퍼지 성능 측정자를 이용한 적응 데이터 마이닝 모델

이 현 숙†

요 약

데이터 마이닝은 방대한 양의 데이터를 다루는 응용영역에서 학습과 함께 연구되어 실세계의 문제를 해결할 수 있는 구체적인 방법을 제시해 주고 있다. 데이터 마이닝을 위한 보편적인 방법으로 사용되어 온 클러스터 분석 방법은 데이터의 양이 많아질수록, 실세계에서 직접 얻은 데이터 일수록 경계가 불분명하고 처리과정에서 많은 오차가 발생하게 되어 직접 적용하고자할 때 고려해야할 점이 많다. 이를 위하여 퍼지 개념이 도입된 퍼지 클러스터링 방법론은 클러스터 타당성문제와 함께 널리 연구되어왔다. 본 논문에서는 클러스터링의 결과가 만들어 내는 오류값을 최소화하는 방향으로 학습하는 비교사 학습신경망에 의하여 클러스터링이 이루어지고 이를 퍼지 성능 측정자에 의하여 평가하면서 최적의 클러스터 수를 찾아가는 적응형 데이터 마이닝 모델을 제안하고자 한다. 또한 뉴스그룹의 텍스트 데이터를 처리하여 문서분류에 활용할 수 있음을 보임으로 제안된 모델의 타당성을 확인하고자 한다.

키워드 : 데이터 마이닝, 퍼지 클러스터링, 비교사 학습 신경망, 클러스터 타당성문제

Adaptive Data Mining Model using Fuzzy Performance Measures

Hyunsook Rhee†

ABSTRACT

Data Mining is the process of finding hidden patterns inside a large data set. Cluster analysis has been used as a popular technique for data mining. It is a fundamental process of data analysis and it has been playing an important role in solving many problems in pattern recognition and image processing. If fuzzy cluster analysis is to make a significant contribution to engineering applications, much more attention must be paid to fundamental decision on the number of clusters in data. It is related to cluster validity problem which is how well it has identified the structure that is present in the data. In this paper, we design an adaptive data mining model using fuzzy performance measures. It discovers clusters through an unsupervised neural network model based on a fuzzy objective function and evaluates clustering results by a fuzzy performance measure. We also present the experimental results on newsgroup data. They show that the proposed model can be used as a document classifier.

Key Words : Data Mining, Fuzzy Clustering, Unsupervised Neural Network, Cluster Validity Problem

1. 서 론

인간과 유사한 기능을 구현하기 위하여 필수적으로 요구되는 학습능력을 부여하기 위한 노력이 계속되어왔다. 인공지능분야의 신경망 이론이나 유전자 알고리즘과 같은 학습 알고리즘이 발표되고 실용화되어 로보틱스, 자동화 분야에 적용되어 활용되기도 하였다. 컴퓨터가 처리하는 정보의 양이 방대해짐에 따라 학습을 통하여 정리된 정보를 추출하고 비슷한 상황에 적절한 정보를 제공하는 일은 더욱 중요시되어 왔다. 학습은 주위환경에 적응하기 위한 기초적인 형태로서 편리한 인터페이스를 가지는 인간의 생활 속의 컴퓨터를 구성하기 위한 기본적인 기능이다. 어떤 관점에서는 학습만이 데이터의 무제한적인 증가로 원하는 정보에 접근하기 어려운 현재 정보시스템의 문제를 해결하기 위한 유일한 방법일 것이다.

데이터 마이닝이란 대용량의 데이터베이스에서 이전에 알려지지 않은(unknown) 유효하고(valid), 활용 가능한(actionable) 정보를 꺼내는 작업을 말한다. 일반적으로 데이터 마이닝은 데이터의 특징, 클래스 및 패턴 등을 발견하는 과정을 말한다. 이러한 데이터 마이닝은 KDD(Knowledge Discovery in Database)와 동격의 과정으로 볼 수 있으며 그 처리주기는 다음과 같다[1-3].

첫째로, 수많은 데이터베이스, 혹은 인터넷의 자료 중에서 자신이 처리하고자 하는 목적을 결정(Objective Determination)하는 과정이다. 두 번째 과정은 선택된 데이터를 준비

† 정 회 원 : 동양공업전문대학 전산정보학부 부교수
논문접수 : 2006년 7월 4일, 심사완료 : 2006년 9월 15일

하는 과정(Data Preparation)이다. 이 과정은 데이터 선택(Data Selection), 데이터 선 처리(Data Preprocessing), 데이터 변환(Data Transformation)의 세 개의 세부과정으로 이루어져 있으며 적용하려는 영역데이터에 따른 여러 처리과정이 필요하다. 세 번째 과정은 이전의 과정에서 변환된 데이터를 가지고 실제로 마이닝을 처리하는 과정이다. 그리고 네 번째 과정은 마이닝된 결과를 분석하는 과정(Analysis of Results)이다. 이 과정에서 분석된 결과에 대하여 데이터 가시화(Data Visualization) 작업을 수행한다. 마지막으로는 마이닝된 결과를 실제 회사에서의 업무나 사용자가 원하는 작업으로 흡수하는 과정이다.

이러한 처리 주기 중에서 세 번째 단계에서 필요한 실제적인 데이터 마이닝을 위하여 정보추출, 정보획득, 패턴인식, 학습이론 등 그동안 여러 분야에서 연구되어 온 알고리즘과 연구결과를 활용할 수 있다. 그 중에서 대표적으로 사용된 기법은 신경망을 이용한 클러스터링 방법론이다[4, 5]. 클러스터링은 비교사 학습(Unsupervised learning) 방법으로 속성이 비슷한 것들끼리 묶어 나누는 것으로, 분석하고자 하는 데이터가 너무 많아 전체를 파악하기 어려울 때, 부분을 살펴 전체의 윤곽을 잡도록 해준다. 이와 같은 클러스터링은 전문가시스템, 패턴인식, 영상처리, 음성인식 등 학습이 필요한 모든 영역에서 포함하는 기본 단계이다. 특히 요즘 인터넷을 통한 검색엔진의 탐색과정에도 활용되어 비슷한 정보를 그룹 핑하여 보여주고 있다. 그러나 기존의 클러스터링 방법론은 대부분 hard partitioning에 의한 방법으로 주어진 데이터 상호간의 경계가 명확하다는 가정에서 각 패턴을 하나의 클래스에 소속시키는 방법이다. 그러나 이 모델은 우리가 다루는 데이터의 경계가 대부분 불명확하므로 실제 데이터 상호간의 군집성을 묘사하기에 부적절하며 주어진 데이터 분포의 성질을 잃어버리는 결과를 가져온다[6].

본 논문에서는 구성된 퍼지 신경망을 통하여 클러스터링 결과가 만들어 내는 오류 값을 요약하는 퍼지 함수의 값이 최소가 되도록 학습의 방향을 유도하는 메카니즘[6]에 의하여 클러스터링을 수행하고 이때 얻은 클러스터링 결과를 퍼지 성능 측정자에 의하여 평가하면서 최적의 클러스터 수를 찾아가는 적응 데이터 마이닝 모델을 제안하고자한다. 또한 뉴스그룹의 텍스트데이터를 처리하여 문서분류에 활용할 수 있음을 보임으로 제안된 모델의 타당성을 확인하고자한다.

2. 연구배경

퍼지 이론은 1965년 Zadeh에 의하여 처음으로 도입되었다[7]. 퍼지이론은 0이나 1중에 어느 하나만을 선택하는 이분법에서의 정보 손실을 막기 위하여, 0과 1사이의 값으로 소속정도를 표현하도록 하는 접근 방법으로 주어진 데이터 구조에 대하여 더욱 정확한 표현방법을 제공해 준다. 이를 개선하기 위하여 Bezdek은 Fuzzy c-Means(FCM) 알고리즘[8]이라고 불리우는 퍼지 분할에 의한 클러스터링 방법을 고안하였다. FCM 알고리즘은 최소자승 기준 함수(least square criterion function)에 퍼지 이론을 적용한 목적함수의 반복 최적화(iterative optimization)에 기반을 둔 방식이다. 이 알고리즘은 hard partitioning에 의한 기존의 클러스터링 방법이 승자 독점(winner take all)형태의 전략을 취하는데 비하여, 각 패턴이 특정 클러스터에 속하는 소속정도를 줌으로서 보다 정확한 정보를 형성하도록 도와준다. 이러한 FCM 알고리즘은 최적 퍼지 분할, 패턴분류와 영상 분할등의 여러 응용에 적용되어 유용한 결과를 얻었으며, 이 방법을 변형한 여러 알고리즘이 개발되어 검증 되고 있다[9]. 이와 같은 퍼지 클러스터링 알고리즘은 그의 타당성을 측정하는 방법과 함께 연구되어 방법의 타당성을 입증할 수 있고 그 응용성을 더욱 확대시킬 수 있다.

2.1 FCM 알고리즘

Fuzzy c-Means(FCM) 알고리즘은 퍼지 개념을 데이터 클러스터링 방법론에 적용하여 식(1)과 같은 최소 자승 오류 함수(least square error functional) J_m의 반복 최적화(iterative optimization)에 기반을 두고 개발되었다.

$$J_m = \sum_{j=1}^{n}\sum_{i=1}^{c}(u_{ij})^m (d_j)^2 \tag{1}$$

여기서 u_{ij}는 주어진 입력 데이터 집합 $X = \{x_1, \ldots, x_n\}$에 대한 퍼지 c 분할을 n×c의 벡터 U로 나타낼때 그의 한 요소로 데이터 x_j의 클러스터 i에 속하는 소속정도를 표현한다. 또한 $(d_{ij})^2 = \| x_j - v_i \|^2$ 이고 $\| \cdot \|$ 은 유클리드 놈을, v_i는 클러스터 i의 중심점을 나타내며 $m \in [1, \infty)$은 퍼지정도를 표시하는 파라메타를 나타낸다. 이때 Bezdek은 m>1인 경우 J_m의 국소적 최소점이 되기 위한 필요조건(충분조건은 아니지만)으로 다음의 식(2)와 식(3)를 유도하였다.

$$v_i = \frac{\sum_{j=1}^{n}(u_{ij})^m x_j}{\sum_{j=1}^{n}(u_{ij})^m}, \quad 1 \le i \le c \tag{2}$$

$(d_{ij})^2 = \| x_j - v_i \|^2 > 0$ 이면 모든 클러스터 i에 대하여

$$u_{ij} = \frac{1}{\sum_{t=1}^{c}(\frac{d_{ij}}{d_{tj}})^{2/(m-1)}} \tag{3a}$$

로 정의하고, $(d_{ij})^2 = \| x_j - v_i \|^2 = 0$인 경우 다음의 조건을 만족하도록 모든 클러스터 i에 대하여 u_{ij}를 정의한다.

$$u_{ij} = 0 \quad \text{if} \, d_{ij}^2 \neq 0 \, \text{and} \sum_{i \in I_k} u_{ij} = 1 \tag{3b}$$

FCM 알고리즘은 단지 식(2)와 식(3)의 반복에 의하여 수렴점을 찾아가는 과정이다[8]. 이 방법의 수렴성과 최적화에

대한 고찰은 계속 진행되어 발전되고 있다[9].

위의 유도된 식 (3)을 이용하여 입력 데이터와 중심점 사이의 거리를 통한 퍼지 소속 함수 값(fuzzy membership value)을 결정하게 된다. 식(3b)를 통하여 알 수 있는 바와 같이 임의의 한 클래스의 중심점과의 거리가 0인 입력 데이터에 대한 그 클래스안의 퍼지 소속 함수 값은 1이 될 것이다. 그리고 식(3)을 고찰하여 알 수 있는 바와 같이 그를 통하여 결정된 퍼지 소속 함수 값은 형성된 각 클래스에 대하여 상대적인 값을 가지며 확률적인 제약을 준수하여 그 클래스에 대한 소속의 확률 치나 공유의 정도로 해석된다. 그러나 퍼지 이론에서 이용하는 소속 함수는 그 클래스에 대한 일치도나 전형성의 정도로서 해석되는 절대적인 값이므로 이를 보완하기 위한 연구도 진행되고 있다.

2.2 클러스터 타당성 측정함수

클러스터 분석 방법을 특정 응용에 적용하기 위해서는 그 결과가 주어진 입력 데이터의 구조를 얼마나 잘 반영하고 있는가를 측정하는 척도가 필요하다. 이에 관련된 연구 영역을 "cluster validity problem"이라하며 이를 위하여 다음과 같이 클러스터 타당성 측정 함수를 정의하였다[8].

• Partition Coefficient(F) :

$$F(U;c) = \frac{\sum_{j=1}^{n}\sum_{i=1}^{c}(u_{ij})^2}{n} \tag{4}$$

• Classification Entropy(H) :

$$H(U;c) = \frac{-\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}\log_a u_{ij}}{n} \tag{5}$$

• Proportion Exponent(P) :

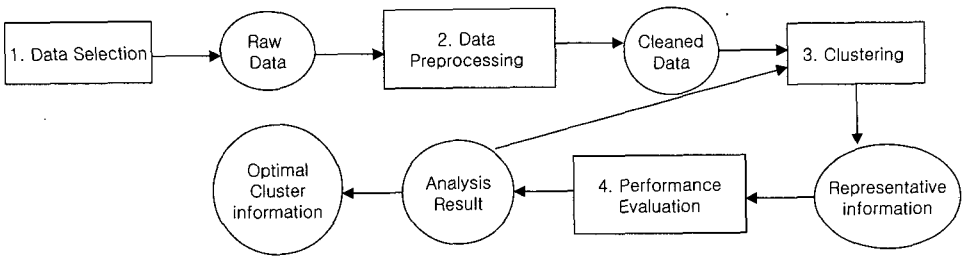
$$P(U;c) = \log_e\left[\prod_{j=1}^{n}\left\{\sum_{l=1}^{\lceil u_l^{-1}\rceil}(-1)^{j+1}\binom{c}{j}(1-ju_l)^{c-1}\right\}\right]$$

$$\text{where } u_l = \bigvee_i u_{ij} \tag{6}$$

위에서 기술한 타당성 측정 함수들은 클러스터링의 결과 형성된 소속 값 안에 포함된 정보를 하나의 값으로 간략하게 요약해 준다. 그러나 이들 타당성 측정 함수들은 그들이 산출하는 값이 주어진 데이터가 가지는 기하학적인 성질을 직접 반영하고 있지 않으며 c의 값이 증가함에 따라 클러스터링의 질과 관련 없이 단조 감소하는 값을 유도해 내는 단점을 가지고 있다. 특히 이와 같은 측정 함수를 클러스터링 알고리즘에 결합시킨 방법은 경험적 성질을 가지며 임의의 임계치를 설정해 주어야 한다. 타당성 측정을 위한 함수가 단조 감소하는 경향을 극복하기 위하여 그 함수에 정규화(normalization)나 통계적 표준화(statistical standardization)를 적용하는 방법이 제안되었다[8]. 이에 Gunderson은 주어진 데이터의 기하학적인 고려를 염두해 두고 separation coefficient를 타당성 측정 함수로서 제안하기도 하였다. Separation coefficient는 기하학적으로 같은 클래스의 데이터는 군집해 있고 다른 클래스 사이의 데이터는 멀리 위치하고 있는지를 알아내기 위하여 고안되었으나 이 방법은 퍼지 클러스터링 알고리즘에 직접 적용할 수 없다. 우선 클러스터링의 결과를 hard한 것으로 변경한 후에 적용해야 하는데 그 변경과정은 여러 방법이 있을 수 있으므로 하나의 결과를 얻을 수 없는 단점을 가지고 있다. Separation coefficient는 최악의 경우를 고려하므로, 평균적인 상태를 고려하기 위한 방법으로 Xie[10]는 타당성 함수 S를 제안하였다. S는 퍼지 c 분할의 클래스안에서의 밀접성과 클래스 사이의 분리정도의 평균을 구하기 위한 측정 함수로서 FCM 알고리즘의 목적함수와도 밀접한 관계가 있음이 고찰되었다.

3. 적응 데이터 마이닝 모델

퍼지 목적함수를 비교사 학습 신경망에 결합한 학습 알고리즘에 의하여 클러스터링이 이루어지고 이를 퍼지 성능 측정자에 의하여 평가하면서 최적의 클러스터 수를 찾아가는 적응형 데이터 마이닝 모델을 제안하고자 한다. 제안된 모델의 대략적인 모형은 기본적인 데이터 마이닝 프로세스를 바탕으로 (그림 1)과 같다. (그림 1)의 과정 중 Data Selection 모듈과 Data Preprocessing 모듈은 우리가 마이닝하려는 목적과 데이터 유형에 따라 처리될 수 있는 부분이다. Preprocessing 을 거친 Cleaned data는 m 개의 속성 값을 가지는 n 개의 데이터 집합으로 제공되어 실제적인 마이닝 과정인 클러스터링을 위한 입력으로 활용된다. 본 논문에서는 클러스터링 모듈과 그 결과 추출해 낸 대표 정보값을 가지고 클러스터링 결과를 평가하는 Performance Evalu



(그림 1) 적응 데이터 마이닝 모델

-ation 모듈을 중점적으로 다룬다. 이때 제안된 성능측정자에 대하여 값을 평가하여 데이터베이스에 저장하고 다음 가정에 대하여 클러스터링 모듈을 다시 수행한다. 즉 예측할 수 있는 여러 가정에 따라 클러스터링과 성능 평가를 반복적으로 수행하여 얻은 결과를 통하여 최적의 클러스터를 찾게 된다. 이에 사용된 모델을 보다 자세히 3.1 절과 3.2 절에서 소개하고자한다.

3.1 목적함수기반 비교사 학습 신경망

제안된 모델에서 클러스터링 모듈을 위하여 FCM 알고리즘의 퍼지 목적함수를 비교사학습신경망에 결합시켜 (그림 2)와 같은 비교사 학습신경망을 구성하였다. 이렇게 구성된 신경망에서 다음의 알고리즘을 통하여 입력 층에 제공된 데이터 ($X_1, \ldots, X_n$)는 대표정보인 클러스터의 중심점 ($v_1, v_2, \cdots, v_c$)을 학습해 간다. 이러한 학습을 통해 형성된 클러스터 층은 정보 사이의 관계를 표현하는 값인($\alpha_1 \ldots \alpha_c$)를 계산하여 그 결과를 다음 학습에 활용한다. 이러한 학습 알고리즘은 클러스터링의 결과가 만들어 내는 오류 값을 요약하는 퍼지 함수를 설정한 후 그 값이 최소가 되도록 학습의 방향을 유도하는 메카니즘에 의해 진행된다. 또한 제안된 방법은 입력과 출력 사이의 관계를 기술하기 어려운 경우도 쉽게 처리하는 비교사 학습신경망의 장점도 함께 가지고 있다.

단계 1 : c, m, ε의 값을 설정하고 입력데이터 셋을 준비한다. c는 클러스터의 수, m은 FCM 알고리즘의 weighting exponent 이다.

단계 2 : 초기 가중치 벡터 $V=(v_1, v_2, \cdots, v_c) \in R_{ck}$와 퍼지 C 분할 U를 0과 1 사이의 난수로 초기화한다.

단계 3 : 다음 식을 이용하여 α_{ij}를 계산하고

$$\eta_i = \frac{1}{\sum_{j=1}^{n} \alpha_{ij}} \text{ 이라고 하자.}$$

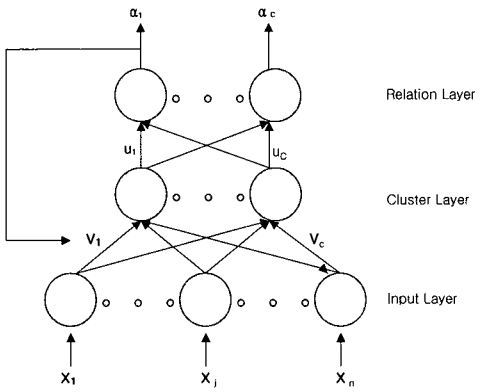
$$\alpha_{ij} = \frac{2m}{m-1} \left\{ \sum_{i=1}^{c} (U_{ij})^{m+1} \left(\frac{\|X_j - V_i\|^2}{\|X_j - V_i\|^2} \right) \right\}^{\frac{m}{m-1}}$$

단계 4 : 다음 식을 이용하여 가중치 벡터를 수정한다.

$$\triangle V_i = \eta \sum_{j=1}^{n} \alpha_{ij} (X_j - v_{ij})$$

단계 5 : FCM 알고리즘을 이용하여 퍼지 C 분할 U를 계산한다.

단계 6 : $diff = \sum_{i=1}^{c} \| V_{i,t+1} - V_{i,t} \|^2 < \varepsilon$ 이면 알고리즘을 끝내고 그렇지 않으면 단계 3으로 가자



(그림 2) 비교사 퍼지 학습신경망

3.2 최적 클러스터를 찾기 위한 성능 측정자

Szu[5]는 패턴분류에 mini-max 필터 개념을 도입하였다. 그는 interclass distance에 반비례하고 intraclass distance에 비례하는 식 (7)과 같은 에너지 함수를 구상하여 그 함수의 값을 최소화시키는 분류알고리즘을 고안하였다.

$$\text{Energy} = \sum \frac{1}{\text{Interclass Distance}} + \sum \text{Intraclass Distance} \quad (7)$$

이러한 개념과 퍼지이론을 적용하여 타당성 측정함수를 설계하여 보자.

n을 주어진 데이터 집합의 데이터 수라 하고, c는 정해진 클러스터의 수이며 각 클러스터는 $C_1, C_2, \ldots, C_c$로 나타낸다. 또한 $d^2(X, Y) = \| X - Y \|^2$ 이며 $\| \ \|$은 유클리드 놈을 나타낸다.

[정의 1] 내부거리(intraclass distance)는 같은 클러스터 안에서의 임의의 두 데이터 사이의 거리로서 전체 데이터에 대한 평균은 다음의 식 (8)과 같이 정의된다. 이때 ω1은 임의의 두 데이터가 같은 클러스터에 속하는 소속정도를 나타낸다. 이와 같이 정의된 C는 퍼지 분할의 밀집성을 나타낸다.

$$C = \frac{2}{n(n-1)} \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} \sum_{i=1}^{c} d^2(X_j, X_k) \omega_1 \quad (8)$$

,where $\omega_1 = \min\{u_{ij}, u_{ik}\}$

[정의 2] 외부거리(interclass distance)는 서로 다른 클러스터에 속하는 임의의 두 데이터 사이의 거리로서 전체 데이터에 대한 평균은 다음의 식 (9)과 같이 정의된다. 이때 ω2는 임의의 두 데이터가 각각 서로 다른 두 클러스터에 속하는 소속정도를 나타낸다. 이와 같이 정의된 D는 퍼지 분할의 분리성을 나타낸다.

$$D = \frac{1}{n^2} \sum_{j=1}^{n} \sum_{k=1}^{n} d^2(X_j, X_k) \omega_2 \quad (9)$$

,where $\omega_2 = \min\{u_{i_1 j}, u_{i_2 k}\}$

이때 x_j의 가장 큰 소속 값을 가지는 클러스터를 i1 이라 하고 i1 이 아닌 클러스터 중 Xk 가 가장 큰 소속 값을 가지는 클러스터를 i2라 하자.

[정의 3] 퍼지 분할의 밀집성 C에 대한 분리성 D의 비율을 클러스터 타당성 척도 G로서 정의한다. 즉 G = D/C.

이렇게 정의된 G는 밀집성과 분리성의 비율로 나타내므로 데이터 분포와 클러스터 사이의 관계를 반영한 값을 산출하며 데이터 집합 사이의 상대적인 비교가 가능하게 하였다. G의 정의로부터 G의 값이 클수록 클러스터 안에서의 밀집성과 클러스터 사이의 분리성이 더욱더 뚜렷한 경우임을 알 수 있다.

이렇게 마련된 타당성 척도 G의 타당성을 확인하기 위하여 널리 알려진 벤치 마크 데이터인 iris data를 준비하여 2장에서 소개한 함수 F, S와 비교하였다. Anderson의 iris data set은 iris 식물의 형태로부터 추출된 4가지 속성의 값을 가지며 각각 50개의 데이터로 구성된 3 종류의 데이터로 구성되어 있다. 그 중 한 가지 종류는 다른 것들로부터 선형분리 가능하고 나머지 두 종류는 선형분리 가능하지 않다. 이 데이터는 여러 분류알고리즘의 성능분석에 사용된 표준 데이터 집합이다. 클러스터의 수 c=2부터 6까지로 하여 실험한 결과는 <표 1>과 같다. 결과에서 보는 바와 같이 제안된 함수 값은 최적 클러스터의 수가 이미 알려진 바와 같이 3으로 나타나지만 다른 함수, F와 S는 2로 나타났다.

〈표 1〉 Iris 데이터에 대한 함수 측정값

클러스터의 수	G	F	S
2	4.89	0.90*	0.05*
3	7.55*	0.79	0.14
4	7.00	0.69	0.61
5	7.01	0.63	0.40
6	6.92	0.58	0.74

4. 실험 및 고찰

제안된 마이닝 기법을 적용하기 위하여 20 Newsgroups [11]에서 5가지의 서로 다른 뉴스 그룹으로부터 각각 20개의 기사를 준비하였다. 100개의 전체 문서에 대하여 전 처리 과정을 통하여 클러스터링 과정에 제공될 데이터를 준비하였다. 전 처리 과정을 통하여 마이닝 작업에 유용한 항(word)을 선택하였다. 처리하고자 하는 텍스트 데이터는 항으로 구성되어 있으며 항을 처리하는 대표적인 방법[12]을 참조하여 특질 선택(Feature Selection)을 하였다. 첫 번째로 적용한 방법은 stop word 라 불리우는 the, and, does, they 등과 같은 대명사, 관사, 접속사, 전치사 등 자주 사용되는 단어들을 제거한다. 다음으로 n번 이상 나타나지 않는 단어들을 제거하였다. 본 실험에서는 n을 3으로 하였다. 이러한 과정을 통해 남은 항들에 대하여 단어 Stemming 처리를 하게 되면 문서에서 추출한 단어의 수를 절반가량 줄일 수 있었다. 이렇게 정리된 데이터에 대하여 원하는 목적의 개념과 높은 관계가 있는 항을 선택하기 위하여 정보 엔트로피의 개념을 가지고 있는 측정자 Mutual Information I(T, w)를 사용하였다[12, 13]. I(T, w)는 Training 문서 T에 대한 word w의 mutual information 으로서 엔트로피(Entropy)를 통해 구해진다. 어떤 항 t가 모든 클래스에 포함되어 있거나 혹은 어떤 문서도 t를 가지고 있지 않으면 그 항은 해당 클래스 C를 나타내는 특질로 적절치 못하다는 것이 기본 개념이다. 이때 I(T, w)를 통해서 순위를 매겨 적절한 특질을 선택할 수 있다. I(T, w)는 다음과 같이 기술할 수 있다. I(T, w)의 정의에서 C는 클래스들의 집합을 나타내고, D는 Training 문서, T(d)∈C는 문서 d가 클래스 C에 속하는지에 대한 true function이다.

$$
\begin{aligned}
I(T,w) &= E(T) - E(T \mid w) \\
&= -\sum_{C \in C} \Pr(T(d)=C) \cdot \log \Pr(T(d)=C) \\
&\quad + \sum_{C \in C} \Pr(T(d)=C, w=0) \cdot \log \Pr(T(d)=C \mid w=0) \\
&\quad + \sum_{C \in C} \Pr(T(d)=C, w=1) \cdot \log \Pr(T(d)=C \mid w=1)
\end{aligned}
$$

이와 같은 단계를 거쳐 I(T,w) 값이 큰 순서로 상위 10위의 단어들, $w_0, \ldots, w_9$,을 선택하게 된다. 이렇게 선택된 단어들은 문서를 분류하는 결정적인 단어가 되므로 이를 이용하여 100개의 각 문서에 대한 이 단어들의 항 빈도수를 구하여 학습에 사용하였다. 이와 같은 전 처리 과정에 의하여 Mutual Information의 값이 상위 10위안에 드는 word, 즉 문서를 분류하는 결정적인 단어들(w_0-w_9)을 선택하였다. 이제 100개의 각 문서(d_0-d_{99})에 대하여 문서분류에 널리 사용되는 w_0-w_9의 term frequency $TF(w_i, d_j)$를 구하여 정규화시킨다. 결과적으로 준비된 데이터 X는 100×10의 w_0-w_9에 대한 normalized term frequency, $NTF(w_k, d_j)$가 된다. 이 값을 입력으로 (그림 2)의 신경망을 구성하고 3.1의 알고리즘을 적용하여 클러스터링을 수행하였다.

$X = (x_0, x_1, \ldots, x_{99})$,
$x_j = (NTF(w_0, d_j), NTF(w_1, d_j), \ldots, NTF(w_9, d_j))$,
$j = 0, 1, \ldots, 99$

다음 <표 2>로부터 클러스터의 수 c=2 부터 8까지의 G의 값으로부터 입력시 준비한 바와 같이 클러스터의 수가 5일 때 최적의 클러스터링을 수행함을 확인할 수 있다.

〈표 2〉 클러스터링 결과 측정된 G의 값

클러스터의 수(c)	2	3	4	5*	6	7	8
측정된 G의 값	3.45	8.75	10.57	11.72	9.35	10.27	8.54

위의 결과를 바탕으로 학습한 결과를 테스트하기 위하여 뉴스그룹의 기사 d를 임의로 추출하여 학습 데이터와 같은 과정으로 NTF(w_i, d), w_0-w_9 값을 구하여 이를 테스트 데이터로 하여 결과를 확인하였다. 테스트 결과 학습한 문서의 경우 평균적으로 100개의 기사 중 97개는 정확하게 자신의 뉴스그룹을 찾았다. 하지만 학습에 사용되지 않았던 뉴스그룹 기사의 경우는 81개 정도 평균적으로 자신의 뉴스그룹을 찾았다. 두 경우 모두 문서 분류에 사용될 정도의 정확성을 가지므로 제안된 방법의 타당성을 입증할 수 있었다. 텍스트 분류의 경우 사람이 분류하여도 어느 그룹에 속하는지 결정하기 어려운 기사도 있다는 것을 고려할 때 의미 있는 결과를 얻었다고 볼 수 있다.

5. 결 론

인터넷과 정보의 홍수 속에 살고 있는 이 시점에서 정보의 검색과 활용은 어려운 연구과제로 남아있다. 이에 데이터 마이닝 기술은 방대한 양의 데이터를 탐색하여 숨겨진 정보와 규칙 또는 요약정보를 얻어 유용하게 정보를 활용하기 위한 분야로서 각광 받고 있다. 이러한 데이터 마이닝 과정 속에 클러스터 분석 방법은 여러 학습 메카니즘과 함께 널리 사용되어왔다. 본 논문에서는 클러스터링의 결과가 만들어 내는 오류 값을 요약하는 퍼지 함수를 설정한 후 그 값이 최소가 되도록 학습의 방향을 유도하여 입력과 출력 사이의 관계를 기술하기 어려운 경우도 쉽게 처리하는 퍼지 신경망의 구조를 보여주고 있다. 이를 바탕으로 퍼지 목적함수를 비교사 학습 신경망에 결합한 학습 알고리즘에 의하여 클러스터링이 이루어지고 이를 퍼지 성능 측정자에 의하여 평가하면서 최적의 클러스터 수를 찾아가는 적응형 데이터 마이닝 모델을 제안하였다. 제안된 모델이 타당하게 접근하고 있음을 확인하기 위하여 뉴스그룹의 텍스트 데이터 집합을 처리하여 문서분류에 활용해 보았다[14, 15].

텍스트 데이터는 기존의 클러스터분석방법에서 벤치마크 데이터로서 다룬 여러 데이터와는 달리 구조화되어 있지 않기 때문에 실험적인 차원에서 활용할 수 있음을 확인하였다. 그러나 실세계의 응용영역과는 달리 아주 작은 데이터 집합을 예제로 다루어 그 타당성만을 확인하였고 검색 엔진 등의 웹 마이닝 기법으로 활용하려면 텍스트 마이닝 등의 관련 연구가 뒷받침되어야한다. 또한 여러 분야의 다양한 형태의 방대한 데이터에 적용하여 실험하는 정교한 분석과정을 통해 실세계에 적용하려는 시도가 계속되어야할 것이다. 이를 통하여 마이닝 과정이 재정립될 것이고 그 결과를 데이터베이스화하여 활용하려는 방안도 구체화될 수 있을 것이다.

참 고 문 헌

[1] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy., "Advances in Knowledge Discovery and Data Minng". AAAI/MIT Press, 1995.

[2] Cabena, Hadjinian, Stadler, Verhees, Zanasi, "Discovering Data Mining From Concept to Implementation", Prentice-Hall, 1997.

[3] Robert Groth, "Data Mining", Prentice Hall PTR, 2000.

[4] Chin-Teng Lin, "Support-Vector-Based Fuzzy Neural Network for Pattern Classification", IEEE Transactions on Fuzzy Systems, Vol.14, No.1, Feb., 2006.

[5] Rabunal, J. Ramon and Dorrado, Julian, "Artificial Neural Networks in Real-life Applicatons", Idea Group, 2005.

[6] Hyun-Sook Rhee and Kyung-Whan Oh, "A Design and Analysis of Objective Function-Based Unsupervised Neural Networks for Fuzzy Clustering", Neural Processing Letters Vol.4, 1996, p.83.

[7] L. A. Zadeh, "Fuzzy Sets", Information and Control 8, 1965.

[8] J. C. Bezdek, "Pattern Recognition with Fuzzy Objective Function Algorithms", Plenum press, New York, 1981.

[9] Jian Yu and Miin-Shen Yang, "Optimality Test for Generalized FCM and Its Application to Parameter Selection", IEEE Transactions on Fuzzy Systems, Vol.13, No.1, Feb., 2005.

[10] Xuanli Lisa Xie and Gerado Beni, "A Validity Measure for Fuzzy Clustering", IEEE Trans. on Pattern Anal. Machine Intell., vol. PAMI-13, no.8, 1991.

[11] KDD 20 Newsgroups Data, http://kdd.uci.edu/databases/20newsgroups

[12] Thorsten Joachims, "A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization", Tech. rep., Carnegie Mellon University, 1996

[13] Soumen Chakrabarti, "Data mining for hypertext: A tutorial survey", SIGKDD Explorations, 2000.

[14] 최윤정, 박승수, "학습방법 개선과 후처리분석을 이용한 자동 문서분류의 성능향상 방법", 정보처리학회논문지, Vol.12-B, No.7, Dec., 2005.

[15] W. N. Street and Y. S. KIm, "Streaming ensemble algorithm(SEA) for large-scale classification", Proc. of 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.377-382, San Francisco, California, 2001.

이 현 숙

e-mail : hsrhee@dongyang.ac.kr

1989년 서강대학교 전자계산 학과(학사)

1991년 포항공대 대학원 컴퓨터공학과(석사)

1997년 서강대학교 컴퓨터학과(박사)

1991년~1997년 한국전자통신연구소(ETRI) 연구원

1997년~현재 동양공업전문대학 전산정보학부 부교수

관심분야 : 소프트컴퓨팅, 패턴인식, 데이터마이닝