응용통계연구 (2009)
22(4), 759–770

혼합 모델 및 다중 가설 검정을 이용한 신호와 잡음의 분류

박해상[1] · 유시원[2] · 전치혁[3]

[1]포항공과대학교 산업경영공학과, [2]NHN 고객만족추진팀, [3]포항공과대학교 산업경영공학과

(2009년 6월 접수, 2009년 7월 채택)

요 약

본 논문은 신호와 잡음이 혼합된 관측치로부터 신호 관측치를 분류하는 문제를 다룬다. 잡음은 가우시안 분포를 따르고 신호는 감마 분포를 따른다고 가정할 때 관측치의 분포는 가우시안과 감마의 혼합 분포를 따르게 된다. EM 알고리즘을 통해 혼합 모델의 모수를 추정하고 신호 및 잡음을 분류하는 것을 다중 가설 검정으로 간주하여 베이즈 오류를 바탕으로 분류를 위한 경계치를 설정한다. 제안하는 방법을 분광 데이터에 근거하여 철강 제품에서 개재물 유무를 검출하는 문제에 적용하였고 별도의 시뮬레이션 데이터를 통해 성능의 우수성을 보였다.

주요용어: 신호, 잡음, 혼합 모델, EM 알고리즘, 다중 가설 검정.

1. 서론

어떤 물질의 특성치를 측정하기 위해 계측 장비를 사용하는 경우 원하는 신호(Signal)뿐만 아니라 잡음(Noise)이 포함되는 것이 일반적이다. 따라서 실험 데이터로부터 잡음을 제거하여 추후 분석을 수행하여야 하는데, 신호의 강도에 대한 분포가 알려져 있다 하더라도 신호와 잡음이 미지의 비율로 혼합되어 있어 이 둘을 분류하는 것은 쉽지 않다.

신호 탐지론(Signal detection theory)에서 많은 경우에 잡음의 분포가 가우시안 분포를 따른다고 가정하여 분석한다 (Abdi, 2007). 신호 또한 가우시안 분포를 따른다고 가정하는 경우가 많으나 신호의 특성이 다양하므로 가우시안 이외의 분포 형태를 고려할 필요가 있다.

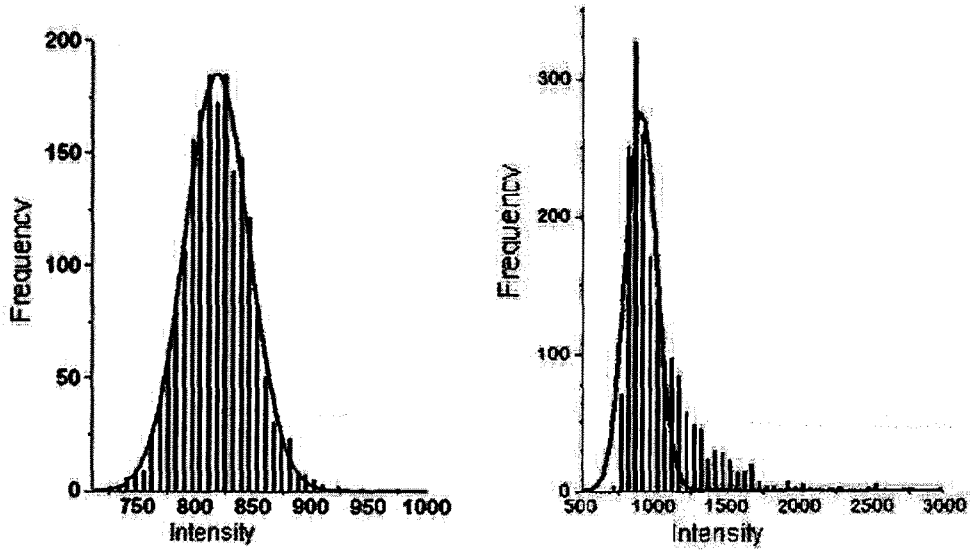
이에 본 연구에서는 데이터의 분포 및 특성에 근거한 보다 효과적인 고정 경계치를 찾는 것을 목적으로, 혼합 모델을 이용한 신호와 잡음을 분류하는 절차를 제안한다. 신호와 잡음이 혼합된 데이터의 기저 분포로 가우시안-감마(Gaussian-Gamma mixture)를 가정한다. EM(Expectation-Maximization) 알고리즘으로 각각의 분포를 추정한 후 추정된 모델의 정보를 이용하여 다중 가설 검정(Multiple testing)에서 사용되는 오발견율(False discovery rate: FDR)과 베이즈 오류(Bayes error)에 기반한 신호와 잡음 분류 방법에 대해 제안한다. 또한 실제 계측데이터를 제안한 방법으로 분류하고 모의 실험을 통해 기존 연구와 성능을 비교하고 제안 방법의 특징에 대해 논의할 것이다.

2. 관련 연구

2.1. 개재물 분석

본 연구는 철강제품에 포함된 개재물(Inclusion)의 유무를 판정하기 위하여 OES(Optical emission spectroscopy) 데이터를 이용하여 분석을 수행한다. OES는 시료에 고전압의 스파크를 가해서 방출되

[3]교신저자: (790-784) 경상북도 포항시 남구 효자동 산31, 포항공과대학교 산업경영공학과, 교수.
E-mail: chjun@postech.ac.kr



(a) 순수 철강 시료 (b) 개재물 포함 시료

그림 **2.1.** 개재물 유무에 따른 OES 데이터의 분포

는 자외선이나 가시영역의 빛을 분광기를 이용하여 측정하는 장치이다. OES 데이터는 개별 스파크에 대해 발생한 신호의 강도(Signal intensity) 값을 가진다. 각 원소 채널 별로 데이터의 값을 분석하면 특정 스파크가 일어난 지점에 어떠한 개재물 유형이 존재하는지 알 수 있다 (Kuss 등, 2005).

그림 2.1(a)는 개재물이 없는 순수한 철강 시료의 OES 데이터를 히스토그램으로 나타낸 것이다. 이 부분이 잡음에 해당하는 데이터로 그림과 같이 가우시안 분포에 근사한 모양을 나타낸다. 반면 개재물이 혼합된 철강 시료에서 검출한 OES 데이터를 히스토그램으로 나타내면 그림 2.1(b)의 우측과 같이 비대칭 분포를 가진다. 이는 잡음으로부터 유래한 가우시안 분포와 개재물 분포가 혼합된 형태이다 (Kuss 등, 2002).

기존 연구를 살펴보면 Kuss 등 (2005)는 데이터의 평균으로부터 표준편차의 3배만큼 떨어진 값을 경계로 하여 그보다 강도가 큰 영역을 신호로 분류하는 방법을 제안하였다. 또한 Shin과 Bae (2003)는 가우시안 분포의 왼쪽 부분을 오른쪽에 대칭시킨 후 겹치는 부분을 잡음에서 유래한 데이터로 간주하여 제거하고, 남은 부분의 면적을 이용하여 신호의 빈도를 구하는 방법을 제안하였다.

그러나 이러한 방법은 휴리스틱 방법으로 신호를 분류하는 것 외에는 데이터에 대한 다른 정보를 알 수 없다. 또한 데이터의 모양 및 신호의 분포가 다양해 질 경우 대처하기 어렵다는 문제점이 있다. 따라서 통계적 방법에 기반한 신호 및 잡음의 분류 과정을 구축할 필요가 있다.

2.2. EM 알고리즘

EM 알고리즘은 불완전 데이터로부터 확률 분포 모수들의 최우 추정치(Maximum likelihood estimator)를 찾기 위한 방법으로 E(Expectation)-step과 M(Maximization)-step을 반복적으로 수행하는 과정을 통한다. 기존의 모수 추정치가 존재할 때 E-step에서는 그 값을 바탕으로 비관측 변수의 기대치를 추정하고, M-step에서는 추정된 비관측 변수를 이용해 우도함수를 최대화하는 새로운 모수를 추정한다

표 **2.1.** m개의 가설 검정 시 발생하는 결과

	Accept null	Reject null	Total
Null true	U	V	m_0
Alternative true	T	S	m_1
Total	W	R	m

(Bishop, 2006). 본 연구에서는 가우시안-감마 혼합 모델을 추정하기 위해 EM 알고리즘을 이용한다.

2.3. 오발견율(FDR)

다중 가설 검정은 두 개 이상의 가설을 동시에 검정하는 것을 말한다. 이 경우 각각의 검정이 제 1종 오류(Type I error)와 2종 오류(Type II error)를 가지기 때문에 검정 전체의 오류율을 측정하는 것이 명확하지 않다. 이 때 사용할 수 있는 방법으로 FWER(Familywise error rate)이 있다 (Hochberg와 Tamhane, 1987). 이는 각 검정마다 제1종 오류를 수준 α에서 통제하는 것이 아니라 전체 검정의 FWER을 수준 α로 통제하는 것이다. 그러나 FWER은 검정의 수가 증가할수록 매우 보수적인 결과를 나타낸다.

Benjamini와 Hochberg (1995)는 다중 가설 검정의 오류 단위로 오발견율(FDR)을 도입하였다. 이는 전체 기각된 귀무가설(Null hypothesis) 중 양성 오류(False positive)의 기대 비율로 정의된다. 표 2.1은 m개의 가설 검정 시 발생하는 결과를 나타낸 표이다. V는 귀무가설 중 기각된 가설의 수를 의미하고 R은 전체가설 중 기각된 가설의 수를 의미한다. 이 때, 오발견율을 나타내면 식 (2.1)과 같다.

$$\mathrm{FDR} = E\left[\frac{V}{R}|R>0\right]\Pr(R>0). \tag{2.1}$$

FDR은 FWER에 비해 제 1종 오류를 덜 엄격하게 통제함으로써 가설 검정 시 보다 큰 검정력을 갖는다. FDR은 기각된 가설이 없을 경우($R=0$) 정의되지 않으므로 Storey (2002)는 식 (2.2)와 같이 pFDR(Positive false discovery rate)을 제시하였다. 'Positive'는 검정 시 양성인 결과가 발생한 경우를 조건으로 함을 의미한다.

$$\mathrm{pFDR} = E\left[\frac{V}{R}|R>0\right]. \tag{2.2}$$

FNR(False nondiscovery rate)은 FDR에 대칭되는 개념으로 기각되지 않은 가설 중 음성 오류(False negative)의 기대 비율로 다음 식 (2.3)과 같이 정의된다. T는 대립가설(Alternative hypothesis) 중 기각되지 않은 가설의 수를 의미하고 W는 전체가설 중 기각되지 않은 가설의 수를 의미한다. pFNR(Positive false nondiscovery rate)은 FNR에서 전체가설 중 기각되지 않은 가설이 존재하는 경우($W\neq 0$)를 조건으로 한다.

$$\mathrm{FNR} = E\left[\frac{T}{W}|W>0\right]\Pr(R>0). \tag{2.3}$$

Benjamini와 Hochberg (1995)은 일련된 p값을 이용하는 방법으로 FDR을 통제하는 방법을 제안하였고 Storey (2002, 2003)는 pFDR, pFNR 개념과 베이지안 오류를 연관지어 분류 이론을 연구하였다.

3. 제안 방법

본 연구에서는 위와 같은 다중 가설 검정과 사후 확률 사이의 연관성을 이용하여 고정 경계치를 갖는 분류 규칙을 설정하고, 그에 따라 신호와 잡음을 분류하는 방법을 제안한다.

3.1. 가정

m개의 측정 데이터가 있고 X_i를 i번째 강도의 관측치라 할 때($i = 1, \ldots, m$) 다음과 같은 다중 가설 검정을 생각한다.

- 귀무가설: X_i가 잡음 분포에서 생성 ($H_i = 0$)
- 대립가설: X_i가 신호 분포에서 생성 ($H_i = 1$)

F_0은 평균 μ_0, 분산 σ_0^2인 가우시안 분포를 따르는 잡음과 관련된 분포이고 G_1은 형태모수(Shape parameter) α_1, 척도모수(Scale parameter) β_1인 감마 분포를 따르는 신호와 관련된 분포이다. 일반적으로 신호 또한 가우시안 분포를 따른다고 가정하나 신호의 특성이 다양하고 분포에 대한 정보를 얻기 힘든 경우가 많다. 감마 분포는 가우시안 분포에 비해 더 다양한 형태를 가정할 수 있어 본 연구에서는 신호가 감마 분포를 따른다고 가정한다. f_0, g_1을 이에 해당하는 서로 독립인 확률 밀도 함수로 정의한다. π_0, π_1을 각각 $H_i = 0$과 $H_i = 1$에 대한 사전 확률이라고 할 때 ($\pi_0 + \pi_1 = 1$) 다음이 성립한다.

$$X_i | H_i = 0 \sim F_0 \tag{3.1a}$$

$$X_i | H_i = 1 \sim G_1 \tag{3.1b}$$

계측 데이터 중 임의로 선택한 데이터는 위 가우시안 분포와 감마 분포의 혼합 분포를 따르므로 X_i의 확률밀도함수는 다음과 같다.

$$f(x) = \pi_0 f_0(x|\mu_0, \sigma_0^2) + (1 - \pi_0) g_1(x|\alpha_1, \beta_1). \tag{3.2}$$

3.2. EM 알고리즘을 이용한 분포 추정

변수 z_i를 식 (3.3)과 같이 정의한다.

$$z_i = \begin{cases} 1, & z_i \text{ belongs to signal,} \\ 0, & \text{otherwise.} \end{cases} \tag{3.3}$$

z_i들을 관측할 수 있다면 신호-잡음 분류일 때는 단순하게 되나, 이는 관측 불능 변수이다. 그러나 z_i들이 주어진다고 할 때 X_i의 확률밀도함수가 유도되므로 $X_i = x_i$들이 관측될 때 관련 모수 θ와 π_0에 대한 로그우도함수(Log likelihood function)는 식 (3.4)가 된다. $\mathbf{x}$는 $\mathbf{x} = [x_1, x_2, \ldots, x_m]'$과 같이 정의된다.

$$l(\theta, \pi_0, z|\mathbf{x}) = \sum_{i=1}^{m} \left[(1 - z_i) \log \left(\pi_0 f_0(x_i|\mu_0, \sigma_0^2) \right) + z_i \log \left((1 - \pi_0) g_1(x_i|\alpha_1, \beta_1) \right) \right]. \tag{3.4}$$

기존의 모수 추정치가 존재할 때 E-step에서는 그 값을 바탕으로 z_i의 기대치를 추정하고, M-step에서는 추정된 $\hat{z}_i$를 이용하여 우도함수를 최대화하는 새로운 모수를 추정하며 이러한 과정을 반복한다. 이를 정리하면 아래와 같다.

Step 0. (Initialization) $\theta = (\mu_0, \sigma_0^2, \alpha_1, \beta_1)$, π_0초기화

Step 1. (E-step) 모수 추정치를 바탕으로 하여 $\hat{z}_i$를 산출

$$\hat{z}_i = \frac{(1 - \hat{\pi}_0) g_1(x_i|\hat{\alpha}_1, \hat{\beta}_1)}{\hat{\pi}_0 f_0(x_i|\hat{\mu}_0, \hat{\sigma}_0^2) + (1 - \hat{\pi}_0) g_1(x_i|\hat{\alpha}_1, \hat{\beta}_1)} \tag{3.5}$$

Step 2. (M-step) $\hat{z}_i$에 기반하여 모수를 새롭게 추정

$$m_0 \leftarrow \sum_{i=1}^{m}(1-\hat{z}_i) \tag{3.6}$$

$$\pi_0 \leftarrow \frac{m_0}{m} \tag{3.7}$$

$$\hat{\mu}_0 \leftarrow \sum_{i=1}^{m}\frac{(1-\hat{z}_i)x_i}{m_0} \tag{3.8}$$

$$\hat{\sigma}_0^2 \leftarrow \sum_{i=1}^{m}\frac{(1-\hat{z}_i)(x_i-\hat{\mu}_0)^2}{(m_0-1)} \tag{3.9}$$

$\hat{\alpha}_1$, $\hat{\beta}_1$을 식 (3.11), (3.12)로부터 산출.

M-step에서 가우시안 분포의 모수는 위의 식으로 추정 가능하나, 감마 분포의 형태모수와 척도모수는 다음 식 (3.10), (3.11), (3.12)를 이용하여 수치적인 기법으로 추정해야 한다.

$$\ln\alpha_1 - \psi(\alpha_1) = \ln\left(\frac{\sum_{i=1}^{m}\hat{z}_i x_i}{\sum_{i=1}^{m}\hat{z}_i}\right) - \frac{\sum_{i=1}^{m}\hat{z}_i\ln(x_i)}{\sum_{i}^{m}\hat{z}_i}, \tag{3.10}$$

여기서 $\psi(\alpha_1)$은 Digamma 함수로써 $\Gamma(\alpha_1)$을 감마함수라 할 때 $\psi(\alpha_1) = \Gamma'(\alpha_1)/\Gamma(\alpha_1)$으로 산출된다. Newton-Ralphson 반복 기법으로 형태모수의 식을 전개하면 k번째 반복시 수치해는 아래와 같다 (Choi와 Wette, 1969).

$$\hat{\alpha}_1^k = \hat{\alpha}_1^{k-1} - \frac{\ln\left(\hat{\alpha}_1^{k-1}\right) - \psi\left(\hat{\alpha}_1^{k-1}\right) - \left[\ln\left(\sum_{i=1}^{m}\hat{z}_i x_i / \sum_{i=1}^{m}\hat{z}_i\right) - \sum_{i=1}^{m}\hat{z}_i\ln(x_i)/\sum_{i=1}^{m}\hat{z}_i\right]}{1/\hat{\alpha}_1^{k-1} - \psi'(\hat{\alpha}_1^{k-1})} \tag{3.11}$$

$\hat{\alpha}_1$이 수렴할 때까지 반복 수행하여 $\hat{\alpha}_1$ 산출 후 이를 이용하여 $\hat{\beta}_1$를 다음 식과 같이 산출한다.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{m}\hat{z}_i x_i}{\hat{\alpha}_1\sum_{i=1}^{m}\hat{z}_i}. \tag{3.12}$$

3.3. 베이즈 오류에 근거한 신호와 잡음 분류

3.2절에서 산출된 z_i값을 바탕으로 신호와 잡음을 분류할 수 있으나 적절한 경계치를 정하는 것이 용이하지 않다. 본 연구에서는 경계치 설정을 위해 베이즈 오류를 도입하고자 한다. 3.1절에서 수립한 귀무가설을 기각하는 영역을 Γ라 두면 기각역과 관련한 pFDR과 pFNR은 다음 식으로 표현된다.

$$\text{pFDR}(\Gamma) = \Pr(H_i = 0 | X_i \in \Gamma) \tag{3.13a}$$

$$\text{pFNR}(\Gamma) = \Pr(H_i = 1 | X_i \notin \Gamma) \tag{3.13b}$$

이를 분류 규칙의 관점에서 보면 $X_i = x_i \in \Gamma$이면 x_i를 신호로 분류하고 $X_i = x_i \notin \Gamma$이면 x_i를 잡음으로 분류하는 것이 된다. 이 때 표 3.1과 같은 오분류에 대한 벌점(Penalty)을 생각하였을 경우 총 오분류 벌점 또는 베이즈 오류는 다음 식이 된다.

$$\text{BE}(\Gamma) = (1-\lambda)\Pr(X_i \in \Gamma, H_i = 0) + \lambda\Pr(X_i \notin \Gamma, H_i = 1). \tag{3.14}$$

표 3.1. 오분류에 대한 벌점

	Classify X_i as noise	Classify X_i as signal
X_i is noise	0	$1-\lambda$
X_i is signal	λ	0

이 베이즈 오류를 최소화 하는 것이 분류의 목적이라고 할 때 이는 pFDR과 pFNR의 상대적 중요도인 가중치 w가 주어졌을 경우, pFDR과 pFNR의 가중 평균을 최소화하는 문제로 정리할 수 있다 (Storey, 2003). 이를 살펴보면 가중치 w를 미리 알고 있고 pFDR과 pFNR의 가중 평균이 $\mathrm{BE}(\Gamma, w) = (1-w)\cdot \mathrm{pFDR}(\Gamma) + w \cdot \mathrm{pFNR}(\Gamma)$로 주어지면 다음 식과 같이 벌점 λ에 대한 기각역의 집합인 B_λ를 구할 수 있다.

$$B_\lambda = \left\{ x : \frac{(1-\pi_0)g_1(x)}{\pi_0 f_0(x) + (1-\pi_0)g_1(x)} \geq \lambda \right\} \tag{3.15}$$

즉, B_λ는 데이터가 신호일 사후 확률(Posterior probability)이 λ보다 크게 되는 영역의 집합이다. 이 때

$$\lambda(w) = \mathrm{argmin}_\lambda \left[\mathrm{BE}(B_\lambda, w)\right] \tag{3.16}$$

라 하면 $\mathrm{BE}(B_{\lambda(w)}, w)$값이 $\mathrm{BE}(\Gamma, w)$의 최소값이 된다 (Storey, 2003). 이에 기초하여 아래와 같은 분류 절차를 제안한다.

Step 0. EM 알고리즘을 통해 혼합 모델의 파라미터 추정 ($\hat{z}_i$, $\hat{\pi}_0$, 가우시안 확률 분포의 파라미터: $\hat{\mu}_0$, $\hat{\sigma}_0^2$, 감마 확률 분포의 파라미터: $\hat{\alpha}_1$, $\hat{\beta}_1$).

Step 1. 분석 목적에 맞게 사용자가 pFDR 및 pFNR의 상대적 중요도인 w 결정.

Step 2. $\lambda(0 \leq \lambda \leq 1)$의 값을 변화시키면서 모든 λ에 대해 가장 근사한 $\hat{z}_i$ 계산.

Step 3. $\hat{z}_i$에 대응하는 x_i를 기각역 B_λ로 추정.

Step 4. pFDR과 pFNR 계산

$$\mathrm{pFDR}(B_\lambda) = \frac{\hat{\pi}_0 F_0(B_\lambda)}{\hat{\pi}_0 F_0(B_\lambda) + (1-\hat{\pi}_0)G_1(B_\lambda)} \tag{3.17}$$

$$\mathrm{pFNR}(B_\lambda) = \frac{(1-\hat{\pi}_0)\,(1-G_1(B_\lambda))}{\hat{\pi}_0(1-F_0(B_\lambda)) + (1-\hat{\pi}_0)(1-G_1(B_\lambda))} \tag{3.18}$$

$F_0(B_\lambda)$: 평균 $\hat{\mu}_0$, 분산 $\hat{\sigma}_0^2$인 정규분포에서 기각역 B_λ에 속할 확률

$G_1(B_\lambda)$: $\mathrm{Gamma}(\hat{\alpha}_1, \hat{\beta}_1)$분포에서 기각역 B_λ에 속할 확률.

Step 5. pFDR과 pFNR의 가중 평균이 최소가 되는 $\lambda(w)$와 그 때의 기각역 $B_{\lambda(w)}$ 찾음

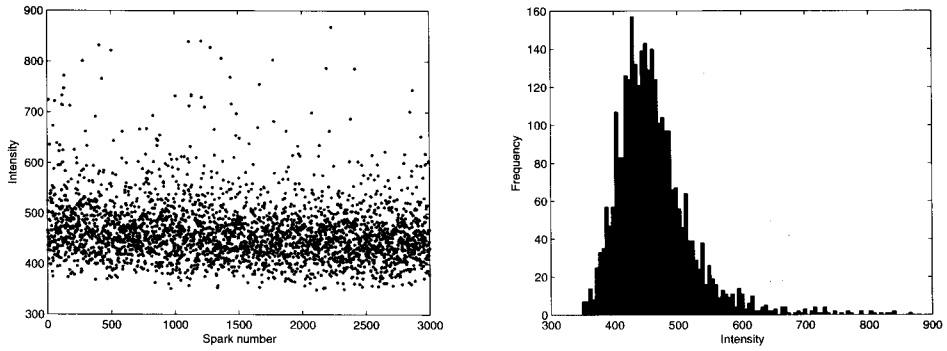
$$\lambda(w) = \mathrm{argmin}_\lambda \left[(1-w)\cdot \mathrm{pFDR}(B_\lambda) + w \cdot \mathrm{pFNR}(B_\lambda)\right]. \tag{3.19}$$

Step 6. 선택된 기각역에 포함되는 데이터를 신호로, 나머지는 잡음으로 판정.

4. 사례 연구 및 모의 실험

4.1. 사례연구: OES 데이터 분석

개재물인 알루미늄(Al)이 포함된 철강 시료에 3,000번의 스파크를 가해서 얻은 실제 OES 데이터에 제안한 방법을 적용하였다. OES 데이터의 강도를 그림 4.1(a)에 점으로 표시하였고 그림 4.1(b)는 이를 히스토그램으로 변환하여 각 강도 구간별 빈도수의 분포를 나타낸 것이다.



(a) OES 데이터의 강도 (b) 히스토그램

그림 4.1. OES 데이터의 분포

표 4.1. OES 데이터의 가우시안-감마 혼합 모델에 대한 모수 추정치

	Gaussian-Gamma mixture	
	잡음	개재물
사전 확률	0.79	0.21
형태 모수	-	40.49
척도 모수	-	13.04
평균	445.60	(528.18)
분산	1367.17	(6889.91)

표 4.2. $\hat{z}$값에 따른 기각 경계치 및 베이즈 오류

$\hat{z}$	0.10	0.14	0.19	0.24	0.29	0.34	0.38	0.43	**0.48**	0.52	0.58	0.62	0.68	0.73	0.78	0.83	0.89
$x_{z(w)}$	465	481	491	498	504	509	513	517	**521**	524	528	531	535	539	543	548	554
Bayes error	0.505	0.477	0.458	0.446	0.437	0.431	0.428	0.426	**0.423**	0.426	0.429	0.431	0.436	0.441	0.448	0.457	0.469

제안한 EM 알고리즘을 사용하여 잡음 및 신호 모델에 대해 모수 추정치를 구한 결과를 표 4.1로 정리하였다.

이를 바탕으로 제안하는 분류 절차에 따라 신호와 잡음을 분류한다. 우선 pFDR 및 pFNR의 중요도를 고려하여 가중치 w를 2/3로 선정하였다. 베이즈 오류를 최소화하기 위해 신호의 사후 확률 경계값인 λ를 0.1부터 0.9까지 0.05간격으로 변화시키며 해당 λ에 가장 근사하는 $\hat{z}$값을 선택하고, 그 값에 해당하는 데이터를 기준으로 기각역을 설정하였다 (표 4.2). 각 기각역에서 pFDR과 pFNR을 구하고 그들의 가중 평균을 계산한 후 그 값이 최소가 되는 $z(w)$ 및 관련 기각역 $\{X \geq x_{z(w)}\}$을 찾는다.

그림 4.2는 $\hat{z}$값의 변화에 따른 베이즈 오류의 변화를 그래프로 나타낸 것이다. 그림 4.2에서 보듯이 $\hat{z}$값이 0.48일 때 베이즈 오류가 최소가 됨을 알 수 있다. 이 때의 기각역 521을 기준으로 데이터가 기각역에 포함되면 신호로, 그렇지 않으면 잡음으로 분류하게 된다. 그림 4.3은 개재물과 잡음이 분류된 결과를 보여준다.

4.2. 모의 실험

보다 객관적인 검증을 위한 모의실험을 통해 제안한 분류 규칙의 성능을 기존 방법과 비교하고자 한다. 신호와 잡음의 분포가 혼합된 상태를 가정하기 위해 잡음의 분포를 가우시안 분포, 신호의 분포를 감마

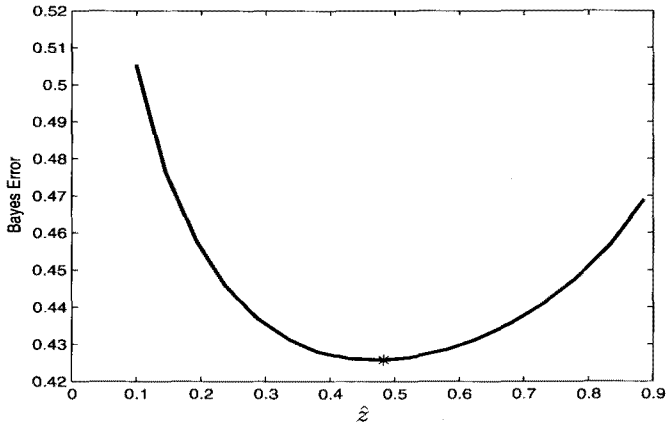


그림 **4.2.** 경계치 변화에 따른 베이즈 오류의 변화

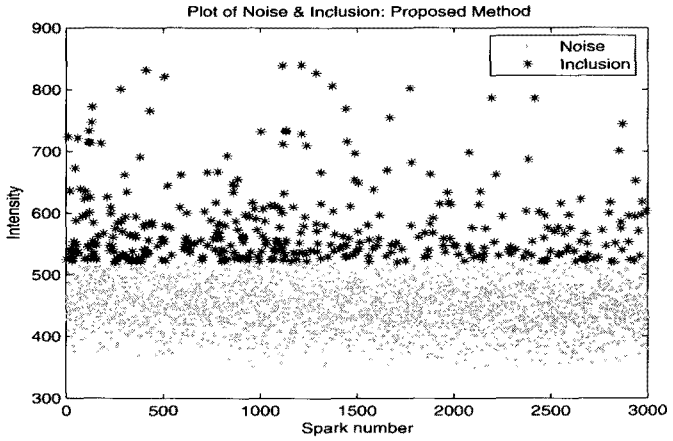


그림 **4.3.** 개재물과 잡음 분류 결과

표 **4.3.** 시뮬레이션을 위한 가우시안-감마 혼합 모델의 설정

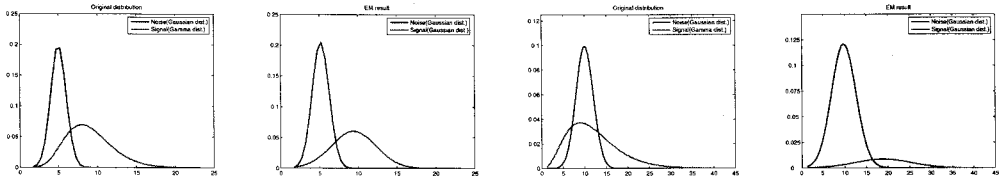
피이크가 잘 구분 되는 경우						피이크가 잘 구분 되지 않는 경우					
Case 1	Noise	Signal	Case 2	Noise	Signal	Case 3	Noise	Signal	Case 4	Noise	Signal
Priors	0.8	0.2	Priors	0.5	0.5	Priors	0.8	0.2	Priors	0.5	0.5
Means	5	(9)	Means	5	(9)	Means	10	(12)	Means	10	(12)
Vars	1	(9)	Vars	1	(9)	Vars	4	(36)	Vars	4	(36)
Shape	-	9	Shape	-	9	Shape	-	4	Shape	-	4
Scale	-	1	Scale	-	1	Scale	-	3	Scale	-	3

분포로 가정하여 데이터를 생성하였다. 두 분포의 피이크가 잘 구분 되는 경우와 그렇지 않은 경우에 결과가 다를 것으로 예상하여 표 4.3과 같이 모수를 설정하였다. 설정한 형태별로 데이터를 각각 3,000개씩 생성하여 모의 실험을 진행하였다.

EM 알고리즘에 의한 모수 추정 결과를 표 4.4에 정리하였다. 피이크가 잘 구분 되는 경우에는 실제값

표 4.4. 가우시안-감마 혼합 모델의 추정 결과

피이크가 잘 구분 되는 경우						피이크가 잘 구분 되지 않는 경우					
Case 1	Noise	Signal	Case 2	Noise	Signal	Case 3	Noise	Signal	Case 4	Noise	Signal
Priors	0.85	0.15	Priors	0.58	0.42	Priors	0.83	0.17	Priors	0.83	0.17
Means	5.03	(9.83)	Means	5.09	(9.50)	Means	9.94	(11.94)	Means	9.52	(17.99)
Vars	1.13	(7.54)	Vars	1.16	(7.94)	Vars	4.20	(37.91)	Vars	7.88	(27.37)
Shape	-	12.81	Shape	-	11.37	Shape	-	3.76	Shape	-	11.83
Scale	-	0.77	Scale	-	0.84	Scale	-	3.17	Scale	-	1.52



(a) Case 2 (b) Case 4

그림 4.4. 실제 분포 및 추정 분포 비교

표 4.5. 각 Case별 정오분류표

피이크가 잘 구분 되는 경우							
Case 1	잡음 판정	신호 판정	Total	Case 2	잡음 판정	신호 판정	Total
Noise	2385	24	2409	Noise	1505	41	1546
Signal	191	400	591	Signal	379	1075	1454
Total	2576	424	3000	Total	1884	1176	3000
피이크가 잘 구분 되지 않는 경우							
Case 3	잡음 판정	신호 판정	Total	Case 4	잡음 판정	신호 판정	Total
Noise	2405	8	2413	Noise	1446	31	1477
Signal	429	158	587	Signal	1060	463	1523
Total	2834	166	3000	Total	2506	494	3000

과 비슷한 결과가 도출되었지만 피이크가 잘 구분 되지 않고 혼합 비율이 5:5인 Case 4 경우에는 분포를 정확히 추정하지 못하는 것을 확인할 수 있다. 그림 4.4는 Case 2 및 Case 4에 대하여 실제 데이터를 생성한 분포(좌)와 EM 알고리즘을 통해 추정한 가우시안-감마 혼합 모델(우)을 보여주고 있다.

각 Case별 데이터에 대해 제안한 방법으로 분류한 결과를 표 4.5에 정리하였다.

기존 분류 규칙인 3σ rule은 데이터의 평균으로부터 표준편차의 3배만큼 떨어진 값을 경계치로 두고 신호와 잡음을 분류하는 방법이다. 제안한 방법과 3σ rule의 성능을 비교하기 위하여 Accuracy, Sensitivity, Specificity를 평가 척도로 사용하고 (Altman과 Bland, 1994) 결과를 표 4.6에 정리하였다. 제안한 방법이 Accuracy 및 Sensitivity 면에서 3σ rule보다 뛰어난 성능을 보임을 확인할 수 있다.

일반적으로 분류 성능 측정 시 Accuracy를 쓰는 경우가 많지만 모의 실험 데이터와 같이 신호와 잡음, 두 집단간의 비율이 크게 차이가 있을 경우에는 적합하지 않다. 따라서 신호와 잡음의 비율에 강건한 척도로 식 (4.1)과 같이 Sensitivity와 Specificity의 기하 평균(Geometric mean)을 사용하도록 제안되고 있다 (Chong과 Jun, 2005).

$$\mathrm{G} = (\text{Sensitivity} \times \text{Specificity})^{\frac{1}{2}} \tag{4.1}$$

표 4.6. 시뮬레이션 데이터에 대한 분류 결과

피이크가 잘 구분 되는 경우					
Case 1	3 σ	Proposed	Case 2	3 σ	Proposed
Accuracy	0.825	0.928	Accuracy	0.527	0.860
Sensitivity	0.113	0.677	Sensitivity	0.025	0.739
Specificity	1	0.990	Specificity	1	0.973
피이크가 잘 구분 되지 않는 경우					
Case 3	3 σ	Proposed	Case 4	3 σ	Proposed
Accuracy	0.824	0.854	Accuracy	0.510	0.636
Sensitivity	0.099	0.269	Sensitivity	0.034	0.304
Specificity	1	0.997	Specificity	1	0.979

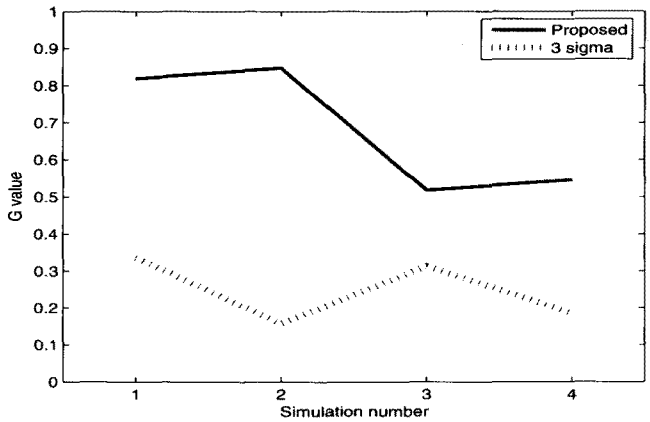


그림 **4.5.** 제안한 방법과 3σ rule의 G 비교

G는 0과 1사이의 값을 가지며 1에 가까울수록 분류가 잘 되었음을 의미한다. 그림 4.5에서 보듯이 제안한 방법의 G값이 3σ rule보다 큰 값을 가짐을 확인할 수 있다.

5. 결론 및 추후 연구

본 논문은 신호와 잡음이 혼합된 대상물에서 이를 분류하는 방법을 제안하였다. 혼합된 관측치에 대하여 가우시안-감마 혼합 모델을 가정하고 EM 알고리즘을 이용하여 분포를 추정하였다. 그 후 다중 가설 검정에서 사용되는 pFDR과 베이즈 오류에 바탕을 둔 분류 규칙을 설정하고 신호와 잡음을 분류하였다. 제안한 방법을 실제 OES 데이터에 적용하여 개재물의 신호를 분류하였고, 모의 실험을 통하여 모델을 추정하고 분류 규칙의 성능을 살펴보았다.

모의 실험에서 제안한 방법이 기존 방법보다 성능이 뛰어남을 보였다. 신호와 잡음 분포의 피이크가 서로 떨어져 잘 구분될 경우에는 실제와 유사한 분포로 추정하였고 분류 성능이 우수함을 확인하였다. 피이크가 서로 중첩될 경우에는 다소 분류 성능이 떨어지나 기존 방법보다는 좋은 것으로 나타났다. 모델 추정 성능을 향상시키기 위하여 세 개 이상의 분포가 혼합되어 있는 경우 및 가우시안, 감마 이외의 여러 분포를 고려하는 연구가 가능할 것이다.

참고문헌

Abdi, H. (2007). *Signal Detection Theory*, In N.J. Salkind (Ed.): Encyclopedia of Measurement and Statistics, Thousand Oaks (CA): Sage.

Altman, D. and Bland, J. M. (1994). Statistics notes: Diagnostic tests 1: Sensitivity and specificity, *British Medical Journal*, **308**, 1552.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society: Series B (Methodological)*, **57**, 289–300.

Bishop, M. C. (2006). *Pattern Recognition and Machine Learning*, Springer, New York.

Choi, S. C. and Wette, R. (1969). Maximum likelihood estimation of the parameters of the gamma distribution and their bias, *Technometrics*, **11**, 683–690.

Chong, I. G. and Jun, C. H. (2005). Performance of some variable selection methods when multicollinearity is present, *Chemometrics and Intelligent Laboratory Systems*, **78**, 103–112.

Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*, Wiley, New York.

Kuss, H. M., Lüengen, S., Müeller, G. and Thurmann, U. (2002). Comparison of spark OES methods for analysis of inclusions in iron base matters, *Analytical and Bioanalytical Chemistry*, **374**, 1242–1249.

Kuss, H. M., Mittelstaedt, H. and Muller, G. (2005). Inclusion mapping and estimation of inclusion contents in ferrous materials by fast scanning laser-induced optical emission spectrometry, *Journal of Analytical Atomic Spectrometry*, **20**, 730–735.

Shin, Y. and Bae, J. S. (2003). Rapid determination of cleanliness for steel by optical emission spectrometer, *IEEE Instrumentation and Measurement Technology Conference (IMTC 2003)*, **2**, 1583–1586.

Storey, J. D. (2002). A direct approach to false discovery rates, *Journal of the Royal Statistical Society: Series B (Methodological)*, **64**, 479–498.

Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value, *The Annals of Statistics*, **31**, 2013–2035.

Separating Signals and Noises Using Mixture Model and Multiple Testing

Hae-Sang Park[1] · Si-Won Yoo[2] · Chi-Hyuck Jun[3]

[1]Department of Industrial and Management Engineering POSTECH;
[2]Customer Satisfaction Improvement Team, NHN;
[3]Department of Industrial and Management Engineering, POSTECH

(Received June 2009; accepted July 2009)

Abstract

A problem of separating signals from noises is considered, when they are randomly mixed in the observation. It is assumed that the noise follows a Gaussian distribution and the signal follows a Gamma distribution, thus the underlying distribution of an observation will be a mixture of Gaussian and Gamma distributions. The parameters of the mixture model will be estimated from the EM algorithm. Then the signals and noises will be classified by a fixed threshold approach based on multiple testing using positive false discovery rate and Bayes error. The proposed method is applied to a real optical emission spectroscopy data for the quantitative analysis of inclusions. A simulation is carried out to compare the performance with the existing method using 3 sigma rule.

Keywords: Signal, noise, EM algorithm, false discovery rate, mixture model, multiple testing.

[3]Corresponding author: Professor, Department of Industrial and Management Engineering, POSTECH, San 31 Hyoja-dong, Pohang 790-784, Korea. E-mail: chjun@postech.ac.kr