Transactions on Machine Learning
and Data Mining
Vol. 5, No. 2 (2012) 63-64
© 2012, ibai-publishing,
ISSN: 1865-6781,
ISBN: 978-3-942952-11-8

Editorial

Carla Brambilla
Consiglio Nazionale delle Ricerche CNR, Istituto di Matematica Applicata e
Tecnologie Informatiche, IMATI, Milano, Italy

The present issue of the International Journal Transaction on Machine Learning and Data Mining contains two papers concerning very interesting applications of data mining methods.

The first paper discusses a general sales forecast model for automobile markets. It is a very relevant topic since strategic planning based on reliable forecast is an essential key ingredient for a successful business management within a market-oriented company, and this is particularly true for the automobile industry, as it is one of the most important sectors in many countries. The model is an additive time series model, which include a calendar component, a seasonal component and a trend component. Estimation of the first two components is treated as an univariate problem and is handled with classical times series tools such as moving averages. On the contrary, estimation of the trend component is treated as a multivariate problem since it is assumed that this component is influenced by exogenous parameters (personal income, unemployment rate, gasoline price and so on), and is performed by data mining methods. Some of the exogenous parameters are correlated, but no dimension reduction is suggested since these methods are usually uneffected by high dimensionality and are indeed very powerful to get insight into internal relationships within large datasets. The data mining methods used include the most successful ones, namely support vector machines, trees and random forests. The markets taken into consideration are the German market and the USA market and the training sales data refer to the periods 1992-2006 and 1970-2005, respectively. The subsequent periods are test periods. Yearly, quarterly and monthly data are considered. The quality of the prediction, as measured by Mean Average Percentage Error (MAPE), is definitely better for quarterly data than for monthly data, as expected since quarterly data are more stable data. However, much better results are obtained for monthly data when using absolute exogenous data instead of a mixture of absolute and relative data. Once more, trees confirm their superiority as far as explicability of the results is concerned. Although the results on the whole are satisfactory, none of the data mining methods applied could predict the special effects occurred in the market after 2008, in particular the 2009 peak of the German market, due to the car-scrap bonus in Germany.

The second paper presents a data mining approach to model pumping systems, in particular to forecast seizings, by means of the so-called First Local Maximum episode-rules (FLM-rules). Since seizings can be provoked by many different causes, it is difficult to establish a preventive maintenance planning, therefore maintenance planning based on faults prediction is very important. FLM-rules are preferred to other episode-based data mining techniques because of their optimal window widths properties. Indeed, when an event is forecast based on these episode-rules, an optimal forecast time interval is automatically computed and it can differ from one rule to another. Forecasting is based on vibratory data, preprocessed in order to derive sequences of types of events along with their occurrence dates. More specifically, at the preprocessing stage for each frequency band the Root Mean Square (RMS) values of the vibration speed are categorized in three classes corresponding to different severity levels of change with respect to a reference value. Each time there is a switch from one level to another, the time spent at the previous discrete level is computed and discretized using four level, ranging from few minutes to more than ten days, and the occurrence time of the switch is recorded. The data to be mined are the sequences of information about switches so obtained. The training data used in the application detailed in the paper are derived from 64 pumping systems of a semi-conductor manufacturer and cover a period of two years. Pumps may undergo several subsequent failures, thus only data recorded before the first seizing are used for learning the rules since learning from potentially degraded systems could bias the model. Results are encouraging; accuracy is quite high both on the training data and on test data (about 98% of seizing was forecast) and there were very few false alarm, which is quite important in a production context. Failures are forecast with a good temporal precision too and it is observed that the more the failure occurrence is close the more the forecast window is precise. The approach presented is now patented.

References

1. Huelsmann, M., Borscheid, D., Friedrich, C. M., Reith,D.: General Sales Forecast Models for Automobile Markets and their Analysis, Transactions on Machine Learning and Data Mining, Vol. 5 (2), 65-86 (2012)
2. Martin, F., Méger, N., Galichet, S., Bécourt, N: Forecasting Failures in a Data Stream Context Application to Vacuum Pumping System Prognosis, Transactions on Machine Learning and Data Minin, Vol. 5 (2), 87-116 (2012)