PREDICTION IN OLAP CUBE

Abdellah Sair[1], Brahim Erraha[1], Malika Elkyal[1], Sabine Loudcher[2]

[1]Laboratory of Industrial Engineering and Computer Science (LG2I),
National School of Applied Sciences -Agadir,
University Ibn Zohr, Morocco
[2]ERIC laboratory, University of Lyon 2, France

Abstract

Data warehouses are now offering an adequate solution for managing large volumes of data. Online analysis supports OLAP data warehouses in the process of decision support and visualization tools offer, structure and operation of data warehouse. On the other hand, data mining allows the extraction of knowledge with technical description, classification, explanation and prediction. It is therefore possible to better understand the data by coupling on-line analysis with data mining through a unified analysis process. Continuing the work of R. Ben Messaoud, where exploitation of the coupling of on-line analysis and data mining focuses on the description, visualization, classification and explanation, we propose extending the OLAP prediction capabilities. To integrate the prediction in the heart of OLAP, an approach based on automatic learning with regression trees is proposed in order to predict the value of an aggregate or a measure. We will try to express our approach using data from a service management reviews to know that it would be the average obtained by the students if we open a new module, for a department at a certain criterion.
Keywords*: online analysis OLAP, data mining, multidimensional data cube, prediction, regression tree, "What-If Analysis".*

1. Introduction

Data warehouses provide adequate solutions to the problem of managing large volumes of data. Indeed a data warehouse is a centralized structure in which is stored a large amount of historical data, organized by topic and consolidated from various sources of information [18].

Specific (multidimensional) structuring, such as star schema or snowflake schema, are designed to make the data warehouse ready for analysis. After that it is the role of OLAP user to browse, explore and analyze data to extract potential knowledge for decision making [17].

However, there is no approach to guide the user in deepening his analysis to the explanation and prediction. The OLAP needs new and innovative automated tools to extract potentially existing knowledge within the data cube.

In [1], this observation has motivated an extension of the OLAP capabilities for visualization, classification and explanation. Coupling methods for data mining with OLAP is an approach that has already proven itself.

[2, 3, 4] address the coupling of the two areas and the problem of multidimensional data mining for navigational aid for further analysis and relationships in data. Early attempts went back to 1997 with the work of Han [5] who introduced the On-Line Analytical Mining terminology (OLAM) and with the creation of DBMiner. However, the references related to these works describe more the functional side and lacks detail on the formalisms and techniques used to deploy data mining methods.

Another approach to the prediction in the OLAP is to move in the context of "What-If Analysis". In decision making, after consultation made in a cube, the user can want to anticipate the realization of future events. This prediction can be placed into the context of the "What-If Analysis "as defined by Golfarelli et al. [6], where the process of projection into the future shows a user-centered approach.

2. Related work

After outlining the various approaches treating the coupling between data mining and online analysis to extend OLAP to prediction, we define criteria for comparing existing proposals and position ourselves among them. We introduce seven criteria including two oriented data mining and the rest to the pairing process. The five criteria of the binding process are:

- ***Multidimensional structure of data***

The inclusion or not of multidimensional modeling allows us to appreciate the rigor of coupling data Mining and OLAP.

- ***Hierarchies of data cubes***

This second criterion looks at whether the hierarchies of the cubes are taken into account when developing the prediction model. For example: recalculation of the model or not when using an OLAP operator as the Roll-Up or Drill-Down.

• ***Optimization of algorithms***
This criterion focuses on the optimization of data mining algorithms• ***Parameterization of the model:***
It seems necessary to give the user the ability to set and to parameter the learning and the data mining method to improve the accuracy of prediction.
• ***Processing of results***
The exploitation of results and more specifically the prediction model is a significant element of a successful integration of data mining in online analysis. And to integrate the prediction in the
OLAP, we need to propose a model, interpret it and associate it with the semantic OLAP. It provides the user new tools while remaining in his analysis environment that he knows.

In a second step, we defined two criteria internal to the conventional process of data mining:
• ***Selection of a subset of the cube***
This process includes a step of selecting predictive (explanatory) attributes, and a step made of splitting facts into learning sample to build the model and in test sample to evaluate the built model.
• ***Validation of the model***
Once the model is built, it must be evaluated and validated. Many techniques exist to validate a model such as cross-validation. The table 1 summarizes the value of criteria for each approach. The proposals are distinguished by:

- (i) The entity predicted,
- (ii) The purpose and results of the method,
- (iii) The learning process.

Table. 1: Comparison of proposed integration of the prediction in OLAP cubes

Proposition	Method of prediction	The entity predicted		Results					Learning process					
		Existing measure	**New measure**	**New cube**	**Completed cube**	**New fact**	**Indicators**	**OLAP Exploitation**	**Hypothesis of the model**	**Algorithmic Optimization**	**Calculation multi-level**	**Data Preparation**	**Subset of the cube**	**Model validation**
Sarawagi et al [2]	log-linear modeling	●		●						●	●			
Cheng [9]	Generalized Linear Model	●			●	●			●	○	●	●	●	
Palpanas et al [10]	Maximum Entropy Information	●		●				●		●			●	○
Chen BC et al [3] 2005	Distributively algebraically decomposable	●		●				●		●	●		●	●
Chen BC et al [19] 2006	Linear regression	●			●	●			●	●		●	●	●
Y.Chen et Pie [12]	Linear regression		●	●					●	●	●	●		
our approach	Regression tree	●			●	●	●	●		○	●	●	●	●

Legend	● : this Criterion	○ : Criterion considered partially or remaining fuzzy

Some works have for objective prediction of a new cube. They propose to generate a new cube using fairly complex models. Thus Cheng [9] uses the generalized linear model to generate a new cube, while Sarawagi et al [2] use a new cube of predicted values to indicate to the user cells with exceptional value or outliers.

While BC Chen et al [3] use a model where the measure indicates a score or a probability distribution associated with the measure value that can be expected in the original cube. Subsequently BC Chen et al [19] uses the prediction model to predict the measure of a new fact.

Finally in Y. Chen and Pei [12], a cubic measure is generated where each value indicates the weight of evidence.

The results provided to the user are often complex to be relevant. Indeed, the user will have the usual difficulties to find areas in the cube that present trends of the most reliable data because he does not have the required skills.

To remedy this problem, we propose a solution that allows the user to predict the value of a measure made according to a new context-defined analysis. We provide the user accurate and understandable results and appropriate indicators for evaluating the quality of the predicted obtained values. We integrate these predictions with existing data in the original cube and we present to the user a completed cube by offering a prediction for some empty cells.

Moreover, our proposal is to predict and not to analysis trends in data.

It should be noted that most of the proposed approaches exploit their results in the philosophy of the OLAP environment; this presents a significant element of a successful integration of data mining in online analysis. We must propose a model, interpret and associate it with the OLAP semantic.

For the learning process, it should be noted that all approaches have considered the criterion of hierarchies of cubes as well when developing the prediction model that when the operating results in the OLAP environment.

In addition we believe that it is important not to have assumptions and constraints on the model because the user does not have all the skills necessary to optimize the prediction accuracy by setting the model development. We propose a method without constraint; data cubes are made along several lines of analysis and dimensions are qualitative variables and the facts are usually measured by continuous quantitative variables. A regression tree meets these characteristics and does not require assumptions about the data.

Our focus is also on optimizations in the algorithms of search. These allow proposing a model for each hierarchical level of a cube. Often, during construction of a predictive model, the data need to be prepared, sampled and selected explanatory variables to use. Once a search method deployed, it must be assessed and validated.

Many techniques exist to perform these steps. It is therefore necessary to incorporate the coupling process, in order to base the models produced on solid foundations.

The proposals of Chen et al [3.19] and those of Sarawagi et al [2], are set to identify subsets interesting in light of a predictive model. Their effort is upstream of the construction of prediction model and consists of searching the data set most relevant to learning in the new fact that the user wants to predict.

We also note that all proposed work is more focused on the data set used for learning than on model validation, where only the work of Chen et al [3.19] investigates the subject. At their first proposal, the model was validated with a sample test and an evaluation function determined by the user. In Chen et al [19], they use cross-validation to evaluate and validate their model.

In our proposal, we include a comprehensive process with a selection phase of the explanatory variables, sharing a stage made for training sample build the model and a test phase to validate the model built.

We want to preserve the philosophy of on-line analysis as Sarawagi et al. [SAM98] propose the prediction when incorporating into a cube.

In our approach, to run the model prediction, the user should not need extensive knowledge on the use of a regression tree that makes it possible, by discrimination of the explanatory variables, to offer a prediction for empty cells. It offers understandable results which are not related to a black box for the user. It provides, by the same token a model to explain existing facts as based on discriminated variables.

3. Our approach

3.1 Objectives

This is to propose a new approach for the prediction of a measurement value of new facts in a data cube by coupling a supervised learning method, regression trees, with online analysis.

Indeed, it is important to associate the semantics of OLAP to data mining method to preserve the philosophy of on-line analysis as Sarawagi et al [2] proposes. To run the model, the user does not need extensive knowledge of regression trees.

Thus our work differ from those of BC Chen et al [3] where the user does not have an available a model to explore as a cube, but the results incorporated into the original cube.

In addition, we hope that our approach provides accurate results and indicators suitable for the user to measure the quality of the predicted values obtained by indicating the degree of validity of a prediction.

Our approach integrates a complete learning process with a data preparation phase, a phase selection of explanatory variables, a validation phase, phase not investigated in depth in previous work.

Our proposal is placed under the "What-If Analysis" as defined by Golfari et al. [6] However; the notion of query "What-If" is distinguished by the works of Imielinski et al [4] and their recovery by Han et al [11].

Finally we want to make the prediction, not the analysis of trends in the data.

3.2 Proposed Approach

To deploy our approach and for the sake of clarification, we use a simple illustrative example of fictitious three dimensional data cube with three: Years (2009, 2010, 2011), Products (E-phone, Mobiles, Camera) and Stores (Store1, Store2, Store3, Store4, Store5). The measure corresponds to the turnover of sales (in Million MAD) products in the stores. The data cube consists of 45 cells (multiplication of the cardinalities of dimensions). It is considered that, on 45 cube cells, nine cells are empty and their value is to predict (see Fig.1 (a)). We return to Figures 1 (b) and (c) when implementing our method.

3.2.1 General Notations

We adopt the definitions of a cube and sub cube of proposed data in [18] and complement to our needs.
C is a data cube with a non-empty set of dimensions $D = \{D_1, D_{i,\ldots}, D_d\}$ and m measurements $M = \{M_1, \ldots, M_q, \ldots, M_m\}$. H_i is the set of hierarchies of dimension D_i. H^i_j is the $j^{ème}$ of hierarchical levels of the dimension D_i. For example, the Products dimension (D_1) contains two hierarchical levels: product code is noted H^1_1 and the level of aggregation of all the products corresponding to the hierarchical level 0 is noted H^1_0.
A^{ij} represents all terms of the hierarchical level H^i_j of the dimension D_i. Code-Year - level (H^1_1) of the Years dimension (D_1) contains three terms: 2009, rated A^{11}_1, 2010, and 2011 marked A^{11}_2, A^{11}_3 rated.
Generally, a cube can represent a set of facts, presenting the values taken by a measure M_q based on all Modalities A^{ij} terms of dimensions $\{D_1, \ldots, D_i, \ldots, D_d\}$ which made to characterize a given level of aggregation H^i_j.

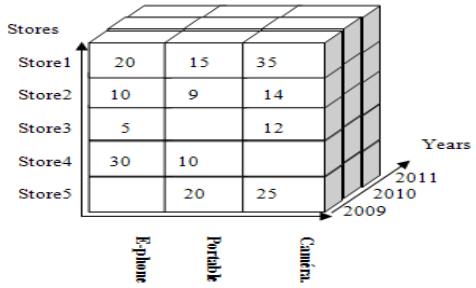

Fig.1 (a)

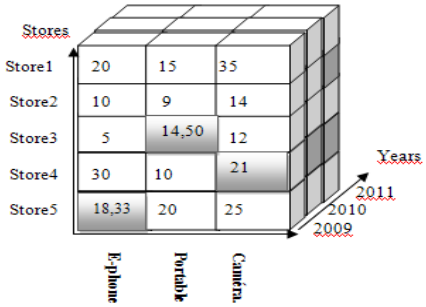

Fig.1 (b)

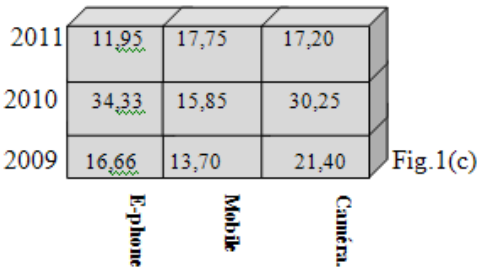

Fig.1(c)

Fig .1: Predicted values within a data cube

From the data cube C, the user selects an analysis context is a sub-cube of the cube C. Let $D' \subset D$, a not empty subset of p dimensions $\{D_1. . ., D_p\}$ of data cube C ($p \leq d$). The p-tuple $(\theta_1, ..., \theta_p)$ is a sub-data cube $\forall i \in \{1,. . ., P\}$, $\theta_i \neq \emptyset$, and there is a single index $j \geq 0$ such that $\theta_i \subseteq A^{ij}$.

A sub-data cube corresponds to a portion of the data cube C. A hierarchical level H^i_j is fixed for each size used $D_i \in D'$ and a subset θ_i non empty terms are selected in this hierarchical level among all the terms A^{ij}.

In the context of analysis $(\theta_1, ..., \theta_p)$, there are n observed facts according to the quantitative measurement M_q defined by the user in a data cube C.

In our illustrative example (see Fig. 1 (a)):

- The context analysis $(\theta_1, \theta_2, \theta_3)$ = (Products, Stores, Years) = ({E-phone, Mobile, Camera}, {Store1, Store2, Store3, Store4, Store5}, {2009, 2010, 2011}) is: (three dimensions)
- Measurement M_q corresponds to turnover of sales, M_q is the variable to predict
- Products, Stores and Years play the role of explanatory variables

3.2.2 Different approaches to the regression tree

Different types of regression trees are proposed in the literature. One of the first approaches is AID (Automatic Interaction Detection) [13]. This approach was taken in [14] where the algorithm CHAID (Chi-Squared Automatic Interaction Detection) is proposed. Breiman et al. [15], offer binary trees with CART (Classification and Regression Tree). Recently, other types of trees have emerged, including Arbogodaï of Zighed et al. [16].

Breiman et al [15] proposed a binary regression tree, called CART, predicting both qualitative and quantitative variables as predictors of qualitative, quantitative or both.

CART is based on the principle of recursive partitioning. At each step, the discriminated explanatory variables are segmented into two new groups of terms or two intervals. When the variable to predict is continuous quantitative prediction obtained is the average of observations belonging to the group or to the interval (leaf of the tree).

The method of a binary tree is therefore to divide the sample into two sub learning thanks to one of the explanatory variables. The operation is repeated separately in each sub-assembly thus formed. The homogeneity of the two groups or intervals is optimized by partitioning criteria. In the case of a continuous quantitative variable to predict the variance of the amalgamation or the interval is used as a measure of homogeneity. At the time of division into two subgroups then we try to minimize intra-group variance and maximize inter-group variance. The quality of the regression can be assessed using standard measures such as squared error.

Learning is implemented in two phases: a first phase, called "expanding", maximizes the homogeneity of the groups on the data set called "growing set". The second phase involves "pruning" of the tree is to minimize the prediction error on another data set, called "pruning set". To determine the number of terminal nodes with the CART algorithm, it therefore let's grow the tree with the stopping criterion a minimum number per node. Then, the pruning of the tree is done using the sample data "pruning set", which allows for a sub tree minimizing the better the prediction error.

3.2.3 Construction and validation of the model prediction

To build a regression tree analysis of the context $(\theta_1, ..., \theta_p)$, we segment it into two bases of random facts: 70% of the facts used for learning and building the model and 30% are reserved to evaluate the resulting model.

Conventionally, the evaluation criteria of a regression tree are the average error rate and reducing the error. The error rate indicates the average deviation between the observed value and the true value of the variable to predict. Over the average of the error approaches 0, the most predictive model is accurate. For our illustrative example, the average error is 0.259 which is acceptable. The reduction of an error $1-R^2$ (with R^2 the coefficient of determination which measures the proportion of variance explained by the model that is to say the quality of the regression) indicates whether the tree predicts better than the default template (the tree reduced to its root) would be used only where the average of the measure to predict the values of the measure. . The prediction

is perfect if this flag is 0 ($R^2 = 1$) when the R^2 is zero or negative, it means that the tree does no better than a tree consisting only of its root, the prediction is then the average of the variable to predict the entire sample.

3.2.4 Interpretation of the prediction model

After building the model, the regression tree decision rules returns λ (λ ≥ 0). All the rules of a model are denoted by R = { $\Re_1$, $\Re_2$, ..., $\Re_\lambda$}.

Definition (Decision Rule): let $\Re$ (X $\Rightarrow$ Y; S; σ) a decision rule ϵ R.

X is a conjunction and / or a disjunction of terms $\subset$ (θ_1, ..., θ_p) and corresponds to the history of the rule. Y is the average value predicted for measuring M_q given X. S is the support of the rule and σ is the standard deviation of M_q, in checking the training set X.

In addition to the two indicators of reliability of the model (average error rate and error reduction), two criteria for assessing the quality of a rule. The first is the relative size S of the facts that support the rule. The second is the standard deviation σ of Mq, which indicates the homogeneity of the facts supporting the rule. More standard deviation σ, the higher the group of facts supporting the rule is heterogeneous.

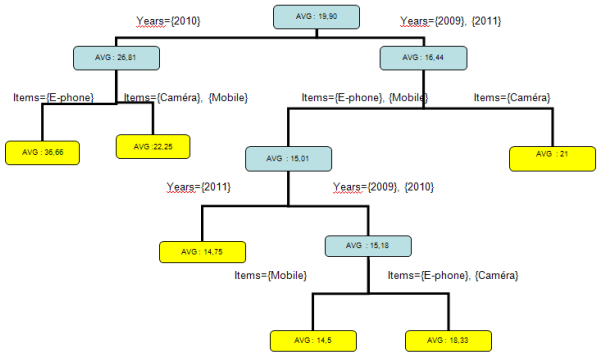

Fig. 2: Regression tree obtained from the analysis context.

In our example, we obtain the regression tree in Figure 2 and the rules following:

•R1 (2010 ∧ (Camera ∨ Mobile) $\Rightarrow$ 22.25; 21.05%;15.17)

•R2 (2010 ∧ E-phone $\Rightarrow$ 36.66; 15.79%; 22.54)

•R3 ((2009 ∨ 2011) ∧ Camera $\Rightarrow$ 21.00; 15.75%; 12.12)

•R4 ((2009 ∨ 2011) ∧ ((E-phone ∨ Mobile) ∧2011) $\Rightarrow$ 14.75; 21.05%; 7.32)

•R5 ((2009 ∨ 2011) ∧ (E-phone ∨ Mobile) ∧ ((2009 ∨ 2010) ∧ (E-phone ∨ camera)) $\Rightarrow$ 18.33; 15.79% ; 12.58)

• R6 ((2009 ∨ 2011) ∧ ((E-phone ∨ Mobile) ∧ ((2009 ∨ 2010) ∧ Mobile)) $\Rightarrow$ 14.5; 10.53%; 7.78)

Each rule corresponds to a terminal leaf of the tree. For example, the rule R1 indicates that if the products camera or mobile are purchased in 2010 so the sales turnover of these products will be 22.25. 21.05% of sales in the file learning fall into this category and the standard deviation is 15.17. The products and the years are the most discriminates variables. They are explanatory of sales results, unlike stores that are not determinants.

3.2.5 Operation of the predictive model in the OLAP environment

- Let (θ1, ..., θp) analysis of the context be defined by the user, indicating all dimensions and conditions.
- Let R= { $\Re_1$, $\Re_2$, ..., $\Re_\lambda$}.the set of prediction rules obtained.

The user designates the cell c = ($\theta 1$,. . . , θp) from the context analysis (θ1, ..., θp), for which it wishes to predict the value of the measure. Each is a singleton containing a single modality for the dimension to which it is attached. There Mq (c) the value of the measure that takes M_q cell c. For each cell c designated by the user, such as M_q (c) = null, i.e. the cell is empty, we search for the rule $\Re \subset$ R as its antecedent X has all its terms included in the all terms describing the cell c.

It is therefore essential to compare all of the terms describing the cell with backgrounds X rules regression tree

For a rule we only look conjunctions X 'of this agreement. If X' $\subset$ ($\theta 1$,. . . , θp) then the average value of the Y prediction rule can be assigned as the value of the measurement of the cell.

We note Mq (c)←Y. The operation is repeated for each cell designated by the user for prediction. (See Algorithm 1)

Algorithm 1: Integration of Prediction in a data cube (R; (θ_1, ..., θ_p)):
1: for each $\Re$ in R do
2: for each cell c do
3: If Mq (c) is empty then
4: Mq (c)←Y
5: end if
6: end for
7: end for

For example, when we targeted the cell described by the terms (2010, E-phone, Store2) for dimensions, respectively: Years, products and stores, R2 (2010 $\wedge$ E-phone $\Rightarrow$ 36.66; 15.79%; 22.54) was selected (2010 $\wedge$ E-phone) $\subset$ (2010, E-phone, Store2). We note that sales of products E-phone in years like 2010 will Turnover 36.66 if the sale is made in the store Store2. For another example, in terms of query such as "What-If" regression tree allows us to know it would be the revenue if we sell the product "camera" in the store "Store4" for 2009?
This integration of the prediction also allows the user to understand the expected values of aggregates for a higher level. Aggregates are recalculated considering the new predicted values. For example by choosing the All for the stores, the sales turnover may be calculated by year and the product (see. Tab.2). Thus, the average annualized revenue expected for all stores, the products "E-phone" in 2010 is 11.62 Million MAD.

Tab.2

	YEAR		
Product Code	2009	2010	2011
E-phone	16,66	11.62	11,95
Mobile	13,70	11.92	17,75
Camera	21,40	13.80	17,20

3.2.6 Visualization of the prediction model in OLAP

A proposed extension for enhancing the predictive model in data cubes consists in using visual indicators to the user. In Figures 1 (b) and 1 (c), we use a shade of gray to a predicted value or an aggregate recalculated from the predicted values. We believe that according to quality criteria of a rule (actual and standard deviation), we can modify the color code. Thus the user can directly interpret the predictions in the data cube.

4. A Case Study

In this section we will test our work on a real data set. We use the data for this study of an exam-management service at our school. 1200 facts are present in the data cube. The cornerstones of analysis of the warehouse are in the sector (Option), the type of degree, the reference, the delegation of bachelor's degree, taught module, sex and mode of access to school. The measure used is the average score in the evaluation of a module.

4.1 Background Analysis

Our analysis context is defined as follows. For dimensions, we use the sector, access mode, sex, type of degree, a statement, rather than bachelor's degree and eventually followed the modules. We propose an analytical representation of the context in the form of Star schema in Figure 3:

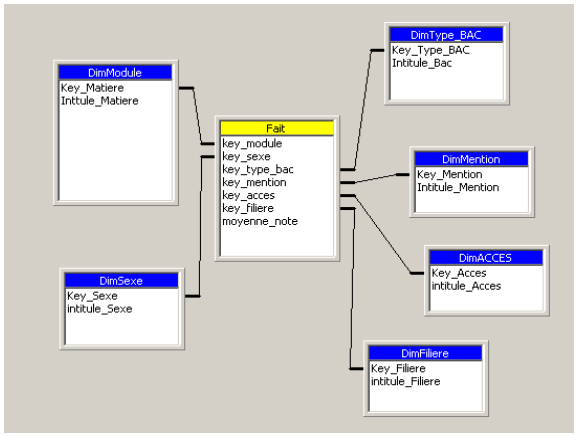

Fig. 3: Representation in the form of star schema of the context analysis.

Thus, in this context of analysis, a user can answer various questions such as what would be the average score of a student enrolled in the department GINFO4, with a Bachelor series' SM' with the grade 'AB', he passed the examination of a new module already taught in other courses.

4.2 Results

We choose to CART Breiman et al [15] as a method of regression tree to build the prediction model in the context of analysis previously defined and used in the software Tanagra.

4.2.1 Software Tanagra

TANAGRA is free software for teaching and research dedicated to data mining. It includes a set of data mining methods from the field of statistical exploratory data analysis, automatic learning.

TANAGRA project aims to provide students and researchers with a platform for data mining easy to use and to conduct studies on real data.

4.2.2 Results Obtained

The regression tree thus constructed has 27 vertices and 14 sheets, the average error of the tree is 0.0469 and the reduction in error is 0.8951. The model is applicable.

The discriminates variables (dimensions) are in the order they appear in the tree: the type of Baccalaureate, access mode, Baccalaureate grade, sex and option.

The tree obtained is shown in Figure 4. Of the 14 rules we present those we obtained for predicting the values of 6 cells describing the average score for students in predicting modules not taught in their option:

- R1((Sexp ∨ C) ∧ (Licence ∨ DUT ∨ Maîtrise) ∧ AB ∧ M⇒12,78; 12,31%; 1,9)
- R2((Sexp ∨ C) ∧ (Licence ∨ DUT ∨ Maîtrise) ∧ AB ∧ F⇒14,05; 1,971%; 1,8)
- R3((Sexp ∨ C) ∧ (Licence ∨ DUT ∨ Maîtrise) ∧ (B ∨ TB ∨ P) ∧ F⇒12,62; 2,67%; 1,77)
- R4((Sexp ∨ C) ∧ (Licence ∨ DUT ∨ Maîtrise) ∧ (B ∨ TB ∨ P) ∧ M ⇒ 13,76; 11,85%; 1,71)
- R5((EL ∨ F1) ∧ M ⇒ 12,11; 7,90%; 2,56)
- R6 ((EL ∨ SM ∨ FM) ∧ (Licence ∨ Maîtrise) ∧ (AB ∨ TB) ∧ F ⇒ 14,23; 5,46%; 2,23)

We note for rules 2 and 3, staff (expressed in frequency) supporting the weak rules, 1.97% for the second rule and 2.67% for the third rule. The predictions are to be taken with caution. We find in the table (see Tab.3) the integration of results from regression tree for six cells. For example, Rule 4 is used for the second row of the table. The condition is verified: if a male student with a bachelor type { Sexp or C}, with honors {B, TB or P}, has joined the school with the access mode {License, Diploma or Master} then its average rating is 13.76.

Similarly, using the set of rules, all empty cells described by the terms Baccalaureate type, access mode, Legal and Gender, can be estimated (see Tab.3).

Tab.3

Type bac	**Mode Access**	**Grade**	**Sex**	**Average mark**
EL	Maîtrise	TB	F	14,23
C	DUT	B	M	13,76
F1	Licence	AB	F	12,11
Sexp	DUT	P	F	12,62
C	Licence	AB	F	14,05
Sexp	Maîtrise	AB	M	12,78

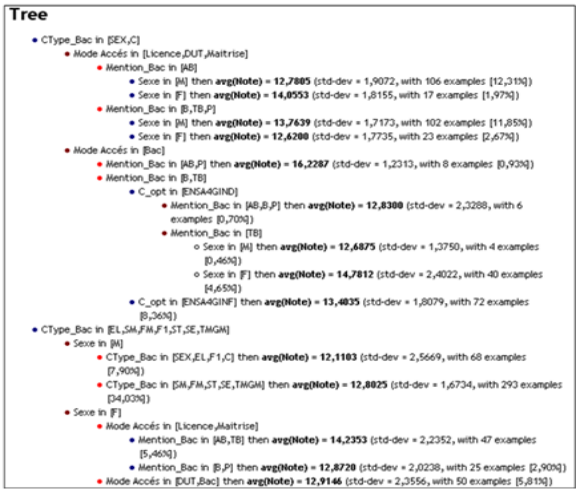

Fig. 4: Regression tree obtained with CART

Each rule corresponds to a terminal leaf of the tree. For example, the rule R1 indicates that if a male student earning a Baccalaureate {C or Sexp} with high grade {TB}, has joined the school with the access mode {License, DUT or Master} then its average rating is 12.78. 12.31% of students present in the training file in this category and the standard deviation of their average is 1.9. The type of degree,

access mode, the reference and sex are the most discriminates variables. They are explanatory of the results of the average score tracking module, unlike modules which are not determinative.

To exploit the predictive module in OLAP environment, we designate the cell for which we wish to predict the value of their position, we look after one of the rules of the regression tree obtained which match all the terms describing the cell c then we assign the average value of the prediction rule as a measurement value of the cell. The operation is repeated for each cell designated for prediction.

This integration of the prediction also allows the user to understand the expected values of aggregates for a higher level. Aggregates are recalculated considering the new predicted values. For example by choosing the all levels for the modules, the average score can be calculated by sex and type of access to school baccalaureate.

	Sex	
type of access	M	F
Maîtrise	13,02	11.62
Licence	13,02	13.55
DUT	12,89	13.29
Baccalauréat	12,94	13.53

5. Conclusion and perspectives

In decision making, after consultation with the user made in a cube can try to anticipate the realization of future events which can assist the user in this task by placing themselves under the "What-If Analysis "user-centered task. Thus, we extend the capabilities of the online analysis by integrating at the heart of OLAP process a prediction technique with regression trees, we propose the analyst to place himself in a predictive approach and through discrimination variables in an explanatory approach.

For this, we want to offer the user an approach that can predict the value of a measurement made according to a new context-defined analysis, which provides accurate results and understandable as well as quality indicators of predicted values.

Our first contribution is a synthesis of various studies that have addressed the subject of the coupling between data mining and analysis online to extend the OLAP prediction. We found that there is methodological work having an orientation OLAP and work more oriented data mining. We believe that both approaches should meet to provide the user with new tools adapted to their needs and philosophy all OLAP exploiting the strengths of data mining.
Our second contribution is to predict the measurement value of new facts using regression tree as a prediction technique and take into account the predictions made in the navigation follow up (higher aggregates recalculated). We propose a formalization of our approach and we illustrate our approach through a simple example. A case study on a real data set demonstrates the feasibility and value of our proposal by providing the user indicators of reliability of decision rules and overall regression tree. We suggest an extension to visual parameters to the user indicating the predicted values of new aggregates, cell values can be set to a higher aggregation level and quality of each of these predictions in the data cube. Thus we have exploited the coupling of on-line analysis and data mining in order to extend the capabilities of the OLAP prediction.

Our work opens several research opportunities. First we want to put our prediction operator an indicator of reliability / quality of the prediction for the case where the tree prediction does not give a more accurate prediction more than the overall average of the variable to predict the sample learning. We also wish to go further in formalizing our operator about its operations in OLAP. So we want to return to the case where the user wants to explore a finer level of aggregation in the light of predictions made at a higher level, in order to take fully into account the concept of hierarchical levels within the OLAP by answering the question should we recalculate the model at each hierarchical level?

References

[1] R. Ben Messaoud. Couplage de l'analyse en ligne et de la fouille de données pour l'exploration, l'agrégation et l'explication des données complexes. PhD thesis, Université Lumière Lyon 2, Lyon, France, Novembre 2006.
[2] S. Sarawagi, R. Agrawal, and N. Megiddo. Discovery-driven Exploration of OLAP Data Cubes. In Proceedings of the 6th International
[3] B.-C. Chen, L. Chen, Y. Lin, and R. Ramakrishnan. Prediction Cubes. In Proceedings of the 31st International Conference on Very Large Data Bases (VLDB'05), pages 982–993, Trondheim,Norway, August - September 2005.ACM Press.
[4] T. Imielinski, L. Khachiyan, and A. Abdulghani. Cubegrades: Generalizing association rules. Tech. Rep., Dept.Computer Science, Rutgers Univ., Aug., 2000.
[5] J. Han. OLAP Mining: An Integration of OLAP with Data Mining. In Proceedings of the 7th IFIP Conference on Data Semantics, Leysin, Switzerland, October 1997.
[6] M. Golfarelli, S. Rizzi, and A. Proli. Designing

what-if analysis: towards a methodology. In Proceedings 9th International Workshop on Data Warehousing and OLAP (DOLAP 2006), pages 51–58, Arlington, USA, 2006

[7] Qiang Yang, Joshua Zhexue Huang, and Michael Ng. A Data Cube Model for Prediction-Based Web Prefetching. Journal of Intelligent Information Systems, 20(1) :11_30, 2003

[8] Joshua Zhexue Huang, Michael Ng, Wai-Ki Ching, Joe Ng, and David Cheung. A Cube Model and Cluster Analysis forWeb Access Sessions. In Revised Papers from the 3rd International Workshop on Mining Web Log Data Across All Customers Touch Points (WEBKDD '01), pages 48_67, San Francisco, CA, USA, August 2002. Springer-Verlag.

[9] Shan cheng. Statistical Approches to Prodictive Modelling in large Databases. Master's thesis, Simon Fraser University, British Columbia, Canada, February 1998.

[10] T. Palpanas, N. Koudas, and A.Mendelzon. Using Data cube Aggregates for Approximate Querying and Deviation Detection. IEEE Transactions on Knowledge and Data Engineering,17(11):1465–1477, November 2005

[11] J. Han, J. Wang, G. Dong, J. Pei, and K. Wang. Cube explorer: online exploration of data cubes. In SIGMOD '02: Proceedings of the 2002 ACM SIG MOD international conference on Management of data, pages 626–626, New York, NY, USA, 2002. ACM

[12] Y. Chen and J. Pei. Regression cubes with lossless compression and aggregation. IEEE Transactions on Knowledge and Data Engineering, 18(12):1585–1599, 2006. Senior Member-Guozhu Dong and Senior Member-Jiawei Han and Fellow-Benjamin W. Wah and Member- Jianyong Wang.

[13] J. N. Morgan and J. A. Sonquist. Problems in the analysis of survey data, and a proposal. Journal of the American Statistical Association, 58(302) :415_434, 1963.

[14] G. V. Kass. An exploratory technique for investigating large quantities of categorical data. Applied Statistics, 29(2):119_127, 1980.

[15] Leo Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. Classification and Regression Trees. 1984

[16] Djamel A. Zighed, Gilbert Ritschard, Walid Erray, and Vasil M. Scuturici.Abogodaï, a new approach for decision trees. In 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 03),Dubrovnik, Croatia, volume 2838 of LNAI, pages 495_506, Heidelberg, Germany, September 2003. Springer.

[17] Kimball R., The Data Warehouse Toolkit, John Wiley & Sons. 1996.

[18] Inmon W.H., Building the Data Warehouse, John Wiley & Sons. 1996.

[19] Bee-Chung Chen, Raghu Ramakrishnan, Jude W. Shavlik, and Pradeep Tamma. Bellwether Analysis: Predicting Global Aggregates from Local Regions. In Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB'06), pages 655_666, Seoul, Korea, September 2006. ACM Press.

Biography of Authors

Abdellah SAIR, PHD Student 'integration of the prediction in Cube OLAP' at the National School of Applied Sciences of Agadir Morocco in collaboration with the ERIC laboratory of university Lyon 2 France, I have a diploma of Superior Studies Specialized in Business Intelligence at the National School of Applied Sciences of Agadir Morocco and I am professor specialty computer at the office of vocational training and promotion of labor. Areas of research are coupling on-line analysis with data mining through a unified analysis in the process of decision support to help Moroccans Universities Systems and application the operator prediction in the heart of the cube's data university environment.

Erraha BRAHIM (PHD), Ability Professor in Computer Science at the National School of Applied Sciences of Agadir And team member of the Laboratory of Industrial Engineering and Computer Science (LG2I), National School of Applied Sciences of Agadir, University Ibn Zohr Morocco.

Malika Elkyal (PHD), Ability Professor in Applied Mathematics at the National School of Applied Sciences of Agadir And team member of the Laboratory of Industrial Engineering and Computer Science (LG2I), National School of Applied Sciences of Agadir, University Ibn Zohr Morocco.

Sabine Loudcher (PHD), Ability Professor in Computer Science at the Department of Statistics and Computer Science of the University of Lyon 2, France. Since 2000, she has been a member of the Decision Support Databases research group within the ERIC laboratory.