논문 2009-3-5

효율적인 베이지안망 학습을 위한 엔트로피 적용

Efficient Learning of Bayesian Networks using Entropy

허고은*, 정용규**

Go-Eun Heo, Yong-Gyu Jung

요 약 베이지안망은 불확실한 상황 하에서 영역지식을 표현하고 예측하기 위한 좋은 도구로 알려져 있다. 그러나 변수가 많아졌을 때 학습이 어렵고 시간의 요구량이 늘어나게 되어 효율적이고 신뢰도 높은 탐색에 문제가 있다. 이를 해결하기 위해서 노드의 순서를 정하여 효율적인 구조학습이 가능하도록 한다. 본 논문에서는 각 상황에 따른 확률의 엔트로피를 계산하여 다양한 변수간의 관계나 상호의존적인 상황에서도 오차를 줄이고 신뢰도를 높일 수 있는 효과적인 분류학습모델을 제시한다. 베이지안망 학습 방법 중 일반적으로 널리 알려져 있는 K2알고리즘에서 각 노드의 엔트로피 수치를 계산하여 엔트로피가 낮은 노드의 순서를 결정하여 결과적으로 빠른 시간 안에 최적화된 베이지안망의 모델을 구성하는 효율적인 학습모델을 제시한다.

Abstract Bayesian networks are known as the best tools to express and predict the domain knowledge with uncertain environments. However, bayesian learning could be too difficult to do effective and reliable searching. To solve the problems of overtime demand, the nodes should be arranged orderly, so that effective structural learning can be possible. This paper suggests the classification learning model to reduce the errors in the independent condition, in which a lot of variables exist and data can increase the reliability by calculating the each entropy of probabilities depending on each circumstances. Also efficient learning models are suggested to decide the order of nodes, that has lowest entropy by calculating the numerical values of entropy of each node in K2 algorithm. Consequently the model of the most suitably settled Bayesian networks could be constructed as quickly as possible.

Key Words : Bayesian, entropy, K2-algorithm, probabilities

Ⅰ. 서 론

베이지안이란 통계학에서 데이터를 분석하는 하나의 방법론이라고 할 수 있다. 데이터를 분석할 때 관측된 데이터만 가지고 분석을 하게 되면 정확도가 떨어질 뿐만 아니라 여러 한계점이 드러나게 된다. 이에 반면 베이지안을 기반으로 데이터를 분석하면 과거에 이미 알려진 사전확률을 고려함과 동시에 각 상황에 따라 분석자의 주관적인 생각까지 함께 분석을 하게 되므로 기존의 분석방법에 비해 훨씬 더 정확한 결론을 얻을 수 있게 된다.[1][2] 하지만 기존의 베이지안망을 기초로 한 모델은 변수가 많아졌을 때 학습이 어렵고 시간의 요구량이 늘어나게 되어 효율적이고 신뢰도 높은 탐색을 하는 데에 문제점이 발생한다. 이러한 문제점을 해결하기 위해 베이지안망의 학습 알고리즘을 통해 효율적인 노드탐색을 위한 방법은 제시되어 왔지만 여전히 노드 순서 추정화에 대한 노드 탐색 효율성의 문제는 남아있다. 본 논문에서는 각 상황에 따른 확률의 엔트로피를 계산하여 다양한 변수간의 관계나 상호의존적인 상황에서도 오차를 줄

*정회원, 을지대학교 의료전산학전공
**종신회원, 을지대학교 의료전산학전공(교신저자)
접수일자 2009.05.24, 수정완료.2009.06.03

이고 신뢰도를 높일 수 있는 효과적인 분류학습모델을 제시하고자 한다. 따라서 베이지안망 학습 방법 중 일반적으로 널리 알려져 있는 K2알고리즘에 각 노드에서의 엔트로피의 수치를 계산하여 엔트로피가 낮은 노드의 순서를 결정하여 결과적으로 빠른 시간 안에 최적화된 베이지안망의 모델을 구성하게 된다.[3]

II. 관련 연구

2.1 베이지안망과 확률분포

베이지안망은 불확실한 상황 하에서 지식을 표현하고 예측하기 위한 도구이다. 그림 1과 같이 사건을 나타내는 확률변수인 노드(node)와 노드간의 인과관계를 나타내는 연결선(edge)로 이루어져 있는 방향성 비순환 그래프(Directed Acyclic Graph) 이다.

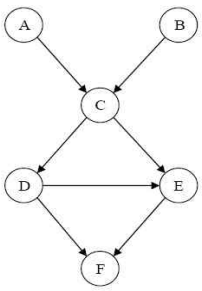

그림 1. 방향성 비순환 그래프
Fig. 1. Directed Acyclic Graph

그림 1에서 노드A와 노드B가 노드C의 부모노드이고, 노드C는 노드A와 B의 자식노드이다.[4] A→C와 같이 각 노드가 부모와 자식 관계에 있을 때, 조건부 확률관계가 성립된다.

불확실한 지식에 근거하여 결정을 내릴 때 조건부확률에 관한 이론들이 사용된다. 조건부 확률은 사건이 어떤 조건에 의해 제한되거나, 이전에 일어난 다른 사건에 의해 영향을 받게 될 때의 확률이다. 예를 들어 임의의 두 사건 X와 Y에 대해 이미 사건 X가 일어난 상황에서 Y가 일어날 확률은 $P(Y|X)$ 로 표현하고 P(X)≠0이라면 (1)과 같이 조건부 확률이 정의된다.[5]

$$P(Y|X) \equiv \frac{P(X \cap Y)}{P(X)} \quad (1)$$

베이지안망은 많은 변수들간의 확률관계를 비교적 축약된 형태로 표현하는데 유용한 모델로 확률적 추론, 예측, 의사 결정 등에 잘 적용될 수 있어 보다 신뢰도 높은 방안을 제시해 주는 효과적인 모델링 방법이다. 각 노드의 조건부 확률과 노드간의 연결선들로 이루어진 베이지안망 구조가 주어지면 변수들에 대한 결합확률분포(Joint Probability Distribution)를 다음 (2)와 같이 나타낼 수 있다. 여기서 Pa(x_i)는 x_i의 부모노드를 의미한다. 그림1에서 A→C의 그래프 구조를 가질 때 노드A는 노드C의 부모노드가 된다.

$$P(x_1, x_2, x_3, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid Pa(x_i)) \quad (2)$$

베이지안망이 주어지면 주어진 베이지안망을 이용하여 다양한 확률 예측 및 추론이 가능하다. 예를 들어 그림1에서 A와 B의 노드 값이 관찰되었다면 이 새로운 관찰 값에 근거하여 나머지 노드들이 일어날 확률들을 예측할 수 있다.[6]

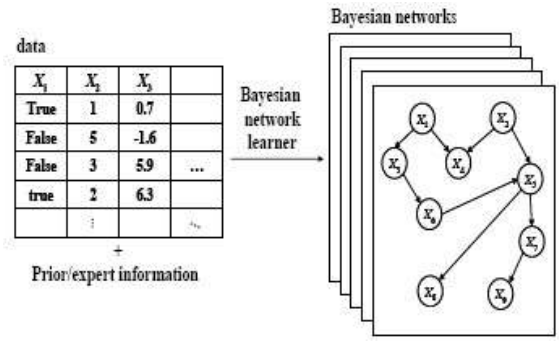

그림 2. 베이지안 네트워크의 학습
Fig. 2. Baysian Networks Learning

그림 2와 같이 베이지안 망의 학습은 기존의 데이터의 확률적인 접근을 통한 사전적인 확률을 고려하여 분석함으로써 효율적이고 추론을 가능하게 한다.

2.2 베이지안망 학습 알고리즘

기존의 베이지안망 학습 알고리즘은 크게 두 가지로 구분할 수 있다. 하나는 각 변수들 사이에 존재하는 조건부 독립성을 이용해서 베이지안망의 구조를 학습하는 알고리즘이다. 조건부 독립성의 검증에는 상호정보량(mutual information), x^2 test 등을 이용한 알고리즘이

있다. 두 번째 접근법은 베이지안망 학습 문제를 최적화의 관점에서 보는 것이다. 이 경우 베이지안망의 구조가 데이터에 적합한 정도를 나타내는 점수를 선정한 후 데이터에 가장 적합한 베이지안망의 구조를 탐색하게 된다. 베이지안망의 적합도를 나타내는 점수로는 MDL(Minimum Description Length)계열, BD(Bayesian Dirichlet) 계열 등이 있으며 데이터의 개수가 많아지면 두 점수는 점근적으로 같아진다. 점수가 선정되면 베이지안망 구조의 학습은 가능한 탐색 공간에서 점수가 가장 좋은 망 구조를 찾는 것이 된다. 그러나 n개의 노드를 가지는 베이지안망 구조 공간의 크기는 약 $n! \times 2^{n(n-1)/2}$ 이며 이 문제는 NP-hard임이 알려져 있다.[7] 의존성 분석 기반과 평가함수 기반의 방법들 모두는 문제점을 지닌다. 첫째, 노드의 순서가 요구된다는 것이다. 기존의 많은 연구들이 노드의 순서를 가정하고 실험을 하였지만 실제 상황에서 입증된 경우는 거의 없다. 둘째, 효율성이 낮다. 비록 노드의 순서를 가정하지 않고 학습할 수 있는 방법이 제시되었지만 노드의 수가 많아지는 대용량의 데이터에는 계산의 복잡도로 인하여 효율적인 학습에 문제가 있다. 이러한 문제점을 극복하기 위해 노드의 순서를 미리 추정하여 이를 바탕으로 성능이 좋은 K2알고리즘 등을 이용하여 베이지안망 구조를 학습한다. 하지만 베이지안망 구조에서의 노드의 순서가 유일하지 않기 때문에 효율적인 노드 탐색을 위한 노드의 순서화가 필요하다.

2.3 K2 학습알고리즘

Cooper와 Herskovits에 의해 제안된 K2알고리즘은 베이지안망 학습 알고리즘으로 가장 잘 알려진 방법이다. 이는 후에 연구된 많은 알고리즘의 기초가 되었다. 이 알고리즘은 베이지안망 G의 조인트 확률과 데이터 D를 이용해 평가한 베이지안 점수 메트릭을 사용하는데 이를 K2 메트릭이라 하며, 가장 잘 알려진 베이지안 망 평가함수이다. K2 메트릭은 다음 (3)으로 나타낸다.

$$P(G,D) = P(G)\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{(r_i-1)!}{(N_{ij}+r_i-1)!}\prod_{k=1}^{r_i}N_{ijk}! \quad (3)$$

여기에서 r_i는 변수 x_i의 가능한 값의 수를 나타내고 q_i는 π_i에 포함된 변수에 의해 조합 가능한 상태의 수를 나타낸다. N_{ijk}는 π_i가 j번째 조합 상태이고 x_i가 그의 k번째 값을 가질 경우를 만족하는 데이터 D에서의 경우의 수를 의미한다. 그리고 $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}!$를 의미한다.[8]

K2 알고리즘의 네트워크 탐색 과정 및 원리는 비교적 간단한 방법으로 모든 노드를 일렬로 배열한 뒤에 앞선 노드가 뒤따르는 노드의 부모가 될 수 있도록 제약하고 노드 연결을 시도하며 점수 메트릭을 최대로 하는 모델을 찾아가는 탐욕적이고 경험적인 알고리즘이다.

III. 엔트로피를 적용한 베이지안망

3.1 엔트로피

엔트로피란 사전적인 의미로 물질계의 열적 상태를 나타내는 물리량의 하나이다. 자연현상은 언제나 물질계의 엔트로피가 증가하는 방향으로 일어나며 이를 엔트로피 증가의 법칙이라고 한다. 가역과정이란 일정한 환경에서 정반응과 역반응이 모두 일어나는 과정을 뜻하며 비가역 과정은 한쪽으로의 반응만 일어나는 과정을 말한다. 실제 생활에서는 가역과정은 존재하지 않으며 모든 움직임은 비가역과정으로 진행된다. 엔트로피는 비가역의 척도를 나타내고자 등장하였으며 비가역과정에서 항상 엔트로피가 증가하는 방향으로 이동하게 된다. 현재 여러 분야에서 엔트로피를 이용한 알고리즘 및 모델이 계속적으로 연구되고 있다. 특히, 학습이나 선택의 문제에서 엔트로피의 개념은 판단 기준이 되는 개념을 수치화하기에 좋은 성질을 가진다. 정보이론으로서 엔트로피에 따르면 주어진 정보가 적을 때 엔트로피가 증가한다. 즉 불확실성이 증가함을 뜻한다. 이러한 무질서의 정도를 베이지안 기반의 환경에 새롭게 적용시켜 수식화하여 나타냄으로써 엔트로피를 최소화 시킬 수 있는 방안을 발견하여 효과적인 분류 예측을 통해 확률적으로 더 정확한 방법을 유도해 낸다.[9][10]

$$entropy(p1,p2,...,p_n) = \\ -p_1\log p_1 - p_2\log p_2 - p_2\log p_2 ... - p_n\log p_n \quad (4)$$

(4)에서 P는 각 사건의 확률을 나타내며 계산을 통해 나타난 값이 엔트로피 즉 혼잡도를 나타낸다. 혼잡도가 최소가 되는 쪽이 신뢰도가 더 높은 방법이다. 엔트로피의 개념을 적용하게 되면 개체가 주어진 작업에 적합할

수록 개체 엔트로피는 낮고, 개체가 주어진 작업에 적합하지 않으면 개체 엔트로피는 높게 측정된다.

3.2 엔트로피와 베이지안망

변수가 많아졌을 때 학습이 어렵고 시간의 요구량이 늘어나게 되어 효율적이고 신뢰도 높은 탐색을 하는데 문제점이 발생하는 베이지안망의 문제점을 해결하기 위해서 노드의 순서를 정하여 효율적인 구조학습이 가능하도록 해야 한다.

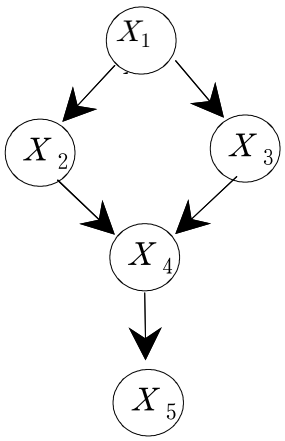

그림 3. 엔트로피에 기초한 베이지안 망
Fig. 3. Entrophy based Baysian Networks

예를 들어 화살표의 방향은 인과적인 연결성을 뜻하므로 그림 3의 경우 베이지안 노드 순서는 $\{x_1,x_2,x_3,x_4,x_5\}$와 $\{x_1,x_3,x_2,x_4,x_5\}$ 모두 가능하다. 이 두 가지 순서 중 시간적으로 더 효율적이면서도 신뢰도가 높은 방안으로 노드의 순서를 결정해야 한다. 여기서 노드의 순서를 결정하는 일은 베이지안망에서 효과적으로 지식을 표현하고 결론을 추론할 수 있게 하는 중요한 방안이 된다.

기존에 베이지안망의 구조학습으로 일반적으로 널리 알려진 K2학습 알고리즘은 노드순서와 데이터를 입력으로 받아서 베이지안망 구조를 출력으로 내어 놓는다. 하지만 노드 순서화를 이용하여 베이지안망 구조를 학습하고자 할 때 노드의 순서가 유일하지 않으므로 노드 탐색의 효율성의 문제가 발생하게 된다. 이를 해결하기 위해 기존의 K2알고리즘에 엔트로피를 적용한다. 각 노드에서의 엔트로피를 구하여 혼잡도를 최소화 시키는 방안으로 최적화된 노드의 연결순서를 만들어 나가는 것이다.

엔트로피는 확률 변수의 불확실성에 대한 수학적 척도로서, 확률이 각각 $P(P(x_1),P(x_2),...,P(x_n)$ 인 $x_1,x_2,...x_n$의 값을 찾는 이산 확률 변수 X에 대해 엔트로피 H(X)에 대한 정의를 내리면 (5)와 같다.[11]

$$H(X) = -\sum_{i=1}^{n} P(x_i)\log P(x_i) \quad (5)$$

일반적으로 이산 확률 변수 X가 균일 분포를 따르는 경우 그 엔트로피의 값은 최대가 되고 반대로 균일화되지 않고 집중되어 분포할수록 엔트로피가 최소가 된다.[12]

3.3 엔트로피를 적용한 알고리즘

K2알고리즘은 베이지안망의 구조 학습에 일반적으로 적용되는 알고리즘으로 표 1에서 MaxParentNum은 K2 알고리즘의 파라미터로서 노드의 부모를 얼마나 허용할지를 결정한다. 이 파라미터는 한 노드에 많은 부모 노드가 연결될 경우 조건부 확률 테이블이 지수적으로 커지는 문제를 막는 역할을 하며, 이를 통해 정확성과 속도의 완급 조절이 가능하다.

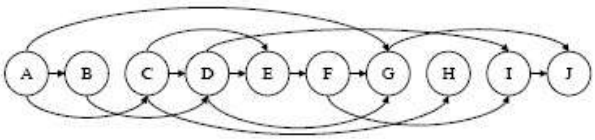

그림 4. 사이클이 생기지 않는 K2알고리즘의 노드구조
Fig. 4. Node structure in Non cycle K2 algorithm

K2 알고리즘은 부모가 되는 노드 x_j의 인덱스 j가 노드 x_i를 가리키는 인덱스 i보다 값이 크도록 제약을 걸어 놓기 때문에 그림 4와 같이 별도의 알고리즘 없이도 사이클(cycle)이 생기지 않고 DAG 구조가 유지된다. 하지만 노드 인덱스를 정의하는 순서에 따라 서로 다른 네트워크 구조를 가질 수 있기 때문에 노드 인덱스의 순서를 최적화해야하는 문제가 발생한다.

구조 학습 단계에서 각 노드의 점수 계산을 위해 사용된 메트릭 함수 g는 (6)과 같으며, (3)의 노드별 점수를 의미한다.

$$g(x_i,Pa(x_i)) = \prod_{j=1}^{q_i} \frac{(r_i-1)!}{(N_{ij}+r_i-1)!} \prod_{k=1}^{r_i} N_{ijk}! \quad (6)$$

이러한 노드 인덱스의 순서를 결정하는 문제를 해결

하기 위해 엔트로피를 적용시킨다.[13]

그림 3에서 노드 x_2와 x_3의 순서를 구하기 위하여 (5)에서 엔트로피의 수치를 계산한다. 만약 x_2의 엔트로피가 x_3의 엔트로피보다 작을 경우 노드 순서에 x_2를 선택한다. 반대로 x_2의 엔트로피가 x_3의 엔트로피보다 클 경우에는 x_3을 선택한다. 이와 같은 방식으로 노드의 순서를 구해 나간다면 노드순서 추정 시에 발생하는 불확실성을 해결할 수 있으며 베이지안망 학습의 문제점인 복잡한 노드에서의 시간 복잡도를 엔트로피를 적용하여 최적화된 노드의 순서를 빠르게 탐색하므로 효율성 있는 학습방안이 된다. 이러한 개념을 K2알고리즘에 엔트로피를 적용시켜보면 표 1과 같이 나타낼 수 있다.

표 1. K2 알고리즘에 엔트로피를 적용한 알고리즘
Table 1. Entrophy algorithm in K2 algorithm

```
procedure K2;
{input: A set of n nodes, an ordering on the nodes,
and upper bound u on the number of parents a node
may have, and a database D containing m class.}
{Output : for each node, a printout of the parents of
the node.}
for I := 1 to n do
 π_i :=∅;
P_old:=f(i,π_i); {this function is computed using
Equation 20.}
OKToProceed:=true;
While OKToProceed and |π_i|<u do
let z be the node in Pred(x_i)-π_i
that maximizes f(i,π_i∪{z});
if z is not unique
then select z that minimizes H(z)
P_new:=f(i,π_i∪{z});
if P_new>P_old then
P_old:=P_new;
π_i:=π_i∪{z};
else OKToProceed:=false;
end {while};
write('Node:',x_i, 'Parent of x_i: ',π_i);
end{for};
end{K2};
```

K2학습 알고리즘은 노드순서와 데이터를 입력으로 받아서 베이지안망 구조를 출력으로 내어 놓는다. 하지만 노드 순서화를 이용하여 베이지안망 구조를 학습하고자 할 때 노드의 순서가 유일하지 않으므로 노드 탐색의 효율성의 문제가 발생하게 된다. k2알고리즘에서 f(i,π_i∪{z})을 최대화시키는 z를 찾아 노드의 구조를 효율적으로 만들어 나가는데 이 때 최대화시키는 z가 유일하다는 보장을 할 수 없다. 효율적인 베이지안망 구조를 만들어 가기 위해서는 이러한 문제점을 해결해야 한다. 이를 해결하기 위해 기존의 K2알고리즘에 엔트로피를 적용한다. 알고리즘의 equation 20은 (6)을 참조하여 계산한다. 각 노드의 최대화된 값인 z에 대하여 엔트로피를 구한 후 엔트로피가 가장 낮은 z를 선택하는 것이다. 최종적으로 엔트로피를 구하여 혼잡도를 최소화 시키는 방안으로 최적화된 노드의 연결순서를 만들어 나가며 효율적인 베이지안망 구조가 이루어진다.

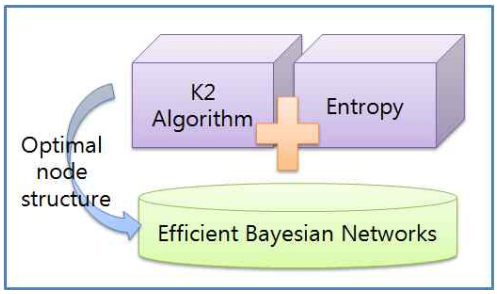

그림 5. 엔트로피 적용 개념도
Fig. 5. Conceptual diagram using entrophy

그림 5와 같이 기존의 K2알고리즘에 엔트로피의 수식을 활용하여 기존의 노드에서 최적의 노드 순서를 발견해 낸다. 이로써 결과적으로 효율적인 베이지안망 학습이 이루어지게 되는 것이다.

Ⅳ. 결 론

물질계의 열적 상태를 나타내는 물리량의 하나인 엔트로피를 이용하여 학습이나 선택의 문제에 적용을 시도하였다. 이는 엔트로피가 판단 기준이 되는 개념을 수치화하기에 좋은 성질을 가지기 때문이다. 정보이론으로서 엔트로피에 따르면 주어진 정보가 적을 때 엔트로피가 증가한다. 즉 불확실성이 증가함을 뜻한다. 이러한 무질서의 정도를 베이지안 기반의 환경에 새롭게 적용시켜 수식화하여 나타냄으로써 엔트로피를 최소화 시킬 수 있는 방안을 발견하여 효과적인 분류 예측을 통해 확률적

으로 더 정확한 방법을 유도해 낸다. 이런 방법으로 기존의 베이지안망이 갖는 문제를 해결한다. 이는 베이지안망이 노드개수의 증가에 따른 시간적 탐색공간의 급격한 증가로 인한 문제를 풀 수 있게 된다. 다만, 이를 검증할 실험을 계속하여 효율적인 학습방법이 되도록 한다.

참 고 문 헌

[1] 김경헌, 베이지안 네트워크에 기초한 백혈병 유전자 데이터의 분석, 공학학사학위논문, 2005.12

[2] 송윤석, 조성배, “로봇의 효과적인 서비스를 위해 베이지안 네트워크 기반의 실내 환경의 가려진 물체 추론”, 정보과학회논문지:컴퓨팅의 실제 제12권 제1호, 2006. 2

[3] 김희택, 조성배, 베이지안 네트워크의 학습에 기반한 모바일 환경에서의 사용자 적응형 음식점 추천 서비스, HCI 2009 학술대회

[4] 구정모, 동적시스템의 신뢰도 평가를 위한 베이지안망의 적용에 관한 연구, 공학석사 학위논문, 2004.2

[5] 도용태, 김일곤, 김종완, 박창현, “인공지능 개념 및 응용”, pp77-82

[6] 하선영, 데이터마이닝을 위한 베이지안망 구조학습, 공학석사 학위논문, 2001.2

[7] 황규백, 장병탁, 대규모 베이지안망 구조 학습 알고리즘, 2001 한국 뇌학회 학술대회 논문집

[8] 황금성, 조성배, “베이지안 네트워크의 학습”, pp 15-27

[9] Ian H.Witten and Eibe Frank, "Data Mining", pp89-97

[10] Ibrahim Dincer and Yunus A. Cengel, "Energy, Entropy and Exergy Concepts and Their Roles in Thermal Engineering", ISSN 1099-4300, 2001.3

[11] Julio Michael, Bayesian Interference and Maximum Entropy Methods in Science and Engineering, 2008

[12] 장정호, 장병탁, 김영택, “최대 엔트로피 기반 문서 분류기의 학습”, 한국정보과학회 추계학술발표논문집 제 26권 제 2호, 1999

[13] 손승현, 김재련, 엔트로피 기반 분할과 중심 인스턴스를 이용한 분류기법의 데이터 감소, 한국산업경영시스템학회, 산업경영시스템학회지 제 29권 제 2호, 2006. 6

저자 소개

허 고 은(정회원)

• 2007년~현재 을지대학교 의료전산학 전공 재학 중

<주관심분야: 데이터마이닝>

정 용 규(정회원)

• 1981 서울대학교 (이학사)
• 1994년 연세대학교 (공학석사)
• 2003년 경기대학교 (이학박사)
• 1999년~현재 을지대학교 교수
• 1994~현재 UN/Cefact TBG3맴버
• 2001~현재 ISO/TC154K 위원장
• 2005~현재 국가표준(KS)심의위원

<주관심분야: 임상데이터마이닝, 의료정보시스템, 국제표준>