Archeologia e Calcolatori
Supplemento 6, 2014, 153-178

COMPUTER SCIENCE PROCEDURES FOR THE LABORATORY OF ANALYTICAL ARCHAEOLOGY AND ARTIFICIAL ADAPTIVE SYSTEMS (LAA&AAS)

1. Introduction. Data mining and Artificial Neural Networks

Data mining is a recently developed research field based on achievements of other disciplines, such as computer science, through research in machine learning, and statistics, through the development of multivariate and computational methods. The techniques of machine learning, in the context of the development of technologies within the field of artificial intelligence-related studies, aim at extracting from data relationships and patterns for the generation of a model of a phenomenon, which allows not only to reproduce the data generating process, but also to generalise what has been observed in relation to data not yet in our possession, to define decision support tools to formulate and validate hypotheses, to simulate scenarios of action. Data mining is an integrated set of analytical techniques, divided into various procedural steps ranging from modeling to exploration and features selection designed to extract unknown a priori knowledge from large datasets, apparently containing no regularity and important relationships

Data mining activity is not limited to the creation of query tools formulated through the SQL language or sophisticated data-retrieval tools aimed at providing multidimensional displays reports. If these tools allow to extract information from the database, the extraction process is strongly dependent on deductive research hypotheses formulated explicitly by the user. Moreover, this extraction is not inspired by modelling and synthesis provided by statistical methodology. In case the number of variables to be analysed simultaneously is high, tens or hundreds, the process of generating a hypothesis and the database analysis for the purpose of verification or falsification is no longer feasible. But data mining cannot be limited to the application of statistical methodology. Compared to the methods developed in machine learning, statistical methods suffer from the strong dependence on the data and a conceptual paradigm of reference. Although these factors have contributed to the definition of consistent and rigorous methods, they have limited the ability to quickly respond to requests advanced by methodological developments of information technology and the development of the applications of machine learning. Statisticians have initially disputed that in data mining there is not a unique theoretical model, but many competing models that are selected on the basis of test data. It fol-

lows that it would always be possible to find a model, although complex, which fits to the data. In addition, the abundance of data could lead to find non-existent relations. Although these considerations are valid, modern data mining methodologies aim to define procedures for model validation and thus for the estimate of its predictive capability, of its function as an effective decision support tool, as a tool for exploration of alternative hypotheses and possible scenarios, especially in the case where the large mass of data does not allow, by itself, to formulate a hypothesis or define a model. In this chapter we propose a graph-based data mining and an artificial neural network-based data mining.

Graphs become increasingly important in modeling complicated structures, such as circuits, images, biological networks, social networks, the Web. Many graphs search algorithms have been developed in chemical informatics, computer vision, video indexing and text retrieval. With the increasing demand on the analysis of large amounts of structured data, graph mining has become an active and important theme in data mining. Much of the data is structural in nature, or is composed of parts and relations between the parts, so a need exists to develop techniques to analyse and discover concepts in structural databases. We will propose a multidimensional data analysis based on a graph theoretic concept, called Minimum Spanning Tree. For many purposes a minimum spanning tree can capture the key essential information of a dataset. Multidimensional datasets can be represented as a minimum spanning tree without losing any essential information.

Three different Minimum Spanning Trees will be showed: in the first two cases, the metrics which defines trees properties are two well-known statistical measures, Linear Correlation and Prior Probability. The properties of the third tree are based on the innovative contribution of a new artificial neural network, the Auto Contractive Map (Auto-CM), designed by P.M. Buscema at the Semeion Research Center of Rome, Italy.

Auto-CM is an unsupervised network, a system that can learn to represent particular input patterns in a way that reflects the statistical structure of the overall collection of input patterns. By contrast with supervised learning or reinforcement learning, there are no explicit target outputs or environmental evaluations associated with each input. The goal of this kind of learning is to build representations of the input that can be used for decision making, predicting future inputs, efficiently communicating the inputs to another machine, etc. Unsupervised learning can be thought of as finding patterns in the data what would be considered pure unstructured noise. Two very simple classic examples of unsupervised learning are clustering and dimensionality reduction, the projection of input of M dimension in an output space with lesser dimension, maintaining topological properties of that input. After introducing the learning equations of Auto-CM networks, we will show a simple

application through a well-known benchmark toy dataset, referring the reader to other chapters of the volume for applications within the ARCHEOSEMA project (Ramazzotti 2012, 2013).

2. Minimum Spanning Tree (MST)

A graph is a mathematical abstraction that is useful for solving many kinds of problems. Fundamentally, a graph consists of a set of vertices, and a set of edges, where an edge is an object that connects two vertices in the graph. More precisely, a graph is a pair (V, E), where V is a finite set and E is a binary relation on V, to which it is possible to associate scalar values defined by a specific metrics. V is called a vertex set whose elements are called vertices. E is a collection of edges, where an edge is a pair (u, v) with u, v belonging to V. In a directed graph, edges are ordered pairs, connecting a source vertex to a target vertex. In an undirected graph, edges are un-ordered pairs and connect the two vertices in both directions, hence in an undirected graph (u, v) and (v, u) are two ways of writing the same edge – Minimum Spanning Trees produced by research experimental software ARCHEOSEMA Lab (see Appendix) – are visualised and manipulated by the free open source software GEPHI v. 0.8.1, an interactive visualisation and exploration platform for all kinds of networks and complex systems, dynamic and hierarchical graphs (for information and tutorials visit: https://gephi.org/).

The graph-theoretic representation is not constrained by any a priori semantic restriction: it does not say what a vertex or edge actually represents. They could be cities with connecting roads, or web-pages with hyperlinks, and so on. These semantic details are irrelevant to determine the graph structure and properties; the only thing that matters is that a specific graph may be taken as a proper representation of the phenomenon under study, to justify attention on that particular mathematical object. An adjacency-matrix representation of a graph is a 2-dimensional VxV array, where rows represent the list of vertices and columns represent edges among vertices. To each element in the array is assigned a Boolean value saying whether the edge (u, v) is in the graph. A distance matrix among V vertices represents an undirected graph, where each vertex is linked with all the others but itself (Tab. 1):

	A	B	C	D	…	Z
A	0	1	1	1	1	1
B	1	0	1	1	1	1
C	1	1	0	1	1	1
D	1	1	1	0	1	1
..	1	1	1	1	0	1
Z	1	1	1	1	1	0

Tab. 1 – Adjacency matrix.

The Minimum Spanning Tree problem is defined as follows: find an acyclic subset T of E that connects all of the vertices in the graph and whose total weight (the total distance) is minimised, where the total weight is given by:

$$d(T) = \sum_{i=0}^{N-1} \sum_{j=j+1}^{N} d_{i,j} \quad \forall d_{i,j} \tag{1}$$

T is called a spanning tree, and the MST is the T whose weighted sum of edges attains the minimum value:

$$\mathrm{Mst} = \mathrm{Min}\{\mathrm{d}(\mathrm{T_k})\} \tag{2}$$

Given an undirected graph G, representing a matrix of distances d, with V vertices, completely linked to each other, the total number of their edges (E) is:

$$E = \frac{V(V-1)}{2} \tag{3}$$

and the number of its possible spanning trees is:

$$T = V^{V-2} \tag{4}$$

The first algorithm for finding a minimum spanning tree was developed by Czech scientist Otakar Borůvka in 1926 (Borůvka 1926). Its purpose was an efficient electrical coverage of Moravia. There are now two algorithms commonly used, Prim's algorithm (Prim 1957) and Kruskal's algorithm (Kruskal 1956). All three are greedy algorithms, following the problem solving heuristic of making the locally optimal choice at each stage with the hope of finding a global optimum. In many problems, a greedy strategy does not in general produce an optimal solution, but nonetheless a greedy heuristic may yield locally optimal solutions that approximate a global optimal solution in a reasonable time. Obviously, the Kruskal algorithm generates one of the possible MSTs. In fact, in a weighted graph more than one MST is possible. Applications of MST include the design of various types of distribution networks in which the nodes represent cities, centers, etc.; and edges represent communication links (fiber glass phone lines, data transmission lines, cable TV lines, etc.), high voltage power transmission lines, natural gas or crude oil pipelines, water pipelines, highways, etc. The objective is to design a network that connects all the nodes using the minimum length of cable or pipe or other resource. The minimum cost spanning tree problem also appears as a sub-problem in algorithms for many routing problems such as the traveling salesman problem.

MST have long been used for data classification in the field of biology (States *et al.* 1993; Xu *et al.* 2001, 2002), for data classification in the field of image processing and pattern recognition (Duda, Hart 1973; Gonzales, Winz 2002; Ma *et al.* 2000) and in many other research fields (Devilliers, Dore 1989; Tapia, Rojas 2004; Assunção *et al.* 2006). A Minimum Spanning Tree is generally considered as a skeleton of a graph. For many purposes a minimum spanning tree can capture the key essential information of a graph.

Multidimensional dataset can be represented as a minimum spanning tree without losing any essential information for the purpose of clustering. From conceptual point of view, the MST represents the energy minimisation state of a structure. In fact, if we consider the atomic elements of a structure as vertices of a graph and the strength among them as the weight of each edge, linking a pair of vertices, the MST represents the minimum of energy needed so that all the elements of the structure preserve their mutual coherence. In a closed system, all the components tend to minimise the overall energy. So the MST, in specific situations, can represent the most probable state for the system to tend. To determine the MST of an undirected graph, each edge of the graph has to be weighted. So we need to define a way to weight each edge whose nodes are the entities of a dataset (records or variables).

3. Two base metrics: linear correlation and prior probability

It is possible to use any algorithm to weight the graph edges, although the final outcome will be in general quite different. It is therefore useful to review briefly some of the most used options in the current practice.

3.1 *Pearson's linear correlation*

First it is necessary to calculate the linear correlation between each pair of variables of the assigned dataset:

$$R_{i,j} = \frac{\sum_{k=1}^{N}(x_{i,k} - \bar{x}_i)\cdot(x_{j,k} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{N}(x_{i,k} - \bar{x}_i)^2 \cdot \sum_{k=1}^{N}(x_{j,k} - \bar{x}_j)^2}}; \tag{5}$$

$$-1 \le R_{i,j} \le 1; \qquad i, j \in [1,2,...,M]$$

where:

$R_{i,j}$ = linear correlation between any couple of variables x_i and x_j of the assigned dataset;
= mean value of any variable I, j;

N = number of records of the assigned dataset;
M = number of variables of the assigned dataset.

The equation (4) will generate a symmetric squared matrix with null diagonal, providing the linear correlation between each variable and any other. Through the following equation (5), the correlation matrix is transformed into a matrix of linear distances among the variables:

$$d_{i,j}^{[R]} = \sqrt{2 \cdot (1 - R_{i,j})} \tag{6}$$

At this point, following the same steps as above, the assigned dataset can be reformulated as an undirected weighted graph, where MST optimisation is applicable.

3.2 *Prior probability*

First we calculate the prior probability of co-occurrence between any couple of variables of the assigned dataset:

$$A_{i,j} = -\ln \frac{\frac{1}{N^2} \cdot \sum_{k=1}^{N} x_{i,k} \cdot (1 - x_{j,k}) \cdot \sum_{k=1}^{N} (1 - x_{i,k}) \cdot x_{j,k}}{\frac{1}{N^2} \cdot \sum_{k=1}^{N} x_{i,k} \cdot x_{j,k} \cdot \sum_{k=1}^{N} (1 - x_{i,k}) \cdot (1 - x_{j,k})} \tag{7}$$

$$-\infty \le A_{i,j} \le +\infty; \quad x \in [0,1]; \quad i, j \in [1,2,...,M]$$

where:

$A_{i,j}$ = Association strength between any couple of variables x_i and x_j of the assigned dataset;
x_i = value of any variable scaled between 0 and 1;
N = number of records of the assigned dataset;
M = number of variables of the assigned dataset.

The transformation of the matrix of association among variables into a non-linear distance matrix is generated by the following equation (8):

$$d_{i,j}^{[A]} = MaxA - A_{i,j} \tag{8}$$

where *MaxA* = Maximum *A* matrix value.

4. Advanced ANN metrics: Auto Contractive Map Network (Auto-CM)

All of the above options have the advantage to be very fast computationally, but their common, serious limit is to define the distance among variables or records by just picking them in couples. That means that each weight explains the association between two variables or records, but it does not take into ac-

count the additional influence that other variables or records could exert on that specific couple. This situation is quite similar, say, to the case of ten children playing all together in a swimming pool. If one would pretend to explain their global behaviour by making the statistics of the interaction between all possible pairs of children, this would amount to skip all of the external constraints that the concomitant positions and movements of the other children are imposing on each given couple at each given moment. By skipping this crucial information, the actual mutual behaviour of each couple will be poorly understood, and a fortiori this will also be the case for the global picture that is built through the aggregation of such partial two-by-two views.The Artificial Neural Network Auto-CM could represent the best choice to compute a complete and a non-linear matrix of weights among variables or among records of any assigned dataset.

Auto-CMs "spatialise" the correlation among datasets entities (record and variables) by constructing a suitable embedding space where a visually transparent and cognitively natural notion such as "closeness" among entities reflects accurately their associations. This "closeness" can be converted into a compelling graph-theoretic representation that picks all and only the relevant correlations and organises them into a coherent picture. Such representation is not actually constructed through some form of cumbersome aggregation of two-by-two associations between couples of entities, but rather by building a complex global picture of the whole pattern of variation (Buscema, Grossi 2007; Buscema *et al.* 2008a; Helgason *et al.* 2009; Buscema, Maurelli 2011). In recent years this technique has been applied in a number of medical settings like Alzheimer disease (Buscema *et al.* 2008b; Licastro *et al.* 2010a, 2010b), Down syndrome (Coppedè *et al.* 2010), gastro-oesophageal reflux disease (Buscema, Grossi 2008c), and myocardial infarction (Street *et al.* 2008) showing the added value of this approach in comparison with traditional statistical techniques. These techniques are novel and therefore not entirely understood so far in all of their properties and implications, and that further research is called for to explore them. But at the same time we are convinced that their actual performance in the context of well-defined, well understood problems provides an encouraging test to proceed in this direction. The Auto-CM is characterised by a three-layer architecture: an Input layer, where the signal is captured from the environment, a Hidden layer, where the signal is modulated inside the Auto-CM, and an Output layer, through which the Auto-CM feeds back upon the environment on the basis of the stimuli previously received and processed (Fig. 1). Each layer contains an equal number of N units, so that the whole Auto-CM is made of 3N units. The connections between the Input and the Hidden layers are mono-dedicated, whereas the ones between the Hidden and the Output layers are fully saturated, i.e. at maximum gradient. Therefore, given N units, the total number of the connections, Nc, is given by:

$$Nc = N (N + 1).$$

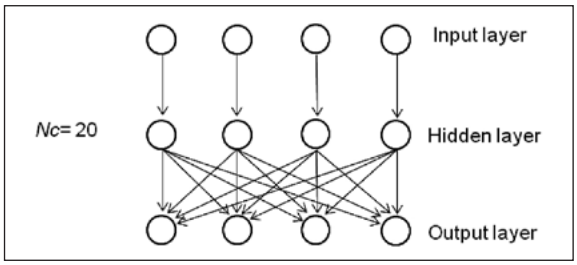

Fig. 1 – An example of an Auto-CM with N = 4.

All of the connections of Auto-CM may be initialised either by assigning a same, constant value to each, or by assigning values at random. The best practice is to initialise all the connections with a same, positive value, close to zero. The learning algorithm of Auto-CM may be summarised in a sequence of four characteristic steps:

1) Signal Transfer from the Input into the Hidden layer;
2) Adaptation of the values of the connections between the Input and the Hidden layers;
3) Signal Transfer from the Hidden into the Output layer;
4) Adaptation of the value of the connections between the Hidden and the Output layers.

We write as $m^{[s]}$ the units of the Input layer (sensors), scaled between 0 and 1; as $m^{[h]}$ the units of the Hidden layer, and as $m^{[t]}$ the units of the Output layer (system target). We moreover define **v**, the vector of mono-dedicated connections; **w**, the matrix of the connections between the Hidden and the Output layers; and *n*, the discrete time that spans the evolution of the Auto-CM weights, or, put another way, the number of cycles of processing, counting from zero and stepping up one unit at each completed round of computation.

In order to specify the steps 1-4 that define the Auto-CM algorithm, we have to define the corresponding signal forward-transfer equations and the learning equations, as follows:

1) Signal transfer from the Input to the Hidden layer:

$$m_{i_{(n)}}^{[h]} = m_i^{[s]} \cdot \left(1 - \frac{v_{i_{(n)}}}{C}\right) \tag{9}$$

where *C* is a positive real number not lower than 1, which we will refer to as the contraction parameter, and where the *(n)* subscript has been omitted from the notation of the input layer units, as these remain constant at every cycle of processing.

2) Adaptation of the connections through the variation, which amounts to trapping the energy difference generated according to equation (9):

$$\Delta v_{i_{(n)}} = (m_i^{[s]} - m_{i_{(n)}}^{[h]}) \cdot \left(1 - \frac{v_{i_{(n)}}}{C}\right) \tag{10}$$

$$v_{i_{(n+1)}} = v_{i_{(n)}} + \Delta v_{i_{(n)}} \tag{11}$$

3) Signal transfer from the Hidden to the Output layer:

$$Net_{i_{(n)}} = \sum_{j=1}^{N} m_{j_{(n)}}^{[h]} \cdot \left(1 - \frac{w_{i,j_{(n)}}}{C}\right) \tag{12}$$

$$m_{i_{(n)}}^{[t]} = m_{i_{(n)}}^{[h]} \cdot \left(1 - \frac{Net_{i_{(n)}}}{C}\right) \tag{13}$$

4) Adaptation of the connections through the variation, which amounts, accordingly, to trapping the energy difference as to equation (13):

$$\Delta w_{i,j_{(n)}} = (m_{i_{(n)}}^{[h]} - m_{i_{(n)}}^{[t]}) \cdot \left(1 - \frac{w_{i_{(n)}}}{C}\right) \cdot m_{j_{(n)}}^{[h]} \tag{14}$$

$$w_{i_{(n+1)}} = w_{i_{(n)}} + \Delta w_{i,j_{(n)}} \tag{15}$$

Even a cursory comparison of (9) and (13) and (10-11), (14-15), respectively, clearly shows how both steps of the signal transfer process are guided by the same (contraction) principle, and likewise for the two weight adaptation steps (for which we could speak of an energy entrapment principle). Notice how the term in (14) makes the change in the connection proportional to the quantity of energy liberated by node in favour of node. The whole learning process, which essentially consists of a progressive adjustment of the connections aimed at the global minimisation of energy, may be seen as a complex juxtaposition of phases of acceleration and deceleration of velocities of the learning signals (adaptations and) inside the ANN connection matrix (Buscema, Grossi 2007). There are a few important peculiarities of Auto-CMs with respect to more familiar classes of ANNs that need special attention and call for careful reflection:

1) Auto-CMs are able to learn also when starting from initialisations where all connections are set at the same value, i.e., they do not suffer the problem of the symmetric connections.
2) During the training process, Auto-CMs always assign positive values to connections. In other words, Auto-CMs do not allow for inhibitory relations among nodes, but only for different strengths of excitatory connections.

3) Auto-CMs can learn also in difficult conditions, namely, when the connections of the main diagonal of the second layer connection matrix are removed. In the context of this kind of learning process, Auto-CMs seem to reconstruct the relationship occurring between each couple of variables. Consequently, from an experimental point of view, it seems that the ranking of its connections matrix translates into the ranking of the joint probability of occurrence of each couple of variables.
4) Once the learning process has occurred, any input vector, belonging to the training set, will generate a null output vector. So, the energy minimisation of the training vectors is represented by a function by means of which the trained connections "absorb" completely the input training vectors. Thus, Auto-CM seems to learn how to transform itself in a "dark body"
5) At the end of the training phase the matrix **w**, then, represents the Auto-CM knowledge about the whole dataset.

For our purpose, the matrix **w** may be transformed into a non-Euclidean distance metric (semi-metric), when we train the Auto-CM with the main diagonal of the **w** matrix fixed at value N. Now, if we consider N as a limit value for all the weights of the w matrix, we can write:

$$d_{i,j} = N - w_{i,j} \tag{16}$$

The new matrix *d* is again a squared symmetric matrix, where the main diagonal entries are null (i.e., they represent the zero distance of each node from itself), and where the off-diagonal entries represent "distances" between each couple of nodes. Each distance between a pair of nodes may therefore be regarded as the weighted edge between these pair of nodes in a suitable graph-theoretic representation, so that the matrix *d* itself may be analysed through the graph theory.

5. Applications: gang toy dataset

5.1 *The dataset*

In this section we propose a validation of the tools discussed. The validation is based on a well-known benchmark introduced by McClelland and Rumelhart (1986, 1988), the *West Side Story* dataset where one has to distinguish members' affiliation in two rival gangs, the Jets and the Sharks, on the basis of certain numbers of identifying characteristics. This is a demanding benchmark in that characteristics are mixed up in a rather tricky way: Jets tend to be in their 20s, single and with a Junior High School education, although no one Jet member actually happens to meet all three criteria at the same time, whereas Sharks tend to be older, married and with a High School education, but again no one Shark happens to meet the three criteria simulta-

	Gang	Age	Education	Status	Profession
ART	Jet	40	Junior School	Single	Pusher
AL	Jet	30	Junior School	Married	Burglar
SAM	Jet	20	College	Single	Bookie
CLYDE	Jet	40	Junior School	Single	Pusher
MIKE	Jet	30	Junior School	Single	Pusher
JIM	Jet	20	Junior School	Divorced	Burglar
GREG	Jet	20	High School	Married	Pusher
JOHN	Jet	20	Junior School	Married	Burglar
DOUG	Jet	30	High School	Single	Bookie
LANCE	Jet	20	Junior School	Married	Burglar
GEORGE	Jet	20	Junior School	Divorced	Burglar
PETE	Jet	20	High School	Single	Bookie
FRED	Jet	20	High School	Single	Pusher
GENE	Jet	20	College	Single	Pusher
RALPH	Jet	30	Junior School	Single	Pusher
PHIL	Sharks	30	College	Married	Pusher
IKE	Sharks	30	Junior School	Single	Bookie
NICK	Sharks	30	High School	Single	Pusher
DON	Sharks	30	College	Married	Burglar
NED	Sharks	30	College	Married	Bookie
KARL	Sharks	40	High School	Married	Bookie
KEN	Sharks	20	High School	Single	Burglar
EARL	Sharks	40	High School	Married	Burglar
RICK	Sharks	30	High School	Divorced	Burglar
OL	Sharks	30	College	Married	Pusher
NEAL	Sharks	30	High School	Single	Bookie
DAVE	Sharks	30	High School	Divorced	Pusher

Tab. 2 – Gang dataset.

Item	Jets	Sharks	Jets(%)	Sharks(%)
20s	9	1	60.00	8.33
30s	4	9	26.67	75.00
40s	2	2	13.33	16.67
Junior School	9	1	60.00	8.33
High School	4	17	26.67	58.33
College	2	4	13.33	33.33
Single	9	4	60.00	33.33
Married	4	6	26.67	50.00
Divorced	2	2	13.33	16.67
Pusher	5	4	33.33	33.33
Bookie	5	4	33.33	33.33
Burglar	5	4	33.33	33.33

Tab. 3 – Gang dataset basic statistics.

neously. Moreover, all members of both gangs are equally likely to operate as pushers, bookies or burglars. We will see how the tools discussed in previous sections could provide an in-depth exploration of the structural properties of the dataset, then very useful information for a number of further tasks, including, for instance, analysis of clusters, scenario simulation and so on.

The Tab. 2 shows this dataset. The basic statistics is reported in Tab. 3.

5.2 *Preprocessing*

Each categorical variable (Gang, Age, Education, Status and Profession) is transformed using dummy variables: each item is transformed in a Boolean value, 1 or 0. This choice is particularly effective when the output is a graph or a neural network based on activations of nodes/entities. Tab. 4 shows the transformed dataset.

	Jet	Sharks	20	30	40	JH	COL	HS	Sing	Marr	Div	Push	Book	Burg
ART	1	0	0	0	1	1	0	0	1	0	0	1	0	0
AL	1	0	0	1	0	1	0	0	0	1	0	0	0	1
SAM	1	0	1	0	0	0	1	0	1	0	0	0	1	0
CLYDE	1	0	0	0	1	1	0	0	1	0	0	0	1	0
MIKE	1	0	0	1	0	1	0	0	1	0	0	0	1	0
JIM	1	0	1	0	0	1	0	0	0	0	1	0	0	1
GREG	1	0	1	0	0	0	0	1	0	1	0	1	0	0
JOHN	1	0	1	0	0	1	0	0	0	1	0	0	0	1
DOUG	1	0	0	1	0	0	0	1	1	0	0	0	1	0
LANCE	1	0	1	0	0	1	0	0	0	1	0	0	0	1
GEORGE	1	0	1	0	0	1	0	0	0	0	1	0	0	1
PETE	1	0	1	0	0	0	0	1	1	0	0	0	1	0
FRED	1	0	1	0	0	0	0	1	1	0	0	1	0	0
GENE	1	0	1	0	0	0	1	0	1	0	0	1	0	0
RALPH	1	0	0	1	0	1	0	0	1	0	0	1	0	0
PHIL	0	1	0	1	0	0	1	0	0	1	0	1	0	0
IKE	0	1	0	1	0	1	0	0	1	0	0	0	1	0
NICK	0	1	0	1	0	0	0	1	1	0	0	1	0	0
DON	0	1	0	1	0	0	1	0	0	1	0	0	0	1
NED	0	1	0	1	0	0	1	0	0	1	0	0	1	0
KARL	0	1	0	0	1	0	0	1	0	1	0	0	1	0
KEN	0	1	1	0	0	0	0	1	1	0	0	0	0	1
EARL	0	1	0	0	1	0	0	1	0	1	0	0	0	1
RICK	0	1	0	1	0	0	0	1	0	0	1	0	0	1
OL	0	1	0	1	0	0	1	0	0	1	0	1	0	0
NEAL	0	1	0	1	0	0	0	1	1	0	0	0	1	0
DAVE	0	1	0	1	0	0	0	1	0	0	1	1	0	0

Tab. 4 – Gang dataset preprocessed.

5.3 *Minimum Spanning Trees*

In this section we show the MST outputs based on the Auto-CM metrics. The Auto-CM weights matrix was calculated by the research experimental software *ARCHEOSEMA Lab* (see Appendix A).

The MST output file is a *.graphml type file and was visualised with the open source software GEPHI v. 0.8.1. We propose two types of trees, each based on different input of the artificial neural network Auto-CM:

1) Variables Tree: input corresponding to the matrix of Tab. 4. Nodes represent variables.
2) Records Tree: the input is transposed matrix of Tab. 4. Nodes represent records. Our trees analysis is based on three specific statistics measures that maintain their meaning even MSTs are acyclic graphs:
3) *Betweenness centrality*. It is a centrality measure of a vertex within a graph and quantifies the number of times a node acts as a bridge along the

shortest path between two other nodes. It was introduced as a measure for quantifying the control of a human on the communication between other humans in a social network by Linton Freeman. In his conception, vertices that have a high probability to occur on a randomly chosen shortest path between two randomly chosen vertices have a high betweenness (Freeman 1977). Betweenness centrality is a more useful measure of the load placed on the given node in the network as well as the node's importance to the network than just connectivity. The latter is only a local effect while the former is more global to the network. In a weighted network the links connecting the nodes are no longer treated as binary interactions, but are weighted in proportion to their capacity, influence, frequency, etc., which adds another dimension of heterogeneity within the network beyond the topological effects.
4) *Mean Weighted Degree of Nodes*: the mean weight of connections per node on the graph.
5) *Modularity*. It is one measure of the structure of networks or graphs. It was designed to measure the strength of division of a network into clusters. Networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules. Modularity is the fraction of the edges that fall within the given groups minus the expected such fraction if edges were distributed at random. The value of the modularity lies in the range [–1/2, 1). It is positive if the number of edges within groups exceeds the number expected on the basis of chance. For a given division of the network's vertices into some modules, modularity reflects the concentration of nodes within modules compared with random distribution of links between all nodes regardless of modules. The resolution coefficient is the main parameter to be set: a value higher than 1.0 (default) provides a minor number of clusters and then bigger ones (Newman 2006, 2007; Reichardt, Bornholdt 2006).

5.4 *Variables MST*

Fig. 2 shows the variables Minimum Spanning Tree. Values of connections between nodes correspond to the values of Auto-CM connections matrix and their thickness is proportional to the value of the connection. We can note the following:

1) *the separation between the two classes*, *Jets and Sharks*, *is clear*; the connections with the class nodes indicate to some extent the prototype of the classes themselves. If we compare the tree with data of Tab. 3, we can see that Age is one of the attributes that most distinguishes the two classes: one Shark is a thirty year old subject with a probability of 70%, while a Jet is twenties with a probability of 60% (note the maximum value of connections in both cases). Similarly, it is more likely that a Shark subject attends the High School, as well as a Jet subject attends the Junior School. Furthermore, a Jet

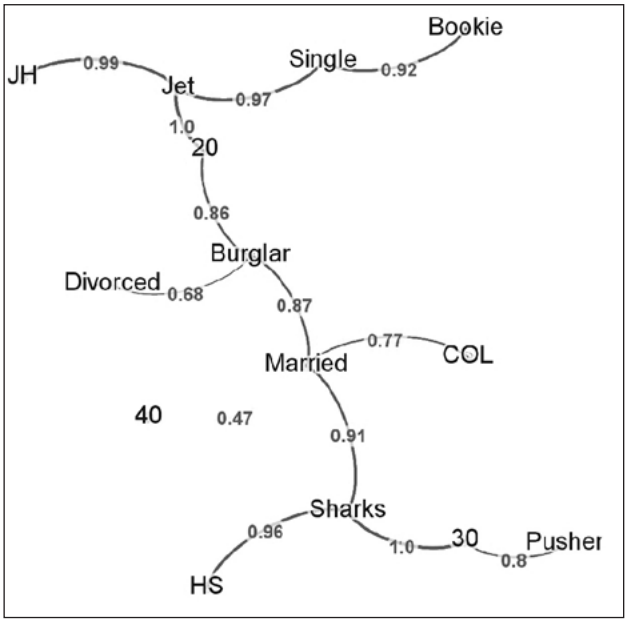

Fig. 2 – Gang variables MST.

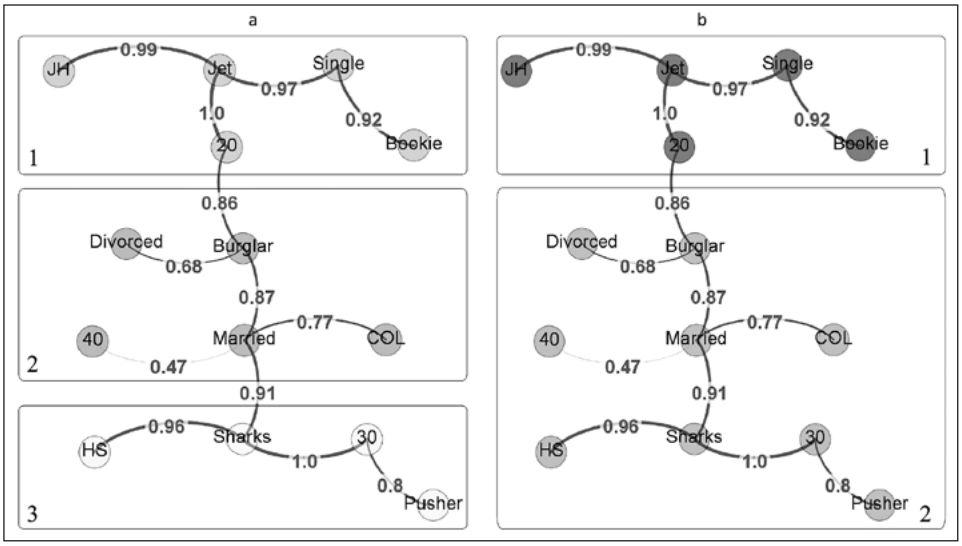

Fig. 3 – Gang variables MST: a) Resolution coefficient = 1; b) Resolution coefficient = 3.

subject is likely to be Single, while the fact that a Shark subject is Married is less likely, as this can be inferred from the fact that the node Married is linked to the node Sharks in a different way compared to the link between Jet and Single: it is in fact also connected to other nodes and can be regarded as a "transition" attribute between two classes;
2) the value of the connection between nodes 40 and Married is very low. In fact in the dataset there are 40 year old subjects who are both Single and Married, so this feature is not discriminating. However, the position of node 40 indicates that this attribute is typically a Sharks attribute. This idea could be deduced also from the distance, in terms of the sum of the linking connections values, of this node from two classes: 40-Sharks = 1,38, 40-Jet = 3.2;
3) the position of the node College points that this attribute is not discriminating as the attributes Age and Education. However, the College-Married connection is typically Sharks.
4) the Bookie attribute is typically Sharks, as a Bookie is typically a Single and this attribute, as already mentioned, is typically Shark.
5) the attribute Burglar is not discriminating. We find this feature in the tree, as a transition node between the two classes. From the dataset we can see that it is more likely that a Burglar is Married and that Divorced is a Burglar. In addition, the meaning of Burglar node as transition attribute is confirmed by its distances from nodes of the classes: Burglar-Jet = 1.86 while Burglar-Sharks = 1.78.

From this first analysis, we can conclude that this Minimum Spanning Tree has not only shown in a compact manner and immediately some relationships that were already inherent in the dataset, but it also has proposed not immediately noticeable solutions, available only through data manipulations. If these manipulations pose no relevant problem when working with small dataset, this feature of the MST based on matrix of connections Auto-CM becomes more relevant when working with dataset of more significant size. Fig. 3 shows GEPHI v. 0.8.1 clustering based on modularity resolution coefficient = 1 and modularity resolution coefficient = 3. Fig. 4 shows the nodes weighted degree statistics (the size of labels is proportional to the weighted degree values) and Tab. 5 reports nodes degree and weighted degree values ordered by weighted degree.

The modularity algorithm based on a resolution coefficient = 3 has detected two clusters, while three clusters have been detected with coefficient = 1. These results once again demonstrate not only the complexity of the dataset, but also the potentiality of the methodology applied. If the proposed solution with a resolution = 3 seems to identify the two classes and confirm our expectations in some way, the solution of the three clusters is definitely the most interesting. The clusters 3 and 1 are those with strong characterisation, while the central cluster 2 can be considered as a transition zone between the two classes's, identifying those attributes whose classes membership is not

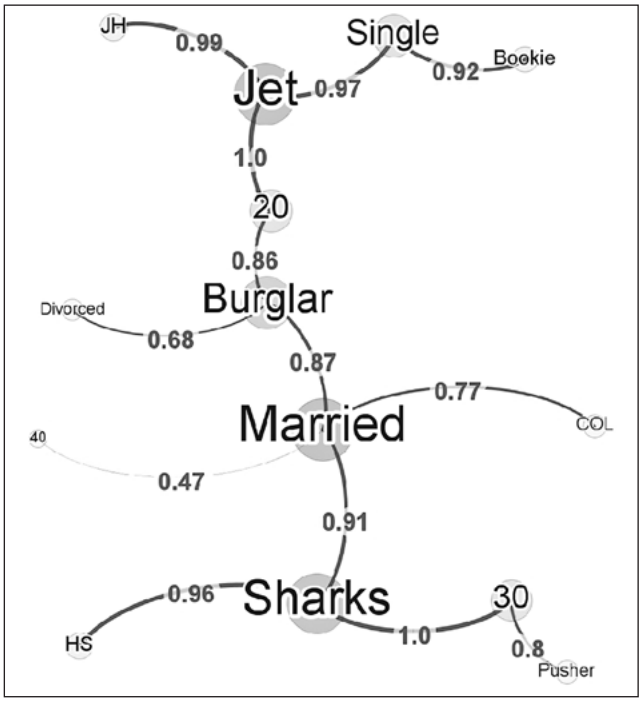

Fig. 4 – Gang variables MST (nodes weighted degree).

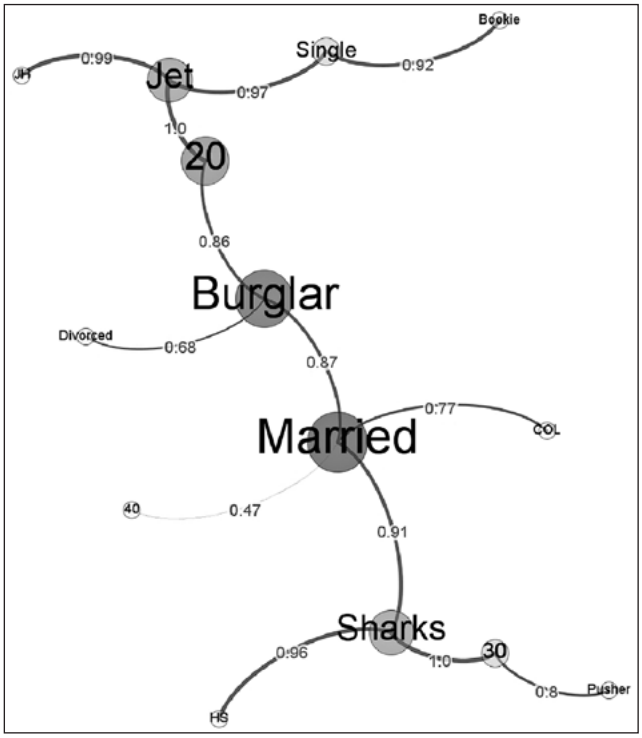

Fig. 5 – Gang variables MST (betweenness centrality).

Node	Degree	Weighted Degree
Married	4	3.02
Jet	3	2.96
Sharks	3	2.87
Burglar	3	2.41
Single	2	1.89
20	2	1.86
30	2	1.80
JH	1	0.99
HS	1	0.96
Bookie	1	0.92
Pusher	1	0.80
COL	1	0.77
Divorced	1	0.68
40	1	0.47

Tab. 5 – Nodes degree and weighted degree values ordered by weighted degree.

Node	Betweenness Centrality
Married	0.65
Burglar	0.60
20	0.46
Jet	0.41
Sharks	0.41
30	0.15
Single	0.15
40	0.00
JH	0.00
COL	0.00
HS	0.00
Divorced	0.00
Pusher	0.00
Bookie	0.00

Tab. 6 – Nodes centrality values ordered by centrality.

so obvious. The weighted degree statistics confirms not only the importance of nodes Jets and Sharks, as cluster centers, but also the importance of node Married as attribute of transition from one class to another (Tabs. 5, 6).

5.5 Records MST

Fig. 6 shows the MST of records. The labels contain, between parentheses, the membership class ID of records. Topological separation of the two classes is evident. The closely related records are very similar and may define additional groups within the main class. In Figs. 7a-b results are analysed in terms of the Modularity Statistics.

The separation of records in two classes is stably supported with values of the resolution from 8 to 4 (Fig. 7a). But even in this case it is much more interesting to analyse the results with lower resolution values which provide the finest hypothesis of possible sub-clusters. In particular, with resolution = 3 (Fig. 7b) we get three clusters: the cluster of Shark remained substantially unchanged (except for the Ike node), while the Jets cluster is divided into two sub-cluster, with the cluster in blue as a cluster of transition between the two classes and nodes Ike and Mike as those records that contain, at this level of resolution, "ambiguous" attributes. With resolution = 2 (Fig. 7c) a change of membership of three Sharks subjects, Neal, Ken and Nick occurs: now they belong to a new cluster together four subjects of the cluster detected by resolution = 3. This new cluster is separated into two sub-clusters with resolution = 1 (Fig. 7d): in this new configuration are 5 clusters. The dataset complexity is still confirmed.

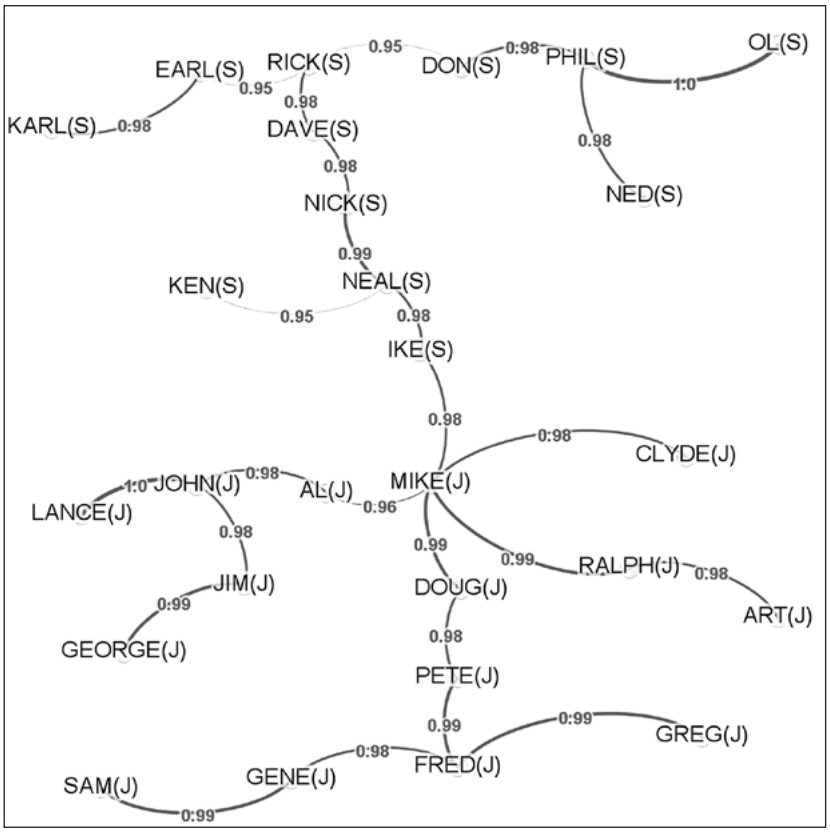

Fig. 6 – Gang records MST.

From comparison of these results we could detect two basic clusters:

1) the cluster composed of Karl, Earl, Rick, Daves, Don, Phil, Ned and Ol: we can assume that it is in the cluster that defines the differentiating characteristics of Sharks subjects;
2) the cluster composed of Doug, Pete, Sam, Gene, Fred and Greg: we can assume that it is in the cluster that defines the differentiating characteristics of Jets subjects.

Fig. 8 shows the graph of Betweenness Centrality statistics (the size of nodes and labels is proportional to the value of betweennes). Tab. 7 lists the values for each node in a descending order.

Fig. 9 shows the graph of Weighted Degree statistics (the size of nodes and of labels is proportional to the value of the degree). Tab. 8 lists the values for each node in a descending order.

Id	Betweenness Centrality
MIKE(J)	0.716923077
NEAL(S)	0.52
IKE(S)	0.507692308
NICK(S)	0.443076923
DAVE(S)	0.409230769
RICK(S)	0.393846154
DOUG(J)	0.323076923
AL(J)	0.270769231
PETE(J)	0.270769231
FRED(J)	0.218461538
JOHN(J)	0.218461538
DON(S)	0.212307692
PHIL(S)	0.150769231
GENE(J)	0.076923077
JIM(J)	0.076923077
RALPH(J)	0.076923077
EARL(S)	0.076923077
SAM(J)	0
GREG(J)	0
LANCE(J)	0
GEORGE(J)	0
KEN(S)	0
ART(J)	0
CLYDE(J)	0
NED(S)	0
KARL(S)	0
OL(S)	0

Tab. 7 – Nodes centrality values ordered by centrality.

Id	Weighted Degree
MIKE(J)	4.900000036
FRED(J)	2.960000038
JOHN(J)	2.960000038
PHIL(S)	2.960000038
NEAL(S)	2.920000017
RICK(S)	2.879999995
NICK(S)	1.970000029
DOUG(J)	1.970000029
PETE(J)	1.970000029
GENE(J)	1.970000029
JIM(J)	1.970000029
RALPH(J)	1.970000029
IKE(S)	1.960000038
DAVE(S)	1.960000038
AL(J)	1.939999998
DON(S)	1.930000007
EARL(S)	1.930000007
LANCE(J)	1
OL(S)	1
SAM(J)	0.99000001
GREG(J)	0.99000001
GEORGE(J)	0.99000001
ART(J)	0.980000019
CLYDE(J)	0.980000019
NED(S)	0.980000019
KARL(S)	0.980000019
KEN(S)	0.949999988

Tab. 8 – Nodes centrality values ordered by weighted degree.

From these analyses it is possible to conclude that the node Mike, first in both statistics, is a very significant node in the tree: Mike is central but is also the node whose the sum of links values to other nodes is the highest, and then we can assign the role of transitional element between the two classes (as the node whose distance from others is the lowest ever).

6. Conclusions

In this chapter we have proposed a data mining technique aiming to explore data in search of consistent patterns, systematic relationships and hidden associations between dataset entities (variables or records). This technique is based on unsupervised artificial neural network and provides a graph data mining methodology: a new paradigm of entities mapping aiming to create a sort of semantic connectivity map in which:

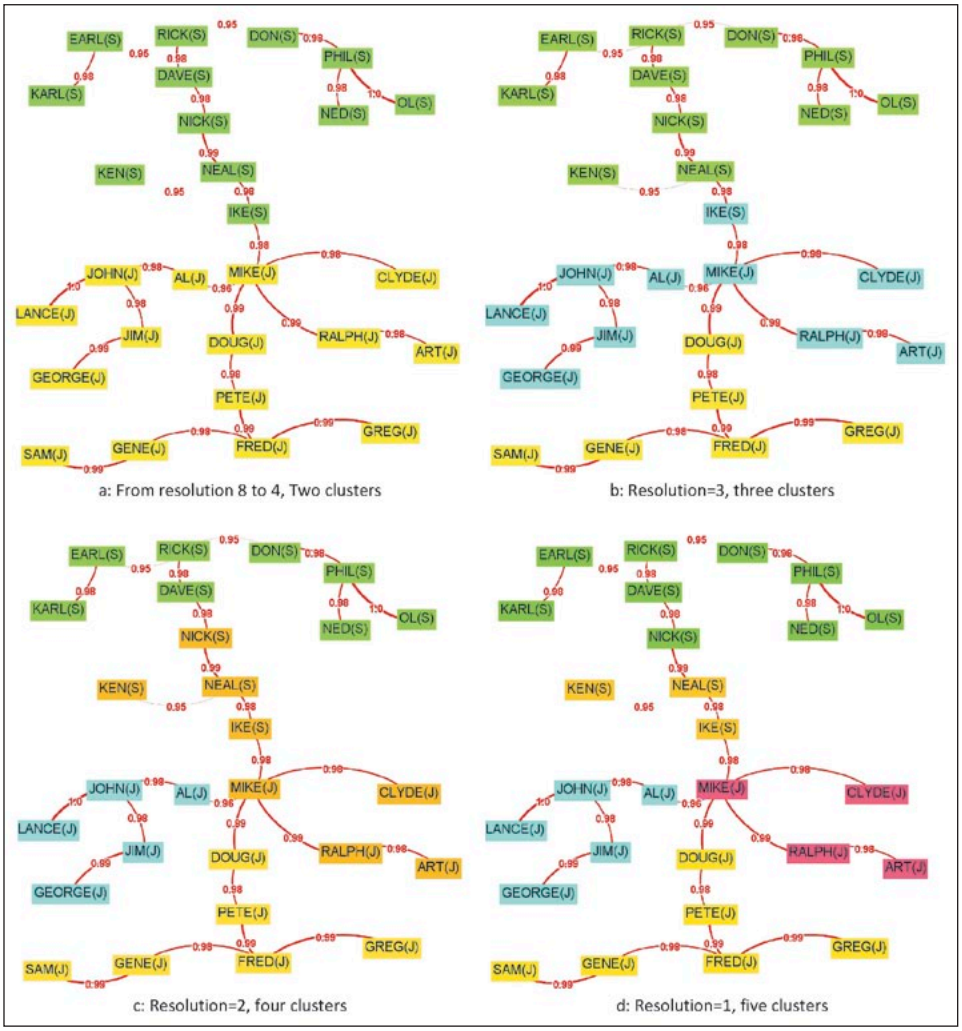

Fig. 7 – Gang records MST (modularity statistics).

1) non-linear associations are preserved;
2) there are explicit connection schemes;
3) the complex dynamics of adaptive interactions is captured.

The weights matrix of the artificial neural network Auto-CM represents the warped landscape of the whole dataset. The MST represents at this point a simple filter to apply to the weights matrix of Auto-CM system to show visually the map of the main connections of the entities of the dataset and

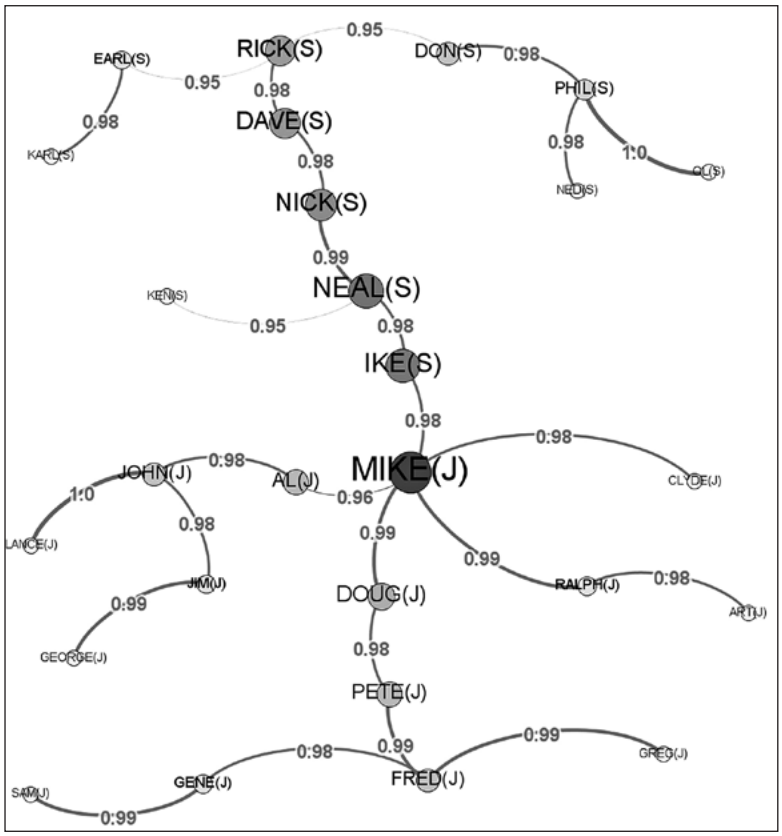

Fig. 8 – Gang records MST (betweenness Centrality statistics).

the basic semantic of their similarities. Data processing of ARCHEOSEMA project datasets, introduced in subsequent chapters of this volume, highlights the potentiality of this methodology also in research fields considerered as traditionally humanistic, for the definition of procedures of construction and validation of an interpretative model, the evaluation of its performance, its use as a valuable tool to support the exploration of alternative hypotheses and scenarios, especially in the case where hypothesis or an interpretative model cannot be directly inferred from data.

For this purpose, alongside non-supervised neural networks here used, supervised neural networks could be used in order to build models of analysis and classification. In supervised training, both the inputs and the outputs are provided. The network then processes the inputs and compares its resulting outputs against the desired outputs. Errors are then propagated back through

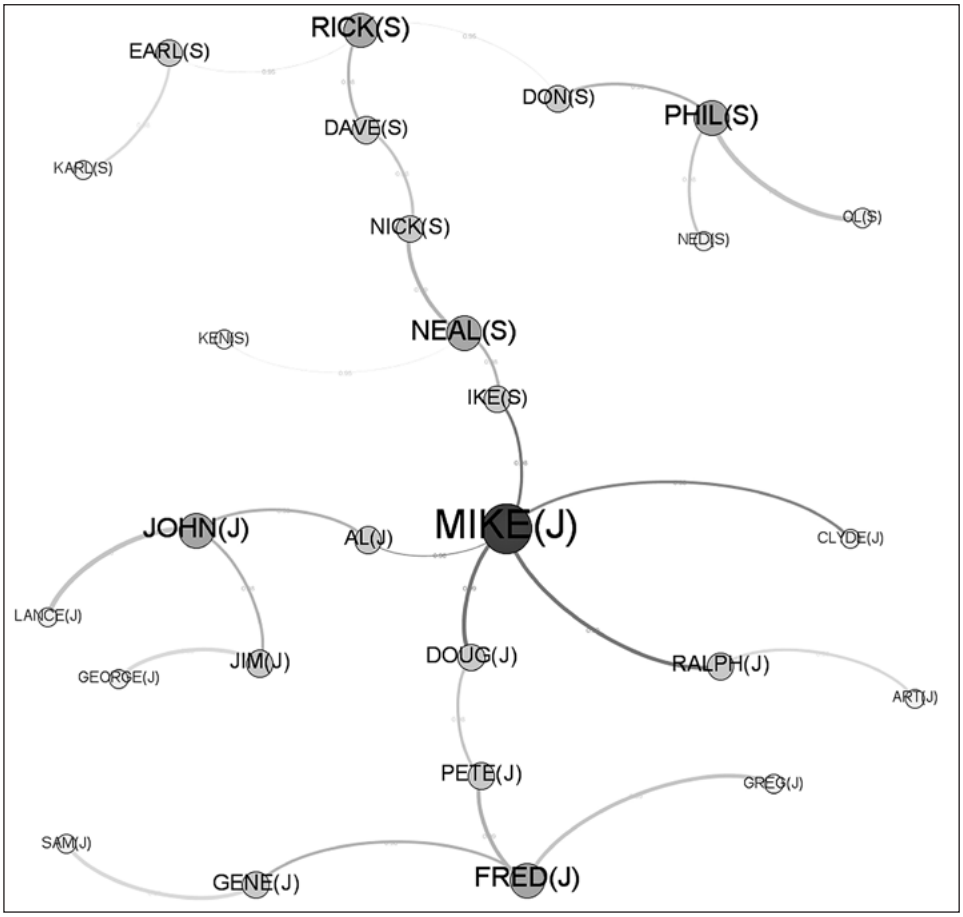

Fig. 9 – Gang records MST (weighted degree statistics).

the system, causing the system to adjust the weights which control the network. This process occurs over and over as the weights are continually tweaked. The set of data which enables the training is called the "training set". During the training of a network the same set of data is processed many times as the connection weights are ever refined. Examples of classic and important research problems that can be addressed with this networks are problems of classification/assignment 1 of n, as long as it is possible to conceptualise and define formally an input dataset, target variables, procedures of data preprocessing and a protocol of validation of the model performance.

It is important however to emphasise that the methodology proposed is not in competition with more traditional data mining techniques, it is

neither an absolute alternative, nor a substitute for the expert knowledge. Researchers should use it as a tool to support their activities and assess the potentiality, even in the context of an interdisciplinary cooperation as increasingly and indispensable prerequisite within the framework of scientific activities starting from the very first stages of data collection, of formalisation of the reference database, of optimisation of stored data access procedures until the formulation of a research problem, the choice of one or more analytical models among those available, procedures for extracting a dataset to represent adequately the problem and to satisfy the requirements of the model. From a strictly epistemological point of view, the validation of the results of research methodologies proposed here will have to meet some basic criteria:

1) In case the results reproduce only what is already known in the literature, the model should be assessed in terms of implementation effectiveness and usability immediacy. If the alternative is a method that involves the application of different procedures of filtered data extraction or different statistical procedures that produce complex reports, then the model proposed here retains a significant advantage.
2) In a initial phase the proposed model cannot completely refute the guidelines of the research community, but necessarily it will contains elements of continuity. The model may suggest, to the extent established by the expert, alternative hypotheses that "explain" the phenomenon under investigation although they are based upon assumptions at the moment unfounded and/or in contradiction to what is supported by the scientific community.
3) Consequences in principle "observable" and "controllable" should be deduced from alternative hypotheses. An hypothesis should suggest some requirements or contraints, strictly experimental or not, for a "predicted" phenomenon by the model to occur or not. Controllability does not mean verifiability: the control procedures may also result in a refutation of the hypothesis as crucial procedure of importance equal to that of a verified hypothesis.
4) If we use different procedures in order to control a hypothesis, it should be necessary to evaluate the degree of convergence of the results. Although total convergence is rarely achievable, comparing the procedure can result in greater awareness of the nature and importance of each procedure parameters and the need to define the problem in a more appropriate way. There is no single and incontrovertible model that returns a unique solution to a problem. On the other hand, there is a research activity that has to deal with the development of different methodologies, to assess its ability to contribute to the development of a discipline through the reformulation of classical problems from a most promising perspective and, in the same time, through the formulation of new and unexpected research problems.

Appendix ARCHEOSEMA Lab software

All elaborations reported were performed with the research software ARCHEOSEMA Lab v 1.0, written in C ++ for Windows. The software has the following functions:

1) Calculation of connections matrices for the metrics Linear Correlation, Prior Probability e Auto Contractive Map Artificial Neural Network;

2) Definition of the Minimum Spanning Tree for each metric;

3) Minimum Spanning Tree output files in a graphml format for the visualisation and manipulation through the open source software GEPHI v. 0.8.1.

A further phase of development involves the implementation of the unsupervised neural network Self Organising Map (Kohonen 2001) and the supervised neural network Back Propagation (McClelland, Rumelhart *et al.* 1986; Werbos 1994).

Massimiliano Capriotti
Semeion Research Center
LAA&AAS
Sapienza Università di Roma

REFERENCES

Assuncão R.M., Neves M.C., Camara G., Da Costa Freitas C. 2006, *Efficient regionalization techniques for socioeconomic geographical units using minimum spanning trees,* «International Journal of Geographical Information Science», 20/7, 797-811.

Borůvka O. 1926, *Příspěvek k řešení otázky ekonomické stavby elektrovodních sítí* (*Contribution to the solution of a problem of economical construction of electrical networks*), «Elektronický Obzor», 15, 153-154.

Buscema P.M., Grossi E. 2007, *A novel adapting mapping method for emergent properties discovery in data bases: experience in medical field*, in Y. Nakamori, Z. Wang, J. Gu, T. Ma (eds.) *Proceedings of International Conference on Systems, Man and Cybernetics, (Montreal 2007)*, Institute of Electrical and Electronics Engineers Omnipress, 3457-3463.

Buscema P.M., Petritoli R., Pieri G., Sacco P.L. 2008a, *Auto-Contractive Maps*, Technical Paper, 32, Roma, Aracne Editrice.

Buscema P.M., Grossi E., Snowdon D., Antuono P. 2008b, *Auto-Contractive Maps: an Artificial Adaptive System for Data Mining. An Application to Alzheimer Disease,* «Current Alzheimer Research», 5, 481-98.

Buscema P.M., Grossi E. 2008c, *The semantic connectivity map: an adapting self-organising knowledge discovery method in data bases. Experience in gastro-oesophageal reflux disease*, «International Journal of Data Mining and Bioinformatics», 2/4, 362-404.

Buscema P.M., Maurelli G. 2011, *Matematica e processi naturali. Nuovi modelli di sistemi artificiali adattivi per l'analisi dei fenomeni complessi*, Roma, Aracne Editrice.

Buscema P.M., Sacco P.L. 2010, *Activation and Competition System and Universe Lines Algorithm*, «2010 Annual Meeting of the North American Fuzzy Information Processing Society», 1-7.

Coppedè F., Grossi E., Migheli F., Migliore L. 2010, *Polymorphisms in folate-metabolizing genes, chromosome damage, and risk of Down syndrome in Italian women: identification of key factors using artificial neural networks*, «BMC Medical Genomics», 3, 42.

Devillers J., Dore J.C. 1989, *Heuristic potency of the Minimum Spanning Tree (MST) method in toxicology*, «Ecotoxicology and Environmental Safety», 17/2, 227-235.

Duda R.O., Hart P.E. 1973, *Pattern Classification and Scene Analysis*, Stanford Research Institute, Menlo Park, California, New York, Wiley-Interscience Publication.

Freeman L. 1977, *A set of measures of centrality based upon betweenness,* «Sociometry», 40, 35-41.

Gonzales R.C., Wintz P. 2002, *Digital Image Processing*, New Jersey, Prentice Hall (2nd ed.).

Helgason C., Buscema P.M., Grossi E. 2009, *Auto-Contractive Maps, H Function and Maximally Regular Graph-Theory in Artificial Adaptive Systems in Medicine*, in *Artificial Adaptive Systems in Medicine: New Theories and Models for New Applications*, E-Book, Chapters 4-5, Bentham Science Publishers.

Kohonen T. 2001, *Self Organizing Map*, Berlin, Springer Series in Information Sciences (3rd extended ed.).

Kruskal J.B. 1956, *On the shortest spanning subtree of a graph and the traveling salesman problem*, «Proceedings of the American Mathematical Society», 7/1, 48-50.

Licastro F., Porcellini E., Chiappelli M., Forti P., Buscema P.M., Ravaglia G., Grossi E. 2010a, *Multivariable network associated with cognitive decline and dementia*, «Neurobiology of Aging», 31/2, 257-269.

Licastro F., Porcellini E., Chiappelli M., Forti P., Buscema P.M., Ravaglia G., Grossi E. 2010b, *Multi factorial interactions in the pathogenesis pathway of Alzheimer's disease: a new risk charts for prevention of dementia*, «Immunity and Ageing», 16 (7 Suppl 1/S4).

Ma B., Hero A., Gorman J., Michel O. 2000, *Image registration with minimum spanning tree algorithm, in Proceedings of the Institute of Electrical and Electronics Engineers International Conference on Image Processing*, Vancouver, 481-484.

McClelland J., Rumelhart D.E. 1986, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. I, Cambridge MA, The MIT Press, 25-31.

McClelland J., Rumelhart D.E. 1988, *Explorations in Parallel Distributed Processing: A Handbook of Models, Programs, and Exercises*, Boston MA, The MIT Press, Macintosh edition, 1990, 39-47.

Newman M.E.J. 2006, *Modularity and community structure in networks*, «Proceedings of National Academy of Sciences USA», 103/23, 8577-8696.

Newman M.E.J. 2007, *Mathematics of networks*, in *The New Palgrave Encyclopedia of Economics*, Basingstoke, Palgrave Macmillan (2nd ed.).

Prim R.C. 1957, *Shortest connection networks and some generalizations*, «Bell System Technical Journal», 36, 1389-1401.

Ramazzotti M. 2012, *ARCHEOSEMA. Un modello archeo-logico per la ricerca teorica, analitica e sperimentale dei fenomeni complessi*, «Archeomatica», 2, 6-10.

Ramazzotti M. 2013, *ARCHEOSEMA. Sistemi Artificiali Adattivi per un'Acheologia Analitica e Cognitiva dei Fenomeni Complessi*, «Archeologia e Calcolatori», 24, 283-303.

Reichardt J., Bornholdt S. 2006, *Statistical mechanics of community detection*, «Physical Review E», 74 (1), 1-14.

Rumelhart D.E., Hinton G.E., Williams R.J. 1986a, *Learning representations by back-propagating errors*, «Nature», 323, 533-536.

Rumelhart D.E., Smolensky P., McClelland J.L., Hinton G.E. 1986b, *Schemata and sequential thought processes in PDP models*, in McClelland, Rumelhart 1986, 7-57.

Miyano S., Shamir R., Takagi T. (eds.) 2001, *Proceedings of the 12th International Conference on Genome Informatics,* Tokyo, Universal Academy Press.

States D.J., Harris N.L., Hunter L. 1993, *Computationally efficient cluster representation in molecular sequence megaclassification*, in *Proceedings of First International Conference on Intelligent System for Molecular Biology*, Bethesda, 387-394.

Street M.E., Grossi E., Volta C., Faleschini E., Bernasconi S. 2008, *Placental determinants of fetal growth: identification of key factors in the insulin-like growth factor and cytokine systems using artificial neural networks*, «BMC Pediatrics», 8, 24.

Tapia E., Rojas R. 2004, *Recognition of on-line handwritten mathematical expressions using a Minimum Spanning Tree construction and symbol dominance*, in *Graphics Recognition. Recent Advances and Perspectives. Lecture Notes in Computer Science,* Vol. 3088, Berlin Heidelberg, Springer-Verlag, 329-340.

Xu Y., Olman V., Xu D. 2001, *Minimum Spanning Tree for gene expression data clustering*, in Miyano, Shamir, Takagi 2001, 24-33.

Xu Y., Olman V., Xu D. 2002, *Clustering gene expression data using a graph theoretic approach: an application of minimum spanning trees*, «Bioinformatics», 18, 536-545.

Werbos, P.J. 1994, *The Roots of Backpropagation. From Ordered Derivatives to Neural Networks and Political Forecasting*, Wiley Series on Adaptive and Learning Systems for Signal Processing, Communication, and Control, New York, Wiley.

ABSTRACT

In this paper the theoretical and methodological aspects of some of the tools applied to the archaeological, geographical and linguistic problems posed by ARCHEOSEMA project will be analysed. In particular, the single steps of the process of generation of outputs, from the initial analysis of the dataset, the subsequent procedures of pre-processing and encoding of the data to the characteristics of the processing algorithms will be described. For this purpose we will use a so-called toy dataset known in the literature. Using the same dataset, we will illustrate the main output produced, Minimum Spanning Tree maps. Along with the use of classical literature measurements, such as the Pearson linear correlation and Prior Probability, both used as metrics for the generation of these outputs, we have tried to show the innovative contribution of a new artificial neural network, the Auto-Contractive Map, designed by P.M. Buscema at the Semeion Research Center.