JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지(J. Korea Inst. Inf. Commun. Eng.) Vol. 20, No. 6 : 1063~1068 Jun. 2016

# 클래스 불균형 데이터를 이용한 나이브 베이즈 분류기 기반의 이상전파에코 식별방법

이한수[1] · 김성신[2*]

# Naive Bayes Classifier based Anomalous Propagation Echo Identification using Class Imbalanced Data

**Hansoo Lee[1] · Sungshin Kim[2*]**

Department of Electrical and Computer Engineering, Pusan National University, Busan 46241, Korea

## 요 약

이상전파에코는 대기 관측을 위해서 사용되는 레이더 전파가 온도나 습도에 의해서 발생하는 이상굴절에 의해서 발생하는 신호로, 지상에 설치된 기상레이더에 자주 발생하는 비기상에코이다. 기상예보의 정확도를 높이기 위해서는 레이더 데이터의 정확한 분석이 필수적이기 때문에 이상전파에코의 제거에 대한 연구가 수행되어 오고 있다. 본 논문에서는 다양한 레이더 관측변수를 나이브 베이지안 분류기에 적용하여 이상전파에코를 식별하는 방법에 대한 연구를 수행하였다. 수집된 데이터가 클래스 불균형 문제를 내포하고 있는 점을 고려하여, SMOTE 기법을 이용하였다. 실제 이상전파에코 발생 사례를 통해, 제안한 방법이 성능을 표출하는 것을 확인하였다.

## ABSTRACT

Anomalous propagation echo is a kind of abnormal radar signal occurred by irregularly refracted radar beam caused by temperature or humidity. The echo frequently appears in ground-based weather radar due to its observation principle and disturb weather forecasting process. In order to improve accuracy of weather forecasting, it is important to analyze radar data precisely. Therefore, there are several ongoing researches about identifying the anomalous propagation echo with data mining techniques. This paper conducts researches about implementation of classification method which can separate the anomalous propagation echo in the raw radar data using naive Bayes classifier with various kinds of observation results. Considering that collected data has a class imbalanced problem, this paper includes SMOTE method. It is confirmed that the fine classification results are derived by the suggested classifier with balanced dataset using actual appearance cases of the echo.

**키워드** : 레이더 데이터 분석, 이상전파에코, 나이브 베이즈 분류기, 클래스 불균형 데이터, SMOTE

Key word : Radar Data Analysis, Anomalous Propagation Echo, Naive Bayes Classifier, Class Imbalanced Data, SMOTE

**Received** 20 May 2016, **Revised** 31 May 2016, **Accepted** 08 June 2016

* **Corresponding Author** Sungshin Kim(E-mail:sskim@pusan.ac.kr, Tel:+82-51-510-2374)
Department of Electrical and Computer Engineering, Pusan National University, Busan 46241, Korea

Open Access http://dx.doi.org/10.6109/jkiice.2016.20.6.1063 print ISSN: 2234-4772 online ISSN: 2288-4165

## Ⅰ. 서 론

기상 예보를 수행함에 있어서 강수의 위치와 이동방향, 강수강도를 예측할 수 있는 데이터를 수집하는 것은 매우 중요한 절차이다. 기상 레이더는 대기 중에 존재하는 대상에 대한 관측 결과를 탐지된 좌표에 저장하는 형식으로 레이더 데이터를 생성한다. 생성된 레이더 데이터에는 반사도, 스펙트럼 폭, 도플러 속도 등과 같은 다양한 정보가 포함되어 있으며, 이는 강수와 관련된 유용한 정보들이 레이더 데이터에 관측되고 저장된다는 것을 의미한다.

하지만 레이더가 탐지영역 내에서 강수 신호만을 선택적으로 관측할 수 없기 때문에, 레이더 데이터 내부에는 불가피하게 비기상에코가 포함되어 있다. 특히 이상전파에코의 경우 대기 중 기온과 습도 분포에 의해서 레이더 빔이 굴절되면서 지표면의 물체들을 대기 중의 반사체로 잘못 인지하여 나타나는 레이더 신호로, 강수량 추정을 수행하는 데 있어서 악영향을 미치는 요소이다 [1]. 이를 자동으로 분류 및 제거하기 위해서 전 세계적으로 다양한 연구가 수행되어 오고 있다.

분류 기법은 패턴 인식에 있어서 중요한 요소 중 하나이며, 다양한 분야에서 여러 가지 성공적으로 개발된 기법들이 적용되어 있다. 대표적인 분류 기법의 종류는 의사 결정 트리, 인공신경망, 베이지안 네트워크, 서포트 벡터 머신 등이 있다. 이들은 모두 지도학습(supervised learning) 범주에 속하는 방법이며, 지도 학습을 이용한 분류기는 주어진 데이터의 클래스 분포가 균일하다고 가정하고 학습을 수행하기 때문에, 충분한 양의 학습 데이터를 확보하는 것은 분류기의 성능 향상에 있어서 중요한 요소가 된다. 하지만 실제 문제에 분류 기법을 적용하기 위해서 측정하는 데이터는 충분한 양의 학습 데이터를 형성하기가 힘들기 때문에 불가피하게 분류기의 성능 저하가 발생할 수 있다 [2].

따라서 본 논문에서는 기상 레이더 관측 데이터에 포함되어 있는 이상전파에코를 분류 및 제거하기 위하여 나이브 베이즈 분류기[3]를 구현하고, 성능 향상을 위해서 over-sampling 기법 중 하나인 SMOTE (Synthetic Minority Over-sampling TEchnique) [4] 기법을 이용하였다. 원시 레이더 데이터를 전처리 및 클러스터링 방법을 이용해 군집화한 후 특성을 추출하고, 실제 사례를 바탕으로 성능을 검증하였다.

본 논문의 구성은 다음과 같다. 2장에서는 이상전파에코의 특성에 대해서 설명한다. 3장에서는 클래스 불균형 문제와 이를 해결하기 위해서 본 논문에서 적용한 SMOTE 기법에 대해서 설명한다. 4장에서는 나이브 베이즈 분류기와 본 논문에서 입력 변수로 사용한 특성에 대해서 설명한다. 5장에서는 실제 이상전파에코 표출 사례를 바탕으로 하여 진행한 실험에 대해 설명한 후 결론 및 향후 연구 방향에 대해서 논한다.

## Ⅱ. 이상전파에코

전파를 이용하는 원격탐사 장비의 특성에 의해 기상 레이더는 대기 상태에 따라 빔 전파 경로가 굴절되는 현상이 발생한다. 대기 중 레이더 빔의 전파는 기온, 기압, 수증기의 분포에 따라 변하게 되며, 전파의 굴절 방향과 굴절 정도에 따라 관측효율이 변하게 된다. 레이더는 대기 상태가 정상적인 상태에서 동작하는 것으로 간주하고 관측을 수행하기 때문에, 전파 굴절이 발생할 경우에는 대기 중 반사체의 고도가 실제보다 높거나 낮게 추정되는 사례가 발생할 수 있다.

레이더 빔의 곡도(degree of curvature)는 굴절률(index of refraction)을 통해서 정의할 수 있다. 계산의 편의성을 위해서 굴절도(refractivity) $N$을 식 (1)과 같이 근사하여 정의할 수 있다.

$$N=(n-1)\times 10^6=\frac{77.6}{T}\times\frac{p+4810e}{T} \tag{1}$$

여기서 $n$은 굴절률(index of refraction)을, $p$는 기압(total pressure)을, $e$는 수증기 분압(partial pressure of water)을, 그리고 $T$는 절대 온도(Kelvin)를 의미한다 [5]. 레이더 빔의 곡률 반경(radius of curvature of radar beam) $r$은 식 (2)에 나타난 것처럼 높이 $h$에 대한 $N$의 변화도를 이용하여 근사치를 구할 수 있다 [6].

$$\frac{r}{R_e}=k_e\approx\frac{1}{1+(dN/dh)/157} \tag{2}$$

여기서 $R_e$는 실제 지구의 반경(true Earth radius)을, $k_e$는 지구 유효 반지름 (effective earth radius factor)을

나타낸다. 지표면 근처에서의 굴절률 변화도는 대략 -39N/km로, 식 (2)에 이를 대입하면 지구 유효 반지름의 값은 $k_e = 4/3$으로 도출할 수 있다.

식 (2)에 나타난 $k_e$를 이용해서 전파의 경로에 따라 발생하는 빔 전파 경로를 정상굴절, 과소굴절, 과대굴절, 빔 갇힘의 네 가지로 구분할 수 있다.

1) 과소굴절 $k_e \geq \frac{4}{3}$

2) 정상굴절 $k_e = \frac{4}{3}$

3) 과대굴절 $0 \leq k_e \leq \frac{4}{3}$

4) 빔 갇힘 $k_e < 0$

정상굴절을 제외한 나머지 굴절된 레이더 전파에 의해서 발생하는 비기상에코를 통칭하여 이상전파에코(anomalous propagation echo)라고 한다. 특히 빔 갇힘 현상에 의해서 레이더가 지표면에 존재하는 물체들을 반사체로 인식하여 관측 결과를 저장할 경우 강한 반사도(reflectivity)를 가지는 에코가 발생할 수 있는데, 이는 정량적 강우량 산정(quantitative precipitation estimation) 과정에서 문제를 발생시킬 수 있다. 따라서 정확한 기상 예보를 위해서는 이를 반드시 제거하여야 할 필요가 있다 [7].

## Ⅲ. 클래스 불균형 문제

지도 학습을 이용한 분류기는 주어진 데이터의 클래스 분포가 균일하다고 가정하고 학습을 수행하기 때문에, 충분한 양의 학습 데이터를 확보하는 것은 분류기의 성능 향상에 있어서 중요한 요소가 된다. 하지만 실제로 측정된 데이터는 클래스의 불균형 분포가 빈번하게 발생하기 때문에 이를 고려하지 않고 학습을 수행한다면 구현된 분류기의 성능이 저하되는 문제점이 발생할 수 있다. 이러한 문제는 특히 측정된 데이터의 특성 공간이 고차원일 경우 부각되는데, 이를 극복하기 위해 over-sampling, under-sampling 등과 같은 다양한 기법들이 개발되어 실제 시스템에 적용되어 오고 있다.

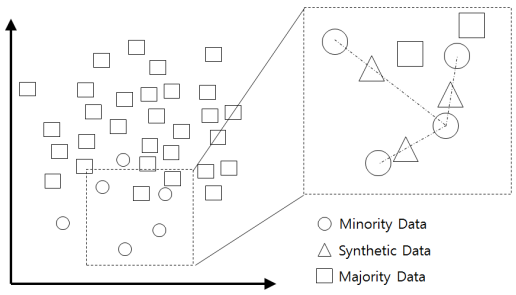


**Fig. 1** Principles of SMOTE method

본 논문에서는 over-sampling 기법 중 하나인 SMOTE (Synthetic Minority Over-sampling TEchnique) 기법을 이용하였다. SMOTE 기법은 over-sampling 기법 중에서 자주 쓰이던 복원추출 기법을 대신하기 위해서 제안된 방법으로, k-nearest neighbor 알고리즘을 소수 클래스 데이터에 적용하여 특성 공간 내에 합성 데이터를 생성하여 다수 클래스와 소수 클래스 간의 수치적 균형을 맞추는 방법이다.

수식 (3)에 나타난 것과 같이, k-nearest neighbor를 이용하여 선정된 각 소수 클래스 데이터와 인접한 데이터 중에서 임의적으로 두 개의 소수 클래스 데이터를 택한 후, 그 사이에 수식 (4)에 나타난 것과 같이 균일 분포 확률 변수(uniformly distributed random variable)를 곱하여 합성 데이터를 생성한다.

$$D_{\text{synthetic}} = D_{\text{origin}} + D_{\text{k}^{\text{th}}\text{neighbor}} \times P_{\text{uniform}}(x) \quad (3)$$

$$P_{\text{uniform}}(x) = \begin{cases} 1, & \text{for } x \in [0,1] \\ 0, & otherwise \end{cases} \quad (4)$$

그림 1은 SMOTE기법의 예시를 그림 형태로 나타낸 것이다. 여기서 원형 기호는 소수 클래스 데이터를, 사각형 기호는 다수 클래스 데이터를, 그리고 삼각형 기호는 합성 데이터를 나타낸다. 그림 1에서 확인할 수 있는 것과 같이 소수 클래스 데이터에 k-nearest neighbor 기법을 적용하여 같은 클래스를 가지는 세 개의 새로운 합성 데이터가 생성된 것을 확인할 수 있다.

## Ⅳ. 나이브 베이즈 분류기

베이즈 정리(Bayes' theorem)는 학습 데이터의 정보

(likelihood)와 사전 확률(prior)를 이용하여 사후 확률(posterior)를 도출하는 방식으로 확률변수들 간의 의존관계를 조건부 확률로서 기술한 확률모델이다 [8, 9]. 베이즈 정리는 수식 (5)에 나타난 것과 같으며, 의사 결정 시스템(decision support system)을 구성하는 데 있어서 지식(knowledge)을 모델링하는데 주로 사용한다.

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \tag{5}$$

여기서 $C$는 데이터의 클래스를, $X$는 $x_1,x_2,\cdots,x_n$를 원소로 가지는 벡터 형태의 데이터를 의미하며, 각 원소는 데이터가 가지는 특성을 의미한다. 베이즈 정리가 분류 기법이 필요한 다양한 분야에서 높은 정확도를 가지기 위해서는 특성 변수간의 정확한 full joint probability distribution이 도출되어야 한다. 하지만, 이 조건을 만족하기 위해서는 충분한 학습 데이터가 요구되지만, 실제 문제를 해결하는 데 있어서 어려움이 존재한다.

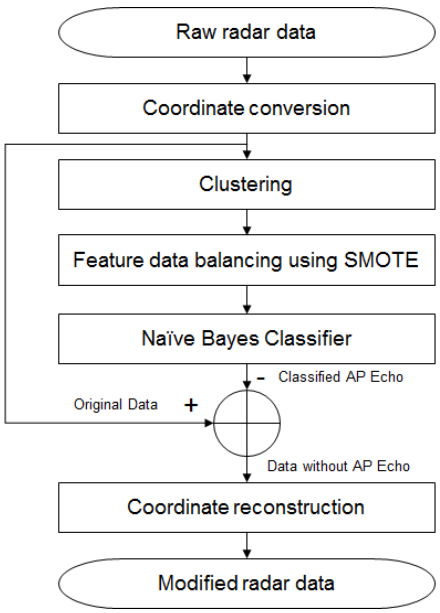


**Fig. 2** System overview

이러한 문제를 해결하기 위해, 주어진 특성들이 서로 독립이라는 가정을 바탕으로 예측을 수행하는 나이브 베이즈 기법이 제안되었다. 이 가정은 대부분의 현실 세계의 문제를 해결하는 데 있어서 명백한 거짓 가정을 바탕으로 만들어졌음에도 불구하고 특히 이진분류(binary classification)의 경우 우수한 성능을 보여주는 데 [10], 이러한 모순은 이진분류의 경우 분류 추정이 함수 추정에서 단지 함수의 부호를 추정하는 것과 같다는 것을 통해서 설명될 수 있다. 이러한 독립 특성 가정 때문의 속성의 수가 많을 때 각 속성의 모수들은 서로 분리해서 학습을 수행할 수 있으며, 이는 학습을 간단하게 만든다 [10, 11].

본 논문에서는 그림 2에 나타난 것과 같이 이상전파에코를 식별 및 제거하기 위한 시스템을 구현하였으며, 대략적인 순서는 다음과 같다. 먼저, 원시 레이더 데이터가 구좌표계로 저장되기 때문에, 직관적인 분석을 위하여 직교좌표계로 변환한다. 그리고 계층적 클러스터링(hierarchical clustering)기법을 변환된 데이터에서 반사도(reflectivity)와 도플러 속도(Doppler velocity) 데이터에 적용하여 클러스터를 만든다. 여기서 생성된 클러스터에서 여섯 가지 통계적, 위치적 특성을 추출하여 나이브 베이즈 분류기에 입력으로 사용한다. 분류기를 통해서 도출된 이상전파에코를 원본 레이더 영상과 비교하여 제거하고, 저장된 데이터의 통일성을 유지하기 위해서 구좌표계로 좌표변환을 수행하게 된다.

본 논문에서 나이브 베이즈 분류기를 위해서 사용한 여섯 가지 특성은 고도(altitude), 평균 반사도(mean reflectivity), 최대 반사도(maximum reflectivity), 최소 도플러 속도(minimum Doppler velocity), 최소 도플러 속도(maximum Doppler velocity), 평균 도플러 속도(mean Doppler velocity)이다. 이는 대체적으로 이상전파에코가 낮은 고도에서 관측되며 불균등한 반사도 분포를 가지고, 0에 가까운 도플러 속도를 가지는 형태로 레이더 영상에 표출된다는 경험 기반 지식(experience-based knowledge)을 토대로 선정한 것이다. 이들 특성을 바탕으로 나이브 베이즈 분류기를 구현하기 위하여 수식 (5)의 likelihood를 수식 (6)과 같은 형태로 도출할 수 있다.

$$P(X|C) = P(x_1, \cdots, x_6|C) = \prod_{k=1}^{6} P(x_k|C) \tag{6}$$

여기서 $x_1$~$x_6$은 위에서 언급한 여섯 가지 특성에 대응되는 변수를 의미하며, 서로 독립적이라는 가정을 하

였기 때문에 각 특성의 조건부 확률의 곱으로 표현할 수 있다. 수식 (5)와 수식 (6)을 합하여 본 논문에서 사용한 여섯 가지 특성을 고려하여  정리하면 수식 (7)과 같이 표현할 수 있다.

$$P(C|X) = \frac{\prod_{k=1}^{6} P(x_k|C)P(C)}{P(X)} \quad (7)$$

## Ⅴ. 실험결과 및 결론

실제 이상전파에코가 발생한 경우의 기상 레이더 데이터를 바탕으로 제안한 알고리즘을 검증하였다. 그림 3은 해안 지역에 인접한 기상 레이더에서 발생한 이상전파에코 발생 사례이며, 그림 3의 (a)는 원본 레이더 영상을, 그림 3의 (b)는 제안한 방법에 의해서 이상전파에코 영역이 제거된 후의 레이더 영상을, 그림 3의 (c)는 이상전파에코만을 표출한 레이더 영상을 나타낸 것이다. 그림 3을 바탕으로 판별하였을 때, 제안한 나이브 베이즈 분류기가 이상전파에코를 성공적으로 제거하는 것을 확인할 수 있다.

그리고 본 논문에서 제안한 SMOTE 기법을 적용한 나이브 베이즈 분류기와 적용하지 않은 나이브 베이즈 분류기의 결과를 비교하기 위해서 세 개의 기상 레이더를 선정하여 실험을 수행하였다. 정확한 비교를 위해서 k-fold cross validation을 수행하였으며, k의 값은 10으로 설정하였다. 표 1에 나타난 것과 같이 일반적인 나이브 베이즈 분류기를 적용하였을 경우에는 평균 68.52%의 정확도를 보였으며, SMOTE 기법을 적용하였을 경우에는 평균 73.30%의 정확도를 보였다. 이 결과를 바탕으로 SMOTE 기법을 적용하여 클래스 불균형 문제를 해결하였을 때, 정확도가 개선된 분류 결과를 얻을 수 있다는 것을 확인할 수 있었다.

**Table. 1** Performance comparison

| Radar site | Naive Bayes | Naive Bayes + SMOTE |
|---|---|---|
| Site # 1 | 69.53% | 75.08% |
| Site # 2 | 67.19% | 71.43% |
| Site # 3 | 68.83% | 73.38% |

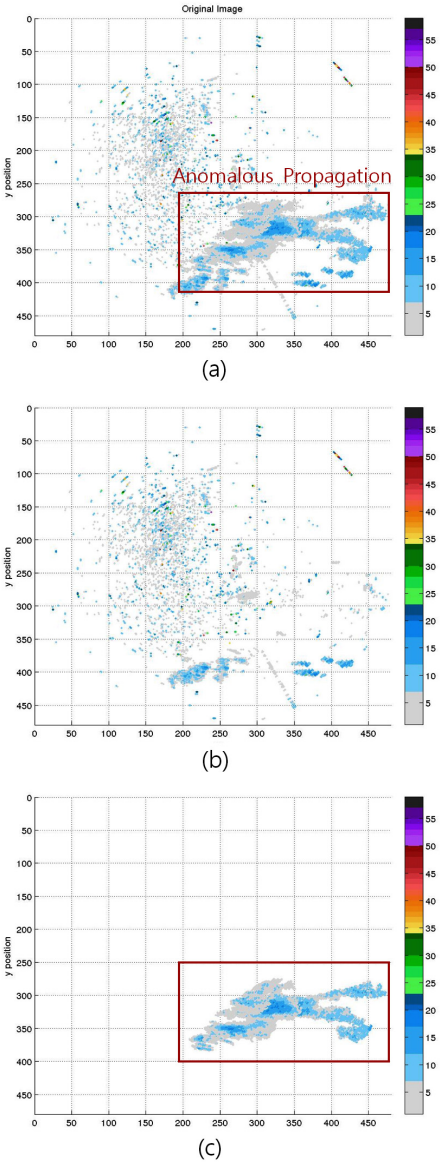


**Fig. 3** Experimental result: (a) Original radar image, (b) modified radar image, (c) identified anomalous propagation echo image

본 논문에서는 기상 예보를 수행하는 과정에서 기상 레이더 데이터에서 발생하는 비강수에코 중 하나이며 강수에코와 그 특성이 유사해 기상 예보 정확도를 높이기 위해서 필수적으로 제거할 필요가 있는 이상전파에코의 식별 및 제거 알고리즘을 나이브 베이즈 분류기를

이용하여 구성하는 방법에 대해 제안하였다. 그리고 클래스 불균형 문제를 해결하기 위하여 SMOTE 기법을 적용하는 방법에 대해서도 연구를 수행하였다. 실제 사례를 바탕으로 실험한 결과 제안한 SMOTE 기법 기반의 나이브 베이즈 분류기가 이상전파에코를 성공적으로 식별하는 것을 확인할 수 있었다. 향후 본 논문의 연구 결과를 토대로 기상 레이더의 위치적 특성에 따른 최적화 방법에 대한 연구를 수행하고자 한다.

## ACKNOWLEDGMENTS

This work was supported by a 2-Year Research Grant of Pusan National University and was supported by BK21PLUS, Creative Human Resource Development Program for IT Convergence .

## REFERENCES

[ 1 ] M. Grecu and W. F. Krajewski, “An efficient methodology for detection of anomalous propagation echoes in radar reflectivity data using neural networks,” *Journal of atmospheric and oceanic technology*, vol. 17, no. 2, pp. 121-129, Feb. 2000.

[ 2 ] Y. Sun, A. K. Wong, and M. S. Kamel, “Classification of imbalanced data: a review,” *International journal of pattern recognition and artificial intelligence*, vol. 23, no. 04, pp. 687-719, Jun. 2009.

[ 3 ] K. P. Murphy, *Machine learning: a probabilistic perspective*, Cambridge, MIT press, 2012.

[ 4 ] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” *Journal of artificial intelligence research*, vol. 16, pp. 321-357, Jun. 2002.

[ 5 ] R. J. Doviak and D. S. Zrnic, *Doppler Radar & Weather Observations*, Cambridge, Academic press, 2014.

[ 6 ] G. Brussaard and P. A. Watson, *Atmospheric modelling and millimetre wave propagation*, New York, Springer Science & Business Media, 1995.

[ 7 ] S. Moszkowicz, G. J. Ciach and W. F. Krajewski, “Statistical detection of anomalous propagation in radar reflectivity patterns,” *Journal of atmospheric and oceanic technology*, vol. 11, no. 4, pp. 1026-1034, Aug. 1994.

[ 8 ] D. Heckerman, “Bayesian networks for data mining,” *Data mining and knowledge discovery*, vol. 1, no. 1, pp. 79-119, Mar. 1997.

[ 9 ] R. E. Neapolitan, *Learning bayesian networks*, New Jersey, Pearson Prentice Hall, 2004.

[10] A. McCallum and K. Nigam, “A comparison of event models for naive bayes text classification,” *in AAAI-98 workshop on learning for text categorization*, vol. 752, Jul. 1998.

[11] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, *Machine learning: an artificial intelligence approach*, New York, Springer Science & Business Media, 2013.

**이한수(Hansoo Lee)**

2010년 부산대학교 전자전기공학부 공학사
2012년 부산대학교 전자전기공학과 공학석사
현재 부산대학교 전자전기컴퓨터공학과 공학박사
※관심분야 : 지능시스템, 데이터마이닝, 예측 및 분류, 빅데이터

**김성신(Sungshin Kim)**

1984년 연세대학교 전기공학과 공학사
1986년 연세대학교 전기공학과 공학석사
1996년 Georgia Institute of Technology, 전기및컴퓨터공학부 공학박사
현재 부산대학교 전기컴퓨터공학부 교수
※관심분야 : 지능시스템, 지능로봇, 고장진단 및 예측