가우시안 혼합 모델에 대한 EM 알고리즘을 이용한 신호와 잡음의 분리

Separating Signals and Noises Using EM Algorithm for Gaussian Mixture Model

유시원, 유한민, 이혜선, 전치혁
포항공과대학교 산업경영공학과
{becrux, inha99, hyelee, chjun}@postech.ac.kr

Abstract

For the quantitative analysis of inclusion using OES data, separating of noise and inclusion is needed. In previous methods assuming that noises come from a normal distribution, intensity levels beyond a specific threshold are determined as inclusions. However, it is not possible to classify inclusions in low intensity region using this method, even though every inclusion is an element of some chemical compound. In this paper, we assume that distribution of OES data is a Gaussian mixture and estimate the parameters of the mixture model using EM algorithm. Then, we calculate mixing ratio of noise and inclusion using these parameters to separate noise and inclusion.

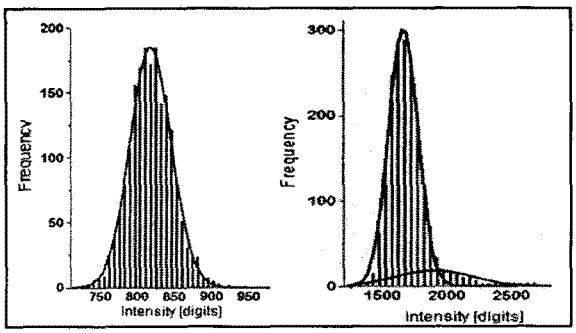
1. 서론

특정 대상물의 측정을 위한 계측장비를 사용하는 경우 원하는 신호뿐만 아니라 잡음이 수반되기 마련이다. 많은 경우 신호 및 잡음의 각 수준은 정규분포를 따르는 것으로 알려져 있다. 그러나 신호와 잡음이 미지의 비율로 혼합된 형태를 갖는 경우 이의 분리가 용이하지 않다.

본 연구에서는 가우시안 혼합모델을 가정한 후 EM 알고리즘을 이용하여 신호 및 잡음을 분리하는 방법을 제안하고 실제 계측데이터에 적용한다. 적용하고자 하는 계측 데이터는 철강제품에 포함된 개재물의 분석을 위한 OES(optical emission spectroscopy) 데이터이다. OES는 시료에 고전압의 스파크를 가해서 원자나 이온의 최외각 전자를 스파크로 들뜨게 한 후 방출되는 자외선이나 가시영역의 빛을 분광기를 이용하여 얻은 스펙트럼선을 읽어서 분석하는 방법이다. OES는 여러 개의 원소 채널을 가지고 있으며 스파크를 일으킬 때마다 각 원소 채널의 신호 강도가 동시에 기록되므로 OES분석을 통해 시료에 존재하는 여러 금속 및 비금속 원소를 동시에 신속하게 검출할 수 있다(Kuss et al., 2005). 각 채널 별로 이들 강도값을 분석하면 특정 스파크가 일어난 지점에 어떠한 개재물 유형이 존재하는지 알 수 있다.

2. 기존연구

[그림 1]의 좌측 그림은 개재물이 없는 순수한 철강 시료를 OES로 측정하여 얻은 신호의 강도를 히스토그램으로 나타낸 것이다. 개재물이 존재하지 않을 때 신호의 강도에 대한 빈도의 분포는 좌측 그림과 같이 가우시안 분포에 근사하며 이는 잡음에 해당한다. 그러나 개재물이 혼합된 철강 시료는 [그림 1]의 우측 그림과 같이 비대칭 분포를 가진다. 이는 신호의 강도가 작은 영역에 존재하는 잡음으로부터 유래한 대칭적인 가우시안 분포와, 강도가 큰



[그림 1] 순수한 철강 시료(좌측)와 개재물이 혼합된 시료(우측)의 OES 신호 강도 분포

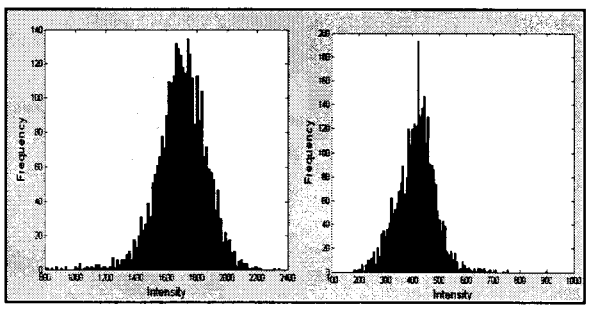
영역의 개재물 분포가 혼합된 형태이다. 개재물의 분포 역시 가우시안 분포를 따른다(Kuss et al., 2002).

기존에는 개재물을 잡음으로부터 분리하기 위해 측정강도에 대한 일정수준의 경계치에 바탕을 둔 방법을 사용하였다. Kuss et al.(2005)는 빈도 분포에서 신호 강도의 평균으로부터 표준편차의 5배만큼 떨어진 값을 경계치로 두어 잡음과 개재물을 분리하는 방법을 제안하였다.

또한 개재물과 잡음이 혼합된 데이터의 히스토그램에서 왼쪽 부분은 정상적인 잡음에서 비롯된 신호로 정상적인 가우시안 분포에 근사한다는 것은 기존 연구를 통해 알 수 있다. 따라서 가우시안 분포의 왼쪽편을 오른편에 대칭시킨 후 겹치는 부분을 잡음에서 유래한 신호로 간주하여 제거하고, 남은 부분의 면적을 계산하여 개재물만의 빈도를 구하는 방법이 제안되었다(Shin and Bae, 2003).

하지만 입자 크기가 작은 개재물의 경우 해당 신호의 강도가 분계점 이하에 존재하는 경우가 발생하기 때문에(Kuss, 2005) 기존의 방법을 사용하여 잡음과 개재물의 신호를 분리할 경우 신호 강도가 낮은 영역에 분포하는 개재물을 검출할 수 없다.

또한 S, Mn등의 원소는 빈도 분포에서 강도가 낮은 쪽에서 잡음 신호의 빈도가 항상 높게 나타난다[그림 2]. 이 경우 개재물에 해당하는 분포보다 잡음 신호의 분포가 더 크고 두텁기 때문에 가우시안 분포의 왼편을 오른편에 겹쳐서 포개지는 부분을 제거하면 개재물의 강도에 해당하는 빈도가 음수가 되는 문제점이 발생한다.



[그림 2] S(좌측)와 Mn(우측)의 OES 신호 강도의 빈도분포

이러한 기존 방법의 문제점들을 개선하기 위해 본 연구에서는 신호의 분포가 가우시안 혼합모델을 따른다고 가정하고 EM 알고리즘을 이용하여 신호 및 잡음을 분리하는 방법을 제안한다.

3. 제안방법

특정 대상물을 측정 할 때에는 계측 장비를 이용하여 대상물에 관련된 신호를 얻어내고자 한다. 계측 장비는 데이터를 측정하는 여러 개의 채널을 갖고 있으며 분석을 위해 데이터를 여러 번 측정한다. 그러나 데이터에는 잡음이 혼합되어 있기 때문에 얻어진 데이터에 혼합된 신호와 잡음을 분리하는 작업이 필요하다. 이 때 관련된 변수를 다음과 같이 정의한다.

c: 채널의 수

m: 각 채널별 측정한 데이터의 수

X_{ij} : i번 채널에서 측정된 j번째 데이터의 값 (i=1,⋯,c; j=1,⋯,m)

측정시 잡음 데이터의 분포가 평균 μ_0, 분산 σ_0^2 인 가우시안 분포를 따르고, i번 채널에서 계측하고자 하는 신호의 데이터는 평균 μ_i , 분산 σ_i^2 인 가우시안 분포를 따르며 이들은 서로 독립이라고 가정한다. 이 때 일반적으로 μ_i 는 μ_0 보다 크다(Kuss, 2005). 또한 π_0 는 임의로 검출한 데이터 값이 신호로부터 왔을 사전 확률이라고 가정한다. f_0 과 f_i 가 관련 모수를 갖는 가우시안 확률분포함수일 때 i번 채널에서 임의로 선택한 데이터는 두 가우시안 분포의 혼합 분포를 따르며 이때의 확률밀도함수는 다음과 같다.

$$f(x) = \pi_0 f_0(x \mid \mu_0, \sigma_0^2) + (1-\pi_0) f_i(x \mid \mu_i, \sigma_i^2)$$

이 경우, 신호를 잡음과 분리하는 문제는 두 개의 군집을 갖는 군집분석 문제로 해석할 수 있으며 모델 기반 군집분석으로 풀이할 수 있다(Fraley and Raftery, 1998). 이 혼합 모델의 우도함수를 최대화하는 관련된 모수들을 추정할 수 있지만 군집 분석의 해를 얻지는 못한다. 즉, 이로부터 데이터 값이 신호인지 잡음인지는 알 수 없다. 그러므로 여기서 새로운 변수 z_{ij} 를 정의한다.

$$z_{ij} = \begin{cases} 1 & \text{data } x_{ij} \textit{ belongs noise} \\ 0 & \text{otherwise} \end{cases}$$

z_{ij} 는 관측할 수 없는 값이지만 z_{ij}의 기대값은 아래의 식으로 예측 가능하다.

$$P\{z_{ij}=1\}=E[z_{ij}]$$

만약 z_{ij} 의 기대값을 예측할 수 있다면 군집의 해를 도출할 수 있다. z_{ij} 의 확률이 높다면 그 관측치가 잡음으로부터 왔음을 의미하기 때문이다. z_{ij} 가 주어졌을 때 x_{ij} 의 확률밀도 함수는 다음과 같다.

$$f(x_{ij}\mid z_{ij},\theta)$$
$$=[f_0(x_{ij}\mid\mu_0,\sigma_0^2)]^{z_{ij}}[f_i(x_{ij}\mid\mu_i,\sigma_i^2)]^{1-z_{ij}}$$

이 때 z_{ij} 의 확률밀도함수는 아래와 같다.

$$f(z_{ij}\mid\pi_0)=\pi_0^{z_{ij}}(1-\pi_0)^{1-z_{ij}}$$

z_{ij}가 관측 가능하지 않으므로 이를 결측치로 보면 (x_{ij},z_{ij})는 불완전 관측치이다. 이런 상황에서 EM 알고리즘은 관련된 모수를 추정하는데 널리 사용된다(Bishop, 2006).

EM 알고리즘은 E-step과 M-step이 반복적으로 시행된다. 기존에 모수 추정치가 존재할 때 E-step에서는 z_{ij} 를 기대치를 이용하여 추정하고, M-step에서는 추정된 $\hat{z}_{ij}$ 를 이용하여 우도함수를 최대화하고 새롭게 모수를 추정한다(Fraley and Raftery, 1998). 이를 정리하면 아래와 같다.

Step 0. $\hat{z}_{ij}$ 를 초기화한다. ($\hat{z}_{ij}$ 에 임의로 0이나 1을 할당한다.)

Step 1. (M-step) $\hat{z}_{ij}$ 에 기반하여 모수를 다음과 같이 추정한다.

$$m_0\leftarrow\sum_{j=1}^{m}\hat{z}_{ij}$$
$$\hat{\pi}_0\leftarrow m_0/m$$
$$\hat{\mu}_0\leftarrow\sum_{j=1}^{m}\hat{z}_{ij}x_{ij}/m_0$$
$$\hat{\sigma}_0^2=\sum_{j=1}^{m}\hat{z}_{ij}(x_{ij}-\hat{\mu}_0)^2/m_0$$
$$\hat{\mu}_i\leftarrow\sum_{j=1}^{m}(1-\hat{z}_{ij})x_{ij}/(m-m_0)$$

($\hat{\mu}_i$가 $\hat{\mu}_0$보다 큰지 확인한다.)

Step 2. (E-step) M-step에서의 모수 추정치를 바탕으로 하여 $\hat{z}_{ik}$를 산출한다.

$$\hat{z}_{ij}=\frac{\hat{\pi}_0 f_0(x_{ij}\mid\hat{\mu}_0,\hat{\sigma}_0^2)}{\hat{\pi}_0 f_0(x_{ij}\mid\hat{\mu}_0,\hat{\sigma}_0^2)+(1-\hat{\pi}_0)f_i(x_{ij}\mid\hat{\mu}_i,\hat{\sigma}_i^2)}$$

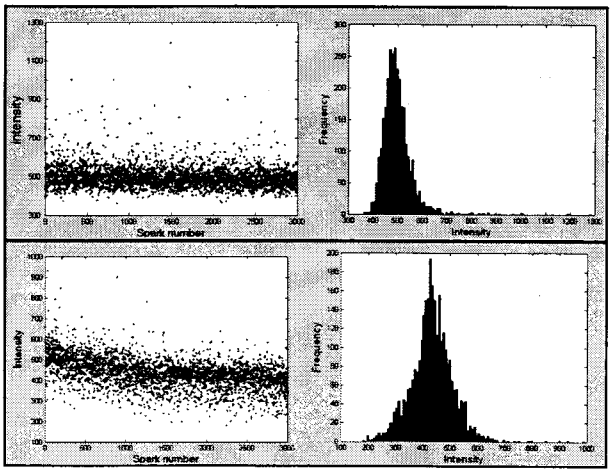
Step 3. 수렴조건을 만족하면 멈추고, 그렇지 않으면 step 1으로 되돌아간다.

Step 0에서 $\hat{z}_{ij}$ 를 초기화하기 위해 잡음 분포의 평균이 신호 분포의 평균보다 작다는 가정을 이용하여, x_{ij}를 오름차순으로 정렬한 후 앞쪽에서부터 일정 비율(예:50%)을 1로 배정하는 방법도 가능하다.

위의 과정을 통하여 전체 데이터 중 $\hat{\pi}_0$의 비율만큼이 잡음으로 판단된다 할 수 있다. 이를 바탕으로 고정 경계치를 이용한 두 가지의 신호 판별 규칙을 정의할 수 있다. 첫째로 $\hat{z}_{ij}$ 를 오름차순으로 배열하여 앞에서부터 $m(1-\hat{\pi}_0)$ 개에 해당하는 데이터 값을 신호로 판단하는 방법이다. 두 번째 방법은 $\hat{c}_0$ 를 경계치로 설정하여 $\hat{c}_0$ 보다 작은 값을 갖는 $\hat{z}_{ij}$ 에 해당하는 데이터 값을 신호로 분류하는 것이다. $\hat{z}_{ij}$ 는 j번째 데이터가 잡음이라고 판단될 사후확률이 되므로 이를 기반으로 위의 판단 규칙을 세울 수 있다.

4. 사례연구

본 실험에서 사용된 데이터는 개재물이 포함된 시료에 3000번의 스파크를 가해서 7개의 원소(Al, Ti, O, Ca, Si, S, Mn)를 검출하는 과정을 10회 반복 실험한 데이터이다(c=7, m=3000). [그림 3]은 그 중 첫 번째 실험에서 얻어진 데이터 중 Al과 S에 대해 정리한 것이다. [그림



[그림 3] OES 데이터 산점도 및 OES 신호 강도의 빈도분포(위: Al, 아래: S)

3]의 좌측 그림은 각 스파크 당 검출된 원소의 강도를 점으로 표시하였으며, 우측 그림은 이를 히스토그램으로 변환하여 각 강도 구간별 빈도수의 분포로 나타낸 것이다.

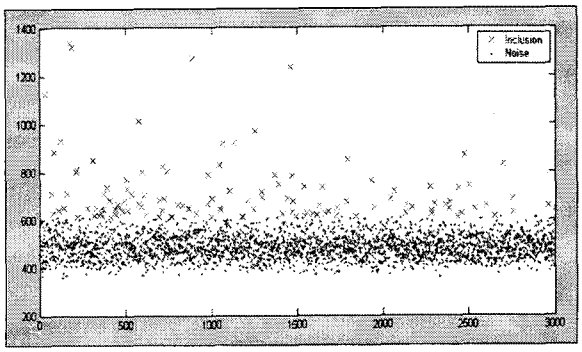
OES 데이터 중 첫 번째 실험 데이터를 제안한 EM 알고리즘을 사용하여 잡음 및 개재물 모델에 대한 모수의 추정치를 구한 결과를 [표 1]로 정리하였다.

[표 1] 각 원소별 모수 추정치

	Al		S	
	잡음	개재물	잡음	개재물
가중치	0.920	0.080	0.746	0.254
평균	488.19	591	464.77	546.46
분산	1611.9	13810	598.05	4570.2
	Ti		O	
	잡음	개재물	잡음	개재물
가중치	0.793	0.207	0.856	0.144
평균	307.18	330.03	99.212	139.37
분산	306.23	1925	254.77	1807.1
	Ca		Si	
	잡음	개재물	잡음	개재물
가중치	0.958	0.042	0.585	0.415
평균	168.09	363.85	1151.2	2870.7
분산	213.25	56866	2.69e+5	9.97e+5
	Mn			
	잡음		개재물	
가중치	0.560		0.440	
평균	1805.7		1957.8	
분산	3173.2		9186.5	

[표 1]의 추정된 모수를 바탕으로 Al에 대해 잡음과 개재물 각각의 모델 및 혼합 모델을 신호 강도의 빈도 분포와 비교한 결과를 [그림 4]로 정리하였다.

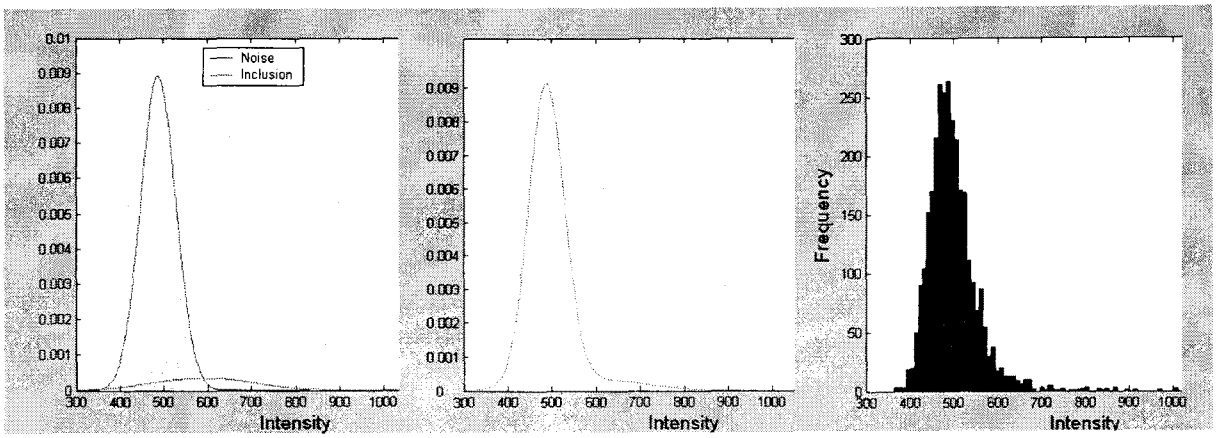
위의 결과에서 잡음과 개재물을 구분하는 경계치 $\hat{c}_0$를 0.5로 설정했을 때 $\hat{z}_{ij}$가 0.5가 되는 지점보다 작은 강도를 갖는 영역은 잡음으로 분류되고, 큰 강도를 갖는 영역은 개재물로 분류할 수 있다[그림 5].



[그림 5] Al의 개재물과 잡음 분리

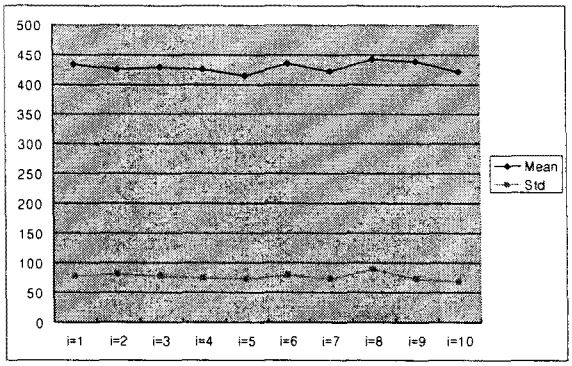
또한 개재물과 잡음의 분포를 알고 있기 때문에 히스토그램의 각 구간별 개재물과 잡음이 어느 정도의 비율로 섞여있는지를 각 분포의 면적의 비를 통해 유추할 수 있으며 각각의 신호 데이터가 잡음일 확률 $\hat{z}_{ij}$이 EM 알고리즘 결과로 주어지므로 신호의 강도가 낮은 부분에서도 개재물일 확률이 높은 신호들을 뽑아낼 수 있을 것이다.

EM 알고리즘으로 알아낸 결과의 안정성을 알아보기 위해 열 개의 실험 데이터에 EM 알고리즘을 반복 적용하고, 잡음 분포의 평균과



[그림 4] Al의 잡음과 개재물 모델(좌), 잡음과 개재물 혼합모델(중간), 히스토그램(우)

분산을 계산해 보았다[그림 6].



[그림 6] S의 잡음 분포의 평균과 분산

각 실험 데이터는 모두 같은 시료를 대상으로 측정된 값이다. [그림 6]에서 실험 데이터에 관계없이 대체적으로 반복 데이터의 평균과 분산이 큰 변동 없이 비교적 유사한 경향을 보이는 것을 확인할 수 있다.

5. 결론

본 연구에서는 신호와 잡음이 미지의 비율로 혼합된 대상물에서 이를 분리하기 위하여 가우시안 혼합모델을 가정하고 EM 알고리즘을 이용하는 방법을 제안하였다. 또한 이를 실제 OES계측 데이터에 적용하여 추정된 분포의 모양을 실제 신호 강도에 대한 빈도수의 분포 모양과 비교한 후 개재물의 신호를 분리해 보았으며, 알고리즘을 평가하는 하나의 척도로 분류 방법의 안정성을 살펴보았다. 본 연구에서는 개재물과 잡음이 각각 가우시안 분포를 따른다고 가정하여 가우시안 혼합 모델을 위한 EM알고리즘을 전개하였으나, 개재물의 분포가 가우시안 분포가 아닌 경우를 가정하여 추후 연구가 가능할 것이다. 또한 EM 알고리즘으로 개재물과 잡음의 분포를 추정한 후 이들을 분류하는 방법으로 고정 경계치를 설정하여 이를 기준으로 분류하는 방법을 이용하였으나, 분포 정보를 바탕으로 보다 다양한 기법을 이용하여 이들의 분류가 가능할 것이다. 마지막으로 현재는 시료에서 개재물의 실제 혼합 비율을 알지 못하여 알고리즘으로 계산한 신호와 잡음의 비율 등 결과값의 정확도를 예측하기 어려우므로 참값을 알고 있는 신호와 잡음의 혼합 데이터를 이용하여 알고리즘의 성능을 보다 정확하게 예측할 수 있을 것이다.

참고문헌

Bishop. M. C. (2006), Pattern Recognition and machine learning, Springer

Fraley, C. and Raftery, A. E. (1998), "How many clusters ? Which clustering method ? Answers via model-based cluster analysis", The Computer Journal, 41(8):578-588.

Kuss, H.-M., S. Lungen, G. Muller, and U. Thurmann (2002), "Comparison of spark OES methods for analysis of inclusions in iron base matters", Anal Bioanal Chem, 374: 1242-1249.

Kuss, H.-M., H. Mittelstaedt, and G. Muller (2005), "Inclusion mapping and estimation of inclusion contents in ferrous materials by fast scanning laser-induced optical emission spectrometry", J. Anal. At. Spectrom. 20, 730-735.

Shin, Y. and J.S. Bae (2003), "Rapid determination of cleanliness for steel by optical emission spectrometer", IEEE Instrumentation and Measurement Technology Conference (IMTC 2003), Vail, Corolado, USA, May 20-22, 2003: 1583-1586.