics.* Wiley-Interscience, New York.
16. C.-Y. Wu, P.A. Rolfe, D.K. Gifford, G.R. Fink, (2010) "Control of Transcription by Cell Size", *PLoS Biology*, Vol. 8, No. 11, 16 p.