이산화 알고리즘을 이용한 계층적 클러스터링의 실험적 성능 평가

원 재 강* · 이 정 찬** · 정 용 규*** · 이 영 호****

목 차

요약

데이터로부터 의미있는 형태의 정보를 얻기 위한 여러 가지 기법들이 개발되어 왔지만, 최근 들어 가장 각광받는 분야 중 하나는 패턴인식과 기계학습 방법이다. 기존의 학습 알고리즘은 대부분 범주 형 속성에 기반 한 규칙 또는 의사 결정 모델을 생성한다. 그런데, 실세계의 데이터는 보통 범주 형 속성 외에도 수치 값을 갖는 속성을 포함하고, 또 많은 경우에 있어 수치 형 속성으로만 구성되기도 한다. 따라서 이러한 경우, 데이터를 학습에 사용하기 위해서는 수치형 속성에 대한 적절한 처리 과정이 필요하다. 본 논문에서는, 수치형 속성의 도메인을 여러 개의 분절된 부분으로 나누어 학습 알고리즘에 사용하는 방법인 이산화 기법을 설명하고 또한 데이터마이닝의 기법으로 사용되는 클러스터링(Clustering)을 사용한다. 클러스터란 대량의 데이터베이스로부터 유사한 레코드 특성을 지닌 작은 그룹으로 여러 개를 분할하는 것으로 패턴 공간에 주어진 유한 개의 패턴들이 서로 가깝게 모여서 무리를 이루고 있는 패턴 집합이다. 그 집합들 중에서 특정한 카테고리를 지정하지 않고 주어진 데이터들에서 어떤 패턴을 추출하여, 비슷한 데이터들을 묶어서 데이터를 분류하는 기법인 클러스터링에 대해 실험한다.

표제어: 패턴인식, 기계학습, 이산화, Discriminant Analysis, Hierarchical Clustering

접수일(2013년 7월 20일), 수정일(2013년 8월 10일), 게재확정일(2013년 8월 20일)

* 경기대학교, 컴퓨터과학과 강사, 06240604@hanmail.net
** 한국정보화진흥원 창의인재부 책임, jclee@nia.or.kr
*** 교신저자, 을지대학교 의료IT마케팅학과 교수, ygjung@eulji.ac.kr
**** 수원대학교 컴퓨터학과 외래교수, yhleepr@gmail.com

1. 서론

데이터 마이닝이란 대용량의 데이터 셋을 분석하기 위하여 새로운 이론, 기법, 분석 툴을 제공하는 전산분야의 새로운 영역중 하나이며, 기계학습, 클러스터 분석, 회귀 분석, 뉴럴 네트워크 등과 관련이 있는 분야이다. 최근에는 대량의 데이터들이 디지털 형태로 제공됨에 따라 이 분야에 대한 연구가 활발해 지고 있다. 최근 정보 산업 분야에서 데이터 마이닝이 주목받고 있는데 그 주된 이유는 데이터의 양적 팽창과 그러한 데이터를 유용한 정보와 지식으로 바꿔야 하는 필요성 때문이다. 이렇게 얻어진 정보와 지식은 기업경영, 생산운영 그리고 시장분석에서부터 공학설계와 과학탐구에 이르기까지 광범위한 응용 분야에 이용될 수 있다. 그 중 최근 들어 가장 각광받는 분야 중 하나는 기계학습 방법이다. 기계학습에서는 훈련데이터인 실험실용 데이터를 사용하여 알고리즘을 만들어내는 작업이다. 그러나 이러한 일련의 기계학습 작업은 현실 세계의 데이터베이스 갱신이 수시로 이루어지는 등 동적학습방법이지만, 현실세계의 데이터베이스에는 적용하기가 곤란하다. 기계학습 알고리즘에 적용되는 실세계 데이터의 속성은 연속형과 범주형의 혼합된 형태를 가지고 있다. 하지만 대부분의 기계학습 알고리즘은 한 가지 형태의 데이터만을 다룰 수 있기 때문에, 이러한 데이터를 기계학습 알고리즘에 적용시키기 위해서는 데이터 속성의 형 변환이 요구된다. 일반적으로 범주형 값을 연속형으로 변환시키는 문제의 복잡성, 그리고 범주형 값이 분류 규칙을 도출하기에 더 용이하다는 장점으로 인해, 연속형을 범주형으로 변환하는 기법인, 이산화(discretization) 방법을 많이 사용한다. 본 논문에서는 계층적 클러스터링 알고리즘을 이용한 추출 방법을 실험한다. 이 방법은 대량의 데이터로부터 의미 있는 규칙들을 발견해 내는 체계적인 방법론이다. 실험은 단계별로 진행된다. 첫 번째는 데이터들을 전처리하는 과정이며 두 번째 단계는 데이터들을 클러스터링 하는 단계이다. 마지막 단계에서는 전 단계에서 클러스터링 된 것들을 이용하여 의미 있는 규칙들을 발견해 내는 것이다.

2. 이론적 배경

2.1 이산화의 정의

이산화는 정의된 데이터의 수치 값을 속성 값으로 변환하는 것으로 많은 알고리즘은 데이터 이산화 과정이 요구 된다. 그림 1은 이산화 과정의 간단한 예를 보이고 있다. 그림에서 화살표는 목적속성 정보를 최대한 유지시킬 수 있는 이산화 경계를 나타내며, 점선은 그것의 일반화된 이산화 경계이다.

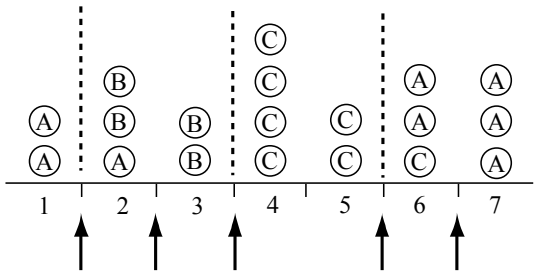

그림 1. 이산화 과정의 예
Fig. 1. An Example of Discretization Process

2.2 이산화 방법의 분류

이산화 방법은 크게 이산화 과정에서 목적 속성 값을 고려하는지 아닌지 여부에 따라 다음 두 가지로 분류될 수 있다.

1. 목적속성에 독립적인 방법(unsupervised method): 일정한 간격으로 구간을 정하는 equal-width intervals 방법, 일정한 데이터 빈도로 구간을 정하는 equal-frequency intervals 방법 등이 있다.
2. 목적속성에 의존적인 방법(supervised method):

속성 값과 목적속성 값과의 상관관계를 구하기 위해 사용된 평가 함수(evaluation function)에 따라 엔트로피 방법, 카이스퀘어 방법, 러프집합(rough set theory) 방법 등이 있다. 일반적으로 목적속성을 고려하지 않은 방법은 정확한 이산화 경계를 갖지 않기 때문에 손실되는 정보의 양이 많다. 그리고 최적의 구간 수를 결정하는 방법 또한 주관적으로 이루어지기 때문에 결정된 구간의 일반성을 보장할 수 없다. 따라서 이러한 문제점을 보완하기 위한 목적속성에 의존적인 이산화 방법들이 연구되어 왔다.

2.3 Equal-Width Binning & Equal-Frequency Binning

Equal-Width Binning과 Equal-Frequency Binning은 비교사, 정적, 전역적 이산화 방법으로 각각 속성의 도메인 구간을 일정한 간격, 인스턴스 수를 갖는 구간으로 분할한다. 먼저 Equal-Width Binning은 속성 도메인의 하한 값과 상한 값을 구한 후, 하한 값과 상한 값 사이의 구간의 k개를 일정한 간격으로 분할한다.

이와 비슷하게 Equal-Frequency Binning은 데이터 집합의 전체 인스턴스 수, n을 구한 후, 각 분할 구간이 n/k개의 인스턴스를 포함하도록 속성 도메인의 하한 값과 상한 값 사이의 구간을 k개로 분할한다. 이때, 각각의 방법은 이산화를 수행하기 전에 미리 적절한 Arity의 값이 입력되어야 한다. 즉, 도메인 구간을 몇 개의 부분으로 분할할 것인가를 사전에 주관적으로 결정해야 한다. 이산화하고자 하는 속성에 대한 정보를 알 경우, 각 속성의 Arity는 속성이 갖는 인스턴스 분포에 따라, 여러 가지 값으로 결정될 수 있을 것이다. 하지만, 속성에 대한 사전 지식이 없는 경우, 이산화 할 속성의 모든 Arity는 보통 단일한 값으로 설정된다.

2.4 1R

1R 알고리즘은 1-level Decision Tree로, 매우 간단한 규칙만으로도 다른 복잡한 학습 방법에 접근한 성능을 보이는 것으로 알려졌다. 1R 이산화 방법은 1R 알고리즘에서 수치형 속성을 다루는 방법으로 교사, 정적, 전역적 이산화 방법이다. 1R 알고리즘은 수치형 속성을 이산화하기 위해, 먼저 속성의 값에 따라 인스턴스의 목적속성 값을 정렬한다. 그리고, 미리 주어진 최소 인스턴스 수에 따라 도메인 구간을 분할한다. 이 때, 분할된 각 부분은, 마지막 구간을 제외하고, 반드시 최소 인스턴스 수 이상의 인스턴스를 포함해야 한다.

2.5 클러스터링 기법

하나의 객체(object)가 여러 속성(attribute)을 갖는다 하고 이러한 객체가 다수 있다고 할 때 클러스터링(clustering)이란 유사한 속성들을 갖는 객체들을 묶어 전체의 객체들을 몇 개의 그룹 또는 군집(cluster)으로 나누는 것을 말한다. 클러스터링 기법은 크게 분할(partitioning) 접근과 계층적(hierarchical) 접근으로 나눌 수 있다. 분할 접근은 범주 함수를 최적화시키는 K개의 분할 영역을 결정해 나가는 방법으로 유클리드 거리(euclidean distance) 측정법에 기반 한다. 숫자 속성(numeric attribute) 데이터를 군집화 하는데 쓰이는 가장 오래되고 잘 알려진 파티션 알고리즘 중에 K-평균(K-means) 알고리즘이라는 것이 있다. 이 알고리즘을 사용하려면 몇 개의 그룹으로 나누기 원하는지 K를 입력해야 한다. 그러면 알고리즘은 일단 K개의 평균점을 지정하고 모든 데이터를 하나씩 보면서 가장 가까운 평균점에 해당되는 그룹에 할당한다. 그 후에 다시 평균점들을 조금씩 바꾸어 나가면서 데이터를 가까운 그룹에 재 할당 한다. 이 과정은 군집 상태를 나타내는 척도 함수가 더 이상 변하지 않을 때까지 반복되며 더 이상 변하지 않게 되면 그 상태

의 그룹들을 군집화의 결과로 정한다.

계층적 접근은 사전에 군집 수 k를 정하지 않고 단계적으로 서로 다른 군집결과를 제공하는 것인데, 이는 다시 집괴법과 분리법(divisive method)으로 나뉜다. 집괴법이란 각 객체를 하나의 군집으로 간주함을 시작으로 유사한 객체들을 묶어 군집으로 만들고 다시 유사한 군집들을 묶어 새로운 군집을 만들어 나가는 과정을 전체의 객체들이 하나의 군집이 되기까지 반복한 후 어떤 규칙에 의하여 최종적인 군집결과를 제공하는 방법이다. 분리법은 집괴법의 역순이라 할 수 있는데, 즉 전체의 객체를 하나의 군집으로 간주함을 시작으로 유사성이 떨어지는 객체들을 분리시켜 다른 군집으로 만들어 나가는 과정을 각 객체가 하나의 군집이 될 때까지 반복한 후 어떤 규칙에 의하여 최종적인 군집결과를 제공하는 것이다.

3. 모형설계

의사결정 테이블은 표 1과 같은 형태로 표현된다. 열은 n개의 속성들과 결정 변수(클래스 속성)로 구성되며, 행은 m개의 데이터 객체들로 이루어진다.

표 1. 의사결정 테이블

Tab. 1. Decision making Table

	속성 1	속성 2	…	속성 n	결정 변수
데이터 객체 1	d 11	d 12	…	d 1n	d 1
데이터 객체 2	d 21	d 22	…	d 2n	d 2
…	…	…	…	…	…
데이터 객체 m	d m1	d m2	…	d mn	d m

표 2의 예제 데이터 집합은 5개의 튜플들로 구성되어 있고, 3개의 속성집합 A1, A2, A3을 가지고 있다. A1 속성에는 a, b, c라는 3가지 속성 값을 가지고 있고, A2속성에는 m, f라는 2가지 속성 값을 가지고 있다. 또한 이 데이터 셋에는 결정변수가 주어져 있지 않다.

표 2. 예제 데이터

Tab. 2. An Example Data

	A1	A2	A3
1	b	f	z
2	a	m	z
3	c	f	y
4	b	m	z
5	c	f	z

본 논문에서는 표 2의 데이터 셋 처럼 결정 변수가 없는 데이터들을 입력 데이터로 사용하여 규칙들을 자동적으로 생성한다. 실험은 그림 2과 같은 단계로 이루어진다.

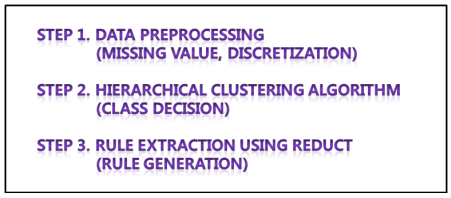

그림 2. 실험과정

Fig. 2. Experimental Process

본 논문에서는 계층적 클러스터링 알고리즘을 사용한다. n개의 데이터들을 클러스터링 하는 문제를 생각해 보자, 처음에는 n×(n-1)/2개의 클러스터간 합병을 고려할 수 잇는데, 이중에서 합병을 했을 경우 가장 높은 평가함수 값을 주는 두 개의 클러스터를 합병한다. 1번째 합병 후에는 (n-1)×(n-2)/2개의 클러서터간 합병을 고려하며, 이중에서 가장 높은 평가함수 값을 주는 두 개의 클러스터를 합병한다. 최종적으로는 주어진 개수의 클러스터가 남을 때까지 위의 과정을 반복한다.

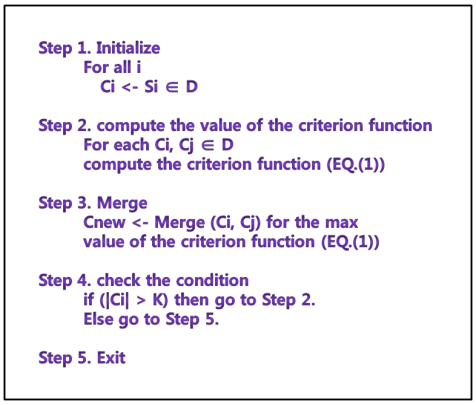

그림 3. 계층적 클러스터링 알고리즘
Fig. 3. Hierachical Clustering Algorithm

Step 1은 초기화 단계로서 데이터베이스 D를 액세스하여 각각의 데이터를 하나의 클러스터로 설정한다. Step 2는 현재 n개의 클러스터가 있다고 하면, n×(n-1)/2개의 평가함수 값을 계산한다. Step 3은 합병 단계로서, Step 2에서 계산한 평가함수 값들 중 가장 큰 값을 주는 두 개의 클러스터를 합병한다. Step 4는 조건 검사 단계로서 클러스터의 개수가 지정된 클러스터 개수보다 크면 Step 2로 간다. 그렇지 않으면 Step 5로 간다. 마지막으로 Step 5는 종료 단계로서 알고리즘을 끝낸다. 데이터 전처리 과정을 거친 데이터들은 결정 변수 값들이 없는 데이터들로 이루어져 있다. 본 논문에서는 이들 데이터들을 계층적 클러스터링 알고리즘을 적용하여 k개의 그룹으로 클러스터링 한 후, 이결과를 결정 변수의 값으로 사용한다.

4. 실험

4.1 실험데이터

실험에 이용된 데이터는 두 가지로 UCI에서 제공하는 Soybean Dataset과 Zoo Dataset을 이용하였으며, 그림 3 Soybean Dataset은 36개의 속성과 47개의 데이터로 이우어져 있는 Dataset이며, 그림 4 Zoo Dataset은 18개의 속성과 101개의 데이터로 이루어진 DataSet이다.

date	plant-stand	precip	temp	hail	crop-hist	area-damaged	severity	seed-숫
august	normal	gt-norm	norm	no	same-lst-yr	scattered	pot-severe	none
september	normal	gt-norm	norm	yes	same-lst-sev-yrs	low-areas	pot-severe	fungicide
july	normal	gt-norm	norm	yes	same-lst-two-yrs	scattered	severe	fungicide
october	normal	gt-norm	norm	yes	same-lst-yr	low-areas	pot-severe	none
august	normal	gt-norm	norm	yes	same-lst-sev-yrs	scattered	severe	none
september	normal	gt-norm	norm	yes	same-lst-two-yrs	scattered	pot-severe	fungicide
july	normal	gt-norm	norm	yes	same-lst-two-yrs	low-areas	pot-severe	none
july	normal	gt-norm	norm	yes	same-lst-yr	scattered	severe	fungicide
october	normal	gt-norm	norm	yes	same-lst-sev-yrs	scattered	pot-severe	fungicide
october	normal	gt-norm	norm	yes	same-lst-yr	scattered	pot-severe	none
october	normal	it-norm	gt-norm	no	diff-lst-year	upper-areas	pot-severe	none
august	normal	it-norm	norm	yes	same-lst-two-yrs	whole-field	pot-severe	fungicide
september	normal	it-norm	gt-norm	yes	same-lst-sev-yrs	upper-areas	pot-severe	none
october	normal	it-norm	norm	no	same-lst-sev-yrs	whole-field	pot-severe	fungicide
july	normal	it-norm	gt-norm	no	diff-lst-year	upper-areas	pot-severe	none
august	normal	it-norm	norm	no	same-lst-yr	whole-field	pot-severe	fungicide
july	normal	it-norm	norm	yes	same-lst-yr	upper-areas	pot-severe	none

그림 4. 첫 번째 실험데이터셋
Fig. 4. First Experiimental Dataset

animal name	hair	feathers	eggs	milk	airborne	aquatic	predator	toothed
Unique	Boolean	Boolean	Boolean	Boolean	Boolean	Boolean	Boolean	Boolean
aardvark	no	yes	yes	no	yes	yes	no	no
antelope	no	yes	yes	no	yes	yes	yes	no
bass	yes	yes	no	yes	yes	no	no	no
bear	no	yes	yes	no	yes	yes	no	no
boar	no	yes	yes	no	yes	yes	no	no
buffalo	no	yes	yes	no	yes	yes	yes	no
calf	no	yes	yes	no	yes	yes	yes	no
carp	yes	yes	no	yes	yes	no	yes	no
catfish	yes	yes	no	yes	yes	no	no	no
cavy	no	yes	yes	no	yes	yes	yes	no
cheetah	no	yes	yes	no	yes	yes	no	no
chicken	yes	no	no	yes	no	yes	yes	yes
chub	yes	yes	no	yes	yes	no	no	no
clam	yes	yes	no	yes	yes	yes	no	yes
crab	yes	yes	no	yes	yes	no	no	yes

그림 5. 두 번째 실험데이터셋
Fig. 5. Second Experiimental Dataset

4.2 실험환경

Microsoft Windows 7 x64 기반으로 XLMiner-Instal lable3_2_8을 이용하였다. 실험데이터의 레코드는 평가판을 사용하여 제한된 양으로 실험하는 한계가 있었다.

4.3 실험결과

첫 번째 실험데이터인 Soybean 47개 Data와 두 번째 실험데이터인 Zoo 101개 Data를 10-fold cross validation 하였고, 훈련데이터는 Soybean 300(30×10)개, Zoo 900(90×10)개이고, 검증 데이터는 Soybean 500(50×10)개, Zoo 1100(110×10)개가 사용된다. 계층적 클러스터링(Hierarchical Clustering)을 위하여 두 Dataset의 각 속성값들을 실수형으로 바꿔 실험하였다.

Parameters/Options	
Draw dendrogram	Yes
Show cluster membership	Yes
# Clusters	4
Selected Similarity measure	Euclidean distance
Selected clustering method	Average group linkage

그림 6. 계층적 클러스터링 매개변수/옵션
Fig. 6. Optional Parameters in Hierachical Clustering Algorithm

5. 결론

데이터들의 양이 많아지고 분류하는 방법들이 각각의 업무에 따라 다양해지면서 그 수많은 데이터들에서 필요한 지식이나 정보들을 추출하는 패턴이 힘들어졌다. 그래서 데이터마이닝 툴인 weka와 xlminer에 대하여 사용법을 분석해보고 비교했다. 그리하여 수많은 데이터들 중에서 필요한 지식이나 정보를 추출하는 패턴을 쉽게 찾을 수 있게 됐다. 두 가지 툴 모두 좋지만 xlminer에 비해 weka가 시각효과가 뛰어나서 결과를 알아보기 쉽게 만들어졌다. 클러스터링을 사용해서 이메일들을 분류하면 사용자가 관리를 편하게 할 수 있다. weka에서 k-means를 적용하여 크기가 다른 데이터를 분석했는데 em 알고리즘도 한번 돌려봤을 때 속도가 k-means 알고리즘이 더 빨랐다. 그리고 애트리뷰트가 많아지고 그 애트리뷰트 안에서도 군집에 영향을 주는 비율도 서로 다르고 영향이 큰 애트리뷰트로 인해서 군집화가 진행 되는 것을 알게 되었다.

본 논문에서는 계층적 클러스터링을 통하여 Zoo Dataset과 Soybean Dataset을 군집화 하였고, 이를 통하여 대량의 데이터로부터 의미 있는 규칙들을 발견하였다. 또한 데이터를 효율적으로 이산화하기 위한 이산화 기법을 제시하였다. 데이터를 이산화 하여 더 직관적이고 이해하기 쉬운 패턴을 생성하도록 한다. 또한 데이터 수를 축소시킴으로써 마이닝 단계의 수행 복잡도를 줄인다. 실험 분석을 통해 제안된 기법이 전체 마이닝 프로세스의 효율성을 높이고 추출된 패턴의 해석성을 높이는 것을 보였다.

참 고 문 헌

〔국내 문헌〕

[1] 강주영 (2009), “데이타 축소와 군집화를 사용하는 시공간 데이타의 이산화 기법”, 정보과학회논문지 : 컴퓨팅의 실제 및 레터, 15(1).

[2] 신민식 (2008), “Gender Differences in Ventricular Arrhythmia Recurrence in Patients With Coronary Artery Disease and Implantable Cardioverter Defibrillators”, 대구대학교 대학원, 박사학위논문.

[3] 이선미, 박래웅 (2009), “임상에서의 데이터 마이닝 개념과 원칙”, 대한의료정보학회.

[4] 최병수 (2011), “데이터마이닝을 위한이산화알고리즘에 대한 비교연구”, 한국통계학회 논문집.

〔국외 문헌〕

[5] Galit, Shmueli, Nitn R. Patel, Peter C. Bruce (2010), “Data Mining for Business Intelligence : Concepts, Technigues, and Applications in Microsoft Office Excel with XLMiner”, Wiley.

[6] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank (2011), Classifier chains for multi-label classification. Machine Learning, 85(3), 333-359.

[7] Jiawei, Han, Micheline Kamber (2001), “Data Mining Concept and Techniques (2nd Edition)”, Morgan Kaufmann.

[8] Lan, H. Witten, Eibe Frank, Mark A. Hall (2010), “Data.Mining Practical Machine LearningTools andTechniques (Third Edition)”, Morgan.Kaufmann.

원 재 강 (Jae Kang Won)

강릉대학교, 경기대학교에서 각각 학사, 석사, 박사학위를 취득하였고 현재 을지대학교, 경기대학교에서 강의를 하고 있다. 연구분야로는 워크플로우이며 해저탐사 등의 분야에서 의사결정지원을 위한 다양한 결정요인을 마이닝기법을 통해 실험하고 연구하고 있다.

이 정 찬 (Jeong Chan Lee)

연성대학교 전자계산학과를 졸업하고 현재 한국정보화진흥원 창의인재부에서 근무하고 있으며, 관심분야로는 대형 네트워크망과 다수의 서버환경에서 정보보안을 연구하고 있다. 특히 트래픽 필터링을 위해 행위기반과 연관규칙을 이용한 패턴분석을 연구하고 적용을 실험하고 있다.

정 용 규 (Yong Gyu Jung)

서울대학교, 연세대학교, 경기대학교에서 각각 학사, 석사, 박사학위를 취득하였고, 현재 을지대학교 의료IT마케팅학과 교수로 재직 중이다. ISO, UN의 전자문서 분야 한국대표위원으로 활동하고 있으며, 의료정보, 전자무역, 해상물류, 금융전산에 Semantic Web, Process Modelling, ebXML 등의 표준기술의 적용에 관심이 많다.

이 영 호 (Young Ho Lee)

수원대학교에서 학사, 석사, 박사를 취득 및 수료하였다. 현재 스마트주소위원회(COCif)에서 한글주소체계 기반 다이얼 서비스 분야 국가표준을 개발하고 있다. 주요 관심분야로는 분산 네트워크, RFID 기반의 상품 관리 및 검색, 휴대전화를 위한 하이브리드형 한글 입력 방식 및 키패드 배열 설계 등에 관심이 많다.

Performance Comparison of Clustering using Discritization Algorithm

Jae Kang Won* · Jeong Chan Lee** · Yong Gyu Jung*** · Young Ho Lee****

ABSTRACT

Datamining from the large data in the form of various techniques for obtaining information have been developed. In recent years one of the most sought areas of pattern recognition and machine learning method is created with most of existing learning algorithms based on categorical attributes to a rule or decision model. However, the real-world data, it may consist of numeric attributes in many cases. In addition it contains attributes with numerical values to the normal categorical attribute. In this case, therefore, it is required processes in order to use the data to learn an appropriate value for the type attribute. In this paper, the domain of the numeric attributes are divided into several segments using learning algorithm techniques of discritization. It is described Clustering with other data mining techniques. Large amount of first cluster with characteristics is similar records from the database into smaller groups that split multiple given finite patterns in the pattern space. It is close to each other of a set of patterns that together make up a bunch. Among the set without specifying a particular category in a given data by extracting a pattern. It will be described similar grouping of data clustering technique to classify the data.

Keywords: Pattern Recognition, Machine Learning, Discretization, Discriminant Analysis, Hierarchical Clustering

* Professor, Kyonggi University, 06240604@hanmail.net
** Researcher, National Information Society Agency, jclee@nia.or.kr
*** Corresponding Author, Professor, Eulji University, ygjung@eulji.ac.kr
**** Professor, Suwon University, yhleepr@gmail.com