

Intelligent Disease Identification based on Discriminant Analysis of Clinical Data

COSTEL SĂRBU^{1*}, HORIA F. POP², RALUCA-³TEFANIA ELEKES¹, GEORGETA COVAC³

¹Babe⁻Bolyai University, Faculty of Chemistry and Chemical Engineering, 11 Arany Janos Str., 400028, Cluj-Napoca, Romania

²Babes-Bolyai University, Faculty of Mathematics and Computer Science, 1 M. Kogalniceanu Str., 400084, Cluj-Napoca, Romania

³Institutul Oncologic "Ion Chiricuta" Cluj-Napoca, 34-36 Republicii Str., 400015, Cluj-Napoca, Romania

Discriminant analysis was applied as a method of disease identification, using data obtained from blood analysis of several patients. The investigated compounds in human blood samples were organic compounds of clinical interest (glucose, triglycerides, cholesterol, creatinine and urea), inorganic compounds (Na, K, Ca, Mg and Fe) and enzymes (Lactate Dehydrogenase (LDH), Alanine Transaminase (ALT), Aspartate Aminotransferase (AST), Alkaline Phosphatase (ALP) and Gamma Glutamyltransferase (GGT)). According to their concentration level the following diseases have been selected for study: hydroelectric disorders, hepatic diseases, lipid disorders, diabetes and renal disorders. Some patients resulted to be healthy. Discriminant analysis was not only used for classifying the patients according to their disease but also for detecting the most important variables that discriminate between the groups. For example it has been found that the greatest contribution to the discriminatory power of the model is given by glucose ($\lambda^ = 0.263$; $F = 44$). The obtained results confirm that clinical analysis combined with the multidimensional interpretation of data gives an interesting and very useful way of disease correlations, interpretations, problem solving and cost effectiveness.*

Keywords: discriminant analysis, clinical data, disease identification, chemical compounds

Modern science and techniques have revolutionized the possibility of rational and objective diagnosis and treating of all kinds of diseases. Together with ultrasonic-, X-ray-, and nuclear magnetic resonance-based methods, clinical laboratory tests provide a sensitive and objective indicator of the patient's condition. Since laboratory tests provide quantitative, reproducible and specific results they are, and must be, an integral part of diagnosis, therapy control and management of a patient disease [1-7].

The primary goal of a clinical chemistry laboratory is to correctly perform analytical procedures that yield accurate and precise information to aid in patient diagnosis. In order to achieve reliable results, the laboratory must include the ability to use basic supplies and equipment correctly and an understanding of fundamental concepts critical to any chemical test [8].

The volume of data generated by the clinical chemistry laboratory is enormous and must be summarized to be most useful to the analyst and clinician. Consequently, the use of multivariate chemometrical methods becomes extremely necessary as a way to handle the very large and complex data sets [9-11]. Principal Components Analysis [12, 13], Cluster Analysis [14] and Discriminant Analysis [14-18] are the most applied multivariate methods for data processing and maximum information extraction.

In this paper discriminant analysis was applied as a method of disease identification, using data obtained from blood analysis of several patients. The investigated compounds in human blood samples were organic compounds of clinical interest (glucose, triglycerides, cholesterol, creatinine and urea), inorganic compounds (Na, K, Ca, Mg and Fe) and enzymes (Lactate Dehydrogenase (LDH), Alanine Transaminase (ALT), Aspartate Aminotransferase (AST), Alkaline Phosphatase

(ALP) and Gamma Glutamyltransferase (GGT)). According to their concentration level the following diseases have been selected for study: hydroelectric disorders, hepatic diseases, lipid disorders, diabetes and renal disorders. Some patients resulted to be healthy.

Discriminant function analysis

Discriminant function analysis or simply, discriminant analysis (DA) is based on the extraction of linear discriminant functions of the independent variables by means of a qualitative dependent variables and several quantitative independent variables [9, 10, 15, 16]. DA can be formulated as follows: let $\mathbf{X} = \{x^1, \dots, x^n\} \subset \mathbf{R}^p$ be a finite set of characteristic vectors, where n is the number of samples (measurements) and p is the number of the original variables (predictors), $\mathbf{x}^i = [x^1, x^2, \dots, x^p]^T$ and y be a nominal characteristic (grouping variable), with k values, each of which characterizes one of the k partition composing the partition substructure of the given data set. The partition of \mathbf{X} into k groups is computationally very similar to analysis of variance (ANOVA/MANOVA), sharing many of the same assumptions and tests; the most important variables are selected, and variables contributing only marginally to the discrimination of groups will be removed. In a similar way as with principal component analysis [9-13], first the total variance/covariance matrix is calculated according to the following expression

$$\mathbf{V} = {}^T\mathbf{XDX}, \quad (1)$$

where \mathbf{X} is the centered data matrix, ${}^T\mathbf{X}$ is the transpose matrix, \mathbf{D} is the diagonal matrix (in most cases is the unity matrix).

Considering a new characteristic defined as $\mathbf{c} = \mathbf{Xu}$, one can calculate its variance by applying the relation (2)

* Tel.: (+40) 264 59 38 77

$$\|c\|^2 = {}^T c D c = {}^T u^T X D X u = {}^T u V u, \quad (2)$$

The total variance V may be decomposed into two components: the between-group variance B and within-group variance W , namely

$$V = B + W, \quad (3)$$

and, as a consequence, the variance of the characteristic c becomes

$$\|c\|^2 = {}^T u V u = {}^T u B u + {}^T u W u. \quad (4)$$

In this case, it is very easy to observe that equation (4) can be rewritten in the following form

$$\frac{{}^T u B u}{{}^T u V u} + \frac{{}^T u W u}{{}^T u V u} = 1, \quad (5)$$

and because any term from the left side is positive, equivalent results will be obtained indifferent of the maximum/minimum condition.

However, in practice the first ratio in equation (5) is maximized

$$\lambda = \frac{{}^T u B u}{{}^T u V u} \quad (0 \leq \lambda < 1) \quad (6)$$

or, in a different form, of a generalized eigenvalue problem:

$$B u = \lambda V u, \quad (7)$$

Let us recall that the matrix V of the total variance is symmetrical and positive definite. As such, this equation may be rewritten to a matrix equation similar to that obtained in the case principal component analysis results

$$V^{-1} B u = \lambda u, \quad (8)$$

where λ and u represent the eigenvalues (known, as well, as characteristic roots) and eigenvectors of the matrix $V^{-1} B$. The vector u^1 , named the first discriminant factor corresponds to the highest value of λ ; the higher this value the higher will be the discriminant power of this factor. After obtaining the first discriminant characteristic $c_1 = X u^1$, in a similar way can be obtained the discriminant characteristic $c_2 = X u^2$, uncorrelated with the first and so on. It appears clearly that eigenvectors corresponding to the matrix $V^{-1} B$ namely u^1, u^2, \dots, u^{k-1} , ranked in decreasing order of the positive values $\lambda_1, \dots, \lambda_2, \dots, \lambda_{k-1}$, are successive solutions of the above matrix equation.

If the vector of the discriminant function is $u = (u_1, \dots, u_2, \dots, u_p)$, then the projection of sample x^i on this axis represents the distance to the origin:

$$c_i = x^i_1 u_1 + x^i_2 u_2 + \dots + x^i_p u_p. \quad (9)$$

The vectors u are called the discriminant factors and the vectors c represent the discriminant scores. The linear function described by equation (9) is called discrimination function.

Finally, we have to emphasize that even if the power of discrimination does not depend on standardization of data, generally standardized data are used.

The quality of discrimination and the selection of the most discriminant independent variables can be evaluated by applying different criteria. The Wilks' lambda F test is

used to test whether the discriminant model as a whole is significant; the larger the lambda, the more likely it is significant. In the same order can be used λ^* statistic defined by the equation 10.

$$\lambda^* = \frac{{}^T u W u}{{}^T u V u} = 1 - \lambda \quad (0 \leq \lambda < 1) \quad (10)$$

The smaller the value of λ^* , the more the model is discriminating.

Concerning the contribution of the independent variables to the discrimination of groups, this can be appreciated either by the assay of the classes homogeneity using statistic F like in the case of ANOVA/MANOVA method, either by using Wilks' lambda for each variable. Wilks' lambda is the standard statistic used to express the significance of the overall discriminatory power of the variables in the model. A value of 1 indicates absolutely no discriminatory power, whereas 0 indicates a perfect discriminatory power. The partial Wilks' lambda describes the unique contribution of each variable to the discriminatory power of the model. The closer the partial lambda is to 0, the better the discriminatory force of the variable. In addition, the tolerance value gives information to the redundancy of the respective variable in the model, and is computed as 1 minus R-squares of the respective variable, with all other variables included in the model. Put in other words, it is the proportion of the variance contributed by respective variable. If variable is completely redundant, the squared tolerance value approaches zero.

This kind of information can be obtained from value of the discriminant coefficients associated to the p descriptive variables, and also from the correlation coefficients between each variable and the vector score. The larger the discriminant coefficient and the closer to 1 the correlation coefficient is, the larger the variable importance for the samples separation in defined groups is. Also, the standardized discriminant coefficients, like beta weights in regression, are used to assess the relative classifying importance of the independent variables.

Experimental part

The studied parameters have been analyzed with spectrophotometric and electrochemical methods using the Johnson & Johnson complex chemical system.

Sodium and potassium have been measured through a *potentiometric method* using an *ion-selective electrode*. *Atomic Absorption Spectrometry* was used to measure the concentrations of Ca, Mg and Fe. The *Automatic Clinical Analyzer* uses complexometric titration with methylthymol blue and a chelating agent to remove Ca interference in its Mg analysis. The concentration of Ca has been determined using the same principle as for Mg. Many automated methods for total calcium are based on the complexometric reaction between Ca and ortho-cresolphthalein complexone, often with 8-hydroxyquinoline added to prevent Mg interference.

Molecular Absorption Spectrometry was used to measure the concentrations of glucose, urea, creatinine, cholesterol, triglycerides, and enzymes.

Results and Discussions

Discriminant Analysis was applied as a method of disease identification using data obtained from blood analysis of several patients. The compounds determined and used to establish the diagnosis are presented in Table 1, together with the normal range of their concentrations in the human blood.

Table 1
NORMAL VALUES OF THE STUDIED COMPOUNDS
IN HUMAN BLOOD

Analyte	Normal range	
	Male	Female
Glucose	75 - 110 mg/dL	65 - 105 mg/dL
Triglycerides	< 150 mg/dL	
Cholesterol	< 200 mg/dL	
Creatinine	0.8 - 1.5 mg/dL	0.7 - 1.2 mg/dL
Urea	7-17 mg/dL	9 - 20 mg/dL
Sodium	137-145 mmol/L	
Potassium	3.6 - 5.0 mmol/L	
Calcium	8.4 - 10.2 mg/dL	
Magnesium	1.6 - 2.3 mg/dL	
Iron	49 - 181 μ g/dL	37 - 170 μ g/dL
AST	15 - 46 U/L	
ALT	11 - 66 U/L	
LDH	313 - 618 U/L	
ALP	38 - 126 U/L	
GGT	8 - 78 U/L	

The training data set consisted of 100 patients and the following 16 characteristics: age, Na, K, Ca, Mg, Fe, glucose, creatinine, urea, cholesterol, triglycerides, ALT, AST, LDH, GGT, ALP. Statistical information on the 16 measured variables is presented in table 2.

The set of 100 studied cases was distributed as follows: 20 are healthy (marked 's'), 20 have lipid disorders ('l'), 20 hepatic diseases ('h'), 20 hydroelectric disorders ('d'), 10 diabetes ('z'), 10 renal disorders ('r').

After application of the standard DA to the data matrix, the variables presented in table 3 were retained in the model. The statistics from this table illustrates the contribution to the discrimination of the components present in human blood according to different parameters.

It is easy to observe that the greatest contribution is given by glucose ($\lambda^* = 0.261$; $F = 44.569$). The next highest are creatinine ($\lambda^* = 0.568$; $F = 11.984$) and triglycerides ($\lambda^* = 0.796$; $F = 4.040$). The smallest contribution was obtained for Mg ($\lambda^* = 0.985$; $F = 0.239$). Also a small contribution brings the age of the patients ($\lambda^* = 0.967$; $F = 0.530$).

A *canonical correlation* analysis has been done, that determined the successive functions and canonical roots. The maximum number of functions will be equal to the

Table 2
STATISTICS OF THE MEASURED VARIABLES

	Mean	Median	Min.	Max.	Range	SD	Skewness	Kurtosis
Age	53.74	56.50	6.00	87.00	81.00	19.20	-0.80	0.21
Na	140.32	140.00	84.00	184.00	100.00	14.51	-0.94	5.23
K	4.27	4.15	2.30	7.50	5.20	0.85	0.61	1.26
Ca	8.86	8.85	6.50	11.00	4.50	0.83	-0.06	0.34
Mg	1.91	1.90	0.90	3.10	2.20	0.40	0.15	1.61
Fe	97.05	87.50	17.00	197.00	180.00	51.79	0.30	-1.18
Glucose	107.15	95.00	70.00	350.00	280.00	48.38	3.48	12.41
Creatinine	1.33	1.00	0.70	6.00	5.30	0.97	3.29	11.05
Urea	16.09	15.00	8.00	53.00	45.00	7.80	2.44	7.41
Cholesterol	164.46	130.00	63.00	500.00	437.00	98.21	1.47	1.56
Triglyceride	142.85	105.00	45.00	478.00	433.00	93.87	1.67	2.06
ALT	59.63	44.00	10.00	520.00	510.00	63.21	4.72	29.21
AST	52.15	35.00	15.00	415.00	400.00	57.06	3.70	17.37
LDH	486.37	474.00	315.00	1050.00	735.00	133.53	1.34	3.27
GGT	57.34	54.00	9.00	304.00	295.00	43.84	2.87	12.06
ALP	86.61	75.00	21.00	314.00	293.00	47.63	2.56	9.62

Table 3
DISCRIMINANT ANALYSIS

Variable	Wilks' λ	Partial λ^*	F-remove	p-level	Toler.	1-Toler.
Age	0.0016	0.9675	0.5303	0.7527	0.8372	0.1628
Na	0.0016	0.9530	0.7793	0.5676	0.4427	0.5573
K	0.0018	0.8367	3.0836	0.0135	0.6180	0.3820
Ca	0.0019	0.7775	4.5207	0.0011	0.7798	0.2202
Mg	0.0015	0.9851	0.2393	0.9439	0.5084	0.4916
Fe	0.0017	0.8993	1.7684	0.1290	0.7627	0.2372
Glucose	0.0058	0.2617	44.5690	0.0000	0.9103	0.0897
Creatinine	0.0027	0.5687	11.9839	0.0000	0.5423	0.4577
Urea	0.0016	0.9406	0.9985	0.4243	0.6472	0.3528
Cholesterol	0.0017	0.8857	2.0394	0.0820	0.3859	0.6141
Triglycerides	0.0019	0.7963	4.0406	0.0026	0.4028	0.5972
ALT	0.0016	0.9475	0.8756	0.5014	0.1147	0.8853
AST	0.0018	0.8401	3.0072	0.0155	0.1150	0.8850
LDH	0.0017	0.9011	1.7346	0.1364	0.8119	0.1881
GGT	0.0016	0.9212	1.3518	0.2515	0.7222	0.2778
ALP	0.0016	0.9210	1.3557	0.2501	0.7695	0.2305

number of groups minus one, or the number of variables in the analysis, whichever is smaller. The eigenvalues (characteristic roots) and the corresponding standardized canonical discriminant function coefficients are showed in table 4.

Table 4
STANDARDIZED COEFFICIENTS FOR CANONICAL VARIABLES

Variable	Root 1	Root 2	Root 3	Root 4	Root 5
Age	0.0738	-0.0980	-0.0142	-0.1767	0.3319
Na	0.0146	-0.3436	0.0343	-0.0883	0.0041
K	-0.2855	0.4399	-0.1582	-0.1279	-0.1913
Ca	0.1713	0.3209	0.1673	0.3296	0.5833
Mg	0.1075	-0.0704	-0.0551	0.1262	0.0653
Fe	0.3612	-0.0588	-0.0973	0.1093	0.2353
Glucose	0.2023	-0.1379	-0.9112	0.3963	-0.1686
Creatinine	-0.6211	0.6562	-0.0349	0.4098	-0.1966
Urea	-0.1406	0.2851	-0.0358	-0.0851	0.0853
Cholesterol	0.3964	0.1270	-0.0404	-0.0051	-0.9208
Triglyceride	0.4523	0.4676	0.2135	0.2948	0.5992
ALT	-0.1129	0.6462	-0.1706	-0.3458	-0.0661
AST	-0.0571	-0.9817	0.5408	0.7718	-0.0916
LDH	-0.0830	-0.0954	0.3538	0.0417	-0.0694
GGT	-0.0714	-0.2289	0.0748	0.3174	0.2268
ALP	-0.0667	-0.1560	0.1174	0.2893	-0.3645
Eigenvalue	6.3514	5.4422	3.8964	1.2905	0.1864
Cumulative Proportion	0.3700	0.6870	0.9140	0.9891	1.0000

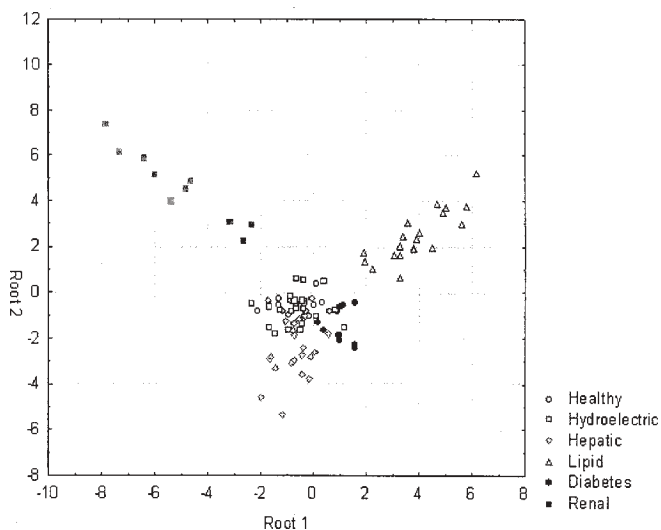


Fig. 1. Scatterplot of canonical scores on the plan described by root 1 and root 2

The first function presented a relatively high eigenvalue (6.351). The eigenvalue drops to 5.442 for the second axis, and further to 3.896 for the third axis.

The highest standardized discriminant coefficients correspond to creatinine (0.621), triglycerides (0.452), cholesterol (0.396), Fe (0.36), K (0.285) in root 1; to AST (0.981), creatinine (0.656), ALT (0.646), triglycerides (0.467) and K (0.439) in root 2. In root 3 the highest values are for glucose (0.911), AST (0.54), LDH (0.353) and again triglycerides (0.213). In root 4 coefficients corresponding to AST (0.771), creatinine (0.409), Ca (0.329) and

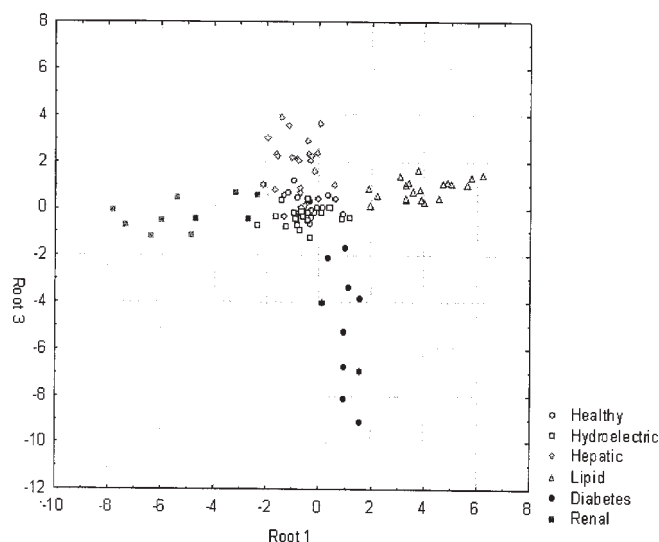


Fig. 2. Scatterplot of canonical scores on the plan described by root 1 and root 3

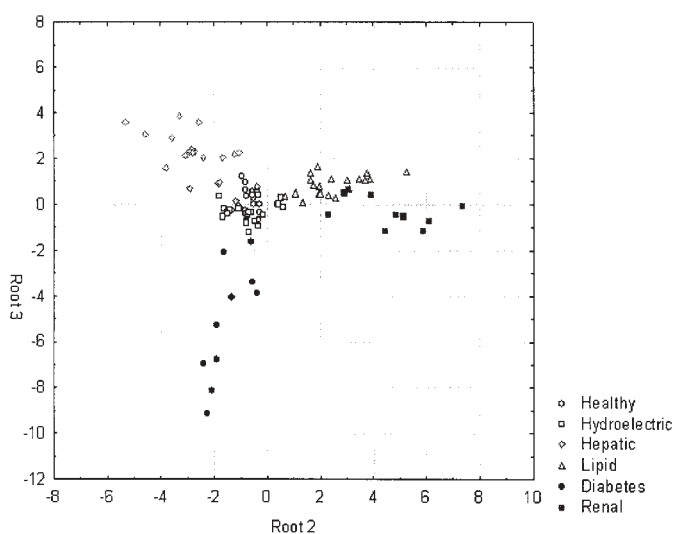


Fig. 3. Scatterplot of canonical scores on the plan described by root 2 and root 3

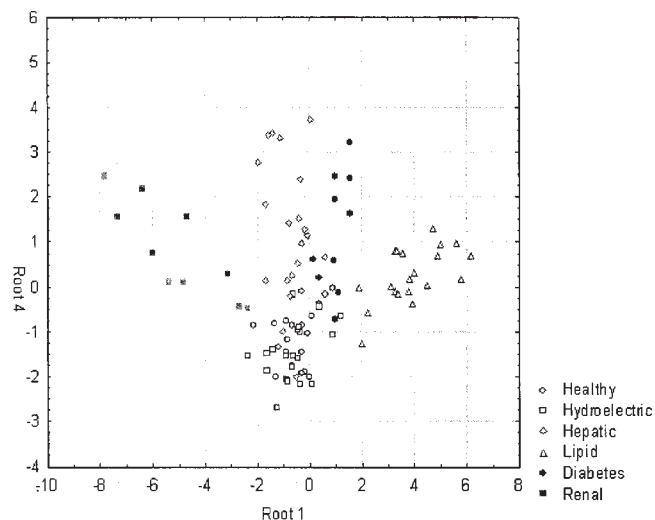


Fig. 4. Scatterplot of canonical scores on the plan described by root 1 and root 4

triglycerides (0.29) are the highest. In root 5 coefficients for cholesterol (0.920), triglycerides (0.599), Ca (0.583), age (0.331), Fe (0.235) are the highest. It can be observed that triglycerides have a major contribution in all roots.

We can also visualize how the functions discriminate between groups by plotting the individual scores for the

discriminant functions. This statement is well supported by the two-dimensional scatterplot using the discriminant scores of the samples along root 1, root 2 and root 3 as one can be seen in the following figures.

A few remarks are in order. We notice an excellent discriminating power for the first root. A projection of all the data on the axis of root 1 shows a clear separation of the Renal and Lipid classes, with all the others grouped together. Similarly, root 2 discriminates between (a) Hepatic, (b) Lipid and Renal, and (c) the remaining classes. Root 3 shows a clear separation of the Diabetes class, a good separation of Hepatic class, with all the others grouped together.

On the other side, combinations of two roots discriminate better than the individual roots by themselves. As such, the projection on roots 1 and 3 shows a clear separation of four of the six classes (Renal, Lipid, Diabetes and Hepatic), with a good separation of the remaining two classes (Healthy and Hydroelectric).

On a different perspective, the apparent intertwining of the classes Healthy and Hydroelectric is consistent with the remark that hydroelectric disorders are found on a regular basis to people considered as generally healthy.

The classification result of this procedure is the *classification matrix*, which shows the number of cases that were correctly classified (on the diagonal of the matrix) and those that were misclassified.

The classification matrix presented in table 5 indicates a satisfactory separation of patients in a good agreement to their origin. The group having lipid disorders, for example, showed that 100% of patients were very well classified. Also the group having renal disorders showed a good separation (90%). The poorest classification was obtained for those having hydroelectric disorders (70%).

Table 5

CLASSIFICATION MATRIX (ROWS: OBSERVED CLASSIFICATIONS; COLUMNS: PREDICTED CLASSIFICATIONS)

Class	Percent	s	d	h	l	z	r
s	90	18	2	0	0	0	0
d	70	6	14	0	0	0	0
h	80	3	1	16	0	0	0
l	100	0	0	0	20	0	0
z	80	1	1	0	0	8	0
r	90	1	0	0	0	0	9
Total	85	29	18	16	20	8	9

Other representative results can be obtained from the *cases classification table*, which describes group membership of the cases.

Conclusions

The obtained results confirm that clinical analysis combined with the multidimensional interpretation of data

gives an interesting and very useful way of disease correlations, interpretations, problem solving and cost effectiveness.

It is interesting to point out the role isolated data plays with the method. A future study should be aimed at robust methods of Discriminant Analysis, targeted around weighted contributions of data samples, possibly through the use of the fuzzy sets theory.

Acknowledgement: This research has been supported by the Romanian National Council for Scientific Research in Higher Education (CNCSIS) through the PNI-IDEI research grant ID_550/2007.

References

- HENRY, J.B., Clinical Diagnosis and Management by Laboratory Methods, 18th ed., W.B. Saunders Company, Philadelphia, 1991
- BISHOP, M.L., DUBEN-ENGELKIRK, J.L., FODY, E.P. Clinical Chemistry. Principles, Procedures, Correlations, 5th ed, Lippincott Williams & Wilkins, Philadelphia, 2004
- MAYNE, P., MAYNE, P.D., Clinical Chemistry in Diagnosis and Treatment, 6th ed, A Hodder Arnold Publication, London, 1994
- KAPLAN, A., PESCE, A.J., KAZMIERCZAK, S.C., Clinical Chemistry: Theory, Analysis and Correlation, 3rd ed, Amer Assn for Clinical Chemistry, Washington, DC, 1997
- COWAN, D., Informatics for the Clinical Laboratory: A Practical Guide (Health Informatics) (Kindle Edition), 1st ed., Springer Verlag, Berlin, 2002
- BERG, D., Advanced Clinical Skills and Physical Diagnosis, 2nd ed., Wiley-Blackwell, Boston, 2004
- MIKKELSEN, S.R., CORTÓN, E. Bioanalytical Chemistry. Inc., John Wiley & Sons, 2004, p. 16
- SCHLEICHER, E., Anal. Bioanal. Chem., **384**, 2006, p. 124
- MANLY, B.F.J., Multivariate statistical methods, Chapman and Hall, London, 1986
- BRERETON, R.G. Applied Chemometrics for Scientists, John Wiley & Sons, Chichester, 2007
- SĂRBU, C., POP, H.F., Fuzzy soft-computing methods and their applications in chemistry, In Lipkowitz, K.B., Larter, R., Cundari, T.R., eds. Reviews in Computational Chemistry, Vol. 20, VCH Publishers, New York, 2004, p. 249
- WOLD, S., ESBENSEN, K., GELADI, P.Q., Chem. Intell. Lab. Syst., **15**, 1987, p. 37
- SĂRBU, C., POP, H.F., Talanta, **65**, 2005, p. 1215
- JAJUGA, K., SOKOŁOWSKI, A., BOCK, H.H., Classification, Clustering and Data Analysis, 1st ed., Springer Verlag, Berlin, 2002
- SCARONI, G., MORET, I., CAPODAGLIO, G., CESCO, P., J. Agric. Food. Chem., **30**, 1982, p. 1135
- DIÁZ-FLORES, J.F., DIÁZ-FLORES, E.F., HERNÁNDEZ CALZADILLA, C., RODRÍGUEZ, E.M., DIÁZ, R.C., SERRA-MAJEM, L., Eur. J. Clin. Nutr., **58**, 2004, p. 449
- KAZMIERCZAK, S.C., GURACHEVSKY, A., MATTHES, G., MURAVSKY, V., Clin. Chem., **52**, 2006, p. 2129
- LI, Q.B., SUN, X.J., XU, Y.Z., YANG, L.M., ZHANG, Y.F., WENG, S.F., SHI, J.S., WU, J.G., Clin. Chem., **51**, 2005, p. 346

Manuscript received: 20.08.2008