韓國經營科學會誌
第29卷 第4號
2004年 12月

연속발생 데이터를 위한 실시간 데이터 마이닝 기법*

김진화** · 민진영***

A Real-Time Data Mining for Stream Data Sets*

Jinhwa Kim** · Jin Young Min***

Abstract

A stream data is a data set that is accumulated to the data storage from a data source over time continuously. The size of this data set, in many cases, becomes increasingly large over time. To mine information from this massive data, it takes much resource such as storage, memory and time. These unique characteristics of the stream data make it difficult and expensive to use this large size data accumulated over time. Otherwise, if we use only recent or part of a whole data to mine information or pattern, there can be loss of information, which may be useful. To avoid this problem, we suggest a method that efficiently accumulates information, in the form of rule sets, over time. It takes much smaller storage compared to traditional mining methods. These accumulated rule sets are used as prediction models in the future. Based on theories of ensemble approaches, combination of many prediction models, in the form of systematically merged rule sets in this study, is better than one prediction model in performance. This study uses a customer data set that predicts buying power of customers based on their information. This study tests the performance of the suggested method with the data set alone with general prediction methods and compares performances of them.

Keyword : Stream Data Sets, Real-Time Data Mining, Merging Rules, Weights on Rules, Predictions

논문접수일 : 2004년 7월 2일 논문게재확정일 : 2004년 9월 2일
* 이 연구는 서강대학교 학술 연구비 지원에 의하여 연구되었음.
** 서강대학교 경영학과
*** 숙명대학교 경영학과

1. 서 론

데이터 마이닝은 방대한 데이터에서 의미 있는 정보를 추출해 내는 데에 유용하게 쓰인다. 많은 데이터 마이닝 방법들이 이미 저장되어 변하지 않는 데이터들에서 정보를 뽑아내는 것에 기초를 두었다. 이러한 방법들은 데이터 마이닝을 시작하기 전에 필요한 데이터가 모두 갖춰져 있어야만 하고 마이닝이 진행되는 동안 시간이 얼마나 지났는지 여부와 상관없이 이 데이터가 변하지 않는다는 특징을 가지고 있다. 그러나 데이터가 시간이 지남에 따라 계속 추가 될 뿐만 아니라 시간의 흐름에 따라 그 성격이 변할 수도 있다면 시간의 순서에 따른 데이터의 변화에 초점을 더 두어야 할 것이다. 따라서 변하지 않는 데이터를 대상으로 데이터 마이닝을 할 때에는 그 데이터에 가장 잘 맞는 방법을 선택하면 이 방법이 분석과정 내내 변할 필요가 없겠지만, 데이터가 시간의 흐름에 따라 변한다면 그것을 분석하는 방법 또한 시간의 흐름에 따라 변화되어야 할 것이다. 예를 들어 어느 회사의 연도별 매출 데이터가 있고 이 데이터들을 통해 과거의 매출액 증가에 영향을 준 어떠한 패턴을 마이닝한다고 가정하자. 이 경우 분석해야 하는 데이터들은 이미 갖추어져 있고, 과거의 시점을 대상으로 하므로 이렇게 갖추어진 데이터는 변화될 필요가 없다. 그러나 여기서 한층 더 나아가서 미래에 연도별 매출액 데이터가 생성될 때 마다 이 데이터를 추가한다고 가정해 보자. 이 경우에 데이터는 이미 누적된 데이터에 계속적으로 추가가 될 것이다. 또한 미래의 매출액을 예측하는 데에 50년 전과 같이 시간적으로 멀리 떨어진 데이터는 그 영향력이 감소하게 될 수도 있고 혹은 데이터의 변화가 주기적으로 일어난다면 영향력이 감소하다가 어느 시점에서는 다시 그 영향력이 증가 될 수 있다는 것을 고려할 필요가 있다. 이렇게 시간이 흐름에 따라 데이터가 추가되고, 과거의 데이터의 중요성이 변하는 예는 어디에나 있다. 또 다른 예로 주식시장에서 주식 가격의 변화에 미치는 요인, 주식 가격의 예측, 시간의 흐름에 따른 소비자들의 구매 변화 등도 이러한 범주에서 분석될 수 있을 것이다. 또한 이렇게 시간이 지남에 따라 계속 증가하는 데이터를 관리하기 위해서는 물리적인 저장 공간 또한 계속 늘어나야 한다. 그러나 방대한 양의 데이터가 계속 늘어나는 경우 거기에 맞추어 계속 저장 공간을 늘리는 것은 물리적인 제한이 따르거나 혹은 엄청난 비용이 예상된다. 이러한 시간의 흐름에 따른 변화와 저장공간, 자원 소요의 문제를 고려하면 연속 발생 데이터에 알맞은 데이터 마이닝 방법이 필요하다. 본 연구는 시간의 흐름에 따라 변하는 데이터를 대상으로 보다 효율적으로 데이터 마이닝을 이용하여 예측하는 방법을 제안 한다.

2. 이론적 배경

데이터 속에 숨겨져 있는 흥미로운 패턴들을 찾아내는 것은 데이터 마이닝에서 중요한 부분을 차지한다. 이 패턴이라는 것은 데이터 베이스 안에서 발견되는 조합 혹은 분류 모델이나 순차적 경향들로서 데이터 마이닝의 기본이 되는 것이다[13]. 이렇게 패턴을 마이닝하는 것은 자료가 어떻게 조합되는지 혹은 어떻게 분류될 수 있는지 밝히고, 더 나아가서는 앞으로 들어오는 자료들을 기존에 생성된 패턴들을 사용해서 효과적으로 예측하려는 목적을 가지고 있다. 이러한 패턴들을 찾아내기 위해 여러 가지 방법들이 제시되었는데 대표적인 것으로 빈번하게 발생하는 패턴을 찾아내는 Apriori 알고리즘이 있다[1, 2]. 이후로 많은 패턴 마이닝에 대한 연구가 있었으며 그 중 많은 연구는 Apriori 알고리즘을 기반으로 하고 있다[9, 15]. Apriori의 패턴이 길어지면 시간과 효율성의 문제가 발생할 수 있다는 단점을 극복하기 위해 Apriori와는 달리 패턴에 대한 정보를 가지고 있다가 이것으로부터 패턴 후보집합의 산출 없이 패턴을 마이닝하는 방법이 있다[3, 14, 17, 19]. 이 중 하나가 분할-정복(divide-and-conquer) 기법을 사용하는 FP-트리(빈번한 패턴 트리)기법이다. 이 방법은 먼저 데이터 베이스를

각각의 항목과 그 항목의 발생빈도수로 구성되는 FP-트리로 구성한다. 그리고는 k-항목 집합마다 데이터베이스를 스캔하여 빈번한 항목집합인지 아닌지 확인하는 대신에, 이미 생성된 FP-트리의 항목과 발생빈도수를 따라가면서 빈번한 항목집합을 찾아낸다.

패턴을 마이닝하는 방법은 자료가 고정되어 있지 않고 시간이 흐름에 따라 변하는 경우 단순히 빈번하게 발생하는 패턴을 찾아내는 것보다 매우 복잡하다. 유한하고 통계적으로 고정되어 있는 데이터와는 다르게 이렇게 시간의 흐름에 따라 변하는 연속적이고 잠재적으로 무한히 발생하는 특징을 가지고 있는 데이터를 연속 발생 데이터라고 한다. 이런 데이터의 예로는 네트웍 traffic 분석, 전화 기록, 소비자의 구매 기록, 웹 클릭 연속 발생 마이닝, 주가 변동의 동적인 기록 등이 있다. 현대에는 많은 데이터가 이러한 연속 발생 데이터의 범주에 속한다고 할 수 있다. 연속 발생 데이터의 정의에 따르면 빈번히 나오던 패턴이 자료가 추가됨에 따라 빈번한 패턴이 아니게 될 수 되고, 빈번히 나오지 않던 패턴이 자료가 추가됨에 따라 빈번하게 나오는 패턴으로 분류될 수 있다. 연속 발생 데이터의 마이닝에서는 패턴의 추출 자체가 문제가 아니라 빈번한 패턴들의 변화를 어떻게 다룰 지가 더욱 중요한 문제가 된다는 것이다. 이렇게 자료가 고정되어 있지 않고 계속 증가하는 경우에 마이닝하는 대상 자료의 변화를 다루는 마이닝 방법이 점진적 데이터 마이닝이라 한다.

점진적 데이터 마이닝의 기본개념은 데이터베이스에 대한 정보를 R이라고 이름 붙여진 형태로 유지하고 있다가 새 데이터가 들어오면 이 데이터에서 r을 뽑아낸 후 이것을 현재 가지고 있는 R과 합쳐서 새로운 R을 만들어내는 것이다. 즉 $R \cup r$을 통해 R를 만들어내는 것이다[13]. 연속 발생 데이터를 분류하거나 예측하기 위해서도 패턴의 마이닝 자체뿐 아니라 패턴들을 추가된 데이터 set에 맞추어 변화시켜야만 이 패턴들을 통해서 효과적인 분류, 혹은 예측을 할 수 있다[4, 7]. 점진적 데이터 마이닝을 통한 연속 발생 데이터의 마이닝은 고정되고 변하지 않는 데이터를 대상으로 마이닝하는 것보다 더 복잡하고 정교한 과정을 요구한다. 패턴을 추출하는 알고리즘은 데이터의 변화를 수용하여 빈번한 패턴을 찾아낼 수 있게끔 변화되어야 한다. 또한 과거에는 데이터의 양이 적었기 때문에 overfitting[1] 문제가 발생되고는 했지만 현대에는 너무 많은 양의 데이터로 underfitting[2] 문제가 생길 수 있다[6]. 점진적 데이터 마이닝은 이러한 점들을 모두 고려해야 한다.

시간의 흐름에 따라 방대한 양의 데이터가 들어오는 연속 발생 데이터 마이닝의 특성을 고려했을 때 연속 발생 데이터의 마이닝이 기존의 통계적으로 고정된 자료를 대상으로 마이닝하는 방법과는 달라야 하는 중요한 이유중의 하나로 저장 공간과 메모리의 문제를 들 수 있다. 통계적으로 고정된 데이터를 분석하는 방법을 사용하면 분석 시점까지의 모든 데이터가 필요하기 때문에 들어오는 데이터를 모두 저장해 놓아야 할 것이다. 그러나 연속 발생 데이터는 잠재적으로 무한한 데이터가 들어오기 때문에 이러한 데이터를 모두 저장한다면 많은 저장 공간이 필요하며 이러한 저장 공간이 다 차고도 연속 발생 데이터가 계속 들어온다면 결국에는 계속해서 저장공간을 늘려주어야만 한다. 또한 한번에 메모리에서 읽을 수 있는 데이터의 양이 대부분 1GB 이하로 고정되어 있으므로 매우 많은 양의 데이터에 대해 분석을 시행하려면 상당한 시간이 소요될 것이다. 이러한 문제 때문에 대부분의 기존 알고리즘이 과거로부터 누적된 많은 데이터를 모두 마이닝에 이용하는 것이 불가능하다[12]. 따라서 방대한 데이터를 저장하는데 필요한 공간상의 문제를 해결하기 위한 연구들이 진행되고 있다[11]. 비록

1) 데이터의 양이 작기 때문에 학습된 모델이 전체적인 데이터의 분포에는 보이지 않음에도 불구하고 실험 데이터의 어떤 특정한 이상치와 결합되어 있는 문제

2) 데이터의 양이 너무나 방대해져서 가능한 데이터중 상당 부분이 사용되지 않게 되어 발생하는 문제

물리적인 저장공간 문제가 해결된다고 하더라도 이러한 방대한 데이터를 일관성 있고 신뢰성 있게 저장하는 것 자체가 커다란 도전이 된다. 또 일관성과 신뢰성을 가지고 저장을 하였다고 하더라도, 데이터를 마이닝하는 방법에 맞도록 적당한 형태로 한 곳에 모으는 것이 불가능할 수도 있다[16]. 이러한 문제들 때문에 전체 데이터가 아니라 데이터에 대한 요약 또는 필요한 정보만을 가지고 있어야 할 필요가 있다. 그러나 데이터에 대한 요약 또는 필요한 정보만을 가지고 있는 것은 데이터의 전부를 사용하는 일괄처리 방법이 낼 수 있는 효과보다 못한 결과를 가져오거나, 중요한 데이터를 간과할 염려가 있다. 이러한 염려들 때문에 전체 데이터 중에서 어떠한 것을 분석대상으로 삼을 것인지에 관한 다양한 연구가 있었다[7, 8].

연속 발생 데이터의 특징인 데이터의 방대함은 이러한 방대한 데이터를 어떻게 효율적으로 마이닝할 것인지, 데이터의 변화를 어떻게 다룰 것인지에 대한 다양한 연구의 필요성을 제기하였다. Ganti 등은 전체 데이터에서 고정된 크기의 샘플만을 사용하여 트리를 구성하여 각 트리 수준에서마다 전체 데이터를 모두 사용하는 기존의 방법보다 빠르게 일단 트리를 구성한 후 전체 데이터를 스캔하여 만들어진 트리를 정제하는 방법을 연구하였다[7]. Domingos와 Hulten은 데이터 set에 대해 하나의 통로로 의사결정 트리를 구성하는 문제를 연구하여 왔다. 이 연구에서 이들은 VFDT(Very Fast Decision 트리 learner)와 CVFDT(Concept-adapting Very Fast Decision 트리 learner)라는 의사 결정 트리를 만들었다[6, 16]. VFDT는 의사결정 트리 알고리즘으로서 노드 결정시 데이터의 일부만 사용하여도 전체 데이터를 사용하는 것과 비교해 큰 손실이 없고 특히 데이터가 추가되는 경우 트리를 구성하는데 필요한 메모리와 시간을 일정하게 유지하면서 데이터를 계속 저장할 필요 없이 데이터가 들어오는 대로 노드를 구성하여 의사결정 트리를 갱신할 수 있기 때문에 연속 발생 데이터를 통해 트리를 구성하는데 매우 유용한 방법이 될 수 있다. 더 나아가서 들어오는 데이터의 성격이 계속적으로 변하는 경우 거기에 맞는 대체 트리를 생성하다가 오래된 트리의 일부가 변화하는 데이터에 적합하지 않고 새로운 트리가 더욱 정확하게 데이터에 들어맞는다고 판단될 때 대체 트리로 오래된 트리를 대신하는 CVFDT가 제안되었다. Giannella 등은 연속 발생 데이터에서 빈번한 패턴을 마이닝하기 위해 Han, Pei, & Yin에 의해서 만들어진 FP-트리와 유사한 형태의 트리로 과거의 빈번한 패턴을 저장하고 이 트리의 각 노드에 대해서 시간 윈도우당 발생 빈도수를 기록하는 테이블이 포함되어 있는 구조를 제시하였다[10]. 이들은 현재 마이닝되는 빈번하게 발생하는 패턴뿐 아니라, 그렇지 않은 패턴들까지도 어느 정도까지는 저장하기 때문에 빈번한 패턴이 그렇지 않은 것으로, 빈번하지 않았던 패턴이 빈번한 것으로 바뀔 수 있도록 하였다. FP-트리는 이밖에도 이것을 기반으로 하여 점진적 데이터 마이닝을 하는 다른 연구들을 낳았다[5]. 이 방법은 과거의 정보와 현재의 정보, 혹은 일반적인 정보와 가장 중요하게 여겨지는 정보를 적절하게 혼합하여 마이닝에 가장 효과적인 정보를 유지하려는 것이다. 결국 연속 발생 데이터의 특성과 사용자의 목적에 맞는 방법을 통해 물리적인 저장 공간과 메모리의 사용, 소요되는 시간을 줄이면서도 방대한 데이터에서 얻을 수 있는 정보의 누락을 최소화하고, 결과적으로는 기존의 전체 데이터에 대해 분석을 행하는 일괄처리 방법과 거의 같은 수준의 효과를 내는 것이 연속 발생 데이터 마이닝의 주요 목표가 되고 있다.

패턴을 마이닝하는 것은 앞에서도 언급했듯이, 자료의 조합과 분류를 밝히는 것 외에 예측하려는 목적 또한 가지고 있다. 예측을 한다는 것은 (x, y)의 형태로 데이터가 구성되어 있다고 생각하고 y는 예측하려고 하는 값이고 x는 속성들의 조합이라고 하면, y = f(x)인 모델을 생성해서 미래에 들어오는 자료의 x를 통해 가능한 정확하게 y를 예측하는 것이다. 연속 발생 데이터의 경우는 자료가 계속 증가되므로 그것을 예측하는 것은 기존에 생성된 패턴

으로 새로 들어오는 자료를 단순히 예측하는 것을 넘어, 새로 들어오는 자료를 더욱 효과적으로 예측할 수 있도록 증가되는 자료를 통해 생성된 예측모델을 갱신하는 것을 포함하게 된다. 또한 연속 발생 데이터의 경우 같은 조건의 데이터가 늘 일정한 예측값을 갖기 보다는 시간의 흐름과 자료의 증가에 따라 패턴이 변하면서 같은 조건의 데이터라도 다르게 예측될 수 있으므로, 이러한 변화를 반영하여 보다 정확한 값을 예측하는 것이 중요하다. 이런 식으로 데이터가 고정되어 있지 않고 변하는 데이터를 분석하여 예측하는 것은 통계적으로 저장된 데이터를 분석하고 예측하는 것보다 훨씬 복잡한 일이다. 예를 들어 예측 모델은 일시적으로 변화하는 패턴을 잡아내야만 하고 오래되어 쓸모없게 된 자료의 영향을 제거해야 한다. 혹은 오래 되었더라도 주기적으로 발생하는 패턴을 잡아내기 위해서 그 영향력이 유지되어야만 할 수도 있다. 따라서 좋은 예측모델은 모델에서 영향을 낮게 미쳐야 되는 자료들과 높게 미쳐야 하는 자료들을 구분해 정확한 예측을 하도록 하는 정확성과 효율성, 그리고 연속 발생 데이터가 계속 증가하는 만큼 점진적 데이터 마이닝으로 인해 지나치게 사용이 복잡해지는 것을 막는 사용의 용이함이 요구된다[18].

데이터가 시간에 따라 그 특성이 변하는 것을 개념 표류(concept drift)라고 한다. 전통적으로는 모든 데이터가 한가지의 개념에 의해서 생성된다고 보았으나 연속 발생 데이터의 경우처럼 데이터가 계속 추가되는 경우에는 데이터가 시간에 따라 다양한 여러 가지 일련의 개념들에 의해서 생성된다고 보는 것이 더 정확할 것이다. VFDT, CVFT, FP-stream등 트리를 사용하는 방법은 새로 데이터가 들어옴에 따라 트리를 갱신하면서 증가하는 패턴에 따라 트리를 구성하고 그렇지 않은 패턴은 자동적으로 버리는 경향이 있으므로 늘 데이터의 새롭게 증가하는 성격에만 초점을 맞추게 된다. 따라서 개념 표류가 일어나도 데이터의 주기적인 변화를 찾아내기 보다는 현재의 추세에 초점을 맞추게 되는 경향이 있다. 따라서 이러한 변화를 잡아내기 위해서는 총체적(ensemble) 접근방법이 필요하다. 총체적 접근 방법은 한명의 전문가보다 여러 명의 전문가가 한 예측을 종합했을 때 더 나은 결과를 가져올 수 있다는 점에 기초한다. 특히 이 여러 명의 전문가가 서로 독립적으로 예측의 잘못을 만들어낼 때에는 그것을 종합했을 때에 더욱 나은 예측을 할 수 있다. 따라서 데이터의 성격이 변화하는 개념 표류가 일어날 경우에는 위에 열거한 트리를 이용하여 하나로 귀결되는 예측모델을 사용하기 보다는 여러 개의 예측모델을 조합하는 것이 나을 수 있다. Wang 등은 데이터베이스의 개념이 변화할 경우 증가되는 데이터 set들에서 분류자(classifier)를 생성하고 각각의 예측 정확도를 따진 후 가장 정확도가 높은 상위 몇 개의 분류자를 종합하여 예측을 하는 방법을 택했다[18]. 그러나 이 방법 또한 상위 몇 개의 분류자를 선택하기 위하여 예측의 정확도를 매번 계산하여야 한다는 번거로움이 있다. 또한 예측 정확도를 계산하는 기준으로 새로 들어오는 데이터 set이 아니라 그 직전 데이터 set이 새 데이터 set과 가장 유사하다고 가정하고 그것을 사용하여 가장 예측 정확도가 높을 것이라 기대되는 상위 몇 개의 분류자를 예측모델로 선택하기 때문에 새 데이터가 직전 데이터와 매우 다르다면 예측 정확도가 떨어질 수 있는 가능성이 있다. 따라서 데이터의 주기적인 개념 변화를 찾아내어 효과적으로 예측하기 위해서는 최신의 패턴만 유지하는 트리나 상위 몇 개의 분류자만 유지하는 것보다 전체 데이터에 대한 정보를 유지하는 방법이 고려되어야 할 것이다.

본 연구에서는 전체 데이터에 대한 정보를 유지하면서 연속 발생 데이터의 효율적인 예측을 하기 위한 방법을 규칙이라는 형태를 이용한다. 보통의 연관 규칙은 X => Y 형태를 띠고 있다. 여기서 X => Y는 데이터베이스의 X 조건을 만족시키는 어떠한 tuple은 Y에서의 어떤 조건을 만족시키는 경향이 있다는 것을 의미한다[1]. 이러한 규칙은 의사결정 트리로부터 규칙으로 쉽게 전환될 수 있다. 의사결정 트리를 유도하는 기본적인 알고리즘은 하향

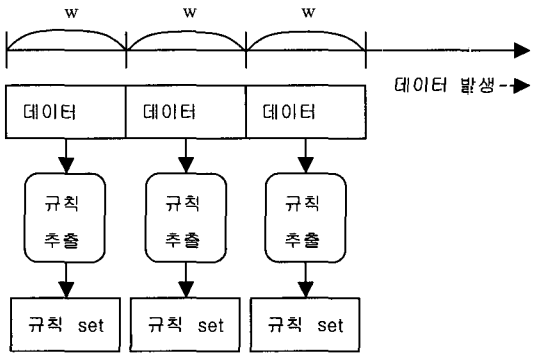
식으로 재귀적인 분할과 정복 방법으로 트리를 구축하는 알고리즘이 있다. 이 중 대표적인 것으로 ID3가 있다. ID3가 속성값을 가지는 속성들에 치우치는 경향이 있다면 명목형 속성 위주로 개선한 것으로 C4.5알고리즘이 있다. C4.5 알고리즘을 구현한 소프트웨어가 SEE5인데 본 연구에서는 SEE5가 만든 트리를 자체적으로 규칙으로 바꾼 것을 사용한다.

이 연구에서는 일정 부분의 자료를 부분적으로 선택하여 분석대상으로 하는 것이 아니라, 전체 데이터에서 추출된 정보를 규칙이라는 형태로 누적하여 사용함으로 데이터의 손실이 없고 과거의 데이터에 대한 정보도 그대로 유지하며 시간의 흐름에 따른 패턴의 변화 또한 처리하였다. 따라서 데이터에 주기적인 변화가 일어나거나 오랜 시간이 지난 후에 과거와 유사한 패턴이 발생하는 경우라도 누적된 정보를 통해 효과적으로 예측하는 것이 가능하게 하였다. 전체 데이터 대신에 각 데이터 set에서 생성된 규칙만을 저장하므로 저장 공간을 현저하게 줄이고, 또한 데이터 set 단위로 분석하기 때문에 전체 데이터를 분석하는데 드는 메모리의 사용을 줄여 효율성을 보장하고, 정보의 마이닝에 소요되는 시간을 줄이면서도 트리 혹은 패턴 자체를 총체적으로 갱신시키는 것보다 훨씬 간편하게 데이터를 마이닝하고 예측 할 수 있다.

3. 연구모형

3.1 규칙 set의 누적

본 연구의 모형은 연속 발생 데이터의 각 시간별로 나누어진 데이터 set에서 정보를 추출하여 체계적으로 결합시킨다. 연속 발생 데이터를 데이터 set 단위로 마이닝해서 규칙 set을 추출하는 과정은 [그림 1]과 같다. w는 데이터 set을 만들어주는 기준이 되는 크기인 윈도우의 크기이다. 연속 발생 데이터가 계속 들어옴에 따라 새로운 데이터 set에서 추출되는 규칙들이 규칙 set으로 만들어지고 이것은 또한 기존의 규칙 set 집합에 추가되므로 시간의 변화에 따라 규칙 set 집합 자체도 계속 변화하게 된다. 연속 발생 데이터가 들어와서 어느 정도 일정 크기가 되면 이것을 하나의 데이터 set으로 보고 이 데이터 set에서 규칙들을 추출하여 규칙 set으로 만든 후 전체 규칙 set 집합에 저장하고 규칙들을 추출한 그 데이터 set 자체는 저장하지 않는다.



[그림 1] 데이터 set 단위의 연속 발생 데이터에서 규칙 set을 추출하는 과정

그리고 또 다시 일정 크기의 데이터가 들어와 미리 정해진 데이터 set의 크기가 되면 다음 데이터 set이 준비되었다고 보고 여기에서도 규칙들을 추출하여 이전에 누적된 규칙 set 집합과 결합한다. 이 과정은 데이터가 추가되는 한 반복된다. 이런 식으로 규칙 set 집합은 전체 데이터를 저장하지 않고도 전체 데이터에 대한 압축된 정보를 규칙이라는 형태로 갖고 있게 된다. 규칙 set 자체는 데이터 set에 비해서 그 크기가 현저히 작기 때문에 규칙 set만을 저장 공간에 유지하는 것은 연속 발생 데이터에서 늘어나는 데이터 자체를 저장하는 것보다 저장 공간을 훨씬 절약해 줄 수 있고 또한 매번 그때까지 쌓인 모든 데이터를 분석해서 패턴을 찾아내지 않아도 되므로 메모리의 사용과 분석에 걸리는 시간의 문제 또한 줄여 줄 수 있을 것이다. 규칙을 추출하여 규칙 set으로 만들고 이 규칙 set을 규칙 set 집합에 추가한 후 이렇게 추가된 규칙 set 집합으로 새로운 데이터 set의 예측되지 않은 목표 class를 예측하는 과정을 <표 1>과 [그림 2]에서 설명하고 있다.

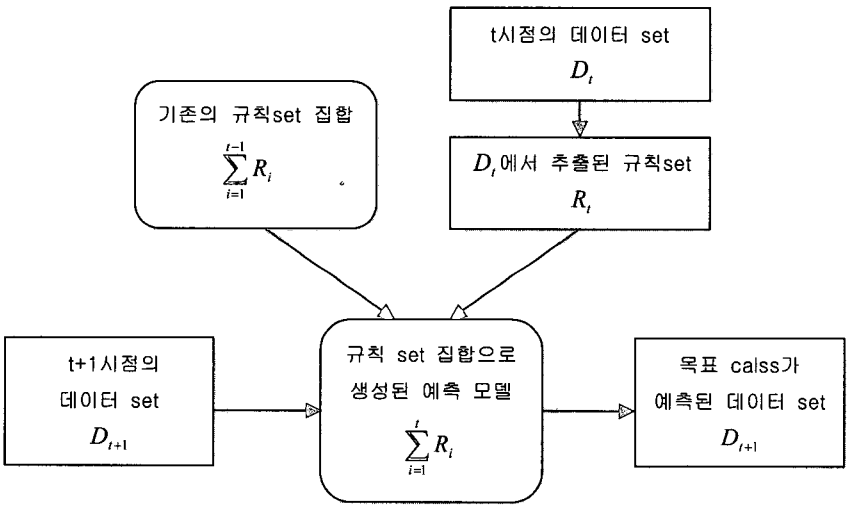
〈표 1〉 변화하는 규칙 set 집합으로 데이터 set을 예측하는 과정

1. t시점의 데이터 set d(D_t)에서 규칙 r들을 추출하여 그 합인 $\sum_{j=1}^{n} r_j$를 t시점의 규칙 set R에 모으고 이를 R_t라 한다. (n = 데이터 set D에서 추출되는 규칙의 수)

   이 때 t시점까지 누적된 규칙 set 집합은 $\sum_{i=1}^{t}\sum_{j=1}^{n} r_{ij}$이 된다.
   (t = t시점까지의 데이터 set의 수, n = 각 데이터 set d에서 추출되는 규칙의 수)
   즉 t시점까지의 데이터 set들에서 추출된 모든 규칙을 누적한다.

   이 규칙 set 집합을 각각의 규칙 set의 합인 $\sum_{i=1}^{t} R_i$라 한다.

2. 데이터 set t+1(D_{t+1})의 예측하고자 하는 목표 class를 $\sum_{i=1}^{t} R_i$을 통해 예측한다.



[그림 2] 예측 과정

여기서는 다음과 같은 두 가지 조건을 미리 가정한다.

조건 1 : 데이터 set의 크기를 미리 정해 두고 그 크기만큼의 데이터가 누적될 경우 하나의 데이터 set으로 만든다.

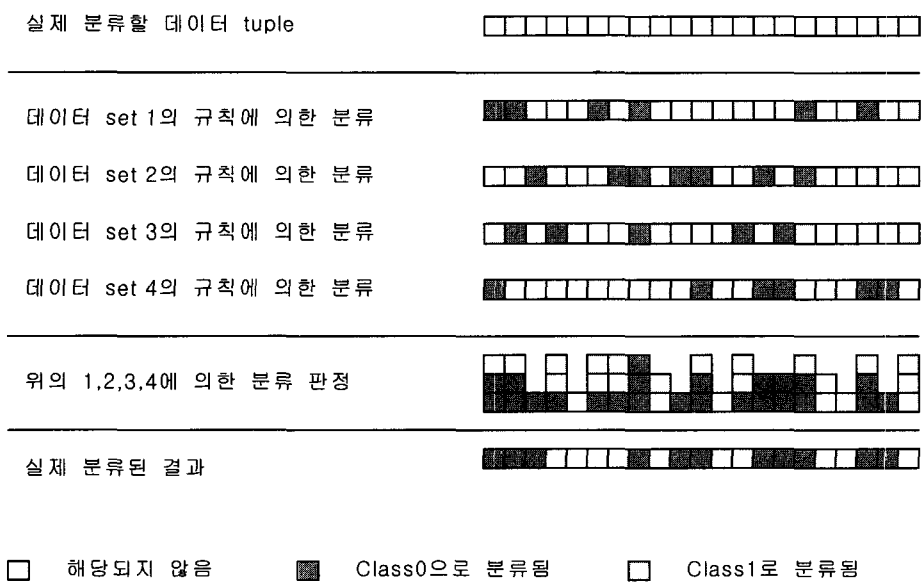
조건 2 : 예측은 하나의 데이터가 아니라 데이터 set 단위 별로 한다.

데이터가 추가됨에 따라 전체 데이터의 성격이 계속적으로 변할 수 있는 연속 발생 데이터의 특성을 고려했을 때 예측하려고 하는 정보가 과거의 데이터와 밀접하게 연관이 되어 있는 경우라면 이러한 정보의 손실을 가져올 수 있다.

규칙 set을 유지하면서 최초 데이터들에서 나온 정보들을 단순히 가지고만 있는 것이 아니라 데이터 set들의 변화되는 정보를 유지하면서도 특이값에 영향을 적게 받도록 할 수 있다. 예를 들어 데이터 set 1, 2, 3에서는 A => 0과 같이 A와 같은 조건을 가진 tuple은 class 0으로 분류되는 규칙을 추출하였는데 데이터 set 4에서는 A=>1와 같이 같은 조건인 A임에도 불구하고 class 1로 분류하는 규칙을 추출하였다. 그 후 데이터 set 5에서 다시 A=>0이라는 규칙을 추출하였다면 이 경우 데이터 set 4는 특정 상황에서 만들어진 특이한 데이터이며 이러한 경우는 다시 발생할 확률이 낮다고 볼 수 있다. 따라서 이러한 특정한 경우에서 받는 영향을 최

소화하는 것이 바람직 할 것이다. 이제 다시 데이터 set 5의 tuple들의 예측 목표 class를 class 0또는 class 1로 예측하려 한다고 하자. 가장 최근 데이터 set을 저장하고 그것의 데이터를 통해 새로운 데이터를 예측하는 방식에서는 데이터 set 4에서 추출된 규칙을 통해 데이터 set 5를 분류하려 할 것이다. 이 경우 A조건을 만족하는 tuple의 경우 class 1로 분류될 것이다. 이 경우에는 데이터 set 5를 예측하는 바탕이 되는 데이터 set 4만을 저장하고 데이터 set 4가 특이한 경우에서 생성된 데이터임을 알 수 있는 비교 가능한 과거 자료가 존재하지 않으므로 데이터 set 4가 특이한 경우에 생성된 자료라고 하더라도 그 영향을 줄일 수가 없을 것이다. 그러나 누적된 규칙 set을 사용하는 경우 데이터 set 1, 2, 3에서 A조건은 class 0으로 분류되어 왔으므로 데이터 set 4에서 추출된 규칙이 A조건을 class 1로 분류하였다고 하더라도 기존의 모든 규칙 set 집합을 살펴보았을 때 class 0과 class 1로 분류되는 비율이 3 : 1로 class 0이 우세하므로 결과적으로 class 0으로 분류한다. 이렇게 연속 발생 데이터와 같이 시간에 따라 변하는 데이터는 최근의 데이터를 분석하여 여기에서 추출된 규칙 set으로 다음 데이터를 예측하거나, 혹은 자료를 통해 전체적으로 포괄된 하나의 규칙 set을 사용하는 것보다 시간의 흐름에 따른 데이터의 변화를 반영할 수 있는 여러 개의 규칙 set이 조합되어 집합을 이룬 예측 모델을 사용하는 것이 예측의 정확도를 높일 수 있다. 이것을 [그림 3]이 설명해주고 있다.

실제 분류할 데이터 tuple 20개로 이루어진 데이터 set이 [그림 3]과 같이 있다고 하자. 이 데이터 set의 tupel들은 각각 데이터 set 1, 2, 3, 4에서 추출된 규칙들에 의해서 class 0 혹은 class 1로 분류되거나 어떠한 tuple은 데이터 set에서 추출된 A조건으로 분류하는 규칙에는 해당되지 않는다. 예를 들어 분류되어야 할 데이터 set의 첫번째 tuple은 데이터 set 1, 4에서 추출되는 규칙에 의해 분리되면 class 0으로 분류되고, 데이터 set 2에서 추출된 규칙에는 해당되지 않으며 데이터 set 3에 의해 추출된 규칙에 의해서는 class 1로 분류된다. 이 과정을 거쳐 최종적으로 이 첫번째 데이터 tuple은 class 0과 class 1이 2 : 1로 결국 class 0으로 분류된다.



[그림 3] 규칙 누적을 이용한 분류의 예

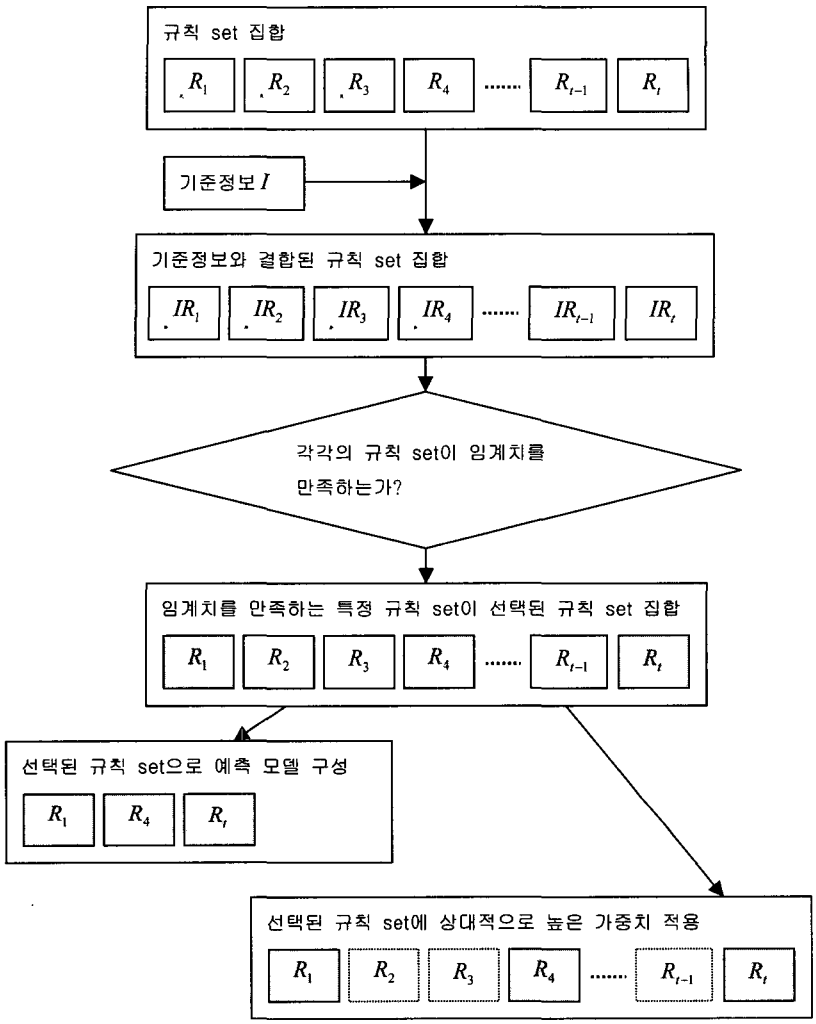
이렇게 규칙이 서로 다른 조건에서 만들어진 데이터 set에서 생성되기 때문에 같은 tuple이라도 어느 데이터 set에서 만들어진 규칙에 의해 분류되느냐에 따라 서로 다른 class로 분류될 수 있지만 그것이 누적되면 결국은 실제 분류되어야 할 값과 같은 값으로 분류될 확률을 높일 수 있다. 따라서 이 방법이 연속 발생 데이터의 경우에 더욱 효율적으로 새로 들어오는 데이터를 예측할 수 있다고 보고 규칙 set들을 누적하지 않는 방법들과 예측 정확도를 비교해 보려고 한다. 규칙 set들을 누적하는 방법과 비교하려고 하는 두 가지 방법으로는 저장공간, 메모리, 시간 등의 제약이나 혹은 오래된 데이터에 대한 흥미의 저하로 인해 최근 데이터만을 가지고 새로 들어오는 데이터를 예측하는 방법과 일반 일괄처리방법처럼 예측하려고 하는 시점 전까지의 모든 데이터를 분석대상으로 삼아 예측모델을 만들고 그것을 통해 예측하는 방법을 택하였다.

3.2 상대적으로 중요 규칙 set 선택, 예측모델 구성

규칙 set을 누적하는 것에서 나아가서 만약 예측하려고 하는 데이터 set의 특징이 지금까지 쌓여있는 데이터 set중에서 어떠한 데이터 set의 특징과 비슷하다고 한다면 그 데이터 set에서 추출된 규칙은 다른 데이터 set에서 추출된 규칙에 비해 상대적으로 중요한 규칙이 될 수 있다. 간단한 예를 들어 소비자의 웹 탐색 행동이 구매로 연결될 것인지를 예측한다고 하자. 만약 소비자의 웹 탐색 행동 데이터를 계절별로 데이터 set으로 나누었다면 현재의 예측하려고 하는 계절과 같은 과거의 계절에 얻어진 데이터에서 추출된 규칙들을 더 중요하게 취급할 수 있을 것이다. 즉 상대적으로 중요한 규칙 set을 선택하였다면, 그 규칙 set에 따라 어떠한 class로 분류되는 빈도수에 다른 데이터 set에서 추출된 규칙 set에 의해 분류되는 것보다 상대적으로 높은 가중치를 주어 선택된 규칙 set의 class 분류 결과에 따라 분류될 가능성을 높이는 것이다. 또한 이렇게 상대적으로 중요한 규칙 set이 일단 선택되었다면 이렇게 선택된 규칙 set만을 가지고 예측 모델을 구성할 것인지, 그렇지 않으면 전체 규칙 set을 사용해서 예측 모델을 구성하되 더 중요하게 여겨지는 규칙 set에 그렇지 않은 규칙 set들보다 상대적으로 높은 가중치를 주어 예측 모델을 구성할 것인지 생각해볼 수 있다

[그림 4]에서 보여지듯이 먼저 각각의 데이터 set에서 추출된 규칙 set 집합을 특정 기준 정보와 결합하게 된다. 이 기준 정보는 어떤 규칙 set이 중요도가 높은지 그렇지 않은지를 판단하는 근거가 되는 것으로서 데이터의 특징에 따라 다른 것을 사용할 수 있다. 임의의 데이터 set을 선택해서 그것과 다른 데이터 set을 비교해도 그 차이가 크지 않을 것이므로 가장 처음 들어오는 첫번째 데이터 set을 데이터가 만들어진 개념의 기본이 되는 데이터라고 가정하고 이것을 각각의 규칙 set을 시험하는 시험 데이터 set으로 하여 이것을 기준정보로 각각의 규칙 set의 예측 정확도를 측정하였다. 이 경우에는 이렇게 측정된 각 규칙 set의 예측 정확도가 [그림 4]에서 표시된 기준정보와 결합된 규칙 set의 집합 안에 포함되게 된다. 이 후 이것이 임계치를 만족하는지 살펴서 임계치를 만족하는 경우 그 규칙 set이 상대적으로 중요도가 높은 규칙 set이라고 판단하였다. 다시 구매력 예측 자료를 예로 들면 미리 정해 놓은 예측 정확도 임계치 수치보다 각 규칙 set의 예측 정확도가 높은지 따져서 임계치보다 높으면 중요도가 높은 규칙 set으로 판단하였다. 즉 그 특정 데이터 set에서 추출된 규칙 set은 다른 데이터 set에서 추출된 규칙 set보다 미래의 자료를 예측하는데 중요한 근거가 된다고 보았다.

다음으로 이렇게 중요도가 높은 것으로 선택된 규칙 set들을 이용해서 어떻게 예측 모델을 만들 것인가의 문제를 살펴볼 수 있다. 본 연구에서 초점을 맞추려고 하는 것은 덜 중요하게 판단된 자료의 영향력을 배제하지 않는다는 것이다. 연속 발생 데이터에서는 시간이 변함에 따라 그 개념이 변화하는



[그림 4] 상대적 중요도가 높은 규칙 set 판단 과정과 예측 모델 구성

개념 표류가 일어나므로 크든 적든 간에 기존에 만들어진 데이터 set과 현재 데이터 set과는 차이가 있다. 따라서 새 데이터 set을 예측하려고 할 때 지금까지 쌓인 데이터 set들 중에서 그것과 더 비슷한 데이터 set이 있고 그렇지 않은 데이터 set이 있다. 그러나 시간의 흐름에 따라 변화가 일어나는 연속 발생 데이터에서 데이터 set단위로 분석, 예측하는 경우 절대적으로 같은 데이터 set이 존재하기 힘들다. 주기적인 패턴과 과거 시간부터 누적된 정보를 반영하기 위한 이유에서 상대적으로 더 중요한 데이터 set이라고 판단되는 데이터 set들 뿐 아니라 그렇지 않은 다른 데이터 set의 영향력 또한 고려하여야 한다. 본 연구는 일반적인 규칙 set을 배제하고 중요도가 높은 규칙 set만으로 예측모델을 구성하는 경우와 일반적인 규칙 set들까지 포함하고 중요도가 높은 규칙 set들에 달리 가중치를 주는 경우 두 가지의 예측력을 비교한다.

4. 구매력 예측 자료에의 적용

본 연구에서 제시한 예측 모델을 소비자 인구통계 데이터를 이용한 구매력 예측 문제에 적용하여

보았다. 이 자료는 시간의 흐름에 따라 자료 흐름에 변화가 일어나기는 하나 그 변화가 크지 않고 매우 오랜 시간에 걸쳐 변화가 천천히 일어나는 자료이다. 따라서 구매력 예측 자료 내의 데이터 set들간에 자료 구조의 차이가 비교적 크지 않다. 이 자료를 대상으로 최근 데이터 set만 유지하여 새 데이터 set을 예측하는 경우, 기존의 모든 데이터를 유지하고 있다가 새 데이터 set을 예측하는 경우, 데이터 set에서 추출된 규칙 set을 누적하여 새 데이터 set을 예측하는 각각의 경우를 비교하였다. 나아가서 누적한 규칙 set 집합에서 중요도가 높은 규칙 set을 판단하여 그것만으로 구성된 예측 모델을 사용하는 경우와, 중요도가 높은 규칙 set뿐 아니라 일반 규칙 set도 포함하여 예측 모델을 사용하는 경우도 비교 한다. 규칙을 추출하기 위하여는 데이터 마이닝의 상용 소프트웨어인 SEE5 1.19를 사용하였다.

4.1 자료의 기초분석과 데이터 set 설정

이 자료는 1990년부터 2000년 사이의 미국 인구조사 자료를 공공 사용 용도에 맞도록 만들어진 micro sample 데이터이다. STATA 6.0 format으로 만들어진 표본 데이터이며 4백 5십만의 남자와 5백만의 여자의 데이터, 총 9백 십만 tuple을 가지고 있다. 이 자료를 토대로 다음과 같이 개인의 특성을 바탕으로 구매력을 측정하는 자료를 구성하였다. 14개의 입력 변수는 <표 2>와 같이 개인의 특성을 나타내고 있다. 예측변수는 개인의 구매력으로서 개인의 구매력이 5만 달러 이하인지 5만 달러를 넘는지를 나타내고 있다. 즉 5만 달러로 구매력의 기준을 잡고 개인의 어떠한 특성에 따라 구매력이 그 이상 혹은 이하가 되는지 예측하기 위한 자료이다. 입력변수에 수입이 있기는 하지만 개인적인 부분이기 때문에 빠져 있는 경우가 많으므로 이것이 구매력이 얼마인지 구분하는 절대적인 기준이 될 수 없다고 가정하였다.

데이터 set설정을 위하여 구매력 예측 자료에서 각각 400개의 tuple로 이루어진 20개의 데이터 set을 순차적으로 추출하였다.

〈표 2〉 구매력 예측 자료의 입력변수와 예측변수

입 력 변 수		예 측 변 수
변 수 이 름	변 수 설 명	
Age	나이	구매력 구분 (5만 달러 이상 혹은 이하)
Workclass	직업구분	
Fnlwgt	기능코드	
Education	교육정도	
Edunum	교육연수	
Ms	결혼상태	
Occupation	직업유형	
Relationship	가족유형	
Race	인종	
Sex	성별	
Gain	수입	
Loss	지출	
Hoursperweek	주당 일하는 시간	
Country	출신 국가	

4.2 실 험

4.2.1 시점 t에서의 데이터 set에서 추출된 규칙으로 시점 t+1의 데이터 set 예측

SEE5를 사용하여 추출된 규칙의 예는 <표 3>과 같다.

〈표 3〉 SEE5를 통해 추출된 규칙의 예

데이터 set	각 데이터 set에서 추출된 규칙의 일부			
1	edunum <= 12	and	gain <= 5013	-> class <= 50K
2	edunum <= 11	and	race = Black	-> class <= 50K
3	edunum > 9	and	ms = Married-civ-spouse	-> class > 50K

〈표 4〉 시점 t에서 R_t의 수

데이터 set	1	2	3	4	5	6	7	8	9	10	
규 칙 수	5	7	5	14	9	10	11	4	4	4	
데이터 set	11	12	13	14	15	16	17	18	19	20	합계
규 칙 수	2	5	6	13	6	6	9	3	6	6	135

〈표 5〉 R_t로 D_{t+1}을 예측 결과

데이터 set	2	3	4	5	6	7	8	9	10	11
정확도(%)	80.2	85	78.2	81.2	82	82.5	79	80.2	86.2	81.5
데이터set	12	13	14	15	16	17	18	19	20	평균
정확도(%)	79.7	79.7	79.7	82.2	86	84.5	83.5	85.2	86	82.24

각 데이터 set 에서 나온 규칙들의 수는 <표 4>과 같다. 데이터 set마다 추출된 규칙의 수가 차이가 나는 이유는 규칙이 그 데이터 set의 데이터분포를 가장 잘 표현해 줄 수 있도록 생성되었고 데이터 set은 저마다 조금씩 다른 개념에서 생성되어 다른 분포를 가지고 있는 데이터를 가지고 있기 때문이다.

20개의 데이터 set에서 총 135개의 규칙이 추출되었다. 다음으로 데이터 set t를 training 데이터 set으로 하고 거기에서 나온 규칙들로 test 데이터 set인 데이터 set t+1을 예측한 후 그 예측 정확도를 구하였다. t시점 전의 데이터 set들은 사용되지 않았다. 이 경우는 저장 공간상의 제약, 혹은 최신 데이터만이 새로운 데이터를 가장 잘 예측할 것이라는 믿음 등의 이유로 인해 전체 데이터에 대한 정보를 모두 저장하고 있기보다는 바로 직전, 다시 말해 가장 최신의 데이터 set만을 유지하고 그것을 통해 새로운 데이터 set을 예측해 보았다. 이 방법의 예측 정확도 평균은 <표 5>에서와 같이 82.24%이다.

4.2.2. t시점까지 누적된 규칙 set으로 t+1시점 데이터 set 예측

데이터 set에서 추출되는 규칙들을 처음부터 모두 누적하여 규칙 set 집합으로 만들어 새로 들어오는 데이터 set을 예측할 경우 예측 정확도를 측정하였다. 규칙 set에서는 한 tuple이 여러 개의 규칙에 해당될 수 있고 또한 서로 다른 class로 예측될 수 있다. 다시 예를 들면 tuple의 edunum(교육연수)의 값이 10이고 gain(수입)의 값은 5000, ms(결혼상태)의 값은 marrid-civ-spouse(공무원과 결혼)이며 race(인종)의 값은 black(흑인)일때 이 tuple은 <표 3>의 세 규칙에 모두 해당된다. 이 경우 첫 번째, 두 번째 규칙이 모두 이 tuple이 지칭하는 특성을 가진 사람의 구매력을 <= 50K(구매력 5만 달러 이하)로 예측했지만 세 번째 규칙의 경우에는 > 50K(구매력 5만 달러 초과)로 예측하였다. 만약 가장 최근 데이터 set에서 추출된 규칙 set으로만 이것을 예측한다면 세 번째 규칙 set이 사용될 것이므로 구매력은 > 50K 일 것이다. 그러나 과거 데이터인 첫 번째, 두 번째 규칙 set까지 고려한다면 class <= 50K와 > 50K가 2 : 1의 빈도이므로 <= 50K로 예측된다. 시험을 위해 default class를 class > 50K로 정한 후 시험하였다. SEE5를 통해 만들어진 규칙들을 누적한 $\sum_{i=1}^{t} R_i$의 개수는 <표 6>과 같으며 이 $\sum_{i=1}^{t} R_i$으로 D_{t+1}을 예측한 것이 <표 7>이다.

어떤 데이터 set을 예측하는 시점에서 단지 그것을 예측하는 규칙의 개수가 많아졌다고 해서 예측이 더 정확해지는 것은 아니다. 규칙들이 제각각 다른 데이터 set에서 추출된 것들이므로 데이터 set의 저마다 다른 특성을 반영하여 같은 tuple이라도 서로 다른 class로 예측할 수도 있기 때문에 오히려 예측 정확도는 더 떨어질 수 있으며 따라서 규칙의 개수가 많다고 해서 예측 정확도를 높일 수는 없다. 많은 경우 규칙이 누적되면서 이것들끼리 모순이

〈표 6〉 시점 t에서 $\sum_{i=1}^{t} R_i$의 수

데이터 set	1	2	3	4	5	6	7	8	9	10
규 칙 수1	5	12	17	31	40	50	61	65	69	73
데이터 set	11	12	13	14	15	16	17	18	19	20
규 칙 수1	75	80	86	99	105	111	120	123	129	135

〈표 7〉 $\sum_{i=1}^{t} R_i$로 D_{t+1}을 예측한 결과

데이터 set	2	3	4	5	6	7	8	9	10	11
정확도(%)	80.5	83.5	82.25	84.75	84.25	83.75	82.75	83.5	86.5	84
데이터 set	12	13	14	15	16	17	18	19	20	평균
정확도(%)	83.25	81	82	85	87.5	85	84	87.75	86.25	84.08

생기는 이유로 같은 tuple이라도 서로 다른 class로 예측하는 일이 생길 수 있지만 결국에는 이것이 t+1시점의 데이터 set을 더욱 정확하게 예측하는 방향으로 정보를 조정하면서 누적된다는 가정을 하고 있고 이것은 예측 정확도가 규칙 set을 누적하지 않은 경우보다 향상이 됨 으로서 뒷받침된다. 예측의 정확도 평균은 84.08%가 되어 단순히 이전 데이터 set에서 추출된 규칙을 통해 다음 데이터 set을 예측하는 경우의 예측 정확도 평균 82.24%보다 예측 정확도가 1.84% 향상된 것을 볼 수 있다. 규칙을 누적하여 다음 데이터 set을 예측하였을 경우 그 전 시점의 데이터로 다음 데이터 set을 예측하였을 경우보다 예측의 정확도가 높아지는 것을 볼 수 있다.

저장 공간상의 문제에 있어 새로 들어오는 데이터 set을 저장하는 공간과는 별도로 이미 들어온 데이터 set 혹은 규칙 set 집합을 저장하는 공간이 필요하다. 다시 말해 데이터 set을 이용하는 경우 이전 데이터 set + 새로운 데이터 set 만큼의 저장 공간이 필요하고 규칙 set 집합을 이용하는 경우에는 규칙 set 집합 + 새로운 데이터 set 만큼의 저장 공간이 필요하다. 각각의 데이터 set file이 49 KB의 저장공간을 차지하고 있으므로 20개의 데이터 set은 총 980 KB의 저장 공간을 차지하고 있다. 규칙 set을 누적한 file의 경우 전체 20개 데이터 set에서 추출된 135개의 규칙을 모두 가지고 있어도 file의 크기는 7 KB에 불과하고 데이터 set 하나의 크기보다 작다. 따라서 데이터 file을 모두 저장하지 않고 그 데이터들에 대한 규칙만을 가지고 있는 file만을 유지하고 있을 경우 저장공간을 현저히 절약할 수 있음을 알 수 있다. 규칙 set 집합이 커짐에 따라 이것이 하나의 데이터 set 크기인 3.35 KB를 넘게 되겠지만 규칙 set 집합이 저장하는 정보의 양이 직전 데이터 set 하나를 저장하는 것보다 많고, 그 크기의 증가가 매우 작으므로 예측 정확도를 높이면서도 저장 공간이 현저하게 절약되고 있음을 알 수 있다.

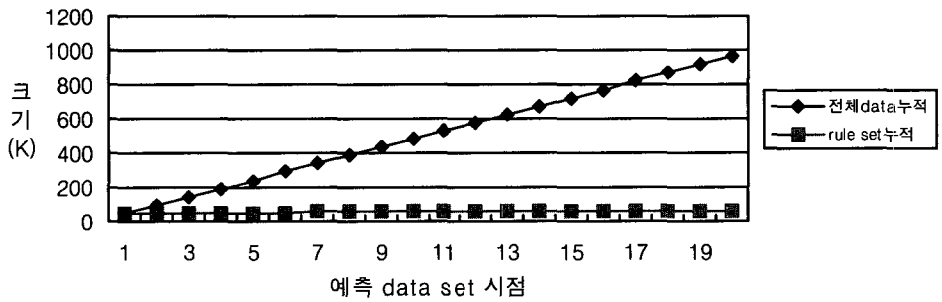
4.2.3. t시점까지의 전체 데이터로 t+1시점의 데이터 set 예측

만약 물리적, 시간적, 비용적인 소모를 감수하고도 예측 시점에서 그때까지의 전체 데이터를 사용하여 예측을 한다고 가정하자. 이 방법은 기존의 일괄처리방법과 같은 방법으로 일괄적으로 모든 데이터를 분석의 대상으로 삼아 모델을 만들어낸 후에 새로운 데이터 set을 예측하는 방법이다. 다시 말해 t+1시점의 데이터 set을 예측하기 위해 예측 모델을 생성할 때 사용되는 데이터는 최초의 데이터부터 t 시점의 데이터까지 그 동안 누적된 모든 데이터이고 이 전체 데이터에서 하나의 규칙 set을 추출하여 예측 모델로 삼은 후 t+1시점의 데이터 set을 예측하는 것이다. 이 방법의 예측 정확도는 <표 8>에 보여지고 있다. 실제 이 방법을 사용하려면 전체 데이터의 크기가 1GB 이하가 되어야 하지만 데이터 마이닝 기법들은 1GB 이상의 크기를 가진 데이터를 이용하기가 어렵다.

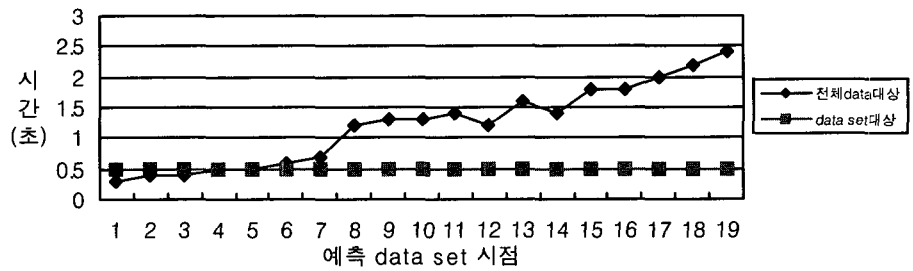
〈표 8〉 전체 데이터에서 추출된 하나의 규칙 set으로 을 예측한 결과

데이터 set	2	3	4	5	6	7	8	9	10	11
정확도(%)	80.2	85.7	83.7	85.7	85.2	84.7	82.7	84.5	87.5	84
데이터 set	12	13	14	15	16	17	18	19	20	합 계
정확도(%)	83.5	81.5	82.5	86.2	86.7	85.7	85.5	87.2	89.2	84.84

이 경우 예측 정확도의 평균은 84.84%로 t+1시점의 데이터 set을 예측하기 위해 t시점의 데이터 set에서 나온 규칙 set만을 사용하는 경우보다는 예측 정확도 평균이 82.24%에서 2.6% 증가했다. 또한 각각의 데이터 set에서 규칙 set을 누적하여 규칙 set 집합으로 예측 모델을 만든 후 t+1시점의 데이터 set을 예측한 경우 그때의 예측정확도 평균인 84.08%보다도 0.76%가 증가했다. 이 경우 전체 데이터 set을 사용하지 않고 직전 데이터 set만을 가지고 모델을 생성하여 예측하는 경우보다는 더욱 많은 정보, 즉 데이터를 사용하므로 예측 정확도가 증가하리라는 것은 쉽게 예상할 수 있다. 그러나 데이터의 손실 없이 이것을 규칙 set 집합이라는 다른 형태의 정보로 바꾸어 누적시키는 방법인, 규칙 set의 집합으로 t+1시점을 예측한 경우보다도 예측정확도가 높아졌다. 그러나 이 경우는 예측의 정확도 평균이 증가하기는 하였지만 그 평균의 증가치가 0.76%로 비교적 높은 수치는 아니다. 저장 공간을 고려하였을 때 규칙 set을 누적하여 규칙 set 집합을 사용하는 경우에는 규칙 set을 추출하기 위하여 늘 한 개의 데이터 set을 저장하고, 누적된 규칙 set 집합을 저장한다. 데이터 set의 크기는 늘 일정하고 누적된 규칙 set 집합은 그 크기가 데이터 set보다 현저하게 작으면서 증가 속도도 빠르지 않다. 그러나 전체 데이터를 누적하는 방식은 들어오는 데이터를 모두 저장해야 하므로 저장 공간이 많이 소요될 뿐 아니라 그 증가 속도도 급격히 증가하게 된다. 이것을 그래프로 나타낸 것이 [그림 5]이다.



[그림 5] 규칙 set을 누적하는 방식과, 전체 데이터 누적방식에 소요되는 저장공간 비교



[그림 6] 규칙 set을 누적하는 방식과, 전체 데이터 누적방식의 규칙 set 추출 시간 비교

규칙 set을 추출하는 데 필요한 시간을 고려한다면 규칙 set 누적 방식의 경우는 일정한 시간이 걸릴 뿐이지만, 누적된 전체 데이터를 대상으로 규칙 set을 추출하는 경우는 저장 공간의 증가와 함께 규칙 set을 추출하는데 걸리는 시간의 증가 속도도 매우 빠르게 증가하는 것을 [그림 6]에서와 같이 볼 수 있다.

본 연구에서는 누적된 데이터의 크기가 1GB이하로 가정하고 이 시험을 하였으나 현실적으로 오늘날 많은 데이터들이 그 전체 누적 크기가 1GB 또는 10GB가 넘는 경우가 많다. 그러므로 <표 8>에서 보여 주듯이 예측 정확도 평균이 84.84% 이기는 하나 누적 데이터 크기가 1GB 이상인 경우 이 값은 큰 의미가 없어진다. 또한 전체 데이터를 사용할 경우 필요한 물리적인 저장 공간의 증가와 규칙 set 추출에 걸리는 시간 증가를 고려하였을 때 어떤 모델을 사용할 것인지는 고려해볼 필요가 있을 것이다. 특히 이 경우 소요되는 저장공간, 시간은 규칙 set 집합을 누적하여 사용하는 경우보다 상당히 빠른 속도로 증가하게 되므로 여기에 따르는 비용을 고려했을 때 예측의 정확도가 비슷할 경우 전체 데이터를 사용하는 경우와 비교하여 그 예측정확도가 사용자가 받아들일 만한 것이라면 규칙 set 집합을 사용하는 경우가 더 바람직할 수 있다.

4.2.4 중요도가 높은 규칙 set만으로 구성된 예측모델

새 데이터 set을 예측하려고 할 때 중요도가 높은 규칙 set을 선택하여 예측 모델을 구성하였다. 구매력 예측 자료는 시간의 흐름에 따른 자료의 변화가 비교적 적은 자료이다. 즉 데이터를 생성하는 개념의 변화가 적고, 따라서 데이터 set들간의 차이가 크지 않은 자료이다. 그러므로 예측 시점마다 중요도가 높은 규칙 set을 결정하기 위하여 새로운 판단기준에 근거하기보다는 기본이 되는 고정된 하나의 test 데이터 set을 마련함으로서 중요도가 높은 규칙 set을 판단하는 근거로 하였다. 이 test 데이터 set을 그 때까지 누적된 규칙 set들로 예측하여 예측 정확도 평균을 낸 결과를 임계치로 삼았고, 예측의 정확도가 임계치를 넘는 경우 이 규칙 set을 중요도가 높은 규칙 set으로 보았다. 이 결과 test 데이터 set으로 시험한 각 규칙 set의 예측도를 <표 9>가 보여주고 있다. 예측하려고 하는 데이터 set

〈표 9〉 test 데이터 set으로 시험한 각 규칙 set의 예측 정확도 결과

데이터 set	1	2	3	4	5	6	7	8	9	10
정확도(%)	84.5	81.2	81.5	81.2	80.7	81.7	83	81.2	83.7	80
데이터 set	11	12	13	14	15	16	17	18	19	평균
정확도(%)	82	81.2	83	85.5	84.5	84.7	84	84.2	83.7	82.7

〈표 10〉 데이터 set을 예측하는 시점에 예측 모델에 포함되는 규칙 set

예측할 데이터 set	예측모델에 포함된 규칙 set	예측할 데이터 set	예측모델에 포함된 규칙 set
2	1	12	1,2,8,10
3	1,2	13	1,2,8,10,12
4	1,2	14	1,2,8,10,12
5	1,2	15	1,2,8,10,14
6	1,2	16	1,2,8,10,14,15
7	1,2	17	1,2,8,10,14,15,16
8	1,2	18	1,2,8,10,14,15,16,17
9	1,2,8	19	1,2,8,10,14,15,16,17,18
10	1,2,8	20	1,2,8,10,14,15,16,17,18,19
11	1,2,8,10		

이 2에서 20까지의 데이터 set이고 예측하는데 사용될 규칙 set은 그 시점 전까지의 데이터 set에서 추출된 규칙 set들이기 때문에 1에서 19까지의 규칙 set들을 시험하였다. 데이터 set 20의 정확도는 그 다음 시점의 예측할 데이터가 없으므로 예측 정확도가 표에 포함되지 않았다.

예측 모델에 포함되는 규칙 set은 <표 10>에서와 같이 그 숫자가 증가한다. 데이터 set 1이 test 데이터 set으로 쓰였고 이것을 기준으로 다른 규칙 set의 예측 정확도를 측정하므로 이 데이터 set 1에서 추출된 규칙 set은 기본이 되는 규칙 set으로 보고 default로 사용한다. 따라서 데이터 set 2를 예측하는 시점에서는 test 데이터 set인 데이터 set 1에서 추출된 규칙 set 1을 사용하기로 한다.

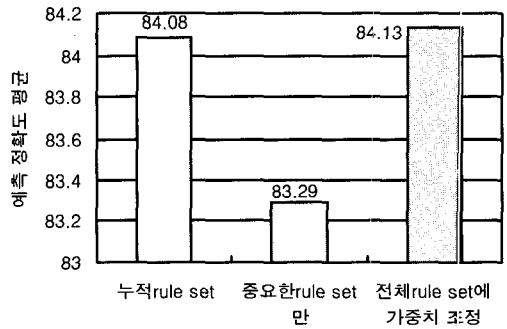
이렇게 중요도가 높은 규칙 set만으로 구성된 예측모델의 정확도는 <표 11>과 같으며 예측 정확도의 평균은 83.29%이다.

4.2.5 전체 규칙 set에 가중치를 적용한 예측모델

이번에는 예측 시점에서 중요도가 높은 규칙 set들로만 예측 모델을 구성하는 것이 아니라 전체 모든 규칙 set 집합으로 예측 모델을 구성하며 또한 중요도가 높은 규칙 set에 가중치를 주어 예측 모델을 구성하였다. 가중치는 시행 착오를 통해서 결정하였다. 실험 결과 가중치를 1.2를 줄 때 가장 예측 성과가 높았고 나머지 경우에도 예측의 정확도 평균이 0.5% 이상 변하지 않았다. 이 예측모델의 평가 결과는 <표 12>와 같다.

이 모델의 예측 정확도의 평균은 84.13%이다. 이것은 중요도가 높은 규칙 set들만으로 예측 모델을 구성하여 예측한 경우의 예측 정확도 83.29%보다 0.84% 향상된 수치이다. 따라서 중요한 규칙 set만으로 예측 모델을 구성한 경우보다 전체 규칙 set 집합에 모든 규칙 set들을 포함하고 중요도가 높은 규칙 set들에는 가중치를 주어 예측 모델을 구성하는 경우가 예측의 정확도가 가장 높았다.

중요도가 높은 규칙 set만으로 예측 모델을 구성한 경우, 전체 규칙 set 집합에 가중치를 달리 적용한 경우, 그리고 규칙 set을 가중치 없이 단순 누적한 경우 등 이 3가지 모델의 예측 정확도 비교를 [그림 7]이 보여준다.



〈그림 7〉 전체 규칙 set, 중요 규칙 set, 선택적 가중치 적용 경우 비교

〈표 11〉 중요도가 높은 규칙 set만으로 구성된 예측 모델로 D_{t+1}을 예측한 결과

데이터set	2	3	4	5	6	7	8	9	10	11
정확도(%)	80.5	83.5	83.25	83.75	84	82.75	81.75	80.25	84	84
데이터set	12	13	14	15	16	17	18	19	20	평균
정확도(%)	81.75	80.5	80.75	83.5	86.5	84	83.75	87.75	86.25	83.29

〈표 12〉 전체 규칙 set 집합에서 중요도가 높은 규칙 set에 가중치를 주어 예측한 결과

데이터set	2	3	4	5	6	7	8	9	10	11
정확도(%)	80.5	83.5	82.25	85.75	85.5	83.75	82.25	83.5	86.25	84
데이터set	12	13	14	15	16	17	18	19	20	평균
정확도(%)	83.75	80.75	82.25	84.5	86.5	85	84.25	88	86.25	84.13

이미 예측의 정확도가 향상된 규칙 set을 누적한 예측 모델을 더욱 효과적인 예측 모델로 만들기 위해서는 누적된 과거 정보를 삭제 없이 모두 누적하면서도 그 중에서 상대적으로 중요한 것을 찾아내어 가중치를 주어야 하고 이것을 사용하여 새로 들어오는 데이터를 더욱 효율적으로 예측할 수 있다는 것을 보여준다.

4.2.6 기타 통계분석 방법과의 비교

SEE5와 거기서 산출된 규칙을 누적한 방법 이외에 로지스틱 회귀분석, 판별분석, 신경망분석을 같은 데이터에 적용하여 보았다. 적용 방법은 SEE5 적용 시와 마찬가지이다. 데이터 set의 구분과 자료는 모두 동일하며 분석방법만 달리하여 직전 데이터 set으로 다음 데이터 set을 예측하였다. 이 결과는 <표 13>에서 보여지는 바와 같다.

19개의 데이터set에 대한 예측 정확도의 평균은 로지스틱 회귀분석이 80.81, 판별분석이 73.43, 신경망분석이 76.09이다. 이들은 모두 SEE5의 예측 정확도 평균인 82.24나 SEE5로 산출된 규칙을 누적해서 예측한 정확도의 평균인 84.13보다 낮은 것을 <표 14>에서 볼 수 있다.

〈표 13〉 기타 통계분석 결과

예측된 데이터 set	로지스틱 회귀분석	판별분석	신경망분석
2	79.12	76.30	74.25
3	83.08	79.30	78.25
4	82.80	73.80	74.75
5	81.81	69.32	73.25
6	84.03	69.80	77.75
7	80.80	77.84	75.50
8	79.30	74.01	72.75
9	80.80	66.30	77.03
10	82.50	65.30	81.75
11	76.50	72.01	74.51
12	80.80	70.50	76.04
13	75.30	75.81	73.25
14	77.50	69.50	73.75
15	81.50	77.83	75.75
16	80.02	76.80	75.50
17	84.04	75.32	76.03
18	82.02	75.80	78.06
19	82.03	77.31	79.75
20	81.50	72.50	78.04
예측정확도 평균	80.81	73.44	76.10

〈표 14〉 기타 통계분석과의 예측 정확도 비교

예측 기법	로지스틱 회귀분석	판별분석	신경망분석	SEE5	규칙누적
예측 정확도 평균	80.81	73.44	76.10	82.24	84.13

5. 결론 및 시사점

본 연구는 나날이 증가하고 복잡해져 가는 데이터를 효과적으로 관리하고 이용하여 정보를 추출하는 모델을 제시하였다. 본 연구에서 제시한 모델의 기여도는 아래의 4가지로 요약할 수 있다.

(1) 예측이 필요한 순간 미리 준비된 예측 모델을 즉시 이용할 수 있다. 기존의 방법으로 예측을 할 경우 현 시점까지 계속 누적되어온 데이터를 이용하여 예측 모델을 만들어야 한다. 허나 만약 누적된 데이터의 사이즈가 1GB 또는 그 이상일 경우 샘플링을 해야 하며 이때 샘플에 포함되지 않은 자료에 담겨있는 정보를 이용하지 못할 수 있다. 본 연구에서 제시하는 모델을 이용할 경우 데이터 크기에 이러한 제약을 받지 않는다.

(2) 새로운 데이터가 추가된 경우 모델의 update가 간편하고 빠르다. 새로운 데이터에서 추출한 규칙 set을 기본 모델에 추가 하면 된다. 허나 기존의 방법을 사용할 경우 새로운 데이터를 추가하여 처음부터 다시 예측 모델을 만들어야 한다. 데이터가 계속 실시간으로 들어오는 상황에서 빈번히 이 예측 모델을 사용해야 할 경우 이 같은 비효율성이 불가피하다.

(3) 데이터 저장 공간을 줄일 수 있다. 오늘날 데이터의 양은 방대하게 증가하고 있다. 만약 이 같은 데이터가 예측을 위해서만 저장이 필요하다면 본 연구에서 제시하듯이 연속으로 들어오는 데이터에서 필요한 정보만을 누적하기 때문에 데이터 저장 공간을 줄일 수 있다.

(4) 예측 정확도를 높일 수 있다. 데이터 크기가 아주 큰 경우 본 연구의 시험 결과에서 보여 주듯이 효율적인 규칙의 누적과 가중치 관리를 통해 다른 통계 분석 방법을 사용한 것보다 더 효율적이고 정확한 예측을 할 수 있다.

t+1시점의 데이터 set을 예측하기 위하여 t 시점의 데이터 set에서 추출한 규칙 set만 사용했을 때 예측 정확도의 평균이 82.24%인데 비해 룰 set을 누적하고 이에 가중치를 부여한 경우 예측 정확도의 평균이 84.08%로 향상되는 것을 볼 수 있다. 처음 시점부터 누적된 모든 데이터를 사용해서 t+1시점의 데이터 set을 예측하여 보았더니 예측 정확도가 소폭 상승한 것을 볼 수 있었다. 그러나 연속 발생 데이터의 경우에는 데이터가 무한히 늘어날 수 있다는 특성상 전체 데이터를 사용한다면 그에 따라 소모되는 메모리, 시간, 저장공간의 증가에 드는 비용이 예측 정확도가 소폭 상승하는 것과 비교할 수 없을 정도로 커질 것이며 이것은 또한 빠른 속도로 계속 증가할 것이다. 결국 누적된 데이터의 사이즈가 아주 커질 경우 이 모든 데이터를 한번에 읽는 것이 불가능해져 sampling이 불가피해진다. 따라서 처음 시점부터 모든 데이터를 누적하고 예측 모델을 생성하는데 드는 저장공간, 메모리, 시간을 고려하여 보았을 때 전체 데이터를 누적하여 사용하는 대신 규칙 set 집합을 누적하여 사용하는 것이 바람직하다.

중요한 규칙 set만으로 예측 모델을 구성한 경우와 전체 규칙 set을 포함하여 예측 모델을 구성하면서 중요한 규칙 set에 가중치를 주는 경우의 예측 정확도 평균은 시험 결과 후자의 경우가 더 좋은 결과를 보여 주었다. 어떠한 규칙 set을 제거하는 것보다는 가중치를 조정하여 전체 규칙 set을 누적하는 예측 모델을 구성하는 것이 예측 정확도가 높았다. 다시 말하면 어떠한 추세에 맞추어 규칙을 삭제 혹은 추가하는 모델을 만드는 것보다는 자료의 선택이나 삭제 없이 각 룰의 영향력만을 조정하는 모델이 더욱 예측 정확도가 높았다.

누적된 규칙 set 집합을 사용했을 때 예측 정확도가 향상된 것은 규칙이 누적되면서 한 tuple이 서로 다른 규칙 set에서 생성된 여러 개의 규칙 조건을 만족시키게 되고 이 결과 더 많이 예측되는 class의 빈도에 따라 결과값이 나오기 때문이다. 다시 말하면 규칙이 누적되면서 데이터의 흐름을 더욱 잘 반영하는 방향으로 과거의 정보를 저장한다고 할 수 있으며, 이러한 경우 단순히 직전 데이터 set에서 얻어진

정보에 의해 예측하는 것보다 과거부터 누적되어 온 정보에 의해 예측하는 것이 예측의 정확도를 높이는 데 기여하였다고 결론 내릴 수 있다. 시간의 흐름에 따라 계속적으로 들어오는 연속 발생 데이터는 시간의 변화에 영향을 많이 받게 되므로 시간의 흐름에 대한 정보를 어떤 식으로 관리하는가가 연속 발생 데이터 마이닝의 예측 정확도를 높이는 중요한 요소가 된다. 어떠한 구체적인 패턴을 가지고 있고 이것의 빈도수 자체를 유지하고 있는 방법의 경우 새로운 패턴이 생기면 이러한 새로운 패턴을 계속 추가시키면서 모든 패턴의 발생 빈도수를 늘 변화시켜주어야 한다. 또한 발생 빈도수에 따라서 빈번하게 발생하지 않는 패턴은 삭제해 주어야 할 것이다. 그러나 규칙을 누적 시키게 되면 그 발생 빈도를 변화시켜 주지 않아도 규칙이 쌓이면서 각각의 규칙에 의해 예측되는 class의 빈도가 변하게 되고 이에 따라 예측 결과값을 결정할 수 있으므로 누적만으로도, 인위적인 패턴의 추가 혹은 삭제 없이 자주 발생하는 패턴을 변화시켜주는 효과를 가져오게 된다. 또한 규칙 set 집합의 저장은 이 규칙이 얻어진 데이터 set 자체를 저장하는 것에 비해 현저하게 저장공간이 감소되므로 연속 발생 데이터와 같이 잠재적으로 무한하게 들어오는 데이터를 마이닝 하는 경우 발생할 수 있는 저장 공간의 문제를 해결할 수 있다. 또한 규칙 set 집합을 정해진 사이즈의 데이터 set에서 마이닝하는 것은 일정한 크기의 데이터 set을 대상으로 하므로 방대한 전체 데이터를 마이닝하는데 드는 시간, 메모리의 문제도 해결할 수 있다.

이러한 규칙 set 집합 마이닝을 더욱 효과적으로 하기 위하여 특정 정보를 기준으로 삼고 그것에 따라 예측하려고 하는 데이터 set과 비슷한 특징을 가지고 있는 기존의 데이터 set을 찾아서 그것을 다른 데이터 set들보다 상대적으로 더 중요한 데이터 set으로 다룰 수 있다. 본 연구에서는 규칙 set이 데이터 set의 정보를 가지고 있음으로 데이터 set 자체 대신 규칙 set들끼리 상대적 중요도를 따져야 할 것이다. 실험 결과를 보면 전체 규칙 set 집합을 사용하되 상대적으로 더 중요하게 판단된 규칙 set과 그렇지 않은 규칙 set들에 가중치를 달리 적용하였을 때가 중요한 규칙 set들로만 예측 모델을 구성한 경우나, 가중치 조정 없이 전체규칙 set 집합을 사용한 경우보다 예측의 정확도가 높았다.

중요한 규칙 set만으로 예측 모델을 구성한 경우 예측 정확도가 떨어지는 것은 상대적으로 덜 중요하다고 판단된 규칙 set의 영향력을 완전히 배제하였기 때문이다. 비록 상대적으로 덜 중요하다고 판단 될 지라도 이것의 영향력을 완전히 배제하는 것은 그 규칙 set이 잠재적으로 가지고 있는 정보의 역량을 완전히 무시하게 되는 결과를 낳는 것이다. 또한 규칙 set 집합을 가중치 적용 없이 누적한 경우보다 가중치를 적용할 때 예측의 정확도가 높아지는 것은 상대적으로 덜 중요한 규칙 set의 영향력을 배제하지 않으면서도 더 중요하다고 판단된 규칙 set의 영향력은 높이는 것이 더 효과적 이라는 것을 검증하였다.

참 고 문 헌

[1] Agrawal, R. and R. Srikant, "Fast algorithms for mining association rules," *VLDB '94*, Sept., 1994.

[2] Agrawal, R. and R. Srikant, "Mining sequential patterns, *Proceedings of 1995 International Conference of Data Engineering*, (1995), pp.3-14.

[3] Ayan, N.F., A.U. Tansel, and M.E. Arkun, "An efficient algorithm to update large item sets with early pruning," *Proceedings of the Fifth CM SIGKDD, International Conference on Knowledge Discovery and Data Mining*, (1999), pp.287-291.

[4] Cheung, D.W., J. Hand, V. Ng, and C.Y. Wong, "Maintenance of discovered association rules in large databases : An incremental updating technique," *Proceedings of the Twelfth International Conferen-*

ce on Data Engineering, (1996), pp.106-114.

[5] Cheung, W. and O.R. Zaiane, "Incremental Mining of Frequent Patterns Without Candidate Generation or Support Constraint," *Proceedings of the 7th International Database Engineering and Applications Symposium*(IDEAS'03), 2003.

[6] Domingos, P. and G. Hulten, "Mining High-Speed Data Streams," *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2000), pp.71-80.

[7] Ganti, V., J. Gehrke, and R. Ramakrishnan, "DEMON : Mining and monitoring evolving data," *Proceedings of the Sixteenth International Conference on Data Engineering*, (2000), pp.439-448.

[8] Ganti, V., J. Gehrke, and R. Ramakrishnan, "Mining Data Streams under Block Evolution," *SIGKDD* Explorations, Vol.3, No.2 (2002), pp.1-10.

[9] Gehrke, J.V., Ganti, R. Ramakrishnan, and W.-L. Loh, "BOAT : optimistic decision tree construction," *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, (1999), pp.169-180.

[10] Giannella, C., J. Han, J. Pei, and X. Yan, "Mining Frequent Patterns in Data Streams at Multiple Time Granularities," Chapter3, Kargupta H., A. Joshi, K. Sivakumar and Y. Yesha(eds.), *Next Generation Data Mining*, MIT Press, 2003.

[11] Greenwald, M. and S. Khanna, "Space-Efficient On-line Computation of Quantile Summaries," *Proceedings of ACM SIG MOD*, 2001.

[12] Guha, S., N. Mishra, R. Motwani, and L. O'Callagan, "Clustering Data Streams," *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, 2000.

[13] Han, J. and M, Kamber, Data "Mining Concepts and Techniques," *Morgan Kaufmann*, 2001.

[14] Han, J., J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," *Proceedings of 2000 ACM-SIGMOD International Conference of Management of Data*(2000), pp.1-12.

[15] Hidber, C., "Online Associattion Rule Mining," *Proceedings of ACM SIGMOD*, (1999), pp.145-156.

[16] Hulten, G., L. Spencer, and P. Domingos, Mining Time-Changing Data Streams *KDD*, 01, 2001.

[17] Pei, J., J. Han, and R. Mao, "CLOSET : An efficient algorithm for mining frequent closed item sets," *Proceedings of 2000, ACM-SIGMOD International workshop of Data Mining and Knowledge Discovery*, (2000), pp.11-20.

[18] Wang, H., W. Fan, P.S. Yu, and J. Han, "Mining Concept-Drifting Data Streams Using Ensemble Classifiers," *Proceedings of ACM-SIGKDD'03*, (August 2003), pp. 24-27.

[19] Zaki, M.J. and C.J. Hsiao, "CHARM : An efficient algorithm for closed itern set mining," *Proceedings of 2002 SIAM*, 2002.