The Journal of Digital Policy & Management 2013 Oct; 11(10): 387-394

러프집합을 통한 취업의사결정 분석시스템

이희태*, 박인규**
평택 기계공업고등학교*, 중부대학교 컴퓨터학과**

Decision Analysis System for Job Guidance using Rough Set

Heui-Tae Lee*, In-Kyoo Park**
Dept. of mechatronics, Pyeongtaek mechanical and technical highschool*
Dept. of Computer Science Joongbu University**

요 약 데이터 마이닝은 예측이나 분석을 위해서 많은 양의 데이터에 존재하는 여러 가지의 관계를 추출하는 과정이라고 할 수 있다. 그러한 데이터에는 매우 많은 변수로 인한 차원의 증가로 인하여 계산상의 어려움이 수반되어지고 변수의 중복성과 중요도에 있어서 다양한 통계적 관계가 존재한다. 따라서 동일하거나 유사한 데이터를 같은 그룹으로 형성하는 클러스터 해석은 데이터 마이닝에서 필수적인 요소이다. 본 연구는 범주형 데이터의 분류에서 발생하는 불확실성의 처리를 위해 러프집합을 이용하여 정보 엔트로피를 이용한 새로운 척도를 정의하고 연구 대상에 대한 유사행동을 분석하는 시스템 구현에 그 의의가 있다. 데이터는 평택공업고등학교에서 채집되었고 이를 토대로 제안된 방법이 학생들의 유사행동에 대한 보다 정확한 결과를 보임을 알 수 있었다. 또한 속성의 개수가 10개 이상인 경우에 기본 방법과의 차이를 보이며 취업의사결정에서 학생들의 의식을 기존 방법보다 효과적으로 반영하였다.

주제어 : 데이터 마이닝, 클러스터 해석, 불확실성, 엔트로피, 러프집합

Abstract Data mining is the process of discovering hidden, non-trivial patterns in large amounts of data records in order to be used very effectively for analysis and forecasting. Because hundreds of variables give rise to a high level of redundancy and dimensionality with time complexity, they are more likely to have spurious relationships, and even the weakest relationships will be highly significant by any statistical test. Hence cluster analysis is a main task of data mining and is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. In this paper system implementation is of great significance, which defines a new definition based on information-theoretic entropy and analyse the analogue behaviors of objects at hand so as to address the measurement of uncertainties in the classification of categorical data. The sources were taken from a survey aimed to identify of job guidance from students in high school pyeongtaek. we show how variable precision information-entropy based rough set can be used to group student in each section. It is proved that the proposed method has the more exact classification than the conventional in attributes more than 10 and that is more effective in job guidance for students.

Key Words : Data Mining, Cluster Analysis, Uncertainty, Entropy, Rough Set

Received 15 August 2013, Revised 17 September 2013
Accepted 20 October 2013
Corresponding Author: In-Kyoo Park(Joongbu Univ.)
Email: fip2441g@gmail.com

ISSN: 1738-1916

1. 서론

데이터마이닝이라는 것은 방대한 양의 데이터 속에서 쉽게 드러나지 않는 유용한 정보를 찾아내는 과정이라고 말할 수 있다. 컴퓨터 과학적인 관점에서는 패턴 인식 기술뿐만 아니라 통계적이나 D적 분석 방법을 이용하여 저장된 거대한 자료로부터 우리에게 유익하고 흥미 있는 새로운 관계, 성향, 패턴 등의 다양한 부가가치 정보를 찾아내는 일련의 과정이라고 정의하고 있다. 또한 MIS 관점에서는 거대한 데이터베이스 혹은 자료에서 유용한 정보를 추출하는 일련의 과정뿐만 아니라 값진 정보를 사용자가 전문적 지식 없이도 사용할 수 있는 의사결정지원시스템의 개발 과정을 통틀어 데이터 마이닝으로 정의하고 있다. 마지막으로 통계적 관점에서는 올바른 의사결정을 지원하기 위한 자료분석(data analysis) 및 모형선택(model selection)으로 정의한다. 이와 같이 데이터 마이닝은 여러 가지의 모든 사용 가능한 근원 데이터를 기반으로 감춰진 지식, 기대하지 못했던 경향 또는 새로운 규칙 등을 발견하고 이를 통하여 어떠한 목적의 의사결정을 하는데 유용한 정보로 활용하고자 하는 것이다.

데이터 마이닝은 데이터에 내재하는 데이터간의 관계, 패턴, 규칙을 발견하는 데 그 목적이 있으며 작업유형에 따라서 사용되는 분석기법(analysis algorithms)으로 통계적인 방법은 모든 데이터를 대상으로 평균이나 표준편차, 선형적인 정규분포를 이용하여 유도된 결론에는 어느 정도의 정보의 손실이 발생할 수 있다[1]. 또한 범주형 데이터에서는 속성의 개수가 많고 그 값이 다양하기 때문에 하나의 객체가 여러 개의 집단으로 분류되는 불확실성으로 인하여 기존의 알고리즘에서는 애매함을 완전하게 처리하지 못하고 있는 실정이다. 이에 대해 범주형(categorical) 데이터를 임의의 기준에 따라서 분류하는 과정에서 발생하는 불확실성을 처리할 수 있는 알고리즘이 필요하다. 따라서 실제적인 데이터는 항상 정규분포를 가정할 수 없으며, 비선형과 불확실성을 내포하기 때문에 의사결정 알고리즘으로 적성데이터를 분석하는 것이 효과적이다. 따라서 불확실하고 애매모호한 성격을 가지는 데이터를 분석하는데 있어서 합리적이고 효과적인 방법은 러프 집합이론(rough set theory)이라 할 수 있다. 따라서 본 논문에서는 러프집합을 이용한 군집화 알고리즘(clustering algorithm)을 제안하고 가변정확도(variable precision) 러프집합의 모형을 기반으로 엔트로피(entropy)에 기초한 새로운 정보 엔트로피(information entropy) 를 정의하고자 한다. 결국 제안된 정보 엔트로피를 기반으로 가변정확도 기반의 러프집합 모형을 이용한 데이터 마이닝 기법을 통하여 학생들의 취업패턴을 도출하여 취업을 지도하시는 담당자가 이를 토대로 학생들에 대한 성향을 파악하여 취업지도에 향상을 유도하고자 한다.

2. 러프집합이론

2.1 러프집합

러프집합(rough set)이론은 1982년 Zdzislaw Pawlak에 의해 개발되었다[2]. 이것은 데이터 테이블에서 군집화(clustering)된 분석을 다루는 것으로 데이터는 현실세계에서 반영되는 연속적인 값(continuous value)의 문제까지 해결 할 수 있는 내용을 나타낸다. 전통적인 접근방법에서 어떤 문제에 대한 해결 방법을 찾고자 할 때 정보단위(information granules)에 대하여 적합하고 세밀한 방법이 필요한데 이에 대한 해결 방법이 러프집합이다.

러프집합이라는 개념에 대한 이론은 부정확한 것들(imprecision, vagueness, uncertainty)에 대한 새로운 D적 접근 방법이라 할 수 있다. 러프집합에서의 기본적 개념은 수렴하면서 식별하기 어려운 관계 속에서 D적 기초를 이용하여 어떤 정보들(data, knowledge)에 대한 연관성에 대한 모든 객체들을 발견하는데 있다. 부정확한 데이터로 부터의 추론, 보다 정확히 말하면 데이터간의 관계를 발견하는 것이다. 이는 통계와 밀접한 관련이 있어 보이나 접근방법에서 전혀 다르다. 데이터의 애매성을 나타내기 위해 확률을 이용하는 대신에 러프집합을 이용한다. U가 관심대상인 객체의 유한집합이고 B는 모든 속성들의 집합 A의 유한집합 즉, B∈A, V는 모든 속성 값의 집합일 경우에 U×A→V는 정보함수이다. 전체집합의 임의의 부분집합 X⊆U를 U내의 개념(concept) 또는 범주(category)라 한다. R이 U의 동치관계이면 U/R은 R의 모든 동치류(equivalence class)들의 집합을 나타내고 이를 R의 범주또는 R의 개념이라고 하며 [x]R은 원소 x∈U를 포함하는 범주를 나타낸다. 지식기반은 U≠∅이 전체집합이라 불리는 유한집합이고 R이 U의 동치관계들의

집합일 때 정보시스템 K=(U, R)을 나타낸다[3].

2.2 러프집합의 근사화

A에 부분집합인 모든 속성들의 집합 B 즉, B⊆A에서 IND(B)는 하나의 이진관계가 되고 B의 식별불가능관계(indiscernibility relation)로써 식(1)과 같이 정의 할 수 있다.

$$IND(B) = (x,y) \in U^2 : \text{for } \forall\ a \in B, a(x) = a(y) \quad (1)$$

$$[x]_{IND(B)} = \bigcap_{a \in B} [x]_R \quad (2)$$

여기서 IND(B)는 동치관계를 이루고 있고 IND(B)를 만족하는 객체 A와 B는 B의 속성에 의하여 식별할 수 없는 관계를 가지고 있다. 또한 식(2)와 같이 U/IND(B) 즉, 동치관계 IND(B)의 모든 동치류의 집합은 K내의 U에 관한 B-기본지식인 동치관계 B의 집합과 관련된 지식을 나타낸다. <표 1>에서 B={C1, C2, C3, C4}의 경우에 U/IND({C1}) = {{1,2,5}, {3,4,6}, {7,8}}과 U/IND(B) = {{1},{2},{3},{4},{5},{6},{7},{8}}이 된다.

〈Table 1〉 Decision table of job guidance

Attr(A) / Obj(U)	C1	C2	C3	C4	D
1	0	0	0	2	1
2	0	0	1	0	1
3	1	0	0	0	1
4	1	1	0	2	2
5	0	1	0	1	2
6	1	1	2	2	3
7	2	1	2	2	3
8	2	2	2	2	3

R이 U의 동치관계일 경우에 지식기반 K(U, R)의 경우 X⊆U와 동치관계 B⊆A을 써서 두 집합 B하한근사와 B상한근사를 각각 식(3)과 식(4)와 같이 정의할 수 있다.

$$\underline{R}(X) = \{x \in U \mid [x]_R \subseteq X\} \quad (3)$$

$$\overline{R}(X) = \{x \in U \mid [x_i]_R \cap X \neq \varnothing\} \quad (4)$$

예를 들어 집합 X={5,6,7,8}에 대하여 속성집합 B={C2, C3}에 대하여 B하한근사는 {6,7,8}이고 B상한근사는 {4,5,6,7,8}이고 X의 B경계영역은 {4,5}이다.

2.3 근사화의 척도

임의의 집합이 부정확하다는 것은 경계영역이 존재하기 때문이다. 또한 집합의 경계영역이 커질수록 그 집합의 정확성은 떨어진다. 이 개념을 보다 D적으로 나타내기 위한 정확성 척도(accuracy measure)가 식(5)에 나타나 있다.

$$\partial_B(X) = \frac{card\ \underline{B}}{card\ \overline{B}},\ X \neq \varnothing \quad (5)$$

정확성 척도 $\partial_B(X)$는 집합 X에 대한 지식의 완전한 정도를 측정한다. 이 범위는 모든 B와 X ⊆U에 대하여 $0 \leq \partial_B(X) \leq 1$이고, $\partial_B(X) = 1$이면 X의 B 경계영역은 공집합이고 집합 X는 B정의 가능하다(B-definable). $\partial_B(X) < 1$ 이면 X는 B경계영역을 가지며 집합 X는 B정의 불가능하다(B-undefinable). 물론 집합 X의 부정확성의 정도는 식(6)과 같이 정의 할 수도 있다.

$$\rho_R(X) = 1 - \partial_B(X) \quad (6)$$

이는 B-러프정도(B-roughness)라 불린다. 정확성에 반대되는 개념인 러프정도는 집합 X에 대한 지식 B의 불완전성의 정도를 나타낸다. 부정확성의 수치는 확률론이나 퍼지집합에서와 같이 미리 가정된 것이 아니고 근사화(approximation)라는 개념을 이용하여 지식의 부정확성을 나타내고 있다. 따라서 부정확성은 객체를 분류하는 능력이라고 할 수 있다. 따라서 지식의 부정확성을 나타내기 위해서 부정확성은 정량적인 개념으로 나타나게 된다. 부정확성의 수치적 특징화는 범주의 정확성을 나타내는데 사용되고 실제적으로 여러 응용에서 매우 유용하다고 입증된 바 있다.

3. 가변정확도 정보엔트로피기반 러프집합

3.1 불확실성과 엔트로피

러프집합에서 다루어지고 있는 불확실성은 식별 불가능한 관계에서 발생하는 불확실성(uncertainty)이다. 즉, 동치류들을 구성하는 객체들을 구별할 수 없다는 것이다.

결국 하한근사와 상한근사라는 근사화를 통하여 이러한 불확실성을 모델링을 할 수 있다. 러프집합의 애매함을 나타내는 정확성(exactness)과 러프정도(roughness)이라는 두 가지의 척도가 있다. 1는 상한근사와 하한근사의 비율을 나타낸다. 또한 러프정도의 척도는 러프집합이 가지고 있는 정보가 완전하지 못한 정도를 나타낸다. 그러나 이러한 척도는 완전한 "러프정도"을 처리하지는 못한다.

예를 들어 다음에서 식별불능 관계를 고려해 보자. X={A_{11}, A_{12}, A_{21}, A_{22}, B_{11}, C_1}, A_1={A_{11}, A_{12}, A_{21}, A_{22}}, {B_{11}, B_{12}, B_{13}}, {C_1, C_2}}, A_2={A_{11}, A_{12}}, {A_{21}, A_{22}}, {B_{11}, B_{12}, B_{13}}, {C_1, C_2}}, A_3={A_{11}}, {A_{12}}, {A_{21}}, {A_{22}}, {B_{11}, B_{12}, B_{13}}, {C_1, C_2}}와 같은 분할은 임의의 X에 대하여 동일한 하한근사와 상한근사를 가지기 때문에 동일한 정확도를 가지고 있다. 그러나 A_1이 가장 불확실성이 높고 A3가 가장 낮다는 것을 알 수 있다. 따라서 보다 효과적인 불확실성에 대한 척도가 필요하다. 정보이론에서 보면 확률이 낮은 사건일수록 더욱 놀랍고 정보량은 크다. 따라서 어떤 사건의 확률을 알고 있을 때 정보량을 어떻게 측정할 것인가로 정의할 수 있다. p(x)는 x의 확률이고 h(x)는 x의 정보량은 식(7)과 같다[4].

$$h(x) = -\log_2 p(x) \tag{7}$$

결국 엔트로피는 랜덤변수 x가 가질 수 있는 모든 값(사건)에 대해 정보량을 평균한 것이다. 본 논문에서는 기존의 정보이론에서 사용되는 엔트로피에 대한 척도를 변형하여 러프 엔트로피(rough entropy:RE)를 제안하여 이러한 단점을 보충하였다. 불확실성에 대해서 엔트로피 이론을 변형하여 많은 접근 방법들이 제시되어왔다. 통계역학에서는 엔트로피의 차뿐만 아니라 엔트로피의 절대적 값을 정의할 수 있다. 어떤 통계적 앙상블(ensemble)을 각 미시적 상태 i 의 확률을 p_i 로 정의할 경우에 N개로 구성된 엔트로피 E는 식(8)과 같이 정의할 수 있다[5].

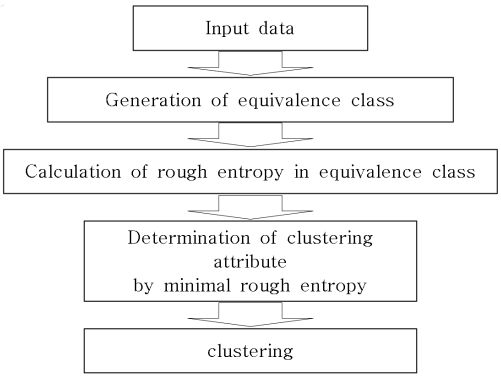
$$E = -k\sum_{i}^{n} p_i \log_2 p_i \tag{8}$$

엔트로피가 커진다는 말은 에너지 무질서도가 높아진다는 말이고 에너지의 질이 떨어진다는 의미가 된다. 따라서 본 논문에서는 러프집합에서의 지식에 존재하는 속성들의 중복성에 관한 불확실성에 대한 문제를 설정하여 임의의 러프집합에서 러프 엔트로피를 식(9)와 같이 정의하였다. 즉, 동치류 X와 Y의 Y에 대한 중복성의 비율과 U에 대한 Y의 비율의 곱이다.

$$RE(X,Y) = -K\log_2 \sum_{j=1}^{n} |X \cap Y_{a_j}| / |Y_{a_j}| \qquad K = card(Y)/card(U) \tag{9}$$

3.2 불확실성과 엔트로피

군집화를 위하여 가변정확도 러프집합(variable precision rough set:VPRS)을 개량한 러프 엔트로피 가변 정확도 러프집합모형(variable precision information -entropy based rough set:VPIERS)을 제안한다. 제안된 방법은 속성에 대하여 러프 엔트로피를 적용하여 속성이 가지는 근사화의 정확도가 항상 0이 아니기 때문에 속성간의 변별력을 유지하는데 보다 안정적이다. 그림 1에 전체적인 군집화를 위한 흐름도가 나타나 있다. 기존의 방법에서는 속성이 가지는 개념간의 중복성을 측정하는 방법으로 속성간의 오류율을 정의하였다. 그러나 이 방법은 개념에 대한 중복성이 일부 동일하게 평가되어 속성간의 변별력이 떨어진다.



[Fig. 1] Clustering algorithm by rough entropy

따라서 이러한 단점을 해결하기 위하여 제안된 러프 엔트로피를 군집화 분석에 적용을 위하여 <표 2>의 예를 고려하자.

〈Table 2〉 Information system for gob guidance

U / A	C1	C2	C3	C4	D
1	0	2	0	1	0
2	0	1	1	0	1
3	1	0	0	0	0
4	1	1	0	1	0
5	2	2	1	0	1
6	2	0	0	0	1
7	2	0	1	1	0
8	1	0	1	1	1

속성의 범주는 C1, C2, C3, C4, D이고, 속성들에 대하여 제안된 방법을 이용하여 가장 신뢰도가 높은 속성을 추출하기 위하여 지식의 속성에 대한 동치류들은 다음과 같다. X('C1'='0') = {1,2}, X('C1'='1') = {3,4,8}, X('C1'='2') = {5,6,7}, U/'C1' = {{1,2}, {3,4,8}, {5,6,7}}, X('C2'='2') = {1,5}, X('C2'='1') = {2,4}, X('C2'='0') = {3,6,7,8}, U/'C2' = {{1,5}, {2,4}, {3,6,7,8}}, X('C3'='0') = {1,3,4,6}, X('C3'='1') = {2,5,7,8}, U/'C3' = {{1,3,4,6}, {2,5,7,8}}, X('C4'='1') = {1,4,7,8}, X('C4'='0') = {2,3,5,6}, U/'C4' = {{1,4,7,8}, {2,3,5,6}}, X('D'='0') = {1,3,4,7}, X('D'='1') = {2,5,6,8}, U/'D' = {{1,3,4,7}, {2,5,6,8}}이다.

각각의 속성간의 의존성의 관계를 조사하여 가장 신뢰도가 높은 속성을 군집화를 수행하는 기준이 되는 속성으로 간주하게 된다. 예를 들어 'C1' 속성에 대하여 'C3' 속성의 러프 엔트로피는 RE('0'|'C3='1')= ({1,2}, {2,5,7,8}) = -(2/8) * Log2(1/2) = 0.173, RE('1'|'C3'='1') = ({3,4,8}, {2,5,7,8}) = -(3/8) * Log2(1/3) = 0.415, RE('2'|'C3'='1') = ({5,6,7}, {2,5,7,8}) = -(3/8) * Log2(2/3) = 0.152, RE('0'|'C3'='0') = ({1,2}, {1,3,4,6}) = -(2/8) * Log2(1/2) = 0.173, RE('1'|'C3'='0') = ({3,4,8}, {1,3,4,6}), = -(3/8) * Log2(1/3) = 0.152, RE('2'|'C3'='0') = ({5,6,7}, {1,3,4,6}), = -(3/8) * Log2(1/3) = 0.415이다. 따라서 이와 같은 엔트로피는 속성간의 불확실성을 나타내기 때문에 이러한 값에서 가장 작은 값을 취함으로써 가장 신뢰도가 높은 속성간의 러프정도는 식(10)에 의하여 구할 수 있다.

$$MRE_{a_j}(a_i = \alpha) = min(RE_{Y|a_j}(X|a_i = \alpha)) \quad (10)$$

MRE('C3'='1') = min(0.173, 0.415, 0.152) = 0.152, MRE('C3'='0')| = min(0.173, 0.152, 0.415) = 0.152이다. 다섯 가지의 속성에 대하여 하나의 속성의 평균 불완전성의 평균을 식(11)에 의하여 구할 수 있다. 즉, 'C1' 속성에 대하여 'C3' 속성의 평균 러프정도는 식(11)과 같이 구해지고, V(a_i)는 '1'와 '0'로 2가 된다. <표 3>에서 가로의 'C3'과 세로의 'C1'의 중복성은 0.152가 된다.

$$MeR_{a_j}(a_i) = K\frac{1}{|V(a_i)|} \quad (11)$$
$$K = MRE_{a_j}(a_i) + ... + MRE_{a_j}(a_{i_{|V(a_i)|}})$$

MeR('C3') = ('C1'(X|'C3'='0') + 'C1'(X|'C3'='1')) / V'C1' = (0.152+0.152)/2 = 0.152이다.

결국 'C1' 에 대한 'C3' 속성의 전체적인 불완전성은 0.152로 나타났다. 이러한 방법으로 모든 속성간의 관계가 <표 3>에 나타나 있다. 또한 식(12)를 이용하여 다섯 가지의 속성에서 발생하는 신뢰도의 평균적인 러프정도(mean roughness)을 구할 수 있다.

$$MMeR(a_i) = \frac{MeR(a_1) + ... + MeR(a_i)}{|V(a_i)|} \quad (12)$$

결국 주어진 다섯 개의 속성에서 군집화속성인 MMMR은 식(13)으로 정의되고 모든 속성의 최소값에 해당하는 속성이 결정된다.

$$MMMR(a_i) = min(MMeR_{a_1}(a_i), .., MMeR_{a_j}(a_i)) \quad (13)$$

〈Table 3〉 MMMR accuracy of 〈Table 2〉

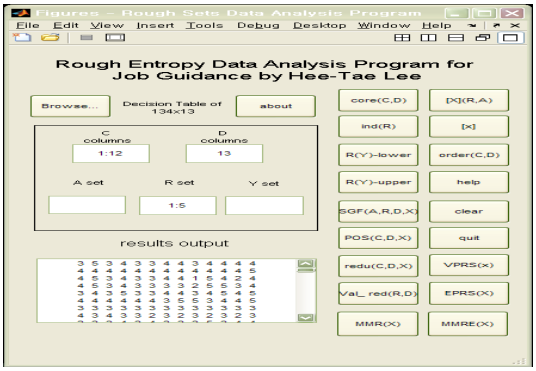
Attr	Rough Entropy Mean Roughness					Mean
	C1	C2	C3	C4	D	
C1		0.1733	0.4621	0.4621	0.4621	0.3899
C2	0.1662		0.5776	0.5776	0.5776	0.4748
C3	0.1520	0.1733		0.3466	0.1438	0.2039
C4	0.1520	0.1733	0.3466		0.1438	0.2039
D	0.1520	0.1733	0.1438	0.1438		0.1533

<표 2>의 지식 시스템에 대하여 [그림 1]의 알고리즘을 이용하여 속성간의 최소평균 정확도를 계산한 결과가 <표 3>에 나타나 있다. 결과에서 알 수 있듯이 속성에 대한 정확도가 가장 높은 속성은 'D' 이다. 따라서 이 속성은 지식 데이터를 분류하기 위한 군집속성으로 간주하게 된다. 추출된 'D' 속성을 기준으로 지식을 분할하게 된다. 결국 'D' 속성에 의한 지식의 동치류는

U/'D'={{1,3,4,7}, {2,5,6,8}}이기 때문에 U의 지식을 {1,3,4,7}와 {2,5,6,8}으로 분할 할 수 있게 된다. 러프엔트로피를 이용한 방법의 경우 평균값이 가장 작은 값이 신뢰도가 가장 양호가기 때문에 군집화 속성으로 검출되었다. 또한 기존의 가변정확도 모형의 경우에는 평균이 큰 값이 신뢰도의 양호함을 나타낸다. 결국 두 가지의 방법의 결과 'D'을 동일한 군집화속성으로 검출하였다. 속성간의 신뢰도를 나타내는 중복성에서 기존의 방법보다 제안된 방법이 변별력이 높다[6].

3.3 군집화 분석 시스템

제안된 기법을 학생들의 취업의사결정에 적용하기 위해 Matlab(Version R2010A)을 이용하여 취업지원 의사결정 시스템(decision support system)을 구현하였다[7,8].



[Fig. 2] Clustering system

〈Table 4〉 Menu of clustering system

Menu	File	Explanation
MAIN	resda	Analysis application
IND(R)	ind*_*	Function of equivalence class
Core(C,D)	core	Ffunction of core
Pos(C,D,X)	pos*.*	Ffunction of positive region
Redu(C,D,X)	redu	Ffunction of reduct
R(Y)-lower	rslower	Lower approximation
R(Y)-upper	rsupper	Upper approximation
SGF(A,R,D,X)	sgf	Significance of attribute
Val_red(R,D)	val_redu	Function of value reduct
Order(C,D)	order	Function of attribute approximation
VPRS(X)	VPRS	Variable precision rough set
VPRSC(X)	vprsc	Variable precision centroiod rough set
EPRS(X)	VEPRS	Variable precision rough entropy set
EPRSC(X)	eprsc	Variable precision centroiod entropy set
MMR(X)	mmr	Min_min_rough set
MMRE(X)	mmre	Min_mean_rough_entropy set
About	rs_about	Help

취업에 관한 데이터의 처리에만 국한되는 게 아니고 모든 분야의 범주형 변수(categorical variable)를 가지는 자료의 분석과 평가에 응용될 수 있도록 구성되었다. [그림 2]에 정보 엔트로피를 이용한 군집화분석 시스템을 나타내었다. 또한 <표 4>에는 분석시스템을 구성하는 파일과 그 적요를 나타내었고 러프집합에 관한 관련연구에서 수식이 가지는 의미를 분석하였다.

4. 결과 고찰

교육의 정보화의 현장 중심, 수요자 중심의 실질적이고도 체계적인 교육 정보화의 실현을 위하여 최신 정보기술인 데이터 마이닝을 활용하여 실업계 고등학교 학생들의 취업 선택과정에 대해 예측하는 모델을 구현하는 것으로 학생들 스스로의 사고는 어떻게 움직이고 있으며, 자신들이 선택한 미래에 대해 얼마나 노력을 하고 있는지를 알아보고자 한다. 학생들이 자신의 적성과 흥미를 정확하게 파악하는 정도가 낮은 상황에서 학생들은 어떠한 판단을 하고 어떻게 움직이고 있는지를 관찰할 필요가 있다.

4.1 데이터 수집

설문 구성은 학생들 스스로가 학교에서 학습하고 있는 C1 전공에 해당하는 전공직무와 취업을 위해 학생들이 무엇을 준비하고 있는가에 대한 구직자들의 노력과 구직에 관한 정보 그리고 구직활동에 어려움이 무엇인가를 조사하였다. 마지막으로 가장 중요한 항목에 해당하는 직장의 중요도에 대하여 갑장 많은 항목을 할당하였다. 취업을 구성하는 여러 가지 변인 인자들의 상관관계를 분석해 보고자 하였다. 전체 153명의 학생에서 19명을 제외한 134명의 설문자료를 데이터베이스로 구축하였다. 여학생의 비중이 전체 7명으로 극히 미비하여 남녀의 구별은 하지 않았다.

4.2 데이터의 전처리

실업계 고등학교 학생들의 취업에 대한 개선으로 여러 가지 제도적인 면과 사회적인 면 그리고 기업적인 면과 같이, 많은 사항이 있지만 무엇보다 학생의 취업의지의 개발과 평가가 가장 절실하다고 보았다. 따라서 <표

5>, <표 6>, <표 7>, <표 8>, <표 9>를 통하여 취업에 관련된 변수들을 다섯 가지의 그룹으로 분류하였다. 전공직무와 관련된 시험, 학교에서 실시하는 구직자들의 노력과 정보원 그리고 구직활동의 어려움에 공히 7가지의 변수를 선정하였고, 직장의 중요도에는 13개의 속성에 대한 변수를 선정하였다.

〈Table 5〉 Attributes of applicant' s major

Attr	Meaning	%1	%2	%3	%4	%5
PP	Poor Preparation	13	39	60	18	4
ME	Melancholy Experience	22	41	49	17	4
WC	Weak Concentration	9	23	59	38	4
CP	Cconfirmmattive Pass	1	27	72	30	4
ME	Misunderstanding of Exam.	11	34	62	26	1
EI	Exminaaiotn Importance	6	25	34	44	25
EE	Embarrassing Exam.	13	36	58	23	4

〈Table 6〉 Attributes of applicant' s endeavors

Attr	Meaning	%1	%2	%3	%4	%5
FL	Foreign Language	3	8	33	33	57
IC	It, Computer	5	15	46	53	15
LA	License Acquisition	2	0	22	46	64
IP	Interview and Personal	1	2	22	41	68
VC	Vocational Carrier	1	3	56	49	25
FT	Faculty Training	1	2	22	50	59
GO	Government Employee	1	5	37	51	40

〈Table 7〉 Attributes of applicant' s sources

Attr	Meaning	%1	%2	%3	%4	%5
IS	Intervention of School	0	6	37	63	28
IA	Intervention of Acquaintance	3	31	48	46	6
IM	Intervention of Mass	8	23	59	36	8
IP	Intervention of Public	3	31	51	45	4
II	Intervention of Internet	4	19	43	57	11
ID	Intervention of Directcall	26	39	48	17	4

〈Table 8〉 Attributes of applicant' s difficulties

Attr	Meaning	%1	%2	%3	%4	%5
PI	Poor Information	11	22	41	51	9
MT	Misunderstanding of Talent	10	26	38	52	8
SF	Short Faculty	15	27	41	42	9
PD	Phisical Defect	52	37	37	5	3
II	Ill Income	22	30	59	20	3
WE	Weak Environments	15	41	55	23	0
SE	Short Employee	23	42	51	16	2

〈Table 9〉 Attributes of applicant' s job

Attr	Meaning	%1	%2	%3	%4	%5
IP	Importance of Pay	0	4	39	69	21
SE	Security of Employee	0	2	45	50	37
SC	Size of Company	0	4	48	60	22
AM	Associative of Major	0	7	44	64	18
DD	Difficulty of Duty	1	8	79	38	8
QB	Quantity of Business	0	4	80	43	7
DE	Development of Employee	0	4	54	49	27
PJ	Profrssional Jump	1	2	47	47	37
ET	Employee Time	2	7	74	36	15
MS	Military Service	20	10	42	32	30
WP	Welfare Program	0	4	48	44	38
TC	Time Commute	3	10	65	38	18
TH	Talent and Hobby	0	7	49	43	35

〈Table 10〉 Clustering result

class no.	applicant major		applicant endeavors		applicant sources		applicant difficulties		applicant job	
	VP	VPRE	VP	VPRE	VP	VPRE	VP	VPRE	VP	VPRE
1	22	22	5	5	8	8	15	15	4	7
2	41	41	15	15	23	23	41	41	48	49
3	50	50	46	46	59	59	55	55	44	43
4	17	17	53	53	36	36	23	23	38	35
5	4	4	15	15	8	8	134	134	134	134
total	134	134	134	134	134	134				

이와 같이 계산된 속성에 의하여 학생들을 분류한 결과가 <표 10>에 나타나 있다. 네 가지의 유형에 대한 분류 결과를 통하여 속성의 개수가 6개와 7개의 경우로써 커다란 변별력을 나타내지 않았음을 알 수 있었다. 그러나 직장의 중요도의 경우에 대한 분석결과를 통하여 속성의 개수가 13개로 구성되어 있기 때문에 변별력이 존재함을 알 수 있었다. 결국 속성간의 애매함은 속성의 개수와 선형적인 관계를 가지고 있다는 것을 확인할 수 있었다.

〈Table 11〉 Clustering attributes of questionnaire' s kinds

Questionnaire's sections	Attr. no.	Clustering attributes
Major	7	Melancholy experience
Endeavors	7	It, computer
Sources	6	Into mass
Difficulties	7	Weak environments
Job	13	Welfare program(VPRS)
		Talent and hobby(VPRERS)

따라서 다섯 가지의 각각의 유형에 대해서 러프집합 분석을 통하여 유형에 따른 분할 속성을 <표 11>과 같이 추출되었다. 이러한 분할 속성을 토대로 학생들을 그룹으로 분할하여 여러 개의 그룹을 형성하여 유사행동을 분석할 수 있다. 따라서 학생들에게 기업체에 대한 취업지도를 수행하는 과정에서 보다 정확하고 정성적인 정보를 가지고 학생들을 지도할 수 있다.

5. 결론

범주형 데이터를 구성하는 속성간의 변별력을 향상시키기 위하여 기존의 엔트로피를 이용하여 새로운 정보엔트로피척도를 정의하였고 학생의 적성과 직업선택의 여러 요인과의 연관성을 파악하여 최적의 직업을 선택하는

데 활용될 수 있는 러프집합을 이용한 군집화 모형을 제안하였다. 또한 정의된 정보엔트로피 척도를 기반으로 데이터를 분할하는 알고리즘을 구현하였다.

"직장의 중요도"의 경우와 같이 속성이 많을 경우(13개) 기존방법의 경우에는 학생들이 후생복지에 관심이 많았는데 반해, 제안된 방법의 경우에는 자기의 적성이나 흥미로 학생들의 관심사가 이동했음을 알 수 있었고 데이터의 분할결과에 대한 기존방법과의 비교우위에서 속성간의 애매성에 대한 변별력의 경우에 만족할 만한 결과를 보였다.

제안된 데이터 마이닝 모형은 취업지도 뿐만 아니라 고교에서 활용할 수 있는 세부 활동별 프로그램에 해당하는 자기이해, 진로상담, 진로이력관리(커리어포트폴리오), 멘토링, 취업캠프, 현장체험 및 견학, 특강 및 강의, 학부모교육, 예비신입생 진로교육 항목으로 확장가능하다. 뿐만 아니라 각각의 항목에서 가장 우세한 분할속성을 이용하여 전체적인 개념을 형성하기 위한 메타데이터로서의 정보를 제시할 수 있을 것으로 사료된다.

REFERENCES

[1] Lin, T. Y. ,and Cercone, N.(eds), Rough sets and data mining-analysis of imperfect data, Boston:Klumer academic publishers, 1997

[2] Pawlak, Z. Rough sets - Theoretical Aspects fo Reasoning about Data, Klumer, 1991

[3] Shampa Semgupta and Asit Kr, Das. Single Reduct Generation based on Relative Indiscernibility of Rough Set Theory, International Journal on SoftComputing(IJSC), Vol.3, No. 1, pp. 107-119, 2012

[4] Shannon, C., L., The mathematical theory of communication, Bell System Technical Journal, Vol. 27, 1948

[5] Beaubouef, T., Petry, F. E. and Arora, G.,, Information-theoretic measures of uncertainty for rough sets and rough relational databases, Information Science, Vol. 109, No. 1-4, pp. 185-195, 1998.

[6] Kanal, L., and Lemmer, J., Uncertainty in artificial intelligence, Amsterdam: North. Holland, 1986..

[7] P.J.Nikumbh, S. K. Mukhopadhyay, Bijan Sarkar and Ajoy Kumer Dutta. Rough Set based Product Mix Analysis, International Journal of Advancements in Technology, Vol. 2, No. 3, pp. 382-398, 2011

I8] Wan Tri Riyadi Yanto et. al., Applying variable precision rough set model for clustering stuent suffering study's anxiety, Expert Systems with Applications, 2012

이 희 태(Lee, Heui Tae)

- 1989년 8월 : 광운대학교 전자계산기공학과(공학석사)
- 2013년 8월 : 중부대학교 정보과학과(공학박사)
- 2012년 3월 ~ 현재 : 평택기계공업고등학교 교사
- 관심분야 : 음성인식,μ-P,러프집합
- E-Mail : useebeer@gmail.com

박 인 규(Park, In Kyoo)

- 1985년 2월 : 연세대학교 전기과 전자계산기 응용(공학석사)
- 1997년 2월 : 원광대학교 전자과 마이크로프로세서 응용(공학박사)
- 1997년 3월 ~ 현재 : 중부대학교 컴퓨터학과 교수
- 관심분야 : 소프트컴퓨팅, 데이터마이닝
- E-Mail : fip2441g@gmail.com