퍼지 및 지능시스템학회 논문지 2005, Vol. 15, No. 3, pp. 312-317

변형된 Support Vector Machine을 이용한 유비쿼터스 데이터 마이닝

Ubiquitous Data Mining Using Hybrid Support Vector Machine

전성해

Sung-Hae Jun

청주대학교 통계학과

요 약

유비쿼터스 컴퓨팅 환경은 정치, 경제, 사회, 문화, 교육 등 대부분의 분야에 많은 영향을 주고 있다. 인터넷에 비해 훨씬 거대한 유비쿼터스 네트워크 환경이 효과적으로 운영되기 위해서는 네트워크에 접속한 다양한 컴퓨터들이 스스로 지능을 가지고 주어진 상황에서 최적의 의사결정을 할 수 있어야 한다. 현재 많은 분야에서 데이터 마이닝은 지능형 시스템 구축을 위한 효과적인 분석도구로 사용되고 있다. 지능화된 유비쿼터스 컴퓨팅 환경의 구현을 위한 유비쿼터스 데이터 마이닝을 위하여 본 논문에서는 변형된 Support Vector Machine 기법을 제안하였다. 유비쿼터스 컴퓨팅 환경에서 상당 부분의 데이터가 센서를 통하여 수집된다. 센서 네트워크를 통하여 수집된 데이터는 상당 부분 잡음을 포함한 데이터이다. 제안 기법은 특히 센서 네트워크를 통한 스트림 데이터의 잡음을 제거하는 데 목적을 두고 있다. 본 논문의 실험에서는 유비쿼터스 센서 네트워크를 나타내는 다양한 분포로부터 시뮬레이션 데이터를 생성하여 제안 방법의 성능 평가를 수행하였다.

Abstract

Ubiquitous computing has had an effect to politics, economics, society, culture, education and so forth. For effective management of huge Ubiquitous networks environment, various computers which are connected to networks has to decide automatic optimum with intelligence. Currently in many areas, data mining has been used effectively to construct intelligent systems. We proposed a hybrid support vector machine for Ubiquitous data mining which realized intelligent Ubiquitous computing environment. Many data were collected by sensor networks in Ubiquitous computing environment. There are many noises in these data. The aim of proposed method was to eliminate noises from stream data according to sensor networks. In experiment, we verified the performance of our proposed method by simulation data for Ubiquitous sensor networks.

Key Words : 유비쿼터스 데이터 마이닝, 우도비 검정통계량, Hybrid Support Vector Machine, 잡음 데이터

1. 서 론

다양한 종류의 컴퓨터가 사람, 사물, 환경 속으로 스며들고 이들이 네트워크로 연결되어 인간의 삶의 질을 향상시키는 유비쿼터스 컴퓨팅(Ubiquitous computing) 환경은 모든 사물과 사람들이 보이지 않는 센서 네트워크로 연결되어 개인화된 맞춤형 서비스를 실시간으로 제공할 수 있는 지능형 환경에 대한 개념을 포함하고 있다[13]. 원래 유비쿼터스는 라틴어에서 유래된 것으로 '도처에 널려 있다', '언제 어디서나 동시에 존재한다'라는 의미이다[25]. 국립국어연구원에서는 2004년 10월에 유비쿼터스의 우리말을 '두루누리'로 결정하였다. 1988년 M. Weiser의 제안에 의해 유비쿼터스 컴퓨팅은 탄생하였다. 그는 '미래의 컴퓨터는 어떠한 컴퓨터가 되어야 할까'라는 문제의식에서 언제, 어디서나, 어떤 형태로든 컴퓨터를 사용할 수 있는 환경, 즉 유비쿼터스 컴퓨팅 개념을 생각하게 되었다. 이는 컴퓨터 사용자가 일보다도 컴퓨터 조작에 더 몰두해야 하는 성가심을 비판하며 인간중심의 컴퓨팅 기술의 비전을 제시한 것이었다[24][29]. 결론적으로 최선의 컴퓨터 시스템이란 사용자가 컴퓨터를 이용하고 있다고 느끼지 않은 상태에서 컴퓨터의 도움을 받아 작업을 할 수 있도록 하는 환경이다. 유비쿼터스 컴퓨팅 환경은 수많은 사람과 사물에 내제된 컴퓨터들이 네트워크로 연결된 복잡한 구조를 이루고 있으며, 이러한 거대한 네트워크 환경에서는 기존의 컴퓨팅 환경처럼 컴퓨터 전문가가 직접 시스템을 제어할 수 있는 여지가 거의 없어지게 된다. 때문에 유비쿼터스 컴퓨팅 환경에 접속한 컴퓨터들이 스스로 판단하고 움직일 수 있는 지능화된 시스템이 요구되어 진다[1][2][17][26]. 시스템의 지능화를 위한 도구(tool)로서 현재 데이터 마이닝이 많이 사용되고 있다[4][5][6][7][8]. 특히 유비쿼터스 컴퓨팅 환경의 지능화를 위한 데이터 마이닝 전략을 본 논문에서는 유비쿼터스 데이터 마이닝이라고 하였다. 오프라인 데이터 마이닝과 웹 마이닝과는 달리 유비쿼터스 데이터 마이닝

접수일자 : 2005년 5월 30일
완료일자 : 2005년 6월 14일

에서는 이질적인 컴퓨터들이 접속하여 매우 다양한 종류의 데이터를 발생시키고, 동시에 의미없는 잡음 데이터도 나타나게 된다[3]. 본 논문에서는 이러한 유비쿼터스 컴퓨팅 환경에서 발생되는 잡음(noise) 데이터 문제점을 해결하기 위하여 변형된(hybrid) Support Vector Machine(HSVM) 모형을 이용한 유비쿼터스 데이터 마이닝 기법을 제안하였다. 제안된 기법의 성능 평가를 위하여 유비쿼터스 컴퓨팅 환경을 시뮬레이션 한 데이터를 이용하여 제안기법의 성능평가를 위한 비교실험을 수행하였다.

2. 관련 연구

2.1 유비쿼터스 잡음 데이터

유비쿼터스 컴퓨팅 환경에서 발생되는 데이터의 상당 부분은 실제 마이닝 작업에서는 필요 없는 것들이다. 이러한 필요 없는 잡음 데이터는 다음의 그림에서 나타내고 있는 지식의 계층구조도를 통하여 정의되어진다[11].

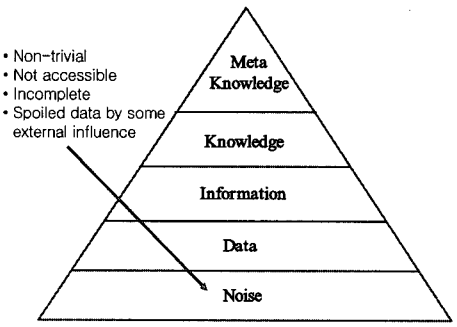

그림 1. 지식의 분류체계
Fig. 1. Hierarchy of knowledge

위의 그림은 데이터로서의 가치가 없는 잡음(noise) 단계부터 데이터 그리고 이를 가공하여 얻게 되는 정보(information) 그리고 최종적인 의사결정에 사용되는 지식(knowledge) 및 메타지식(meta knowledge)에 이르는 지식의 계층구조(hierarchical structure)를 보여주고 있다. 유비쿼터스 컴퓨팅 환경에서 많은 부분에서 데이터 수집(gathering)이 센서(sensor)를 통하여 이루어진다. 센서를 통하여 데이터를 수집하게 되는 경우에는 센서 자체의 결합 뿐만 아니라 빛과 온도와 같은 외부 요인 등에 의해 불완전한 잡음들이 수집되는 데이터에 다량이 포함되게 된다. 때문에 효과적인 유비쿼터스 데이터 마이닝을 위해서는 센서 네트워크로부터의 모니터링 데이터에서 잡음을 제거하는 효과적인 필터링(filtering) 전략이 필요하게 된다. 현재 연구되어지고 있는 대부분의 유비쿼터스 데이터 마이닝은 이러한 센서 네트워크로부터 받아들인 데이터 스트림에 대한 마이닝 기법에 대한 연구가 대부분이다[3][9][10][19][21][23]. 하지만 이 연구들은 센서에 수집된 데이터에 잡음이 포함되지 않거나 극히 적은 양만 있는 것으로 가정하고 마이닝을 진행하고 있다. 하지만 실제 데이터는 그렇지 않다. 본 논문은 현재 연구되는 스트림 데이터의 마이닝 작업이 이루어지기 위한 잡음에 대한 처리에 대한 효과적인 방법으로서 변형된 통계적 학습모형을 제안하였다. 센서 네트워크에서 발생되는 잡음은 매우 다양하게 발생되기 때문에 형태 또한 매우 다양하고 마이닝 모형이 이전에 학습해 보지 못하고 필터링에 적용되는 경우도 생기게 될 것이다. 때문에 이러한 경우에는 잡음의 제거 능력이 떨어지게 된다. 즉 잡음과 비잡음(non-noise)을 제대로 구별하는 능력이 떨어져 효과적인 마이닝 모형 구축이 어렵게 된다[15][22]. 따라서 본 노문에서 이러한 문제점을 해결하기 위하여 다음 절에서 변형된 통계적 학습모형을 이용한 유비쿼터스 데이터 마이닝 기법을 제한하였다.

2.2 검정통계량

주어진 데이터의 각 속성(attribute)은 하나의 확률 변수(random variable)가 된다. 예를 들어 학습 데이터 개체 $z=(x, y)$에서 x와 y는 각각 확률 변수이다. 확률 변수들 간의 연관성 유무, 또는 연관성의 정도를 확인하는 통계적 기법이 우도비 검정(Likelihood Ratio Testing: LRT)이다[16]. 변수 x와 y에 대한 LRT를 위한 가설(hypothesis)은 다음 식과 같다.

$$H_0 : \mu_x - \mu_y = 0 \quad vs \quad H_1 : \mu_x - \mu_y \neq 0 \tag{1}$$

식 (1)의 가설에서 귀무가설(null hypothesis) H_0는 두 변수 x와 y 사이에 관련성이 없다는 것이고 대립가설(alternative hypothesis) H_1는 두 변수 간에 연간성이 존재한다는 것을 나타낸다. LRT는 H_0와 H_1중의 하나를 결정하는 통계적 가설 검정 기법 중의 하나이다. 본 논문의 센서 네트워크로부터의 잡음 데이터를 제거하는데 사용될 HSVM의 각 입력변수에 대한 가중치 정보로 사용할 LRT 검정 통계량(power) T는 다음식과 같이 표현된다.

$$T(x_1, x_2, \cdots, x_l) = \frac{\sqrt{l} \cdot \overline{x}}{\sqrt{\sum_{i=1}^{l} (x_i - \overline{x})^2/(l-1)}}$$
$$where, \quad \overline{x} = \frac{1}{l} \sum_{i=1}^{l} x_i \tag{2}$$

식 (2)에서 T-값이 클수록 두 확률 변수 사이의 연관성은 크게 된다. T-값은 t-확률분포(probability distribution)상의 기각역 정보를 제공하는 값이며 입력변수의 척도(scale)에 따라 값의 범위가 달라진다. 때문에 서로 다른 척도를 가진 입력변수(input variable)들에 대한 목표변수(target variable)에 대한 유의성을 비교하기 위해서는 이 값을 유의확률(p-value)로 바꾸어 주어야 한다. 이 값은 $-\infty$에서 $+\infty$까지의 범위를 같게 되는 T-값을 0에서 1사의 확률 값으로 척도에 관련 없이 표준화시켰기 때문에 서로 다른 척도를 가진 입력변수들에 대해서도 동등한 비교가 가능하게 된다. T 값과 유의확률은 반비례한다. 즉 유의 확률 값이 작을수록 두 변수간에 서로 유의한 차이가 없다는 귀무가설을 기각하게 되어 해당 입력변수가 잡음과 비잡음의 레이블을 가진 목표변수에 상관성이 높다고 할 수 있다. 따라서 본 논문에서는 유의 확률 값이 작은 입력 변수에 대하여 큰 가중치를 부여하는 전략을 취하였다. 즉, 본 논문의 실험에서는 각 입력 변수들에 대하여 두개의 레이블을 갖는 목표 변수와의 T 통계량 값을 구하여 해당 입력변수에 곱함으로써 가중치를 결정하였다.

2.3 Support Vector Machine

본 논문의 잡음 정제 모형 구축을 위하여 사용하는 변형 이전의 통계적 학습 모형으로 SVM을 이용하였다[1, 13]. 클래스 레이블들을 가진 목표변수 y와 입력벡터(input vector) x로 구성된 데이터 집합 S는 다음 식과 같은 데이터 구조로

표현된다[27][28].

$$(y_1, x_1), (y_2, x_2), \cdots, (y_l, x_l), \quad x_i \in R^N, \; y_i \in \{-1, 1\} \tag{3}$$

대부분의 분류모형 구축의 경우에 입력공간(input space)에서 서로 다른 클래스 레이블을 분류하는 정확한 초평면(hyperplane)을 찾는 것은 매우 제한적이기 때문에 바로 분류 모형을 사용하기가 어렵다. 이러한 상황에서 해결 방안은 입력공간을 더 높은 차원의 특징 공간(feature space)으로 사상(mapping)시키고, 이 특징 공간에서 최적의 초평면을 찾는 것이다. $z = \phi(x)$를 입력공간 R^N에서 특징 공간 Z로의 사상 ϕ를 갖는 특징 공간 벡터로 표현하면, (w, b)의 쌍으로 이루어진 다음의 초평면을 구해야 한다.

$$w \cdot z + b = 0 \tag{4}$$

식 (4)의 초평면 식을 구하게 되면 다음의 식 (5)의 함수에 의해 개개의 x_i들을 분류할 수 있게 된다.

$$f(x_i) = sign(w \cdot z_i + b) = \begin{cases} 1 & \text{if } y_i = 1 \\ -1 & \text{if } y_i = -1 \end{cases} \tag{5}$$

여기서 $w \in Z$이고 $b \in R$이다. 특히, 집합 S는 (w, b)의 쌍이 존재하면 선형분류 가능(linearly separable)이라고 하고 다음의 부등식이 S의 모든 원소들에 대해 성립한다.

$$\begin{cases} (w \cdot z_i + b) \geq 1, & \text{if } y_i = 1 \\ (w \cdot z_i + b) \leq -1, & \text{if } y_i = -1 \end{cases} \quad i = 1, 2, \cdots, l \tag{6}$$

선형분류 가능한 집합 S는 두 개의 서로 다른 클래스 레이블들의 학습 데이터의 사영(projection)들 사이의 마진(margin)을 최대화 하는 유일한 최적 초평면을 구할 수 있다. 만약 집합 S가 선형 분류 가능이 아니면 분류규칙 위반(classification violations)이 SVM 형식에서 허용되어져야 한다. 선형분류 가능이 아닌 데이터를 다루기 위하여 음이 아닌 변수 ξ_i를 도입하여 아래 식과 같이 식 (6)을 일반화한다.

$$y_i(w \cdot Z_i + b) \geq 1 - \xi_i, \quad i = 1, 2, \cdots, l \tag{7}$$

식 (7)에서 ξ_i는 식 (6)을 만족하지 않는 x_i들이다. 그러므로 $\sum_{i=1}^{l} \xi_i$는 오분류(misclassification)의 양을 나타내는 측도로서 고려된다. 따라서 최적 초평면을 구하는 문제는 아래의 문제에 대한 해(solution)가 된다.

$$\begin{aligned} &minimize \; \frac{1}{2} w \cdot w + C \sum_{i=1}^{l} \xi_i \\ &subject \; to \; y_i(w \cdot z_i + b) \geq 1 - \xi_i \end{aligned} \tag{8}$$

여기서 $\xi_i \geq 0$이고 $i = 1, 2, \cdots, l$이다. C는 상수(constant)이며 조정 모수(regularization parameter)이다. 이 모수의 조정으로 마진 최대화와 분류 규칙 위반 사이의 균형을 맞출 수 있게 된다. 식 (8)에서 최적 초평면을 찾는 것은 다음의 라그랑지 변환(Lagrangian transformation)을 통하여 풀 수 있는 문제가 된다.

$$\begin{aligned} &maximize \; W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j z_i \cdot z_j \\ &subject \; to \; \sum_{i=1}^{l} y_i \alpha_i = 0 \qquad 0 \leq \alpha_i \leq C, \; i = 1, 2, \cdots, l \end{aligned} \tag{9}$$

여기서 $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_l)$는 식 (7)의 제한 조건과 관련된 음이 아닌 라그랑지 승수(multiplier)들의 벡터이다. Kuhn-Tucker 정리는 SVM 이론에서 중요한 역할을 한다. 이 정리에 의하여 식 (9)의 해 $\overline{\alpha_i}$는 다음을 만족한다.

$$\begin{aligned} &\overline{\alpha_i}(y_i(\overline{w} \cdot z_i + \overline{b}) - 1 + \overline{\xi_i}) = 0 \\ &(C - \overline{\alpha_i})\,\overline{\xi_i} = 0, \; i = 1, 2, \cdots, l \end{aligned} \tag{10}$$

식 (10)의 첫 번째 식으로부터 구한 해 $\overline{\alpha_i}$는 식 (7)의 등호를 만족시킨다. $\overline{\alpha_i} > 0$인 x_i를 support vector라고 부른다. 분류가 가능하지 않은(nonseparable) 경우에는 support vector는 두 가지의 형태로 존재한다. $0 < \overline{\alpha_i} < C$인 경우의 support vector x_i는 $y_i(\overline{w} \cdot z_i + \overline{b}) = 1$과 $\xi_i = 0$을 만족하고, $\overline{\alpha_i} = C$인 경우의 ξ_i는 널(null)이 아니고 대응되는 support vector x_i는 식 (6)을 만족하지 않는다. 이 support vector 들은 오차(error)로서 간주된다. $\overline{\alpha_i} = 0$에 대응되는 x_i는 결정 마진(decision margin)과 떨어져서 정확하게 분류된다. 최적 초평면 $\overline{w} \cdot z + \overline{b}$를 구축하기 위하여 다음의 식과 스칼라 $\overline{b}$가 필요하다.

$$\overline{w} = \sum_{i=1}^{l} \overline{\alpha_i} y_i z_i \tag{11}$$

이것은 식 (10)의 첫 번째 식의 Kuhn-Tucker 조건에 의해 결정된다. 결정 함수(decision function)는 식 (5)와 식 (11)에 의해 다음식과 같이 일반화된다.

$$f(x) = sign(w \cdot z + b) = sign\left(\sum_{i=1}^{l} \alpha_i y_i z_i \cdot z + b \right) \tag{12}$$

ϕ에 대한 어떠한 지식(knowledge)도 없기 때문에 식 (9)와 식 (12)의 계산은 불가능하다. 하지만 SVM은 ϕ에 대해서 알 필요가 없다. 단지 커널(kernel)이라 불리우는 $K(\cdot, \cdot)$가 다음과 같은 식에 의해 특징 공간 Z에 데이터의 내적(dot product)을 계산한다.

$$z_i \cdot z_j = \phi(x_i) \cdot \phi(x_j) = K(x_i, x_j) \tag{13}$$

Mercer의 정리를 만족하는 함수들은 내적 계산이 가능하고 따라서 커널로써 사용이 가능하다. SVM 분류기(classifier)를 구축하기 위하여 아래와 같은 차수(degree) d의 다항(polynimial) 커널을 사용한다.

$$K(x_i, x_j) = (1 + x_i \cdot x_j)^d \tag{14}$$

따라서 비선형 분류 가능 초평면은 다음 식의 해로서 구해진다.

$$\begin{aligned} &maximize \; W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\ &subject \; to \; \sum_{i=1}^{l} y_i \alpha_i = 0 \quad 0 \leq \alpha_i \leq C, \; i = 1, 2, \cdots, l \end{aligned} \tag{15}$$

그리고 최종적인 결정 함수는 다음과 같다.

$$f(x) = sign(w \cdot z + b) = sign\left(\sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b \right) \tag{16}$$

본 논문의 제안 모형은 LRT에 의해 가중치가 적용된 입력 변수들을 이용하여 잡음과 비잡음의 두 개의 레이블을 갖은 목표 변수의 클래스를 분류하게 된다.

3. 변형된 Support Vector Machine 모형을 이용한 유비쿼터스 데이터 마이닝

3.1 모형의 구조

본 논문에서 제안하는 모형은 LRT 과정과 HSVM 과정의 2단계 프로세스로 구성되었다. 우선 원래의 유비쿼터스 센서 네트워크 추출 데이터에 대하여 목표 변수와 입력 변수들 사이의 유의한 연관성의 정도를 검정통계량의 유의확률 값 계산결과에 의해 연관성의 순위를 결정하고, 이를 통하여 각 입력 변수에 대한 목표 변수의 가중치를 결정하였다. 다음으로 HSVM 단계에서, 기존의 SVM에서는 모든 입력 변수들이 목표 변수에 대하여 동일한 중요도를 가지지만 본 논문에서 사용한 LRT 과정에 의해 각 입력 변수의 가중치가 새롭게 부여되어 잡음 제거 모형이 구축되었다. 다음 그림은 제안 모형의 절차를 도식화하였다.

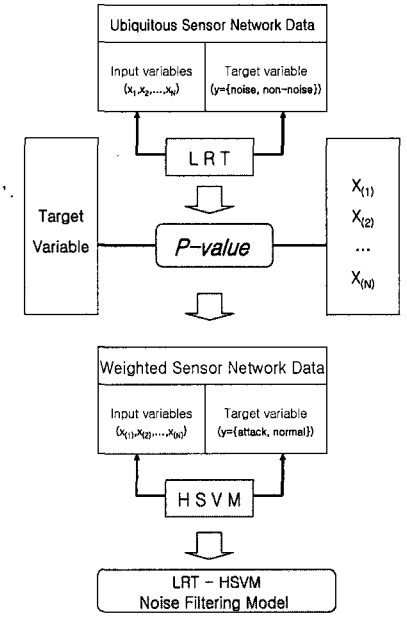

그림 2. 제안모형의 절차
Fig. 2. Process of proposed method

위 그림에서 $x_{(i)}$는 유의 확률에 의하여 목표 변수에 i번째로 유의한 입력 변수를 의미한다. 즉, 전체 입력 변수들 중에서 i번째로 큰 가중치 값을 갖게 되는 변수이다.

3.2 제안 잡음 제거 모형의 알고리즘

제안 모형의 알고리즘은 다음과 같이 3단계로 이루어진다.

단계 1 : (LRT step)

① 목표변수(y)와 입력변수들($x_1, x_2, \cdots, x_N$)간의 LRT 시행 → T-검정통계량 계산

$$T(x_1, x_2, \cdots, x_l) = \frac{\sqrt{l} \cdot \overline{x}}{\sqrt{\sum_{i=1}^{l}(x_i - \overline{x})^2/(l-1)}}$$

$$where, \quad \overline{x} = \frac{1}{l}\sum_{i=1}^{l} x_i$$

② 입력변수들간의 척도 영향을 없애기 위하여 T-값을 유의 확률(p-value)로 변환

$$p-value = P(t > T_i)$$

→ p-value의 크기 비교를 통한 입력 변수들의 서열화

$$(\{x_1, x_2, \cdots, x_N\} \rightarrow \{x_{(1)}, x_{(2)}, \cdot, x_{(N)}\})$$

$x_{(i)}$: i번째 가중치 크기를 갖는 입력 변수

③ p-value에 반비례하여 가중치 부여

단계 2 : (HSVM step)

잡음 정제 모형의 구축

$$f(x) = sign(w \cdot x + b), \quad z = \phi(x)$$
$$x = (x_{(1)}, x_{(2)}, \cdots, x_{(N)})$$

(w, b) : parameters

$\phi(\cdot)$: 입력 공간에서 특징 공간으로의 함수(mapping)

단계 3 : (Application step)

새로운 잡음 데이터의 $f(x)$ 계산

→ $f(x)$의 값에 의해 잡음과 비잡음을 분류

잡 음 : $f(x) = 1$ if $w \cdot z + b > 0$

비잡음 : $f(x) = -1$ if $w \cdot z + b \leq 0$

4. 실험 및 결과

본 논문에서 제안하는 모형의 성능 평가를 위한 시뮬레이션은 입력변수들 간의 연관성이 강하게 존재하는 경우와 그렇지 않은 경우로 나누어서 수행하였다. 왜냐하면 유비쿼터스 센서 네트워크에서 발생되는 잡음 데이터의 구조가 이와 같은 두 가지 경우로 구분될 수 있기 때문이다. 따라서 본 논문에서는 난수발생 분포로서 다변량 정규분포(multivariate normal distribution)를 이용하였다. 다변량 정규분포를 따르는 난수(random number)를 발생시키기 위해서는 다변량 확률변수(random variables)를 발생시켜야 한다. 이 방법은 일반적인 일변량(univariate) 정규분포를 따르는 난수를 발생시키는 것과 비슷하지만 다음 식과 같이 표준정규 난수의 d-차원(dimension) 벡터로서 난수를 생성한다[18][20].

$$x = R^T z + v \tag{17}$$

위 식에서 z는 (d×1)차원의 표준정규분포(standard normal distribution)를 따르는 확률변수이다. v는 평균(mean)을 나타내는 (d×1)차원 벡터이고 R은 다음 식을 만족하는 (d×d)차원의 행렬(matrix)이다.

$$R^T R = \Sigma \tag{18}$$

위 식에서 Σ는 분산-공분산(variance-covariance) 행렬이다. 본 논문의 실험을 위한 입력변수들 간의 연관성은 이 행렬을 조정하여 시뮬레이션 데이터를 생성하였다. 다음 그림은 본 논문의 실험을 위하여 사용된 2개의 분산-공분산 행렬이다.

$$\begin{pmatrix} 1 & 0.09 & 0.21 & 0.13 \\ 0.09 & 1 & 0.16 & 0.19 \\ 0.21 & 0.16 & 1 & 0.23 \\ 0.13 & 0.19 & 0.23 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 0.73 & 0.59 & 0.48 \\ 0.73 & 1 & 0.68 & 0.83 \\ 0.59 & 0.68 & 1 & 0.91 \\ 0.48 & 0.83 & 0.91 & 1 \end{pmatrix}$$

(a) (b)

그림 3. 두 개의 분산-공분산 행렬
Fig. 3. Two variance-covariance matrices

위 그림에서 (a)는 입력 변수들 간의 연관성이 크지 않은 경우이고 (b)는 입력변수들 간의 연관성을 크게 한 경우이다. 또한 비잡음 데이터의 평균벡터는 0 벡터로 했고 나머지 잡음 데이터를 위한 평균벡터는 0 벡터가 아닌 임의의 벡터 값을 주었다. 실험을 위하여 비잡음은 한 개의 분포로부터 추출되게 하였고 잡음 데이터는 5개의 서로 다른 평균벡터를 갖는 5개의 분포로부터 각각 생성되게 하였다. 제안 방법과의 성능 평가를 위한 기존의 분류 기법은 SVM, 3개의 의사결정나무(decision trees) 모형, 베이지안 네트워크(Bayesian networks), 그리고 로지스틱 회귀모형(logistic regression)을 이용하였다. 다음 표는 이들 모형과의 비교실험의 결과를 나타내고 있다.

표 1. 비교모형들의 오분류율

Table 1. Misclassification ratios of comparative methods

비교 모형	오분류율
HSVM	0.0615
SVM	0.0915
CHAID	0.2587
CART	0.2688
C4.5	0.2311
Logistic	0.1356

위의 표에서 CHAID(chi-square automatic detection detector), CART(classification and regression tree), 그리고 C4.5는 각각 카이제곱 검정(chi-square test), 지니 지수(gini index), 그리고 엔트로피(entropy)에 기반한 의사결정나무 모형이다[12]. 오분류율에 의한 비교 결과에서는 제안 모형이 의사결정나무 모형에 비해서는 4배이상 정확히 분류하고 있고, 현재 이진 분류기로서 좋은 성능을 보이고 있는 SVM에 비해서도 오분류율이 작게 나타나고 있음을 알 수 있다. 그런데 위의 결과를 보면 분류나무 모형의 오분류율이 다른 모형들에 비해 특히 좋지 않게 나왔다. 이는 연속형 입력 변수를 분류 나무 모형에 적용하기 위하여 범주화 하는 과정에서 정보의 손실이 발생하였기 때문으로 볼 수 있다. 따라서 분류나무 모형은 연속형의 센서 네트워크 잡음 데이터의 처리에는 적절치 않다고 판단되어 리프트 값(lift value)에 대한 모형 비교에서는 이 모형은 배제하였다. 분류 모형에서는 리프트 값보다도 오분류율이 모형 성능 평가에서 우선시되기 때문이기도 하다. 일반적으로 분류에 대한 리프트 값은 다음과 같이 구할 수 있다[12].

$$LV = \frac{\%Resp}{BL-LV} \tag{19}$$

위 식에서 %Resp는 해당 클래스의 전체 수에 대한 해당 클래스에서 목표 변수의 특정 레이블의 빈도에 대한 백분율을 나타낸다. 또한 BL-LV(base line lift value)은 분류 모형이 적용되지 않은 원래의 학습 데이터에 대한 리프트 값이다. <표 4>는 이들 비교 모형들의 리프트 값의 결과를 나타내고 있다.

표 2. 비교모형들의 리프트 값

Table 2. Lift values of comparative methods

비교 모형	lift value
HSVM	2.68
SVM	2.11
Logistic	1.65

위의 결과를 보면 HSVM 모형의 리프트 값이 2.68로 나왔다. 이는 이 모형에 의해 잡음의 가능성에 대한 분류 결과 상위 10%의 예측이 모형을 구축하기 전에 비해 2.68배 성능 향상을 보이고 있는 것이다. 이에 비해 다른 모형들의 리프트 값은 제안 모형에 비해 작음을 알 수 있다.

5. 결론 및 향후 연구과제

본 논문에서는 유비쿼터스 환경의 센서 네트워크로부터 수집되는 데이터에 포함된 잡음을 제거하기 위한 전략으로서 변형된 SVM 모형을 제안하였다. 개개의 잡음 형태를 모두 모형화하는 대신 본 논문에서는 비잡음 데이터이외의 모든 데이터는 잡음으로 분류하는 이진 분류 전략을 제안하였다. 제안 전략은 기존의 학습에 의해 알려진 잡음에 대한 패턴을 추출하고 모형화 하여, 또다시 같은 잡음이 발생하는 경우에 이를 탐지할수 있을 뿐만 아니라 알려지지 않은 새로운 잡음에 대해서도 적응성을 가지고 제거해 낼 수 있는 모형을 제안하였다. 센서를 통한 데이터에 포함된 잡음의 패턴이 점점 복잡화됨에 따라 기존의 방법으로는 잡음에 대한 제거에 한계를 보이고 있고 자동화, 분산화 컴퓨팅에 의해 하나의 위치에서만 잡음이 발생하지 않고 동시에 여러 곳에서 발생됨에 따라 제안 모형의 사용은 좀 더 효과적일 전략이 될 수 있을 것이다. 센서 네트워크로부터의 잡음 데이터를 탐지하여 모형화 하기 위하여 통계적 가설검정 기법인 LRT에 의한 유의 확률값을 이용하여 각 입력변수의 가중치를 구하고 이를 SVM의 입력변수로 적용한 HSVM 모형을 개발하였다. 실험을 통하여 제안된 모형이 기존의 분류 모형들보다 잡음의 분류에 대한 적응성이 뛰어남을 확인하였다. 앞으로 HSVM의 커널함수와 평활상수의 최적 선택방법에 대한 연구를 통하여 제안모형의 추가적인 성능향상을 기대할 수 있을 것이다.

참 고 문 헌

[1] 백성욱, "유비쿼터스 컴퓨팅 환경을 위한 지능형 미디어 기술", *정보과학회지*, 21권, 5호, pp. 36-42, 2003.

[2] 유준재, "유비쿼터스 컴퓨팅 플랫폼", *충북 IT 신기술 워크샵 발표자료집*, pp. 9-73, 2004.

[3] 장세이, 우운택, "유비쿼터스 컴퓨팅 환경을 위한 센싱 기술과 컨텍스트-인식 기술의 연구 동향", *정보과학회지*, 21권, 5호, pp. 18-28, 2003.

[4] 전성해, 류제복, 이승주, "Data Mining Approach to Supporting Hoarding in Mobile Computing Environments", *한국통계학회 2003 춘계 학술대회 논문집*, pp. 13-18, 2003.

[5] G. D. Abowd, E. D. Mynatt, "Charting Past, Present, and Future Research in Ubiquitous Computing", ACM Transactions on Computer-Human Interaction, vol. 7, no. 1, pp. 29-58, 2000.

[6] R. Alonso and H. F. Korth, "Database system issues in nomadic computing", ACM SIGMOD International Conference on the Management of Data, 1993.

[7] P. K. Chrysanthis, "Transaction processing in mobile computing environment", IEEE Workshop on

Advances in Parallel and Distributed Systems, 1993.
[8] M. H. Dunham and A. S. Helal, "Mobile computing and databases: Anything new?", *SIGMOD Record*, vol. 24, no. 4, pp. 5-9, 1995.
[9] M. M. Gaber, S. Krishnaswamy, A. Zaslavsky, "A Wireless Data Stream Mining Model", Third International Workshop on Wireless Information Systems, 2004.
[10] M. M. Gaber, S. Krishnaswamy, A. Zaslavsky, "Ubiquitous Data Stream Mining", Eighth Pacific-Asia Conference on Knowledge Discovery and Data Mining, Sydney, 2004.
[11] J. Giarratano, G. Riley, *"Expert System, Principles and Programming"*, PWS Publishing Company, 1998.
[12] J. Han, M. Kamber, *"Data Mining: Concepts and Techniques"*, Morgan Kaufmann Publishers, 2001.
[13] R. Hunter, *"World Without Secrete: Business, Crime, and Privacy in The Age of Ubiquitous Computing"*, John Wiley and Sons, 2002.
[14] J. Jing, A. Helal and A. K. Elmagarmid, "Client server computing in mobile environments", *ACM Computing Surveys*, 1999.
[15] T. Joerding, K. Meissner, "Intelligent multimedia presentations in the Web: fun without annoyance", *Computer Network and ISDN Systems*, no. 30, pp. 649-650, 1998.
[16] R. A. Johnson, D. W. Wichern, *"Applied Multivariate Statistical Analysis"*, Prentice Hall, 1992.
[17] G. Kuenning, G. Popek, "Automated hoarding for mobile computers", ACM Symposium on Operating Systems Principles, 1997.
[18] W. L. Martinez, A. R. Martinez, *"Computational Statistics Handbook with Matlab"*, Chapman & Hall/Crc, 2002.
[19] K. Mase, "Intelligent Interfaces for Information Agents: Systems, Experience, Future Challenges", *Lecture Note in Artificial Intelligence*, No. 2446, pp. 10-13, 2002.
[20] S. M. Ross, *"Simulation"*, 2nd edition, Academic Press, 1997.
[21] Y. Saygin, O. Ulusoy, A. Elmagarmid, "Association Rules for Supporting Hoarding in Mobile Computing Environments", IEEE 10th International Workshop on Research Issues on Data Engineering, 2000.
[22] U. Shardanand, P. Maes, "Social Information filtering Algorithms for Automating 'word of Mouth'", CHI'95, 1995.
[23] T. A. Soe, S. Krishnaswamy, S. W. Loke, M. Indrawan, D. Sethi, " AgentUDM: A Mobile Agent Based Support Infrastructure for Ubiquitous Data Mining", 18th International Conference on Advanced Information Networking and Application, 2004.
[24] M. Weiser, "The computer for 21st Century", *Scientific American*, vol. 265, no. 3, pp. 94-104, 1991.
[25] M. Weiser, "Some computer science issues in ubiquitous computing", *Communications of ACM*, vol. 36, no. 7, pp. 75-84, 1993.
[26] H. Wittig, C. Griwodz, "Intelligent media agents in interactive television systems", International Conference on Multimedia Computing and Systems, pp. 182-189, 1995.
[27] V. N. Vapnik, *"The Nature of Statistical Learning Theory,"* New York, Springer-Verlag, 1995.
[28] V. N. Vapnik, *"Statistical Learning Theory,"* New York : Wiley, 1998.
[29] http://www.ubiq.com/weiser.

저 자 소 개

전성해(Sung-Hae Jun)
1993년 : 인하대 통계학과 (학사)
1996년 : 인하대 통계학과 (이학석사)
2001년 : 인하대 통계학과 (이학박사)
2003년 : 서강대학교 컴퓨터학과 (공학박사 수료)
2003년~현재 : 청주대학교 통계학과 전임강사

관심분야 : 데이터마이닝, 데이터공학, 유비쿼터스 컴퓨팅
Phone : 043-229-8205
Fax : 043-229-8432
E-mail : shjun@cju.ac.kr