http://dx.doi.org/10.14400/JDC.2016.14.3.209

야외활동 의사결정을 위한 가중치 기반 기상정보 분석 알고리즘

이무훈, 김민규
한국외국어대학교 차세대도시농림융합기상사업단

Meteorological Information Analysis Algorithm based on Weight for Outdoor Activity Decision-Making

Moo-Hun Lee, Min-Gyu Kim
Weather Information Service Engine Institute, Hankuk University of Foreign Studies

요 약 최근 경제성장과 더불어 삶의 질이 향상됨에 따라 야외활동이 증가되었으며, 야외활동의 진행여부 의사결정은 기상여건과 밀접한 관계를 갖고 있다. 현재 이러한 야외활동 의사결정은 기상청의 일기예보와 주관적인 경험에 의해 결정되어지고 있다. 따라서, 야외활동 의사결정을 위해 기상정보를 기반으로 객관적 근거를 제시할 수 있는 분석 방법이 필요하다. 논문에서는 데이터마이닝을 기반으로 기상정보를 분석하여 야외활동 의사결정을 지원할 수 있는 기상정보 분석 알고리즘을 제안한다. 또한, 프로야구 일정 히스토리와 자동기상관측장비의 관측 자료를 데이터마이닝의 분류 알고리즘을 적용하여 실험을 수행하고, 제안한 알고리즘의 향상된 성능을 검증하였다.

주제어 : 기상정보, 데이터마이닝, 분류 알고리즘, 의사결정지원 시스템, 자동기상관측장비

Abstract Recently, the outdoor activities were increased in accordance with economic growth and improved quality of life. In addition, weather and outdoor activities are closely related. Currently, Outdoor Activities decisions are determined by the Korea Meteorological Administrator's forecasts and subjective experience. Therefore, we need the analysis method that can provide a basis for the decision on outdoor activities based on meteorological information. In this paper, we propose an algorithm that can analyze meteorological information to support decision-making outdoor activities. And the algorithm is based on the data mining. In addition, we have constructed a baseball game schedule with automatic weather system's observation data in the training data. We verified the improved performance of the proposed algorithm.

Key Words : Meteorological Information, Data Mining, Classification Algorithm, Decision Support System, AWS(Automatic Weather System)

1. 서론

최근 경제성장과 더불어 삶의 질이 향상됨에 따라 야외활동에 대한 관심이 높아졌다. 특히, 주5일제가 실시되면서 주말에 야외에서의 여가활동을 즐기는 인구가 확산되었다. 야외활동은 기상여건과 밀접한 관계를 갖고 있

* 이 연구는 기상청 차세대도시농림융합스마트기상서비스개발(WISE) 사업(KMIPA-2012-0001-1)의 지원으로 수행되었습니다.
Received 8 December 2015, Revised 15 January 2016
Accepted 20 March 2016, Published 28 March 2016
Corresponding Author: Mingyu Kim(WISE Institute)
Email: gisdev107@gmail.com

ISSN: 1738-1916

기 때문에 기상청의 일기예보를 기반으로 야외활동 계획 및 진행 여부를 의사결정(decision-making)하고 있다. 이에 맞추어 기상청은 주5일 근무제 시대에 맞는 여가활동 정보를 제공하기 위해 매주 화요일에 주말예보를 실시하고 있으며, 주말 기상조건을 한눈에 알 수 있는 야외활동 지수를 제공하고 있다. 야외활동 지수 서비스는 중부, 남부와 같은 저해상도로 주말에 대해서만 매우 나쁨, 나쁨, 보통, 좋음, 매우 좋음 형태로 제공하고 있다. 하지만, 이러한 정보는 중부, 남부와 같은 100km 이상의 저해상도로 제공되고 있기 때문에 어느 특정 장소에서만 발생되는 기상 상황을 파악하여 야외활동 의사결정하기에 힘든 실정이다. 따라서 이러한 야외활동 의사결정을 위한 고해상도 기상정보(meteorological information) 분석 방법이 필요하며, 실시간 관측 자료(observation data)를 기반으로 하는 초단기 기상정보를 분석하기 위한 알고리즘이 필요하다[1,2].

국내에서는 1980년대 후반부터 슈퍼컴퓨터를 활용한 수치예보를 도입하고, 단기 및 주간예보를 위한 전지구 예보모델, 지역예보모델, 국지모델, 응용 및 통계모델 등을 활용하고 있다. 최근에는 이러한 수치모델의 예측 성능을 고도화하기 위해 데이터마이닝(data mining) 기법을 도입하고 있는 실정이며, 응용기상과 관련하여 데이터마이닝 알고리즘을 활용하는 사례가 늘고 있다. 하지만, 기존 기상정보 분석과 관련된 데이터마이닝 알고리즘은 응용기상의 특정 도메인에 한정적으로 활용되어 도메인마다 다른 학습 모델을 실험하고 적용하는 정도이며, 현업을 위해서는 범용적으로 활용 가능한 학습 모델 구축과 정확도 향상이 필요한 실정이다.

본 논문에서는 데이터마이닝을 기반으로 고해상도의 초단기 분석을 통해 야외활동 의사결정을 지원하기 위한 기상정보 분석 알고리즘을 제안한다. 제안 알고리즘은 복수개의 학습 모델을 적용/평가하여 가중치를 부여한 학습 알고리즘으로 기존 단일 학습 모델을 적용한 기상분석 알고리즘보다 향상된 정확도로 범용적인 학습 모델을 구축할 수 있을 것이다.

실례로 프로야구 경기 일정 히스토리와 자동기상관측장비(AWS; Automatic Weather System)의 관측 자료를 기반으로 분류 알고리즘(classification algorithm)을 적용하여 실험을 수행하고, 제안한 알고리즘의 성능을 검증하였다.

본 논문의 구성은 다음과 같다. 2장에서는 기존 기상정보분야 데이터마이닝 알고리즘, 분류 알고리즘, 웨카(weka)에 대해 소개하고, 3장에서는 제안한 기상정보 분석 알고리즘을 설명한다. 4장에서는 제안 알고리즘의 성능 평가 및 분석에 대해 설명하고, 마지막으로 5장에서 결론을 기술한다.

2. 데이터마이닝 알고리즘

2.1 기상분야 데이터마이닝 알고리즘

최근에는 기상 분야의 도시기온 예측, 강수확률 예측, 건조특보 예측, 안개 예측, 관측자료 품질관리 등에 인공신경망, SVM(Support Vector Machine)과 같은 데이터마이닝 기법이 사용되고 있다. Hyeon 등[3]은 AWS 지점별 기상데이터를 기반으로 진화적 회귀분석 기법을 적용하여 단기 풍속 예보를 개선하였고, Lee 등[4]은 도시기온과 토지이용 유형의 관계성을 분석하고 신경망 및 회귀분석을 활용하여 도시기온 예측모형을 구축하는 연구를 수행하였다. Kang 등[5]은 한반도 영역을 대상으로 수치예보자료, AWS의 관측강수를 이용하여 권역별 강수발생확률을 예측할 수 있는 인공신경망 모형을 제시하였으며, Hall 등[6]은 텍사스 지역의 강수확률 및 강수정량 예측을 위해 인공신경망을 활용한 연구를 수행하였다. Luk 등[7]은 세 가지 종류의 인공신경망을 이용하여 단기 강우 예측을 연구하였으며, Liu 등[8]은 홍콩 지역의 기상 데이터를 이용하여 역전파 신경망 기반으로 단기 강수 예측 연구를 수행하였다.

기존의 기상분야에서 적용된 데이터마이닝 알고리즘은 단일 알고리즘을 적용하여 기상 예측 모델의 개선을 위해 사용되었으나 학습 알고리즘 자체에 대한 정확도 개선에 대한 연구는 진행되지 않았다. 따라서 정확도 향상을 위한 학습 알고리즘 개선은 응용기상 예측 모델의 예측 정확도 향상을 위해 필요하다.

2.2 분류 알고리즘

논문에서는 제안 알고리즘의 성능 비교를 위해 5가지의 대표적인 분류 알고리즘을 사용하였다.

첫 번째, 트리 분류기(tree classifier)는 어떤 항목에 대한 관측 값과 목표 값을 연결시켜주는 예측 모델로써

결정 트리(decision tree)를 사용하고 있으며, 의사결정 규칙과 그 결과들을 트리 구조로 도식화한 의사결정 지원 도구의 일종이다. 대표적인 결정 트리 알고리즘은 ID3, C4.5, CART, CHAID가 있다[13,14]. 웨카의 J48은 C4.5를 웨카에서 재구현한 것이다.

두 번째, 규칙 분류기(rule classifier)는 결정 트리 방식에서 파생된 알고리즘이며 ripper 알고리즘은 결정 나무와 유사한 규칙을 생성하는 알고리즘이다. JRip은 ripper를 웨카에서 구현한 것으로 규칙집합의 휴리스틱(heuristic) 전역 최적화 알고리즘을 포함하고 있다[15,16].

세 번째, 레이지 분류기(lazy classifier)는 훈련 인스턴스들을 저장하되 분류 작업 전까지 실질적인 작업을 하지 않는 알고리즘이다. IBk는 k-최근접 이웃 분류기(k-NN; k-nearest neighbor classifier)로 여러 가지 서로 다른 검색 알고리즘을 이용해 최근접 이웃을 찾는 알고리즘이다 [17,18]. 거리 측정 알고리즘은 유클리드(euclidean), 체비쇼프(chebyshev), 맨하탄(manhattan), 민코우스키(minkowski) 거리 알고리즘이 있다[19,20].

네 번째, 함수 분류기(functional classifier)는 수학식으로 표현할 수 있는 분류 알고리즘이다. SMO는 SVM을 활용한 알고리즘으로 훈련시키기 위한 순차적 최소 최적 알고리즘을 구현한 것으로 다항식이나 가우시안 커널(Gaussian Kernel)과 같은 커널 함수를 활용하고 있다[21].

다섯 번째, 베이지안 분류기(bayesian classifier)는 특성들 사이의 독립을 가정하는 베이즈 정리(Bayes' theorem)를 적용한 확률 분류기의 일종이다[22]. 베이즈 정리는 두 확률 변수의 사전 확률과 사후 확률 사이의 관계를 나타내는 정리이며, 나이브베이즈(NaiveBayes)는 확률을 기반으로 한 단순 베이지안 분류기를 웨카에서 구현한 것이다.

2.3 웨카

데이터마이닝은 대규모로 저장된 데이터 안에서 체계적이고 자동적으로 통계적 규칙이나 패턴을 찾아내는 것으로, 데이터 분석을 통해 분류(classification), 군집화(clustering), 연관성(association)을 분석하는 분야에 활용되고 있다. 분류는 일정한 집단에서 특정 정의를 통해 분류 및 구분을 추론한다. 군집화는 구체적인 특성을 공유하는 군집을 찾는 것으로, 미리 정의된 특성에 대한 정보를 가지지 않는다는 점에서 분류와 차이점을 가지고 있다. 연관성은 동시에 발생한 사건간의 관계를 정의한다[9,10].

웨카(Weka; Waikato Environment for Knowledge Analysis)는 뉴질랜드 와이카토(waikato) 대학의 이안 위튼(Ian Witten) 교수팀이 개발한 자바(Java) 기반 오픈소드(open source) 데이터마이닝 도구이다. 무료이지만 상업용 프로그램보다 다양한 분석 알고리즘을 제공하고 있어 연구자들에게 인기 있는 분석 도구이다. 1999년부터 시작되어 꾸준하게 업그레이드되고 있으며, 새롭게 연구된 알고리즘들을 추가하고 있어 최신 기법을 테스트해 볼 수 있으며, GPL(General Public License)을 따르고 있고 소스코드가 공개되어 있어 알고리즘 수정도 가능한 유용한 도구이다[11,12].

3. 기상정보 분석 알고리즘

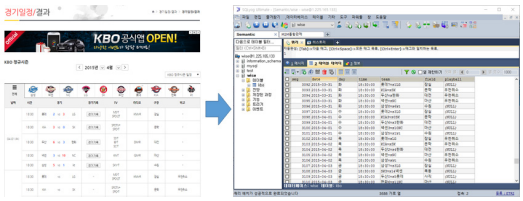
3.1 학습 데이터 구축 및 분석

〈Table 1〉 Observation data of AWS

Column Name	Comment
Station_ID	observation point number
Station_Name	observation point name
Station_Type	observation point type
Station_Latitude	observation point latitude
Station_Longitude	observation point longitude
Sky_Code	Sky status code
Sky_Name	Sky status name
Precipitation_Type	precipitation type
Precipitation_SinceOnTime	1 hour accumulated precipitation
Rain_SinceOnTime	1 hour accumulated rainfall
Rain_SinceMidNight	1 day accumulated rainfall
Rain_Last10Min	last 10 minute accumulated rainfall
Rain_Last15Min	last 15 minute accumulated rainfall
Rain_Last30Min	last 30 minute accumulated rainfall
Rain_Last1Hour	last 1 hour accumulated rainfall
Rain_Last6Hour	last 6 hour accumulated rainfall
Rain_Last12Hour	last 12 hour accumulated rainfall
Rain_Last24Hour	last 24 hour accumulated rainfall
Temperature_TC	present temperature
Temperature_Max	maximum temperature
Temperature_Min	minimum temperature
Wind_WDir	wind direction
Wind_WSpd	wind speed
Humidity	relative humidity
Pressure_surface	spot atmospheric pressure
Pressure_Sealevel	sea-level pressure
Lightining	lighting
TimeObservation	observation time

본 논문에서 사용된 학습 자료는 크게 두 가지를 활용하였다. 첫 번째는 기상관측 자료로써 SKP(SK Planet)에서 구축한 자동기상관측장비(AWS)의 관측 자료를 활용하였다. 자동기상관측장비에서 수집되는 관측 자료는 1분단위로 데이터베이스(database)에 축적되며 이를 활용하여 야외활동 의사결정을 위한 학습 데이터로 사용하였다. <Table 1>은 자동기상관측장비의 데이터베이스 카탈로그(catalog)를 정리한 것이다.

두 번째는 구체적인 실례를 기반으로 기상관측 정보 분석을 수행하기 위해 대표적인 야외 스포츠 종목인 프로야구를 대상으로 경기 일정 정보를 수집하여 우천취소 경기와 정상경기로 구분하여 그 사례에 해당하는 기상조건을 분석하였다. [Fig. 1]은 KBO(Korean Baseball Organization)의 공식 홈페이지에서 2015년도 경기 일정 정보를 수집하여 데이터베이스로 구축하는 과정을 설명하고 있다.

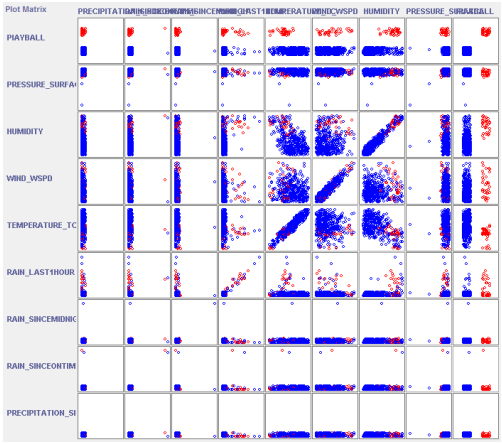


[Fig. 1] Game records of KBO League

기상 조건에 따른 프로야구 우천취소 데이터를 분석하기 위해 시계열(time series)로 경기 일정과 AWS 관측자료의 매핑을 수행하여 학습 데이터를 구성하였다. 학습 데이터 구성 기간은 2015년 프로야구 시즌이 개막된 3월 28일부터 10월 4일까지 구축하였으며, 가용할 수 있는 AWS 관측자료를 기반으로 서울 소재의 잠실과 목동 인근 12개소의 관측 자료를 사용하여 학습 데이터를 구성하였다. [Fig. 2]는 활용된 잠실과 목동 경기장 인근 관측지점 각 6개소를 설명하고 있다.

[Fig. 2] Location of AWS(Jamsil & Mok-dong Stadium)

기상 관측 자료는 앞에서 설명한 AWS 관측자료 중 일 누적 강우량, 1시간 누적 강수량/강우량, 현재기온, 일 최고/최저 기온, 10분/15분/30분/1시간/6시간/12시간/24시간 이동누적 강우량, 풍향, 풍속, 상대습도, 현지기압, 해면기압이 분석에 사용되었다. 현재 국내 프로야구 우천취소는 KBO 경기 감독관의 판단에 따라 경기 시작 2-3시간 전 운동장 상태를 확인 후 취소 결정이 된다. 따라서 경기 시작 시간의 2시간 30분 전 기상 관측 자료를 분석에 사용하였다. [Fig. 3]의 산점도(Scatter Diagram)는 대표적인 기상 변수 간의 관계를 설명하고 있다.



[Fig. 3] Scatter Diagram

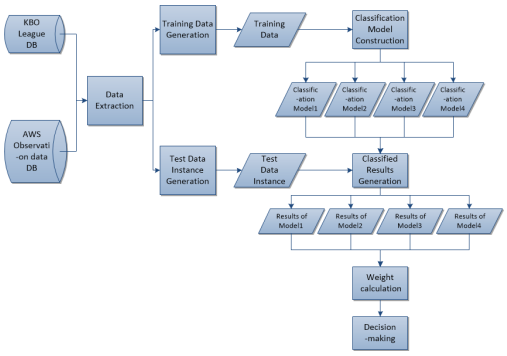
[Fig. 4]에서는 구축된 데이터베이스로부터 생성한 학습 데이터의 ARFF(Attribute-Relation File Format) 형태를 보여주고 있다. ARFF는 웨카에서 활용되는 데이터 포맷이다.

```
@relation Seoul_Baseball_Stadium

@attribute PRECIPITATION_SINCEONTIME numeric
@attribute RAIN_SINCEONTIME numeric
@attribute RAIN_SINCEMIDNIGHT numeric
@attribute RAIN_LAST10MIN numeric
@attribute RAIN_LAST15MIN numeric
@attribute RAIN_LAST30MIN numeric
@attribute RAIN_LAST1HOUR numeric
@attribute RAIN_LAST6HOUR numeric
@attribute RAIN_LAST12HOUR numeric
@attribute RAIN_LAST24HOUR numeric
@attribute TEMPERATURE_TC numeric
@attribute TEMPERATURE_TMAX numeric
@attribute TEMPERATURE_TMIN numeric
@attribute WIND_WDIR numeric
@attribute WIND_WSPD numeric
@attribute HUMIDITY numeric
@attribute PRESSURE_SURFACE numeric
@attribute PRESSURE_SEALEVEL numeric
@attribute PLAYBALL {y,n}

@data
0,0,0,0,0,0,0,0,0,0,16.2,16.8,3.5,47.2,1,21.6,1014.18,1007.5,y
0,0,0,0,0,0,0,0,0,0,15.4,17.1,2.8,131.1,4,8,20.9,1014.57,1009.6,y
0,0,0,0,0,0,0,0,0,0,15.2,16,5,330.6,1.4,25.9,1014.58,1009,y
0,0,0,0,0,0,0,0,0,0,16,16,5,222.8,1.3,21.7,1014.45,1008.8,y
0,0,0,0,0,0,0,0,0,0,16.1,16.8,3.2,238.6,2.1,19.9,1014.23,1006.1,y
0,0.1,0,0,0,0,0,0,0,0.1,14.9,16.5,7.4,162.3,1.6,58.1,1015.32,1008.6,y
0,0,0,0,0,0,0,0,0,0,15.6,16.8,6.9,84.5,1.6,55.2,1015.67,1010.7,y
0,0,0,0,0,0,0,0,0,0,14.9,15,8,17.4,2.8,58.5,1015.59,1010,y
0,0,0,0,0,0,0,0,0,0,16.6,15,8,309.3,1.5,52.1,1015.44,1009.8,y
0,0,0,0,0,0,0,0,0,0,15.5,16.3,7.2,149.5,1.7,56,1015.35,1007.2,y
0,0,0,0,0,0,0,0,0,0,14.6,17.2,7.6,330.3,1.1,78.5,1012.08,1007.1,y
```

[Fig. 4] Training Data Format(ARFF)

3.2 의사결정 알고리즘

논문에서는 앞 절에서 구축한 학습 데이터를 기반으로 대표적인 분류 알고리즘을 통해 프로야구 경기의 우천취소와 기상 조건 간의 관계를 분석하여 우천취소 결정을 할 수 있는 분류기를 제안하고자 한다.



[Fig. 5] Training procedures

[Fig. 5]는 논문에서 제안하는 알고리즘 절차를 설명하고 있다. 3.1절에서 설명한 바와 같이 프로야구 경기 일정 데이터와 AWS 관측 데이터로 학습 데이터를 구축하고 2.3절에서 설명하고 있는 분류 알고리즘을 적용하여 각 분류기로 학습을 수행한다. 각 분류기가 학습을 통해 분류 정확도를 평가하고, 식 1에 의해 전체 분류기에 대한 분류기1의 정확도 가중치를 계산할 수 있다. 각 분류기의 가중치가 산정되면 실제 분류 대상이 입력되었을 때 각 분류기가 분류를 수행하고, 식 2에 의해 미리 산정된 분류기 가중치에 의해 분류 대상의 클래스 가중치를 산정할 수 있다. 이 때 산정된 값이 0.5 이상일 경우 우천취소로 분류하게 된다.

- Classifier 1의 *Accuracy Weight*

$$AW_{C1} = \frac{a_{c1}}{a_{c1} + a_{c2} + a_{c3} + \cdots + a_{ck}} \quad \text{(식 1)}$$

- *Class Weight*
  - V_{ct} : classifier 1의 분류 값 (y=1, n=0)
  - *Class Weight* < 0.5 → n(우천취소)로 분류

$$CW = V_{c1} \times AW_{C1} + V_{c2} \times AW_{C2} + V_{c3} \times AW_{C3} + \cdots + V_{ck} \times AW_{Ck} \quad \text{(식 2)}$$

<Table 2>는 각 분류기의 분류 정확도 가중치에 따른 분류 대상의 클래스 가중치 산정의 예를 설명하고 있다. 6개의 분류기를 사용했을 경우 3개의 분류기는 정상경기, 3개의 분류기는 우천취소로 분류하였을 경우 가중치 계산을 통해 정확도 높은 분류기가 분류한 결과가 반영되는 것을 확인 할 수 있다.

〈Table 2〉 Class weight calculation

classifier	class	class permutation	accuracy	accuracy weight	class weight
J48	n	0	0.979	0.173	0.000
JRip	n	0	0.984	0.174	0.000
IBk	y	1	0.958	0.169	0.169
SMO	y	1	0.920	0.163	0.163
NavieBayes	y	1	0.915	0.162	0.162
etc	n	0	0.901	0.159	0.000
total			5.657	1.000	0.494

대표적인 분류 알고리즘의 실험 성능은 4장에서 자세히 설명하고 있다. 대체적으로 일정 수준 이상의 분류 성능을 보여주고 있지만, 향상된 분류 성능을 위해 논문에서는 여러 개의 학습 모델을 기반으로 가중치를 계산하고 그 가중치에 따라 분류 결과를 생성함으로써 분류 정확도(accuracy)를 향상 시킬 수 있었다. 또한, 새롭게 수집되는 관측 자료를 학습 데이터에 반영함으로써 새로운 기상 조건에 대응할 수 있는 학습 모델을 생성할 수 있으며, 하나의 분류기가 아닌 다중 분류기를 활용함으로써 분류의 정확도를 향상 할 수 있다.

4. 성능 평가 및 비교 분석

4.1 실험 환경

논문에서 제안하고 있는 기상정보 분석 알고리즘의 성능 분석을 위해 야구장 인근 2km 이내 지점의 AWS 관측 자료를 수집하고, 2015년 야구장 경기 일정 자료를 수집하여 학습 데이터를 구성하였다. 학습 데이터는 810개의 인스턴스(정상경기 720, 우천취소 90), 19개의 속성(attribute), 2개(정상경기, 우천취소)의 클래스(class)로 구성되어 있으며, Weka에서 활용되는 ARFF 포맷으로 구축하였다. 구성된 학습 데이터를 각 분류 알고리즘에 대해 10-폴드 교차검증(10-fold cross validation)으로 성능 평가를 수행하였다. 비교분석 대상 분류 알고리즘은 Weka에서 제공하는 5개의 분류 알고리즘(J48, JRip, IBk, SMO, NavieBayes)을 활용하였으며, 제안한 알고리즘에

도 5개의 분류 알고리즘의 분류기 가중치를 산출하여 실험을 수행하였다. 실험에 사용된 Weka는 3.7.13 버전의 라이브러리를 활용하였고, 이클립스(eclipse)에서 Java 1.8 기반으로 각 분류기의 모델을 빌드하여 실험을 수행하였다.

4.2 성능 평가 방법

분류 알고리즘의 성능 평가 방법은 TP Rate, FP Rate, Precision, Recall, F-Measure, ROC Area를 활용하여 수행하였다. [Fig. 6]과 같이 실제 정답을 정답으로 분류한 것, 오답을 정답으로 분류한 것, 정답을 오답으로 분류한 것, 오답을 오답으로 분류한 것을 나타내는 Confusion Matrix로부터 식 3, 4, 5, 6, 7과 같은 수식에 의해 분류 알고리즘의 성능을 비교분석하였다.

Confusion Matrix				
Positive	Negative	←Predicted outcome		
a	b	←Classified as		
387 (TP)	5 (FN)	a = y	True	Actual Value
4 (FP)	30 (TN)	b = n	False	

[Fig. 6] Confusion Matrix

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (식\ 3)$$

$$TP\ Rate(Recall) = \frac{TP}{TP + FN} \quad (식\ 4)$$

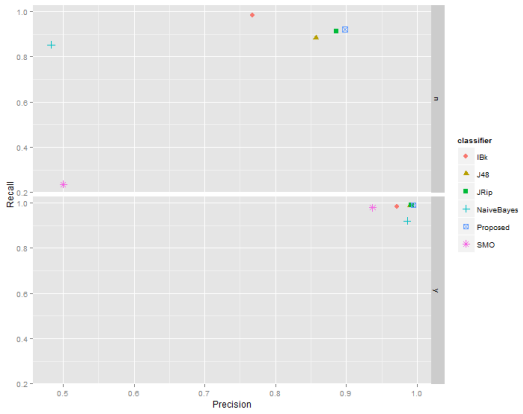
$$FP\ Rate = \frac{FP}{FP + TN} \quad (식\ 5)$$

$$Precision = \frac{TP}{TP + FP} \quad (식\ 6)$$

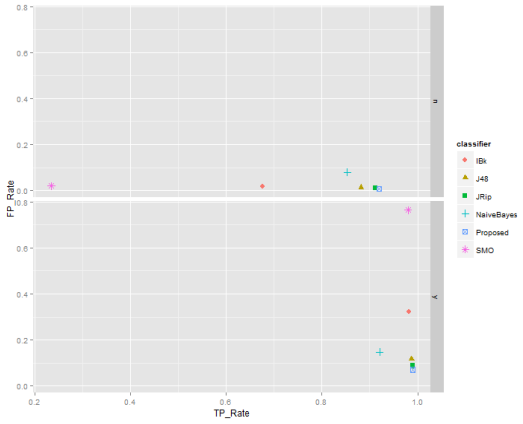
$$F - Measure = \frac{2(Precision * Recall)}{Precision + Recall} \quad (식\ 7)$$

4.3 비교 분석

[Fig. 7]과 [Fig. 8]은 각 알고리즘의 성능 지표별 성능을 설명하고 있다. [Fig. 7]의 경우 정상경기(class Y)와 우천취소(class N)에 대한 각 분류기의 정확율(precision)과 재현율(recall)에 대한 성능을 설명하고 있다. 정확율과 재현율은 값이 클수록 좋은 성능을 표현하는 것이다. 따라서 차트의 오른쪽 상단에 위치하고 있는 제안 알고리즘, J48, JRip 등이 정상경기와 우천취소에 대해 좋은 성능을 보여주고 있다. SMO와 NaiveBayes 같은 경우 우천취소 분류에 대해 정확율 및 재현율이 낮은 성능을 보여주고 있다.



[Fig. 7] Performance for Precision and Recall



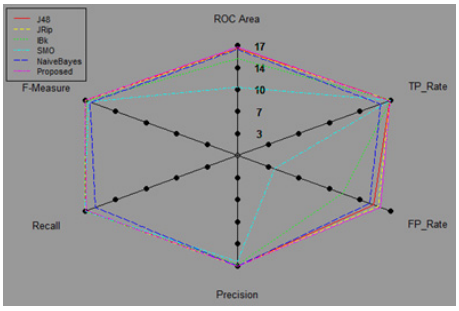
[Fig. 8] Performance for TP and FP rate

[Fig. 8]의 경우 정상경기와 우천취소에 대한 각 분류기의 TP Rate과 FP Rate에 대한 성능을 설명하고 있다. TP Rate은 식 4에서 설명하고 있는 것처럼 재현율과 같은 의미로써 높은 값을 나타낼수록 좋은 성능을 의미한다. 반면 FP Rate은 식 5에서 설명하고 있는 것처럼 오답을 정답으로 분류하고 있는 비율이므로 값이 낮을수록 좋은 것이다. 따라서 TP/FP Rate은 차트에서 오른쪽 하단에 위치할수록 좋은 성능의 분류 알고리즘이 된다. 제안 알고리즘과 J48, JRip, NaiveBayes가 좋은 성능을 나타내고 있다.

<Table 3>과 [Fig. 9]는 정상경기에 대한 각 분류기의 분류 성능을 비교분석한 것이다.

〈Table 3〉 Experimental results of Class Y

	TP_Rate	FP_Rate	Precision	Recall	F-Measure	ROC Area
J48	0.987	0.882	0.99	0.987	0.989	0.965
JRip	0.99	0.912	0.992	0.99	0.991	0.96
IBk	0.982	0.676	0.972	0.982	0.977	0.864
SMO	0.98	0.235	0.937	0.98	0.958	0.607
Naive Bayes	0.921	0.853	0.986	0.921	0.953	0.951
Proposed	**0.99**	**0.93**	**0.995**	**0.99**	**0.994**	**0.98**

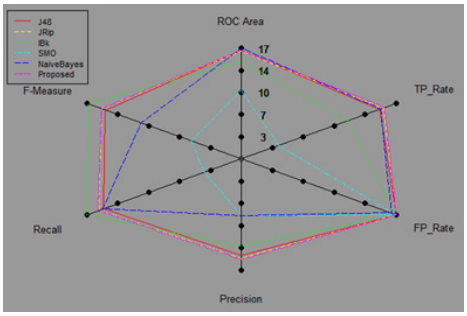


[Fig. 9] Diamond plot of Classification Performance (Class Y)

<Table 4>와 [Fig. 10]은 우천취소에 대한 각 분류기의 분류 성능을 비교분석한 것이다.

〈Table 4〉 Experimental results of Class N

	TP_Rate	FP_Rate	Precision	Recall	F-Measure	ROC Area
J48	0.882	0.987	0.857	0.882	0.87	0.965
JRip	0.912	0.99	0.886	0.912	0.899	0.96
IBk	0.676	0.982	0.767	0.982	0.977	0.864
SMO	0.235	0.98	0.5	0.235	0.32	0.607
Naive Bayes	0.853	0.921	0.483	0.853	0.617	0.95
Proposed	**0.92**	**0.992**	**0.898**	**0.92**	**0.901**	**0.972**



[Fig. 10] Diamond plot of Classification Performance (Class N)

[Fig. 9]와 [Fig. 10]의 다이아몬드 플롯(diamond plot)에서 살펴보듯이 가장 바깥쪽에 있는 제안 알고리즘이 향상된 성능을 보여주고 있다. FP rate의 경우 값이 낮을수록 좋은 성능을 의미하는 것이지만 [Fig. 9]와 [Fig. 10]에서는 역산출하여 모든 값들이 클수록 좋은 성능을 나타내도록 표현 하였다. 모든 성능 지표에서 제안 알고리즘이 확연한 차이를 보이지는 않지만 기존 알고리즘보다 조금씩 우수한 성능을 보여주고 있음을 실험을 통해 검증하였다.

5. 결론 및 향후 연구과제

본 논문에서는 분류 알고리즘을 기반으로 야외활동 의사결정을 지원하기 위한 기상정보 분석 알고리즘을 제안하였다. 기존 단일 학습 모델을 적용한 알고리즘과 비교하여 제안 알고리즘은 복수개의 학습 모델을 기반으로 가중치를 계산하고 그 가중치에 따라 분류 결과를 생성함으로써 기존 알고리즘보다 분류 정확도를 향상시킬 수 있으며, 범용적인 학습 모델로 다양한 응용기상에 활용할 수 있을 것이다. 실례로 프로야구 경기 일정 히스토리와 자동기상관측장비의 관측 자료를 기반으로 분류 알고리즘을 적용하여 실험을 수행하고, 제안 알고리즘의 성능을 검증하였다. 제안 알고리즘은 타 야외활동에 관련된 일정 데이터를 수집하여 AWS와 같은 기상관측 자료와 시계열로 매핑하고 복수개의 학습 모델을 적용함으로써 골프, 캠핑, 관광, 야회행사 등과 같이 기상여건에 영향을 받는 야외활동에 적용 가능할 것이다.

향후 연구과제에서는 과거 기상관측 자료를 더 수집하여 다양한 기상 조건에 대한 야외활동 간의 관계를 분석하고 상관성을 규명할 예정이다.

ACKNOWLEDGMENTS

This work was funded by the Weather Information Service Engine Program of the Korea Meteorological Administration under Grant KMIPA-2012-0001-1.

REFERENCES

[1] Seong-Hoon Lee, “Actual Cases and Analysis of IT Convergence for Green IT”, Journal of the Korea Convergence Society, Vol. 6, No. 6, pp. 147-152, 2015.

[2] Young-Suk Chung, Rack-Koo Park, Jin-Mook Kim, “Study on predictive modeling of incidence of traffic accidents caused by weather conditions”, Journal of the Korea Convergence Society, Vol. 5, No. 1, pp. 9-15, 2014.

[3] Byeongyong Hyeon, Yonghee Lee, Kisung Seo, “Evolutionary Nonlinear Regression Based Compensation Technique for Short-range Prediction of Wind Speed using Automatic Weather Station”, The Transactions of the Korean Institute of Electrical Engineers, Vol. 64, No. 1, pp. 107-112, 2015.

[4] Seul-Gi Lee, Sung-Gwan Jung, Woo-sung Lee, Kyung-hun Park, “A Predictive Model for Urban Temperature using the Artificial Neural Network”, Journal of the Korea Planning Association, Vol. 46, No. 1, pp. 129-142, 2011.

[5] Boosik Kang, Bongki Lee, “Prediction Probability of Precipitation Using Artificial Neural Network and Mesoscale Numerical Weather Prediction”, Journal of the Korean Society of Civil Engineers, Vol. 28, No. 5B, pp. 485-493, 2008.

[6] T. Hall, H. E. Brooks, C. A. Doswell, “Rrecipitation forcasting using a neural network”, Weather and Forecasting, Vol. 14, No. 3, pp. 338-345, 1999.

[7] K. C. Luk, J. E. Ball, A. Sharma, “An application of artificial neural networks for rainfall forecasting”, Mathematical and Computer Modelling, Vol. 33, No. 6, pp. 638-693, 2001.

[8] J. N. K. Liu and R. S. T. Lee, "Rainfall forecasting from multiple point sources using neural networks", In Proceedings of IEEE International Conference on Systems, Man, and Cybernetics, Vol. 3, pp. 429 - 434, 1999.

[9] Michael Mayo, “Random convolution ensembles”, Advances in Multimedia Information Processing – PCM 2007, 8th Pacific Rim Conference on Multimedia, Lecture Notes in Computer Science 4810, pp. 216-225. Springer, 2007.

[10] RJ Durrant, A Kaban, “Random projections as regularizers: learning a linear discriminant from fewer observations than dimensions”, Machine Learning, Vol. 99, No. 2, pp. 257-286, 2014.

[11] Sally Jo Cunningham, Eibe Frank, “Market basket analysis of library circulation data”, Proc 6th International Conference on Neural Information Processing, Vol. 2, pp. 825-830, 1999.

[12] Ian H. Witten, Eibe Frank, Len Trigg, Mark Hall, Geoffrey Holmes, and Sally Jo Cunningham, “Weka: Practical machine learning tools and techniques with Java implementations”, Proceedings of the ICONIP/ANZIIS/ANNES'99 Workshop on Emerging Knowledge Engineering and Connectionist-Based Information Systems, pp. 192-196, 1999.

[13] Ross Quinlan, “C4.5 Programs for Machine Learning”, Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[14] Jan N van Rijn, Geoffrey Holmes, Bernhard Pahringer, and Joaquin Vanschoren, “Algorithm selection on data streams”, Proc 17th International Conference on Discovery Science, pp. 325-336, Springer, 2014.

[15] William W. Cohen, “Fast Effective Rule Induction”, International Conference on Machine Learning, pp. 115-123. 1995.

[16] Indrė Žliobaitė, Albert Bifet, Bernhard Pfahringer, and Geoff Holmes, “Active learning with drifting streaming data”, IEEE Transactions on Neural Networks and Learning Systems, Vol. 25, No. 1, pp. 27-39, 2014.

[17] N. Bhatia, “Survey of Nearest Neighbor Techniques”, International Journal of Computer Science and Information Security, Vol. 8, No. 2, 2010.

[18] D. Aha, D. Kibler, “Instance-based learning algorithms”. Machine Learning. pp. 37-66, 1991.

[19] Deza, M.; Deza, E. “Encyclopedia of Distances”, Springer-Verlag, pp.94, 2009.

[20] David M. J. Tax, Robert Duin, and Dick De Ridder, "Classification, Parameter Estimation and State Estimation: An Engineering Approach Using MATLAB", John Wiley and Sons. pp. 440, 2004.
[21] S.S. Keerthi, S.K. Shevade, C. Bhattacharyya, K.R. K. Murthy, "Improvements to Platt's SMO Algorithm for SVM Classifier Design", Neural Computation. Vol. 13, No. 3, pp. 637-649. 2001.
[22] George H. John, Pat Langley, "Estimating Continuous Distributions in Bayesian Classifiers", Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo, pp. 338-345, 1995.

이 무 훈(Lee, Moohun)

・2002년 8월 : 한남대학교 컴퓨터공학과(공학사)
・2004년 8월 : 한남대학교 컴퓨터공학과(공학석사)
・2013년 2월 : 한남대학교 컴퓨터공학과(공학박사)
・2008년 10월 ~ 2015년 2월 : 한국전자통신연구원 선임연구원
・2015년 2월 ~ 현재 : 한국외국어대학교 차세대도시농림융합기상사업단 선임연구원
・관심분야 : 데이터마이닝, 정보검색
・E-Mail : macbethe@gmail.com

김 민 규(Kim, Mingyu)

・2012년 2월 : 남서울대학교 GIS공학과(공학사)
・2014년 8월 : 인하대학교 지리정보공학과(공학석사)
・2015년 2월 ~ 현재 : 한국외국어대학교 차세대도시농림융합기상사업단 연구원
・관심분야 : 공간빅데이터, 공간데이터베이스, 데이터마이닝
・E-Mail : gisdev107@gmail.com