How to Discover Hidden Knowledge According to Different Type Data Set: A Guideline to Apply the Right Hybrid Information Mining Approach

Roberto Paiano
Department of Engineering for Innovation, University of Salento
Piazza Tancredi, n7, 73100 Lecce LE, Italy
Phone: +39 0832 29 11 11
roberto.paiano@unisalento.it

Stefania Pasanisi
Department of Engineering for Innovation, University of Salento
Piazza Tancredi, n7, 73100 Lecce LE, Italy
Phone: +39 0832 29 11 11
stefania.pasanisi@unisalento.it

Abstract

The use of advanced data analysis techniques is now of considerable importance in order to allow the complex extraction of previously unknown and potentially useful implicit information on the data. Interest in this area has grown appreciably since these techniques had to meet the challenges introduced by the enormous proliferation of data triggered by the big data era. This implied, in the last few years, on developing advanced analysis techniques or improving existing ones by constantly introducing new techniques. The selection of an appropriate algorithm for a specific problem is very difficult and often the only solution is to proceed by trial and error. This paper intends to investigate which analysis technique should be used on a particular data set, based on the characteristics of this data set. We present three case studies, each of them concerns a very specific domain (Educational, Health and Safety) that is represented by a particular type of data set. The results establish a possible relationship between the analysis techniques implemented, that is Clustering analysis, Association Rule and Neural Network and the data set type analyzed.

Keywords: Information Mining, Data set characteristics, Big Data, Rich Data, Data Mining, Exploratory Data Analysis.

1. Introduction

The field of data exploration is very wide; this field includes several disciplines such as data mining, machine learning, data science, big data analytics and artificial intelligence. These disciplines are so interlaced and overlapped that it is very difficult, if not impossible, to delimit the boundaries or establish a hierarchy. In other words, these fields are symbiotic, and a combination of these approaches can be used as a tactic to produce more efficient and sensitive results. A hybrid information mining approach to data analysis, combining different techniques of data exploration, improve knowledge discovery and the information extraction (Pasanisi & Paiano, 2018). Interest in this area has grown considerably since these techniques had to meet the challenges introduced by the enormous proliferation of data triggered by the big data era. Big data, in fact, are changing our whole perspective on data extraction and interpretation due to the high volume, speed and variety of data, traditional modeling or manual inspection becomes impossible (Swathi & Seshadri, 2017). The methods of machine learning and data mining and their scientific basis provide results that are conceivable for extracting accurate information from data. So, to manage huge information (Big data), machine learning techniques are needed (Hammer et al., 2014). The use of advanced data analysis techniques (advanced analytics) is therefore of considerable importance today in order to allow the complex extraction of previously unknown and potentially useful implicit information on the data. The extraction and analysis carried out by means of automatic and semi-automatic systems, of large amounts of data allow the discovery and identification of significant patterns, models, relationships in the data, in terms of meaningful and immediately usable information. Many researchers have therefore concentrated, in the last few years, on developing advanced analysis techniques or improving existing ones by constantly introducing new techniques. With such a wideness of data analysis and exploration techniques, the need arises to determine which technique to use in a given situation trying to reply at the question: 'which algorithm should be the first choice for my problem?' According to various theories and authors, there is no better technique for all situations, in fact the characteristics of data sets, such as size, class distribution, or noise, can affect the performance of classifiers (Peng et al., 2011). It is therefore imperative to selectively apply appropriate techniques using a "trial and error" basis. The selection of an appropriate algorithm for a specific problem is very difficult: there is therefore no systematic research on which the analysis technique should be used on a particular data set, based on the characteristics of this data set. Moreover, it is useful to remember the No Free Lunch (NFL) theorem (Wolpert & Macready, 1997): "If algorithm A outperforms algorithm B on some cost functions, then loosely speaking, there must exist exactly as many other functions where B outperforms A". This theorem suggests that a more useful strategy is to gain an understanding of the data set characteristics that enable different learning algorithms to perform well, and to use this knowledge to assist learning algorithm selection based on the characteristics of the data set. A first issue of data mining and knowledge discovery is to correctly manage data considering different data types. It is necessary to distinguish between two ample categories of Big Data: those that are semantically poor, for instance sensor readings, and those that are more complex, multi-faceted, hierarchical, in a word, semantically rich. A characterization of the data on the basis of the semantic concepts and size is possible: small amounts of semantically rich data, large amounts of data semantically poor, large amounts of semantically rich data (Pasanisi & Paiano, 2016). Rich data types can be categorized into: non-dependency and dependency data. The non-dependency data is the most commonly encountered type, which refers to data without specified dependencies between data instances. In other words, data instances are assumed independent and identically distributed. Examples of non-dependency data include multidimensional data, text data, and image data. In practice, data can be more complex, and there exists dependency between data instances. Dependency data can be correlated with temporal, spatial, sequential, and social relationships such as time-series, sequence, graph, multi-media, and social-media data.

In this paper, we have therefore conducted a series of experiments, through which we examined the inherent relationship that may exist between the characteristics of a data set and the performance of a technique of analysis and exploration of data. We have chosen to test three analysis techniques related to three different methods of analysis: descriptive method, local method, forecasting method according to our hybrid information mining approach for knowledge discovery discussed in (Pasanisi & Paiano, 2017) and (Pasanisi & Paiano, 2018). As a descriptive method, the Cluster Analysis technique was applied, for the local method the Association Rule was applied while for the forecast method an artificial neural network was applied. To evaluate the performance of these techniques, numerous experiments have been conducted. The experiments were conducted on three case studies related to three different application domains. The efficiency of an algorithm depends in some way on the data set and the domain to which it is applied. Under certain conditions, a Machine Learning algorithm may be more powerful than another (Das & Behera, 2017). It is still a complex issue to determine which algorithm is how strong or how weak in relation to which data set. We have applied this approach to three distinct domains: Educational domain, Road Safety domain, Health domain. The corresponding data sets, therefore, are very different and present peculiar characteristics in terms of semantic wealth, number of attributes, size, inter-correlation between the data, qualitative and quantitative characteristics. The results of the experiments allowed us to evaluate the effectiveness of the approach and derive useful considerations for the evaluation of a data set in its structure, in its qualitative and quantitative characteristics and in its semantic richness or not. These considerations offer suggestions on which analysis technique is more performant for a given data set or domain.

The paper's structure is the following: related works are described in section 2, the Section 3 details the data set type and characteristics, the Section 4 describes data mining's techniques that are applied, in section 5 we show experiments and results. Finally, in section 6 and 7 the dissertation and conclusion.

2. Background

In the literature, little recent research has focused on these issues. Previous research in this context has been conducted, for example, to find a relationship between classification techniques and data set characteristics. It has been seen that the characteristics of the data set considerably influence the performance of the classification methods and show that the choice of the best classification algorithm depends on the specified data set (Oreski et al., 2017). In (Wang et al., 2012) a recommendation method for classification algorithms based on the data set characteristics is presented. The implemented method uses structural and statistical information-based feature vectors to characterize each data set. The experimental results demonstrate that this recommendation method is effective. Furthermore, the method to characterize data sets is better than the traditional and the problem complexity-based methods. In (Kiang, 2003) the conducted experiments suggest that data characteristics impact considerably the classification performance of the methods. The proposed results can aid in the design of classification systems in which several classification methods can be employed to increase the reliability and consistency of the classification. The authors put at highlight that while different data characteristics may affect performance of the classification methods to different degrees, the level of bias in individual data characteristics may also affect the performance to different degrees. However, the study needs more elaborate experiments which are required to examine the possible interactions among factors, and the effect of varying levels of biases on the outcome. An empirical study (Temizel, 2017) compared three clustering algorithms, which are k-means, SOM (Self-Organizing Maps), and agglomerative clustering using Dunn (Dunn, 1974) and SD (Halkidi et al., 2000) validity index methods that validate the clusters by taking into consideration the data set characteristics. The study shows that neither validity methods are reliable when the data points are scattered uniformly and clusters, if any, are not distinguishable, but they work well when there are distinct clusters. This implies that the data set characteristics affect the performance of validity indices and can mislead us to choose the right cluster number. Validity indices should be tested on data sets having different characteristics such as a set including no clusters. As a future work of this study, the authors want to improve these validity indices.

3. Data set characteristics

This section may be divided in subheadings. It should provide a concise and precise description of the experimental results, their interpretation as well as the experimental conclusions that can be drawn.

3.1. Educational domain

In the educational domain, we have analyzed educational learning Experiences. The main focus of the analysis and data exploration was the correct assessment of learning projects, exploring and discovering the hidden relationships between the educational tools (in particular the use of technology educational tools) and the learning outcomes and didactic benefits. This task is a complex task because there are several aspects (such as contents, technologies, organizations etc.) that must be considered and many actors (learners, teachers, pedagogues, etc.) each one with specific requirements to be met. The analysis on this domain was conducted within the case study Exploratory Portal learning4all (L4All), for EDOC@Work3.0 project. By Exploratory Portal, we intend a highly interactive environment, where the exploration can take place through a number of strongly interconnected (and interdependent) interactions (Di Blas & Paolini, 2013). The L4ALL is characterized by a repository of shared meaningful learning experiences that have made important use of technology to innovate and improve teaching process. L4All is an exploratory portal that hosts almost 300 objects describing educational experiences in which the use of technology was relevant. Each object implies various information items: an abstract, some structured data, one or more reports, interviews, documents produced within the experiment, etc. All the objects are classified by pedagogy experts according to a complex taxonomy consisting of 39 attributes' categories and more than 300 attributes. Categories and attributes are organized into widgets sustaining both selection and exploration. Simple selection or complex selection operations, with Boolean operators, are possible. Each widget shows the value of the attributes for the current state of the data set with different visualization. Thanks to advanced Human -Computer Interaction mechanisms, the portal can support exploration activities in the cycle <selection, feedback, selection>. Based on L4All, a number of scientific investigations by different research groups took place: on the relation between different forms of group-work and inclusion, on digital storytelling and related benefits, etc. (Di Blas & Paolini, 2013), (Falcinelli & Laici, 2012), (Falcinelli, 2012).

3.2. Health domain

In medical fields that deal with chronicity, there is the need to plan and implement effective and efficient management of chronicity with the main goal to improve quality of life. Our work focuses on the context related to cardiovascular risk and the pathologies involved in particular, diabetes, hypertensive disease, heart failure, cardio-circulatory disease, cerebral vascular diseases, and circulatory system diseases. In the literature there are multiple risk factors for heart diseases, and we have found 12, which are useful to healthcare professionals in recognizing patients with high risk of heart disease (Halkidi et al., 2000). An important factor that we have considered in our analysis is the individual risk score. The individual risk score is a simple tool for assessing the probability of developing a first major cardiovascular event (myocardial infarction or stroke) over the following 10 years, when the values of eight risk factors are known (Rosenblatt, 1961). In this way, the total evaluation of all higher risk factors is taken into account, not based on the mere summation of risk factors' values, but considering the multi factorial cause of cardiovascular disease (Rosenblatt, 1961). There are 6 CVD Risk classes that identify 6 range of risk score based on severity. A LHO (a local healthcare organization) is interested in finding out how certain variables are associated with the onset of CVD risk and in assessing the efficacy of a patient's therapeutic path. To achieve this aim, LHO was enrolled in 2008 a sample population of 5134 healthy patients (that is, patients who have never had a cardiovascular event) recording several variables of particular interest to CVD risk and then monitored their health status for 10 years. The data related to the treatment and therapy of patients from 2008 to 2017 were collected: specialist medical services, hospitalizations, consumed medicines, and new diagnoses for the diseases considered. The purpose is to stratify the population of patients, to identify patterns of high-risk patients, and to find groups of patients who have worsened after 10 years from the initial enrollment. We have collected and integrated information related to all services (hospital, diagnostic, specialist, therapy, etc.) with the aim of obtaining useful information to intercept indications on the diagnostic–therapeutic assistance pathways (PDTA) of patients, which are difficult to obtain from the simple registration of a care contact. The administrative flows useful for the purposes are those related to the patient registry, exemptions for pathology, hospital discharge cards, outpatient specialists, and pharmaceuticals. A data set exists in the database of LHO that contains the following variables of particular interest to CVD risk: gender, age, smoking habit, onset of diabetes, onset of hypertension, individual risk score, systolic blood pressure, serum cholesterol, High -Density Lipoprotein cholesterol. Furthermore, the presence of the following diseases has been considered: cardiac disease, cardio-circulatory disease, heart failure, cerebrovascular disease, and circulatory system disease. The distribution of CVD risk score at enrollment of patients is the following:

- CVD1 (score < 5%) = 3414 patients
- CVD2 (score 5-10%) = 931 patients

- CVD3 (score 10-15%) = 391 patients
- CVD4 (score 15-20%) = 195 patients
- CVD5 (score 20-30%) = 143 patients
- CVD6 (score >30%) = 60 patients

In table 1, the initial distribution of diseases at time of patients' enrollment and then, the final distribution of diseases, after 10 years from patients' enrollment are shown.

Table 1. Initial and final disease distribution

Distribution	**Initial**		**Final**	
Disease	**Yes**	**No**	**New Diagnoses**	**Total Number Disease patients**
diabetes	465	4669	135	600
hypertension	1686	3448	54	1740
cardiac disease	682	4452	338	1020
cardio-circulatory disease	104	5030	413	517
other CV disease	26	5108	36	62

3.3. Road Safety domain

In this case study, we evaluate the data concerning car accidents occurred in the year 2016 in the Lecce city (Puglia region, Italy). On the basis of the available data we ask ourselves if we can identify sub-sets of road accidents with similar characteristics or if we can identify hidden relations between the attributes that characterize this data set or, also, if it is possible to create a predictive model of accident risk on the streets. The data set is divided into streets. Each record is a road where the cumulative data of different attributes, in one year, of the accidents on that road are reported. The number of present roads is equal to 466, while the number of road accidents is 1147. The data set contains the following attributes on the road accidents: street name, vehicle type (Car, Motorcycle / Moped, / Trailer, Velocipede), driver sex, age range (up to 18, from 18 to 30, from 31 to 50, over 50), season, weekday, type of collision (frontal collision, lateral, pedestrian investment, vehicle against obstacle, roadway exit.), road surface (dry, wet, slippery), time hours (we have identified the following 6 time slots 3-6,7-10,11-14,15-18,19-22,23-2), weather conditions (Clear, Cloudy, Rain); paving of road and so on, for a total of 58 attributes for 466 instances.

4. Data Mining techniques

The choice of analytical methods to be used according to the exploration objectives we have defined, is an important phase of data exploration process. There are many techniques and methods and each of them has been implemented in a myriad of algorithms. The type of analysis we want to do, which in turn depends on the objectives set and the data set we have available, influence the choice of methods. The principal classes of statistical methods are three: descriptive methods, predictive methods and local methods (Giudici, 2005). The descriptive methods also called unsupervised or indirect methods, aim to group the data on the basis of relationships "non-notables" a priori or with an exploratory analysis. Predictive methods, also called supervised methods, aim to find relationships between features and targets, in order to identify classification or prediction. Finally, the local methods identify particular characteristics and relationships on subsets of the data set. To test all methods, we have chosen, among the multiple algorithms existing, Cluster Analysis as descriptive method, Artificial Neural Network as predictive method and Association Rules among the local methods. Below we describe the methods chosen for our analysis.

4.1. Cluster Analysis

The Clustering algorithms allow performing segmentation operations on the data, that is to identify homogeneous patterns, which have regularities in them able to characterize and differentiate from the other patterns.

The Clustering technique finds groups of elements, in which the elements within a group are similar to each other and different from the elements of another group (El-Halees, 2009). This technique is considered the most important unsupervised learning technique. Main concept of this technique is that of distance: near elements, in accordance to a certain metric, must be included in the same cluster while distant elements must be inserted in a different cluster. We have chosen to use the K-means algorithm that find a k number of disjoint groups from a data set where the value of k is set early. A known clustering algorithm is K-means, it finds a k number of disjoint groups from a data set where the value of k is set early. There are two steps: in the first step, define k centroids, one for each cluster; in the second step, associate each point of the data set to the nearest centroid through a metric distance between the data points and the centroids. When the first pass is completed, it is necessary to recalculate the new centroids and the new distances producing a loop. In this loop, the k centroids may modify their position in a step by step way. When the centroids do not change position anymore, the convergence criterion for clustering will be satisfied (Nazeer & Sebastian, 2009). Therefore, the algorithm is iterative and needs two inputs: a metric of distance between patterns (Euclidean distance is generally considered) and the initial number of clusters k, where k is found by trials. The aim is to minimize the function (1):

$$C_{KM} = min_{\mu_1,\ldots,\mu_k} \sum_{i=1}^{k} \sum_{x_j \in S_j} \left\| x_j - \mu_j \right\|_2^2 \qquad (1)$$

where μ_j, $j = 1, \ldots,$ k, indicate k centroids and S_j are k clusters. The centroid is identified by mediating the position of the points belonging to the cluster. The quality of the final clusters depends strongly on the values of the initial centroids (Nazeer & Sebastian, 2009).

4.2. Association Rules

Association rules is a technique that show the hidden relationship between data attributes. The association rules algorithms are unsupervised, as we start from the principle of not knowing anything about our data set and wanting to try to obtain information discovery. Among the different algorithms to find the rules of associations between features, Apriori Association algorithm discovers in the data set the best combination of attributes. Apriori (Agrawal & Srikant, 1994) is an iterative method, which proceeds by creating so-called item sets, or sets of rules for which the conditions of support and confidence are verified. Support and confidence are the metrics to evaluate the quality of association rule in two steps: First, minimum support is applied to find all frequent item sets in a database. Second, these frequent item sets, and the minimum confidence constraint are used to form rules. Support is an indication of item how frequently it occurs in data set. The algorithm finds rules of the type: "If antecedent then (likely) consequent", where antecedent and consequent are item sets which are sets of one or more items. For a rule A=> B, its support is the percentage of transaction in data set that contain 'A U B' (means both A and B); confidence indicates the number of times the statements found to be true (Suresh & Ramanjaneyulu, 2013).

4.2.1. Apriori Algorithm

The Apriori principle "any subset of a frequent item sets is a frequent item sets" is applied for the calculation of frequent item sets. Given a K-item sets with k elements in input. Apriori looks for large item sets considering the k-item sets for increasing k values. Steps executed:

1. count item occurrences to calculate large 1-item sets
2. iterate until no new large 1-item sets are found
3. (k+1) length candidate item sets are identified from length k large item sets
4. candidate item sets containing non-large subsets of k length are not considered

5. count Support of each candidate item sets by scanning the data set
6. remove candidate item sets that are small

The output generated are item sets that are "large" and that satisfy the min support and min confidence thresholds (Agrawal & Srikant, 1994).

4.3. Neural Network

An artificial neural network is a computational system that emulates the biological neural networks in the way the human brain processes information. A Multilayer Perceptron (MLP) is the type artificial neural network most known that use, for training, a supervised learning technique called back-propagation (Rosenblatt, 1961). An MLP is formed of at least three layers of nodes and each node is a neuron that uses a nonlinear activation function (except for the input nodes). The first layer (input layer) is in direct contact with the input data; the intermediate layer (hidden layer) has no direct contact with the outside, as it receives data from the input layer and sends it to the output neuron layer; the last layer (output layer) receives data from the neurons of the hidden layer and interfaces with the output. There is no precise method in the literature to determinate the number of neurons and layers for the hidden layer, but usually we proceed by trials comparing between architectures to choose the optimal one. Hidden neurons, therefore, receive information from the input neurons through the weights w, and produce the output $h_k = f(x, w_k)$, where $f(\cdot)$ is the activation function of the neuron. The output neurons, in turn, receive the data from the hidden layer, apply the appropriate z weights, and then produce the output $y_j = g(h, z_j)$. By combining the two functions, the output of the j_{th} neuron is therefore

$$y_j = g\left(\sum_k h_k z_{kj}\right) = g\left(\sum_k z_{kj} f\left(\sum_i x_i w_{ik}\right)\right) \qquad (2)$$

This equation shows how the mapping of the inputs in an MLP is highly nonlinear. However, it is necessary to find the weights that characterize the network. Among the different types of MLP, we have chosen the supervised learning MLP in which in the training set we have defined a target, so it is possible to define an error function with respect to the optimal output.

5. Hybrid information mining approach

The proposed approach aims (Pasanisi & Paiano, 2017), (Pasanisi & Paiano, 2018) at combining these techniques for an effective Knowledge Discovery. The Clustering algorithms allow, in fact, to perform segmentation operations on the data, which is to identify homogeneous patterns, which have regularities within them able to characterize them and differentiate them from other patterns. Knowledge patterns, on a complex data set, make it possible to explore data within the most interesting cluster and this can facilitate the correct interpretation of the results of the exploration. The relevant properties that allowed to build the clusters can be used for a new phase of "pre-processing" of the data in order to apply in a more meaningful way another analysis technique, Association Rule. The algorithm for the search of the associative rules between attributes will be applied to the entire data set appropriately conFigured according to the results of the clustering. Moreover, the algorithm will be applied also to the subset or the knowledge patterns identified through clustering. As we will see, this will allow us to better understand the relationships between the attributes because they are not related to the whole data set but to a specific pattern, that has certain properties and characteristics. Ultimately, we will use both the results obtained with these analyzes (knowledge patterns, relevant properties of the pattern, relations between attributes) to conFigure the data set and apply a Neural Network capable of predicting certain behaviors, situations or states.

5.1. Technology used

Clustering, Association Rule and Artificial Neural Network have been analyzed on the Weka workbench (Weka software v3.8.1). Weka is Waikato Environment for Knowledge analysis and it is a suite of machine learning algorithms for data mining task. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization (Witten et al., 2016). Weka provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query. It is not capable of multi-relational data mining, but there is separate software for converting a collection of linked database tables into a single table that is suitable for processing using Weka (Reutemann et al., 2004). Weka contains several applications: Explorer, Experimenter, KnowledgeFlow, Workbench, and SimpleCLI. In the application Explorer, Weka provides comprehensive sets of data pre-processing tools, learning algorithms and evaluation methods, graphical user interfaces and an environment for comparing learning algorithms. Explorer contains "cluster" tab for finding groups of similar instances in a data set. Some implemented schemes are: K-means, EM, Cobweb, X-means, FarthestFirst, etc. Explorer contains "classify" tab for finding classification scheme and predictive models. Classified implemented algorithms are: NaiveBayes, Multilayer Perceptron, ZeroR, J48, RandomForest, etc. Another tab is "Associate" for finding hidden relationships between attributes through the algorithms Apriori, FPGrowth, FilteredAssociator. The present evaluation options are: cross-validation, learning curve, hold-out and it is possible also to iterate over different parameter settings (Witten et al., 2016)

6. Experiments and Results

Data pre-processing is a salient step in the data mining process. The knowledge discovery during the training phase is more difficult if there is much irrelevant and redundant information present or noisy and unreliable data. This task can take the considerable effort and amount of processing time. For each data set domain, the pre-processing phases conducted, are data selection, data cleaning, data discretization, data integration, data filtering.

After the first phases of information retrieval and pre-processing for each data set, we uploaded a data set obtained in csv format on tool Weka.

To apply Clustering technique, we have implemented the Weka "SimpleKMeans" algorithm from tab "Cluster", tuning on the available parameters on which to act, as for example distancefunction, initializationMethod, maxIterations, numClusters (k), numExecutionSlots, and seed. In the Weka SimpleKMeans algorithm, we used the Euclidean distance measure to compute distances between instances and clusters. The best clustering was made by changing the classification parameters. We tested SimpleKMeans, on our data set, tuning with different values of k, to find the optimal centroids. In general, as you know, there is no method for determining the exact value of k, but an accurate estimate can be obtained, for example, monitoring the value of the sum of squared error (SSE) for some values of k (for example 2, 4, 6, 8, etc.).

To apply Association Rule technique, we have implemented the Weka "Apriori" algorithm from tab "Associate", interacting with a wide range of options that allow us to find the best results through appropriate parameter tuning. The real work for association rule learning is in the interpretation of results. From looking at the "Associator output" window, we can see the rules learned from the data set. The algorithm is conFigured to stop at 10 rules by default and ends when there are at least 10 rules with the minimum confidence established or when the support has achieved a 10% lower bound. It is possible to conFigure it to find and report more rules by changing the "numRules" value.

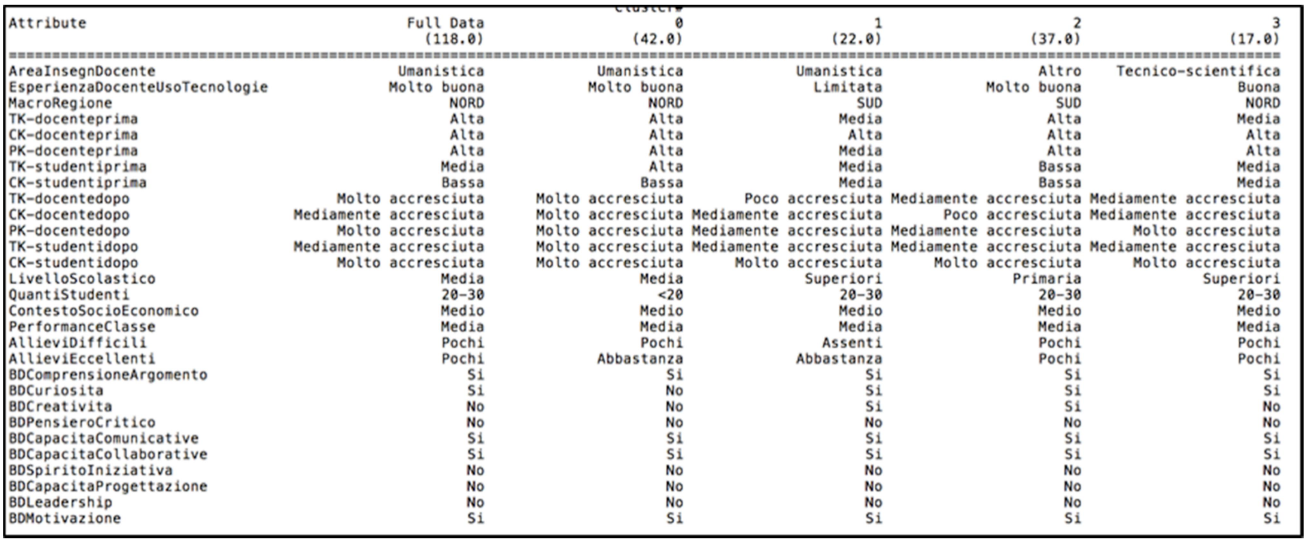
To apply Neural Network technique, we have implemented the Weka "Multilayer Perceptron" algorithm from tab "Classify". Also, in this case it is possible to interact with different parameters to train the network and find an effective function that connects the inputs with the outputs of the network. Some of the parameters with which to interact are the number of neurons of the hidden layer, learning rate, momentum, number of training periods.

Below are the results of the three analysis techniques for each case study.

7. Case Study 1 – L4ALL portal: educational learning experiences data set

7.1. Cluster Analysis results

Our aim was to identify, using data mining models, patterns knowledge inside of the facets of the exploratory portal. To achieve this purpose, we have applied the clustering algorithm on experiences of the exploratory portal. The portal is schema-driven through a modeling of taxonomy, the data and the portal layout on Excel. Thus, the starting data set consists of the general scheme of the complex taxonomy on which the modeling of experiences is based. The proposed general scheme consists of two excel files: one related to the data and one related to the annexes of the experiences. The data file consists of "widget"(only one sheet) that defines the overall layout and the number of columns in which subdivide the widgets in the interface; "define Widget” (one sheet for each facet) that defines the structure of each widget, the labels displayed for each widget; "widget label" (one sheet for each facet) that defines the data of the experience. The connection between the sheets is through the widget id. The schema presented defines all aspects of the data for our case study. From this starting data set, we have extracted and built the data set on which to apply the data mining clustering technique. We selected the relevant facets for our purpose. So, starting with the 39 initials facets we extracted 23 facets for a total of 42 types of attribute and, after the operations of cleaning, enrichment and coding, we have a total of 118 instances. Then we uploaded a data set of the facets of the experiences in csv format on the Weka tool. Testing SimpleKMeans algorithm with different values of k, we estimated in k =8 and seed=10 and number of epochs = 500 the best conFigureuration of cluster. We obtained the following clustered instances (Figure 1):



```
Attribute                                  Full Data                      0                      1                      2                      3
                                             (118.0)                 (42.0)                 (22.0)                 (37.0)                 (17.0)
=================================================================================================================================================
AreaInsegnDocente                         Umanistica             Umanistica             Umanistica                  Altro    Tecnico-scientifica
EsperienzaDocenteUsoTecnologie           Molto buona            Molto buona               Limitata            Molto buona                  Buona
MacroRegione                                    NORD                   NORD                    SUD                    SUD                   NORD
TK-docenteprima                                 Alta                   Alta                  Media                   Alta                  Media
CK-docenteprima                                 Alta                   Alta                   Alta                   Alta                   Alta
PK-docenteprima                                 Alta                   Alta                  Media                   Alta                   Alta
TK-studentiprima                               Media                   Alta                  Media                  Bassa                  Media
CK-studentiprima                               Bassa                  Bassa                  Media                  Bassa                  Media
TK-docentedopo                     Molto accresciuta      Molto accresciuta       Poco accresciuta Mediamente accresciuta Mediamente accresciuta
CK-docentedopo                Mediamente accresciuta      Molto accresciuta Mediamente accresciuta       Poco accresciuta Mediamente accresciuta
PK-docentedopo                     Molto accresciuta      Molto accresciuta Mediamente accresciuta Mediamente accresciuta      Molto accresciuta
TK-studentidopo               Mediamente accresciuta      Molto accresciuta Mediamente accresciuta Mediamente accresciuta Mediamente accresciuta
CK-studentidopo                    Molto accresciuta      Molto accresciuta      Molto accresciuta      Molto accresciuta      Molto accresciuta
LivelloScolastico                              Media                  Media              Superiori               Primaria              Superiori
QuantiStudenti                                 20-30                    <20                  20-30                  20-30                  20-30
ContestoSocioEconomico                         Medio                  Medio                  Medio                  Medio                  Medio
PerformanceClasse                              Media                  Media                  Media                  Media                  Media
AllieviDifficili                               Pochi                  Pochi                Assenti                  Pochi                  Pochi
AllieviEccellenti                              Pochi             Abbastanza             Abbastanza                  Pochi                  Pochi
BDComprensioneArgomento                           Si                     Si                     Si                     Si                     Si
BDCuriosita                                       Si                     No                     Si                     Si                     Si
BDCreativita                                      No                     No                     Si                     Si                     No
BDPensieroCritico                                 No                     No                     No                     No                     No
BDCapacitaComunicative                            Si                     Si                     Si                     Si                     Si
BDCapacitaCollaborative                           Si                     Si                     Si                     Si                     Si
BDSpiritoIniziativa                               No                     No                     No                     No                     No
BDCapacitaProgettazione                           No                     No                     No                     No                     No
BDLeadership                                      No                     No                     No                     No                     No
BDMotivazione                                     Si                     Si                     Si                     Si                     Si
```

Figure 1. L4ALL - SimpleKMeans results

In Table 2 a characterization of the identified clusters is presented, representing patterns of knowledge where a significant exploration is possible:

Table 2. Knowledge’s pattern

Cluster	**Characterization**
C0	teaching area: **Humanistic** - expertise with technology of teacher: **Very good** - macro-region: **Nord** - school level: **Secondary** - social-economic context: **Average** - class performance: **Average**
C1	teaching area: **Humanistic** - expertise with technology of teacher: **limited** - macro-region: **South** school level: **College** - social-economic context: **Average** - class performance: **Average**
C2	teaching area: **Other** - expertise with technology of teacher: **Sufficient** - macro-region: **South** school level: **Primary** - social-economic context: **Average** - class performance: **Average**
C3	teaching area: **Technical-scientific** - expertise with technology of teacher: **Good** - Macro-region: **North** school level: **College** - social-economic context: **Average** - class performance: **Average**

Through the knowledge's patterns, the user can explore the information within the more interesting cluster facilitating the correct interpretation of exploration results and, furthermore, can use the relevant properties of each cluster to refine the information search on the entire data set in order to conduct a more effective general exploration.

7.2. Association Rules results

We conducted several experiments using the Apriori algorithm on different groups of data. Starting from the entire data set, we found meaningful rules for the general data set, and we subsequently applied the algorithm on the basis of the appropriate filter operation to the following subsets of data: C0-C3 cluster. We present below the results of some of the experiments.

Abbreviations legend:

- EB: Educational Benefits
- CK: Content Knowledge
- PK: Pedagogy Knowledge
- TK: Teaching Knowledge

Experiment 1. General data set (118 instances)

Best rules found with confidence range 0.91 – 0.96:

1. EB Curiosity = Yes 68 ==> EB Understanding Argument = Yes 65
2. EB Spirit Initiative = No EB Design Ability = No 76 ==> EB Leadership = No 71
3. EB Spirit Initiative = No 83 ==> EB Leadership = No 77
4. EB Spirit Initiative = No EB Leadership = No 77 ==> EB Design Ability = No 71
5. EB Understanding Argument = Yes EB Spirit Initiative = No 73 ==> EB Design Ability = No 67
6. EB Understanding Argument = Yes EB Spirit Initiative = No 73 ==> EB Leadership = No 67
7. EB Capacity Design = No EB Motivation = Yes 72 ==> EB Understanding Argument = Yes 66
8. EB Spirit Initiative = No 83 ==> EB Design Capacity = No 76
9. CK teacher first = High 78 ==> EB Understanding Argument = Yes 71
10. EB Communications Capacity = Yes EB Collaborative Capacity = Yes 75 ==> EB Understanding Argument = Yes 68

Experiment 2. Cluster 0 data set (42 instances)

Best rules found with confidence range 0.93 – 0.97:

1. PK teacher first =High 29 ==> CK student after=Highly increased 28
2. TK teacher after =Highly increased EB Understanding Argument=Yes 29 ==> CK student after=Highly increased 28
3. TK teacher after =Highly increased EB Motivation=Yes 29 ==> CK student after=Highly increased 28
4. EB Communication Capacity=Yes EB Motivation=Yes 29 ==> EB Collaborative Capacity=Yes 28
5. CK teacher first=High EB Leadership=No 28 ==> CK student after=Highly increased 27
6. TK teacher after =Highly increased 35 ==> CK student after=Highly increased 33
7. CK teacher first=High 31 ==> CK student after=Highly increased 29
8. TK teacher first=High 30 ==> CK student after=Highly increased 28
9. PK teacher after=Highly increased 30 ==> CK student after=Highly increased 28
10. EB Spirit Initiative =No 30 ==> EB Design Capacity=No 28

Experiment 3. Cluster 1 data set (22 instances)

Best rules found with confidence range 0.94 – 0.95:

1. EB Communication Capacity=Yes 20 ==> EB Understanding Argument=Yes 19
2. TK student after=On average increased 18 ==> EB Understanding Argument=Yes 17
3. EB Collaborative Capacity=Yes 18 ==> EB Understanding Argument=Yes 17

4. EB Motivation=Yes 18 ==> EB Understanding Argument=Yes 17
5. EB Collaborative Capacity=Yes 18 ==> EB Communication Capacity=Yes 17
6. EB Motivation=Yes 18 ==> EB Communication Capacity=Yes 17
7. EB Design Capacity=No 17 ==> EB Understanding Argument=Yes 16
8. EB Communication Capacity=Yes EB Collaborative Capacity=Yes 17 ==> EB Understanding Argument=Yes 16
9. EB Understanding Argument=Yes EB Collaborative Capacity=Yes 17 ==> EB Communication Capacity=Yes 16
10. EB Communication Capacity=Yes EB Motivation=Yes 17 ==> EB Understanding Argument=Yes 16

Experiment 4. Cluster 2 data set (37 instances)
Best rules found with confidence range 0.92 – 0.95:
1. CK teacher first=High 22 ==> EB Understanding Argument=Yes 21
2. CK student first=Low EB Collaborative Capacity=Yes 21 ==> EB Understanding Argument=Yes 20
3. EB Design Capacity=No EB Motivation=Yes 21 ==> CK student first=Low 20
4. EB Curiosity=Yes EB Design Capacity=No 21 ==> EB Leadership=No 20
5. EB Motivation=Yes 27 ==> EB Understanding Argument=Yes 25
6. EB Curiosity=Yes 25 ==> EB Understanding Argument=Yes 23
7. EB Collaborative Capacity=Yes 25 ==> EB Understanding Argument=Yes 23
8. EB Understanding Argument=Yes EB Design Capacity=No 25 ==> CK student first=Low 23
9. CK student first=Low EB Leadership=No 25 ==> EB Design Capacity=No 23
10. CK student first=Low EB Motivation=Yes 24 ==> EB Understanding Argument=Yes 22

Experiment 5. Cluster 3 data set (17 instances)
Best rules found with confidence 1:
1. EB Design Capacity=No 15 ==> EB Leadership=No 15
2. EB Understanding Argument=Yes EB Design Capacity=No 14 ==> EB Leadership=No 14
3. TK teacher first=Average 13 ==> EB Leadership=No 13
4. EB Motivation=Yes 13 ==> EB Understanding Argument=Yes 13
5. EB Creativity=No 13 ==> EB Design Capacity=No 13
6. EB Creativity=No 13 ==> EB Leadership=No 13
7. EB Spirit Initiative =No 13 ==> EB Design Capacity=No 13
8. EB Spirit Initiative =No 13 ==> EB Leadership=No 13
9. EB Creativity=No EB Leadership=No 13 ==> EB Design Capacity=No 13
10. EB Creativity=No EB Design Capacity=No 13 ==> EB Leadership=No 13

The algorithm provides the relationships between attributes. It is noted that: the rules identified in the general data set can be better researched in the knowledge patterns identified by the clusters. In this way, we not only discover that there is a semantic rule that binds the attributes, but we can also know to which kind of pattern this rule it belongs, by further refining the discovered knowledge.

7.3. Artificial Neural Network results

In this case, the multilayer perceptron neural network has the aim to create a predictive model of educational benefits according to the type of educational experience. The use of the neural network did not produce satisfactory results: the model was not able to learn from the data a function that connected inputs to outputs probably because the input data are semantically complex and multifaceted and therefore difficult to understand from the model. Therefore, it was not possible to apply this method to this data set.

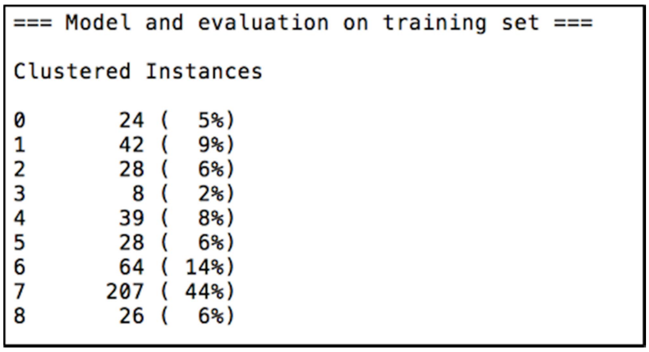
8. Case Study 2 – CVD risk data set

We presented the results of the hybrid approach in this area in our previous article (Pasanisi & Paiano, 2018).

9. Case Study 3 – Road safety data set

9.1. Cluster Analysis results

The best conFigureuration of the SimpleKMeans algorithm is obtained setting the following parameters: k = 9 and seed = 150 and number of epochs = 500. We obtained the following clustered instances in Figure 2 and Figure 3:

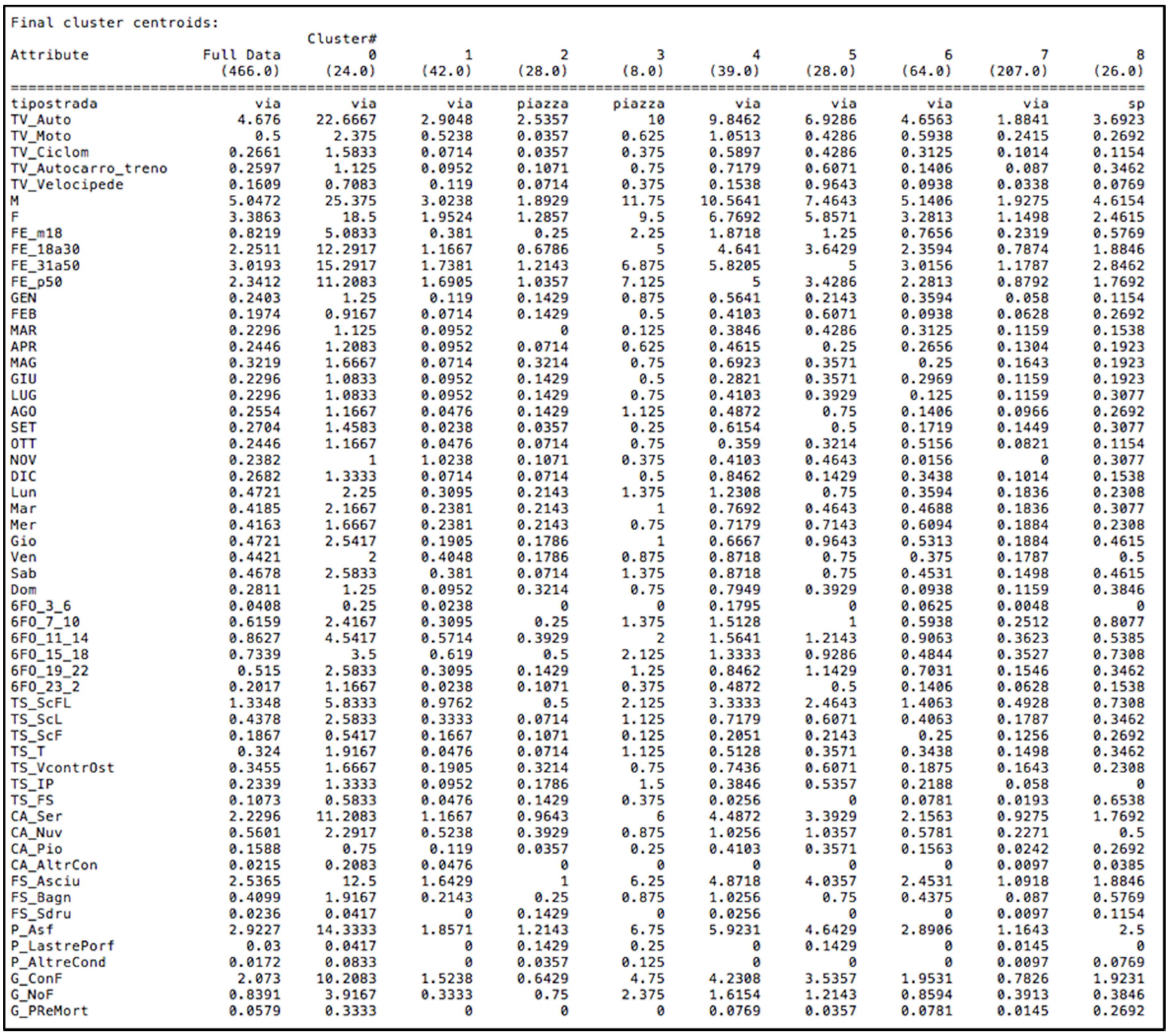


```
=== Model and evaluation on training set ===

Clustered Instances

0       24 (  5%)
1       42 (  9%)
2       28 (  6%)
3        8 (  2%)
4       39 (  8%)
5       28 (  6%)
6       64 ( 14%)
7      207 ( 44%)
8       26 (  6%)
```

Figure 2. Road Safety – clustered instances.



```
Final cluster centroids:
                                Cluster#
Attribute            Full Data         0         1         2         3         4         5         6         7         8
                       (466.0)    (24.0)    (42.0)    (28.0)     (8.0)    (39.0)    (28.0)    (64.0)   (207.0)    (26.0)
========================================================================================================================
tipostrada                 via       via       via    piazza    piazza       via       via       via       via        sp
TV_Auto                  4.676   22.6667    2.9048    2.5357        10    9.8462    6.9286    4.6563    1.8841    3.6923
TV_Moto                    0.5     2.375    0.5238    0.0357     0.625    1.0513    0.4286    0.5938    0.2415    0.2692
TV_Ciclom               0.2661    1.5833    0.0714    0.0357     0.375    0.5897    0.4286    0.3125    0.1014    0.1154
TV_Autocarro_treno      0.2597     1.125    0.0952    0.1071      0.75    0.7179    0.6071    0.1406     0.087    0.3462
TV_Velocipede           0.1609    0.7083     0.119    0.0714     0.375    0.1538    0.9643    0.0938    0.0338    0.0769
M                       5.0472    25.375    3.0238    1.8929     11.75   10.5641    7.4643    5.1406    1.9275    4.6154
F                       3.3863      18.5    1.9524    1.2857       9.5    6.7692    5.8571    3.2813    1.1498    2.4615
FE_m18                  0.8219    5.0833     0.381      0.25      2.25    1.8718      1.25    0.7656    0.2319    0.5769
FE_18a30                2.2511   12.2917    1.1667    0.6786         5     4.641    3.6429    2.3594    0.7874    1.8846
FE_31a50                3.0193   15.2917    1.7381    1.2143     6.875    5.8205         5    3.0156    1.1787    2.8462
FE_p50                  2.3412   11.2083    1.6905    1.0357     7.125         5    3.4286    2.2813    0.8792    1.7692
GEN                     0.2403      1.25     0.119    0.1429     0.875    0.5641    0.2143    0.3594     0.058    0.1154
FEB                     0.1974    0.9167    0.0714    0.1429       0.5    0.4103    0.6071    0.0938    0.0628    0.2692
MAR                     0.2296     1.125    0.0952         0     0.125    0.3846    0.4286    0.3125    0.1159    0.1538
APR                     0.2446    1.2083    0.0952    0.0714     0.625    0.4615      0.25    0.2656    0.1304    0.1923
MAG                     0.3219    1.6667    0.0714    0.3214      0.75    0.6923    0.3571      0.25    0.1643    0.1923
GIU                     0.2296    1.0833    0.0952    0.1429       0.5    0.2821    0.3571    0.2969    0.1159    0.1923
LUG                     0.2296    1.0833    0.0952    0.1429      0.75    0.4103    0.3929     0.125    0.1159    0.3077
AGO                     0.2554    1.1667    0.0476    0.1429     1.125    0.4872      0.75    0.1406    0.0966    0.2692
SET                     0.2704    1.4583    0.0238    0.0357      0.25    0.6154       0.5    0.1719    0.1449    0.3077
OTT                     0.2446    1.1667    0.0476    0.0714      0.75     0.359    0.3214    0.5156    0.0821    0.1154
NOV                     0.2382         1    1.0238    0.1071     0.375    0.4103    0.4643    0.0156         0    0.3077
DIC                     0.2682    1.3333    0.0714    0.0714       0.5    0.8462    0.1429    0.3438    0.1014    0.1538
Lun                     0.4721      2.25    0.3095    0.2143     1.375    1.2308      0.75    0.3594    0.1836    0.2308
Mar                     0.4185    2.1667    0.2381    0.2143         1    0.7692    0.4643    0.4688    0.1836    0.3077
Mer                     0.4163    1.6667    0.2381    0.2143      0.75    0.7179    0.7143    0.6094    0.1884    0.2308
Gio                     0.4721    2.5417    0.1905    0.1786         1    0.6667    0.9643    0.5313    0.1884    0.4615
Ven                     0.4421         2    0.4048    0.1786     0.875    0.8718      0.75     0.375    0.1787       0.5
Sab                     0.4678    2.5833     0.381    0.0714     1.375    0.8718      0.75    0.4531    0.1498    0.4615
Dom                     0.2811      1.25    0.0952    0.3214      0.75    0.7949    0.3929    0.0938    0.1159    0.3846
6FO_3_6                 0.0408      0.25    0.0238         0         0    0.1795         0    0.0625    0.0048         0
6FO_7_10                0.6159    2.4167    0.3095      0.25     1.375    1.5128         1    0.5938    0.2512    0.8077
6FO_11_14               0.8627    4.5417    0.5714    0.3929         2    1.5641    1.2143    0.9063    0.3623    0.5385
6FO_15_18               0.7339       3.5     0.619       0.5     2.125    1.3333    0.9286    0.4844    0.3527    0.7308
6FO_19_22                0.515    2.5833    0.3095    0.1429      1.25    0.8462    1.1429    0.7031    0.1546    0.3462
6FO_23_2                0.2017    1.1667    0.0238    0.1071     0.375    0.4872       0.5    0.1406    0.0628    0.1538
TS_ScFL                 1.3348    5.8333    0.9762       0.5     2.125    3.3333    2.4643    1.4063    0.4928    0.7308
TS_ScL                  0.4378    2.5833    0.3333    0.0714     1.125    0.7179    0.6071    0.4063    0.1787    0.3462
TS_ScF                  0.1867    0.5417    0.1667    0.1071     0.125    0.2051    0.2143      0.25    0.1256    0.2692
TS_T                     0.324    1.9167    0.0476    0.0714     1.125    0.5128    0.3571    0.3438    0.1498    0.3462
TS_VcontrOst            0.3455    1.6667    0.1905    0.3214      0.75    0.7436    0.6071    0.1875    0.1643    0.2308
TS_IP                   0.2339    1.3333    0.0952    0.1786       1.5    0.3846    0.5357    0.2188     0.058         0
TS_FS                   0.1073    0.5833    0.0476    0.1429     0.375    0.0256         0    0.0781    0.0193    0.6538
CA_Ser                  2.2296   11.2083    1.1667    0.9643         6    4.4872    3.3929    2.1563    0.9275    1.7692
CA_Nuv                  0.5601    2.2917    0.5238    0.3929     0.875    1.0256    1.0357    0.5781    0.2271       0.5
CA_Pio                  0.1588      0.75     0.119    0.0357      0.25    0.4103    0.3571    0.1563    0.0242    0.2692
CA_AltrCon              0.0215    0.2083    0.0476         0         0         0         0         0    0.0097    0.0385
FS_Asciu                2.5365      12.5    1.6429         1      6.25    4.8718    4.0357    2.4531    1.0918    1.8846
FS_Bagn                 0.4099    1.9167    0.2143      0.25     0.875    1.0256      0.75    0.4375     0.087    0.5769
FS_Sdru                 0.0236    0.0417         0    0.1429         0    0.0256         0         0    0.0097    0.1154
P_Asf                   2.9227   14.3333    1.8571    1.2143      6.75    5.9231    4.6429    2.8906    1.1643       2.5
P_LastrePorf              0.03    0.0417         0    0.1429      0.25         0    0.1429         0    0.0145         0
P_AltreCond             0.0172    0.0833         0    0.0357     0.125         0         0         0    0.0097    0.0769
G_ConF                   2.073   10.2083    1.5238    0.6429      4.75    4.2308    3.5357    1.9531    0.7826    1.9231
G_NoF                   0.8391    3.9167    0.3333      0.75     2.375    1.6154    1.2143    0.8594    0.3913    0.3846
G_PReMort               0.0579    0.3333         0         0         0    0.0769    0.0357    0.0781    0.0145    0.2692
```

Figure 3. Road Safety – SimpleKmeans results

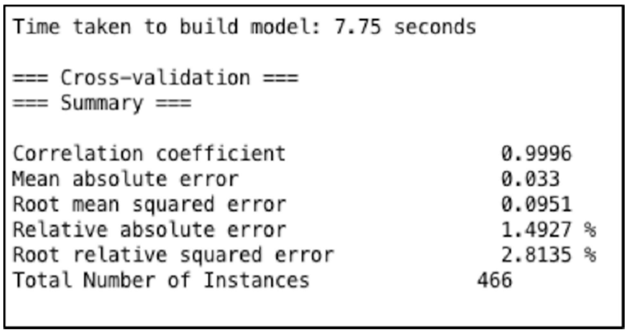
These results shows that the algorithm has performed a significant grouping. For example in the cluster 0 we have 24 instances, where there are the most dangerous roads with the greatest number of road accidents; in the cluster 1, with 42 instances, there are the roads with several characterizations on the number road accidents, with respect to the months (min number from May to October and in December; max number in November), with respect to the day of week (min number in Friday), with respect to type of collision (min number with road impact), with respect to weather conditions (min number with weather serene); in the cluster 2 (28 instances) there are principally squares and max number of road accidents with slippery road surface; and so on.

9.2. Association Rules results

We conducted several experiments using the Apriori algorithm on different groups of data. In this case, all the rules found by the Apriori algorithm are not semantically meaningful. As we know the associative rules describe correlations of events and can be seen as probabilistic rules: two events are correlated when they are frequently observed together. The affinity analysis process has not determined (interesting) objects that have such common characteristics. This means that the search for frequent item sets, which is a sub-problem that arises from the research of Association Rules, has not produced usable results in terms of interesting insights.

9.3. Artificial Neural Network results

We applied the MultilayerPerceptron algorithm with the goal of predicting the number of road accidents. The optimal conFigureuration for the network was made by setting the following parameters: learning rate L = 0.3, momentum M = 0.2, and number of epochs (training time) N = 1000. Figure 4 shows the result from the application of the Multilayer Perceptron algorithm.



```
Time taken to build model: 7.75 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.9996
Mean absolute error                      0.033
Root mean squared error                  0.0951
Relative absolute error                  1.4927 %
Root relative squared error              2.8135 %
Total Number of Instances              466
```

Figure 4. Road safety – MLP results

Figure 5 shows the comparison graph between real values and expected values:

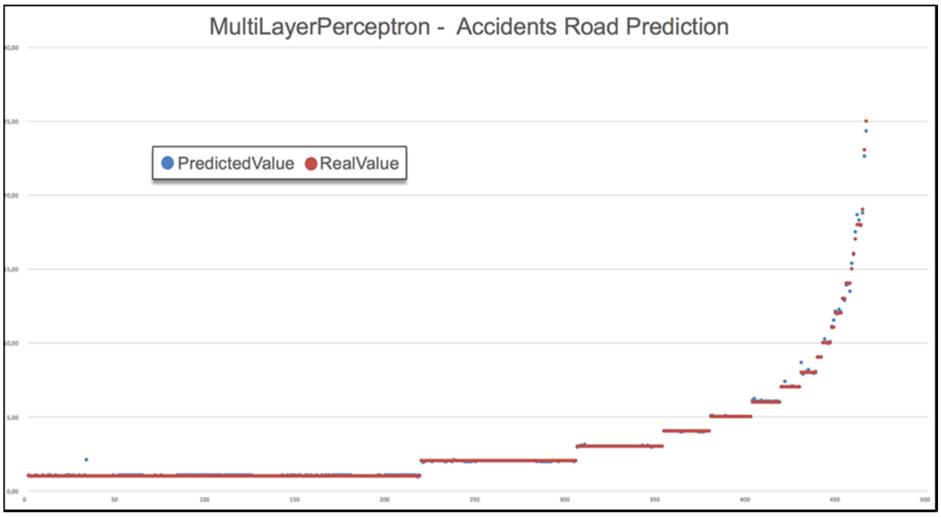


Figure 5. Road Safety – MLP results: Predictive Value vs Real Value

To analyze the performance of the several experiments executed, we evaluated the following measures: Correlation Coefficient, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). MAE is a statistical coefficient calculated as the sum of absolute errors divided by the number of predictions. It measures the nearness of the predicted model with respect to the actual model. In our case, we have MAE = 0.033. RMSE is calculated as the square root of the mean squared error; thus, it evaluates the differences between values predicted by a model and the values really observed. Lower values of RMSE and MAE therefore mean better prediction and accuracy. In our case, we have an RMSE score of 0.0951. To validate this training, we repeated the experiment, applying two test options: "Percentage Split" and "K-Cross-validation". In the first method, the prediction results are evaluated on a test set that is a part of the original data. The split is 66%, which means that 66% of the data go for training and the remaining 34% for testing. With K-cross validation, we set $K = 10$; this means that the original data set is partitioned into ten equal size subsets. We employ 9 partitions for learning and leave 1 partition out for testing. The cross-validation process is then repeated 10 times, each time leaving a different partition for testing. As can be seen from the results, the network is well trained. The network learned an input-output relationship and provided correct forecast values.

10. Discussion

From the experiments conducted, we can make several considerations on the most appropriate analysis technique based on the type of starting data set.

We presented three case studies, each of which concerns a very specific domain that is represented by a particular type of data set. In the first case study, educational domain, we analyzed a data set related to the educational learning experiences. The data set consists of a total of 300 instances (experiences), 42 attribute categories (and more than 300 attributes). The attributes are of the following types: categorical, integer, text, sequential, time-series. The data set is multivariate, with a strong semantic richness and described by a complex taxonomy.

The analysis of this data set has allowed to have interesting results in terms of search for clusters and associative rules between the attributes, while it was not possible to train a neural network in order to predict what the educational benefits can be obtained on the type of conducted teaching experience. Probably the strong semantics inherent in the data set, does not allow the MultilayerPerceptron algorithm to find a function able to relate the input data with the output.

In the second case study, concerning the health domain, we analyzed the data set related to the risk of cardiovascular disease. The CVD risk data set consists of 5134 instances (patients) described by 75 categorical, integer and real attributes. This data set is multivariate and with little semantic richness. The analysis techniques implemented in this case, led to significant results in all three methods of analysis, identifying: significant patient clusters, hidden relationships between attributes and a predictive model on the aggravation of patients. In the third and last case study of the safety domain, we explored the data set relating to road accidents. The data set consists of 466 instances (roads) and 58 attributes that describe the streets and the type of traffic accident reported. The attributes are simple and categorical or integer (mostly binary). It's a multivariate, but semantically poor data set. The application of the analysis techniques has brought good results with the SimpleKMeans clustering algorithm, which allowed to group the roads in a significant way according to road accidents. Through the MultilayerPerceptron artificial neural network, we obtained a predictive model for the number of road accidents assuming the characteristics involved. Instead, the Association Rule technique through the Apriori algorithm has not produced meaningful and semantically intuitive rules; probably this result is due to the poor semantics of this data set. In table 3 below, we have summarized the results obtained. Authors should discuss the results and how they can be interpreted in perspective of previous studies and of the working hypotheses. The findings and their implications should be discussed in the broadest context possible. Future research directions may also be highlighted.

Table 3. Different type's data set vs analysis techniques

Data set Name	**Educational Learning Experiences**	**CVD Risk**	**Road Safety**
Area	Learning	Health	Safety
Instances numbers	300	5134	466
Attributes numbers	39 categories of attributes	75	58
Attributes characteristics	Categorical, Integer, Text, Sequential, Time-Series	Categorical, Integer, Real	Categorical, Integer
Data set characteristics	Multivariate and with strong semantic richness	Multivariate and with little semantic richness	Multivariate and semantically poor
Associated analysis techniques	Clustering, Association Rule	Clustering, Neural Network, Association Rule	Clustering, Neural Network

This guidance might be useful to determine which technique is most appropriate for a new task.

11. Conclusions

This study contributes to research in data exploration and data mining by providing guidance on the selection of the most appropriate analysis techniques through a comparative analysis. Therefore, our research examines experimentally how the characteristics of the data set or different types of data sets affect the precision and complexity of the exploratory analysis and therefore the discovery of hidden knowledge. Several characteristics define a data set's structure and properties. These include the number and types of the attributes or variables, the strong or poor semantic richness. Through the experiments conducted in three case studies related to completely different domains and with very different data sets, we could establish a relationship between the analysis techniques implemented, that is Clustering analysis, Association Rule and Neural Network and the analyzed data set. We have found that the technique of cluster analysis can be said to be independent of the type of data set and in particular independent of the presence or absence of semantics, since in all three case studies we have obtained satisfactory results and semantically significant groupings of the instances. The Association rule technique depends in some way on the presence of semantics because the best results were in the presence of semantic richness, while in the opposite case no meaningful rules were found. The predictive artificial neural networks train well in the presence of data sets with a simple and semantically poor structure, while they cannot find a function that links the inputs to the outputs in case of complex data sets. Our aim has been to determine a guide governing when certain algorithm should be recommended. Based on the results presented here, future research effort should focus on investigate to more types of analysis techniques and algorithms. We plan to explore further the possible relationships between the analysis and exploration techniques and data set type.

Acknowledgments

The authors would like to thank the Professors Paolo Paolini and Luca Mainetti, the scientific managers of EDOC@WORK3.0 (Education and Work on Cloud) Italian national research project, the Health Organization of Salento (ASL Lecce) and the Local Police of the Lecce city (Salento region, Italy) for making the data available. A particular thank to the manager Rodolfo Rollo for the inspiring conversations.

References

Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In Proc. 20th int. conf. very large data bases, VLDB (Vol. 1215, pp. 487-499).

Das, K., & Behera, R. N. (2017). A survey on machine learning: concept, algorithms and applications. International Journal of Innovative Research in Computer and Communication Engineering, 5(2), 1301-1309.

Di Blas, N., & Paolini, P. (2013, March). Technology and group work: inclusion or diversification of talents?. In Learning & Teaching with Media & Technology. ATEE-SIREM Winter Conference Proceedings, 7Á9 March (p. 218Á231).

Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. Journal of cybernetics, 4(1), 95-104.

El-Halees, A. (2009). Mining students data to analyze e-Learning behavior: A Case Study.

Falcinelli, F. (2012, November 20-30). Evidence-Based Research About the Impact of ICT on Italian Schools: The Cl@ssi2.0 Project. Online Educa Berlin 2012. Berlin.

Falcinelli, F. & Laici, C. (2012). Teaching with ICT: The Policultura and Moodle Didactic Format Experimented in Schools, IJCEE, January-March 2012, Vol. 2, No. 1.

Giudici, P. (2005). Applied data mining: statistical methods for business and industry. John Wiley & Sons.

Halkidi, M., Vazirgiannis, M., & Batistakis, Y. (2000, September). Quality scheme assessment in the clustering process. In European Conference on Principles of Data Mining and Knowledge Discovery (pp. 265-276). Springer, Berlin, Heidelberg.

Hammer, B., He, H., & Martinetz, T. (2014). Learning and modeling big data. In ESANN (pp. 343-352).

Kiang, M. Y. (2003). A comparative assessment of classification methods. Decision Support Systems, 35(4), 441-454.

Nazeer, K. A., & Sebastian, M. P. (2009, July). Improving the Accuracy and Efficiency of the k-means Clustering Algorithm. In Proceedings of the world congress on engineering (Vol. 1, pp. 1-3).

Oreski, D., Oreski, S., & Klicek, B. (2017). Effects of data set characteristics on the performance of feature selection techniques. Applied Soft Computing, 52, 109-119.

Paiano, R., & Pasanisi, S. (2017). A New Challenge for Information Mining. BRAIN. Broad Research in Artificial Intelligence and Neuroscience, 8(2), 63-80.

Pasanisi, S., & Paiano, R. (2018). A Hybrid Information Mining Approach for Knowledge Discovery in Cardiovascular Disease (CVD). Information, 9(4), 90.

Peng, Y., Wang, G., Kou, G., & Shi, Y. (2011). An empirical study of classification algorithm evaluation for financial risk prediction. Applied Soft Computing, 11(2), 2906-2915.

Reutemann, P., Pfahringer, B., & Frank, E. (2004, December). A toolbox for learning from relational data with propositional and multi-instance learners. In Australasian Joint Conference on Artificial Intelligence (pp. 1017-1023). Springer, Berlin, Heidelberg.

Rosenblatt, F. (1961). Principles of neurodynamics. perceptrons and the theory of brain mechanisms (No. VG-1196-G-8). CORNELL AERONAUTICAL LAB INC BUFFALO NYJ.

Song, Q., Wang, G., & Wang, C. (2012). Automatic recommendation of classification algorithms based on data set characteristics. Pattern recognition, 45(7), 2672-2689.

Suresh, J., & Ramanjaneyulu, T. (2013). Mining Frequent Item sets Using Apriori Algorithm. Int. J. Comput. Trends Technol, 4, 760-764.

Swathi, R., & Seshadri, R. (2017, June). Systematic survey on evolution of machine learning for big data. In Intelligent Computing and Control Systems (ICICCS), 2017 International Conference on (pp. 204-209). IEEE.

Temizel, T. T., Mizani, M. A., Inkaya, T., & Yucebas, S. C. (2007, November). The effect of data set characteristics on the choice of clustering validity index type. In Computer and information sciences, 2007. iscis 2007. 22nd international symposium on (pp. 1-6). IEEE.

Witten, Ian, H., et al. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. IEEE transactions on evolutionary computation, 1(1), 67-82.

Roberto Paiano (b. June 14, 1958) graduated in Electronic Engineering at the University of Bologna. He worked in IBM for 10 years. He was team leader at IBM RNSL and Project Manager at the CORINTO Consortium (National Research Consortium about Object-Oriented Technology). He was member of the IEEE. Currently, he is assistant professor at University of Salento (Italy). He has authored papers about information systems, Web modelling and design, metrics for the Web development. His current research interests are: the methodology of design of Web information systems, the automatic code generation using Open-Source Frameworks and Information Systems modelling.

Stefania Pasanisi (b. October 6, 1978) graduated in Automation Engineering at the University of Salento (Italy) in April 2009. After the degree she worked for five years in the company (Lecce) on projects for observational study and experimental project in the medical field and for design and development software and web applications. Since November 2014 she is a PhD student in Engineering of Complex Systems at the University of Salento (Italy). Her main research areas include advanced semantics exploration techniques on dynamic and complex information spaces, Exploratory Computing Technique and Data Mining. She participates to several research projects and she is (co-) author of several scientific papers.