ISSN:0975-9646

Machine to Machine (M2M) Open Data System for Business Intelligence in Products Massive Distribution oriented on Big Data

Angelo Galiano[#], Alessandro Massaro[#], Donato Barbuzzi[#], Leonardo Pellicani[#], Giuseppe Birardi[#], Bachir Boussahel[#], Francesca De Carlo[#], Veronica Calati[#], Giovanni Lofano[#], Laura Maffei[#], Marilena Solazzo[#], Vito Custodero[#], Gaetano Frulli[*], Egidio Frulli[*], Francesca Mancini[*], Leonardo D'Alessandro[+], Francesco Crudele[+]

[#]*Dyrecta Lab srl,*
Via Vescovo Simplicio, 45 – 70014 Conversano (BA) – Italy

[*]*Fruman Rappresentanze srl,*
Via Demetrio Marin, 35 - 70125 Bari (BA) – Italy

[+]*INGEL srl,*
Via Mantova, 23 - 70014 Conversano (BA) – Italy

Abstract— **In this work we present useful Graphical User Interfaces (GUIs) suitable for clustering, classification, and sales predictions of the "Grande Distribuzione Organizzata" GDO (major multiples/large-scale retail channel, chain store/mass distribution) products. This Open Data Machine to Machine (M2M) system combines all products, provided both from an open scenario and from a proprietary database, in order to cluster and to classify new data for the creating of predictive rules. The experimental results, carried out on massive simulated data, demonstrate the efficacy of GUIs for the application of Business Intelligence rules.**
For the purpose, the massive data are stored by means of Cassandra DB a Big Data System. Instead, the GUIs implementation, related to the main tasks of Machine Learning, are designed in Orange tool. Finally, the prediction results are validated respect to most know tool in literature: Weka.

Keywords— **GDO, Machine to Machine Systems, Data mining, GUI, Prediction, DBMS, Big Data.**

I. INTRODUCTION

Data mining techniques [1]-[6] are commonly used for statistical and market basket analyses, as well as for business intelligence activities. Important aspects, in the use of data mining approaches, are related to the interfacing with data systems, including: M2M [7], Open Database (DB) and data source systems [8]. Concerning massive product distribution channels, it is important to design particular GUIs (or widgets) able to provide data mining results in real time, such as the clustering [9], classification [9], prediction [10], statistical trends [11], market basket analysis [12] and so on.

A critical aspect related to the data processing in these GUIs is the importing of products. So, the GUIs will be able to standardize the different attributes in unique values in order to extract Business Intelligence rules by data. This point requires the automatically identification (or clustering) of the products in a generic typology useful for data mining processing. In this way it is necessary to create a standard format interfacing respect to the data, for example, provided by electronic balances or points of sales (POS), local open databases, local proprietary database, external open database. For instance, local proprietary DB such as Magento [13] and Danea [14] are considered as Data Base Management System (DBMS). While, for open external DB are considered those external systems with different features and classes, for example wine, beer, vegetables and so on. Of course, a correct correlations by them provides the first step for promoting actions [15], and for product facing [16],[17]. In particular promoting actions could be adopted in real time in order to correct the marketing procedures, as well as the correct positioning of products for the sales optimization (through market basket analysis).

For the processing of massive data, it is necessary to import data in a Big Data system. An example of powerful big data system is Cassandra DB [18] which is used in this work. In this paper we will be discuss: (i) an application of the DB in a defined system architecture and contest; (ii) the GUI interface for product clustering/classification; (iii) calculus of sales prediction by means of linear regression; (iv) validation of prediction results; (v) GUI interface for market basket analysis.

II. SYSTEM ARCHITECTURE AND DATABASE DESIGN

In this paper, the system architecture is sketched in Figure 1. Firstly, the features of the products are organized in a structured unique format in order to run the clustering task by means of a data mining tool. All data will be grouped in 4 clusters indicated as: "food", "no food", "beverage" and "fresh". The entire DB (named Open) is stored in Cassandra, which it contains all loaded and standardized data: the data will be processed by analytics and/or data mining tools, and the outputs will be plotted by

data mining tools or by means of platforms linked with Cassandra DB such bas Cloud9Charts suitable for filtering data and to perform structure query.

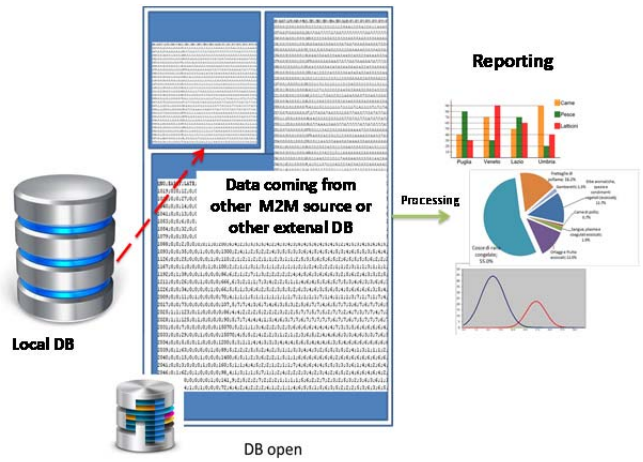

Figure 1. Basic architecture of partial opened DB system applied to GDO

In Figure 2 is illustrated a basic architecture of the proposed system. The database has been designed by DB Designer 4 tool. In the tables, the primary key and foreign key identify different relationship between products, clients and promotional activities available. DB Open has been designed in order to collect different information suitable for business intelligence planning: sensitive data such as promoting actions, warehouse data, customers data, product orders, product data and so on could improve the business intelligence performances by following proper criteria of processing.

In appendix are reported the NoSQL scripts related to the designed tables. This script provides useful information in order to understand the structure of the whole database.

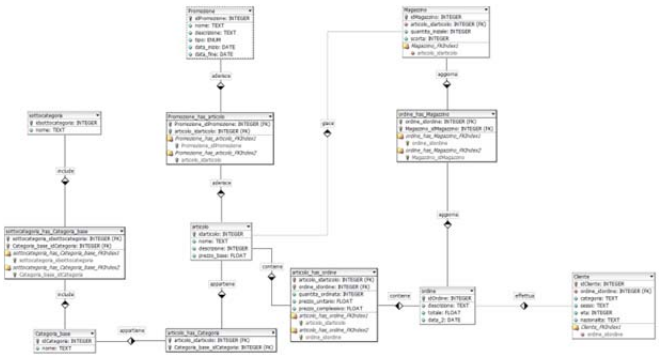

Figure 2. Example of a generic database architecture for GDO products

III. GDO Product Clustering/Classification

In this section we will describe the procedure to implement product clustering and classification. In particular, Figure 3 shows all steps for clustering and classification tasks, using GUIs implemented by Orange Canvas tool. As first step, the data coming from Open DB are preprocessed by the "Concatenate" and "Select Attributes" widgets in order to facilitate the clustering process performed by the "k-Means Clustering" widget. The k-Means algorithm is typically used for collect data respect to a reference point named centroid. All clustered data are shown in a table through the use of "Data Table" widget. Moreover, this widget transfers the selected data to the classification section, where the "Data Sampler" widget splits data for training and test phases. So, the "Prediction" widget will classify new data in the cluster previously generated. Naive Bayes and k–Neraest Neighbours (k-NN) Classifiers are considered: both are classification algorithms based on Euclidean distance and on Bayes probability, respectively. Finally, the output data of the classifiers are plotted and evaluated by means of the particular widgets: more specifically, the "Scatter Plot" function plots the date and the "Test Learner" evaluates the classifiers. The results of the evaluations are shown by the "Confusion Matrix", by the "ROC Analysis" and by the "Scatter Plot" widgets.

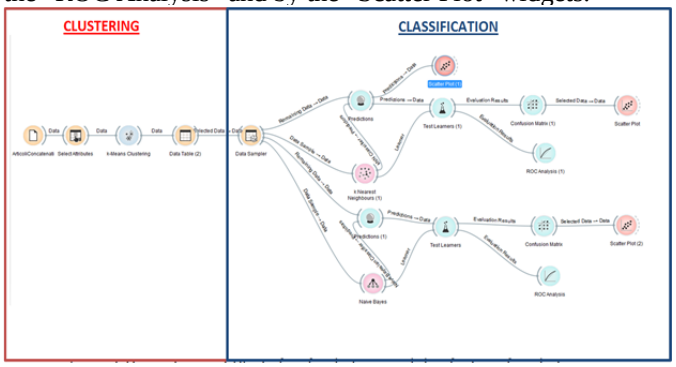

Figure 3. Orange Canvas: GUI interface for clustering and classification of products

In Figure 4 is illustrated the Confusion Matrix concerning the Naïve Bayes classification of the four classes: C1, C2, C3, C4 (food, no food, beverage, fresh, respectively). This Confusion Matrix indicates that all products are correctly classified. The correct classifications are reported along the main diagonal of the Confusion Matrix. We observe that the Naïve Bayes approach is most accurate if compared with k-NN which provides only one classification error.

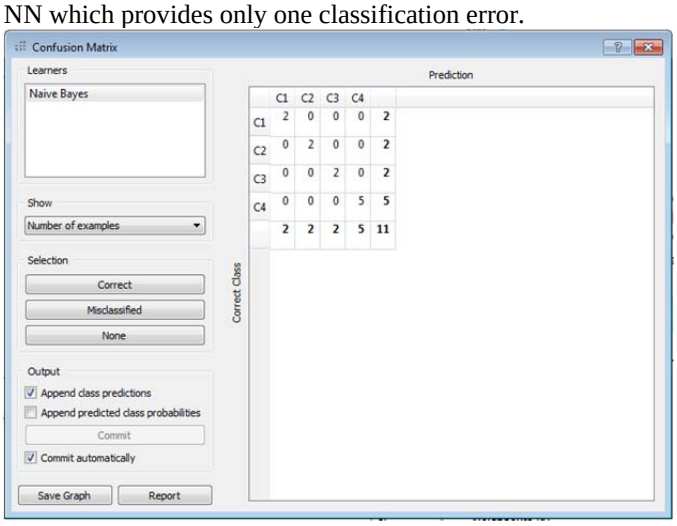

Figure 4. Orange Canvas: Confusion Matrix related Naive Bayes data processing

The ROC curve help to evaluate graphically the performance of the classifiers.

In Figure 5 is illustrated the ROC curve concerning evaluation of the Naïve Bayes classifier thus proving the efficiency of the specific classifier.

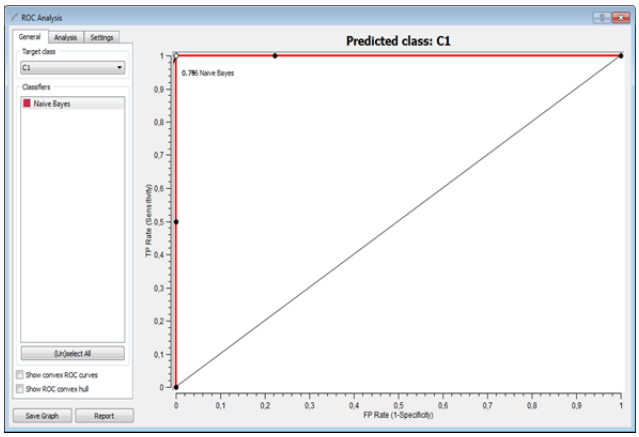

Figure 5. Orange Canvas: ROC Curve related Naive Bayes classifier

IV. Linear Regression and Sales Prediction

Sales prediction has been performed by means of the linear regression model, which is represented by the following equation:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

where

- i is the index which changes with the observations, i=1,…,n;
- Y_i is the dependent variable;
- X_i is the independent variable;
- $\beta_0 + \beta_1 X_i$ is the regression function;
- β_0 is the intercepts of the regression function;
- β_1 is the angular coefficient of the regression function;
- u_i is the statistical error.

In Figure 6 is illustrated the Orange Canvas GUI interface, related to the prediction estimation of a store. More specifically, three months of historical sales data are concatenated into an unique table and these are processed by means of the following widgets: “Linear Regression”, “Prediction” and “Test Learner”. The processed results were analyzed, successively, for a comparative analysis with Weka linear regression. The same method could be applied for more historical data addressing the analysis on Big Data system.

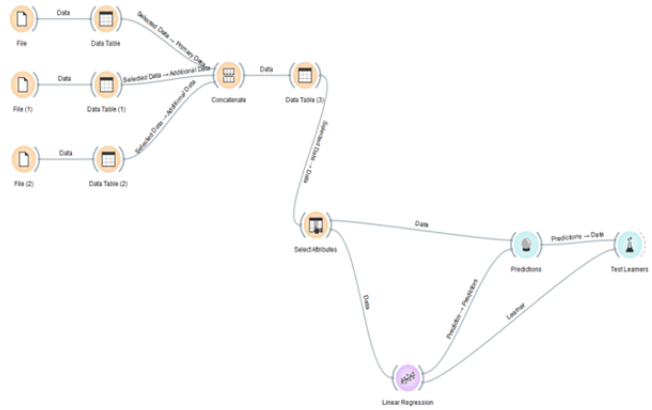

Figure 6. Orange Canvas: GUI interface for prediction by means of linear regression approach

In Figure 7 is illustrated a screenshot related to prediction results of a product (pasta) correlated with the sales of other products. We assume that the number 1 indicates the sales of a product over a certain quantity and it is related to the previous month. For example, concerning the first row, the prediction indicates good sales of pasta for good sale of Beers, Sausages and Cakes and cookies (and not for Drugs, Wines, Care and welfare and House care). We observe that the prediction results take into account of the correlation with the sales of the other products. This correlation could be enhanced also by the market basket analysis. The results are validated by comparing the same simulation of Orange Canvas with the linear regression of the Weka tool. In table 1 are reported the calculus parameters (Root Mean Squared Error, Relative Absolute Error, Mean Absolute Error, Relative Squared Error) for both the simulations: the very low variations confirm the validation of the linear regression results.

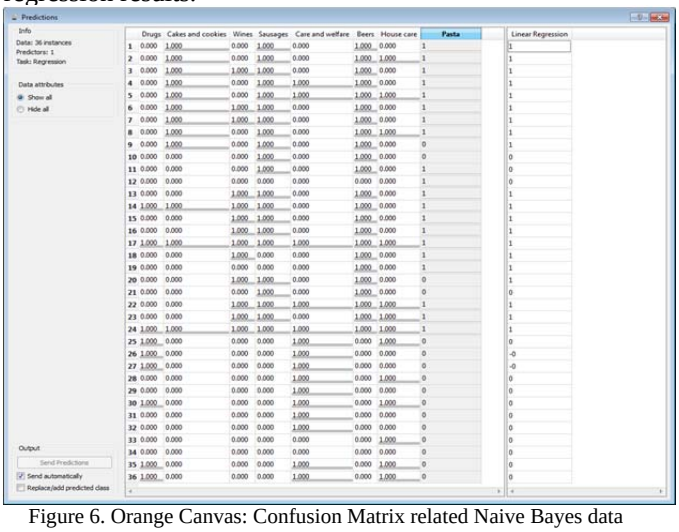

Figure 6. Orange Canvas: Confusion Matrix related Naive Bayes data processing.

TABLE I
Comparison between calculus parameters of Orange Canvas and Weka tools

Calculus parameters	Orange Canvas	Weka
Root Mean Squared Error	0.3806	0.3952
Relative Absolute Error	0.5636	0,543376
Mean Absolute Error	0.2783	0.2717
Relative Squared error	0.7660	0.784802

V. GUI for Market Basket Analysis

Market basket analysis represents a supporting study to the prediction. Some parameters such as support, confidence, lift, leverage, strength and coverage could provide important information about relationships of the sales. In Figure 7 is illustrated the Orange Canvas GUI for the market basket analysis, using the Association Rules. We assumed as data input of the GUI the products coming from three different files obtained by exporting data from the Open DB and indicating the sales of three store (for example of the same group) in the current month. The data have been previously uploaded in the Cassandra DB. The

three data tables of the files are concatenated by means of the "Concatenate" widget in order to merge data. The widgets "Association Rules" and "Association Rules Explorer" will estimates all the parameters as reported in Figure 8.

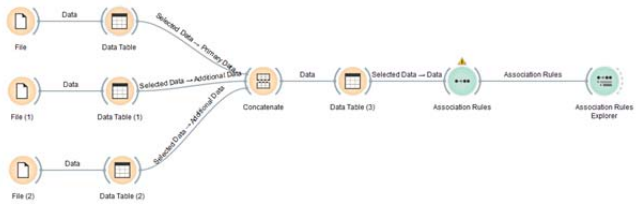

Figure 7. Orange Canvas: GUI of calculus of market basket analysis parameters

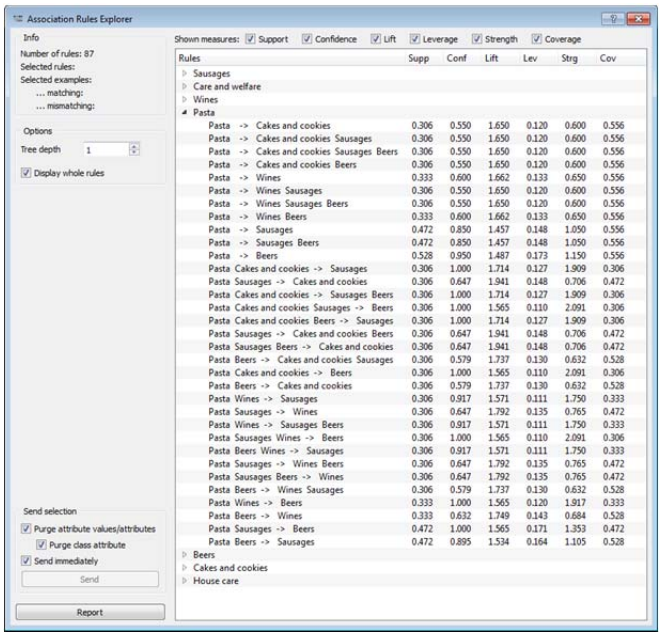

Figure 8. Orange Canvas: screenshot of market basket parameters.

VI. Conclusion

New GUIs are presented in order to obtain Business Intelligence rules in real time on massive data. More specifically, the two main goals of this work are: (1) to store massive data, provided in an open scenario, on Cassandra DB, (2) to extract useful information by massive data to optimize the sales strategies, using the open source tool ORANGE Data Mining. Experimental results, carried out on simulated data, show the efficacy of the designed GUIs for the following tasks: Clustering, Classification, Association Rules for the goal to increase sales. Finally, all possible data could be applied to the system in an Open and Big Data environment. This allows to perform into an unique systems all the data processed in order to optimize the analyses and predictions of the sales. The paper exhibits the most important procedures for GDO using data mining.

Acknowledgment

The work has been developed within the framework of Research Project titled: "Sistema machine to machine di immissione dati di tipo open data per analisi di mercato real time" – (Machine to Machine System for data input in open data for real time market analysis).

References

[1] A. H. Wahbeh, Q. A. Al-Radaideh, M. N. Al-Kabi, and E. M. Al-Shawakfa, "A Comparison study between data mining tools over some classification methods," IJACSA International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence, pp. 18-26, 2011.

[2] Y. Ramamohan, K. Vasantharao, C. K. Chakravarti, and A. S. K. Ratnam, "A study of Data Mining Tools in Knowledge Discovery Process," International Journal of Soft Computing and Engineering IJSCE , vol.2, no.3, pp. 191-194, 2012.

[3] S. H. Begum, "Data Mining Tools and Trends- an Overview," International Journal of Emerging Research in Management & Technology, pp. 6-12, 2013.

[4] M. Verma, and D. Mehta, "A Comparative study of Techniques in Data Mining," International Journal of Emerging Technology and Advanced Engineering," vol. 4, no. 4, pp. 314-321,2014.

[5] S., Impedovo, & D. Barbuzzi, "Instance Selection for Semi-Supervised Learning in Multi-Expert Systems: A Comparative Analysis". Journal of Next Generation Information Technology, 5(4), 61, 2014

[6] D., Barbuzzi, G., Pirlo, & D. Impedovo, "About retraining rule in multi–expert intelligent system for semi–supervised learning using SVM classifiers". International Journal of Signal and Imaging Systems Engineering, 7(4), 245-251, 2014.

[7] J. Holler, V. Tsiatsis, C. Mulligan, S. Avesand, S. Karnouskos, D. Boyle, "From Machine-to-Machine to the Internet of Things: Introduction to a New Age of Intelligence" Elsevier 2014.

[8] J. M. Hellerstein, M. Stonebraker, and J. Hamilton, "Architecture of a Database System," Foundations and Trends in Databases Vol. 1, No. 2, pp. 141–259, 2007.

[9] S-H. Liao, P-H Chu, P.Y Hsiao, " Data Mining techniques and applications- A Decade Review from 2000 to 2011," Expert Systems with Applications vol. 39, pp. 11303–11311, 2012.

[10] D. Das and M. S. Uddin, "Data Mining and Neuronal Network Techniques in Stock Market Prediction: A Methodological Review," International Journal of Artificial & Applications (IJASA), vol. 4, no.1, 2013.

[11] C. Rygielski, J-C. Wang, and D. C. Yen, "Data Mining Techniques for Customer Relationship Management," Technology in Society vol. 24, pp. 483-502, 2002.

[12] L. C. Annie M.C., A. Kumar D., "Market Basket Analysis for a Supermarket based on Frequent Itemset Mining," International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012.

[13] https://magento.com/

[14] https://www.danea.it/

[15] F. Dallari e D. Milanato, «Progettazione e controllo delle promozioni commerciali,» Logistica, pp. 37-41, 2013.

[16] BUTTLE F. (1984), "Retail Space Allocation", International Journal of Physical Distribution & Materials Management, 14 (4), 3-23.

[17] SABBADIN E. (1991), Merchandising, Packaging e Promozione, le Nuove Dimensioni della Concorrenza Verticale, Franco Angeli, Milano.

[18] http://cassandra.apache.org/

APPENDIX - NOSQL code

```
CREATE KEYSPACE fruman_dyrecta WITH
REPLICATION ={ 'class' :
'NetworkTopologyStrategy',
'datacenter1' : 3 };

CREATE TABLE auto_increment (
counter_value counter,
nome_tabella varchar,
PRIMARY KEY (nome_tabella) );

CREATE TABLE articolo (
idarticolo varint ,
nome TEXT ,
descrizione varint ,
prezzo_base FLOAT ,
PRIMARY KEY(idarticolo) );

CREATE TABLE articolo_has_Categoria (
articolo_idarticolo varint ,
Categoria_base_idCategoria varint ,
PRIMARY KEY(articolo_idarticolo,
  Categoria_base_idCategoria) );

CREATE TABLE articolo_has_ordine
( articolo_idarticolo varint ,
ordine_idordine varint ,
quantita_ordinata varint ,
prezzo_unitario FLOAT ,
prezzo_complessivo FLOAT , PRIMARY
KEY(articolo_idarticolo,
ordine_idordine) );

CREATE TABLE Categoria_base
  ( idCategoria varint , nome TEXT ,
  PRIMARY KEY(idCategoria) );

CREATE TABLE Cliente (
idCliente varint ,
ordine_idordine varint ,
categoria TEXT ,
sesso TEXT ,
eta varint ,
nazionalita TEXT ,
PRIMARY KEY(idCliente) );

CREATE TABLE Magazzino (
idMagazzino varint ,
articolo_idarticolo varint ,
quantita_iniziale varint ,
scorta varint ,
PRIMARY KEY(idMagazzino) );

CREATE TABLE ordine (
idOrdine varint ,
descrizione TEXT ,
totale FLOAT ,
data_2 timestamp ,
PRIMARY KEY(idOrdine) );

CREATE TABLE ordine_has_Magazzino (
ordine_idordine varint ,
Magazzino_idMagazzino varint ,
PRIMARY KEY(ordine_idordine,
  Magazzino_idMagazzino) );

CREATE TABLE Promozione (
idPromozione varint ,
nome TEXT ,
descrizione TEXT ,
tipo varchar ,
data_inizio timestamp ,
data_fine timestamp ,
PRIMARY KEY(idPromozione) );

CREATE TABLE Promozione_has_articolo (
Promozione_idPromozione varint ,
articolo_idarticolo varint ,
PRIMARY KEY(Promozione_idPromozione,
  articolo_idarticolo) );

CREATE TABLE sottocategoria (
idsottocategoria varint ,
nome TEXT ,
PRIMARY KEY(idsottocategoria) );

CREATE TABLE
  sottocategoria_has_Categoria_base (
sottocategoria_idsottocategoria varint ,
Categoria_base_idCategoria varint ,
PRIMARY
  KEY(sottocategoria_idsottocategoria,
  Categoria_base_idCategoria) );
```