Journal of Environmental Science International
25(4); 603~613; April 2016

pISSN: 1225-4517 eISSN: 2287-3503
http://dx.doi.org/10.5322/JESI.2016.25.4.603

ORIGINAL ARTICLE

데이터마이닝 기법들을 통한 제주 안개 예측 방안 연구

이영미* · 배주현 · 박다빈
㈜에코브레인

A Study on Fog Forecasting Method through Data Mining Techniques in Jeju

Young-Mi Lee*, Joo-Hyun Bae, Da-Bin Park
Eco Brain Co., Ltd., Jeju 63309, Korea

Abstract

Fog may have a significant impact on road conditions. In an attempt to improve fog predictability in Jeju, we conducted machine learning with various data mining techniques such as tree models, conditional inference tree, random forest, multinomial logistic regression, neural network and support vector machine. To validate machine learning models, the results from the simulation was compared with the fog data observed over Jeju(184 ASOS site) and Gosan(185 ASOS site). Predictive rates proposed by six data mining methods are all above 92% at two regions. Additionally, we validated the performance of machine learning models with WRF (weather research and forecasting) model meteorological outputs. We found that it is still not good enough for operational fog forecast. According to the model assesment by metrics from confusion matrix, it can be seen that the fog prediction using neural network is the most effective method.

Key words : Fog prediction, Data mining, R, Tree models, Conditional inference tree, Random forest, Multinomial logistic regression, Neural network, Support vector machine, Confusion matrix

1. 서 론

안개는 지면이나 해면에 인접한 층에서 수증기가 응결하여 대기 중에 부유하는 현상으로 시정이 1 km 미만인 경우를 말하며, 냉각에 의해 형성되는 안개로는 복사안개, 이류안개, 활승안개 등이 있고, 수증기 첨가에 의해 형성되는 안개로는 증기안개, 전선안개 등이 있다. 내륙에서는 복사안개가 주로 발생하는 반면 연안과 해상에서는 이류안개인 해무가 주로 발생하며, 주변의 종관적 특성뿐만 아니라 해양학적 특성의 영향을 받아서 그 발생빈도가 계절적·지역적으로 큰 차이가 나타난다(Leiper, 1994).

안개는 사회 · 경제 · 인간생활 전반에 걸쳐 영향을 미치는 중요한 기상현상으로 특히 고속도로나 공항 주변에서 발생하는 안개는 차량 및 항공기 운행 등 운수업에 막대한 영향을 준다. 도로교통공단 교통사고종합분석센터(2014)의 기상상태별 치사율 분석결과를 살펴보면, 안개 낀 날의 경우 100건당 10.6명이 사망한 것으로 나타나,

Received 21 March, 2016; **Revised** 29 March, 2016;
Accepted 4 April, 2016
***Corresponding author** : Young-Mi Lee, Eco Brain Co., Ltd., Jeju 63309, Korea
Phone : +82-70-7018-3082
E-mail : leeym@ecobrain.net

맑은 날(2.2명/100건), 흐린 날(3.7명/100건), 빗길(2.9명/100건) 등에 비해 2～4배 높은 것으로 분석되었다.

Sohn(2010)은 고산, 서귀포, 제주 등의 관측소를 포함하고 있는 제주 지역에서는 여름, 봄 순으로 안개가 많이 발생하였고, 가을과 겨울에는 안개가 거의 발생하지 않았음을 밝혔다. 제주 지역은 운량, 상대습도, 강수량의 값이 클수록 그 해의 안개 발생빈도가 증가하였고, 일교차와 해면기압은 반대로 그 값이 작을 때 안개 발생이 많았다. 이는 제주도에서 발생한 안개 사례의 70%는 서쪽의 고산과 남쪽의 서귀포에서 발생한 것이며, 전선 및 저기압 통과에 연관된 것이 60%라고 한 Heo and Ha (2004)의 연구와 일치하는 내용이다.

제주도 포함 한반도 주변에 나타나는 안개의 경우 이류안개가 가장 일반적이며, 이류안개는 주로 따뜻하고 습한 공기가 충분히 차가운 지표 위를 지날 때, 습윤한 공기가 포화점까지 냉각되면서 응결되어 형성되는 안개이다. 이에 우선적으로 제주도에서 일사량 관측이 이루어지고 있는 제주(184지점)와 고산(185지점) 지역의 automatic synoptic observation system (ASOS) 기상관측자료들을 분석하여 제주도 안개 특성과 발생 메카니즘을 밝히고자 하며, 이 결과들을 기계 학습시켜서 안개 발생 여부를 예측해 내는 데이터 마이닝 기법들의 우수한 안개 예측모형을 개발하고자 한다.

데이터마이닝 기법에 의해 안개 예측을 한 연구로는 대표적으로 Kim(2009)에 의한 자기구성신경망을 이용한 안개 예측이 있으며, 그 외 회귀분석과 의사결정나무 등에 의해 이루어진 바가 있지만 다양한 기법들의 비교를 통한 검증 연구가 미비한 상황이다. 이에 본 연구에서는 다양한 데이터마이닝 기법을 이용하여 우리나라 제주 지역을 중심으로 위험 기상 요소인 안개 발생에 대한 예측을 개선함으로써 안전한 도로 주행과 수상레저활동, 선박 항해 등이 이루어질 수 있도록 기여하고자 한다.

2. 재료 및 방법

2.1. 데이터 마이닝 이론

데이터 마이닝은 많은 데이터 가운데 숨겨져 있는 유용한 상관관계를 발견하여 미래에 실행 가능한 정보를 추출해 내고 의사 결정에 이용하는 과정을 말한다. 즉 데이터에 고급 통계 분석과 모델링 기법을 적용하여 유용한 패턴과 관계를 찾아내는 과정이다(David, 1998).

데이터마이닝기법을 위한 소프트웨어로는 Intelligent Miner, 마이크로소프트사의 SQL (structured query language) Server, 오라클의 Data Mining, 테라데이터의 Warehouse Miner, SAS (statistical analysis system)의 Enterprise Miner, SPSS (statistical package for social science)가 있다. 최근 주목받고 있는 R과 WEKA (waikato environment for knowledge analysis)는 오픈소스 형태로 사용할 수 있다(Jiawei et al., 2011).

본 연구에서는 다양한 기계적 학습을 위한 새로운 패키지들의 추가 지원이 가능하고, 차후 다른 기상요소들 간의 예측 시스템 연계를 위해 CUI (character user interface) 방식의 R을 사용한 데이터마이닝 기법들에 의해 안개 발생 여부의 예측 정확도를 높이고자 하였다.

안개 발생 메카니즘을 학습시키기 위해 사용한 데이터 마이닝 기법들로는 의사결정나무(tree models)와 조건부 추론 나무(conditional inference tree), 앙상블 학습기법인 랜덤포레스트(random forest), 다항 로지스틱 회귀 분석(multinomial logistic regression), 신경망(neural network), 서포트 벡터 머신(support vector machine, SVM) 등이다.

R은 기계 학습, 통계, 금융, 생물정보학, 그래픽스에 이르는 다양한 통계 패키지를 갖추고 있으며, 본 연구에서 사용되는 패키지에는 rpart, ctree, randomForest, nnet, kernlab 등이 있다(Table 1).

Table 1. Statistical packages used in this study

Model	R Package	Model	R Package
Tree Models	rpart	Neural Network	nnet
Conditional Inference Tree	party	Multinomial Logistic Regression	nnet
Random Forest	randomForest	Support Vector Machine	kernlab

먼저 의사결정나무는 의사결정규칙을 나무 구조로 도표화하여 분류와 예측을 수행하는 분석 방법이다. 이 방법은 분류 또는 예측 과정이 나무구조에 의한 추론 규칙에 의해서 표현되기 때문에 분석자가 그 과정을 쉽게 이해하고 설명할 수 있다는 장점이 있다(Ko, 2011).

반면, 조건부 추론 나무는 변수와 반응값(분류) 사이의 연관관계를 측정하여 노드 분할에 사용할 변수를 선택하는데, 의사 결정 나무에서 노드를 반복하여 분할하면서 발생하는 문제인 다중 가설 검정을 고려한 방법이다.

Breiman(2001)에 의해 개발된 랜덤포레스트는 앙상블(ensemble) 학습 기법을 사용한 모델로 입력변수로부터 여러 개의 모델을 학습한 다음, 예측 시 여러 모델의 예측 결과들을 종합 사용하여 정확도를 높이는 기법을 말한다. 랜덤포레스트에서의 변수 중요도는 변수가 정확도와 노드 불순도 개선에 얼마만큼 기여하는 지로 측정되기 때문에 변수 선택 방법 중 필터 방법이라고도 한다(Breiman, 2001; Ramón and Sara, 2006).

로지스틱 회귀분석은 단순회귀분석과 다중회귀분석이 선형으로 가정하는데 비해 로지스틱 회귀분석은 S자형으로 가정하며, 일반적으로 종속변수의 범주가 두 개인 경우에 적용된다. 또한 분류가 두 개가 아닌 여러 개가 될 수 있는 경우, 다항 로지스틱 회귀분석을 사용한다(Annette and Adrian, 2008).

신경망은 경험을 통해 자기 자신의 규칙을 만들 수 있는 구조와 능력을 가진 인간의 두뇌를 모형화하여 문제해결 능력을 갖게 한 수학적 모델의 하나이다. 이 모델은 입력층, 은닉층, 출력층으로 구성되어 있고 신경망은 학습(training)을 통하여 최적화된다(Bishop, 1999). 실제 본 연구에서 안개 예측을 위해 사용되었던 3개의 은닉층을 가지는 신경망 모형 블록도의 한 예를 Fig. 1에 나타내었다.

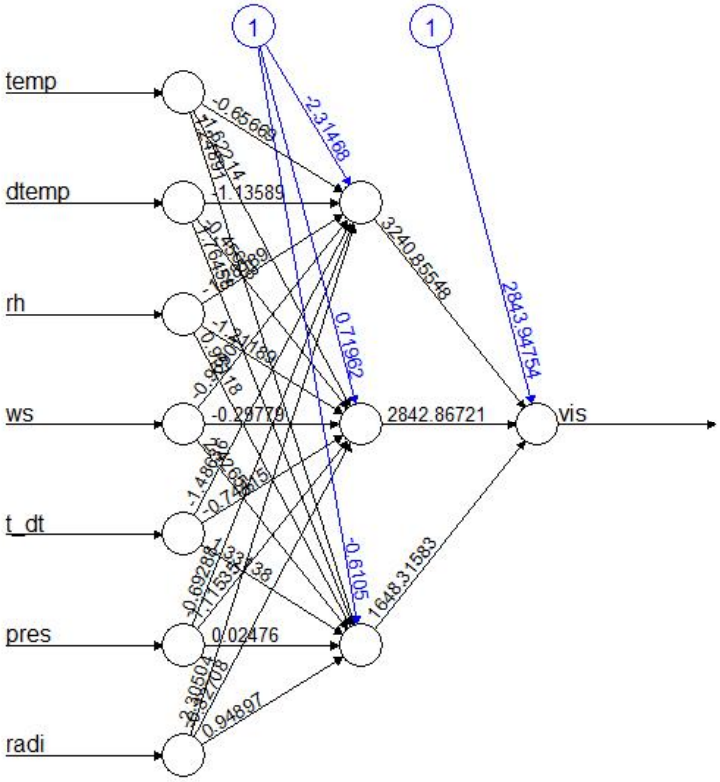


Fig. 1. Diagram sample of neural network.

마지막으로 서포트 벡터 머신은 입력되는 데이터를 두 집단으로 분리하고 분석하는 학습 알고리즘으로, 데이터 분리를 위해 데이터의 반대 집단에서 가장 멀리 떨어진 서포트 벡터(support vector)를 찾아서 두 집단으로 나누는 기준인 초평면을 정하고, 여백(margin)을 계산한다. 이론과는 다르게 동적 시스템에서 측정된 데이터를 입력으로 사용하는 경우에, 실제 입력 데이터들은 선형 분리가 불가능한 상황인 경우가 많다. 이러한 경우, 서포트 벡터 머신 학습을 위해서는 고차원의 특정 공간에서 데이터를 분리하는 커널 함수를 사용하여 비선형 문제를 해결해야 한다(Cherkassky and Ma, 2004; Shin et al., 2009). 즉, 고차원 벡터의 개념을 가진 커널 함수를 적용할 필요가 있으며, 특정 공간상의 벡터(x_i, x_j)를 매개변수로 갖는 커널 함수를 $K(x_i, x_j)$라 할 때, 식 (1)과 같이 표현할 수 있다.

$$K(x_i, x_j) = \Phi(x_i)\Phi(x_j) \quad (1)$$

커널 함수에 따라 동일한 데이터를 분석한 결과가 서로 상이하므로 데이터의 종류와 분포에 따라 적절한 커널 함수를 선택하여 사용하거나 필요에 따라 변형하여 사용한다(Cristianini and Taylor, 2000). 본 연구에서는 비선형 데이터의 패턴 구분에 가장 널리 사용 되고 있는 가우시안 RBF (radial basis function) 커널 함수를 사용하여 비선형 문제를 해결하고 기존 데이터를 분석 하고자 하며(Cherkassky and Ma, 2004), 이 연구에서 사용할 RBF 커널은 식 (2)와 같이 표현되어진다.

$$K(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2}) \quad (2)$$

2.2. 분석 및 모델 평가 방법

학습과 모델 구축, 또한 예측된 결과의 검증을 위해 제주도 연안지역에서의 안개를 예측하기에 적합한 제주와 고산 ASOS 지점의 3시간 간격 기상관측자료를 이용하였다(Table 2). 2005년 1월 1일 ~ 2014년 12월 31일 자료는 학습과 통계 예측 모형 도출을 위해 사용되었고, 2015년 1월 1일 ~ 2015년 12월 31일까지 자료로 각 모델을 통해 예측된 안개 결과를 검증하였으며, 학습 시 구성되는 입력자료에 결측값이 있는 경우는 그 전체 레코드를 제거함으로써 우수한 학습을 위한 필터링된 데이터 셋을 구성하였다. 각 데이터 마이닝 기법으로 예측된 모델을 통해 검증 기간 동안의 안개 발생 유무를 예측하였고 실제 안개 발생 유무와의 검증이 이루어졌다.

본 연구에서는 다양한 데이터 마이닝 기법을 통해 안개 발생을 예측하고 검증하고자 R을 통한 6가지의 통계 패키지를 구성하였다. 제주와 고산지역의 기상관측자료 중 시정값을 도출하여 1 km 미만일 때를 안개가 발생한 경우(fog), 1 km 이상일 때를 안개가 발생하지 않은 경우(non_fog)로 설계하여 분석하였다. 2005년부터 2014년의 기상관측자료를 통해 안개 발생 유무 즉, fog와 non_fog를 예측하는 통계 예측 모형을 도출하고, 2015년의 시정값과 기상관측자료를 통해 검증하였다. 기상관측자료는 기온(이하 temp), 노점온도(이하 dtemp), 상대습도(이하 rh), 풍속(이하 ws), 현지기압(이하 pres), 일사량(이하 radi) 그리고 기온과 노점온도의 차(기온노점차, 이하 t-dt)의 총 7개 요소이다.

2015년 관측값을 기계학습모델들 결과와 검증해볼 뿐만 아니라, 실제 안개 예측을 위해서 각 통계모델에 사용되어지는 기상입력자료 생성을 위하여 중규모 기상모델인 WRF ver 3.5를 사용하여 기상장을 생성시키고 그 자료들을 입력값으로 사용하여 실제 72시간 이후까지 안개를 예측할 수 있도록 하였다. 초기 및 경계조건은 NCEP/NCAR (national centers for environmental prediction/national centers for atmospheric research)에서 제공하는 FNL (final analyses)자료를 이용하였다.

Table 2. Information of observation station

Observation station		Station information	
Name	Number	Latitude	Longitude
Jeju	184	33.51411 N	126.5297 E
Gosan	185	33.29382 N	126.1628 E

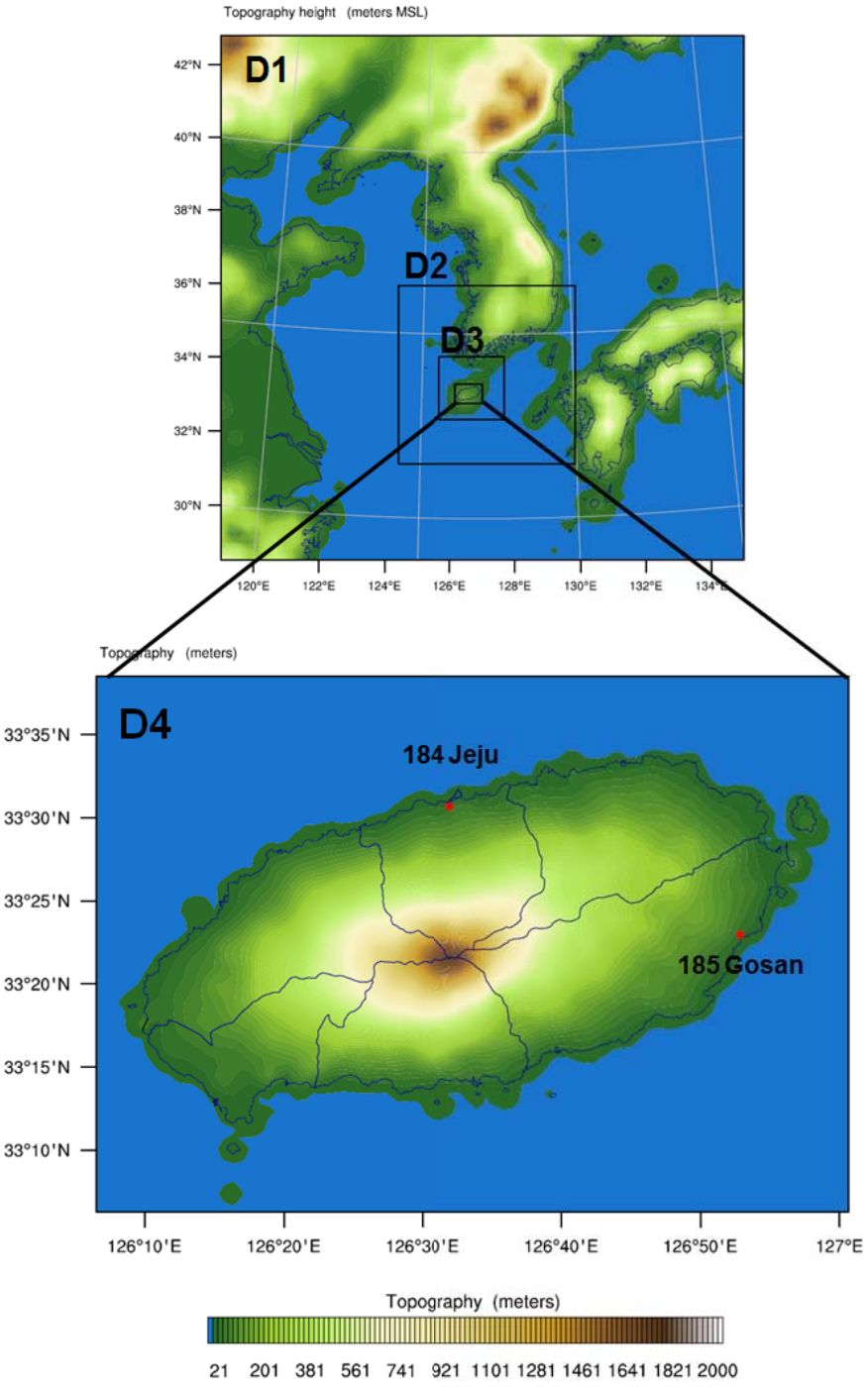


Fig. 2. The nested model domains. The bottom figure is enlarged details of the target areas (D4) and the red points indicate the automatic synoptic observing system (ASOS) observed by Korea Meteorology Administration (KMA).

Table 3. Confusion matrix

	Observed Value(Y)	Observed Value(N)
Predicted Value(Y)	True Positive(TP)	False Positive(FP)
Predicted Value(N)	False Negative(FN)	True Negative(TN)

Table 4. Typical metrics that can be calculated from confusion matrix

Metric	Calculation
Precision	$\frac{TP}{TP+FP}$
Accuracy	$\frac{TP+TN}{TP+FP+FN+TN}$
Recall	$\frac{TP}{TP+FN}$
Specificity	$\frac{TN}{FP+TN}$
FP Rate	$\frac{FP}{FP+TN}$
F1 Score	$2\times\frac{precision \cdot recall}{precision+reall}$
Kappa	$K=\frac{accuracy-P(e)}{1-P(e)}$ $P(e)=\frac{(TP+FP)*(TP+FN)*(FN+TN)*(FP+TN)}{(TP+FP+FN+TN)^2}$

FNL자료는 1°×1°의 공간 해상도를 가지며 6시간 간격의 시간 해상도를 가진다. 모델링 영역은 Two-way nesting기법을 이용하여 동아시아영역(27 km, D1), 영·호남권영역(09 km, D2), 제주 및 한반도 남해안영역(03 km, D3), 그리고 제주도영역(01 km, D4)으로 총 4개의 도메인으로 구성하였다(Fig. 2). 대상지역의 지형과 지표면 상태를 현실적으로 반영하기 위하여 고해상도 지형고도(shuttle radar topography mission 3sec, SRTM 3sec)와 토지피복자료(environmental geographic infor-mation system, EGIS)를 사용하였다. 모델링 수행 시 사용된 물리옵션 중 구름미세물리 모수화 과정은 WSM (WRF single momentum) 6-Class graupel scheme을 적용하였고, 복사과정에서 장파복사기법과 단파복사기법은 RRTM (rapid radiative transfer model)과 Dudhia scheme을 사용하였다. 행성 경계층(planetary boundary layer, PBL)은 YSU (yonsei university) scheme을 사용하였으며, 적운 모수화과정은 Kain-Fritsch scheme을 동아시아영역(D1)에만 적용하였다.

본 연구에서 개발된 다양한 모델들 중 어떤 모델이 좋은지의 평가 방법으로는 평가 메트릭을 활용하였다. 분류가 Y, N 두 종류가 있다고 할 때 분류 모델에서의 모델 평가 메트릭은 모델에서 구한 분류의 예측값과 데이터의 실제 분류인 실제 값의 발생 빈도를 Table 3과 같은 혼동행렬(confusion matrix)로부터 계산한다. 혼동 행렬에서 true positive (TP)에 해당하는 셀은 실제 값이 Y이고, 예측도 Y였던 경우의 수이며, false positive는 실제 값은 N이었는데 예측이 Y로 된 경우의 수를 기록한다. 같은 방식으로 false negative (FN), true negative (TN)도 기록할 수가 있다.

Table 4는 혼동 행렬로부터 계산할 수 있는 대표적인 메트릭이다. 본 연구에서는 각 메트릭을 계산해서 모델의 우수한 정도를 평가하고자 한다. Precision은 Y로 예

측된 것 중 실제로도 Y인 경우의 비율이고, Accuracy는 전체 예측에서 옳은 예측의 비율이다. Recall은 Hit Rate라고도 불리며, 실제로 Y인 것들 중 예측이 Y로 된 경우의 비율이다. Specificity는 실제로 N인 것들 중 예측이 N으로 된 경우의 비율이며 FP Rate는 Y가 아닌데 Y로 예측된 비율로 이 값의 경우는 0에 가까울수록 모델이 우수하다고 볼 수 있다. F1점수는 Precision과 Recall의 조화평균으로 한쪽만 클 때보다 두 값이 골고루 클 때 큰 값을 가진다. 마지막으로 코헨의 Kappa는 두 평가자의 평가가 얼마나 일치하는지 평가하는 값으로 두 평가자의 평가가 우연히 일치할 확률을 제외한 뒤의 점수이다 (Seo, 2014; Yaser et al., 2012).

3. 결과 및 고찰

3.1. 분석 결과

먼저 2015년의 연별 및 월별 안개 발생 횟수를 확인하였다. 2015년 안개 발생 횟수는 제주와 고산지역이 각각 28회, 89회로 고산지역이 안개 발생 횟수가 더 많았으며, 월별 안개 발생 횟수의 경우 제주와 고산지역 모두 4월에 각각 12회, 24회로 가장 많았고, 반면에 안개가 한 번도 발생하지 않은 달도 존재하였다.

R을 통한 6가지 통계 패키지를 활용하여 2015년 제주와 고산지역의 안개 발생 유무를 예측하고 2015년에 대해 검증하여 예측률을 비교하였다. 입력변수는 temp, dtemp, rh, ws, pres, radi 그리고 t-dt로 7개 모든 변수를 사용하였으며, 6개 데이터 마이닝 기법 모두 동일하다 (Table 5).

제주와 고산지역 모두 데이터 마이닝 기법 중 실제 안개 발생을 가장 정확히 예측한 다항 로지스틱 회귀분석의 제주지역 안개 발생 유무 예측률은 98.87%, 고산지역은 93.28%로 다른 분석 기법과 유사하게 높은 예측률을 나타내었다. 의사결정나무와 신경망 모델의 경우 고산지역의 안개 발생 유무만을 예측할 뿐, 제주지역의 안개 발생을 제대로 예측하지 못하였다. 이는 제주와 고산의 안개 발생 사례가 적어서 안개 발생 메카니즘 학습이 제대로 이루어지지 못했음을 의미하며, 안개 사례가 많은 내륙지역의 관측자료들을 통해 복사안개, 활승안개 등의 추가 발생 메카니즘 학습을 동반한 통합적 모델 생성이 시급함을 시사한다. 서포트 벡터 머신 또한 전체 예측률 자체는 높았으나, 두 지역 모두 안개 발생을 정확히 예측하지는 못하였다. 전체 예측률은 높으나 안개 발생을 제대로 예측하지 못한 결과를 보임에 대한 전반적인 모델 평가는 3.2절에 자세히 설명하고자 한다.

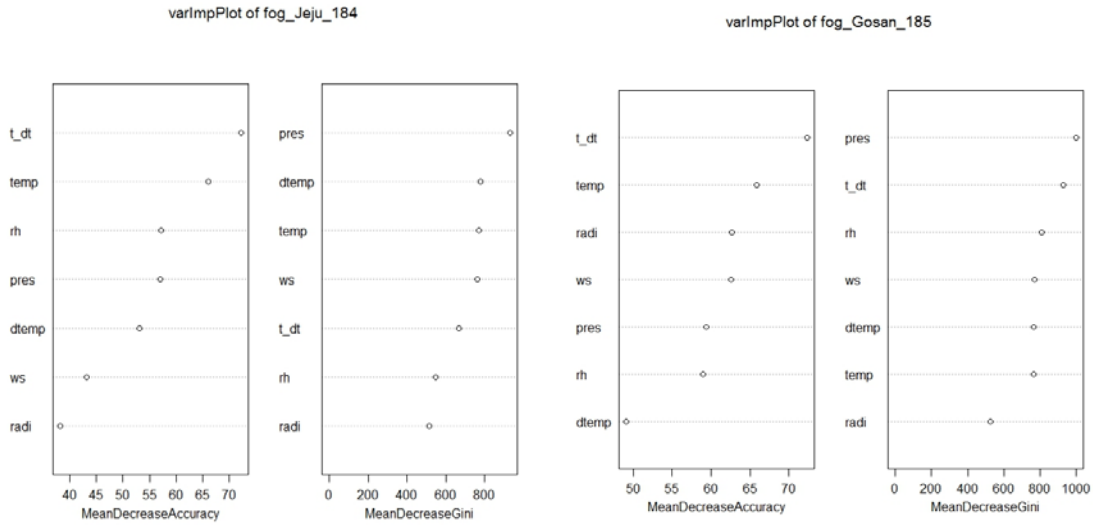
제주와 고산지역에서 안개 발생 유무를 예측한 데이터 마이닝 기법을 분석하고 입력변수들 간의 안개 예측 효과를 파악하였다. 다항 로지스틱 회귀분석은 안개 발생 유무를 예측하는 상관계수 중 pres에 의한 영향이 두 지역에서 모두 가장 높았으며, radi에 의한 영향이 가장 낮았다.

조건부 추론 나무의 경우 제주와 고산지역 각각 36개와 59개의 Terminal node를 가졌으며, rh가 첫 번째 기준으로 분류되었고 다음으로는 pres가 두 번째 기준으로 분류되어 두 입력변수가 가장 중요한 변수로 고려되었음을 확인하였다.

랜덤포레스트의 경우는 변수 정확도와 노드 불순도를 통해 입력변수 중요도를 측정한다. 변수 정확도를 의미

Table 5. Predictive rate in each model

Model	Input Variable	Prediction Rate (%)		
		Jeju(184)	Gosan(185)	Gosan(WRF)
Tree Models	temp, dtemp, rh, ws, pres, radi, t-dt	99.04	92.93	83.33
Conditional Inference Tree		98.18	93.24	83.33
Random Forest		97.67	92.11	81.25
Neural Network		99.04	94.04	80.00
Multinomial Logistic Regression		98.87	93.28	81.67
Support Vector Machine		99.04	96.93	83.33



(a) (b)

Fig. 3. MeanDecreaseAccuracy and MeanDecreaseGini of RandomForest Models at (a) Jeju(184) (b) Gosan(185).

하는 MeanDecreaseAccuracy는 두 지역 모두 t-dt가 가장 높은 값을 보여 중요도가 큰 변수로 인식되었고, 제주의 경우는 radi, 고산의 경우는 dtemp가 낮은 값을 보였다. 노드 불순도 개선을 의미하는 MeanDecreaseGini의 경우 pres가 두 지역 모두에서 가장 중요하였으며, radi의 불순도가 가장 낮은 값으로 분석되었다(Fig. 3).

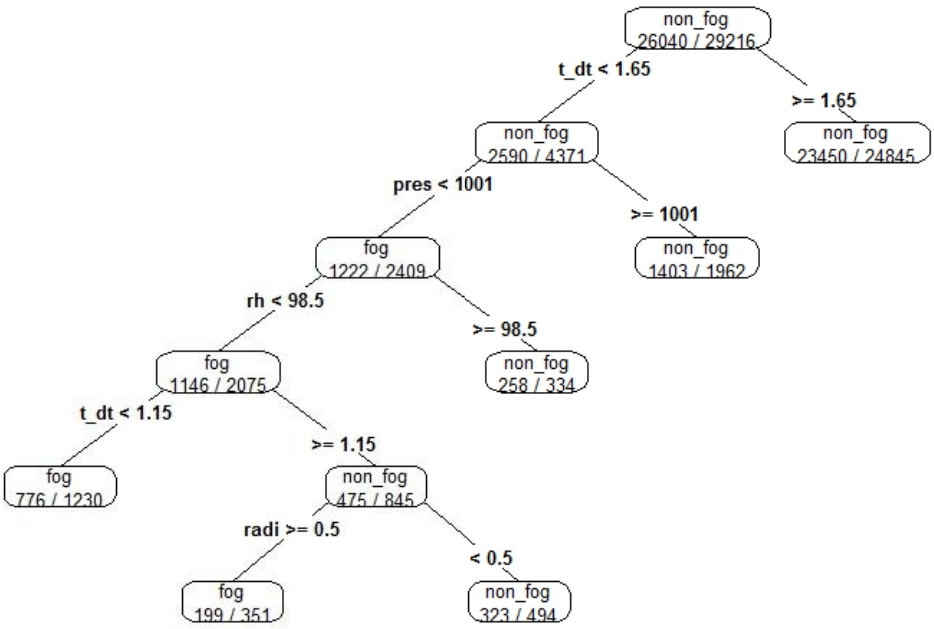
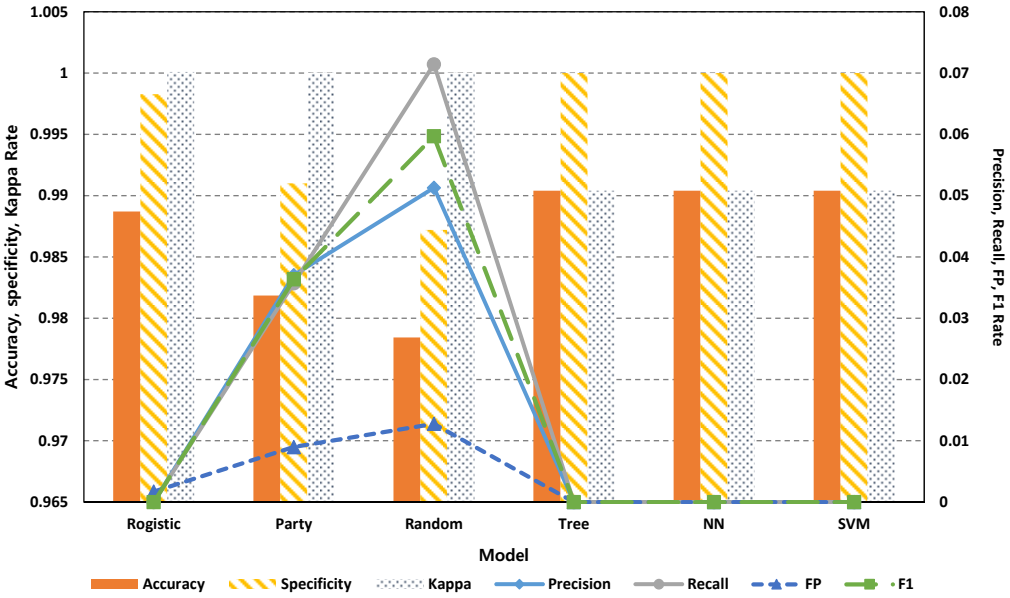
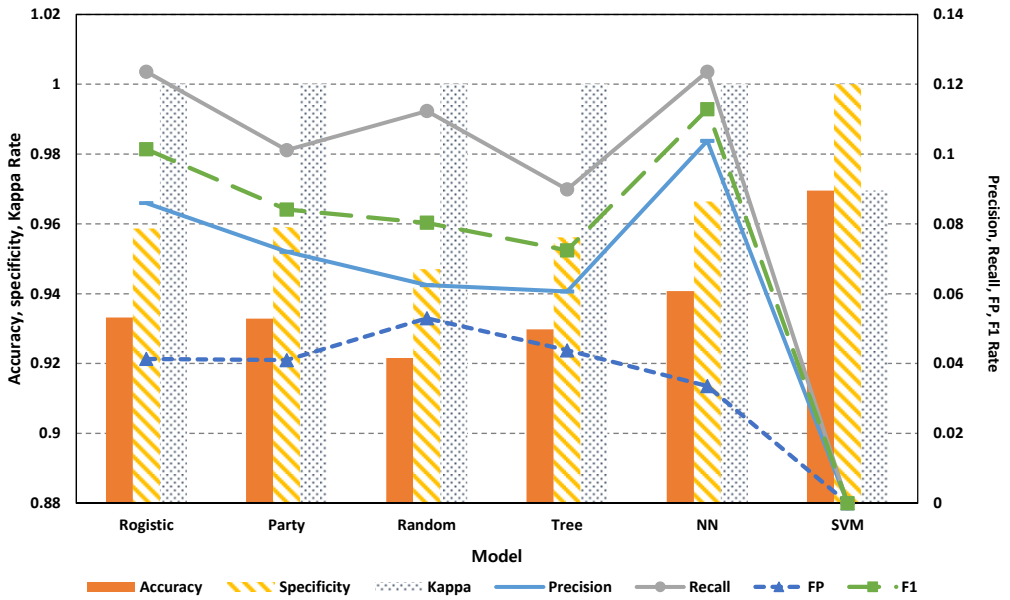


Fig. 4. Classification map of tree models at Gosan(185).



(a)



(b)

Fig. 5. Model assesment using metrics from confusion matrix at (a) Jeju(184) (b) Gosan(185).

의사결정나무를 통해 고산지역에서의 안개 발생 유무 예측 결과를 확인하였다. 의사결정나무의 경우 조건부 추론 나무와 마찬가지로 입력변수의 분류를 통해 예측을 시도했으며, 본 연구에서 진행한 의사결정나무의 분류도를 Fig. 4에 나타내었다. 7개의 입력변수 중 t-dt를 통해 가장 먼저 안개 발생 유무를 분류하였으며, 기온과 노점온도 차가 1.65 이하일 때는 습윤한 정도가 크기 때문에 첫 번째 분류 척도로 사용되었다.

구축한 신경망 모델구조에서는 3개의 은닉층(hidden layer)을 가지고 28 가중치로 분석하였으며, 각 은닉층의 노드 수에 따른 훈련 횟수를 100번으로 고정하였다. 예측 모델의 예측값과 실측값의 정확도는 고산지역이 92.24%로 나타났다.

서포트 벡터 머신은 Vanilla kernel 함수를 기본으로 하여 예측 및 검증을 실시하였는데, 제주 지역의 Training error값이 0.095, 고산 지역은 0.109로 분석되었으며, 두 지역 모두 예측률은 높게 나타났으나 안개 발생 유무는 예측하지 못했다.

안개 발생이 가장 많았던 고산지역의 2015년 4월 안개 발생 유무를 wrf output을 이용하여 예측하였다. 입력변수와 분석 모델은 동일한 6개의 데이터 마이닝 기법을 활용하였으며, 예측 결과를 Table 5에 관측값 검증 결과와 함께 나타내었다. 예측률은 6개 모델 모두 80% 이상으로 나타났으나, 안개 발생 유무를 제대로 예측하지 못하여, 예측률 자체를 판단하는 데 있어 특별한 의미를 가지지 못하는 것으로 판단된다. 이는 수치모델 결과에서 습도장을 제대로 모사하지 못함에 따른 결과로 보이며, 우수한 습도장 예측을 위한 자료동화 등의 부가적인 작업이 수행되어야 할 것으로 사료된다.

3.2. 모델 평가

본 연구에서 개발한 데이터 마이닝 모델 6개의 성능 비교를 위해서 혼동 행렬로부터 계산한 메트릭들의 결과를 Fig. 5에 나타내었다. 위의 그래프는 제주이고 아래는 고산 관측값과 예측값을 사용한 메트릭 결과들이다. 그림에서 막대그래프들은 왼쪽 주축의 값으로 비교하고, 꺾은 선형그래프들은 우측 보조축의 값으로 비교분석하면 된다.

흔히 정확도로 사용되어지는 Accuracy의 경우는 제주에서는 의사결정나무, 신경망, SVM이 0.99로 가장 높은 값을 보였으나 코헨의 Kappa수가 높으면서, Precision, Recall, F1 점수가 제일 높은 것은 랜덤포레스트로 나타났다. 종합적으로 분석한 결과로는 조건부추론나무가 정확도는 0.98로 FP Rate는 0.009로 낮음으로 인해 안개발생이 없는데 안개로 예측되는 비율도 낮으면서 안개 발생 유, 무 전체를 예측하는 정확도는 높은 것으로 보인다.

고산의 경우도 살펴보면, 제주에 비해 안개 발생 예측률을 알 수 있는 Precision, Recall 수치가 더 높은 것으로 보이며, 이는 제주에 비해 고산의 안개 발생 빈도가 높아서 안개 발생 메카니즘에 대한 학습이 더 잘 이루어진 것으로 보인다. Accuracy는 0.97로 서포트 벡터 머신이 가장 높았으나, Precision, Recall, Specificity 그리고 F1도 높고 FP Rate는 낮으며, 정확도 Accuracy도 0.94로 두 번째로 높은 신경망의 결과가 종합적으로 봤을 때 가장 좋은 것으로 나타났다.

기계 학습의 경우, 안개 발생이 많을수록 우수한 모델을 생성시킬 수 있으나, 제주도 연안에 위치한 제주, 고산의 경우에는 안개 발생 빈도가 낮아 충분한 학습이 이루어졌다고 볼 수는 없지만, 이처럼 관측값과의 검증에 의하면 신경망에 의한 안개 발생 예측이 가장 효과적이라고 볼 수 있다.

4. 결 론

안개 발생 메카니즘을 분석하고 도로, 관광 등의 기상정보로 제공하고자 안개 예측 정확도 개선을 위해 다양한 데이터마이닝 기법을 이용하여 기계학습을 실시하였다. 본 연구에서 시행한 학습으로는 의사결정나무, 조건부 추론 나무, 랜덤포레스트, 다항 로지스틱 회귀 분석, 신경망, 서포트 벡터 머신의 6가지 기법들이다.

제주도의 제주와 고산지역의 10년간 관측값으로 학습하여 만든 모든 모델들의 2015년 관측값과의 검증에 의하면 모델 모두 92%를 넘는 예측률을 보여주었고, 안개 발생이 가장 많았던 고산의 2015년 4월의 수치모델 결과를 입력장으로 사용한 예측률 또한 80%을 넘는 것으로 분석되었다. 최종적으로 모델들의 종합적 평가를 위해 혼동 행렬로부터 계산한 메트릭 평가를 시행하였고, 모두 Accuracy는 높은 편이나 다른 통계분석 결과들을 통합적으로 봤을 때 신경망이 데이터 마이닝 기법 중 안

개 발생 유무 예측에 가장 효과적이라고 판단된다.

결론적으로 제주 지역에서의 안개 발생 유무 예측을 위해서는 신경망 기법을 활용하면 빠르고 정확한 안개 예측이 가능하리라 본다. 하지만 제주도의 경우는 연안보다도 중산간 지역에서의 복사안개 및 활승안개 발생 빈도가 높은 것으로 나타나 실제 도로기상 예보로 활용 가능한 모형을 개발하기에는 기계 학습할 관측자료 수집에 한계가 있다. 향후 중산간 지역의 시정 관측 등을 추가함으로써 확보된 기상 자료들의 데이터베이스 구축을 통해 제주 지역에서 안개로 인해 발생하는 교통사고 등의 위험을 줄이기 위한, 더 개선된 데이터 마이닝 기법의 안개 예측모델을 개발하고자 한다.

감사의 글

이 논문은 2015년도 지역주력산업육성(R&D) 기술개발 사업인 「운전자 환경 반응형 CEV(Connected Electricity Vehicle) 서비스 개발」 사업(R0003890)으로 산업통상자원부의 지원을 받아 연구한 논문임.

REFERENCES

Annette, J. D., Adrian, G. B., 2008, An Introduction to generalized linear model 3rd ed., Chapman Hall/CRC, 149-163.

Bishop, M. C., 1999, Neural networks for pattern recognition, Oxford University Press, 140-148.

Breiman, L., 2001, Random forests, Machine Learning, 45(1), 5 - 32.

Cherkassky, V., Ma, Y., 2004, Practical selection of SVM parameters and noise estimation for SVM regression, Neural Networks, 17(1), 113-126.

Cristianini, N., Taylor, J., 2000, An Introduction to support vector machines, Cambridge University Press, 93-124.

David, J. H., 1998, Data mining: Statistics and more, The American Statistician, 52(2), 112-118.

Heo, K. Y., Ha, K. J., 2004, Classification of synoptic pattern associated with coastal fog around the korean peninsula, Korean Meteorological Society, 40(5), 541-556.

Jiawei, H., Micheline, K., Jian, P., 2011, Data mining: Concepts and techniques 3rd ed., Morgan Kaufmann, 23-26.

Kim, M. J., 2009, Self-organizing neural network of fog prediction, Master's Dissertation, Yonsei University, Seoul.

Ko, Y. S., 2011, The construction methodology of a rule-based expert system using CART-based decision tree method, The Journal of the Korea Institute of Electronic Communication Sciences, 6(6), 849-854.

Leiper, D. F., 1994, Fog on the U.S. west coast: A Review, Bull. Amer. Meteor. Soc., 75(2), 229-240.

Ramón, D. U., Sara, A. A., 2006, Gene selection and classification of microarray data using random forest, BMC Bioinformatics, 7(1), 3.

Road Traffic Authority traffic accidents Comprehensive Analysis Center, 2014, Weather conditions specific mortality analysis.

Seo, M. G., 2014, Data processing using the R & industry analysis, Gilbut Inc., 447-450.

Shin, J., Park, H., Seok, K. H., 2009, Variance function estimation with LS-SVM for replicated data, Journal of Korean Data and Information Science Society, 20, 925-931.

Sohn, H. J., 2010, Characteristic analysis of long-term variability of fog occurrence in South Korea, Master's Dissertation, Kongju National University, Korea.

Yaser, S. A., Malik, M. I., Lin, H. T., 2012, Learning from data, AML Book, 32-39.