MINING DATA QUALITY IN COMPLETENESS

(Research Paper)

Shouhong Wang
University of Massachusetts Dartmouth, USA
swang@umassd.edu

Hai Wang
Saint Mary's University, Canada
hwang@smu.ca

Abstract: Completeness is an important attribute of data quality. This paper discusses the measures of completeness of data in a data set. It proposes a data mining model based on self-organizing maps (SOM) to visualize the patterns of missing values in a data set to assess the data quality in completeness.

Key Words: Data Quality, Completeness, Missing Values, Data Mining, Self-Organizing Maps.

INTRODUCTION

Information quality is important to organizations [9, 15]. People use information attributes as a tool for assessing information quality. In this approach information quality is measured based on users' as well as experts' opinions on the information attributes. The commonly known information attributes for information quality including accuracy, objectivity, believability, reputation, access, security, relevancy, value-added, timeliness, completeness, amount of data, interpretability, ease of understanding, concise representation, and consistent representation [5, 6, 10, 13]. As the boundary between data and information can never be unambiguous, these attributes can also be applicable to data quality. In this paper we focus on the data quality in terms of completeness.

Commonly, one can rarely find a data set that contains complete entries. According to "10 potholes in the road to information quality" [12], the common causes of incompleteness include:

- data are produced using subjective judgments, leading to omission;
- systemic errors in information production lead to lost data;
- access to data may conflict with requirements for security, privacy, and confidentiality; and
- lack of sufficient computing resources limits access.

The extent of damage of missing data is unknown when it is virtually impossible to return the data source for completion. The incompleteness of data is vital to data quality [3, 4].

There have been several traditional approaches to handling missing values in data analysis including eliminating from the data set those records that have missing values [8] and imputations [2]. However, these traditional approaches can lead to biased analysis results and invalid conclusions. Research [14] has pointed out that the properties of missing values must be taken into account in assessing the data quality of a data set.

One of a simple index to measure data quality in completeness is the missing rate which is equal to the number of incomplete entities over the number of total entities. However, this simple index ignores the pattern of missing values. For instance, suppose we have two data sets, A and B, with the same missing

rate. In data set A, missing values spread almost evenly over all attributes. In data set B most missing values occur in only some attributes. The quality of data set A is better than that of data set B, since the complete data in data set A are representative while the complete data in data set B are biased. This paper is to discuss the patterns of missing values in data sets, and proposes a model to detect patterns of missing values based on self-organizing maps (SOM) [7]. An experiment on real-world data is employed to demonstrate the usefulness of the proposed model in assessing data quality in completeness.

PATTERNS OF MISSING VALUES

Missing at Random (MAR)

Missing values often randomly distributed throughout the sample space. There is no particular assumption on the reason of value missing. Few correlations among the missing values can be observed if the missing values have the MAR pattern. Since values are missing at random, the missing values distribute almost equally towards each attribute. The quality of the entire data set is homogeneous. Accordingly, complete data can be considered representative for the entire data set.

Missing in Cluster (MIC)

Data are often missing in some attributes more than in others. Also, missing values in those attributes can be correlated. It is difficult to use statistical techniques to detect multi-attribute correlations of missing values. The quality of data with this pattern of missing values is less homogeneous than that with MAR. Applications of any analytical results based on the complete data set should be cautious, since the sample data are biased in the attributes with a large number of missing values.

Systematic Irregular Missing (SIM)

Data can be missing highly irregularly, but systematically. There might be too many missing correlations between the attributes, but these correlations are too tedious to analyze. An implication of SIM is that the data with complete entities are unpredictably under-representative. The quality of data with this pattern of missing values is much less homogeneous than that with MAR and also less controllable than that with MIC. Applications of any analytical results based on the complete data set are highly questionable.

When the data set is high-dimensional, statistical tests for these missing patterns might become powerless due to the curse of dimensionality. On the other hand, the interest of a data pattern depends on the purpose of the data analysis and does not solely depend on the estimated statistical strength of the pattern [11].

SELF-ORGANIZING MAPS

Self-Organizing Maps (SOM)

The self-organizing maps (SOM) method based on Kohonen neural network [7] has become one of the promising techniques in cluster analysis for data mining. SOM-based cluster techniques have advantages over statistical methods in cluster analysis to discover patterns of missing values. The SOM method does not rely on any assumptions of statistical tests, and is considered as an effective method in dealing with

high-dimensional data. Since real world data often do not have regular multivariate distributions, traditional statistical methods have their limitations in these cases. On the other hand, the SOM method has demonstrated its flexibility and usefulness in cluster analysis because of the relaxation of statistical assumptions [1, 7]. The SOM method provides a base for the visibility of clusters of high-dimensional data. This feature is not available in any other cluster analysis methods. It is extremely important for our study since we are interested in not only individual clusters, but also the global patterns of data related to missing values.

SOM map the high-dimensional data onto low-dimensional pictures, and allow human to view the clusters. Two-layer SOM are used in this research. The nodes at the lower layer (input nodes) receive inputs presented by the sample data points. The nodes at the upper layer (output nodes) will represent the organization map of the input patterns after the unsupervised learning process. Every low layer node is connected to every upper layer node via a variable connection weight. The unsupervised learning process in SOM can be briefly described as follows. The connection weights are assigned with small random numbers at the beginning. The incoming input vector presented by a sample data point is received by the input nodes. The input vector is transmitted to the output nodes via the connections. The activation of the output nodes depends upon the input. In a so-called "winner-take-all" competition, the output node with the weights most similar to the input vector becomes active. In the learning stage, the weights are updated following Kohonen learning rule [7]. The weight update only occurs for the active output node and its topological neighbors. In this one-dimensional output case, we assume a linear neighborhood. The neighborhood starts large and slowly decreases in size over time. Because the learning rate is reduced to zero, the learning process will eventually converge. The weights will be organized such that nodes that share a topological resemblance are sensitive to inputs that are similar. The output nodes in SOM will thus be organized and represent the real clusters in the self-organizing map. The function of SOM is illustrated in Figure 1.

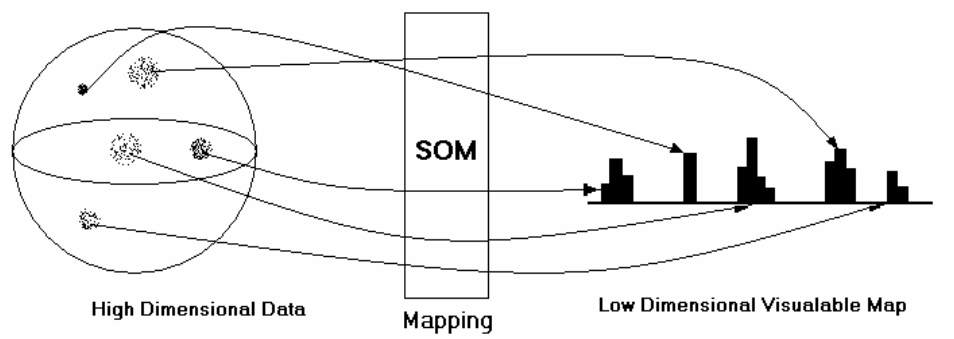

Figure 1. SOM Maps High Dimensional Clusters onto Low Dimensional Map

Detecting Patterns of Missing Values

The method of visualizing and detecting patterns of missing values is described as follows. The first step of this method is to convert entities with missing values into shadow entities by assigning a value 1 for the data item with missing value and 0 for the data item without missing values. For example, suppose we have an entity with five data items [4, M, 2, 3, 5] where M indicates a missing value. Its shadow entity is [0, 1, 0, 0, 0].

Shadow entities are then presented to the SOM and they are self-organized on the map represented by the output nodes of the SOM. The map is then depicted graphically. The one-dimensional map is almost the

same as histogram in statistics. The horizontal axis represents locations of output nodes of the SOM. The height of a bar indicates the number of data points which activate the output node of the SOM at the corresponding location.

The three patterns of missing values in data set can be illustrated in Figure 2. Figure 2(a) shows a case where no significant cluster is generated by the SOM. Apparently, this pattern shows MAR. The cluster circled in Figure 2(b) represents the pattern MIC. It indicates that a substantial number of entities have missing values in certain attributes concurrently. If one traces these shadow entities back to their original multivariate data, specific attributes that are correlated in missing values can be found. Figure 2(c) shows SIM patterns where there are too many irregular bumpy clusters.

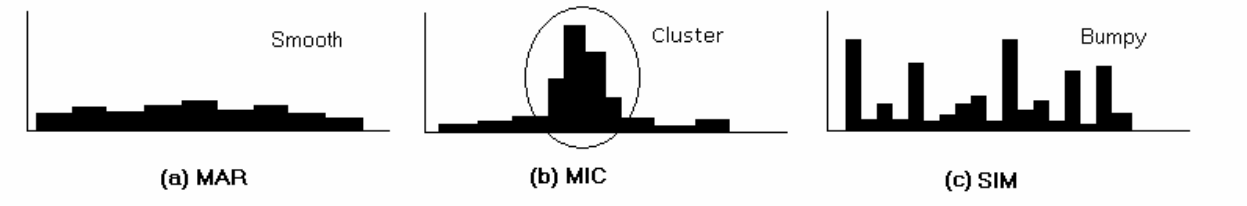

Figure 2. Patterns of Missing Values in Data set

To validate the patterns of missing values visualized from SOM, one might make changes to the parameters of the SOM network, including the neighborhood function and its change rate, and the learning rate and its change rate, to regenerate several self-organizing maps. The conventional split-half validation method can also be used to validate the patterns detected. Using this pattern discovery method, the experiment process and validation are conducted simultaneously, and an adequate result can always be obtained.

AN EXPERIMENT WITH REAL-WORLD DATA

Student evaluation of instruction methods are widely used at universities to evaluate the teaching performance of instructors. The data used in this experiment came from a student evaluation of instruction survey at a Canadian university. In this case, twenty questions describe the characteristics of an instructor's performance. Each question is rated on a five-point scale for students to answer. A high mark for a question indicates a positive answer to the question. Usually, many students do not answer all the questions for whatever reasons. The quality of the survey data in completeness should be measured by missing rates as well as the patterns of missing values. This experiment places the focal point on the patterns of missing values.

The experiment is to compare the data quality of the data sets from two faculties. We depict the map in two parts. The upper part represents missing patterns of an evaluation data set from one faculty, and the lower part represents that of the other faculty.

Sample data of 2000 incomplete entities (1000 for each faculty) were collected for this experiment. These entities with missing values were converted into shadow entities for SOM. In making trials of SOM, we started with the extremely large amount of output nodes in accordance with the number of the training samples (i.e., 2000 in this case), and then gradually reduced the number of output nodes to find a clear feature map. Our SOM program was implemented in C++, and had a data conversion interface with

Microsoft Excel for graphics visualization. Figure 3 shows the result using the SOM with 200 output nodes, 200 nodes for the initial neighborhood, the initial learning rate of 0.01, and 2000 learning iterations. The map presented a clear comparison of the patterns of missing values in the two data sets. Apparently, the data quality of the data corresponding to the upper part is better than the other.

The original data entries corresponding to the shadow data entries of the SOM were extracted and were further analyzed to identify the nature of the incompleteness in the survey. Our focal point was placed on the difference between the two data sets. It was found from the clustered data that three questions were often skipped concurrently by students who were responsible for survey data set-2. These three questions were: "Tests and assignments are reasonable measures of student learning"; "Where appropriate, student work is graded promptly"; and "Test and assignments provide adequate feedback on student progress." Such correlation of missing data and the pattern of missing data are hard to discovery by using other methods. The implication of the incompleteness of data in this case was that any conclusion based only on the complete data of the survey data set-2 would be biased.

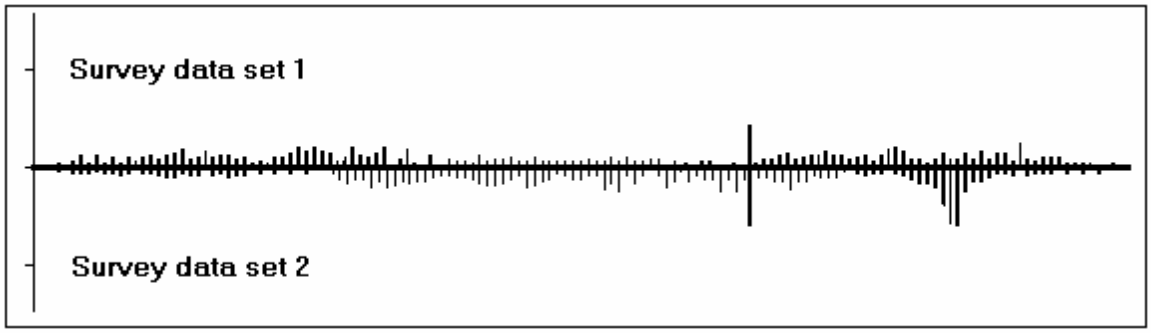

Figure 3. The Self-Organizing Maps for the Missing Values

CONCLUSIONS

In this paper, we have discussed the patterns of missing values in the context of data quality in completeness. As a data mining tool, SOM has been considered a useful technique in visualized clustering analysis when statistical clustering methods are difficult to apply. This paper proposes a SOM-based visualized pattern discovery method for missing values in data set. Through the real-world case, it has been shown the usefulness of this method.

Completeness is one of the important attributes of data quality. The ultimate objective of data quality assessment is to fully understand the characteristics of the data set and determine strategies for the data analyses. The proposed method has certain advantages in assessing data quality. This data mining method requires few assumptions, but provides visual presentations of patterns of missing values so that the user of the method is allowed to detect the quality of data and initiate pertinent strategies for validating data analysis results based on the data quality.

REFERENCES

[1] Deboeck, G., and Kohonen, T. *Visual Explorations in Finance with Self-Organizing Maps*, London, UK: Springer-Verlag, 1998.
[2] Dempster, A. P., Laird, N. M., and Rubin, D. B., Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society,* B39(1), 1997, 1-38.
[3] English, L. P., "Help for data quality problems -- A number of automated tools can ease data cleansing and help improve data quality," *InformationWeek*, Oct 7, 1996, 53.
[4] English, L. P., Information quality for business intelligence and data mining: Assuring quality for strategic information uses, 2005. <http://support.sas.com/news/users/LarryEnglish_0206.pdf> [retrieved April 1, 2007].
[5] Garvin, D. A., *Managing Quality*, The Free Press, New York, 1988.
[6] Huang, K. T., Lee, Y. W., Wang, R. Y., *Quality Information and Knowledge*, Prentice-Hall, New York, 1999.
[7] Kohonen, T. *Self-Organization and Associative Memory*, 3rd Ed. Berlin: Springer-Verlag, 1989.
[8] Little, R. J. A., and Rubin, D. B., *Statistical Analysis with Missing Data, 2nd Ed*. New York: John Wiley and Sons, 2002.
[9] Salaun, Y. and Flores, K., "Information quality: meeting the needs of the consumer," *International Journal of Information Management*, 21(1), 2001, 21-37.
[10] Salmela, H., "From information systems quality to sustainable business quality," *Information and Software Technology*, 39(12), 1997, 819-825.
[11] Silberschatz A, and Tuzhilin A., What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering*, 5(6), 1996, 970-974.
[12] Strong, D., Lee, Y. W., and Wang, R. Y., 10 potholes in the road to information quality, *IEEE Computer*, 30(8), 1997, 38-46.
[13] Tozer, G., *Metadata Management for Information Control and Business Success*, Artech House, Norwood, MA, 1999.
[14] Wang, H, and Wang, S., Data mining with incomplete data, in *Encyclopedia of Data Warehousing and Mining*, John Wang (Ed.), Idea Group Inc.: Hershey, PA, 2005, pp.293-296.
[15] Wang, R. Y., Lee, Y. W., Pipino, L. L., and Strong, D. M., "Manage your information as a product," *Sloan Management Review*, 39(4), 1998, 95-105.