Industrial Engineering
& Management Systems
Vol 15, No 1, March 2016, pp.110-121
ISSN 1598-7248 | EISSN 2234-6473 |
http://dx.doi.org/10.7232/iems.2016.15.1.110

Discovering Relationships between Skin Type and Life Style Using Data Mining Techniques: A Case Study of Korea

Taeheung Kim, Jihyun Ha, Jong-Seok Lee*
Department of Industrial Engineering, Sungkyunkwan University, Suwon, Korea

Younhak Oh
Nielsen Company Korea, Seoul, Korea

Yong Ju Cho
Korea Institute of Industrial Technology, Cheonan, Korea

(Received: March 1, 2016 / Accepted: March 8, 2016)

ABSTRACT

With the growing interest in skincare and maintenance, there are increasing numbers of studies on the classification of skin type and the factors influencing each type. This study presents a novel methodology by using data mining, for the determination of the relationships between skin type, lifestyle, and patterns of cosmetic utilization. Eight skin-specific factors, which are moisture, sebum in U-zone (both cheeks), sebum in T-zone (forehead, nose, and chin), pore, melanin, wrinkle, acne, hemoglobin, were measured in 1,246 subjects living in South Korea, in conjunction with a questionnaire survey analyzing their lifestyles and pattern of cosmetic utilization. Using various multivariate statistical methods and data mining techniques, we classified the skin types based on the skin-specific values, determined the relationship between skin type and lifestyle, and accordingly sorted the subjects into clusters. Logistic regression analysis revealed gender-related differences in the skin; therefore, separate analyses were performed for males and females. Using the Gaussian Mixture Modeling (GMM) technique, we classified the subjects based on skin type (two male and four female). Using the ANOVA and decision tree techniques, we attempted to characterize the relationship between each skin type and the lifestyles of the subjects. Menstruation, eating habits, stress, and smoking were identified as the major factors affecting the skin.

Keywords: Skin Type, Life Style, Data Mining, Gaussian Mixture Model, Decision Tree

* Corresponding Author, E-mail: jongseok@skku.edu

1. INTRODUCTION

Rapid changes in modern society have led to several changes in sociocultural trends. In addition, the progress of the health industry has steadily improved life expectancy and increased the efforts exerted towards the maintenance of a sound body and mind. The development of the trends related to general well-being is an expression of the desire for quality improvement of health, in addition to disease management. In particular, the quality improvement of skin is a major factor that has been attracting growing interest for its aesthetic value. According to a survey conducted by Datamonitor in 2013, the cosmetic industry has seen steady growth since 2001,

with skincare leading this upward trend.[1] Additionally, there has been a recent, growing interest in skin type, as demonstrated by the point of sale expert service provided by many skincare brands, which offer consumers products tailored to their individual skin types.[2]

There are various methods for the classification of skin type. Initially, a research study had proposed the classification of skin types based on the response of skin to sun exposure and a congenital factor (Fitzpatrick, 1998). Based on the sun exposure method, the skin type was classified between types I through VI, depending on the degree of burn and tanning. The skin type was also classified by skin tone according to the genetic factor. In recent times, four-dimensional classification criteria: dry and oily; sensitive and resistant; pigmented and non-pigmented; and wrinkled and tight, have been established (Baumann, 2008). Many other methods have been developed for skin typing. However, the typical skin typing method adopted in the cosmetic industry and its associated research areas, is the classification based on the sebum excretion rate (SER), which classifies skin types as normal, dry, oily/greasy, and combined (Fur *et al*., 1999; Kim *et al*., 2011). Skin types, especially the facial skin types, cannot be permanently fixed but does undergo changes according to the major intrinsic and extrinsic variables such as those in climate, stress, and menstrual status (Baumann, 2008). Therefore, most research has been focused on the identification of the factors influencing skin type.

According to the research findings in Fur *et al*. (1999), there is no consistent criterion to distinguish the skin of Caucasian women into normal, dry, greasy, and combined skin types. However, a larger quantity of sebum was observed in greasy and combined skin types than the others, which implies the significance of sebum production in the determination of the cosmetic skin type (Fur *et al*., 1999; Kumagai *et al*., 1985). In addition, dryness of the skin was confirmed to be unrelated to the quantity of skin surface lipids (Fur *et al*., 1999; Pierard, 1987; Rurangirwa *et al*., 1987). Skin melanin, hemoglobin, and light scattering properties were quantitatively analyzed, and it was found that melanin is a major factor for the assessment of a particular type of skin, while the scattering properties provide information about the skin structure and morphology (Zonios *et al*., 2001). The experiments on Chinese subjects have verified the influence of gender differences on skin surface pH, sebum content, and stratum corneum hydration, with sebum content particularly, showing a marked difference among different genders (Man *et al*., 2009). The research in Luebberding *et al*. (2014) has studied the influence of gender differences on skin elasticity, focusing on the age-dependent changes in Caucasian men and women. It was discovered that skin elasticity shows different changing patterns in males and females, with menopause triggering a rapid reduction in skin elasticity. From this, it could be inferred that skin properties demonstrate gender-related differences.

External environments also affect the skin properties. Wendling and Dell'Acqua (2003) conducted a study involving people living in the Swiss Canton of Valais, which is located at an extremely high altitude, and discovered that they showed a low skin hydration viscoelasticity. This could be attributed to environmental factors of the region, such as elevated sun irradiation and low humidity. In addition, Youn *et al*. (2005) have found in their study with Korean subjects living in a four-season climate zone that sebum secretion increases during the summer, as compared to the other seasons. Several studies have also focused on the different skin-aging patterns between Europeans and Asians (Nouveau-Richard *et al*., 2005; Tsukahara *et al*., 2003). Among the Asian countries, differences in skin aging and facial wrinkles between the Japanese, Chinese, and Cantonese populations (Tsukahara *et al*., 2007), and the differences in hydration, transdermal water loss, sebum level, pores, pH, elasticity, wrinkles, skin brightness, and sensitivity among the people in Korea, Indonesia, Thailand, and Malaysia have been previously investigated.[3] In particular, Galzote *et al*. (2013) have measured the skin properties in women from China, Korea, Japan, and the Philippines and investigated the impact of age and skincare habits on the same. This study indicated that fewer wrinkles appeared, with respect to the subjects' age brackets, with the earlier use of skincare products. Moreover, in a study involving aged (≥ 65) Japanese subjects, Asakura *et al*. (2008) have discovered that smoking and use of sunscreen correlated to the condition of the skin.

A large number of diverse skin-related studies have been conducted in South Korea. Kim *et al*. (2011) have observed a significant association between facial pores, sebum secretion, and skin elasticity, and a direct association between acne lesion count and facial pores, in men. However, the acne lesion count and age were discovered to not significantly affect facial pore formation. Park *et al*. (1999) have presented two skin typing systems: determination of the sebum excretion rate (SER) and skin surface relief (SSR). SER and SSR are influenced by hormones and skin surface morphology, respectively. Therefore, we have assumed that this easily measurable skin typing method would provide the consumers of cosmetic products with a simple tool for choosing skin type-specific products. Kim *et al*. (2011)

1) 2013 Consumer Survey Data-Personal Care, Available at http://www.datamonitor.com/store/Product/2013_consumer_survey_data_personal_care?productid=CM00270-002 (accessed 4 January 2015).

2) GCI Magazine, Market for Personalized Skincare Continues to Show Potential, Available at http://www.gcimagazine.com/marketstrends/segments/skincare/Market-Personalized-Skin-Care-Potential-275465071.html (accessed 4 January 2015).

3) Korea Health Industry Development Institute (KHIDI) (2013), Beauty Industry Analysis 2013 Available at http://www.khidi.or.kr/board/view?linkId=100680&menuId=MENU00085 (accessed 4 January 2015).

have confirmed that lifestyle and skincare patterns such as the frequency of alcohol consumption and home care, respectively, affect the health of the skin. They have also identified factors such as the body mass index (BMI), residential type, lifestyle (amount of water intake, beverage intake, and alcohol consumption frequency), eating habits (prioritized food, fruit intake frequency, dairy intake frequency, seaweed intake frequency, and fast food intake frequency), and skincare patterns (facial cleansing frequency, method of water usage during facial cleansing, skincare, color makeup, etc.) that influence the skin types. They have suggested that the skin pH is influenced by age, BMI, amount and frequency of alcohol consumption, exercise frequency and duration, stress management methods, prioritized food types, legume and meat/seafood/egg intake frequencies.

This study aims to explore and determine the relationship between skin type and lifestyle, by applying a data mining technique yet not attempted in the dermatology sector. As mentioned above, skin types are currently being classified based on the independent examination of the individual skin characteristics and the setting of threshold values. In contrast, this study adopted a cluster analysis, where the subjects were grouped into clusters, with respect to the interrelated skin-specific measurement values. Based on the results, we performed a classification modeling, where we represented the identified clusters as a response variable, and the lifestyle-related variables as independent variables. The data on current lifestyle trends were collected from the subjects by means of a questionnaire. Among the various classification techniques, we adopted the decision tree analysis to determine the differences in skin types based on the lifestyles of the subjects. As the logistic regression analysis performed prior to clustering and classification modeling revealed a gender-based difference in skin characteristics, both clustering and classification modeling were conducted separately for male and female subjects.

2. MATERIALS AND METHODS

2.1 Subjects and Measurements

In this survey, we included 1,246 adults aged 16-75 (mean 41±16 standard deviation (SD); female: 934, male: 312). The variables were largely divided into two categories: 8 variables represented the skin characteristics measured by using measurement devices, and 18 variables represented the results of the life style questionnaire, regarding the use of basic and hair/scalp cosmetics. Table 1 outlines the descriptive statistics of the variables pertaining to skin characteristics, while Table 2 summarizes of the results of the lifestyle questionnaire.

In order to ensure the accuracy of data analysis, we subjected the data to a pre-processing step. We deleted 32 subjects either having missing values or outlying. After preprocessing the data, 1,214 subjects were included in the data mining analysis. The variables of the questionnaire were converted to values between 0-1 for scale adjustment, and a binary derived variable was added for the female subjects, in order to separately examine the effects of menstruation on the status of the skin. This assisted in the addition of the factor, 'irregular menstruation' to the existing 'menopausal status' and 'regular menstruation' factors.

Table 1. Descriptive statistics of skin measurements

Variable	Mean	Median	StDev
Moisture	60.07	60	12.51
SebumU	3.93	1	9.29
SebumT	10.69	2	16.12
Pore	48.48	47	15.76
Melanin	45.78	52	27.47
Wrinkle	46.11	51	16.44
Acne	32.70	20	29.70
Hemoglobin	25.76	17	22.47

2.2 Data Mining Techniques

Logistic regression analysis was adopted in this study in order to determine the presence of gender differences in the manifestation of skin characteristics. We also used a Gaussian mixture model to group the subjects according to the measured skin characteristics, and the decision tree model was used for lifestyle assessment for each group.

Logistic regression is a linear model primarily used as a classification method when the response variable is categorical (Sharma, 1995). It does not require the proposition of hypotheses for the distribution of independent variables. The probability of a given object to belong to a particular class based on independent variables is expressed by

$$p = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)} \quad (1)$$

where the regression coefficients (β_j 's) are estimated by the method of maximum likelihood estimation for the given data. The advantage of logistic regression lies in the fact that the relationship between a set of independent variables and a response variable can be analyzed by using the estimated coefficients (Lee *et al*., 2014). A large value of coefficients would mean that the corresponding independent variables exert a strong influence on the response variable. On the contrary, if the value of the coefficients is approximately zero, they are hardly believed to influence the response variable. The gender-related significance of the data derived from the study was compared against a logistic regression model. When

Table 2. Life style and use of cosmetics

Category	Questionnaire item	Likely (Yes)	Medium Likelihood	Unlikely (No)
Life style				
Eating habits	Blue fish intake	191	460	413
	Daily vegetable consumption	364	621	261
	Regular meals	226	448	572
	Avoiding fast food	372	461	412
Skincare habits	Use of skin-type basic cosmetics	384	438	423
	Double facial cleansing	390	290	564
	Use of functional cosmetics	449	366	429
	Skin care by experts	62	173	1,008
	Skincare at home	128	346	768
Living habits	Sufficient sleep and rest	237	478	527
	Sufficient water intake	232	494	516
Health status	Healthy without any particular complaints	578	532	135
	Exposure to stress	441	650	155
	Menstruation or absence thereof	659	NA	464
	Regular menstruation	409	NA	250
Tobacco and alcohol	Smoking	150	NA	1,091
	Drinking	519	NA	724
Use of cosmetics				
Basic cosmetics	Cleanser		816	
	Foam cleanser		756	
	Deep cleanser		143	
	Skin lotion		757	
	Lotion		1,004	
	Essence		660	
	Facial cream		657	
	Eye cream		365	
	Pack/Mask		282	
	Sunscreen		687	
Cosmetics for hair and scalp care	Shampoo		1,178	
	Rinse		682	
	Treatment product		314	
	Hair growth/Hair restoration products		17	
	Dandruff remover		19	

the logistic regression analysis was performed with one of the measured variables as an independent variable and the gender as the response variable and if the coefficient check against the independent variable does not yield a significant value, this measured value could be concluded to be unaffected by gender difference. If the coefficient proves significant, a gender-dependent model can be set up, as the measured variables are believed to have significant gender differences.

The Gaussian Mixture Model (GMM) is a clustering technique, where it is postulated that there are k number of pre-determined clusters, with each cluster following the multivariate Gaussian distribution (McLachlan and Peel, 2004). The mixture of k components is thus given by

$$p(\mathbf{x}) = \sum_{i=1}^{k} p_i f_i(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{2}$$

where a priori probability p_i is the faction of the objects in cluster i such that $\Sigma p_i = 1$, and

$$f_i(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{((2\pi)^m |\boldsymbol{\Sigma}_i|)^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^{\mathrm{T}} \Sigma_i^{-1} (\mathbf{x}-\boldsymbol{\mu}_i)\right\} f_i(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{3}$$

where m is the dimension of the data objects, and $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are the parameters that are said to be estimated by using the expectation-maximization algorithm. Because of its efficacy in estimating the density or distribution hidden in data, and in interpretation of models, the

GMM is widely used by researchers in a variety of academic disciplines. GMM can also be applied efficiently during the pre-treatment stage, as it can be used simultaneously for clustering and the detection of outliers (Roberts, 1999). In this study, we have attempted to determine the optimal number of clusters by applying GMM to the data (representative of the measurement of skin characteristics), and simultaneously classify the skin type of the subjects (by grouping them into several clusters according to their skin characteristics).

C4.5 is one of the most frequently used decision tree analysis methods (Quinlan, 1993; Wu *et al*., 2008), and it uses information entropy as a selection criterion for the splitting of variables. The completed tree model is expressed as a set of easy-to-analyze rules, and it does not require any hypothesis with respect to data distribution (Tang *et al*., 2005), similar to logistic regression. Like other tree techniques, C4.5 uses a pruning algorithm to prevent the side-effect of over fitting of the final model with the data. The reduced error pruning method was used in this study.

3. RESULTS

3.1 Gender Difference Validation

Prior to cluster analysis by using the 8 measured variables pertaining to skin conditions, a pre-testing process was conducted to determine whether each measured variable shows a gender-dependent difference in distribution. In order to implement the test, we performed a logistic regression analysis by setting each of the 8 measured variables as an independent variable (X), and the gender as the response variable. Specifically, 8 models were generated by setting the likelihood of the subject belonging to the male class set to be p. This is expressed by

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X \tag{4}$$

Wald χ^2 statistic is primarily employed to test the statistical significance of the estimated regression coefficient $\hat{\beta}_1$, and the χ^2 statistic extracted from each model is listed in Table 3.

Table 3. Significance test of individual coefficients

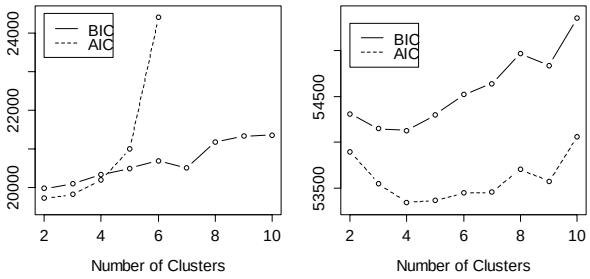
Variable	χ^2 statistic	p-value
Moisture	1.14	0.2858
SebumU	55.93	< 0.001
SebumT	69.77	< 0.001
Pore	109.4	< 0.001
Melanin	57.26	< 0.001
Wrinkle	15.43	< 0.001
Acne	62.23	< 0.001
Hemoglobin	4.6	0.0310

According to the χ^2 statistic, pore is the variable showing the largest gender difference. At a significance level $\alpha = 0.05$, all variables except moisture were found to be significant, which led us to conclude that all future analyses must be conducted separately for males and females. For all subsequent analyses, therefore, the data collected was classified with respect to the gender and subjected to separate analyses.

3.2 Skin Type Classification

As briefly mentioned above, we used GMM clustering, in order to group the subjects, and attempted a skin type classification by analyzing the 8 skin-related features in each group. In order to determine the optimal number of clusters, we observed the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) while increasing the number of clusters. In general, the optimal number is determined as the one representing the smallest AIC and BIC values. Figure 1 illustrates the resulting AIC and BIC values according to different numbers of clusters by GMM clustering of male and female data. Accordingly, we determined the number of male and female clusters at 2 and 4, respectively.

Table 4 shows the results of GMM grouping of male and female subjects into clusters, based on the obtained number of the clusters, as described above. Subjects belonging to cluster 0 were those that could not be grouped into any specific cluster based on the GMM and were hence considered to be outliers. For example, male subjects (n = 49) belonging to cluster 0 have a very low probability of belonging to either cluster 1 or 2. Therefore, subjects belonging to cluster 0 were excluded from subsequent analyses.



(a) male (b) female

Figure 1. AIC and BIC values at varying number of clusters.

Table 4. Number of subjects grouped into clusters

Cluster	Male		Female	
	Count	Proportion	Count	Proportion
1	160	0.52	234	0.25
2	93	0.30	52	0.06
3	NA	NA	422	0.46
4	NA	NA	153	0.17
0	49	0.18	51	0.06

Table 5. Number of subjects grouped into clusters

Variable	Male		Female	
	F-stat	*p*-value	*F*-stat	*p*-value
Moisture	0.05	0.8229	2.16	0.0915
SebumU	**144.89**	**< 0.0001**	**791.46**	**< 0.0001**
SebumT	**416.06**	**< 0.0001**	**590.03**	**< 0.0001**
Pore	**11.77**	**0.0007**	**3.91**	**0.0086**
Melanin	0.03	0.8528	**12.41**	**< 0.0001**
Wrinkle	0.03	0.8543	1.51	0.2107
Acne	**7.68**	**0.0060**	**211.33**	**< 0.0001**
Hemoglobin	0.09	0.7676	**12.54**	**< 0.0001**

ANOVA (analysis of variance) was performed to determine the significant variables characterizing each cluster. Specifically, after establishing the results of clustering as a factor and skin-specific values as response variables, we attempted to determine whether the different clusters show significantly different mean values on individual skin-specific values. The results of this testing are summarized in Table 5.

The highlighted cases in the above table represent the skin-specific values with significant inter-cluster difference, at a significance level of $\alpha = 0.01$. For male subjects, the skin-specific values of sebumU, sebumT, pore, and acne, showed statistically significant inter-cluster differences, while in female subjects the sebumU, sebumT, pore, melanin, acne, and hemoglobin factors demonstrated statistically significant differences among the four clusters. In order to determine the characteristics of each cluster, box plots were drawn for the variables displaying all significant differences. Figure 2 displays all box plots.

The box plots indicate that the division of male subjects into the clusters was intuitively reasonable. The sebumU, sebumT, pore, and acne values of the male cluster 1 were observed to be lower than those of male cluster 2. In other words, cluster 1 was classified as a dry-skin subject group, while cluster 2 comprised of oily-skin subjects. As displayed in Table 4, over half of all women subjects belonged to the female clusters 1 and 3, both high-melanin clusters. When compared to the other female clusters, the female cluster 3 showed a low-acne tendency. The female cluster 2, comprising of only 6% of all subjects, was considered to belong to the oily-skin cluster (high sebumU, sebumT, pore, and hemoglobin values). The female cluster 4 was characterized by high sebumT as compared to all other clusters.

3.3 Identification of the Relationships between Skin Type and Life Style

Under the assumption that the clusters determined in our cluster analysis represented the skin types of the individual subjects, we performed ANOVA and a decision tree analysis, in order to determine the relationship between skin type and life style.

Table 6. ANOVA results for life style and use of cosmetics; *p*-values reported from F-tests

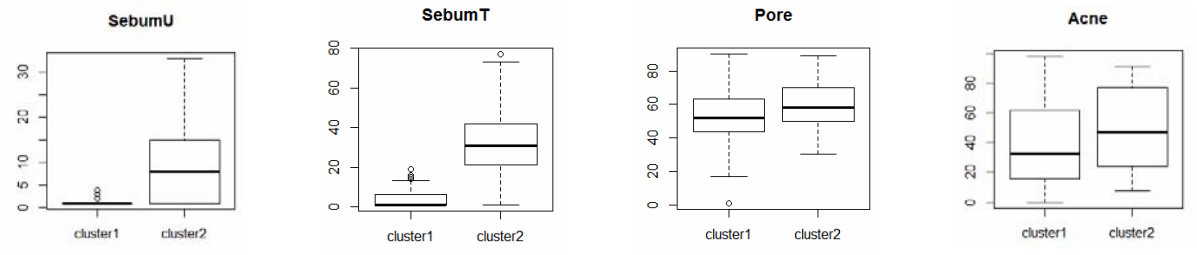
Category	Questionnaire item	Male	Female
	Age	**0.0329**	**0.0001**
	Life style		
Easting habit	Blue fish intake	0.6558	0.0944
	Daily vegetable consumption	0.0696	0.0167
	Regular meals	0.3808	0.3972
	Avoiding fast food	0.6987	0.0440
Skincare habit	Use of skin-type basic cosmetics	0.9902	0.3367
	Double facial cleansing	0.9760	**0.0006**
	Use of functional cosmetics	0.3010	0.0760
	Skin care by experts	0.0601	0.0387
	Skincare at home	**0.0312**	0.0668
Life style	Sufficient sleep and rest	0.8471	0.2728
	Sufficient water intake	0.6458	0.0161
Health status	Healthy, without any particular complaints	0.5973	0.0476
	Exposure to stress	0.1318	**< 0.0001**
	Menstruation or absence thereof	NA	**< 0.0001**
	Regular menstruation	NA	**0.0002**
	Irregular menstruation	NA	**0.0031**
Tobacco alcohol	Smoking	0.1656	0.0155
	Drinking	0.2929	0.1304
	Use of cosmetics		
Basic cosmetics	Cleanser	0.7384	0.5015
	Foam cleanser	0.5441	**< 0.0001**
	Deep cleanser	0.9020	**< 0.0001**
	Skin lotion	0.3122	0.3425
	Lotion	0.4778	0.0497
	Essence	0.6206	0.5788
	Facial cream	0.8757	0.8439
	Eye cream	0.3035	0.4888
	Pack/Mask	0.5427	0.0723
	Sunscreen	0.4379	0.5343
Cosmetics for hair and scalp care	Shampoo	0.7351	0.8063
	Rinse	0.5890	0.3800
	Treatment product	**0.0096**	**0.0011**
	Hair tonic/Hair growth products	0.4979	0.1761
	Dandruff remover	0.2126	0.1747

Similar to the ANOVA conducted in the previous subsection, the clusters were defined as factors and the elements of the questionnaire regarding life style and the utilization of cosmetics were defined as response variables. These values were then tested (F-test) for inter-cluster differences in life style and cosmetics use. Of the obtained values, only the *p*-values have been outlined in

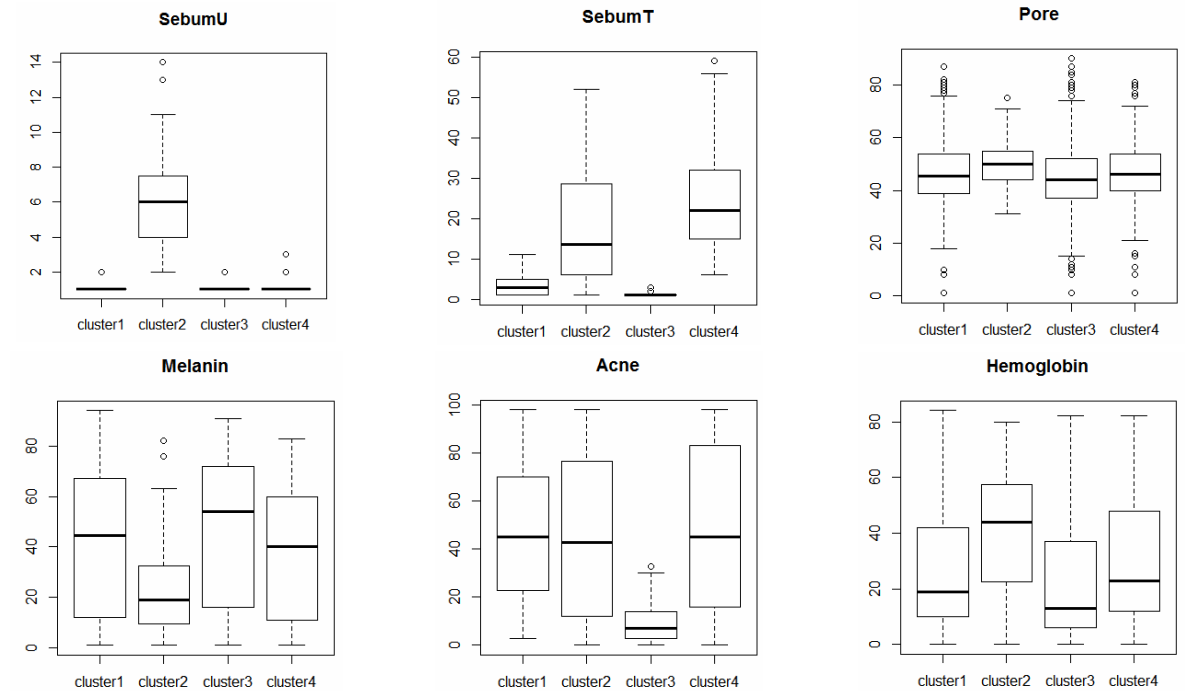
Table 6.

In the above table, the highlighted values for the male subjects represented the life style and usage of cosmetics displaying significant inter-cluster differences at $\alpha = 0.05$. In case of female subjects, the significance level α was set to 0.01. The reason for assigning a higher value of significant level for male subjects, was that no variables with significant differences could be derived at a significance level of $\alpha = 0.01$. As seen in the above table, significant differences were observed in 'age', 'skincare at home', and 'treatment product' between two male clusters. In female subjects, similar to male subjects, 'age' showed significant difference among four female clusters. We observed that additional 8 factors, including 'double facial cleansing', 'exposure to stress', 'menstruation or absence thereof', 'regular menstruation', 'smoking', 'foam cleanser', 'deep cleanser', and 'treatment product', also showed significant differences among the 4 clusters. As in the ANOVA described above, the box plots and bar charts in Figure 3 were drawn, highlighting the significant differences between the characteristics of each cluster with respect to lifestyle and the use of cosmetics. In Figure 3, the bar charts consisting of 1 and 0 are the cases of a binary variable (1 for yes and 0 for no), whereas the bar charts consisting of more than two values are the cases measured in Likert scale for frequency (the larger value, the more frequent).

An analysis of the plots revealed the following characteristics for each cluster. The subjects of male cluster 1 used treatment products at a greater frequency than those of cluster 2, while the subjects of cluster 2 comprised of older subjects who used regular home remedies for skincare, as compared to those of cluster 1. The female cluster 1 comprised of subjects who used treatments at a higher frequency as compared to the other clusters. The female cluster 2 comprised of younger subjects, at the peak of their menstrual cycles, did not smoke, and had no particular diseases. In addition, these subjects used foam cleanser and deep cleanser, double facial cleansing, and specialized skincare strategies more



(a) male

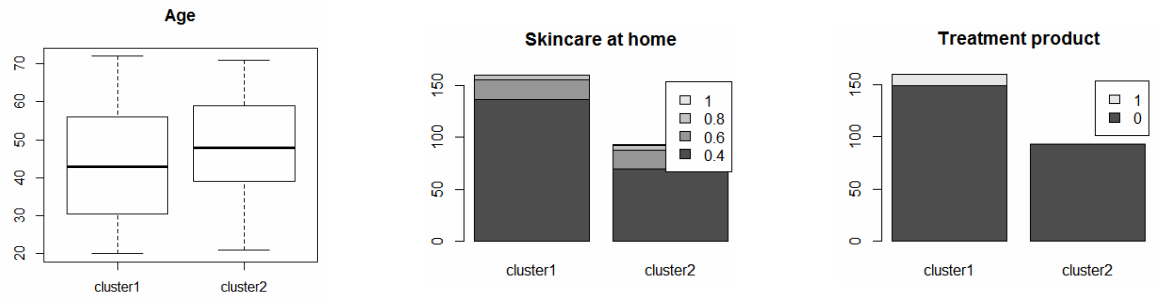


(b) female

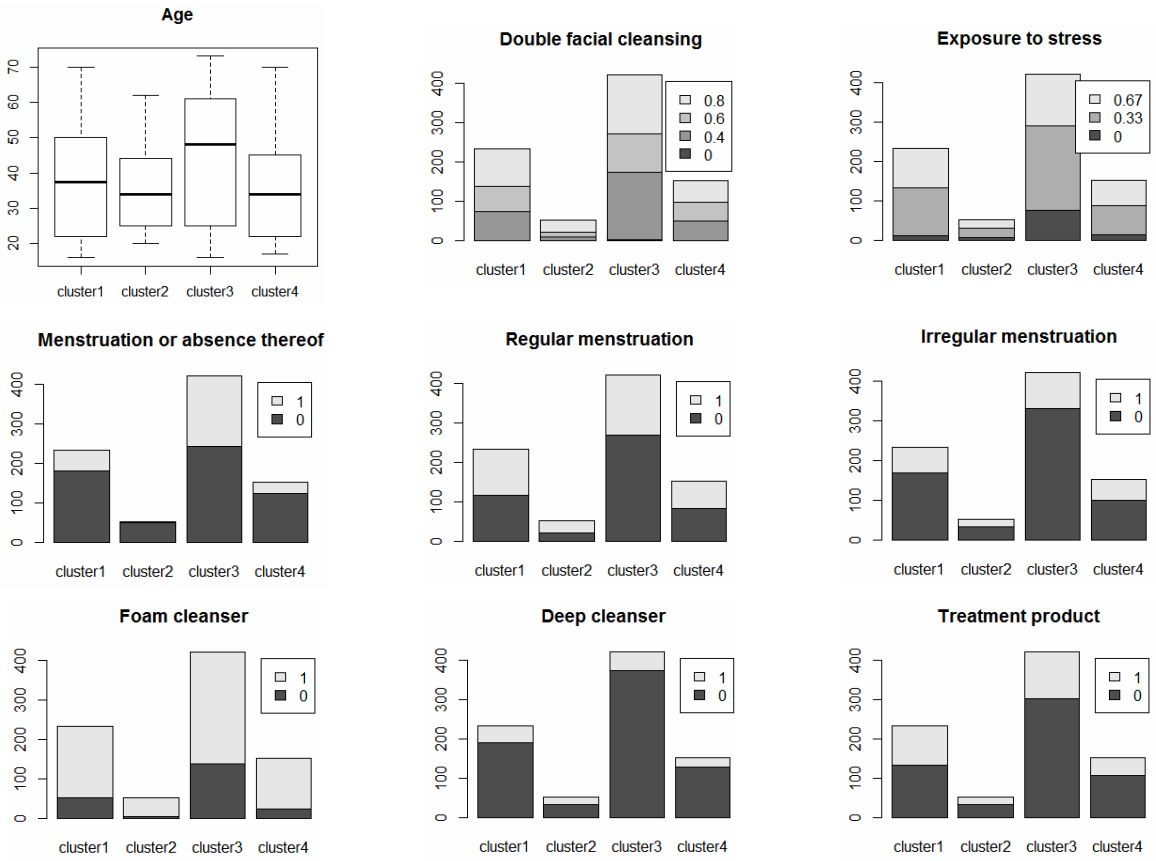
Figure 2. Box plots of significant variables.

frequently, i.e., the cluster comprised of women who paid particular attention to skincare. The female cluster 3 comprised of older subjects, who demonstrated a higher consumption rate of vegetables as compared to fast food, and showed lower stress levels, with a high likelihood of menopause due to age, and demonstrated the tendency to use casual lotions as compared to foam cleansers. The female cluster 4 comprised of women who showed a high rate of smoking and insufficient water intake.

In addition to determining the significant mean difference by the ANOVA F-test, we also employed a decision tree method, which is one of the most commonly used classification methods in data mining, to determine lifestyles and patterns of cosmetic use for each cluster, i.e., skin type. As mentioned above, we used the C4.5 algorithm, and the final model was obtained by using the reduced error pruning method. As a result, the final trees were found to have a training error of approximately 26% in both male and female datasets. This means that approximately 74% of subjects in the training datasets are correctly classified by the trees. Figures 4 and 5 show the decision trees created with data obtained from male and female subjects, respectively. Note that, in this tree building, the target variable contained the cluster labels that were created by GMM based on skin-specific values, and the independent variables indicate the life style and cosmetic usage of the subjects.



(a) male



(b) female

Figure 3. Box plots and bar charts of significant variables.

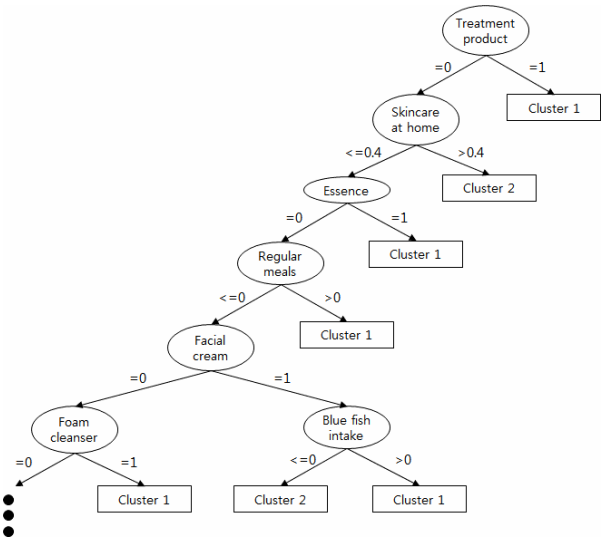


Figure 4. Decision tree obtained from male data.

As observed in the two figures, the female tree intended to classify four clusters has more complicated tree structure than the male tree classifying two clusters. Therefore, more various independent variables influencing the classification of each cluster were selected in the female tree, which is consistent with the obtained results of the ANOVA. As seen in Figure 4, the use of treatment product was revealed to occupy the top root node of the male tree, as the factor with the largest influence on the classification of the two clusters. The leaf nodes in the male tree show that the subjects often using skincare services generally belong to cluster 2. The female tree in Figure 5 indicates that the menopausal status occupies the root node and is thus identified as the most influential variable. Well-balanced eating habits, consistent vegetable intake, and the use of functional cosmetics also express a relatively high frequency of appearance in the tree. In other words, these are the factors that determine the formation of the female clusters, i.e. skin types.

Table 7. If-then rules for individual clusters

Tree model	If	Then
Male	- Treatment	Cluster 1
	- No treatment, no regular skincare, use of essence - No treatment, no regular skincare, no use of essence, well-balanced diets	
	- No treatment, regular skincare	Cluster 2
Female	- Menstruation, moderate exposure to stress, use of lotions, regular meals	Cluster 1
	- Menstruation, high exposure to stress, relatively high tobacco consumption, usage of creams, specialist skincare	
	- Menstruation, moderate exposure to stress, no use of lotion	Cluster 2
	- Menopausal status	Cluster 3
	- Menstruation, exposure to stress, relatively high tobacco consumption, usage of creams, but no specialist skincare	
	- Menstruation, exposure to stress, relatively high tobacco consumption, no usage of creams	Cluster 4

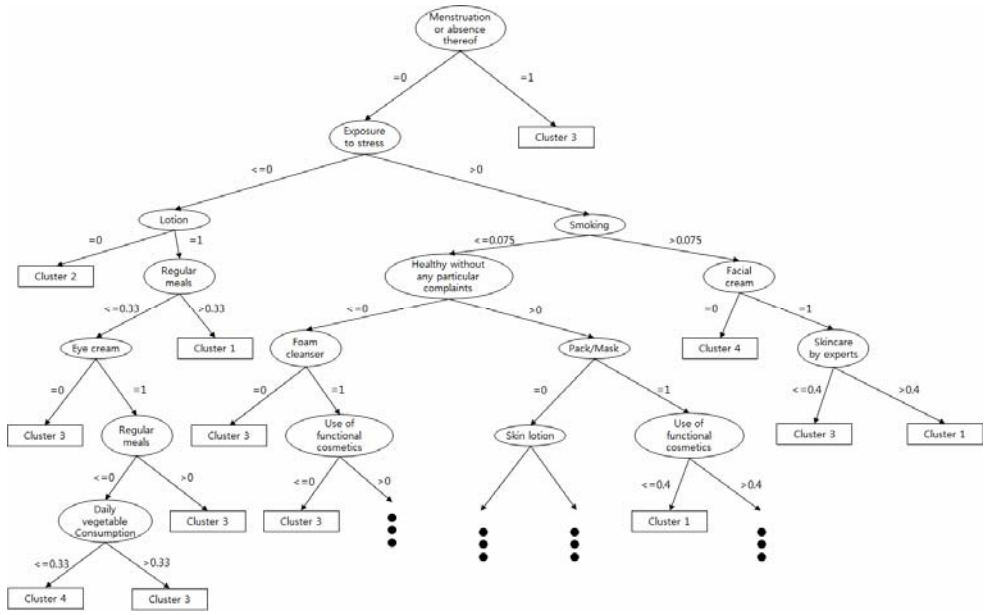


Figure 5. Decision tree obtained from female data.

Table 7 shows the interpretations of the male and female trees, arranged in accordance with the 'if-then' rules.

4. DISCUSSION

In this section, we will provide an overview of the effects of lifestyle including cosmetic usage on the skin condition based on the results of analysis presented in the previous sections. The subjects were classified into clusters by cluster analysis, associating the general skin conditions quantitatively measured by a device. In addition, the life styles of the different clusters were analyzed by using the ANOVA and the decision tree analysis technique. Table 8 summarizes the cluster-specific characteristics, integrating the results of all the analyses.

The subjects belonging to the male cluster 1 were relatively young with less visible prevalence of acne, which is associated with low interest in skincare. On the other hand, the subjects belonging to the male cluster 2 were exposed to workplace stress, which seems associated with a higher frequency of skincare at home. Compared to female subjects, however, the interest in skincare does not involve specialist skincare, but casual skincare at home for the male cluster 2. The subjects comprising the female cluster 1 were relatively young, and displayed regular eating habits. The high sebum and hemoglobin levels seen in the subjects belonging to the female cluster 2 indicate their interest in regular eating and skincare. The subjects in the female cluster 3 mainly comprised of retired housewives (middle-aged or older), demonstrating high interest in skin and health care. Lastly, the subjects in the female cluster 4 showed a higher rate of smoking and a low water intake, demonstrating a low interest in health care, which results in poor skin conditions.

Table 8. Overview of the cluster characteristics

	Male
Cluster 1	Relatively young Low sebum level and acne prevalence Health/skincare does not require much attention.
Cluster 2	Relatively old and employee status High sebum level and acne prevalence Health and skincare requires some attention.
	Female
Cluster 1	Relatively young Irregular eating habits High melanin concentration
Cluster 2	Relatively young and healthy High sebum, pore, and hemoglobin levels Regular eating habits must be developed. Attention must be paid to skin health/care.
Cluster 3	Elevated interest in health for reasons of age High melanin levels and low acne prevalence Great interest in skincare
Cluster 4	Smoking tendency and insufficient water intake High sebum level

5. CONCLUSIONS

Based on the data collected from both a questionnaire survey and skin measurement of 1,246 subjects, we conducted analyses by using the multivariate statistical analysis and data mining techniques, in order to determine the relationship between skin type and life style. Because of the gender-dependent significant differences in skin type as determined by logistic regression, we conducted individual analyses for male and female subjects. Based on the data measuring the skin conditions, we defined two clusters for male subjects and four clusters for female subjects, by using the GMM clustering technique. We then subjected the variables to ANOVA and C4.5 technique (a decision tree analysis technique), in order to determine the relationships between lifestyle of the subjects and the characteristics of each cluster, i.e., the skin type, as well as the regular patterns associated with the same. All the results of quantitative analysis were integrated and arranged according to the possible similarities between the skin-type characteristics of the subjects grouped together in each cluster and their association with the life styles of the subjects.

Most existing studies have been conducted by using a univariate approach, where the skin types were classified by independent examination of the individual skin-specific values, and the setting of threshold values. Therefore, this study is significantly different in that the skin types were determined by adopting a multivariate approach, where the measured skin-specific values for all subjects, and their correlations were simultaneously taken into account. To achieve this, we used data mining techniques. Specifically, our study is distinctive in that the decision tree method was employed to derive the correlations between skin type and life style, as well as the rules of classification pertaining to each skin type, based on the demographic elements and the self-reported life style and skincare status. This study is also important for researching the specific skin characteristics of Koreans.

A limitation of this study is that we did not conduct a longitudinal study to determine the intra-subject changes of skin-specific values over time. Future research can be directed to the observation and analysis of the changes in skin type, by applying a study design where the life-style-related conditions of the subjects of the same skin type are artificially varied, in order to derive more accurate relationships between lifestyle and skin type. Another limitation of this study is that the changes in skin conditions with respect to the seasonal factors in Korea, i.e., in a temperate climate zone with four seasons, were not analyzed. We believe that the careful analysis of these aspects in future research could assist in a more

comprehensive and accurate determination of the skin characteristics.

ACKNOWLEDGMENTS

This work is supported by the National Strategic R&D Program for Industrial Technology (10043869, Development of service platform for Personalized Quasi-drug and Cosmetic to individual skin and hair), funded by the Ministry of Trade, Industry and Energy (MOTIE).

REFERENCES

Asakura, K., Nishiwaki, Y., Milojevic, A., Michikawa, T., Kikuchi, Y., Nakano, M., and Takebayashi, T. (2008), Lifestyle factors and visible skin aging in a population of Japanese elders, *Journal of Epidemiology/Japan Epidemiological Association*, **19**(5), 251-259.

Baumann, L. (2008), Understanding and greating various skin types: the Baumann Skin Type Indicator, *Dermatologic Clinics*, **26**(3), 359-373.

Fitzpatrick, T. B. (1989), The validity and practicality of sun-reactive skin types I through VI, *Archives of Dermatology*, **124**(6), 869-871.

Fur, I. L., Lopez, S., Morizot, F., Guinot, C., and Tschachler, E. (1999), Comparison of cheek and forehead regions by bioengineering methods in women with different self-reported cosmetic skin types, *Skin Research and Technology*, **5**(3), 182-188.

Galzote, C., Estanislao, R., Suero, M. O., Khaiat, A., Mangubat, M. I., Moideen, R., and Wang, X. (2013), Characterization of facial skin of various Asian populations through visual and non-invasive instrumental evaluations: Influence of age and skincare habits, *Skin Research and Technology*, **19**(4), 454-465.

Kim, J. G., Park, B. S., and Kim, J. S. (2011), Relationships of food habits and life style and skin health of young females, *Korean Association of Human Ecology*, **20**(2), 449-465.

Kumagai, H., Shioya, K., Kawasaki, K., Horii, I., Koyara, J., Nakayama, Y., Mori, W., and Ohta, S. (1985), Development of a scientific method for classification of facial skin types, *Journal of Society of Cosmetic Chemists of Japan*, **19**(1), 9-19.

Lee, W., Lee, J., Lee, H., Jun, C.-H., Park, I.-S., and Kang, S.-H. (2014), Prediction of hypertension complications risk using classification techniques, *Industrial Engineering and Management Systems*, **13**(4), 449-453.

Luebberding, S., Krueger, N., and Kerscher, M. (2014), Mechanical properties of human skin in vivo: a comparative evaluation in 300 men and women, *Skin Research and Technology*, **20**(2), 127-135.

Man, M. Q., Xin, S. J., Song, S. P., Cho, S. Y., Zhang, X. J., Tu, C. X., Feingold, K. R., and Elias, P. M. (2009), Variation of skin surface pH, sebum content and stratum corneum hydration with age and gender in a large Chinese population, *Skin Pharmacology and Physiology*, **22**(4), 190-199.

McLachlan, G. and Peel, D. (2004), *Finite mixture models*, John Wiley and Sons.

Nouveau-Richard, S., Yang, Z., Mac-Mary, S., Li, L., Bastien, P., Tardy, I., Bouilon, C., Humbert, P., and De Lacharrière, O. (2005), Skin ageing: A comparison between Chinese and European populations: A pilot study, *Journal of Dermatological Science*, **40**(3), 187-193.

Park, S. G., Kim, Y. D., Kim, J. J., and Kang, S. H. (1999), Two possible classifications of facial skin type by two parameters in Korean women: Sebum excretion rate (SER) and skin surface relief (SSR), *Skin Research and Technology*, **5**(3), 189-194.

Pierard, G. E. (1987), What Does 'Dry Skin' Mean?, *International Journal of Dermatology*, **26**(3), 167-168.

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

Roberts, S. J. (1999), Novelty detection using extreme value statistics, *IEE Proceedings-Vision, Image and Signal Processing*, **146**(3), 124-129.

Rurangirwa, A., Pierard-Franchimont, C., Le, T., Ghazi, A., and Pierard, G. E. (1987), Corroborative evidence that dry skin is a misnomer, *Bioengineering and the Skin*, **3**(1), 35-42.

Sharma, S. (1995), *Applied multivariate techniques*, John Wiley and Sons.

Tang, T.-I., Zheng, G., Huang, Y., Shu, G., and Wang, P. (2005), A comparative study of medical data classification methods based on decision tree and system reconstruction analysis, *Industrial Engineering and Management Systems*, **4**(1), 102-108.

Tsukahara, K., Fujimura, T., Yoshida, Y., Kitahara, T., Hotta, M., Moriwaki, S., Witt, P. S., Simion, F. A., and Takema, Y. (2003), Comparison of age-related changes in wrinkling and sagging of the skin in Caucasian females and in Japanese females, *Journal of Cosmetic Science*, **55**(4), 351-371.

Tsukahara, K., Sugata, K., Osanai, O., Ohuchi, A., Miyauchi, Y., Takizawa, M., and Kitahara, T. (2007), Comparison of age-related changes in facial wrinkles and sagging in the skin of Japanese, Chinese and Thai women, *Journal of Dermatological Science*, **47**(1), 19-28.

Wendling, P. A. and Dell'Acqua, G. (2003), Skin biophysical properties of a population living in Valais,

Switzerland, *Skin Research and Technology*, **9**(4), 331-338.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., and Steinberg, D. (2008), Top 10 algorithms in data mining, *Knowledge and Information Systems*, **14**(1), 1-37.

Youn, S. W., Na, J. I., Choi, S. Y., Huh, C. H., and Park, K. C. (2005), Regional and seasonal variations in facial sebum secretions: a proposal for the definition of combination skin type, *Skin Research and Technology*, **11**(3), 189-195.

Zonios, G., Bykowski, J., and Kollias, N. (2001), Skin melanin, hemoglobin, and light scattering properties can be quantitatively assessed in vivo using diffuse reflectance spectroscopy, *Journal of Investigative Dermatology*, **117**(6), 1452-1457.