A Survey on Data Mining Techniques for BI

Santhosh Pawar
Department of CS
KSWU Vijayapura

Abstract-*Here we discuss a data mining approaches and tools used for the business intelligence and inventory analysis. It describes the use of data mining approaches in business intelligence (BI), which coupled with data warehouse to employ data mining technology to provide accurate and up-to-date information for effective decision support system in business. The list of methodologies used in the literature is analyzed. This will help to provide out-of-stock forecasts at the store/product level. The inventory management and supply chain management using data mining techniques will improve the business by providing effective demand analysis, demand forecasting and appropriate decision support for the business. This survey concludes the further improvements which can improve the accuracy in demand analysis and forecasting in the inventory and supply chain management system. This also provides the challenges of those processes in terms of data size and uncertain business environments.*

Keywords-Inventory Management, Data Mining, Big Data, Neural Networks, Hybrid Intelligent System,..

I. INTRODUCTION

Business intelligence (BI) is the collection of software applications, which is used to find the intelligent decision support for business from the raw data of the organization. This will lead to the successful business. The usage of data mining and big data, business intelligence became very successful [1]. Due to this vast data size, the analysis should need data mining support. Business intelligence consists of different types like supply chain management, inventory management, sales and marketing analytics, finance analysis etc. The evaluation of BI in such process is enhanced since 1970. The recent trend in BI is using big data, social and mobile cloud analytics [2]. Due to the technology advancement, the tools and functionalities are upgraded. This type of tools helps to unlock the value of structured and unstructured data and to driven innovation, competition and productivity in the organization. The latest tools are integrated into the cloud platform to perform latest up to date functionalitiesin BI.

Table 1.0 shows the evaluation of BI tools and its functionality.

Business Intelligence tools functionality	**USE**
Mainframe based reporting tools/4GL	Traditional decision support system
RDBM's based reporting and analytics	Management/Execution decision support system
Rise of OLAP/Data warehouse, Business analytics and BI tools	Historical data analysis comparison and decision support
Enterprise performance management and predictive analytics	Measuring Key performance indicators of the organization and to predict about future
Big data, social, mobile & cloud based analytics	To unlock the value of structured and unstructured data and to drive innovation, competition, and productivity in the organization

The business intelligence process consists of several functionalities, for every process data mining techniques were used. In this paper, we reviewed various techniques and algorithms used in the business intelligence with its challenges

II. LITERATURE REVIEW

INVENTORY DEMAND ANALYSIS AND DEMAND FORECASTING:

Forecasting product stock based on the market condition is a challenging task due to the complex nature of the data. The massive automatic data collection is the recent trend, now a day's huge volume data are stored in the data repository. This is performed systematically obtaining many measurements, not knowing which ones will be relevant to the phenomenon of interest [3]. Authors handled the high dimensional data by using traditional statistical methodology, assuming many observations and a few well-chosen variables are not designed to cope with this kind of explosive growth of

dimensionality of the observation vector. The increasing availability of data is thus creating new challenges for the market modeler. There are, essentially, three different approaches to address this problem. The first approach is concerned with finding the most influential subset of predictors; the second approach builds predictive models based on summaries of the predictor variables and the third approach is penalized (L-1) likelihood method which automatically selects influential variables via continuous shrinkage. Authors in [4] proposed a best subset selection, and it is a popular class of the dimension reduction methods concerned with finding the most influential subset of predictors in predictive modeling from a much larger set of potential predictors. The best subset problem belongs to the class of NP-hard problems known as induction of minimal structure. When the number of potential predictors is large, the selection process cannot be solved exactly within an acceptable amount of computation time. Consequently, heuristic optimization algorithms have evolved, including iterative improvement algorithms such as stepwise regression, forward and backward feature selection algorithms and stochastic search methods. The stochastic search methods like Genetic algorithms [5], simulated annealing [6], to solve larger scale combinatorial problems. However, the expensive computational cost still makes best subset selection procedures infeasible for high-dimensional data analysis. The information summary approach to forecasting with high dimensional data is based on the assumption that the relevant information is captured by a small number of factors common to the predictor variables. A popular technique is proposed in [7], which combines the potentially relevant predictors into new predictors is principal components. For example, basing the forecast model on data summaries in the form of principal components allows information from all the predictors to enter into the forecasts. In literature all find that diffusion factors based forecasts have smaller mean-squared errors than forecasts based upon simple auto-regressions and more elaborate structural models. A criticism of factor augmented regressions is that the factors are estimated without taking into account the dependent variable. Thus, when only a few factors are retained to represent the variations of whole explanatory variable space, they might not have any predictive power for the dependent variable whereas the discarded factors might be useful. Later in paper [8], authors used penalized L-1 likelihood methods have been successfully developed to cope with high dimensionality. Penalized L-1 regression is called LASSO in the ordinary regression setting which has received much attention due to its convexity and encouraging sparsity solutions. It minimizes the usual sum of squared errors, with a bound on the sum of the absolute values of the coefficients. There has been much work in recent years, applying and generalizing the LASSO and L1-like penalties to a variety of problems [9]. Authors in [10] proposed a fast and efficient least angle regression (LARS) algorithm for variable selection, a simple modification of which produces the entire LASSO solution path. A linear combination of L-1 and L-2 penalties is called elastic net, which encourages some grouping effects. Authors in [11] introduced an adaptive lasso in a finite parameter setting. L-1 type regularization does not eliminate the conflict between consistent model selection and prediction. With high dimensionality, important predictors can be highly correlated with some unimportant ones, and the maximum spurious correlation also grows with dimensionality. But LASSO tends to arbitrarily select only one variable among a group of predictors with high pair wise correlations. This may results in some unimportant predictors that are highly correlated with the important predictors being selected by LASSO while important predictors are missed. Though existing research has provided evidence that promotions of one product can influence the sales of another because of both intra- and inter-category effects, most of the existing literature has focused on developing explanatory models, using a set of features assumed product relationships to test the significance of the cross brand/category promotional affects. These theoretical findings can be applied in a real forecasting system to help retailers improve the decision accuracy at inventory level. This is a very different problem than those only concerned with explanation and hypothesis testing. When we build forecasting models for tens thousands of SKUs in a store, a problem size many retailers face, most of these existing theoretical models lose their feasibility. Even basic least square regression will not be applicable because the dimensionality of cross category promotion explanatory variables is potentially much larger than the sample size. In practice, it also cannot easily identify which product complements/substitutes another. The basic inventory and sales forecasting methods are univariate forecasting models which are based on time series techniques that analyze past sales history in order to extract a demand pattern that is projected into the future [12]. Authors in [13] found that the simple time series techniques perform well for periods without promotions. However, for periods with promotions, models with more inputs improve accuracy substantially. Therefore, univariate forecasting methods are usually adopted as a benchmark model in many. In order to improve inventory sales forecasting in the presence of promotions, many studies have integrated the focal product's promotional variables into their forecasting models.

In practice, many retailers use a base-times-lift approach to forecast product sales at the inventory level. The approach is a two-step procedure which initially generates a baseline forecast from a simple time series models and then makes adjustments for any incoming promotional events. The

adjustments are estimated based on the lift effect of the most recent price reduction and/or promotion, and also the judgments made by brand managers. These judgmental adjustments are common in practice, expensive and potentially prone to systematic errors. Studies have shown that statistical models usually performed better than the expert adjustments. In the recent literature, authors in [14] focus on how to make the adjustment more effectively, while others discuss how to integrate statistical forecasts and managers‟ judgment. Another stream of studies uses a model-based forecasting system to forecast product sales by directly taking into account the promotional information. These methods are usually based on multiple regression models or data mining technologies whose exogenous inputs correspond to the focus product's own promotion features. For example, in literature a promotion-event forecasting system called PromoCast is reported, which uses a static cross-sectional regression analysis of inventory-store sales under a variety of promotion conditions, with store and chain specific historical performance information. The limitation of these studies is they do not overlook the potential importance of price reductions and promotions of other influential products, nor do they include time series dynamics Forecasting product sales integrating influential products‟ promotional information has also been explored by previous researchers. The key similarities of this study are that both studies aim to improve the forecasting accuracy for retailers at inventory level by integrating extra promotional information from other products. At the same time, there are some important differences. First, this paper considers both intra- and inter-category promotional interactions, while only considered intra-category competition. Second, authors used a ―general to specific‖ approach to manually select explanatory variables for every inventory one by one. Though theoretically showing that integrating intra-category competitive information could improve inventory forecasting accuracy, the approach is in fact inapplicable: in a real grocery forecasting system, it is impossible to manually manipulate individual forecasting models for tens thousands of items in a store. Instead, we propose to use a multi-stage variable selection and model estimation strategy based on LASSO regression; the total process is fully automatic and therefore can be easily integrated into a forecasting system. Third, they pooled the inventory sales from 83 stores to simulate a chain level forecasting situation. This does not help a chain manager allocate inventory stocks at the store level, because of the heterogeneity among stores. Furthermore, the price and promotional indexes are both aggregated across multiple stores; this may weaken the explanatory power of these variables. In our research, we focus on store level sales forecasting, using the raw inventory level information to build a forecasting model without any aggregation. This is the forecasting situation directly links to a chain or store manager's weekly stocking allocation decisions. But this is a more challenging problem, for the data at the disaggregate level contains more noise than at the aggregate level. Fourth, authors considered inventory prediction from 6 categories in their empirical study. It is a large scale empirical study compared to previous existing researches; most of them usually consider only tens of items in empirical study. This research empirically examines the forecasts on different categories of product out of sample forecasting. At such a scale, we need to weigh the complexity of the model and the corresponding computing efficiency. Therefore, our results will be more realistic, robust and useful in inventory level decisions. ANNs are popular artificial intelligence-based data mining tools due to their superior prediction performance. For example, ANNs are widely used in forecasting many applications like ATM cash demand, wind speed, inventory level etc., Intelligent inventory management system:

III. RECENT INVENTORY FORECASTING TECHNIQUES

Inventory and Demand forecasting are key business functions for retailers and manufacturers. Demand forecasting helps retailers to identify underachieving stores where the potential sales of a product appear to exceed the actual sales of the product. Product allocations assist manufacturers to allocate products to stores and accounts. The main aim of inventory forecasting is to minimize the inventory loading problem. A common practice of inventory forecasting is to predict the demand for a particular item in the future and reserve the appropriate amount of items, based on the forecasting results. However, inventory data is a type of time series with large volume, long time span, and less regularity. These features bring up two challenges to inventory forecasting: 1) Implementing an accurate interpretable inventory prediction; 2) Modeling the relationships among multiple time series data sets and predicting their future values simultaneously. Although, in recent years, there has been an explosion of interest in mining time series, traditional approaches such as auto-regression (AR), linear dynamical systems (LDS), Kalman filter (KF) cannot solve above challenge directly. Hence, more accurate and optimal forecasting methods are expected. In paper [15], authors proposed iMiner , which is a new inventory forecasting models such as dynamic prediction model and joint prediction model to solve the prediction problems.

IV. INVENTORY ANOMALY DETECTION

Anomaly detection in time series is one of the fundamental issues in data mining that addresses various problems in different domains such as intrusion detection in

computer networks, irregularity detection in healthcare sensory data and fraud detection in insurance or securities. Risk management in business is a major concern. Especially in the current scenario, demand analysis, inventory management and effective decision support system is more important. In inventory management, the anomaly should be detected at the right time. It should give a timely message or alert about the abnormal inventory data to the business administrator. In the paper [16], authors developed a computational intelligence approach for price exploitation detection. The anomaly detection on product prices is identified using hidden markov models. Authors compared the adaptive hidden Markov model with anomaly states (AHMMAS) method with the existing KNN and SVM. Authors in [17] proposed a Contextual Anomaly Detection (CAD) method for complex time series that is applicable for identifying stock market anomalies. The method considers not only the context of a time series in a time window but also the context of similar time series in a group of similar time series. Authors designed and implemented a comprehensive set of experiments to evaluate the proposed method on different sectors with daily and weekly frequencies. The results indicate that the CAD method outperforms kNN and Random Walk in identifying anomalies in time series grouped by sectors. Although many anomalies have been established (relatively high recall), our method still flags false positives (low precision).

V. CONCLUSION

This paper gives a summary about the techniques and methods used for the inventory management. The inventory management with high dimensional data flow and uncertain flow is a challenging one. After the literature analysis, this paper suggests different ideas for future work, so this summarizes that as follows:

i. Application of effective data mining techniques can improve the inventory level forecasting and business intelligence
ii. The development of a novel fully-automatic algorithm that is capable of selecting key explanatory variables from a very large data set.
iii. The focus on retail store level modeling and forecasting at inventory level for thousands of products in order to capture dynamic promotional effects.
iv. The inclusion of both intra- and inter-category information.
v. The examination of comparative results for a large number of inventory levels forecasting over a large number of categories.

This study concludes with Data Mining and big data are the most useful functional compilation for Business Intelligence and inventory managements. The implementation of effective data mining with business decision support will lead the current business scenario in a top level.

REFERENCES

[1] Chen, Hsinchun, Roger HL Chiang, and Veda C. Storey. "Business intelligence and analytics: From big data to big impact." MIS quarterly 36.4 (2012).
[2] Tan, Wei, M. Brian Blake, Iman Saleh, and Schahram Dustdar. "Social-network-sourced big data analytics." IEEE Internet Computing 17, no. 5 (2013): 62-69.
[3] Donoho, David L. "High-dimensional data analysis: The curses and blessings of dimensionality." AMS Math Challenges Lecture 1 (2000): 32.
[4] John, George H., Ron Kohavi, and Karl Pfleger. "Irrelevant features and the subset selection problem." Machine learning: proceedings of the eleventh international conference. 1994.
[5] Melab, Nordine, Sébastien Cahon, El-Ghazali Talbi, and Ludovic Duponchel. "Parallel GA-Based Wrapper Feature Selection for Spectroscopic Data Mining." In ipdps. 2002.
[6] Meiri, Ronen, and Jacob Zahavi. "Using simulated annealing to optimize the feature selection problem in marketing applications." European Journal of Operational Research 171.3 (2006): 842-858.
[7] Stock, James H., and Mark W. Watson. "Forecasting using principal components from a large number of predictors." Journal of the American statistical association 97.460 (2002): 1167-1179.
[8] Tibshirani, Robert. "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society. Series B (Methodological) (1996): 267-288.
[9] Tibshirani, Robert. "Regression shrinkage and selection via the lasso: a retrospective." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73.3 (2011): 273-282.
[10] Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004). Least angle regression. Annals of Statistics, 32(2), 407- 451 .
[11] Cao, Feng, et al. "Density-based clustering over an evolving data stream with noise." Proceedings of the 2006 SIAM international conference on data mining. Society for Industrial and Applied Mathematics, 2006.
[12] Raju, Puthankurissi S., Subhash C. Lonial, and W. Glynn Mangold. "Differential effects of subjective knowledge, objective knowledge, and usage experience on decision making: An exploratory investigation." Journal of consumer psychology 4.2 (1995): 153-180.

[13] Ali, Özden Gür, Serpil Sayın, Tom Van Woensel, and Jan Fransoo. "inventory demand forecasting in the presence of promotions." Expert Systems with Applications 36, no. 10 (2009): 12340-12348.

[14] Nikolopoulos, A., D. Papafotiou, N. Nikolopoulos, P. Grammelis, and E. Kakaras. "An advanced EMMS scheme for the prediction of drag coefficient under a 1.2 MW th CFBC isothermal flow—Part I: Numerical formulation." Chemical Engineering Science 65, no. 13 (2010): 4080-4088.

[15] Zhou, Qifeng, Bin Xia, Wei Xue, Chunqiu Zeng, Ruyuan Han, and Tao Li. "An Advanced Inventory Data Mining System for Business Intelligence."

[16] Cao, Yi, Yuhua Li, Sonya Coleman, Ammar Belatreche, and Thomas Martin McGinnity. "Adaptive hidden Markov model with anomaly states for price manipulation detection." IEEE transactions on neural networks and learning systems 26, no. 2 (2015): 318-330.

[17] Golmohammadi, Koosha, and Osmar R. Zaiane. "Time series contextual anomaly detection for detecting market manipulation in stock market." Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on. IEEE, 2015