Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2009 Proceedings

Americas Conference on Information Systems (AMCIS)

2009

Identifying Diabetic Patients: A Data Mining Approach

Thomas Porter
The University of North Dakota, tporter@und.edu

Barbara Green
The University of North Dakota, bgreen@und.edu

Follow this and additional works at: http://aisel.aisnet.org/amcis2009

Recommended Citation

Porter, Thomas and Green, Barbara, "Identifying Diabetic Patients: A Data Mining Approach" (2009). *AMCIS 2009 Proceedings*. 500.
http://aisel.aisnet.org/amcis2009/500

This material is brought to you by the Americas Conference on Information Systems (AMCIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in AMCIS 2009 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

Identifying Diabetic Patients: A Data Mining Approach

Thomas Porter
The University of North Dakota
tporter@und.edu

Barbara Green
The University of North Dakota
bgreen@und.edu

ABSTRACT

Mounting amounts of data made traditional data analysis methods impractical. Data mining (DM) tools provide a useful for alternative framework that addresses this problem. This study follows a DM technique to identify diabetic patients. We develop a model that clusters diabetes patients of a large healthcare company into different subpopulation. Consequently, we show the value of applying a DM model to identify diabetic patients.

Keywords:

Data Mining, Healthcare, Information Theory, Inventory Theory.

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) or when it is being authenticated (Dretske 1981; Machlup 1980; Vance 1997).

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 and Spiegler (2007) and employ concepts from other fields, such as Operations Research, Inventory Management, and Information Theory. We mainly concentrate on the preprocessing steps (i.e., data discretization and data transformation). Our main goal in this study is to create a core DM application that helps identifying the causes of diabetics.

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.

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, Operations Research, Information Theory and Inventory Management.

Data Mining

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 and Spiegler, 2007). 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.

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 pplication 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).

Data Representation

When following the DM process, we use binary databases as used by Spiegler and Maayan (1985) and Erlich et al. (2003). In those databases, data appears in a binary form rather than the common alphanumeric format. The binary model views a database as a two-dimensional matrix where the rows represent objects and the columns represent all possible data values of attributes. The matrix's entries are either '1' or '0' indicating that an object has or lack the corresponding data values. We note that binary databases require that data appears as discrete. Therefore, in order to comply with this requirement we later discretize any continuous or alphanumeric attribute. We stress that when we transform data into a binary format, we maintain data integrity. That is, no information loss is tolerated in the binary conversion process.

Information Theory Concepts

In addition to binary data representation, this study also employs some techniques from information theory. For a complete review of information theory and its application see Witten and Frank (2000). Information theory, first set up by Shannon (1948), is a discipline in applied mathematics involving the quantification of data with the goal of enabling as much data as possible to be reliably stored on a medium or communicated over a channel. The measure of information is known as information entropy. This section follows closely with Ben-Zvi and Spiegler (2007).

The entropy H(X) of a discrete random variable X is defined by

$$H(X) = -\sum_{x} p(x)\log p(x) \tag{1}$$

where p(x) denotes the probability that X will take on the value x, and the summation is over the range of X.

The joint entropy H(X,Y) of pair of discrete random variables X and Y with joint distribution p(x,y) is given by:

$$H(X,Y) = -\sum_{x}\sum_{y} p(x,y)\log p(x,y) \tag{2}$$

The mutual information I(X:Y) is the relative entropy between X and Y and is defined as follows:

$$I(X:Y) = H(X) - H(X,Y) = -\sum_{x}\sum_{y} p(x,y)\log\frac{p(x)p(y)}{p(x,y)} \tag{3}$$

Mutual information represents the reduction in the uncertainty of X that is provided by knowing the value of Y.

When natural logarithms are used, and I(X:Y) is estimated from a sample of *n* observations, then the following result is obtained:

$$2nI(X:Y) = -2n\sum_{x}\sum_{y} p(x,y)\log\frac{p(x)p(y)}{p(x,y)} = L^2 \tag{4}$$

L^2 is known as the likelihood ratio statistic and is asymptotically chi-square distributed.

For a more comprehensive review on information theory, the reader is referred to Cover and Thomas (2006) and Gallager (1968).

We later use the above concepts of entropy, mutual information and the likelihood ratio statistic for the main DM diabetic application.

Information as Inventory

Some studies (e.g., Eden and Ronen, 1990; Ronen and Spiegler, 1991; Kalfus et al., 2004) suggest that information, as a resource, should be viewed and treated as inventory. While considering inventory theory, we also employ modern production and manufacturing concepts. Such a view of information is in fact consistent with the analogy of data processing and production management. The idea of the above studies was to use modern inventory techniques, and apply them to the information system area.

We follow the same supposition and apply a production problem that is referred to as “Multiple Lot sizing in Production to Order” (MLPO). This problem is extensively discussed in literature (e.g., Ben-Zvi and Grosfeld-Nir, 2007; Grosfeld-Nir et al., 2006). Grosfeld-Nir and Gerchak (2004) state that the need to incorporate production and inventory decision models has been recognized by industrial engineers and management scientists since the 1950’s. However, the issue was usually posed as selecting the optimal reject (shrinkage) allowances in determining production or order lot sizes. While some additional work was done in the 1970’s, the area has witnessed a vigorous explosion of activity in the mid 1980’s, which is still going strong. Some fundamental results were obtained, and new and more general models formulated and analyzed. This resurgence of interest in the subject can be attributed, at least in part, to the renewed interest by manufacturing industries in understanding the logistical implications of producing defective items.

In this study we refer to a serial multistage production system and assume the system is facing a certain demand. This demand needs to be fulfilled by producing products. We assume that the cost of producing one unit on machine *k* is β_k . Naturally, the production process is imperfect and each input unit has a success probability θ_k to be successfully processed on machine *k* (this complies with the Bernoulli distribution). Now, assuming we may rearrange the order of production, moving the machines forward or backwards, we can sequence the processing machines to achieve an optimal arrangement (cost wise). Studies show that this arrangement can be achieved when the ratio $\frac{\beta_k}{1-\theta_k}$ is increasing.

THE MODEL

In this section we develop the DM model, following several pre-processing activities of the DM process. We assume a dataset is represented as a finite data table with *n* rows labeled as objects $\{x_1,x_2,\ldots,x_n\}$ and *d* columns labeled as attributes which characterize the objects $\{a_1,a_2,\ldots,a_d\}$. The entry in row *x* and column *a* has the value f(x,a).

Data Discretization

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 define the following information function (Info):

$$\text{Info}([a,b]) = H(\frac{a}{a+b},\frac{b}{a+b}) \tag{5}$$

Using the formulas in (1) and (5) we can calculate the information measure for certain values of *a* and *b* (e.g., a=2, b=7):

$$\text{Info}([2,7]) = -2/7 \text{ x } \log 2/7 - 7/2 \text{ x } \log 7/2 = 0.425 \tag{6}$$

The 0.425 bits we obtained represents the amount of information given at a certain examined data point. This procedure may be applied for each possible data point, where *a* and *b* represent the number of values at the data point. 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.

Data Transformation

The goal of data transformation is to transform the current data representation into an appropriate format which can be used directly as a binary database. This section follows closely with Ben-Zvi and Spiegler (2007).

For each object, we form a binary representation vector, which represents the values of its attributes in a binary format, as follows:

The domain of each attribute a_j (j=1,2,…,d) is all its possible values, where p_j is the domain size (i.e., its exclusive possible values).

We denote the kth value of attribute a_j (j=1,2,…,d; k=1,2,…,p_j) by $a_{j,k}$. We can now represent the domain attributes vector of all possible values of all *d* attributes as:

$$(a_{1,1},a_{1,2},\ldots,a_{1,p1},a_{2,1},a_{2,2},\ldots,a_{2,p2},\ldots,a_{d,1},a_{d,2},\ldots,a_{d,pd})$$

We define the binary representation vector for each object i (i=1,2,…,n) in the following form:

$$x_{i,j,k} = \begin{cases} 1 & \text{,if for object i, the value of attribute j is } a_{j,k} \\ 0 & \text{,otherwise} \end{cases}$$

where i=1,2,…,n; j=1,2,…,d; and k=1,2,…,p_j

$x_{i,j,k}$ is the corresponding value for the k^{th} value of attribute j ($a_{j,k}$) for object i. $x_{i,j,k}$ may obtain either 1 or 0, indicating that a given object has or lacks a given value $a_{j,k}$ for attribute j. Then, the binary representation vector, for object i, is given by

$$(x_{i,1,1}, x_{i,1,2}, \ldots, x_{i,d,pd})$$

In the next section we introduce the core DM procedure.

THE DATA MINING PROCEDURE

The DM algorithm makes an evaluation of the data. We randomly allocate a value $\beta_{j,k}$ ($j=1,2,\ldots,d$; $k=1,2,\ldots,p_j$) to each data item (entries). This would be considered the item's weight. The weights are limited to values between 0 and 1, where the sum of all weights allocated must equal to 1.

Next we need to process the data. For that, 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 $L_{j,k}$ ($j=1,2,\ldots,d$; $k=1,2,\ldots,p_j$). To be consistent with the production system parameters, we transform the likelihood ratio statistic into a chi-square probability, denoted by $\theta_{j,k}$ ($j=1,2,\ldots,d$; $k=1,2,\ldots,p_j$). 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 $\frac{\beta_k}{1-\theta_k}$ ratio numerator to arrange the data entries by the increasing ratio of $\frac{1-\beta_{j,k}}{1-\theta_{j,k}}$.

The final 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.

THE DIABETIC 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	74
2	136
3	421
4	859
5	1235
6	2285
7	4268
Total	9278

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

For this study we used a database of 9278 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	74	5	5.6
2	136	12	10.2
3	421	35	31.6
4	859	60	64.4
5	1235	90	92.6
6	2285	165	171.4
7	4268	335	320.1

Table 2. Predicted and Actual Number of Diabetic Patients.

METHOD EVALUATION AND COMPARISON

Next, we evaluated the results of our DM algorithm and compared them with the traditional analysis methods. Many studies make this comparison when introducing new methods or DM techniques (see for example, Lim et al. 2000; Wilson et al. 2006). Yet, 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.

UTILIZING DISCOVERED KNOWLEDGE TO IDENTIFY DIABETIC PATIENTS

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.

CONCLUSIONS

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

1. Alavi, M., Leidner, D., (2001) Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues, *MIS Quarterly*, 25, 1, 107–136.
2. 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.
3. Ben-Zvi, T., and Grosfeld-Nir, A. (2007) Serial Production Systems with Random Yield and Rigid Demand: A Heuristic, *Operations Research Letters,* 35, 2, 235-244.
4. Ben-Zvi T. and Spiegler, I., (2007) Data Mining and Knowledge Discovery: An Analytical Investigation, Proceedings of the *13th Americas Conference on Information Systems (AMCIS)*, Keystone, Colorado.
5. Berson, A., Smith, S., and Thearling, K. (1999) *Building Data Mining Applications for CRM*, McGraw-Hill Companies.
6. 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.
7. Chan, C., and Lewis B. (2002), A Basic Primer on Data Mining, *Information Systems Management*, 19, 4, 56-60.
8. Chen, L., Sakaguchi, T., and Frolick, M.N. (2000) Data Mining Methods, Applications, and Tools, *Information Systems Management*, 17, 1, 65-70.
9. Chung, H.M., Gray, P. (1999), Data mining, *Journal of Management Information Systems*, 16, 1, 11–16.
10. Cover, T.M., and Thomas, J.A. (2006) *Elements of information theory*, 2nd Edition. New York: Wiley-Interscience.
11. 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.
12. Dretske, F. (1981), *Knowledge and the Flow of Information*, MIT Press, Cambridge, MA.
13. Drucker, P.E. (1995), The Post Capitalistic Executive, in P.E. Drucker (ed.), *Management in a Time of Great Change*, New York: Penguin.
14. Eden, Y., and Ronen, B. (1990) Service Organization Costing: A Synchronized Manufacturing Approach, *Industrial Management*, 32, 5, 24-26.
15. 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.
16. 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.
17. 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.

18. Fayyad, U., and Uthurusamy, R. (2002) Evolving Data Mining into Solutions for Insights, *Communications of the ACM,* 45, 8, 28-31.

19. Gallager, R. (1968) *Information Theory and Reliable Communication*, New York: John Wiley and Sons.

20. Grosfeld-Nir, A., Anily, S., and Ben-Zvi, T. (2006) Lot-Sizing Two-Echelon Assembly Systems with Random Yields and Rigid Demand, *European Journal of Operational Research*, 173, 2, 600-616.

21. 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.

22. Hand, D. J., Mannila H., and Smyth, P. (2001) *Principles of Data Mining*, MIT Press.

23. 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.

24. Kalfus, O., Ronen, B., and Spiegler I. (2004) A Selective Data Retention Approach in Massive Databases, *Omega,* 32, 2, 87-95.

25. 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.

26. Lim, T.S., Low, W.Y., and Shih, Y.S. (2000) A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms, *Machine Learning*, 40, 3, 203-229.

27. Machlup, F. (1980) *Knowledge: Its Creation, Distribution, and Economic Significance*, Princeton University Press, Princeton, NJ.

28. 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.

29. Metaxiotis, K. (2006) Healthcare Knowledge Management, In Schwartz, D. (Ed.), *Encyclopedia of Knowledge Management*, Hershey, PA: Idea Group Inc., 204-210.

30. Rao, R. B., Krishnan, S., and Niculescu R. S. (2006) Data Mining for Improved Cardiac Care, *SIGKDD Explorations*, 8, 1, 3-10.

31. Ronen, B., and Spiegler, I. (1991) Information As Inventory: A New Conceptual View, *Information & Management,* 21, 4, 239-247.

32. Scime, A. *Web Mining: Applications and Techniques*, Idea Group Publishing, 2004.

33. Shannon, C.E. (1948) A Mathematical Theory of Communication, *Bell System Technical Journal*, 27, 379–423 and 623–656.

34. Spiegler, I. (2000) Knowledge Management: a New Idea or a Recycled Concept, *Communications of the AIS*, 3, 14, 1-24.

35. Spiegler, I. and Maayan, R. (1985) Storage and retrieval considerations of binary data bases, *Information Processing & Management,* 21, 3, 233-254.

36. 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.

37. Vance, D. M. (1997) Information, Knowledge and Wisdom: The Epistemic Hierarchy and Computer-Based Information System, in B. Perkins and I. Vessey (Eds.), Proceedings of *the Third Americas Conference on Information Systems*, Indianapolis, IN.

38. Wilson. R.I., Rosen, P.A., Al-Ahmadi, M.S., (2006) Knowledge Structure and Data Mining Techniques, In Schwartz, D. (Ed.), *Encyclopedia of Knowledge Management*, Hershey, PA: Idea Group Inc., 523-529.

39. Witten, I.H., and Frank, E., (2000) *Data Mining – Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann Publishers.