http://dx.doi.org/10.14400/JDPM.2013.11.11.343

데이터 마이닝을 위한 제어규칙의 생성

박인규
중부대학교 컴퓨터학과

The Generation of Control Rules for Data Mining

In-Kyoo Park
Dept. of Computer Science Joongbu University

요 약 러프집합에서는 동치류와 근사공간의 개념을 이용하여 데이터 마이닝 분야에서 중복되는 정보로부터 특징점을 효율적으로 추출하여 최적화된 제어규칙을 유도할 수 있다. 이러한 추출과정에서 가장 중요하게 고려되어져야 할 부분은 많은 속성에 대한 감축이다. 본 논문에서는 속성간의 관계에서 러프엔트로피를 이용하여 가장 신뢰도가 우수한 속성을 구할 수 있는 정보이론적인 척도를 제시한다. 제안된 방법은 러프엔트로피를 기반으로 불필요한 속성을 제거함으로써 유용한 리덕트를 생성하고 이들에 대한 코어를 형성한다. 결과적으로 원시정보의 내용은 변하지 않으면서 지식감축을 통하여 간소화된 제어규칙을 구축할 수 있음을 보인다.

주제어 : 러프집합, 엔트로피, 데이터마이닝, 리덕트, 코어

Abstract Rough set theory comes to derive optimal rules through the effective selection of features from the redundancy of lots of information in data mining using the concept of equivalence relation and approximation space in rough set. The reduction of attributes is one of the most important parts in its applications of rough set. This paper purports to define a information-theoretic measure for determining the most important attribute within the association of attributes using rough entropy. The proposed method generates the effective reduct set and formulates the core of the attribute set through the elimination of the redundant attributes. Subsequently, the control rules are generated with a subset of feature which retain the accuracy of the original features through the reduction.

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

1. 서론

정보의 수요가 날로 커져감에 따라서 다양한 분야에서 처리되어야 할 정보량의 증가와 다양성을 초래하였고, 인터넷을 통한 이러한 정보에 대한 정열과 검색은 잠정적인 중요성을 가지게 되었다. 이러한 이유로 인하여 정보량을 줄이고 관련이 있는 정보만을 채집하는 일이 매우 복잡하게 되었다. 이러한 문제에 대한 해결책으로 데이터베이스에 존재하는 많은 정보로부터 관련된 정보를 채집할 수 있는 유용한 도구로 러프집합이론이 좋은 결실을 맺고 있다[7,10]. 러프집합 이론의 특징으로 첫째, 데이터에 숨겨져 있는 사실(facts)을 해석할 수 있다. 둘

Received 1 November 2013, Revised 20 November 2013
Accepted 20 November 2013
Corresponding Author: In-Kyoo Park(Joongbu Univ.)
Email: fip2441g@gmail.com

ISSN: 1738-1916

째, 데이터에 대한 추가적인 정보가 필요 없다. 셋째, 최소한의 지식으로 전체 데이터를 나타낼 수 있다. 샤논에 의하여 개발된 정보이론은 다양한 분야에서 정보의 량을 측정하는 도구로 실제로 널리 활용되고 있다. 엔트로피를 이용하여 어떤 정보를 표현하는 방식은 규칙베이스 시스템에서 데이터베이스, 퍼지 데이터베이스 쿼리(query), 데이터 할당과 분류와 같은 분야에서 활용되고 있다[4,6]. 또한 이를 토대로 어떠한 의사결정에 필요한 정보를 정형화함으로써 유용한 정보를 도출할 수 있고, 어떤 객체나 상태, 의견 과정들 사이의 차이점을 추적할 수 있는 유용한 도구이다[1,2,3,8].

그러나 정보시스템은 객체의 속성 값이 크리스프(crisp)하거나 실수 값을 가지는 경우와 속성 값의 알갱이성(granularity)으로 인한 부정확함 또는 속성 값의 중복과 같이 식별 불가능성과 애매성으로 인한 정보의 불확실성(uncertainty)을 내포하기 마련이다. 일반적으로 러프집합에서는 근사영역을 통한 정확도(accuracy)와 거침(roughness)을 이용하여 불확실성을 해결한다. 따라서 동치류의 알갱이성이 거칠어지면 특정한 객체에 대한 정보의 양이 줄어들게 된다. 그러나 하한근사 영역의 객체들에 대한 식별불가능성 관계에 있는 알갱이 성으로 인한 불확실성에 대한 문제해결이 필요하다. 러프집합에서는 근사공간을 이용한 정확도와 거침을 이용하여 데이터의 불완전성을 처리하였으나, 중복되거나 누락된 데이터의 처리등 모든 불완전 정보를 다루기에는 부족하다. 본 논문에서는 기존의 정보이론에서 이용되는 엔트로피개념을 이용하여 식별불가능성으로 인한 불완전성을 해결하기 위해 러프집합의 불확실성에 대한 새로운 척도를 제시한다.

2. 러프집합 이론

2.1 러프집합

관심 대상인 객체의 유한집합 U≠∅(전체집합)을 가정하자. 전체집합의 임의의 부분집합 X⊆U를 U내의 범주(category)라 하고, U내의 범주들의 임의의 집합을 U에 대한 지식(knowledge)라고 하자. 지식기반 시스템은 전체집합과 R이 U의 동치관계들의 집합일 때 식(1)과 같이 정의할 수 있다.

$$K=(U,R), X\subseteq U, P\subseteq R \tag{1}$$

속성들의 모든 부분집합 P⊆R와 임의의 원소 x_i, x_j ∈U라면 P의 식별불가능관계(indiscernibility)인 이진 관계 IND(P)는 식(2)와 같이 정의한다.

$$IND(P)=\{(x_i,x_j)\in U\times U:\ \forall p\in P\ \ p(x_i)=p(x_j)\} \tag{2}$$

여기서 x_i, x_j는 정보시스템 K에서 속성 P의 집합에 의하여 식별불능이다. 또한 P(x)는 객체 x에 할당된 속성 P의 값으로서, IND(P)는 모든 P⊆A에 대하여 U에서 식별 불가능한 동치관계(equivalence relation)가 되며 식(3)과 같은 관계가 된다.

$$IND(P)=\cap_{p\in P} IND(p)$$
$$IND(K)=\{IND(P):\ \varnothing\neq P\subseteq R\} \tag{3}$$

IND(K)는 K의 모든 원소관계를 포함하고 동치관계끼리의 교집합에 대하여 닫혀있는 동치관계들의 최소의 집합이다. P⊆R이고 P≠∅이면 ∩P(P에 해당되는 모든 동치관계들의 교집합)도 역시 동치관계가 되며, 이를 IND(P)로 나타내고 x의 동치 클래스는 식(4)와 같다.

$$[x]_{IND(P)}=\cap_{R\in P}[x]_R \tag{4}$$

러프집합의 이론에 의하여 $\underline{P}X$와 $\overline{P}X$를 각각 X의 P-하한근사(lower approximation)와 P-상한근사(upper approximation)라고 하며 식(5)와 같이 정의할 수 있다.

$$\underline{P}X=\cup\{x\in U|P(X):\ P(X)\subseteq X\}$$
$$\overline{P}X=\cup\{x\in U|P(X):\ P(X)\cap X\neq\varnothing \tag{5}$$

하한근사는 집합 X에 확실히 속하는 모든 원소들로 구성되고 상한근사는 집합X와 교집합이 공집합이 아닌 원소들로서 X에 속할 가능성이 있는 원소들로 구성된다. 그리고 하한근사와 상한근사와의 차는 경계영역으로서 어느 쪽에도 속하지 않는 애매한 영역으로 식(6)과 같이 정의할 수 있다.

$$BND(X) = \overline{P}X - \underline{P}X \quad (6)$$

집합 X에 대한 정확성 척도는 식(7)과 같다. 집합 X에 대한 지식의 불완전성의 정도를 나타내는 불완전성 척도는 식(8)과 같다.

$$\alpha_R(X) = \frac{card\ \underline{R}X}{card\ \overline{R}X}, X \neq \varnothing \quad (7)$$

$$\rho_R(X) = 1 - \alpha_R(X) \quad (8)$$

식(6)은 경계영역에서 발생되는 불완전성을 나타내는 방법이지만 식별불가능 관계에 의한 불확실성을 완전히 처리하지 못한다.

2.1 리덕트와 코어

지식의 감축에 있어 가장 기본적인 역할을 하는 것은 리덕트(rediuct)와 코어(core)의 핵심적인 두 개념이다. 첫째는 리덕트(reduct)라는 개념으로써 속성집합 A에 대한 리덕트 B는 A의 최소집합으로 객체 X를 B의 기본범주로 속성집합 A와 동일하게 분류할 수 있다. 둘째는 코어(core)라는 개념으로써 A에 대한 B의 코어는 A에 가장 필수적인 부분으로 B의 기본범주로 객체 X를 분류하는데 결정적으로 관련이 있다.

일반적으로 임의의 의사 결정표에는 많은 리덕트가 존재하게 되고 경험적으로 볼 때 규칙과 리덕트와의 연관성이 있는 것을 볼 수 있다. 따라서 결정표에 대한 리덕트를 발견하는 많은 방법이 개발되어 있다. 대표적으로 식별행열(discernibility matrix)을 들 수 있는데, 이를 이용하여 결정표에 존재하는 리덕트를 발견할 수 있다. IND(B)-IND(B-a)인 관계가 성립하면 B에 속하는 속성 A는 B에서 불필요하다고 할 수 있다. 만약에 관계가 성립하지 않으면 속성 A는 B에서 필수 불가결한 속성이 된다. B에 속하는 모든 속성 A가 B에서 필수 불가결하면 B는 직교(orthogonal)이다. B'가 직교 즉, 상호 독립적(independent) 이고 Ind(B)=Ind(B')이면 B'⊆B는 B의 리덕트이다. B에 존재하는 모든 필수 불가결한 속성들의 집합은 B의 코어(core)가 되고 식(9)와 같이 CORE(B)라고 나타낸다.

$$CORE(B) = \cap \cap_{R \in RED(B)} R \quad (9)$$

여기서 RED(B)는 B의 모든 리덕트의 집합으로서 속성들에 대한 최소한의 집합으로 모든 객체들을 속성들에 의하여 분별할 수 있다.

3. 러프 엔트로피

러프집합을 활용하여 데이터베이스에서 지식을 추출하는데 활용되고 있는 분야에는 특징점 추출, 데이터의 감축, 데이터의 이산화(discretization)등과 같이 많은 응용범위를 가지고 있다. 러프집합을 이용하여 임의의 데이터집합에서 불필요한 특징을 제거함으로써 전체적으로 집합의 차원을 줄여줌으로써 문제를 보다 단순화하여 필요한 특징만 추출할 수 있다. 데이터의 양이 날로 증가함으로 인하여 각각의 제어대상에 대하여 어떤 속성이 필요하고 그렇지 않은지를 정확하게 측정하는 것은 매우 어려운 일이다. 이러한 이유로 데이터의 양을 먼저 감소시키는 것이 중요하며 이를 위해서는 데이터에 대하여 중요한 특징을 추출하기 위한 동정(identification)이 선행되어야 한다. 이로 인하여 중요한 속성들을 추출함으로써 잡음과 같은 불필요한 데이터를 제거하여 적합한 데이터에 대한 제어규칙을 설정할 수 있다[9]. 따라서 러프집합 이론의 리덕트와 같은 동정을 통하여 제어 대상에 관계하는 데이터에서 추출한 특징은 전체 데이터에 대한 특징과 동일하게 된다. 결국 제어에 대한 규칙의 수와 제어시간을 최적화 할 수 있다.

러프집합에서 다루어지고 있는 불확실성은 식별 불가능한 관계에서 발생하는 불확실성이다. 즉, 동치류들을 구성하는 객체들을 구별할 수 없다는 것이다. 결국 하한근사와 상한근사라는 근사화를 통하여 이러한 불확실성을 모델링을 할 수 있다. 앞 절에서도 언급한 바와 같이 러프집합의 애매함을 나타내는 정확성(accuracy)과 거침(roughness)라는 두 가지의 척도가 있다. 전자는 상한근사와 하한근사의 비율을 나타낸다. 또한 거침의 척도는 러프집합이 가지고 있는 정보가 완전하지 못한 정도를 나타낸다. 그러나 이러한 척도는 완전한 거침을 처리하지는 못한다. 예를 들어 식(10)에서 러프집합의 하한근사와 상한근사는 각각 식((11)과 식(12)로 근사화 할 수 있

다[5]. 이와 같은 근사를 통하여 식(13), 식(14), 식(15)와 같은 분할을 얻을 수 있다.

$$X=\{A_{11},A_{12},A_{21},A_{22},B_{11},C_1\} \quad (10)$$

$$\underline{R}X=\{A_{11},A_{12},A_{21},A_{22}\} \quad (11)$$

$$\overline{R}X=\{A_{11},A_{12},A_{21},A_{22},B_{11},B_{12},B_{13},C_1,C_2\} \quad (12)$$

$$A_1=\{A_{11},A_{12},A_{21},A_{22},B_{11},B_{12},B_{13},C_1,C_2\} \quad (13)$$

$$A_2=\{A_{11},A_{12},A_{21},A_{22},B_{11},B_{12},B_{13},C_1,C_2\} \quad (14)$$

$$A_3=\{A_{11},A_{12},A_{21},A_{22},B_{11},B_{12},B_{13},C_1,C_2\} \quad (15)$$

위의 분할은 임의의 X에 대하여 동일한 하한근사와 상한근사를 가지기 때문에 동일한 정확도를 가지고 있다. 그러나 A_1이 가장 불확실성이 높고 A_3가 가장 낮다는 것을 알 수 있다. 따라서 보다 효과적인 불확실성에 대한척도가 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(16)과 같다.

$$h(x)=-\log P(x) \quad (16)$$

결국 엔트로피는 랜덤변수 X가 가질 수 있는 모든 값(사건)에 대해 정보량을 평균한 것이다. 본 논문에서는 기존의 정보이론에서 사용되는 엔트로피이론을 변형하여 많은 접근 방법들이 제시되어 왔다. 어떤 통계적 앙상블(ensemble)을 각 미시적 상태 i의 확률을 P_i로 정의할 경우에 N개로 구성된 앙상블의 엔트로피 E는 식(17)과 같이 정의할 수 있다.

$$E=-k\sum_{i}^{n}P_i\log_2 P_i \quad (17)$$

본 논문에서는 러프집합에서의 지식에 존재하는 속성들의 중복성에 관한 불확실성에 대한 문제를 설정하여 임의의 러프집합에서 러프 엔트로피를 식(18)과 같이 정의한다. 즉, 동치류 X와 Y의Y에 대한 중복성의 비율과 U에 대한 Y의 비율의 곱이다.

$$H(X_i|Y)=-K\log_2\sum_{i=1}^{n}\frac{|X_i\ \cap Y|}{|Y|} \quad (18)$$
$$K=card(Y)/card(U)$$

4. 적용사례

4.1 문제의 설정

어떤 기업체의 신입사원 채용 여부를 판단하는 정보시스템을 <Table 1>과 같이 구성하였다. 여기서 구분은 신입사원 지원 대상자이며, 조건 속성에 해당하는 평가조건은 대학성적(x), 창의성(y), 사회성(z) 및 인성(w)의 3개 영역으로 구분하였다. 그리고 결정속성에 해당하는 판단결과(D)는 평가 조건에 따라 결정되었다. <Table 1>에서 객체 9와 10은 동일한 조건 속성에 대하여서로 다른 결론 속성을 가지므로 <Table 1>에 존재하는 데이터는 일관성을 가지고 있지 않다고 할 수 있다.

〈Table 1〉 Decision table of (x,t,z,w,D)

index	condition				decision
	x	y	z	w	D
1	A	P	3	A	1
2	A	P	1	S	1
3	P	P	1	A	1
4	P	R	3	A	2
5	A	R	2	A	2
6	P	R	3	P	3
7	S	R	3	P	3
8	S	N	3	P	3
9	S	N	2	S	2
10	S	N	2	S	1

〈Table 2〉 Reduced decision Table of 〈Table 1〉

index	condition				decision
	x	y	z	w	D
1	A	P	3	A	1
2	A	P	1	S	1
3	P	P	1	A	1
4	P	R	3	A	2
5	A	R	2	A	2
6	P	R	3	P	3
7	S	R	3	P	3
8	S	N	3	P	3

이 경우에 결론부의 속성에 부합하는 하한근사와 상한근사를 통하여 이러한 비일관성의 데이터를 처리할 수

있다. <Table 1>에서 주어진 데이터의 기본 범주는 {1}, {2}, {3}, {4}, {5}, {6}, {7}, {8}, {9}, {9, 10}로 구성된다. <Table 1>에 존재하는 세 가지의 개념 즉, {1,2,3,10}, {4,5,9}, {6,7,8}에 대하여 10개의 객체에 대한 하한근사는 {1, 2, 3, 4, 5, 6, 7, 8}이고 상한근사는 {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}이다. 따라서 경계지역은 {9,10}이 된다. 따라서 객체 9와 10을 제거한 데이터는 <Table 2>와 같이 구성된다.

4.2 제어규칙의 발생

지식 시스템은 객체에 대한 성질들의 집합으로 볼 수 있기 때문에 속성-값(attribute-value)표로 구성되는 지식 시스템을 구축할 수 있다. 또한 이러한 집합을 이용하여 집합을 구성하는 데이터와 데이터 감축에 대한 종속성을 전술한 식을 이용하여 구할 수 있다. 결국 집합을 이용하여 어떤 식의 정규형(canonical form) 표현을 나타낼 수 있고, 식의 진위(true/false)를 가리기 위하여식별 불가능성을 도입할 수 있다. 이비 식별 불가능성을 이용한 방법은 데이터의 감축과 분석을 위한 알고리즘으로 사용되고 있다. 이러한 데이터 표는 의사결정 논리라는 하나의 모델로 볼 수 있고 또한 이를 이용하여 지식 시스템에서 이용 가능한 데이터로부터 결론을 도출하여 우리가 원하는 추론을 수행하기 위하여 리덕트와 코어를 이용하여 제어규칙을 발생할 수 있다.

〈Table 3〉 Rough entropy of attribute x

rule of attribute x	rough entropy
(x=A)⇒(D=1)	-3/8ln(2/3)=0.152
(x=P)⇒(D=1)	-3/8ln(1/3)=0.412
(x=P)⇒(D=2)	-2/8ln(1/2)=0.173
(x=A)⇒(D=2)	-2/8ln(1/2)=0.173
(x=P)⇒(D=3)	-3/8ln(1/3)=0.412
(x=S)⇒(D=3)	-3/8ln(2/3)=0.152

〈Table 4〉 Rough entropy of attribute y

rule of attribute x	rough entropy
(y=P)⇒(D=1)	-3/8ln(1/3)=0.412
(y=R)⇒(D=2)	-2/8ln(2/2)=0.575
(y=R)⇒(D=3)	-3/8ln(2/3)=0.152
(y=N)⇒(D=3)	-3/8ln(1/3)=0.142

〈Table 5〉 Rough entropy of attribute z

rule of attribute x	rough entropy
(z=3)⇒(D=1)	-3/8ln(1/3)=0.142
(z=1)⇒(D=1)	-2/8ln(2/2)=0.152
(z=3)⇒(D=2)	-2/8ln(1/2)=0.173
(z=2)⇒(D=2)	-2/8ln(1/2)=0.173
(z=3)⇒(D=3)	-3/8ln(3/3)=0.86

근사화를 통하여 확보한 일관성은 데이터에 대하여 제어규칙을 발생한다. 먼저 <Table 2>의 데이터에 대한 리덕트는{x,z,w}, {x,y,w}, {y,z,w}이고, 코어는 w가 되어 가장 중요한 속성이 된다. 따라서 w를 제외한 다른 속성에 대하여 속성의 중요도(confidence factor)를 계산하여 데이터를 보다 감축할 수 있다. 조건부와 결정부간의 의사결정규칙 $x \Rightarrow D$의 중요도는 x를 포함하는 객체의 수에 대하여 $x \cup D$를 포함하는 객체의 수에 대한 러프 엔트로피로 정의한다. 따라서 <Table2>에서 w를 제외한 x, y와 z속성에 대한 러프 엔트로피는 식(18)에 의해 각각 <Table 3>, <Table 4>, <Table 5>와 같이 계산할 수 있다.

〈Table 6〉 Reduced decision Table of 〈Table 3〉

index	condition		decision
	y	w	D
1	P	A	1
2	P	S	1
3	P	A	1
4	R	A	2
5	R	A	2
6	R	P	3
7	R	P	3
8	N	P	3

〈Table 7〉 Reduced decision Table of 〈Table 3〉

index	condition		decision
	y	w	D
1	P	A	1
2	P	S	1
4	R	A	2
6	R	P	3
8	N	P	3

결국 x, y, z속성별 러프 엔트로피는 각각 0.355, 0.255와 0.3인 곳을 알 수 있다. 따라서 y속성이 가장 적은 값

을 가짐으로써 가장 중요한 속성이라는 것을 알 수 있다.

〈Table 8〉 Core of 〈Table 3〉

index	condition		decision
	y	w	D
1	P	-	1
2	P	-	1
4	R	A	2
6	-	P	3
8	-	P	3

결국 원래의 속성집합 {x, y, z, w}에서 {y, w}속성으로 감축이 이루어졌다. 이에 대한 데이터는 <Table 6>과 같이 구축되어진다. <Table 6>에서 속성 값의 중복을 제거하면 <Table 7>을 얻을 수 있다. <Table 7>에서 코어를 이용하여 규칙의 일관성을 유지할 수 있다. w속성의 A를 제거하면 데이터의 일관성을 확보할 수 있다. 또한 y속성의 R을 제거하면 역시 일관성을 유지 할 수 있다. 결국 <Table 8>과 같이 객체에 대하여 코어를 확보하여 최적의 감축을 수행할 수 있다. 결국 <Table 9>와 같이 러프 엔트로피를 이용하여 데이터의 감축을 수행하여 제어규칙은 IF y ⇒ P THEN D ⇒ 1, IF y ⇒ R 와 w ⇒ A THEN D ⇒ 2 와 IF w ⇒ P THEN D ⇒ 1이 된다.

〈Table 9〉 Final control rules of 〈Table 3〉

index	condition		decision
	y	w	D
1	P	-	1
4	R	A	2
6	-	P	3

5. 결론

본 논문에서는 조건부의 속성과 결론부의 속성간의 연관관계에서 연관정도를 결정하는 조건부 엔트로피를 정의하였다. 제안된 러프 엔트로피를 이용하여 임의의 데이터에 존재하는 불필요한 속성을 제거함으로써 지식을 효율적으로 감출할 수 있었다. 제안된 방법은 지식의 감축, 데이터 마이닝과 다른 음성인식과 같은 속성감축과 같은 속성 감축과 특징점 추출과 같은 분야에 적용될 수 있다. 앞으로 제안된 방법을 보다 방대한 양의 데이터 베이스에 적용하여 기존의 방법들과 비교우위를 논할 필요가 있을 것으로 사료되어 진다.

참 고 문 헌

[1] 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.

[2] Hand, D.J., Blunt, G., Kelly, M.G. & Adams, N.M., "Data mining for fun and profit, Statistical Science, vol. 15, pp. 111-131, 2000

[3] Hand, D.J., Mannila, H., & Smyth, P. "Principles of Data Mining", Cambridge, MA:MIT Press, 2001

[4] Han, Jiawei, Kamber, Micheline, "Data Mining: Concepts and Techniques", San Franciso CA, USA, Morgan, Kaufmann, Publishers, 2001.

[5] Pawlak, Z., "Rough sets", International Journal of Information Sciences, 11, pp. 341-356, 1982

[6] Pawlak, Z., "Using Variable Precision Rough Set for Selection and Classification of Biological Knowledge Integrated in DNA Gene Expression", Jouranl of Integrative Bioinformatics, Vol. 9, No. 3, pp.1-17, 2012

[7] Pal S.K., Skowron, "Rough Fuzzy Hybridization: A new trend in decision making", Springer Verlag, Berlin, 1999

[8] R. Vashist, M.L. Garg, "Rule Generation based on Reduct and Core: A Rough Set Approach", International Journal of Computer Applications, Vol. 29, No. 9, pp. 0975-8887, Sept. 2011

[9] Ramakrishnan., Naren and Grama, Ananth Y,, "Data Mining: From Serendipity to Science", IEEE Computer August Vol. 34-37, 1999

[10] Williams, Grahm J. and Simoff, Simeon J. "Data Mining Theory, Methodology, Techniques and Applications(Lecture Notes in Computer Science/Lecture Notes in Artificial Intelligence)", Springer, 2007

박 인 규(Park, In Kyoo)

- 1985년 2월 : 연세대학교 공학석사
- 1997년 2월 : 원광대학교 공학박사
- 현재 : 중부대학교 컴퓨터학과 교수
- 관심분야 : 소프트컴퓨팅, 데이터마이닝
- E-Mail : fip2441g@gmail.com