International Journal of Electrical and Computer Engineering (IJECE)
Vol. 9, No. 3, June 2019, pp. 2094~2102
ISSN: 2088-8708, DOI: 10.11591/ijece.v9i3.pp2094-2102

Data mining techniques application for prediction in OLAP cube

Asma Lamani, Brahim Erraha, Malika Elkyal, Abdallah Sair
Laboratory of Industrial Engineering and Computer Science (LG2I), National School of Applied Sciences, University Ibn Zohr Agadir, Morocco

Article Info

Article history:

Received Jun 5, 2018
Revised Nov 28, 2018
Accepted Dec 25, 2018

Keywords:

Automatic learning
Clustering
Data mining
OLAP
Prediction

ABSTRACT

Data warehouses represent collections of data organized to support a process of decision support, and provide an appropriate solution for managing large volumes of data. OLAP online analytics is a technology that complements data warehouses to make data usable and understandable by users, by providing tools for visualization, exploration, and navigation of data-cubes. On the other hand, data mining allows the extraction of knowledge from data with different methods of description, classification, explanation and prediction. As part of this work, we propose new ways to improve existing approaches in the process of decision support. In the continuity of the work treating the coupling between the online analysis and data mining to integrate prediction into OLAP, an approach based on automatic learning with Clustering is proposed in order to partition an initial data cube into dense sub-cubes that could serve as a learning set to build a prediction model. The technique of data mining by regression trees is then applied for each sub-cube to predict the value of a cell.

Copyright © 2019 Institute of Advanced Engineering and Science.
All rights reserved.

Corresponding Author:

Asma Lamani,
Laboratory of Industrial Engineering and Computer Science (LG2I),
National School of Applied SciencE,
University Ibn Zohr,
Agadir, Morocco.
Email: asma.lamani@gmail.com

1. INTRODUCTION

Data warehouses are databases of information specifically structured for analysis and decision making [1]. The data are subject-oriented, integrated, time-variant, and non-volatile organized in a multidimensional way [2]. The star schemas initially produce data cubes suitable for analysis. Secondly, it is to the user to browse, explore and analyze a cube's data to extract relevant information for decision making. This is an online analysis using OLAP technology, a data cube is a multidimensional representation of the data, each cell in a data cube represents an aggregated fact described by analysis axes. These correspond to the dimensions of the cube. The fact is observed by a measure which is associated with an aggregation function (SUM, AVG, MAX, MIN). A dimension can be organized in hierarchy, therefore the facts can be observed according to different levels of granularity. The information is therefore aggregated in the cube according to the user needs.

In another side, data mining can extract knowledge from data and has a wide variety of methods with different analytical objectives. In a decision-making process, a user observes the OLAP cube facts in order to extract useful information. This also allows the user to anticipate the realization of future events. Indeed, the OLAP technology is limited to exploratory tasks and does not provide automatic tools to help and guide the user in the deepening of his analysis, to explain values of cells, existing associations in the multidimensional data, or to predict values in the data cube.

While datamining techniques are effective approaches to guide the analyst and extract new knowledge. Thus a new problematic OLAP has appeared. Since the end of the 90s, several works [3]-[5] propose to associate the principles of the OLAP with the methods of data mining to enrich the online analysis and no longer limit it to a simple exploration or a simple data visualization. Coupling online analysis and data mining is then referred to terms as OLAM (On-Line Analytical Mining) [3], OLAP Intelligence, Multidimensional Mining.

As part of the coupling between online analysis and data mining, we propose in this work a decision support process that combines OLAP technology with unsupervised classification techniques and prediction techniques to integrate the prediction in cube. In fact, to have more precision in the prediction, it is essential to work on a homogeneous dataset, that's why we propose to partition the initial cube in dense sub-cubes by applying methods of clustering then applying a regression tree technique for the construction of the prediction model.

2. RELATED WORK

Several approaches have been proposed for coupling data mining and online analysis to extend OLAP to prediction. In the work of Riad Ben Messaoud [6], the author has defined three coupling approaches, a process of transforming multidimensional data into two-dimensional data, the second approach is based on the exploitation of tools offered by multidimensional database management systems, and the third is to evolve the data mining algorithms to adapt them with the types of data handled by the cubes.

As part of the coupling, new proposals are emerging. They consist of using statistical and machine learning methods for prediction in order to enrich the capabilities of online analysis. Sarawagi and al [7] use prediction by building a cube of predicted values from the initial data cube, the learning base is the original cube, and the model is based on a log-linear regression. Deviations between the two cubes can indicate to the user exceptional values. These exceptional cells are then signaled to the user when navigating the data cube with three indicators that also show him interesting paths to explore.

The work of Cheng [8] is aimed at predicting new facts. So he proposes to generate a new cube using a generalized linear model. The resulting cube corresponding to the prediction model. Han and al [9] proposes to predict a new fact measure by identifying subsets of interesting data. The predictive model is a cube where the measure indicates a score or a probability distribution associated with the measure value that can be expected in the original cube, resulting cube corresponds to the model to be used for prediction.

Y. Chen and Pei's proposal [10] consists of building cubes based on linear regression. From the initial data cube, a cubic measure is generated where each value indicates the weight of evidence. Continuing on the work of Sarawagi, which focuses on the aid to navigation and also on the explanation of the facts, A.Sair [11] pushes the limits of exploratory navigation by injecting prediction techniques at the heart of OLAP processes, this work is based on the integration of a complete learning process in OLAP for online data mining. A complete process then includes a selection phase of the explanatory variables, a fact sharing phase in a learning sample and a test sample. Next, a learning phase and a validation phase are executed. A.Sair [11] proposes an approach based on automatic learning with regression trees in order to predict the value of an aggregate or a measure.

Other work involves applying methods for partitioning OLAP cubes. The research of R.Missaoui and C.Goutte [12] proposes to analyze the potential of a probabilistic modeling technique, called “non-negative multi-way array factorization”, for approximating aggregate and multidimensional values. Using such a technique, they compute the set of components (clusters) that best fit the initial data set and whose superposition approximates the original data. The generated components can then be exploited for approximately answering OLAP queries such as roll-up, slice and dice operations.

3. OUR APPROACH

Our work is part of approach of the coupling between data mining and online analysis to predict the measured value for non-existent facts or facts with a missing value. The idea is to partition, using methods of clustering, an initial data cube into dense sub-cubes that could serve as a learning set to build a prediction model. The choice to using dense sub-cubes is justified by the quality of the information obtained by these dense sub-cubes, and it will be more interesting to search in the predictive model of sub-cube which contains the cell designated by the user than look through of all data cube.

In this work, we discuss the first part concerning the application of clustering for the partitioning of the initial cube; we first make an experimental study of the clustering methods and then apply the chosen method on our real cube. In the second part, we apply the regression tree method for the construction and validation of the prediction model for each sub-cube;

Finally, we proceed to the prediction of the value of the cell designated by the user through the model of the sub-cube in which is the selected cell.

3.1. Clustering methods

With the increase of the information obtained during the work of information processes, treatment becomes difficult. The need for an initial treatment of the information for its structuring, the isolation of the characteristic features, generalization, sorting appears.

For this purpose, classification and clustering processes are used to perform the required information treatment for later analysis by a specialist. Partitioning observations into groups of similar objects makes it possible to simplify the further treatment of data and decision-making by applying to each cluster its method of analysis. Clustering is the process of grouping similar objects into different groups, or more precisely, the partitioning of a data set into subsets, so that the data in each subset according to some defined distance measure [13].

The starting point of our approach is to conduct experiments on three different algorithms, the first is based on a hierarchical method, we use HAC algorithm, the second is based on the distance, we use the K-means algorithm, and the last is a model-based method, the EM algorithm. These algorithms require to specify the number of clusters as input parameters.

The choice of the clustering method used in our study is based on the evaluation of the quality of the result. Indeed the evaluation of a clustering always contains a part of subjectivity and that it is impossible to define a universal criterion which would allow an unbiased evaluation of all the results produced by all the methods of clustering. However, a number of criteria exist and are used recurrently by many researchers to compare the results obtained. Since there are a large number of possible clustering results for the same dataset, the goal is to evaluate whether one of these results is better than another.

Both algorithms have been implemented on the same dataset to analyse their performances, by taking same number of clusters (3 clusters) and same number of iterations. After implementation of these algorithms, the following results have been obtained (Table 1).

Table 1. Comparative Results of Both Algorithms

Algorithm	Computation time (ms)	Error Ratio
EM	1514	0,21
Kmeans	814	0,34
HAC	980	0,4

In this comparative study found that EM algorithm gives the better performance as compare to K-Means and HAC with minimum error rate. As result of our experiment, the EM algorithm seems to be the most strongest for clustering, it allows the processing of huge databases and offer high accuracy. The EM algorithm is defined as: [14]

Given a statistical model which generates a set X of observed data, a set of unobserved latent data or missing values Z, and a vector of unknown parameters θ , along with a likelihood function (1)

$$L(\theta; X; Z) = p(X, Z \mid \theta) \quad (1)$$

The maximum likelihood estimate (MLE) of the unknown parameters is determined by the marginal likelihood of the observed data (2):

$$L(\theta; X) = p(X|\theta) = \sum_Z p(X, Z \mid \theta) \quad (2)$$

The EM algorithm seeks to find the maximum likelihood estimate (MLE) of the marginal likelihood by iteratively applying the following two steps: Expectation step (E step): Calculate the expected value of the log likelihood function, with respect to the conditional distribution of Z given X under the current estimate of the parameters θ (t):

$$Q\left(\theta \middle| \theta^{(t)}\right) = E_{Z \mid X, \theta^{(t)}} [\log L(\theta; X; Z)] \quad (3)$$

Maximization step (M step): Find the parameter that maximizes this quantity:

$$\theta^{(t+1)} = arg \max_{\theta} Q(\theta \mid \theta^{(t)}) \quad (4)$$

The main issue is therefore to select the correct number of clusters. The choice of the number of clusters has often been studied as a model selection problem. In this case, the algorithm is usually run multiple times independently with a different number of clusters. The results are then compared based on a selection criterion that allows you to choose the best solution.

The probabilistic modeling framework offers tools for selecting the appropriate model complexity. One solution is to rely on information criteria such BIC (Bayesian Information Criterion) [15]. These criteria are generally based on strong statistical bases and apply naturally to probabilistic clustering methods .So we choose to use the model selection criterion BIC to automatically select the number of clusters.

$$\text{BIC} = -2 \ln(L) + \ln(N)k \quad (5)$$

with L is the likelihood function, N is the number of observations, K is the number of clusters to be estimated.

3.2. Regression tree

A classification or regression tree is a prediction model that can be represented as a decision tree. Regression trees are for dependent variables that take continuous or ordered discrete values, with prediction error typically measured by the squared difference between the observed and predicted values.

A regression tree is built in an iterative way, by dividing in each step the population into two or k subsets. The division is carried out according to simple rules on explanatory variables, by determining the optimal rule which makes it possible to construct two or more most differentiated populations in terms of values of the variable to be explained.

The evaluation criteria of a regression tree are the average error rate and the reduction of error. The error rate indicates the average deviation between the observed value and the true value of the variable to predict. If the error rate is close to 0 then this means that the prediction model (the tree) is accurate.

The reduction of error: 1-R2, with R2 the coefficient of determination which measures the proportion of variance explained by the model that is to say the quality of the regression. Among the methods for constructing a regression tree, the two most common techniques are CART [16] and AID [17].

In our case, we use CART to build the regression tree. A decision tree built with the CART algorithm can work with all types of variables: qualitative, ordinal and continuous quantitative. This method makes it possible to create decision rules mixing all types of information.

The general principle of CART is to partition recursively the input space in a binary way, then determine an optimal subset for the prediction. Building a CART tree is done in two steps. A first phase is the construction of a maximal tree, which maximizes the homogeneity of the groups on the dataset, and a second phase, called pruning, which builds a sequence of optimal sub-trees pruned from the maximal tree.

3.3. General notion

We take the definitions proposed in [3] of a data cube.

C is a data cube with:

a non-empty set of d dimensions $\boldsymbol{D} = \{\boldsymbol{D_1}, \boldsymbol{D_i}, \ldots, \boldsymbol{D_d}\}$

and m measurements $M = \{M_1, \ldots, M_q, \ldots, M_m\}$.

$\boldsymbol{H_i}$ is the set of hierarchies of dimension Di.

$\boldsymbol{H_{ij}}$ is the ***jth*** of hierarchical levels of the dimension $\boldsymbol{D_i}$.

$\boldsymbol{A_{ij}}$ represents all terms of the hierarchical level $\boldsymbol{H_{ij}}$ of the dimension $\boldsymbol{D_i}$.

From the data cube ***C***, the user selects an analysis context is a sub-cube of the cube ***C***. To do this we introduce the definition of a data sub-cube.

Let $\boldsymbol{D' \subseteq D}$ a non-empty set of p dimensions $\boldsymbol{\{D_1, \ldots, D_p\}}$ of data cube $\boldsymbol{C\ (p \leq d)}$. The P-tuple $(\boldsymbol{\theta_1, \theta_2, \ldots, \theta_p})$ is a data sub-cube of ***C*** along ***D'*** if $\boldsymbol{\forall i \in \{1, \ldots, p\}, \theta i \neq \emptyset}$ and there exists a unique $\boldsymbol{j \geq 0}$ such as $\boldsymbol{\theta_i \subseteq A_{ij}}$.

A sub-data cube corresponds to a portion of the data cube **C**. A hierarchical level $\boldsymbol{H_{ij}}$ is fixed for each dimension $\boldsymbol{D_i \in D'}$ and a subset $\boldsymbol{\theta_i}$ non-empty terms are selected in this hierarchical level among all the terms $\boldsymbol{A_{ij}}$.

3.4. Interpretation of the prediction model

Our starting point is a data cube with n observed facts according to the quantitative measurement M_q defined by the user in a data cube C. Unlike the approach of [11] and [18], where the user selected a context of analysis, in our approach, the user designates directly the cell c to predict and determines the sub-cube C_i which contains the cell and predict the measurement M_q of the cell through the predictive model built on the sub-cube.

We use EM Alogorithm to partion the data cube into a k dense sub-cubes C_i

$C = \{C_1, C_2, \ldots C_k\}$.

The sub-cube designated, can be considered as region carrying information and can be considered as a training set to build the prediction model.

CART allows the creation of binary tree based on supervised learning methods, the explanatory variables are the dimensions of the sub-cubes and the variable to be predicted represents the corresponding M_q measurement. We build for each subcube C_i in C a regression tree that returns decision rules denoted: $R_i = \{R_i^1, R_i^2, ..R_i^k\}$. Each rule corresponds to a terminal leaf of the tree.

The user directly designates the cell to predict and determines the sub-cube containing the selected cell, and then predicts the measurement of the cell through predictive model built on the cube.

Let c be the empty cell selected by the user.

$M_q(c)$ is the measurement value of cell c.

Let us seek the sub-cube C_i containing cell c.

We are looking for the rule R_i^j derived from the prediction model built from the sub-cube C_i.

(see Algorithm 1)

Algorithm 1:

For $M_q(c) = Null$ do
For each sub-cube $C_i \subset C$
If $c \in C_i$ then
For each $R_i^j \subset R_i$ do
$M_q(c) \leftarrow Y$
End for
End If
End for
End for

To deploy our approach and for the sake of clarification, we use a simple illustrative example of fictitious three dimensional data cube with three D = {Time, Product, Stores}.. The measure corresponds to the number of sales products in the stores. The hierarchy of the stores dimension has 2 levels: Branch and country. In the same way, the Products dimension consists of three levels: product, range and type. In addition, Time dimension is organized following 2 levels: month and year. The data cube consists of 1069 cells which is a detailed representation of the cube with a lower level of granularity for each dimension: (Month, product, Branch).

We use EM algorithm with BIC as a method of clustering to partition our cube. The obtained clusters are shown in Table 2.

Table 2. Number of Facts in Obtained Clusters

	Number of facts (cells)
Sub-cube 1	181
Sub-cube 2	90
Sub-cube 3	390
Sub-cube 4	54
Sub-cube 5	148
Sub-cube 6	23
Sub-cube 7	117
Sub-cube 8	66

In our example, the user designates the cell c, for which it wishes to predict the value of the measure. Cell c belongs to the sub-cube 7, applying the predictive model constructed from the sub-cube 7, we obtain the regression tree in Figure 1 and the following rules:

- ***R1 (***Montreal ∨ Toronto ∨ HongKong ∨ Lyon ∨ Londres ∨ Tokyo) → 78.69)
- ***R2*** ((Mexico ∨ Madrid ∨ Paris2 ∨ Francfort ∨ Marseille ∨ Benelux ∨ Chamonix ∨ Orléans) ∧ (Adapt ∨ Armoire ∨ Balance ∨ Corde ∨ Masq ∨ Pat ∨ Polo ∨ Psavon ∨ Sac ∨ Short ∨ Tabl ∨ Tab ∨ Téléc) → 106.82)
- ***R3*** ((Mexico ∨ Madrid ∨ Paris2 ∨ Francfort ∨ Marseille ∨ Benelux ∨ Chamonix ∨ Orléans) ∧ (Blouson ∨ Chaise ∨ Chaussure ∨ Combin ∨ Four ∨ Jean ∨ Pant ∨ Survet ∨ Sweat bic ∨ Sweat br ∨ Tshirt) → 497.66)

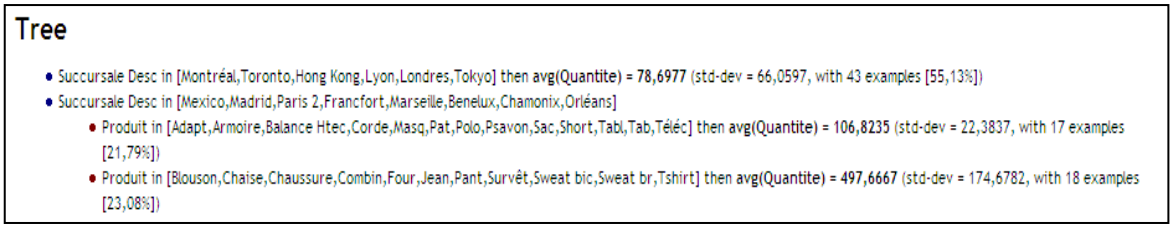

Figure 1. Regression tree obtained

For example, when we targeted the cell described by the terms (Avril2009, Téléc, Paris 2) for dimensions, respectively: Month, products and Branch, R2 was selected. We note that sales measure of products Téléc in Paris 2 will be 106.82 regardless of Month of sale.

For another example, to complete an empty cell of the cube, we want to know what would be the number of sales of the product "Jean" in the Branch "Marseille" for Mars2009 ? The quantity of sales can be predicted by applying rule R3 and the result will be 497.66

4. CASE STUDY AND RESULTS

We test our work on a set of real data. We use for this study the data of the urbanism authorizations service of an urban municipality. The dimensions of analysis of the warehouse analysis are: authorization type (Permit to construct, Permit to demolish, etc. …), subdivision, nature of the project and filing date; ***D***={Authorization Type, subdivision, nature Project, filing Date}. The measure used is the authorization demand's treatment duration (number of days).

The hierarchy of the AuthorizationType dimension and natureProject dimension has 1 level, the subdivision dimension consists of 2 levels: district and subdivision. In addition, filing Date dimension is organized following 2 levels: month and year. The Figure 2 shows the cube.

We select a context of analysis with 16757 cells, which is a detailed representation of the cube with a lower level of granularity for each dimension: (Authorization Type, subdivision, nature Project, month Filing Date). EM algorithm with BIC partition our cube in 10 sub-cubes. The obtained clusters are shown in Figure 3.

The user wants to know the treatment duration for a new filled authorization demand, then the user designates the cell ***c***, for which it wishes to predict the value of the measure. Cell ***c*** belongs to the sub-cube C_4, applying the predictive model constructed from the sub-cube C_4, We use CART as a method of regression tree to build the prediction model with the average error is 0.074. Figure. 4 represent a Regression tree obtained. We obtain the following rules:

- R_4^1(AC → 53)
- R_4^2(Branch ∧ (Juil ∨ Aout) → 14,83
- R_4^3(Branch ∧ (Sept ∨ Oct ∨ Nov ∨ Dec ∨ Jan ∨ Fev ∨ Mars ∨ Avr ∨ Mai ∨ Jui) → 11,01

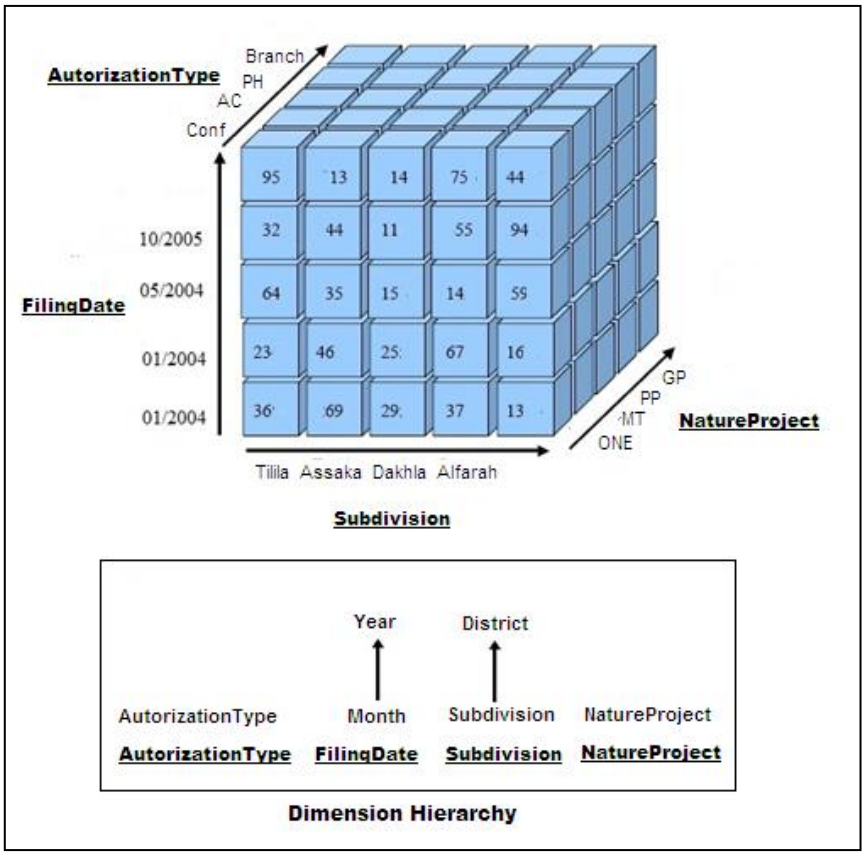

Figure 2. Initial data cube

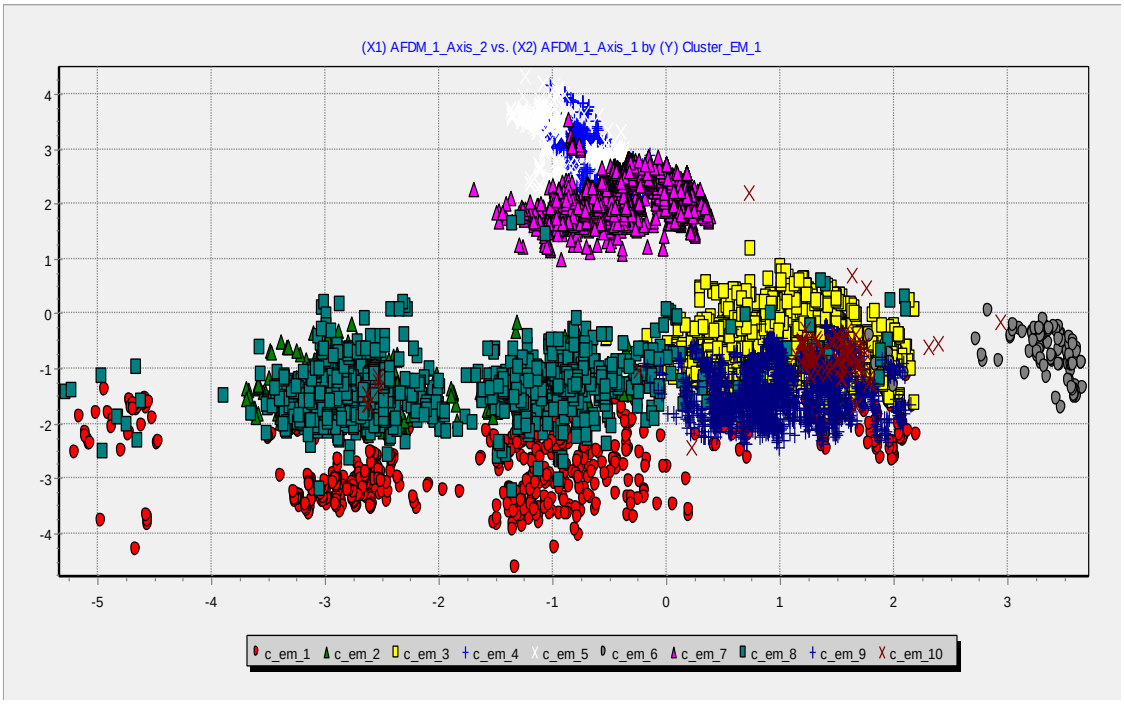

Figure 3. Produced clusters by the EM algorithm

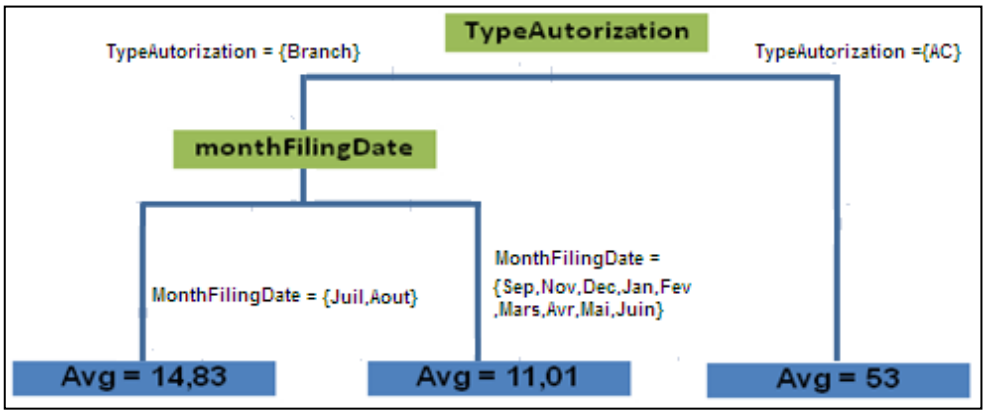

Figure 4. Regression tree obtained from sub-cube C4

Using the set of rules, the empty cell, can be estimated see Table 3. To integrate the predictive module in the OLAP environment, the user designates the cell for which he wishes to predict the value, we designate the rule of the regression tree obtained which correspond to all the terms describing the cell, then we assign the average value of the prediction rule as the measurement value of the cell. This integration of prediction allows the user to understand the expected values of aggregates for a higher level of granularity. The aggregates are recalculated considering the new predicted values.

Table 3. Predicted Values

Type Autorization	Month	Durée de traitement prévue
Branch	Juillet	14,83
AC	Janvier	53

5. CONCLUSION

This research will be used to extract useful and practical knowledge to support the decision-making process from data warehouses containing a huge volume of information, which will enable users in the decision-making systems to achieve a high level of performance by providing them with new elements to understand existing relationships or phenomena in the data and allowing them to anticipate the realization of events according to a number of conditions.

Our first contribution is a synthesis of the various works that have covered the subject of the coupling data mining and online analysis for the prediction, and the work that has treated the subject of clustering and partitioning OLAP cubes. Our second contribution is to offer a new approach for the prediction in OLAP cubes, which provides accurate and understandable results, Our goal is to allow the analyst to predict the value of a measure for a new fact and thus complete the cube using the coupling of online analysis and data mining, And integrate the learning process: apply an unsupervised learning method: Clustering, and a supervised learning method: Regression tree.

REFERENCES

[1] Kimball R., "The Data Warehouse Toolkit," John Wiley & Sons, 1996.
[2] Inmon W. H., "Building the Data Warehouse," John Wiley & Sons, 1996.
[3] Han J., "OLAP Mining: an Integration of OLAP with Data Mining," *Proceedings of the 7th IFIP Conference on Data Semantics, Leysin, Switzerland,* 1997.
[4] G. Sathe and S. Sarawagi, "Intelligent rollups in multidimensional OLAP data," *Proceedings of the 27th International Conference on Very Large Data Bases, Morgan Kaufmann Publishers Inc*., pp. 531-540, 2001.
[5] S. Goil and A. Choudhary, "PARSIMONY: an infrastructure for parallel multidimensional analysis and data mining," *J. Parallel Distrib. Comput.*, vol/issue: 61(3), pp. 285-321, 2001.
[6] R. B. Messaoud, "Couplage de l'analyse en ligne et la fouille de données pour l'exploitation, l'agrégation et l'explication des données complexes," PhD thesis, Université Lumière Lyon 2, Lyon, France, 2006.
[7] Sarawagi S., *et al*., "Discovery-driven Exploration of OLAP Data Cubes," *Proceedings of the 6th International Conference on Extending Database Technology (EDBT'1998), Valencia, Spain: Springer,* pp. 168-182, 1998.
[8] Cheng S., "Statistical Approaches to Predictive Modeling in Large Databases," Master's thesis, Simon Fraser University, British Columbia, Canada, 1998.

[9] J. Han, *et al.*, "Cube explorer: online exploration of data cubes," *SIGMOD "02: Proceedings of the 2002 ACM SIG MOD international conference on Management of data, New York, NY, USA*, pp. 626-626, 2002.
[10] Y. Chen and J. Pei, "Regression cubes with lossless compression and aggregation," *IEEE Transactions on Knowledge and Data Engineering*, vol/issue: 18(12), pp. 1585-1599, 2006.
[11] A. Sair, *et al.*, "Prediction in OLAP Cube," *IJCSI International Journal of Computer Science Issues*, vol/issue: 9(3), 2012.
[12] R. Missaoui, *et al.*, "A Probabilistic Model for Data Cube Compression and Query Approximation," *Proceedings of the Coupling OLAP and data mining for prediction 15 10th ACM International Workshop on Data Warehousing and OLAP (DOLAP'2007), Lisbon, Portugal : ACM Press*, pp. 33-40, 2007.
[13] T. S. Madhulatha, "An overview on clustering methods," Alluri Institute of Management Sciences, Warangal.
[14] A. P. Dempster, *et al.*, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol/issue: 39(1), pp. 1-38, 1977.
[15] E. Wit, *et al.*, "'All models are wrong...': an introduction to model uncertainty," *Statistica Neerlandica*, vol/issue: 66(3), pp. 217-23.
[16] L. Breiman, *et al.*, "Classification and Regression Trees," 1984.
[17] J. N. Morgan and J. A. Sonquist, "Problems in the analysis of survey data, and a proposal," *Journal of the American Statistical Association*, vol/issue: 58(302), pp. 415-434, 1963.
[18] A. B. Niemczuk, *et al.*, "Vers l'intégration de la prédiction dans les cubes OLAP," Laboratoire ERIC, Université Lumière Lyon 2.