AUTOMATYKA • 2008 • Tom 12 • Zeszyt 2

Waldemar Wójcik*, Konrad Gromaszek*

The Use of Data Mining Approach to Prediction Control Strategies for Industrial Processes

1. Introduction

Data mining stands one of the stages of Knowledge Database Discovery (KDD) process. The origins of the data mining techniques are based on such science disciplines like statistics (statistical multidimensional analysis) or machine learning. This idea combines regularities finding (hidden for human, because of time limitations) with computer's calculation speed in large amount of data [5].

Nowadays, data mining techniques met a very intensive interest of the business, providing the key strategic tools relevant to Business Intelligence implementations.

The idea is not new; it is rather based on efficient usage of the information already stored in the enterprise for fasten and more exact (strategic) decision making.

Data mining techniques may be used for several business problems. Depending on the problem character, they can be divided into several tasks: classification, segmentation, basket analysis, regression, prediction, association analysis, anomaly detection, sequence analysis, time-series analysis, text categorization, advanced insights discovery and others. In this work prediction task is considered, with DMX language examples, because it is the aim of 80% implementations.

2. SQL Server and Analysis Services

Being complex, mining techniques require stable and infallible framework. There are several data mining software vendors with leadership of Oracle, Microsoft and IBM. The Microsoft SQL Server 2005 seems to be very interesting in this area, as it stands complete business intelligence solution. Apart from relational database management system (RDBMS)(OLTP database engine), it also offers Integration Services, Analysis Services as well as Reporting Services. Whole framework hold forth to transfer data between different systems, summary reports creation as well as advanced warehouses implementation.

* Katedra Elektroniki, Politechnika Lubelska w Lublinie

The common practice shows that several stages of production are handled by different databases. Typically, relational database (OLTP), records are assigned to individual transactions. It contains measured output and control signal values as well as repository states.

Despite of continuous transaction processing from different measuring devices and actuators, such relational model is unable to specify what is the total efficiency or emergency during the particular production interval. Answer for this question requires time consuming summaries in many tables, often using very complex join operations. The main disadvantage is, that the whole long-lasting computation needs to be repeated after questions about efficiency from several departments.

The reports of this kind are improved by the Analysis Services, that store pre-processed data in the warehouse as well as they are more suitable for preparing reports. The typical database consists of fact table with aggregated values, named measures which are divided into time, departments etc. However, the fact table of OLAP (On-Line Analytical Processing) database contains keys (not values) for measures tables, apart from aggregated values.

Data contained in the fact table constitute multidimensional cube, or hypercube precisely, because there can be more than three dimensions in an OLAP system. The arrangement of data into cubes avoids a limitation of relational databases which are not well suited for near instantaneous analysis of large amounts of data.

Usually star schema is used for linking the fact table with measures tables. Such a structure allows for simple creation of queries and reports, using Multidimensional Expressions and Data Mining Extensions languages.

Analysis Services use previously acquired and pre-processed data by the Integration services, but often they become source for Reporting Services and other applications [2, 4, 8, 9].

3. Prediction using DMX language

3.1. The concept

Data mining is used for prediction analysis by many big companies, mainly for marketing research. The technology lets a retailer use collections of data gathered from customers with similar purchases to predict, for instance, the most preferable goods bought by particular age female at 25–30 at his local supermarket.

This technology can be used for similar purpose, but in completely different area. Many industrial automation systems acquire on-line process data using SCADA (Supervisory Control and Acquisition) systems, that allow for process supervising by the operators. The typical SCADA system workstation block scheme is presented in the fig.1a, where on-line database and historical data repository was distinguished.

Although human stands the “perfect controller” on one hand, he is failure susceptible, because of his physiology on the other hand. Not optimal (even suboptimal in some cases)

controller sets with it's negative process influence refers to final product quality, and increases the total costs.

The solution of that problem could be an advisory system for process operators, that use "faultless operator" behavior from knowledge base already stored in the system (see Fig. 1b). It's primal aim ought to be optimized control set prediction [3].

Crucial functional purposes of considered advisory system:

- storage, processing and integration data from different sources;
- data analysis, ability to implement Business Intelligence (BI) implementation, hierarchical views and several data mining techniques deploy;
- presentation and distribution of achieved results.

Notice, that all above can be solved by the internal services of SQL Server 2005 standard version. Scheme of the proposal advisory system is presented in the Figure 1b, where apart standard elements analysis SQL Server database are localised.

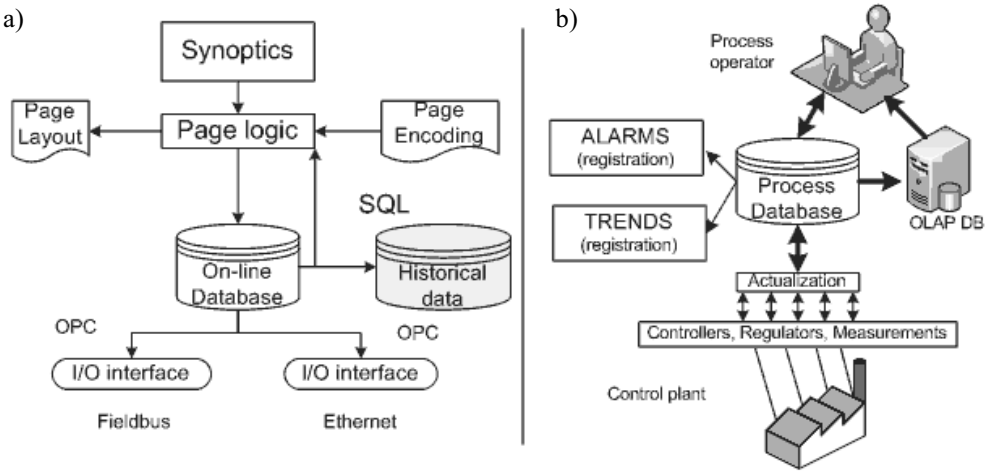

Fig. 1. Process operator workstation block chart (a);
OLAP DB extended SCADA system scheme concept (b)

The information gathered (and later mined) are contained in the warehouse embedded trained data model. Gathering data and making predictions on trained data models is improved by Microsoft's special query language for data mining, called DMX.

This approach requires the additional, SCADA independent platform for real-time efficiency improvement. The crucial question is about the contribution of the "faultless operator" OLAP database and regular process operator in the SCADA system? It seems to be dependent form several factors like: process character, dynamics, sampling frequency and data granulation.

3.2. Brief description of the process

The beet slicer set stands the first chain of sugar production process, which has influence on the whole process as well as economical income of the enterprise. Compensation of variable demand from diffuser and keeping cassets at the appropriate level is the objective task for slicing process.

Beet slicer works under stochastic disturbances such as inhomogeneity of material (beets), contaminations and temperature, so that it's proper work depends on knives waste level as well as drums velocity (which are controlled outputs).

Currently both slicers work under about fifty percent of their abilities, though they are equipped with modern control drives allowing them for point of work rapid changes. It means, that half o their potential is wasted.

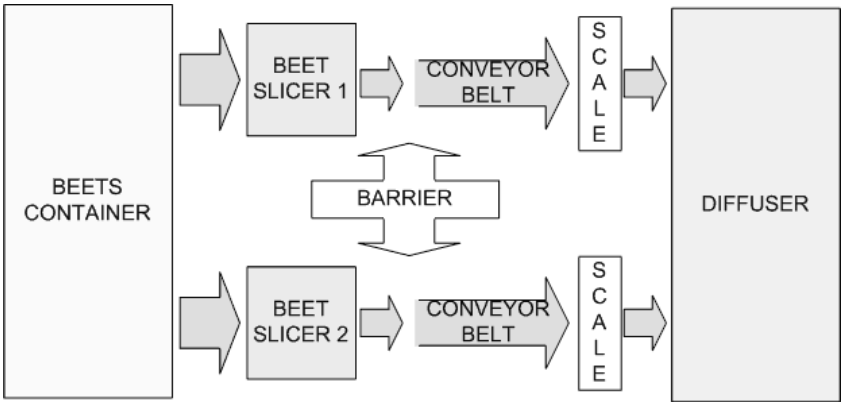

Fig. 2. Block diagram of the slicing process

On one hand the appropriate quality level of cassets depends on drums velocity. But on the other hand stochastic disturbances such as inhomogeneity of material (beets), contaminations and temperature lead to knives blunt as well make productivity lower in the effect or even standstills. The block diagram of the slicing process is presented in the Figure 2.This is a serious problem, and it exists even though there are some preventers like pneumatic knives cleaning system or dirtiness snatch set are installed.

3.3. Creating and training a mining model

The prior action in this stage is as set up a connection to Analysis Services 2005 and create a new mining model. The purpose of the model is to predict the velocity set for beet slicer control based on some sugar production process. The DMX statement for model creation is following:

```
String CreateModel = "Create mining model BSVelocity_Prediction" +
"(SampleID long key," +
"BS_DateTime text discrete, BS_No text discrete," +   //which slicer
"BS_Charge real, BS_Efficiency real," +
"BS_VelocityCV real," +        //measured velocity
"BS_VelocitySP real predict)" +      //setpoint
"Using Microsoft_Decision_Trees";
OleDbCommand CMD = new OleDbCommand(CreateModel, conn);
CMD.ExecuteNonQuery();
```

There is also added the content type in the above statement, after the declaration of the data type for each column. This action inform the algorithm applying to the mining model (in this example, Microsoft Decision Trees) how the data in the columns are distributed. The last line of the `Create` statement uses the `predict` keyword, telling the algorithm that all other columns will predict the outcome of `BS_VelocitySP` column for the model. The appropriate stages of the mining model are presented in the Figure 3.

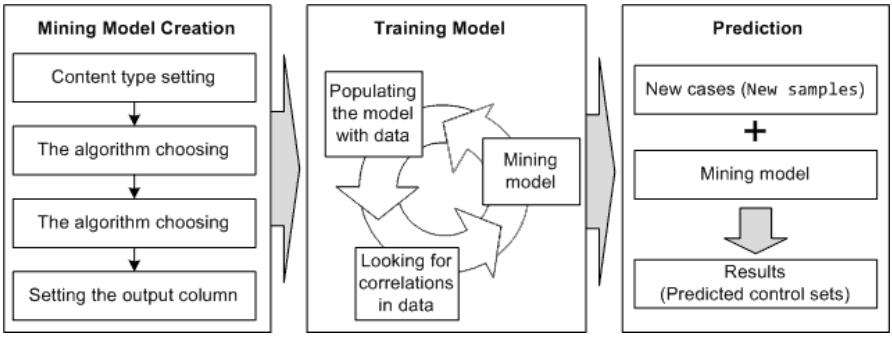

Fig. 3. Three stages of prediction framework preparation

Training the model consists of two stages: the data mining algorithm input cases testing and looking for correlations in the data. After correlations identification, the model is repopulated with these new patterns. Model processing starts over as new data is introduced into the model. This results in more precise predictions, because the patterns are revised over time. To populate the model with data, the DMX Insert statement is used:

```
String PipeData2Model = "INSERT INTO VelocitySP_Prediction"
+ "(SampleID, BS_DateTime, BS_No, BS_Charge,"
+ " BS_Efficiency, BS_VelocityCV,"
+ " BS_VelocitySP) OpenQuery(btsldbsource, 'Select sampleid, datetime, "
+ " bs_no, charge, efficiency, vel_cv, vel_sp FROM btsl')";
OleDbCommand CMD = new OleDbCommand(PipeData2Model, conn);
CMD.ExecuteNonQuery();
```

Above query seems to be roadmap between a table called `Samples` in a SQL Server database, defined by the datasource `btsldbsource`, and the mining model. `OpenQuery` is a DMX function for performing DMX queries against relational databases from inside an OLE-DB session connection. After delivering data to the model, the algorithm may be used to test the cases as well as to identify patterns.

3.4. Control set prediction

Prepared model may be used to predict the best control set of velocity (setpoint), with appropriate efficiency level and current load of the beet slicer. For example, velocity control set is going to be determined with efficiency greater than 65%, efficiency range 40–70%, with a 80% probability or better. DMX `Select` query statement is used to make predictions on the model:

```
String PredictModel = "Select T.SampleID, VelocitySP_Prediction"
+".BS_VelocitySP From VelocitySP_Prediction NaturalPredictionJoin"
+ " OpenQuery(BTSL, 'select * from NewSamples) As T"
+ " Where T.BS_Efficiency > 65 And T.BS_Charge Between 40 And 70"
+ " And PredictProbability(BS_VelocitySP, «60») >0.8";
OleDbCommand CMD = new OleDbCommand(PredictModel, conn);
OleDbDataReader myReader; myReader = CMD.ExecuteReader();
while (myReader.Read()) { //return data } myReader.Close();
```

Considered query introduces new cases to the mining model from BTSL datasource, containing the table `NewSamples`. The DMX function `NaturalPredictionJoin` allows to join the data from the `NewSamples` table and model without any additional specifications, because both tables have the same columns. Function, `PredictProbability` is used in conjunction with the `Where` clause to produce desired results [6, 7].

3.5. Industrial usage assessment

The companies ought to adapt to the market changes very quickly to be competitive, maximizing their profit with simultaneous lowering production costs. The reasonable and cheap solution is optimization of the most important stages of the production process. Possessing appropriate information in proper time is the key element allowing to bring the optimization through. It gives a stable framework for wise and precise decision making.

Although data mining techniques are mainly addressed to IT branch, banking and stock markets, there are many arguments for industrial usage.

Described solution is intended particularly both to the enterprises that trying to keep up the market and also to these with modern processing lines. The local sugar industry may be a very good example, while it is strongly influenced by the French and German consortiums. To remain competitiveness Polish sugar industry may improve thanks to processing lines modernization, but it is very expensive and time consuming solution.

Alternative idea – efficiency improvement by the production important stages optimization, using data mining seems to be necessary.

4. Conclusion

Data mining techniques becomes the key element of the modern business. Although the idea is not new, new technologies and implemented standards make a contribution to their growing popularity. Regarding to mining model usage SQL Server 2005 is a breakthrough in this area. Thanks to the DMX language either programmers or database administrators are able to create Data Mining Systems in simple way.

Although economical and business publications are very fruitful of data mining approaches, the described problem is presented rather weak in the international publications. Nethertheless some industrial appliances of data mining technology were considered in [1].

Industrial usage of data mining techniques opens new possibilities in decision making not only for top level management, but also for advisory or control systems. Several prediction, classification or even anomaly detection algorithms implementation may become lucrative tool for industrial process appropriate stages optimization, that combines diagnosis and control functions.

References

[1] Duebel C., *Application of Data Mining Techniques to Industrial Processes for Improved Business Performance*. APACT Conference, 2003.
[2] Elmasri R., Navathe S., *Wprowadzenie do systemów baz danych*. Addison-Wesley, Helion, 2005.
[3] Findeisen W., *Technika regulacji automatycznej*. PWN 1978.
[4] Garcia-Molina H., Ullman J.D., Widom J., *Systemy baz danych. Pełny wykład*. WNT, 2006.
[5] Jacobson R., Misner S., *Microsoft SQL Server 2005 Analysis Services. Step by step*. Promise, 2005.
[6] Smith J., *Data mining with C# and ADO.NET*. www.devsource.com 2003.
[7] Tang Z., *Data Mining with SQL Server 2005*. John Wiley & Sons, 2005.
[8] Zawadzki M., *SQL Server 2005*. MIKOM, 2005.
[9] Żarski A., *Data mining using SQL Server 2005*. www.codeguru.pl 2006.