Available online at www.sciencedirect.com
ScienceDirect
International Journal of Forecasting 25 (2009) 456–460
www.elsevier.com/locate/ijforecast

Discussion

Mining the past to determine the future: Comments

Sven F. Crone

Department of Management Science, Lancaster University, United Kingdom

Abstract

In forecasting, data mining is frequently perceived as a distinct technological discipline without immediate relevance to the challenges of time series prediction. However, Hand (2009) postulates that when the large cross-sectional datasets of data mining and the high-frequency time series of forecasting converge, common problems and opportunities are created for the two disciplines. This commentary attempts to establish the relationship between data mining and forecasting via the dataset properties of aggregate and disaggregate modelling, in order to identify areas where research in data mining may contribute to current forecasting challenges, and vice versa. To forecasting, data mining offers insights on how to handle large, sparse datasets with many binary variables, in feature and instance selection. Furthermore data mining and related disciplines may stimulate research into how to overcome *selectivity bias* using reject inference on observational datasets and, through the use of experimental time series data, how to extend the utility and costs of errors beyond *measuring performance*, and how to find suitable time series benchmarks to evaluate computer intensive *algorithms*. Equally, data mining can profit from forecasting's expertise in handling nonstationary data to counter the *out-of-date-data* problem, and how to develop *empirical evidence* beyond the fine tuning of algorithms, leading to a number of potential synergies and stimulating research in both data mining and forecasting.

1. Introduction

Hand (2009) foresees the creation of new opportunities in predictive modelling when the two areas of *forecasting* based on large masses of data, and using the *tools of data mining* come together. Not only does he thus imply that data mining (DM) and forecasting are complementary, or at least compatible, disciplines, but he also 'forecasts' that a trajectory of common research topics exists that may be extrapolated to meet in the foreseeable future. Possibly not to his surprise, his view was not unanimously received with enthusiasm by the forecasting community, where the derogatory connotation of DM still prevails for many with a rigorous statistical or econometrical upbringing. This view is not only echoed in the commentary by Price (2009), who openly admits that "it was never considered to be good", and recent papers by Armstrong (2006) in this journal, but equally shared by researchers in computer science and machine learning, despite their mutual passion for algorithms associated with all three disciplines. From my

DOI of original article: 10.1016/j.ijforecast.2008.09.004.
E-mail address: s.crone@lancaster.ac.uk.

0169-2070/$ - see front matter © 2009 Published by Elsevier B.V. on behalf of International Institute of Forecasters.
doi:10.1016/j.ijforecast.2009.05.022

personal experience (caught between the domains of forecasting and DM in applying artificial neural networks (NN)), forecasters falsely remain sceptical regarding the value of DM and its relevance to their domain, decrementing DM purely to the application of computer intensive methods such as NN, a particular predictive task such as classification, or a general lack of theory in model building, while in fact it has become much more.

This commentary seeks to first establish an (inherently subjective) link between predictive DM and forecasting via the properties of the underlying data, on which DM is anchored. Once similarities are established in Section 2, we can discuss where forecasting can learn from DM (Section 3) and vice versa (Section 4), regarding the common challenges identified by Hand (2009), and where these opportunities have been missed. Given the limitations of my own experience, I will restrict the discussion to the business domain of forecasting and predictive DM, and omit other prominent areas, such as DNA microarray DM for genome and disease discovery, association rule analysis and text mining.

2. Forecasting versus data mining

Data mining, introduced as 'the science of extracting useful information from large data sets' (Hand, Mannila, & Smyth, 2001), is a relatively new discipline, originating at the interface of statistics, machine learning, pattern recognition and computer science (Hand, 1998). Historically, the notion of finding and predicting useful patterns from data has been a statistical endeavour. As a response to the (possibly premature) claim of DM for this field, a number of survey articles have attempted to distinguish DM from other disciplines, and in particular how DM differs from statistics (Chatfield, 1995; Hand, 1998), and how traditional 'algorithmic' approaches differ from the 'statistical learning' methods employed in DM (Breiman, 2001; Jain, Duin, & Mao, 2000). Chen, Han, and Yu (1996) contrast DM techniques from an informatics and database perspective. However, no attempts have been made to distinguish DM from forecasting or market modelling nor to find synergies between them, despite their close relationship in predictive decision making. Both Hand's introductory and detailed definitions of DM (Hand et al., 2001), and his proposed opportunities in merging DM and forecasting, emphasise the properties of large datasets that constitute DM, rather than particular algorithms or applications of DM. Although Hand has elaborated on this extensively (Hand, 1997, 1998), I will seek to summarise this notion further, to distinguish and reconcile between forecasting and DM.

The characteristics of the datasets define the preference of models and algorithms for DM and forecasting. Datasets of DM often contain millions of records used for predictive modelling at an individual level (e.g. of individual customer accounts), each characterised by dozens of nominal variables translated into hundreds of binary attributes in modelling. The resulting size of the datasets, the number and heterogeneous scale of attributes constitute some of the particular challenges in DM. Predicting an individual's decision (e.g. responding to a direct mailing, defaulting on a loan) relates to a nominal dependent variable of class membership through classification. In contrast, cumulating the binary decisions of many individuals to an aggregate level suggests that forecasting frequently employs a dependent variable of metric scale, and hence regression. Although DM's emphasis remains on classification, as is reflected in the top 10 algorithms in data mining (Wu & Kumar, 2008), neither the scale of the dependent variable nor the algorithm employed accurately discriminate between forecasting and DM tasks (i.e. regression and classification, respectively). Predictions of aggregate demand may be modelled as either regression or classification (e.g. by downscaling a regression of stock market price into a classification of a rise-or-fall prediction). Also, a disaggregate prediction may predict an individual's response using a binary variable of class membership (e.g. of being a 'good' or 'bad' credit risk), the probability of a 'bad' class membership, or the actual profit/loss generated from the credit decision in the form of regression (Finlay, 2008). Furthermore, computer intensive algorithms such as NN are capable of modelling both regression and classification, further limiting a discrimination. Consequently, we will instead employ the level of disaggregation to distinguish between DM and forecasting, as it induces the defining dataset properties.

Given the disparate datasets, and distinct aggregate versus disaggregate modelling, how can DM and

forecasting converge, as Hand has suggested? One obvious transition follows the convergence of the dataset properties. Time series data may be disaggregated, not only with regard to the level of product hierarchies or regions, but through an increase in the frequency at which a time series is sampled in discrete time: from low frequency time series of yearly, quarterly and monthly data to high frequency data of weekly, daily and intraday data (Engle, 2000; Granger, 1998).

With the increasing frequency, additional datapoints are generated for a given history (e.g., a history of three years in order to model yearly seasonality, regardless of whether we are using monthly or hourly data), increasing the size of datasets. More importantly, additional data properties emerge that impact the time series at smaller time intervals, and whose magnitude may be either masked or compensated for on longer intervals. For observations with different forms of seasonality, e.g. in analysing retail sales, a low time frequency of yearly data eradicates the effect of any seasonality. At higher frequencies, a monthly time series may display a single seasonality of month-in-the-year, and daily time series may exhibit further day-of-the-week, day-of-the-month (e.g. pay-day), week-in-the-year and week-in-the-quarter seasonality. Similarly, higher time frequencies may exhibit local time trends, level shifts and outliers that would otherwise not require explicit modelling.

In addition, the effect of an external event (e.g., the calendar effects of Christmas, Easter and bank holidays, marketing activities such as promotions, pay day or extreme weather on retail sales) will hardly be noticeable in quarterly or monthly data, but its relative impact on smaller time intervals of a week, day or hour is more pronounced. As seasonal patterns and the relative effects of events increase with more frequent recording intervals, they require explicit modelling as explanatory (dummy-) variables to capture calendar effects. Modelling these exogenous effects – in addition to lead and lag effects of different forms – extends the dataset characteristics towards large scale datasets of observational data with many variables of a heterogeneous, often binary, scale, not unlike the dataset properties that constitute DM.

The resulting challenges can already be observed in retail forecasting (and electrical load forecasting), where retailers such as Tesco need to forecast the demand for 10,000s of products across thousands of stores every day, resulting in millions of forecasts that need to automatically incorporate calendar events, marketing activity and weather. As an increasing time frequency leads to a convergence of the dataset properties of forecasting and DM, opportunities are thus created for forecasting to use the expertise of DM in modelling large datasets.

3. Data mining lessons for forecasting

Large datasets of high-frequency data pose novel challenges, e.g. in specifying models using statistical tests. Given the abundance of data and the curse of dimensionality, most variables and lags become statistically significant, leading to overparameterised and non-parsimonious models with long computation times. As datasets in DM and forecasting share similar properties as they converge, DM may contribute established approaches in tackling issues of forecasting large, high-frequency datasets. Amongst others, issues of feature selection amongst the many binary (and often multicollinear) explanatory variables, the use of wrappers and filters in model building, learning from imbalanced datasets where interesting features are underrepresented, or different approaches to combine individual models in ensembles to deal with randomness in the data have received substantial attention in DM, machine learning and statistics. Furthermore, it seems plausible that some of the existing problems of DM identified by Hand (2009) may also be encountered by forecasting. Here DM may help to identify lessons learnt on selectivity bias, measuring accuracy and algorithms that may stimulate future research in forecasting. While some of these may even be extended to conventional time series of lower frequency, they require a more through review beyond the scope of this commentary.

4. Forecasting lessons for data mining

Hand (2009) notes that the issues created by large datasets are threefold: (a) searching through the vast datasets; (b) issues of data quality; and (c) apparent structure arising by chance (see also Hand, Blunt, Kelly, & Adams, 2000). However, he neglects to draw explicit attention to one important aspect which is

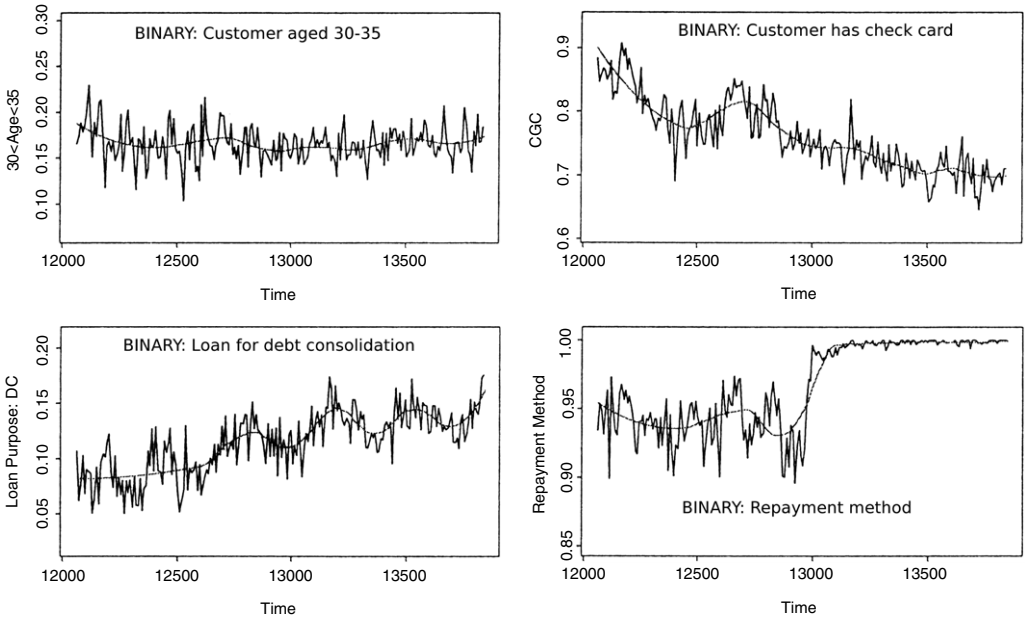

Fig. 1. Weekly averages of binary variables across four years of data (Hand et al., 2000)

at the forefront of interest to forecasters: change, in the form of nonstationarity (population drift in DM) through local and global time trends, seasonality, cyclical changes or heteroscedasticity, reflected in *out-of-date-data*.

The use of *out-of-date data* is typical of most DM research and practice: The DM sector relies largely on algorithms to build non-dynamic models, treating cross sectional data accumulated over time as stationary, tracking performance and rebuilding the models on a regular basis to adjust for nonstationarities (Hand, 2009). Contrary to the assumption of Price (2009), most DM neglects the information contained in changes over time, and even fails to provide time-stamps to the disaggregate transactions, making changing population structures undetectable (Hand, 1998). In credit scoring, e.g., where data is frequently gathered over a period of two years to determine defaults, all instances are presumed to have occurred at the same time, regardless of seasonality or a rising credit crisis, and people that default after 23 months are classed together with those who default after only 4 weeks, despite possibly quite different characteristics and implications for model building.

However, empirical data, on an aggregated level for forecasting or a disaggregated level for DM, is mostly nonstationary: Fig. 1 shows four binary variables describing personal loan applicants, aggregated to weekly averages. The graphs show how some attributes of credit applicants such as customer age remain stationary, while other attributes change with time: using a check card shows a clear downwards trend, repayment method (1d) a level shift and heteroscedasticity due to a policy change of the bank, and the purpose of using a loan for debt consolidation shows an upwards trend with superimposed yearly seasonality (Hand et al., 2000). Inferences made on data collected at one time will have limited applicability later on, requiring dynamic modelling of both dependent and independent variables. Just as the forecasting datasets are gradually enriched with explanatory variables to derive better decisions on time series of higher frequency, the DM datasets of the future will be extended along the time domain, recording the development of explanatory variables of individual customers over time. Already, the popular Recency-Frequency-Monetary-(RFM) approach in direct marketing (Reinartz & Kumar, 2003) aims to capture evolving populations and time dependencies in a

simplified version, and provides further evidence of the importance of explicitly reflecting time. To date, only a few sophisticated models of *dynamic DM* exist to cope with evolving data (e.g., Ganti, Gehrke, & Ramakrishnan, 2001; and Park, Piramuthu, & Shaw, 2001), even in the academic literature. And it is here that DM could draw on the experience of forecasting in modelling nonstationary systems and dataset population drift.

5. Conclusions

With most advances in DM being reported elsewhere, it is not surprising that the merit of DM research has eluded forecasters. In essence, the two disciplines of DM and forecasting face the same challenges of building predictive models in a nonstationary and hierarchical reality, with exogeneous factors and stochastic and chaotic influences that have an impact on disaggregated models of an individual's responses, as well as aggregated models of a product or service. They merely employ different datasets that steer their modelling decisions in different directions. Given the shared objective of predictive analytics and the possible convergence of datasets when (a) DM datasets are extended into the time domain to capture nonstationarities, and/or (b) forecasting datasets are recorded at higher frequencies with additional explanatory variables, potential synergies may arise.

To summarise: DM can learn from forecasting's experience in modelling non-stationary data, while forecasting can draw upon DM's expertise in large and sparse data sets of heterogeneous scales. It is here that the convergence of datasets can promise synergies in learning from the two disciplines.

Acknowledgments

I am grateful to the editor-in-chief Rob Hyndman for the invitation to comment on Prof. Hand's presentation and paper. Many of my arguments draw upon discussions with Robert Fildes as to what constitutes DM, preceding a recent paper on forecasting and operations research (Fildes, Nikolopoulos, Crone, & Syntetos, 2008), and with Scott Armstrong regarding the forecasting principles website. Fortunately we have not reached a consensus, and I am truly thankful for the stimulating discussions.

References

Armstrong, J. S. (2006). Findings from evidence-based forecasting: Methods for reducing forecast error. *International Journal of Forecasting*, *22*(3), 583–598.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199–215.

Chatfield, C. (1995). Model uncertainty, data mining and statistical-inference. *Journal of the Royal Statistical Society, Series A – Statistics in Society*, *158*, 419–466.

Chen, M.-S., Han, J., & Yu, P. S. (1996). Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering*, *8*(6), 866–883.

Engle, R. F. (2000). The econometrics of ultra-high-frequency data. *Econometrica*, *68*(1), 1–22.

Fildes, R., Nikolopoulos, K., Crone, S. F., & Syntetos, A. A. (2008). Forecasting and operational research: A review. *Journal of the Operational Research Society*, *59*(9), 1150–1172.

Finlay, S. M. (2008). Towards profitability: A utility approach to the credit scoring problem. *Journal of the Operational Research Society*, *59*(7), 921–931.

Ganti, V., Gehrke, J., & Ramakrishnan, R. (2001). DEMON: Mining and monitoring evolving data. *IEEE Transactions on Knowledge and Data Engineering*, *13*(1), 50–63.

Granger, C. W. J. (1998). Extracting information from mega-panels and high-frequency data. *Statistica Neerlandica*, *52*(3), 258–272.

Hand, D. J. (1997). Intelligent data analysis: Issues and opportunities. *Advances in Intelligent Data Analysis*, *1280*, 1–14.

Hand, D. J. (1998). Data mining: Statistics and more? *American Statistician*, *52*(2), 112–118.

Hand, D. J. (2009). Mining the past to determine the future: Problems and possibilities. *International Journal of Forecasting*, *25*(3), 441–451.

Hand, D. J., Blunt, G., Kelly, M. G., & Adams, N. M. (2000). Data mining for fun and profit. *Statistical Science*, *15*(2), 111–126.

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. Cambridge, Mass: MIT Press.

Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(1), 4–37.

Park, S. C., Piramuthu, S., & Shaw, M. J. (2001). Dynamic rule refinement in knowledge-based data mining systems. *Decision Support Systems*, *31*(2), 205–222.

Price, S. (2009). Comments on "Mining the past to determine the future: Problems and possibilities" by David J. Hand. *International Journal of Forecasting*, *25*(3), 452–455.

Reinartz, W. J., & Kumar, V. (2003). The impact of customer relationship characteristics on profitable lifetime duration. *Journal of Marketing*, *67*(1), 77–99.

Wu, X. D., & Kumar, V. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*(1), 1–37.