Interactive Knowledge Discovery from Marketing Questionnaire Using Simulated Breeding and Inductive Learning Methods

Takao TERANO
Graduate School of Systems Management,
The University of Tsukuba, Tokyo
3-29-1 Otsuka, Bunkyo-ku, Tokyo 112, Japan
terano@gssm.otsuka.tsukuba.ac.jp

Yoko ISHINO
Interdiciplinary Course on Advanced
Science and Technology, The University of Tokyo
4-6-1 Komaba, Meguro-ku, Tokyo 153, Japan
ishino@ai.rcast.u-tokyo.ac.jp

Abstract

This paper describes a novel method to acquire efficient decision rules from questionnaire data using both simulated breeding and inductive learning techniques. The basic ideas of the method are that simulated breeding is used to get the effective features from the questionnaire data and that inductive learning is used to acquire simple decision rules from the data. The simulated breeding is one of the Genetic Algorithm (GA) based techniques to subjectively or interactively evaluate the qualities of offspring generated by genetic operations. In this paper, we show a basic interactive version of the method and two variations: the one with semi-automated GA phases and the one with the relatively evaluation phase via the Analytic Hierarchy Process (AHP). The proposed method has been qualitatively and quantitatively validated by a case study on consumer product questionnaire data.

Introduction

Marketing decision making tasks require the acquisition of efficient decision rules from noisy questionnaire data. Unlike popular learning-from-example methods, in such tasks, we must interpret the characteristics of the data without clear features of the data nor pre-determined evaluation criteria. This causes serious KDD problems. Traditionally, statistical methods have been used for these analyses, however, conventional techniques in statistics are too weak because they usually assume the linearity of the models and the form of distributions of the data. During the statistical analysis, emphasis has been placed on understanding trends after identifying target data. Furthermore, marketing requires use of quantitative as well as qualitative analysis. Unfortunately, there are no statistical tools to facilitate to satisfy both requirements simultaneously.

Based on the above background, this paper proposes a novel method to acquire efficient decision rules from questionnaire data. In the following sections, we will describe a method to solve the feature selection problem in inductive learning by the simulated breeding (Dawkins 1986, Sims 1992), genetic algorithms (Goldberg 1989), and the AHP (Saaty 1980) methods. As a result, it will be possible to develop a decision tree with comparatively smaller number of features and which incorporates human subjective evaluations.

Problem Description

First, the techniques used in the method are summarized as follows: (1) Simulated breeding is one of the GA-based techniques to evolve offspring via user interaction based on human preference without explicit evaluation functions; (2) As inductive learning tool, we adopts C4.5 (Quinlan 1993) a noise tolerant successor of ID3, which gives a decision tree or a set of rules from data with attributes-value pairs; and (3) The Analytic Hierarchy Process (AHP)(Saaty 1980) hierarchically decomposes a given problem into its smaller constituent parts and then evaluates the weights of these sub-problems by pair-wise comparison judgements.

Next, in a saturated market domain such as oral care products, marketing decision analysts as domain experts must determine the promotion strategies of new products according to the abstract image of the products to be produced. However, in the task domain, although we can only gather noisy sample data with complicated models, it is critical to get simple but clear rules to explain the characteristics of the products in order to make decisions for promotion.

Third, the difficult points of the research are that 1) the questionnaire data intrinsically involve noises, 2) a distribution of data cannot be previously assumed, 3) selection of appropriate features of the data is inevitable, because of the difficulty in interpreting the results of analysis incorporating all the various features, and 4) we do not know how to define the evaluation criteria in advance for effective explanation.

Forth, the focuses of this research are 1) to classify noisy questionnaire data with multiple features, 2) to select necessary and sufficient features to explain the characteristics of the data, and 3) to generate effective interpretations provided by decision trees or a set of decision rules.

Algorithm for Acquiring Decision Rules

```
Step 1: Initialization
Select plural number of image words to be explained
by decision rules.
Randomly select $m$ sets of individuals with $l$ selected
features

Repeat Steps 2-4 until
an appropriate decision tree or a set of decision rules is
obtained.
    Step 2: Inductive Learning
    Apply an inductive learning program to the selected
    $m$ individuals.
    Obtain the corresponding decision trees or sets of
    decision rules.
    step 3: Interactive Evaluation
    From among the obtained decision trees or decision
    rules, two are selected by a user based on the simplicity,
    correctness & reliability, and the understandability
    of the decision tree.
    step 4: Application of Genetic Operations
    Select the best two offspring as parents, apply
    uniform-crossover operations to them in order to get
    new sets features, then generate corresponding offspring.
```

Figure 1: SIBILE-I Algorithm

The procedure of the basic method, SIBILE-I is shown in Figure 1 (Terano, et al. 1995). We call both the algorithm and the system equipped with it SIBILE[1].

In Step 1, we define a set of target concepts to be explained by decision trees. By defining and explaining plural image words simultaneously, we try to solve multi-objective optimization problems. Then, we generate the initial population. The m and l respectively represent the number of individuals and the length of their chromosomes. The number m in simulated breeding is set to very small compared with standard GA-based applications. The chromosomes to represent the features are coded in binary strings, in which a '1' (respectively '0') means that a feature is (not) selected for inclusion in the inductive learning process in Step 2.

In Step 2, the data acquired from the questionnaire is aggregated, each of which has the corresponding features in it. Then the m sets of the data are processed by inductive learning programs.

In Step 3, a user or a domain expert must interact with the system. This is a highly knowledge-intensive task. The domain expert judges them based on simplicity, understandability, accuracy, reliability, plausibility, and applicability of the represented knowledge.

The trees selected in Step 3 are set as parents, and in Step 4, new product characteristics are determined by genetic operations. The GA techniques we have adopted are based on the Simple GA found in (Goldberg 1989). The corresponding chromosomes of the selected decision trees become parents for genetic operations. We apply uniform-crossover operations to them in order to get new sets of features to broaden the variety of offspring.

Steps 2 to 4 are repeated until an appropriate decision tree or set of decision rules is obtained. As are illustrated in (Dawkins 1986), the steps required to obtain the appropriate results are very small. In our experiments, it usually takes only less than 10 steps.

Two Variations

Algorithm with Interactive- and Automated-Phases

As stated in the previous section, Step 3 of SIBILE-I requires highly knowledge intensive tasks and times. Furthermore, in the steps in SIBILE-I, the once omitted features of the data will not appear anymore, because the algorithm does not employ mutation operations in GAs. To improve this, we develop a half-automated version: SIBILE-II: first, in the interactive phase, we subjectively evaluate decision trees or sets of decision rules to get the biases of features in the data as is used in the previous section, then in the automated phase, using the biases of the features given in the interaction, genetic operations are applied to develop offspring with *effective* features.

Algorithm with relative evaluation phase via the AHP

This variation: SIBILE-III facilitates the interactive evaluation of offspring represented by decision trees and/or sets of decsion rules. The user feels it convenient to evaluate them pairwisely, instead of comparing them entirely. The evaluation result can be validated by the consistency indices used in the AHP. The variation is easy to implement: we only add the following two sub-steps in Step 3 of SIBILE-I: 1) the interactive pairwise comparison phase and 2) the weight computation phase from the pairwise comparison matrix in the AHP.

[1] 'Sibyl' in old French, which stands for Simulated Breeding and Inductive LEarning.

Experimental Results

To validate the effectiveness of the proposed method, we have carried out intensive experiments from a practical case study on consumer product questionnaire data. This section describes the experimental results.

Methods

Questionnaire data to investigate the features of new products in a manufacturing company was used as a case study of the proposed method. The experimental methods are summarized as follows.

- **Questionnaire used:**
  Questionnaire survey conducted with 2,300 respondents by a manufacturing company in 1993 regarding oral care products.

- **Domain Expert:**
  The resulting knowledge was evaluated by a domain expert who is concerned with marketing analysis on the task domain at the manufacturing company. She has been required to interactively and subjectively evaluate the quality of the discovered knowledge from the viewpoints of simplicity, understandability, accuracy, reliability, plausibility, and applicability of the knowledge.

- **Experimental Methods and Implementation:**
  - 16 image words were selected to define product image. Respondents of the questionnaire evaluated how well each of the 16 image words fit the categories (*Fit*, *Moderate*, and *Does not Fit*, which will be respectively denoted as O, M, and X in the following) of the toothpaste brand they mainly use.
  - 16 features words were selected for the evaluation. Respondents of the questionnaire evaluated whether they were *satisfied* or *not satisfied* with their toothpaste brand with regards to each of the 16 features. Therefore, the size of the search space is 2^{16}, which seems small to use Genetic Algorithms, however, it is enough large for using Simulated Breeding. For example, refer to (Bala et al. 1995).

Results

This subsection presents the results of two experimental results (the one for SIBILE-I and the other for SIBILE-II/III) for the selected images: *innovative* and *effective*. In the experiments, we have tried to discover the knowledge to represent both of the two image words simultaneously. Prior to the experiments, as an initial investigation, we applied C4.5 programs to the data with all 16 features. As a result, we have got a huge *pruned* decision tree with 113 nodes, which was impossible for even the experienced expert to correctly interpret.

```
CHARACTERISTIC = YES: O (293.0/117.3)
CHARACTERISTIC = NO:
|   LIQUID = YES: O (48.0/21.9)
|   LIQUID = NO:
|   |   COMBINATION = YES: O (120.0/59.3)
|   |   COMBINATION = NO:
|   |   |   FREQUENT-CH = YES: O (198.0/115.3)
|   |   |   FREQUENT-CH = NO:
|   |   |   |   MAKER-VALUE = NO: M (1191.0/595.4)
|   |   |   |   MAKER-VALUE = YES:
|   |   |   |   |   RECOMMENDATION = YES: O (33.0/14.5)
|   |   |   |   |   RECOMMENDATION = NO: M (417.0/236.5)
```

Figure 2: Resulting Decision Tree from Experiment 1

The final results of the decision tree is shown in Figure 2. It took 7 generations or user interaction to obtained the desired results. The decision tree is represented in the form of standard outputs of C4.5 programs.

Results on SIBILE-II/-III are also shown in Figure 3 of the set of decision rules.

```
RESULTING DECISION RULES:


Rule 6:                         Rule 7:
       medical-type = YES              medical-type = YES
       liquid = YES                    maker-value = YES
       family-use = YES                ->  class 0 [61.5%]
       ->  class 0 [85.7%]
                                Rule 5:
Rule 10:                               characteristics = YES
       characteristics = NO            maker-value = NO
       liquid = YES                    family-use = NO
       maker-value = YES               ->   class 0 [58.5%]
       family-use = YES
       ->  class 0 [77.7%]  Rule 9:
                                       medical-type = YES
Rule 4:                                family-use = NO
       characteristics = YES           ->   class 0 [57.0%]
       liquid = NO
       family-use = YES     Rule 13:
       ->  class 0 [63.0%]             liquid = YES
                                       maker-value = NO
Rule 2:                                family-use = NO
       characteristics = YES           ->   class 0 [53.9%]
       maker-value = YES
       ->  class 0 [62.2%]
```

Figure 3: Resulting Decision Rules from Experiment 2

Discussion

The above experimental results have been evaluated by both quantitative and qualitative ways. Since one of the objectives of this research is to support the creativity of marketing analysts, there is more than one right answer to the questions we are investigating and there are several potential answers left uncovered. With this in mind, the interpretations of the simulation results are described below.

Simplicity of the Decision Rules As depicted in the decision trees obtained, toothpaste with the images of both *innovative* and *effective* were explained by the seven features in the first experiment and the five features in the second experiment. We have got much simpler decision rules than the tree with all 16 features generated by C4.5 programs. It is remarkable that the sizes of the trees do not dramatically change as the generation proceeds. This suggests that in the task domain, the size of the trees does not necessarily become a good measure to evaluate resulting decision rules.

Understandability of the Resulting Rules The decision tree obtained for first experiment explain why the image characteristics both innovative and effective fit the data. For example, from the tree in Figure 2, the user can easily derive the strategy :

> *Develop a line-up of toothpaste with technical characteristics other brands do not have, a liquid toothpaste, and a combination toothbrush/toothpaste brand.*

This strategy is confirmed by the other domain experts to be similar to the company's actual strategy for its brand which was not on the market at the time the questionnaire survey was conducted.

Accuracy Comparison Accuracy does not overcome the other measures in SIBILE, however, it is one of the important measure which can be evaluated among the other methods. Table 1 shows the accuracy comparison results of the tree with all features and the resulting tree generated by the experiment. Furthermore, we have compared the accuracy of resulting decision trees by SIBILE with the other statistical methods: the linear discrimination method (LD)in *SAS* package and the automatic interaction detection (AID) in *S* package. The experimental results are also summarized in Table 1.

Table 1: Accuracy of SIBILE, C4.5, LD, and AID

Methods	C4.5	LD	AID	Sibile	LD	Sibile	LD
Selected Features	All	All	All	Same 1.7.4	Same 1.7.4	Same 2.3.4	Same 2.3.4
Total Accuracy	57.3%	41.4%	56.0%	51.4%	40.6%	52.4%	33.9%
Class O Accuracy	51.8%	48.2%	61.3%	41.3%	37.2%	45.3%	50.4%
Class M Accuracy	81.2%	34.5%	69.4%	77.5%	43.8%	76.1%	8.2%
Class X Accuracy	0.0%	43.5%	0.9%	0.0%	40.0%	0.0%	66.5%

The first three columns, the next two columns, and the final two columns respectively indicate the results using all 16 features, the results using selected features in the experiment 1, and the results using selected features in the experiment 2. The data C4.5 with features 1.7.4 and 2.3.4 mean the results which have been selected by the proposed method.

In our task domain, the total accuracy of the resulting rules and the accuracy for Class O are critical to get decision knowledge. Keep this in mind, the figure suggest that the proposed method shows the same level of accuracy among the other method, in spite that the resulting rules are so simple.

Concluding Remarks

The main contributions of the research to KDD are (1) that the combinatorial feature selection problem in inductive learning can be resolved by simulated breeding, which is characterized by subjective and interactive evaluations of offspring generated by genetic operations, (2) that the effectiveness of the proposed method SIBILE has been validated by a case study on practical questionnaire data, and (3) that we have shown the Alife oriented techniques such as simulated breeding can be applied to practical knowledge discovery problems.

The pre-requisites of the proposed method are quite simple and the algorithm is easy to implement. Therefore, we conclude the proposed method is applicable to other task domain problems.

References

Bala, J. W.; Huang, J.; Vafaie, H.; De Jong, K; and Wechsler, H. 1995: Hybrid Learning Using Genetic Algorithms and Decision Trees for Pattern Recognition. in *Proceedings of 14th International Joint Conference on Artificial Intelligence*, 719-724.

Dawkins, R. 1986. *The Blind Watchmaker.* W. W. Norton.

Goldberg, D. E. 1989. *Genetic Algorithms in Search, Optimization and Machine Learning.* Addison-Wesley.

Quinlan, J. R. 1993. *C4.5: Programs for Machine Learning.* Morgan-Kaufmann.

Saaty, T. L. 1980. *The Analytic Hierarchy Process - Planning, Priority Setting, Resource Allocation.* McGraw-Hill.

Sims, K. 1992. Interactive Evolution of Dynamical Systems. in Varela, F. J., Bourgine, P. (eds.) : *Toward a Practice of Autonomous Systems - Proc. 1st European Conf. Artificial Life*, MIT Press: 171-178.

Terano , T.; Ishino, Y.; and Yoshinaga, K. 1995: Integrating Machine Learning and Simulated Breeding Techniques to Analyze the Characteristics of Consumer Goods. in Biethahn, J., Nissen, V. eds.1995. *Evolutionary Algorithms in Management Applications*, Springer-Verlag, 211-224.

Weiss, S. M., and Kulikowski, C. A. 1991. *Computer Systems that Learn.* Morgan-Kaufmann.