Journal of the Korea Academia-Industrial cooperation Society
Vol. 22, No. 8 pp. 287-293, 2021

https://doi.org/10.5762/KAIS.2021.22.8.287
ISSN 1975-4701 / eISSN 2288-4688

데이터의 중요 특성 선택과 전처리에 기반한 분류체계의 적용

조현우
대구대학교 융합산업공학과

Application of Classification Method Based on Selection of Important Features and Filtering of Measurement Data

Hyun-Woo Cho
Department of Industrial Convergence Engineering, Daegu University

요 약 다수의 변수가 존재하는 데이터 분석을 수행하는 경우 중요성이 낮거나 타 변수와 중복되어 불필요한 정보를 제공하는 변수들을 제거함으로써 구축된 데이터 모델의 성능을 향상시킬 수 있다. 이러한 특성 선택 (feature selection)은 과거부터 여러 다양한 목적과 많은 실제 문제에서 활발히 사용되고 있는데 그중에서도 특히 데이터에 기반한 클래스의 분류 (classification) 문제에서는 특성 선택 사용 시 집단 또는 클래스 간 구별력이 높아지는 장점이 있다. 본 연구에서는 서포트 벡터 머신 (support vector machine, 이하 SVM)에 기반한 특성 선택 기법과 전처리 필터링을 결합한 분류 체계를 제안하였다. 특성 선택을 수행하기 이전에 전처리 과정으로서 데이터의 필터링을 진행하였으며 분류 성능을 높이는 데 필수적이며 기여도가 큰 특성들만을 추출하여 이를 모델 데이터로 구성하였다. 분류 체계의 성능을 검증하기 위하여 사례 연구로서 3개의 데이터 세트에 대하여 SVM에 기반한 두 종류의 특성 선택 방법의 분류 정확도를 비교하였다. 그 결과 사용된 3개 데이터 세트 모두에서 제안된 분류 체계의 성능이 기존 방법 대비 향상되었으며 가우시안 커널 함수를 사용하였을 때 상대적으로 높은 정확도를 보여주었다.

Abstract The performance of data-based empirical models can be improved by removing variables or features that are of low importance or strongly correlated with other features. Such a feature selection has been used for various purposes, and it helps increase the discriminating power of classification models between groups or classes. This study presents a classification scheme that combines a feature selection technique based on support vector machines and preprocessing filtering of raw measurement data. Before performing the feature selection task, raw data is filtered as a pre-processing step to remove unwanted variation. In the next feature selection, important features with high contribution to classification are extracted and used as classification training data instead of raw data. A case study on three cases was conducted to verify the performance of the classification scheme. The classification accuracy of the proposed scheme was compared with other schemes. As a result, the accuracy of the proposed classification scheme improved significantly in the three cases. In addition, it showed relatively higher classification performance when the Gaussian kernel was used in the three cases.

Keywords : Classification, Data Analysis, Feature Selection, Filtering, Support Vector Machine

This research was supported by Daegu University Research Grant, 2018.
*Corresponding Author : Hyun-Woo Cho(Daegu Univ.)
email: hwcho@daegu.ac.kr
Received April 26, 2021 Revised July 5, 2021
Accepted August 5, 2021 Published August 31, 2021

1. 서론

분류 (classification), 군집화 (clustering), 회귀 분석 (regression analysis) 등 데이터 분석 과정에서 발생하는 주된 문제는 데이터에 포함된 변수 또는 특성이 많을 때 주로 발생하게 된다[1,2]. 데이터 내에 있는 변수는 동등하게 중요하지 않으며 심지어 불필요하거나 중복된 정보를 제공하는 경우가 많아 이를 제거하는 것이 향후 데이터 분석에 도움이 될 수 있다. 분류에 있어 이러한 특성 선택 (feature selection)의 목적은 분류에 기여도가 높은 소수의 선택된 특성들을 찾아내어 이를 분류에 사용함으로써 전체 특성을 사용할 때 대비 분류 클래스 간 구별력을 높이고 최종적으로 분류 정확도를 높이는 데 있다. 최근 관심이 커지고 있는 빅데이터, 인공지능, 기계학습 분야에서도 수집되는 변수의 수는 센서와 사물인터넷, 사회 관계망 등의 확산으로 급격히 늘어나는 상황이므로 이러한 특성 선택의 중요성 역시 커지고 있다.

방법론적인 측면에서는 데이터의 비선형성이 증가하고 대상 문제의 복잡도가 높아지면서 비선형 방법론의 활용도 늘어나고 있다. SVM을 비롯하여 커널 주성분 분석 (kernel principal component analysis), 커널 부분 자승법(kernel partial least square), 커널 독립 성분 분석 (kernel independent component analysis) 등의 비선형 기법이 활발히 활용되고 있다[2-6]. 커널 주성분 분석과 커널 독립 성분 분석은 데이터의 설명 (description)과 해석 (interpretation)에 장점을 가지며 커널 부분 자승법은 데이터의 독립변수들로 종속 변수를 예측하기 (prediction) 위해 사용되고 있다[7-9]. SVM은 분류 문제가 발생하는 이미지나 영상 데이터, 개인이나 기업의 신용 데이터, 텍스트 마이닝과 결합한 문서나 사용자 리뷰 데이터 등 다양한 영역에 적용되고 있다[7,8]. 이러한 SVM의 분류 문제 적용에 있어서 특성 선택을 마이크로 어레이 (micro array) 데이터에 구현한 재귀적 특성 제거 (recursive feature elimination) 기법은 대표적인 사례라 할 수 있다[6]. 여기에서는 DNA 마이크로 어레이에 기록된 광범위한 패턴의 유전자 발현 데이터에서 분류에 적합한 유전자들만을 선택하였다. 선택된 유전자 집합은 백혈병과 대장암 사례에서 중복된 유전자를 제거하여 분류 성능을 향상시켜 주었으며 생물학적으로 암과 관련이 높음을 보여주었다[6].

이 논문에서는 비선형 분류 방법인 SVM에 기반한 특성 선택 기법과 전처리 필터링 과정에 기반한 분류 체계를 제안한다. 분류 성능을 검증하기 위하여 3개의 공개 데이터 세트를 활용하여 이에 대한 사례 연구를 수행하고자 한다. 제안된 분류 체계에서는 특성 선택을 통해 데이터의 차원을 줄이기 이전에 데이터의 불필요하거나 중복되는 정보를 필터링하는 전처리 과정이 추가된다. 이를 위하여 직교 신호 보정 (orthogonal signal correction)을 적용하여 데이터의 클래스 멤버십 (class membership)에 직교하거나 상관관계가 없는 데이터 부분을 제거하게 되는데 특성 선택 이전 단계에서 수행된다. 다양한 분야에서 여러 목적으로 사용되는 분류 문제는 많은 변수 또는 특성을 처리해야 하는 경우가 대부분이다. 이 때 모든 변수나 특성을 분류에 포함하지 않고 분류 성능을 높이는 데 필수적이며 기여도가 높은 특성들의 집합을 별도로 구성한다면 분류 정확도의 향상을 기대할 수 있을 것이다. 이러한 분류 체계의 성능 평가를 위하여 3개의 데이터 세트에 대해 그 분류 정확도를 비교하였다. 우선 SVM 방법론에 대한 간략한 소개와 SVM 기반 특성 선택 기법들, 그리고 전처리 기법을 살펴본 후 제안된 분류 체계의 사례 연구 결과를 제시한다.

2. 방법론

SVM은 분류나 회귀 분석에 응용할 수 있도록 데이터에 대한 지도 학습 (supervised learning)을 수행하는 기법이다[6]. 기본적으로 대상 데이터가 고차원의 특성 공간 (feature space)로 투영되고 최대 마진 (maximum margin)을 갖도록 최적 결정 함수 (optimal decision function)이 주어진다. 이러한 최적 결정 함수는 아래의 부등식을 만족하는데

$$y_i(w\Phi(x_i)+b)-1 \geq 0 \qquad (1)$$

이는 다시 아래의 듀얼 문제 (dual problem)으로 변환될 수 있다. 따라서 SVM에서의 지도 학습이란 α_i, b, 그리고 서포트 벡터를 결정하는 것으로서 아래와 같이 풀기 쉬운 듀얼 문제로 주어진다[6].

$$L_d = \sum \alpha_i - \frac{1}{2}\sum \alpha_i \alpha_j y_i y_j \Phi(x_i)\Phi(x_j) \qquad (2)$$

Eq. (2)에서 내적은 $K(x_i,x_j)=\Phi(x_i)\Phi(x_j)$로 주어지는 커널 함수로 대체되며 해는 $\alpha_i > 0$인 서포트 벡터에 대해 계산된 $w=\sum \alpha_i y_i \Phi(x_i)$으로 주어진다.

Fig. 1의 예에서 보이듯이 커널 함수를 사용하여 내적 계산과 비선형 매핑 (nonlinear mapping)을 대체하는데, 이는 데이터를 선형적으로 분리되는 고차원 특성 공

간으로 투영하는 것을 의미하며 이를 다시 원래의 공간으로 되돌리면 비선형 결정 경계면을 얻을 수 있다. 특성 또는 변수 선택 기법은 데이터 시각화 (visualization), 차원 축소 (dimension reduction), 해석 등의 장점을 가지고 있다. 앞에서 설명한 SVM에서의 특성 선택 기법인 재귀적 특성 제거 방법의 경우에는 데이터와 클래스 정보에 대하여 비용 함수 (cost function) J를 아래와 같이 정의할 수 있다 [6].

$$J = (1/2)\alpha^T H\alpha - \alpha^T e \quad (3)$$

위의 식에서 α는 라그랑주 승수 (Lagrange multiplier), e는 1로 구성된 벡터, $H_{hk} = y_h y_k K(x_h, x_k)$, 그리고 $K(x_h, x_k)$는 커널 함수를 나타낸다.

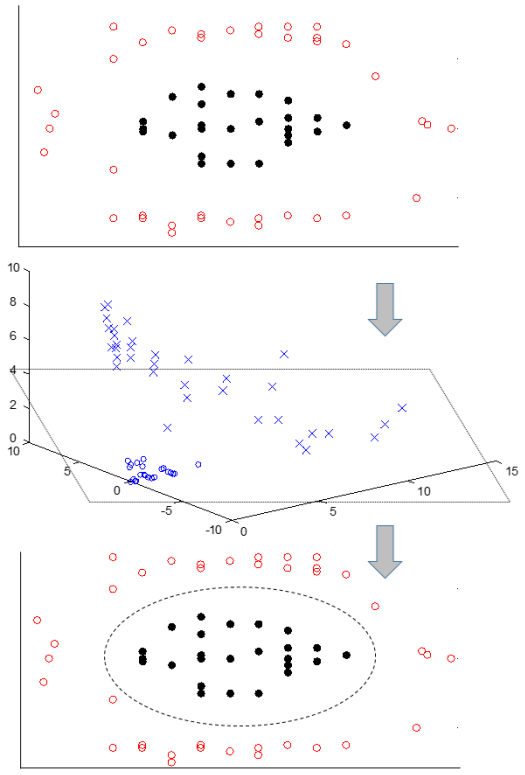

Fig. 1. Simple illustration of SVM and kernel function

J의 변화량 중 알고리즘에서 제거되는 i번째 변수로 인해 생긴 것을 계산하기 위해 α은 일정하다고 가정하며 i번째 변수의 제거를 $(-i)$로 표시하면 아래의 결과를 얻을 수 있는데

$$H(-i)_{hk} = y_h y_k K(x_h(-i), x_k(-i)) \quad (4)$$

여기에서 $K(x_h(-i), x_k(-i))$는 커널 함수를 나타낸다. 이를 통하여 아래와 같은 민감도 함수 $DJ(i)$를 얻을 수 있다:

$$\begin{aligned} DJ(i) &= J - J(-i) \\ &= (1/2)\alpha^T H\alpha - (1/2)\alpha^T H(-i)\alpha. \end{aligned} \quad (5)$$

SVM의 재귀적 특성 제거 알고리즘은 아래의 절차를 s가 빈 배열이 될 때까지 반복하여 수행된다[6].

(1) 새로운 샘플의 구성 $X = X_{all}(:, s)$

(2) SVM의 α계산 및 순위 기준 구함

(3) $DJ(i) = (1/2)\alpha^T H\alpha - (1/2)\alpha^T H(-i)\alpha$

(4) $f = \arg\min_i DJ(i)$만족하는 변수 f를 찾음

(5) r의 업데이트 및 s에서 f 제거

$r = [s(f), r], s = s - s(f)$.

직교 신호 보정 기법은 대상 데이터의 불필요한 정보를 제거하기 위한 전처리 알고리즘이다[10]. 데이터의 클래스 멤버십에 대한 정보를 가진 종속 변수에 직교하거나 상관관계가 없는 데이터 부분이 독립 변수 데이터에서 필터링 된다. 직교 신호 보정의 첫 번째 단계는 독립 변수 데이터에 대해 첫 번째 주성분 스코어 (score) 벡터 t를 계산한다. 이러한 t 스코어 벡터는 클래스 멤버십에 직교화 (orthogonalized) 되어 다음과 같이 보정 벡터 t^*를 생성한다[10]:

$$t^* = I - Y(Y^T Y)^{-1} Y^T t. \quad (6)$$

그런 다음 부분 최소 제곱법 가중치 (weight) 벡터 w가 $Xw = t^*$에서 계산되고 새로운 스코어 벡터 $t = Xw$가 계산된다. 이러한 프로세스는 t가 수렴될 때까지 반복된다. 마지막으로, 로딩 벡터 (loading) p가 계산되고 보정치인 tp^T가 오리지널 데이터에서 차감되어 잔차 (residual)를 구할 수 있다. 다음 컴포넌트는 이러한 방식으로 지속적으로 계산하게 된다.

앞에서 살펴본 SVM의 재귀적 특성 제거 기법과는 다른 새로운 특성 선택 기법은 변수의 랭킹 (ranking)을 결정하기 위하여 최초 모든 변수로 학습한 이후 특성 선택 기준을 계산하게 된다[11]. 이를 통해 제거된 특성을 제외한 나머지 변수들로만 재학습한 후 다시 최소 특성 선택 기준을 가진 변수를 제거하는 과정을 통해 모든 변수에 대한 랭킹을 결정할 수 있다. 여기서 $X_{all} = [x_1, \ldots, x_l]'$로 주어지는 학습 데이터, 클래스 레이블 y, $s = [1,2,...,n]'$일 때 선형 커널에 대한 절차는 우선 학습 데이터를 $X = X_{all}(:, s)$과 같이 구성한다. $W = \nabla g(X) = \sum_{i \in SV} \alpha_i y_i x_i$와 같이 SVM 학습을 통해 이를 계산한 후 최소 w_j를 가진 특성을 제거하여 업데이트하는 과정을 반복하게 된

다. 비선형 커널일 때의 절차는 선형 커널과 유사하고 단지 $d = \sum_{i \in SV} P_i$ 값이 최소가 되는 특성을 제거하며 아래의 수식을 변형하여 계산하는 차이만 있을 뿐이다.

$$W = \sum_{i \in SV} \alpha_i y_i \nabla_x K(x_i, x) \quad (7)$$

$$P_i = \left| \nabla g(x_i) \right| / \| \nabla g(x_i) \|^2 \quad (8)$$

3. 사례 및 결과

본 연구에서 제안된 분류 체계의 성능을 평가하기 위하여 세 가지 종류의 공개 데이터를 활용하여 사례 연구를 진행하였다. Case 1은 결장암 데이터로서 62개의 인스턴스로 구성되며 각 인스턴스에는 2,000개의 유전자 데이터가 있다. 데이터 세트는 22개의 정상 조직 샘플과 40개의 결장 종양 샘플로 이루어져 있어 전체 결장암 데이터는 (62 x 2,000) 차원의 매트릭스로 주어진다 (Fig.2). 결장암 데이터는 학습 데이터 (training data)와 테스트 데이터 (test data)가 별도로 설정되어 있지 않기 때문에 두 그룹으로 무작위로 분할되었다.

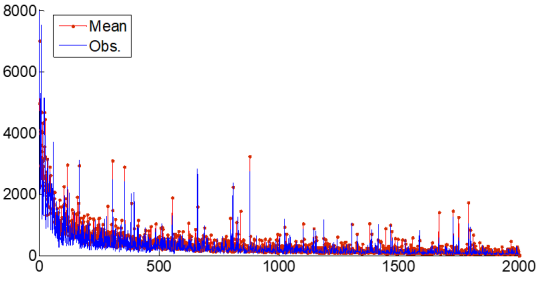

Fig. 2. Variable plot for case 1

Case 2는 지구 대기의 전리층 데이터로서 351개 인스턴스를 가지는데 각 인스턴스는 전리층으로부터의 레이더 반사 측정값을 나타낸다. 225개의 인스턴스는 "양호"라고 표시되며 126개의 인스턴스는 그 신호가 전리층을 통화하기 때문에 “나쁜”으로 레이블되어 있다. Case 3는 수중 음파 탐지 데이터로서 수중에서 수집된 소나 반환값을 나타낸다. 전체 1,200개의 반환값 중 강도 측면에서 208개의 반환값을 선택하여 사용하였으며 Case 1과 Case 2에서 하였듯이 무작위로 학습 데이터와 테스트 데이터를 구분하였다 (Fig. 3). 3개의 데이터 세트에 전처리 과정으로 직교 신호 보정 필터링을 적용하였는데 본 연구에서는 신뢰도가 높다고 알려진 직접 직교 신호 보정 알고리즘을 사용하였다[12]. 직교 신호 보정 필터링을 포함한 모든 실험은 매트랩 프로그램 환경에서 수행하였으며 선형과 비선형 커널 중 우수한 가우시안 커널을 사용하여 결과를 표시하였다.

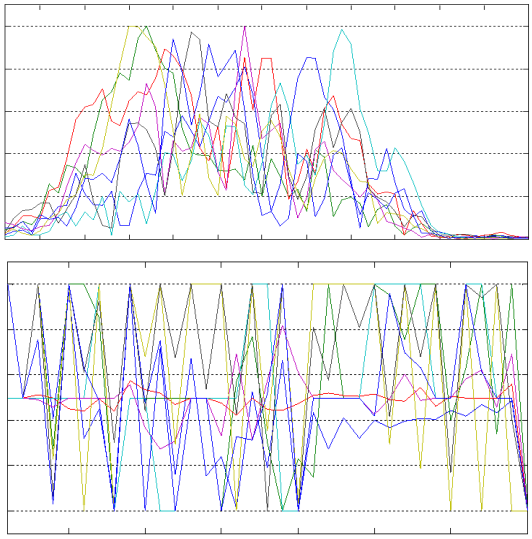

Fig. 3. Variable plot for case 2 (upper) and case 3 (lower)

Table 1은 Case 1에 대한 분류 정확도 결과를 선택된 특성의 수에 따라 보여주고 있다. 비교되는 방법론 (“compared”)과 제안된 방법론 (“proposed”) 각각에 대하여 선형 및 가우시안 커널에 대해 결과치가 표시되어 있다. 표에서 보여주듯이 전체적인 분류 성능관점에서 보면 가우시안 커널을 사용하는 것이 선형 커널을 사용할 때 보다 더 우수한 것으로 나타났다.

Table 1. Classification accuracy for case 1

No.	Linear Kernel		Gaussian Kernel	
	Compared	Proposed	Compared	Proposed
1	0.69	0.69	0.71	0.66
2	0.74	0.74	0.66	0.66
3	0.71	0.84	0.66	0.73
4	0.73	0.82	0.65	0.74
5	0.73	0.73	0.66	0.76
10	0.73	0.73	0.71	0.76
49	0.76	0.76	0.85	0.87
68	0.77	0.77	0.84	0.90
92	0.81	0.81	0.82	0.89
100	0.81	0.81	0.84	0.89

여기서 선형 커널을 사용하였을 때 분류 정확도 범위는 0.69~0.81로 주어지며, 비교된 방법과 제안된 방법은 동일한 분류 정확도 결과값을 보여주고 있다. 한편, 가우시안 커널을 사용하는 경우 분류 정확도는 비교 방법론에서는 0.65~0.85이고 제안된 방법론에서는 0.66~0.90를 보이고 있다.

구체적으로 살펴보면, 제안된 방법론의 최대 분류 정확도는 전체 변수 중 68개의 선택된 특성을 사용할 때 0.90이라는 결과값을 얻었다. 전체 2,000개의 유전자 데이터 중 분류에 중요하다고 선택된 유전자 특성들만으로 얻어진 결과로서 2,000개의 유전자 중 대부분이 분류에 기여하지 못하거나 중복된 결과로 판단된다. Table 1에서 비교 방법론의 경우 가우시안 커널 사용 시 최대 분류 정확도는 49개의 특성을 사용하였을 때의 0.85임을 알 수 있다. 여기서 비교 방법론은 제안된 방법론 (68개 특성) 보다 적은 수의 특성인 49개를 사용하였지만 최대 분류 정확도는 제안 방법론 대비 낮아진 (0.90 vs. 0.85) 것이다. 또한 49개의 동일한 특성을 사용할 때에도 제안 방법론이 비교 방법론보다 높은 분류 정확도를 (0.87 vs. 0.85) 보이고 있다.

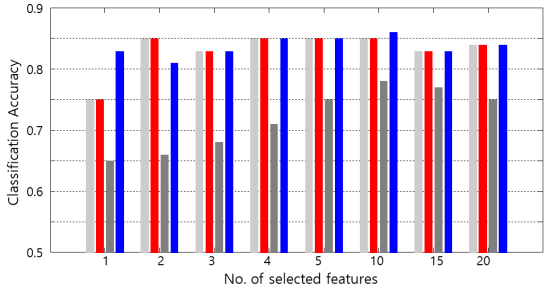

Fig. 4. Results for case 2 with number of selected features

Fig. 4는 Case 2에 대한 분류 정확도 결과값이 선택된 특성의 수에 따라 표현되어 있다. Case 1과 동일하게 선형 및 가우시안 커널에 대하여 비교 및 제안 방법론을 적용하였다. 선택된 특성의 수에 대해 4개의 바 그래프가 존재하는데 왼쪽부터 선형커널의 비교 방법론, 선형 커널의 제안 방법론, 가우시안 커널의 비교 방법론 (이하 가우시안-비교), 그리고 가우시안 커널의 제안 방법론 (이하 가우시안-제안)이 표시되어 있다.

전체적으로 Case 2에 대한 최대 분류 정확도는 "가우시안-제안" 방법론에서 10개의 선택된 특성을 사용할 때 0.86이 달성되었다. 반면에 비교 방법론은 가우시안 커널과 동일한 10개 특성을 사용하였을 때 0.78의 최대 분류 정확도를 보여 차이가 큼을 알 수 있다.

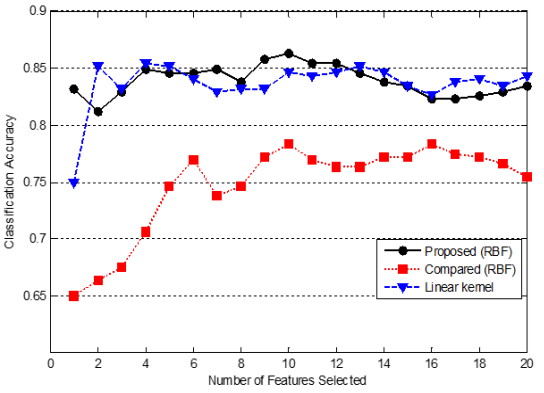

Fig. 5. Performance comparison for case 2

가우시안-제안 방법론은 선택된 특성의 수와 관계없이 가우시안-비교 방법론보다 좋은 분류 성능을 보여주고 있다. 한편, Case 1과는 다르게 Case 2 결과에서는 선형 커널 방법론의 정확도가 가우시안 커널 대비 크게 저하되지 않으며 오히려 가우시안-비교보다는 우수하고 가우시안-제안 방법론과 비슷함을 알 수 있다 (Fig. 5). Fig. 5에서 선형 커널의 결과값은 비교 및 제안 방법론에서 동일하기에 하나의 선으로 표시되어 있다.

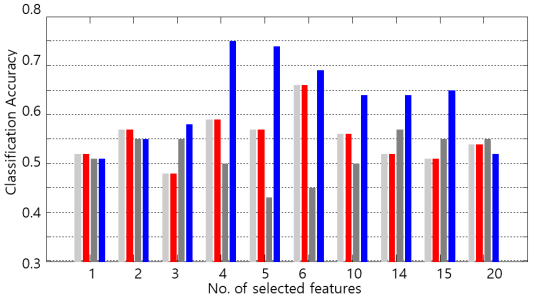

Fig. 6. Results for case 3 with number of selected features

Fig. 6에서는 Case 3에 대한 분류 정확도 결과를 선택된 특성의 수에 따라 나타내고 있다. 앞에서 기술하였던 Case 1과 Case 2와 같은 방식으로 선형 및 가우시안 커널을 사용하여 결과를 도출하였다. 전체 분류 정확도 측면에서 최고값은 가우시안-제안 방법론에서 4개의 선택된 특성을 사용할 때 0.75를 얻을 수 있었다. 두 번째로 높은 분류 정확도는 선형 커널에서 선택된 6개 특성

사용 시 0.66을 얻을 수 있었다. 가우시안-비교 방법론에서의 최대 분류 정확도는 14개 특성 선택 시 얻어진 0.57로 다른 방법론 대비 낮게 구해졌다. 이러한 결과는 Case 2에 대하여 제시되었던 Fig. 5와 동일하게 작성된 Fig. 7에서 확인할 수 있다.

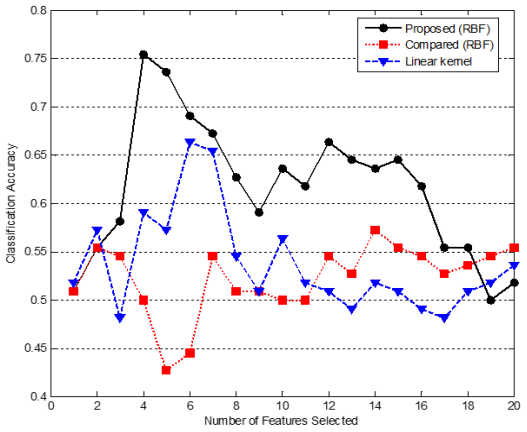

Fig. 7. Performance comparison for case 3

4. 결론

본 논문에서는 SVM에 기반한 특성 선택과 전처리 필터링을 결합한 분류 체계를 제안하고 이를 세 개의 데이터 분류 문제에 적용하였다. 최초의 데이터에 직교 신호 보정 필터링을 수행하여 데이터의 잡음을 제거하였고 이후에는 모든 변수를 사용하는 대신 랭킹이 높아 기여도가 높은 특성들의 집합을 별도로 구성하였다. 세 개의 데이터 세트에 대하여 분류 정확도를 비교한 결과 제안된 방법의 성능이 기존 방법 대비 향상되었음을 확인할 수 있었다. 가우시안 커널의 사용 시 보다 높은 정확도를 얻었으며 선형 커널도 데이터에 따라서는 유사한 성능을 보여주었다. 향후 연구로서 제안된 분류 체계를 복잡한 실제 공정의 측정 데이터에 실시간으로 적용한다면 유효한 결과를 도출할 수 있을 것이다. 이런 상황에서는 데이터의 잡음이 크고 공정 변수의 상호 연관성이 커 실제 모니터링에 중요한 특성값들은 극히 소수에 국한되게 된다. 따라서 실시간 데이터에 기반한 공정 모니터링은 공정의 정상과 이상 상태를 실시간으로 분류하는 문제로 쉽게 변환되는 장점을 가질 수 있다. 또한 이를 정상과 비정상의 두 개의 클래스 분류문제에서 다중 클래스 분류로 확장한다면 설비나 공정의 이상 원인 진단에서도 적용이 가능할 것이다.

References

[1] K. Tidriri, N. Chatti, S. Verron, T. Tiplica, "Bridging data-driven and model-based approaches for process fault diagnosis and health monitoring: A review of researches and future challenges", *Annual Reviews in Control*, Vol.42, pp.63-81, 2016.
DOI:https://doi.org/10.1016/j.arcontrol.2016.09.008

[2] B. Schölkopf, A. J. Smola, K. Müller, "Nonlinear component analysis as a kernel eigenvalue problem", *Neural Computation*, Vol.10, pp.1299-1319, 1998.
DOI:https://doi.org/10.1162/089976698300017467

[3] S.J. Qin, "Survey on data-driven industrial process monitoring and diagnosis", *Annual Reviews in Control*, Vol.36, pp.220-234, 2012.
DOI:https://dx.doi.org/10.1016/j.arcontrol.2012.09.004

[4] G. Baudat, F. Anouar, "Generalized discriminant analysis using a kernel approach", *Neural Computation*, Vol.12, pp.2385-2404, 2000.
DOI:https://doi.org/10.1162/089976600300014980

[5] D. M. J. Tax and R. P. W. Duin, "Support vector data description", *Machine Learning*, Vol.54, pp.45-66, 2004.
DOI: https://doi.org/10.1023/B:MACH.0000008084.60811.49

[6] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, "Gene selection for cancer classification using support vector machines", *Machine Learning*, Vol.46, pp.389-422, 2002.
DOI:https://doi.org/10.1023/A:1012487302797

[7] H. Li, Y. Liang, Q. Xu, "Support vector machines and its applications in chemistry", *Chemometrics and Intelligent Laboratory Systems*, Vol. 95, pp. 188-198, 2009.
DOI: https://doi.org/10.1016/j.chemolab.2008.10.007

[8] A.J. Smola, B.A. Schölkopf, "A tutorial on support vector regression", *Statistics and Computing* , Vol.14, pp.199-222, 2004.
DOI:https://doi.org/10.1023/B:STCO.0000035301.49549.88

[9] B.M. Henrique, V.A. Sobreiro, H. Kimura, "Literature review: machine learning techniques applied to financial market prediction", *Expert Systems with Applications*, Vol.124, pp. 226-251, 2019.
DOI: https://doi.org/10.1016/j.eswa.2019.01.012

[10] S. Wold, H. Antti, F. Lindgren, J. Öhman, "Orthogonal signal correction of near-infrared spectra", *Chemometrics and Intelligent Laboratory Systems*, Vol.44, pp.175-185, 1998.
DOI:https://doi.org/10.1016/S0169-7439(98)00109-9

[11] H. Cho, S. Baek, E. Youn, M. Jeong, A Taylor, "A two-stage classification procedure for near-infrared spectra based on multi-scale vertical energy wavelet thresholding and SVM-based gradient-recursive feature elimination", *Journal of Operational research Society*, Vol.60, pp.1107-1115, 2009.
DOI:https://doi.org/10.1057/jors.2008.179

[12] J. Westerhuis, S. Jong, A. Smilde, "Direct orthogonal signal correction", *Chemometrics and Intelligent Laboratory Systems*, Vol.56, pp.13-25, 2001.
DOI:https://doi.org/10.1016/S0169-7439(01)00102-2

조 현 우(Hyun-Woo Cho) [정회원]

- 2003년 8월 : 포항공과대학교 기계산업공학부 (공학박사)
- 2003년 8월 ~ 2007년 8월 : 포항공과대학교, 조지아텍, 테네시주립대 연구원
- 2007년 9월 ~ 2011년 2월 : 삼성전자, 삼성디스플레이 책임연구원
- 2011년 3월 ~ 현재 : 대구대학교 융합산업공학과 교수

〈관심분야〉
공정모니터링, 빅데이터분석, 인공지능