지능정보연구 제17권 제4호 2011년 12월(pp.109~130)

지식 누적을 이용한 실시간 주식시장 예측

김진화
서강대학교 경영학과
(jinhwakim@sogang.ac.kr)

홍광헌
서강대학교 경영학과
(khhong@sogang.ac.kr)

민진영
The Fox School of Business and Management, Temple University
(saharamin@gmail.com)

연속발생 데이터는 데이터의 원천으로부터 데이터 저장소로 연속적으로 축적이 되는 데이터를 말한다. 이렇게 축적된 데이터의 크기는 시간이 지남에 따라 점점 커진다. 또한 이러한 대용량 데이터에서 정보를 추출하기 위해서는 저장공간, 시간, 그리고 많은 자원이 필요하다. 이러한 연속발생 데이터의 특성은 시간이 지남에 따라 축적된 대용량 데이터의 이용을 어렵고 고비용이 되게 한다. 만약 정보나 패턴을 추출할 때 누적된 전체 발생 데이터 중에서 최근의 일부만 사용 한다면 적은 일부 표본의 사용의 문제로 인하여 전체 데이터 사용에서 발견될 수 있는 유용한 정보의 유실이 있을 수 있다. 이러한 문제점을 해결하기 위해서 본 연구는 연속발생 데이터를 발생 시점에서 계속 모으기 보다 이러한 발생되는 데이터에서 규칙을 추출하여 효율적으로 지식을 관리하고자 한다. 이 방법은 기존의 방법에 비하여 적은 양의 데이터 저장공간을 필요로 한다. 또한 이렇게 축적된 규칙집합은 미래에 예측을 위해서 언제든 실시간 예측을 할 수 있게 준비가 된다. 여러 예측 모델을 결합시키는 방법인 앙상블 이론에 의하면 본 연구가 제시하는 데로 체계적으로 규칙집합을 시간에 따라 융합시킬 경우 더 나은 예측 성과가 가능하다. 본 연구는 주식시장의 변동성을 예측하기 위하여 주식시장 데이터를 사용하였다. 본 연구는 이 데이터를 이용해 본 연구가 제시하는 방법과 기존의 방법의 예측 정확도를 비교 하였다.

논문접수일 : 2011년 08월 09일 게재확정일 : 2011년 08월 16일
투고유형 : 학술대회우수논문 교신저자 : 김진화

1. 서론

주식시장에서의 주식 가격의 움직임을 예측하려는 시도는 본격적으로 주식이 시장에서 거래되기 시작한 시점부터 진행되어 왔으며, 현재에도 많은 시도가 진행되고 있음을 부인할 없다. 주식가격의 움직임을 예측하려는 분석은 크게 기본적 분석과 기술적 분석으로 분류할 수 있다. 기본적 분석은 기업의 내재적 가치에 영향을 미치는 요소들을 분석함으로써 주식의 움직임을 예측하려는 시도로서, 기업의 내적인 요인들에 대한 분석과 함께 기업의 외적인 요인들에 대한 분석인 산업분석과 경제분석을 포괄적으로 진행하는 것이 일반적이다. 기술적 분석은 과거의 주식가격의 움직임과 주식의 거래와 관련되는 변수들을 이용하여 미래 주식가격의 변동을 예측하려는 시도를 일컫는다. 이러한 분석은 주식의 수요와 공급에 대한 변화는 주식시장의 움직임과 관련되는 변수에 관찰될 것

* 이 논문은 2009년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2009-327-B00214).

이고, 이러한 변수를 분석함으로써 궁극적으로 미래 주식 가격의 변화를 파악할 수 있다는 믿음에 기초하고 있다. 이러한 믿음에 기초하여 미국을 비롯한 여러 나라의 주식시장 참여자들은, 기술적 분석 방법에 의한 실증분석을 통하여 주식시장의 움직임에 대한 예측을 생산해 내고 있다.

과거 주식가격의 변화를 통하여 미래 주식가격의 움직임을 예측한다는 것이 어려울 뿐만 아니라, 비록 주식의 수익률이 예측을 통하여 얻어진다 하더라도 이는 경제적으로나 통계적으로 유의적이지 않다는 전통적인 주장과는 달리, 주식가격의 변화는 단순한 랜덤워크를 따르는 것이 아니고 주식의 수익률을 측정하는 기간에 따라서 수익률은 음 또는 양의 자기 상관관계(Serial-correlation)를 나타내고 있음을 보임으로써 어느 정도의 예측이 가능하다는 새로운 주장이 대두 하게 되었다(Jegadeesh, 1990). 그는 두 달의 시차를 두고 수익률을 측정하였을 경우에는 음의 상관 관계가 관찰되지만, 더 이상의 시차를 두고 측정한 수익률에서는 양의 자기상관 관계가 관찰되었음을 발견하였다. 또한 다른 연구에서의 주식시장 자료를 이용하여 보고한 결과에 따르면, 수익률을 월별로 측정하였을 경우에는 일반적으로 양의 자기 상관관계를 관찰 할 수 있으나, 수익률을 3년에서 5년의 시차를 두고 측정하였을 경우에는, 음의 자기 상관관계를 보인다는 것이다(Cutler et al., 1990).

주식시장의 예측 가능성에 대한 경제적 유효성(economic significance)을 평가하는 방법으로서는, 예측 가능성을 주장하는 실증적 분석의 결과에 기초한 투자 전략을 구사하였을 경우에 실제적으로 경제적으로 충분히 유효한 초과 수익률을 얻을 수 있는지를 살펴보는 것이다. 이러한 접근방법의 장점은, 투자결정이 순전히 과거에 모든 투자자들에게 제공될 수 있었던 과거의 정보에 기초하고 있다는 사실에 있다. 그러나, 이와는 상반되게 어떤 요소에 대한 정보가 주식시장의 움직임을 예측하는데 중요한 영향을 미치는지에 대해서는 설명하지 못하고 있다는 결정적인 단점을 보이고 있다. 이러한 단점을 극복하기 위한 하나의 방안으로, 주식시장의 움직임에 영향을 미칠 수 있으리라고 예측되는 요소들을 사전적으로 추출하여 이러한 요소들에 대한 과거의 정보를 이용하여 투자전략을 수립한 후에 그 전략이 체계적으로 초과수익률을 창출할 수 있을 만큼 경제적으로 유효한지를 살펴보는 것이다. 이러한 투자전략을 구사함에 있어서 유의해야 할 점은 후 판단 편의에 빠질 수 있다는 것이다. 즉 사전적으로는 요소들의 중요성을 파악하지 못하다가, 사후적으로 전체자료에 기초한 연구결과를 파악한 후에는 그 요소들에 대해 중요성을 부여하게 되고 이러한 요소에 기초하여 투자전략을 수립하고자 하는 오류를 의미한다. 그러므로 이러한 편의가 없이 주식수익률의 예측을 시도하고자 하는 투자전략을 수립하고 구사하는 것이 중요하다. 즉, 투자자들에게 제공될 수 있는 모든 정보와 모든 과거자료를 이용하여 투자전략을 수립한다는 것은 현실적으로 불가능하며, 또한 모든 자료에 기초하여 측정된 변수의 추정치를 구한다는 것도 어렵다(Pesaran and Timmerman, 1995).

이러한 관점에서 볼 때, 데이터 속에 숨겨져 있는 어떠한 흥미로운 패턴들을 찾아내는 것은 데이터마이닝에서 중요한 부분을 차지한다. 이 패턴이라는 것은 데이터 베이스 안에서 발견되는 조합 혹은 분류 모델이나 순차적 경향들로서 데이터마이닝의 기본이 되는 것이다(Han and Kamber, 2001). 패턴을 마이닝하는 방법은 자료가 고정되어 있지 않고 시간이 흐름에 따라 변하는 경우 단순히 빈번하게 발생하는 패턴을 찾아내는 것보다 매우 복잡해지게 된다. 유한하고 통계적으로 고정되어 있

는 데이터와 반대되는 개념으로, 이렇게 시간의 흐름에 따라 변하는 연속적이고 잠재적으로 무한히 발생하는 특징을 가지고 있는 데이터를 연속 발생 데이터라고 한다. 이런 예로는 네트웍 트래픽 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 되고, 빈번히 나오지 않던 패턴이 자료가 추가됨에 따라 빈번하게 나올 수도 있다는 것이다. 다시 말하면 연속 발생 데이터의 마이닝에서는 패턴의 추출 자체가 문제가 아니라 빈번한 패턴들의 변화를 어떻게 다룰지가 더욱 중요한 문제가 된다는 것이다. 이렇게 자료가 고정되어 있지 않고 계속 증가하는 경우에 마이닝하는 대상 자료의 변화를 다루는 마이닝 방법이 점진적 데이터마이닝이다. 점진적 데이터마이닝이 성공적으로 수행되기 위한 기본 조건으로는, 시간이 지남에 따라 지속적으로 증가하는 데이터를 관리하기 위한 물리적인 저장 공간 또한 계속 늘어나야 한다는 점이다. 그러나 방대한 양의 데이터가 계속 늘어나는 경우 거기에 맞추어 계속 저장 공간을 늘리는 것은 물리적인 제한이 따르거나 혹은 엄청난 비용을 수반할 것이다. 이러한 시간의 흐름에 따른 변화와 저장공간, 자원 소요의 문제까지 고려해서 연속 발생 데이터에 알맞은 마이닝 방법이 선택되면 이로부터 자료간의 조합이나 관련성을 밝히고 또한 효과적으로 새로운 데이터를 예측할 수 있게 되는 것이다. 따라서 여기서는 시간의 흐름에 따라 변하는 데이터를 대상으로 보다 효율적으로 데이터마이닝을 하는 방법을 적용하여 새로운 데이터가 들어왔을 때에 효과적으로 예측하는 방법을 살펴보려 한다.

2. 기존연구

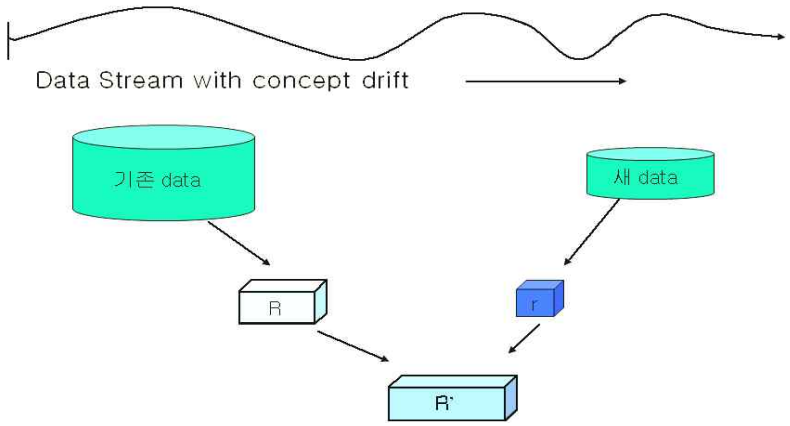
점진적 데이터마이닝의 기본개념은 데이터베이스에 대한 정보를 R이라고 이름 붙여진 형태로 유지하고 있다가 새 데이터가 들어오면 이것을 기존의 R과 형태가 같은 정보를 뽑아낼 수 있도록 만들고 거기에서 r을 뽑아낸 후 이것을 현재 가지고 있는 R과 합쳐서 새로운 R을 만들어내는 것이다. 즉 R∪r을 통해 R'를 만들어내는 것에 있다(Han and Kamber, 2001). 연속 발생 데이터를 분류하거나 예측하기 위해서도 패턴의 마이닝 자체뿐 만이 아니라 패턴들을 데이터 set의 추가에 맞추어 변화시켜야만 이 패턴들을 통해서 효과적인 분류, 혹은 예측을 할 수 있게 되는 것이다. 따라서 이 경우에도 점진적 데이터 패턴 마이닝이 중요하다(Cheung et al., 1996; Domingos and Hulten, 2000). 점진적 데이터마이닝을 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 작았기 때문에 문제가 발생되고는 했지만 현대에는 데이터의 사이즈가 너무 커지는 데에서 문제가 생길 수 있다(Domingos and Hulten, 2000). 시간의 흐름에 따라 방대한 양의 데이터가 들어오는 연속 발생 데이터마이닝의 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. 연속 발생 데이터는 잠재적으로 무한한 데이터가 들어오기 때문에 이러한 데이터를 모두 저장한다면 많은 저장 공간이 필요하며 이러한 저장 공간이 다 차고도 연속 발

생 데이터가 계속 들어온다면 결국에는 계속해서 저장공간을 늘려주어야만 한다. 또한 한 번에 메모리에서 읽을 수 있는 데이터의 양이 고정되어 있으므로 매우 많은 양의 데이터에 대해 분석을 시행하려면 상당한 시간이 소요될 것이다. 이러한 문제 때문에 알고리즘이 과거에 읽어진 지나치게 많은 데이터를 모두 기억하는 일이 불가능해진다. 따라서 방대한 데이터를 저장하는데 필요한 공간상의 문제를 해결하기 위한 연구들이 행해지고 있다(Giannella et al., 2003).

연속 발생 데이터의 특징인 데이터의 방대함은 또한 이러한 방대한 데이터를 어떻게 효율적으로 마이닝할 것인지, 데이터의 변화를 어떻게 다룰 것인지에 대한 다양한 연구를 낳았다(Cheung and Zaiane, 2003; Domingos and Hulten, 2000). Domingos and Hulten(2000)은 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 의사결정 트리를 만들었다(Cutler et al., 1990; Greenwald and Khanna, 2001). 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. 이들 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 저장공간과 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 데이터마이닝의 주요 목표가 되고 있다.

패턴을 마이닝하는 것은 앞에서도 언급했듯이, 자료의 조합과 분류를 밝히는 것 외에도 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 패턴으로 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 다루어 정확한 값을 예측하는 것이 중요하다. 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 데이터마이닝으로 인해 지나치게 사용이 복잡해지는 것을 막는 사용의 용이함이 요구된다(Wang et al., 2003). 아래 <그림 1>은 같이 스트림 데이터를 이용한 incremental mining의 개념을 보여 주고 있다.

본 연구에서는 주식 시장 자료를 통해 실험을 한다. 주식시장을 연구한 기존의 연구 에서는 Neural Network 등의 예측방법을 택하였다. 본 연구에서는 생성된 규칙을 통해서 새로 들어오는 데이터의



<그림 1> 스트림 데이터와 Incremental Mining

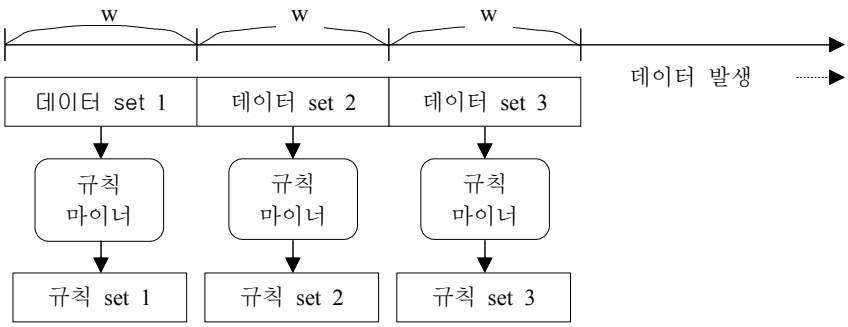
알려지지 않은 목표 class의 값이 예측된다고 가정하고 이러한 규칙 자체를 변경하거나 삭제하지 않고 누적함으로서 전체 데이터에 대한 정보를 유지하고 이것을 통해 효과적인 예측을 할 수 있는지 연구하려 한다. 이 경우에는 일정 부분의 자료를 부분적으로 선택하여 분석대상으로 하는 것이 아니라, 전체 데이터에서 추출된 정보를 규칙이라는 형태로 누적하여 사용하고 어떠한 정보를 취사 선택하지 않으므로 데이터의 손실이 없고 과거의 데이터에 대한 정보도 그대로 유지하므로, 시간의 흐름에 따른 패턴의 변화 또한 처리할 수 있을 것이다. 따라서 데이터에 주기적인 변화가 일어나거나 오랜 시간이 지난 후에 과거와 유사한 패턴이 발생하는 경우라도 누적된 정보를 통해 효과적으로 예측하는 것이 가능할 것이다.

이러한 방법이 효과적으로 시행되어 전체 데이터를 다 사용하는 기존의 일괄처리 방법이 제공하는 것, 혹은 그 이상의 효과를 가져 온다면, 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 갱신시키는 것보다 훨씬 간편하게 데이터를 마이닝하고 예측 할 수 있을 것이다.

3. 연구모형

3.1 규칙 set 누적

보통의 연관 규칙은 앞에서도 얘기했듯이 X => Y 형태를 띠고 있다. 연속 발생 데이터의 각 데이터 set에서는 이러한 규칙들을 추출하여 규칙 set을 만들 수가 있다. 연속 발생 데이터를 데이터 set 단위로 마이닝해서 규칙 set을 추출하는 과정은 <그림 2>와 같다. w는 데이터 set을 만들어주는 기준이 되는 크기인 윈도우 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 될 것이다. 여기에서는 연속 발생 데이터가 들어와서 어느 정도 일정 크기가 되면 이것을 하



<그림 2> 데이터 set 단위의 연속 발생 데이터에서 규칙 set을 추출하는 과정

나의 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. 그리고 또 다시 일정 크기의 데이터가 들어와 미리 정해진 데이터 set의 크기가 되면 다음 데이터 set이 들어왔다고 보고 여기에서도 규칙들을 추출하여 규칙 set을 확정 한 후 이전에 만들어진 규칙 set 집합을 수정하고 역시 그 데이터 set 자체를 저장하지 않는다. 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.

저장 공간이 제약되어 있다면 새로운 데이터를 효과적으로 예측하기 위해 전체 데이터를 저장하기 보다는 어떠한 선택 기준에 따라 일정한 데이터만을 저장하는 것을 선호할 수 있을 것이다. 예를 들어 늘 최신 데이터가 새로운 데이터를 예측하는데 중요하게 여겨진다면 연속 발생 데이터가 계속 들어옴에 따라 저장공간의 제약 혹은 흥미도의 저하로 인해 오래된 데이터들은 버리고 비교적 최신의 데이터들만을 유지할 수밖에 없을 것이다. 그렇게 오래된 데이터들이 버려진다면 예측하려고 하는 데이터 set과 시간적으로 떨어져 있는 데이터들에서 추출되는 정보는 예측에 이용되지 못할 것이다 그러나 데이터가 추가됨에 따라 전체 데이터의 성격이 계속적으로 변할 수도 있는 연속 발생 데이터의 특성을 생각해 보았을 때 예측하려고 하는 정보가 과거의 데이터와 예기치 않게 밀접하게 연관이 되어 있는 경우라면 이러한 정보의 손실은 원치 않는 결과를 낳게 될 것이다. 예를 들어 어떠한 이유로 인해 늘 한 해 분의 데이터만을 저장하고 있다면 계절적 수요가 발생할 경우에도 과거 정보를 가지고 있지 않으므로 그것이 계절적 수요, 즉 특정 시기에 빈번하게 발생하는 패턴이라는 것을 인지하지 못하게 되어 현재 데이터의 흐름을 적절하게 따라잡지 못하게 될 것이다. 따라서 새로운 데이터를 예측할 때 마다 어느 정도 시점 이전의 정보는 가지고 있지 않게 되는 특정 시점의 데이터들에서 추출된 규칙들로 새로 들어오는 데이터를 예측하는 방법을 사용하거나 저장 공간과 메모리, 시간이 지나치게 많이 드는 전체 데이터를 가지고 새로 들어오는 데이터를 예측하는 방

법을 사용하는 것보다 상대적으로 크기가 작으면서도 최초 데이터들에 대한 정보까지도 가지고 있는 규칙 set 집합으로 새로 들어오는 데이터를 예측하는 것이 더욱 효율적일 것이다.

4. 주식 시장 자료에의 적용

본 연구의 실험을 위해 주식시장 자료를 수집하였다. 이 자료는 시간의 흐름에 따라 매우 변화가 심한 자료로, 따라서 데이터 set들간 자료 구조의 차이가 크다. 여기서는 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 누적한 규칙 set들로 새 데이터 set을 예측하는 경우를 비교하려 한다. 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 비교하려 한다.

4.1 자료의 기초분석과 데이터 set 설정

1998년 1월 2일에서 2003년 11월 28일까지 일별로 1541개의 tuple을 가지는 데이터로서 Slow %K, Slow %D, ROC, William %R, CCI를 입력 변수로 종합주가지수의 변화를 예측하는 데이터이다. 예측 변수로 사용한 종합주가지수의 변화는 지수를 다음날과 비교하여 다음날의 지수가 오를 때는 1, 같거나 내릴 때는 0으로 표시하였다. 변수에 대한 설명은 <표 1>과 같다(최세일, 1999; Achelis,

<표 1> 입력변수, 예측변수 설명

입력변수		
Slow %K	현재의 주가가 해당기간 동안의 주가 범위 중 어느 위치에 있느냐를 분석하여 향후 주가 방향을 예측하는 기법	$\frac{(\text{당일 종가})-(\text{해당기간 최저가})}{(\text{해당기간 최고가})-(\text{해당기간 최저가})}\times 100$
Slow %D		$\frac{\sum_{i=1}^{n-1}\%K_{t-i}}{n}$
ROC (Price Rate Of Change)	일정 시점의 가격 변화율을 백분율로 나타낸 것	$\frac{(\text{당일 종가})}{(\text{해당기간 종가})}\times 100$
William %R	과매도, 과매수를 측정하기 위한 모멘텀 지표	$\frac{(\text{해당기간 최고가})-(\text{당일 종가})}{(\text{해당기간 최고가})-(\text{해당기간 최저가})}\times(-100)$
CCI(Commodity Channel Index)	이동 평균으로부터 주가의 변동을 측정하는 지표	$\frac{M-m}{d\times 0.015}$ M : (고가+저가+종가)/3 m : M의 해당기간(n) 단순 이동 평균 값 d : M과 m사이 편차절대값을 단순 이동 평균한 값 $\frac{\sum_{t=1}^{n}\lvert(M-m)\rvert}{n}$
예측변수		
종합주가지수 변화(0, 1)	종합주가지수의 변화 여부	T시점에서 t+1시점과 비교하여 종합주가지수가 오르면 1, 떨어지거나 변화 없으면 0

1995).

1998년 1월 2일에서 100일 동안의 데이터를 100개의 tuple을 가진 데이터 set 1로 하고 그 이후 100일의 자료를 또 다른 100개의 tuple을 가진 데이터 set 2로 설정하는 방법으로 1541개의 데이터 tuple을 100개 단위로 순차적으로 데이터 set을 설정하였다. 이와 같이 데이터 set을 나눈 결과 100개의 tuple을 가진 데이터 set 1에서 15까지 15개의 데이터 set과 나머지 41개의 tuple을 가진 16번째의 데이터 set, 도합 16개의 데이터 set이 나왔다.

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 예측하는 경우이다. SEE5를 사용하여 추출된 규칙의 예는 <표 2>와 같다.

<표 2> SEE5를 통해 추출된 규칙의 예

William %R <= -13.27	→ class 0
William %R <= -15.01	→ class 0
William %R > -92.31	→ class 1

4.2 데이터 set t에서 추출된 규칙 set으로 데이터 set t+1 예측

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 방법의 예측 정확도를 나타내는 것이 <표 3>이다. 앞에서 규칙을 추출하여 규칙 set으로 만드는 과정에서 명시하였듯이 R_t는 t시점에서의 규칙 set, D_t는 t시점에서의 데이터 set, D_{t+1}는 t+1시점에서의 데이터 set, $\sum_{i=1}^{t} R_i$는 t시점까지의 규칙 set들의 집합을 의미한다.

<표 3> R_t로 D_{t+1}을 예측

데이터set	2	3	4	5	6	7	8	9
정확도(%)	70	67	69	50	65	56	62	64
데이터set	10	11	12	13	14	15	16	평균
정확도(%)	66	62	59	51	59	62	58.5	61.37

예를 들어 데이터 set 5의 예측 정확도를 살펴보면 데이터 set 5에서 추출된 규칙으로 데이터 set 6의 데이터를 예측했을 경우 100개의 tuple중 65개는 실제 그 tuple이 가지고 있는 분류 class값과 같은 값을 예측했으며 35개는 그렇지 못했다는 것을 의미한다. 위와 같이 두 번째에서 열 여섯번째까지의 데이터 set을 예측한 예측 정확도의 평균을 구하였더니 61.37%가 나왔다.

4.3 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 높일 수 있다는 가정에 따라 규칙을 누적하고 예측 정확도를 측정하였다. 이 시험을 위하여 SEE5에서 추출된 규칙의 형태를 JAVA를 이용하여 짠 프로그래밍에 알맞게 변형하고 예측할 데이터 set을 이 프로그램의 입력자료로 준 후 예측의 정확도가 출력되도록 하였다.

SEE5를 통해서 다음과 같이 추출된 규칙을 사용하여 만들어진 규칙 set 집합에서는 한 tuple이 여러 개의 규칙에 해당될 수 있고 또한 서로 다른 class로 예측될 수 있다

<표 4> 규칙 set 집합의 예

데이터 set	각 데이터 set에서 추출된 규칙의 일부	
1	if ((William %R <= -13.27))	class0.hit = class0.hit+1
2	if ((William %R <= -15.01))	class0.hit = class0.hit+1
3	if ((William %R > -92.31))	class1.hit = class1.hit+1

<표 4>는 각 데이터 set에서 추출된 규칙 set의 일부를 보여주고 있다. 만약 현재 예측하려는 주식시장 자료 tuple의 William %R의 값이 -30일 경우 데이터 set 1, 2에서는 위와 같이 -30이 데이터 set 1의 규칙인 -13.27보다 작아서 class 0으로 예측하고, 데이터 set 2의 규칙을 보면 -30이 -15.01보다 역시 작기 때문에 또한 class 0으로 예측한다. 하지만 데이터 set 3의 규칙을 보면 -30이 -92.31보다 커서 class 1로 예측하고 있다. 이렇게 한 tuple이 서로 다른 class로 예측된다고 해도 이 tuple의 값은 class 0으로 두 번, class 1로 한 번의 값이 예측되었기 때문에 각각의 class로 예측된 빈도수를 비교하면 2 : 1로 class 0이 우세하므로 이 tuple은 class 0으로 예측 될 것이다. 이 경우에 데이터 set 3의 규칙이 기존의 데이터 분포와는 다르게 특이값의 영향 하에서 만들어졌다고 하더라도 과거의 규칙과의 조합을 통해 그 영향을 줄일 수 있는 것이다. 예를 들어 현재 예측하려고 하는 tuple이 데이터 set 4의 tuple이고 직전 데이터 set에서 추출된 규칙만을 사용하는 방법을 써서 데이터 set 3에서 추출된 규칙만을 사용한다고 하자. 이 경우에는 William %R 값이 -30을 가지고 있는 tuple이 데이터 set 3의 규칙에만 해당되기 때문에 class 1로 예측될 것이고 규칙을 누적 시킨 위와 방법과는 다른 결과를 낳게 될 것이다. 이 방법은 만약 데이터가 급격히 변해서 과거 데이터는 모조리 쓸모 없게 된 경우에는 의미가 있겠지만 단 한 번의 변화로 데이터의 성격이 모두 변했다고 하는 것은 섣부른 판단일 수 있으며 또한 이것이 특이값에 의한 결과였을 경우는 잘못된 예측을 낳게 될 것이다. 결국 이것은 최근 데이터에 중점을 두고 과거의 데이터는 무시하는 결과를 낳게 되는데 시간에 따라 변화하면서 계속적으로 들어오는 연속 발생 데이터의 마이닝에 있어서는 과거 데이터를 무시하는 경우 예측도를 높이는데 도움이 되지 않는 결과를 낳을 수 있다. 또한 과거 데이터가 주기적으로 변한다면 과거 데이터의 중요성을 간과할 수 없을 것이다. 따라서 과거의 데이터에서 얻어진 정보도 예측에 반영하기 위해 누적된 규칙 set 집합을 사용하여 예측하였다. 이 과정에서 만약 class 0과 class 1로 예측된 빈도수가 같다면 default값을 정하여 예측하도록 하였다. 여기서는 0이 종합주가지수가 떨어지거나 변하지 않는 것을 의미하고 1이 종합주가지수가 오르는 것을 의미하므로 class 0과 class 1이 같은 빈도수로 나와 어느 한쪽으로 예측할 수 없다는 것은 지수가 오르거나 떨어지지 않는다는 것, 즉 변하지 않는다는 것과 의미상 더 가깝다고 판단할 수 있다고 생각되었기 때문에 default class를 class 0으로 정하였다.

여기서는 규칙을 누적하기 때문에 시간이 지날수록 예측 시점에서 쓰이는 규칙의 수는 늘어나게 된다. 그러나 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 무작정 규칙의 개수가 많기만 해서는 예측 정확도를 높일 수 없을 것이다. 여기서는 규칙이 누적되면서 이것들끼리 모순이 생겨 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 향상됨으로서 뒷받침된다. $\sum_{i=1}^{t} R_i$으로 D_{t+1}을 예측한 것이 <표 5>이다.

<표 5> $\sum_{i=1}^{t} R_i$로 D_{t+1}을 예측

데이터set	2	3	4	5	6	7	8	9
정확도(%)	60	72	69	67	64	64	65	65
데이터set	10	11	12	13	14	15	16	평균
정확도(%)	72	77	63	70	75	71	70.7	68.31

규칙을 누적하여 $\sum_{i=1}^{t} R_i$로 D_{t+1}을 예측하였더니 예측 정확도의 평균이 68.31%가 되었다. 이것은 단순히 R_t로 D_{t+1}을 예측하였을 때 예측 정확도의 평균이 61.37%였던 것에 비해 6.94%가 향상된 수치이다. 따라서 t+1시점의 데이터 set을 예측하기 위하여 단순히 t시점에서 추출된 규칙 set을 적용하는 것보다 t시점까지 누적된 규칙을 적용하는 것이 연속 발생 데이터와 같이 시간의 순서에 따라 데이터의 개념이 계속적으로 변화하면서 들어오는 데이터의 경우 예측 정확도를 높여줄 수 있다는 것을 보여주고 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. 데이터 set 자체를 저장하는 경우 데이터 set의 평균 크기가 3.35K로 15개 데이터 set을 모두 저장하려면 50.4K가 필요한 데 비해 규칙 set 집합은 규칙 43개가 누적된 마지막 규칙 set 집합의 경우에도 1.63K의 저장 공간만을 차지한다. 이것은 이전까지의 데이터 set 모두를 저장하지 않고 직전 데이터 set만을 유지한다고 했을 경우에도 데이터 set하나의 크기인 3.35K보다 작은 크기이다. 규칙 set 집합이 커짐에 따라 이것이 하나의 데이터 set크기인 3.35K를 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.

4.4 t시점까지의 전체 데이터로 t+1시점의 데이터 set 예측

규칙을 누적시킨 규칙 set 집합의 t+1시점의 데이터 예측 정확도 평균은 68.31%로 단순히 t시점에서의 데이터 set에서 규칙을 추출하여 예측했을 때의 예측 정확도 평균 61.37%보다 6.94%나 향상되었다. 만약 물리적, 시간적, 비용적인 소모를 감수하고도 예측 시점에서 그때까지의 전체 데이터를 사용하여 데이터 손실을 줄이려고 한다고 생각해보자. 이 방법은 기존의 일괄처리방법과 같은 방법으로 일괄적으로 모든 데이터를 분석의 대상으로 삼아 모델을 만들어낸 후에 새로운 데이터 set

을 예측하는 방법이다. 다시 말해 t+1시점의 데이터 set을 예측하기 위해 예측 모델을 생성할 때 사용되는 데이터는 최초의 데이터부터 t시점의 데이터까지 그 동안 누적된 모든 데이터이고 이 전체 데이터에서 하나의 규칙 set을 추출하여 예측 모델로 삼은 후 t+1시점의 데이터 set을 예측하는 것이다. 이것을 시험하여 그 예측 정확도를 표로 나타낸 것이 <표 6>이다.

<표 6> 최초부터 t시점까지의 모든 데이터에서 추출된 하나의 규칙 set으로 D_{t+1}을 예측

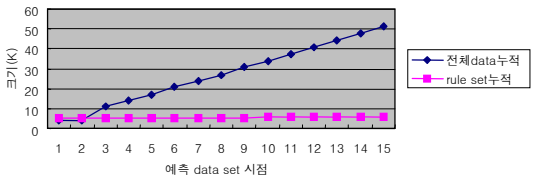
데이터set	2	3	4	5	6	7	8	9
정확도(%)	70	73	70	66	67	62	55	75
데이터set	10	11	12	13	14	15	16	평균
정확도(%)	75	64	66	72	74	65	68.3	68.15

예측 정확도의 평균은 68.15로 t+1시점의 데이터 set을 예측하기 위해 t시점의 데이터 set에서 나온 규칙 set만을 사용하는 경우보다는 예측 정확도 평균이 61.37%에서 6.78% 증가했지만 각각의 데이터 set에서 규칙 set을 누적하여 규칙 set 집합으로 예측 모델을 만든 후 t+1시점의 데이터 set을 예측한 경우의 예측정확도 평균인 68.31%보다는 0.16%가 오히려 감소하는 것을 볼 수 있다.

t시점의 데이터 set에서 나온 모델로 t+1시점의 데이터 set을 예측하는 방법이나, 각각의 데이터 set에서 추출된 규칙 set들을 누적 시킨 규칙 set 집합으로 t+1시점의 데이터 set을 예측하는 방법의 경우에는 규칙을 추출해서 모델을 만들어 내어야 하는 대상이 되는 데이터가 정해진 크기의 데이터 set으로 그 크기가 항상 일정하다. 따라서 데이터 set을 저장하는데 소요되는 저장공간과, 그것에서 규칙을 추출해내는데 걸리는 시간과 요구되는 메모리는 전체 데이터가 시간이 지남에 따라 점점 증가되는 것과 상관없이 평균적으로 일정한 정도를 유지한다. 그러나 전체 데이터에서 예측 모델을 생성해내야 하는 경우는 규칙 set을 추출해야 하는 대상 데이터가 시간이 지남에 따라 증가되어 점점 커지므로 데이터를 저장하는데 소요되는 공간이나, 거기에서 규칙들을 추출하여 예측 모델을 생성하는데 걸리는 시간과 메모리도 데이터가 증가되는 것과 같이 증가할 수 밖에 없을 것이다. 일정 크기의 데이터 set에서 규칙을 추출하여 누적하는 경우와 그때까지 누적된 전체 데이터에서 규칙 set을 추출하는 경우 소요되는 저장 공간의 크기 비교는 <그림 3>과 같다. 규칙 set을 추출하여 예측하는 경우에는 규칙 set을 추출하는 대상이 되는 데이터 set과 더불어 그 시점까지 누적된 규칙 set 집합을 저장하여야 하지만 규칙 set 집합의 크기가 증가하더라도 그 증가율이 현저히 낮다는 것을 볼 수 있다.

저장공간, 메모리사용, 모델 생성 시간에 드는 시간과 비용을 고려했을 때 규칙 set 집합을 누적시켜 예측한 경우가 전체 데이터를 사용하여 예측한 경우보다 예측 정확도가 받아들여질 수 있을만한 범위 내에서 감소한다면 그때까지 쌓인 전체 데이터를 대상으로 예측 모델을 구성하는 것 대신에 규칙 set을 누적하여 사용하는 것을 생각해 볼 수 있을 것이다. 그러나 여기에서는 규칙 set을 누적 시켰을 경우가 전체 데이터를 사용하여 만든 예측 모델을 사용한 경우보다 예측 정확도가 그 크기가 작기는 하지만 오히려 올라가는 것을 볼 수 있었다. 따라서 이 경우에는 시간적, 물리적 비용면에 있어서도 이점을 가지고 있을 뿐더러 예측 정확도에 있어서도 규칙 set을 누적 시킨 경우가 더 나은 결과를 보여주는 것을 볼 수 있었다. 따라서 규칙 set을 누적 시켜 새로 들어오는 데이터 set을 예측한 경우가 직전 데이터 set으로 예측하

거나 전체 데이터를 통해 생성된 하나의 규칙 set으로 예측하는 경우보다 효과적일 것이라는 가정을 만족한다.



<그림 3> 규칙 set을 누적하는 방식과, 전체 데이터누적방식에 소요되는 저장공간 비교

4.5 각 방법의 예측 정확도 차이에 대한 통계적 검정

앞에서는 실험에서 실시한 각각의 예측 모델 구성 방법들의 우수성을 비교하기 위하여 예측의 정확도 평균이라는 대표 값을 설정하고 그것을 비교하였다. 이제 이 실험 방법 별 예측 정확도 평균의 차이가 실제로 통계적으로 의미 있는 것인지 살펴보기 위하여 통계적 검정을 실시하였다. 통계적 검정 방법은 각 방법들이 모두 같은 data set에 대한 예측력을 측정하므로 대응표본 t-test를 사용하였으며 rule set 집합 방식에 초점을 맞추어 <표 7>과 같이 비교하였다.

<표 7> 대응표본 t-test 결과

대응한 예측방법	유의확률
Rule set 집합-최근 rule set	0.005
Rule set 집합-누적 data	0.923

Rule set 집합을 사용했을 때와 최근의 rule set만을 사용한 경우의 표본인 data set들의 예측 정확도 평균 차이는 유의확률 0.005로 유의수준 0.05보다 작으므로 통계적으로 유의하다. 따라서 rule set 누적방식이 최근 rule set만을 사용하여 예측한 경우보다 예측 정확도 평균이 61.37%에서 68.31%로 향상된 것은 rule set 누적 방식이 최근 rule set으로 새 data set을 예측하는 것보다 더 정확하게 새 data set을 예측하기 때문이라고 할 수 있다. Rule set을 누적하여 예측하는 경우와 예측 시점까지 전체 누적된 data를 사용하여 예측하는 경우는 유의확률이 0.923으로 유의수준 0.05보다 커서 두 방법의 차이로 인한 표본의 평균 차이가 없다고 할 수 있다. 앞에서 두 방법이 평균 차이가 크게 없다면, 즉 예측의 정확도가 비슷하다면 누적된 전체 data를 사용하는 경우 필요한 자원이 훨씬 크게 되므로 rule set을 누적하는 방법을 사용하는 것이 더 효율적일 것이라는 것은 언급한 바 있다. 따라서 소요되는 자원을 생각한다면 rule set 집합을 사용하는 것이 최근 rule set을 사용하는 경우보다는 물론이고 예측시점까지 누적된 전체 data를 사용하는 것보다도 더 우수하다고 할 수 있다.

4.6 중요도가 높은 규칙 set만으로 예측모델 구성

주식 시장 자료는 변동이 심한 자료이기 때문에 중요한 규칙 set을 판단하는 근거가 되는 기준정보도 고정되어 있지 않고 예측 시점마다 변해야 할 것이다. 여기서는 예측을 해야 하는 데이터 set 자체를 기준으로 하였다. 예측하려고 하는 시점에서는 이 데이터 set의 예측 목표 class의 실제 값이 알려져 있지 않으므로 이것을 시험 데이터 set으로 삼아 예측 할 수는 없다. 그러나 기존의 데이터 set이 예측하려 하는 데이터 set과 비슷한 구조를 가지고 있다면 기존의 하나의 데이터 set에서 추출된 규칙 set의 각 규칙을 거쳐가는 tuple들의 합이 새 데이터 set의 실제 tuple수와 비슷해야 할 것이다. 예를 들어 기존의 데이터 set이 100개의 tu-

ple을 가지고 있다면 여기에서 추출된 규칙 set은 100개의 tuple을 기준으로 만들어졌기 때문에 이 데이터 set의 tuple들은 규칙 set 안의 규칙 중 한 개에 해당하게 될 것이다. 왜냐하면 주식 시장 자료는 독립 변수들이 연속형 수치로 이루어져 있는 자료 구조이므로 tuple들이 규칙들에 의해서 특정한 기준으로 나뉘어지기 때문에 tuple 하나가 여러 개의 규칙, 즉 여러 개의 기준에 모두 걸리지는 않기 때문이다. 따라서 각각의 규칙에 해당된 tuple 수를 합하면 실제의 tuple 수와 차이가 없을 것이다. 만약 새로 예측하려는 데이터 set이 이 데이터 set과 비슷한 구조를 가지고 있다면 tuple 한 개는 여기에서 추출된 규칙 set 안에 있는 한 개의 규칙에 해당되고 다른 규칙들에는 해당되지 않을 것이다. 규칙들이 tuple들을 어떠한 기준에 의해 나누어 주었기 때문이다. 제대로 기준이 설정되었다면 하나의 tuple은 한 개의 기준에만 걸려야 할 것이다. 따라서 규칙들을 거쳐간 총 tuple 수는 실제 tuple수와 비슷하게 될 것이다. 그러나 만약 새로운 데이터 set이 기존의 데이터 set과 비슷하지 않은 자료 구조를 가지고 있다면 새 데이터 set의 한 tuple이 기존 데이터 set에서 추출된 규칙 set 안의 규칙 하나 이상에 해당될 수 있다.

Tuple들을 나누어주는 기준이 되는 규칙들이 그것이 추출된 데이터 set의 tuple들을 나눠주는데는 적절한 기준이 되었지만 새로운 tuple들을 나누어주기에는 부적절해서 하나의 tuple이 여러 개의 규칙, 즉 하나 이상의 기준에 해당될 수 있기 때문이다. 이렇게 되면 하나의 tuple이 서로 다른 class로 예측될 경우가 생길 수 있는 것이다. 따라서 여기서는 중요한 규칙 set을 판단하는 근거로서 새 데이터 set을 기존의 규칙 set들에 통과시켜보고

<표 8> D_{t+1}의 tuple이 $\sum_{i=1}^{i} R_i$을 거쳐간 각각의 합계와 각 시점에서의 평균

		규칙 set												평균
		1	2	3	5	7	8	9	10	11	13	14	15	
D_{t+1}	2	117												
	3	111	103											107
	4	105	103	124										111
	5	121	106	107										111
	6	125	107	110	108									113
	7	123	110	108	105									112
	8	120	104	113	103	100								108
	9	118	107	115	107	100	100							108
	10	115	111	124	109	100	100	114						110
	11	111	103	130	110	100	100	100	114					109
	12	124	108	109	107	100	100	102	106	100				106
	13	116	111	119	106	100	100	105	106	100				108
	14	125	109	110	104	100	100	102	105	100	104			106
	15	113	103	138	101	100	100	100	117	100	104	108		108
	16	44	42	56	42	41	41	42	45	41	42	45	45	44

각 규칙 set을 거쳐간 tuple의 총 개수를 따져 그것이 각 데이터 set을 거쳐간 tuple들의 개수의 평균 안에서 차이가 나면 그 규칙 set이 새 데이터 set의 자료구조와 상대적으로 비슷한 데이터 set에서 추출되었다고 보고 중요한 규칙 set이라고 판단하였다. <표 8>은 그때까지의 규칙 set $\sum_{i=1}^{t-1} R_i$를 거쳐간 D_{t+1}의 tuple 수를 나타내고 그것들의 평균이 R_t가 상대적으로 중요한 규칙 set인지 그렇지 않은지를 결정하는 임계값이 되고 있다. 데이터 set 4, 6, 12에서는 추출된 규칙 set이 없었으므로 해당 부분은 표에서 빠져 있다. 회색으로 선택된 규칙 set은 데이터 set D_{t+1}을 예측하려고 할 때에 조건을 만족하여 선택된 규칙 set을 나타낸다. 규칙 set 1의 경우는 데이터 set 2의 tuple들이 규칙 set 1을 거쳐간 tupel 수 평균을 낼 규칙 set이 그 자신뿐이므로 평균을 내지 않았고 tuple 수가 117로 비교적 크기 때문에 상대적 중요도가 높은 규칙 set에서는 제외하였으나 데이터 set 2를 예측할 규칙 set이 오직 규칙 set 1뿐이므로 다음 시점에서 더 규칙 set이 모이기 전까지는 규칙 set 1을 기본으로 쓰도록 하였다.

위 과정을 통해 각 데이터 set시 중요한 규칙 set으로 예측된 것은 <표 9>과 같다. <표 9>에서 보여지듯이 중요도가 높은 규칙 set을 판단하는 근거가 되는 기준이 현재 예측하려고 하는 데이터 set D_{t+1}이므로 예측 시점마다 상대적으로 중요도가 높은 규칙 set 집합이 바뀌게 되고 이들만으로 예측한 예측 정확도는 <표 10>과 같다.

<표 10> 중요도가 높은 규칙 set만으로 예측

데이터set	2	3	4	5	6	7	8	9
정확도(%)	60	67	69	67	65	65	64	63
데이터set	10	11	12	13	14	15	16	평균
정확도(%)	72	72	66	66	77	72	70.7	67.71

예측 정확도 평균은 67.71%로 규칙 set을 전체 누적시킨 경우의 예측 정확도 평균 68.31%보다 0.6% 오히려 떨어진 것을 볼 수 있다.

4.7 중요도가 높은 규칙 set 가중치 달리 조정, 전체에 포함

이제 앞에서 시험한 중요도가 높은 규칙 set만으로 구성한 예측모델과, 전체 규칙 set 집합을 유지하되 중요도가 높은 규칙 set과 그렇지 않은 규칙 set에 가중치를 달리 조정하여 준 것을 비교하려 한다. 이 과정에서 상대적 중요도가 높지 않은

<표 9> 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	10	5, 7, 8
3	2	11	2, 7, 8, 9
4	1, 2	12	7, 8, 9, 10, 11
5	2, 3	13	5, 7, 8, 9, 10, 11
6	2, 3, 5	14	5, 7, 8, 9, 10, 11, 13
7	2, 3, 5	15	2, 5, 7, 8, 9, 11, 13, 14
8	2, 5, 7	16	1, 2, 5, 7, 8, 9, 11, 13
9	2, 5, 7, 8		

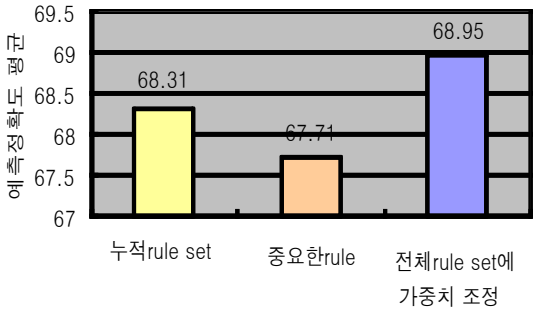
규칙 set이라고 판단된 규칙 set은 D_{t+1}의 tuple이 하나의 규칙 set을 거쳐가면서 한 개 이상의 규칙에 걸리는 경우가 많은 규칙 set이다. 하나의 tuple이 여러 개의 규칙에 해당된다면 서로 다른 class로 예측하는 여러 개의 규칙에 걸리는 경우가 생길 수 있기 때문에 같은 규칙 set안에 있는 규칙들끼리도 서로 모순되는 예측을 할 확률이 늘어난다고 할 수 있다. 따라서 이 경우에는 같은 규칙 set안에서도 좀 더 올바른 예측을 할 가능성이 많은 규칙과 그렇지 않은 규칙을 구분하여 줄 필요가 있다. 이것을 구분하는 기준으로 각 규칙의 confidence를 사용하였다. Confidence는 탐색된 연관성의 확실성 정도를 평가해주는 신뢰도로서 X => Y와 같은 규칙에서 X를 포함하는 트랜잭션이 Y도 포함할 확률인 조건부 확률이다. 그러므로 1보다 작고 규칙 set안의 각각의 규칙이 얼마만큼 정확하게 tuple의 목표 class를 예측하는지 알려주는 척도가 된다. 따라서 상대적으로 중요도가 높은 규칙 set은 그 안의 각각의 규칙이 동일한 가중치를 가지고 tuple이 예측될 class의 빈도를 규정하지만 그렇지 않은 규칙 set의 규칙들은 1보다 작은 수치의 confidence를 가지고 예측할 class의 빈도를 규정하게 된다. 상대적 중요도가 높은 규칙 set의 규칙들에 주는 가중치는 시행착오를 통해 계산하였으며 가중치를 어떻게 주느냐에 따라 예측의 정확도 평균이 0.5% 미만의 차이를 보였다. 그 중에서 각 규칙에 1.5, 즉 그 규칙이 예측하는 class로 예측될 빈도수를 1.5로 주었을 때 가장 높았고 예측 정확도 평균은 68.95%로 상대적 중요도가 높은 규칙 set만으로 예측 모델을 구성한 경우의 예측 정확도 평균 67.71%보다 1.24% 높아진 것을 볼 수 있었다. 각 예측 정확도는 <표 11>와 같다.

중요도가 높은 규칙 set에 가중치를 어떻게 주느냐, 중요도가 상대적으로 낮은 규칙 set의 가중치를 어떻게 조정하느냐에 따라 예측의 정확도가 달라졌으나 시험 결과 상대적으로 중요한 규칙 set들에 가중치를 주어 전체 규칙 set에 포함시킨 경우가 중요도가 높은 규칙 set만으로 예측 모델을 구성한 경우의 예측 정확도 평균 67.71%보다 68.95%로 예측 정확도가 더 높게 나왔다. 따라서 중요도가 높은 규칙 set만으로 예측 모델을 구성한 경우보다 그렇지 않은 규칙 set까지 모두 예측 모델에 포함하되 중요도가 높은 규칙 set과 그렇지 않은 규칙 set에 달리 가중치를 조정하여 준 경우가 더욱 예측 정확도가 높게 나올 것이라는 가정이 만족되었다. 이것을 비교한 것이 <그림 4>이다.

앞의 실험에서 살펴본 결과 규칙 set을 누적할 경우 최근 규칙 set만으로 예측하거나, 누적된 전체 데이터에서 하나의 규칙 set을 사용하여 예측한 경우보다 예측의 정확도가 높아지는 것을 확인하였다.

<표 11> 중요도가 높은 규칙 set과 그렇지 않은 규칙 set에 가중치를 달리 주어 예측

데이터set	2	3	4	5	6	7	8	9
정확도(%)	70	67	72	74	65	65	65	58
데이터set	10	11	12	13	14	15	16	평균
정확도(%)	72	73	67	67	76	70	73.2	68.95



<그림 4> 전체 규칙 set, 중요 규칙 set, 선택적 가중치 적용 경우 비교

여기서 더 나아가서 이것을 더욱 효과적인 예측 모델로 만들기 위해서는 누적된 과거 정보를 삭제 없이 모두 누적하면서도 그 중에서 상대적으로 중요한 것을 찾아내어 가중치를 준다면 더욱 효율적으로 새로 들어오는 데이터를 예측할 수 있다는 것을 실험을 통해 확인하였다.

4.8 실험 결과 요약

t+1시점의 데이터 set을 예측하기 위하여 t시점의 데이터 set에서 추출한 규칙 set만 사용했을 때 예측 정확도의 평균은 61.37%였으나 처음부터 t시점까지 누적된 데이터 set에서의 규칙 set 집합을 사용하였을 경우 이 예측 정확도가 68.31%로 6.94% 향상되는 것을 볼 수 있었다. 처음 시점부터 누적된 모든 데이터를 사용해서 t+1시점의 데이터 set을 예측하여 보았더니 여전히 규칙 set 집합 사용의 경우보다 예측 정확도가 떨어졌다. 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. 따라서 처음 시점부터 모든 데이터를 누적하고 예측 모델을 생성하는데 드는 저장공간, 메모리, 시간을 고려하여 보았을 때 예측 정확도의 저하가 사용자가 받아들일 만한 정도의 것이라면 전체 데이터를 사용하는 대신 규칙 set 집합을 사용하는 것을 생각해 볼 수 있을 것이다. 여기서는 자원을 절약하면서도 예측 정확도가 약간 상승하는 것을 볼 수 있었다. 이것들의 예측 정확도 평균을 비교한 것이 <표 12>이다.

중요한 규칙 set만으로 예측 모델을 구성한 경우와 전체 규칙 set을 포함하여 예측 모델을 구성하되 중요한 규칙 set에 가중치를 주는 경우의 예측 정확도 평균은 <표 13>과 같다. 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 영향력만을 조정하는 모델이 더욱 예측 정확도가 높았다는 것이다.

<표 12> 최근데이터 set, 누적규칙 set 집합, 전체 데이터로 예측한 것의 비교

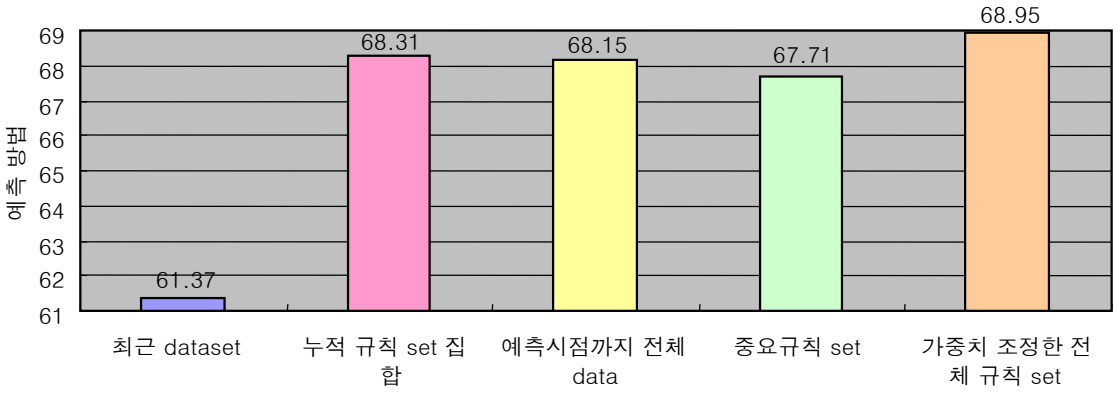
예측기준	최근 데이터 set	누적 규칙 set 집합	예측시점까지의 전체 데이터
예측 정확도	61.37	68.31	68.15

<표 13> 중요 규칙 set, 중요 규칙 set에 가중치 적용 예측모델 예측도 비교

예측모델 구성기준	중요한 규칙 set	중요한 규칙 set에 가중치 준 전체 규칙 set
예측 정확도	67.71	68.95

누적된 규칙 set 집합을 사용했을 때 예측 정확도가 향상된 것은 규칙이 누적되면서 한 tuple이 서로 다른 규칙 set에서 생성된 여러 개의 규칙의 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 높이는데 기여하였다고 결론 내릴 수 있다. 실행에서 시행된 예측방법들을



<그림 5> 예측방법에 따른 예측 정확도 평균 비교

종합한 것이 <그림 5>에 나타나 있다

5. 결론 및 시사점

5.1 결 론

시간이 흐름에 따라 계속적으로 들어오는 연속 발생 데이터는 시간의 변화에 영향을 많이 받게 되므로 시간의 흐름에 대한 정보를 어떤 식으로 가지고 있느냐는 연속 발생 데이터마이닝의 예측 정확도를 높이는 중요한 요소가 될 것이다. 여기서는 그것을 규칙으로 설정하였다. 어떠한 구체적인 패턴을 가지고 있고 이것의 빈도수 자체를 유지하고 있는 방법의 경우 새로운 패턴이 생기면 이러한 패턴을 계속 추가 시키면서 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 있을 것이다. 또한 규칙 set 집합을 마이닝하는 것은 일정한 크기의 데이터 set을 대상으로 하므로 방대한 전체 데이터를 마이닝하는데 드는 시간, 메모리의 문제도 해결할 수 있을 것이다.

5.2 연구의 한계 및 미래 연구 방향

위의 자료에서는 시험의 편의상 데이터 set의 크기를 크지 않게 설정하였으므로 데이터 set의 크기를 매우 크게 하여 시험을 해 보는 것이 필요할 것으로 보인다. 또한 데이터 set의 개수를 상당히 늘려서 규칙 set 집합 크기가 늘어남에 따라 예측 정확도의 변화를 어떻게 변하는지 연구할 수 있을 것이다. 가중치를 주는 경우에 여기서는 예측 대상이 되는 데이터 set의 tuple이 각 규칙 set을 거쳐가는 빈도수의 합을 가중치를 주는 기준으로 삼았다. 그러나 기준이 되는 정보를 어떠한 것으로

하느냐에 따라 상대적으로 중요한 데이터 set이라고 규정 지어지는 경우가 무수히 다르게 생길 수 있을 것이다. 이러한 기준을 설정하는 것은 자료에 대한 분석이 선행되어야 하며 잘못 설정되었을 경우 예측의 정확도를 떨어뜨리거나 높이지 못하는 결과를 낳을 수 있기 때문에 신중하게 결정되어야 할 것이다. 그러므로 자료에 따라서 이 기준을 정하는 가장 효율적인 방법을 찾아 볼 수 있을 것이다. 또한 여기서는 시행착오법을 통하여 여러 가지 가중치를 설정하여 시험하였으나 데이터 set의 크기와 규칙 set에 있는 규칙의 수, 기준이 되는 정보 등을 더 연구하여 어떤 경우에 어느 정도 크기의 가중치가 가장 효과적이라는 것을 밝혀내는 것이 필요할 것이다.

참고문헌

최세일, 최세일 “주식시장 기술적 지표 분석”, 진리탐구, 1999.

Achelis, S. B., “Technical analysis from A to Z”, Chicago : Probus Publishing, 1995.

Agrawal, R. and R. Strikant, “Mining sequential patterns”, *Proceedings of 1995 International Conference of Data Engineering*, (1995), 3~14.

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), 287~291

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 Conference on Data Engineering*, (1996), 106~114

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*, 2003.

Cutler, D. M., J. M. Poterba and L. H Summers, “Speculative Dynamics and the Role of Feedback Traders”, *American Economic Review*, Vol.80(1990), 63~68.

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), 71~80.

Domingos, P. and G. Hulten, Mining High-Speed Data Streams, KDD, Boston, ACM Press, 2000.

Ganti, V., J. Gehrke, and R. Ramakrishnan, “DEMON : Mining and monitoring evolving data”, *Proceedings of the Sixteenth International Conference on Data Engineering*, (2000), 439~448

Ganti, V., J. Gehrke, and R. Ramakrishnan, “Mining Data Streams under Block Evolution”, *SIGKDD Explorations*, Vol.3, No.2 (2002), 1~10.

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), 169~180.

Giannella, C., J. Han, J. Pei, and X. Yan, “Mining Frequent Patterns in Data Streams at Multiple Time Granularities”, *Next Generation Data Mining*, MIT Press, 2003.

Greenwald, M. and S. Khanna, “Space-Efficient On-line Computation of Quintile Summa-

ries", Proceedings of ACM SIGMOD, Santa Barbara, 2001.

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.

Han, J. and M. Kamber, Data Mining-Concepts and Techniques, Morgan Kaufmann, 2001.

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), 1~12.

Hidber, C., "Online Associattion Rule Mining", *Proceedings of ACM SIGMOD*, (1999), 145~156.

Hulten, G., L. Spencer, and P. Domingos, "Mining Time-Changing Data Streams", *KDD 01 San Francisco*, C.A., 2001.

Jegadeesh, N., "Evidence of Predictable Behavior of Security Returns", *Journal of Finance*, Vol.45, No.3(1990), 881~898.

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

Pesaran, M. H. and A. Timmerman, "Predictability of Stock Returns : Robustness and Economic Significance", *Journal of Finance*, Vol.50, No.4(1995), 1201~1228

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

Abstract

A Real-Time Stock Market Prediction Using Knowledge Accumulation

Jinhwa Kim* · Kwang Hun Hong* · Jin Young Min**

One of the major problems in the area of data mining is the size of the data, as most data set has huge volume these days. Streams of data are normally accumulated into data storages or databases. Transactions in internet, mobile devices and ubiquitous environment produce streams of data continuously. Some data set are just buried un-used inside huge data storage due to its huge size. Some data set is quickly lost as soon as it is created as it is not saved due to many reasons. How to use this large size data and to use data on stream efficiently are challenging questions in the study of data mining.

Stream data is a data set that is accumulated to the data storage from a data source continuously. The size of this data set, in many cases, becomes increasingly large over time. To mine information from this massive data, it takes too many resources such as storage, money and time. These unique characteristics of the stream data make it difficult and expensive to store all the stream data sets accumulated over time. Otherwise, if one uses only recent or partial of data to mine information or pattern, there can be losses of valuable information, which can be useful.

To avoid these problems, this study suggests a method efficiently accumulates information or patterns in the form of rule set over time. A rule set is mined from a data set in stream and this rule set is accumulated into a master rule set storage, which is also a model for real-time decision making. One of the main advantages of this method is that it takes much smaller storage space compared to the traditional method, which saves the whole data set. Another advantage of using this method is that the accumulated rule set is used as a prediction model. Prompt response to the request from users is possible anytime as the rule set is ready anytime to be used to make decisions. This makes real-time decision making possible, which is the greatest advantage of this method. Based on theories of ensemble approaches, combination of many different models can produce better prediction model in performance. The consolidated rule set actually covers all the data set while the traditional sampling approach only covers part of the whole data set.

This study uses a stock market data that has a heterogeneous data set as the characteristic of data varies over time. The indexes in stock market data can fluctuate in different situations whenever

* School of Business, Sogang University
** The Fox School of Business and Management, Temple University

there is an event influencing the stock market index. Therefore the variance of the values in each variable is large compared to that of the homogeneous data set. Prediction with heterogeneous data set is naturally much more difficult, compared to that of homogeneous data set as it is more difficult to predict in unpredictable situation.

This study tests two general mining approaches and compare prediction performances of these two suggested methods with the method we suggest in this study. The first approach is inducing a rule set from the recent data set to predict new data set. The seocnd one is inducing a rule set from all the data which have been accumulated from the beginning every time one has to predict new data set. We found neither of these two is as good as the method of accumulated rule set in its performance. Furthermore, the study shows experiments with different prediction models. The first approach is building a prediction model only with more important rule sets and the second approach is the method using all the rule sets by assigning weights on the rules based on their performance. The second approach shows better performance compared to the first one. The experiments also show that the suggested method in this study can be an efficient approach for mining information and pattern with stream data.

This method has a limitation of bounding its application to stock market data. More dynamic real-time steam data set is desirable for the application of this method. There is also another problem in this study. When the number of rules is increasing over time, it has to manage special rules such as redundant rules or conflicting rules efficiently.

Key Words : Stock Market Prediction, Stream Data, Data Mining, Knowledge Accumulation

저 자 소 개

김진화

서강대학교 영문학 학사, 경영학 학사, University of Wisconsin에서 전산학 석사, 경영학 석사, 경영학 박사를 취득하였다. 1998년부터 2003년까지 Oklahoma State University, School of Business에서 MIS분야 조교수로 재직하였으며, 2003년부터 현재까지 서강대학교 경영학과에서 경영정보학 분야 교수로 재직 중이다. 지능정보시스템학회지, 전자거래학회지, Entrue Journal of Information Technology, International Journal of Information and Operations Management Education의 편집위원을 지냈으며 현제 Creative Business Review의 편집위원장을 맡고 있다. 주요 연구관심분야는 Data Mining, Customer Relations Management, Decision Support Systems, Artificial Intelligence, Future Forecasting, 창조경영 등이다.

홍광헌

미국 University of Wisconsin에서 경영학 박사학위를 취득한 후에, University of Houston, Saginaw Valley State University 등에서 강의를 하다가 2004년 9월 서강대학교에 부임한 이래 현재까지 재직 중이다. Asia-Pacific Journal of Financial Studies, 재무연구(한국재무학회 발행), 경영학연구(한국경영학회 발행) 의 편집위원으로 활동하는 등 활발한 학술활동을 수행하고 있으며, 대구경북경제자유구역청의 투자자문위원으로도 활동하고 있으며, Marquis 사에서 발행하는 Who's Who in the World 인명사전(2010년, 2011년, 2012년)에 등재되었다.

민진영

KAIST경영대학에서 박사학위를 취득하였으며 현재 Temple University에서 Post-Doc 연구원으로 있다. 주요 관심분야는 Social Media, Human and Computer Interaction Technology and Virtual Environment, Digital Ecosystem이다. 경영학연구, 경영과학지 등의 국내 학술지에 논문을 게재한 바 있다.