Journal of The Korea Society of Computer and Information
Vol. 21 No. 5, pp. 167-175, May 2016

www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2016.21.5.167

Pattern Recognition for Typification of Whiskies and Brandies in the Volatile Components using Gas Chromatographic Data

Sungmin Myoung*, Chang-Hwan Oh**

Abstract

The volatile component analysis of 82 commercialized liquors(44 samples of single malt whisky, 20 samples of blended whisky and 18 samples of brandy) was carried out by gas chromatography after liquid-liquid extraction with dichloromethane. Pattern recognition techniques such as principle component analysis(PCA), cluster analysis(CA), linear discriminant analysis(LDA) and partial least square discriminant analysis(PLSDA) were applied for the discrimination of different liquor categories.

Classification rules were validated by considering sensitivity and specificity of each class. Both techniques, LDA and PLSDA, gave 100% sensitivity and specificity for all of the categories. These results suggested that the common characteristics and identities as typification of whiskies and brandys was founded by using multivariate data analysis method.

▸ Keyword : Discrimination, gas chromatographic data, multivariate data analysis, pattern recognition

I. Introduction

식품 및 음료를 포함하는 다양한 제품의 규격준수 여부 확인을 위한 판별문제(discrimination issue)는 매우 중요한 이슈 들 중 하나이다 [1].

위스키, 브랜디, 리큐르 같은 증류주는 지역(국가)별로 서로 다른 규격을 가지고 있는데 유럽 상공회의소(European Community Council) 규정 1576/89 에서 법적으로 규정한 위스키의 규격은 식용 맥아곡물(malted cereal)을 당화시키고 효모를 사용하여 발효시킨 것으로 최소 3년간 700ℓ를 넘지 않는 나무통에 숙성시킨 것으로, 위스키의 도수는 부피비로 최소 40%로 규정하고 있다. [2-3].

계량분석화학(chemometrics)에서의 기본적 목표 중 하나는 식품과 음료를 특징짓기 위한 방법으로 화학적 표현자(chemical descriptor)를 이용한다 [4]. 위스키 식별을 위해 선택된 화학적 표현자들의 경우 대부분이 휘발성분이기 때문에, 이러한 휘발성 동족체(volatile congener)에 대한 계량분석화학적 판별 기법은 일반적으로 기체 크로마토그래피(gas chromatography; GC)를 이용한다. GC는 일반적으로 여러 증류주들(진, 브랜디, 코냑, 럼, 데낄라, 위스키 등)에서 분석목적으로 사용되어져 왔다 [4-5].

본 연구에서는 여러 증류주들 중 위스키 및 브랜디의 판별에 초점을 두었는데, 특히 싱글몰트(single malt) 위스키, 블랜디드(blended) 위스키 및 브랜디(brandy)를 대상으로 한다.

싱글몰트 위스키는 한 종류의 곡물(보리나 호밀)을 이용하여 단일 주조장에서 만들어진 위스키를 의미한다. 최소 3년 이상을 참나무통에 숙성시켜야 하며, 주로 아일랜드와 스코틀랜드에서 생산된다. 블랜디드 위스키의 경우 싱글몰트와 다른 곡물을 사용한 위스키를 서로 혼합(블랜딩)하여 제조한 주류를 의미한다. 브랜디는 독일어 Weinbrand와 함께 정의되어 있으며, 과일(포도/

• First Author: Sungmin Myoung, Corresponding Author: Chang-Hwan Oh
*Sungmin Myoung (smmyoung@jwu.ac.kr), Dept. of Health Adminisaration, Jungwon University
**Chang-Hwan Oh (och35@semyung.ac.kr), Dept. of Oriental Medical Food and Nutrition, Semyung University
• Received: 2016. 02. 19, Revised: 2016. 03. 19, Accepted: 2016. 05. 04.
• This research was supported by Ministry of Food and Drug Safety, South Korea in 2015.

사과 등)을 발효시켜서 증류한 술을 말하는데, 1000리터가 넘지 않는 오크통에서 최소 6개월 이상 숙성하도록 규정하고 있다.

본 연구의 목적은 싱글몰트, 블랜디드 및 브랜디를 특징짓는 휘발성분들을 탐색하고 분류/판별하는 것이다. 이를 위해서 GC를 이용하여 검출한 휘발성분 자료를 대상으로 위스키의 특성 및 유형분류를 위해 다변량 분석기법들 중 데이터 마이닝 기법을 이용하고자 한다.

다변량 분석은 시료사이의 차이점과 유사성(similarity)을 인식하는데 유용하게 이용되는 기법이다. 주로 이러한 크로마토그래피 자료를 분석할 때 가장 많이 이용하는 방법으로 주성분 분석(principle component analysis: PCA)을 들 수 있는데, 이 방법은 서로 연관되어져 있는 측정자료들을 보다 작은 수의 상관관계가 존재하지 않는 새로운 변량의 자료로 전환하는 방법으로서, 총괄적 지표로 해석하려는 통계학적 기법이다 [6-7]. 데이터마이닝은 자동화/반자동화 도구를 이용하여 대용량 자료로부터 의미 있는 규칙 또는 패턴을 발견하는 것을 목적으로 데이터를 탐색 및 분석하는 과정이라 정의한다 [8]. 계량분석화학 분야에서 주로 이용되는 데이터마이닝 기법은 판별분석(discriminant analysis), 신경망기법(neural network) 등이 주로 이용되었으며, 최근에서는 부분최소제곱 판별분석(Partial Least Square Discriminant Analysis, PLSDA)이 많이 제시되는 상황이다 [6].

위스키 및 브랜디 유형에 대한 분류 및 판별을 수행한 선행연구는 다음과 같다. Wilson 등은 위스키에서 퓨젤오일(fusel oil)과 알코올성분에 대한 판별분석 및 패턴분석을 수행하였으며 [9], González 등은 아이리쉬 위스키를 대상으로 고급알코올(higher alcohol) 성분에 대한 분류기법을 제안하였다 [1]. 또한, Aylott 등은 휘발성분 및 페놀릭성분을 각각 GC-MS 및 HPLC로 분석하여 5개 요인이 위스키 진품여부 판정에 중요한 요인이라고 제시한 바 있다 [5]. Ledauphin 등은 207개의 휘발성분을 대상으로 4개 유형의 브랜디(Armagnac, Cognac, Calvados, Mirabelle)에 대한 판별분석을 PLS-DA을 이용하여 수행하였다 [10]. 일반적으로 휘발 성분은 증류주의 제조 공정 중에 사용되는 방법 또는 원산지를 특징짓는데 널리 사용되고 있는데, Ledauphin 등의 연구를 제외하면 대부분 증류주에 대한 통계적 분류는 주로 전체 휘발성분이 아닌 고급알콜 성분 (퓨젤알콜)의 정량적 데이터를 대상으로 하고 있으며, 싱글몰트/블랜디드 위스키를 비롯한 브랜디까지 총 증류주를 대상으로는 아직까지 그 통계적 분류가 시도된 적은 없었다.

이에 본 연구에서는 3개 유형의 증류주(싱글몰트 위스키 44종, 블랜디드 위스키 20종, 브랜디 18종)의 휘발성분을 GC-MS를 이용하여 확인하고 GC 분석을 통해 얻어진 휘발성분의 정규화된 상대적 피크 면적(normalized peak area ratio)을 도출한 후, 이를 대상으로 다음과 같은 분석을 수행하고자 한다.

첫째, 3개 유형의 위스키에 대한 패턴분석을 주성분 분석을 이용하여 확인하며, 둘째, 3개 유형에 대한 여러 방법의 데이터마이닝 기법을 적용하여 판별/분류를 수행하고 이에 대한 타당도를 확인한다.

II. Material and Methods

1. Standards and Samples

본 연구에 사용한 3개 유형의 위스키 시료들(병)은 국내 주요 마트 혹은 백화점에서 직접 구매하였다. 총 82개의 위스키 표본이 분석에 이용되었으며, 싱글몰트 위스키 44개, 블랜디드 위스키 20개, 브랜디 18개로 구성되었다. 각 시료별 식별코드를 부여하였다: 싱글몰트 위스키는 S01~S44, 블랜디드 위스키는 B01~B20, 브랜디는 R01~R18로 명명하였다. 시료는 20ml씩 소분하여 냉동보관하였으며, 병 개봉후 1 개월 이내에 실험/분석하였다.

2. Analytical Procedure

모든 시료는 초순수급 물(HPLC grade water)로 알코올 함량을 20%로 조정한 후 액액분배(Liquid-liquid extraction; LLE) 하였으며, 시료 10㎖를 바이알(vial)에 넣고 NaCl 약 3g을 첨가하여 포화시킨 후 다이클로로메테인(Dichloromethane: DCM) 1 ㎖ 를 넣고, 와류 믹서(vortex mixer)를 이용해 혼합하였다. 냉장실에서 30분간 보관하여 분리된 DCM층을 분획한 후 PTFE 실린지 필터(syringe filter)로 여과하여 GC에 주입하였다. 휘발성분의 정성은 GC-MS(6890 GC, Agilent /time of flight mass spectrometer: TOF-MS, LECO)를 이용하였다. Transfer line 온도는 225℃, 이온원(ion source) 온도는 220℃로 설정하였다. 위스키 시료에 대한 가스 크로마토그램은 Fig 1.과 같다.

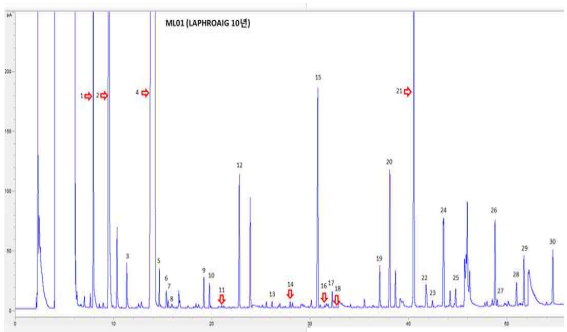

Fig. 1. A gas chromatogram of a single malt whisky (LAPHROAIG 10 years)

상대적 면적은 GC-FID(7890A, Agilent Technologies) data를 활용하였으며, 이동상은 질소 1 ㎖/min, 검출기 온도는 250℃로 설정하였다. GC-MS 및 GC-FID 모두 비분할 방식(splitless mode)으로 시료 1㎕를 주입하였다. 두 기기 모두 분석용 컬럼은 DB-WAX(30m × 0.25mmI.D., 0.25㎛ film thickness, Agilent Technologies)를 이용하였다. 검출된 각 성분은 질량 스펙트럼을 라이브러리를 검색하거나, n-paraffin 혼합물 (C_{10}~C_{26})을 기준으로 산출한 Retention index(RI)를 표준품 혹은 참고문헌 상의 수치 등과 비교하여 정성하였고, 각

RI 별 상대적 면적 값을 통계분석자료로 이용하였다.

3. Statistical Analysis

통계분석은 각각의 위스키 시료를 대상으로 GC-MS 및 GC-FID 분석으로 얻은 30개의 휘발성분에 대하여 상대적 피크면적을 도출하고 데이터 행렬(data matrix) 형태(82×30)로 기록하였다. 데이터 행렬에서 각 행은 위스키 및 브랜디 시료들을 의미하며, 열은 휘발성분의 고유지표(RI)에 대한 정규화된 상대적 피크 면적(normalized peak area ratio)이다.

각 고유지표별 성분명은 Table 1. 에 제시하였으며, 성분명이 'unknown' 인 것은 정성되지 않은 것을 의미한다.

본 연구에서는 통계적 다변량 분석기법 중 데이터 마이닝 기법을 적용할 것인데, 주요 기법은 비 지도학습 패턴인식방법으로 주성분분석 및 군집분석, 지도학습 패턴인식 방법으로 선형판별분석 및 부분최소제곱 판별분석을 제시할 것이다. 모든 통계적 분석은 R-package 3.2.1 및 R-studio 0.99를 이용하였고, 유의수준은 5%로 설정하였다.

Table 1. Volatile Composition of Whisky

peak	Retention Index	compound
p1	1051	1-Propanol
p2	1105	iso-Butanol
p3	1151	1-Butanol
p4	1226	iso-Amyl alcohol
p5	1236	Ethyl hexoate
p6	1254	3-Methyl-3-buten-1-ol
p7	1258	1,1-Diethoxyisobutane
p8	1268	Acetyl acetate
p9	1347	Ethyl 2-hydroxypropionate
p10	1360	n-Hexanol
p11	1390	3-Octenoic acid
p12	1435	Ethyl octanoate
p13	1517	Benzaldehyde
p14	1564	n-Octanol
p15	1639	Benzeneacetaldehyde
p16	1656	unknown
p17	1677	Diethyl butanedioate
p18	1688	Dodecanyl acetate
p19	1812	2-Phenylethyl acetate
p20	1843	Ethyl dodecanoate
p21	1914	Phenylethyl alcohol
p22	1952	unknown
p23	1972	Dodecan-1-ol
p24	2006	unknown
p25	2046	Diethyl hydroxybutanoate
p26	2176	unknown
p27	2184	(Z)-11-Hexadecenal
p28	2251	unknown
p29	2277	unknown
p30	2382	unknown

3.1 주성분 분석

주성분 분석(principle component analysis: PCA)는 "다변량 자료분석에서 모든 방법의 근원" 이라고 주장할 정도로 중요한 방법이다 [11]. 주성분 분석의 목적은 차원축약(dimension reduction)이며, 일반적으로 성분(component) 이라고 불리우는 선형잠재변수(linear latent variable)들을 계산하는데 빈번하게 적용되는 방법이다. 또한 새로운 좌표계를 계산하는 방법으로 알려져 있는데, 여기서 새로운 좌표계는 직교(orthogonal)하면서 가장 정보를 많이 가지고 있는 차원이 사용되는 잠재변수에 의해 형성되어 진다.

이러한 주성분분석으로 나타나는 잠재변수는 고차원 변수공간에서 개체들 간의 거리를 최적으로 나타낸다. 그리고 이 방법은 탐색적 자료분석(exploratory data analysis)이며 독립변수들(x-변수)들로만 적용될 수 있다 [11]. 또한 주성분 분석으로 인한 차원 축약은 주로 다음과 같은 분야에서 사용된다.

- 산점도에 의한 다변량 자료의 시각화
- 높은 상관관계를 가지는 x-변수들을 다른 분석에 이용하기 위해 상관이 존재하지 않는 잠재변수로 변환하려는 경우
- 오차(noise)로부터 (몇 개의 잠재변수들로 나타내어지는) 관련 정보의 분리
- 화학적 프로세스를 한 개 또는 몇 개의 '특징적' 변수들로 특성화 하는 여러 변수들의 조합

3.2 군집분석

군집분석에서 군집(cluster)은 '집중화된 집단(concentrated group)'의 의미를 가진다 [11-12]. 이는 일반적으로 변수공간(variable space)에서 개체들을 지칭할 뿐만 아니라, 반대의 경우, 즉, 개체 공간에서 변수들을 지칭할 수도 있으며, 둘 다 동시에 지칭할 수도 있다.

개체들의 관점에서 설명하면, 군집분석은 집단에 대한 어떠한 정보도 없는 상태에서, 개체들의 집중화된 집단(군집)을 식별하는 것이며, 심지어는 군집의 개수도 일반적으로 알려져 있지 않다 [13-14].

일반적으로 많이 이용하는 군집분석 알고리즘은 아래의 4가지로 알려져 있다.

- 분할적(partitional) 방법 ~ 각 개체는 반드시 한 개의 군집으로만 할당되게 하는 방법
- 계층적(hierarchical) 방법 ~ 한 군집안에 부분군집을 허용하는 형태. 즉, 전체 데이터를 한 군집으로 하고 이를 부분군집으로 분할한 후, 각각의 부분군집을 다시 또 분할하는 계층적 형태. 나무형태로 나타날 수 있음.
- 퍼지 군집(fuzzy clustering) 방법 ~ 각각의 개체가 각 군지에 속할 가중값 또는 가능성을 0과 1사이의 숫자로 표시해주는 형태
- 모형에 기초한(model-based) 방법 ~ 다른 군집들이 다변량 정규분포(multivariate normal distribution)와 같은 특정모형을 가정하는 경우

3.3 선형판별분석

판별분석은 집단 간의 차이를 식별하는 목표변수(target variable) 또는 반응변수(response variable)와 각 개체가 소속되어 있는 집단을 나타내는 다변량 자료(독립변수라고 정의함), 여기에서는 화학적 지표(chemical descriptor)를 그 대상으로 한다. 판별분석은 일반적으로 두 가지 과정으로 나눌 수 있는데, 첫 번째는 판별과정(discrimination process)으로서, 주어진 관찰값들로부터 전체집단을 특성에 따라 서로 다른 성격을 가지는 부분집단으로 분류하기 위하여 기준이 되는 판별함수를 추정 및 해석하는 과정을 의미한다. 두 번째는 분류과정(classification process)으로, 소속집단이 알려지지 않은 새로운 개체를 판별과정에서 유도된 기준을 활용하여 어느 부분집단으로 분류하는 과정이다 [15-16]. 본 연구에서는 집단을 분류하기 위한 목표변수(반응변수)는 위스키 유형(싱글몰트, 블랜디드, 브랜디)를 의미하며, 각 집단에 영향을 주는 독립변수는 30개의 휘발성분을 의미한다.

판별분석 과정에서 집단 간의 차이를 구별하는데 충분한 변수들을 선택하는 과정을 변수선택(variable selection)이라 하며, 변수선택의 기준은 집단 내 분산에 대한 집단 간 분산의 증감을 고려하는 F-통계량을 이용하는 것이 주로 고려된다. 이러한 기준으로 변수를 선택하는 방법으로 전진적 도입(forward inclusion), 후진적 제거(backward elimination), 단계적 방법(stepwise method)이 있다 [15].

3.4 부분최소제곱 판별분석

부분최소제곱 판별분석(partial least square discriminant analysis: PLSDA)은 계량분석화학(chemometrics) 분야에서 시작하여 최근에는 생물통계, 식품연구, 의학, 사회과학 등에서 적용되고 있는 방법이다 [16-17]. 부분최소제곱을 이용한 판별분석의 장점은 독립변수들 사이에 상관관계가 높아 다중공선성(multicolinearity) 이 존재하거나 데이터에 노이즈가 많이 존재하는 경우라도 기존 다중회귀분석 또는 선형판별분석과 비교하여 신뢰성이 높은 모형을 얻을 수 있다는 장점이 있다.

이론적으로는 능형회귀(ridge regression) 및 주성분회귀(principle component regression)와 관련이 있으며, 기본적인 개념은 다음과 같다 [18].

독립변수와 종속변수의 변화를 측정한 데이터를 대상으로 형성된 데이터 공간에 대하여 서로 직교하는 새로운 축을 정의하여 저차원의 특성 공간으로 투영한 결과로 특성벡터가 나타나는데, 이들 간에 최적의 상관관계를 구하는 것이다.

모형화를 하기 위해서는 두가지 단계를 거치는데, 첫 번째 단계는 목표변수를 연속으로 처리하고 PLS는 원래 예측변수의 선형결합인 구조적 잠재변수(latent variable)로 사용한다. 두 번째 단계는 첫 번째 단계에서 나타난 것을 가지고 분류분석방법을 이용하는데, 주로 일반화 선형 모형(generalized linear model)을 적용하며 로그우도함수는 Newton-Rhapson 알고리즘을 사용하여 최대화한다 [19].

PLSDA에서 변수선택방법은 여러 알고리즘이 있지만 본 연구에서는 VIP(variance importance on PLS projections)값을 기준으로 선정한다. 일반적으로 VIP가 1.0이상의 값을 갖는 변수값들을 모형에 포함시키는 것을 원칙으로 한다 [20].

III. Results

1. Pattern analysis in volatile components

각 위스키 유형별 30개의 휘발성분에 대한 피크비율 값들의 분포를 상자-수염 그림(Box and whisker plot)으로 도식화한 결과는 다음 Fig 2.와 같다.

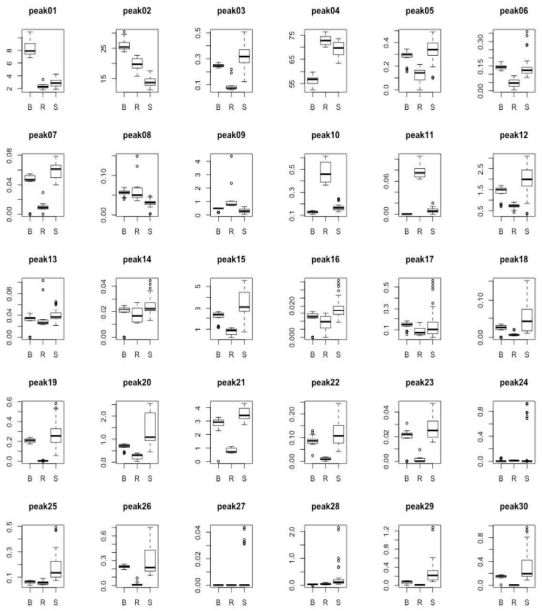

Fig. 2. Box and whisker plot for 30 volatile compound peak in extract of whiskies

Fig 2.에서 확인한 결과 몇몇 휘발성분 피크를 제외하고 거의 대부분의 피크들이 위스키별 유형(B: 블랜디드, R: 브랜디, S: 싱글몰트)별로 분포들이 다르게 나타남을 확인할 수 있었다. 그리고 몇몇 휘발성분 들에서 이상치(outlier)가 나타났는데, 이러한 경우는 자료의 오류라기 보다는 위스키 고유의 특징이라고 고려할 수 있다.

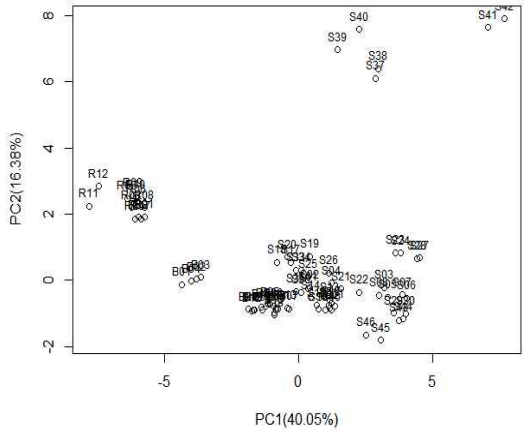

Fig. 3. PC scores plot of the studies samples for the first two PCs

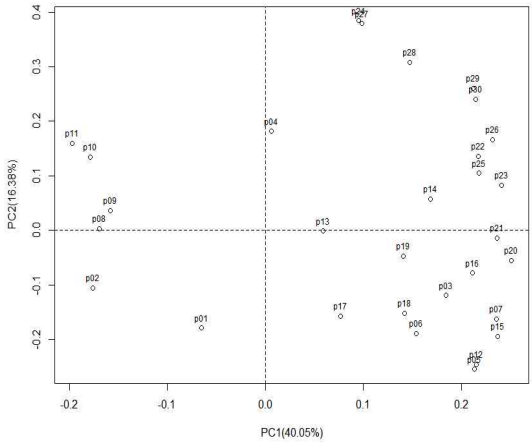

Fig. 4. PC loading plot of the studies samples for the first two PCs

자료의 구조와 휘발성분의 판별효율성을 시각화하기 위하여 주성분분석에 기초한 도식방법이 적용되었다. 30개의 휘발성분에 대한 2개의 주성분(PC1 and PC2)이 선택되어졌으며, Fig 3.과 같다. 여기서 PC1의 기여율은 0.4005 이고, PC2의 기여율은 0.1638로서 누적기여율이 0.5643으로 적절한 값으로 나타났다. 일반적으로 기여율이 높을수록 자료를 대표하는 종합적 정보로서의 가치가 크다고 판단할 수 있다.

Fig 3.을 통하여 확인할 수 있듯이, 3~4개의 위스키 군집형태로 나타났는데, 싱글몰트 위스키의 경우 PC1점수가 비슷하게 나타났으며, 특히 S37~S42는 PC1점수와 PC2점수가 큰 형태, 즉, 따로 떨어져 있는 덩어리(isolated clump)로 나타났다. 이는 S37~S42 주류가 일본산 싱글몰트 위스키로서 나머지 스코틀랜드산 싱글몰트 위스키와 다르기 때문으로 기인한다.

블랜디드 위스키 B01~B20의 경우 PC1점수가 낮은 쪽, PC2점수도 낮은쪽에 분포로 나타났으며, 싱글몰트 위스키에 가까운 쪽, 즉 PC2가 낮은쪽 으로 나타났다.

이는 블랜디드 위스키 자체가 싱글몰트와 다른 곡물을 사용한 위스키를 서로 혼합하여 블랜딩한 형태이기 때문에 싱글몰트 쪽과 비슷한 패턴을 가지는 것이라 판단된다. 브랜디의 경우에는 PC1점수는 싱글몰트 위스키에 비해 낮은 쪽에, PC2점수는 높은쪽에 위치하여 큰 편차가 나타남을 알 수 있었다.

2개의 주성분 PC1과 PC2와 관련된 변수들의 적재값(loading)을 Fig 4.에 제시하였다. 적재도표로 확인해 볼 때, 싱글몰트위스키와 블랜디드 위스키/브랜디를 구별하는데 가장 영향을 주는 휘발성분은 dodecan-1-ol(p23), diethyl hydroxy butanoate(p25), unknown(p22, p26)로 나타났다.

1-Propanol(p01), iso-butanol(p02)은 브랜디에, n-hexanol(p10), 3-octenoic acid(p11), ethyl 2-hydroxypropionate(p09), acetyl acetate(p08)은 블랜디드 위스키의 주성분 분석을 통해서 두 위스키를 분류하는데 기여하는 성분으로 고려되어질 수 있다.

이와 같이 주성분분석을 통해서 이러한 휘발성분들이 위스키의 유형을 분리 및 분류할 수 있는 지표들로 나타낼 수 있다는 것을 의미한다 [1]. 이를 평가하기 위하여, 각 위스키 시료들을 대상으로 계층적 군집분석(hierarchical cluster analysis)을 수행하였다. 유사성 척도(similarity measurement)로서 유클리디안 거리(euclidean distance)를 이용하였고, 군집연결방법으로 Ward 방법을 적용하였다.

군집분석결과를 나타내는 나무구조그림(dendrogram)은 Fig 5.와 같다.

최대 거리의 약 20% 내에서 3개의 군집이 형성되었으며, 이는 싱글몰트, 블랜디드 위스키, 브랜디의 3 유형의 위스키와 일치하였다. 이러한 군집분석을 통해 GC를 통한 휘발성분 peak를 이용하여 위스키의 유형을 분리하는 것은 적절하다고 판단할 수 있다.

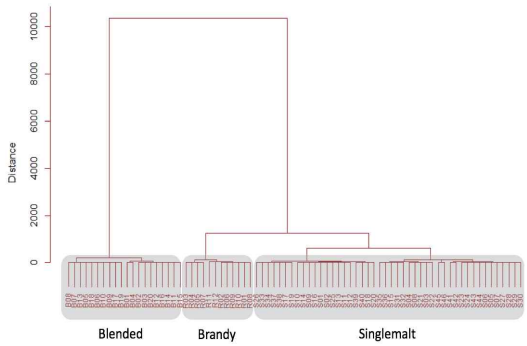

Fig. 5. Hierarchical Clustering of Whiskies using GC volatile compounds

2. Classification of whiskeys using LDA and PLSDA

위에서 기술하였듯이, 주성분분석은 자료를 축소요약하여 종합적 지표(주성분)을 추출하는 것이 목적이지만, 판별분석은 어떠한 시료가 어떠한 증류주 유형(싱글몰트 위스키, 블랜디드 위스키 및 브랜디)으로 분류되는 지를 다변량통계기법으로 판별하는 것이 주 목적이다 [21]. 특정한 휘발성분의 피크비율이 어떤 증류주 유형에 유사한지의 여부를 선형판별분석을 통하여 확인하였다. 판별함수는 Fisher의 방법을 이용하였으며, 변수선택방법으로 단계적(stepwise)방법을 고려하였다. 변수선택 기준은 집단 내 분산과 집단 간 분산의 F-통계량을 이용하였다.

30개의 휘발성분 피크비율 중 단계적 변수선택으로 선택된 피크는 13개의 성분이 선택되었다. 선택된 13개의 성분을 대상으로 판별분석을 수행한 결과 2개의 판별함수에 대한 계수가 추정되었다.

판별점수1=-117.82 -0.82×P01 + 1.47×P02 + 1.38×P04 - 20.22×P05 + 25.54×P07 + 2.36×P09 + 135.13×P11 + 134.11×P16 - 6.03×P19 + 6.24×P20 - 121.40×P23 + 3.21×P28 - 4.94×P30

판별점수2=-28.47 + 1.23×P01 + 0.72×P02 + 0.07×P04 + 6.96×P05 - 49.18×P07 + 1.46×P09 + 95.04×P11 + 155.14×P16 + 1.44×P19 - 0.93×P20 - 146.70×P23 + 2.54×P28 - 5.21×P30

판별함수1에 대한 고유값(eigenvalue)을 확인한 결과 총 판별력의 76.8%로서, 집단간의 분리에 대해서 비교적 2개의 판별함수가 잘 나타낸다고 할 수 있었다.

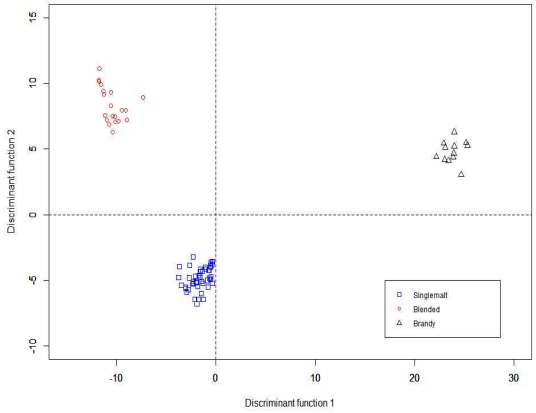

Fig. 6. Discriminant scatter plot of whisky samples

위스키 시료들에 대하여 위에서 추정한 2개의 판별함수에 휘발성분 피크값을 넣었을때의 판별점수에 관한 도표를 Fig.6에 제시하였다. Fig 6.에서와 같이 각 판별점수가 위스키 유형별로 명확하게 분리됨을 확인할 수 있었다.

PLSDA에 대한 모델링 전에, 30개의 휘발성분 피크비율에 대한 변수선택을 위하여 VIP(Variable importance on PLS projections)값을 기준으로 고려하였다. 각 휘발성분 변수에 대한 VIP 값은 1이상의 값을 갖는 경우 해당 변수를 PLSDA 독립변수에 포함시킨다. 이를 기준으로 선택된 독립변수는 7개로 선정되었으며, 선택된 휘발성분은 Table 2와 같다.

Table 2. Selected volatile component of the classifiers

LDA	PLSDA
P01(1-Propanol)*	
P02(iso-Butanol)*	
P04(iso-Amyl alcohol)*	
P05(Ethyl hexoate)	
P07(1,1-Diethoxyisobutane)*	
P09(Ethyl 2-hydroxypropionate)	
	P10(n-Hexanol)
P11(3-Octenoic acid)*	
P16(unknown)	
P19(2-Phenylethyl acetate)	
P20(Ethyl dodecanoate)	
	P21(Phenylethyl alcohol)
P23(Dodecan-1-ol)	
P28(unknown)	
P30(unknown)	

* common selected volatile component

선형판별분석에서 단계적변수선택법을 통하여 선택된 휘발성분은 13개였으며, PLSDA에서 VIP를 이용해 선택된 휘발성분은 7개였다.

그 중 두 분류기법에서 공통적으로 선택된 휘발성분은 5개로서 1-propanol(P01), iso-butanol(P02), iso-amyl alcohol (P04), 1,1-diethoxyisobutane(P07), n-hexanol(P10)이었다.

7개의 휘발성분을 대상으로 PLSDA를 적용한 결과 전체 독립변수들의 변동의 43.14%가 PLSDA의 첫 번째 성분으로 설명되어지며, 두 번째 성분은 47.82%로서 총 90.96%를 설명한다. Fig. 7은 위스키 유형별로 PLSDA 점수를 도식화한 결과로서, 3개의 유형이 명확하게 정의됨을 확인할 수 있었다.

위에서 제시한 분류모형의 성능을 판단하기 위하여 민감도(sensitivity)와 특이도(specificity)를 Table 3. 에 제시하였다.

민감도는 전체 데이터 N 개 중 분류모형에 의해 올바르게 분류된 비율을 의미하며, 정분류율(true positive rate)이라고도 한다. 1-특이도는 오분류율(false positive rate)이라 한다.

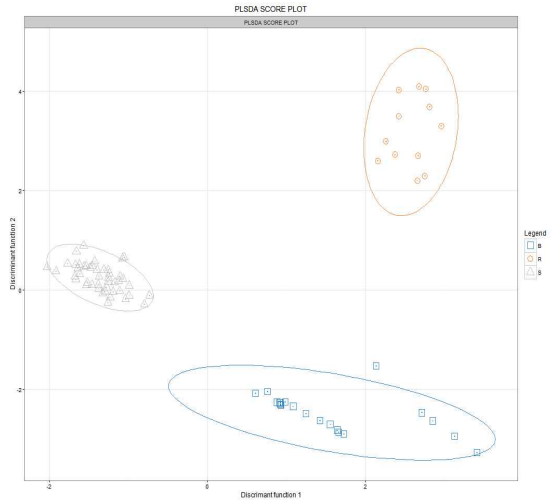

Fig. 7. PLSDA score plot of the volatile composition of whiskies

Table 3.을 확인한 결과 LDA, PLSDA 모두 이상적인 결과(민감도=1.0, 특이도=1.0)로 나타났다. 두 개의 분류모형 중 적절한 모형을 판단하기 위해서는 ROC 곡선을 그려 AUC등으로 더욱 적절한 모형을 판단할 수도 있겠지만, 두 모형 모두 동일한 민감도, 특이도를 가지기 때문에 모수절약의 원칙(principle of parsimony)을 적용하여 13개의 휘발성분으로 분류한 모형보다는 7개의 모형으로 분류한 것이 적절하다고 고려하였다.

Table 3. Classification Performance of Different Pattern Recognition Techniques

Category	Discriminant Techniques			
	LDA		PLSDA	
	sens	spec	sens	spec
Single malt Whisky	1.0	1.0	1.0	1.0
Blended Whisky	1.0	1.0	1.0	1.0
Brandy	1.0	1.0	1.0	1.0

PLSDA에서 선택되어진 7개의 휘발성분들 중에서 위스키 유형을 가장 잘 분리하는 휘발성분들을 확인하기 위하여 이용하는 가장 직관적인 방법은 각 7개의 휘발성분들의 변수-변수 도표를 통하여 확인하는 것이다 [1].

변수-변수 도표를 통해서 확인한 결과 1,1-diethoxy-isobutane(P07)과 n-hexanol(P10)이 위스키 유형을 잘 분리할 것이라고 나타났다 (Fig. 8).

Fig 8.에서 3개의 위스키 유형들이 각각 동일한 패턴을 가짐을 알 수 있으며, 싱글몰트와 블랜디드 위스키는 브랜디에 비해서 상대적으로 비슷한 패턴을 가짐을 확인하였다. 이와 같은 방법으로 만약 미지의 술을 직관적인 도표로 고려하여 분류할 수 있다면, 신제품 개발과 품질평가 및 개선 등에 유익할 것이라 판단된다.

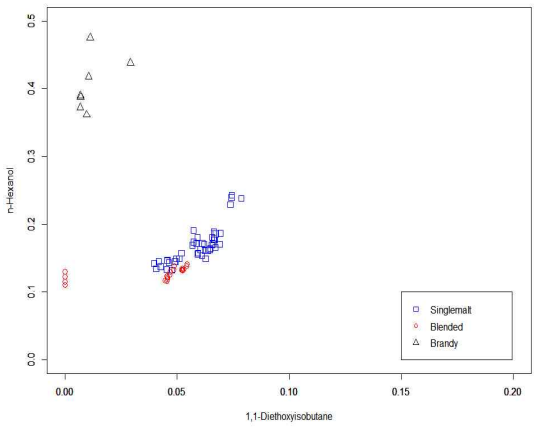

Fig. 8. Scatter plot of the studied whiskies using as coordinate variables the contents in 1,1-diethoxyisobutan e and n-hexanol

IV. Conclusions

본 연구는 3개 유형의 위스키(싱글몰트 44종, 블랜디드 20종, 브랜디 18종)에 대하여 GC로 휘발성분을 분석하고 다변량 통계기법을 적용하였다. 이를 통하여 위스키의 유형별 휘발성분의 주요 패턴을 확인하였고, 분류기법을 통하여 3종의 위스키에 대한 최적 변수선택 및 분류모형을 제시하였다.

GC를 통해서 도출된 30개의 휘발성분 피크면적 비율을 대상으로 주성분분석을 통하여 휘발성분의 패턴을 확인하였으며, 휘발성분이 위스키 유형을 분류할 수 있는 지표의 가능성 여부를 군집분석을 통하여 확인하였다. 또한 선형판별분석(LDA) 및 부분최소제곱 판별분석(PLSDA)의 대표적인 분류기법을 통하여 3개의 위스키 유형에 대해 분류하고, 위스키를 분류하는데 가장 기여하는 휘발성분을 변수선택기법으로 선정하였으며, 분류된 모형에 대한 성능평가를 민감도, 특이도로 확인한 결과 두 방법 모두 우수한 결과로 나타났다. 또한 선정된 휘발성분들의

변수-변수 도표를 통해 직관적으로 분류할 수 있는 가능성을 확인하였다. 이상의 연구결과는 추후 위스키의 유형 혹은 위조에 대하여 선정된 휘발성분을 이용하여 품질관리에 응용할 수 있는 가능성을 보여주었다.

그러나 생물을 원료로 생산되는 위스키의 특성 및 다수의 증류소로부터 생산되는 원액을 혼합하는 블랜디드 위스키의 특성상 보다 많은 다양성을 파악하기 위한 추가적인 연구가 필요할 것으로 판단되며, 이를 해결하고 유형을 판별하기 위해서는 보다 다양한 방법들이 동원되어야 하고 데이터에 대한 축적이 필요할 것이다. 또한 이를 이용하여 분류모형에 적용하였을 때 나타나는 분류의 정확도 확인 등이 필요할 것이라 판단된다.

본 연구에서는 분류기법으로 2가지 방법만 제시하였지만, 추후 인공신경망(artificial neural network), SVM(Support Vector Machine), k-최근접 이웃 알고리즘(k-nearest neighbor algorithm) 등의 여러 분류기법을 적용하고 이에 대한 모형평가를 통하여 최적의 분류모형을 제시하는 것이 위스키 유형에 대한 판별 정확성을 보다 향상시킬 수 있을 것이라 기대된다.

REFERENCES

[1] D. González-Arjona, G. López-Perez, V. González-Gall ero and A. Gustavo, "Supervised pattern recognition procedures for discrimination of whisky from gas chromatography/mass spectrometry congener analysis" J. Agric. Food, Chem., Vol. 47, No. 2, pp. 20-46, Jun. 2014.

[2] C. Oh, B. Kim, S. Ahn and K. Kim, "Development of physicochemical methods for the characterization of adulterated foods" Food Science and Industry, Vol. 47, No. 2, pp. 20-46, Jun. 2014.

[3] Council Regulation, "(EEC) No. 1576/89" Off. J. Eur. Commun, pp.1-, 1989.

[4] D. González-Arjona, V. González-Gallero, F. Pablos and A. Gustavo, "Authentication and differentiation of irish whiskeys by higher-alcohol congener analysis" Analytica Chimica Acta, Vol. 381, No. 2, pp. 257-264, Mar 1999.

[5] R. Aylott, A. Clyne, A. Fox and D. Walker, "Analytical strategies to confirm Scotch whisky authenticity" Analyst, Vol. 119, No. 8, pp. 1741—1746, Aug 1994.

[6] J. Choi, K. Bang, K. Han and B. Noh, "Discrimination analysis of the geographical origin of foods" Korean J. Food Sci. Technol., Vol. 44, No. 5, pp. 503—525, May 2012.

[7] B. Noh and D. Lee, "New product development by using principle component analysis" Food Sci. Indus., Vol. 29, pp. 2-12, Jan 2012.

[8] M. Berry and G. Linoff, "Data Mining techniques: For marketing, sales and customer relationship management" Wiley, New York, pp. 80-100, 2011.

[9] L. Wilson, J. Ding and A. Woods, "Gas chromatographic determination and pettern recognition analysis of methanol and fusel oil concentrations in whiskeys" J. Assoc. Off. Anal. Chem, Vol. 74, pp. 248-256, Mar 1991.

[10] J. Ledauphin, C. Le Milbeau, D. Barillier and D. Hennequin, "Differences in the volatile compositions of French labeled brandies(Armagnac, Calvados, Cognac, and Mirabelle) using GC-MS and PLS-DA" J. Agric. Food. Chem., Vol. 58, pp. 7782-7793, Sep 2010.

[11] K. Varmuza and P. Filzmoser, "Introduction to Multivariate Statistical Analysis in Chemometrics" CRC Press, pp. 74-80, 2009.

[12] Y. Cho, S. Moon and K. Ryu, "Clustering analysis by customer feature based on SOM for predicting purchase pattern in recommendation system" Journal of the Korean Society of Computer and Information, Vol. 19, No. 2, pp. 193-200, Feb 2014.

[13] D. Massart and K. Leonard, "The interpretation of analytical chemical data by the use of cluster analysis" Wiley, pp. 15-40, 1983.

[14] B. Ripley, "Pattern recognition and neural networks" Cambridge University Press, pp. 150-178, 2007.

[15] K. Kim and M. Jeon, "Multivariate Statistical Data Analysis" Free Academy, pp. 213-221, 1999.

[16] S. Wold, M. Sjostrom and L. Eriksson, "PLS-Regr ession: a Basic Tool of Chemometrics" Chemometrics Intell. Lab. Syst., Vol. 58, pp. 109-130, Apr. 2014.

[17] I. Han, M. Kim, C. Lee, W. Cha, B. Ham, J. Jeong, H. Lee, C. Chung and C. Han, "Application of partial least squares method to a terephthalic acid manufacturing process for product quality control" Korean J. Chem. Eng., Vol. 20, pp. 977-984, Sep. 2003.

[18] I. Han and H. Shin, "Modelling of a PEM fuel cell

stack using partial least squares and artificial neural networks" Korean Chem. Eng. Res., Vol. 53, pp. 236-242, Feb. 2015.

[19] S. Kim, "Classification with categorical data using partial least squares" Sungkyunkwan University, pp. 8-9, 2011.

[20] I. Chong and C. Jun, "Performance of some variable selection methods when multicolinearity is present" Chemometrics Intell. Lab. Syst., Vol. 78, pp. 103-112, Jan. 2005.

[21] D. Lee, H. Park, K. Kim, T. Lee and B. Noh, "Determination and Multivariate Analysis of Flavour Components in the Korean Folk Sojues Using GC-MS" Korean J. Food Sci. Technol., Vol. 26, pp. 750-758, Jun 1994.

Authors

Sungmin Myoung received the Ph.D. degrees in Biostatistics and Computing from Yonsei University, Korea, in 2006. Dr. Myoung joined the faculty of the department of Health Administration at Jungwon University, Chungbuk, Korea, in 2009. His interest includes medical informatics, biostatistical data analysis, and medical data mining.

Chang-Hwan Oh received the B.S., M.S. and Ph.D. degrees in Food Engineering from Yonsei University, Korea, in 1986, 1988 and 1993, respectively.

Dr. Oh joined Korea Food and Drug Administration as a second deputy, Seoul, Korea, in 1996. He is currently a Professor in the Department of Oriental Medical Food & Nutrition, Semyung University. He is interested in the analysis of the food and herbal drug materials for the confirmation of their quality and safety.