가우시안 혼합 모델을 이용한 하드 디스크 결함 분포의 패턴 분류

전재영*, 김정헌, 문운철+, 최광남**
중앙대학교 컴퓨터공학부, 전자전기공학부+
jyjun@vimlab.cau.ac.kr*, jeongheon.kim@vimlab.cau.ac.kr,
ucmoon@cau.ac.kr+, knchoi@cau.ac.kr**

Pattern Classification of Hard Disk Defect Distribution Using Gaussian Mixture Model

Jae-Young Jun*, Jeong-Heon Kim, Un-Chul Moon+, Kwang-Nam Choi**
School of Computer Science & Engineering, Chung-Ang University,
School of Electrical & Electronics Engineering, Chung-Ang University+

요 약

본 논문에서는 하드 디스크 드라이브(Hard Disk Drive, HDD) 생산 공정 과정에서 발생할 수 있는 불량 HDD의 결함 분포에 대해서 패턴을 자동으로 분류해주는 기법을 제시한다. 이를 위해서 표준 패턴 클래스로 분류되어 있는 불량 HDD의 각 클래스의 확률 모델을 GMM(Gaussian Mixture Model)로 가정한다. 실험은 전문가에 의해 분류된 실제 HDD 결함 분포로부터 5가지의 특징 값들을 추출한 후, 결함 분포의 클래스를 표현할 수 있는 GMM의 파라미터(Parameter)를 학습한다. 각 모델의 파라미터를 추정하기 위해 EM(Expectation Maximization) 알고리즘을 사용한다. 학습된 GMM의 분류 테스트는 학습에 사용되지 않은 HDD 결함 분포에서 5가지의 특징 값을 입력 값으로 추정된 모델들의 파라미터 값에 의해 사후 확률을 구한다. 계산된 확률 값 중 가장 큰 값을 갖는 모델의 클래스를 표준 패턴 클래스로 분류한다. 그 결과 제시된 GMM을 이용한 HDD의 패턴 분류의 결과 96.1%의 정답률을 보여준다.

1. 서 론

제조 공정의 산출물에 대한 분석은 공정상의 문제 분석 및 조정에 대한 지시를 결정하는 중요한 정보다. HDD(Hard Disk Drive) 제조 공정에서는 HDD상의 결함 분포 분석을 통해 문제 부품 교체 및 재작업 지시를 결정할 수 있다. 따라서 HDD의 특성에 관련된 결함 분포의 패턴 분류 연구는 실제 공정에서 직접 적용 될 수 있는 응용 분야이다. HDD제조 공정에서는 각 공정 단계에서 HDD의 읽기, 쓰기, 찾기 등의 검사를 수행하여 HDD의 이상 유무를 확인한다. 이 때, 정상적인 읽기, 쓰기, 찾기가 수행되지 않는 섹터는 결함으로 처리되며, 이러한 정보는 HDD 자체에서 소프트웨어적으로 처리되어 데이터의 저장에 사용하지 못하도록 처리한다. 이러한 결함의 개수가 주어진 임계치를 초과하게 되면, 그 HDD는 불량품으로 처리된다.

수리공정에서 HDD 결함 분포의 패턴 분류는 불량의 종류를 나누는 중요한 정보를 제공한다. 수리사가 분류하는 결함 분포의 표준 패턴 클래스(Standard Pattern Class)는 루프(Loop)형, 분침(Watch)형, 아크(Arc)형, 방사(Radial)형, 찍힘(Spot)형, 전체(Whole)형의 여섯 가지이다. 결함 분포의 자동 패턴 분류에는 다양한 알고리즘이 적용되어 왔다. 디스크 상의 두 동심원 사이의 공간을 정해진 회전 각 별로 등분 한 후, 나누어진 구간별로 결함 발생 빈도 히스토그램(Histogram)구하여, 분석을 수행하였다. 결함 분포의 패턴을 분류하기 위하여, 5가지의 특징을 선정한 뒤, 각 특징들이 표준 패턴 클래스에 속할 가능성(Possibility)을 종합적으로 추론하여 최종적인 패턴을 선정한 퍼지 추론(Fuzzy Inference)을 사용하였다[1]. 다층 퍼셉트론(Multi-Layer Perceptorn) 신경망의 오차 역전파 알고리즘을 통하여 학습 한 후 입력된 결함 분포 데이터의 패턴을 분류하는 연구가 진행 되었다. 점을 찾기 위한 연구는 Mean Shift기법을 이용하여 영상의 안장 점(Saddle Point)을 찾는 방법이 제안되었다[2]. 개선된 k-means 알고리즘을 이용하여 유사한 의미를 갖는 점을 군집화 후, 그리드(Grid)를 생성

※ 이 논문은 2007년도 정부(과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임 (R01-2005-000-10940-0)
* 학생회원 : 중앙대학교 컴퓨터공학부
** 종신회원 : 중앙대학교 컴퓨터공학 교수
접수일 : 2008년 4월 18일,
완료일 : 2008년 5월 26일

하는 연구가 진행되었다[3].

본 연구에서는, 이러한 HDD 결함 분포의 패턴 분류에 관한 연구 결과를 제시한다. 실험을 위해 표준 패턴 클래스의 확률 모델을 GMM(Gaussian Mixture Model)으로 가정하였다. GMM은 주어진 표본 데이터 클래스의 분포 밀도를 복수개의 가우시안 확률 밀도 함수로 모델링하는 방법이다[4][5]. 이를 위하여, 실제 불량 HDD 결함 분포의 패턴 분석을 바탕으로, 패턴 분류의 근거가 되는 5가지의 특징들을 선정하였다. 그리고 전문가에 의해 미리 분류된 표준 클래스의 데이터들을 가지고 각 클래스의 GMM의 파라미터를 추정하는 학습 과정을 수행하였다. GMM의 파라미터를 추정하기 위해 EM(Expectation Maximization) 알고리즘이 사용되었다[6]~[9]. 그 후에 학습에 사용되지 않은 불량 HDD 데이터들을 입력값으로 하여 추정된 모델의 파라미터 값에 의해 가장 적합한 모델을 검색하고, 검색된 모델이 속한 표준 패턴 클래스를 해당 데이터의 클래스로 선정하였다. 그 결과 제안된 5가지의 특징들과 알고리즘은 불량 HDD를 표준 패턴 클래스로 분류함에 있어서 만족할 만한 성능을 보여 주었다.

2. 패턴 분류를 위한 클래스 선정 및 특징 추출

2.1 표준 패턴 클래스 선정

디스크는 표면상에 동심원들로 구성된 실린더(Cylinder)와 각 실린더 내의 물리적인 최소 저장 단위인 섹터(sector)들로 구성되어 있다. 고속으로 회전하는 디스크와 디스크의 중심에 수직인 방향으로 왕복 운동하는 HAS(Head Suspension Assembly)의 제어에 의해 고밀도로 집적된 디스크 섹터에 접근하여 데이터의 입출력이 수행된다.

불량으로 판정된 HDD는 수리공정으로 이송되고, 수리공정의 작업자들은 해당 HDD의 가장 안쪽의 실린더(Maintenance Cylinder, MC)에 기록된 결함들의 정보에 접속하여 결함 분포를 관찰한다. 관찰 결과 수리사들은 결함 분포가 루프(Loop)형, 분침(Watch)형, 아크(Arc)형, 방사(Radial)형, 찍힘(Spot) 형, 전체(Whole)형의 여섯 가지 표준 클래스에 해당하는 지를 판단한다. 결함 분포의 표준 클래스에 근거하여 후속 테스트가 결정된다. 예를 들면, 결함 분포가 루프형인 경우에는 불량 헤드(Head)로 인한 불량일 확률이 높기 때문에, 헤드에 추가적으로 읽기/쓰기를 수행한다. 읽기/쓰기 에러 발생 횟수가 임계치를 넘어선 경우에는 헤드 불량으로 선정하여, 해당 HDD는 다시 클린 룸으로 투입되어 헤드를 교체한 후, 생산 공정에 다시 투입된다.

2.2. HDD 결함 분포의 전처리

본 연구에서는, HDD결함 분포의 전처리(Pre processing)를 통해 극 좌표(Polar Coordinate) 상으로 나타내었다. 그 결과, 반지름은 헤드로부터 반지름이 1 에서 최대 224까지의 구간으로 실제 측정 반지름은 74 에서 224까지의 151개 값으로 측정 되었으며, 각도는 1° 씩 360 구간으로 각각 이산화 시켜서 구역화 하였다. 각 구역에 기준 이상의 결함들이 존재하는 경우, 그 구역에는 결함이 존재하는 것으로 처리하였다. 이와 같이 전처리를 거치면 일반적으로 500개 이하의 결함들이 다음과 같은 형태로 나타내어진다.

$$d_i = (r_i, \theta_i) \quad (1)$$

여기서 d_i 는 i 번째 결함, r_i 와 θ_i 는 각각 결함의 거리 및 각도이다. 이 때, HDD의 3시 방향을 기준으로 하여 반 시계 방향으로 각도가 증가하는 것으로 설정하여 결함의 각도를 나타내었다.

표 1. 전처리 된 결함 패턴 데이터의 예

결함	위치 (r_i, θ_i)
d_1	(45, 30°)
d_2	(130, 40°)
d_3	(85, 110°)
d_4	(90, 320°)

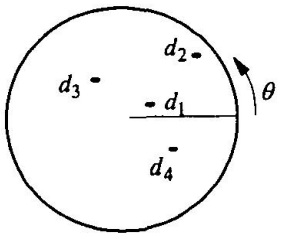

그림 1. 표1의 결함 패턴

표 1은 전처리 된 패턴 데이터의 한 예로서, 총 4개의 결함으로 구성된 패턴을 나타내었다. 각 결함의 위치를 극 좌표 형태로 나타내었으며, 거리와 각도는 각각 정수 단위로 이산화 되어 있음을 알 수 있다. 그림 1은 표1의 결함 패턴 데이터의 결함 분포를 나타낸다.

그림 2는 전처리된 표준 패턴의 전형적인 데이터 그림들이다.

2.3 패턴 분류를 위한 5가지 특징

전처리된 입력 패턴이 표준 패턴 클래스 중에서 어느 클래스에 해당하는지를 판단할 수 있는 특징들에 대한 분석을 실시하였다. 분석 결과, 패턴 분류를 위하여 5가지 특징들이 선정 되었다.

본 연구에서는 산발적인 결함들을 제거하기 위하여 극 좌표를 횡축은 각도, 종축은 반지름 값을 갖는 직교 좌

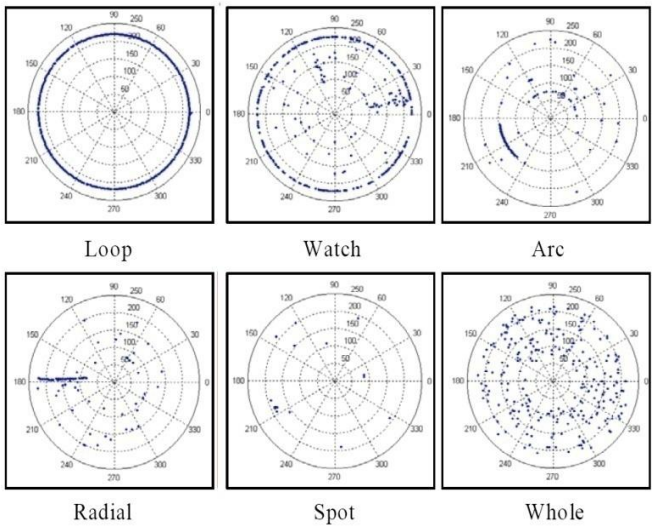

그림 2. 전처리 된 패턴 클래스

표(Cartesian Coordinate)로 변환하여 표현하였다. 반지름과 각도로 이산화 시켜 격자 형태로 표현된 직교 좌표 위에, 종축과 횡축이 각각 5칸의 크기를 갖는 5*5 정사각형 형태의 이동 창(Moving window)을 이용하여 결함 개수가 3개 이하인 이동창의 결함은 제거하는 필터링(Filtering) 방식을 1차로 사용하였다. 2차 필터링으로 30*30 정사각형 형태의 이동창을 이용하여 결함 개수가 8개 이하인 이동창의 결함은 제거하는 필터링 적용 시켰다.

가. 필터링 후 결함의 개수

첫 번째 특징은, 직교 좌표로 표현된 HDD 결함 분포 데이터에서 이동 창을 이용한 필터링 방식을 적용 한 후 제거되지 않은 결함의 개수이다.

$$n = filtering(n_0) \tag{2}$$

위 식에서 n 은 필터링 후 결함의 개수, n_0는 필터링 전 결함의 개수이다.

나. 필터링 전후의 결함 개수 변화율

두 번째 특징으로는, 필터링 전과 후의 결함개수의 변화율로 선정하였으며 다음과 같은 식으로 계산된다.

$$n_r = {}^{n}/_{n_0} \tag{3}$$

여기서 n_r은 필터링 전후의 결함 개수 변화율을 나타낸다.

다. 중심으로의 거리 r 성분의 점유도

이는 HDD 결함 분포의 r 성분으로 계산한 점유도이다. 결함 분포에서 r 의 범위가 74에서 224까지 151개의 구간에서 나타났다. r 성분의 범위에서 결함을 포함한 반지름은 다음과 같이 계산된다.

$$r_s = \left({}^{r_n}/_{151}\right) \times 100 \tag{4}$$

여기서 r_s는 r 성분의 점유도, r_n은 결함이 존재하는 반지름의 개수를 나타낸다.

라. 각도 θ 성분의 평균 간격

이 특징을 추출하기 위해서, 각도와 반지름에 대한 정수 값으로 이산화 되어있는 HDD 결함 데이터 중에서 반지름 값들은 제외한다. 결함 데이터에서 반지름을 제외하고 남은 각도 값을 내림차순으로 정렬하여 각도 값들의 차를 구한다. 이 과정을 통해 얻어진 각도 값들의 차를 사용하여 평균 값 μ_θ 구하게 된다.

마. 유사 거리 내에 존재하는 결함 개수들의 표준편차

이 특징을 계산하기 위해, 먼저 직교 좌표로 표현된 HDD 결함의 한 점에서 일정 거리 내에 존재하는 점들의 개수를 구한다. 그 다음에 다른 모든 점에 대해서도 해당 점에서 일정 거리 내에 존재하는 점들의 개수를 구한다. 마지막으로 HDD 결함의 모든 점에서 구한 값들의 표준 편차 σ_d를 구한다.

3. GMM을 이용한 패턴 분류

3.1 GMM

GMM은 주어진 표본 데이터 집합의 분포 밀도를 단 하나의 확률 밀도 함수로 모델링 하는 방법을 개선한 밀도 추정 방법으로 복수 개의 가우시안 확률 밀도 함수로 데이터의 분포를 모델링 하는 방법이다. 이는 K 개의 가우시안 확률 밀도 함수의 선형 결합으로 식 (5)과 같이 표현 된다.

$$P(x|\theta) = \sum_{i=1}^{M} p(x|\omega_i, \theta_i)\, P(\omega_i) \tag{5}$$

여기서, $p(x|\omega_i, \theta_i)$ 는 데이터 x 에 대하여 ω_i 번째 모델 파라미터 θ_i 로 이루어진 확률 밀도 함수를 의미하며, $P(\omega_i)$ 는 혼합 가중치로 각 확률 밀도 함수의 상대적인 중요도를 의미한다.

3.2 GMM의 학습을 위한 EM 알고리즘

GMM의 학습이란 표본 데이터 집합이 주어질 경우 데이터의 로그-우도(Log-Likelihood)를 최대로 하는 가우시안 모델들의 파라미터들을 추정하여 구하는 문제를 말한다. 일반적으로 GMM은 EM 알고리즘으로 최적 모델을 추정하여 결정한다.

가우시안 모델은 다음과 같이 표현된다.

$$p(x|\omega_j,\theta) = p(x|\mu_j,\sigma_j^2) = \frac{1}{(2\pi)^{1/2}|\sigma_j|^{1/2}} exp\left(-\frac{(x_n-\mu_j)^2}{2{\sigma_j}^2}\right) \quad (6)$$

파라미터 θ를 이루는 각 모델의 평균(μ_j)과 분산(σ_j), 가중치($\alpha_j = P(\omega_j)$)를 구하기 위해 MLE (Maximum Likelihood Estimation)알고리즘을 사용한다.

MLE 알고리즘을 사용하여 파라미터를 구하는 일반식은 다음과 같다.

$$\hat{\theta} = arg\,max\left[\sum_{n=1}^{N} log \sum_{j=1}^{M} p(x_n|\theta_j)\ P(\omega_i)\right] \quad (7)$$

파라미터를 구하기 위해 편미분을 이용한 MLE 방법으로 추정치를 유도한다. 그 후 최대 로그-우도를 구하기 위해 편미분을 0으로 두고 정리하면, 파라미터의 추정치를 얻을 수 있다.

평균 μ_j는 다음 식으로 유도된다.

$$\frac{\partial}{\partial \mu_j}[\cdot] = 0\,,\ \hat{\mu}_j = \frac{\sum_{n=1}^{N} P(\omega_j|x_n,\theta)\,x_n}{\sum_{n=1}^{N} P(\omega_j|x_n,\theta)} \quad (8)$$

분산 σ_j^2는 다음 식으로 유도된다.

$$\frac{\partial}{\partial \sigma_j^2}[\cdot] = 0\,,\ \hat{\sigma}_j^2 = \frac{1}{d}\frac{\sum_{n=1}^{N} P(\omega_j|x_n,\theta)\|x_n-\hat{\mu}_j\|^2}{\sum_{n=1}^{N} P(\omega_j|x_n,\theta)} \quad (9)$$

혼합 가중치 α_j는 다음 식으로 유도된다.

$$\frac{\partial}{\partial \alpha_j}[\cdot] = 0\,,\ \hat{\alpha}_j = \hat{p}(\omega_j) = \frac{1}{N}\sum_{n=1}^{N} P(\omega_j|x_n,\theta) \quad (10)$$

유도 과정에서 학습 데이터 집합 x_n이 주어질 때, j번째 혼합 모델의 사후 확률(Posterior Probability)는 다음과 같이 구할 수 있다.

$$P(\omega_j|x_n,\theta) = \frac{p(x|\mu_j,\sigma_j^2)P(\omega_i)}{p(x|\theta)} \quad (11)$$

MLE 과정에서 유도된 파라미터의 식 (8)~(10)은 우항에 최종 모델의 파라미터들이 포함되어 있다. 따라서 이 식을 통해 로그-우도를 최대로 하는 값을 구하기 위해 EM 알고리즘을 반복해서 수행한다. EM알고리즘을 통한 GMM 파라미터 학습에는 혼합 모델(Mixture Model)의 수를 사용자가 입력해야 한다. EM 알고리즘은 입력된 혼합 모델의 수에 맞게 파라미터θ를 임의로 설정하는 단계로부터 시작된다. 그 다음 θ 와 사후확률 $P(\omega_j|x_n,\theta)$의 추정치를 향상시키기 위해 E-step과 M-step을 반복 수행한다.

4. 실험 및 결과

HDD 결함 분포를 분류하기 위해 제안된 5가지의 특징과 알고리즘을 평가에 실제 제조공정에서 발생한 불량 HDD를 대상으로 실험 했다. 실험을 위해 사용된 HDD 결함 분포 데이터는 459개가 사용되었으며, 이는 전문가에 의해서 6개의 표준 패턴 클래스로 미리 분류된 데이터들이다. 먼저 실험 데이터들을 가지고 제안된 5가지의 특징들을 추출하였다. 특징 값들은 0 에서 1 간격으로 정규화 되었다

GMM학습을 위해서 실험에 사용된 459개의 데이터 중에서 306개의 데이터가 사용되었다. 학습은 5가지의 특징으로 표현된 학습 데이터를 사용하여 클래스 별로 정의된 혼합 모델의 파라미터들을 추정한다. 그 다음, 추정된 패턴 클래스 별 혼합 모델의 파라미터를 이용하여 학습에 사용되지 않은 153개의 HDD결함 분포 데이터들을 분류하였다. 분류는 입력된 데이터들의 특징을 통해 각 가우시안 모델의 사후 확률을 계산한 뒤, 그 중 가장 큰 확률 값을 가지는 표준 클래스의 모델을 입력된 결함 패턴의 클래스로 선정하였다. 표 2는 153개 HDD 분포 데이터들의 패턴 분류 결과를 보여준다. 분류 결과는 147개가 일치하여 96.1%의 적중률을 나타내었다.

표 2. 패턴 분류 결과

표준 클래스	총 개수	성공	실패	성공률
Arc	10	9	1	90.0 %
Loop	7	7	0	100 %
Spot	61	59	2	96.7 %
Watch	11	10	1	90.9 %
Radial	2	2	0	100 %
Whole	62	60	2	96.8 %
계	153	147	6	96.1 %

5. 결론

본 연구에서는 GMM 을 HDD 결함 분포의 패턴 분류에 적용한 연구 결과를 제시 하였다. GMM은 주어진 표본 데이터 클래스의 분포 밀도를 복수개의 가우시안 확률 밀도 함수로 모델링 하는 방법이다. 이를 위하여, 실제 불량 HDD 결함 분포의 패턴 분석을 바탕으로, 패턴 분류의 근거가 되는 5가지의 특징들을 선정하였다. 그

후 전문가에 의해서 미리 분류된 패턴 데이터를 이용하여 GMM을 EM 알고리즘을 통하여 학습하였다. 실험 결과 96.1%의 적중률을 갖는 분류 성능을 확인하였으며, 본 연구 결과는 수작업에 기초한 수리 판정 작업의 자동화에 유용하게 응용될 수 있을 것으로 기대된다.

참고 문헌

[1] 문운철, 권현태, “퍼지추론을 이용한 HDD (Hard Disk Drive) 결함 분포의 패턴 분류”, 대한전기학회 논문지, 54D권, 6호, pp. 383-389, 2005

[2] D. Comaniciu, “Image Segmentation Using Clustering With Saddle Point Detection”, IEEE Int. Conf. on Image Processing, Vol.3, pp.297-300, 2002

[3] M. Berger and I. Rigoutsos, “An Algorithm for Point Clustering and Grid Generation”, IEEE Trans. on Systems, Man and Cybernetics, Vol.21, No.5, 1991

[4] C. Fraley and A. Raftery, “How Many Clusters? Which Clustering Method? Answers Via Model-based Cluster Analysis”, The Computer Journal, Vol.41, No.8, pp. 578-588, 1998

[5] M. Figueiredo and A. Jain, "Unsupervised Learning of Finite Mixture Models", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.24, No.3, pp. 381-396, 2002

[6] R. Duda, P. Hart and D. Stork, Pattern Classification (2nd Edition), Wiley-Interscience, 2000

[7] A. Dempster, N. Laird, and D. Rubin, “Maximum-likelihood From Incomplete Data Via The EM Algorithm”, Journal of the Royal Statistical Society, Series B, Vol.39, No.1, pp. 1-38, 1997

[8] M. Jordan and R. Jacobs, “Hierarchical Mixtures Of Experts And The EM Algorithm”, Neural Computation, Vol.6, No.2, pp. 131-214, 1994

[9] G. McLachlan and T. Krishnan, “The EM Algorithm and Extensions”, Wiley, 1996