Intelligent Recommender System for High Dimensional Transaction Data Set with Complex Relationships among the Variables

Article *in* Indian Journal of Science and Technology · May 2016

DOI: 10.17485/ijst/2016/v9i20/94686

2 authors, including:

Jun Woo Kim
Dong-A University
36 PUBLICATIONS **81** CITATIONS

SEE PROFILE

ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, Vol 9(20), DOI: 10.17485/ijst/2016/v9i20/94686, May 2016 ISSN (Online) : 0974-5645

Intelligent Recommender System for High Dimensional Transaction Data Set with Complex Relationships among the Variables

Jun Woo Kim[1] and Soo Kyun Kim[2*]

[1]Department of Industrial and Management Systems Engineering, Dong-A University, Busan, South Korea; kjunwoo@dau.ac.kr

[2]Department of Game Engineering, Paichai University, Daejeon, South Korea; kimsk@pcu.ac.kr

Abstract

Background/Objectives: This paper aims to develop a novel intelligent recommender system suitable for high dimensional data where multiple factor variables influence on multiple response variables. **Methods/Statistical Analysis:** This paper suggests that the structure of the cause-and-effect relations among the variables can be represented in a simple form called structured association network model (SANM). Based on the SANM and conventional data mining techniques such as association rule mining and naïve Bayesian classifier, the proposed recommender system computes three novel recommendation scores for each response variables, and the variables with high scores can be selected for recommendation. **Findings:** For illustration, the proposed recommender system has been applied to a mass health examination result data set. Owing to its simple structure, a SANM for a given data set can be easily obtained by simply identifying the factor and the response variables, and the experiment results revealed that the proposed system can identify the recommendable items for the individuals more effectively than the traditional classification techniques such as naïve Bayesian classifier. Consequently, we can conclude that the proposed recommender system can deal with high dimensional transaction data set in more effective manner than the traditional approaches where the underlying semantic relationships among the variables are not considered. **Applications/Improvements:** The proposed recommender system is useful for evaluation of the potential risks of specific diseases. Moreover, the recommendation scores can also be used as a tool for feature construction.

Keywords: Association Rule Mining, Data Mining, Recommendation, Structured Association Network Model, Transaction Data

1. Introduction

The role of the recommender systems is to identify the items relevant to the particular users by taking their context into account. To this end, the recommender systems typically calculate the recommendation scores of the candidate items and choose the items with higher scores[1,2]. In recent, such recommendation systems have become quite useful tools for the individuals facing the massive amount of information generated and gathered by the modern information systems, and they have been adopted in a wide range of application domains including online shopping[3,4], marketing[5,6], finance[7], education[8] and healthcare[9], etc.

In order to calculate the recommendation scores and identify the relevant items, the relationship between the individual users' features and the possible items must be analyzed in advance. In this context, data mining techniques had been widely adopted in developing the recommender systems[10]. From data mining perspective, the recommendation techniques can be categorized into the two groups, supervised approach and unsupervised approach. In supervised approach, the classification techniques such as decision tree[11] and case based reasoning[12], or forecasting techniques such as linear regression[13] are used to calculate the recommendation

* Author for correspondence

scores for the candidate items. Although these applications have performed well in specific application domains, the supervised approach has several important limitations: (i) Conventional classifiers and forecasting models can handle only one response (dependent) variable. (ii) Hence, multiple models are required if there are two or more candidate items. (iii) It is hard to consider the relationships among the candidate items themselves.

On the other hand, in unsupervised approach, the recommendation scores are calculated by using the techniques which does not assume a single response variable, such as association rule mining[8]. However, the association rule based recommender systems also pose several problems: (i) The vast amount of the association rules can be extracted by the conventional association rule mining algorithms. (ii) Some association rules are redundant in that they consist of similar itemsets, and we can have trouble in choosing the association rule suitable for individual users[14]. (iii) Rule base for the extracted association rules and triggering rules have to be carefully designed.

Moreover, those limitations of the traditional recommendation approaches become evident when they are applied to the high dimensional data set where the multiple factor (independent) variables influence on the multiple response variables. To address these issues, this paper aims to develop a novel recommender system especially suitable for calculating the recommendation scores for the multiple response variables (items) within the high dimensional transaction data set. To this end, we assume that the cause-and-effect relations within the given data set can be represented in the simple form called the structured association network model (SANM), and the three novel recommendation scores based on SANM and the conventional data mining techniques are developed in this paper.

The rest of this paper is organized as follows. Section 2 provides a brief literature review on the association rule mining and the unsupervised recommendation strategy. Section 3 introduces the concept of SANM and describes the overall framework of the proposed intelligent recommender system. Section 4 provides the experiment results obtained by applying the proposed system to a mass health examination result data set. Finally, we present the discussions and the concluding remarks in Section 5.

2. Research Backgrounds

2.1 Association Rule Mining

Association rule mining is a well-known data mining technique for extracting the useful association rules from the given transaction data set with binary attributes called items. The association rule is defined as a rule in the form of "A → B", where both A and B are sets of items, itemsets. Note that the LHS (A) and the RHS (B) are mutually exclusive, and they are called antecedent and consequent of the association rule, respectively. An association rule suggests the positive correlation between the antecedent and the consequent, that is, "A → B" means that if the items of A are included in a transaction, the transaction is likely to contain the items of B also[15,16].

The usefulness of an association rule can be evaluated by calculating the performance measures such as support, confidence and lift, and the objective of the association rule mining is to discover all useful association rules[17]. In practice, this objective is generally achieved by using association rule mining algorithms such as Apriori and its variants[18].

Inherently, the useful association rules are suitable for the purpose of recommendation. For example, consider an useful association rule "{a,b} → {c}", where a, b and c denote the products (items) sold by a retailer. This rule states that "if a customer buys the product a and b, then he or she probably buys the product c also". Hence, it will be reasonable for the retailer to try to find the customers who bought both a and b, and recommend the product c to them.

2.2 Recommender System for Transaction Data Set

Let D denote a transaction data set and S_{item} denote the set of the discrete items $d_1, d_2, ..., d_n$. For a single transaction t, the objective of the recommender system is to identify the recommendable elements of $S_{item} - t$. Hence, it is straightforward that an useful association rule "A → B" may be helpful for this purpose if $A \subset t$ and $B \subset S_{item} - t$. In addition, the performance measures of such association rule can be used to compute the recommendation scores for the elements of $S_{item} - t$[19].

Conventional association rule mining algorithms such as Apriori exhaustively extracts all association

rules that satisfy the pre-specified thresholds for the performance measures, and they can be computationally burdensome if the dimensionality of given transaction data is high. Therefore, many recommender system based on the association rule extract the useful association rules in an offline manner, and the extracted rules generally form the rule base for future recommendation. For an individual transaction t, such recommender systems try to retrieve the relevant association rules from the rule base by comparing their antecedents and t, and compute the recommendation scores for the items in S_{item} - t by using the performance measures of the retrieved rules. Consequently, the items with high recommendation scores may be identified and recommended[8,19,20].

Although this recommendation strategy is quite intuitive and has performed well in traditional e-Business domains, it still poses several problems. Firstly, the conventional association rule mining algorithms often produce too many association rules, so the rule base and its interfaces for retrieval of the relevant rules must be carefully designed. Secondly, an individual transaction must coincide with the antecedents of the extracted association rules in order to trigger them. For example, a transaction t_0 = {a, b} can trigger the association rule "{ a, b} → {e}" in the rule base, and the item e may be selected for recommendation. However, the association rule "{a, b, c} → {f }" in the rule base is not triggered by the transaction t_0 since the item c is not included in t_0. Similarly, a transaction t_1 = {a, b, c} cannot trigger the association rule "{a, b} → { e }" and "{c } → {e }", although the item e seems to be partly recommendable. Thirdly, the previous association rule based recommender systems did not consider the underlying semantic relationships within the given data sets. For example, let's assume that the set of items S_{item} is partitioned into two subsets, S_a and S_b, where the items of S_a denote the features of raw materials while the items of S_b denote the features of the consequent products. Then, we may be particularly interested in the association rule "A → B" where $A \subset S_a$ and $B \subset S_b$, however, such cause-and-effect relations have been not considered in previous recommender systems. Moreover, the items within a high dimensional transaction data set can have quite complex relationships.

In order to address these issues, this paper aims to develop a novel recommender system for high dimensional transaction data set with semantic relationships. Compared to the previous association rule based recommender systems, the proposed system can be distinguished by three features. Firstly, we need not to extract the useful association rules and construct the rule base in advance. Secondly, we need not to determine the threshold values for the performance measures of association rules since the proposed system does not rely on the useful association rules. Finally, the semantic relationships among the items can be modeled by using SANM and explicitly considered by the proposed system.

3. Proposed Recommender System

Suppose that the set of items within a high dimensional transaction data set D can be partitioned into the two mutually exclusive subsets, the set of the factor variables $F = \{f_1, f_2, ..., f_n\}$ and the set of the response variables $R = \{ r_1, r_2, ..., r_m\}$. Note that the terms factor variable and response variable are synonyms for independent variable and dependent variable, respectively, and we assume that the f_i s (i =1, 2, …, n) have some influences on the r_js (j=1, 2, …,m).

Let $F_t = \{f_i \mid f_i \in F, f_i \in t \}$ and Let $R_t = \{r_j \mid r_j \in R, r_j \in t \}$ for a transaction t. Then, the objective of this paper is to develop a recommender system that can be used to compute the recommendation scores for the items in $R - R_t$, considering the relationships among the variables.

3.1 Structured Association Netowork Model (SANM)

The relationships among the variables can be highly complex, however, this paper suggests that such relationships can be simplified by using SANM as shown in Figure 1. The SANM consists of the factor variables, the response variables and the three types of the cause-and-effect relations including intra-association among the response variables, intra-association among the factor variables and inter-association between the factor and the response variables. The intra-associations indicate that a single factor (response) variable can influence on other factor (response) variables. That is, the occurrence of a factor (response) variable can cause the occurrences of other factor (response) variables if they are positively correlated. Similarly, the inter-association means that the occurrences of the factor variables can bring about the occurrences of the correlated response variables.

Therefore, the intra-association and the inter-association among the given variables should be carefully analyzed in order to calculate the recommendation scores of the response variables in R - R_t.

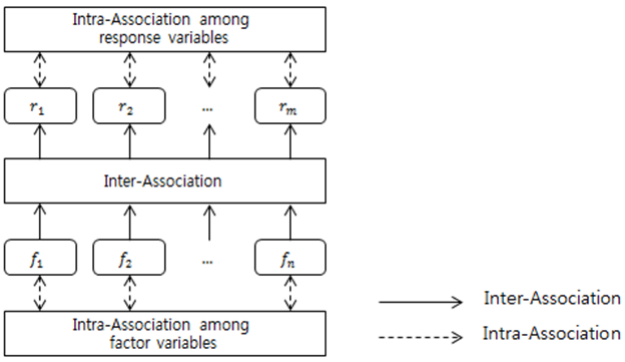

Figure 1. Structured association network model.

Although the concept of the SANM is quite simple, it contains comprehensive information about the relationships among the variables. In addition, the strengths of the associations are not considered in SANM, and we can obtain a SANM for a given transaction data set by simply identifying the factor and the response variables.

3.2 Recommendation Score based on Inter-Association and Response Intra-Association

This paper proposes three recommendation scores based on SANM, considering the three types of relations among the items. The first one is $RS_{INTER/R}$, which is based on the inter-association and the intra-association of response variables. For an unobserved response item y ($y \in R - R_k$), the value of $RS_{INTER/R}$ is computed by using the similarities between the profiles of the response items.

Let e denote a response item included in given transaction t($1 \le e \le m$). Then, the profile vector of e, P_e is generated as follows.

$$P_e = [I_{1e}, I_{2e}, ..., I_{ne}], \quad (1)$$

where I_{ie} (i=1, 2, …, n) denotes the performance measure of the association rule "$\{f_i\} \rightarrow \{e\}$". Note that the I_{ie} s can be obtained by using the well-known confidence or lift.

The profile vector P_y of an unobserved response item y also can be obtained in similar manner, and the influence of e on y, L_{ey} is calculated as follows.

$$L_{ey} = S_{cos}(P_e, P_y) \cdot J(e, y) \quad (2)$$

where $S_{cos}(P_e, P_y)$ denotes the cosine similarity[17] between the two profile vectors, P_e and P_y. In addition, J(e, y) is the Jaccard similarity coefficient[22] between the response items e and y, which is given in (3) where $E = \{t \mid t \in D, e \in t\}$ and $y = \{t \mid t \in D, y \in t\}$. That is, high L_{ey} value will be obtained if the profiles of e and y are similar to each other and those items co-occur frequently.

$$J(e, y) = \frac{|E \cap Y|}{|E \cup Y|}, \quad (3)$$

Finally, the $RS_{INTER/R}$ of an unobserved response variable y is obtained as follows.

$$RS_{INTER/R}(y) = \max_{e \in R_t} L_{ey} \quad (4)$$

The influence L_{ey} represents the effects of the factor variables on a specific response variable, and the $RS_{INTER/R}$ indicates that an unobserved response variable y is recommendable if any observed response variable e has a large influence on y.

3.3 Recommendation Score based on Inter-Association

The second recommendation score, RS_{INTER} is based on the inter-association. In more detail, $RS_{INTER/R}$ considers the one-to-one relationships between one factor variable and one response variable, and it is calculated as follows.

$$RS_{INTER}(y) = \max_{x \in F_t} I_{xy} \quad (5)$$

That is, RS_{INTER} of y is high if and only if there is any observed factor variable x that has highly positive correlation with y. The primary benefit of RS_{INTER} is that it allows the key factor variables to directly affect the recommendation score of y, while the irrelevant factor variables are not taken into account. On the contrary, the drawback of this recommendation score is that it does not consider the many-to-one relationships among the variables, however, this can be partly compensated by $RS_{INTER/R}$ and $RS_{INTER/F}$.

3.4 Recommendation Score based on Inter-Association and Factor Intra-Association

The third recommendation score, $RS_{INTER/F}$ considers the inter-association and intra-association of the factor

variables. There can be some potential factor variables that are not observed yet but likely to occur for a given transaction t soon. $RS_{INTER/F}$ indicates that an unobserved response item y is also recommendable if any potential factor variable has highly positive correlation with y, and it is calculated as follows.

$$RS_{INTER/F}(y) = \max_{v \in F - F_t} P(v \mid F_t) \cdot I_{vy}, \quad (6)$$

where $P(v \mid F_t)$ is the conditional probability of an unobserved factor variable v given that the other factor variables in $F - F_t$ are observed. Similarly with the well-known naive Bayesian classifier[21], we can obtain $P(v \mid F_t)$ as follows.

$$P(v \mid F_t) = \frac{P_v}{P_v + P_{\neg v}}, \quad (7)$$

where

$$P_v = P(v) \times \prod_{x \in F_t} P(x \mid v), \quad (8)$$

and

$$RS_{INTER/R}(y) = \max_{e \in R_t} L_{ey}. \quad (9)$$

That is, the $RS_{INTER/F}$ of y is high if and only if there is any unobserved factor variable likely to occur and it has highly positive correlation with y. We can see that $RS_{INTER/F}$ allows the observed factor variable x to indirectly affect the recommendation score of unobserved response variable y via the potential factor variable v . Hence, this third recommendation score may be helpful especially for identifying the relevant unobserved response variables from a long-term perspective.

4. Experiment Results

This section illustrates how the proposed recommender system works. To this end, the system has been applied to a mass health examination result data set and used to compute the three recommendation scores for a specific disease.

4.1 Description of the Data

The initial data is collected from a mass health examination for 278 high school students in Korea, which is designed to investigate the health status and diagnose the common diseases. Each variable in the initial data relates to a disease, a symptom, or a characteristic of life style. The diseases are detected by the medical tests, and such test results are contained within the initial data, however, the data give no information about the undetected diseases except that they are not detected yet. Here, we aim to address this problem by providing additional information about the undetected diseases to the individual examinees. Especially, we focus on the variables related with the dental health and compute the recommendation scores for a common dental disease.

In order to apply the proposed recommender system, we have to identify the factor and the response variables that form the SANM. In this paper, the variables for the life styles relevant to the dental health are used as the factor variables, while the variables for the dental diseases and the subjective symptoms related with dental health are selected as the response variables. That is, we assume that the individual's life style can affect the symptoms and the dental diseases, and the factor and the response variables used in the numerical experiment are listed in Table 1 and Table 2. Note that some of those variables have been binarized before applying the proposed recommender system, while the others have been represented in the binary format inherently.

Table 1. Factor Variables

Variable	Description
f_1	Dental clinic visit during last year
f_2	Preference for sugary foods or carbonated beverage
f_3	Disuse of fluoride toothpaste
f_4	Irregular meals
f_5	Non-preference for milk or milk products
f_6	Non-preference for vegetables or fruits
f_7	Preference for sugary or salty foods
f_8	Preference for fast foods
f_9	Poor hand washing after going out
f_{10}	Brushing teeth less than once a day

Among the 11 response variables in Table 2, the dental caries, r_1 was selected as the target variable y. In addition, 20 records of the data set have been reserved for the test of the proposed recommender system, while the other 258 records have been used as the input data that provides the information necessary for the system to compute the recommendation scores. We can see that the roles of these two data sets are very similar to those of the test set and training set, however, it is worth noting that the proposed system is designed for recommendation,

not for classification or prediction. That is, our objective is to identify the examinees that need to be careful for the dental caries, not to approximate the values of y in the reserved data set accurately.

Table 2. Response Variables

Variable	Description
r_1	Dental caries
r_2	Dental caries risk
r_3	Oral hygiene
r_4	Gingival bleeding
r_5	Tartar
r_6	Tooth fracture
r_7	Dental pain triggered by cold or hot foods and beverage
r_8	Dental pain
r_9	Bleeding gums
r_{10}	Glossalgia
r_{11}	Halitosis

The reserved data set is listed in Table 3, where we can see that 50% of the examinees have dental caries while the others do not. Moreover, our objective is to compute the recommendation scores for the dental caries of the 20 examinees.

4.2 Recommendation Scores for Variable 'Dental Caries'

In order to make recommendations by using the proposed recommender system, we have to generate the profiles of the response variables, P_es ($1 \le e \le 11$) at first. The profiles obtained by using eq. (1) are listed in Table 4, where the confidence is used as the performance measure I_{ie} (i=1, 2, …,10).

Next, the cosine similarity $S_{cos}(P_e, P_1)$ and the Jaccard similarity J(e,1) in Eq. (2) have been computed as shown in Table 5 and Table 6 ($2 \le e \le 10$). Then, we can obtain $RS_{INTER/R}$ by using Table 3~6 and Eq. (4), and RS_{INTER} by using Table 3~4 and Eq. (5).

In computing $RS_{INTER/P}$, the posterior probability $P(v \mid F_t)$s have been obtained by using Table 3 and the conventional data mining software[23]. Consequently, the three recommendation scores for the dental caries of each examinee can be computed as shown in Table 7. Note that RS_{TOTAL} in the fifth column of Table 7 represents the sum of $RS_{INTER/R}$, RS_{INTER} and $RS_{INTER/R}$.

In Table 7, we can see that $RS_{INTER/R}$ =0 for examinee 6, 10, 15 and 20. This is because those 4 examinees do not have any symptoms or diseases concerning the response

Table 3. Reserved Data Set

id	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8	r_9	r_{10}	r_{11}
1	1	0	1	0	1	1	1	0	0	0	**0**	0	0	0	0	0	0	1	0	0	0
2	1	1	0	0	0	0	1	0	0	0	**0**	0	0	0	0	0	1	0	0	0	1
3	1	0	1	0	0	0	1	0	0	1	**0**	0	0	0	0	0	0	1	1	1	0
4	0	0	0	0	0	0	1	0	1	0	**0**	0	0	0	0	0	0	0	1	0	0
5	0	0	1	0	1	0	1	0	0	0	**0**	0	1	0	1	0	0	0	0	0	0
6	0	0	1	0	0	1	1	0	1	0	**0**	0	0	0	0	0	0	0	0	0	0
7	0	0	1	0	0	1	1	0	0	0	**0**	1	1	0	0	0	0	0	0	1	0
8	1	0	1	0	0	0	1	0	0	0	**0**	0	0	0	0	0	1	0	0	0	0
9	1	0	0	0	0	0	1	0	0	0	**0**	0	0	0	0	1	0	0	0	0	0
10	0	0	1	0	0	0	1	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0
11	0	1	1	0	0	0	1	1	0	0	**1**	0	1	0	0	0	1	0	1	0	1
12	0	0	1	1	0	0	1	0	0	0	**1**	0	1	0	0	0	0	0	1	0	1
13	0	0	1	1	1	0	1	1	1	0	**1**	0	0	0	0	0	1	0	0	0	0
14	0	0	1	0	0	0	0	1	0	0	**1**	0	1	0	0	0	0	0	0	0	0
15	0	0	1	1	0	0	1	1	0	0	**1**	0	0	0	0	0	0	0	0	0	0
16	1	0	1	0	1	0	1	0	0	0	**1**	0	1	0	0	0	0	0	0	0	0
17	1	1	1	1	0	0	1	1	0	1	**1**	0	0	0	0	0	0	0	1	0	0
18	0	0	1	0	1	1	1	0	1	0	**1**	0	0	0	0	0	0	0	0	0	1
19	1	1	1	0	1	1	1	0	0	0	**1**	0	1	0	0	0	0	0	0	0	0
20	0	0	1	1	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0

Table 4. Response Variable Profiles based on Confidence

Profile	I_{1e}	I_{2e}	I_{3e}	I_{4e}	I_{5e}	I_{6e}	I_{7e}	I_{8e}	I_{9e}	I_{10e}
P_1	0.2683	0.3429	0.2903	0.3000	0.2734	0.3097	0.2765	0.3488	0.3889	0.3939
P_2	0.0488	0.1286	0.1014	0.1300	0.1172	0.1327	0.0968	0.0233	0.1250	0.2121
P_3	0.0976	0.1714	0.1244	0.1300	0.1172	0.1416	0.1152	0.2093	0.1806	0.1818
P_4	0.0244	0.0714	0.0507	0.0600	0.0469	0.0442	0.0415	0.0233	0.0972	0.0000
P_5	0.0854	0.1286	0.0876	0.1100	0.1016	0.0973	0.0829	0.1163	0.0694	0.0606
P_6	0.2195	0.2429	0.1843	0.1700	0.1797	0.1947	0.2120	0.2093	0.2083	0.3030
P_7	0.2927	0.3000	0.2765	0.2400	0.3047	0.3097	0.2857	0.2326	0.3056	0.3030
P_8	0.1707	0.0714	0.0876	0.1100	0.1016	0.0973	0.1152	0.1628	0.1389	0.2121
P_9	0.1707	0.1857	0.2028	0.2300	0.2109	0.2035	0.1935	0.2326	0.2639	0.2121
P_{10}	0.0610	0.0429	0.0461	0.0500	0.0625	0.0088	0.0553	0.0000	0.0278	0.0303
P_{11}	0.2317	0.2286	0.1889	0.1500	0.1797	0.2212	0.1889	0.2093	0.2917	0.2424

Table 5. Cosine similarity among the response variable profiles

Response variable	r_2	r_3	r_4	r_5	r_6	r_7	r_8	r_9	r_{10}	r_{11}
r_1	0.9329	0.9921	0.8698	0.9614	0.9908	0.9875	0.9603	0.9931	0.8367	0.9919

Table 6. Jaccard similarity measure among the response variables

Response variable	r_2	r_3	r_4	r_5	r_6	r_7	r_8	r_9	r_{10}	r_{11}
r_1	0.1149	0.3784	0.1351	0.1687	0.2227	0.2069	0.1279	0.1415	0.0494	0.1485

variables r_2, r_3, ..., r_{11}. Among the 4 examinees, examinee 6 and 10 does not have the dental caries ($r_1 = 0$), while $r_1 = 1$ for the other two, as shown in Table 3. This indicates that the risks of the dental caries for the examinee 6 and 10 are currently very low because of their good dental health status. Although the examinee 15 and 20 have got some cavities in the past, it seems that they are caring about their dental health in appropriate manner and the risk of additional dental caries is currently low. In addition, these discussions reveal the complicated aspects of the interpretation of the recommendation results. Recommendation scores are not about approximation of the target variable, as opposed to the outputs of classification.

Among the examinees with $r_1 = 0$, the examinee 3, 4 and 6 have high RS_{INTER} values. This means that those examinees have life styles that can cause the dental caries. For example, we can see that the examinee 3 brushes teeth less than once a day ($f_{10} = 1$), as shown in Table 3. Therefore, they have to take care of their dental health appropriately, although they currently have no dental caries. Similarly, the examinee 13, 17 and 18 also have inappropriate life styles, and they have developed cavities already.

Table 7. Recommendation scores for dental caries

id	$RS_{INTER/R}$	RS_{INTER}	$RS_{INTER/F}$	RS_{TOTAL}
1	0.1228	0.3097	0.1332	0.5657
2	0.2043	0.3429	0.2073	0.7545
3	0.1405	0.3939	0.0940	0.6284
4	0.1405	0.3889	0.1288	0.6582
5	0.3754	0.2903	0.1214	0.7871
6	0.0000	0.3889	0.1826	0.5715
7	0.3754	0.3097	0.1717	0.8568
8	0.2043	0.2903	0.1053	0.5999
9	0.2256	0.2765	0.1907	0.6928
10	0.0000	0.2903	0.1085	0.3988
11	0.3754	0.3488	0.1041	0.8283
12	0.3754	0.2903	0.1539	0.8196
13	0.2043	0.3889	0.2515	0.8447
14	0.3754	0.3488	0.2063	0.9305
15	0.0000	0.3488	0.1920	0.5408
16	0.3754	0.2903	0.1118	0.7775
17	0.1405	0.3939	0.1933	0.7277
18	0.1473	0.3889	0.1899	0.7261
19	0.3754	0.3429	0.1800	0.8983
20	0.0000	0.2903	0.2295	0.5198

In the fourth column of Table 7, we can see that the value of $RS_{INTER/F}$ for examinee 2 is high, although he or she has no dental caries. This can be explained by the high probability of f_3, 'disuse of fluoride toothpaste' $P(f_3 \mid \text{examinee } 2) = 0.71$). That is, the examinees similar in terms of life styles tend not to care about their toothpaste, even though this examinee currently uses the fluoride toothpaste. Therefore, it is recommendable to remind the examinee 2 of the importance of the toothpaste. Similarly, the examinee 13, 14 and 20 with $r_1 = 1$ also have high $RS_{INTER/F}$ values, and this suggests that their potential inappropriate life styles have caused the dental caries already.

In order to decide if a single variable (item) is recommendable for an individual examinee, we need the decision thresholds for the recommendation scores. In this paper, a decision threshold is simply calculated as (average(-) + average(+))/2, where the average(-) and the average(+) denote the average recommendation scores of the negative cases with y = 0 and the positive cases with y = 0, respectively. The decision thresholds for r_1, the dental caries, are summarized in Table 8. In Table 8, we can see that average(-) < average(+) for all recommendation scores, and this indicates that the proposed recommender system appropriately evaluates the risk of the target variable on the whole.

Table 8. Decision thresholds for dental caries

	$RS_{INTER/R}$	RS_{INTER}	$RS_{INTER/F}$	RS_{TOTAL}
Average(–)	0.1789	0.3281	0.1444	0.6514
Average(+)	0.2369	0.3432	0.1812	0.7613
Decision threshold	0.2079	0.3357	0.1628	0.7064

Next, we have applied the proposed recommender system to the classification for the target variable r_1, in order to further investigate the performance of the system. To this end, the reserved data set in Table 3 and the remaining input data set have been used as the test set and the training set, respectively. In more detail, a single recommendation score has been used for classification at a time, and an individual case of the test set is classified as a positive case (r_1=1) if and only if the calculated score is greater than or equal to the associated decision threshold. In addition, we also have applied the naïve Bayesian classifier, one of the most widely used classification method, to those data sets for comparison.

The classification results are summarized in Figure 2, where accuracy(-), accuracy(+) and accuracy_total denote the accuracies for the negative cases, the positive cases and the entire cases, respectively. As aforementioned, the proposed system is designed for recommendation, not for classification, however, we can see that the classification performance of the system is superior to that of the naïve Bayesian classifier.

As shown in Figure 2, accuracy(-) of the naïve Bayesian classifier is 100%, which means that all negative test cases are correctly classified, however, accuracy(+) of this conventional classifier is 0%. Indeed, the naïve Bayesian classifier have classified all test cases as the negative ones, and its overall accuracy is 50%. Again, this poor performance of the naïve Bayesian classifier shows the complex relationships within the mass health examination result data set.

On the contrary, the overall accuracy of the classifier based on $RS_{INTER/R}$ is 65%, superior performance to that of the naïve Bayesian classifier, where its accuracy(-) and accuracy(+) are 80% and 50%, respectively. Hence, we can see that the classifier based on $RS_{INTER/R}$ is more suitable for the negative test cases. The overall accuracies of the classifier based on RS_{INTER} and the classifier based on $RS_{INTER/F}$ are also 65%, same performance with the classifier based on $RS_{INTER/R}$. However, their accuracy(-) and accuracy(+), 60% and 70%, indicate that they are more suitable for the positive test cases.

Furthermore, the classifier based on RS_{TOTAL} has classified 75% of the test cases correctly, and this is the highest overall accuracy value in Figure 2. Note that RS_{TOTAL} is the total sum of $RS_{INTER/R}$, RS_{INTER} and $RS_{INTER/R}$, where no weights for those three basic recommendation scores are considered. So, it is expected that more complicated combination of the three recommendation scores, such as the weighted sum, can produce better performances.

As a result, we can draw several conclusions: (i) The recommendation scores proposed in this paper is helpful not only for recommendation but also for classification. (ii) The performances of the proposed recommender system are competitive against the conventional classification methods. For example, the simple classifiers based on the proposed recommendation scores have shown superior performances to the naïve Bayesian classifier. (iii) Different recommendation scores can be suitable for different cases. For example, one recommendation score can be especially suitable for the positive or negative cases. (iv) The performance of the recommender system can be enhanced by combination of the basic recommendation scores. For example, the classifier based on the total sum

of the scores RS_{TOTAL} has shown the best performance in the experiment.

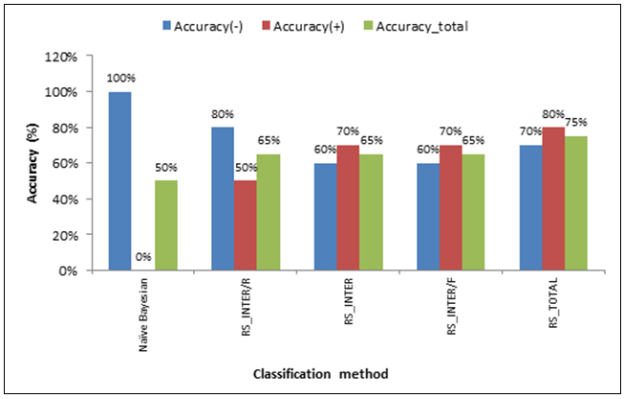

Figure 2. Classification performances (accuracy).

5. Conclusions

The variables within the high dimensional data sets often have quite complex relationships difficult to analyze. The first contribution of this paper is framework called SANM that provides a simple view for the variables within a given data set. The SANM consists of the factor variables, the response variables, and the three types of cause-and-effect relations among them. Due to its simple structure, a SANM for a given data set can be constructed by simply identifying the factor and the response variables, and it is expected that this framework can be used in a wide range of application domains.

Recommendation is one of such application of SANM, and the second contribution of this paper is a novel intelligent recommender system especially suitable for analyzing the high dimensional transaction data set with complex relationships among the variables. The proposed recommender system consists of three recommendation scores, $RS_{INTER/R}$, RS_{INTER} and $RS_{INTER/F}$, which are calculated based on SANM and the data mining techniques such as association rule mining and naïve Bayesian classifier.

From the technical perspective, the proposed recommender system has following benefits: (i) In contrast to the many previous recommender systems based on data mining or artificial intelligence techniques, we need not to construct any knowledge base or rule base. This enables the practitioners to develop the recommender system more conveniently. (ii) While the conventional data mining techniques involves complex and time-consuming procedures for extracting the rules or patterns from the given data set, the calculations of the proposed recommendation scores are relatively simple and intuitive. (iii) The complex relationships among the variables can be dealt with appropriately by adopting the SANM, and we need not to concern about the structure of the cause-and-effect relations within the given data set.

For illustration, we have applied the proposed system to a mass health examination result data set, and the experiment results led us to draw following conclusions: (i) The recommendation scores proposed in this paper is helpful not only for recommendation but also for classification. (ii) Compared to the conventional classification methods, the proposed system has competitive performances. (iii) Different recommendation scores can be suitable for different cases. (iv) The performance of the recommender system can be enhanced by combination of the basic recommendation scores.

Despite the promising features of the proposed recommender system, we still have several further topics for future research. Firstly, the recommender system cannot be applied to the data sets with continuous variables or multilevel categorical variables, since it has been designed to analyze the transaction data sets with binary variables. This problem can be partly addressed by using binarization, however, it is well known that such procedure can cause the loss of information and hamper efficient analysis. Therefore, we plan to extend the current recommender system so that it can be directly applied to various types of variables. Secondly, $RS_{INTER/R}$ in this paper is 0 if a given case has no positive response variables, as shown in Table 7. That is, sparse responses can degrade the performance of the proposed system, and this should be addressed by applying the revised procedures for calculation of the recommendation scores. Finally, the experiment results in this paper suggest that the proposed recommendation scores can also be used for classification analysis. In other words, the recommendation scores can be used as a feature construction method, and we will further investigate the performances of such classifiers based on the proposed recommendation scores.

5. Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation

of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2012R1A1A1044834).

6. References

1. Lu J, Wu D, Mao M, Wang W, Zhang G. Recommender system application developments: a survey. Decision Support Systems. 2015 Jun; 74(C):12-32.
2. Lee S, Lee E, Park J-M. An optimal sorting algorithm for mobile devices. Indian Journal of Science and Technology. 2015 Apr; 8(8):226-32.
3. Kim KJ, Ahn H. A recommender system using GA K-means clustering in an online shopping market. Expert Systems with Applications. 2008 Feb; 34(2):1200-209.
4. Kim JW, Ha SH. Price comparisons on the internet based on computational intelligence. Plos-One. 2014 Sep; 9(9):e106946.
5. Felfernig A, Teppan E, Gula B. Knowledge-based recommender technologies for marketing and sales. International Journal of Pattern Recognition and Artificial Intelligence. 2007 Mar; 21(2):333-55.
6. Ngai EW, Xiu L, Chau DC. Application of data mining techniques in customer relationship management: a literature review and classification. Expert Systems with Applications. 2009 Mar; 36(2):2592-602.
7. Musto C, Semeraro G, Lops P, de Gemmis M, Lekkas G. Personalized finance advisory through case-based recommender systems and diversification strategies. Decision Support Systems. 2015 Sep; 77(C):100-111.
8. Hsu MH. A personalized English learning recommender system for ESL students. Expert Systems with Applications. 2008 Jan; 34(1):683-88.
9. Duan L, Street WN, Xu E. Healthcare information systems: data mining methods in the creation of a clinical recommender system. Enterprise Information Systems. 2011 Jan; 5(2):169-81.
10. Bodadilla J, Ortega F, Hernando A, Gutierrez A. Recommender systems survey. Knowledge-Based Systems. 2013 Jul; 46:109-32.
11. Vialardi C, Bravo Agapito J, Shafti LS, Ortigosa A. Recommendation in higher education using data mining techniques. Spain: Proceedings of the 2nd International Conference on Educational Data Mining. 2009 Jul; p. 1-10.
12. Bridge D, Goker MH, McGinty L, Smyth B. Case-based recommender systems. The Knowledge Engineering Review. 2005 Sep; 20(3):315-20.
13. Azaria A, Hassidim A, Kraus S, Eshkol A, Weintraub O, Netanely I. Movie recommender system for profit maximization. China: Proceedings of the 7th ACM Conference on Recommender Systems. 2013 Oct; p. 1-8.
14. Ramezani A, Dehkordi MN, Esfahani FS. Hiding sensitive association rules by elimination selective item among RHS items for each selective transaction. Indian Journal of Science and Technology. 2014 Jun; 7(6):826-32.
15. Agrawal R, Imielinski R, Swami R. Mining associations between sets of items in massive databases. USA: Proceedings of the ACM-SIGMOD 1993 International Conference on Management of Data. 1993; p. 1-10.
16. Agrawal R, Srikant R. Fast algorithms for mining association rules. Chile: Proceedings of the International Conference on Very Large Databases. 1994; p. 1-13.
17. Tan PN, Steinbach M, Kumar V. Boston: Addison-Wesley: Introduction to data mining. 1st edn. 2005.
18. Hipp J, Guntzer U, Nakhaeizadeh G. Algorithms for association rule mining - a general survey and comparison. ACM SIGKDD Explorations Newsletter. 2000 Jun; 2(1):58-64.
19. Lin W, Alvarez SA, Ruiz C. Efficient adaptive-support association rule mining for recommender systems. Data Mining and Knowledge Discovery. 2002 Jan; 6(1):83-105.
20. Garcia E, Romero C, Ventura S, de Castro C. A collaborative educational association rule mining tool. The Internet and Higher Education. 2011 Mar; 14(2):77-88.
21. Yager RR. An extension of the naive Bayesian classifier. Information Sciences. 2006 Mar; 176(5):577-88.
22. Yin Y, Yasuda K. Similarity coefficient methods applied to the cell formation problem: a comparative investigation. Computers and Industrial Engineering. 2005 May; 48(3):471-89.
23. Weka, data mining software in Java. Date accessed: 03/21/2016: Available from: http://www.cs.waikato.ac.nz/ml/weka.