A data mining approach for bill of materials for motor revision

Francesco Maiorana
Department of Electrical, Electronic and Computer Engineering, University of Catania. Viale Andrea Doria, 6, Catania, Italy
Email: francesco.maiorana@dieei.unict.it

Angelo Mongioj
ENI - Refining and Marketing
Via Laurentina, 449 - 00142 Roma, Italy
Email: angelo.mongioj@eni.com

***Abstract*—Supply *chain management is a core* business process and is today considered the focus of competitive analysis. Business enterprises are data overloaded and, hence, using data mining techniques to transform the vast amount of data into meaningful information can be extremely beneficial. We will present a data mining approach for inventory forecasting and planning a Bill Of Materials in a highly competitive environment such as an Italian car racing team. By exploiting clustering algorithms and by using statistical techniques to identify the optimal number of clusters this work presents a method to optimally cluster a multi-year dataset containing the products used in car revision after each rally competition during a three-year period. The Bill Of Materials was used as input for the Material Requirements Planning.**

I. Introduction

SUPPLY chain management (SCM) has become a central aspect in modern manufacturing. In [1] the author defines SCM as “the management of upstream and downstream relationships with suppliers and customers to deliver superior customer value at less cost to the supply chain as a whole”. One of the key aspects is minimizing cost without reducing quality and assuring in time delivery of the products. SCM has to optimize several activities from obtaining raw materials to the delivery of the final product to the customers [2].

The difficulty of decision making in this fields arises from the globalization of competition in a fast changeable market with changing customer demand. In the meanwhile a fast growing amount of information is collected at different stages in the SCM and, hence, a semi-automatic and integrated approach is becoming more and more necessary. In this scenario data-mining tools and techniques can offer a valuable instrument to transforming vast amounts of data into meaningful information. Several reviews are present in literature that survey the state of the art of data-mining application in SCM such as [3-6].

In a recent study [7] the authors go even further by asserting that there is a “significant need for research on universal data integration and data storage concepts for data mining in manufacturing to generate versatile pre-configured and truly process-centric data mining applications that can be adapted to heterogeneous manufacturing environments and different branches”.

In this scenario inventory management, or rather the organization that leads to the availability of items to the costumers, is considered one of the most important segments of SCM and must be performed in difficult conditions due to the increased fluctuation in demand, lead time, difficulties in choosing the best production scheduling and the demand information distortion due to the bullwhip effect [8]. In this segment of SCM, data mining has been exploited to improve performance also.

In [9] data mining techniques have been applied for supply chain inventory forecasting. In particular, their work focuses on out of stock prediction of spare parts inventories in different store locations. They use clustering algorithms for grouping together different stores that exhibit a similar aggregate sale patter, then algorithms such as decision trees and neural networks are used to build a more accurate out of stock forecasting model.

Similarly, in [10] the authors used data mining techniques in two case studies. In particular, by applying a Multi Layer Perceptron (MLP) and a Time Delay Neural Network (TDNN) the authors developed a data mining model, obtaining a 50% reduction of the inventory cost of drugs in a large medical distribution company.

In [11] the authors present an algorithm for finding the minimum weighted symmetric difference between two Bill Of Materials (BOM) trees. The comparison result is used as a distance metric for a clustering algorithm in order to group different BOMs into families of products.

This work will present a data mining approach for clustering the product used in the revision process performed after each car racing competition by a leading Italian team. The engine revision after each competition requires a careful planning of component supply. The mechanical pieces are in some cases expensive since they are built with special requirements in order to maximize performance at the expense of their duration. Moreover, their life span is short due to the varying car regulations as well as to the frequent design and production of new engines

978-83-60810-51-4/$25.00 © 2012 IEEE

with increased performance in order to maximize winning probabilities.

I. Beside the major driving force of competing for winning that requires the highest level of quality there is the necessity to reduce cost and expenses in order to maximize profit. In this highly competitive environment the only possibility for the two opposite driving forces to coexist is by reducing cost through reducing inefficiency, without any loss in quality. In this scenario this work reports on experiments with a hierarchical clustering algorithm whose performance in terms of classification results are maximized by means of the Marriott statistical criterion that is able to give an indication of the adequacy of the clustering process. By experimenting with different algorithm parameters and by finding the optimal number of clusters for the chosen parameter, it is possible to obtain an optimal clustering of the components that are particularly significant in the revision process and allows us to obtain a BOM that can be used as input for the Material Requirements Planning (MRP).

This work is organized as follows: section 2 presents the data mining tool used and in particular the hierarchical clustering algorithm and the parameters that affect its performance, together with the Marriot criterion used to find the optimal number of classes that leads to the best clustering; section 3 presents the case study while section 4 draws some conclusions and highlights future work.

II. The clustering algorithm

Data mining is a well-established discipline that encompass knowledge from various disciplines ranging from artificial intelligence to statistics. It is a process that allows us to extract useful information and patterns from large amounts of raw data. According to [12] a data mining process consists of the following phases:

- Goal definition.
- Selection, organization and pre-treatment of data.
- Exploratory data analysis and eventual transformation.
- Design and choice of the analysis process and tools.
- Data elaboration and analysis.
- Evaluation and model comparison with a final choice for the best model.
- Results interpretation and usage in the decision process.

Among the data analysis method, clustering techniques have been widely adopted in many fields such as knowledge discovery from text [13] where the authors used K-means to cluster sets of documents and extract meaningful associations among the biological entities characterizing the associated classes. In [14] the authors designed a parallel version of a variant of a Self Organizing Map (SOM) and deployed and tested it on a grid or cloud [15] infrastructure. The input noise robustness of the SOM algorithm on sparse dataset was also evaluated in [16]. A strategy for choosing the optimal number of clusters was designed in [17].

For a survey on clustering algorithms, the reader can reference [18] where the reader can find information on the different type of algorithms and techniques that can be used for clustering. According to [19] cluster analysis is the "task of organizing a set of objects into meaningful groups. The groups can be disjoint, overlapping or organized in some hierarchical fashion". What "meaningful" means is usually application dependent. Usually "meaningful" means groups with a maximized similarity between the objects of each group and a minimized similarity between pairs of objects belonging to different groups. The measure of similarity varies from one algorithm to another.

Clustering algorithms can be classified on the basis of:

1. The underlying methodology: partitional or agglomerative.
2. The structure of the final solution: hierarchical or nonhierarchical.
3. The characteristics of the space in which they operate: the features or the similarity.
4. The type of cluster they discover: globular or transitive.

Partitional clustering algorithms start either with a cluster containing all the elements and find subsequent clusters by a sequence of repeated bisections or with a set of k clusters that are refined by the process. The process iteratively searches for the best division according to a specific division rule that usually tries to maximize differences among groups and minimize the differences inside the elements of the group. The number of clusters is usually determined by the user a priori but it can also be automatically derived. Partitioning algorithms can be viewed as optimizing algorithms that aims at optimizing at each iteration an objective function usually given by a combination of intra-cluster similarity and inter-cluster dissimilarity.

Agglomerative algorithms start with a number of clusters equal to the number of elements to cluster; at each iteration two clusters are merged according to a defined criterion, until a stopping condition is met. The objective function is optimized at each iteration by looking at all the possible pairs of clusters, leading to a locally optimized solution. In this way the algorithm iteratively finds the two closest elements, joins them by computing the center of the new cluster and iteratively proceeds until all the elements are joined in one class.

Hierarchical clustering constructs a tree-like partition. With this method the elements that are merged or divided will remain merged or divided until the end of the classification process. Hierarchical clustering algorithms can be agglomerative or divisive.

The clustering algorithm can operate on the object's feature space or on a derived similarity space. The similarity space representation is usually obtained by computing a similarity measure between all the pairs of objects on the basis of their features.

What differentiates globular and transitive clusters is the relationship between the cluster's object and the dimension of the feature space:

- For globular clusters there is a subset of the original space dimension (a subspace, or a fraction of the features used for clustering) in which a large percentage of the objects belonging to the same cluster agree (this is a typical problem in document classification).
- For transitive clusters it is also possible to find a subspace of the features used for clustering, but in this case the elements of the cluster share a very small number of features. However, it is possible to find “a strong path” or a sequence of subspaces that links A and B in which each pair of clusters in the chain shares many features of the subspace.

In this work we have used a hierarchical agglomerative clustering algorithm. According to [20] the main parameters of this type of algorithm are the following:

- The measure of proximity or distance used to find the closest pair of elements or clusters. Among the different types of distance we may recall the Euclidean distance, the cosine distance, the city block distance and so on.
- The method used to compute the distance between the new formed cluster and the other elements such as:
    - Single linkage is based on the shortest distance among the elements of the group and the other units.
    - Complete linkage is based on the furthest distance among the elements of the groups and the other units.
    - Centroid: the distance between the centroids of two clusters.
    - Weighted or un-weighted average distance.
- The way by which the representative of each new formed cluster is chosen.

The approach used was to compute the entire dendogram, i.e. the complete hierarchical tree representing the successive aggregation.

One of the most difficult problems to cope with in clustering analysis is to find the optimal number of clusters. Different approaches have been used in literature such as in [17] where the authors using distance based metrics present a tool that automatically gives hints to the user guiding him in adding or deleting neurons in a SOM. In [21] the authors use an information theory criterion based on entropy estimation in a partitional hierarchical clustering algorithm. In [22] the authors take a dissimilarity matrix between patterns as the basic measure for characterizing clusters. The statistical distribution of dissimilarity increments between neighboring patterns within a cluster was modeled by exponential density and used to characterize context and to derive a new cluster isolation criterion. The criterion was integrated into a hierarchical agglomerative clustering algorithm.

We chose to use, on the computed dendogram, the Marriot statistical criterion to choose the optimal number of clusters i.e the best point to cut the dendogram.

The goal of an optimal clustering algorithm is to find the best clustering in order to maximize the variance between the different classes and minimize the variance among the elements of the same class.

The variance is given by:

$$\frac{\sum_{i=1}^{N}(x_i - M(x))^2}{n-1} \qquad (1)$$

Where M(x) is the mean of the N elements in the cluster. The numerator can be written as:

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\left|x_{ij} - M\left(x_j\right)\right|^2 + \sum_{i=1}^{N}\sum_{j=1}^{N}\left|M\left(x_j\right) - M\left(x\right)\right|^2 \qquad (2)$$

where the first addendum is the variance inside the group while the second addendum represents the external variance. Increasing the number of elements in the cluster increases the internal variance; diminishing the number of clusters increases the variance among the cluster, i.e. the external variance. The Marriott criterion searches for the minimum of:

$$M = g^2 \frac{\det\left|W\right|}{\det\left|T\right|} \qquad (3)$$

where the numerator is the determinant of the internal variance-covariance matrix (Within group, W) and the denominator is the determinant of the sum of the variance-covariance matrix (Between groups B) and Within groups (W), and finally g is the number of clusters.

III. Data Analysis

We used a dataset containing information about the product used in the revision process performed after each car racing competition by the Fiat Powertrain Technologies (FPT), a leading Italian automotive firm involved in word wide car competitions.

All data analyses were performed by means of an in house tool developed using Matlab.

We use a data set containing information about car component replacement after each car competition in a three-year time span. The dataset consists of roughly 10^4 rows with 21 variables. Among the variables characterizing the dataset we find:

- Customer/Supplier: this is the code of the customer.
- Operation code: code identifying the type of operation.
- Date: the date of the operation.
- Article code: the code identifying each article.
- Article description: contains the description of the article.
- Quantity: the quantity associated with the operation.
- Type of engine: a code that identifies the type of engine.

Table 1 summarizes the number of modalities of some dataset variables. The variables were analyzed in terms of their variability, homogeneity and heterogeneity, skeweness and the complete set of tools for exploratory data analysis as reported in [23].

TABLE I.
THE NUMBER OF MODALITIES OF SOME DATASET VARIABLES

Variable	**Modality**
Customer/Supplier	15
Operation code	15
Article code	799
Type of engine	214
Motorization	11
Document number	1592

The first step was a data transformation. The values of qualitative variables were replaced by their cardinal number which represents the modality of the value. The dataset was then normalized. We experimented both with a min-max normalization and with a normalization with null mean and standard deviation 1. With both types of normalization procedure we arrived at the same conclusion.

We then performed a univariate, bivariate and a multivariate analysis. Fig. 1 shows the Pareto diagram regarding the article code. From the diagram we can observe that 25% of article codes represent 80% of the observations. All the main exploratory data analysis steps, as described in [24], were performed.

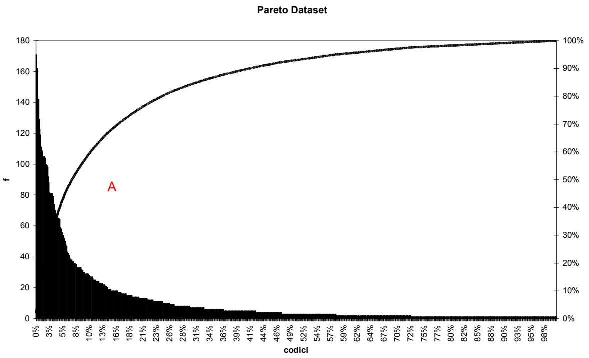
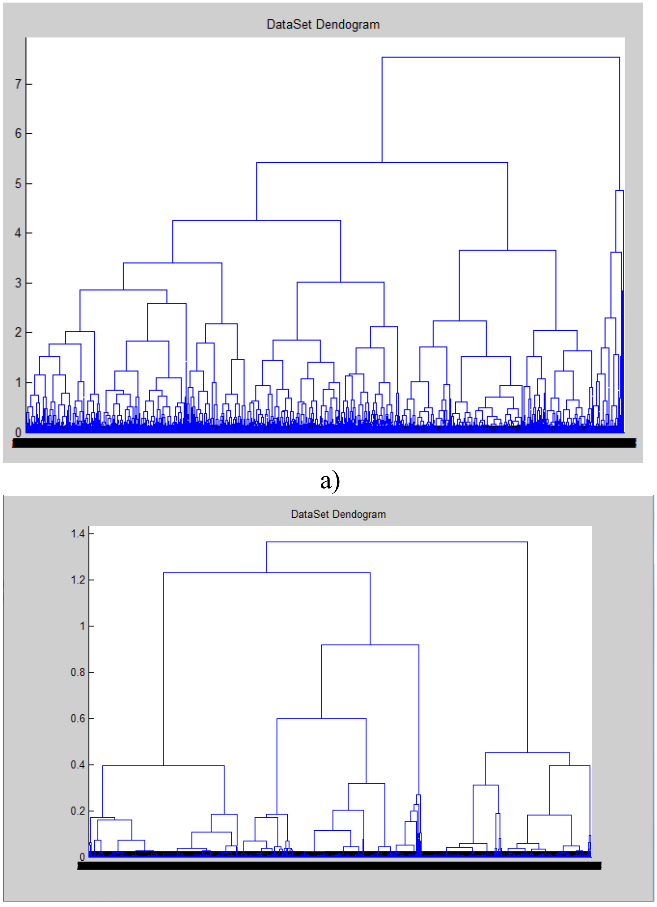


Fig. 1: Pareto diagram relative to the variable Article code.

The following step was a hierarchical agglomerative clustering algorithm using different distance functions and different linkage criterion. Fig. 2 shows the results of the hierarchical agglomerative clustering.



a)

b)

Fig. 2: Hirarchical clustering algorithm results: with Euclidean distance a); and with cosine distance b).

In particular Fig. 2 a) shows the results with Euclidean distance and Fig. 2 b) shows the results with cosine distance. In both cases we used the complete linkage method.

We than applied the Marriott criterion to iteratively find the optimal number of clusters. Table 2 reports the value of the Marriott index found for different number of clusters with Euclidean distance and complete linkage.

Table 3 reports the number of elements in each cluster for different type of linkage method and the optimal number of clusters obtained by applying the Marriott criterion in each case.

The result of the classification process is depicted in Fig. 3, which shows the first two normalized coordinates of the dataset used by the clustering algorithm. Fig. 3 a) shows the optimal clustering result in five classes with the complete linkage method; fig. 3 b) shows the clustering result in eight classes with the average linkage.

TABLE II.
MARRIOTT INDEX FOR DIFFERENT NUMBER OF CLUSTERS

Number of clusters	Marriott index
2	7.96 E-05
3	2.79 E-08
4	2.55 E-08
5	9.11 E-11
6	1.11 E-10
7	2.48 E-10
8	3.20 E-10

TABLE III.
NUMBER OF ELEMENTS IN EACH CLUSTER FOR DIFFERENT LINKAGE METHODS

Linkage	**N° cluster**	**Number of elements in the cluster**
Single	7	444;7;14;4;8299;281;1
Complete	5	2316;3341;21;450;2924
Average	8	2507;2868;259;2665;22;429;281;21
Weighted	5	1101;3550;3929;21;451

Several experiments were also performed with different distance measures. The best clustering result was obtained with five classes.

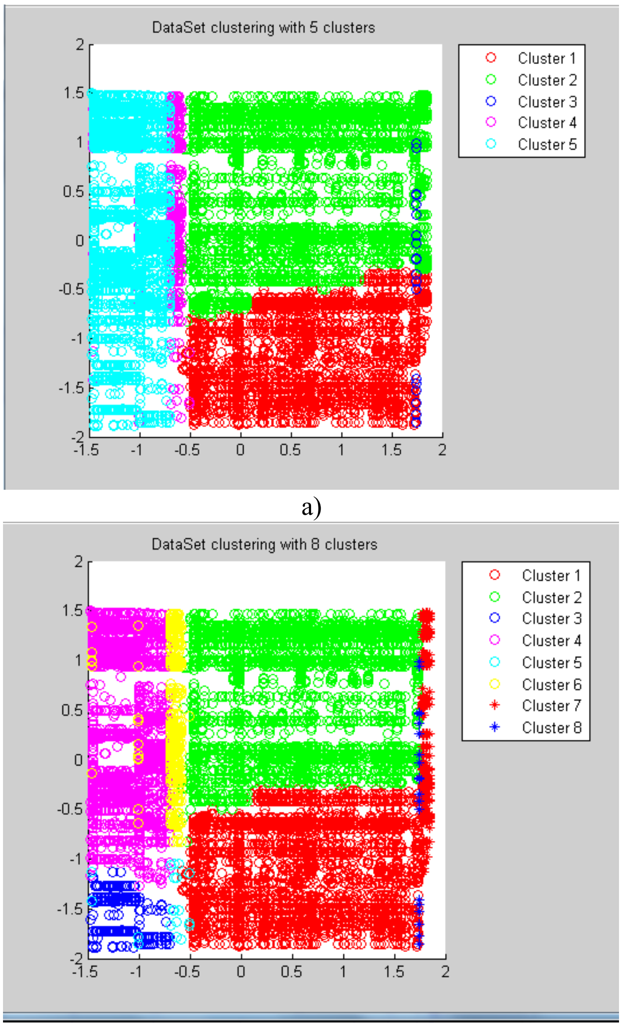
The clustering process allowed us to identify the class of components with a high frequency of substitution. The result was almost in line with an analysis of the revision process where it was possible to confirm the nature of highly deteriorable pieces. In a similar manner in the chosen optimal clustering, the two classes with fewest element contains pieces with a very low usage frequency.

The result is similar to [9] where the authors suggest reducing the inventory for popular items and increasing the inventory for less popular or unpopular items. The results were used to plan a reliable BOM, that was used in MRP. A further analysis is nevertheless required in order to apply the same conclusion in the presented scenario due to the differences in requirements. The clustering results can be used to apply different BOM strategies for each class in the chosen optimal clustering.

IV. CONCLUSION

This work has presented a data mining approach for inventory forecasting and BOM planning in a highly competitive environment such as the FTP Italian car racing team. By exploiting clustering algorithms and by using the Marriott statistical criterion to identify the optimal number of clusters, we optimally clustered a dataset containing the products used in car revision after each rally competition during a three-year period.

As future work we plan to extend the analysis to wider datasets by applying and comparing results with other clustering algorithms and using neural networks such as SOM deployed in cloud or grid architecture, if this is required by the size of the dataset.



a)

b)

Fig. 3: clustering result representation with different linkage methods: complete a); average b).

It is also possible to use a shared memory design model [25-27] in order to organize projects around data mining problems allowing the acquisition of a variegated knowledge of different solutions, techniques and algorithms. In further work it is also possible a deepening of user interface design in order to allow a better visualization of the results even using graphical and multimedial elements [28-29].

REFERENCES

[1] M. Christopher, "*Logistics and Supply Chain Management*" (4th Edition), Financial Times Prentice Hall, 2011.

[2] E. Roghanian, S. J. Sadjadi, and M. B. Aryanezhad, "A probabilistic bi-level linear multi-objective programming problem to supply chain planning," Applied Mathematics and Computation, vol. 188, no. 1, pp. 786-800, 2007.

[3] J. A. Harding, M. Shahbaz, Srinivas et al., "Data Mining in Manufacturing: A Review," Journal of Manufacturing Science and Engineering, vol. 128, no. 4, pp. 969-976, 2006.

[4] R. Carbonneau, K. Laframboise, and R. Vahidov, "Application of machine learning techniques for supply chain demand forecasting," European Journal of Operational Research, vol. 184, no. 3, pp. 1140-1154, 2008.

[5] A. Choudhary, J. Harding, and M. Tiwari, “Data mining in manufacturing: a review based on the kind of knowledge,” Journal of Intelligent Manufacturing, vol. 20, no. 5, pp. 501-521, 2009.

[6] M. Ko, A. Tiwari, and J. Mehnen, “A review of soft computing applications in supply chain management,” Applied Soft Computing, vol. 10, no. 3, pp. 661-674, 2010.

[7] C. Gröger, F. Niedermann, B. Mitschang, “Data mining-driven manufacturing process optimization,”, Proceedings of the World Congress on Engineering 2012, vol III WCE 2012, July 4 - 6, 2012, London, U.K

[8] A. Ancarani, C. Di Mauro, D. D'Urso, “An experimental investigation of the effects of supply uncertainty on supply chain performance,” POMS 22 nd Annual Conference, Reno, Nevada, U.S.A., 2011.

[9] N. Stefanovic, D. Stefanovic, and B. Radenkovic, "Application of Data Mining for Supply Chain Inventory Forecasting Applications and Innovations in Intelligent Systems XV," R. Ellis, T. Allen and M. Petridis, eds., pp. 175-188: Springer London, 2008.

[10] A. Dhond, A. Gupta, and S. Vadhavkar, “Data mining techniques for optimizing inventories for electronic commerce,” in Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, Boston, Massachusetts, United States, 2000, pp. 480-486.

[11] C. J. Romanowski, and R. Nagi, “On comparing bills of materials: a similarity/distance measure for unordered trees,” Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, vol. 35, no. 2, pp. 249-260, 2005.

[12] J. Han, M. Kamber, *Data Mining: Concepts and Techniques*, 2nd ed. The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor, 2006.

[13] A. Faro, D. Giordano, F. Maiorana, C. Spampinato, C, “Discovering Genes, Diseases Associations from Specialized Literature Using the GRID,” *IEEE Transactions on Information Technology in Biomedicine*, vol.13, no. 4, pp. 554-560, 2008.

[14] A. Faro, D. Giordano, F. Maiorana, “Mining Massive Datasets by an Unsupervised Parallel Clustering on a GRID: Novel Algorithms and Case Study,” *Future Generation Computer Systems*, vol. 27, no. 6, pp. 711-724, 2011 .

[15] F. Maiorana, G. Fazio, “Knowledge Discovery from Text on a Cloud Architecture and its Application to Bioinformatics,” in *Proc. 9th International Conference on Biomedical Engineering, IASTED* , 2012.

[16] A. Faro, D. Giordano, F. Maiorana, “Input Noise Robustness and Sensitivity Analysis to Improve Large Datasets Clustering by Using the GRID”. *Discovery Science, Lecture Notes in Computer Science* vol. 5255, pp. 234-245: Springer Berlin/Heidelberg, 2008.

[17] A Faro, D Giordano, F Maiorana, "Discovering complex regularities from tree to semi–lattice classifications", International journal of computational intelligence, vol.2, no. 1, pp. 34-39, 2005-

[18] R. Xu, D. Wunsch, “Survey of Clustering Algorithms”, IEEE Transactions on Neural Networks, vol. 16, no. 3, May 2005.

[19] Y. Zhao, G. Karypis, “Data clustering in life science”, Molecular Biotechnology, vol 31, no.1, pp. 55-80, 2005.

[20] P. Giudici, S. Figini, *Applied Data Mining for Business and Industry*”, 2nd Edition, John Wiley & Sons, 2009.

[21] M. Aghagolzadeh, H. Soltanian-Zadeh, B. N. Araabi et al., "Finding the Number of Clusters in a Dataset Using an Information Theoretic Hierarchical Algorithm." pp. 1336-1339.

[22] A. L. N. Fred, and J. M. N. Leitao, “A new cluster isolation criterion based on dissimilarity increments,” Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 25, no. 8, pp. 944-958, 2003.

[23] F. Maiorana, A. Mongioj, M. Vaccalluzzo, “A data mining E-learning tool: description and case study,” Proceedings of the World Congress on Engineering 2012, vol I WCE 2012, July 4 - 6, 2012, London, U.K

[24] F. Maiorana, “A Teaching Experience on a Data Mining Module,” Proceedings of Information Systems Education & Curricula Workshop, Wroclaw, Polland, September 9-12,2012.

[25] A. Faro, D. Giordano, “StoryNet : an Evolving Network of Cases to Learn Information Systems Design,” *IEEE Proceedings SOFTWARE*, vol.145, no. 4, pp. 119-127, 1998.

[26] A. Faro, D. Giordano, “Concept Formation from Design Cases: Why Reusing Experience and Why Not,” *Knowledge Based Systems Journal*, vol.11, no. 7, pp. 437-448. Elsevier, 1998.

[27] A. Faro, D. Giordano, “Design memories as evolutionary systems: socio-technical architecture and genetics,” *IEEE Proc. Int. Conf. on Systems, Man and Cybernetics*, Washington, D.C. USA., vol.5, pp. 4288-4293, IEEE, 2003.

[28] D. Giordano, “Evolution of interactive graphical representations into a design language: a distributed cognition account,” *International Journal of Human-Computer Studies* Vol. 57, no. 4, pp. 317-345, 2002.

[29] A. Faro, D. Giordano, “Ontology, esthetics and creativity at the crossroad in information systems design,” *Knowledge-Based Systems*, vol.13, no. 7, pp. 515-525, Elsevier, 2000.