ISSN 1229-1889(Print)
ISSN 2287-9005(Online)
J Korean Soc Qual Manag Vol. 47, No.3: 537-552, September 2019
https://dx.doi.org/10.7469/JKSQM.2019.47.3.537

고차원 범주형 자료를 위한 비지도 연관성 기반 범주형 변수 선택 방법

이창기 · 정욱†

동국대학교 경영대학

Association-based Unsupervised Feature Selection for High-dimensional Categorical Data

Changki Lee · Uk Jung†

College of Business Administration, Dongguk University

ABSTRACT

Purpose: The development of information technology makes it easy to utilize high-dimensional categorical data. In this regard, the purpose of this study is to propose a novel method to select the proper categorical variables in high-dimensional categorical data.

Methods: The proposed feature selection method consists of three steps: (1) The first step defines the goodness-to-pick measure. In this paper, a categorical variable is relevant if it has relationships among other variables. According to the above definition of relevant variables, the goodness-to-pick measure calculates the normalized conditional entropy with other variables. (2) The second step finds the relevant feature subset from the original variables set. This step decides whether a variable is relevant or not. (3) The third step eliminates redundancy variables from the relevant feature subset.

Results: Our experimental results showed that the proposed feature selection method generally yielded better classification performance than without feature selection in high-dimensional categorical data, especially as the number of irrelevant categorical variables increase. Besides, as the number of irrelevant categorical variables that have imbalanced categorical values is increasing, the difference in accuracy between the proposed method and the existing methods being compared increases.

Conclusion: According to experimental results, we confirmed that the proposed method makes it possible to consistently produce high classification accuracy rates in high-dimensional categorical data. Therefore, the proposed method is promising to be used effectively in high-dimensional situation.

● Received 17 June 2019, 1st revised 30 June, accepted 1 July 2019
† Corresponding Author(ukjung@dongguk.edu)

※ 이 연구는 2018년도 한국연구재단의 국제협력사업의 지원(2017K2A9A2A06016127)과 2019학년도 동국대학교 논문게재장려금 지원으로 이루어졌음.

Key Words: Feature Selection, High-dimensional Categorical Data, Association-based Dissimilarity, Distance Metric, Unsupervised Learning

1. 서 론

최근 기업들은 데이터 획득 경로의 다양화, 데이터 습득 비용의 감소, 데이터 저장 기술의 발달로 인해 엄청난 양의 데이터를 보유하는 것이 가능하게 되었다. 이로 인해 자신들이 보유한 거대한 크기의 데이터를 품질경영에 도움이 되는 유용한 정보로 활용하기 위해 다양한 노력을 기울이고 있다(Cheong et al, 2017; Ree, 2017). 빅데이터 시대의 도래로 인해 연구자들은 연구 목적에 따라 정형화된 자료를 수집하는 환경에서 벗어나 끊임없이 생산 및 수집된 다양한 형태의 자료를 분석해야 하는 환경을 맞이하게 되었다. 이 과정에서 자료를 구성하는 변수의 수가 매우 많은 고차원의 자료가 흔하게 발생하며 이와 더불어 연속형 변수로 측정된 자료뿐만 아니라 범주형 변수로 측정된 자료가 수집되는 경우도 흔하게 되었다. 이런 고차원 범주형 자료는 연속형 변수로 측정된 자료에만 국한된 많은 분석기법을 그대로 적용되기 어렵고, 고차원의 자료에서 자주 발생하는 '차원의 저주'와 계산 비용의 증가를 유발하는 문제를 안고 있다.

빅데이터를 통해 의미 있는 정보를 찾는 과정에서 두 관측치(Observation) 사이의 거리(Distance)를 측정하는 방법은 매우 중요한 역할을 한다(Jia et al. 2016). 예를 들어 분류(Classification) 문제에서 널리 알려진 기법인 k-최근접 이웃 분류(k-nearest neighborhood classification)는 관측치의 부류(Class)를 분류하기 위해 한 관측치로부터 k번째 가까운 거리에 존재하는 관측치들의 부류(Class) 정보를 이용한다. 연속형 변수로 측정된 자료의 경우 관측치와 관측치 사이의 거리는 유클리디언 거리, 민코프스키 거리 등을 통하여 계산된다. 이에 반해 범주형 변수로 측정된 자료의 경우 이와 같은 거리 측정 방법을 적용할 수 없다. 따라서 범주형 변수로 측정된 자료를 one-hot encoding 방법을 통해 이진 벡터로 변환하여 두 관측치 간 거리를 계산하거나, 해밍 거리(Hamming 1950)를 이용하여 두 관측치 간 거리를 계산하기도 한다. 해밍 거리는 단순하게 범주형 변수 값(Categorical values)이 같으면 0, 다르면 1로 구분하여 불일치하는 변수 값들의 수를 두 관측치 간의 거리로 간주한다. 그러나 해밍 거리는 두 관측치 간의 거리를 지나치게 단순화시켜 계산하는 단점을 가지고 있어 정보의 손실을 감내하여야 한다. 이외에도 고차원의 범주형 자료를 one-hot encoding으로 변환하는 경우 변수의 수가 더욱 증가하여 관측치의 수보다 변수의 수가 지나치게 많아지게 될 가능성이 커져 이는 다시 '차원의 저주(Curse of dimensionality)' 문제를 초래하게 된다.

이러한 배경 하에서 해밍 거리의 단점을 극복하고자 범주형 변수로 이루어진 두 관측치 사이의 거리를 계산하는 다양한 방법이 많은 연구자에 의해 제시되었다(Goodall 1966; Smirnov 1968; Burnaby 1970; Lin 1998; Stanfill and Waltz, 1986; Cost and Salzberg 1993; Cheng et al., 2004; Le and Ho, 2005, Xie, 2010). Le and Ho(2005)는 두 범주형 변수 값(Categorical values) 사이의 비유사도(Dissimilarity)를 다른 변수 값들의 조건부 확률 분포 간 거리를 이용하여 계산하는 간접적인 방법을 제시하였으며 이 방법은 범주형 변수들 사이의 연관성이 높은 경우에 연관성을 고려하지 않는 방법들보다 우수한 성능을 보였다. 그러나 고차원 자료의 경우 적절한 변수(Relevant variables)와 불필요 변수(Redundancy variables)가 뒤섞여 있는 경우가 잦다. 이로 인해 모든 변수를 이용하여 조건부 확률 분포 간 거리를 계산하는 Le 와 Ho (2005)의 방법은 고차원 자료에서 여전히 '차원의 저주' 문제를 초래하고 계산 비용이 증가한다는 단점 또한 가지고 있다. 이 때문에 고차원 범주형 자료의 경우 거리 학습에 유의미한 변수를 적절하게 선택하는 변수 선택 (Feature selection) 방법이 필요하게 된다.

변수 선택은 고차원의 자료에서 적절한 변수만을 선택하는 기법을 의미하고 변수 선택을 통해 변수의 수를 감소시킴으로써 '차원의 저주'를 해결할 수 있으며 계산 비용 또한 줄일 수 있다. 변수 선택 기법은 부류 정보 이용 여부에 따라 지도 변수 선택 방법과 비지도 변수 선택 방법으로 나뉜다. 지도 변수 선택 방법은 부류 정보와 높은 연관성을 가진 변수들은 적절한 변수로 정의하여 변수를 선택하는 방법을 의미하며 많은 연구자에 의해 다양한 방법들이 제안되었다(Yu and Liu 2003, Liu et al., 2009; Vergara and Estevez 2013). 이에 비교해 비지도 변수 선택 방법은 상대적으로 적은 관심을 받아왔다. Mitra et al. (2002)는 변수간 유사성을 이용한 비지도 변수 선택 방법을 제안하였으며, Dy 와 Brodley (2004)는 Expectation-Maximization clustering 방법을 이용한 방법을 제안하였다. 그러나 이 두 연구 모두 수치형 자료에만 적용 가능한 방법으로 고차원 범주형 자료에 적용할 수 없다. 이에 본 연구에서는 부류 정보가 없는 고차원의 범주형 자료에 적용 가능한 변수 선택 방법을 제시하고자 한다.

본 연구의 구성은 다음과 같다. 제2장에서는 범주형 자료의 관측치 간 거리를 측정하는 방법과 변수선택 기법에 관한 선행 연구를 기술하였다. 제3장은 본 연구에서 제안하는 연관성 기반 범주형 변수 선택 방법을 서술했으며, 제4장에는 제안된 방법의 효과를 검증하기 위한 실험 설계 및 실험 결과에 관하여 기술하였다. 마지막으로 제5장에는 본 연구의 결론을 서술하였다.

2. 선행 연구

2.1 범주형 관측치 간 거리 측정 방법

범주형 변수로 측정된 두 관측치 간 거리를 계산하는 가장 간단한 방법은 해밍거리이다. 그러나 해밍 거리는 범주형 변수들 사이의 관계를 지나치게 단순하게 계산하는 단점을 가지고 있다. 이에 범주형 변수로 구성된 두 관측치 사이의 거리를 측정하기 위한 다양한 방법들이 제안되었다(Gooall, 1966; Smirnov, 1968; Burnaby, 1970; Stanfill and Waltz, 1986; Lin, 1988; Cost and Salzberg, 1993; Cheng et al., 2004; Le and Ho, 2005; Xie, 2010). 두 범주형 관측치 사이의 거리를 측정하는 방법은 크게 부류(Class) 정보를 이용하는 지도 학습 방법(Supervised learning method)과 부류 정보를 이용하지 않는 비지도 학습 방법(Unsupervised learning method)으로 나뉜다.

Stanfill과 Waltz (1986)는 부류 정보를 잘 구분할 수 있는 범주형 변수 값에 높은 가중치를 주는 Value Difference Metric(VDM)를 제안하였다. 그러나 VDM은 두 변수 값의 차이가 대칭적이지 않은 문제점이 있어 Cost와 Salzberg (1993)는 이를 수정 보완한 Modified Value Difference Model (MVDM)을 제안하였다. VDM과 MVDM 모두 조건부 확률을 이용하여 두 변수 값의 차이 (즉, 거리)를 정의하였다. 이에 반해 Xie (2010)는 경사하강법을 이용해 범주형 변수 값에 실수(real number)를 대응시키는 방법을 제안하였다. Xie (2010)가 제안한 방법은 임의의 실수 값을 범주형 변수 값에 대응시킨 다음 k-최근접 이웃 분류 수행하고 이를 통해 예측한 부류 값과 실제 부류 정보와의 오차를 계산한 뒤 경사하강법을 통해 대응된 실수 값을 조금씩 수정하여 오차를 최소화하는 실수 값을 찾는 방법이다. 이와 유사하게 Cheng et al (2004)은 경사하강법을 이용한 Adaptive dissimilarity matrix를 제안하였다. 그들은 임의의 실수 값을 수정하는 Xie (2010) 방법과는 다르게 경사하강법을 통해 두 변수 값 사이의 비유사도를 직접 수정하는 방법이다. 그러나 언급한 모든 방법은 부류 정보가 있어야 적용이 가능한 지도 학습 방법이다.

앞서 논의한 지도 학습 방법과 달리 부류 정보를 이용하지 않는 비지도 학습 방법 또한 다양하게 제안되었다

(Gooall, 1966; Smirnov, 1968; Burnaby, 1970; Lin, 1988; Le and Ho, 2005). Goodall(1966)은 두 변수 값 사이의 유사도(두 변수가 동일한 같은 경우 유사도가 가장 높음)를 계산 하는 방법으로 희소하게 발생하는(빈도가 낮은) 변수 값이 동일한 경우에 상대적으로 더 높은 유사도를 부여하는 방식을 제안했다. 그러나 Goodall(1966)이 제안한 방법은 두 변수 값이 불일치하는 경우를 0의 유사도를 가지도록 함으로 인해 상이한 두 변수 값의 차이 정도가 지나치게 단순화 된다는 단점이 있었다. 이를 극복하고자 Smirnov(1968)는 두 변수 값이 불일치하는 경우에 0이 아닌 값을 부여하는 방법을 제안하였다. 이 방식은 두 변수 값이 동일한 경우에는 그것이 희소한 변수 값일 경우에 상대적으로 더 높은 유사도를 부여하고 두 변수 값이 불일치하는 경우에는 불일치하는 두 변수 값을 제외한 나머지 변수 값들을 고려한 유사도를 부여한다. 확률적 방법에 기반한 Goodall(1966)과 Smirnov(1968)의 방법은 두 변수 값이 일치하는 경우에 더 중점을 둔 방법이다. 그러나 Burnaby(1970)과 Lin (1988)은 정보이론(Information theory)에 기반을 두어 두 변수 값이 불일치하는 경우에 보다 초점을 둔 방법을 제안하였다. Burnaby(1970)는 두 변수 값이 일치하는 경우에는 간단하게 유사도 값을 1로 계산하고, 희소한 두 변수 값이 불일치하는 경우는 비유사도 값을 높게 부여함으로써 상대적으로 유사도 값이 낮아지는 방법을 제안했다. Lin (1988)은 Burnaby(1970)의 방법과는 다르게 자주 발생하는 두 변수 값이 일치하는 경우 1보다 큰 유사도 값을 부여하였으며, 희소하게 발생하는 두 변수 값이 불일치하는 경우 더 큰 비유사도 값을 부여하는 방법을 제시하였다. 그러나 앞서 설명한 모든 비지도 학습 방식은 범주형 변수들 사이의 관계(즉, 연관성)를 고려하지 않는 단점을 가지고 있다. 이에 Le 와 Ho(2005)는 변수들 간의 연관성을 고려한 연관성 기반 비유사도(Association-based dissimilarity) 방법을 제안하였다. Le 와 Ho(2005)는 두 변수 값 사이의 비유사도를 다른 변수 값들의 조건부 확률 분포 간 차이로 계산하였으며 범주형 변수 사이에 연관성이 존재할 때 연관성을 고려하지 않는 방법들에 비해 더 우수한 성능이 나타남을 보였다.

다음은 본 연구에서 제안한 변수 선택 방법의 효과성을 살펴보기 위해 선택한 두 범주형 관측치 간의 거리를 계산하는 방법에 관해 서술하고자 한다. 본 연구에서 관측된 n개의 관측치로 이루어진 데이터 셋을 $\mathbf{x} = \{\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_n}\}$으로 표기하며, 각 관측치는 p차원의 범주형 확률 변수로 측정되었다. 임의의 한 범주형 확률 변수 A_i는 r_i개의 범주형 변수 값을 가지며 $A_i = \{a_{i1}, a_{i2}, \cdots, a_{ir_i}\}$로 표현된다. 범주형 확률 변수에서는 수치형 확률 변수와 달리 임의의 두 변수 값 사이의 차이가 정의되어 있지 않으므로 두 관측치 사이의 거리를 계산하기 전에 먼저 두 범주형 변수 값 사이의 차이인 비유사도를 정의하였다. 임의의 한 범주형 확률 변수(A_i)의 두 변수 값 a_{ij}와 a_{ik} 사이의 비유사도는 $d(a_{ij}, a_{ik})$로 표현했으며. 두 관측치 $\mathbf{x_g}$와 $\mathbf{x_h}$ 간 거리는 $d(\mathbf{x_g}, \mathbf{x_h})$로 표현한다.

2.1.1 해밍 거리 (Hamming distance)

해밍 거리는 범주형 변수로 측정한 두 관측치 사이의 거리를 계산하는 가장 대표적이고 단순한 방법이다. 해밍거리에서 임의의 한 범주형 확률 변수(A_i)의 두 변수 값 a_{ij}와 a_{ik} 사이의 비유사도는 아래와 같이 계산한다.

$$d(a_{ij}, a_{ik}) = \begin{cases} 0 & \text{if } a_{ij} = a_{ik} \\ 1 & \text{if } a_{ij} \neq a_{ik} \end{cases}, \qquad \text{수식 (1)}$$

여기서 $\forall i \in \{1,2,\cdots,p\}$이며, $\forall j,k \in \{1,2,\cdots,r_i\}$ 이다. 즉, 해밍거리에서 두 변수 값의 비유사도는 동일한 값을 가지면 0이고 다른 값을 가지면 1이다.

2.1.2 연관성 기반 비유사도 (Association-based dissimilarity)

Le and Ho(2005)는 연관성 기반의 비유사도를 계산하기 위해 임의의 확률 변수 A_i의 변수 값 a_{ij}가 주어졌을 때 다른 변수 $A_{i'}\,(i' \neq i)$의 변수 값 $a_{i's}$의 조건부 확률은 $p(a_{i's}|a_{ij})$로 표기하고, 임의의 확률 변수 A_i의 변수 값 a_{ij}가 주어졌을 때 다른 확률 변수 $A_{i'}$의 변수 값들의 조건부 확률 분포를 $\boldsymbol{P}(A_{i'}|A_i = a_{ij})$로 표기한다. 연관성 기반 비유사도에서 임의의 한 범주형 확률 변수가 가지는 두 변수 값 사이의 비유사도는 임의의 범주형 확률 변수의 변수 값이 주어졌을 때, 다른 범주형 확률 변수의 조건부 확률 분포 차이로 아래와 같이 계산한다.

$$d(a_{ij}, a_{ik}) = \sum_{i'=1}^{\hat{p}} \psi(\boldsymbol{P}(A_{i'}|A_i = a_{ij}), \boldsymbol{P}(A_{i'}|A_i = a_{ik}))\ \ (i' \neq i), \qquad \text{수식 (2)}$$

여기서 $\forall i, i' \in \{1,2,\cdots,\hat{p}\}$, $\forall j,k \in \{1,2,\cdots,r_i\}$이며, $\psi(\cdot,\cdot)$은 확률 분포 간의 거리를 계산하는 함수이다. 확률 분포 간 거리를 계산하는 함수는 다양한 학자들에 의해 제안되었다(Lin 1991; Kullback and Leibler 1951; Chakraborty 2008). 본 연구에서는 확률 분포의 거리를 구하는 함수로 Hellinger 거리를 사용 하였다(Chakraborty 2008). Hellinger 거리는 아래의 수식 (3)을 통해 구한다.

$$\psi(\boldsymbol{P}(A_{i'}|A_i = a_{ij}), \boldsymbol{P}(A_{i'}|A_i = a_{ik})) = \sqrt{\sum_{s=1}^{r_{i'}} (\sqrt{p(a_{i's}|a_{ij})} - \sqrt{p(a_{i's}|a_{ik})})^2} \qquad \text{수식 (3)}$$

여기서 $\forall i, i' \in \{1,2,\cdots,\hat{p}\}$, $\forall j,k \in \{1,2,\cdots,r_i\}$, $\forall s \in \{1,2,\cdots,r_{i'}\}$이며, $p(\cdot|\cdot)$은 조건부 확률이다. Hellinger 거리는 0과 1 사이의 값을 가지게 되며 두 분포가 같은 경우 0을 다른 경우 1을 가진다.

연관성 기반 비유사도의 계산 과정을 간단한 예를 통해 살펴보고자 한다. 아래의 Table 1과 같이 데이터가 총 8개의 관측치와 2개의 범주형 변수(Gender와 Color)로 구성되어 있다고 하자.

Table 2. A synthetic small dataset

Synthetic small dataset	
Gender	Color
Female	White
Female	Black
Male	Black
Female	Black
Male	Black
Male	White
Female	Black
Female	White

Table 1를 통해 아래의 Table 2와 같이 교차 빈도표와 교차 확률표를 구할 수 있다.

Table 3. The co-occurrence & conditional probability between Gender and Color

	Co-occurrence table			Conditional probability		
	white (w)	black (b)	sum	$p(w\mid\cdot)$	$p(b\mid\cdot)$	sum
Female (F)	2	3	5	2/5	3/5	1
Male (M)	1	2	3	1/3	2/3	1

수식 (2)과 (3)를 이용하여 (F,M)의 비유사도를 구하면 아래와 같다.

$$d(\mathrm{F,M}) = \frac{1}{\sqrt{2}}\sqrt{(2/5-1/3)^2+(3/5-2/3)^2} = 0.067,$$

다음으로 두 변수 값의 비유사도을 이용한 두 관측치 사이의 거리는 아래의 수식 (4)를 통해 계산된다. 즉, 수식 (1) 또는 수식 (2)를 통해 계산된 비유사도 값의 합이 두 관측치 사이의 거리이다.

$$d(\mathbf{x_g}, \mathbf{x_h}) = \sum_{i=1}^{\hat{p}} d(a_{ij}, a_{ik}), \qquad \text{수식 (4)}$$

여기서 $\forall\, g,h \in \{1,2,\cdots,n\}$이다.

2.2 변수 선택 (Feature selection)

변수 선택은 p 개의 변수를 가지는 변수 집합 F에서 중요한 변수 일부만을 선택하여 부분 집합 S를 구성하는 작업이다(Oh, 2008). 변수 선택의 핵심 조건은 원본 데이터 변수의 수를 감소시킴에도 불구하고 원본 데이터의 중요한 정보를 잃지 않는 것이다. 변수 선택을 통해 얻는 가장 큰 이점은 계산 시간의 감소와 함께 차원의 저주를 피함으로써 일반화 능력을 갖추는 것이다. 변수선택 방법은 부분 집합을 생성하고(부분 집합 생성기) 생성된 부분 집합을 평가하는(분별력 측정기) 두 가지 함수로 구성된다. 변수선택 방법은 계산 시간을 고려하지 않는다면 가능한 모든 부분 집합을 생성하고 가장 우수한 평가를 받은 부분 집합을 선택하는 단순한 문제이지만, 변수의 개수가 늘어남에 따라 계산 시간이 기하급수적으로 증가한다. 따라서 이를 개선하기 위한 많은 변수 선택 기법들이 제안되었다.

변수 선택 기법은 작동 원리에 따라 크게 임베디드 (Embedded), 래퍼(Wrapper), 필터(Filter) 방법으로 분류 된다(Guyon et al., 2003). 임베디드 방법은 부분 집합의 생성 및 선택 과정이 모형학습 과정에 포함이 되어있다. 대표적인 방법으로 C4.5(Quinlan, 2014)나 LASSO(Tibshirani, 1996)가 이에 해당한다. 래퍼 방법은 특정 모형의 성능 향상에 가장 이상적인 변수의 조합 찾는 방법이다. 회귀분석의 전진 선택법 (Forward selection), 후진 제거법 (Backward elimination), 단계 선택법 (step-wise selection)이 래퍼 방법에 해당한다. 래퍼 방법은 다른 방법에 비해 정확도가 높은 장점이 있으나 과적합 문제와 계산 비용이 많이 든다는 단점이 있다(Liu et al., 2009). 마지막으로 필터 방법은 모형학습과 독립적으로 변수 집합 F에서 적합도 평가지표 (goodness measure)를 이용하여 기준 조건을 충족하는 변수를 선택함으로써 부분 집합을 구성하는 방법이다. 평가지표로는 주로 상관 계수(Correlation coefficient)을 사용하거나 정보이론의 엔트로피(Entropy)를 이용한다. 필터 방법은 래퍼 방법과 비교하여 성능이 다소 낮으나 계산 시간이 작으며 과적합(Overfitting)을 피할 수 있다는 장점이 있다.

변수 선택 기법을 분류하는 또 다른 기준으로는 부류 정보(Class information)의 이용 여부이다. 부류 정보를 이

용하는 기법인 지도 변수 선택(Supervised feature selection)과 부류 정보를 이용하지 않는 기법인 비지도 변수 선택(Unsupervised feature selection)으로 나뉜다(Luis 2000; Part Punpiti 2014; Park and Kim 2014). C4.5, Lasso, 단계 선택법 등의 방법들은 모두 부류 정보를 이용하는 지도 변수 선택 방법이다. 지도 변수 선택은 비지도 선택 방법에 비해 상대적으로 다양한 방법들이 논의되었으나(Yu and Liu 2003, Liu et al., 2009; Vergara and Estevez 2013). 이에 반해 비지도 변수 선택 방법은 상대적으로 적은 관심을 받아 왔다(Luis 2000; Part Punpiti 2014). 비지도 변수선택 방법은 일반적으로 지도 선택 방법에 비해 과적합(Overfiting) 경향이 작은 장점을 가지고 있다(Guyon et al., 2003). 그러나 많은 수의 비지도 변수 선택 방법은 범주형 자료에 적용이 불가능 하다. 이에 본 연구에서는 부류 정보가 없는 고차원 범주형 자료에 적용 가능한 연관성 기반 변수 선택 방법을 제안하고자 한다.

3. 제안 방법: 연관성 기반 범주형 변수 선택 방법

본 연구의 목적은 부류 정보가 없는 고차원 범주형 데이터에서 적절한 변수를 선택하는 방법을 제안하는 것이다. 본 연구는 계산 시간이 긴 고차원 데이터를 다루고 있어서 변수 선택 원리는 계산 비용이 적은 필터 방법 선택하였다. 따라서 본 연구에서 제안하는 변수 선택 방법은 비지도 필터 방식이다. 필터 변수 선택 기법은 적합도 평가 지표(Goodness measure)를 기준으로 기준 조건을 충족하는 변수를 선택하여 적합(Relevance) 부분 집합을 구성한 뒤 적합 부분 집합에서 잉여 변수(Redundancy variables)를 제거하는 단계로 구성되어있다(Yu and Liu, 2003).

3.1 적합도 평가 지표 (The goodness measure) 선정: 정규화된 조건부 엔트로피 $D(A_{i'}, A_i)$

적합도 평가 지표를 정의하기 위해서는 적합 변수에 대한 사전 정의가 필요하다. 대부분의 지도 변수 선택 기법들은 부류 정보와 연관성이 높은 변수를 적합 변수(Relevance variables)로 정의한다(Blum and Langley, 1997). 그러나 본 연구에서는 부류 정보가 없는 상황이므로 이와 같은 방법으로 적합 변수를 정의할 수 없다. Luis (2009)는 군집 분석을 위한 변수 선택 기법을 제시하였는데 그의 연구에서는 변수들 간 연관성이 높은 변수를 적합 변수로, 변수들 간 연관성이 낮은 변수들은 불필요한 변수로 정의하였다. 본 연구에서는 Luis (2009)의 적합 변수 정의 방식에 따라 적합 변수와 부적합 변수를 정의한다.

변수와의 연관성을 정의하는 일반적인 방법은 피어슨 상관 계수(Pearson correlation coefficient)이다. 그러나 피어슨 상관 계수는 비선형적 관계를 반영하지 못하며, 범주형 변수에는 적용할 수 없다. 이런 단점을 극복하기 위해 정보이론의 엔트로피 개념을 기반으로 한 변수 간 연관성 측정 방법이 제시되었다(Yu and Liu, 2003). 따라서 본 연구에서는 변수 간 연관성을 엔트로피 개념을 사용하여 정의한다.

엔트로피는 확률 변수 A_i가 가질 수 있는 모든 사건에 대한 정보량을 평균한 값이다. 엔트로피 값이 크다는 것은 사건의 불확실성이 크다는 것을 의미하고 통계적으로는 모든 사건이 동일한 확률을 가지는 것을 의미한다. 확률 변수 A_i의 엔트로피는 아래의 수식 (5)를 통해 계산한다(Vergara and Estévez, 2014).

$$H(A_i) = -\sum_{j=1}^{r_i} P(a_{ij}) \log_2 P(a_{ij}) \qquad \text{수식 (5)}$$

다음으로 임의의 두 범주형 확률 변수 A_i와 $A_{i'}$에 대한 조건부 엔트로피는 아래의 수식 (6)을 통해 계산된다

(Vergara and Estévez, 2014).

$$H(A_{i'}|A_i) = \sum_{j=1}^{r_i} p(a_{ij}) H(A_{i'}|A_i = a_{ij}) \quad \text{수식 (6)}$$
$$= -\sum_{j=1}^{r_i} p(a_{ij}) \sum_{s=1}^{r_{i'}} p(a_{i's}|a_{ij}) \log_2 p(a_{i's}|a_{ij})$$

여기서 $\forall i, i' \in \{1,2,\cdots,p\}$, $\forall j \in \{1,2,\cdots,r_i\}$, $\forall s \in \{1,2,\cdots,r_{i'}\}$이며 $p(a_{i's}|a_{ij})$는 조건부 확률이다. 만약 두 범주형 확률 변수 A_i와 $A_{i'}$가 완전히 종속이면 $H(A_{i'}|A_i)$는 0되고 독립이면 $H(A_{i'}|A_i)$는 $H(A_{i'})$이 된다. 따라서 조건부 엔트로피는 아래의 수식 (7)의 관계식을 만족한다(Vergara and Estévez, 2014).

$$0 \leq H(A_{i'}|A_i) \leq H(A_{i'}) \quad \text{수식 (7)}$$

조건부 엔트로피 $H(A_{i'}|A_i)$를 정규화하기 위해 이를 상한 값 $H(A_{i'})$으로 나눈 값을 $D(A_{i'}, A_i)$로 표기하면 아래의 수식 (8)와 같다.

$$0 \leq D(A_{i'}, A_i) \leq 1 \quad \text{수식 (8)}$$

여기서 $D(A_{i'}, A_i)$는 $H(A_{i'}|A_i)/H(A_i)$이며, $D(A_{i'}, A_i)$ 값이 0이면 $A_{i'}$는 A_i에 종속, 1이면 독립이다.

3.2 적합 변수(Relevance variables) 선정

본 장에서는 앞서 정의한 정규화된 조건부 엔트로피 $D(A_{i'}, A_i)$를 통한 평가 기준을 정의하고 이를 이용하여 적합 변수를 선정하고자 한다. 임의의 범주형 확률 변수 A_i가 다른 변수들에 대해 가지는 의존도는 아래의 수식 (9)와 같이 정의된 개별 독립성 점수(Individual independence score)로 측정 할 수 있다.

$$ID(A_i) = \frac{1}{p-1} \sum_{i'=1}^{p-1} D(A_{i'}, A_i), \quad \text{수식 (9)}$$

여기서 $ID(A_i)$은 범주형 확률 변수 A_i와 다른 범주형 확률 변수 $A_{i'}$들의 정규화 된 조건부 엔트로피의 평균값이다. 독립성 점수 $ID(A_i)$가 1이면 다른 모든 변수들과 독립이고, 0이면 모든 변수들과 종속이다. 정의된 독립성 점수 $ID(A_i)$와 적합도 평가 지표 $D(A_{i'}, A_i)$와의 관계를 통하여 적합 변수 선정 여부를 결정하는 평가 기준과 이를 이용한 개별 적합 변수 집합(Individual relevant variables set)을 아래의 수식 (10)와 같이 정의하였다.

$$\delta(A_i) = \{A_{i'} \in F \neg A_i | D(A_{i'}, A_i) \leq \theta \cdot ID(A_i)\} \quad \text{수식 (10)}$$

여기서 $F \neg A_i$는 확률 변수 A_i를 제외된 변수 집합을 의미하며 θ는 데이터의 특성을 조정 해주는 값을 의미한다. 수식 (10)이 의미하는 바는 A_i가 다른 전체 변수와 가지는 평균적인 연관성 $ID(A_i)$와 임의의 한 확률 변수 $A_{i'}$가 A_i와 가지는 연관성 $D(A_{i'}, A_i)$를 상대적으로 비교하는 것이다. 예를 들어 대부분의 범주형 변수가 A_i와 독립이고

소수의 변수(A_i와 $A_{i'}$)만이 종속인 경우, $ID(A_i)$는 1에 가까운 수를 가지지만 $D(A_{i'}, A_i)$는 0에 가까운 수를 가진다. 따라서 $ID(A_i)$와 $D(A_{i'}, A_i)$의 값을 비교를 통해 A_i에 대해 $A_{i'}$가 적합한 변수 인지 여부를 판정한다. 그러나 $ID(A_i)$는 범주형 확률 변수 A_i와 다른 범주형 변수들의 정규화 된 조건부 엔트로피의 평균값이기 때문에 $ID(A_i)$를 보정 없이 $D(A_{i'}, A_i)$와 비교할 경우 평균적인 수의 범주형 변수가 $\delta(A_i)$에 속하게 되는 문제점을 가지고 있다. 이를 해결하고자 본 연구에서는 전체 범주형 자료의 독립 점수 (Categorical data independence score)인 θ를 통해 개별 독립성 점수를 보정한 $\theta \cdot ID(A_i)$를 개별 적합 변수의 평가 기준으로 정의하였다. θ는 아래의 수식 (11)을 통해 계산된다.

$$\theta = \frac{number\ of\ independent\ variable\ pairs}{p \cdot (p-1)} \quad \text{수식 (11)}$$

θ는 전체 변수 쌍의 조합 가운데 독립적인 변수 쌍의 비율을 의미하며 전체 범주형 자료를 구성하는 범주형 변수의 독립정도를 의미한다. 여기서 θ가 0이면 모든 범주형 변수들은 서로 종속이며, 1이면 모든 범주형 변수들은 서로 독립이다. 범주형 변수의 사이의 독립성을 확인하는 방법으로 본 연구에서는 유의수준 0.01에서 χ^2 검정을 시행하였다.

수식 (10)을 통해 계산된 개별 적합 변수 집합 $\delta(A_i)$를 이용하여 적합 변수 집합(Relevance variables set, F_R)은 아래의 수식 (12)과 같이 정의할 수 있다.

$$F_R = \{A_i |\ \delta(A_i) \neq \varnothing\} \quad \text{수식 (12)}$$

3.3 잉여 변수(Redundancy variables) 제거

이번 장에서는 앞서 정의된 적합변수 집합 $F_R = \{A_1, \cdots, A_{p'}\}$에서 잉여변수(Redundancy variables)를 제거하는 방법에 대해 다룬다. 잉여변수는 적합 변수들 사이의 연관성을 통해 정의된다(Yu and Liu, 2003). 본 연구에서는 다른 적합 변수들 사이에 의존관계가 적은 변수를 잉여 변수로 보고 이를 제거하여 최적의 집합 $S = \{A_1, A_2, \cdots, A_{\hat{p}}\}$을 찾는 방법을 제안한다. 앞서 수식 (10)를 통해 계산된 $\delta(A_i)$는 기준 조건 ($D(A_{i'}, A_i) \leq \theta \cdot ID(A_i)$)을 만족하는 개별 적합 변수 집합을 의미한다. 따라서 집합의 크기(Cardinality)를 이용하면 임의의 범주형 변수 A_i와 의존 관계를 가지고 있는 변수의 수를 계산할 수 있다. $card(\delta(A_i))$는 임의의 범주형 변수 A_i와 의존관계를 가지고 있는 변수의 수를 의미한다. 예를 들어 $card(\delta(A_i)) = 3$이 의미하는 바는 임의의 변수 A_i와 의존 관계를 가지고 있는 범주형 변수가 총 3개임을 의미한다. $\delta(A_i)$ 집합의 크기와 적합 변수 집합의 크기 $card(F_R)$를 이용하면 임의의 변수 A_i가 다른 적합변수와 의존 관계를 맺고 있는 비율(Dependency ratio, $\mu(A_i)$)를 아래의 수식 (13)를 통해 계산 할 수 있다.

$$\mu(A_i) = \frac{card(\delta(A_i))}{card(F_R)} = \frac{card(\delta(A_i))}{p'} \quad \text{수식 (13)}$$

수식 (13)를 이용한 최적의 부분 집합 S는 아래의 수식 (14)과 같이 정의할 수 있다.

$$S = \{A_i | \mu(A_i) \geq T\} \quad \text{수식 (14)}$$

수식 (14)은 적합 변수 집합 F_R에 속한 범주형 변수들 중 임계치 상수 T 이상의 다른 변수들과 의존 관계를 맺고 있는 변수만을 선택하는 것을 의미한다. 임계치 상수 T는 0부터 1사이의 값으로 임계치 상수 T가 1이면 적합 변수 집합 F_R에 속한 모든 범주형 변수들과 의존 관계를 맺고 있는 범주형 변수만을 최적 부분 집합 S에 속하는 변수로 선택함을 의미하고 임계치 상수 T가 0이면 적합 변수 집합 F_R에 속한 모든 범주형 변수들이 최적 부분 집합 S의 변수로 선택됨을 의미한다. 본 연구에서는 임계치 상수 T를 중앙값인 0.5로 하였으며, 이는 적합 변수 집합F_R에 속한 범주형 변수 가운데 다른 범주형 변수들과 절반 이상의 의존 관계를 맺고 있는 변수만을 선택함을 의미한다. 아래의 Figure 1은 본 연구에서 제안한 변수선택 방법을 도식화한 것이다.

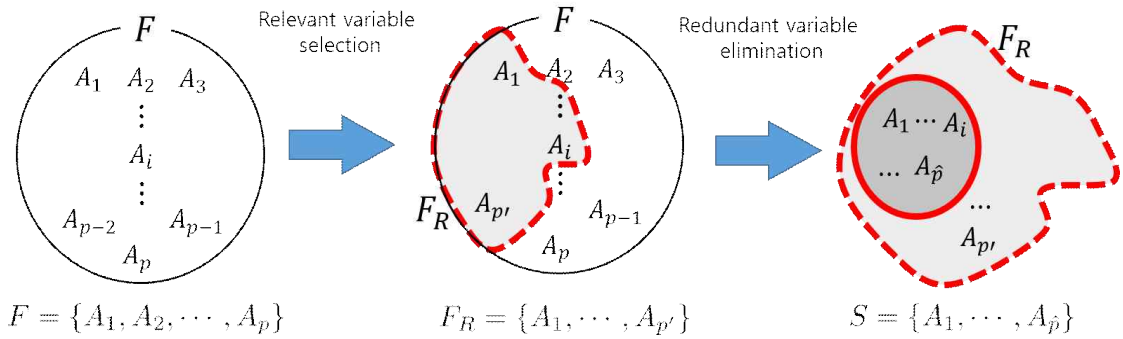

Figure 3. Summary of feature selection procedures

4. 실험 설계

4.1 실험 데이터

본 장에서는 시뮬레이션을 통해 임의로 고차원 범주형 자료를 생성하고 3-최근접 이웃 분류를 통해 기존의 방법 거리 측정 방법과 제안된 변수 선택을 적용한 거리 측정 방법의 성능을 비교하여 제안된 방법의 효과성을 살펴보고자 한다. 본 연구에서 제안한 방법의 효과성을 알아보기 위해 사용한 두 가지의 거리 방법은 해밍 거리(Hamming distance, HD)와 연관성 기반 비유사도(Association based dissimilarity, ABD)이다. 본 연구에서 제안한 변수선택 방법을 적용한 해밍거리를 HHD(High-dimensional Hamming distance)로 표기하고, 변수선택 방법을 적용한 연관성 기반 비유사도를 HABD(High-dimensional Association-based dissimilarity)로 표기하였다. 고차원의 범주형 데이터를 생성하기 위해 UCI repository로부터 실제 범주형 변수로 구성된 자료(Breast cancer)와 임의로 생성한 범주형 자료를 결합하였다. 실제 데이터인 Breast cancer 자료는 총 683개의 관측치와 9개의 범주형 변수로 구성되어 있으며 부류 정보는 2가지 값을 가진다. 임의로 생성한 범주형 변수는 범주형 변수의 수, 범주형 변수 값, 범주형 변수 값의 균형/불균형에 따라 총 6개의 사례를 제시하였다. 인위적으로 생성한 범주형 변수는 50개부터 50씩 증가

시켜 600개의 변수를 생성하였고, 한 변수가 가질 수 있는 변수 값은 2부터 4까지 사례에 따라 다양하게 구성했다. 또한 각 변수의 변수 값의 빈도는 균형 혹은 불균형 상태로 다르게 생성하여 다양한 상황에 따른 제안 변수 선택 방법의 효과를 살펴보고자 하였다. 아래의 Table 3는 실제 데이터와 인위적으로 생성한 데이터의 특징을 정리한 것이다.

Table 4 Summary of Breast cancer data and synthetic generated categorical data

	Number of categorical variables	Number of values in each categorical variable	Balanced/ Imbalanced
Breast cancer	9	Depends on each categorical variable	Depends on each categorical variable
Case 1	From 50 to 600	2	Balanced (1/2, 1/2)
Case 2	From 50 to 600	3	Balanced (1/3. 1/3, 1/3)
Case 3	From 50 to 600	4	Balanced (1/4, 1/4, 1/4, 1/4)
Case 4	From 50 to 600	2	Imbalanced (1/10, 9/10)
Case 5	From 50 to 600	3	Imbalanced (1/10, 1/10, 8/10)
Case 6	From 50 to 600	4	Imbalanced (1/10, 1/10, 1/10, 7/10

4.2 실험 결과

본 장에서는 실제 데이터 (Breast cancer)와 각 시뮬레이션 사례를 결합하여 범주형 변수의 수를 증가시킴에 따라 변수 선택 여부에 따른 성능의 차이를 k-최근접 이웃 분류의 정확도를 통해 살펴보았다. 아래의 Table 4와 Figure 2는 각 사례 별로 k-최근접 이웃 분류의 정확도를 계산하고 도식화 한 것이다. 시뮬레이션 결과 모든 사례에서 변수선택 과정이 생략된 기존의 거리 측정 방법은 (HD와 ABD) 상호 독립적인 (즉, 연관성이 작은) 범주형 변수의 수가 늘어남에 따라 분류 정확도가 하락하는 것을 확인할 수 있었다. 즉 분석에 불필요한 변수의 수가 늘어감에 따라 유의미한 변수들을 선택하는 단계가 필요함을 의미한다. 또한, ABD의 경우 변수 값의 분포가 불균형한 경우 (Case 4, Case 5, Case 6)에서 변수의 수가 늘어감에 따라 정확도가 더욱 심하게 하락하였다. 이에 반해 범주형 변수 값의 수에 따른 정확도의 차이는 크게 나타나지 않는 것으로 보인다. 다음으로 HHD와 HABD는 범주형 변수의 증가하더라도 동일한 정확도를 나타냈다. 이는 불필요한 변수의 수가 증가하더라도 본 연구에서 제안한 변수 선택 단계에서 분류 분석에 유의미한 변수만을 선택함을 의미한다. 아래의 Table 4는 각 사례에서 계산 된 최대, 최소, 평균 정확도와 그 표준편차를 의미한다.

5. 결 론

정보 기술의 발달로 인해 기업들은 다양한 형태로 측정된 빅데이터를 수집함에 따라, 거대한 규모의 자료를 분석해야 하는 상황에 놓이게 되었다. 수집된 데이터가 수치형으로 표현된 고차원 데이터를 다루는 방법은 많은 연구자에 의해 다양한 방법들이 제안되었다. 그러나 대다수의 고차원 자료를 다루는 기법들은 수치형 자료에만 국한하여 적용될 수 있기 때문에 고차원 범주형 자료에서 다루는 기법에 대한 논의는 상대적으로 적은 관심을 받아왔다. 변수 선택 방법은 대표적인 고차원의 자료를 다루는 기법이다. 본 연구에서는 부류 정보가 없는 고차원 범주형 자료에서 적절한 변수를 선택하는 방법을 제안하였다.

본 연구에서는 적절한 변수를 다른 변수와 연관성이 높은 변수로 정의하고 이를 선택하는 변수 선택 방법을 제안하였고 그 효과성을 살펴보기 위하여 변수간 연관성이 존재하는 실제 데이터와 실제 데이터와 연관성이 적은 독립적인 범주형 변수를 결합한 고차원 범주형 데이터를 생성한 뒤 k-최근접 이웃 분류 분석을 시행하였다. 분석 결과 본 연구에서 변수 선택 방법을 적용한 거리 측정 방법이 변수 선택 방법을 적용하지 않은 거리 측정 방법들과 비교하여 더 정확하게 분류 문제를 해결할 수 있음을 확인하였다.

본 연구의 시사점은 다음과 같다. 범주형 자료로 구성된 고차원의 데이터 또한 적절한 변수 선택이 필요하며 부류 정보를 이용하지 않는 범주형 변수 선택 방법을 통해 분류 문제 해결의 정확도를 높이는 것이 가능하다. 따라서 변수 선택과 변수간의 연관성 정보를 이용하는 방법을 통해 차원의 저주를 해소함과 동시에 계산 비용의 감소와 일반화의 능력을 갖출 수 있다. 하지만 본 연구에서는 엔트로피 개념에 기초하여 두 범주형 확률 변수 간의 연관성만을 고려한 한계점을 지니고 있다. 따라서 향후 추가적인 후속 연구 주제로는 2개 이상의 범주형 변수 사이의 교호 작용을 고려한 변수 선택 방법을 개발하거나, 범주형 변수와 연속형 변수가 혼재된 경우에 차원 축소 방법과 관측치 간 거리 측정 방법을 개발하는 것이다. 범주형 변수와 연속형 변수가 혼재되어 있는 경우에는 각 형태별 거리 측정 방식이 혼합되어 하나의 거리 값으로 결정되는 과정에서 상대적 가중치의 결정 문제와 분류 결과의 원인을 탐색하는 진단(Diagnose)의 문제가 나타날 것으로 기대된다. 그 이외에도 임계치 상수 T의 변화에 따른 본 연구에서 제안한 변수 선택 기법의 성능의 차이를 살펴 보는 것도 흥미로운 연구 주제로 보인다. 이런 후속 연구들은 빅데이터를 활용한 품질경영 영역에서 범주형 자료를 더욱더 다양하게 활용하는 흥미로운 주제가 될 것으로 보인다.

Table 5. Experiment results (Maximum, minimum accuracy rate, average of accuracy rate and standard deviation of accuracy rate)

Scenarios	The accuracy of 3-NN classification	Hamming distance (HD)	High dimensional Hamming distance (HHD)	Association-based dissim-ilarity (ABD)	High dimensional association based dissimilarity (HABD)
Breast cancer + Case 1	Max	91.51%	97.07%	97.36%	97.95%
	Min	85.07%	97.07%	82.28%	97.95%
	Average	87.70%	97.07%	87.66%	97.95%
	S.D.	2.36%	0.00%	5.76%	0.00%
Breast cancer + Case 2	Max	92.68%	97.07%	97.22%	97.95%
	Min	85.07%	97.07%	80.09%	97.95%
	Average	87.85%	97.07%	85.20%	97.95%
	S.D.	2.36%	0.00%	5.77%	0.00%
Breast cancer + Case 3	Max	94.00%	97.07%	96.49%	97.95%
	Min	86.09%	97.07%	80.53%	97.95%
	Average	88.84%	97.07%	85.03%	97.95%
	S.D.	2.39%	0.00%	5.06%	0.00%
Breast cancer + Case 4	Max	95.75%	97.07%	97.51%	97.95%
	Min	84.33%	97.07%	75.84%	97.95%
	Average	90.03%	97.07%	84.71%	97.95%
	S.D.	3.71%	0.00%	7.48%	0.00%
Breast cancer + Case 5	Max	93.56%	97.07%	96.05%	97.95%
	Min	83.75%	97.07%	72.77%	97.95%
	Average	88.32%	97.07%	81.41%	97.95%
	S.D.	3.09%	0.00%	7.27%	0.00%
Breast cancer + Case 6	Max	94.58%	97.07%	95.90%	97.95%
	Min	81.26%	97.07%	71.01%	97.95%
	Average	86.38%	97.07%	79.22%	97.95%
	S.D.	3.88%	0.00%	7.52%	0.00%

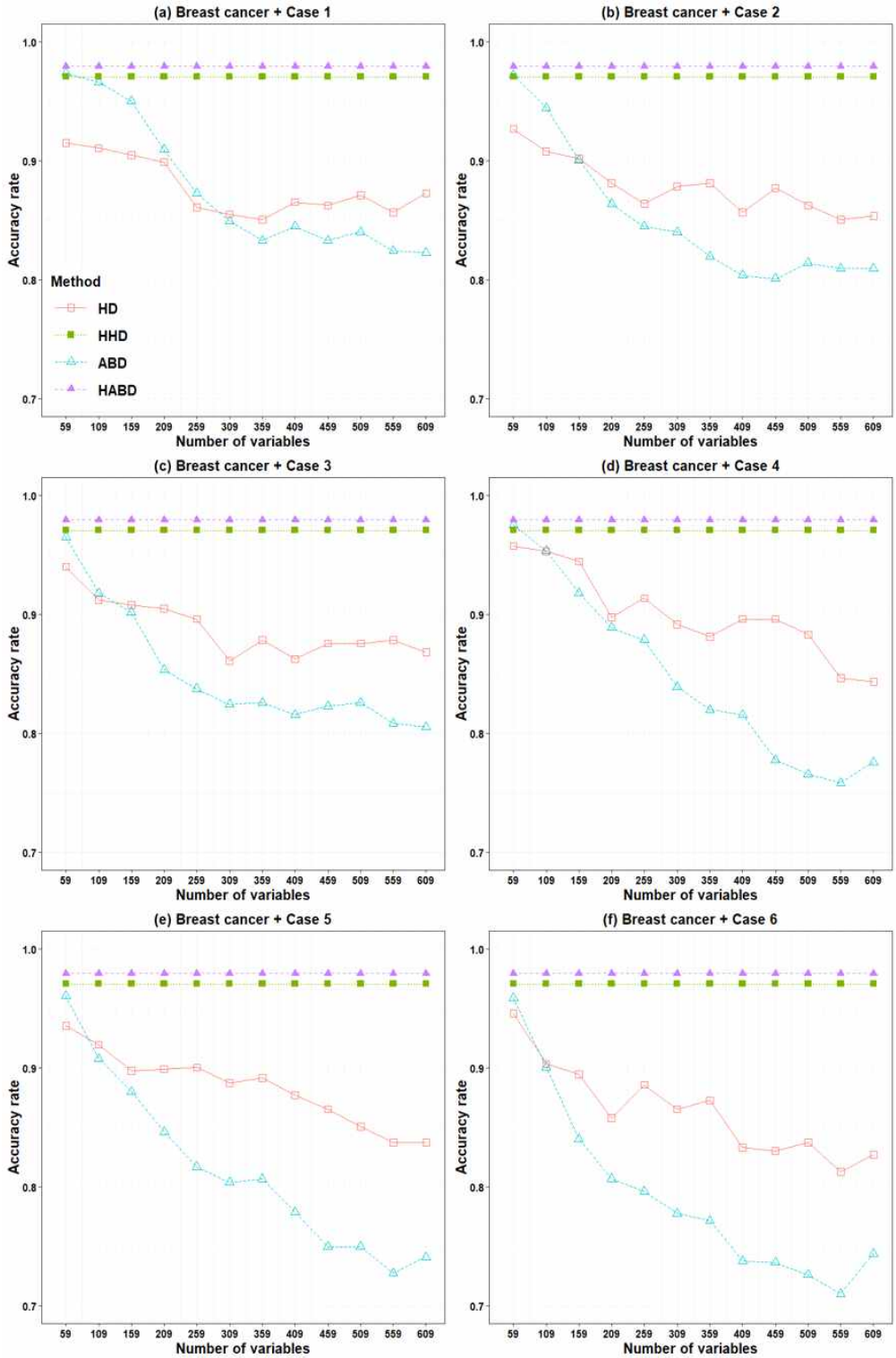

Figure 4. Results of simulations

REFERENCES

Blum, A. L., and Langley, P. 1997. "Selection of Relevant Features and Examples in Machine Learning." Artificial intelligence 97(1-2):245-271.

Burnaby, T. P. 1970. "On a Method for Character Weighting a Similarity Coefficient, Employing the Concept of Information." Journal of the International Association for Mathematical Geology 2(1):25-38.

Chakraborty, D. D. 2008. "Statistical Decision Theory. Estimation, Testing and Selection." Investigación Operacional 29(2):184-185.

Cheng, V., Li, C. H., Kwok, J. T., and Li, C. K. 2004. "Dissimilarity learning for nominal data." Pattern Recognition 37(7):1471-1477.

Chong, H. R., Hong, S. H., Lee, M. K., and Kwon, H. M. 2017. "Quality Management on the 4th Industrial Revolution." Journal of the Korean Society for Quality Management 45(4):629-648.

Cost, S., and Salzberg, S. 1993. "A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features." Machine learning 10(1):57-78.

Goodall, D. W. 1966. "A New Similarity Index Based on Probability." Biometrics, 882-907.

Guyon, I., and Elisseeff, A. 2003. "An Introduction to Variable and Feature Sselection." Journal of Machine Learning Research, 3(Mar), 1157-1182.

Hamming, R. W. 1950. "Error Detecting and Error Correcting Codes." Bell System Technical Journal, 29(2):147-160.

Jia, H., Cheung, Y. M., and Liu, J. 2016. "A New Distance Metric for Unsupervised Learning of Categorical Data." IEEE Transactions on Neural Networks and Learning Systems 27(5):1065-1079.

Kullback, S., and Leibler, R. A. 1951. "On Information and Sufficiency." The Annals of Mathematical Statistics 22(1):79-86.

Le, S. Q., and Ho, T. B. 2005. "An Association-based Dissimilarity Measure for Categorical Data." Pattern Recognition Letters 26(16):2549-2557.

Lin, J. 1991. "Divergence Measures Based on the Shannon Entropy." IEEE Transactions on Information Theory 37(1):145-151.

Lin, D. 1998. "An Information-theoretic Definition of Similarity." In Icml 98(1998):296-304.

Liu, H., Sun, J., Liu, L., and Zhang, H. 2009. "Feature Selection with Dynamic Mutual Information." Pattern Recognition 42(7):1330-1339.

Mitra, P., Murthy, C. A., and Pal, S. K. 2002. "Unsupervised Feature Selection Using Feature Similarity." IEEE Transactions on Pattern Analysis and Machine Intelligence 24(3):301-312.

Il seok Oh. 2008. Pattern recognition, Kyobo book.

Park, Y. J., and Kim, S. B. 2014. "Unsupervised Feature Selection Method Based on Principal Component Loading Vectors." Journal of Korean Institute of Industrial Engineers 40(3):275-282.

Quinlan, J. R. 2014. C4. 5: Programs for Machine Learning. Elsevier.

Ree, S. 2017. "Proposal of Korean Quality Management in the 4th Industrial Revolution." Journal of the Korean Society for Quality Management 45(4):739-760.

Smirnov, E. S. 1968. "On Exact Methods in Systematics." Systematic Biology 17(1):1-13.

Stanfill, C., and Waltz, D. L. 1986. "Toward Memory-based reasoning. Commun." ACM, 29(12):1213-1228.

Tibshirani, R. 1996. "Regression Shrinkage and Selection Via the Lasso." Journal of the Royal Statistical Society: Series B (Methodological) 58(1):267-288.

Vergara, J. R., and Estévez, P. A. 2014. "A Review of Feature Selection Methods Based on Mutual Information." Neural Computing and Applications 24(1):175-186.

Xie, J., Szymanski, B., and Zaki, M. 2010. Learning Dissimilarities for Categorical Symbols. In Feature Selection in Data Mining:97-106.