International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:5, No:1, 2011

Using Data Mining Techniques for Finding Cardiac Outlier Patients

Farhan Ismaeel Dakheel, Raoof Smko, K. Negrat, Abdelsalam Almarimi

***Abstract*—**In this paper we used data mining techniques to identify outlier patients who are using large amount of drugs over a long period of time. Any healthcare or health insurance system should deal with the quantities of drugs utilized by chronic diseases patients. In Kingdom of Bahrain, about 20% of health budget is spent on medications. For the managers of healthcare systems, there is no enough information about the ways of drug utilization by chronic diseases patients, is there any misuse or is there outliers patients. In this work, which has been done in cooperation with information department in the Bahrain Defence Force hospital; we select the data for Cardiac patients in the period starting from 1/1/2008 to December 31/12/2008 to be the data for the model in this paper. We used three techniques for finding the drug utilization for cardiac patients. First we applied a clustering technique, followed by measuring of clustering validity, and finally we applied a decision tree as classification algorithm. The clustering results is divided into three clusters according to the drug utilization, for 1603 patients, who received 15,806 prescriptions during this period can be partitioned into three groups, where 23 patients (2.59%) who received 1316 prescriptions (8.32%) are classified to be outliers. The classification algorithm shows that the use of average drug utilization and the age, and the gender of the patient can be considered to be the main predictive factors in the induced model.

***Keywords*—**Data Mining, Clustering, Classification, Drug Utilization..

I. Introduction

IN modern times, cardiac disease has emerged as the leading cause of death worldwide, particularly in developed countries. The World Health Organization (WHO) (reported that 16.7 million deaths in 2003 (29.2% of total global deaths) were caused by some form of cardiovascular disease. Though the rate of cardiac disease is highest in developed countries, developing countries are seeing an increase in the occurrence of cardiac disease, as well as a corresponding rise in the number of heart-related deaths [1].

The World Health Organization (WHO) estimates that by 2010, cardiac disease will surpass AIDS as the leading cause of death in developing countries [2].

Most of the care for adult cardiac patients is provided by primary health care physicians. It is estimated that about 18.5% of the adult population in the Kingdom of Bahrain are diagnosed as suffering from cardiac diseases [3]. Optimal management of the disease requires continuous medical surveillance with periodical adaptation of the treatment regimen to the clinical state of the patients. The disease has significant clinical and economic impacts both on patients and the health care system. As a rule, many of the patients are using simultaneously a combination of drugs for extended periods of time in order to control their disease. Thus, the costs involved in care of these patients are substantial.

Currently, in many ambulatory health care systems, the extensive use of computers and computerized data-bases (DB) for routine services is quite common. In addition, medical information systems (MIS) and medical decision support system (MDSS) are also used in order to standardize and improve the quality of health care.

Data mining (DM) is the core stage of knowledge discovery in databases (KDD), which is a "non-trivial extraction of implicit, novel, and potentially useful information from data" [4]. It applies machine learning and statistical methods in order to discover areas of previously unknown knowledge. As a rule, the KDD process involves the following steps: data selection, data pre-processing, transformation, DM (induction of useful patterns), and interpretation of results.

Two common data mining tasks are: (a) *cluster analysis* aimed at organizing a given dataset into groups (clusters) of similar objects or characteristics and *(b)* classification aimed at predicting the class of objects whose class label is unknown. One of the most common classification models is the decision tree, which is a tree-like structure where each internal node denotes a test on a predictive attribute and each branch denotes an attribute value. A leaf node represents predicted classes or class distributions [5]. An unlabeled object is classified by starting at the topmost (root) node of the tree, then traversing the tree, based on the values of the predictive attributes in this object. Decision-tree techniques assume that the data objects are described by a fixed set of attributes, where each predictive attribute takes a small number of disjoint possible values and the target (dependent) variable has discrete output values, each value representing a class label.

There are several known algorithms of decision tree induction: ID3 - which uses information gain with statistical pre-pruning, C4.5, an advanced version of ID3, and probably the most popular decision-tree algorithm [6], CART, which minimizes a cost-complexity function, See5 - which builds several models and uses unequal misclassification costs, and

Farhan Ismaeel is with the Gulf University, Kingdom. (Phone: +973-336331352; e-mail: farhanawni@yahoo.com).
Raoof Smko, is with the College of Electronic Technology, Libya (Phone: +218-917201245; e-mail: raoofsmko@yahoo.com).
K. Negrat is with College of Electronic Technology, Libya. (Phone: +21891711715; e-mail dr_negrat@yahoo.com).
Abdelsalam Almarimi is with College of Electronic Technology, Libya (Phone: +218927409426; e-mail: belgasem_2000@yahoo.com).

IFN – Info-Fuzzy Network which utilizes information theory to minimize the number of predictive attributes in a decision-tree model [7] [8]. In [7], the IFN algorithm is shown empirically to produce more compact models than C4.5, while preserving nearly the same level of classification accuracy.

Cluster analysis or clustering, is a data mining technique aimed at grouping the data objects into classes or clusters so that objects within one cluster have high similarity to each other, but are very dissimilar to objects in all other clusters [4]. Clustering is a form of *unsupervised learning*, since it does not rely on class-labeled training examples. One of the most popular partitioning clustering methods is *k-means*, which splits a set of *n* objects into *k* clusters. Several methods for determining the optimal number of clusters have been proposed [9].

A time series database (TSDB) may include various types of variables with at least one temporal dimension. A large portion of today's databases, especially in financial, scientific, and medical domains, can be considered as time series data. Recently there has been a growing interest in mining time series databases, with attempts to cluster, classify [10], maintain, and index temporal data.

Medical applications of data mining include prediction of the effectiveness of surgical procedures, medical tests, and medications as well as discovery of relationships between clinical and pathological data [11]. Clinical databases store large amounts of information about patients and their medical conditions. Data mining techniques applied to these databases discover relationships and patterns which are helpful in studying progression and management of diseases.

II. Data Acquisition and Data Cleaning

There are many Databases are used by BDF (Bahrain Defense Force) Royal Medical Services in Kingdom of Bahrain. In this work, we used the data from (AI Care Medical Information System (CMIS)) as the main source for data mining process. All medications prescribed and given to patients in the pharmacy are recorded in central CMIS, primarily for administrative and financial control. For each prescription the following data is recorded: date, patient ID, Age, Gender, Nationality, Region, Visit Date, and the Medicine Details, as shown in Fig. 1. We have extracted from the CMIS a data set of all cardiac-related prescriptions issued to adult patients (ages 25-65) registered during the period of January 2008 to November 2008.

The drug prescriptions TSDB (Time-Series DB) that was extracted from the CMIS database initially included 15,806 records of cardiac drugs prescribed for 1603 patients included in our study.

In the data cleaning stage, duplicate records were removed and missing attributes completed by using appropriate data. Records with missing drug name and/or date attributes were completely removed from the dataset.

Based on recommendations made by an cardiac diseases specialist, we decided to consider a patient as an "cardiac" if she/he received six or more prescriptions of medications commonly used for treating cardiac diseases during the one-year period of study. The number of patients remained in the drug data set after applying this restriction was 1603, with 15,806 prescriptions.

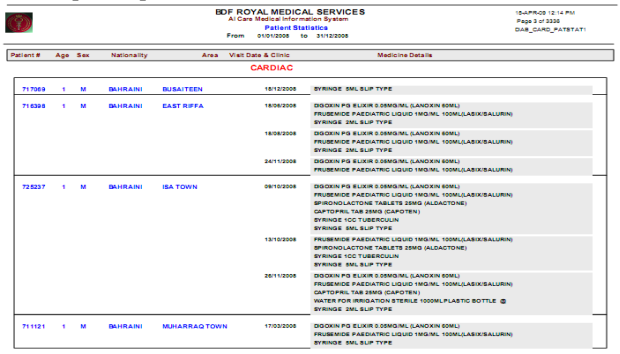
BDF ROYAL MEDICAL SERVICES
AI Care Medical Information System
Patient Statistics
From 01/01/2008 to 31/12/2008
CARDIAC
Patient # | Age | Sex | Nationality | Area | Visit Date & Clinic | Medicine Details
717069 | 1 | M | BAHRAINI | BUSAITEEN | 18/12/2008 | SYRINGE 5ML SLIP TYPE
716398 | 1 | M | BAHRAINI | EAST RIFFA | 18/06/2008 | DIGOXIN PG ELIXIR 0.05MG/ML (LANOXIN 60ML)
FRUSEMIDE PAEDIATRIC LIQUID 1MG/ML 100ML(LASIX/SALURIN)
SYRINGE 2ML SLIP TYPE
18/06/2008 | DIGOXIN PG ELIXIR 0.05MG/ML (LANOXIN 60ML)
FRUSEMIDE PAEDIATRIC LIQUID 1MG/ML 100ML(LASIX/SALURIN)
SYRINGE 2ML SLIP TYPE
24/11/2008 | DIGOXIN PG ELIXIR 0.05MG/ML (LANOXIN 60ML)
FRUSEMIDE PAEDIATRIC LIQUID 1MG/ML 100ML(LASIX/SALURIN)
722237 | 1 | M | BAHRAINI | ISA TOWN | 09/10/2008 | DIGOXIN PG ELIXIR 0.05MG/ML (LANOXIN 60ML)
FRUSEMIDE PAEDIATRIC LIQUID 1MG/ML 100ML(LASIX/SALURIN)
SPIRONOLACTONE TABLETS 25MG (ALDACTONE)
CAPTOPRIL TAB 25MG (CAPOTEN)
SYRINGE 1CC TUBERCULIN
SYRINGE 5ML SLIP TYPE
13/10/2008 | FRUSEMIDE PAEDIATRIC LIQUID 1MG/ML 100ML(LASIX/SALURIN)
SPIRONOLACTONE TABLETS 25MG (ALDACTONE)
SYRINGE 1CC TUBERCULIN
SYRINGE 5ML SLIP TYPE
26/11/2008 | DIGOXIN PG ELIXIR 0.05MG/ML (LANOXIN 60ML)
FRUSEMIDE PAEDIATRIC LIQUID 1MG/ML 100ML(LASIX/SALURIN)
CAPTOPRIL TAB 25MG (CAPOTEN)
WATER FOR IRRIGATION STERILE 1000ML PLASTIC BOTTLE @
SYRINGE 2ML SLIP TYPE
711121 | 1 | M | BAHRAINI | MUHARRAQ TOWN | 17/03/2008 | DIGOXIN PG ELIXIR 0.05MG/ML (LANOXIN 60ML)
FRUSEMIDE PAEDIATRIC LIQUID 1MG/ML 100ML(LASIX/SALURIN)
SYRINGE 5ML SLIP TYPE

Fig. 1 A sample page of the data taken from CMIS for cardiac patients

III. Clustering Patients by Drug Utilization Data

We started the data mining process by using the *K*-means clustering algorithm to identify groups of patients having similar utilization patterns of cardiac drugs over the period of 12 months. The features we used to characterize each patient were the number of prescriptions issued per each month. All records (time series) had 12 monthly values for drug utilization; hence, we clustered vectors having 12 dimensions. As shown in Figure 2, the average monthly drug utilization is about 84% (746) of patients consume less than two prescriptions per month.

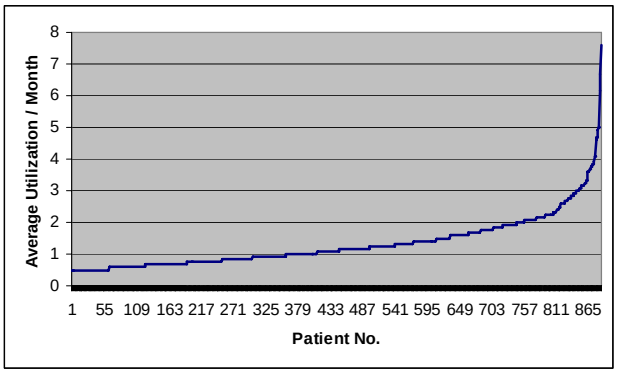

Fig. 2 Average Monthly Utilization per Patient (Sorted in ascending order)

It is not possible to indicate the exact number of clusters, so we used k=2 to k=10 for the purpose of finding the optimal number of clusters. We calculated the inter-cluster and the intra-cluster dissimilarity using the Euclidean distance between two vectors [12].

Clustering validation indices measure how well a dataset is clustered with different settings. In this work, we used the following validation techniques to determine the best number

of clusters *k*: Dunn Validity Index and Silhouette Validation Method.

Dunn Validity Index: [13][14] This index is seeking for compact and well separated clusters. For any partition of clusters, where c_i and c_j are *i* and *j* clusters of such partition, the Dunn parameter *D* is calculated by the following equation:

$$D = \min_{1\le i\le n}\left\{\min_{\substack{1\le j\le n\\ i\neq j}}\left\{\frac{d(c_i,c_j)}{\max_{1\le k\le n}\{d'(c_k)\}}\right\}\right\} \quad (1)$$

In Eq. (1) above, $d(c_i, c_j)$ is the distance between clusters *i* and *j* (inter-cluster); $d'(c_k)$ is the distance between objects in cluster *k* (intra-cluster); and *n* is the total number of clusters. Eq. (1) is aimed at minimizing the intra-cluster distance and maximizing the inter-cluster distance. Thus, the best number of clusters should maximize this equation.

Silhouette Validation Method ([13] [14]). In this method, each cluster is represented by a type of silhouette, which is the ratio between its compactness and distribution. The technique computes the silhouette width for every object, the average width for each cluster, and the average silhouette width for the entire dataset.

The formula we use to construct the object silhouette is:

$$S(i) = \frac{(b(i) - a(i))}{\max\{a(i), b(i)\}} \quad (2)$$

Where *a(i)* is the average distance (dissimilarity) of an object *i* to all other objects in its cluster and *b(i)* is the average distance of an object *i* to all other objects in the closest cluster.

Based on Eq. (2), $-1<S(i)<1$. If $b(i)\gg a(i)$ then $S(i)\rightarrow 1$, the object is clustered as best as possible. If $b(i)=a(i)$ then $S(i) = 0$, the object is indifferent and can be allocated in another cluster. if $b(i)\ll a(i)$ then $S(i)\rightarrow -1$, the object is misclassified and located between the clusters, with a high tendency to the other cluster. The average of all silhouette widths for all objects in the dataset is the average *S(i)* for all items. The optimal number of silhouette widths is taken as the largest *S(i)*, which indicates the best clustering. The main idea of both indices (Dunn and Silhouette) is that the inter-similarity will be as small as possible and the intra-similarity large but while Dunn's method uses the inter-similarity per cluster, the Silhouette method constructs its formula by inter-cluster similarity per object.

We performed ten clustering runs for each value of *k*; each run started with different objects as initial centroids. In each run, we calculated the objects in the clusters, the centroid of each cluster, the average inter-cluster similarity, the intra-cluster similarity of each object inside its cluster, the average intra-cluster similarity of each cluster, the Dunn index for each run, the Silhouette index of each object, and, finally, the total Silhouette index of each run.

The results of two clusters and ten clusters were considered poor by the domain experts. For two clusters, all patients were trivially partitioned into two groups of high and low utilization, which has not provided any new insight into the distribution of drug usage. For ten clusters, in almost every run we obtained a cluster of one object, which made us believe that the actual number of multi-object clusters is not greater than nine.

As can be seen in Figure 3 both validation indices tend to become smaller as *k* is increased. Thus, we have concluded that the best number of clusters for this dataset should be three.

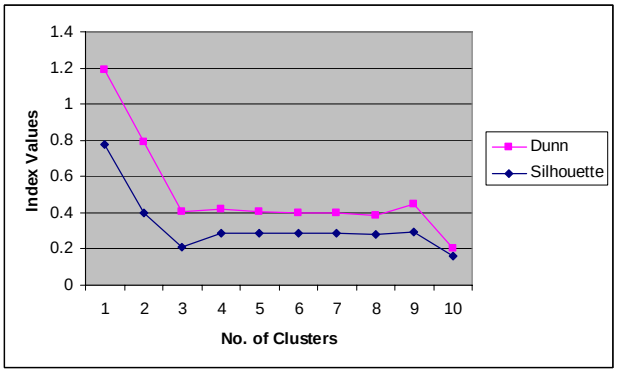

Fig. 3 Validation indices as function of *k*

The results from this clustering have shown that the utilization time series (representing cardiac patients) should be partitioned into the following three clusters: 23 patients (2.59%) with very high utilization (denoted as Cluster 1, centered at 54), 200 patients (22.5%) with medium utilization (denoted as Cluster 2, centered at 26), and 664 patients (74.91%) with low utilization (denoted as Cluster 3, centered at 11). The average monthly number of prescriptions was 4.47, 1.95, and 0.93 for patients in Clusters 1, 2, and 3, respectively. Although the clusters were constructed from vectors of 12 months, the cluster assignment of all objects but one matched their average monthly utilization. The most interesting cluster is the smallest one for obvious reasons. In addition, the largest cluster shows that on average a cardiac patient gets more than 4 prescriptions per month, while about 50% of patients in this cluster get only two prescriptions every month or even less than that. In the next subsection, we are applying classification algorithms to these clusters with the purpose of discovering the most influential features that affect the monthly drug utilization for each patient.

IV. Classification of Clustered Patients

After selecting the *k* value, and grouping the patients into three clusters according to there drug utilization. A second stage for this work is to apply a supervised classification algorithm to the main data set. We applied a direct decision tree algorithm to the data with predefined parameters as

shown in Table 1. We used the variables available in the data set for this model which are: Age of the patient, Sector Area, Gender, Nationality, and the average drug utilization per month.

TABLE 1 THE MODEL SUMMARY FOR THE CLASSIFICATION STAGE

Specifications	Growing Method	CHAID
	Dependent Variable	Average Utilization
	Independent Variables	Age, Gender, Nationality, Sector
	Validation	None
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50
Results	Independent Variables Included	Age, Gender
	Number of Nodes	8
	Number of Terminal Nodes	5
	Depth	2

Fig. 4 shows the output decision tree, which is consists of three levels with 8 nodes including 5 of them acting as terminal nodes. In the root node (Node.0 in figure (4)), all the (1603) records appear, with mean equal 0.822 and a standard deviation of 0.833. The remaining two variables will make the other two levels of the decision tree. In the first level, there are three nodes according to the grouping of the Age of the patient variable. The age is divided into three categories: first category includes the patients with ages less than or equal 64 years as shown in Node.1 in figure (4), the number of patients in this node are 647 patients (40.4% from total number of patients), the mean is 0.427 and the standard deviation is 0.511. In Node.2 (figure (4)), the second category of ages between 64 and 73 years are included. In this node, 169 patients (10.5%) appear with mean to be 0.748 and the standard deviation is 0.758. The final category of the age is in Node.3, the patients with ages more than 73 years are included in this node. The number of patients is 787 (49.1%), the mean is 1.162 and the standard deviation is 0.912. In the second level, another variable included for prediction process; this variable is the gender of the patient. Node.1 and Node.2 are divided according to this variable into another two nodes, while this variable has no effect on Node.3 (due to the data set). Node.4 contains 185 patients (11.5%), the mean is 0.307 and the standard deviation is 0.403. Node.5 contains 462 patients (28.8%), the mean is 0.475 and the standard deviation is 0.542. Node.6 contains 56 patients (3.5%), the mean is 0.579. Node.7 contains 113 patients (7%), the mean is 0.831 and the standard deviation is 0.776. The means of these nodes actually indicated the predicted average drug utilization per month for the patients included in this node. So, the classification process serves us to predict the drug utilization for any patient according to these variables included in the analysis.

Fig. 4 shows part of the decision tree, including two levels of branching, depending on the age and the gender of the patients. Table 2 shows the result tree in table form, which is another representing for the results.

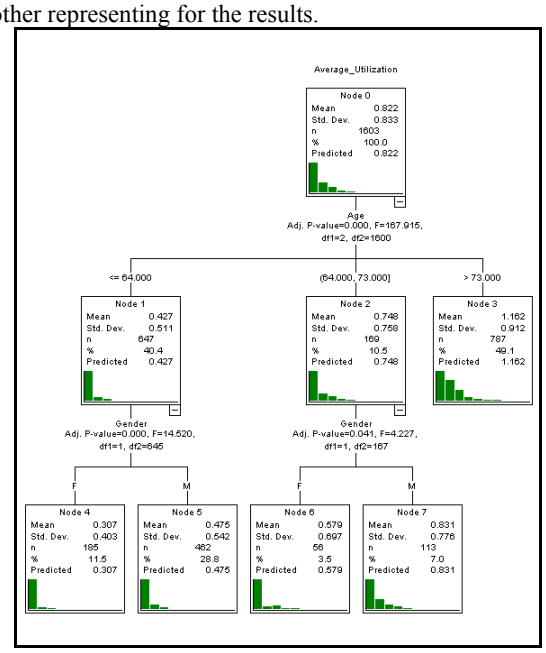

Fig. 4 The Classification Tree of the Model

TABLE II TREE TABLE FOR THE CLASSIFICATION MODEL

Node	Mean	Std. Deviation	N	Percent	Predicted Mean	Parent Node
0	.8217	.83327	1603	100.0%	.8217	
1	.4269	.51120	647	40.4%	.4269	0
2	.7475	.75805	169	10.5%	.7475	0
3	1.162	.91219	787	49.1%	1.1621	0
4	.3071	.40318	185	11.5%	.3071	1
5	.4749	.54153	462	28.8%	.4749	1
6	.5788	.69715	56	3.5%	.5788	2
7	.8311	.77590	113	7.0%	.8311	2

V. CONCLUSIONS AND FUTURE WORKS

Application of data mining methods can help physicians and health management organizations to monitor efficiently the utilization of drugs by chronic patients. Automated

identification of patients at risk of suboptimal treatment or over utilization can help a physician in better controlling such patients.

In this paper, we have partitioned cardiac patients into three clusters of: high usage (about (2.59%) of the patients), medium usage (22.5%), and low usage (74.91% of the patients). Decision tree classification algorithm was applied to the clustered data. According to the classification algorithm, the average drugs utilization per month is the major predictor. Minor prediction factors that were identified are: Age, gender, nationality, sector area.

This model can predict the class of any new patient according to the prediction factors identified by the model, and hence, we can determine the average drug utilization for these patients which has a major effect on estimating health budget, and find the outlier cases inside the data.

As ideas for future work, we can improve the classification process by using more efficient algorithms; such as CLIP4, INF, and C4.5. Also, it is a great idea to implement the work on huge size data. Working with huge data (in amount and in the time period) means more accurate prediction model.

It is interesting point for physicians and health management organizations to deal with specific drug utilization, and finding the trends and the patterns of such drugs. This idea focuses on another side of physicians and health management organizations views.

REFERENCES

[1] G. Y. H. Lip, K. Peter *"New oral anticoagulant drugs in cardiovascular disease"*, Thrombosis and Haemostasis. ISSN: 0340-6245. 2010 July.
[2] World Health Organization, "*The World Health Report 2006 - working together for health*", http://www.who.int/whr/2006/en/index.html. 2006.
[3] Ministry of Health – Kingdom of Bahrain. Annual Report of 2008.http://www.moh.gov.bh/PDF/Publications/Statistics/HS2008/PDF/CH03-vital%20stat_2008.pdf
[4] J. Han, M. Kamber, *Data Mining: Concepts and Techniques*, 2nd Edition, Morgan Kaufmann, 2006.
[5] T. Mitchell, *Machine Learning*, McGraw Hill, 1997.
[6] J.R. Quinlan: *C4.5, Programs for MachineLearning*, Morgan Kaufmann, 1993.
[7] M. Last and O. Maimon, "A Compact and Accurate Model for Classification", *IEEE Transactions on Knowledge and Data Engineering* 2004; 16, 2: 203-215.
[8] O. Maimon and M. Last, *Knowledge Discovery and Data Mining – The InfoFuzzy Network (IFN) Methodology*, Kluwer Academic Publishers, Massive Computing, Boston, December 2000.
[9] M. Halkidi, Y. Batistakis, M. Vazirgiannis, "On Clustering Validation Techniques", *J. Intell. Inf. Syst.* 2001; 17, 2-3: 107-145.
[10] M. Last, Y. Klein, A. Kandel, "Knowledge Discovery in Time Series Databases", *IEEE Transactions on Systems, Man, and Cybernetics* 2001; 31, 1: 160-169.
[11] J.C. Prather, D.F. Lobach, L.K. Goodwin, J.W. Hales, M.L. Hage, W.E. Hammond, "Medical Data Mining: Knowledge Discovery in a Clinical Data Warehouse", *Proc AMIA Annu Fall Symp*. 1997:101-5.
[12] Krzysztof J. Cios, Witold Pedrycz, Roman W. Swiniarski, and Lukasz A. Kurgan "Data Mining: A Knowledge Discovery Approach" ISBN-13: 978-0-387-33333-5; 2007 Springer.
[13] J.C. Dunn, "Well Separated Clusters and Optimal Fuzzy Partitions", *J. Cybern*. 1974; 4: 95-104.
[14] F. Azuaje, "A Cluster Validity Framework for Genome Expression Data", *Bioinformatics* 2002; 18: 319-320.