9-2010

DATA MINING CLUSTERING IN HEALTHCARE

Bai Patel
New Jersey Institute of Technology, USA, bai.patel@njit.edu

lin Chang
Shenyang University of Technology, China, lin.chang@sut.edu.cn

DATA MINING CLUSTERING IN HEALTHCARE

Bai Patel, New Jersey Institute of Technology, USA, bai.patel@njit.edu
*Lin Chang, Shenyang University of Technology, China, lin.chang@*sut.edu.cn

Abstract

The accumulating amounts of data are making traditional analysis methods impractical. Novel tools employed in Data Mining (DM) provide a useful alternative framework that addresses this problem. This research suggests a technique to identify certain patient populations. Our model examines the patient population and clusters certain groups. Those subpopulations are then classified in terms of their appropriate medical treatment. As a result, we show the value of applying a DM model to more easily identify patients.

Keywords: Data Mining, Healthcare, Information Theory.

1 INTRODUCTION

The exponential growth of information and technology in recent years necessitates a more thorough understanding of stored data and information. Information and data are being accumulated in pace never seen before and traditional methods of handling those huge amounts are just not sufficient. This is particularly true in the healthcare industry. A search for a resolution yielded many potential solutions. One popular approach that is frequently being used in industry and that was proven quite efficient in analyzing data is Data Mining (DM). Today, DM tolls are widely used to understand marketing patterns, customer behavior, examine patients' data, and detect fraud.

This research follows DM procedures and presents a model that transform data and information into knowledge in the healthcare industry. Several authors in the information systems field studied data, information and knowledge (Alavi and Leidner 2001). The dominant view in the field is that data is raw numbers and facts. Information is processed data, or "data endowed with relevance and purpose" (Drucker 1995). Information becomes knowledge when it adds insight, abstractive value, better understanding (Spiegler 2000).

Spiegler (2000) described a model that relates data to information to knwoedge using various terms and concepts. The author stated that all are considered states in the transformation process of knowing. Tuomi (2000), on the other hand, presented a reverse model where knowledge served as the bases for information and data. The author claimed that knowledge was the result of cognitive processing initiated by an inflow of new stimulation and it can become information when it is articulated and presented in the form of text, words, or other representative forms. When incorporating both models together the result is a cycle that begins with the application of structured tacit (implicit, cognitive) knowledge; this, in turn, yields information; finally, if one adds a fixed representation and interpretation to the generated information, the outcome is data, that can be used as raw material to produce information knowledge again.

We follow this taxonomy and aim to generate knowledge to improve decision making. Specifically, we produce knowledge related to diabetes. Diabetes is considered one of the most frequent diseases in the United States. Identifying diabetic patients is therefore very important. To that end, we follow the notions of Ben-Zvi (2009) and employ concepts from related fields, such as Operations Research. We mainly concentrate on the preprocessing steps of DM and examine different applications.

The study is organized as follows: First, we review related literature. Then, we introduce the components of our model, propose several techniques for pre-processing activities and present the application with a patient database. Finally, we interpret the results and summarize the study.

2 LITERATURE REVIEW

This study applies and integrates various concepts from different fields. We now explore the different fields which are relevant to this study. We cover Data Mining and Operations Research models.

DM is one of the emerging methods in the information systems field in the past decade. When looking for its formal definition, it can be associated with the process of extracting knowledge and insights from vast quantities of data in an efficient manner (Chung and Gray, 1999; Khan et al., 2006). However, DM is not just the application of specific algorithms for extracting structure from data or information, DM also includes data pre-processing procedures. It is associated with data cleaning, incorporating appropriate prior knowledge, and proper interpretation of the mining results (Ben-Zvi, 2009). Integrating those activities together is what can be regarded as the main core of extracting knowledge out of data, what makes DM so useful.

When using DM, we mainly refer to applying statistical techniques to discover and present information in a form that are easily comprehensible (Fayyad, Piatetsky-Shapiro and Smyth 1996). DM can be applied

to different tasks related to decision-making. Those tasks include decision support, forecasting, estimation, and uncovering and understanding relationships among data elements. Chan and Lewis (2002) state that DM may help organizations achieve business, operational, and scientific goals by revealing and analyzing hidden patterns in their data — existing data from operational systems that may consume many gigabytes or terabytes of storage and may be stored on a variety of operating system platforms. The authors also claim that the challenge many organizations face is detecting these patterns in a reasonable timeframe and at an acceptable cost. When examining the actual application that have used DM, one can get the impression that this is exactly where DM can play an important role, by presenting the researcher a cost-effective balance question.

The DM methods being used today are taken from diverse fields as statistics, machine learning and artificial intelligence (Fayyad and Uthurusamy 2002; Hand et al. 2001; Khan et al. 2006). Most popular methods include regression, classification and clustering. Regression is a statistical method that makes prediction of a certain dependent variable according to the values of other independent variables. It is very useful in cases where the desired result is a concrete continuous value. Classification is learning function that maps (classifies) a data item into one of several predefined classes (Fayyad, Piatetsky-Shapiro and Smyth 1996). With classification, the predicted output (the class) is categorical; a categorical variable has only a few possible values, such as yes–no, high–middle–low, etc. (Chan and Lewis 2002). Chan and Lewis (2002) state that regression and classification are related to one another. They claim that a regression problem can be turned into a classification problem by bracketing the predicted continuous variables into discrete categories, and a classification problem can be turned into a regression problem by establishing a score or probability for each category. The most frequently used techniques with those methods are decision tress, naïve-bayes, K-nearest neighbor and neural networks.

When considering clustering, one refers to the task of segmenting a diverse group into a number of similar subgroups or clusters (Chan and Lewis 2002). Unlike what happens in classification, there are no predefined classes or groups. The clustering algorithms work according to similarities that can be found in the data itself, without any predefined rules. When comparing classification and clustering, one needs to realize that even the resulted groups in clustering are not necessarily well-defined, and it is up to the miner himself to label the final clusters, according to the clustered data (Spiegler and Maayan, 1985; Erlich et al., 2003). For a more comprehensive review on segmenting, the reader is referred to Cover and Thomas (2006).

Today, DM is applied in panoply of successful applications in many industries and scientific disciplines (Melli et al. 2006); for example, financial institutes (Chen et al., 2000), insurance agencies (Apte et al., 2002), marketing contexts (Berson et al., 1999; Davenport et al., 2001) and web mining (Scime, 2004). One important DM application is in healthcare. DM can potentially improve organizational processes and systems in hospitals, advance medical methods and therapies, provide better patient relationship management practices, and improve ways of working within the healthcare organization (Metaxiotis 2006). You may use DM to make utilization analysis, perform pricing analysis, estimate outcome analysis, improve preventive care, detect questionable practices and develop improvement strategies (Chae et al. 2003; Chan and Lewis 2002). For concrete healthcare applications, the reader is referred to Rao et al. (2006), Apte et al. 2002 and Hsu et al. 2000).

Furthermore, we also apply a concept of using production problems in data mining and use this example to leverage our technique. This production problem is extensively discussed in literature (e.g., Ben-Zvi and Grosfeld-Nir, 2010; Eden and Ronen, 1990; Grosfeld-Nir and Gerchak, 2004; Ronen and Spiegler, 1991; Kalfus et al., 2004).

Today's reality mandates companies to try and find better and efficient way to produce their products. Since every time a manufacturer attempts to produce, the product might end up defective, one might invest more efforts in inspection processes, and thus, increase the chance of success. On the other hand, investing a lot of efforts in inspection may be costly for the long-term. Therefore, a manufacturer must come up with the right balance between efficient production and costly inspections.

Our model follows the following scenario: a certain order needs to satisfied according to certain specifications (that is, the demand is considered "rigid"). Production is conducted on a single machine in batches. Each production process includes a fixed cost and a variable (per unit) cost. As the outcome is

unobservable, after the production process ends the manufacturer needs to initiate an inspection process. Inspection also entails costs per unit inspected. Inspection is conducted on a one-to-one basis. Once enough conforming units to satisfy the demand are found in the batch, the entire process is over. There is no value to the remaining units. If after the inspection process terminates the demand is still not satisfied, the manufacturer needs to produce more units to satisfy the remaining demand. The process then continues until the entire demand is satisfied.

This problem is referred to in literature as Multiple Lotsizing in Production to Order (MLPO). We note that the most important feature of this problem is its infinity capacity. Unlike other models of single-attempt demand, where unfulfilled demand is being satisfied by some sort of penalty and then the problem terminates, in this scenario, endless production and inspection processes might be required. In addition, while there are several studies on single-attempt production models with random yields, MLPO models are more rare and there not a lot of publications dealing with rigid demand. However, we do recognize the importance of those models, and therefore, focus on the inspection process.

We concentrate on inspection as its importance has dramatically increased in the past 10 to 15 years. A policy of "zero defectives" has been adopted by many manufacturing enterprises. The results have shown that this policy leads many times to somewhat expensive quality assurance inspection procedures, which make the production procedure more costly.

We refer to a serial multistage production system and assume the system is facing a certain demand and the cost of producing one unit on machine *k* is β_k . Production is imperfect and each input unit has a success probability θ_k to be successfully processed on machine *k* (Bernoulli distribution). In Figure 1 we illustrate an example of such production system. Now, if one has the option of sequencing the processing machines, then it can be shown that it is optimal (cost wise) to arrange the machines so that the ratio between the production cost and the success probability, θ_k, is increasing.

3 THE MODEL

The model we develop is binary, and therefore, it can be applied to only discrete attributes. Therefore, for continuous data we follow the algorithm suggested by Fayyad and Irani (1993) and restrict the possibilities to at least two-way, or binary, interval split for any continuous attribute.

We conduct an interval split (if at all) at the point where the information value is smallest. Once the first interval split is determined, the splitting process is repeated in the upper and lower parts of the range, and so on recursively. We use a significance level of 5% as a reasonable threshold as a stopping criteria.

To appropriately process the data, we utilize the MLPO production scenario. We sequence the data items (entries) according to their allocated weights and their amount of mutual information with respect to the dependent variable. Using (4), each attribute is allocated a likelihood ratio statistic $L_{j,k}$ (j=1,2,…,d; k=1,2,…,p_j). To be consistent with the production system parameters, we transform the likelihood ratio statistic into a chi-square probability, denoted by $\theta_{j,k}$ (j=1,2,…,d; k=1,2,…,p_j). Note that in the MLPO problem β_k represent costs (which are sequenced in increasing order) while in our model $\beta_{j,k}$ represent importance (how important the specific data item is). Therefore, to be consistent with the mathematical result, we perform the simple transformation of $1-\beta_{j,k}$ in the MLPO ratio numerator to arrange the data entries by the increasing ratio.

The result of this assessment constitutes a clustering of the data into a number of groups that have significantly different weights. We can define each group by the weight it was assigned, which can, in turn, represent the combinations of values of the independent variables. This clustering may be used to predict the likelihood of the dependent variable's event occurrences.

Next, we employ the Vector space model (VSM) is a classic technique of information retrieval that transforms textual data into an algebraic vector. Let *n* be the number of documents in a corpus, and let *m* be the total number of different words after preprocessing (such as, removing stop words and numbers, stripping non-word tokens, extracting stem etc.), VSM is also called the dictionary or bag of words.

Every unique term (word) from the collection of analyzed documents forms a separate dimension in the VSM. In its simplest form a model of a document d is a vector of length m whose i-th entry indicates whether or not the i-th word of the dictionary occurs in d. Note that documents are considered as vectors in the m -dimensional space of all dictionary entries.

The main advantage of our technique over methods utilizing term frequency distribution only is that phrases are usually more informative than unorganized set of keywords, and can be directly used to label the discovered clusters, which in other clustering algorithms becomes a problem. This method treats documents as a set of phrases (sentences) not just as a set of words. The sentence has a specific, semantic meaning (words in the sentence are ordered). Suffix tree model considers a document as a sequence of words, not characters. A revised definition of suffix tree is follow:

The same classic is used to describe the suffix tree clustering document model, but the leaf nodes have been revised by adding document frequency term.

The internal nodes of the suffix tree are drawn as circles, labelled with characters a through f for further reference. Each internal node represents a phrase and a base cluster. Each leaf node is drawn as rectangle and designates a suffix of a document. It keeps the frequencies information that different documents traversing. The first number in each rectangle indicates the string from which document that suffix originated; the second number represents the position in that string where the suffix starts; the third number represents the traversed times of the relevant document.

Each internal node represents an overlap phrase shared by at least two suffixes. The more internal nodes shared by two documents, the more similar the documents tend to be.

4 THE CLUSTERING APPLICATION

When considering the healthcare industry, we may find several interesting and challenging applications for DM. Following our analytical formulation, we now present a real-life application for identifying diabetic patients in a small US town. The main objective of this application is to recognize what causes diabetics. We were able to obtain a patient database and conduct an analysis seeking to identify which patients have high probability of being diabetic. Thus, we may gain some insights on the disease and its causes.

Group	No. of Patients
1	12
2	56
3	789
4	123
5	564
6	662
7	218
8	23
9	87
10	878
Total	3,412

Table 1. The Resulted Groups (Clusters) of the Data Mining Procedures.

For this study we used a database of 3,412 with several relevant attributes. We note that most attributes are defined as numeric and therefore may take any possible numeric number. This, of course, makes the original database impractical for the needs of this study and the model we developed. However, following the described transformation of the data, with the appropriate pre-processing operations, we

applied the DM procedures detailed above to obtain a database we can analyze. As a result, the patient population was divided into distinct groups (clusters) defined in Table 1.

It seems that the following characteristics were important to distinguish between the groups: age, race, family disease history, patients with family history of diabetes and body weight.

The next step was to validate the DM procedure. We used the dataset and followed the procedures conducted with the patient list to cluster the validation dataset into the seven groups. The results are presented in Table 2. The results show that the actual distribution of diabetic patients does not deviate significantly from the prediction made based on the DM results.

Patient Group	No. of Patients	Diabetic Patients	
		Actual	Predicted
1	12	10	16
2	56	56	47
3	789	12	8
4	123	89	98
5	564	54	65
6	662	94	95
7	218	125	102
8	23	2	1
9	87	58	50
10	878	169	171

Table 2. Predicted and Actual Number of Diabetic Patients.

Next, we aim to evaluate the results of our DM algorithm and to compare them with the traditional analysis methods. However, no established criteria can be found in literature for deciding which methods to use in which circumstances. We tested the benchmark methods using the dataset of the previous section and compared the results obtained by the various methods. We measured whether the different methods were able to make the correct predictions (diabetic and non-diabetic patients).

Our findings show that using a clustering method with a single linkage technique and a Euclidean Distance as a criterion produces the best result. This method was able to identify 80% of the diabetic cases. The second best method was our suggested clustering technique with 77% of correct predictions. The other methods also produced relatively good results: Classification was able to predict 75% of diabetic cases. Regression was the worst method with only 71% accuracy. We believe that this lack of accuracy was due to the fact that we are dealing with a discrete variable (the diabetic variable) and regression usually produces good results with continuous numeric variables.

In the next section we discuss the interpretation and outcomes of our application.

5 DISCUSSION AND CONCLUSIONS

Our method provides many useful insights:

First, our method is making use of concepts from other close field, like Operations Research and Inventory Management. The use of Information Theory is particularly interesting as this theory relates also to the Information Systems field. When incorporating those concepts together we were able to show that our method is relatively good compared to other traditional methods. Therefore, one outcome is establishing our method as a valid method for DM.

Second, we used to the DM procedure to gain knowledge about diabetes. We conclude that the following variables can serve as good indicators for identifying potential diabetic patients: family history, body

weight and age. This may become a powerful predictive tool for any organization seeking to perform a more precise and informed patient selection process to identify diabetic patients. Although we do not attempt to generalize the results to the entire population in the United States, we believe that our findings represent the different population distribution and the causes of diabetics we found in this study are valid. Obviously, each organization (e.g., hospitals) will have its own set of variables that determines the causes of diabetics (according to its own measures). However, we expect that the nature of the significant variables is similar across institutions with similar patient populations.

This study showed the benefits of using DM in the healthcare domain. We made a theoretical contribution, as we exhibit a formal presentation of the DM process, while integrating several concepts from other disciplines. We believe that the results that we shoed in this study can help decision makers in determining a health policy related to diabetes. However, although the presented method was proven to be quite good, it also has its limitations. First, we were not able to cluster the population into different risk-related populations. This was due to the low probability of being a diabetic patient – 7.5% for the entire patient population. Second, we were not able to subcategorize the different variables that we found critical for identifying diabetic patients. For example, we cannot state that people over 40 have a larger probability of catching the disease or that people who are considered fat are in a high risk group. We leave those determinations for future inquiry. In addition, the data we used was taken from relational datasets. The applicability of our model to other types of databases is yet to be studied.

References

Alavi, M., Leidner, D., (2001) Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues, MIS Quarterly, 25(1), 107–136.

Apte, C., Liu, B., Pednault, E.P.D., and Smyth, P. (2002) Business Applications of Data Mining, Communications of the ACM, 45(8), 49-53.

Ben-Zvi T. (2009) Data Analysis: A Roadmap to Better Decision-Making: Methods, Techniques and Applications, VDM Verlag.

Ben-Zvi, T. and Grosfeld-Nir, A. (2010) Multistage production systems with random yields and rigid demand, International Journal of Manufacturing Technology and Management, 20(1), 286–299.

Berson, A., Smith, S., and Thearling, K. (1999) Building Data Mining Applications for CRM, McGraw-Hill Companies.

Chae, Y., Kim, H., Tark, K., Park, H., and Ho, S. (2003), Analysis of Healthcare Quality Indicators Using Data Mining and Decision Support Systems, Expert Systems with Application, 24(2), 167-172.

Chan, C., and Lewis B. (2002), A Basic Primer on Data Mining, Information Systems Management, 19(4), 56-60.

Chen, L., Sakaguchi, T., and Frolick, M.N. (2000) Data Mining Methods, Applications, and Tools, Information Systems Management, 17(1), 65-70.

Chung, H.M., Gray, P. (1999), Data mining, Journal of Management Information Systems, 16(1), 11–16.

Cover, T.M., and Thomas, J.A. (2006) Elements of information theory, 2nd Edition. New York: Wiley-Interscience.

Davenport, T.H., Harris, J.G., and Kohli, A.K. (2001) How Do They Know Their Customers So Well?, MIT Sloan Management Review, 42(2), 63-73.

Drucker, P.E. (1995), The Post Capitalistic Executive, in P.E. Drucker (ed.), Management in a Time of Great Change, New York: Penguin.

Eden, Y., and Ronen, B. (1990) Service Organization Costing: A Synchronized Manufacturing Approach, Industrial Management, 32(5), 24-26.

Erlich, Z., Gelbard, R., and Spiegler, I. (2003) Evaluating a Positive Attribute Clustering Model for Data Mining, Journal of Computer Information Systems, 43(3), 100-108.

Fayyad, U. M., and Irani, K. (1993) Multi-Interval Discretization of Continuous-Valued Attributes for Classification Leaning, Proceedings of the 13th International Joint Conference on Artificial Intelligence, 1022-1027.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. (1996) The KDD Process for Extracting Useful Knowledge from Volumes of Data, Communications of the ACM, 39(11), 27–34.
Fayyad, U., and Uthurusamy, R. (2002) Evolving Data Mining into Solutions for Insights, Communications of the ACM, 45(8), 28-31.
Grosfeld-Nir, A., and Gerchak, Y. (2004) Multiple Lotsizing in Production to Order with Random Yields: Review of Recent Advances, Annals of Operations Research, 126(1), 43-69.
Hand, D. J., Mannila H., and Smyth, P. (2001) Principles of Data Mining, MIT Press.
Hsu, W., Lee, M., Liu, B., and Ling, T. (2000) Exploration Mining in Diabetic Patient Databases: Findings and Conclusions, In Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2000), Boston, August 20-23, ACM Press, New York, 430-436.
Kalfus, O., Ronen, B., and Spiegler I. (2004) A Selective Data Retention Approach in Massive Databases, Omega, 32(2), 87-95.
Khan, S., Ganguly, A.R., and Gupta, A. (2006) Creating Knowledge for Business Decision Making, In Schwartz, D. (Ed.), Encyclopedia of Knowledge Management, Hershey, PA : Idea Group Inc., 81-89.
Melli, G., Zaïane, O.R., and Kitts, B. (2006) Introduction to the Special Issue on Successful Real-World Data Mining Applications, SIGKDD Explorations, 8(1), 1-2.
Metaxiotis, K. (2006) Healthcare Knowledge Management, In Schwartz, D. (Ed.), Encyclopedia of Knowledge Management, Hershey, PA: Idea Group Inc., 204-210.
Rao, R. B., Krishnan, S., and Niculescu R. S. (2006) Data Mining for Improved Cardiac Care, SIGKDD Explorations, 8(1), 3-10.
Ronen, B., and Spiegler, I. (1991) Information As Inventory: A New Conceptual View, Information & Management, 21(4), 239-247.
Scime, A. (2004) Web Mining: Applications and Techniques, Idea Group Publishing.
Spiegler, I. (2000) Knowledge Management: a New Idea or a Recycled Concept, Communications of the AIS, 14(3), 1-24.
Spiegler, I. and Maayan, R. (1985) Storage and retrieval considerations of binary data bases, Information Processing & Management, 21(3), 233-254.
Tuomi, I. (2000) Data is More Than Knowledge: Implications of the Reversed Hierarchy for Knowledge Management and Organizational Memory, Journal of Management Information Systems, 16(3), 103-117.
Witten, I.H., and Frank, E., (2000) Data Mining – Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann Publishers.