2000

A Customer Segmentation Mining System on the Web Platform

Yao-Tsung Chen
National Sun Yat-Sen University, ytchen@venus.mis.nsysu.edu.tw

Bingchiang Jeng
National Sun Yat-Sen University, jeng@mail.nsysu.edu.tw

A Customer Segmentation Mining System on the Web Platform

Yao-Tsung Chen, Department of Information Management, National Sun Yat-sen University, Taiwan, ytchen@venus.mis.nsysu.edu.tw

Bingchiang Jeng, Department of Information Management, National Sun Yat-sen University, Taiwan, jeng@mail.nsysu.edu.tw

Abstract

We will introduce a knowledge discovery system developed on the World Wide Web platform in this paper. Its algorithm is based on Fuzzy Inductive Learning Method (FILM), which can segment consumers' behavior from a set of customer data with noises. In a visualization way, the system will present the acquired knowledge as a set of IF-THEN rules that can be run on top of an expert system. Moreover, the system will provide advices in response to a user's request through the network and a friendly user interface. At last, we evaluate the function of the system by training it with a transaction database provided by a local automobile dealer.

Introduction

Since the emergence of computers in business applications, people use computers to store various kinds of data. In particular, after the database systems were invented, the data is getting much more bigger. Mining knowledge from this huge data set is any managers' wish. The procedure to do this has been called knowledge discovery or data mining. One of the most popular applications of knowledge discovery is customer segmentation, which some researchers (Simoudis, 1996) defined as "the process of analyzing data about actual customers or general consumers to identify characteristics and behaviors that can be exploited in the marketplace". Customer segmentation is important because the sales strategy nowadays is changed from production-oriented to customer-oriented (Berkman and Gilson, 1986). Any corporation that wants to survive in the competing age must consider to collecting this kind of knowledge.

Much work has been done on this topic in recent years. Statisticians used statistical methods to analyze data in very early time (Simoudis, 1996). In recent years, using statistic or other quantitative methods to discover a potential market from huge consumer data was widely adopted. These kinds of applications were usually called database marketing. According to Donnelley Marketing Inc.'s annual survey of promotional practices, 56% of manufacturers and retailers currently have or are building a database, an additional 10% plan to do so, and 85% believe they'll need database marketing to be competitive past the year 2000 (Berry, 1994). However, because there are limitations of statistical methods and the models developed using by regression techniques could only account for certain characteristics in the data, these methods were not able to explain the meaning of data sufficiently (Simoudis, 1996).

Machine learning techniques in the artificial intelligent area have been also introduced in this research. Based on different assumptions, they were able to explain knowledge behind a set of data more sufficiently, and also represent knowledge as a set of easily understandable rules. More over, rules can be integrated into an expert system as a knowledge base. Any user can take its advice by answering some questions. Some companies have already used such kind of knowledge to help managers to make decisions.

Although much work has been done on research of customer segmentation, little attention has been paid to the fuzzy nature of this problem. For example, if 30 thousands dollars is a threshold to classify a customer's income, then one with US$30000 annually income is considered high-pay, while another with US$29999 will be considered low. This is unfair. So one of the two purposes in this research is to use a kind of fuzzy technique called FILM (Jeng and Jeng, 1993; Jeng et al., 1997) to solve this problem.

Another purpose is to explore the potential of applying this technique to Electronic Commerce (EC). Since EC environments are increasingly mature, more and more trades now are done over Internet. A system that can show information of other customers' buying behavior might be beneficial for both buyers and sellers. For buyers, they can consult an online expert system from a desktop PC to find what other buyers do in a certain product. For sellers, they may utilize this knowledge to design an even more attractive program to promote its products. Amazon online bookstore is a good example for this idea that has a primitive system to suggest related books to customers. For this reason, we will implement the system on the Web platform.

The remainder of this paper is organized as follows. Section 2 describes design concepts and principles, especially FILM. Section 3 gives an overview of the system's architecture. Section 4 describes a real example

of using this system in an automobile dealer's database. The last section concludes the paper with a short discussion.

Figure 1. Crisp partition of attribute space

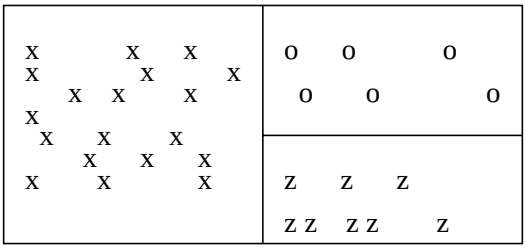
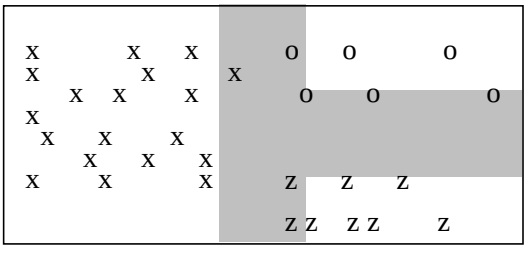


Figure 2. Fuzzy partition of attribute space



What is FILM

Regular induction learning methods have a known shortcoming that the hurdle values for attribute segmentation are crisp. This is inconsistent with human information processing. The crisp nature of the hurdle values also affects the robustness of the induced decision trees. Because attribute segmentation is determined by training cases, the resulting knowledge model based on crisp rules will be very sensitive to noises in the training data. For example, in the credit card application, an applicant making US$30000 annually might be considered good, while another person making US$29999 will be considered bad. As shown in Figure 1, the cutting lines are so sharp that a data point near boundary might be miss-classified into a wrong category if a tiny perturbation exists.

This above problem is, in fact, well-known, but is not handled very well for a long time. Fuzzy Inductive Learning Method, abbreviated as FILM (Jeng and Jeng, 1993; Jeng et al., 1997), is an improvement in this problem. FILM integrates the fuzzy set theory into a tree induction process to overcome the above deficiency. A major advantage of this fuzzy approach is that it makes the classification process be more flexible and the resulting tree be more accurate due to the reduced sensitivity to slight changes of hurdle points. The empirical studies (Jeng and Jeng, 1993; Jeng et al., 1997) have confirmed this hypothesis by showing that FILM greatly improve existing methods including discriminant analysis and ID3 (Quinlan, 1979).

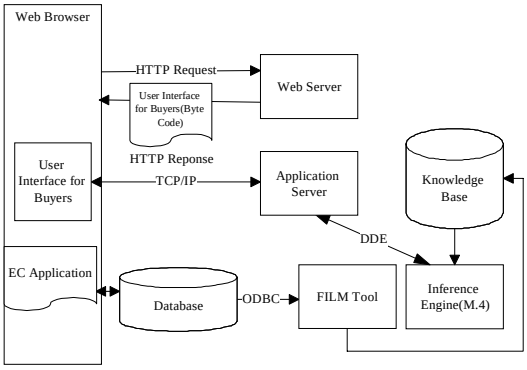
Prior to FILM, there are other studies that also tried to introduce fuzzy sets to handle the hurdle values of continuous variables. Most of the empirical evaluation however, does not show good improvement on doing this. Their approaches fuzzify the attribute variables too earlier before induction learning. Hence it not only solves the problem of crisp hurdle values but also fuzzifies the correctness of a decision tree. The secret of FILM to avoid this unwanted side-effect is to let the fuzzification process independent of induction learning. FILM will use a crisp decision tree generated by either ID3 or other methods as a base one and then applies the fuzzification operations to improve it. The overall process is as follows:

1. Use ID3 or a similar algorithm to create a regular decision tree from a set of training data.
2. Fuzzify the hurdle values in each branch of the decision tree.
3. Re-calculate the class memberships of each tree leaf by re-do the training process against the modified decision tree (i.e. "fuzzify the class" that each tree leaf belongs to). The result is a tree in which the hurdle values and leaf classifications are both fuzzy.
4. When applies the tree to determine the classification of a new case, the inference result must be defuzzified back to a single conclusion.

The purpose of the last step is to convert the result back to a single value in order to be consistent with most of the applications needs; otherwise we can save this step. Coincidentally, Uthurusamy et al. (1991) defined the concept of inconclusiveness and illustrated why ID3-like algorithms are bound to result in overspecialized classifiers when trained on inconclusive data. According to their definition, "a training set of examples that are known to be noise free is said to be inconclusive if there exists no set of rules using only the given attributes that classifies all possible examples perfectly". They proposed that the proper way of handling inconclusive data is to resort to probabilistic rather than categorical classification rules, and **a rule can predict more than one outcome.** They also addressed that most of real-world databases are inclusive.

The transaction data in electronic commerce applications are usually inclusive because customers belong to the same category may not necessarily buy the same product. For this reason, we will save the last step of FILM to keep multiple fuzzy conclusions in each leaf node of a decision tree.

Figure 3. System Architecture



System Overview

In order to apply FILM into customer segmentation system on the Web platform, we must design an Internet-based architecture. The architecture of this system is shown in Figure 3. The graphic user interface and network communication protocols of a client were each implemented as a Java Applet running on top of a Web browser. To keep the Web server's architecture simple, there was no CGI or ASP program written to communicate with the FILM tool. Instead, the FILM tool and the application server, which handles all protocols except HTTP, are both independent programs and implemented in C language. Both of them run on a machine other than that of the Web server. The reason is that passing parameters via CGI is less flexible; also the workload of the Web server can be decreased.

The architecture shown in Figure 3 is not unique for Internet applications. There are other approaches also proposed. Which of them is the best has no conclusion at this moment. In the following sections, four major modules: the FILM tool, the application server, the inference engine and the user interface for buyers are described in detail.

FILM Tool

This module includes two major components: the FILM engine and the user interface for sellers. The FILM engine is the implementation of FILM. It also transforms a fuzzy decision tree into uncertainty rules. Class memberships in a leaf node of the fuzzy tree were each mapped to a confidence factor in a rule. Some friendly text is also embedded into rules to make buyers more comfortable when they use the system.

The user interface for sellers visualizes a fuzzy decision tree generated by the FILM engine, which is presented intuitively to sellers. In practice, sellers in EC can see the discovered knowledge including a complete decision tree and cutting information in each node through this interface.

Application Server

Independent of the Web server, the application server that plays as a middleman is a program between the browser and the inference engine. It communicates with the user interface for buyers on one hand, and the inference engine on the other hand. We implemented two communication protocols to realize this idea: (1) Dynamic Data Exchange (DDE) and (2) TCP/IP. DDE is an inter-process protocol on the Microsoft Windows platform, which was used to communicate with the inference engine. TCP/IP is a standard Internet protocol, which was used to communicate with the user interface for buyers. The design concept and implementation detail of this module is referred to Lee and Chen's research (Lee and Chen, 1997a; 1997b; 2000).

Inference Engine

We adopt M.4 (Cimflex Teknowledge) as the inference engine in this system. It supports forward/backward inference mechanisms, uncertainty rules, and DDE interface that can be communicated with other programs on the Microsoft Windows platform.

User Interface for Buyers

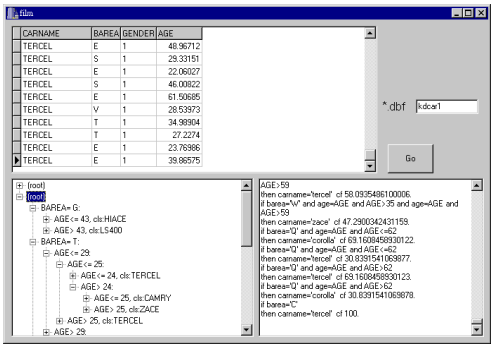
Buyers can take an advice from the expert system, which applies the rules generated by the FILM tool, through this interface. It receives events from users' input such as mouse clicking and text editing, and then encodes them into a message stream which will be transferred to the inference engine through Internet. The message stream from the inference engine is also decoded by this interface, and then the decoded messages will be shown on the right place.

An Example

In this section, we will present a physical application of the system. The customer data set was obtained from a local automobile dealer. It contains a half-month of data (202 records) on September 1999. The only useful attributes in each record are customers' age, birth place and gender, and from which we want to predict a customer's purchase decision from 11 types of cars. The "type of car" field must be placed in the first column in the data set as shown in the top pane in Figure 4

If a sales manager of the automobile dealer wants to segment customers' purchase behavior from the data set, he must choose the data set name, and press the *Go* button. Then the discovery process begins. As the result of the discovery process, two decision trees are shown in the bottom-left pane. The first node represents the root of a regular decision tree; the second node represents a fuzzy decision tree. Each tree can be expanded into any layer at the manager's will. The bottom-right pane shows the corresponding rules of each decision tree.

Figure 4. User interface for a sales manager



On the other hand, if a customer also wants to take an advice from the system, he can access the user interface for buyers, which is shown in Figure 5. There are four buttons on the top of the window. Buyers can click on *Go* to start the inference from the beginning; *Restart* to review the inference result without answering any questions again; *Stop* to stop the inference anytime to change the answers from beginning; *Cache* to see existing data in the working memory of the expert system. The other panes in the window are for displaying output messages such as the inference process, conclusions, status flags and questions, respectively.

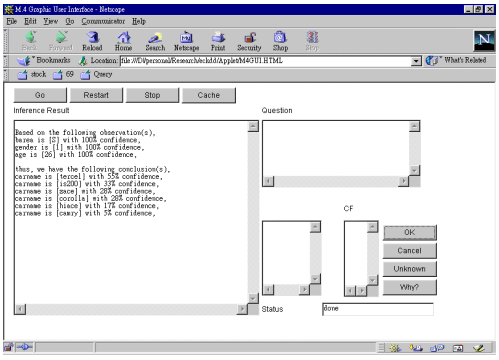
A customer can input his answers in response to the system's questions via the text box in the bottom-right corner of the window. There are more than one ways to input the data. If it is a numeric data type, the answer will be typed in; otherwise it will be selected from a list of items. In either case, a confidence factor will be associated and should be given by the customer. After a customer type in his answer, he can press *OK* to send out, *Cancel* to re-type or *Unknown* to answer unknown. He can also press *Why?* button to ask the inference engine for reasons about the questions.

An example advice provided by the system is shown in Figure 5. A customer who is 26 years old male and born in Kaohsiung will receive a list of other buyers' favor: Tercel(55%), Ls200(33%), Zace(28%), Hiace(17%), and Camry(5%).

Conclusion

In this paper, we have presented a customer segmentation mining system with a fuzzy technique called FILM and illustrated an example to explore the potential of applying this system to EC. The example shows that buyers can consult an online expert system from a desktop PC to find what other buyers do in a certain product and sellers may utilize this knowledge to design an even more attractive program to promote its products.

Figure 5. User interface for buyers



We recognize that our data is not sufficient to show the benefits of this system; interactions between buyers and sellers through this system are also not clear in the simple example. Further work will be in the following directions: (1) a complete experiment to demonstrate how buyers and sellers can be beneficial from this system. (2) a further investigation of interaction changes between buyers and sellers after introducing this system.

Acknowledgement

This research was supported in part by the National Science Council of Taiwan under contracts NSC-88-2416-H-110-044.

Reference

Berkman, H. W., and Gilson C., *Consumer Behavior : Concepts and Strategies*, Boston,1986.

Berry, J., A potent new tool for selling: database marketing, *Business Week* (Sep 5), 1994, pp. 37-41.

Jeng, B., and Jeng, Y.-M., A fuzzy tree induction learning method, in: *Proceeding of The 1st Asian Fuzzy Systems Symposium*, Singapore, 1993.

Jeng, B., Jeng, Y.-M., and Liang, T.-P., FILM: a fuzzy inductive learning method for automated knowledge acquisition, *Decision Support Systems* (21), 1997, pp. 61-73.

Lee, C., and Chen, Y.-T., An embedded visual programming interface for intelligent information retrieval on the Web, in: *Proceeding of IEEE Knowledge & Data engineering Exchange Workshop 1997*, CA, 1997.

Lee, C., and Chen, Y.-T., An embedded visual programming interface for intelligent information retrieval on the Web, in: *Proceeding of National Computer Symposium 1997 Republic of China*, Tai Chung, 1997, pp. F69-F74.

Lee, C., and Chen, Y.-T., Distributed visual reasoning for intelligent information retrieval on the Web, *Interacting with Computers* (12), 2000, pp. 445-467.

Quinlan, J. R., Discovering rules from large collections of examples: a case study, in: D. Michie (eds.), *Expert Systems in the Micro Electronic Age*, Edinburgh University Press, Edinburgh, 1979.

Simoudis, E., Reality check for data mining, *IEEE Expert* (11), 1996, pp. 26-33.

Uthurusamy, R., Fayyad, U. M., and Spangler, S., Learning useful rules from inconclusive data, in: G. Piatetsky-Shapiro and W.J. Frawley (eds.), *Knowledge Discovery in Databases*, AAAI Press, California, 1991, pp. 141-157.