산업기술연구(강원대학교 산업기술연구소 논문집), 제33권 A호, 2013.
Journal of Industrial Technology, Kangwon Natl. Univ., Korea, No. 33 A, 2013.

SVM을 이용한 고속철도 궤도틀림 식별에 관한 연구

A Study on Identification of Track Irregularity of High Speed Railway Track Using an SVM

김 기 동* 황 순 현**
Kim, Ki-Dong Hwang, Soon-Hyun

Abstract

There are two methods to make a distinction of deterioration of high-speed railway track. One is that an administrator checks for each attribute value of track induction data represented in graph and determines whether maintenance is needed or not. The other is that an administrator checks for monthly trend of attribute value of the corresponding section and determines whether maintenance is needed or not. But these methods have a weak point that it takes longer times to make decisions as the amount of track induction data increases.

As a field of artificial intelligence, the method that a computer makes a distinction of deterioration of high-speed railway track automatically is based on machine learning. Types of machine learning algorism are classified into four type: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning.

This research uses supervised learning that analogizes a separating function form training data. The method suggested in this research uses SVM classifier which is a main type of supervised learning and shows higher efficiency binary classification problem. and it grasps the difference between two groups of data and makes a distinction of deterioration of high-speed railway track.

키워드 : *인공지능, 기계학습, 지지 벡터 기계*
Keywords : *artificial intelligence, machine learning, support vector machine(SVM)*

1. 서론

한국고속철도(KTX)는 2004년 4월 개통을 하여 평균 시속 300km의 속도로 운행하고 있다. 이러한 고속선의 궤도는 일반노선에 비해 상대적으로 열차의 동적하중이 크게 증폭되고, 비슷한 정도의 궤도틀림이 발생한 경우 궤도틀림이 일반철도보다 빠른 속도로 진전된다. 궤도 틀림의 진전은 궤도의 수명과 열차 운행의 안전에 직결되는 문제이기 때문에 체계적인 궤도 유지관리가 매우 중요하다[8].

고속선 궤도검측은 총 0km~300km 구간을 25cm 단위로 검사하고 검측은 약 두 달에 세 번 간격으로 시행하고 있다. 이때 보수구간 선정방식은 검측항목이 임계치를 초과한 구간이 존재하거나 궤도틀림의 정도가 증가하는 추세를 보이는 구간에 대하여 선정하고 있다. 하지만 300km 구간의 검측데이터는 1,200,000건이 발생한다. 이때 여러 개의 검측항목을 확인하고 궤도의 이상 유무를 판

* 강원대학교 산업공학과 교수, 공학박사, 교신저자
** 강원대학교 대학원 산업공학과 석사과정

단하는 것은 오랜 시간이 소요되는 단점을 가지고 있다.

본 논문에서는 보수구간의 검측데이터와 미 보수구간의 검측데이터 사이의 차이점을 파악하고, 궤도의 이상유무를 판별하기 위하여 SVM(Support Vector Machine)이라는 기계학습방법을 이용한다. 보수 이전에 검측된 데이터를 SVM 분류기를 이용하여 학습을 시키고 이후의 검측데이터에 적용하여 두 개의 클래스(궤도이상 유/무)로 분류한다. 이는 궤도 속성 값들이 임계치를 넘지 않았더라도 궤도의 이상 유무를 판단 할 수 있도록 한다. 또한 검측데이터 속성값과 SVM과의 적합성을 판단하기 위해 다양한 커널을 사용하여 판별된 값을 비교하였다.

2. 관련 연구현황

2.1 SVM(Support Vector Machine) 분류기

SVM의 이론은 1979년에 통계학적인 학습이론을 기초로 하여 Vapnik에 의해 처음 제안되었다[12]. 기본적인 SVM은 이진분류문제에 널리 이용되며, hyperplane을 중심으로 한쪽은 positive 클래스, 다른 한쪽은 negative 클래스로 나눈다. SVM의 가장 기본적인 아이디어는 두 개의 범주(positive, negative 클래스)를 구성하는 데이터들을 가장 잘 분리해 낼 수 있는 초평면(hyperplane)을 찾는 것이다[13].

$(x_i, y_i)(i=1,...,N)$를 훈련데이터 집합 S, 데이터 $x_i \in R^N$가 클래스 $y_i \in \{-1,1\}$에 속한다고 하면, 두 데이터 집합의 경계가 되는 초평면은 다음의 수식 (1)로 표현된다. 즉, SVM은 두 개의 클래스 사이의 마진(margin)을 최대로 할 수 있는 초평면을 찾아 두 데이터를 이진분류 하는 방식이다. SVM의 이진분류 방식은 다음의 그림 1과 같다.

$$w \cdot x + b = 0,\ w \in R^N, b \in R \quad (1)$$

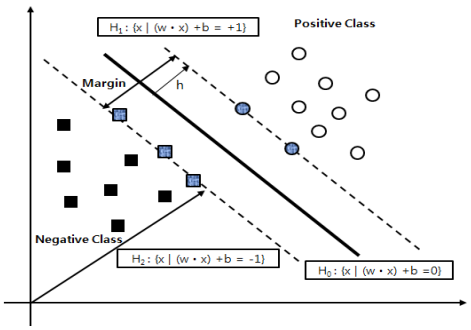

그림 1 SVM을 이용한 데이터 set의 이진 분류

여기서, w는 초평면의 법선 벡터로 초평면의 방향을 나타내고, b는 위치를 나타내며, x는 N차원의 입력벡터를 나타낸다. 학습문서 집합 $S=(x_i,y_i)$에서 입력데이터 x_i가 클래스에 속하면 y_i는 +1의 값을 가지고, 속하지 않으면 -1의 값을 가지게 되며 이는 수식 (2)~(3)으로 표현된다. 결국 SVM은 최적의 w와 b를 찾는 문제이다.

$$w \cdot x_i - b \geq +1, \quad \text{if } y_i = 1 \quad (2)$$
$$w \cdot x_i - b \geq -1, \quad \text{if } y_i = -1 \quad (3)$$

위의 수식을 단일항의 식으로 나타내면 다음 수식 (4)와 같다.

$$y_i(w \cdot x_i) + b \geq 1, \quad \text{if } i = 1,2,...,N \quad (4)$$

일반화 손실 없이 두 개의 범주를 분리해내는 선형 분류기라 할지라도 좁은 마진폭을 갖는 선형 분류기의 경우 예상 리스크가 높기 때문에 최대 마진(γ)을 갖는 선형 분류기보다 낮은 일반화 성능을 보이게 된다. 다음의 그림 2는 마진(γ)의 변화에 따른 일반화 성능을 차이를 나타내었다.

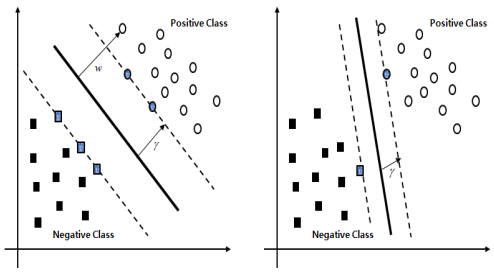

그림 2 최대 마진 분류기

선형 SVM에서 최대 마진을 찾기 위해서는 $\|w\|$가 최소가 되는 값을 결정해야 한다. 최적화 문제를 해결하기 위해서는 라그랑지 함수(Lagrange function)를 이용하여 1차 영역(primal)과 2차 영역(dual)으로 나눌 수 있다. 1차 영역에서는 1차 변수 w와 b에 대해서는 최소화 되어야 하며, 라그랑지 승수(Lagrange multipliers) $\alpha_i \geq 0$에 대해서는 최대화 되어야 한다. 이를 수식으로 표현하면 다음의 (5)와 같다.

$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha[y_i(w^T \cdot x_i + b) - 1)] \quad (5)$$

w와 b에 대한 $L(w,b,\alpha)$는 각각에 대한 미분으로 구할 수 있다.

$$\frac{\partial}{\partial b}L(w,b,\alpha)=\sum_{i=1}^{n}\alpha_i y_i=0 \quad (6)$$

$$\frac{\partial}{\partial w}L(w,b,\alpha)=w-\sum_{i=1}^{N}y_i\alpha_i x_i=0 \quad (7)$$

최대 α_i를 구하기 위해서는 라그랑지 함수 $L(w,b,\alpha)$를 2차 영역(dual)의 목적함수 $Q(\alpha)$로 표현할 수 있다. 이는 수식 (6)과 (7)을 수식 (5)에 대입하여 구할 수 있고 α_i를 결정하는 수식은 (8), 조건은 수식 (9)와 같다.

$$\max\ Q(\alpha)=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}a_i a_j y_i y_j x_i^T x_j \quad (8)$$

$$\sum_{i=1}^{N}\alpha_i y_i=0,\ \alpha_i\geq 0(i=1,...,N) \quad (9)$$

수식(8)의 목적함수는 2차 계획법(QP: quadratic programming)을 이용하여 최대 α_i값을 구한다. 데이터를 두 개의 범주로 선형분리가 불가능할 경우에는 잘못 분류된 지점을 허용하는 완화변수 ξ_i를 이용한 소프트 마진 분류기를 사용한다. 이때, 초평면 (w,b)과 마진(γ)상에서 데이터 (x_i,y_i)에 대한 완화 변수 ξ_i는 다음의 수식 (10), (11)과 같다.

$$\xi((x_i,y_i),(w,b),\gamma)=\xi_i \quad (10)$$

$$\xi_i=\max(0,\gamma-y_i(w\bullet x_i+b)) \quad (11)$$

이때, $\xi_i>\gamma$일 경우 데이터 (x_i,y_i)는 잘못 분류된 것을 의미하며, 완화 변수 ξ_i는 얼마나 많은 데이터들이 초평면에서 마진(γ)의 범위를 벗어나 있는지를 측정하는 척도가 된다[11]. 다음의 그림 3은 선형으로 분리가 불가능한 데이터를 소프트 마진 분류기를 이용하여 분류하는 방법을 나타내었다. 이때, 잘못 분류된 두 개의 데이터는 완화 변수 ξ_i와 ξ_j값이 마진(γ)보다 큰 값을 가지게 되고, 제대로 분류된 나머지 데이터들은 0값을 가지게 된다.

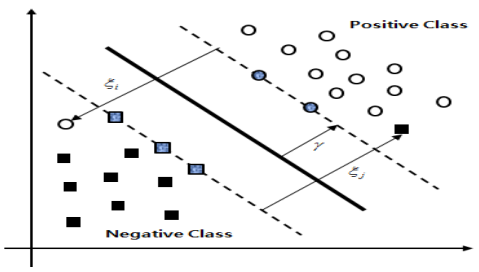

그림 3 소프트 마진 분류기

결국 완화 변수 ξ_i와 오류 패널티 변수 C값을 도입한 소프트 마진 분류기는 선형으로 분리할 수 없는 분류 문제도 어느 정도 해결할 수 있게 해준다[10]. 일반화 성능의 향상을 위해서는 마진(γ)과 학습오류를 조정해야 하고, C(오류 패널티 변수) 값을 최소화해야 한다. 이를 수식으로 표현하면 (12), 조건은 (13)과 같다.

$$\min\ \frac{1}{2}\|w\|^2+C\sum_{i=1}^{N}\xi_i \quad (12)$$

$$y_i((w\bullet x_i)+b)\geq 1-\xi_i\ (i=1,...,N),\ \xi_i\geq 0 \quad (13)$$

라그랑지 승수(Lagrange multipliers) α_i를 이용하여 최대화하는 문제로 변형하여 수식으로 표현하면 (14), 조건은 (15)와 같다.

$$\max\ Q(\alpha)=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j x_i^T x_j \quad (14)$$

$$0\leq\alpha_i\leq C(i=1,...,N),\ \sum_{i=1}^{N}\alpha_i y_i=0 \quad (15)$$

2.2 기존의 고속철도 궤도상태 판별 연구

궤도 상태를 판별에 관한 연구는 다양한 방법의 접근이 시도 되어왔다. 강기동 외(2000), 강기동(2004), 이준석(2010)은 필터를 이용하여 검측데이터의 노이즈를 제거하고 궤도틀림을 판단하는 방법을 연구했고[1][2][6], 윤석주 외(2010)는 웨이브렛 전달 함수를 이용하여서 궤도틀림을 추정하는 방법을 연구하였다[7]. 또한 강기동(2005)은 장파장을 분석하여 궤도틀림을 산정하는 방법을 연구하였으며[3], 김상수 외(2005)는 종거법, 레일형상 스캐닝법등 다양한 궤도상태 분석 방법을 연구했다[4]. 다음의 표 1은 기존의 연구에서 궤도 상태를 판별하는 방법에 대하여 도시하였다.

표 1 기존의 궤도상태 판별에 관한 연구

방법	연구자
필터를 이용하여 검측데이터의 노이즈 제거	강기동 (2004)
	이준석 (2010)
	강기동 외 (2000)
웨이브렛 전달 함수로 궤도틀림 추정	윤석주 외 (2010)
궤도틀림 파장 분석	강기동 (2005)
종거법, 레일형상 스캐닝법 등	김상수 외 (2005)

3. SVM을 이용한 궤도이상상태 분류

3.1 기존의 궤도이상구간 판별 방법

궤도검측차량 EM120은 면틀림(면좌, 면우), 줄틀림(방향좌, 방향우), 수평, 궤간의 6가지 기본 검측항목을 제공한다. 이중 수평과 궤간은 면틀림과 줄틀림에 의해 산정된 값이므로 궤도의 상태는 면틀림과 줄틀림에 가장 영향을 많이 받고 있다고 볼 수 있다[5]. 이때, 실제 검측은 25cm단위로 총 300km의 구간을 대상으로 실시하며 상행/하행 총 2번의 검측으로 검측데이터는 총 2,400,000건이 발생한다.

기존의 궤도 이상구간 선정방식은 크게 세 가지로 구분할 수 있다. 첫 번째는 검측데이터를 통한 판별 방식이다. 이는 다시 2가지로 나뉠 수 있는데 단일 검측데이터를 통한 판별과 historical 검측데이터를 통한 판별이 있다. 두 번째는 보수주기가 짧은 주의 구간을 육안으로 확인하고 판별하는 방식이다. 이는 고속철도의 궤도를 직접 가서 확인하고 이상이 있을 시에 체크리스트에 추가하여 보수하는 방법이다. 세 번째는 고속철도에 탑승하여 판별하는 방식이다. 이는 실제 고속철도 탑승하여 차량 내의 소음, 진동 등을 파악하여 이상이 있는 구간을 체크리스트에 추가하여 보수하는 방법이다. 다음의 표 2는 기존의 고속철도 궤도이상구간 판별 방식에 대하여 도시하였다.

표 2 기존의 고속철도 궤도이상 판별방식

구분	방법
검측데이터를 통한 판별	단일 검측데이터를 통한 판별 후 보수구간으로 선정 (임계치 초과 구간)
	Historical 검측데이터를 통한 판별 후 보수구간으로 선정 (속성값이 꾸준히 상승)
이상구간을 육안으로 판별	고속철도(KTX)의 이상궤도를 육안으로 판별 후 보수구간으로 선정
고속철도에 탑승하여 판별	고속철도에 탑승하여 차량 내의 소음, 진동 등을 파악 후 이상이 있는 구간을 보수구간으로 선정 (순회점검)

3.2 SVM을 이용한 궤도상태 분류 구조

전체 분류구조는 크게 학습과정과 궤도 상태 분류 과정으로 나뉜다. 학습 과정에서 전처리 과정을 통해 KTX 보수이력 데이터에서의 보수 날짜와 해당구간을 기준으로 그 이전에 검측된 데이터에서 해당구간의 검측 값들을 추출하고, 이진 벡터 프로파일을 생성한다.

궤도 상태 분류과정은 SVM 분류기의 다항식 커널과 RBF 커널을 이용하여 학습과정을 거치고, 생성된 프로파일과 검증문서 벡터간의 관계를 파악하여 해당 범주를 결정한다. 본 연구의 SVM을 이용한 궤도상태 분류 구조는 다음의 그림 4와 같다.

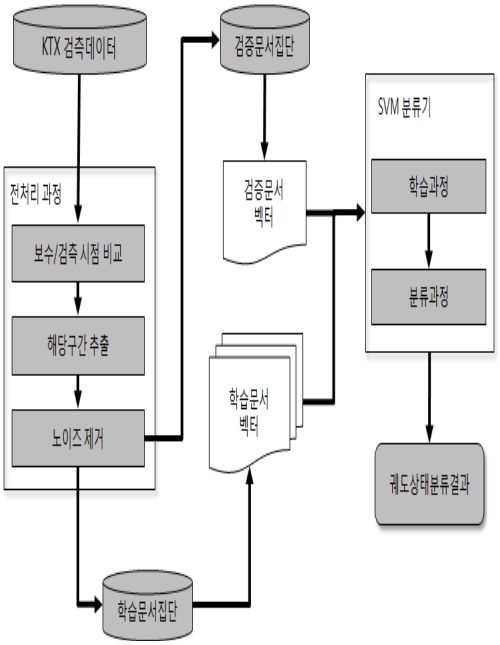

그림 4 SVM을 이용한 궤도상태 분류 구조

3.3 전 처리 및 학습과정

KTX 검측데이터는 긴 구간의 검측과 1년여의 시간으로 인해 검측단위별 데이터의 수가 천만 개 이상 존재하는데 알고리즘의 시간 복잡도를 줄임과 동시에 성능을 높이기 위해서는 검측 데이터집단 내에서 의미 있는 검측 구간의 데이터를 선정하는 것은 중요한 사항이다.

3.3.1 보수/검측 시점 비교

본 논문에서는 보수이력 전의 검측이력 데이터를 추출 시 세 가지 사항을 고려하고 있다. 첫 번째는 보수구간은 보수이전에 시행된 검측데이터를 이용하는 것이다. 두 번째 고려사항은 보수날짜와 보수 전에 시행된 검측작업의 종료날짜가 40일 이상 차이가 나는 경우의 검측데이터는 사용하지 않는 것이다. 그 이유는 보수가 2달에 3번 이루어졌을 때 약 20일에 한번 씩 검측이 시행된 것인데 이전 40일 동안 검측데이터가 존재하지 않는다는 것은 중간에 시행된 검측작업의 데이터 누락이 의심되기 때문이다. 세 번째 고려사항은 해당 월의

검측이 하루 만에 이루어지지 않고 이틀에 걸쳐 이루어졌다면, 검측작업 사이에 일어난 보수작업은 실험데이터에서 제외하는 것이다. 그 이유는 검측작업의 중간에 이루어진 보수작업이 어떠한 검측데이터를 기준으로 시행되었는지 불분명하기 때문이다.

3.3.2 해당구간 추출

해당 보수날짜에 적용될 보수이력데이터가 선정되었으면, 보수가 이뤄진 구간과 검측구간을 매칭시켜서 속성 값을 추출해야한다. 이때 SVM의 training을 위한 50,000개의 데이터(궤도이상 유 50% / 무 50%)와 training된 모델의 test를 위한 250,000개의 데이터(궤도이상 유 50% / 무 50%)를 추출해낸다.

전월의 검측수치를 바탕으로 보수구간이 결정되면 해당 구간의 각각의 속성치를 하나의 데이터 셋으로 구성한다. 각각의 데이터 셋은 보수구간의 속성 값들의 training을 위해서 25,000개와 test를 위해서 총 125,000개로 구성된다. 그리고 미 보수구간의 training과 test를 위해서 보수작업이 이루어지지 않은 구간에 대해서 각각 25,000개와 125,000개의 데이터 셋을 구성한다. 다음의 그림 5는 해당구간의 각각의 속성 값을 데이터 셋으로 형성하는 방법을 나타내었다.

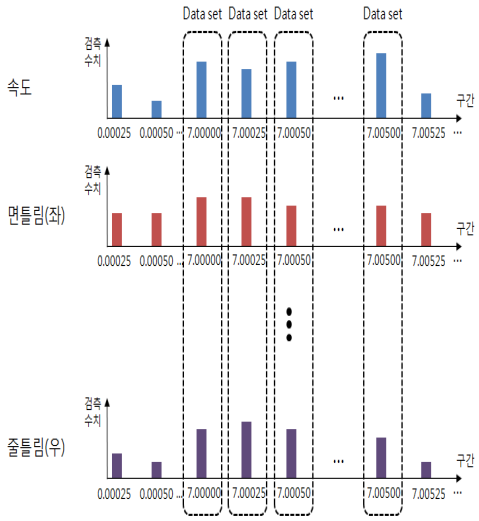

그림 5 SVM의 데이터 셋 구성

3.3.3 노이즈 제거

해당구간 추출과정을 통해서 얻어진 검측데이터는 다음의 표 3과 같은 임계치를 가지고 있다. 이는 검측차량 EM120의 검측항목인 속도, 면틀림(면좌, 면우), 줄틀림(방향좌, 방향우), 수평, 궤간을 나타내며 각각의 검측항목들은 각각의 최대/최소 임계치를 지니고 있다.

표 3 검측항목별 임계치

검측명	최대치	최소치
속도	150	0
면틀림(좌)	7	-7
면틀림(우)	7	-7
궤간	6	-6
수평	90	-90
줄틀림(좌)	25	-25
줄틀림(우)	25	-25

KTX의 2008년도 데이터는 2011년도에 종료된 KTM-SYS 최적화 프로그램이 적용되기 이전의 검측자료로써 임계치가 넘어간 속성 값이 해당 구간에 존재하더라도 관리자의 판단 실수로 인해서 보수가 이루어지지 않은 경우가 발생된다. 그러므로 추출된 해당구간의 속성 값이 임계치가 넘어갔지만 보수가 이루어지지 않은 경우에는 해당구간을 학습데이터 / 테스트데이터에서 제외한다.

4. 실험 및 결과 분석

4.1 실험데이터

궤도 상태 분류를 위한 데이터는 총 2가지 종류가 필요하다. 첫 번째는 SVM에서 학습과정을 위한 데이터로 2008년 2월 28에 검측작업이 시행되었다. 이는 보수날짜 2008년 3월 1일~20일까지를 기준으로 40일 이내에 검측된 데이터로 데이터 추출기준에 위배되지 않는다. 보수 구간은 전 검측작업의 구간과 매칭 시켰으며, 이때 학습과정 데이터는 다양한 경우의 수가 고려되어야 하기 때문에 여러 구간의 검측자료를 랜덤샘플링을 통해 추출해냈다. 또한 보수가 된 구간을 제외하고 보수되지 않은 구간 역시 랜덤샘플링으로 속성 값을 추출하여 총 50,000개의 데이터 셋을 구성하였다.

두 번째는 SVM의 분류과정을 위한 데이터로 2009년 9월 25일과 26일 총 이틀에 걸쳐서 검측작업이 시행되었다. 이는 보수 날짜는 2008년 10월 1일~20일까지를 기준으로 하여 40일 이내에 검측된 데이터로 데이터 추출기준에 위배되지 않는다. 보수 구간은 전 검측작업의 구간과 매칭 시켜서 50,000개의 데이터 셋을 총 5개의 test 문서로 만든다. 이때, 각각의 test 문서는 보수가 발생된 구간전체의 속성 값을 추출한다. 그 이유는 해당구간의 전체에 문제가 발생되지 않았어도 작업구간으로 산정되는 경우가 발생되기 때문인데, 이는 보수

차량이 보수를 실시하면서 문제가 발생될 수 있는 구간까지 함께 보수작업을 실시하기 때문이다. 보수가 이루어지지 않은 구간에 대한 속성 값은 보수날짜와 다음 검측 날짜를 기준으로 추출한다. 이때, 정확하게 보수작업이 이뤄지지 않은 구간을 추출하기 위해서 검측이 이뤄진 날짜인 9월 25일과 26일에 사이에 시행된 보수작업에도 포함되지 않고 다음 검측날짜인 10월 21일까지도 보수가 일어나지 않은 문제가 없는 구간들을 선정한다.

표 4 실제 실험 데이터 구성

	Training Data	Test Data	
검측 날짜	2008년 2월 28일	2008년 9월 25일 , 26일	
검측 구간	0km ~ 265.997km (단위25cm)	0km ~ 265.92275km (단위 25cm)	
보수 날짜	2008년 3월 1일 ~ 20일	2008년 10월 1일 ~ 20일	
보수 구간	전체 구간의 다양한 보수구간에 대한 랜덤 샘플링	실험 1:	20.51550km ~ 70.49200km
		실험 2:	70.49225km ~ 95.12300km
		실험 3:	182.81725km ~ 195.46175km
		실험 4:	200.00025km ~ 225.47725km
		실험 5:	225.47750km ~ 242.71000km
미 보수 구간	다음 검측작업 전까지의 다양한 미 보수구간에 대한 랜덤샘플링	다음 검측작업 전까지의 다양한 미 보수구간	
총 데이터	50,000개	250,000개	

4.2 성능평가 방법

KTX의 검측데이터가 특정 범주에 속하는지 그렇지 않은지 결과를 나타내주는 분류 정확도 평가는 각 범주에 대한 분류 결과를 표현하는 2 x 2 분할표를 사용하여 측정할 수 있다[6].

표 5 분류 결과 2 x 2 분할표

	보수구간 데이터	미 보수구간 데이터
보수구간으로 분류	a	b
미 보수구간으로 분류	c	d

표 5에서 a는 궤도의 보수를 실시한 구간 데이터가 제대로 보수구간으로 할당된 경우(positive examples)의 수이고 b는 미 보수 구간데이터를 보수구간으로 잘못 할당된 경우(negative examples)의 수이다. c는 보수구간 데이터가 미 보수 구간으로 할당된 경우의 수이고, d는 미 보수 구간 데이터를 미 보수 구간으로 할당된 경우의 수이다. 본 연구의 분류성능 평가 척도로는 총 3가지를 사용하며, 이들의 공식은 다음의 수식 (16)~(18)과 같다.

$$\text{보수구간 분류 정확률} = \frac{a}{a+c} \tag{16}$$

$$\text{미 보수구간 분류 정확률} = \frac{b}{b+d} \tag{17}$$

$$\text{전체 정확도} = \frac{a+d}{a+b+c+d} \tag{18}$$

4.3 파라미터 결정을 위한 사전실험

SVM은 학습과정이 시행되기 전에 사용자가 직접 파라미터를 결정해야하며 파라미터 값의 변화에 따라 SVM 분류기의 성능에 차이가 발생한다. 미리 결정해야 하는 파라미터는 학습과정에서 마진폭과 분류 오류 사이의 타협점을 찾아주는 오류 패널티 변수 C값과 비선형 SVM에 적용하는 커널함수의 파라미터이며, 본 연구에서는 SVM을 이용한 궤도 이상 판별 실험에 앞서 이에 적합한 C값과 커널함수의 파라미터 값을 결정하기 위해 사전실험을 수행하였다.

4.3.1 C값을 결정하기 위한 실험

C값은 마진폭과 분류 오류 사이의 타협점을 찾아주는 역할을 담당하며, 분류 할 수 없는 데이터에 대한 오류 페널티 값이다. 일반적으로 C값이 0으로 수렴 할수록 학습데이터의 정확한 분류보다는 마진의 최대화에 중점을 둠으로써 아주 넓은 마진폭을 갖는 간단한 모형을 형성하게 된다.

하지만 C값이 커질수록 최적의 초평면을 구축하여 학습집단의 모든 데이터를 정확하게 분류하려는 경향이 있다. 그러나 C값을 매우 크게 정의하게 되면 입력데이터에 대해서 정확하게 분류할

수 있다고 하더라도 오류가 포함된 선형 분리가 가능하지 않은 데이터에 대해 분류 성능을 보장할 수 없게 된다. 즉, 실험을 통하여 적당한 C값을 선정하는 것이 모형 복잡도를 통제하고 분류 성능을 향상시키는 효과를 가지게 된다.

본 연구에서 가장 적합한 C값을 찾기 위하여 선형 SVM을 이용하여 C값의 변화를 주며 실험하였고, 가장 좋은 성능을 보인 C값을 탐색하였다. 다음의 표 6는 파라미터 결정을 위한 사전실험을 나타낸다.

표 6 C값 결정을 위한 사전실험

	보수 구간 분류율	미 보수구간 분류율	전체 정확도
C = 1	46.748%	96.508%	71.628%
C = 10	23.976%	100.000%	61.988%
C = 100	86.932%	84.908%	85.920%
C = 150	68.164%	97.820%	82.992%
C = 1000	30.172%	93.584%	61.878%
C = 10000	11.624%	96.108%	53.866%

적합한 C값을 찾기 위한 사전실험에서 선형(Linear)커널에서 C값이 100일 때 가장 좋은 결과를 보였으며 이는 최적의 초평면을 구축하여 학습 집단의 데이터를 정확하게 분류하고 일반화 성능을 향상시켰기 때문으로 판단된다. 이때, C값과 다항식커널함수 파라미터와 RBF 커널함수 파라미터의 조합에 의해 결과값의 차이가 발생될 수 있으므로 모든 조합에 대한 실험을 통해 최적의 파라미터를 찾는다.

4.3.2 커널함수의 파라미터를 결정하기 위한 실험

SVM에서 커널함수는 선형분리가 불가능한 경우에도 자질벡터를 고차원 자질공간으로 사상시킴으로써 선형분리가 가능하게 하는 역할을 한다. 다시 말하여 비선형 분류문제를 해결하기 위해 저차원의 입력데이터 x를 보다 고차원 공간의 값 Φ(x)로 매핑 시키는 것이다[9]. 즉, 커널함수는 $(k(x_i,x_j)=\Phi(x_i),\Phi(x_j))$로 표현할 수 있다.

SVM 분류기의 성능평가를 위한 실험에서 다항식(Polynomial) 커널함수와 RBF(Radial Basis Function) 커널함수를 이용하며, 커널함수의 파라미터 값을 결정하기 위해 사전실험에서 결정된 C값인 100을 적용하여 사전실험을 실시하였다. 이때 사용되는 커널함수의 수식은 (19)~(20)과 같다.

다항식 커널함수: $k(x,y)=((x \cdot y)+1)^d$ (19)

RBF 커널함수: $k(x,y)=\exp(-\gamma\|x-y\|^2)$ (20)

다음의 표 7은 C값과 다항식 커널함수의 d(degree) 값을 변화시키면서 사전실험을 실시한 결과이다. 실험결과를 통해서 C값이 100이고, d값이 6일 경우 가장 우수한 분류 성능을 보인다는 것을 알 수 있다.

표 7 다항식 커널함수의 파라미터 결정을 위한 사전실험

		보수구간 분류율	미 보수구간 분류율	전체 정확도
C= 100	d = 2	52.560%	99.848%	76.204%
	d = 3	14.432%	99.956%	57.194%
	d = 4	32.136%	99.252%	65.694%
	d = 5	64.400%	94.124%	79.262%
	d = 6	64.128%	95.924%	80.026%
	d = 7	16.636%	47.244%	31.940%

다음의 표 8은 C값과 RBF 커널함수의 γ(gamma)값을 변화시키면서 사전실험을 실시한 결과이다. 실험결과를 통해서 C값이 10이고, γ값이 0.07일 경우 가장 우수한 분류 성능을 보인다는 것을 알 수 있다.

표 8 RBF 커널함수의 파라미터 결정을 위한 사전실험

		보수 구간 분류율	미 보수구간 분류율	전체 정확도
C= 10	γ = 0.01	77.592%	99.996%	88.794%
	γ = 0.04	78.320%	99.996%	89.158%
	γ = 0.07	79.104%	99.996%	89.550%
	γ = 0.1	78.860%	99.992%	89.426%
	γ = 0.5	9.852%	100.000%	54.926%
	γ = 1	2.808%	100.000%	51.404%

4.4 SVM 분류기 성능평가

본 연구의 실험은 2008년 2월 28일의 검측데이터 중 25,000개를 보수구간 클래스인 +1로 나머지 25,000개를 미 보수구간 클래스인 -1로 데이터를 생성하여 총 50,000개의 데이터로 학습과정을 거쳤다. 이 과정을 통해서 생성된 training 모델을 기준으로 삼아서 2008년 9월 25일, 26일에 검측된 데이터에 대하여 분류를 실시한다. test 데이터는 training 데이터와 마찬가지로 총 50,000개의 데이터로 구성하며, 2008년도 10월 1일 사이에 발생된 각기 다른 보수구간을 대상으로 총 5개의 데이터로 생성하였고, 사전 실험으로 결정된 파라미터를 바탕으로 다항식커널과 RBF커널을 이용하여 실험을 실시하였다.

4.4.1 다항식(Polynomial) 커널의 실험결과

사전실험을 통해서 얻은 파라미터 값인 d(6)과 C(100)을 이용하여 총 5개의 서로 다른 구간을 포함하는 실험데이터로 분류율을 평가하였다. 다항식 커널의 보수구간 분류율은 평균 76.5362%, 미 보수구간 분류율은 평균 96.0864%, 전체정확도는 평균 86.3128%를 나타냈다. 다음의 표 9는 다항식 커널을 이용한 궤도상태 분류의 전체 결과이다.

표 9 다항식 커널을 이용한 실험결과

	보수구간 분류율	미 보수구간 분류율	전체 정확도
실험 1	76.468%	96.184%	86.326%
실험 2	72.996%	96.068%	84.532%
실험 3	93.764%	96.176%	94.970%
실험 4	75.340%	96.080%	85.710%
실험 5	64.128%	95.924%	80.026%
평균	76.5362%	96.0864%	86.3128%

4.4.2 RBF 커널의 실험결과

사전실험을 통해서 얻은 파라미터 값인 γ(0.07)과 C(10)을 이용하여 총 5개의 서로 다른 구간을 포함하는 실험데이터로 분류율을 평가하였다. RBF 커널의 보수구간 분류율은 평균 91.7976%, 미 보수구간 분류율은 평균 99.9952%, 전체정확도는 평균 95.8964%를 나타냈다. 다음의 표 10은 RBF 커널을 이용한 궤도상태 분류의 전체 결과이다.

표 10 RBF 커널을 이용한 실험결과

	보수구간 분류율	미 보수구간 분류율	전체 정확도
실험 1	88.060%	100%	94.030%
실험 2	99.976%	99.992%	99.984%
실험 3	96.492%	99.992%	98.242%
실험 4	95.356%	99.996%	97.676%
실험 5	79.104%	99.996%	89.550%
평균	91.7976%	99.9952%	95.8964%

4.4.3 실험결과에 대한 분석

총 5회의 실험에서 보수구간 분류율이 100%가 나온 경우는 존재하지 않는다. 이는 크게 2가지의 이유로 설명할 수 있는데 첫 번째는 실제로 보수가 일어난 구간에 검측항목의 수치가 낮은 구간이 포함되어서 미 보수 구간으로 분류한 경우이고, 두 번째는 보수구간을 미 보수 구간으로 잘못 분류한 경우이다.

첫 번째 이유에 대한 자세한 설명은 다음과 같다. 보수 시행 시에 짧은 구간의 경우는 검측항목이 높은 구간을 선별하여 인력이 투입되어 작업을 진행하지만 구간이 긴 경우에는 보수장비가 투입되는데 이때, 장비는 궤도 위에서 직진하여 보수를 시행하는 특성을 지니고 있고, 보수해야 하는 2개의 구간에서 첫 번째 구간의 마지막 지점과 2번째 구간의 시작지점 사이의 거리가 짧다면 해당구간의 사이에 검측항목의 수치가 낮더라도 모두 보수를 시행하였기 때문이다. 다음의 그림 6은 보수로 선정된 구간과 실제 보수구간의 차이가 발생하는 이유를 그림으로 나타내었다.

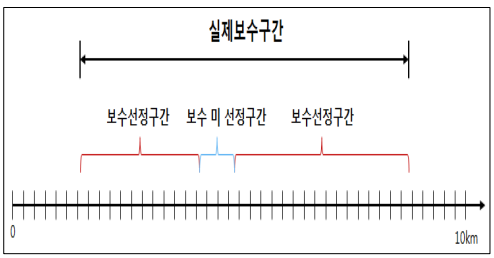

그림 6 보수 선정구간과 실제 보수구간의 미일치

두 번째 2가지의 이유로 발생될 수 있다. 첫 번째는, 검측차량인 EM120은 25cm단위의 측정을 하기위하여 검측차량의 바퀴가 돌아가는 수를 고려하는데 레일 위에서의 바퀴의 마찰력은 지면에 비해서 낮기 때문에 정확한 25cm단위로 검측을 시행하였다고 할 수 없으므로 보수가 시행된 구간과 검측지점이 다를 수가 있다. 두 번째로 실제보수작업은 EM120을 통하여 검측된 데이터 이외에 직원들이 직접 눈으로 레일의 틀림정도를 파악하여 체크리스트를 작성하고 추가적으로 실제 KTX에 탑승하여 체크된 지점의 객차 내 진동, 소음 등을 파악하여 보수구간을 결정하기 때문에 실제 보수구간의 검측자료가 낮게 나왔더라도 보수가 시행되었을 수가 있다.

5. 결론 및 추후 연구 방향

궤도 유지 보수 작업은 궤도의 틀림을 바로잡고 궤도를 안정화하는 중요한 작업이다. 이는 고속주행을 하는 열차의 특성상 궤도이탈 등의 문제를 미연에 방지할 수 있는 방법이다. 현재 고속철도의 궤도이상 판별은 검측차량을 통하여 검측된 항목들을 토대로 하여 해당 구간의 매 월 검측항목이 증가하는 추세를 보이는 구간을 선정하거나, 관리자의 경험적 지식을 토대로 선정하는 방식으로 이루어지고 있다. 이는 300km의 상/하행 총 2번의 검측으로 발생되는 2,400,000건의 데이터를 완벽하게 파악하여 보수구간을 결정하기에 무리가 있다.

따라서 본 논문에서는 궤도 상태분류를 위하여 SVM을 이용하였고, 각 구간별 검측치를 판단하여 해당 구간의 궤도 이상 유무를 판별하였다. 이때 최적화된 파라미터 값 설정을 위해서 오류 패널티 C값과 다항식 커널의 d(degree)값, RBF 커널의 γ(gamma)값을 조합해서 실험을 실시하였고, 가장 좋은 파라미터 조합을 찾아내고 SVM에서 비선형 문제를 풀기위한 2가지 커널인 Polynomial, RBF을 이용하여 판별된 데이터에 대한 분류율을 비교한 결과 RBF 커널이 본 문제에서 가장 효율적인 커널임을 보였다.

추후연구로서 SVM의 학습데이터의 수를 증가시키고, 고속철도 궤도의 속성별(터널구간, 교량구간 등)로 구분하여 해당 구간의 궤도가 이상이 발생한 경우를 각각의 학습데이터로 선정하여 궤도 속성별로 구간의 이상발생을 분류하는 방법에 대한 연구가 필요할 것이다. 또한 이상이 발생된 구간에 대해서 이상 판별 정확도가 낮은 부분에 대해 정확도를 높이는 방안에 대해 연구하며, 궤도틀림이 진전된 구간 사이에 궤도틀림의 정도가 낮은 구간이 속해 있을 경우에 해당구간을 보수 혹은 미 보수구간으로 판별하는 방법에 관한 연구가 진행되어야 할 것이다.

참 고 문 헌

[1] 강기동, 손기준, “고속철도 궤도선형 검측 자료 분석을 통한 궤도상태의 이해”, *한국철도학회*, 제4권, pp.451-454, 2000.

[2] 강기동, “고속철도 궤도검측 자료 분석기법에 관한 연구”, *한국철도학회*, 제7권, 제4호, pp.291-295, 2004.

[3] 강기동, “고속철도의 장파장 궤도틀림 분석에 대한 연구”, *한국철도학회*, 제8권, 제2호, pp.111-115, 2005.

[4] 김상수, 김영모, 한영재, 박춘수, “궤도 검측시스템의 현황과 응용”, *한국소음진동공학회, 추계학술대회 논문집*, pp.139-142, 2005.

[5] 심윤섭, “고속선 궤도품질 평가 방법론 및 틀림진전 예측에 대한 연구”, *강원대학교 산업공학과 석사학위 논문*, 2010.

[6] 이준석, 최성훈, 김상수, 박춘수, “고속철도차량의 축상 진동가속도에서 파장대역별 궤도불규칙 추정에 관한 연구”, *한국소음진동공학회, 춘계학술대회 논문집*, pp.385-386, 2010.

[7] 윤석준, 최배성, 이형진, 김만철, 최성훈, 신수봉, “웨이브렛 전달함수를 이용한 궤도틀림 추정”, *한국철도학회, 춘계학술대회 논문집*, pp.330-337, 2010.

[8] 한국철도공사, 고속선 궤도관리 의사결정지원 시스템 개발 : 2차년도 중간보고서, 2008.

[9] Burges, C., “A Tutorial on Support Vector Machines for Pattern Recognition,” *Data Mining and Knowledge Discovery*, Vol.2, No.2, pp.995-974, 1998.

[10] Cortes, Vapnik V.N., “Support Vector Networks”, *Machine Learning*, Vol.20, No.2, pp. 273-297, 1995.

[11] Cristianini, Nello, and John Shawe-Tayor, *An intoriduction to Support Vector Machines and other kernel-based learning methods*, Cambridge University Press, 2000.

[12] Mitchell, T.M., *Machine Learning*, McGraw-Hill, 1997.

[13] Vapnik, V.N., *The nature of Statistical Learning Theory*, Spinger-Verlag, 2000.