한국지리정보학회지 6권 1호 : 12~23(2003)

일반화 기반 분류기법을 이용한 산불예측시스템 설계 및 구현*

김상호[1※] · 김대진[1] · 류근호[1]

Design and Implementation of Forest Fire Prediction System using Generalization-based Classification Method*

Sang-Ho KIM[1※] · Dea-Jin KIM[1] · Keun-Ho RYU[1]

요 약

정보산업의 급속한 발전은 축적되어 있는 대규모의 데이터로부터 보다 가치 있는 정보 생성 및 정확한 데이터 분석 능력을 요구하고 있다. 특히 데이터마이닝 기법을 이용하여 주어진 데이터간의 연관관계를 도출하고, 얻어진 패턴을 바탕으로 미래를 예측하는 방법은 주목을 받고 있다. 이 연구에서는 속성중심 귀납방법과 분류규칙을 통합한 일반화 기반의 분류기법을 제안하였고, 간결한 모델의 구축 및 규칙 추출을 수행하였다. 또한 일반화 기반 분류 예측시스템에 산불데이터를 적용하여, 기상 데이터와 산불발생 사이의 관련성을 분석하고 효율적인 예측을 수행하였다. 이 연구에서 제시한 기법은 반복적으로 발생하는 자연재해에 대한 분석 및 예측, 에너지의 수요량 예측 등과 같이 실생활의 중요한 부분들에 다양하게 응용할 수 있다.

주요어: 데이터마이닝, 분류기법, 일반화, 규칙갱신, 산불예측, 시각화

ABSTRACT

The expansion of internet and the development of communication technology have brought about an explosive increasement of data. Further progress has led to the increasing demand for efficient and effective data analysis tools. According to this demand, data mining techniques have been developed to find out knowledge from a huge amounts of raw data. This paper suggests a generalization based classification method which explores rules from real world data appearing repeatedly.

Also, it analyzed the relation between weather data and forest fire, and efficiently predicted through it as a prediction model by applying the suggested generalization based classification method to forest fire data. Additionally, the proposed method can be utilized variously in the important field of real life like the analysis and prediction on natural disaster occurring repeatedly, the prediction of energy demand and so forth.

2003년 1월 20일 접수 Received on January 20, 2003 / 2003년 3월 7일 심사완료 Accepted on March 7, 2003
* 이 연구는 한국과학재단 RRC(ICRC) 지원으로 수행되었음
1 충북대학교 컴퓨터과학과 Department of Computer Science, Chungbuk National University
※ 연락저자 E-mail: shkim@dblab.chungbuk.ac.kr

KEYWORDS: Data Mining, Classification Method, Generalization, Rule Updating, Forest Fire Prediction, Visualization

서 론

정보산업의 급속한 발전은 축적되어 있는 대규모의 데이터로부터 보다 가치 있는 정보 생성과 정확한 데이터 분석 능력을 요구하고 있다. 대용량 데이터로부터 예측을 위한 유용한 정보를 고객에게 제공하기 위해서는 관계, 패턴, 규칙탐사를 통해 모형을 수립하여야 한다.

복잡한 현실세계의 문제 해결을 위해 모델을 만들고 이를 이용하여 결과를 미리 예측하거나 결과에 대한 분석할 수 있는 다양한 방법들이 제시되고 있다. 데이터마이닝 기법을 사용하면 주어진 데이터들의 연관관계를 추출하고, 얻어진 패턴을 바탕으로 모델을 수립할 수 있다(Han과 Fu, 1996). 특히, 패턴탐사나 시계열 분석, 분류기법 등은 미래를 예측하는 방법으로 주목을 받고 있다. 분류기법은 입력 데이터를 분석하여 각 클래스에 대한 정확한 표현이나 모델을 수립하는 방법이고, 이를 위해 통계(statistics), 신경망(neural network), 결정트리(decision tree) 등이 연구되어 왔다(Cortes 등, 1995; Lange, 1996).

기존의 통계적 기법은 가설 및 검증과정을 통해 비교적 신뢰성이 있는 예측 모델을 만들어내지만 많은 데이터를 기반으로 사용자의 요구사항을 빠르게 응답하기에 적합하지 못하다. 또한 신경망 분석기법은 복잡한 구조를 가진 데이터들 사이의 관계나 패턴을 찾아내는 비선형 모형의 하나로, 모델 설계에서 결과를 얻기까지 많은 어려움이 따르며, 설명력이 부족하다는 단점을 가진다(Wang, 1999).

일반화 기반의 분류 탐사는 일반화 과정을 거친 데이터로부터 흥미로운 지식을 추출하는 셋-기반의 데이터 분석기법으로 상위레벨에서의 규칙 발견을 도우며, 분석데이터를 압축함으로 I/O 비용을 줄인다. 따라서 이 논문에서는 반복적으로 발생하는 실세계의 데이터로부터 지식을 탐사하는 일반화 기반 분류 기법을 제시하며, 규칙들의 능동적 갱신으로 미래의 사건을 효과적으로 예측할 수 있는 모델을 설계한다.

이 논문은 1993년~1995년 간의 강원도지역의 산불 이력 데이터를 적용하여, 산불발생과 기상요인간의 관련성을 추출하고 기상데이터에 대한 산불 발생 예측을 수행하는 일반화 기반 분류 예측 시스템을 제안한다. 제안하는 시스템은 기존 데이터베이스에 구축된 시·공간 정보를 이용하여 규칙을 생성하고 향후 전망이나 기대 예측치를 관찰하며, 또한 규칙 갱신을 효율적으로 수행하는데 목적이 있다.

이와 같은 연구를 통해 반복적으로 발생하는 자연 재해에 대한 분석 및 에너지 수요량 예측 등의 이력데이터를 이용한 실생활의 중요한 부분들에서 다양이 활용할 수 있다.

연구자료 분석 및 연구범위

1. 산불 데이터의 분석

우리나라의 산불은 날씨가 가장 건조하며 낙엽이 많이 쌓여 있는 춘기와 추기에 집중적으로 발생하고 있다. 전 국토의 65%가 산지이며, 불에 잘 타는 침엽수임상이 42%로 가장 많고 잡목이 우거져 있으며, 가연성 낙엽이 많이 쌓여있기 때문에 일단 산불이 발생하면 수평적으로나 수직적으로 빠르게 확산된다.

산불은 대부분 인위적인 실화로 등산객, 성묘객, 무속행위자 등의 입산자의 부주의에 의한 실화와 논·밭두렁 소각에 의한 산불발생이 연평균 66%로 가장 많은 비중을 차지하고 있다. 이외에도 어린이 불장난, 쓰레기 소각, 담뱃불 실화, 군사훈련 등의 순으로 산불이 발생하고 있다.

최근 5년(1996~2000)간 우리나라의 연평균 산불발생현황은 발생건수의 경우 1997년 이후 감소추세를 보이다가 1999년부터 급격한 증가 추세를 보이고 있으며, 5년간 연평균 발생건수는 472건에 피해면적은 6,958ha로 나타났다.

우리나라의 산불발생은 계절풍의 영향을 많이 받는 3~4월에 가장 많이 발생한다. 이 시기는 건조한 날씨가 지속되고 바람이 많이 불어 산림 내의 가연물이 건조해지기 쉬워 산불이 많이 발생한다.

최근 5년(1996~2000)간 계절별 산불 발생 건수를 보면 봄철에 301건으로서 65%가 발생하였으며, 여름철에 7건에 1%, 가을철에 30건에 6%, 겨울철에 13건에 28%가 각각 발생하였다. 또한 최근 5년간 30ha 이상 대형산불 발생지역을 보면 강원(59%), 경남(16%), 경북(11%), 충남(5%)과 부산·전남·경기(각 3% 순)으로 나타났다. 강원도 지역의 경우, 건당 피해 면적이 우리나라의 통계에 두 배인 21ha로서 피해 규모가 큰 것을 알 수 있다(이시영 등, 2001). 또한 임상이 단순한 종들이 밀집되어 있고, 지형 조건이 험준하여 소화에 어려움이 많으며 가연성이 높은 소나무림이 광범위하게 차지하기 때문에 산불이 발생하면 대형화하는 경향을 보이는 것이 특징이다.

산불은 일차적인 원인으로 다양한 인위적 요인에 의하여 시작되지만 이와 같은 실화가 산림에 피해를 줄 정도로 발화·연소·확산되기 위해서는 강우량, 풍속, 상대습도 및 가연성물질의 습도 등 연소환경을 구성하는 기상적 요인과 깊은 상관관계를 가진다(최관과 한상열, 1996 ; 조명희 등, 2001).

2. 연구 범위

일반화 기반 분류 예측시스템의 적용대상 지역은 산불 다발지역인 강원도 지역으로, 그림 1과 같다. 강원도의 7개 시와 11개 군의 각 산불발생 이력데이터를 규칙탐사를 위한 훈련 데이터 셋(training data set)으로 선택하였다. 훈련 데이터 셋의 제약사항은 다음과 같다.

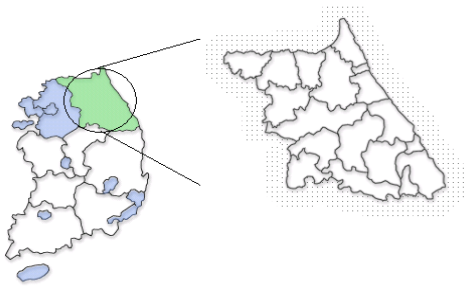

FIGURE 1. The location map of research area

① 제약 사항 1 : 강원도 지역의 시·군을 모델로 선정하였으며, 검색의 기본단위로 관서별 각 시·군 단위로 선정하였다.
② 제약 사항 2 : 훈련 데이터 셋은 1996년부터 2000년의 자료로 제한하였으며, 테스트 데이터로는 1993년부터 1995년까지의 산불발생 이력 데이터로 선정하였다.
③ 제약 사항 3 : 기상 데이터에 대한 기상인자로는 최고온도, 최저온도, 평균온도, 평균풍속, 상대습도에 관한 분석을 다루었다.

일반화 기반 분류기법을 이용한 산불 발생 패턴 분석

1. 일반화 기반 분류기법

일반화 기반 분류기법(classification method based on generalization)은 속성 중심 귀납 기법(attribute-oriented induction method)과 결정 트리 기법(decision tree method)을 통합하여 분류 규칙을 탐사하는 기법이다(Kamb 등, 1997).

속성 중심 귀납기법은 관계형 데이터베이스의 온라인 데이터의 부분집합들을 일반화하고, 일반화된 데이터로부터 흥미 있는 지식을 추출하는 셋 기반의 데이터 분석기법이다(나

민영, 1998). 데이터 일반화 수행과정은 속성 제거, 개념 트리 탐색, 속성 한계치 제어, 카운트 및 다른 집단 함수를 적용하여 진행된다. 일반화된 데이터는 일반화된 릴레이션으로 표현되고, 다른 연산 및 변화과정이 수행되어 다른 형태의 지식으로 변화되거나 매핑되어 규칙이 도출된다. 그리고 일반화 된 릴레이션에 대해 결정 트리 기법을 수행하여, 상위레벨에서의 흥미로운 규칙을 발견한다(Han 등, 1996b).

결정트리 기법은 관심 대상이 되는 집단을 몇 개의 소집단으로 분류하거나 예측을 수행하기 위해서 자주 사용되는 분석기법 중의 하나이다. 특히, 의사결정규칙(decision rule)이 트리구조로 표현되기 때문에 분류 또는 예측을 수행하는 다른 방법들에 비해서 연구자가 분석과정을 쉽게 이해하고 설명할 수 있다. 결정트리 분석은 예측모형 자체로 사용될 뿐만 아니라 이상치(outlier)를 검색하거나 분석에 필요한 변수 또는 교호효과(interaction)를 찾아내는데 많이 이용된다(최종후 등, 1998). 일반화 기반의 규칙발견은 다중 레벨 분류기법과 레벨 조정기법이 통합되어 DBMiner에서 대용량 데이터베이스에서 속성 기반의 유도와 분류기법을 통합함으로 분류의 정확성을 향상시키기 위해 개발되었다(Han과 Fu, 1996).

이러한 결정트리와 일반화의 통합은 상위 추상레벨에서의 흥미로운 규칙 발견을 돕는다(Han 등, 1996a). 또한, 대용량 데이터베이스로부터 튜플(tuple) 수를 줄임으로써 규칙 추출의 어려움을 해결할 수 있으며, 간결한 규칙 생성을 유도한다.

2. 일반화 기반 분류 예측 알고리즘

1) 일반화 및 규칙생성

다음의 알고리즘 1은 저수준 데이터 셋의 입력에 대한 개념 상승 과정을 나타낸다. 먼저 입력되는 데이터 집합의 각 튜플(tuple)들을 LT_j 라 하고, 각 튜플들의 속성을 $LT_i.A_j$ 라고 할 때, 정보량(information gain)을 계산하고 각 속성들의 우선순위를 결정하여 분리기준을 선택한다. 이때 정보량이 작은 속성값에 대한 제거 작업과 정렬을 SortByInfoGain 함수에서 수행한다. GetDigit 함수에서는 문자 데이터를 수치화 하여 일반화 과정을 수행한다. 그리고 문자 및 수치 데이터에 대한 분류 및 수치화를 거쳐 각 클래스에 누적된 빈발 카운트를 $GT_{classLevel}$.CNT를 이용하여 계산하고 저장한다. 결과로 생성된 일반화 패턴들은 규칙베이스에 저장된다.

```
Function: Generalization and rule-pattern
          creation
Input :   LT(low level data set)
Output : GT(generalized rule-pattern set)
-------------------------------------------
for(i = 0; I < GetMaxList(LT); i++)
{
LTi = SortByInfoGain(LTi);
classLevel = GetLevel(LTi);
for(j = 0; j < GetMaxAttr(LTi); j++)
    {
     if(IsDigit(LTi.Aj))
     GTclassLevel.Ai += LTi.Aj;
     else
     GTclassLevel.Ai += GetDigit(Ai);
     }
GTclassLevel.CNT++;
 }
for(i = 0; i < GetMaxClass(LT); i++)
    {
     for(j = 0; j < GetMaxAttr(GT); j++)
     GTi.Aj = LTi.Aj/GTi.CNT;
     }
```

ALGORITHM 1. Generalization and rule-pattern creation

2) 분류 및 예측

알고리즘 2는 새로운 테스트 데이터를 이용하여 클래스를 발견하는 과정을 나타내는 알고리즘이다. 먼저 테스트 데이터의 속성 순서에 따라 속성 값의 순서를 결정하여 가능한 클래스들을 찾는다. 그 다음 정보량이 낮은 속성에 대해서는 수집된 클래스에 대한 제거 작업을 수행한다. 이때 사용자가 명시한 임계치 범위를 이용하며, 해당 클래스의 저장된 패턴과 비교하여 결과를 제공한다. 알고리즘의 출력 값은 복수의 클래스가 될 수 있으며 이때 사용자가 명시한 클래스 레벨이 출력된다.

```
Function: Discover prediction class
Input: TD(test data sample),
usrDefinedLevel(user defined class level),
GT(generalized rule-pattern set)
Output: prediction class
-----------------------------------------
TD = SortByRuleAttr(TD);
for(i = 0; i < GetMaxAttr(TD); i++)
{
if(IsNotDigit(TDi))
TDi = GetDigit(Ai);
for(j = 0; j < GetMaxClass(GT);j++)
{
if(j > 0)
{
if(TDj > GTj + ε.j && TDj < GTj -ε.j)
DelClassLevel(j);
}
else if(TDj < GTj + ε.j &&
                                TDj > GTj -εj)
{
SetClassLevel(j);
}
}
return GetClassLevel(usrDefinedLevel);
 }
```

ALGORITHM 2. Classification and prediction

일반화 기반 분류 예측 시스템

제시한 일반화 기반 분류 기법을 이용하여 예측수행을 위한 시스템을 설계한다. 예측 시스템의 구조 및 각 모듈의 특성을 기술하고, 분류규칙의 탐사과정과 제시한 예측 시스템의 특징을 설명한다.

1. 시스템 구조

분류 규칙 탐사를 위해 설계된 시스템의 전체적인 구성도는 그림 2와 같다.

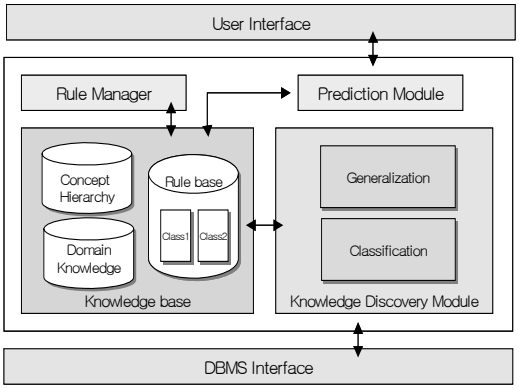

FIGURE 2. Prediction system using generalization based classification method

이 시스템의 세부 구성 요소들은 크게 이력 데이터 셋에 대한 패턴을 추출하는 분류 모듈과 입력 된 테스트 데이터에 대하여 예측 결과를 제공하는 예측 모듈로 구성된다.

첫째, 지식 탐사 모듈(knowledge discovery module)은 일반화 알고리즘과 분류기법 알고리즘을 결합한 모듈이다. 먼저, 선 처리된 이력 데이터를 일반화 과정을 통해 일반적인 특성이나 요약된 고수준 개념으로 변경한 후, 분류기법 알고리즘을 수행하여 지식을 탐사하고 그 결과를 지식베이스에 저장한다. 둘째, 지식베이스(knowledge base)는 지식 탐사를 위해 필요한 모든 배경지식과 지식탐사 모듈에 의해 탐사된 규칙을 저장하는 저장소로 도메인

지식, 규칙 베이스, 개념 계층으로 구성된다. 도메인 지식은 일반화를 위해 사용자가 입력한 데이터에 대한 개념 계층 등을 포함한다. 규칙 베이스는 사용자가 정의 한 클래스의 개수에 대해 각 클래스 별로 규칙을 저장한다. 셋째, 예측 수행 모듈(prediction module)은 사용자가 입력한 입력정보에 대하여 지식베이스에 저장된 각 클래스별 규칙들과 비교하여 예측의 결과를 사용자 인터페이스를 통하여 제공한다. 넷째, 규칙 관리자(rule manager)는 새로운 이벤트가 발생할 경우 규칙 베이스를 갱신하고 관리한다. 즉, 새로운 이벤트가 발생되면 추가되는 데이터에 대해서 지식탐사를 점진적으로 수행하고 규칙 베이스에 저장된 기존의 규칙과 새로 탐사한 규칙을 비교하여 규칙을 추가하거나 변경한다. 마지막으로 사용자 인터페이스(user interface)는 사용자의 입력변수에 따른 예측을 수행하고자 할 때, 입력변수는 예측 수행 모듈에 의해 지식베이스에 저장되어 있는 각 클래스별 규칙과 비교한 후 가시화 도구를 이용하여 예측의 결과를 제공한다.

2. 각 모듈 설계 및 수행과정

다음은 일반화 기반 분류 예측 시스템의 수행과정을 설명하며, 주요 절차로써 규칙 생성, 예측 수행, 지식베이스 갱신 과정으로 나뉘어진다. 세부 설명은 다음과 같다.

1) 규칙 생성 과정

관련 데이터에 대한 규칙을 생성하는 과정은 데이터 수집, 일반화 과정, 규칙 생성, 지식베이스에 저장 과정을 거치게 된다. 수행과정은 그림 3의 ① ~ ④와 같으며, 각 과정의 구체적 내용은 다음과 같다.

① 과정 : 데이터 수집

작업과 관련된 데이터로서 운영 데이터베이스로부터 추출한 훈련 집합과 사용자 및 전문가에 의해 미리 정의된 개념 계층을 입력한다.

② 과정 : 일반화 모듈

훈련 데이터 집합에 개념 계층을 적용하여 일반화 데이터 집합을 생성한다. 이때 개념 계층의 상승정도는 미리 정의되거나 또는 사용자의 요구에 의해 결정된다. 상위 레벨로 모두 일반화 수행 후 중복된 튜플은 카운팅 하여 제거시킨다.

③ 과정 : 규칙 생성 모듈

일반화 과정을 거친 데이터 집합을 사용자의 임의에 따라 m개의 클래스로 나누며, 분리기준을 선택하기 위하여 엔트로피를 계산하여 각 클래스별 사례집합의 후보 속성에 대한 정보 값을 구한다. 최대 정보를 가진 후보 속성을 선택하여 결정 트리를 생성한다.

④ 과정 : 규칙 저장

클래스별 추출된 각 패턴들은 속성별 도메인 값과 카운팅 값을 포함하여 지식베이스에 저장한다.

2) 예측 수행 과정

예측 수행 과정은 그림 3의 ⑤ ~ ⑦과 같으며, 사용자의 입력이 있게 되면 예측모듈에 의해 비수치 데이터에 대한 일반화 작업이 이루어진다. 일반화 된 데이터는 속성 값에 따른 할당된 클래스 값을 얻게 된다. 예측모델은 지식베이스의 해당 클래스의 패턴과 비교 및 검색을 수행하며, 테스트 데이터와 비교한 결과 예측치를 가시화 작업을 통해 보여준다.

3) 지식베이스 갱신 과정

지식베이스 갱신 과정은 그림 3의 ⑧과 ② ~ ④과 같으며, 새로운 훈련 데이터 집합이 발생할 경우 일반화 모듈과 규칙 생성 모듈의 과정을 거쳐 패턴을 추출하게 된다. 지식베이스에 저장되어있는 해당하는 클래스의 패턴을 갱신한다.

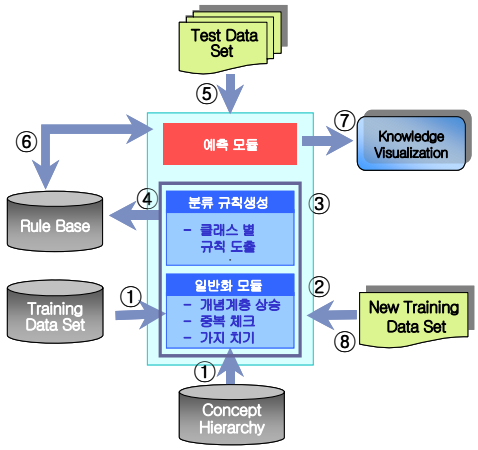

FIGURE 3. The flowchart of prediction system

3. 일반화 기반 분류 예측 시스템의 특징

일반화 기반 분류 예측시스템이 갖는 특징은 다음과 같다.

첫째, 규칙을 추출하기 위해 관련된 공간, 비공간 데이터를 추출하고 정제하며 데이터의 변형 및 조정을 통하여 표준화한다. 둘째, 공간 데이터에 대한 의미 있는 패턴을 추출하기 위한 일반화 기반 분류 알고리즘을 이용한다. 전처리 과정을 거친 데이터들은 미리 제시된 배경지식을 이용하여 상위 레벨로 일반화되며, 분석 데이터를 압축함으로써 I/O작업과 탐사 과정을 효율적으로 수행한다. 이러한 일반화 기법을 적용함으로써 상위레벨에서의 흥미로운 지식을 추출이 가능하다. 셋째, 각 클래스별로 추출된 규칙들은 규칙베이스에 관리되며, 새로운 이벤트 발생 시 규칙 갱신의 용이성을 제공하기 위해 전체 데이터베이스를 재스캔하지 않고, 해당 클래스의 데이터만을 스캔한다. 또한, 예측의 정확도를 유지하기 위해 규칙베이스를 주기적으로 갱신한다. 넷째, 추출된 공간데이터에 대한 패턴은 실시간으로 입력되는 변수들에 의해 예측되며 가시화 도구를 통하여 제공한다. 따라서 전문적 지식이 없이도 사용자에게 예측과 결과의 이해를 용이하도록 한다.

산불 예측 시스템 구현

일반화 기반 분류 예측 시스템에 산불 이력 데이터를 적용하여 산불데이터의 패턴을 도출하고 이를 이용하여 예측을 수행하는 산불 예측 시스템을 구현하였다.

1. 구현 환경

일반화 기반 분류 예측 시스템은 효율적인 GUI 개발을 위해 객체지향 언어인 Visual Basic 6.0을 사용하였으며, 예측의 가시화를 위해 Map Object 2.0을 사용하였다. Windows NT상에서 MS Access 7.0을 사용하여 데이터베이스를 구축하였고 응용 프로그램과 연결하였다. 구현 내용은 산불 이력 데이터와 날씨 데이터에 대한 연관성을 분석하여 패턴을 추출하고 이를 이용하여 관련 기상데이터의 입력에 의한 산불을 예측 수행하는 시스템이다.

2. 일반화 정보 추출을 위한 배경지식

다음은 기상데이터의 기온과 습도에 대한 배경지식 및 산불발생 규모에 따른 일반화를 위한 개념계층을 나타낸다.

표 1은 평균온도(average temperature : AT)에 대한 배경지식으로 2.5℃ 간격으로 구분된 14개의 평균온도이다. 이러한 배경지식은 비공간 속성의 일반화 과정 중 상위수준으로 통합될 때 병합된 속성 값에 따라 일반화된 값으로 매핑하는 기준으로 사용된다. 예를 들어 병합된 평균온도 값이 7.6℃ ~ 10.0℃ 사이에 존재하면 AT8으로 일반화되는 것이다.

그 외에 표 2, 표 3에서의 최고온도(high temperature: HT)와 최저온도(low temperature : LT) 역시 2.5℃ 간격으로 구분되어 각각 14개, 18개로 일반화된 배경지식을 나타낸다. 표 4에서 평균풍속(average wind: AW)은 10㎧ 간격

TABLE 1. Concept hierarchy for average temperature

Class	Range
AT1	-10.0 ~ -7.6
AT2	- 7.5 ~ -5.1
AT3	- 5.0 ~ -2.6
AT4	- 25 ~ -1
AT5	0 ~ 2.5
AT6	2.6 ~ 5.0
AT7	5.1 ~ 7.5
AT8	7.6 ~ 10.0
AT9	10.1 ~ 12.5
AT10	12.6 ~ 15.0
AT11	15.1 ~ 17.5
AT12	17.6 ~ 20.0
AT13	20.1 ~ 22.5
AT14	22.6 ~ 25.0

* AT : Average Temperature(℃)

TABLE 2. Concept hierarchy for high temperature

Class	Range
HT1	0 ~ 2.5
HT2	2.6 ~ 5.0
HT3	5.1 ~ 7.5
HT4	7.6 ~ 10.0
HT5	10.1 ~ 12.5
HT6	12.6 ~ 15.0
HT7	15.1 ~ 17.5
HT8	17.6 ~ 20.0
HT9	20.1 ~ 22.5
HT10	22.6 ~ 25.0
HT11	25.1 ~ 27.5
HT12	27.6 ~ 30.0
HT13	30.1 ~ 32.5
HT14	32.6 ~ 35.0

* HT : High Temperature(℃)

TABLE 3. Concept hierarchy for low temperature

Class	Range
LT1	-20.0 ~ -17.6
LT2	-17.5 ~ -15.1
LT3	-15.0 ~ -12.6
LT4	-12.5 ~ -10.1
LT5	-10.0 ~ -7.6
LT6	-7.5 ~ -5.1
LT7	-5.0 ~ -2.6
LT8	-2.5 ~ -1
LT9	0 ~ 2.5
LT10	2.6 ~ 5.0
LT11	5.1 ~ 7.5
LT12	7.6 ~ 10.0
LT13	10.1 ~ 12.5
LT14	12.6 ~ 15.0
LT15	15.1 ~ 17.5
LT16	17.6 ~ 20.0
LT17	20.1 ~ 22.5
LT18	22.6 ~ 25.0

* LT : Low Temperature (℃)

TABLE 4. Concept hierarchy for average wind

Class	Range
AW1	0 ~ 1.0
AW2	1.1 ~ 2.0
AW3	2.1 ~ 3.0
AW4	3.1 ~ 4.0
AW5	4.1 ~ 5.0
AW6	5.1 ~ 6.0
AW7	6.1 ~ 7.0
AW8	7.1 ~ 8.0
AW9	8.1 ~ 9.0
AW10	9.1 ~ 10.0

* AW : Average Wind (㎧)

으로 10개의 범위로 구분되었으며, 표 5에서는 평균습도(average humidity: AH)를 10% 간격으로 8개의 일반화된 배경지식을 나타냈다.

표 6은 산불의 피해면적에 따른 산불규모를 대형, 중형, 소형 등과 같이 3단계로 구분하여 나타내었다.

TABLE 5. Concept hierarchy for average humidity

Class	Range
AH1	10.1 ~ 20.0
AH2	20.1 ~ 30.0
AH3	30.1 ~ 40.0
AH4	40.1 ~ 50.0
AH5	50.1 ~ 60.0
AH6	60.1 ~ 70.0
AH7	70.1 ~ 80.0
AH8	80.1 ~ 90.0

* AH : Average Humidity (%)

TABLE 6. Classification for the rage of forest fire

산불 분류	피해규모(ha)
대 형	30 이상
중 형	1 ~ 30
소 형	1 미만

훈련 데이터 셋은 지금까지 나열된 기상자료의 배경지식과 산불규모에 따른 분류에 의해 일반화된다.

3. 사용자 인터페이스 메뉴 구성 및 수행

일반화 기반 분류 예측 시스템은 크게 세 부분으로 나누어 설명할 수 있다. 첫째로 데이터베이스를 연결하고 관련 데이터 셋을 메모리에 적재하는 부분, 둘째로 적재된 데이터 셋에 대해 일반화 및 분류기법을 수행하는 부분, 셋째로 추출된 패턴을 이용하여 예측을 수행하고 결과를 가시화하는 부분이다.

구현된 시스템의 초기화면은 그림 4와 같다. 패턴을 추출하고자 하는 이력데이터베이스를 선택하도록 구현되었다. 시스템에는 사전에 정의되어 있는 데이터베이스를 적재하거나 사용자 정의에 의한 데이터베이스를 선택하여 연결할 수 있다. 그림 5는 메뉴를 이용하여 데이터 셋을 선택하는 화면이다. 데이터베이스가 연결되기 이전에는 데이터 셋을 적재할 수 없다.

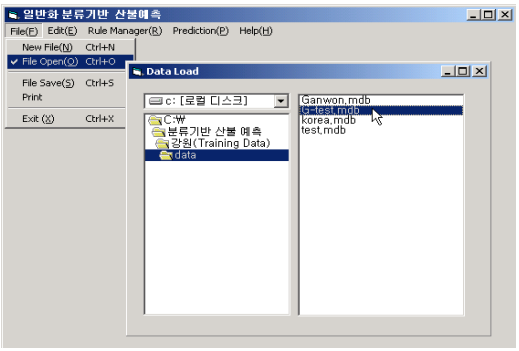

FIGURE 4. Main interface of prediction system

FIGURE 5. Data set

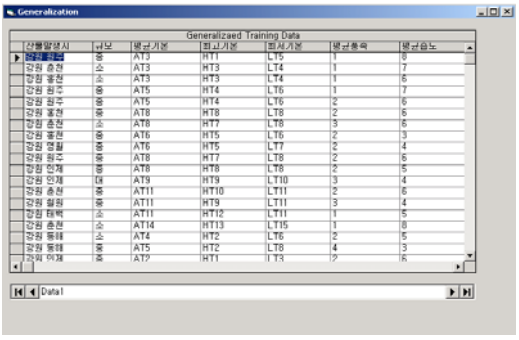

FIGURE 6. The function of generalization

그림 5는 선택한 데이터 셋을 적재시킨 결과 화면이다. 그리고 그림 6은 불러온 데이터 셋에 대해 미리 정의된 배경지식을 이용하여 일반화 한 결과를 보여주고 있다. 먼저 관련 데이터를 불러들인 화면과, 일반화를 적용한 후의 화면을 보여주고 있다.

일반화 과정을 거친 데이터 셋은 분류기법을 통해 지식을 추출하게 된다. 이 논문에서는 산불이력 데이터의 발생 규모에 따른 대형, 중형, 소형산불에 대한 패턴을 추출하였다. 다음은 1996년~2000년 간의 강원도지역 산불 발생 데이터로 분류기법을 수행한 결과 다음과 같은 지식이 도출되었다. 지식베이스는 도출된 규칙과 count값을 저장한다.

"AT9, HT9, LT10, AW3, AH3 일 경우 대형 산불이 발생한다 "

"AT8, HT5, LT8, AW2, AH4 일 경우 중형 산불이 발생한다"

"AT8, HT8, LT9, AW1, AH5 일 경우 소형 산불이 발생한다"

그림 7은 일반화 기반 분류 예측기의 메뉴에서 prediction을 선택한 후의 첫 화면으로 우리나라의 각 시·군단위로 나뉜 shape 파일이 로드된다. 그리고 예측을 원하는 지역을 선택할 수 있도록 하였다.

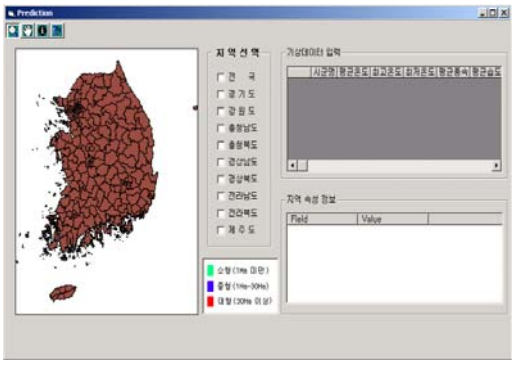

FIGURE 7. The initial image of prediction system

그림 8은 강원도 지역을 선택한 후에 강원도 임의지역의 기상자료를 입력 한 후 예측수행 결과 화면이다.

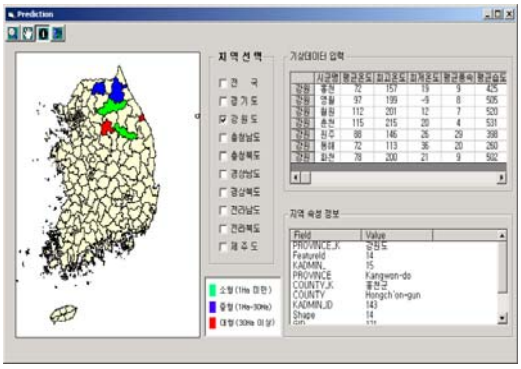

FIGURE 8. Prediction of forest fire

결과 및 고찰

이 연구를 통해 접근한 일반화 기반 분류 예측 시스템의 궁극적인 목표는 이력 데이터 베이스로부터 새롭고 유용한 지식을 발견하고 이를 이용하여 실시간으로 예측을 수행하는 데 있다. 일반화 기반 분류기법으로 구축된 예측 모델의 실효성 검증을 위해서 모델에 포함된 과거 기상 데이터를 입력하여 산불의 발생 위험 예측을 수행한 후, 실제의 산불 발생 기록과 비교한다. 또한, 3년간(1993년~1995년) 산불이 가장 많이 발생하는 1월에서 5월까지의 강원지역 기상데이터를 입력하여 산불 예측을 수행하였다. 그리고 사례의 수 변화에 따른 예측성능과 학습한 패턴의 양에 따라 미치는 영향에 대하여 성능 평가를 수행하였다.

그림 9는 예측 모델의 테스트 사례 개수에 대한 예측정확도를 측정하였다. 데이터는 1996년~2000년 간의 강원도 산불 발생 사례를 임의로 추출하여 적용하였다. 수행 결과 약간의 변동을 보였지만, 대체적으로 안정적인 예측 정확도를 유지하며 테스트 데이터 수에 대한 예측률에는 커다란 영향을 미치지 않았다.

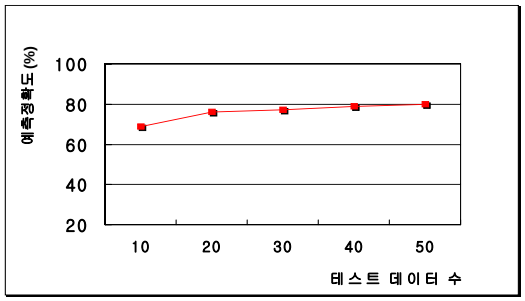

FIGURE 9. Prediction accuracy of various test data set

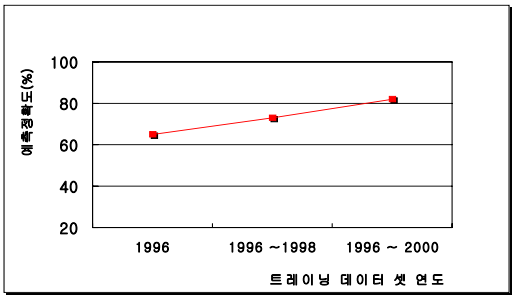

FIGURE 10. Prediction accuracy of various training data set

그림 10은 지식베이스를 생성 시 사용하는 훈련데이터의 양에 따른 예측 정확도를 측정한 결과이다. 이는 표본의 복잡도를 측정하여, 유효한 일반화 성능을 보장하기 위해 필요한 평가로써 먼저 1996년도의 사례만을 기반으로 예측을 수행하였을 경우는 65%의 비교적 낮은 예측률을 보였으나, 1996년 ~ 2000년의 사례를 기반으로 예측을 수행하였을 경우 82%의 예측정확도를 보임으로써, 점차 지식베이스에 훈련되는 사례의 수가 커질수록 예측정확도가 증가함을 볼 수 있었다.

결 론

데이터마이닝의 범위가 확대되고 작업에 대한 요구사항이 다양해지고 있으며 그 기법 또한 다양해지고 있다. 이러한 기법들은 각 작업의 특성에 알맞게 사용되어 효과적인 의사결정 뿐 아니라 사업의 이익 증대 또한 높여야할 것이다. 기존의 연구에서는 특정시점 혹은 특정기간의 데이터 셋을 기반으로 모델을 구축하여 새로운 지식을 추출하였고, 이러한 지식들은 각 분야의 특성에 맞게 의사결정에 사용되어 왔다. 그러나 이러한 규칙들은 사용되어진 데이터 셋에 대하여 일정 시기가 지나면 유효성을 떨어져, 새로운 데이터 셋에 대한 또 다른 모델을 구축해야만 하는 실정이다.

따라서 이 연구에서는 반복적으로 발생하는 실세계의 데이터로부터 지식을 탐사하는 일반화기반 분류기법을 제시하였고, 이벤트가 발생함에 따라 점진적으로 규칙들을 갱신함으로써 고정된 규칙이 아닌 보다 능동적인 규칙을 유지하며, 이를 이용하여 미래의 이벤트를 효과적으로 예측할 수 있는 모델을 설계 및 구현하였다. 즉 구축한 모델을 동적으로 갱신하며 규칙에 대한 재사용성을 추가함으로 효율적인 예측 뿐 아니라 모델 재구축에 대한 비용도 절감할 수 있도록 하였다. 또한, 추출된 규칙들은 시각화 도구로 제공하여 최종사용자에게 보다 쉽게 이해하고 이용할 수 있도록 하였다. 제시된 일반화 기반 분류기법은 산불데이터에 응용하여, 기상 데이터와 산불발생 사이의 관련성을 분석하고 이를 통해 구축된 예측 모델로 효율적인 예측을 수행하였다.

이 연구에서 제시하고 있는 기법은 반복적으로 발생하는 자연재해에 대한 분석 및 예측, 에너지의 수요량 예측, 원유탐사에 있어서 지층의 변화를 분석 등 실생활의 중요한 부분들에서 다양하게 활용할 수 있다.

향후 연구될 방향으로는 이벤트에 따라 갱신되는 규칙들에 대한 이력관리에 대한 연구가 필요하며, 예측이 수행되는 과정에서 규칙이 중복되는 경우에 대한 연구가 요구된다. 또한, 패턴에 대한 예측 수행 뿐 아니라 적용 데이터에 대한 관리 시스템으로 확장이 필요하다. KAGIS

참고문헌

나민영. 1998. 데이터 마이닝. 2000년대의 DB 응용기술 9월호 특집.

이시영, 한상열, 안상현, 김진열, 오정수. 2001. 산불발생인자의 지역별 유형화. 산불예측 및 생태계 보전 심포지엄. 3-17쪽.

조명희, 조윤원, 백승렬, 오정수. 2001. GIS를 이용한 산불 현황정보 검색시스템 개발. 대한원격탐사학회 춘계학술대회 발표집 3:49-55.

최 관, 한상열. 1996. 기상자료를 이용한 산불 발생확률모형의 개발. 한국임학회지 85: 15-23.

최종후, 한상태, 강현철, 김은석. 1998. AnswerTree를 이용한 데이터마이닝 의사결정 나무 분석. SPSS 아카데미. 서울. 154쪽.

Cortes,C., H. Drucker, D. Hoover and V. Vapnik. 1995. Capacity and complexity control in predicting the spread between borrowing and lending interest rates. Proceedings of the 1st Int. Conf. KDD'95. pp.51-56.

Han, J. and Y. Fu. 1996. Exploration of the power of Attribute-oriented Induction in data mining. In: U.M. Fayyad et al.(ed.). Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, pp.399-421.

Han, J., Y. Fu, W. Wang, J. Chiang, W. Gong, K. Koperski, D. Li, Y. Lu, A. Rajan, N. Stefanovic, B. Xia and O.R. Zaiane. 1996a. DB Miner: A system for mining knowledge in large relational databases. KDD. pp.250-255.

Han, J., M.S. Chen and P.S. Yu. 1996b. Data mining: An overview from database perspective. IEEE TKDE 8(6):866-883.

Kamb, M., L. Winstone, W. Gonh, S. Cheng and J. Han. 1997. Generalization and decision tree induction: efficient classification in data mining. Proceedings of RIDE'97. pp.111-120.

Lange, R. 1996. An empirical test of the weighted effect approach to generalized prediction using neural nets. Proceedings of the 2nd Int. Conf. KDD'96. pp.183-188.

Wang, W. 1999. Predictive modeling based on classification and pattern matching method. Beijing Polytechnic University Papers. pp.36-50.