KEER2010, PARIS | MARCH 2-4 2010

INTERNATIONAL CONFERENCE ON KANSEI ENGINEERING AND EMOTION RESEARCH 2010

KANSEI IMPRESSION ANALYSIS USING FUZZY C4.5 DECISION TREE

Masataka TOKUMARU * [a], and Noriaki MURANAKA[a]

[a]*Kansai University, Japan*

ABSTRACT

This paper proposes a method for investigating Kansei impressions of products using of a Fuzzy C4.5 decision tree. A decision tree is built based on the results of a questionnaire that determines users' product impressions and helps to uncover the major factors that affect the product's attractiveness. In this paper, we analyzed the appearance of running shoes and investigated the factors that determine particular user preferences. The survey, conducted through a questionnaire, about impressions of the shoes' appearance asks the subjects about their intuitive impressions. Therefore, answers to the questions are very subjective and vague. The subjectivity and vagueness of the answers cause difficulty in constructing a reliable decision tree. In this paper, we propose a method for extracting some peculiar answers of the questions based on rotation estimation, using which we obtained a reliable decision tree and adopted the general public's Kansei impression.

Keywords: fuzzy decision tree, impression analysis, running shoes, Kansei information

1. INTRODUCTION

Product design requires careful consideration of many factors. However, it is difficult for many developers to design a product that will satisfy everyone. For some designers and manufacturers, it is difficult to create a product that is beneficial for all users because it will be used by a variety of people who have different aesthetics and preferences. Therefore, some technical methods for investigating people's impressions of a product are proposed in order to uncover the major factors that determine a product's usability and attractiveness [1–5]. These methods are a part of preference-based design and ergonomic engineering, and they are based on a questionnaire survey about consumers' impressions of various types of commodities. These methods visualize the relationship among these impressions using a map with two or three dimensions. In previous studies,

*Corresponding author: 3-3-35 Yamate-cho, Suita-shi OSAKA 564-8680 JAPAN : toku@kansai-u.ac.jp

various methods such as factor analysis and rough sets have investigated the relationship between quantifiable factors (such as product specifications) and the over all impression of the products (such as product usability and attractiveness). However, people generally have multiple responses to a product and react differently to its various components.

The key to discovering what makes a product usable or attractive for a large number of people lies in investigating the relationship between the consumers' overall impression of the product and their response to individual components. For a particular product, it is important to know why some people feel it is usable and others do not. Every individual has a set of responses to the product's components and features that differ from that of others' because individuals have different physiques and subjective responses. Previous methods have ignored these differences in people and failed to investigate such relationships. To investigate these relationships, we have proposed a new method for analyzing product impressions [6]. This method analyzes the results of product-impression questionnaires while taking into account subjects' subjective reactions. Unlike previous methods, it can reveal some general rules that describe which of the reactions determine their overall impression of a product. In our previous paper, we investigated the factors that determine the ease-of-hitting of golf clubs and obtained some reliable rules affecting user impressions, using a fuzzy decision tree built from questionnaire results.

In this paper, we analyze impressions of the appearance of running shoes. This investigation uses pictures of a number of running shoes. Many testers answered questions about their impressions of the shoes' appearance. The questionnaire includes questions concerning various aspects of running shoes and overall impressions such as "I want this shoe." The purpose of this analysis is to discover what kind of shoe is wanted by many users. The questionnaire on the impressions of the appearance of the shoes asks the subjects about their intuitive impression. Therefore, answers to the questions are more subjective and vague than answers to questions about ease-of-hitting of golf clubs. The subjectivity and vagueness cause difficulty in constructing a reliable decision tree. This analysis aims to investigate important factors that influence attractiveness in the appearance of running shoes. In this case, it is desirable that the created decision tree and rules affecting user impressions should reflect many testers' answers. Moreover, it is desirable that these rules can be true of other people in general, besides the tester. For obtaining reliable rules, we execute performance estimation of a fuzzy decision tree by rotation estimation and remove some answers to questions that reduce the reliability of the general rules.

2. KANSEI ANALYSIS USING A FUZZY C4.5 DECISION TREE

2.1. Data Accumulation in Relation to Products' Kansei Impressions

To establish a relationship between overall impressions and specific product attributes, we conducted a questionnaire survey that collected a variety of product impressions from consumers. This subsection describes a method of product-impression analysis using a seven-point scale semantic differential (SD) method, which is generally used for analyzing such survey results. When using the SD method, the following procedures are required: First, we define the various qualities (called "attributes") that describe general or detailed impressions of the overall product and its components, in order to execute the questionnaire survey. For example, in an impression analysis of cell phones, the attributes may be “easy or hard to use” or “want or do not want.” These attributes express an overall impression of the phones. We then include attributes for specific components to describe detailed impression of the phones such as “buttons are easy or hard to push,” “the body is too narrow or too fat,” and “the LCD screen feels small or large.” Once we have established the correct set of attributes, we collect various types of cell phones and a group of test subjects. Each test subject should use each of the phones and then fill out the questionnaire that records his or her impressions. Finally, a Fuzzy C4.5 decision tree is built that takes into account the data accumulated from the questionnaire results.

2.2. Algorithm for Building the Fuzzy C4.5 Decision Tree

C4.5 is an algorithm proposed by R.Quinlan in 1993 [7] for building a decision tree. The C4.5 decision tree divides data items into subsets based on an attribute. If an attribute maximizes the gain ratio when dividing data into categories, it is considered useful for producing a decision tree. A set of data items is represented as T and an attribute as X_p. The maximum value of the index p is the number of attributes. When investigating the attractiveness of running shoes, data items are categorized into classes such as “I want this shoe,” “I do not want this shoe,” and “neutral.” Each class is presented as C_i, and the maximum value of the integer index i is determined by the number of categories into which the classes divide the data. When an attribute X_p takes n types of categories, the set of data items T is divided into n subsets T_1, T_2, ... ,T_n. If the questionnaire sheet uses the seven-point scale SD method, each attribute takes seven values. C4.5 requires all data items to be divided into categories. In addition, each data item should be included in one of the non-overlapping subsets. Our previous report divided the data, placing seven values into three classes [8]. The values of 1, 2, and 3 were classified into a category named “+,” the value of 4 into “0,” and the values of 5, 6, and 7 into “$-$.” However, this approach was limited as it could not take into account the differences between the values of 1, 2, and 3. To solve this problem, we adopt Fuzzy C4.5, which enables subsets for categorizing data items, to be represented as fuzzy sets. In Fuzzy C4.5 decision tree, each data point belongs to one or more classes with specific possibilities. Moreover, our other previous report verified that the optimized membership functions using non-normalized fuzzy sets were more reliable than normalized fuzzy sets [6].

Fig. 1 shows the membership functions used to calculate the possibility of belonging to each “preference” data class. Here, the possibility of belonging to class C_i of the data item s is presented as $\mu(C_i, s)$. For example, a data item s with a “preference” value of 3 belongs to the class C_1, which means “I want this shoe” with the possibility of $\mu(C_1, s) = 0.6$, class C_2 means “neutral (think neither)” with the possibility $\mu(C_2, s) = 0.3$, and class C_3 means “I do not want this shoe” with the possibility $\mu(C_3, s) = 0$.

A common C4.5 algorithm classifies data into categories based on information gain ratio. An entropy *info*(S) of a set of data items, represented as $S = \{s_1, s_2, ..., s_x\}$, is given by (1). Here, *freq*(C_i, S) is the sum of the possibilities of belonging to class C_i of all the examples and is given by (2). The maximum value of the integer index x is the number of data items belonging to S. $|S|$ is the sum of the sums of the possibilities for each class of all the examples included in the set S.

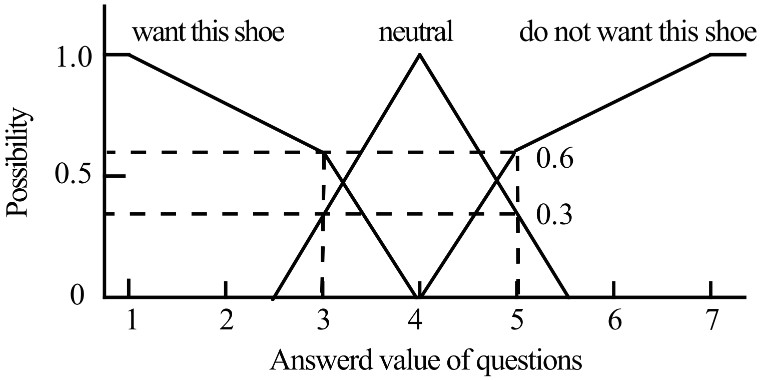

Figure 1: Membership functions used for dividing data items into three subsets.

The integer index k is the number of categories into which the classes divide the data.

$$\mathit{info}(S) = -\sum_{i=1}^{k} \left\{ \frac{\mathit{freq}(C_i, S)}{|S|} \times \log_2 \frac{\mathit{freq}(C_i, S)}{|S|} \right\} \text{bit} \tag{1}$$

$$\mathit{freq}(C_i, S) = \sum_{h=1}^{x} \mu(C_i, s_h) \tag{2}$$

An entropy $\mathit{info}_{X_p}(T)$, where examples belonging to the set T are divided into some subset T_j $(j : 1 - n)$ by an attribute X_p, is given by (3). Entropy $\mathit{info}(T_j)$ is given by (4).

$$\mathit{info}_{X_p}(T) = \sum_{j=1}^{n} \left\{ \frac{|T_j|}{|T|} \times \mathit{info}(T_j) \right\} \tag{3}$$

$$\mathit{info}(T_j) = \sum_{i=1}^{k} \left\{ \frac{\mathit{freq}(C_i, T_j)}{|T_j|} \times \log_2 \frac{\mathit{freq}(C_i, T_j)}{|T_j|} \right\} \tag{4}$$

An attribute X_p divides data into fuzzy sets T_j $(j : 1-n)$ and gives a possibility grade $\mu(T_j, s_h)$ $(h : 1 - x)$. The sum of the possibilities belonging to class C_i for the examples belonging to the subset T_j, represented as $\mathit{freq}(C_i, T_j)$, is given by (5).

$$\mathit{freq}(C_i, T_j) = \sum_{h=1}^{x} \{\mu(C_i, s_h) \times \mu(T_j, s_h)\} \tag{5}$$

The information gain of an attribute X_p shown as $\mathit{gain}(X_p)$ is calculated by (6). This describes a reduction in information entropy where the set T is divided into subsets T_j $(j : 1-n)$ by attribute X_p.

$$\mathit{gain}(X_p) = \mathit{info}(T) - \mathit{info}_{X_p}(T) \tag{6}$$

In a standard ID3 decision tree, developed before C4.5, an attribute dividing data into some subsets is determined based on the information gain. Unfortunately, this more often selects an attribute with many subsets than one with few subsets. C4.5 solves this problem by applying an amount of split information and an information gain ratio to select attributes for categorizing data. The amount of split information *split info*(X_p) is calculated by (7).

$$\textit{split info}(X_p) = -\sum_{j=1}^{n} \frac{|T_j|}{|T|} \times \log_2\left(\frac{|T_j|}{|T|}\right) \tag{7}$$

Here, a set T is divided into n subsets by an attribute X_p. The information gain ratio, *gain ratio*(X_p), is calculated by (8).

$$\textit{gain ratio}(X_p) = \textit{gain}(X_p)/\textit{split info}(X_p) \tag{8}$$

In C4.5 decision tree, an attribute maximizing the information gain ratio, *gain ratio*(X), is selected to categorize data included in a set T into subsets $T_1, T_2, ..., T_n$. Furthermore, in each of the subsets $T_1 - T_n$, data included in the subset is categorized into some smaller subsets by an attribute other than the attribute adopted as the upper branch node. When most of the data included in the subset belongs to the same class, or when few data are included in the subset, the subset becomes a leaf node belonging to the class.

2.3. Optimization of Decision Tree Structure Based on Classification Error

The size and structure of a decision tree generally depend on the conditions required for building a leaf node. For impression analysis, it is preferable to have a small decision tree with a simple structure, because it is difficult to extract useful rules from large and complex decision trees. On the other hand, oversimplifying a decision tree may decrease its reliability. This paper employs a system of classifying data errors to optimize the decision tree structure. Two conditions determine the construction of real nodes:

- Maximum class occupancy ratio (*MCOR*)
- Minimum data content ratio (*MDCR*)

MCOR represents the ratio of the data belonging to one class to the data belonging to all classes. When a node of the tree contains a data set T, *MCOR* is calculated by (9).

$$\textit{MCOR} = \frac{\max\{\textit{freq}(C_1, T), \textit{freq}(C_2, T), ..\textit{freq}(C_k, T)\}}{\sum_{i=1}^{k} \textit{freq}(C_i, T)} \tag{9}$$

When the value of *MCOR* in a node of the tree is higher than the threshold level, $\textit{class}_{max}$, the node stops dividing the data and becomes a leaf node with class C_i maximizing *freq*(C_i, T).

MDCR is the ratio of the number of data items contained in a node to the number of data items contained in the tree. For example, when the total number of data items is denoted by X, and the number of data items contained in the subset T, except for data items s_h with possibility $\mu(T, s_h) = 0$, is denoted by $N(T)$, *MDCR* is calculated by (10).

$$\textit{MDCR} = \frac{N(T)}{X} \tag{10}$$

When the value of *MDCR* in a node is lower than the threshold level, $\textit{data}_{min}$, the node stops dividing data and becomes a leaf node with class C_i maximizing *freq*(C_i, T_j).

2.4. Establishing Rules from Constructed Decision Tree

This subsection describes a method for extracting useful knowledge about product impressions from our C4.5 decision tree. Fig. 2 shows an example of a built decision tree. An attribute, X_p, is assigned to each leaf node, where p is the attribute number that categorizes data. For the decision tree shown in Fig. 2, the attribute X_5 has two categories represented as "$+$" and "$-$", and the attribute X_3 has three categories represented as "$+$", "0", and "$-$." Data is classified into three classes, represented as C_1, C_2, and C_3. Fuzzy rules are generated from the built decision tree. In Fig. 2, the far left end of the leaf node defines fuzzy rule R1, shown as (11), and the far right end of the leaf node defines fuzzy rule R2, shown as (12).

$$\text{R1: } X_5(+) \ \mathit{and} \ X_2(+) \ \mathit{and} \ X_3(+) \ \mathit{then} \ C_1 \tag{11}$$

$$\text{R2: } X_5(-) \ \mathit{and} \ X_7(-) \ \mathit{and} \ X_6(+) \ \mathit{and} \ X_2(-) \ \mathit{then} \ C_3 \tag{12}$$

Each of the characteristics in (11) and (12) describes a value of attribute X_p. In this way, we establish some fuzzy rules for obtaining useful knowledge about product impressions.

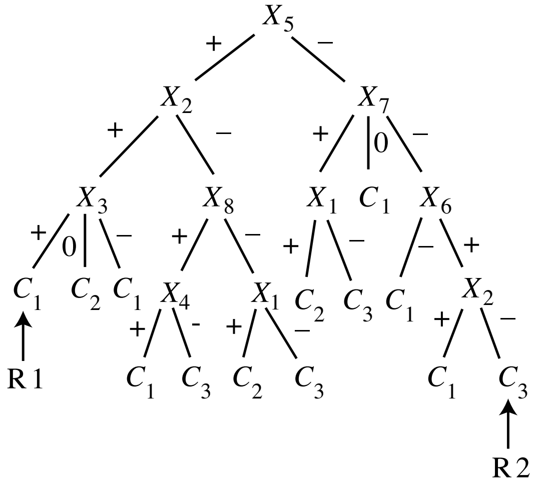

Figure 2: An example of decision tree.

3. IMPRESSION ANALYSIS OF RUNNING SHOES

3.1. Questionnaire Survey about Impression of Running Shoes

This section describes the impression analysis conducted on running shoes. We used 30 pictures of running shoes available on the market, manufactured by a number of companies. The subjects of the test were 100 American runners. Fifteen were male athletic runners, 27 were male fun and fitness runners, 15 were female athletic runners, and 43 were female fun and fitness runners. The subjects answered 18 questions about their impressions of the running shoes using the seven-point scale SD method. The questionnaire included items concerning various aspects of the shoes, for example, "looks futuristic or retro," which related to impressions of the shoes' appearance, and "looks breathable or highly waterproof," which related to impressions of functional aspects of a shoe's design. It also included the items "I want this shoe" or "I do not want this shoe," which measure the subjects' preference for each running shoe.

In the test, those pictures were displayed on the wall of the park. Each subject sequentially saw the pictures and answered questions about their impression of the appearance of the shoes. Consequently, we obtained 2,543 data items (evaluations) for various running shoes. Each item was presented as 18 attribute−value pairs, each of which represents the evaluation of the running shoe with respect to each attribute. Therefore, by using a Fuzzy C4.5 decision tree, we may obtain some rules to categorize the data items into multiple disjoint categories according to the value of each attribute, and classify them based on subjective evaluation (preference).

3.2. Performance Estimation of Fuzzy Decision Tree

The structure of the Fuzzy C4.5 decision tree is usually affected by the conditions for building branch nodes. Therefore, we optimized the values of $class_{max}$ or $data_{min}$, described in subsection 2.3 of section 2. We divided 2,543 data items into groups according to the types of subjects and optimized these values in their respective groups based on resubstitution classification error [9]. Table 1 shows the groups, the number of data items included in each group, and the optimized value of $class_{max}$ and $data_{min}$.

Table 1: Group and Optimized Values

Group	Data Items	$class_{max}$	$data_{min}$	RMSE
Male	1,164	0.5	0.05	1.42
Female	1,379	0.55	0.05	1.44
Athletic	772	0.55	0.1	1.49
Fun and Fitness	1,771	0.6	0.1	1.47
Male Athletic	403	0.55	0.1	1.43
Female Athletic	369	0.65	0.1	1.50
Male Fun and Fitness	761	0.60	0.05	1.43
Female Fun and Fitness	1,010	0.7	0.05	1.40

This optimization uses all data items for constructing and evaluating the decision tree. However, the performance estimation of the decision tree should be verified against unknown data items that are not used for constructing the tree. When a decision tree is constructed by using comparatively fewer data items, the tree can classify data items into correct classes. However, it cannot classify unknown data correctly. To verify the performance estimation of the decision tree, the number of data items used for constructing the tree should be optimized. Therefore, after dividing the data items belonging to each group into subsets, we executed the randomized version of rotation estimation. The estimation process in the case of the male group (1,164 data items) is as follows.

1. Randomly select 100 data items from the 1,164 data items accumulated by the survey. These are defined as "training data." Also randomly select 100 data items from the remaining 1,064 data items (those not used as training data). These 100 data items are called "evaluation data." A combination of data from the training data set and evaluation data set is called an "example set" in this paper.

2. Prepare 100 example sets for experimentation; each example set must have different training and evaluation data.

3. Construct a decision tree using the training data from one of the example sets. Execute a classification test using the decision tree. This divides the evaluation data included in the example set into classes.

4. Determine the classification error rate by averaging the root mean square error (RMSE) between the classification value of the evaluation data and the values estimated by the decision tree.

5. In the same way, construct decision trees and perform evaluations using the remaining ninety nine example sets. The average, maximum, and minimum values of the classification error rate are determined by the results of the 100 example sets.

6. Increase the training data by 100. Repeat operations 1−5.

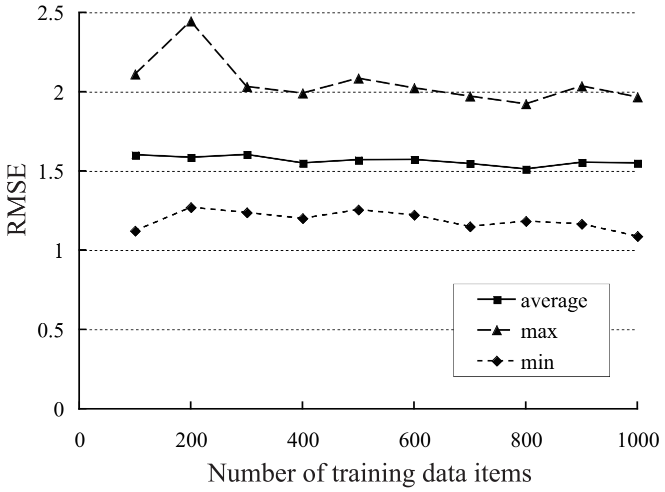

Figure 3: Change in classification error rate (all data items).

Fig. 3 shows an example of experimental results for the Male group. The results show that changing the number of training data has little influence on the reliability of the decision tree, and the constructed decision tree does not possess high reliability. In Fig. 3, when the number of training data items is 800, the average value of the classification error rate is about 1.5. This means that the constructed decision tree has the ability to estimate the answers to questions using the seven-point scale SD method with an error margin of 1.5. For example, when the correct value of the answer in an unknown data item is 4, the constructed decision tree estimates the value at an error rate of 2.5−5.5.

The fact that increasing the number of training data items does not increase the reliability of the decision tree suggests the variety of Kansei impressions of subjects and the impossibility of discovering general rules covering all people. However, if some peculiar answers can be ignored, useful rules adapted to many people may be obtained by the constructed decision tree. To obtain useful rules from the decision tree, the average RMSE should be less than at least 1.0. We investigated the classification error rate, determined by the average RMSE between the values estimated by the decision tree and the classification value of the 80 evaluation data items except for 20 data items that had a bigger RMSE. Fig. 4 shows the resulting change in the classification error rate, for the 80 evaluation data items. The average RMSE is about 1.0. This suggests that the constructed decision tree performs well, estimating 80% of the general public's Kansei evaluations.

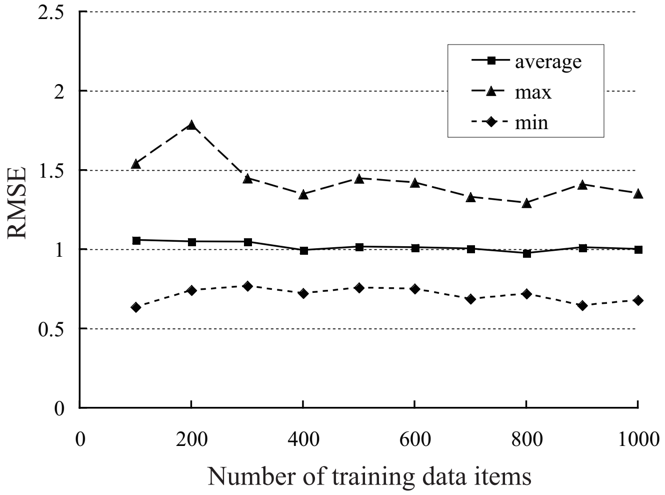

Figure 4: Change in classification error rate (excluding 20% of the data items).

To construct a more reliable decision tree, we propose a method for extracting peculiar data items. The extraction process is as follows:

1. Randomly select 800 training data items and 100 evaluation data items from the 1,164 data items in the Male group.
2. Construct a decision tree using the training data. Execute a classification test using the decision tree. This divides the evaluation data into classes.
3. Select the 20 evaluation data that have RMSE values bigger than those of the others and set them as "taboo data."
4. Repeat operations 1−3 1,000 times.
5. Extract data items set as taboo data more than ten times in 1,000 trials.

After the operation, 374 data items of 1,164 Male runners' data items were extracted as taboo data. Using the remaining 790 data items, we executed rotation estimation. When the number of training data items was 500 and the number of evaluation data items was 100, the average classification error rate was 0.75. This means that the decision tree constructed without taboo data possesses high reliability, with a generality rate of about 68%.

3.3. Conclusion

This paper has proposed a method to investigate Kansei impressions using a Fuzzy C4.5 decision tree. We conducted an experiment to test impression analysis for Kansei impressions of the appearance of running shoes. The experimental results clearly demonstrated that increasing the number of training data items did not increase the reliability of the decision tree. However, the results also showed that excluding some peculiar data items increased the reliability of the decision tree and could provide useful rules for adapting the results for many people's Kansei. In our future work, we should validate the attribute selection measure [10] and consider how to extract stable rules from the constructed decision tree.

REFERENCES

[1] Y.Uchida, T.Yoshikawa, T.Furuhashi, E.Hirao, and H.Iguchi. Evaluation of products by analysis of user-review using hk graph. In *Proc. of Joint 4th International Conference on Soft Computing and Intelligent Systems and 9th International Symposium on advanced Intelligent Systems (SCIS & ISIS 2008)*, pages 376–379, 2008.

[2] P.J.Tsai and S.Nagasawa. Applied research on the product planning of cosmetics for men. *Kansei Engineering International*, 3(4):15–24, 2002.

[3] K.B.Lee and R.R.Grice. Developing a new usability testing method for mobile devices. In *Proc. of 2004 International Professional Communication Conference*, pages 115–127, 2004.

[4] T.Tsuchiya and Y.Matsubara. Non-linear data analysis on kansei engineering and design evaluation by genetic algorithm. *Kansei Engineering International*, 6(4):55–62, 2006.

[5] J.Lu, X.Deng, P.Vroman, and X.Zeng. Fuzzy multi-criteria group decision support system for nonwoven based cosmetic product development evaluation. In *Proc. of 2008 IEEE International Conference on Fuzzy Systems*, pages 1700–1707, 2008.

[6] M.Tokumaru and N.Muranaka. Product-impression analysis using fuzzy c4.5 decision tree. In *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 13(6):122–127, 2009.

[7] J.R.Quinlan. *C4.5: Program for Machine Learning*. Morgan Kaufmann Publishers, 1993.

[8] M.Tokumaru, N.Muranaka, and S.Imanishi. Fuzzy decision tree analysis to investigate what has an influence on ease-of-use and preference with products. In *Proc. of The 8th World Multi-Conference on Systemics, Cybernetics and Informatics, Vol.1 Information Systems, Technologies and Applications*, pages 433–438, 2004.

[9] P.A.Devijver and Kittler, J. *Pattern Recognition: A Statistical approach*. Prentice-Hall International, 1982.

[10] X.Wang and C.Borgelt. Information measures in fuzzy decision trees. In *Proc. of the IEEE International Conference on Fuzzy Systems*, pages 85–90, 2004.