Prediction Operator By Regression Tree

Abdellah Sair[1], Brahim Erraha[1], Malika Elkyal[1]
[1]Laboratory of Industrial Engineering and Computer Science (LG2I),
National School of Applied Sciences - University Ibn Zohr
Agadir, Morocco

Sabine Loudcher[2]
[2]ERIC laboratory, University of Lyon 2
Lyon, France

***Abstract*: The objective of this paper is to extend the capacities of the online analysis, by associating the OLAP with data mining. Indeed the online analysis OLAP supports data warehouse in the process of decision support and offers tools for visualization, structuring and exploitation the data of the data warehouse of whereas the 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. To integrate the prediction in the middle 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 out our approach by exploiting data from a credit department of a bank to know that it would be the rate of solvency of a customer if he asks a credit for a new product launched in his city according to a certain criterion.**

Keywords: online analysis OLAP, data mining, multidimensional data cube, prediction, regression tree, "What-If Analysis".

I. Introduction

Business intelligence is designed to collect, organize, store and analyze information to help decision making. Into this context, Inmon introduces the data warehouses [13].

A warehouse is a collection of data, subject-oriented, integrated, not nonvolatile and historized, organized as support of the process of the decision aid. The data are extracted, cleaned, transformed into a single format which prepares them for analysis.

Conceptually, the data are modeled in a multidimensional with indicators to observe (measures) and analysis axes (dimensions). At the logical level, particular structuring Multidimensional were designed, such that the star schema, snowflake or constellation, in order 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 [13].

However, 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, associations existing in the multidimensional data, to predict values in the data cube. Moreover, one of the rules founders specifies that the OLAP must extract and manage the missing values in the multidimensional representation of the cube. No standard operator OLAP allows making this management.

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.

Thus, these various observations allow us to say that it is necessary to make evolve the OLAP to other possibilities of analysis. Among all the possible extensions, we choose to tackle the problem of the research of the prediction in a cube. Our idea is to combine the OLAP with the data mining and to thus extend the possibilities of the OLAP.

Indeed, in the context of data mining, the user can ask questions less precise ("what is the turnover of sales of Clothes for next year in Morocco?") and the user needs less knowledge about the field. The methods of data mining are numerous and aim different objectives (description, structuring, explanation, prediction,). They automatically seek useful potentially information to answer the questions of the user.

[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.

To predict the value of a missing measurement in a cell, we turned to regression trees. This prediction method does not assume any assumptions about the data and is adapted to the context of predicting the value of the measurement (typically quantitative) by those of the dimensions (often qualitative).

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. [5], where the process of projection into the future shows a user-centered approach.

II. Related work

In this section, we propose to combine the various works which covered the topic of the coupling between the data mining and online analysis for to extend OLAP to the prediction.

Some works have for objective prediction of a new cube. They propose to generate a new cube using fairly complex models. Thus Cheng [6] 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 [15] uses the prediction model to predict the measure of a new fact.

Finally in Y. Chen and Pei [8], 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 [15], 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.

III. Our approach

A. 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. [5] However; the notion of query "What-If" is distinguished by the works of Imielinski et al [4] and their recovery by Han et al [7].

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

B. *Proposed Approach*

Our approach is based on a learning step automatic (machine learning) and uses the regression tree. This approach is concretized by the operator of prediction on line by tree of regression (Online Prediction by Regression Tree).

To deploy our approach and for more clearness, we use a simple illustrative example of a cube of data with three dimensions: Years (2010, 2011), Product (shoes, Clothes, Bags) and Country (France, Spain, Morocco, Belgium, Portugal). Measurement corresponds to the turnover of the sales (in Million Dhs) of the products in the countries.

The cube of data is composed of 30 cells. It is considered that the user wants to predict the measurement of the blank cells (those which in are grayed in the figure Fig.1 (A)). We return to the figures Fig.1 (b) and (c) when deploying our method.

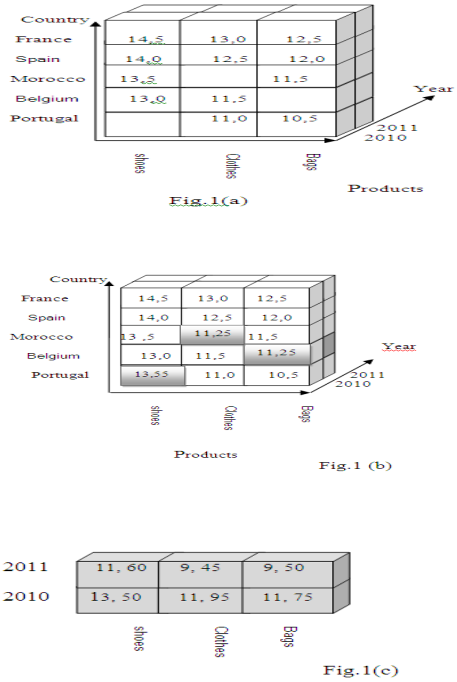
1) General Notations

We adopt the definitions of a cube and sub cube of proposed data in [14] and complement to our needs.

C is a data cube with a non-empty set of dimensions D = {D_1, D_i, ..., D_d} and m measurements M = {M_1, ..., M_q, ..., 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 H^1_1 noted the level of aggregation of all the products corresponding to the hierarchical level 0 H^1_0 noted.

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 two terms: 2010, rated A^{11}_1, and 2011 rated A^{11}_2.

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, ..., D_i, ..., D_d} which made to characterize a given level of aggregation H^i_j.



	shoes	Clothes	Bags
France	14,5	13,0	12,5
Spain	14,0	12,5	12,0
Morocco	13,5		11,5
Belgium	13,0	11,5	
Portugal		11,0	10,5

Fig.1(a)

	shoes	Clothes	Bags
France	14,5	13,0	12,5
Spain	14,0	12,5	12,0
Morocco	13 ,5	11,25	11,5
Belgium	13,0	11,5	11,25
Portugal	13,55	11,0	10,5

Fig.1 (b)

	shoes	Clothes	Bags
2011	11, 60	9, 45	9, 50
2010	13, 50	11, 95	11, 75

Fig.1(c)

Figure 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 ≤ d). The p-tuple (θ_1, ..., θ_p) is a sub-data cube $\forall i \in$ {1,. . ., P}, $\theta_i \neq \emptyset$, and there is a single index j ≥ 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 (θ_1, ..., θ_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 (θ_1, θ_2, θ_3) = (Products, Country, Years) = ({Shoes, Clothes, Bags}, {France, Spain, Morocco, Belgium, Portugal}, {2010, 2011}) is: (three dimensions)
- Measurement M_q corresponds to turnover of the sales, M_q is the variable to predict

- Products, Country and Years play the role of explanatory variables

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) [9]. This approach was taken in [10] where the algorithm CHAID (Chi-Squared Automatic Interaction Detection) is proposed. Breiman et al. [11], offer binary trees with CART (Classification and Regression Tree). Recently, other types of trees have emerged, including Arbogodaï of Zighed et al. [12].

Breiman et al [11] 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 lets 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) *Construction and validation of the model prediction*

To build a regression tree analysis of the context (ɵ$_1$, ..., ɵ$_p$), 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 0.241 which is acceptable. The reduction of an error $1\text{-}R^2$ (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. In this case, we say that the explanatory variables (dimensions) do not allow predicting the measurement.

In the end, the built model (regression tree) during the learning phase must be evaluated at the test phase. The learning tree is used on the sample test to predict the value of measurement for each observation or fact. If the average error and the reduction error in test phase are weak and close to those obtained in phase, then the model is validated.

4) *Interpretation of the prediction model*

After building the model, the regression tree decision returns λ rules ($\lambda \geq 0$). All the rules of a model is denoted by $R = \{ \mathfrak{R}_1, \mathfrak{R}_2, ..., \mathfrak{R}_\lambda\}$.

Definition (Decision Rule):

let $\mathfrak{R}$ $(X \Rightarrow Y; S; \sigma)$ a decision rule $\in$ R.

X is a conjunction and / or a disjunction of terms $\subset$ (ɵ$_1$, ..., ɵ$_p$) 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.

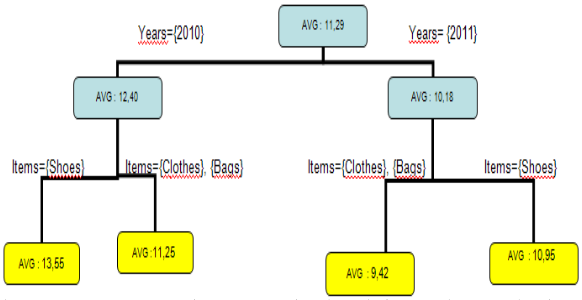


Figure 2. Regression tree obtained from the analysis context.

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

- *R*1(*2011* ∧ (*Clothes* ∨ *Bags*) ⇒ 9,42 ; 33% ; 0,83)
- *R*2(2011 ∧ *Shoes* ⇒ 10,95 ; 17% ; 0,91)
- *R*3(*2010* ∧ (*Clothes* ∨ *Bags*) ⇒ 11,25; 33% ; 0,84)
- *R*4(*2010*∧ *Shoes* ⇒ 13,55 ; 17% ; 0,64)

Each rule corresponds to a terminal leaf of the tree. For example, the rule R1 indicates that if the products clothes or Bags are purchased in 2011 so the turnover sales of these products will be 9.42, 33% of sales (facts) present in the training file in this category and the standard deviation is 0.83. The products and the years are the most discriminating variables. They are explanatory of the turnover of sales, unlike Country that are not determinants.

5) *Operation of the predictive model in the OLAP environment*

- Let (ɵ1, ..., ɵp) 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 (ɵ1, ..., ɵp), 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, ie 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; (ɵ1, ..., ɵp) ) :
for each $\Re$ in R do
    for each cell c do
        If $M_q$ (c) is empty then
            Mq (c)←Y
        end if
    end for
end for
```

For example, when we targeted the cell described by the terms (2010, Shoes, Portugal)) for dimensions, respectively: Years, products and country,

(*R*4(*2010*∧ *Shoes* ⇒ 13,55 ; 17% ; 0,64) was selected

, (2010 ∧ Shoes) $\subset$ (2010, Shoes, Portugal).. We note that sales of products Shoes in years like 2010 will Turnover of sales 13, 55.

if the sale is made in Portugal. For another example, in terms of query such as "What-If" regression tree allows us to know it would be the of sales if we sell the product “Clothes” in “Morocco” for 2010?. Indeed we see that the sales of Clothes in Morocco in 2010 would have like turnovers of sales 11, 25 MDhs.

This integration of the prediction also allows the user to understand the predicted values of aggregates for a higher level. Aggregates are recalculated considering the new predicted values. For example by choosing the level all for countries, the turnover of sales can be calculated by year and by product (see Tab.1).Thus, the average of the turnover of sales planned in year 2010 for all country and for the Shoes products is 13, 50 MDhs.

TABLE.1 PREDICTED VALUES OF AGGREGATES FOR A HIGHER LEVEL

	Year	
Code Product	2010	2011
Shoes	13, 50	11, 60
Clothes	11, 95	9, 45
Bags	11, 75	9, 50

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.

IV. A Case Study

In this paragraph, we experience our work on a real dataset. We use the data from a loans department of a bank in Morocco. The facts correspond to the customers who have benefited the various appropriations suggested by the bank. 1000 facts are present in the data cube.

A. *Background Analysis*

Our analysis context is defined as follows. For dimensions, we use the sex, the activity, the yearly income, the marital status, the sex and eventually followed the types of allotted appropriations.

The Measure used is the rate of solvency of customers during refunding of his appropriation. We propose an analytical representation of the context in the form of star schema in Fig.3:

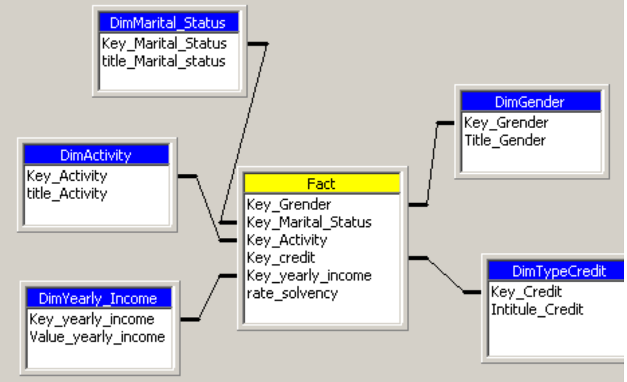


Figure 3. Representation in the form of star schema of the context analysis.

Thus, in this context of analysis, a user can answer various questions, as for example : which would be the rate of solvency of an unmarried customer of female sex, exerting in the sector ` health' like activity and gaining an annual salary `$30K-$50K' if he asked a appropriation according to his criteria ?

B. *prediction model*

We choose the learning algorithm AID [9] as method of tree of regression to build the model of prediction in the context of analysis previously definite. This last is segmented as follows: 70% of the facts constitute the sample of learning, either 700 aggregate facts and the 30% remainders, 300 aggregate facts, constitute the sample test. We parameterize the method with a minimal number on each top fixed at 5 aggregate facts and a maximum number of levels in the tree to 8.

The tree of regression built with the sample of learning comprises 9 tops and 5 sheets (fig 4). On the sample test, the average error of the tree is of 5, 1 and reduction of the error is of 0, 81 then the model must thus be exploited with precaution.

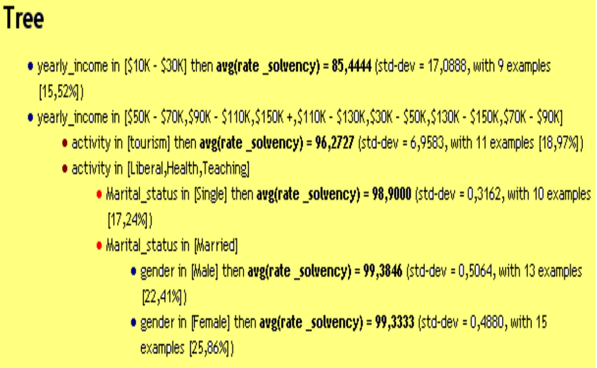
C. *Interpretation of the prediction model*

The discriminating explanatory variables (dimensions) are, in the order of their appearance in the tree: yearly income, activity, marital status and gender. Other dimensions are not thus variables explaining the rate of solvency to be predicted.

We obtain the 5 following rules; each one corresponds to a terminal sheet of the tree:

- R1($10K - $30K ➔85,44; 15,52%; 17)
- R2(($30K - $50K v $50K - $70K v $70K - $90K v $90K - $110K v $110K - $130K v $130K - $150K v $150K +) ^ (tourism) ➔96,27; 18,97%; 6,95)
- R3(($30K - $50K v $50K - $70K v $70K - $90K v $90K - $110K v $110K - $130K v $130K - $150K v $150K+) ^ (Liberal v Health v Teaching) ^ Single ➔98,90; 17,24%; 0,32)
- R4(($30K - $50K v $50K - $70K v $70K - $90K v $90K - $110K v $110K - $130K v $130K - $150K v $150K+) ^ (Liberal v Health v Teaching) ^ Maried ^ Male ➔98,38; 22,41%; 0,51)
- R5(($30K - $50K v $50K - $70K v $70K - $90K v $90K - $110K v $110K - $130K v $130K - $150K v $150K+) ^ (Liberal v Health v Teaching) ^ Maried ^ Femele ➔99,33; 25,86%; 0,49)

As example one interprets the R4 rule as follows: if the yearly income of a customer is worth ($30K - $50K or $50K - $70K or $70K - $90K or $90K - $110K or $110K - $130K or $130K - $150K or $150K+), its activity is (Liberal or health or teaching), of male sex and married then the rate of solvency of the customer requesting a credit will be 98,38. 22, 41% of the individuals of the whole of learning support this rule and the standard deviation is of 0,51.



Tree

- yearly_income in [$10K - $30K] then **avg(rate _solvency) = 85,4444** (std-dev = 17,0888, with 9 examples [15,52%])
- yearly_income in [$50K - $70K,$90K - $110K,$150K +,$110K - $130K,$30K - $50K,$130K - $150K,$70K - $90K]
    - activity in [tourism] then **avg(rate _solvency) = 96,2727** (std-dev = 6,9583, with 11 examples [18,97%])
    - activity in [Liberal,Health,Teaching]
        - Marital_status in [Single] then **avg(rate _solvency) = 98,9000** (std-dev = 0,3162, with 10 examples [17,24%])
        - Marital_status in [Married]
            - gender in [Male] then **avg(rate _solvency) = 99,3846** (std-dev = 0,5064, with 13 examples [22,41%])
            - gender in [Female] then **avg(rate _solvency) = 99,3333** (std-dev = 0,4880, with 15 examples [25,86%])

Figure 4. Regression tree on the data cube obtained with AID

D. *Exploitation OLAP of the prediction model*

We find in table (cf Tab.2) the integration of the results obtained starting from the tree of regression for the 5 cells. As example, rule 1 is used for the first line of the table. The checked condition is: if a customer of married male sex having an annual income `$10k-$30' and exerting in the sector `Tourism' like activity, then his rate of solvency will be 85,44.

Similarly, by using the whole of the rules, all the blank cells described by the yearly income, activity, marital status and the Sex can be estimated (cf Tab.2).

TABLE.2. PREDICTED VALUES FOR THE RATE OF SOLVENCY

Yearly_income	**Activity**	**Marital_status**	**Gender**	**rate of solvency**
$10k-$30	Tourism	Maried	Male	85,44
$30K-$50K	Health	Single	Female	99,33
$110K-$130K	Tourism	Maried	Female	96,27
$10k-$30	Teaching	Single	Female	85,44
$150k+	Liberal	Maried	Male	99,38

Each rule corresponds to a terminal leaf of the tree. for example the R2 rule indicates that if the annual income of a customer is worth ($30K - $50K or $50K - $70K or $70K - $90K or $90K - $110K or $110K - $130K or $130K - $150K or $150K+), its activity is {Tourism}, some is its sex and its marital status then the rate of solvency of the customer will be 96,27. 18,97% of the individuals of the whole of learning support this rule with the standard deviation is of 6,95.

Yearly_income dimensions, activity, gender and marital_status , are most discriminant variables. They are explanatory of the results the rate of solvency of the customer for a credit requested, contrary to the types of credit that are not determining.

To exploit the predictive model in environment OLAP, we indicate the cells for which we wish predict the value of their measurement, after we seek among the rules of the tree of regression obtained that which corresponds to the whole of the modalities describing the cell C then we affect the average value of the rule of prediction like value of the measurement of the cell. The operation is reiterated for each cell designated for prediction.

This integration of the prediction also makes it possible the user to apprehend the values envisaged of the aggregates for a higher hierarchical level (cf Tab.3).. The aggregates are recomputed considering the new predicted values. For example by choosing the All level for the types of credit, the average of the rates of solvency can be calculated according by sex, and type of activity of the customers.

TABLE.3. VALUES ENVISAGED OF THE AGGREGATES FOR A HIGHER HIERARCHICAL LEVEL

	Gender	
Activity	Male	Female
Tourism	89,78	91,92
Health	99,05	99,22
Liberal	98,66	98,00
Teaching	99,25	99.66

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 inorder 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] M. Golfarelli, S. Rizzi, and A. Proli. Designing what-if analysis: towards amethodology. In Proceedings 9th International Workshop on Data Warehousing and OLAP (DOLAP 2006), pages 51–58, Arlington, USA, 2006

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

[7] J. Han, J. Wang, G. Dong, J. Pei, and K. Wang. Cubeexplorer: 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

[8] 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.

[9] 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.

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

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

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

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

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

[15] 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.

AUTHORS PROFILE

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 AgadirAnd 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.