분류기법을 이용한 예측 시스템 설계

김대진[0] 이준욱 류근호
충북대학교 데이터베이스연구실
{djk, junux, khryu}@dblab.chungbuk.ac.kr

Design of Prediction System based on Classification Method

Dea Jin Kim[0] , Joon Wook Lee , Keun Ho Ryu
Database Laboratory, Chungbuk National University

요 약

정보화시대에 들어서면서 나날이 급증하는 데이터에 대한 재가용성을 위한 많은 연구가 이루어지고 있다. 이러한 연구들은 의사결정지원, 예측, 추정 등의 분야에서 적용되고 있으나, 실생활에 활발히 적용되기까지 앞으로 많은 연구 및 개발이 요구된다. 이 논문에서는 수집된 데이터로부터 패턴을 추출하여 예측결과를 제공할 수 있는 시스템 모델과 모델에 적합한 점진적 규칙갱신 알고리즘을 제안하였다. 제안하는 예측 모델의 특징은 새로 입력되는 정보에 대한 반복 학습시 수치데이터에 대한 평균근사치 할당방법을 적용하여 규칙갱신을 용이하게 하였으며 각 클래스의 수치데이터에 대한 분류를 용이하도록 하였다.

1. 서론

데이터 자동화 수집도구는 많은 데이터를 생산하였으며 수집된 데이터로부터 유용한 정보를 추출하고자 하는 필요성은 데이터마이닝 또는 통계분석 기법들을 더욱 세련되게 발전시켰다.

데이터마이닝의 여러 가지 기법들은 크게 예측적 관점과 데이터 셋을 설명하는 묘사적 관점으로 볼 수 있다. 예측적 관점의 데이터마이닝 기법들에는 분류 기법과 값 예측 등이 이에 속하며, 묘사적 관점의 기법들에는 연관 규칙 탐사와 클러스터링 기법 등이 속한다[1].

입력데이터를 분석하여 각 클래스에 대한 정확한 표현이나 모델을 개발하기 위한 분류기법에 관한 연구는 이미 통계(Statistics)[2], 신경망(Neural network)[3], 결정트리(Decision tree)와 같은 기법을 통해 연구되어 왔다[4,5].

기존의 통계적 기법은 가설 및 검증과정을 통해 비교적 신뢰성이 있는 예측 모델을 만들어 내지만 많은 데이터를 기반으로 사용자의 요구사항을 빠르게 응답하기에 적합하지 못하며, 데이터마이닝의 신경망 분석기법은 매우 복잡한 구조를 가진 데이터들 사이의 관계나 패턴을 찾아내는 유연한 비선형 모형의 하나로, 정확한 예측모델을 세우기 위해서는 모델설정에서 고려해야할 부분이 많기 때문에 결과를 얻어내기까지 많은 어려움이 따르며, 설명력이 부족하다는 특징을 가진다.

따라서, 이 논문은 과거에 수집된 데이터를 분석하여 미래의 상황을 예측하여 의사결정에 반영될 수 있는 예측기법 제안과 이를 효율적으로 구현할 수 있도록 설계한다. 여기 에서는 분류기법에서 주로 사용되는 결정트리 기법을[6,7] 응용하여 사례기반 규칙패턴을 생성한다 이 방법은 기존의 예측기법에서 힘들었던 새로 입력되는 정보의 반복학습에 대한 규칙갱신 기법 및 수치데이터의 보다 정확한 분류를 위한 분류기법을 제안하고 이에 적합한 시스템을 설계한다. 이 논문에서 제안하는 시스템은 기존 데이터베이스에 구축된 정보를 이용하여 규칙을 생성하여 향후 전망이나 기대 예측치 관찰할 수 있고, 향후 규칙갱신을 효율적으로 수행하는데 목적이 있다. 제안하는 예측시스템의 특징은 다음과 같다.

- 새로 입력되는 정보에 대한 규칙갱신을 용이하게 하기 위해 전체 데이터베이스를 재 스캔하지 않고 해당 클래스의 데이터에 대한 데이터스캔을 수행한다.

- 수치 데이터에 대한 예측결과를 정확하게 하기 위해 기존에 결정트리기법 에서 수행했던 각 항목의 일정한 범위를 갖는 도메인 값으로의 일반화 대신 평균 근사치 할당 방법을 적용하기 때문에 미묘한 항목 값에 따른 수치데이터에 대한 분류정확도를 기대할 수 있다.

이 논문의 구성은 다음과 같다. 2장에서는 분류 기반 예측 시스템을 설계하고 그 특징 및 주요 모듈을 설명을 한다. 3장에서는 시스템 설계에 쓰인 세부 모듈에 이용되는 알고리즘을 기술한다. 마지막 4장에서는 본 논문에

대한 결론 및 향후 연구 계획에 대해 설명한다.

2. 분류 기반 예측 시스템

2.1 시스템 구조

공간 데이터에 대한 분류작업을 수행할 시스템의 구조는 다음 그림 1과 같다. 이력 데이터 셋에 대한 패턴을 추출하는 분류모듈과 테스트 데이터가 입력되어, 지식저장소를 검색 후 예측 결과를 제공하는 예측 모듈로 구성된다.

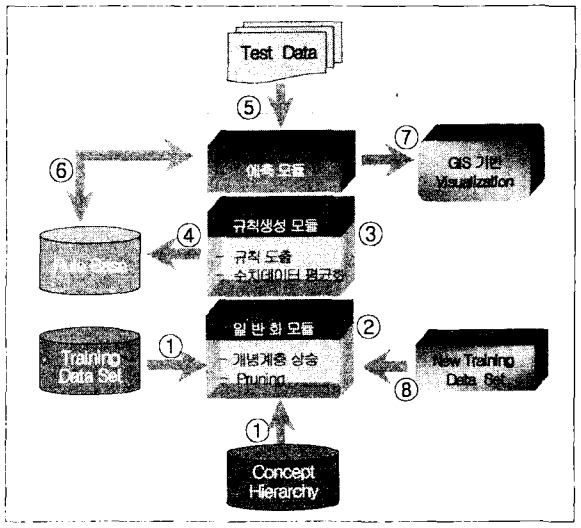

그림 1. 분류기반 예측 시스템 구조

규칙 생성 과정으로 분류모듈은 일반화 모듈과 규칙 생성 모듈로 구성되며, 예측 모듈을 이용한 테스트 데이터에 대한 예측 수행 과정, 지식 저장소 갱신 과정을 가진다. 각 모듈의 처리과정은 다음과 같다.

2.2.1 규칙 생성 과정

규칙 생성 과정은 그림 1의 ①~④에서 보여준다.

① 데이터 수집

작업에 관련된 데이터를 운영 데이터베이스로부터 추출한 트레이닝 셋과 사용자 및 전문가에 의해 정의된 개념계층을 입력한다.

② 일반화 모듈

데이터 셋에 개념계층을 적용하여 일반화 데이터 셋을 생성한다. 이때 개념계층의 상승정도의 수준은 사용자가 결정한다. 개념계층을 상승시킨 후 중복된 튜플은 카운팅을 하고 제거시킨다.

③ 규칙 생성 모듈

일반화 과정을 거친 데이터 셋은 사용자의 임의에 따라 m개의 클래스로 나뉘며, 엔트로피를 이용하여 각 클래스별 사례집합의 후보 어트리뷰트에 대한 정보값을 계산한다. 최대 정보를 가진 후보 어트리뷰트를 선택하여 결정트리를 생성한다.

④ 클래스별 추출된 각 패턴들은 어트리뷰트별 도메인값과 평균값을 포함하여 규칙 저장소에 저장한다.

2.2.2 예측 수행 과정

예측 수행 과정은 그림 1의 ⑤~⑦에서 보여주며, 사용자의 입력이 있게 되면 예측모듈에 의해 비수치 데이터에 대한 일반화 작업이 이루어진다. 일반화 된 데이터는 속성값에 따른 할당된 클래스값을 얻게 된다. 예측모델은 지식저장소의 해당 클래스의 패턴과 비교 및 검색을 수행한다. 테스트 데이터에 한 결과 예측치를 가시화 작업을 통해 보여준다.

2.2.3 지식 저장소 갱신 과정

지식 저장소 갱신 과정은 ⑧과 ②~④의 과정이며, 새로운 트레이닝 데이터 셋이 발생할 경우 일반화 모듈과 규칙 생성 모듈의 과정을 거쳐 패턴을 추출하게 된다. 지식 저장소에 저장되어있는 해당하는 클래스의 패턴을 갱신한다.

3. 주요 알고리즘

3.1 일반화 및 규칙생성

다음의 알고리즘 1은 저수준 데이터 셋의 입력에 대한 개념상승 하는 과정을 나타낸다.

먼저 입력되는 데이터 셋의 각 튜플들을 LT_j 라 하고, 각 튜플들의 속성을 $LT_i.A_j$ 라고 할 때, 먼저 정보량을 계산하여 각 속성들의 우선순위 결정하고, 정보값이 작은 속성값에 대한 제거작업을 SortByInfoGain 함수에서 수행하여 정렬과정을 거친다. GetDigit 함수에서는 문자 데이터에 대한 수치화를 통한 일반화 과정을 수행한다. 각각 정보량에 의한 각 속성값에 대한 평균치를 계산하기 위해 문자 및 수치 데이터에 대한 분류 및 수치화를 거쳐 각 클래스에 누적된 빈발 카운트 값 $GT_{classLevel}.CNT$을 이용한다. 결과로 나오는 일반화된 패턴들은 규칙베이스에 저장된다.

알고리즘 1. 일반화 및 규칙생성 알고리즘

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))
```

```
      GT_classLevel.A_i += LT_i.A_j;
    else
      GT_classLevel.A_i += GetDigit(A_i);
  }
  GT_classLevel.CNT++;
}
for(i = 0; i < GetMaxClass(LT); i++)
{
  for(j = 0; j < GetMaxAttr(GT); j++)
    GT_i.A_j = LT_i.A_j/GT_i.CNT;
}
```

3.2 분류 및 예측

알고리즘 2는 새로운 샘플 테스트 데이터를 이용하여 클래스를 발견하는 과정을 나타내는 알고리즘이다. 이 알고리즘은 규칙 인스턴스에 대한 사례기반알고리즘으로서, 기존 트리 기반 알고리즘과 유사하게 먼저 정보량이 많은 순서로 분류되지만, 각 속성에 대한 사용자의 임계치 값을 이용하게 된다.

먼저 샘플 테스트 데이터의 어트리뷰트 순서에 따라 속성 값의 순서를 결정하여 가능한(feasible) 클래스들을 찾는다. 이때 해당 클래스의 패턴과 비교 시, 평균 근사치로 할당하여 보다 정확한 결과를 제공한다. 그 다음 정보량이 낮은 어트리뷰트에 대해서는 수집된 클래스에 대한 제거 작업을 수행하게 되며 이때 사용자가 명시한 임계치 범위를 이용한다.

이 알고리즘의 출력 값으로 복수의 클래스가 될 수 있으며 이때 사용자가 명시한 클래스 레벨이 출력된다.

알고리즘 2. 분류 및 예측 알고리즘

```
Algorithm : 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(TD_i))
    TD_i = GetDigit(A_i);
  for(j = 0; j < GetMaxClass(GT);j++)
  {
    if(j > 0)
    {
      if(TD_j > GT_j + ε_j && TD_j < GT_j - ε_j)
        DelClassLevel(j);
    }
    else if(TD_j < GT_j + ε_j && TD_j > GT_j - ε_j)
    {
      SetClassLevel(j);
    }
  }
  return GetClassLevel(usrDefinedLevel);
}
```

향후 발생될 새로운 트레이닝 데이터(NT)에 대한 알고리즘은 알고리즘 1 과 유사하며 규칙갱신은 해당 클래스에 대하여 다음과 같은 식을 이용하여 해당 클래스 레벨의 속성 값을 갱신할 수 있다.

$$GT_i.A_j = (LT_i.A_j \times GT_i.CNT + NT_i) / GT_i.CNT + 1$$

5. 결론

이 연구에서는 데이터마이닝의 예측적 기법을 응용한 예측 모델을 제안하였다. 제안하는 예측 모델은 기존에 수집된 이력데이터를 이용하여 규칙을 생성하며, 새로 입력되는 테스트 데이터를 통하여 미래에 발생 가능한 상황을 예측하도록 하였다. 또한 새로 입력되는 사실 값 즉, 새로운 트레이닝 데이터는 규칙베이스에 데이터베이스 전체의 재 스캔 없이 갱신 가능하도록 하였다. 이러한 작업이 수행 가능하도록 점진적 규칙갱신 알고리즘을 제시하였다. 이 알고리즘은 수치 및 문자 데이터에 대한 일반화 과정을 분리하여 규칙베이스를 점진적으로 갱신 가능하도록 수치데이터를 평균 근사치로 일반화하여 미묘한 클래스의 속성 값에 대하여 효율적인 갱신 작업이 수행하도록 하였다.

현재 우리는 논문에서 제시한 내용을 근거로 예측시스템의 구현과 공간 이력 데이터를 적용함으로써 시스템의 테스트 및 이에 대한 성능 평가를 진행하고 있다.

참고문헌

[1] Wei Wang, Predictive Modeling Based On Classification and Pattern matching methods, B.Sc. Beijing Polytechnic University, 1992

[2] C.Cortes, H. Drucker, D. Hoover, and V. Vapnik. Capacity and complexity control in predicting the spread between borrowing and lending interest rates. In Proc. 1st Int. Conf. KDD'95, pages 51-56, Montreal, Canada, August 1995

[3] Rense Lange. An empirical test of the weighted effect approach to generalized prediction using neural nets. In Proc, 2nd Int, Conf. KDD'96, pages 183-188, Portland, Oregon, August 1996

[4] C.Apte and S. Hong. Predicting equity returns from securities data and mininal rule generation. In Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, 1995

[5] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MITPress, 1996

[6] J.R. Quinlan. Induction of decision trees. Machine Learning, 1:81 - 106, 1986

[7] J.R. Quinlan. C4.5: Programs for Machine Lerning. Morgan Kaufmann. 1993