2011

DATA MINING CLUSTERING: A HEALTHCARE APPLICATION

Dominick Doust

Zack Walsh

DATA MINING CLUSTERING: A HEALTHCARE APPLICATION

Abstract

The exponential growth of data in recent years necessitates the development of new methods that can handle massive amounts of stored data and information. This is particularly true in the healthcare industry. A popular approach that was proven efficient in analyzing data is Data Mining (DM). DM aims to find patterns in databases and to facilitate decision-making according to those patterns. We suggest that the use of DM can help physicians and healthcare administrators make better decisions and even save lives.

Keywords: Data Mining, Diabetes, Healthcare, Information Theory, Inventory Theory.

1 INTRODUCTION

Mounting amounts of data make traditional data analysis methods impractical. Data mining (DM) tools provide a useful alternative framework that addresses this problem. This research follows the DM process and presents a mathematical model of transforming data and information into knowledge in the healthcare industry. This knowledge is then used to improve decision making. Specifically, we use the obtained knowledge to identify potential diabetic patients. To that end, we borrow ideas from related, applicable fields (e.g., Operations Research, Inventory Management, and Information Theory). We present several preprocessing steps (i.e., data discretization transformation), and also exhibit interpretation of the obtained data patterns. As diabetes is considered one of the leading causes of deaths in the US, with more than 23 million, approximately 7% of the population, suffering from the disease (National Center for Health Statistics, 2009), we aim to use DM techniques to identify diabetic patients. Given the fact that a third of all Americans with diabetes is undiagnosed (National Center for Health Statistics, 2009), understanding the factors that influence the disease is crucial. We develop a mathematical model that clusters the patients of a large healthcare institute into different subpopulation. Consequently, we show the merit and value of applying a well-structured model to identify the probability of a patient to become diabetic. Our practical goal is to create a core DM algorithm that helps identifying the causes of diabetes (type 2).

The study is organized as follows: First, we review related literature. Then, we introduce the model, propose several techniques for pre-processing activities and present the DM algorithm for extracting patterns from data. Next, we conduct an investigation with a patient database and evaluate the obtained results. Finally, we report our interpretation of the outcomes and summarize the study.

2 LITERATURE REVIEW

This study applies and integrates various concepts from several fields (Data Mining, Operations Research, Information Theory and Inventory Management). This section summarizes relevant literature in those fields.

2.1 Data Mining

DM is considered one of the new methods that have appeared in the information systems field in the past decade. It often appears in literature as a synonym for the process of extracting knowledge and insights from vast quantities of data in an efficient manner (Chung and Gray, 1999; Khan et al., 2006). Beyond the application of specific algorithms for extracting structure from data under acceptable computational efficiency limitations, DM includes data pre-processing, data cleaning, incorporating appropriate prior knowledge, and proper interpretation of the mining results (Ben-Zvi and Spiegler, 2007; Pal and Mitra, 2004). All those activities are essential to ensure that useful knowledge is derived from the data.

One of the most important applications of DM 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). Moreover, Hospitals are using 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). For example, Rao et al. (2006) present a DM application that improves the quality of health care and reduces the costs surrounding cardiovascular disease. The authors state that this application appears to have saved lives. Other healthcare applications may be found in Apte et al. (2002), Hsu et al. (2000) and Wang et al. (2005).

In their comprehensive review, Jain et al. (2009) state that the availability of a vast collection of data mining methods and techniques can easily confuse a user attempting to select an algorithm suitable for the problem at hand. Nevertheless, researchers such as Chung and Gray (1999) point out to several drawbacks traditional data mining methods and techniques hold; for example, a limited capacity to handle multifaceted datasets with high dimensionality. The algorithm we present in this work successfully addresses those challenges.

2.2 Data Representation and Information Theory Concepts

When executing the DM process, we employ the concept of binary database (see Spiegler and Maayan, 1985; Erlich et al., 2003), where 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 lacks the corresponding data values. We note that when transforming regular alphanumeric data into a binary format, we maintain data integrity. That is, no information loss is tolerated in the binary conversion process.

In addition to binary data representation, this study also employs some techniques from information theory. 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. 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).

We later use the above concepts of entropy, mutual information and the likelihood ratio statistic when conducting data discretization and the DM algorithm.

2.3 Information as Inventory – An Operations Research Perspective

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, in line with modern production and manufacturing concepts. Such a view of information is in fact consistent with the analogy of data processing and production management. Their idea is to use modern inventory techniques, and apply them to the information system area.

For this aim, we apply an Operations Research 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 and Gerchak, 2004; Grosfeld-Nir et al., 2006).

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 $\frac{\beta_k}{1-\theta_k}$ is increasing.

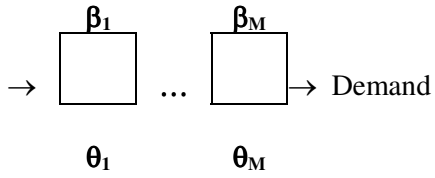

Figure 1. An Example of an MLPO Production System with M machines.

3 THE MODEL

In this section we develop our 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).

3.1 Data Discretization

The DM algorithm we develop deals only with 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.

Employing the information theory technique introduced in (1), we define the following information function (Info):

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

Using the formulas in (1) and (5) we can calculate the information measure for certain values of a and b (e.g., a=4, b=5):

$$\text{Info}([4,5]) = -4/9 \text{ x } \log 4/9 - 5/9 \text{ x } \log 5/9 = 0.298 \tag{6}$$

The 0.298 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 loss at 5% level is considered statistically reasonable.

3.2 Data Transformation

The goal of data transformation is to transform the current data representation into an appropriate format which can be used directly by the DM algorithm.

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 k^{th} 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.

4 THE DATA MINING ALGORITHM

The DM algorithm consists of the following three procedures: (1) data assessment and evaluation; (2) partitioning; and (3) grouping.

We begin the algorithm with data assessment and evaluation. This procedure determines which attributes are more critical than others and establishes the sequence of operation. As attributes were reduced and transformed in preprocessing procedures, we allocate a value $\beta_{j,k}$ (j=1,2,…,d; k=1,2,…,p_j) to each attribute $a_{j,k}$ (j=1,2,…,d; k=1,2,…,p_j), representing the attribute's weight. The weights are limited to values between 0 and 1, where the sum of all weights allocated must equal to 1 ,i.e.,

$$\sum_{j=1}^{d}\sum_{k=1}^{p_j}\beta_{j,k} = 1 \qquad (7)$$

Now, the algorithm determines the attributes' processing sequence. For this aim we utilize the MLPO production scenario. We sequence the attributes 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 (which, respectively, ought to be

sequenced in decreasing order). Therefore, we perform the simple transformation of $1-\beta_{j,k}$ in the MLPO $\frac{\beta_k}{1-\theta_k}$ ratio numerator to arrange the attributes by the increasing ratio of $\frac{1-\beta_{j,k}}{1-\theta_{j,k}}$.

To ease understanding of this algorithm procedure, we now present it in pseudo code:

For j=1 to d
  For k=1 to p_j
  $X_{j,k,total}=0$
  For i=1 to n
    $x_{j,k,total}=x_{j,k,total}+x_{i,j,k}$
  Next i
  $\beta_{j,k}=\frac{x_{j,k,total}}{n}\beta_j$
  Calculate $L_{j,k}$
  Transform $L_{j,k}$ into a chi-square probability $\theta_{j,k}$
  Next k
Next j
Arrange $\beta_{j,k}$ and $\theta_{j,k}$ by the increasing ratio of $\frac{1-\beta_{j,k}}{1-\theta_{j,k}}$

The core of the algorithm follows:

Partitioning step:

- Use the first sequenced variable to split the population sample into two partitions, corresponding to its two possible values: "0" and "1".
- Repeated this procedure for each sequenced attribute until no further splitting is justified; a justification is determined by a likelihood ratio statistic. If the likelihood ratio statistic is greater than the critical value of the chi-square distribution for a given significance level, then it is justified to partition the population into two subpopulations corresponding to the two values of the selected independent variable: "0" and "1". Otherwise, no partition is warranted.
- If partitioning is justified, repeat this procedure for each of the two newly created subpopulations.
- Terminate the procedure when all remaining subpopulations are terminal.

Grouping step:

- Segment the subpopulations created by the partitioning procedure into groups that are most similar in terms of the probabilities associated with the dependent variable.
- Rank the subpopulations in ascending order of the dependent variable's occurrence probabilities as estimated from the sample.
- Using equation (3), calculate the loss of information about the dependent variable that would result if the two subpopulations were to be combined into a single subpopulation for each pair of subpopulations ranked adjacently.
- Identify the pair resulting in the smallest loss.
- Calculate the likelihood ratio statistic using equation (4), where the sample size *n* equals to the number of observations in the two samples to be combined.
- If and only if the statistic is smaller than the critical value of the chi-square distribution for a given probability of false rejection, then group the two subpopulations and merge the corresponding samples.

- Repeat the process until the smallest loss of mutual information becomes statistically significant. That is, the best pair of subpopulations being considered for grouping is significantly different. Then, the grouping procedure terminates.

Figure 2 presents an illustration of the first two steps of the algorithm: Data Assessment and Partitioning. Figure 3 illustrates the grouping procedure.

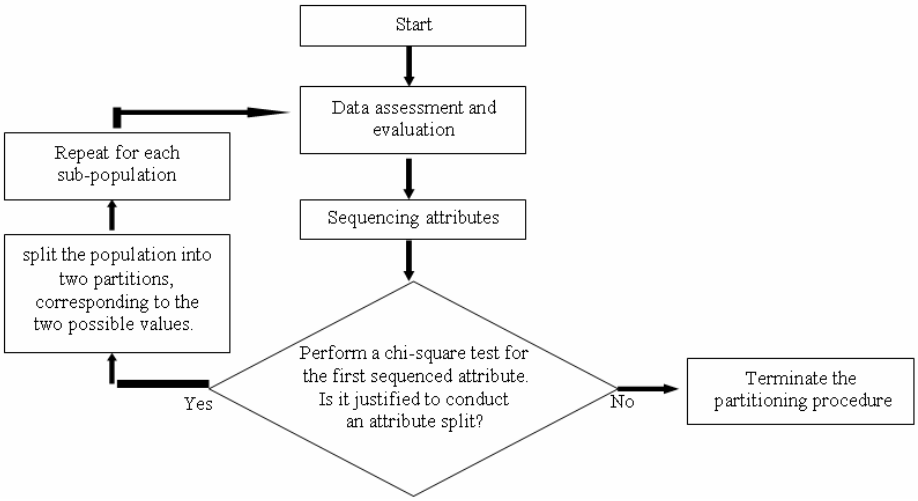

Figure 2. The Flow of the First Two Data mining Algorithm's Steps: Data Assessment and Partitioning.

The subpopulations remaining when the algorithm terminates constitute a clustering of the population into a number of groups that have significantly different occurrence probabilities with regard to the dependent variable. Each group is defined in terms of combinations of values of the independent variables. This clustering may be used to predict the likelihood of the dependent variable's event occurrence among the database's inflowing "new" objects and may carry out a certain policy for decision makers.

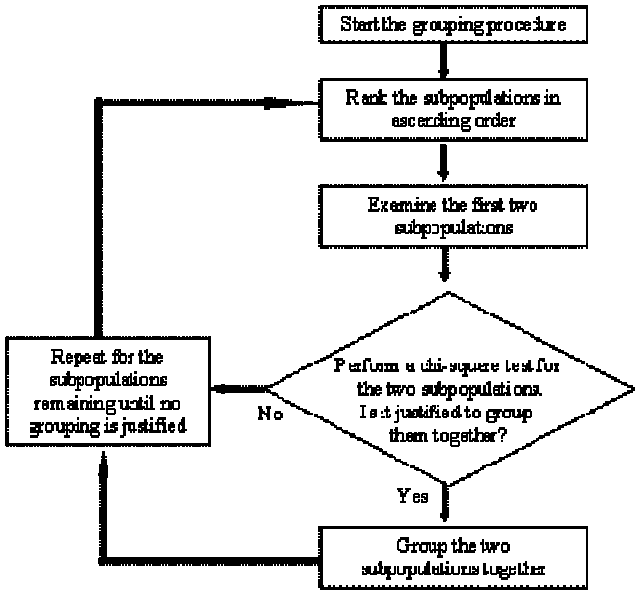

Figure 3. The Flow of the Grouping Procedure.

5 THE DIABETES APPLICATION

The healthcare industry offers many interesting and challenging applications for DM. Following our analytical formulation, we now present a real-life application for identifying diabetic patients (type 2) of a large US healthcare institute. The main objective of this application is to recognize the probability of a patient to become diabetic. We aim to profile a patient database and to 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. From a utilitarian perspective, the company is interested in improving its ability to trace patients with certain potential diseases and charge them accordingly.

The healthcare institute we worked with mainly treats cancer; however, about 10% of its patients have diabetes. We were provided with a patient dataset, but without social security numbers or any other forms of identification. We divided the dataset into two datasets: the first dataset included 13592 patients and the second dataset included 1359 patients. We used the main dataset for our analysis and the second one for validation. Both datasets included 49 (relevant) attributes (See the appendix for a complete list of attributes). We note that most attributes are defined as numeric and therefore may take many possible values. Therefore, we discretized the variables and used the diabetes variable as a target attribute.

Following pre-processing operations, we applied the DM algorithm detailed above. As a result, the patient population was divided into four distinct groups (clusters). In Table 1 we examine each cluster and summarize the main (interesting) characteristics of each group.

Characteristic	Group 1	Group 2	Group 3	Group 4	All Groups
Group size	86	505	2488	10513	13592
Diabetes probability	48.8%	27.9%	10.4%	4.7%	6.9%
Women	35%	37%	42%	39%	40%
Younger than 20	1%	0%	1%	2%	2%
Older than 50	75%	71%	56%	42%	46%
Black	25%	21%	16%	12%	13%
Mediterraneans	6%	4%	2%	1%	1%
Low activity level	62%	42%	35%	27%	29%
Family history of diabetes	78%	51%	24%	11%	15%
Less than 110 pounds	3%	3%	2%	2%	2%
Over 190 pounds	67%	45%	28%	11%	16%
Blood Pressure Over 160	58%	34%	21%	8%	12%
Triglycerides Over 200	34%	14%	5%	2%	3%
HDL: Less than 35	15%	10%	4%	2%	3%

Table 1. The Resulted Data Mining Algorithm Groups (Clusters) Characteristics.

When comparing the population with high diabetes probability (groups 1 and 2) with the population with low diabetes probability (group 4), it seems that people with the following characteristics are more likely to become diabetic:

- Age: Patients over 50
- Race: Black
- Race: Mediterraneans
- Activity level on record: Low activity level
- Family disease history: Patients with family history of diabetes
- Body weight: Over 190 pounds
- Blood Pressure: Over 160
- Triglycerides: Over 200

- HDL: Less than 35

We validate our algorithm on the second dataset. We followed the procedures conducted with the full patient list to cluster the validation dataset into the four groups, identified by the algorithm. The results were smoothed using an iterative proportional fitting procedure (see Ben-Zvi and Spiegler, 2007) to ensure that the total number of diabetic patients was equal to the actual total. Predicted and actual values 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 algorithm results.

Patient Group	No. of Patients	Diabetic Patients	
		Actual	Predicted
1	10	5	4.8
2	48	14	13.4
3	256	25	26.6
4	1045	50	49.1
Total	1359	94	93.9

Table 2. Predicted and Actual Number of Diabetic Patients for the Validation Dataset.

6 METHOD EVALUATION AND COMPARISON

Next, we evaluated the results of our DM algorithm and compared them with traditional analysis methods. Considerable research has been conducted to compare performance of different DM techniques on various data sets (e.g., Lim et al. 2000; Wilson et al. 2006). Although there are several popular metrics, such as Receiver Operating Characteristic (ROC) or Area Under the ROC Curve (AUC), 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 by a measurement called "goodness of fit". This is a classification correct rate that is defined as the number of successful predictions (diabetic and non-diabetic patients) divided by the total number of observations:

$$\text{Goodness} - \text{of} - \text{fit} = \frac{\text{Number of successful predictions}}{\text{Number of observations}} \tag{8}$$

We purposely chose this metric because of the imbalanced groups. That is, we use this metric to examine the ability of different algorithms and methods to differentiate between a relatively small group of people with diabetes and a much larger group of healthy people.

In Table 3 we summarize the results of all considered methods in descending order of the goodness-of-fit measure. We used the following methods: (1) linear regression; (2) logistic regression; (3) clustering (using a single linkage technique and a Euclidean Distance as a criterion) and (4) classification (using decision trees C4.5).

Method	**Goodness-of-fit Measure**
Suggested Technique	*83.6*
Clustering	79.4
Classification	77.9
Logistic Regression	69.2
Linear Regression	61.8

Table 3. Summarizing Results of the Different Examined Methods.

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

7 DISCUSSION AND CONCLUSIONS

We demonstrated in this study the powerful capabilities of the model we developed and presented its benefits within the healthcare domain. We made a theoretical contribution, as we exhibit a formal presentation of activities for the DM process, while integrating several applicable concepts from other disciplines.

Our method provides several types of useful insights:

First, our method, incorporating several variables, was shown superior to all other compared traditional methods. Therefore, as a result of our analysis we suggest using the following variables for identifying potentially diabetic patients: family history, activity level, body weight, age, blood pressure, level of triglycerides, HDL and race. Our method may be used as a predictive tool for the institute to perform a more precise and informed patient selection process and to accept patients with low diabetes probability; those less likely to get the disease.

Second, although this study does not attempt to generalize the results to the entire healthcare community, the emerged significant groups are representative of different population types. Obviously, each health institution (e.g., hospitals, clinics) will have its own set of variables that describes its patients. We presume applying the methodology suggested in this research in different institutions will yield different results; however, we expect that the nature of the significant variables is similar across institutions with similar patient populations in the US.

Those findings, together with lessons of experience, can be integrated to create useful knowledge for clinical, management and policy decisions. Taking on the DM approach can also enhance communication and interaction between different health services and health policymakers. We believe that this is a well-suited method to explore and characterize some of the less understood relationships between the different variables or factors that cause or affect diabetes. By furthering our understanding of the different factors that impact diabetes, this research can allow healthcare professional to: (1) establish a national health policy; (2) achieve effective diabetes diagnose and help patient care; (3) facilitate exercising various prevention strategies; and (4) bring down health costs. We leave this further discussion for future research.

Although the presented method is proven to be quite good, it also has its limitations: (a) discretization of continuous numeric data and construction of discrete data intervals may lead, in some cases, to information loss. The 5% significance level we used may not be enough for certain applications, especially since the probability of being a diabetic patient for the entire patient list was only 6.9%; and (b) the presented dataset is based on relational datasets the company maintains. The applicability of the model and the algorithm to other types of databases other institutions may retain is yet to be explored. Varying techniques may lead to different results: we cannot state that there is one best technique for data analysis. Future research can therefore concentrate on determining which technique is suitable for the problem at hand.

References

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., and Grosfeld-Nir, A. (2007) “Serial Production Systems with Random Yield and Rigid Demand: A Heuristic”, Operations Research Letters 35(2), 235-244.

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.

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.

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.

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.

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.

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.

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.

Jain, A.K., Murty, M.N., and Flynn, P.J. (2009), "Data Clustering: A Survey", ACM Computer Surveys, 31(3), 264-323.

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.

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.

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

National Center for Health Statistics (2009), Health, United States, 2008: with chartbook. Hyattsville, Maryland: Department of Health and Human Services.

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.

Pal, S.K. and Mitra, P. (2004) Pattern recognition algorithms for data mining: scalability, knowledge discovery and soft granular computing, Chapman and Hall/CRC Computer Science and Data Analysis Series.

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

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

Wang, J.T.L., Zaki, M.J., Toivonen, H.T.T. and Shasha, D.E. (2005) Data Mining in Bioinformatics: Advanced Information and Knowledge Processing Series, Springer Verlag.

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.

Appendix

The following table contains the relevant variables from the main dataset:

Attribute Description	**Data Type**
Patient SSN	Unique Identifier
Age	Numeric
Birth Place	Qualitative
Current Residence	Qualitative
Gender	Binary
Race	Qualitative
Enrolment date	Numeric
Current Status	Numeric
Premiums	Numeric
Activity Level on record	Qualitative
Family disease history	Qualitative
Body weight on record	Numeric
Blood pressure on record	Numeric
CBC (Complete Blood Count)	Numeric
WBC (Cells White Blood)	Numeric
RBC (Red Blood Cells)	Numeric
HCT	Numeric
MCV (Volume Mean Cell)	Numeric
MCH (Hemoglobin Mean Cell)	Numeric
RDW (Distribution Width Red Cell)	Numeric
PLT (Platelets)	Numeric
Hb (Hemoglobin)	Numeric
Ferritin	Numeric
Transferrin	Numeric
Differential	Numeric
Neut	Numeric
LYMPH (Lymphocyte)	Numeric
MONO (Monocyte)	Numeric
BASO (Basophil)	Numeric
EOS (Eosinohil)	Numeric
Albumin	Numeric
Glucose	Numeric
Electrolytes	Numeric
BUN (Notrogen Blood Urea)	Numeric
CRP (Protein C-Reactive)	Numeric
ESR (Sedimentation Rate Erythrocytes)	Numeric
Triglycerides	Numeric
Cholesterol Total	Numeric
HDL (Lipoprotein High Density)	Numeric
LDL (Lipoprotein Low Density)	Numeric
ALP	Numeric
SGPT	Numeric
AST	Numeric
GGT	Numeric
Bilirubin	Numeric
PT (Prothrombin Time)	Numeric
PPT	Numeric
APTT	Numeric
INR (Ratio Normalized International)	Numeric
Diabetic (YES/NO)	Binary