데이터마이닝 기법을 이용한 국지기상예보칙 작성 방안 연구

최재훈*, 이상훈**
*국방대학교 전산정보학과 석사과정
**국방대학교 전산정보학과 교수
e-mail: ricechoi@korea.com

A Study on Creation Plan of the Local Weather Prediction Method Using Data Mining Techniques

Jae-Hoon Choi*, Sang-Hoon Lee**
*Dept. of Computer Science & Imformation, Korea National Defense University
**Prefessor on Dept of Computer Science & Imformation, Korea National Defense University

요 약

데이터 마이닝 기법 중 회귀분석 기법과 의사결정나무 분석 기법을 이용하여 국지기상예보칙을 작성하는 방안을 연구하였다. 회귀분석기법을 이용하여 예보값에 영향을 미치는 예보요소를 도출하고, 도출된 예보요소를 회귀분석 기법과 의사결정나무 분석 기법에 적용하여 예보칙을 작성하였다.

1. 서론

모두가 잘 알고 있는 삼국지연의의 적벽대전, 주유군의 화공에 의해 조조군이 패하기에는 제갈양의 남동풍이 있었다. 이 남동풍은 제갈양이 하늘에 빌어 일으킨 바람이 아니라 평소의 세밀한 관측에 의한 예보에 의한 바람이었다. 이런 정확한 기상예보로 인해서 주유와 제갈양은 백만의 조조군을 무찌를 수 있었다. 이와 같이 기상예보는 군작전에 결정적인 영향을 미치고 있다. 기상의 영향은 정밀 유도무기 등을 사용하는 현대전에서는 더욱 중요하다 할 수 있다.

공군 기상 예보는 단지 공군의 비행작전 뿐만 아니라 전군의 작전에 지대한 영향을 미치고 있다. 전산학의 발전에 따라 기상학에도 많은 발전이 있었는데, 기상청과 공군 기상전대에서는 단기·중기·장기예보를 함에 있어, 슈퍼컴퓨터에서 수행하는 기상수치예보 시스템을 이용함으로서 보다 객관적이고 우수한 예보를 생산하고 있다.

그러나 군 비행 작전 등에 필수적인 국지기상 예보칙 연구에는 수동적인 방법을 사용하는 경우가 많다. 이에 따라 수 십 년간의 데이터를 분석하기에는 과도한 시간과 인력이 요구되며, 기상데이터 분석에 있어 효율성, 정확성, 객관성이 다소 부족한 실정이다.

따라서 국지기상 연구에 전산학적 기법을 도입하여 국지기상 예보칙을 보다 우수하게 발전시키기 위한 연구가 필요한데, 이에 적합한 전산학적 기법이 데이터마이닝이다. 데이터마이닝 기법에 적용하여 강수 등과 같은 고·저기압과 관련된 종관 기상현상을 예보하는 것은 어렵겠지만, 지역적으로 나타나는 현상인 악시정(안개) 등에 대한 국지 기상예보칙은 몇 가지 예보요소를 이용하며 비교적 단순한 분석과정으로 예보를 결정하기 때문에 적용할 수 있다.

본 논문에서는 악시정 예보칙을 결정하는 입력변수를 회귀분석을 통하여 분석하여, 예보값을 가장 잘 설명할 수 있는 예보요소의 집합을 도출하고, 도출된 예보요소를 회귀분석 기법과 의사결정나무 분석 기법에 적용하여 악시정 예보칙을 작성하는 방안을 연구하고자 한다.

본 논문의 구성은 국지 기상예보칙에 대한 소개, 데이터마이닝 특히 회귀 분석 기법과 의사결정나무 분석 기법에 대한 이론적 배경과 회귀분석을 통한 데이터 분석, 회귀분석 기법 및 의사결정 나무 분석에 의한 악시정 예보칙 및 결론으로 이루어져 있다. 분석에 사용된 데이터마이닝 도구는 SAS

Enterprise Miner 4.0(이하 SAS E-miner) 이다.

2. 국지기상 예보칙

가. 국지기상 예보칙 개요

국지기상 예보칙은 강풍, 뇌우, 강수, 악시정, 최고·최저 기온, Low CIG(Low Ceiling, 저고도 차폐)등에 대한 예보칙으로서 각 기상대의 지리적 요건 등을 포함한 기후요소들을 통한 예보이다. 따라서, 각 기상대 별로 독창적이면서도 경험적인 요소가 많이 포함되어, 종관 예보와 달리 각 기상대 별로 사용하는 예보요소가 다르다.

본 논문에서는 수원 기상대의 추계 악시정 예보칙에 대해서 연구하였다.

나. 수원기상대 악시정 예보칙

악시정은 주로 안개로 인한 시정 불량 현상으로 수원기지의 경우 대부분 복사무 형태를 띤다. 수원 비행장은 기지 지형적 특성으로 인해 유입된 해무와 전선무도 존재하지만 항공작전 지원에 필요한 것은 복사무 예보이다.

수원의 악시정 예보는 일반적으로 사용되는 최저습도, 최고기온, 전일 강수, 야간 바람 등에 추가하여 기압계 패턴 등을 이용함으로서 정확률을 높였는데, 악시정 예보칙의 절차를 간단히 표현하면 그림 1과 같다.

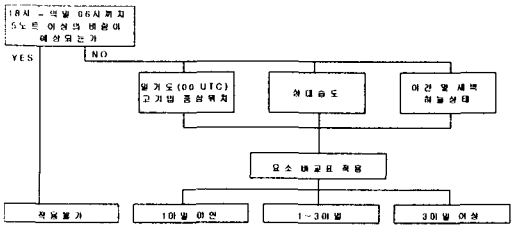


그림 1. 예보 검목 절차

3. 데이터마이닝 소개

가. 데이터마이닝 정의

데이터마이닝이란 자동화 되고 지능을 갖춘(automated and intelligent) 데이터베이스 분석기법으로 90년대 초반부터 지식발견(KDD: Knowledge Discovery in Databases), 정보 발견(information discovery), 정보수확(information harvesting) 등의 이름으로도 소개되어 왔는데 일반적으로 "대량의 데이터로부터 새롭고 의미있는 정보를 추출하여 의사결정에 활용하는 작업"이라 정의된다[2].

데이터마이닝은 대용량(massive)의 관측 가능한 자료(observational data)를 다루며, 일반적인 기법들로 연관성 규칙 발견(Association Rule Discovery), 회귀 분석(Regression Analysis), 군집 분석(Cluster Analysis), 의사결정나무(Decision Tree), 인공신경망(Artificial Neural Network) 등이 있다.

나. 회귀 분석 개요

회귀분석은 변수 상호간의 관계를 표본으로부터 추정하는 방법으로 목표변수가 입력변수에 의해 어떻게 설명 또는 예측되는지를 알아보기 위해 자료를 적절한 함수식으로 표현하여 분석하는 통계적 방법이다[2].

선형방정식에 의해 자료를 표현하는 것을 선형회귀분석(Linear Regression Analysis)이라 하고, 비선형 방정식에 의해서 표현하는 것을 비선형회귀분석(Nonlinear Regression Analysis)이라 한다. 또 입력변수가 하나인 경우의 단순회귀(Simple Regression) 분석과 입력변수가 여러 개인 경우의 다중회귀(Multiple Regression) 분석으로 구분된다[1].

일반적으로 목표변수의 설명에 불필요한 변수들이 모두 들어 있는 완전모형(Full Model)보다는 필요한 변수들만 들어 있는 축소모형(Reduced Model)이 보다 바람직한 회귀모형이라 하겠다. 따라서, 입력 변수를 선택해야 할 필요가 있는데, 이러한 방법에 전진선택법(Forward Selection), 후진소거법(Backward Elimination), 단계적 방법(Stepwise Method) 등이 있다[1].

다. 의사결정나무 분석 개요

의사결정나무(Decision Tree)는 의사결정규칙(Decision Rule)을 나무구조로 도표화 하여 분류(Classification)와 예측(Prediction)을 수행하는 분석방법이다. 이 방법은 분류 또는 예측의 과정이 나무구조에 의한 추론규칙(Induction Rule)에 의해서 표현되기 때문에, 다른 방법들(신경망, 판별분석, 회귀분석 등)에 비해서 연구자가 그 과정을 쉽게 이해하고 설명할 수 있다는 장점을 가지고 있다. 연구자는 나무구조로부터 어떤 변수가 목표변수의 분류에 영향을 많이 주는지 그리고 어떤 경우에 목표변수가 분류되는지를 쉽게 파악할 수 있다[1].

목표변수가 이산형인 경우에는 각 범주에 속하는 빈도(Frequency)에 기초하여 분리가 일어나며, 이를 분류나무(Classificaton Tree)를 구성한다고 한다. 이때 사용되는 분리기준은 카이 제곱 통계량(Chi-Square Statistic)의 P-값, 지니 지수(Gini Index), 엔트로피 지수(Entropy Index)등이며, 목표변수가 연속형인 경우에는 목표변수의 평균(Mean)

에 기초하여 분리가 일어나는데 이때 사용되는 분리 기준으로는 분산분석에서의 F통계량, 분산의 감소량(variance Reduction)이다[1].

4. 회귀분석을 통한 기상 데이터 분석

가. 입력 데이터 구성 및 변형

분석에 이용한 데이터는 기상대의 자료를 이용하여 10년치 자료로 구성하였다. 지상 관측자료 및 단열선도자료, 일기도 등의 자료를 이용하였다.

전체 입력데이터의 구성은 표 1와 같다.

표 1. 입력 데이터 구성

변수	내용	변수	내용	변수	내용
Patt	기압계패턴	TTd!!	매시 온도·노점온도 차	Wea!!	매시 현천
RHmin	일 최저습도	Vis!!	매시 시정	CCL	대류응결고도
Tmax	당일 최고기온	FogTime Yes	전일 안개지속시간	LCL	상승응결고도
RainYes	전일 강수유무	SandTi meYes	전일 황사지속시간	Cig1,4	1시, 4시 실링고도
Wdir 1, 4	1시, 4시 풍향	Wvelmax	최대풍속	SSI	안정도지수
Trange	기온 일교차	RH!!	매시 상대습도	TdTmin	15시 노점온도·최저기온 차

나. 단순 선형 회귀분석 실행 및 결과

이 절에서는 입력 변수 전체를 각각 단순선형 회귀분석 기법으로 분석하고 그 결과를 검토하였다. 목표변수는 vis07, 즉 7시 시정값이다. 분석결과는 그림 2에서 나타난다. 결정계수 값은 회귀식이 얼마나 목표변수를 잘 나타내는 지를 알 수 있는 값이다[1].

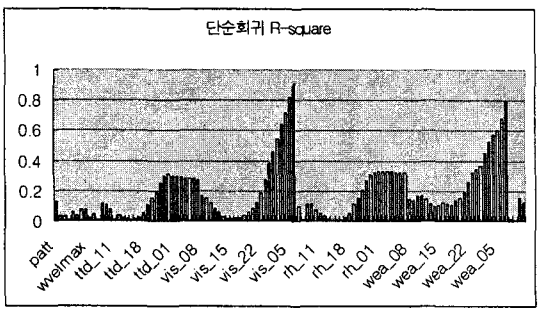


그림 2. 단순선형 회귀 모형의 분석결과(결정계수)

다. 다중 선형 회귀분석 실행 및 결과

다중 선형 회귀분석을 실행함에 앞서 입력변수를 결정해야 한다. 데이터 선택 방법은 표 3와 같다.

표 2 입력 데이터 세트 구성

I - i ~ v	단순선형 회귀분석 결과를 이용한 데이터 선정
II - i ~ vii	기존의 국지 예보칙에 사용되는 변수 및 임의적 선정
III - i ~ iii	회귀분석 자체에서 지원되는 데이터 선별법이용(전진 선택법, 후진소거법, 단계적 방법)
IV	Using a Variable Selection Node

표 2와 같이 구성된 데이터 세트를 다중 선형 회귀 분석 결과에 대해서 결정계수값을 비교값으로 설정하여 그래프로 나타내면 그림 3과 같다.

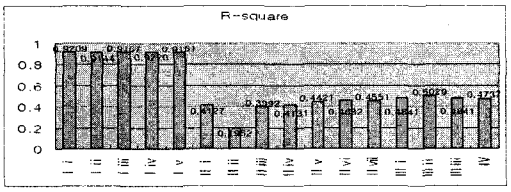


그림 3. 다중 선형 회귀 분석의 입력 데이터 세트별 결과

분석 결과, 선형 회귀 분석 기법을 이용하여 예보요소를 선정하는 데 있어 가장 효율적인 방법은 전진선택법과 단계적 방법이다. I 계열의 결정계수 값이 가장 높지만 이는 15시 이후의 요소를 이용한 것으로 예보에 사용하기는 불가능하다. 최종적으로 선택된 예보칙 입력변수는 표 3과 같다[4].

표 3. 예보칙 입력변수

구분	입력 변수
III - i (전진선택법), III - iii (단계적 방법)	cig2, fogtimeyes, patt, tdtmin, trange, vis08, wdirmean, wdir_01, wdir_04, wvelmax
III - ii (후진소거법)	cig2, patt, tdtmin, trange, vis_08, vis_14, wdirmean, wdir01, wdir04, wea15, wvelmax

5. 악시정 예보칙

가. 회귀분석 기법을 이용한 예보칙

회귀분석의 결과에서 회귀계수를 이용함으로써 예측 모형식을 만들 수 있다. 이 절에서는 전진선택법에 의해 선정된 입력변수를 이용한 회귀모형을 이용하였다. 반응확률에 대하 예측 모형식은 아래와 같다.

VIS07 = 19.5312 + VIS_08×0.1328 +·····+CIG4_1 ×(-19.6642) + CIG4_2 ×6.3638 + CIG4_3×17.3238 + CIG4_4×(-2.6610) + FOGTIMEYES×(-0.00755)

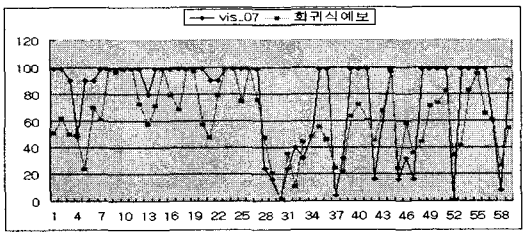


그림 4 회귀분석 예보법에 의한 예보값과 실측값 분포

그림 4는 위의 예측 모형식을 이용하여 악시정 예보칙을 만들어 이를 2002년 추계 자료를 이용하여 검증한 내용이다.

표 5 회귀분석 예보법 오분류표(2002년 추계)

실황 \ 예보	3MILE 이상	1~3MILE 미만	1MILE 미만	계
3MILE 이상	41	4	0	45
1~3MILE 미만	1	8	1	10
1MILE 미만	0	3	1	4
계	42	15	2	59

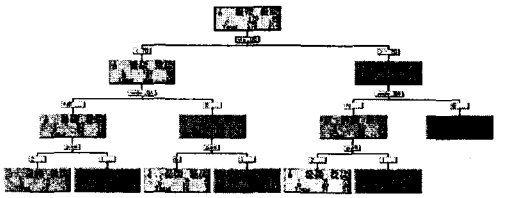
정확율 : 50 + 59 × 100 = 84.7 %

그림4의 결과를 오분류표에 적용하여 예보칙의 정확율을 검토하였다. 분류 기준은 공군 비행 작전

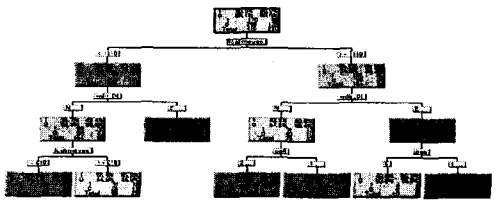
의 시정에 따른 비행 기준, 즉, 시계 비행 규칙(VFR), 계기 비행 규칙(IFR), 비행불가(B/M)로 선정하였다. 오분류를 살펴보면 예보 정확율은 약 85%로서, 수원기상대에서 2002년 자료를 이용하여 자체 악시정 예보법을 검증한 결과에 비해 상당히 높은 수치를 기록하였다.

나. 의사결정나무 분석기법을 이용한 예보칙

앞의 3.다 절에서 언급한 것과 같이 하나의 예보 입력 요소에 대해서 목표변수를 이산형 변수로 설정한 Tree가 3개, 연속형 변수로 설정한 Tree가 2개로, 총 5개의 Tree를 분석하였다. 분석결과 엔트로피 지수를 분석기준으로 한 Tree가 가장 우수한 정확율을 보였으며, 목표변수를 연속형 변수로 설정한 Tree는 예보칙으로 적합지 않았다.



가. Tree1



나. Tree2

그림 6 예보칙에 사용된 Tree

의사결정나무 분석 기법을 이용한 예보칙은, 전진선택법에 의해 선정된 입력변수(Tree2)와 후진소거법에 의한 입력변수(Tree1)에 대해 엔트로피 지수를 분석 기준으로 생성된 Tree를 이용하여, 이 두개의 Tree에 관측치를 대입하여 결과값을 확인하고, 두 Tree에서 결과값이 같은 경우는 그 값으로 예보값을 결정하고, 다를 경우는 확률이 높고 경우의 수가 많은 값으로 결정하였다. 이러한 방법에 의해 2002년 추계 자료를 이용하여 분석하였다.

표 6 Tree 예보칙 오분류표

실황 \ 예보	3 MILE 이상	3 MILE 미만	계
3 MILE 이상	43	2	45
3 MILE 미만	6	8	14
계	49	10	59

정확율 : 51 ÷ 59 × 100 = 86.4%

Tree 예보칙의 정확율은 86.4%로서 비교적 높은 것으로 산출되나, 회귀분석 예보칙을 3마일 이상과 미만으로 구분한 결과, 91.5%의 정확율을 보임으로서, 회귀분석 예보칙에 비해서는 다소 정확율이 떨어졌다.

6. 결 론

본 연구에서는, 예보칙을 연구함에 있어, 보다 과학적이고 객관적인 방법으로 국지예보를 실시하고 많은 데이터를 쉽고 효율적으로 분석하기 위해서 전산학적 기법을 도입하였다. 국지기상 예보칙 중, 수원기상대 악시정 예보칙에 대한 예보요소를 선정함에 있어 전산학적 기법을 도입하고자 회귀분석 기법과 의사결정나무 분석 기법을 이용하였다.

회귀분석 기법을 이용하기 위해서 기상대에서 보유하고 있는 원시 데이터를 변환하였으며, 이 변환된 데이터를 단순 선형 회귀 분석과 다중 선형 회귀 분석기법으로 분석하여 예보요소를 도출하였으며, 도출된 예보요소를 회귀분석기법과 의사결정나무 분석기법에 적용하여 예보칙을 작성하였다.

이러한 예보칙 작성방안은 몇 가지 장점을 가지고 있다. 첫째, 종전 기상대 예보칙에 비해 높은 정확율을 보였으며, 둘째, 예보칙 개발이 용이하다는 것이다. 이는 동일한 방식으로 전 기상대의 예보칙을 작성할 수 있으며 예보칙 작성에 필요한 인원, 시간 자원을 줄일 수 있다는 것이다. 셋째, 기후의 변화에 따른 예보칙 수정이 용이하다는 것이다. 이는 예보생산지 주위의 기후조건 변화에 따른 예보칙 수정이 용이하며, 엘니뇨, 라니냐 등과 같은 전세계적인 이상 기후현상에 대해서도 즉시 대처할 수 있다는 것이다.

위에서 제시한 장점과 같이 데이터마이닝을 이용한 예보칙 작성 방안은 예보의 정확률과 효율성을 증대시키는 데 크게 기여할 수 있다.

향후, 악시정 뿐만이 아닌 다른 예보요소에도 이러한 방법을 적용하고, 연관성 분석이나 신경망 분석과 같은 다른 데이터마이닝 기법을 적용할 수 있는 방안에 대해 추가적인 연구가 필요하다.

참고문헌

[1] 데이터마이닝-방법론 및 활용, 강현철외 4 공저, 자유아카데미, 2002. 8.

[2] 데이터마이닝, 장남식 외 2 공저, 대청미디어, 99. 10.

[3] 통계자료분석방법, 김종섭 저, 학문사, 98. 3.

[4] 최재훈, 이상훈, 회귀분석을 이용한 국지기상 예보요소 분석, 군사과학기술학회, 2003.