Proceedings of the 2012 International Conference on Industrial Engineering and Operations Management
Istanbul, Turkey, July 3 – 6, 2012

Application of Data Mining Techniques in Drug Consumption Forecasting to Help Pharmaceutical Industry Production Planning

Rouzbeh Ghousi, Saharnaz Mehrani and Morteza Momeni
Department of Industrial Engineering
Iran University of Science and Technology, Narmak
Tehran 16846-13114, Iran

Sina Anjomshoaa
Graduate School of Management and Economics
Sharif University of Technology, Azadi
Tehran 11365-11155, Iran

Abstract

No doubt that production planning and inventory control in an erratic and dynamic drug consumption environment would be considered as a fundamental need of pharmaceutical industries to enhance the commercial competitive advantage. There are many hidden factors that can affect drug consumption. One way for forecasting the medicine consumption and production planning is identification of these latent and effective factors. By means of data mining techniques, hidden relationships between the variables can be identified. Meanwhile, effective variables can be utilized to predict the dependent ones. Therefore, data mining techniques such as a combination of association rule and prediction algorithms have been found helpful in forecasting. However, using data mining methods requires access to historical report of drug usage, along with purchasing features and characteristics of the buyers. The purpose of this study provides a convenient and rapid method to predict different drugs consumption. Based on the purpose, a comparison on the predictive algorithms is performed to result in one algorithm with higher accuracy. Furthermore, the efficiency of prediction algorithms, including regression, artificial neural networks, and decision trees is compared. In this investigation, the empirical study is on a sample 3-year dataset consists of a combination of drug distribution and customers demography information.

Keywords
Pharmaceutical Industry, Data Mining, Forecasting, Production Planning.

1. Introduction
Pharmaceutical industry is one of the most critical level of pharmaceutical supply chain which has a noticeable impact on every society's hygiene and treatment part. In addition, Production planning is one of the main operations of Pharmaceutical industries. With regard to the uncertainty space, forecasting method could have a significant role in decision making as it was also mentioned by Armstrong in 2001 [1]. As it was illustrated in some investigation [2-4], forecasting methods are employed in order to improve decisions related to production planning. Also, Demand forecasting influences many functional areas within an organization such as revenue planning, production planning, resource allocation, and etc [5]. Depending on the produced product and the organization stakeholders, it may be focused on one of the named functional areas more than the others.

Pharmaceutical industries have sizable data sets that have not been used to make any information that could help decision makers. Data mining methods are used to analyze large observational data sets, find unsuspected relationships [6], and discover patterns and trends [7].

Drug demand forecasting in Pharmaceutical domain is a complex task with respect to several effective involved factors [8]. There are many factors that might be affective in consumption amount of a special kind of drug, including: 1) Pharmacologic category and applications of various kinds of medicine, 2) Seasonal variations which imply the very high effect of seasonality on certain drugs, 3) Geographic diversity of the customers with respect to

the therapy used by doctors, distribution of specialist doctors, patients' characteristics depending on the geographic and also cultural attributes, 4) Being a new product, and 5) Drug price. All these different concepts have in common that they recognize that users can play an important role in innovation processes and more than that, users can actually be as source of innovation [9].

2. Data mining

Data mining as an interdisciplinary field which gets together the techniques from machine learning, pattern recognition, statistics, databases, and visualization [10], is used to analyze large observational data sets, find unsuspected relationships [6], and discover patterns and trends [7]. Due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge [11], data mining has attracted a great deal of attention in various domains. In spite of various utilizations of data mining techniques in previous investigations, there are many domains which still have not used data mining methods. The most common tasks, which data mining can accomplish, are as follows: description, estimation, prediction, classification, clustering, and association [7]. Depending on the target of investigation which contains one or a compound of the named tasks, data mining has several defined algorithms and methods.

2.1 Classification and Prediction

Classification is the process of finding a model that describes and distinguishes data classes or concepts. Due to the similarities between prediction and classification techniques, classification methods can also be used in prediction tasks [7]. Therefore, Prediction techniques contain traditional statistical methods of estimation such as simple linear regression and correlation, and multiple regression, in addition to data mining methods such as artificial neural network (ANN), decision tree, and k-nearest neighbor methods for classification. In a study, which was done by Kim [12], the performance of some the named methods were compared depending on the problem situation around the number and types of independent variables and the sample size, and finally it was found that data mining techniques such as ANN performance improve faster than that of statistical prediction methods. Moreover, some studies around forecasting method comparison were done in some special cases [13-15]. For instance, Chang [13] compared ANN algorithm method with two others, including decision trees and a hybrid model of ANN and decision trees, for forecasting stock prices. Based on the average accuracy of those three methods, the study discovered ANN as a more stable method in forecasting the stock prices, which are unpredictable because of the many intervening factors. In another study, Londhe et al. [14], compared data-driven modeling techniques for river flow forecasting, containing artificial neural networks (ANN), genetic programming (GP) and model trees (MT), and the results showed that all the techniques performed well, but GP performed a little better than both two others in accuracy. In a study by Kim et al. [15], there was an assessment on the efficiency of data mining techniques including: ANN, support vector machine (SVM), and DT in comparison with the conventional logistic regression (LR) statistical model, and the conclusion illustrated that the data mining techniques outperform the conventional statistical methods, and also the accuracy of DT is higher than the others in this case.

2.1.1 Artificial neural network (ANN)

Artificial neural networks, or ANN, as a technique that simulates the learning process of biological neural networks has become a popular prediction tool. ANN is a structure of many connected neurons which are arranged in layers in systematic ways. The connections between neurons have weights associated with them depending on the amount of influence one neuron has on another. There are some advantages in using neural networks in some special problem situations. For instance, due to containing many neurons and also the assigned weight to each connection, artificial neural networks is quite robust with respect to noisy and erroneous data sets [7]. In this study, relying on the huge amount of drug consumption data in addition to the uninformative examples in the data set, ANN was found efficient in forecasting and model simulation of drug usage.

Due to the ability of ANN in estimating nonlinear functions and capturing complex relationships in data sets, there were some of its applications in medical domain especially for diagnosing the cancers [14, 16, 17, 18]. One of those common usage was for breast cancer, which Setiono et al.[18] after extensive experiments with back propagation ANN (BPN) over its Wisconsin dataset, extracted rules with accuracy of 94% on the test data. Cynthia Taylor [19] found multi-layer perceptron type of ANN as a valuable asset in freeway traffic data prediction. This application was due to ANN ability in handling nonlinear systems with respect to highly dynamic traffic data. In a study which was for predicting stock price direction, D.Senol et al. [20] illustrated the efficiency of ANN, despite the beliefs around unpredictable behavior of stock price direction. Gjertson et al. [21] presented a comparative study of two

methods including an ANN and a scoring nomogram calibrated from Cox regression coefficients in case of predicting the fate of clinical transplants, which finally, ANN exhibited better performance characteristics than nomogram.

2.1.2 Decision tree

Decision tree classification method makes a tree structure from the data set. In each node of the tree, an attribute test is performed to separate the data into the subsets. The records within each branch (each subset of the node) are more homogenous than in the previous one [7]. Finally, terminal nodes hold the class labels. One of the main reasons of applying model trees in comparison with some methods such as ANN is their understandable rules representation [25]. Classification and regression tree (C&RT) is a one kind of popular decision tree algorithms, which also have been used in this study. One of the advantages that makes it different from some other decision tree methods such as C5.0, QUEST, CHAID and ID3 is its binary splitting [22]. By partitioning the data into two subsets in each node, C&RT method allows for easy interpretation and analysis [23].

3. Experimental Data analysis and results

3.1 Business understanding

Drug demand forecasting in pharmaceutical domain is investigated in this study. For this prediction, it is necessary to have historical records of the drug consumption. Due to the accumulated large quantities of data in drug distribution centers, their databases would be so efficient in this kind of prediction. This empirical study is on a sample 3-year data set with 407000 records of a drug distribution center of Tehran capital of Iran, which contained the drug distribution information of Tehran eastern areas. Moreover, this database contains various types of drugs with their purchase information.

3.2 Data understanding

Variable set applied as independent variables in prediction model building are shown in table 1. The independent variables are different in forecasting a drug usage, due to the disease it is used for. Thus, some of the mentioned variables were selected for forecasting the consumption of each drug, separately. Having more correlations with the amount of drug consumption, and fewer with each other, the variables were selected. In addition to that, Appriori, as a data mining association rule algorithm, was employed to uncover variables having relationship with the amount of drug consumption with more extreme confidence index amount.

Table 1: Independent variables for drug consumption amount prediction

Variable name	Description
Drug purchasing specifications:	
Year	Year extracted from the drug purchasing date.
Season	Season extracted from the drug purchasing date.
Month	Month extracted from the drug purchasing date.
Customer ID	Indicates the pharmacy, clinic, or hospital, which purchased the drug.
Company ID	Indicates the drug production company.
Disease	Indicates the disease the drug is used for.
Price	Price of each unit of the drug.
Area num	The area customer located in.
Demographic specifications of the consumption area:	
Population	Population of the area in the specified year.
Men ratio	Proportion of the number of men to the population in the year of purchasing the drug.
Educated women	Proportion of the educated women to the population in the year of purchasing the drug.
Educated men	Proportion of the educated men to the population in the year of purchasing the drug.
Marriage ratio	Proportion of the number of marriages to the population in the year of purchasing the drug.
Divorce ratio	Proportion of the number of divorced marriages to the population in the year of purchasing the drug.

3.3 Data preparation

Data preparation phase included two steps, purifying data to eliminate data set errors and merging data in order to reduce their volume. In purifying the data, according to the manual data entrance process, there is a probability of data error. In this data set, some errors were found and corrected, such as mistaken date data format, and some missed pharmacies and medical centers specifications. Also in order to eliminate abnormal data of “total sales” variable, the k-means clustering algorithm based on this variable is applied to the data to have close value data in similar cluster. Afterwards in each cluster, the records of any data with numerical values of the variable "Total sales" more than slope from the mean value of the cluster, have been deleted. In clustering, the criteria to select the optimal number of clusters, is the minimum clusters sum of squares error (SSE) criterion.

Because of large volume of data and also high dispersion of the numerical variables, it was needed to merge them. In order to reduce data volume, values of “total sales” for the records with the same values of "Medicines", "year", "month" and "manufacturer", have been summed. As a result, records were reduced from 226354 to 6248. The new variable called "drug consumption", is the consumption per person in each state.

3.4 Data analysis

The data mining techniques employed in this section are ANN, and C&RT. Before applying the mentioned algorithms, each example of the data set was randomly partitioned into two subsets [12]. One set that contained 70% of the records was for training and generating the model while the remainder was for testing the generated model. Depending on numeric nature of the prediction field, root-mean-square deviation (RMSE) and normalized RMSE (NRMSE) considered as the comparison scales. See equations (1) and (2) below for RMSE and NRMSE.

$$(1) \qquad RMSE = \sqrt{\frac{\sum_R (t_r - o_r)^2}{n_r}}$$

$$(2) \quad NRMSE = \frac{RMSE}{\max(t_r) - \min(t_r)}$$

Where R is the set of records, t_r is the target output for the record r, o_r is the estimated amount by the model, and n_r is the number of the records.

The ANN applied in this examination was a multilayer feed-forward network, which also known as perceptron. Also, the method employed for revising the network connection weights was back-propagation algorithm. The network contained one hidden layer while the number of neurons in each has been altered through the examination till finding the optimum, which led to higher accuracy and the best ANN structure. There should be a balance between prevention of overfitting and the model accuracy [7]. By more number of neurons in each hidden layer, the risk of overtraining will increase, in addition to the network efficiency in identifying the complex patterns. The transfer function in each connection between neurons is sigmoid function. The learning rating and momentum were set to 0.3 and 0.9, respectively. The mentioned amount for learning rate and momentum were adjusted after some experimentation with various values [7].

For DT analysis, first of all, the tree diagram of each drug was drawn. To avoiding the overly complex model, the maximum number of tree levels considered five. Due to the numeric nature of the target field, least squared deviation (LSD) was used to measure the impurity of the model and find the best split in each node. After generating the tree, to prevent from over-fitting and improve understandability, it was needed to prune the branches that do not contribute significantly to the model accuracy [24]. In pruning the tree, the risk of misclassification was tried to be close to that of before pruning.

It is needed to say that in this study, all the mentioned algorithms were performed by “Weka miner” software.

3.5 Analysis results

The results of monthly drug consumption amount forecasting are shown in table 2. The results are only for drugs of more important diseases and for those with increasing consumption trend. From NRMSE results and their low amounts for both ANN and DT methods, it can be understood that they are stable for forecasting drug consumption with those mentioned variables. As it is illustrated by RMSE scale, in most cases, DT was superior.

Table 2: Monthly drug usage amount forecasting results

Drug	Disease	NRMSE		RMSE	
		ANN	DT	ANN	DT
Ferrous Sulfate	Anemia	0.01933	0.00887	29841.92	13693.59
CO-Trimoxazole	Digestive putrefaction	0.00895	0.00820	232.58	213.09
Amoxicillin	Antibiotic	0.01691	0.01066	2652.61	1671.74
Clotrimazole	Antibiotic (for women)	0.01997	0.02088	559.03	584.66
Aminophylline	Asthma	0.01083	0.00968	0.1300	0.1161
Rifampin	Tuberculosis	0.12080	0.02452	271.33	264.85

4. Conclusions

In this paper, the main purpose is to diagnose the appropriate method for forecasting the amount of different drugs monthly consumption. With respect to existence variables, data mining techniques are more efficient than statistical ones. For achieving this purpose, employed methods were ANN and DT algorithms. After an empirical study, it was discovered that both ANN and DT are stable methods in this kind of forecasting situation with noisy and erroneous drug usage data and involving several effective factors, while DT would perform a little better than ANN in most of these cases.

References

1. Armstrong, J. S., 2001, Principles of Forecasting: A Handbook for Researchers and Practitioners, Kluwer, Boston.
2. Moskowitz, H., 1972, "The value of information in aggregate production planning – A behavioral experiment," AIIE Transactions, 4, 290–297.
3. Moskowitz, H., Miller, J. G., 1975, "Information and decision systems for production planning," Management Science, 22, 359–371.
4. Goodwin, P., 2005, "Providing support for decisions based on time series information under conditions of asymmetric loss," European Journal of Operational Research, 163, 388–402.
5. Arthur G. C., 2006, Forecasting for the Pharmaceutical Industry: Models for New Product And In-market Forecasting And How to Use Them, first Edition.
6. Hand, D., Mannila, H., and Smyth, P., 2001, Principles of Data Mining, MIT Press, Cambridge, MA.
7. Larose, T. D., 2005, DISCOVERING KNOWLEDGE IN DATA, Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
8. Ruud, E. H. M., Smits and Wouter P.C. B., 2008, "The role of users in innovation in the pharmaceutical industry," Drug Discovery Today, Volume 13, Numbers 7/8 April.
9. www.decisioncraft.com
10. Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., and Zanasi, A., 1998, Discovering Data Mining: From Concept to Implementation, Prentice Hall, Upper Saddle River, NJ.
11. Han, J., Kamber, M., 2006, Data Mining: Concepts and Techniques, 2rd edition, Elsevier Inc, Morgan Kaufmann Publishers.
12. Kim, Y. S., 2008, "Comparison of the decision tree, artificial neural network, and linear regression methods based on the number and types of independent variables and sample size," Expert Systems with Applications, 34, 1227–1234.
13. Chang, T. S., 2011, "A comparative study of artificial neural networks, and decision trees for digital game content stocks price prediction," Expert Systems with Applications, 38, 14846–14851.
14. Londhe, Sh., Charhate, Sh., 2010, "Comparison of data-driven modelling techniques for river flow forecasting," Hydrological Sciences Journal, 55(7), 1163–1174.
15. Kim, S., Kim, W., Park, R. W, 2011, "A Comparison of Intensive Care Unit Mortality Prediction Models through the Use of Data Mining Techniques," Healthcare Informatics Research, December, 17(4), 232-243.
16. Floyd, J. C. R., Lo, J. Y., Yun, A. J., Sullivan, D. C., and Kornguth, P. J., 1994, "Prediction of Breast Cancer Malignancy Using an Artificial Neural Network," Cancer, 74 (1994) 2944–2998.
17. Wilding, P., Morganb, M. A., Grygotisa, A. E., Shoffnera, M. A., Rosatoc, E. F., 1994, "Application of backpropagation neural networks to diagnosis of breast and ovarian cancer," Computer applications for early detection and staging of cancer, (15 March 1994), 77, 145–153.

18. Setiono, R., Huan, L., 1995, “Understanding neural networks via rule extraction,” Proceedings of the International Joint Conference on Artificial Intelligence, Morgan Kauffman, San Mateo, CA, 480–487.
19. Taylor, C., 1995, “Freeway Traffic Data Prediction Using Neural Networks,” IEEE, 0-7803-2587-7195.
20. Senol, D., Ozturan, M., 2008, “Stock Price Direction Prediction Using Artificial Neural Network Approach: The Case of Turkey,” Journal of Artificial Intelligence, 1 (2), 70-77.
21. Gjertson, D. W., Clark, B. D., 2008, “For an Always Promising Transplant Prediction, Call ANN”, Transplantation, 86(10), 1349–1350.
22. Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), “Classification and regression trees,” Belmont, CA: Wadsworth.
23. Ture, M., Tokatli, F., and Kurt, I. (2009), “Using Kaplan–Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4 5 and ID3) in determining recurrence-free survival of breast cancer patients,” Expert Systems with Applications, 36(2P1), 2017–2026.
24. Mingers, J., 1989, “An Empirical Comparison of Pruning Methods for Decision Tree Induction,” Machine Learning, 4, 227-243.
25. Shahidi, A. E., Mahjoobi, J., 2009, “ComparisonbetweenM50 model tree and neural networks for prediction of significant wave height in Lake Superior,” Ocean Engineering, 36, 1175–1181.