Clustering Application for prediction in OLAP Cube

Asma Lamani¹, Brahim Erraha², Malika Elkyal³

(Laboratory of Industrial Engineering and Computer Science (LG2I),
National School of Applied Sciences -Agadir, University Ibn Zohr, Morocco)

Abstract:
Data warehouses are designed to help users to manage large volumes of data. Online analysis OLAP provides a simple and quick visualization of the information and multidimensional view of data. On the other hand, Data mining is a set of techniques which allows the extraction of knowledge from data warehouses.

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 on-line analysis and data mining, an approach based on automatic learning and clustering was proposed in order to integrate the prediction in the heart of OLAP.

Keywords — online analysis OLAP, data mining, multidimensional data cube, prediction, clustering.

I. INTRODUCTION
Data warehouses are collections of data organized to support a process of decision, and provide an appropriate solution for managing large volumes of data.

The Online analysis is a technology that complements data warehouses to make data usable and understandable by users in the most natural way possible, by providing tools for visualization, exploration and navigation in data cubes.

On the other hand, data mining enables knowledge discovery in data with different methods of description, classification, explanation and prediction.

With the increase in processed information obtained during the work of information processes, its processing becomes difficult. The need for initial processing of information for its structuring, the isolation of characteristic traits, generalization, sorting appears.

For this purpose, the classification and clustering processes are used which make it possible to completely carry out the processing of the required information, for its subsequent analysis by a specialist.

The partitioning of observations into groups of similar objects makes it possible to simplify the subsequent processing of data and decision-making by applying to each cluster its method of analysis ("divide and govern" strategy).

Thus, grouping is used to detect novelty. Atypical objects, which are not homogeneous, are separated, and the homogeneous objects are grouped together in the same cluster.

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 [1].

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.
Our 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.

II. RELATED WORK

Coupling methods for data mining with OLAP is an approach that has already proven itself. Without being exhaustive, we can cite the work of S. Cheng [2], the work of Sarawagi [3], related work of B.C. Chen [4], [5], and work of Palpanas [6], [7].

In this context, Sarawagi and al [3] builds a new 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.

R. Ben Messaoud[8] define three coupling approach, 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.

Continuing the work of R. Ben Messaoud, and to integrate the prediction in OLAP, an approach based on automatic learning with regression trees is proposed by A.Sair [9] in order to predict the value of an aggregate or a measure.

III. OUR APPROACH

Our approach is to integrate prediction into the OLAP environment for decision support. Our goal is to allow the analyst to predict the value of a measure for a new fact according to a defined analytical context 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.

In this article, 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.

A. General notions

We take the definitions proposed in [8] of a data cube and a data sub-cube. C is a data cube with:
- a non-empty set of d dimensions $D = \{D_1, D_2, \ldots, D_d\}$
- and m measurements $M = \{M_1, \ldots, M_q, \ldots, M_m\}$.
- $H_i$ is the set of hierarchies of dimension $D_i$.
- $H_{ij}$ is the jth of hierarchical levels of the dimension $D_i$.
- $A_{ij}$ represents all terms of the hierarchical level $H_{ij}$ of the dimension $D_i$.

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

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

https://www.ijctjournal.org
Our starting point is a context of analysis \((\theta_1, \ldots, \theta_p)\), with \(n\) observed facts according to the quantitative measurement \(M_q\) defined by the user in a data cube \(C\).

We apply the methods of clustering on cube that contains measures of sales income, these measures can be studied on 3 dimensions: \(D = \{\text{Time, Product, Stores}\}\). 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 figure 1 (fig 1) shows the cube of sales.

We choose a context of analysis with 1069 cells, which is a detailed representation of the cube with a lower level of granularity for each dimension: (Month, product, Branch).

![Initial data cube](image)

**B. Clustering methods**

Far from wanting to make a complete state of art of different approaches of clustering, we present the basic concepts. The first concept is the structure of results, depending on the applied method; obtained clusters can be hard or fuzzy sets. Some objects may be not classified and some clusters may overlap and the result can be a hierarchical model.

The second distinction to do concern the strategy used to build clusters, the methods can be classified into four groups: distance-based methods, grid-based methods, model-based methods and hierarchical methods.

We conducted 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 choose the K-means algorithm, and the last is a model-based method, the EM algorithm.

The experimental implementation of these algorithms has been realized using the software TANAGRA.

TANAGRA is a 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. [12]

**C. Results Obtained**

1) **HAC Algorithm**

Initialization:
- Initial class = \(n\) elements.
- Calculation of the distance matrix of the elements 2 to 2.

Iteration of the following steps:
- Grouping the two closest elements.
- Updating the table of distances by replacing the two grouped classes by the new one and recalculating its distance from the other classes.

End of iteration: aggregation of all elements in a single class.

The highest jump in the obtained dendrogram determines the number of clusters.

2) **K-means Algorithm**

K-means is known as a partitional clustering method that allows to classify a given data set through \(k\) clusters already fixed, the main idea is to define \(k\) centroids, one for each cluster, and then assign each fact to the closest cluster, until the centroids no longer move.

The best grouping is the partition of the data set that minimizes the sum of squares of distances between data and the corresponding cluster centroid.

The application of K-means algorithm requires giving the number of partitions you want to have in
result, so we choose to partition our dataset in 3 clusters.

3) EM Algorithm:

The EM algorithm is defined as: [11]

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 \( \theta \), along with a likelihood function:

\[
L(\theta; X, Z) = p(X, Z|\theta)
\]  

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

\[
L(\theta; X) = \mathbb{E}_{Z|X, \theta} \left[ \log \left( L(\theta; X, Z) \right) \right] - \mathbb{E}_{Z|X, \theta} \left[ \log \left( \pi(X, Z) \right) \right]
\]  

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 \( \theta \) (t):

\[
Q^t(\theta|\theta^{(t)}) = \mathbb{E}_{Z|X, \theta} \left[ \log L(\theta; X, Z) \right]
\]  

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

\[
\theta^{(t+1)} = \arg \max_{\theta} Q^t(\theta|\theta^{(t)})
\]  

The main issue is therefore to select the correct number of clusters. 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 ».[14]

\[
BIC = -2 \ln(\mathcal{L}) + \ln(N)k
\]

With \( \mathcal{L} \) is the likelihood function, \( N \) is the number of observations, \( K \) is the number of clusters to be estimated.

The selected model is the one that minimizes the BIC criterion; in our example, the partition in 8 classes seems most appropriate.

4) Results:

According to Figures 2, 3 and 4, we observe that all elements of the produced clusters by the three algorithms are well classified; we can notice that the groups are homogenous.

In HAC algorithm, the number of clusters cannot be known in advance. The system takes the dataset as an input, and gives a cluster tree in output.

HAC processes data in pairs for sorting and creating partitions which makes it heavy according to the large databases.

K-means algorithm is the most used for huge databases. However, this latter requires specifying the number of clusters as input.

The EM algorithm seems to be the most strongest for classification, it allows the processing of huge databases, and through the application of the BIC, we can estimate the number of partitions which guarantees the best representation of the data.

<table>
<thead>
<tr>
<th>Sub-cube</th>
<th>HAC</th>
<th>K-means</th>
<th>EM with BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>528</td>
<td>271</td>
<td>181</td>
</tr>
<tr>
<td>2</td>
<td>177</td>
<td>552</td>
<td>90</td>
</tr>
<tr>
<td>3</td>
<td>364</td>
<td>246</td>
<td>390</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>54</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>148</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>117</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>23</td>
<td></td>
</tr>
</tbody>
</table>
D. A Case Study

In this section, we will 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. 16757 facts are present in the data cube.

The cornerstones of analysis of the warehouse analysis are: authorization type (Permit to construct, Permit to demolish, etc. ...), subdivision, district, nature of the project and filing date;

The measure used is the authorization demand's treatment duration (number of days).

Our analysis context is defined as follows: For dimensions, we use authorization type, subdivision, nature of the project, and filing date.

We choose EM algorithm with BIC as a method of clustering to partition our cube. The obtained clusters are shown in Figure 5 and table 2.

IV. CONCLUSIONS

As part of this work, we offer a new approach for the prediction in OLAP cubes.

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.

After applying the Clustering for partitioning our initial cube in dense sub-cubes, we will proceed in the next stage to create and validate a prediction model for each sub-cube using the solution based on automatic learning with regression trees proposed in the work of A.Sair [9].

We determine the sub-cube in which the empty cell designated by user, the cell value is then predicted using the model of the sub-cube.

REFERENCES


[13] Distributional Similarity Models, Regina Barzelay, EECG Department, MIT, October 15, 2004

[14] Wit, Ernst; Edwin van den Heuvel; Jan-Willem Romeyn, «’All models are wrong...’: an introduction to model uncertainty ». Statistica Neerlandica. 66 (3):217–236