ISSN:0975-9646

Decision Theoretic in SQL to Prepare Data Sets for Data Mining Analysis

Venkata Changal Rao , T. Rajesh, Sk. Karimulla

CSE, Sri Sunflower College of Engineering & Technology,
Lankapally, Krishna Dist,A.P, India.

Abstract- The data set for analysis is generally the most time consuming task in a data mining project, requiring many complex SQL queries, joining tables and aggregating columns. Existing SQL aggregations have limitations to prepare data sets because they return one column per aggregated group. In general, a significant manual effort is required to build data sets, where a horizontal layout is required. In this paper we proposed a decision-theoretic framework for evaluating data mining systems, which employ classification methods, in terms of their utility in decision-making. The decision-theoretic model provides an economic perspective on the value of "extracted knowledge," in terms of its payoff to the organization, and suggests a wide range of decision problems that arise from this point of view. The relation between the *quality* of a data mining system and the amount of investment that the decision maker is willing to make is formalized. We propose two ways by which independent data mining systems can be combined and show that the combined data mining system can be used in the decision- making process of the organization to increase payoff. Examples are provided to illustrate the various concepts, and several ways by which the proposed framework can be extended are discussed. Keywords: Classification, data mining, data mining economics, decision-making, knowledge discovery systems. Aggregation, data preparation.

I. INTRODUCTION

Several agencies, businesses, and nonprofit organizations in order to support their short and long-term planning activities are searching for a way to collect, store, analyze, and report data about individuals, households, or businesses. relational database, especially with normalized tables, a significant effort is required to prepare a summary data set [16] that can be used as input for a data mining or statistical algorithm [17], [15]. Most algorithms require as input a data set with a horizontal layout, with several records and one variable or dimension per column. That is the case with models like clustering, classification, regression and PCA; consult [10], [15]. Each research discipline uses different terminology to describe the data set. In data mining the common terms are point-dimension. Statistics literature generally uses observation-variable. Machine learning research uses instance-feature. The data acquisition systems (such as minicomputers, microprocessors, transducers, and analog-to-digital converters) that collect, analyze, and transfer data are in use in various mid-range and large organizations [2], [4]–[7]. Over time, more and more current, detailed, and accurate data are accumulated and stored in databases at various stages. This data may be related to designs, products, machines, materials, processes, inventories, sales, marketing, and performance data and may include patterns, trends, associations, and dependencies. The data collected contain valuable information that could be integrated within the organization strategy, and used to improve organization decisions. The large amount of data in current databases, which contain large number of records and attributes that need to be simultaneously explored, makes it almost impractical to manually analyze them for valuable decision-making information. The need for automated analysis and discovery tools for extracting useful knowledge from huge amounts of raw data suggests that knowledge discovery in databases (KDDs) and data mining methodologies may become extremely important tools in realizing the above objectives. Some researchers often define data mining as the process of extracting valid, previously unknown, comprehensible information from large databases in order to improve and optimize organization decisions [5], [23]. Other researchers use the term KDD to denote the entire process of turning low-level data into high-level knowledge, where data mining is considered as a single step in the process that involves finding patterns in the data. To avoid confusion, we choose the later definition. The KDD process is defined in [5] as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data." According to [2], although data mining is at the core of the KDD process, it is just one step in the overall KDD process, and it usually takes about 15 to 25% of the overall effort. The KDD process often includes the following important stages [5]. The first step involves understanding the application domain in which the data mining is applied and the goals of the data mining process. The second step includes selecting, integrating, and checking the target data set. The target data set may be defined in terms of the records as well as the attributes of interest to the decision-maker. The third step is data preprocessing. This includes data transformation, handling missing or unknown values, and data cleaning (this can be done by applying algorithms in order to remove unreliable and erroneous data). In the fourth step, data mining for extracting patterns from data takes place. This involves model and hypothesis development and the selection of appropriate data mining algorithms. The fifth step involves interpreting and presenting results for the decision-maker. Fayyad *et al.* [5] distinguish between two main categories of data mining (fourth step above):

verification-oriented and *discovery- oriented*. Verification-oriented techniques focus mainly on testing preconceived hypotheses (generated by the decision- maker) and on fitting models to data. Discovery-oriented methods focus on autonomously finding new rules and patterns and are classified as *descriptive* or *predictive*. Descriptive methods include *visualization* techniques (e.g., scatter plots and histograms) and *clustering* (e.g., identifying subgroups of silicon wafers that have a similar yield). Predictive methods include *regression* and *classification*. Regression is concerned with the analysis of the relationships between attribute values within the same record and the automatic production of a model that can predict attribute values for future records. Classification methods assign records to predetermined classes. For example, medical patients may be classified according to the outcome of their treatment; thus, the most effective treatments for new patients can be identified. In this paper, our focus is on data mining systems that employ classification methods. While much research has been conducted on making optimal cost-sensitive classification decisions [24], there is virtually no rigorous and formal research related to the question of *action ability*—the ability of the "extracted knowledge" to suggest concrete and profitable action by the decision- makers [8], [9], [14], [16]. The difficulty of determining the value of "mined data" and the tangible benefits resulting from investing for an organization to investment in the KDD process keeps many organizations from fully exploiting the affluence of data that is generated and collected during daily operations. The importance of this question increases even more when considering that the market for data mining has grown from $50 million in 1996 to $800 million in 2000 [7]. Moreover, many organizations use data mining as a strategic tool in order to become more competitive. The purpose of this paper is to develop a framework for evaluating data mining systems, which use classification methods, in terms of their value in decision-making. Our framework is based on the belief that the question of evaluating data mining systems can only be addressed in a *utilitarian* framework, that is, the patterns extracted by the data mining system are effective only to the extent that the derived information leads to action that increases the *payoff* of the decision-maker (see [8] and [18] for a similar view). The decision-theoretic framework developed in this paper connects the organization's strategic objectives with KDD investment and data mining quality. This helps in understanding how KDD benefits change as a function of the deployment cost of the KDD process, what should be the optimal investment in KDD, and what is the nature of the relationship between the organizational strategy and data mining quality. Our modeling approach also enables us to address the question of evaluating different data mining processes when making decisions.

II. Optimal KDD Process

The basic model presented in Section II describes an environment in which the decision-maker creates a decision rule to optimize the expected payoff given a confusion matrix, a payoff function, and prior probabilities of actual classes. In this section, we extend the basic model by defining a subspace of square and symmetric confusion matrices whose values reflect the quality of classification as a function of its cost. Larger investments in the KDD process will typically provide the decision-maker with classification of a higher quality. For example, the decision-maker would like to know how much to invest in a KDD process in order to support a credit screening application for credit cards. In order to increase the expected payoff, the credit company would like to base its decision whether to approve an applicant or not based on a data mining process with low class-conditional error rates. The quality of the data mining process, however, may become higher as the financial investments increases. the decision-maker is willing to invest more for a KDD process if this would ensure a better process in terms of the relationship "*more effective*" presented in Definition 1. With a better KDD process, the decision-maker increases the expected payoff, or at least does not worsen it. The following results examine the relationship between the cost and the effectiveness of . It will be shown that as the decision-maker invests more in the KDD process, a "*more effective*" confusion matrix is obtained (yielding no less expected payoff *regardless of payoff or prior probabilities information*).

III. Composite Classification

The decision-maker already employs a KDD process associated with a confusion matrix . By Theorem 2, the decision- maker can improve the quality of the overall process by investing in an independent KDD process associated with a confusion matrix . Moreover, Theorems 3 and 4 show that investing more in the second KDD process renders the overall neural network classifier and a decision tree classifier, which is used to classify examples
coming from the *same* set of actual credit risk classes: *low risk, medium risk, or high risk*. The outputs (i.e., predicted classes) of these classifiers are then combined (as discussed below), and the decision-maker chooses whether to approve an applicant or not based on the composite classification. Data mining systems are probabilistically independent if the probability of deciding by one data mining system that an example of actual class belongs to class does not depend on the classification produced by the other data mining system. Optimized in KDD. The basic model presented in Section II describes an environment in which the decision-maker creates a decision rule to optimize the expected payoff given a confusion matrix, a payoff function, and prior probabilities of actual classes. In this section, we extend the basic model by defining a subspace of square and symmetric confusion matrices whose values reflect the quality of classification as a function of its cost. Larger investments in the KDD process will typically provide the decision-maker with classification of a higher quality. For example, the decision-maker would like to know how much to invest in a KDD process in order to support a credit screening application for credit cards. In order to increase the expected payoff, the credit company

would like to base its decision whether to approve an applicant or not based on a data mining process with low class-conditional error rates. The quality of the data mining process, however, may become higher as the financial investments increases. Next, we show how to incorporate the investment cost within the basic decision-theoretic framework introduced. Let denotes the investment cost of the KDD process, and assume that the data mining measure of performance is defined in terms of a confusion matrix of size , where The reason that the diagonal elements, as well as the off-diagonal elements of are equal, reflects the uniform prior probabilities that the decision-maker assigns to the class-conditional errors. Note also that the off-diagonal elements decreases when the investment cost increases, thus reflecting the fact that a larger investment will provide the decision-maker with classification of a higher quality. In practice, the function may be determined by applying a procedure of fitting a parametric function to historical data relating the error rates of similar data mining processes to their investment costs . Often, the decision-maker already employs a KDD process associated with a confusion matrix . By Theorem 2, the decision- maker can improve the quality of the overall process by investing in an independent KDD process associated with a confusion matrix . Moreover, Theorems 3 and 4 show that investing more in the second KDD process renders the overall The decision-maker's goal is to maximize the *expected net payoff* by stating an optimal decision rule , as well as an optimal investment cost in the second KDD process. One of the most successful applications of data-mining is performed in "database marketing" [19]–[21]. Database marketing is a method that enables marketers to develop customized marketing strategies based on extracted patterns derived from customer databases [19]. For example, by employing database marketing, local retailers can reach customers with the "best fit" offer and products at the right time and geographical area. As another example, telephone companies have identified and segmented high-valued customers (called "power users"). Data mining is, then, used to determine which terms and products to offer to people in this high-valued segment. In this section, we present an example in which the marketers of a chain store wish to develop a marketing strategy that utilizes the knowledge resulting from data mining. To increase the number of sales and the amount of customer satisfaction, the marketers want to make sales promotion offers by direct mails to selected customers. In order to increase the expected payoff, the marketers determine which customer classes in a list to mail to based on patterns extracted from customer information originating from sales transactions.

IV. DISJOINT DATA MINING SYSTEMS

We have shown how several independent data mining processes that are used to classify examples coming from the *same* set of actual classes can be combined using the Cartesian operator into one data mining process. In this section, we address cases where the decision-maker applies independent data mining processes that are used to classify examples coming from *disjoint* sets of actual classes. We say that such data mining processes are *disjoint*. To illustrate, we might use a decision tree classifier used to classify examples coming from the actual "credit risk" classes *good* or *bad* and a neural network classifier used to classify examples coming from the actual "credit usage" classes *heavy* or *light*. We show how to combine two data mining processes used to classify examples coming from disjoint and independent sets of actual classes by introducing the *doubly Cartesian* operator. We prove that the effectiveness of the doubly Cartesian process is *not less* than any of the component data mining processes from which the doubly Cartesian process is formed. We also show that improving one of the data mining processes provides more effective process for the decision-maker, regardless of the quality of the component data mining processes.

QUERY RESULTS

Possible queries on a database of transactions could be selection and projection which may also involve statistical operations, and maybe a temporal extension to those. In terms of data mining, users would like to know what are the maximal set of items purchased having a count greater than a threshold value. These types of queries cannot be answered by standard querying tools. But, queries such as what is the count of transactions where milk and bread are purchased together can be answered by the standard querying tools. Users may also ask queries that return a set of transactions in the database. Given the time-stamp of each transaction, users may want to write queries with a temporal dimension such as how many customer transactions for April, 2002 contain both milk and bread. Among statistical operations, min, max, and average does not make sense in a database of binary values where quantities of items sold, or price information is not involved. Therefore, we will not consider these types of queries in our discussion. Queries that return a set of transactions may be considered as micro-queries and the queries that return only statistical information can be considered as macro-level queries. Given a set of rules Rh that are hidden from the database D, we can construct a view of the database Dv, where Dv is a subset of D which consists of the transactions not modified by the hiding process. IDs of modified transactions could be released so that Dv could be easily constructed. Macro-level queries whose results are a subset of Dv return correct results. The rest of the queries may return incorrect results. This is also true for queries that involve a temporal dimension. In order to improve the correctness of temporal queries, the hiding process may be biased over older transactions, this way ensuring the correctness of queries over more recent transactions. Queries that contain count operation return correct results if they are issued over items that are not used by the hiding process. We can also give a maximum error range for the count queries which can be used by the user to have a rough idea of the error in the returned count values. This maximum error could be the highest support reduction percentage among the items used by the hiding strategies. In terms of data mining, the user would like to obtain association rules from the database.

In the previous section, we have already discussed the side effects of the hiding process in terms of the hidden or newly appearing association rules.

V. CONCLUSION

The process of extracting valid, previously unknown, comprehensible information from large databases and using it to make crucial organization decisions. Global, national, and even local organizations are driven by information, which is uncovered by the data mining process. Nowadays, data mining has become an essential core of KDDs and therefore, their quality must be improved as much as possible in order to guarantee successful KDD processes. Although evaluating the quality of the data mining process is one of the most pressing challenges facing KDD research today, few organizations have effective ways of managing data mining quality, which is so important to their competitiveness. This paper considers data mining quality as a main goal to achieve, instead of a sub-product of database creation and KDD development processes. To this end, we developed a decision theoretic approach for evaluating data mining systems, which employ classification methods, in terms of their utility in decision making. The decision-theoretic model was developed in order to provide an economic perspective on the value of "extracted information," in terms of its payoff to the organization, and to suggest a wide range of decision problems that arise from this point of view. In the decision-based approach, the decision- maker observes predicted classes as determined by the data mining classification system and chooses actions accordingly. The decision-maker wishes to maximize the expected payoff by choosing an optimal decision rule.

REFERENCES

[1] J. Marschak, "The economics of information systems," in *Frontiers of Quantitative Economics,*M. Intrilligator, Ed. Amsterdam, The Netherlands: North-Holland, 1971, pp. 43–107.

[2] T. M. Mitchell, *Machine Learning*. New York: McGraw-Hill, 1997.

[3] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan-Kaufmann, 1988.

[4] G. Piatetsky-Shapiro and C. J. Matheus, "The interestingness of deviations," in *Proc. Knowledge Discovery Data Mining*, 1994, pp. 25–36.

[5] L. H. Setiono and H. Liu, "Effective data mining using neural networks," *IEEE Trans. Know. Data Eng.*, vol. 8, pp. 957–961, Dec. 1996.

[6] A. Silberschatz and A. Tuzhilin, "What makes patterns interesting in knowledge discovery systems," *IEEE Trans. Know. Data Eng.*, vol. 8, pp. 970–974, Dec. 1996.

[7] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, pp. 81–106, 1986.

[8] M. J. Berry and G. Linoff, *Data Mining Techniques: For Marketing, Sales, and Customer Support*. New York: Wiley, 1997.

[9] J. R. Bult and T. Wansbeek, "Optimal selection for direct mail," *Marketing Sci.*, vol. 14, no. 4, pp. 378–381, 1995.

[10] S. H. Ha and S. C. Park, "Application of data mining tools to hotel data mart on the intranet for database marketing," *Expert Syst. Applicat.*, vol. 15, no. 1, pp. 1–31, July 1998.

[11] N. Ahituv and B. Ronen, "Orthogonal information structures—A model to evaluate the information provided by second opinion," *Dec. Sci.*, vol. 19, pp. 255–268, July 1988.

[12] R. Brachman and T. Anand, "The process of knowledge discovery in databases: A human-centered approach," in *Advances in Knowledge Discovery and Data Mining*, U. Fayyad, G. Piatetsky-Shapiro, S. P. Amith, and R. Uthurusamy, Eds. Cambridge, MA: MIT Press, 1996, pp. 37–58.

[13] J. S. Demski, *Information Analysis Reading*. Reading, MA: Addison-Wesley, 1972.

[14] U. Fayyad, "Data mining and knowledge discovery: Making sense out of data," *IEEE Expert*, vol. 11, pp. 20–25, Oct. 1996.

[15] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery: An overview," in *Advances in Knowledge Discovery and Data Mining*, U. Fayyad, G. Piatetsky-Shapiro, S. P. Amith, and R. Uthurusamy, Eds. Cambridge, MA: MIT Press, 1996, pp. 1–36.

[16] , "The KDD process for extracting useful knowledge from volumes of data," *Commun. ACM*, vol. 39, no. 11, pp. 27–34, Nov. 1996.

[17] M. Goebel and L. Gruenwald, "A survey of data mining and knowledge discovery software tools," *SIGKDD Explorations*, vol. 1, no. 1, pp. 20–33, June 1999.

[18] J. Kleinberg, C. Papadimitriou, and P. Raghavan, "A microeconomic view of data mining," *Knowl. Disc. Data Mining*, vol. 2, no. 4, pp. 311–324, Dec. 1998.

[19] B. M. Masand and G. Piatetsky-Shapiro, "A comparison of approaches for maximizing business payoff of prediction payoff," *in Conf. Proc. Knowledge Discovery Data Mining*, pp. 195–201, 1996.

[20] G. Luo, J.F. Naughton, C.J. Ellmann, and M. Watzke. Locking protocols for materialized aggregate join views. *IEEE ransactions on Knowledge and Data Engineering (TKDE)*, 17(6):796–807, 2005.

[21] C. Ordonez. Horizontal aggregations for building tabular data sets. In *Proc. ACM SIGMOD Data Mining and Knowledge Discovery Workshop*, pages 35–42, 2004.

[22] C. Ordonez. Vertical and horizontal percentage aggregations. In *Proc. ACM SIGMOD Conference*, pages 866–871, 2004.

[23] C. Ordonez. Integrating K-means clustering with a relational DBMS using SQL. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 18(2):188–201, 2006.

[24] C. Ordonez. Statistical model computation with UDFs. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 22, 2010.

[25] C. Ordonez. Data set preprocessing and transformation in a database system. *Intelligent Data Analysis (IDA)*, 15(4), 2011.

[26] C. Ordonez and S. Pitchaimalai. Bayesian classifiers programmed in SQL. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 22(1):139–144, 2010.

[27] S. Sarawagi, S. Thomas, and R. Agrawal. Integrating association rule mining with relational database systems: alternatives and implications. In *Proc. ACM SIGMOD Conference*, pages 343–354, 1998.