BioMedical Engineering OnLine

BioMed Central

Research

Open Access

Intelligent data analysis to interpret major risk factors for diabetic patients with and without ischemic stroke in a small population

Fikret Gürgen*[1] and Nurgül Gürgen[2]

Address: [1]Dept of Computer Eng., Boaziçi University TR-80815 Bebek-Istanbul/TURKEY and [2]Neurology Division, Lütfiye Nuri Burat Hospital Sultançiftlii Istanbul/TURKEY

Email: Fikret Gürgen* - gurgen@boun.edu.tr; Nurgül Gürgen - gurgen@boun.edu.tr

* Corresponding author

Published: 4 March 2003

BioMedical Engineering OnLine 2003, **2**:5

Received: 20 November 2002

Accepted: 4 March 2003

This article is available from: http://www.biomedical-engineering-online.com/content/2/1/5

Abstract

This study proposes an intelligent data analysis approach to investigate and interpret the distinctive factors of diabetes mellitus patients with and without ischemic (non-embolic type) stroke in a small population. The database consists of a total of 16 features collected from 44 diabetic patients. Features include age, gender, duration of diabetes, cholesterol, high density lipoprotein, triglyceride levels, neuropathy, nephropathy, retinopathy, peripheral vascular disease, myocardial infarction rate, glucose level, medication and blood pressure. Metric and non-metric features are distinguished. First, the mean and covariance of the data are estimated and the correlated components are observed. Second, major components are extracted by principal component analysis. Finally, as common examples of local and global classification approach, a k-nearest neighbor and a high-degree polynomial classifier such as multilayer perceptron are employed for classification with all the components and major components case. Macrovascular changes emerged as the principal distinctive factors of ischemic-stroke in diabetes mellitus. Microvascular changes were generally ineffective discriminators. Recommendations were made according to the rules of evidence-based medicine. Briefly, this case study, based on a small population, supports theories of stroke in diabetes mellitus patients and also concludes that the use of intelligent data analysis improves personalized preventive intervention.

Introduction

In modern medicine, large amounts of data are generated, but there is a widening gap between data collection and data comprehension. It is clear that a single human can not process all of the data available and make a rational decision of basic trends. Thus, there is a growing pressure for intelligent data analysis techniques to facilitate the creation of knowledge to support clinicians in making decisions [1–3].

Understanding the major risk factors of a diesease is an important factor for clinicians in prevention strategy. The attending physician plays an important role providing information to reduce those risk factors. It is up to the physician whether to warn patients at risk about the major causes of a particular disease and the degree of risk that they are facing. Consider the example of a 66-year-old person who does not know about stroke (also in this study, ischemic stroke) but wants to know the risk of having certain medical test results outside of normal. Explaining the relative risk of stroke given the test results and given the evidence of previous cases interpreted with the aid of intelligent data analysis methods will make the situation clearer.

Stroke is an important health issue worldwide and expressing and interpreting risk factors provides vital epidemiological information [4–13]. This study discusses a computational method for highlighting the major risk factors of a small population of diabetic patients with and without non-embolic stroke by performing dependency analysis with local and global classification aspects. For this purpose, the follow-up data of 22 diabetic patients with ischemic stroke (non-embolic) and 22 diabetic patients without stroke were collected over several years [6]. Average population age was 66.2 ± 9.9 For the stroke group, age was 66.2 ± 9.9 (mean and s.d.) years and 61 ± 6.1 (mean and s.d.) years for the control group. Diabetes mellitus (DM) is diagnosed by a fasting glucose level higher than 140 mg/dl and random glucose levels higher than 200 mg/dl in repeated measurements. The study population of 44 patients was chosen with these glucose levels. Then, a set of tests were applied to construct the parameters of each feature vector. The tests include age, gender, duration of diabetes, cholesterol, high density lipoprotein (HDL), triglycerides levels, neuropathy, nephropathy, retinopathy, peripheral vascular disease, myocardial infarction rate, fasting and random glucose levels (FGL and RGL), medication, and systolic and diastolic blood pressures. The feature vectors thus contain both metric and nonmetric components. For example, a blood cholesterol level test is a metric component that can be processed with mathematical operations. On the other hand, retinopathy is a nonmetric component that provides a nominal scale to label or to identify retinal conditions.

This article discusses a limited number of cases of stroke in diabetic patients with primary risk factors for which preventive measures exist. Our purpose is to use existing knowledge for developing prevention strategies based on evidence-based medicine [13] that fit the patient and comply with current scientific concepts. Thus, this system transforms data into biomedical information that, together with expert knowledge, is helpful for decision making in patient care.

A small population study: Diabetic patients with and without non-embolic stroke

Using various diagnostic tests and searching for evidence of disease are generally routine for monitoring patients with diabetes mellitus (DM), and are also critical for patients [4,5]. Measuring major risk factors of ischemic stroke in patients with DM can also provide reasons for the attending physician to initiate preventive measures adapted to particular case. Typical microvascular complications are neuropathy, nephropathy, retinopathy; macrovascular complications are coronary artery diseases (CAD), and peripheral vascular diseases (PVD) [6,7]. Abnormal test results of cholesterol, HDL, triglycerides levels, FGL and RGL and sistolic and diastolic blood pressures are considered risk factors of nonembolic-ischemic stroke for DM patients [5–12]. Various studies have addressed the relationship between DM and stroke, and the probable risk factors [4–12]. Even though ischemic cerebravascular events are generally caused by macrovascular changes [7–9], in DM the cerebral ischemia is caused by small artery occlusion [6]. In DM, age, hypertension, MI, PVD are the known risks of stroke [4,5] but few studies are available to indicate the relationship between neuropathy, nephropathy, retinopathy, and stroke.

Means, standard deviations and correlations are estimated from the observed population. It is known that small sample sizes reduce the statistical power of the study and generally result in wide confidence interval of estimated parameters. Thus, identical studies on larger samples may identify more important differences that have gone undetected here. Also, deviation from the normal may cause errors in parameter estimations. As compensation for the small size of the study population, we first observe the statistics with an assumption of normal distribution, then attempt to employ nonparametric classification techniques such as k-NN and MLP with the jackknife method. Although there exist ways of checking the fit: for example, how well the normality of distribution for each component can be evaluated by various means such as the Kolmogorov-Smirnov test, our aim is to build a decision support mechanism. Regardless of the limitations imposed by technical artifacts and sample size constraints, our results describe several important findings that emerged from this preliminary investigation.

Mean and standard deviation of metric components: cholesterol, HDL, triglycerides levels, FGL and RGL, and systolic and diastolic blood pressures are presented in Table 1. Correlation estimates of pairwise relationships of the components were calculated, as presented in Tables 3 and 4. Nonmetric components were counted as the percentage of existing or nonexisting cases, as given in Table 2.

Investigating principal factors of data

Among the factor analysis methods [13–16], principal component analysis (PCA) is used to obtain independent factors from data. PCA is based on obtaining eigenvalues and eigenvectors from the covariance matrix of data. The covariance matrix is transformed into a diagonal matrix by applying an orthogonal transformation, with the factors arranged in relative order of importance. The estimation of the covariance matrix, $\hat{\Sigma}$ is

$$\hat{\Sigma} = \frac{1}{n}\sum (x-\mu)(x-\mu)^T \quad (1)$$

Table 1: Statistics of metric components of DM patients with and without ischemic stroke

	Ischemic stroke		No stroke	
	Mean	**St. Dev.**	**Mean**	**St. Dev.**
Age	66.2	9.9	61	6.1
Cholesterol	203.4	54.5	217.8	50.9
Triglyceride	150.5	97.8	1556.0	89.9
HDL	41.0	6.5	44.1	6.8
RGL	206.6	86.2	196.2	53.0
FGL	187.3	75.3	180.3	62.5
Systolic	143.6	26.3	143.2	14.6
Diastolic	83.6	12.9	85	11.0
DM (month)	97.9	91.1	186.3	87.7

Table 2: Statistics of non-metric components of DM patients with and without ischemic stroke

	first class	second class
Gender		
Female	6(%27.3)	8(%36.4)
Male	16(%72.7)	14(%63.6)
Medicine		
Used	11(%50)	17(%77.3)
Not used	11(%50)	5(%22.7)
Neuropathy		
Exist	15(%68.2)	19(%86.3)
Nonexist	7(%31.8)	3(%13.7)
Nephropathy		
Exist	7(%31.8)	9(%40.9)
Nonexist	15(%68.2)	13(%59.1)
PVD		
Exist	6(%27.3)	7(%31.8)
Nonexist	16(%72.7)	15(%68.2)
Retinopathy		
Exist	4(%18.2)	2(%9.1)
Nonexist	18(%81.8)	20(%90.9)
MI		
Exist	4(%18.2)	1(%4.5)
Nonexist	18(%81.8)	21(%95.5)

Here, the x_i's are constituent vectors for sample space, and $\hat{\mu}$ is the mean value. PCA compresses information to eigenvalues, $\lambda_1,\lambda_2,...,\lambda_n$ and eigenvectors, $u_1,u_2,...,u_n$ of $\hat{\sum}$. The matrix $x^T u_i$ and the eigenvectors u_i, i = 1,...,p are needed for recovering data. The reconstruction error R^2 is due to the neglecting small eigenvalues:

$$R^2 = trace[D] - trace[D_M] = \sum_{n=M+1}^{N} \lambda_n \qquad (2)$$

where D is the diagonal eigenvalue matrix and R^2 is the error arising from the eigenvalues λ_n, i = M + 1, ... N that are left out.

By PCA analysis of all components, we find the first macrovascular components of Table 1: cholestrol, trigliseride, glucose, and blood pressures to be the principal components. From a decision making point of view, we observe that these components are the most informative with regard to in the population study. In fact, this finding confirms the reports of major risk factors reported in the literature [4–13]. Among the other components, age and

Table 3: Results of k-NN method

	All Components	Principle Components
	% for total	
k = 1	52.23	52.3
k = 3	65.9	65.9
k = 5	68.2	68.2
k = 7	63.6	63.6
k = 10	68.2	68.2
k = 20	63.6	63.6
k = 40	29.5	29.5

Table 4: Results of MLP method

Epsilon = 0.1 alpha = 0.8 iteration = 10000	
Number of units (k)	**Success rate (%)**
3	60
5	66
7	65

gender are designated during data collection and have less information. Microvascular changes, neuropathy, nephropathy, retinopathy, are also assessed by only two distinctive levels (existence-nonexistence) to show the subgroups of DM subjects. Thus, they too have limited information that can be used for decision making. PCA is found to be valuable for summarizing the trends of features of high-risk patients with DM.

Decision making by principal factors

It is known that there is no standard approach to establish the optimum decision criteria for a generic problem. Nonparametric classification approaches [19–22] overcome the difficulty of accurate estimation of underlying distributions from small population size. These approaches are used without the assumptions about the form of the density function. Their decision procedures bypass probability estimation and go directly to decision functions. Due to the small population, we here employ nonparametric techniques instead of parametric ones such as K-Means, vector quantization (VQ) and gaussian mixture model (GMM) which, otherwise, are very useful with the large population. Fuzzy modeling methods could be used to also enhance the decision ability [23].

Among various nonparametric techniques, we choose a local k-NN and a global polynomial classifier, MLP. The concept of locality and globality is related to the location of information that is extracted for class decision. The k-NN rule employs local information, in contrast, an MLP extracts global information.

It is known that the nearest neighbor in the feature space gives the half of the classification information obtained by the Bayes classifier [19]. On the other hand, polynomials have excellent properties as classifiers. Thus, polynomial classifiers are universal approximators relative to the optimal Bayes classifier [20]. Typical polynomial classifiers have either been based on either statistical methods or minimizing a mean-squared error criterion. The focus has been on linear and second-degree polynomial classifiers, but both of these methods traditionally have had limitations. Classical statistical methods estimate the mean and covariance for a parametric model [19–22]. These methods with a large number of features or multivariables lead to large intractable problems [17,18].

With the k-NN rule, the class decision of an unknown sample is based on the majority of the nearest k samples. In other words, a local voting process decides the class of the unknown by the highest vote. This is, in fact, an approximation of Bayes Decision Theory in a local environment [19]. This process can further be extended to weighted voting and gaussian weighted voting (Parzen window). To recognize pattern v, k minimum distance samples are computed among all the samples. The dis-

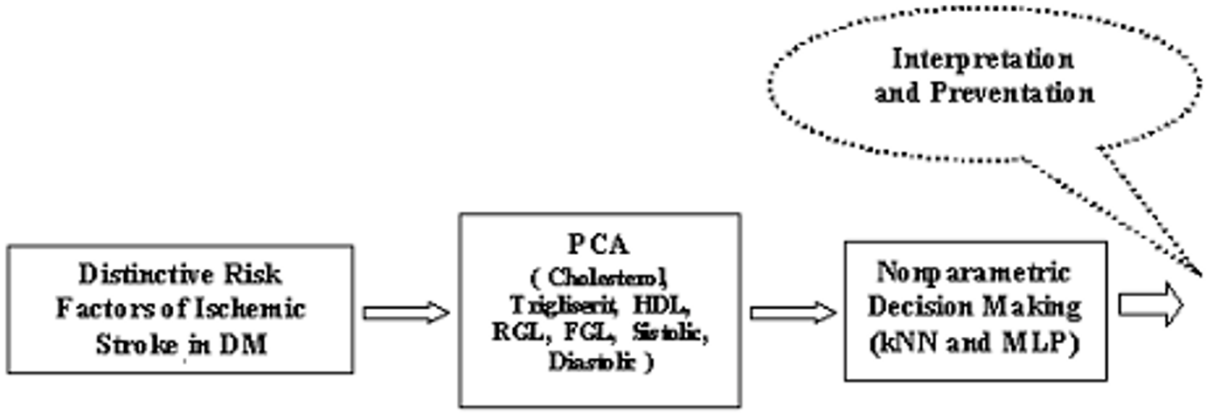

Figure 1
Distinctive risk factors of DM patients with and without stroke

tance can be computed by various norms: minkowski-norm-based distances, e.g. euclidian as second order, covariance-based distance or mahalonobis, entropy-based distance or kullback-leibler distances. The simplest and the most suitable process for a small sample size is the euclidian metric, which is defined as

$$dj = |v - v_j|^2 = (v - v_j)^T(v - v_j) \quad (3)$$

where dj defines the distance between patterns v and v_j. The k-NN rule classifies v by assigning it a label that is the most frequently represented among the k nearest samples:

$$ki = \max \{k_1,....,k_L\}\ x \in w_i \quad (4)$$

$$k_1 + ... + k_L = k$$

Here, k_i is the number of neighbors belong to class w_i (i = 1,....L) class among the k nearest neighbors.

The global MLP [22] is a parallel, feedforward structure that consists of input, output, and hidden layers. Each layer has sigmoidal units interconnected through weight connections. The MLP is trained with a supervised back propagation (BP) algorithm to efficiently compute partial derivatives of an approximating function F(w; x). The network has an adjustable weight vector, w, that is computed with respect to all training data for a given value of input vector x and output vector y. The weights are adjusted to fit a set of surfaces to the input space. The surfaces are constructed with sigmoids by using the best linear regression concerning the cluster membership. The mean square (MSE) (Eq. 5) error function, which is the difference between the network's output and the supervisor output, is minimized to find the cluster membership:

$$MSE = \sum_q \left(y_q - F\left(w; x_q \right) \right)^2 \quad (5)$$

In k-NN rule, an arbitrary discriminant function is constructed for class decision. The nearest neighbor contributes the half of the classification information. An MLP can generate any nonlinear discriminant function of input by incorporating multiple constraints. Each sigmoidal unit contributes to the global discriminant function by a linear constraint.

Results

The study population contains data relating to 44 diabetic patients' collected by neurology specialists. The data have a total of 16 metric and non-metric components. The computer-based system uses two stages: first, the PCA method reduces the features to seven dimensions. The best discriminators for class membership are then searched by nonparametric classifiers, k-NN and MLP. Multi-parameter classifiers were found to significantly improve upon the classification performance of single parameter designs. Instead of single parameter based

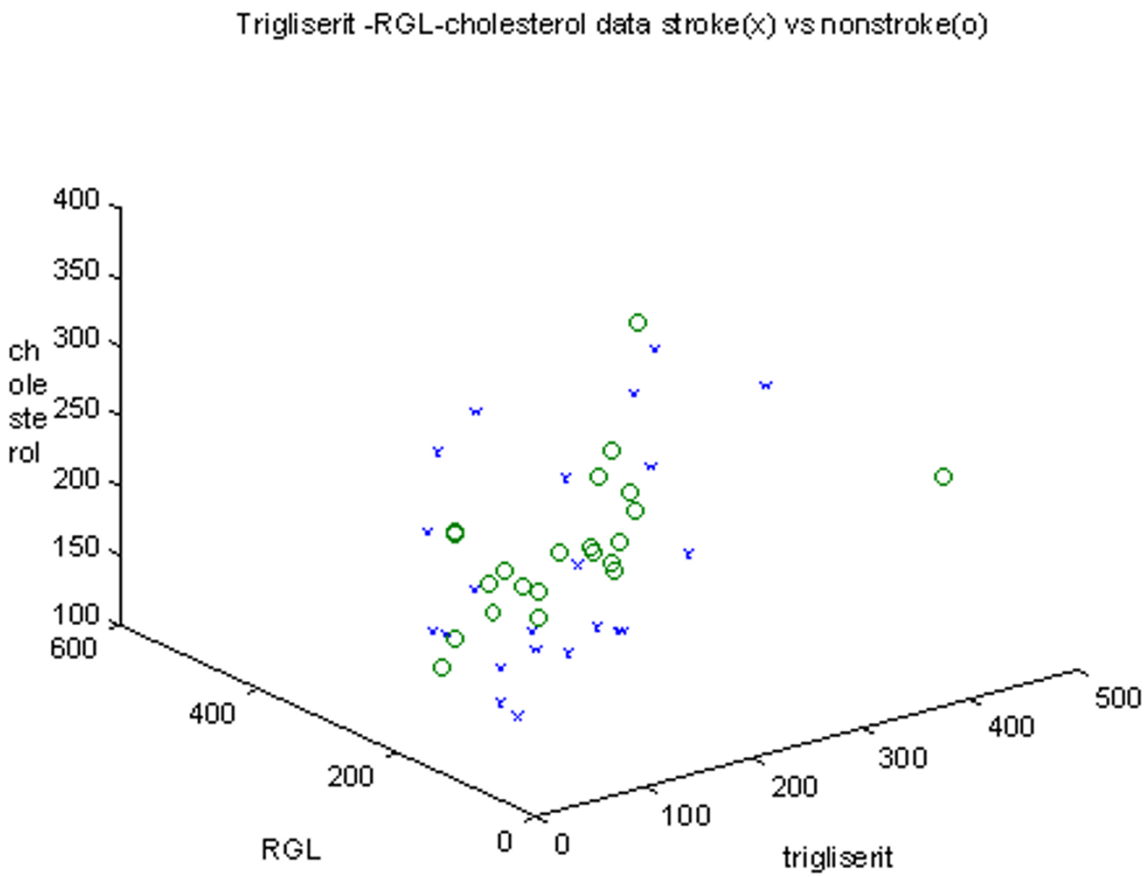

Figure 2
Visual information of triglyceride-RGL-cholestrol

conclusions, we employed the decision produced by major risk factors. The method gives the study stronger arguments on the distintive factors. The small sample population size has an acknowledgable potential effect on the statistical power of the study, but we use classification techniques like k-NN and MLP to help overcome this drawback.

The results are shown in Table 3 and Table 4. In case of k-NN, training and test sample sets are equivalent. An average classification score of 68.2% is obtained for k-NN. But in the case of MLP, we implement an hold-n-out method (jacknife, cross-validation) which is proven to be the most common method of network construction and validation, train-and-test, in cases where data are scarce. Test results are repeated with different sets using 10 test datum points and the remaining data for training. The average performance of five MLP experiments is found to be 66%. Both classifiers identify the macrovascular changes as the best discriminators for this case study of stroke.

Figure 2 shows the component clusters and clearly provides a convincing picture to the physician and patients as the end-users. This easy-to-use graphic offers users information about the limits of the major risk factors.

Discussions and Conclusion

The aim of this study is to support physicians with intelligent data analysis techniques by providing what may be thought of as an indirect assistance tool. This paper presents a computer-based decision support system for physicians to improve prevention measures for ischemic stroke in DM patients. PCA helps to identify the basic trends that distinguish two cases: ischemic stroke and no stroke in DM. In this study, major distinctive factors are found to be cholesterol, triglycerides level, fasting and random glucose levels (FGL and RGL), and systolic and

diastolic blood pressures. Physicians, furthermore, can suggest dietary supplements, drugs, exercises, etc. for the prevention of ischemic stroke for diabetic patients. Graphics of the components also support the discussion with clues for prevention.

We also plan to improve the information of diagnosis with a certain risk assessment factor. A grading for the risk of ischemic stroke for diabetic patients will support the prevention measures according to evidence-based medicine.

The major concern of the study is that the investigators could not have a bigger size of study population. This raises concerns about the meaningfulness of results but the authors believe that the methods used here clearly useful to overcome this handicap. Also the results of the study are supported by the risk factor findings of the other stroke and diabet studies in the literature [4–13]. When there is a large database, this will provide more convincing statistical power to the study. In this case, the general statistics will confirm the results and parametric decision making techniques such as expectation-maximization (EM) algorithm would be utilized.

The other limitation is the inability to directly account for certain microvascular factors in the decision process due to imprecision of their records: for example, retinopathy has only two identifying labels as exist or non exist. In fact, there are typically four cases that can be distinguished by the physician but the limited data size prevents us to include finer details. This may cause an error. The lack of precision, in general, reduces the effect of the feature in decision making.

Among the various other tests, we also believe that the doppler ultrasound measurements may present more detailed information on this case study. The areas of arteries and veins can be monitored for deciding the stages of the disease, and the condition of occlusions can be important for the preventive follow up.

The other weak point of this investigation may be the quality issues of the recording such as the use of non-standardized procedures, the use of potentially low grade equipment or the measurement quality. However, as best supported by a group of physicians' observations, this was not the case.

As a conclusion, this case study points at a friutful line of enquiry in intelligent medical data analysis and the content of the work can further be extended to the other areas of stroke diagnosis and prevention.

Acknowledgement

We would like to express our thanks to Boaziçi University BAP 03A103 for their support. We would also thank to Editor of Biomedical Engineering Online Prof. Alwin Wald.

References

1. Brause and Rudiger **Medical Data Analysis.** *Springer-Verlag* 2000,
2. Hand DJ, Kok JN and Bertholt MR **Advances in Intelligent Data Analysis.** *Springer-Verlag* 1999, 89-93
3. Gurgen F **Neural-Network-Based Decision Making in Diagnostic Applications.** *IEEE Engineering in Medicine and Biology* 1999, 89-93
4. Barnett HJM, Mohr JP, Stein BM and Yatsu FM **Stroke: Pathophysiology, Diagnosis and Management.** *Secon Edition, Churchill Livingstone Inc* 1992,
5. Bogousslavssky J and Caplan L **Stroke Syndromes.** *Cambridge University Press* 1995,
6. Aydýn N, Esgin H, Yilmaz A, Gözeten F and Utku U **Diabetes Mellitus'lu Non-Embolik Stroklu Olgularda Retinopati ve Dier Risk Faktörleri.** *In Proceeding of theTürk Beyin Damar Hastalýklarý Dernei 3. Sempozyumu* 1999,
7. Silvestrini M, Rizzato B, Placidi F, Baruffaldi R, Bianconi A and Diomedi M **UKPDS 60: Risk of Stroke in Type 2 Diabetes Estimated by the UK Prospective Diabetes Study Risk Engine.** *Stroke* 2002, **33:**1776-1781
8. Horenstein RB, Smith DE and Mosca L **Cholesterol Predicts Stroke Mortality in the Women's Pooling Project.** *Stroke* 2002, **33:**1863-1868
9. Muntner P, Garrett E, Klag MJ and Coresh J **Trends in Stroke Prevalence Between 1973 and 1991 in the US Population 25 to 74 Years of Age.** *Stroke* 2002, **33:**1209-1213
10. Petty GW, Brown RD Jr, Whisnant JP, Sicks JD, O'Fallon WM and Wiebers DO **Ischemic Stroke Subtypes : A Population-Based Study of Functional Outcome, Survival, and Recurrence.** *Stroke* 2000, **31:**1062-1068
11. David SH and Bell MB **Stroke in the diabetic patient. Diabetes Care.** *Stroke* 1994, **17:**213-219
12. Jose B and Love BB **Diabetes and stroke.** *Medical Clinics of North America* 1993, **77(1):**95-109
13. Diana BP and Hilarey B **Retinopathy as a Risk Factor for Non-embolic Stroke in Diabetic Subjects.** *Stroke* 1995, **26:**593-596
14. Abbott RD and Donahue RP **Diabetes and Risk of Stroke.** *JAMA* 1987, **257:**949-952
15. Mortel KF, Meyer JS, Sims PA and McClintic K **Diabetes Mellitus as a Risk Factor for Stroke.** *Southern Medical Journal* **83:**904-911
16. Sackett DL, Rosenberg WM, Gray JA, Haynes RB and Richardson WS **Evidence-based Medicine: What it is and what it isn't.** *Mr Med J* 1996, **312:**71-72
17. Sharma S **Applied Multivariate Techniques.** *John Wiley* 1996,
18. Rencher AC **Methods of Multivariate Analysis.** *John Wiley* 1995,
19. Duda RO and Hart PE **Pattern Classification and Scene Analysis.** 1972,
20. Devroye L, Gyorfi L and Lugosi G **A Probabilistic Theory of Pattern Recognition.** *NewYork: Springer Verlag* 1996,
21. Fukunaga K **Introduction to Statiscal Pattern Recognition.** *Academic Press* 1990,
22. Schürmann J **Pattern Classification, A Unified View of Statistical and Neural Approaches.** *John Wiley & Sons Inc* 1996,
23. Takagi T and Sugeno M **Fuzzy identification of systems and its applications to modelling and control.** *IEEE Trans. Systems Man and Cybernetics* 1985, **15(1):**116-132