사례기반학습을 이용한 주식 데이터 예측 방법

김주현* · 전민수* · 정용규**

목 차

요 약

현재 국내에선 많은 수의 사람들이 주식관련업계에 종사하고 있으며 주식관련 정보와 관련 산업은 점점 발전해 가고 있다. 따라서, 주식을 예측하는 프로그램이 많이 나왔으며, 또한 정확한 수치화를 통해 주식을 예측하고자 하는 노력들이 더해지고 있다. 그러나 주식예측 결과는 아직 불안정하고, 근거가 없는 것이 현실이다. 본 논문에서는 방대한 량의 주식 데이터를 가지고, 주식의 변동 폭에 많은 영향을 끼치는 항목들을 조사하고, 가중치를 구하고자 한다. 이는 기존에 주식에 관련된 수치와 종목별의 분류와 다른 방법이다. 실험결과에 따른 체계적인 주식 데이터의 객관성 있는 분류를 제시하고자 한다.

표제어: 통계분석기법, 주식 예측, 사례기반학습, 가중치부여

접수일(2011년 8월 30일), 수정일(2011년 9월 15일), 게재확정일(2011년 9월 27일)
* 을지대학교 의료산업학부 의료전산학전공
** 을지대학교 의료IT마케팅학과교수, ygjung@eulji.ac.kr

1. 서론

국내 주식관련 산업은 날로 발전해 나가고 있다. 직접적으로 주식산업에 1차적으로 종사하는 인원을 제외하고도 순수 주식시장에 투입되는 자본은 외국인과 국내인을 포함하여 점점 성장해가고 있으며 그 날의 주가 등락폭에 따라 큰 이익을 창출하기도 하고 손실을 보기도 한다. 자연스러운 결과로 이러한 주가에 대한 등락폭의 원인에 대해 관심이 쏠리고 있으며 그에 부응하여 주식 예측 프로그램을 이용하거나 또는 전문인에게 돈을 맡기기도 한다. 하지만 주식의 등락을 결정하는 요인에는 무수히 많은 요소들이 있기 때문에 이러한 방식 또한 완벽한 방법이 아니어서 손실을 보기도 한다.

본 논문에서는 주식을 예측하기 이전에 주식의 등락에 영향을 끼치는 데이터는 무엇인지 알기 위하여 통계기법을 이용하여 체계적인 분류를 해 나갈 것이다. 이러한 분류를 통해 주식에 관련된 수많은 정보 중에 어떠한 정보가 신뢰할 만한 것인지 알 수 있을 것이며 나아가 보다 정확한 예측을 내리는데 도움이 될 것이다.

2. 관련연구

인간은 과거의 사례나 경험에 비추어 현재의 문제를 인식하고 유형화하여 해결책을 내놓을 수 있다. 이를 이해서는 현재의 당면문제가 과거에 나타났던 사례나 경험과 일치하는지 여부를 살피고 일치하면 과거에 나타났던 사례나 경험과 일치하는지 여부를 살피고 일치하면 과거의 해결책에 비추어 답을 낸다. 만약 정확히 일치하지 않더라도, 법관이 지난 사건에 대한 판례를 통해 당면한 사건에 대해 판단을 내리듯, 과거의 사례나 경험은 현재의 문제에 부분적인 해결책을 제시할 수 있다.

사례기반 추론은 이러한 인간의 지적 활동을 모델화한 것으로, 과거 문제로부터 얻은 상황 경험이나 지식을 사례 데이터베이스로 구축하여 어떠한 상황이나 문제가 발생하면 기존의 사례 데이터베이스에서 똑 같거나 가장 유사한 사례를 선택하여 그 사례가 제시하는 해결책으로 현 문제에 대한 답을 제시한다.

이는 규칙베이스 전문가 시스템에서의 정방향 추론과 같이, 주어진 지식베이스 공간에서 추론하는 것과는 다르다. 규칙베이스 전문가시스템의 추론방식은 사용자로부터 문제에 대한 세부사항이나 혹은 질문과 대답을 통해 일력을 받아 이를 토대로 규칙베이스에서의 관련된 규칙간의 연쇄를 통해 해답에 도달한다.

사례기반 추론은 다음의 접근법을 가진다. 과거의 사례를 바탕으로 문제를 해결하기 때문에 비록 문제가 복잡하더라도

[그림 2-1] 규칙베이스 전문가시스템에서의 규칙의 연쇄

[Fig. 2-1] Chain of Rules in Expert System Rule Base

이미 해결된 사례를 통해 해답을 빨리 도출할 수 있다. 그러므로 지식이 잘 파악되지 않은 대상영역에 있어서도 사례로서 추론을 가능하게 한다. 그리고 정확히 일치되는 사례를 발견할 수 없다면 가장 유사한 사례를 변형하여 새로운 문제를 해결하도록 할 수 있으며 이렇게 해결된 사례는 다시 새로운 사례로서 저장되게 된다. 사례기반 추론에서의 추론과정은 다음과 같이 표현될 수 있다.

[그림 2-2] 사례기반 추론 과정
[Fig. 2-2] Case-based Reasoning

3. 자료의 정리

3.1 자료의 수집

주식 데이터의 수집은 코스닥 종목 중 거래량 상위랭크 Top20까지의 종목을 선정하였다. 코스피 우량주보다 코스닥 주식종목의 경우 등락의 리스크가 크기 때문에 사례기반학습의 적용에 따른 효과가 클 것이라 기대한다.

〈표 3-1〉 실험데이터
〈Tab. 3-1〉 The Experimental Data

1	에코솔루션	11	코데즈컴바인
2	테라리소스	12	이지바이오
3	어울림엘시스	13	인트로바이오
4	티케이케미칼	14	미주제강
5	엔케이바이오	15	세운메디칼
6	헤파호프	16	시노펙스그린테크
7	엘앤피아너스	17	어울림정보
8	삼천당제약	18	아가방컴퍼니
9	국순당	19	셀트리온
10	한일사료	20	예당

3.2 변수의 선정

주식 데이터의 변수선정에는 흔히 말하는 주식의 등락에 영향을 끼치는 요소들로 선정하였다. 첫 번째 변수는 거래량으로써 거래량의 높고 낮음에 따라 데이터마이닝 되어지는 것이 아닌 거래량의 한계 거래량을 두어 그 거래량의 수치안에 들어갈 때에만 등락의 영향을 끼치는 것으로 설정하였다. 두 번째는 최근거래 주식종목의 일주일전의 등락폭으로써 이는 거래량과 연산하여 데이터마이닝 되어진다. 세 번째로는 주식종목의 일주일전 뉴스이다. 이는 주관적인 데이터로써 수치화하기 까다로운 면이 있지만 주식의 변동에 가장 큰 영향력을 끼칠 수 있을 것이다.

3.3 가중치의 부여

(1) 거래량

주가에 영향을 미칠 것으로 예상되는 변수로 주식의 거래량을 들 수 있다. 여기서

제시하는 가정치는 3,000,000~6,000,000 주로써 당일 주식의 거래량 = 3,000,000~6,000,000일 경우 가중치부여 +3이 되어진다. 그와 반대로 그 이하일 경우에는 +1의 가중치가 부여되며 그 이상일 경우 -1의 가중치가 부여된다.

〈표 3-2〉 거래량의 따른 가중치 부여의 분류

〈Tab. 3-2〉 Weighting Based on the Classification of Volume

구분	거래량	가중치
A	0~3,000,000	+1
B	3,000,000~6,000,000	+3
C	6,000,000~	-1

(2) 과거 등락 수치

두 번째 주가에 영향을 주는 변수로 과거 등락 수치를 들 수 있다. 주식 종목의 일주일간의 등락의 폭에 따라 데이터를 수집하게 된다. 이러한 데이터의 수집이 아래의 수식에 해당하는 경우, 즉 그 날의 주식가격의 상승이 전날대비 3% 이상일 경우 가중치(+1)가 부여하게 되며 이와 연계되어 자동적으로 뉴스 데이터를 수집하게 된다.

$$\text{종목의 가격} \times 3/100 > \text{종목의 등락폭} \quad (1)$$

(3) 뉴스

주가에 영향을 미치는 세 번째 변수로 뉴스를 들 수 있는데, 뉴스의 가중치가 부여될 경우는 두 가지 경우로 나뉜다. 앞에서 언급한 일주인 전의 등락폭에 해당되는 경우 가중치의 부여가 더욱 크게 이루어지며 그것과 달리 단순 뉴스 데이터의 경우 키워드 검색을 통해 가중치가 부여되게 된다. 이런 가중치 부여 절차를 보면 다음과 같다.

[그림 3-1] 키워드에 따른 가중치 부여

[Fig. 3-1] Keyword, According to the Weighting

4. 실험

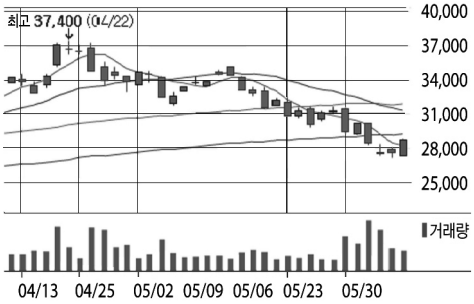
4.1 제안 알고리즘

위에서 제시한 주식 데이터의 주가 영향 변수의 가중치를 구하는 과정으로, 실험을 위해 제안한 알고리즘은 아래와 같다.

① 특정 주식 등락 데이터 선정
② 상승/하락 주식으로 분류 후 가장 변동이 큰 변동 주식의 시점과 시기를 결정(기준점 부여)
③ 등락 기준 시점을 중심으로 일주일 전의 주식 관련 뉴스 수집
④ 등락 기준 시점의 해당 주식의 키워드 도출하고 키워드 가중치를 부여한다.
⑤ 도출해낸 키워드로 수집된 뉴스와 검색 및 매칭
⑥ 매칭 후 결과를 분류하여 가중치를 부여한다.
⑦ 키워드/뉴스 가중치 데이터를 저장한다.

4.2 종목 및 기준 선정

아래 그림은 '하이닉스' 종목의 증권 정보의 등락 추이를 알아보기 위한 차트이다. 일반적으로 여러 가지 분석차트들이 존재하지만 과거 주식 데이터의 변화 추이를 쉽게 알아 볼 수 있는 봉형 차트를 이용하였다. 선정된 차트의 추이를 분석하여 상승 및 하락 폭이 큰 시기를 판단하고 기준점을 정하도록 한다.



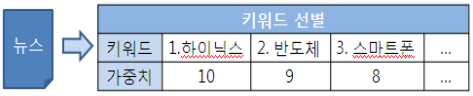
[그림 4-1] 하이닉스 차트
[Fig. 4-1] Hynix Chart

4.3 관련 뉴스 수집

기준 선정 단계에서 도출된 기준점의 날짜를 기점으로 일주일전의 주식 뉴스 데이터들을 수집한다. 수집되어지는 뉴스들은 한국경제, 아시아경제, 이데일리 등 공인된 정보매체의 뉴스만을 유효 데이터로서 인정한다.

4.4 키워드 도출

키워드 도출은 뉴스 데이터를 대상으로 한다. 종목 및 기준이 선정된 후 초반 키워드를 생성하는 과정은 종목의 이름(회사명)을 기준 키워드로 선정 후 이와 관련된 뉴스를 가져와 그 정보들에 중첩되는 단어들의 중첩 횟수를 근거로 키워드를 선별한다. 기준 키워드로 만들어진 종목이름을 포함해서 생성된 키워드들의 중첩 횟수는 가중치로 이어지며, 가중치가 정해진 이 키워드는 이후 새로운 키워드 도출 시 검색과 비교를 위한 기준점으로 사용되어 진다. 키워드는 지속적으로 데이터베이스에 저장되며 키워드 가중치는 유지되거나 가감되어진다.



[그림 4-2] 도출된 키워드
[Fig 4-2] Derived Keywords

4.5 검색 및 매칭

전 단계의 과정에서 생성된 키워드는 관련 뉴스 수집에서 수집된 뉴스 데이터와 검색/매칭에 사용된다. 뉴스는 키워드와 매칭 되는 횟수를 기준으로 분류되어 가중치가 산정된다. 가중치가 부여된 뉴스는 주식 등락 기준점을 만든 이후 뉴스 수집 선정기준인 7일 안에 포함되는 뉴스에 한해 데이터와 가중치가 유지 및 가감되며 포함이 되는 않는 뉴스 데이터는 유효성이 없는 데이터라 판단되어 삭제된다.

5. 실험결과

키워드 도출과 뉴스 데이터 매칭을 통

하여 주식과 연관성이 있는 두 가지의 가중치 데이터가 추출된다. 키워드 데이터는 가중치에 따라 분류가 되며 가중치가 같은 값의 키워드가 생성된다면 키워드의 중요도는 같다고 본다. 키워드 가중치는 주식 등락의 여부 혹은 뉴스 데이터의 중요도에 대한 연관성이 높은 데이터이다.

No.	키워드	가중치
1	하이닉스	10
2	반도체	9
3	다이오드	8
4	스마트폰	7
5	물가 안정	6
⋮	⋮	⋮

[그림 5-1] 키워드 가중치 데이터
[Fig. 5-1] Keyword Weight Data

매칭수	키워드	가중치
5	하이닉스, 낙폭 과도 평가 …	5
	하이닉스 상승반전, 주가 …	5
	코스피 나흘째 하락 '21 …	5
	키움증권 연 25% 수익 …	5
4	하이닉스 매각구조 이번 …	4
⋮	대신 이영주, 수익률 70 …	4
	⋮	⋮

[그림 5-2] 뉴스 가중치 데이터
[Fig. 5-2] News weight data

뉴스 데이터 매칭은 뉴스와 키워드의 매칭수에 따라 가중치가 부여되며 키워드와 마찬가지로 가중치가 높으면 주식 등락의 여부 혹은 뉴스 데이터의 중요도에 연관성이 높은 뉴스라고 분류한다. 이렇게 분류되어진 뉴스 데이터는 키워드 데이터와 다르게 주식 등락 기준점에 따른 뉴스 수집 선정일수에 포함이 되지 않는 뉴스는 삭제되며 포함되는 뉴스는 데이터 및 가중치의 유지와 가감이 가능하다.

6. 결론

본 논문에서는 사례기반학습을 통한 가지치기와 가중치 부여 등을 이용하여 주식 종목의 등락에 영향을 끼치는 변수들은 무엇인지 우선순위를 선정하여 주식의 매수 매도에 조금이나마 도움이 되려 하였다. 그에 따른 절차로서 자료의 수집과 변수의 설정 그리고 가중치의 부여를 들 수 있다. 이를 실제적으로 2주 동안 관찰하여 사례기반에 따른 가중치부여가 실제의 상황과 비교하여 얼마나 일치하는가를 보았다. 하지만 본 논문에서는 변수의 설정의 폭이 좁다는 점과 주식등락의 원인에 대한 경우의 수가 생각보다 다양하다는 점에 대해 조금 더 연구해야 할 필요성을 느꼈으며 차후 실제 사례에 적합한 가중치 부여가 이루어지게 되면 향후 주식 예측 프로그램을 통해 알고리즘을 적용할 계획이다.

참고문헌

[1] 신동호, 장병탁(1998), 점진적 프로토타입 구성에 의한 사례 기반 학습의 특성 분석, 한국정보과학회 춘계학술

대회 논문집, 25(1), 267-269.

[2] 박문서, 성기훈, 이현수, 지세현, 김수영(2010), 사례기반추론을 이용한 초기 단계 공사비 예측 방법, 한국건설관리학회논문집, 11(4), 22-31.

[3] 박선일(2009), 수의학 관련 자료에 대한 통계분석 기법, 한국임상수의학회 춘계학술대회.

[4] 박우진, 김명현, 민경수, 오혜란, 임채미(2007), 주식 가격 변동 예측을 위한 다단계 뉴스 분류시스템, 정보관리학회지(Journal of the Korean Society for Information Management), 24(2), 5-195.

[5] 박형준, 홍다혜, 김문현(2007), 주식 예측을 위한 은닉 마코프 모델의 이용, 성균관대학교 정보통신공학부 인공지능 연구실.

[6] 윤석영(2004), 주식수익률을 기초로 클러스터 기법을 이용한 주식스타일 분류에 관한 연구, 연세대학교 석사학위논문.

[7] 이원태(2011), 소셜미디어의 사회적 영향 및 전망, 한국인터넷 방송통신학회.

[8] 이인형(2007), 국내 주식 시장에서의 스타일 분류와 활용에 관한 연구, 한국증권학회.

Stock Prediction Method using Case-based Learning

Ju-Hyun Kim* · Min-Soo Jeon* · Yong-Gyu Jung**

In recent years, a number of people are going more and more to develope and engage in stock and equity-related information and related industry. a lot of stock expection programs came out, but they are still unstable, especially predictions methods and reality is unfounded. In this paper, we have vast amounts of stock data, shares many of the changes affecting the width of the survey items, and should seek the weights. This is related to existing stock levels and categories and is another way. Results based on a systematic classification of the stock data that would like to introduce objectivity

Key word: Smart Phones, Intrinsic Value, Network Value, User Satisfaction, Repurchase Intention

* Eulji University Department of Medical Industry, Medical Computer Science
** Eulji University IT Department of Medical Marketing, Professor, ygjung@eulji.ac.kr

저 자 소 개

김 주 현 (Ju-Hyun Kim)

을지대학교 의료산업학부 의료전산학전공에 재학하고 있으며, 의료정보시스템에 관심이 많으며, 특히 의료의사결정지원을 위한 다양한 결정요인을 마이닝기법을 통해 실험하고 연구하고 있다.

전 민 수 (Min-Soo Jeon)

을지대학교 의료산업학부 의료전산학전공에 재학하고 있으며, 의료정보시스템에 관심이 많으며, 특히 의료의사결정지원을 위한 다양한 결정요인을 마이닝기법을 통해 실험하고 연구하고 있다.

정 용 규 (Yong-Gyu Jung)

서울대학교, 연세대학교, 경기대학교에서 각각 학사, 석사, 박사 학위를 취득하였고, 현재 을지대학교 의료IT마케팅학과 교수로 재직중이다. ISO/TC154, UN/Cefact의 한국대표위원으로 활동하고 있으며, 의료정보, 전자무역, 해상물류, 금융전산에 Semantic Web, Process Modelling, ebXML 등의 표준기술의 적용에 관심이 많다.