December 2007

Data Mining and Knowledge Discovery: An Analytical Investigation

Tal Ben-Zvi
Stevens Institute of Technology

Israel Spiegler
Tel-Aviv University

Data Mining and Knowledge Discovery: An Analytical Investigation

Tal Ben-Zvi (*Corresponding Author*)
Wesley J. Howe School of Technology Management, Stevens Institute of Technology, Castle Point on the Hudson, Hoboken, NJ 07030, USA
E-mail: tal.benzvi@stevens.edu

and

Israel Spiegler
Tel-Aviv University, Faculty of Management, Tel-Aviv 69978, Israel
E-mail: spiegler@post.tau.ac.il

Abstract

In recent years, the exponentially growing amount of data made traditional data analysis methods impractical. Knowledge discovery in databases (KDD) provides a framework for alternative methods that address this problem. In this research we follow the KDD process, develop a mathematical model of transforming data and information into knowledge and create a clustering data mining algorithm. To that end, we employ ideas from related, applicable fields (e.g., Operations Research, Inventory Management, and Information Theory). Consequently, we show the merit and value of applying a well-structured model to knowledge acquisition.

Keywords: Knowledge Discovery Process, Data Mining, Binary Representation, 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. A unifying and general approach is that of *knowledge discovery in databases* (KDD), namely discovering patterns in databases. KDD consists of several steps and its goal is to derive useful insights and knowledge from data.

This research follows the KDD process and presents a mathematical model of transforming data and information into knowledge. To that end, we employ ideas, detailed later, from related, applicable fields (e.g., Operations Research, Inventory Management, and Information Theory). We mainly focus on preprocessing steps (data discretization, data reduction and transformation), data mining and interpretation. Also, we create a clustering data mining algorithm and present an empirical affirmation.

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 data mining algorithm for extracting patterns from data. Next, we conduct an investigation with a real-life database and evaluate the obtained results. Finally, we report our interpretation of the outcomes and summarize the study.

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.

The KDD Process and Data Mining

The term "KDD" was first conceived at the first KDD workshop in 1989 (Piatetsky-Shapiro 2000). We use that term as defined by Fayyad et al. (1996): *the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.* KDD consists of the following steps: data selection and preparation, cleansing and pre-processing operations, data reduction and transformation, data mining and interpretation of discovered patterns. This approach regards data mining as a step in the KDD process (see Fayyad et al., 1996; Imberman and Tansel, 2006; Natarajan and Shekar, 2006), and we follow that supposition.

Data mining is the application of specific algorithms for extracting structure from data. Contemporary data mining methods combine innovative computational technologies with analytical techniques 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, clustering, and so on. Today, data mining is applied in panoply of successful applications in many industries and scientific disciplines (Melli et al. 2006). It is used in healthcare settings (Metaxiotis, 2006), 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) —to name a few.

The additional steps in the KDD process, such as data pre-processing, data cleaning, incorporating appropriate prior knowledge, and proper interpretation of the mining results, are essential to ensure that useful knowledge is derived from the data.

Data Representation

When executing steps of the KDD process, we employ the concept of binary database (see Spiegler and Maayan, 1985; Erlich et al., 2003), where data appear 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 transformation is designed for attribute values that are discrete. As later explored, we can discretize any continuous or alphanumeric attribute. Also, 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.

Information Theory Concepts

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

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 when conducting data discretization, data reduction and data mining.

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, 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.

Later, when executing the data mining algorithm, we conduct data assessment and evaluation. For this, we make use of the following Operations and Inventory Management scenarios to arrange the dataset's attributes: A production process is imperfect, due to capacity limitations and technical or environmental factors. Operations managers must meet demands and deal with costs. This production problem is referred to as "Multiple Lot sizing in Production to Order" (MLPO) and is extensively discussed in literature (e.g., Ben-Zvi and Grosfeld-Nir, 2007; Grosfeld-Nir and Gerchak, 2004; Grosfeld-Nir, Anily and Ben-Zvi, 2006; Pentico, 1994).

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.

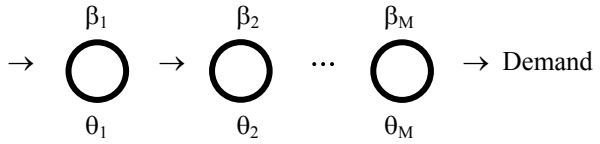

Figure 1. An Example of an MLPO Production System with M machines.

The Model

In this section we develop our model, following several pre-processing activities of the KDD 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 data mining algorithm, detailed later, 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=2, b=3):

$$\text{Info}([2,3]) = -2/5 \times \log 2/5 - 3/5 \times \log 3/5 = 0.971 \tag{6}$$

The 0.971 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 Reduction

This procedure reduces the target dataset of insignificant data to a size appropriate for data mining. We first identify those attributes values that have the most significant effect on the dependent variable. Using the mutual information concept, the procedure identifies the independent attributes that provide the largest amount of mutual information with respect to the dependent variable. Employing the likelihood ratio statistic, it is possible to test the null hypothesis that the dependent variable and the independent attributes are mutually unrelated. If the likelihood ratio statistic is greater than the critical value of the chi-square distribution for a given significance level (probability of false rejection), then we reject the null hypothesis and conclude that the independent attribute does indeed affect the distribution of the dependent variable. In such a case, it is justified to select the attribute, as it is a "relevant" attribute. If we fail to reject the null hypothesis, we may conclude that the relationship between the two is not statistically significant and therefore we may omit the attribute from our dataset (an "irrelevant" attribute). This procedure is repeated for each attribute. We use a 5% significance level for this procedure.

By the end of this procedure, we obtain a dataset containing *n* objects and *d* attributes ($d \le D$). This reduction eases the execution of following KDD process steps, primarily the data mining step.

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 data mining 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 of the knowledge discovery process: the data mining step.

Data Mining

The data mining 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.,

$$\sum_{j=1}^{d}\sum_{k=1}^{p_j}\beta_{j,k}=1 \tag{7}$$

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 $\frac{1-\beta_{j,k}}{1-\theta_{j,k}}$.

The core of the algorithm follows: we use the first sequenced variable to split the population sample into two partitions, corresponding to its two possible values: "0" and "1". After the first portioning, the procedure is repeated 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 (probability of false rejection), then we conclude that the independent variable does affect the distribution of the dependent variable. In this case, it is justified to partition the population into two subpopulations corresponding to the two values of the selected independent variable: "0" and "1". If partitioning is justified, we repeat this procedure for each of the two newly created subpopulations. If, for a particular subpopulation, we fail to reject the null hypothesis that the independent variable providing the largest mutual information and the dependent variable are independent of each other, then we conclude that the relationship between the two is not statistically significant, providing no grounds for partitioning. This procedure terminates when all remaining subpopulations are terminal.

Finally, we segment the subpopulations created by the partitioning procedure into groups that are most similar in terms of the probabilities associated with the dependent variable, while minimizing the resulting loss of mutual information. We rank the subpopulations in ascending order of the dependent variable's occurrence probabilities as estimated from the sample. Next, for each pair of subpopulations ranked adjacently, the loss of information about the dependent variable (that

would result if the two subpopulations were to be combined into a single subpopulation) is calculated. The calculation may be executed using (3). The pair resulting in the smallest loss is identified. Then, calculating the likelihood ratio statistic using (4), where the sample size *n* being equal to the number of observations in the two samples to be combined, we test whether the loss of information is significant. If the statistic is smaller than the critical value of the chi-square distribution for a given probability of false rejection, then we fail to reject the hypothesis that the dependent variable and the indicator variable are independent of each other. Accordingly, we proceed to group the two subpopulations and merge the corresponding samples. This process is repeated until the smallest loss of mutual information becomes statistically significant. This indicates that the best pair of subpopulations being considered for grouping is significantly different, and hence, the grouping procedure terminates.

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.

Model Application

The education domain offers many interesting and challenging applications for data mining. Following our analytical formulation, we now present a real-life application for MBA alumni of a large business school, accredited by the Association to Advance Collegiate Schools of Business (AACSB). The main objective of this application is to test and evaluate the student selection process and its effectiveness. We aim to profile the MBA alumni and to conduct a performance analysis seeking to identify distinction students for the MBA program and improve the admission process. From a utilitarian perspective, the faculty is interested in improving the quality of its students by selecting better students.

Pre-processing Procedures

The dataset we used for this study obtained 1053 MBA alumni (graduating in 2000-2005) and 368 attributes (See Appendix A for a complete list of attributes). Although most attributes are defined as discrete numeric attributes, we discretized attributes taking many possible values as well. Therefore, we discretized the following attributes: (1) Age; (2) Undergraduate GPA; (3) GMAT Total Score; (4) GMAT Verbal Score and (5) GMAT Quantitative Score. We used the MBA GPA as a target attribute, following the school's criterion of graduation with distinction, i.e., MBA GPA equal or greater than 90 and discretized the attributes accordingly.

Next, we followed the data reduction procedure detailed above to obtain the following 10 attributes which the target attribute MBA GPA depends on: (1) Age; (2) Student's Gender; (3) Minor Specialization; (4) Undergraduate GPA; (5) Undergraduate Major Subject; (6) Undergraduate Minor Subject; (7) Undergraduate Institution Name; (8) GMAT Total Score; (9) GMAT Verbal Score; and (10) GMAT Quantitative Score. Then, we applied the data transformation procedure to obtain 116 attributes. The full list of the modified attributes is presented in Appendix B.

Applying the Data Mining Algorithm

Following pre-processing operations, we applied the data mining algorithm detailed above. As a result, the student population was divided into four distinction groups (clusters) defined in Table 1.

Comment: There is no particular significance in the fact that Groups 2 and 4 are of the same size.

Group	No. of Observations	Distinction Prob. (%)
1	23	100.0
2	371	37.7
3	189	13.8
4	371	7.8
Total	954	22.9

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

Validating the Algorithm

We validate our algorithm on a dataset comprised of MBA alumni graduated during the first semester of the 2006 academic year. This dataset includes 43 students. We followed the procedures conducted with the full MBA dataset to cluster the validation dataset into the four distinction groups, identified by the algorithm in the previous section. The results were smoothed using an iterative proportional fitting procedure to ensure that the total number of distinction students was equal to the actual total. Predicted and actual values are presented in Table 2. The results show that the actual distribution of distinction students does not deviate significantly from the prediction made based on the algorithm results ($\chi^2 = 1.6$).

Distinction Group	No. of Observations	Distinction	
		Actual	Predicted
1	1	1	0.7
2	14	4	3.5
3	10	1	0.9
4	18	0	0.9
Total	43	6	6

Table 2. Predicted and Actual Number of Distinction Students for the Validation Sample.

Method Evaluation and Comparison

Next, we evaluated the results of our data mining algorithm and compared them with traditional analysis methods. Considerable research has been conducted to compare performance of different data mining techniques on various data sets (e.g., 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 by a measurement called "goodness of fit". We define the goodness-of-fit measure as the number of successful predictions (distinction and non-distinction students) divided by the total number of observations:

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

In Table 3 we summarize the results of all considered methods in descending order of the goodness-of-fit measure (the results of our proposed method are marked in *italic*). We used the following methods: (1) linear regression (achieving distinction by predicting the MBA GPA); (2) logistic regression (using a similar technique as linear regression); (3) clustering (using a single linkage technique and a Euclidean Distance as a criterion) and (4) classification (using decision trees). Also, we examined the method currently being used by the school. This method incorporates only Undergraduate GPA and GMAT Quantitative Score.

We note that the algorithm's running time is exponential and is a function of the initial data dimensionality. Running time of all methods was comparable.

Method	Goodness-of-fit Measure	Relevant Statistical Data
Proposed Method	*95.8*	
Logistic Regression	83.7	Chi square value=245.124; Cox and Snell R^2=0.227; Nagelkerke R^2=0.367
Linear Regression	81.4	F value=7.03; R^2=0.325; adjusted R^2=0.278
Current Method	81.4	
Clustering	79.1	
Classification	79.1	

Table 3. Summarizing Results of the Different Examined Methods.

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

Utilizing Discovered Knowledge

Our method provides several types of useful insights:

First, according to Table 3, the current used method can correctly identify approximately only 81% of distinction and non-distinction students. Therefore, we may conclude that the current method should be re-evaluated by the school's education committee.

Second, our method, incorporating more variables, was shown superior to all other compared traditional methods. Therefore, we suggest using more variables for the admission process (e.g., GMAT score - both verbal and quantitative, the undergraduate institution and the undergraduate major). Our method may be used as a predictive tool for faculty to perform a more precise and informed student selection process and to accept qualified students; those more likely to succeed in the MBA program and achieve distinction.

Third, although this study does not attempt to generalize the results to all other higher education institutes, the significant distinction groups are representative of the different types of management or business school students. Obviously, each institution (or faculty) will have its own set of variables that describes the distribution of distinction students. 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 student populations.

Although the presented method is proven to be quite good, it also has its limitations: (a) discretization of continuous 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; and (2) the presented dataset is based on relational datasets. The applicability of the model and the algorithm to other types of databases (e.g., multimedia) is yet to be explored.

Conclusions

We reported in this study a mathematical model and an application with impressive results. We demonstrated the powerful capabilities of the model and presented its benefits within the application domain. We made a theoretical contribution, as we exhibit a formal presentation of activities in the KDD process, while integrating several applicable concepts from other disciplines. We believe that the combination of theoretical research and practical considerations discussed herein will augment existing research of knowledge discovery and foster the expansion of its business applications. Yet, varying techniques may lead to different results: we cannot state that there is one best technique for data analysis. The issue is therefore to determine which technique is suitable for the problem at hand. Future research should focus, therefore, on the development of architecture that allows easy synthesis or integration of the wide range of methods and techniques to address contemporary applications.

References

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

Ben-Zvi, T., and Grosfeld-Nir, A. "Serial Production Systems with Random Yield and Rigid Demand: A Heuristic", *Operations Research Letters* (35:2), 2007, pp. 235-244.

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

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

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

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

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

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

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

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. "The KDD Process for Extracting Useful Knowledge from Volumes of Data", *Communications of the ACM* (39:11), 1996, pp. 27–34.

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

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

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

Grosfeld-Nir, A., and Gerchak, Y. "Multiple Lotsizing in Production to Order with Random Yields: Review of Recent Advances", *Annals of Operations Research* (126:1), 2004, pp. 43-69.

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

Imberman S., and Tansel A.U. "Frequent Itemset Mining and Association Rules", In Schwartz, D. (Ed.), *Encyclopedia of Knowledge Management*, Hershey, PA: Idea Group Inc., 2006, pp. 197-203.

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

Khan, S., Ganguly, A.R., and Gupta, A. "Creating Knowledge for Business Decision Making", In Schwartz, D. (Ed.), *Encyclopedia of Knowledge Management*, Hershey, PA : Idea Group Inc., 2006, pp. 81-89.

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

Melli, G., Zaïane, O.R., and Kitts, B. "Introduction to the Special Issue on Successful Real-World Data Mining Applications", *SIGKDD Explorations* (8:1), 2006, pp. 1-2.

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

Natarajan, R., and Shekar, B. "Interesting Knowledge Patterns in Databases", In Schwartz, D. (Ed.), *Encyclopedia of Knowledge Management*, Hershey, PA: Idea Group Inc., 2006, pp.297-304.

Pentico, D.W. "Multistage Production Systems with Random Yield: Heuristics and Optimality", *International Journal of Production Research* (32), 1994, pp. 2455-2462.

Piatetsky-Shapiro, G. "Knowledge Discovery in Databases: 10 Years After" *SIGKDD Explorations* (1:2), 2000, pp. 59-61.

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

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

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

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

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

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

Appendixes

Appendix A

The following table contains the full attribute table of the MBA dataset: The complete dataset consists of 368 attributes.

No. of Attribute(s)	Attribute Name / Description	Data Type
1	Student ID	Numerical - Discrete
1	Age	Numerical - Discrete
1	Birth Place	Qualitative
1	Current Residence	Qualitative
1	Student's Gender	Binary
1	First Semester of MBA	Numerical - Integer
1	Academic Status	Numerical – Integer
1	MBA GPA	Numerical – Continuous
1	Total Courses Weight	Numerical – Integer
1	Major Specialization	Numerical – Integer
1	Minor Specialization	Numerical – Integer
1	Highest Degree hold by the Student	Numerical – Integer
1	Undergraduate GPA	Numerical – Continuous
1	Undergraduate Major Subject	Qualitative
1	Undergraduate Minor Subject	Qualitative
1	Undergraduate Institution Name	Qualitative
1	GMAT Total Score	Numerical – Integer
1	GMAT Verbal Score	Numerical – Integer
1	GMAT Quantitative Score	Numerical – Integer
1	Father Birth Year	Numerical – Integer
1	Father Birth Country	Qualitative
1	Mother Birth Year	Numerical – Integer
1	Mother Birth Country	Qualitative
80	Undergraduate Course Name	Numerical – Integer
80	Undergraduate Course Code	Numerical – Integer
80	Undergraduate Course Grade	Numerical – Integer
35	MBA Course Code	Numerical – Integer
35	Semester of taking the MBA Course	Numerical – Integer
35	MBA Course Grade	Numerical – Integer

Appendix B

The following is the attribute list of the MBA alumni reduced and transformed dataset:

1. Age: <23
2. Age: 23,24
3. Age: 25
4. Age: >25
5. Student's Gender: Male
6. Student's Gender: Female
7. MBA GPA: Distinction
8. Minor Specialization: Organizational Behavior
9. Minor Specialization: Finance-Accounting
10. Minor Specialization: Marketing Management
11. Minor Specialization: Management of Technology and Information Systems
12. Minor Specialization: Strategy and Entrepreneurship
13. Minor Specialization: Operations Research and Decisions
14. Minor Specialization: None
15. Undergraduate GPA: ≤76.5
16. Undergraduate GPA: >76.5 and ≤79
17. Undergraduate GPA: >79 and <81
18. Undergraduate GPA: ≥81 and ≤83
19. Undergraduate GPA: >83 and ≤84
20. Undergraduate GPA: >84 and ≤85
21. Undergraduate GPA: >85 and ≤86
22. Undergraduate GPA: >86 and ≤86.5
23. Undergraduate GPA: >86.5 and ≤87.5
24. Undergraduate GPA: >87.5 and <88
25. Undergraduate GPA: ≥88 and <88.5
26. Undergraduate GPA: ≥88.5 and <89
27. Undergraduate GPA: ≥89 and <91
28. Undergraduate GPA: ≥91 and ≤91.5
29. Undergraduate GPA: >91.5 and <93
30. Undergraduate GPA: ≥93 and ≤94
31. Undergraduate GPA: >94 and <95
32. Undergraduate GPA: ≥95 and <96
33. Undergraduate GPA: ≥96 and ≤97
34. Undergraduate GPA: >97
35. Undergraduate Major/Minor: Accounting
36. Undergraduate Major/Minor: Agriculture Economics
37. Undergraduate Major/Minor: Art
38. Undergraduate Major/Minor: Behavioral Sciences
39. Undergraduate Major/Minor: Biology
40. Undergraduate Major/Minor: Chemistry
41. Undergraduate Major/Minor: Communication
42. Undergraduate Major/Minor: Computer Sciences
43. Undergraduate Major/Minor: Criminology
44. Undergraduate Major/Minor: Dental Medicine
45. Undergraduate Major/Minor: East Asia studies
46. Undergraduate Major/Minor: Economics
47. Undergraduate Major/Minor: Education
48. Undergraduate Major/Minor: Electronics
49. Undergraduate Major/Minor: Engineering
50. Undergraduate Major/Minor: English
51. Undergraduate Major/Minor: Film
52. Undergraduate Major/Minor: French
53. Undergraduate Major/Minor: General Studies
54. Undergraduate Major/Minor: Geography
55. Undergraduate Major/Minor: Hebrew Studies
56. Undergraduate Major/Minor: Insurance
57. Undergraduate Major/Minor: International Relations
58. Undergraduate Major/Minor: Law Studies
59. Undergraduate Major/Minor: Linguistics

60. Undergraduate Major/Minor: Literature
61. Undergraduate Major/Minor: Management
62. Undergraduate Major/Minor: Mathematics
63. Undergraduate Major/Minor: Middle East Studies
64. Undergraduate Major/Minor: Nutrition
65. Undergraduate Major/Minor: Pharmacy
66. Undergraduate Major/Minor: Philosophy
67. Undergraduate Major/Minor: Physics
68. Undergraduate Major/Minor: Political Science
69. Undergraduate Major/Minor: Psychology
70. Undergraduate Major/Minor: Social Sciences
71. Undergraduate Major/Minor: Sociology
72. Undergraduate Major/Minor: Statistics
73. Undergraduate Major/Minor: None
74. Undergraduate Institution: Tel-Aviv University
75. Undergraduate Institution: The Hebrew University
76. Undergraduate Institution: The Technion
77. Undergraduate Institution: Ben-Gurion University
78. Undergraduate Institution: Bar-Ilan University
79. Undergraduate Institution: Haifa University
80. Undergraduate Institution: The Open University
81. Undergraduate Institution: College of Management
82. Undergraduate Institution: The Academic College
83. Undergraduate Institution: The Interdisciplinary Center
84. Undergraduate Institution: Other Colleges
85. Undergraduate Institution: Institutes outside Israel
86. GMAT Total Score: ≤440
87. GMAT Total Score: >440 and ≤520
88. GMAT Total Score: 530
89. GMAT Total Score: ≥540 and ≤570
90. GMAT Total Score: 580
91. GMAT Total Score: 590,600
92. GMAT Total Score: 610
93. GMAT Total Score: 620,630
94. GMAT Total Score: 640
95. GMAT Total Score: ≥650 and ≤680
96. GMAT Total Score: 690,700
97. GMAT Total Score: ≥710 and ≤740
98. GMAT Total Score: >740
99. GMAT Verbal Score: <12
100. GMAT Verbal Score: 12,13,14,15
101. GMAT Verbal Score: 16,17,18,19,20
102. GMAT Verbal Score: 21,22,23
103. GMAT Verbal Score: 24,25
104. GMAT Verbal Score: 26,27
105. GMAT Verbal Score: 28
106. GMAT Verbal Score: 29,30
107. GMAT Verbal Score: 31,32
108. GMAT Verbal Score: 33,34
109. GMAT Verbal Score: 35-42
110. GMAT Verbal Score: >42
111. GMAT Quantitative Score: <44
112. GMAT Quantitative Score: 44
113. GMAT Quantitative Score: 45,46
114. GMAT Quantitative Score: 47
115. GMAT Quantitative Score: 48,49
116. GMAT Quantitative Score: >49