The Korean Journal of Applied Statistics
2022, Vol. 35, No. 5, 593–601

DOI: https://doi.org/10.5351/KJAS.2022.35.5.593
Print ISSN1225-066X / Online ISSN 2383-5818

Fraud detection support vector machines with a functional predictor: application to defective wafer detection problem

Minhyoung Park[a], Seung Jun Shin[1,a]

[a]Department of Statistics, Korea University

Abstract

We call "fruad" the cases that are not frequently occurring but cause significant losses. Fraud detection is commonly encountered in various applications, including wafer production in the semiconductor industry. It is not trivial to directly extend the standard binary classification methods to the fraud detection context because the misclassification cost is much higher than the normal class. In this article, we propose the functional fraud detection support vector machine (F^2DSVM) that extends the fraud detection support vector machine (FDSVM) to handle functional covariates. The proposed method seeks a classifier for a function predictor that achieves optimal performance while achieving the desired sensitivity level. F^2DSVM, like the conventional SVM, has piece-wise linear solution paths, allowing us to develop an efficient algorithm to recover entire solution paths, resulting in significantly improved computational efficiency. Finally, we apply the proposed F^2DSVM to the defective wafer detection problem and assess its potential applicability.

Keywords: fraud detection, functional data, piece-wise linear solution paths, support vector machine

1. 서론

웨이퍼(wafer)는 반도체의 품질을 결정하는 가장 핵심적인 부품이다. 따라서 웨이퍼의 품질관리 혹은 불량 탐지는 반도체 생산 공정에서 매우 중요한 과정이다. 통상적으로, 반도체 공장에서 생산되는 웨이퍼가 불량일 가능성은 낮다. 하지만, 한번 불량이 발생했을 때 적시에 탐지되지 않으면 회사는 장기적으로 큰 손실을 입게 된다. 본 논문에서 웨이퍼 불량, 신용카드 위조, 컴퓨터 침입 등 빈번히 발생하지는 않지만 한번 발생하게 되면 상대적으로 큰 손실을 가져오는 사례들을 탐지하는 문제를 고려하고자 하며, 이를 부정 탐지(fraud detection)라 한다 (Bolton과 Hand, 2002). 부정 탐지는 정상과 부정을 분류하는 문제이기에 이항분류로 볼 수 있다. 하지만, 부정 탐지의 문제에서 부정사례는 정상사례에 비해 상대적으로 매우 적고 오분류의 비용이 월등히 크기 때문에, 통상적인 이항분류 기법을 그대로 적용하는데는 문제점이 있다. 일반적인 이항분류 기법은 두 집단의 상대적 중요도 혹은 크기가 어느정도 균형을 이루고 있음을 가정하고 분류의 정확도를 최대화하는 방식으로 분류기를 학습시킨다. 만약 부정 탐지의 문제에서 분류정확도만을 기준으로 삼게되면 모든 부정사례(예, 불량 웨이퍼)를 정상으로 분류하게 되는 경우가 발생하며, 이는 매우 바람직하지 않다. 한가지 대안으로, 집단 간의 비용을 고려하여 집단 간의 가중치 부가를 통해 어느정도의 불균형을 해소하는 경우도 있지만, 각 집단의 가중치를 정확히 측정하는 것은 불가능하다. 이외에도 부정 개체를 과대추출(oversampling)하거나 정상 개체를 과소추출(undersampling)하는 방식으로 불균형을 해결하는 방법도 널리 쓰인다 (Chawla 등, 2002;

This work is supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (Grant No. NRF-2018R1D1A1B07043034).

[1] Corresponding author: Department of Statistics, Korea University, 145 Anam-Ro, Seongbuk-Gu, Seoul 02841, Korea. E-mail: sjshin@korea.ac.kr

Published 31 October 2022 / journal homepage: http://kjas.or.kr
© 2022 The Korean Statistical Society. All rights reserved.

Feng 등, 2021). 하지만, 인위적인 재표본을 통해 불균형을 해소하는 경우 해당 방법에 대한 이론적인 분석이 쉽지 않다는 단점이 있다. 본 연구에서는 부정 개체에 대한 분류정확도를 연구자가 원하는 수준으로 만족시킬 수 있는 방법을 활용하여 부정 탐지 문제를 해결하고자 한다.

부정 탐지의 문제를 보다 정확히 기술하기 위해 다음과 같은 이항분류 자료 $(y_i, \mathbf{x}_i) \in \{-1, 1\} \times \mathbb{R}^p$, $i = 1, \ldots, n$를 고려하자. 두 집단을 $I_+ = \{i : y_i = 1\}$와 $I_- = \{i : y_i = -1\}$로 각각 표기하고, 집단내 개체수를 $n_+ = |I_+|$과 $n_- = |I_-|$로 표현하자. 여기서, $n_+ + n_- = n$이다. 일반성을 잃지않고, 반응 변수 y_i가 1의 값을 가지는 경우를 부정 혹은 불량, -1의 값을 가지는 경우를 정상이라고 하자. 부정 탐지 문제에서는 정상 개체의 수가 불량 개체의 수보다 월등히 크다고 가정한다 (즉, $n_+ \ll n_-$). 부정탐지에서의 최종 목표는 불량 혹은 부정을 정확히 탐지하면서도 정상개체를 최대한 잘 분류하는 분류함수 f를 찾는 것이다. 이 때, 분류함수 f는 주어진 공변량 $\mathbf{x}_i$에 대해 $f(\mathbf{x}_i) > 0$이면 y_i를 불량으로, 그렇지 않으면 정상으로 분류하는 함수를 의미한다. 따라서 부정 탐지 문제는 다음과 같은 일반화된 최적화 문제로 표현할 수 있다.

$$\min_{f \in \mathcal{F}} \sum_{i \in I_-} \mathbb{1}\{f(\mathbf{x}_i) \geq 0\}, \quad \text{subject to } f(\mathbf{x}_i) > 0, \forall i \in I_+. \tag{1.1}$$

여기서 $\mathcal{F}$는 분류함수 f의 공간을 의미하며, $\mathbb{1}\{A\}$는 조건 A를 만족하면 1 그렇지 않으면 0의 값을 가지는 지시함수를 나타낸다. 위의 최적화 문제 (1.1)은 매우 직관적이고 단순해 보이지만, 주어진 표본으로부터 해를 구하는 것이 불가능하다. 지시함수의 사용으로 인해 목적함수가 비연속일 뿐만 아니라, 무한 차원상의 공간에 존재하는 f에 대해 최적화를 해야하기 때문이다. 이를 해결하기 위해, 지지 벡터기계(Support Vector Machines; SVM) (Vapnik, 1999)에 활용되는 경첩 손실(hinge loss) 함수로 지시함수를 대체하고, f가 재생커널힐베르트(Reproducing Kernel Hilbert Space; RKHS) (Wahba, 1990) 공간 상의 원소임을 가정하자. 여기에 통상적인 L_2-정규화를 추가하면 부정 탐지를 위한 지지벡터기계를 정의할 수 있다. 본 논문에서는 이를 부정 탐지 지지 벡터기계(Fraud Detection Support Vector Machines; FDSVM)라 한다. 부정 탐지 지지 벡터기계, FDSVM에 대한 자세한 내용은 2절에서 참고하자.

부정 탐지의 문제는 시간, 공간, 파장 등과 같은 함수형 공변량(Functional Predictor)을 수반하는 경우가 종종 있다 (Woodall 등, 2004). 예를 들어 웨이퍼 불량의 경우 시간에 따른 공정이 진행됨에 따라 불량 문제가 발생하게 되는데, 측정 시간마다 관측된 자료를 단순히 다변량 자료로 인식하게 되면 측정의 간격이 짧을수록 공변량의 차원이 증가하게 되어 차원의 저주로 인한 과적합(over-fitting)이 발생하기 쉽다. 자료가 포함하는 정보의 크기 관점에서 본다면 측정 간격이 짧아진다는 것은 자료 속에 내재된 정보의 크기가 증가함을 의미하는데 이를 제대로 활용하지 못하는 것은 매우 바람직하지 못하다. 만약, 공변량을 함수라 정의한다면 주어진 관측치들을 함수형 자료의 함수값으로 볼 수 있다. 이렇게 되면 앞서와 같은 문제점을 자연스럽게 해결할 수 있다. 측정값이 많을 수록 함수형 공변량을 보다 정확하게 관찰하게 되므로 분류기의 성능이 증가하기 때문이다. 함수형 자료에 대한 보다 자세한 내용은 Ramsay와 Silverman (2002)를 참조하자.

이항 분류문제에서 함수형 공변량을 다루는 방법은 활발히 연구되어져 왔으며, 그 예로는 함수형 선형 판별분석(Functional Linear Discriminant Analysis; FLDA) (James와 Hastie, 2001), 함수형 일반화 선형모형 (Functional Generalized Linear Models; FGLM) (James, 2002), 함수형 K-최근접이웃거리 (Biau 등, 2005), 그리고 함수형 지지 벡터기계(Functional Support Vector Machines; FSVM) (Rossi와 Villa, 2006; Park 등, 2008)등이 있다. 본 논문에서는 Park 등 (2008)이 제안한 함수형 자료에 대한 선형커널을 활용하여 FDSVM을 함수형 자료로 확장하고, 이를 함수형 부정 탐지 지지 벡터기계(Functional Fraud Detection Support Vector Machines; F^{2}DSVM)라고 부르기로 한다.

본 논문의 구성은 다음과 같다. 2절에서는 본 연구의 기반이 되는 FDSVM를 소개하고, 3절에서는 함수형 공변량의 활용을 위한 커널(kernel)을 제안하고 F^{2}DSVM을 자세히 기술하고자 한다. 4절에서는 실제 반도체 생산과정에서 파생된 웨이퍼 자료를 통해 F^{2}DSVM의 성능을 확인하고, 5절에서는 연구 내용에 대한 요약과 결론을 제시하고자 한다.

2. FDSVM

전술한 바와 같이 식 (1.1)를 직접 풀 수 없다. 이를 해결하기 위해 RKHS를 활용하고, 지시함수를 경첩손실 함수로 대체한 뒤 통상적인 L_2-정규화를 적용한 것이 FDSVM이다. 즉, FDSVM은 다음과 같은 최적화 문제로 표현된다.

$$\min_{f\in\mathcal{H}_K} \sum_{i=1}^{n} [1 - y_i f(\mathbf{x}_i)]_+ + \frac{\lambda}{2}\|f\|^2_{\mathcal{H}_K}, \quad \text{subject to } f(\mathbf{x}_i) \geq 1,\ \forall i \in I_+. \tag{2.1}$$

여기서 $\mathcal{H}_K$는 주어진 커널함수 $K : \mathbb{R}^p \times \mathbb{R}^p \mapsto \mathbb{R}$에 의해 생성되는 RKHS를 나타내며, $\lambda > 0$는 자료에 대한 적합성과 분류기 f의 복잡성 사이의 균형을 제어하는 조율모수이다. 정규화에 사용된 $\|f\|^2_{\mathcal{H}_K}$는 주어진 RKHS 상에서 측정한 f의 L_2-노름(norm)을 의미한다. RKHS이론에 따르면, 식 (2.1)의 해는 다음과 같은 유한 차원의 형태로 주어진다는 것이 잘 알려져 있으며, 이를 대표자 정리(Representer Theorem) (Kimeldorf와 Wahba, 1971)라 한다.

$$f(\mathbf{x}) = \alpha_0 + \frac{1}{\lambda}\left\{\sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i)\right\}. \tag{2.2}$$

이제, 식 (2.2)를 식 (2.1)에 대입하고 라그랑지(Lagrange)이론을 적용하면, 식 (2.1)을 다음과 같은 이차 계획법 형태의 쌍대 문제(Dual Problem)로 나타낼 수 있으며, R에서 제공하는 `solve.QP{quadprog}`와 `ipop{kernlab}` 같은 상용 패키지를 활용하여 해를 구할 수 있다.

$$\begin{aligned} &\underset{\alpha}{\text{argmax}} \sum_{i=1}^{n}\alpha_i - \frac{1}{2\lambda}\left\{\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)\right\}, \\ &\text{subject to } \sum_{i}^{n}\alpha_i y_i = 0 \quad \alpha_i \geq 0,\ \forall i \in I_+; \quad 1 \geq \alpha_i \geq 0,\ \forall i \in I_-. \end{aligned} \tag{2.3}$$

FDSVM은 반응 변수가 양의 값(y = +1)을 가진 자료와 값이 없는 자료에서 양의 자료를 식별하기 위한 PUSVM (Positive-Unlabeled Support Vector Machines; PUSVM) (Yao 등, 2009)과 목적함수의 구성이 동일하므로, 기술적으로는 PUSVM으로 이해할 수도 있다. 하지만 PUSVM과 FDSVM과는 다른 상황에서 활용되는 방법이므로, 학습된 모형을 평가하는 방법이 상이하다. 따라서 본 연구에서는 PUSVM 대신 FDSVM이라고 칭하고자 한다.

3. F^{2}DSVM: extension to functional predictor

앞의 2절을 통해 FDSVM을 정의하였다. 공변량이 함수형인 경우로 확장하기 위해서는 함수 자료들 간의 비유사성을 측정할 수 있는 적절한 커널 함수를 정의해야한다. 함수형 공변량에 대한 커널 함수만 정의하면, 2절에서 소개한 FDSVM을 함수형 자료에 손쉽게 확장 적용할 수 있다.

본 논문에서는 Park 등 (2008)가 제안한 함수형 자료에 대한 선형 커널을 활용하고자 한다. 반응 변수 $y_i \in \{-1, 1\}$와 함수형 공변량 $\mathbf{x}_i(\mathbf{t}_i) = (x_i(t_{i1})), \ldots, x_i(t_{id_i}))^T$, $i = 1, \ldots, n$로 구성된 자료가 주어졌다고 하자. 공변량이 함수형인 경우, 실제 자료는 함수 자체가 아닌 이산형으로 표현되는 함수값들로 주어진다. 예를 들어, 차원이 각각의 측정 시간이라면 함수형 공변량은 이러한 시간에 따른 측정된 여러 이산형 함수값들을 모아 하나의 곡선으로 생각하는 것이다. 따라서 주어진 이산형 함수값들을 기저시스템에 기반한 함수 형태로 재표현하기 위해 함수형 공변량 $x_i(t)$가 L_2공간의 원소라 가정하자. 주어진 기저(basis) 시스템 $\{\phi_m\}_{m=1}^{\infty}$에 대해

$x_i(t) = \sum_{m=1}^{\infty} c_{i,m}\phi_m(t)$로 표현할 수 있으며, $\mathbf{c}_i$의 추정량 $\hat{\mathbf{c}}_i$는 다음과 같이 추정할 수 있다.

$$\hat{\mathbf{c}}_i = (\hat{c}_{i,1}, \ldots, \hat{c}_{i,M})^T = \underset{\mathbf{c}_i}{\operatorname{argmin}} \sum_{j}^{d_i} \left\{ x_i\left(t_j\right) - \sum_{m=1}^{M} c_{i,m}\phi_m\left(t_j\right) \right\}^2. \tag{3.1}$$

주로 사용되는 기저 시스템에는 푸리에(Fourie), B-스플라인(B-Spline) 등이 있으며, 기저 시스템이 선택되면 충분히 큰 M에 대해 실제 관찰된 $\mathbf{x}_i$에서 $\hat{x}_i(t) = \sum_{m=1}^{\infty} \hat{c}_{i,m}\phi_m(t),\ i = 1, \ldots, n$를 추정할 수 있다.

Park 등 (2008)은 함수 예측변수 간 유사성을 측정하기 위해 다음과 같은 선형 커널을 제안하였다.

$$K\left(\mathbf{x}_i, \mathbf{x}_j\right) = \langle\, \hat{x}_i . \hat{x}_j \rangle\ = \hat{\mathbf{c}}_i^T \Phi \hat{\mathbf{c}}_j. \tag{3.2}$$

여기서, $\Phi = (\int_0^T \phi_i(t)\phi_j(t)dt)_{i,j=1,\ldots,M}$는 M개의 기저함수로 이루어진 기저시스템 하에서의 내적행렬을 나타낸다. 식 (3.2)는 함수 예측 변수의 선형 커널로 간주될 수 있다. 선형 커널 외에 가우시안(RBF) 커널과 같은 비선형 커널 함수를 활용할 수 있으나, 과적합이 발생할 우려가 있다고 알려져 있다 (Park 등, 2008). 이제, 앞절에서 설명한 FDSVM에 식 (2.3)을 대입하면 F^{2}DSVM을 정의할 수 있다.

FDSVM과 마찬가지로 F^{2}DSVM의 성능은 조율모수 λ의 값에 따라 민감하게 변화한다. 따라서 분류기의 학습을 위해서는 조율모수 λ를 신중하게 선택해야 한다. 이를 위해 통상적으로 사용하는 방법은 사전에 선택된 몇 개의 서로다른 조율모수에 대해 각각 F^{2}DSVM을 훈련시킨 후 성능을 비교하여 최적의 λ를 선택하는 것이다. 이를 격자 검색(grid search)이라 한다. 격자 검색은 매우 직관적인 방법이라 많이 활용되긴 하지만, 주어진 λ에 대해 분류기를 반복시켜야하기 때문에 계산량이 상당하다. 뿐만 아니라 사전에 결정된 조율모수 후보, 즉 격자의 선택에 영향을 받는다는 단점이 있다. 이를 해결하기 위한 방법으로 자취해(solution path)를 구하는 방법을 고려할 수 있다. 일반적인 SVM의 경우 쌍대 모수(dual parameter) α가 λ에 대해 조각별 선형성을 만족함이 알려져 있으며, 이를 활용하면 효율적인 자취해 알고리즘을 구현할 수 있다 (Hastie 등, 2004). FDSVM도 SVM을 기반으로 제안된 방법이기 때문에 동일한 형태의 자취해를 구할 수 있다 (Yao 등, 2009). F^{2}DSVM은 함수형 공변량을 활용한 새로운 형태의 커널을 활용했다는 점 이외에는 FDSVM과 동일한 방법이므로 역시 자취해를 구할 수 있으며, 이를 바탕으로 조율모수를 매우 효율적으로 선택할 수 있다.

조율모수의 선택에 있어 가장 중요한 것은 F^{2}DSVM의 성능을 평가하는 척도를 결정하는 것이다. 서론에서 언급했다시피, 부정 사례가 정상 사례보다 월등히 적을 경우, 분류정확도를 활용하여 분류기의 성능을 평가하는 것은 부적절하다. 부정 탐지의 상황에서는 특이도(정상 탐지)보다 민감도(부정 탐지)가 훨씬 중요하다는 사실을 상기하면 원하는 목표 수준의 최소 민감도를 제어하면서 특이도를 극대화하는 것이 최적의 분류기라 할 수 있으며, 본 논문에서는 이를 활용하여 최적의 조율모수 λ를 선택하기로 한다.

4. Wafer data에 대한 적용

본 절에서는 F^{2}DSVM의 성능을 나타내기 위해 실제 자료를 적용한 결과를 관찰하였다. 사용한 자료는 반도체 웨이퍼의 식각(etching) 공정과 관련된 함수 자료이다 (Olszewski와 Thomas, 2001). 개별 웨이퍼의 예측 변수(405nm 파장 방출)는 일정 시간 동안 연속적인 함수 자료를 형성하고 있다. 반응 변수는 전문가가 최종적으로 결정한 웨이퍼 상태이다. 총 1,103개의 웨이퍼 자료 중 1,067개는 정상($y = -1$)이고 127개는 비정상($y = +1$)이다. 측정 시간은 각 웨이퍼에 따라 다르기 때문에, 가장 긴 시간(측정 시간 198)을 기준으로 모든 웨이퍼를 동일한 측정 시간으로 조정하였다. 이 경우 측정 시간을 통일하면서 생성된 값이 없는 자료는 방출 유휴 시간의 최빈값(= 2)로 채우고 첫 번째 방출이 발생한 지점을 기준으로 모든 웨이퍼의 측정 시작점을 통일하였다.

Figure 1은 전체 1,103개의 전처리된 정상 및 비정상 웨이퍼 405nm 파장 방출값(y축)을 측정 시간(x축)에 따라 그린 것이다. 정상 웨이퍼와 비정상(부정) 웨이퍼 모두 파장 방출값 2,000정도까지 상승하는 첫 번째 정상점(peak)까지는 패턴에 큰 차이가 없다. 반면 파장 방출값이 1,300선까지 상승하는 두 번째 정상점에서는

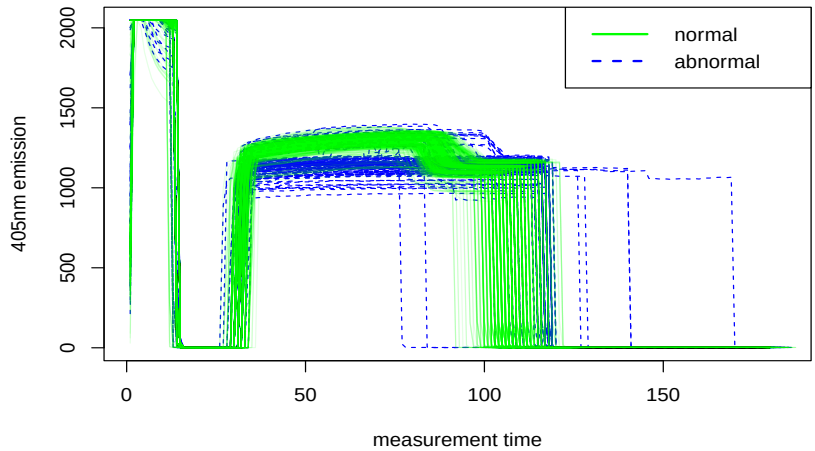


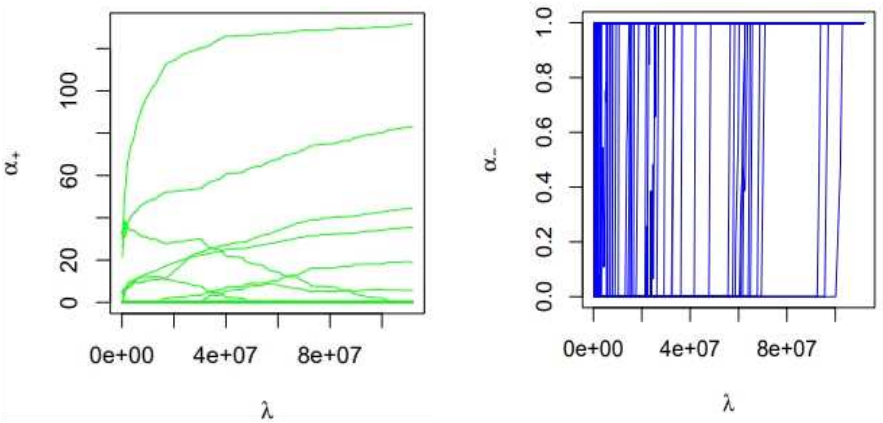
Figure 1: *The total 1,194 wafer data of the 405nm emission during measurement time.Normal wafers in green solid lines and abnormal wafers in blue dotted lines.*

Table 1: Comparison of averaged computing time between the grid search with quadratic programming (QP) for F^{2}DSVM and its path algorithm over fifty repetitions with random partitioning of the wafer data

Training Size (Proportion)	Computing Time (in seconds)		Ratio (QP / Path)	Grid size
	QP	Path		
59 (5%)	0.341 (0.037)	0.053 (0.005)	6.487	73.96 (8.126)
118 (10%)	1.909 (0.112)	0.132 (0.007)	14.444	200.02 (11.700)
297 (25%)	51.561 (0.865)	0.586 (0.010)	88.061	563.54 (7.594)

The number of knots satisfying the piecewise linearity calculated through the solution path algorithm was used as a grid: The computational improvement of the proposed path algorithm becomes substantial as the sample size increases. Standard errors are given in parentheses.

패턴의 차이가 나타난다. 일반적으로 두 번째 방출 정상점에서의 진행 시간이 너무 길거나 너무 짧을 경우, 방출 패턴이 일반적이지 않은 계단 함수를 이룰 경우 등이 부정 사례로 판정됨을 알 수 있다. 위의 탐색적 분석을 통해 자료가 계단 형태의 함수로 나타남을 확인하였고, 함수자료의 표현을 위해 상수 스플라인 기저를 사용하기로 하였다. 스플라인을 활용하는 경우 매듭(knot)점의 선택이 매우 중요하다. 즉, 부정 사례와 정상 사례를 명확히 구분 지을 수 있는 중요한 지점 인근에 많은 매듭을 배치하면 분류기의 성능이 향상될 것이다. 본 연구에서는 매듭점의 위치는 등간격의 크기를 늘려가며 교차 타당성 검증을 통해 최적의 매듭 간격을 찾아보았다. 이러한 접근법은 단순하기는 하지만, 계산이 오래 걸린다는 단점이 있다. 하지만 F^{2}DSVM의 경우 자취해 알고리즘을 활용할 수 있기 때문에 계산 시간을 크게 단축할 수 있다. 자취해 알고리즘의 계산 효율성을 확인하기 위해 웨이퍼 자료에 F^{2}DSVM를 적합시킬 때, 자취해 알고리즘을 적용한 경우(Path)와 단순한 격자 검색을 적용하여 이차계획법으로 푼 경우(QP)를 계산 속도 측면에서 비교해 보았다 (Table 1). 비교의 정확도를 위해 훈련자료와 검증자료의 임의 분할을 50회 반복하고 계산 시간의 평균을 측정하였다. Table 1을 보면, 훈련자료의 크기가 증가함에 따라 자취해 알고리즘의 계산 효율성이 상대적으로 증가함을 알 수 있다. 자취해 알고리즘은 F^{2}DSVM의 쌍대해가 조율모수에 대해 조각별 선형이라는 점을 활용하므로 많은 계산을 필요로하는 수치적 최적화가 전혀 필요하지 않기 때문이다. Figure 2는 실제 웨이퍼 자료에 F^{2}DSVM를 적합시킬 때 계산된 쌍대해의 자취해를 그린 것이다. 좌측 그림 (a)는 부정 개체에 대응하는 쌍대해, $\{\alpha_i, i \in I_+\}$,



(a) (b)

Figure 2: *Trajectories of dual solution as a function of λ for the training set of wafer data we analyzed in Section 4. The left plot represent trajectories of α_i $i \in I_+$ and the right plot shows α_j $j \in I_-$ respectively.*

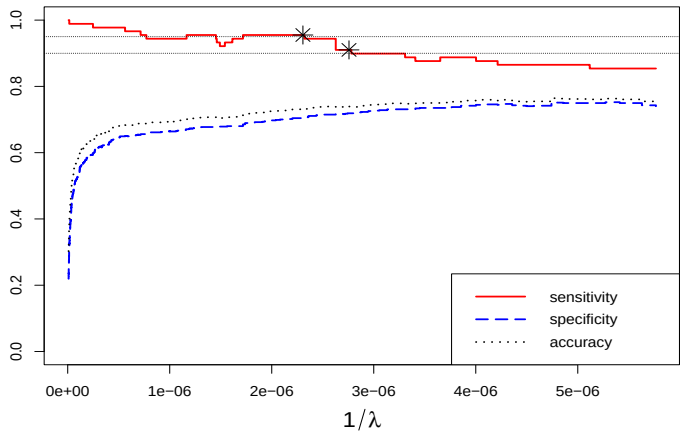


Figure 3: *Trajectories of sensitivity and specificity as a function of $1/\lambda$ for the test set of wafer data. For the left plot, the dotted horizontal lines represent the target level of sensitivities, 0.95 and 0.9 respectively. The two highlighted points are the selected values of $1/\lambda$ that maximize specificity at given sensitivity levels 0.95 and 0.9 respectively.*

우측 그림 (b)는 정상 개체에 대응하는 쌍대해 $\{\alpha_i, i \in I_-\}$를 조율모수의 변화에 따라 나타낸 것이다. 모든 쌍대해가 조율모수에 대한 조각별 선형성을 만족하고 있음을 알 수 있다.

이제, 총 1,194개의 웨이퍼 자료를 무작위로 훈련 자료 70%와 검증자료 30%로 나누었다. 훈련 자료를 활용하여 F^2DSVM의 분류함수를 학습하고, 검증자료를 대입하여 목표 민감도 수준 0.95을 초과하는 $1/\lambda$값 중 최상의 특이도를 나타내는 지점을 선택하였다 (Figure 3참조). 최적의 매듭 간격과 수를 찾기위해 매듭점 사이의 간격을 1부터 18까지 늘려가면서 독립적으로 50회씩 반복하여 비교하였으며, 8의 등간격으로 25개의 매듭점을 배치하는 것이 가장 우수한 성능 보임을 확인하였다. 따라서 이후의 모든 실험에서 상수 스플라인의 등간격 매듭 갯수를 25로 고정하였다. 성능 비교를 위해, SVM과 함수형 SVM(Functional Support

Table 2: The averaged of sensitivity, specificity, accuracy, and G-mean over 50 iterations with random sampling (70% for the train and 30% for the test) from wafer data are reported. Funtional predictors obtained by employing the equally spaced constant basis (knots interval is 8). The numbers in parentheses are standard errors.

Method	Target level	Sensitivity	Specificity	Accuracy	G-mean
SVM	-	0.897 (0.007)	0.992 (0.001)	0.982 (0.001)	0.943 (0.003)
FSVM	-	0.906 (0.010)	0.983 (0.005)	0.975 (0.004)	0.943 (0.006)
FLR	-	0.911 (0.005)	0.805 (0.003)	0.817 (0.003)	0.856 (0.003)
F^{2}DSVM	0.90	0.946 (0.003)	0.717 (0.008)	0.742 (0.007)	0.823 (0.005)
	0.95	0.975 (0.001)	0.602 (0.019)	0.643 (0.017)	0.760 (0.014)

Vector Machines; FSVM), 함수형 로지스틱 회귀모형(Functional Logistic Resgession; FLR)도 적용해 보았으며, 비교의 정확성을 높이기 위해 임의분할을 50회 반복하여 그 평균성능을 비교하였다. 그 결과는 Table 2와 같다. F^{2}DSVM의 경우 조율모수를 선택함에 있어 $1/\lambda$가 증가함에 따라 민감도가 1부터 점진적으로 감소하는 형태를 띄므로 언제나 연구자가 원하는 수준의 민감도를 달성 가능하지만 다른 모델들은 그렇지 못하다. 따라서 SVM, FSVM 그리고 FLR의 조율모수를 선택할 때는 특이도와 민감도에 대한 기하평균(Geometric mean; G-mean)을 활용하여 모형의 성능을 평가하였다. G-mean을 통해 조율모수를 선택하더라도 세 방법은 부정 탐지의 문제를 통상적인 이항 분류 문제의 틀 안에서 풀었기 때문에, 목표한 민감도 수준을 만족할 수 없었다. G-mean이나 정확도의 측면에서는 SVM이나 FSVM이 더 좋은 성능을 보이고 있으나, 원하는 수준의 부정 탐지 능력을 보장해야한다는 측면에서 그 활용이 제한적이다. FLR는 SVM이나 FSVM보다 민감도 부분에서 조금 더 나은 성능을 보여주지만, 다른 기준에서는 모두 뒤떨어지는 경향을 보였으며, 민감도를 여전히 조절할 수 없다는 한계를 지니고 있다. 나아가, 본 연구에서 활용한 웨이퍼 자료의 경우 그 불균형의 정도가 심하지 않다고 볼 수 있는데, 만약 부정의 비율이 더 적은 상황이라면 통상적인 이항분류 기반의 접근은 전체적인 정확도를 높이기 위해 부정의 탐지를 전혀하지 못할 가능성이 높다. 반면, F^{2}DSVM의 경우 언제나 연구자가 원하는 최소 민감도의 수준을 달성할 수 있기 때문에, 부정 탐지의 상황에서는 훨씬 더 적합하다고 할 수 있다.

5. Conclusion

부정 탐지의 문제는 종종 함수형 공변량을 동반한다. 함수자료에 대한 특성을 무시하는 경우, 더 많은 정보를 활용할 수록 차원의 저주로 인해 모형 성능이 저하되는 모순에 빠질 수 있다. 함수자료의 이항 분류에서 널리 활용되는 FSVM의 방법은 비록 함수형 공변량의 특성을 잘 활용하고 있지만, 여전히 부정 탐지의 관점에서는 그 활용이 제한적이다. 본 연구에서는 부정 탐지 지지벡터기계의 아이디어를 함수형 공변량에 확장적용한 F^{2}DSVM을 제안하였다. 제안된 방법은 기존의 FDSVM과 마찬가지로, 쌍대해의 조각별 선형성을 활용하여 자취해를 효율적으로 계산할 수 있을 뿐만 아니라, 사용자가 원하는 수준의 민감도 설정을 통해 부정 탐지에서 활용 가능하다. 실제 웨이퍼 자료 분석 결과는, 매듭점의 위치 선정 등에 있어 배경지식이 부족할 때에도 교차타당성 검증에 사용할 수 있을만큼 충분히 빠른 성능을 보여주었으며, 원하는 수준의 민감도를 조절할 수 있기에 함수 자료의 부정 탐지에 널리 활용 가능한 방법임을 확인하였다.

References

Biau G, Florentina B, and Wegkamp MH (2005). Functional classification in H ilbert spaces, *IEEE Transactions on Information Theory*, **51**, 2163–2172.

Bolton RJ and Hand DJ (2002). Statistical fraud detection: A review, *Statistical Science*, **17**, 235–255.

Chawla NV, Bowyer KW, Hall LO, and Kegelmeyer WP (2002). SMOTE: synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research*, **16**, 321–357.

Feng Y, Zhou M, and Tong X (2021). Imbalanced classification: A paradigm-based review, *Statistical Analysis and Data Mining: The ASA Data Science Journal*, **14**, 383–406.

Hastie T, Rosset S, Tibshirani R, and Zhu J (2004). The entire regularization path for the support vector machine, *Journal of Machine Learning Research*, **5**, 1391–1415.

James GM (2002). Generalized linear models with functional predictors, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**, 411–432.

James GM and Hastie TJ (2001). Functional linear discriminant analysis for irregularly sampled curves, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63**, 533–550.

Kimeldorf G and Wahba G (1971). Some results on Tchebycheffian spline functions, *Journal of Mathematical Analysis and Applications*, **33**, 82–95.

Olszewski and Thomas R (2001). *Generalized Feature Extraction for Structural Pattern Recognition in Time-series Data*, Carnegie Mellon University, Pennsylvania.

Park C, Koo JY, Kim S, Sohn I, and Lee JW (2008). Classification of gene functions using support vector machine for time-course gene expression data, *Computational Statistics & Data Analysis*, **52**, 2578–2587.

Ramsay JO and Silverman BW (2002). *Applied Functional Data Analysis: Methods and Case Studies*, Springer , New York.

Rossi F and Villa N (2006). Support vector machine for functional data classification, *Neurocomputing*, **69**, 730–742.

Vapnik V (1999). *The Nature of Statistical Learning Theory*, Springer science & business media, Berlin.

Wahba G (1990). *Spline Models for OIbservational Data*, SIAM, Pennsylvania.

Woodall WH, Spitzner DJ, Montgomery DC, and Gupta S (2004). Using control charts to monitor process and product quality profiles, *Journal of Quality Technology*, **36**, 309–320.

Yao L, Tang J, and Li J (2009). Entire solution path for support vector machine for positive and unlabeled classification, *Tsinghua Science and Technology*, **14**, 242–251,

Received July 21, 2022; Revised August 10, 2022; Accepted August 14, 2022

불량 웨이퍼 탐지를 위한 함수형 부정 탐지 지지 벡터기계

박민형[a], 신승준[1,a]

[a]고려대학교 통계학과

요 약

빈번하지는 않지만 한번 발생하면 상대적으로 큰 손실을 가져오는 사례를 통칭하여 부정 사례(Fraud)라고 부르며, 부정 탐지의 문제는 많은 분야에서 활용된다. 부정 사례는 정상 사례에 비해 상대적으로 관측치가 매우 적고 오분류의 비용이 월등히 크기 때문에 일반적인 이항분류 기법을 바로 적용할 수 없다. 이러한 경우에 활용할 수 있는 방법이 부정 탐지 지지 벡터기계(FDSVM)이다. 본 논문에서는 공변량이 함수형일 때 활용 가능한 함수형 부정 탐지 지지 벡터기계(F^2DSVM)를 제안하였다. 제안된 방법을 사용하면 함수형 공변량을 가진 데이터에서 사용자가 목표하는 부정 탐지의 성능을 만족시키는 제약하에서 최적의 예측력을 가지는 분류기를 학습시킬 수 있다. 뿐만아니라, 통상적인 SVM과 마찬가지로, F^2DSVM도 자취해의 조각별 선형성을 보일 수 있으며 이를 바탕으로 효율적인 자취해 알고리즘을 활용할 수 있고 분류기의 학습 시간을 크게 단축시킬 수 있다. 마지막으로, 반도체 웨이퍼 불량 탐지 문제에 제안된 F^2DSVM을 적용해 보았고, 그 활용 가능성을 확인하였다.

주요용어: 부정 탐지, 함수자료, 조각별 선형 자취해, 지지 벡터기계

이 논문은 정부 (미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2018R1D1A1B07043034).

[1]교신저자: (02841) 서울특별시 성북구 안암로 145, 고려대학교 통계학과. E-mail: sjshin@korea.ac.kr