8-15-1997

Integrating Logistic Regression with Knowledge Discovery Systems

D.J Berndt
University of South Florida, berndt@bsn.usf.edu

R.K Satterfield
University of South Florida, rsatterf@bsn01.bsn.usf.edu

Integrating Logistic Regression with Knowledge Discovery Systems

D. J. Berndt and R. K. Satterfield
Information Systems and Decision Sciences
College of Business Administration
University of South Florida
{berndt@bsn.usf.edu, rsatterf@bsn01.bsn.usf.edu}

Introduction

Interest in data warehousing continues to grow as more systems support large-scale data collection [Inmon 1992] [Mattison 1996]. The primary role of a data warehouse is to support decision making, executive information systems, and knowledge discovery. However, there is a dearth of efficient systems that help managers make use of these very large data repositories. Inmon [1992] defined data warehousing as follows, clearly highlighting the role of decision support, and therefore the need for easy-to-use systems that will assist managers in gaining the greatest advantage from data resource investments.

A *data warehouse* is a subject oriented, integrated, nonvolatile, time variant collection of data in support of management's decisions.

Our focus is on knowledge discovery systems [Fayyad et al. 1996]. In particular, we draw a distinction between discovery techniques that are based on user expertise (i.e., domain knowledge) and those that rely on more generic search processes, such as neural networks or statistical techniques [Cheng and Titterington 1994; Elder and Pregibon 1996]. Domain knowledge can originate with users or be derived via automated discovery techniques. Some discovery techniques focus on database attributes and the property of generalization that allows us to organize our knowledge hierarchically [Walter 1980; Cai et al. 1991; Dhar and Tuzhilin 1993; Berndt 1995]. Figure 1 shows a hierarchy of concepts from a small student loan database. If a rule-based discovery process is used, the problem lies in discovering relationships between attributes at an appropriate level of generalization.

In this paper we discuss integrating domain knowledge and statistical techniques. We rely on domain knowledge when available, but use other techniques when domain knowledge is lacking. Logistic regression is proposed as a supplementary approach to assessing attribute importance without relying explicitly on domain knowledge.

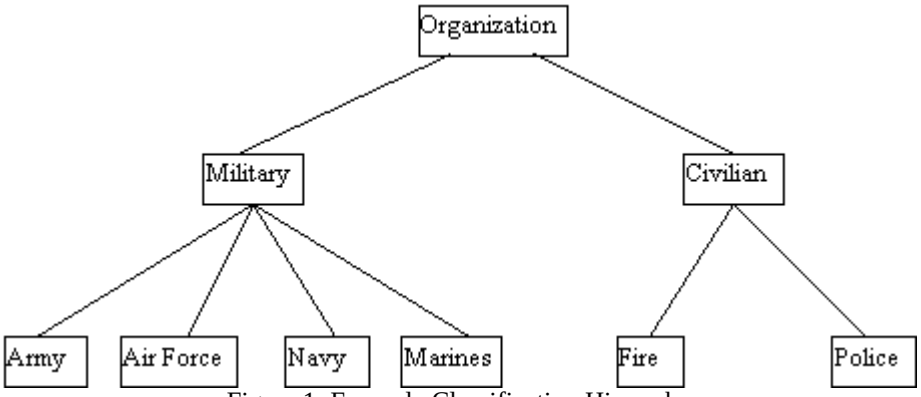

Figure 1: Example Classification Hierarchy

The Role of Domain Knowledge

Analysts using knowledge discovery systems are certainly guided by their knowledge of the task at hand. By capturing some of that expertise knowledge discovery systems can better focus the search process. Results from such systems should also more closely match the interests of users. Therefore, the problem-specific domain knowledge or expertise is important because:

1. domain knowledge reduces the size of the search space and,
2. discovered knowledge is more likely to be "interesting" since it reflects user-defined concepts.

However, domain knowledge can be a two-edged sword. It is useful in focusing the search process, but this focus may also bias the result toward user expectations. There is a trade-off between the relevance domain knowledge can provide and the lack of new ideas resulting from a search process wedded to past experience.

A second concern is the cost incurred in acquiring and maintaining the store of domain knowledge. Explicit domain knowledge requires labor-intensive maintenance supported by a user interface and representation language. These demands are exacerbated in the typical case where multiple users must coordinate their activities.

Classification Hierarchies

Figure 1 shows that attributes can be organized as classification hierarchies. This is a powerful method of organizing domain knowledge for use in knowledge discovery. For instance, the AX system uses several types of links, including classification links, to organize domain knowledge as concept networks [Berndt 1995]. These structures then provide a context within which search processes may operate. Hierarchical structures let us reduce the range of attribute values, for instance, from specific military organizations to a military/civilian dichotomy. Therefore, a logistic regression model can be simplified by generalizing attributes into fewer categorical values. Continuous attributes such as salary are amenable to similar treatment. For instance, a salary attribute can be made

discrete by forming ranges, and these ranges may be coalesced to form more general categories (e.g., "high income").

Search

With a foundation of domain knowledge we can use experience-based search techniques to explore potential rules supported by the data warehouse. For example, we might discover a rule that holds for police departments and through generalization or specialization try the rule with respect to other military organizations, i.e. the Navy. In general, the search process generates candidate rules for evaluation against the underlying database through the use of domain knowledge. When there is little domain knowledge logistic regression can be used to select potentially interesting attributes for exploration. Ideally several assessment methods would be available to form composite measures of "interestingness" and determine future search directions. Such interest measures at the attribute level can be presented to the user, or used in fully automated search processes.

Logistic Regression

Logistic regression is an iterative technique especially suited to problems characterized by binary variables or variables with small numbers of categorical values [Hosmer and Lemeshow 1989]. A potential problem with logistic regression is that convergence may require a substantial computational effort, and may never materialize in some circumstances. We propose using the logistic regression output to develop relative rankings of attribute importance with respect to a particular dependent variable. These rankings may be valuable when the domain knowledge is too sparse to effectively guide the search process.

Standard logistic regression procedures form a contingency table in memory over which many iterations are performed in determining a proper set of coefficients. While this contingency table may be very large, its size can be estimated from the set of attributes and the range of their values. An important characteristic from a data warehousing perspective is that the contingency table is populated using a single sweep over the database. The logistic regression technique then relies on the manipulation of a virtual memory-resident table, avoiding additional database access. Data retrieval can be further reduced by basing the technique on a sample of the database rather than exhaustive analysis.

Non-Convergence and Separation

In logistic regression certain properties of the data may lead to infinite parameter estimates and non-convergence. The notion of "separation" is one of the pitfalls that can cause such non-convergence. *Complete separation* occurs when attribute values are perfectly correlated with outcomes. This results in some cells in the contingency table being zero, leading to undefined odds ratios. *Quasi-complete separation* occurs when most attribute values exhibit similar behavior. This problem is less onerous, but still may cause excessive iterations or non-convergence. Currently, statistical routines test for

separation after a threshold number of iterations are performed, issuing a warning if separation is detected. The statistical analyst will usually begin an informal process of dropping variables to alleviate the problem.

The notion of separation may not be as difficult an obstacle to our use of logistic regression as it is for more traditional model building approaches. We are applying the technique to form attribute rankings, and thus need not pursue problematic logistic regression calculations in their entirety. We can form relative rankings for most attributes while "pruning" those that cause separation. Such pruning operations could be based on marking zero entries as the contingency table is being formed. These entries lead to separation problems during iterations. If separation is anticipated, we can collapse the contingency table through inexpensive means and continue with the simpler set of variables. Additionally, separation is most likely associated with small data collections, not typically encountered in a true data warehouse environment.

Conclusions

In order to integrate logistic regression into our current knowledge discovery system, we plan to pursue the following.

- Use the properties of classification hierarchies to collapse the number of variable categories.
- Implement a custom logistic regression module that employs "pruning" to avoid problems with complete or quasi-complete separation.
- Incorporate the logistic-based attribute rankings in the knowledge discovery process by giving the information to the user and allowing any automated search techniques to utilize the rankings.

In summary, we believe there can be synergies in combining techniques based on domain knowledge with various statistical tools. Logistic regression is particularly well-suited to this role since it can readily provide relative rankings of attribute importance in knowledge discovery processes with reasonable resource requirements. We will verify the usefulness of such a combination through manual simulation of the search process using a statistical package to calculate the coefficients for input into a knowledge discovery system. We will also develop a custom logistic regression module for inclusion within our prototype knowledge discovery system.

References

[Berndt 1995] D. J. Berndt. AX: searching for database regularities using concept networks. *Workshop on Information Systems and Technology*, December, 1995.

[Cai, Cercone, and Han 1991] Y. Cai, N. Cercone, J. Han. Attribute-oriented induction in relational databases. In G. Piatetsky-Shapiro and W. J. Frawley, editors, *Knowledge Discovery in Databases*, AAAI Press, 1991.

[Cheng and Titterington 1994] B. Cheng and D. M. Titterington. Neural networks: a review from a statistical perspective. *Statistical Science*, 9(1), 1994.

[Dhar and Tuzhilin 1993] V. Dhar and A. Tuzhilin. Abstract-driven pattern discovery in databases. *IEEE Transactions on Knowledge and Data Engineering*, 5(6), December, 1993.

[Elder and Pregibon 1996] J. Elder and D. Pregibon. A statistical perspective on KDD. In U. Fayyad et al., editors, *Advances in Knowledge Discovery and Data Mining*, AAAI Press, 1996.

[Fayyad et al. 1996] From data mining to knowledge discovery. *AI Magazine*, 17(3):37-54, Fall 1996.

[Hosmer and Lemeshow 1989] D.W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. John Wiley & Sons, Inc. 1989.

[Inmon 1992] W. H. Inmon. *Building the Data Warehouse*. John Wiley & Sons, Inc. 1992.

[Mattison 1996] R. Mattison. *Data Warehousing: Strategies, Technologies, and Techniques*. McGraw-Hill, 1996.

[Walker 1980] A. Walker. On retrieval from a small version of a large database. In *Proceedings of the VLDB Conference*, 1980.