Application of Data Mining techniques to relate Cardiovascular Risk and Coronary Calcium

F N Lujan, L J Cymberknop, M Alfonso, W Legnani, R Armentano Feijoo

National Technological University – Buenos Aires Regional Faculty,
Medrano 951 (C1179AAQ) Capital Federal, Argentina

E-mail: lujan.facundo@hotmail.com

Abstract. Introduction: Knowledge Discovery in Databases (KDD) constitutes a process that allows data sets to be modeled and analyzed in an automated and exploratory manner. In this sense, data mining can be considered the main core of this procedure. Objective: In this study, a classification of clinical subjects (cluster) based on the comparison of parameters associated to cardiovascular risk factors was performed by means of KDD-based algorithms. Materials and Methods: the K-means algorithm, Hierarchical Agglomerative Clustering and Kohonen's Self-organizing Maps were applied to the database in order to obtain relationships based on the dissimilarity of its constitutive fields. Results: Four different clusters were obtained, represented by a group of well-defined clustering rules. Conclusion: KDD can be used to extract relevant data from clinical databases, which are strongly correlated with well-known cardiovascular risk markers.

1. Introduction

Knowledge Discovery in Databases (KDD) consists in the automated and exploratory analysis and modeling of big data repositories. KDD is an organized process aimed at identifying useful and novel patterns from complex datasets. In this sense, Data Mining (DM) constitutes the core of this process since it includes inference algorithms used to explore data and to understand, analyze and make predictions about the phenomena involved [1]. A subset of these algorithms may be defined as Clustering Algorithms. They have various applications ranging from data compression and vector quantization [2] to pattern discovery and recognition [3], among others. Their implementation yields a differentiated set of clusters (resulting from the application of specific rules). The notion of what constitutes a proper cluster depends on its application and there are various methods to obtain them. Such methods, in turn, must meet different criteria, both ad-hoc and systematic [4].

In biomedical terms, coronary diseases are the main cause of death worldwide. According to the world health statistics 2014 published by the World Health Organization (WHO), coronary heart disease (ischemic) is the leading cause of premature death worldwide [5]. These diseases affect the vascular conduits that irrigate the cardiac muscle (myocardium), where the ischemic event is produced by atherosclerotic obstructions of the arterial walls. This event prevents blood supply to the cardiac muscle, thus causing cell death (in ongoing situations) [6].

In the cardiovascular health, the need to obtain information for making decisions has become critical [7] together with the increasing importance of having accurate markers from a Digital Clinical History System [8]. In this regard, DM algorithms have been successfully applied in the prediction of clinical events in patients with chronic diseases [9], in the evaluation of the effectiveness of specific treatments for certain types of cancer [10], and in increasing the level of accuracy of medical evaluations (making differential diagnoses) [11], among other applications.

The main aim of this work was to obtain a series of clusters from a database (generated for the prediction of atherosclerotic events) with a specific set of clinical parameters and risk factors, by applying DM techniques. This implementation and correlation with sophisticated parameters of cardiovascular evaluation was accompanied by an analysis based on inclusion rules, which was not influenced by subjective evaluations from healthcare professionals.

2. Materials and Methods

2.1. Clustering Algorithms

The term clustering refers to the set of techniques and tools used for fractioning or partitioning data in a database. In order to do this, each cluster is grouped with those elements which are most alike and, at the same time, with those which are most different from the elements of the other clusters. In turn, the term centroid refers to the point equidistant (in Euclidean terms) to all the objects belonging to such cluster [12]. In this work, this type of algorithms has been used to establish a taxonomy of subjects from clinical parameters obtained in a non-invasive way.

2.2. k-means algorithm

Let us consider a dataset $X = \{x_1, x_2, \ldots, x_n\}$, $x_n \in \Re^d$, where the objective is to partition the dataset in M disjunct clusters $C_1, C_2, \ldots, C_M$. The k-means algorithm determines the local minimum solution to the clustering error, defined as the sum of the Euclidean distance between each point of data x_n and the center m_k of the cluster X_n to which it belongs. Analytically, the clustering error is defined as:

$$E(\boldsymbol{m_1}, \ldots, \boldsymbol{m_M}) = \sum_{i=1}^{N} \sum_{k=1}^{M} I(\boldsymbol{x_i} \in C_k) \, \|\boldsymbol{x_i} - \boldsymbol{m_k}\|^2 \quad (1)$$

where $I(Y) = 1$ if Y is true and 0 in another case.

2.3. Hierarchical Agglomerative Clustering

The idea behind Agglomerative Hierarchical Clustering (HAC) is to start with each object in a cluster of its own and then repeatedly merge the closest pair of clusters until we end up with just one cluster containing everything [13]. The basic algorithm is given in Table 1.

Table 1. Hierarchical Agglomerative Clustering Basic Algorithm

1. Assign each object to its own single-object cluster. Calculate the distance between each pair of clusters.
2. Choose the closest pair of clusters and merge them into a single cluster (so reducing the total number of clusters by one).
3. Calculate the distance between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all the objects are in a single cluster.

If there are N objects there will be N −1 mergers of two objects needed at Step 2 to produce a single cluster. However the method does not only produce a single large cluster, it gives a hierarchy of clusters as we shall see. In this case, the linkage criterion used was the decrease in variance for the cluster being merged[14].

2.4. Kohonen's Self-organizing Maps

Kohonen's self-organizing maps (SOM) are important neural network models for dimension reduction and data clustering. "Self-Organizing" is because no supervision is required. SOMs learn through unsupervised competitive learning on their own. "Maps" is because they attempt to map their weights to conform to the given input data. The nodes in a SOM network attempt to become like the inputs presented to them. In this sense, this is how they learn. SOM can learn from complex, multidimensional data and transform them into a topological map of much fewer dimensions typically one or two dimensions.

2.5. Determining the number of clusters

As is known, determining the number of clusters is one of the major problems in the use of clustering algorithms. In this context, this problem has been addressed in two directions. On the one hand, we have used the "elbow method" to determine k, where k is the number of clusters to seek the k-means algorithm.

The elbow method is based on the observation that increasing the number of clusters can help to reduce the sum of within-cluster variance of each cluster. This is because having more clusters allows one to capture finer groups of data objects that are more similar to each other. However, the marginal effect of reducing the sum of within-cluster variances may drop if too many clusters are formed, because splitting a cohesive cluster into two gives only a small reduction. Consequently, a heuristic for selecting the right number of clusters is to use the turning point in the curve of the sum of within-cluster variances with respect to the number of clusters. Technically, given a number, $k > 0$, we can form k clusters on the data set in question using a clustering algorithm like k-means, and calculate the sum of within-cluster variances, var(k). We can then plot the curve of var with respect to k. The first (or most significant) turning point of the curve suggests the "right" number [15].

As seen in Figure 1, the first (and most important) peak corresponds to four clusters. Thus, it is used $k = 4$ for implementing the k-means algorithm.

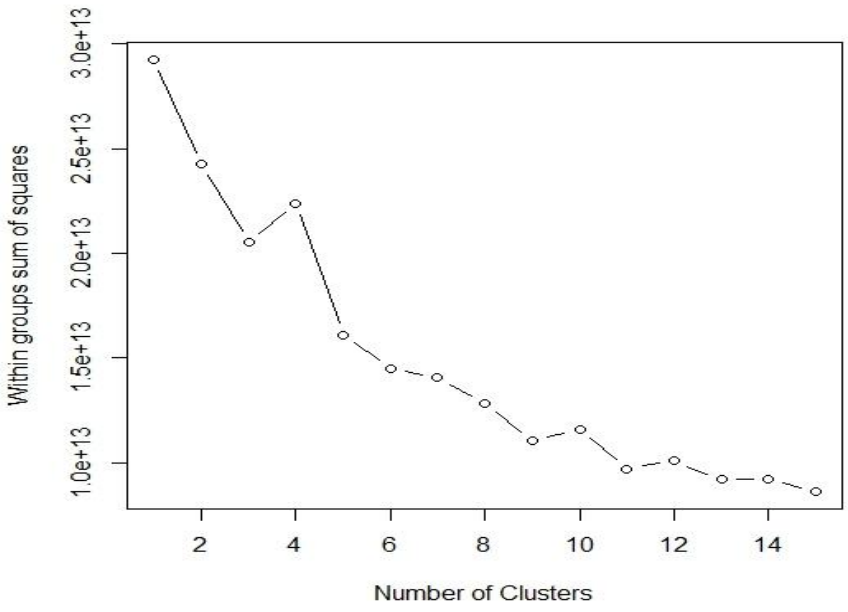

Figure 1. Number of clusters - Within groups sum of squares

On the other hand, the method for determining the number of clusters defined by the HAC algorithm has been implemented as part of the algorithm itself, taking into consideration the biggest jump in the dendrogram.

2.6. C4.5 algorithm

Once every cluster had been defined, the C4.5 algorithm was applied in order to obtain inclusion rules. This algorithm, developed by Ross Quinlan [16], is an extension of the ID3 algorithm previously developed by him. C4.5 is used to generate a *decision tree*, which takes an object or situation described by a set of attributes as input and then provides a "true/false" decision. Therefore, this set of rules defines the *natural* behavior of such input. In this study, the classification field obtained from the application of the *k-means* algorithm was defined as the input and the remaining attributes were defined as "descriptors".

2.7. TANAGRA Software

Within the currently available options, the Tanagra plat-form (Entrepôts, Représentation et Ingénierie des Connaissances, Lyon, France) was chosen because it is a free soft-ware tool used for academic and DM specialized research purposes. This platform allows multiple supervised learning algorithms to be

implemented. The main objective of the Tanagra project is to provide students and researchers with an entirely accessible DM algorithmic structure, which makes it easy to overcome the intrinsic programming problems in this domain.

2.8. Database

We have a database belonging to the French PCV-METRA cholesterol follow-up program (Prévention Cardio-Vasculaire en Médicine du Travail), a comprehensive evaluation of cardiovascular risk factors and non-invasive detection of infra-clinical atherosclerosis during a day of hospitalization. The subjects involved in the study were contacted between 1995 and 1996. None of them had a history (or symptoms) of cardiovascular disease. The database used was composed of 618 records and the clinical parameters evaluated by the algorithm are detailed in Table 2. In addition, the existing data include a quantification of cardiovascular risk based on the Framingham model (CVR) and the determination of coronary arterial calcium (CAC) [17]. Under such terms, the mean values for both parameters were obtained for the group of subjects belonging to each cluster.

Table 2. Definition of attributes

Attribute	Abbreviation	Type	Min	Max	Average	Std-dev	Std-dev /avg
Age	AGE	Numeric Ordinal	30	70	48,0162	7,6317	0,1589
Systolic Blood Pressure	SBP	Continue	100	236	137,2540	18,7956	0,1369
Diastolic Blood Pressure	DBP	Continue	50	146	87,8528	12,0556	0,1372
Low density lipoproteins	LDL	Continue	72,375	352,575	188,0516	38,2125	0,2032
High density lipoproteins	HDL	Continue	20,625	127,5	48,0803	12,2573	0,2549
Smoking habits	SMO	Binary	0	1	0,3916	0,4885	1,2475
Diabetes	DIAB	Binary	0	1	0,0518	0,2218	4,2828

All attributes described above have been converted to numeric-continuous power in order to be properly processed by the data mining algorithms. Furthermore, in order to standardize the attributes are applied to the function:

$$f(x) = \frac{x - average(x)}{\sigma} \quad (1)$$

3. Results

After the k-means algorithm, detailed in previous sections was applied (McQueen method [18]), 4 clusters composed of 187, 153, 32 and 246 data records respectively, were obtained. They are characterized by the calculated centroids, whose values are shown in Table 3. The algorithm was executed in 5 Trials and 40 as the maximum number of iterations. The 5 most significant inclusion rules for each cluster may be observed in Table 4. Thus, the mean values of CVR and CAC corresponding to each cluster are described in Table 5.

As it can be observed, the 4 clusters are clearly defined, represented by their inclusion rules, which are met by more than 80% of the cases studied.

Table 3. Clusters represented by Centroids after the application of an algorithm for data mining

Attribute	C1	C2	C3	C4
AGE	45.711	52.588	53.156	46.256
SBP	128.829	158.673	149.063	128.801
DBP	82.551	100.386	94.375	83.240
LDL	191.282	175.449	173.688	195.302
HDL	45.293	53.213	42.949	47.674
SMO	1.000	0.307	0.250	0.000
DIAB	0.000	0.000	1.000	0.000

Table 4. Inclusion Rules

Cluster No.	Rule	Effectiveness
1: Smokers	Non-diabetic Smoker Normal to moderate blood pressure	186 cases
2: Hypertensives	Non-diabetic High blood pressure	109 cases
3: Diabetics	Diabetic	32 cases
4: Normals	Non-diabetic Non-smoker Normal blood pressure	233 cases

Table 5. Mean values corresponding to Cardiovascular Risk (CVR) and Coronary Arterial Calcium (CAC)

Attribute	C1	C2	C3	C4
CVR	0.169	0.179	0.284	0.109
CAC	46.989	104.007	100.625	43.439

On the other hand, the HAC algorithm has been executed as another entry point in the data analysis process. The resulting output was four clusters with centroid that can be seen in Table 6. Furthermore, in order to compare the results of K-means and HAC, a comparative table of the classification obtained by each method has been prepared (see Table 7).

Table 6. HAC Clusters Represented By Centroids After The Application Of An Algorithm For Data Mining

Attribute	C1	C2	C3	C4
AGE	53.156	53.500	45.560	45.513
SBP	149.063	156.642	128.696	129.154
DBP	94.375	99.216	82.435	83.467
LDL	173.688	177.290	191.418	194.650
HDL	42.949	53.900	45.053	47.157
SMO	0.250	0.309	1.000	0.000
DIAB	1.000	0.000	0.000	0.000

Table 7. HAC & K-Means Clusters Cross Table

		K-MEANS				
		C1	***C2***	***C3***	***C4***	***Sum***
HAC	***C1***	0	0	32	0	32
	C2	4	144	0	14	162
	C3	183	1	0	0	184
	C4	0	8	0	232	240
	Sum	187	153	32	246	618

As a result of and by means of the cluster obtained, a hierarchy of attributes was generated, which determines inclusion of each data record in its corresponding cluster (Table 8).

Table 8. Attribute hierarchy

Attribute
Diabetes
Smoking habits
Systolic blood pressure
Diastolic blood pressure

Finally, SOM algorithm was ran with a 2x2 MAP structure, an starting learn rate of 0.2

Table 9. SOM Clusters Represented By Centroids After The Application Of An Algorithm For Data Mining

Attribute	C1	C2	C3	C4
AGE	45,663	46,366	52,471	53,156
SBP	129,144	128,829	158,242	149,063
DBP	82,711	83,142	100,346	94,375
LDL	192,992	195,246	173,450	173,688
HDL	45,203	47,690	53,299	42,949
SMO	1,000	0,000	0,307	0,250
DIAB	0,000	0,000	0,000	1,000

Table 10. SOM & K-Means Clusters Cross Table

		K-MEANS				
		C1	***C2***	***C3***	***C4***	***Sum***
SOM	***C1***	184	3	0	0	187
	C2	0	2	0	244	246
	C3	3	148	0	2	153
	C4	0	0	32	0	32
	Sum	187	153	32	246	618

4. Discussion

CVR estimation is mainly used for raising public aware-ness of the occurrence of diseases with high morbidity and mortality rates as is the case of cardiovascular diseases, over a 5 to 10-year timespan. In the present clinical practice, CVR is estimated by means of stratified tables where cardiovascular risk factors are evaluated through a scoring system for men and women separately. Equations based on the Framingham model, the SCORE algorithm and the PROCAM model, among others, constitute some of the standardized predictors for the evaluation of CVR [19]. Coronary and cerebrovascular events, as a result of atherosclerosis development, occur in a sudden and asymptomatic manner. The reduction of the occurrence of coronary and cerebrovascular events depends on early modifications of factors such as tobacco, diet and sedentary lifestyle, together with periodic controls of cholesterol, glycemia and blood pressure [20].

In addition, in this work, DM techniques were applied with the aim to cluster subjects from a clinic database (patients who attended a cardiovascular check-up) without the direct intervention of healthcare professionals. Thus, we have implemented three different clustering algorithms in order to assess the existence of a classification / hierarchy in the data. Then these results were compared to identify any relationship between them.

Thereby, observing the tables centroids K-means algorithm, HAC and SOM (Table 3, Table 6 and Table 9, respectively), and then the comparison chart between these algorithms (Table 7 and Table 10), one can observe a strong relationship between clusters.

In clinical terms, it may be inferred that there are 4 clearly differentiated clusters. Cluster 4 corresponds to healthy subjects (non-smokers, non-diabetics, with normal blood pressure). Cluster 1 is composed of smokers mainly. Cluster 3 is identified with diabetic subjects and, finally, Cluster 2 is made up of hypertensive subjects.

These results are in line with the clinical diagnoses that any general practitioner would provide [21]. However, this study shows that cluster 4, characterized by low risk subjects (see Table 5, CVR 10.9%) are also subjects with low artery calcium levels, which is a specific marker for the presence of atherosclerosis. The hypertensive group (cluster 2) mostly composed of non-smokers (70%) has a higher CVR (17.9%) as a result of their high blood pressure but with a CAC 2.5 times higher than that of normal subjects. Cluster 1, composed of 100% smokers has a CVR similar to that of hypertensives. Nonetheless, they show a slight increase in CAC, which would imply a low predisposition for coronary disease. Finally, cluster 3, composed of 100% diabetics, has the highest CVR (28.4%) as well as a high marker for atherosclerosis (CAC 100).

It was observed that, when using 3 clusters or more, there is always a group exclusively composed of diabetics. In this sense, the analysis of the results shows that diabetic subjects have physiopathological characteristics significantly different from those of a non-diabetic subject. Finally, the methodology selected has the advantage that it is not biased by the opinion of the healthcare professional, and, as a result, it is free from subjectivities that could be present in the analysis and

examination performed by the physician [22]. As a disadvantage, it must be acknowledged that this analysis requires that the number of groups to be identified by the algorithm be defined beforehand; and therefore, the ideal number of clusters is obtained through successive runs.

5. Conclusion

An original way to obtain relevant information from clinical databases is showed as feasible in this work. Such information is strongly correlated to frequently used cardio-vascular risk markers, as a comparison technique. The basic clustering analysis done here is enough to aim futures lines of research. However, further studies are required in order to optimize the determination of the number of groups as well as a better correlation between the attributes (cardiovascular risk factors) and the evaluation of such risk. Among other improvements, exists clustering schemes that could contribute in a best way to classify the clinical cases based upon sophisticated clustering algorithms.

References

[1] Bramer M 2013 *Principles of data mining* (Berlin: Springer)
[2] Gersho A and Gray R M 1992 *Vector quantization and signal compression* (Berlin: Springer)
[3] Duda R O and Hart P E 1973 *Pattern classification and scene analysis* vol 3 (New York: Wiley)
[4] Mikut R 2008 *Data Mining in der Medizin und Medizintechnik* vol 22 (Karlsruhe: KIT Scientific Publishing)
[5] World Health Organization 2014 *Estadísticas Sanitarias mundiales 2014 – una mina de información sobre salud pública mundial* Available: http://apps.who.int, last visit: 07/06/15
[6] Cabrera Fischer E 2013 *Bases de la fisiología para ingeniería* (Buenos Aires: CEIT).
[7] Bellazzi R and Zupan B 2008 *Predictive data mining in clinical medicine: current issues and guidelines (*International journal of medical informatics 77(2) 81-97)
[8] Luna D Soriano E González Bernaldo de Quirós F 2007 *Historia Clínica Electrónica* vol 27 Nº 2 (Buenos Aires: Revista del Hospital Italiano de Buenos Aires)
[9] González Bernaldo de Quirós F Luna D Baum A Plazzotta F Otero C Benítez S 2012 *Incorporación de tecnologías de la información y de las comunicaciones en el Hospital Italiano de Buenos Aires* vol 27 Nº 2 (Buenos Aires: Revista del Hospital Italiano de Buenos Aires)
[10] Reparaz D Merlino H Rancan C Rodríguez D Britos P V and García Martínez R 2008 *Determinación de la eficacia de la braquiterapia en tratamiento de cáncer basada en minería de datos* (Buenos Aires: X Workshop de Investigadores en Ciencias de la Computación)
[11] Kuo W J Chang R F Chen D R and Lee C C 2001 *Data mining with decision trees for diagnosis of breast tumor in medical ultrasonic images*. (Neatherlands: Kluwer Academic Publishers - Breast cancer research and treatment, 66(1), 51-57)
[12] Forgy E 1965 Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications (United Kingdom: Biometrics 21, 768)
[13] Kaufman L Rousseeuw P J 1990 *Finding Groups in Data: An Introduction to Cluster Analysis* (1 ed.) (New York: John Wiley. ISBN 0-471-87876-6)
[14] Ward J H Jr 1963 *Hierarchical Grouping to Optimize an Objective Function* (Massachusetts: Journal of the American Statistical Association 58 236–244)
[15] Bholowalia P and Kumar A 2014 *EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN* (Pennsylvania: International Journal of Computer Applications, 105(9))
[16] Quinlan J R 1993 *C4. 5: programs for machine learning* vol 1 (Massachusetts: Morgan kaufmann)
[17] Chironi G Simon A Megnien J-L Sirieix M-E Mousseaux E Pessana F Armentano R L 2011 *Impact of coronary artery calcium on cardiovascular risk categorization and lipid-lowering drug eligibility in asymptomatic hypercholesterolemic men* (Netherlands: International Journal of Cardiology 151: 2. 200-204)
[18] MacQueen, J. 1967 *Some methods for classification and analysis of multivariate observations*. (California: Proceedings of the fifth Berkeley symposium on mathematical statistics and

probability (Vol. 1, No. 14, pp. 281-297))
[19] Lloyd-Jones D M 2010 *Cardiovascular Risk Prediction Basic Concepts, Current Status, and Future Directions* (Philadelphia: Circulation 121, 1768–1777)
[20] World Health Organization 2007 *Prevention of Cardiovascular Disease. Guidelines for assessment and management of cardiovascular risk* Available: http://apps.who.int, last visit: 07/06/15
[21] Srinivas K Rani B K and Govrdhan A 2010 *Applications of data mining techniques in healthcare and prediction of heart attacks* (Indian: International Journal on Computer Science and Engineering (IJCSE), 2(02), 250-255)
[22] Soni J Ansari U Sharma D and Soni S 2011 *Predictive data mining for medical diagnosis: An overview of heart disease prediction* (Indian: International Journal of Computer Applications, 17(8), 43-48)