한국지능시스템학회 논문지 2011, Vol. 21, No. 6, pp. 730-736
http://dx.doi.org/10.5391/JKIIS.2011.21.6.730

시계열 자료의 데이터마이닝을 위한 패턴분류 모델설계 및 성능비교

Pattern Classification Model Design and Performance Comparison for Data Mining of Time Series Data

이수용* · 이경중**+
Lee Soo Yong, Lee Kyoung Joung

연세대학교 인문예술대학 교양교직과
**+ 연세대학교 보건과학대학 의공학부

요 약

본 연구는 순차적인 시계열 자료들에서 가장 최근의 추세가 반영될 수 있는 패턴분류 모델을 설계하였다. 의사결정을 지원하는 데이터마이닝 패턴분류 모델을 설계할 때 통계 기법과 인공지능 기법을 융합한 모델들이 기존의 모델보다 우수함을 입증하였다. 특히 퍼지이론과 융합된 패턴분류 모델들의 적중률이 상대적으로 더 향상되었다. 예를 들어, 통계적 이론을 기반으로 한 SVM모델과 퍼지소속함수와의 결합, 혹은 신경망과 FCM을 결합한 모델들의 성능이 우수하였다. 실험에서 사용한 패턴분류 모델들은 BPN, PNN, FNN, FCM, SVM, FSVM, Decision Tree, Time Series Analysis, Regression Analysis 등이다. 그리고 데이터베이스는 시계열 속성을 지닌 금융시장의 경제지표 DB(한국, KOSPI200 데이터베이스)와 병원 응급실의 부정맥환자에 대한 심전도 DB(미국 MIT-BIH 데이터베이스)들을 사용하였다.

키워드 : KOSPI, KOSPI200, KODEX200, 부정맥, 심실빈맥(VT), 심실세동(VF)

Abstract

In this paper, we designed the models for pattern classification which can reflect the latest trend in time series. It has been shown that fusion models based on statistical and AI methods are superior to traditional ones for the pattern classification model supporting decision making. Especially, the hit rates of pattern classification models combined with fuzzy theory are relatively increased. The statistical SVM models combined with fuzzy membership function, or the models combining neural network and FCM has shown good performance. BPN, PNN, FNN, FCM, SVM, FSVM, Decision Tree, Time Series Analysis, and Regression Analysis were used for pattern classification models in the experiments of this paper. The economical indices DB with time series properties of the financial market(Korea, KOSPI200 DB) and the electrocardiogram DB of arrhythmia patients in hospital emergencies(USA, MIT-BIH DB) were used for data base.

Key Words : KOSPI, KOSPI200, KODEX200, Arrhythmia, Ventricular Fibrillation, Ventricular Tachycardia

1. 서 론

현실세계에서 미래를 예측하는 것은 항상 인간의 관심을 끌어 온 흥미로운 분야이다. 이를 위해 끊임없는 노력을 해 왔으며 대부분 과거의 경험에 의한 지식과 정보를 기반으로 미래를 예측하고 의사결정을 하였다. 이런 노력들은 결국 합리적인 사고방식에 의존하여 수학 및 통계학 그리고 인공지능의 예측방법을 찾게 되었다. [4,9]. 날마다 급변하는 복잡계의 금융시장의 주요 지표들이 주가, 환율, 금리 등의 소수 요인들에 의해 결정되는 것이 아니라 경제적 요인, 사회적 요인, 심리적 요인, 그리고 국제적 관계 등의 요인들이 복잡계를 이루고 있는 것이 사실이다. 또한 금융시장의 추세들이 카오스적이며 비정상적인 경우들이 발생되어 예측을 어렵게 하는 상황에서 투자자들이 객관적인 판단없이 기대 수익률에만 관심을 둔 비합리적인 투자를 하고 있는 실정이다. 실질적인 투자성과는 기대수익률뿐만 아니라 위험에 좌우되기에 위험관리에 대한 정확한 측정이 요구된다. 이에 주식시장에서는 주가가 가지는 패턴을 찾아 투자에 적용함으로써 경제적 이익을 얻고자 하였기에 이를 위한 다양한 투자전략과 주가예측기법들이 개발되었다.[2,4,7].

또한 인간의 생명을 유지하는 가장 중요한 기능 중에 하나는 심장의 역할이다. 인체의 구성요소들은 유기적으로 연계되어 있으며 생명에 대한 지표들은 주식시장에서 파생되는 다양한 지표보다 오히려 더욱 다양한 복잡계를 이루고 있기에 인체의 구성 역시 카오스적이어서 생명의 예측이 쉽지 않다. 심전도(electrocardiogram: ECG) 신호는 동방결절에서 발생하는 전기적인 신호에 의해 심장이 규칙적으로 수축과

접수일자 : 2011년 11월 3일
완료일자 : 2011년 12월 3일
본 연구는 “지식경제부”, “한국산업기술진흥원”, “강원광역경제권 선도산업지원단”의 “광역경제권 선도산업 육성사업”으로 수행된 연구결과입니다.
+ 교신저자 : 이경중 (연세대학교 의공학부)

이완할 때 체표면에서 기록한 신호이다. 심전도 신호 중 정상적인 범위내의 빈도를 벗어나거나 정상빈도를 갖더라도 패턴상의 이상이 있을 때를 부정맥(arrhythmia)이라 한다. 따라서 병원에서 심장질환인 부정맥 환자들의 경우도 주식시장에서 패턴을 찾아 투자전략을 위한 포토폴리오를 설계하듯이 부정맥 환자의 경우도 심장작동에 대한 심전도의 패턴인식에 의존하여 응급환자를 관리한다. [3,5,8,10,11,16,17]

주식시장의 주가예측 및 인체의 생체신호에 대한 예측이 어렵다는 연구결과의 대부분은 주가와 예측변수들이 기존의 이론에 바탕을 둔 선형관계를 가진다는 가정에 따른 결과이다. 그러나 기존의 선형에 기초한 이론으로 설명할 수 없는 포함한다는 가능성을 고려한다면 기존의 선형연구는 예측가능성에 대한 충분한 검증이라 볼 수 없으며 이러한 관점에서 통계적 모형과 인공지능 기법을 이용한 모형들을 상호 비교 분석하는데 목적이 있다.

인간의 정보처리 방식을 모방한 신경회로망은 적응학습기능, 대규모 병렬처리, 그리고 함수의 근사화 및 일반화의 장점을 가지고 있어 기존의 방법으로 해결하기 어려웠던 적응형 예측 및 제어, 음성인식, 패턴인식, 최적화 등 다양한 분야에서 활발히 응용되고 있다. 한편 인간의 사고 및 추론과정을 수학적으로 표현하여 전문가의 지식이나 불확실한 데이터를 처리하는데 효과적인 퍼지논리는 가전제품 및 공정제어 분야에서 우수한 성능을 보이고 있다.

한편 지금까지 산업 및 기업의 예측 및 진단 모델들은 전통적인 통계기법 즉 단순회귀 또는 다중회귀 모델, 연립방정식, 단일 다변량 모델 등에 근거하기 때문에 모델의 비적응성의 제약적인 가정의 요구, 처리의 복잡성 및 예측의 비정확성 등의 단점이 있다. 그동안 예측모델을 위해 회귀분석, 시계열분석, 판별분석, 군집분석, 위사결정나무, 신경회로망, 퍼지논리, 사례기반추론, 유전자알고리즘, 혼돈이론, SVM 등을 적용하고자 하는 시도가 많이 있다. 최근의 세계적인 기술개발 추세는 기존의 단일문제 해결을 위한 접근방식에서 점차 다양한 기술들을 결합하여 문제 해결을 추구하고 있으며, 신경회로망과 퍼지논리, 유전자 알고리즘과 혼돈이론, 그리고 퍼지이론과 SVM 과 같은 기술들을 각각의 장점들을 최대한 이용하여, 단점들을 서로 보완할 수 있는 융합형 모형의 개발이 절실히 필요해 지고 있다.

본 연구에서는 회귀분석(Regression Analysis), 시계열분석(Time series Analysis), 군집분석(HCM: Hard c-means clustering, FCM: Fuzzy c-means clustering), 의사결정나무(Decision Tree), 신경회로망(BPN: Back Propagation Neural Network, PNN: Probability Neural Network, FNN: Fuzzy Neural Network), 퍼지소속함수, SVM(Support Vector Machines), FSVM(Fuzzy Support Vector Machines) 등을 이용하여 금융지수인 KOSPI지수 패턴과 생체신호인 심전도신호의 부정맥 패턴을 예측하는 모형을 구축하고 성능을 서로 비교한다. 시계열 속성을 갖는 데이터베이스는 금융시장에서 파생되는 한국주식시장의 경제지표 데이터베이스와 병원의 부정맥 환자에 대한 미국 MIT-BIH의 심전도 데이터베이스를 사용하였다.

2. 시계열 데이터 패턴

2.1 KOSPI 지수

금융시장의 시계열데이터에 대한 향후 상승패턴과 하락패턴을 분류하는 것은 현물시장뿐만 아니라 선물시장과 옵션시장에서 매우 중요하다. 특히 금융선물시장에서 주가지수의 선물의 경우 만기일 전에 투기거래(speculation trading), 헤지거래(hedge trading), 차익거래(arbitrage trading), 스프레드거래(spread trading) 등의 반대매매를 통해 대부분 거래를 청산하기 때문에 KOSPI200지수의 향후 패턴분류는 의사결정에서 가장 중요한 기준이 된다.

1990년대에 들어 경제 전반에 걸친 자율화 및 개방화의 추진으로 구가불안성이 높아짐에 따라 투자자들이 위험관리기법을 고도화하여 운용수익을 안정적으로 확보하고 증시 안정화에도 기여할 수 있도록 1996년 5월 3일에 KOSPI200선물시장이 개설되었으며 이어서 1997년 7월 7일에 KOSPI200옵션도 도입되었다. KOSPI200 주가지수는 1,000여개 이상의 상장주식 종목 중에서 종목별 대표성과 지수조작 방지를 통한 공정성의 확보를 위하여 주요 우량종목 200개로 구성되었다. 이들 200종목의 시가총액은 전체 주직시장 시가총액의 70% 이상을 차지하고 있다. 기준 시점은 1990년 1월3일이며, 이날의 지수를 100포인트로 하여 기준으로 한다.

$$KOSPI200 = \frac{\text{비교시점에서 } KOSPI200\text{구성종목 시가총액 합계}}{\text{기준시점에서 } KOSPI200\text{구성종목 시가총액 합계}} \times 100$$

그림 1. 패턴분류 구간의 캔들스틱
Fig.1 Candle stick of pattern classification section

데이터베이스는 1999년 9월 3채주부터 2002년 9월 2채주까지 24개변수(국내 금융지수: KOSPI200-시가, 고가, 저가, 종가, 거래량, 거래대금, KOSPI-시가, 고가, 저가, 종가, 거래량, 거래대금, 원달러환율, 회사채수익률/ 미국 금융지수: DOW-시가, 고가, 저가, 종가, 거래량, NASDAQ-시가, 고가, 저가, 종가, 거래량)의 일별 데이터들을 이동평균법에 따라 주별데이터 로 변환 후 표준정규분포를 따르도록 전처리 하여 구성하였다.

주별데이터(157개)를 훈련데이터 101개(65%)와 테스트데이터 55개(35%)로 나누었다. 주어진 데이터베이스에서 목적변수 4개들을 T-test를 이용하여 각 변수선정을 구별하였다. 시가(open)는 10개, 고가(high)는 10개, 저가(low)는 10개, 그리고 종가(closed)는 8개가 선정되었다.

본 연구에서는 한국주식시장에서 KOSPI200지수의 주별데이터에 대한 시가(open price), 고가(high price), 저가(low price), 종가(closed price)를 분류하는 분류모델들을 비교하였다. 실험을 위해 설계한 분류모델은 BPB, SVM, Fuzzy SVM, Decision Tree, Regresion

Analysis, Time Series Analysis 들을 이용하였다.

2.2 ECG 신호

심전도(electrocardiogram: ECG) 신호는 동방결절에서 발생하는 전기적인 신호에 의해 심장이 규칙적으로 수축과 이완할 때 체표면에서 기록한 신호이다. 심전도 신호 중 정상적인 범위내의 빈도를 벗어나거나 정상빈도를 갖더라도 패턴상의 이상이 있을 때를 부정맥(arrhythmia)이라 한다. 부정맥 중에서 특히 심실빈맥(ventricular tachycardia: VT)와 심실세동(ventricular fibrillation: VF)은 심장의 무질서한 전기적 활동으로 인해 심근수축이 동시에 이뤄지지 않게 되어 발생한다. 이로 인해 심장의 혈액 공급이 중단되어, 신체 기관 및 뇌에 산소공급이 중단되게 된다. 뇌에 산소공급이 중단되면 뇌는 손상을 입게 되고, 결국은 몇 분 안에 뇌의 기능이 정지하여 급성 심장사(sudden cardiac death: SCD)에 이르게 된다. 심실빈맥과 심실세동은 환자의 생명을 위협하는 가장 치명적인 부정맥으로 즉각적인 치료를 하지 않을 경우 환자는 바로 사망하게 되므로 사전에 심실빈맥과 심실세동의 패턴인식이 매우 중요하다.

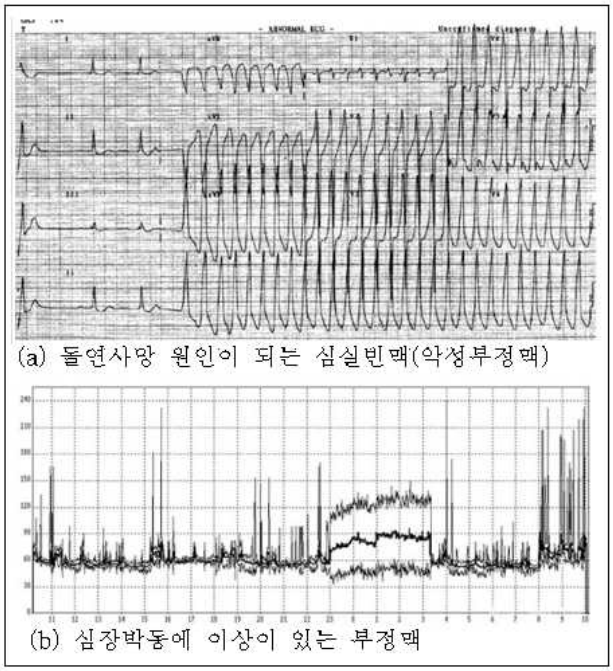

그림 2. 심전도 신호
Fig. 2. ECG (ElectroCardioGram signal)

그림 2는 정상리듬(normal sinus rhythm: NCR), 심실빈맥(ventricular tachycardia: VT), 심실세동(ventricular fibrillation: VF) 등 심전도의 부정맥의 예를 나타낸 것이다. 그림에서 보는 바와 같이 규칙성이 없는 기이한 양상을 보이며, QRS군이나 T파를 감별할 수 없이 불규칙한 진동파로 나타나는 특성이 있다. 한편 심실빈맥은 정상리듬과 심실세동의 중간 정도의 단계로서, 주기와 주파수 특성에서 심실세동과 유사한 특성을 갖고 있다.

심전도 신호 중 부정맥에 관한 파형의 측정분석은 파형의 크기 분석법[16], 비선형 분석법[14], 시주파수 분석법[1], 신경회로망[5,13,16,17], 퍼지추론[16,17,19], SVM [5,16,17] 등을 이용한 기존의 검출방법들이 있다. 심전도 신호의 전처리를 위하여 푸리에변환[12], 웨이브릿변환[15] 또는 힐버트변환[3] 등을 사용한다. 본 실험에서는 웨이블릿변환의 레벨에 의한 주파수 변환 및 상세계수를 이용하여 특징점을 추출하였다.

심실세동의 패턴분류를 위해 심전도 파형의 크기 분석, 비선형적 측정, 주파수 변환 등의 분석 방법들을 사용한다. 파형의 크기 분석방법은 측정하는 환경에 영향을 받기 때문에 신뢰도가 낮으며, 비선형 측정방법은 필터링과 잡음에 민감하다. 따라서 본 논문에서는 리듬기반의 정보 및 웨이브렛 변환계수를 선택하였다. 이를 위하여 MIT-BIH 부정맥 데이터베이스 중에서 다양한 부정맥이 포함되어 있는 10개의 레코드를 사용하였다.

실험에 사용된 10개의 레코드에는 심실빈맥 8,810개와 심실세동 7,941개 등의 부정맥 데이터이외에 정상 및 다른 데이터를 포함하여 총 89,766개의 데이터를 포함하고 있다. 학습 데이터는 4,136개(4.61%)이며, 테스트데이터는 85,630개(95.39%)로 구성하였다.

부정맥 패턴분류기에 입력할 입력특징을 추출하기 위하여 웨이브렛 변환기반의 대역통과 필터링, R점 검출, 입력특징 추출구간 설정을 수행한다. 또한 리듬기반의 정보 및 웨이브렛 변환계수를 선택하였다. [10, 16]

3. 퍼지소속함수와 SVM

3.1 *SVM*

*SVM*의 기본 원리는 훈련 패턴들을 고차원의 특징공간으로 사상(mapping) 시킨 후 두 분류사이의 여백(margin)을 최대화 시키면서 오분류률을 최소화 하는 의사결정함수(hyperplane)를 찾는 것이다. 하지만 *SVM*은 사상에 대한 정보 없이 특징공간에서 커널(kernel) 함수를 활용하여 최적의 의사결정함수를 찾는다. 최적의 의사결정함수는 지지벡터들의 결합으로 표현된다.

부호화된 학습집합 $S=\{(y_i,x)|i=1,2,...n\}$이 주어졌을 때, 각 훈련데이터 $x_i \in R^N$ 은 두개로 부호화된 부분 중 반드시 한 곳에 속하게 되며 부호는 $y_i \in \{-1,+1\}$이다.입력공간에서 최적의 결정함수를 탐색하는 것이 쉽지 않으므로, 입력공간의 차원보다 더 높은 차원의 특징공간으로 입력공간을 사상(mapping) 시키면 기대하는 최적의 의사결정함수를 탐색할 수 있게 된다.

$z=\phi(X)$이 R^N에서 특징 공간 Z로의 사상일 때, $W \circ Z+b=0$를 만족하는 (W,b)를 의사결정함수라 한다. 이 때 $W \in Z$, $b \in R$이고, X_i는 다음 함수에 의해 분리된다.

$$f(x_i)=sign(W \circ Z_i+b)=\begin{cases}+1, & y_i=+1\\ -1, & y_i=-1\end{cases} \quad (1)$$

선형분리 되지 않는 데이터들을 처리하기 위하여, 완화변수(slack variable) $\xi_i \geq 0$을 오분류 척도(measure)라 정의하면 (1)은 (2)로 수정된다.

$$y_i(W \circ Z_i + b) \geq 1 - \xi_i, \ (i = 1,...,l) \quad (2)$$

여기서 ξ_i는 의사결정함수를 만족하지 않는 X_i에 대한 오분류(misclassification) 척도이므로 $\sum_{i=1}^{l}\xi_i$는 훈련집합 S 에 대한 오분류 척도가 된다. 따라서 최적의 의사결정함수(hyperplane)는 (3)으로 표현 된다.

$$min \ \frac{1}{2} W \circ W + C \sum_{i=1}^{l} \xi_i$$
$$subject \ to \ y_i(W \circ Z_i + b) \geq 1 - \xi_i \quad (3)$$

라그랑지 승수 $\alpha = (\alpha_1, ..., \alpha_l)$를 도입하고, 커널다항식 K가 $K(x_i, x_j) = (1 + x_i \circ x_j)^d = \phi(x_i) \circ \phi(x_j) = Z_i \circ Z_j$을 만족하면 최적의 의사결정함수는 Mercer의 정리에 의해 (4)와 (5)로 표현할 수 있다.

$$max \ W(\alpha) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j K(x_i, x_j) \quad (4)$$

$$subject \ to \ \sum_{i=1}^{l} y_i\alpha_i = 0, \ (0 \leq \alpha_i \leq C, \ i = 1,...,l)$$

$$sign(W \circ Z + b) = sign(\sum_{i=0}^{l}\alpha_i y_i K(x_i, x_j) + b) \quad (5)$$

3.2 퍼지소속 함수와 $FSVM_l$, $FSVM_n$, $FSVM_a$

시계열 데이터들이 추세를 형성할 때 가장 최근 데이터들의 패턴에 영향을 많이 받는다면 순차적 성질을 갖는 퍼지소속 함수를 정의하여 각 훈련패턴에 적용하면 학습할 때 모든 훈련 데이터들이 획일적으로 취급되지 않고 순차 데이터들의 패턴에 영향을 받도록 학습시킬 수 있다. 따라서 SVM은 순차적 속성을 반영하는 퍼지소속 함수와 결합하여 확장된 SVM 패턴분류기 모델의 설계가 필요하다. Lin[6]은 퍼지소속 함수를 SVM의 완화변수(slack variable)에 적용하는 $FSVM$을 제안했지만, 다양한 시계열 속성을 갖는 데이터베이스에 퍼지소속 함수의 적용을 일반화할 필요가 있다. 따라서 본 논문에서는 Lin이 제시한 퍼지소속 함수를 일반화하여 새롭게 정의하고 시계열 속성을 지닌 심전도 데이터베이스의 부정맥 패턴분류를 위한 실험을 통해 일반화된 퍼지소속 함수를 적용한 $FSVM$이 패턴분류 성능을 향상시켰다.

$FSVM$의 특징은 SVM이 비선형 분류문제를 해결할 때 퍼지소속 함수와 결합한 훈련 데이터를 사용함으로 오분류 단위인 완화 변수들이 퍼지소속 함수의 영향을 받아 의사결정곡면의 기울기를 조정할 때 유연성을 지니도록 조정하는 한다.

부호화된 학습집합 $S = \{(y_i, x_i, s_i) | i = 1, ..., l\}$이 주어졌을 때, 각 훈련데이터 $x_i \in R^N$ 은 두개로 부호화된 목표데이터 $y_i \in \{-1, +1\}$의 한 곳에 반드시 속하며, 임의의 작은 수 σ에 대하여, $\sigma \leq s_i \leq 1$를 만족하는 $s_i \in R$ 를 퍼지소속 함수라 한다.

퍼지소속 함수값 s_i는 벡터 x_i가 한$Class$에 속하는 정도를 표시한 속성이고, $\xi_i \geq 0 \ (i = 1, ..., l)$는 SVM에서 오분류에 대한 오차의 척도이므로 $s_i\xi_i$는 서로 다른 가중치를 갖는 새로운 오차의 척도로 변환된다. 따라서 에서 최적의 의사결정함수는 (6)을 만족하는 해이다.

$$min \frac{1}{2} W \circ W + C\sum_{i=1}^{l} s_i\xi_i$$
$$subject \ to \ y_i(W \circ Z_i + b) \geq 1 - \xi_i \quad (6)$$

훈련 데이터가 시계열 속성을 갖는 경우, 가장 최근의 훈련 데이터가 패턴분류기의 학습효율을 높이는 중요한 데이터라면, 학습할 때 가장 순차적으로 최근 데이터부터에서 과거의 데이터로 갈수록 학습에 차등을 주는 가중치가 필요하다. 이는 모든 학습 데이터들을 획일적으로 학습에 영향력을 행사하도록 모델링하는 것보다 오분류률을 크게 낮출 수 있다. 따라서 시계열 속성을 갖는 데이터의 경우, 식(6)의 의사결정곡면은 식(3)보다 패턴분류 성능을 향상 시킨다.

Lee[7.8.9.10]는 시계열 속성을 갖는 퍼지소속 함수 $s_i = f(t_i), s_i = f_n(t_i)$, 그리고 $s_i = f_a(t_i)$들을 제안하였다. 위 소속 함수들이 적용된 FSVM을 각 각 $FSVM_l$, $FSVM_n$, 그리고 $FSVM_a$ 로 표시한다.

퍼지소속 함수의 하한(lower bound)를 $\sigma \in R^+$ 를 선택하자. 양의 정수 n과 시간 $t_i \ (i = 1,...,l)$에 대하여, $s_1 = f_n(t_1) = \sigma$이고 $s_l = f_n(t_l) = 1$ 일 때, 퍼지소속 함수를 표-1과 같이 표시한다.

표-1. 퍼지소속 함수 $f(t_i)$, $f_n(t_i)$, $f_a(t_i)$
Table-1.fuzzy membership function $f(t_i)$, $f_n(t_i)$, $f_a(t_i)$

s_i	퍼지소속함수	Type	FSVM 표시
$f(t_i)$	$\frac{1-\sigma}{t_l - t_1}t_i + \frac{t_l\sigma - t_1}{t_l - t_1}$	1차형	$FSVM_l$
$f_n(t_i)$	$(1-\sigma)\left(\frac{t_i - t_1}{t_l - t_1}\right)^n + \sigma$	다항식형	$FSVM_n$
$f_a(t_i)$	$\frac{1}{1+\exp(a - 2a\frac{t_i - t_1}{t_l - t_1})}$	지수형	$FSVM_a$

4. 실험 및 결과 고찰

4.1 KOSPI지수

KOSPI200 지수의 향후 추세는 한국주식시장에서 현물시장을 비롯하여 선물시장과 옵션시장에서자금운용의 포토폴리오를 구성할 때 매우 중요한 기준이 되는

지표이다. KOSPI200 지수의 시가(OPEN), 고가(HIGH), 저가(LOW), 종가(CLOSED)에 대한 향후 주별 상승 또는 하락의 패턴을 분류하는 데이터마이닝 분류기 모델을 설계하고 시뮬레이션 결과들을 비교하였다.

실험을 위한 패턴분류기들은 인공지능 기법으로 신경회로망(BPB, PNN, FNN), SVM(rbf, poly), FSVM(rbf, poly) 등을 설계하였고, 통계 기법으로 의사결정나무(D.T.), 회귀분석(Regr.), 시계열분석(T.S.) 등을 설계하였다. 표5에서 한국금융시장의 주가지수 KOSPI200 패턴분류 모델의 성능을 비교하였다.

Adeli-Hung 알고리즘을 이용한 퍼지신경회로망(FNN)의 경우, HCM에 의한 군집분석으로 찾은 중심점과 FCM에 의한 군집분석으로 찾은 중심점들을 각기 실험했을 때, FCM(Fuzzy c-means clustering)에 의한 FNN 분류모델의 적중률이 높았다.

표-2 KOSPI200 분류 적중률 (%)
Table-2 Classification hit rate of KOSPI200 (%)

Clustering	Signal	FNN
Hard c-means clustering	+ / -	65
Fuzzy c-means clustering	+ / -	75

통계기법인 회귀분석, 의사결정나무, 시계열분석, 그리고 SVM 등의 분류모형들과 비교한 결과이다. 시계열분석(Time series Analysis)은 ARIMA(0,1,0)인 확률보행 모형(Random Walk Model)로 식별되었다. 따라서 현시점의 값은 바로 전시점 자료에 오차를 더해지는 모형이므로 상승/하락에 대한 2원 분류를 랜덤하게 예측하므로 좋은 분류적중률을 기대하기 어렵다.

표-3 KOSPI200 분류 적중률 (%)
Table-3 Classification hit rate of KOSPI200 (%)

Regression			Decision Tree			TA	SVM	
BW	FW	Enter	Gini	Chi^2	Entropy		Poly	RBF
43.6	43.6	49.1	41.8	49.1	50.9	43.6	58.0	55.0

표-4 KODEX200 분류 적중률 (%)
Table-4 Classification hit rate of KODEX200 (%)

Classifier	OPEN	HIGH	LOW	CLOSED
BPN	62.7	68.6	74.6	61.2
SVM	60.3	64.7	70.6	58.5
FSVM$_{33}$	63.2	66.2	70.6	64.7

4.2 ECG 신호

부정맥 패턴분류 모형은 *SVM*, $FSVM_l$, $FSVM_n$, $FSVM_a$ 등으로 4원 분류 모형을 구현하였다. 다원분류기 모형을 위하여 일대다 정책(one-against-all)과 일대일정책(one-against-one) 중에서 일대다정책으로 4원 분류를 $4\times(4-1)$개의 2원 분류 모델로 구성하고 결과는 다수결 원칙으로 분류선택을 하였다. 파라미터는 $d=1$, $std=1$, 그리고 $C=10$으로 고정하였으며, 커널의 선택은 Linear, RBF, Polynomial들로 각기 설계하였다. 실험에 의하면 초기 응급상황을 예견할 수 있는 심실빈맥(VT)의 경우 $FSVM_n$(99.00%), *SVM*(98.75%), $FSVM_a$(98.06%), 그리고 $FSVM_l$(97.35%) 순서대로 높은 분류적중률을 나타내었다. $FSVM_n$(커널:rbf)의 분류적중률은 *SVM*, $FSVM_l$, $FSVM_a$보다 2.75%, 6.33%, 25.29% 우수하였다. 표6은 심전도(ECG) 신호에 대한 *SVM*, $FSVM_l$, $FSVM_n$, $FSVM_a$ 패턴분류 모델의 실험결과이다.

5. 결 론

(I). KOSPI200: 본 논문은 금융시장의 KOSPI200 지수의 시가, 고가, 저가, 종가에 대해 데이터마이닝 패턴분류기들의 분류 성능을 비교하였다. 전체 분류적중률 평균은 61.4%이다. 시가의 경우 76.3%로 높은 적중률을 보였는데 이는 시가의 경우 KOSPI200 종가와 DOW 종가에 종속적임을 확인할 수 있었다. 시가의 전

표 5. 한국금융시장의 주가지수 KOSPI200 패턴분류 모델의 성능비교 (단위: 분류 적중률 %)
Table 5. Performance comparison of KOSPI200 pattern classification (unit: hit rate %)

Model	BPN	PNN	FNN	SVM		FSVM		D.T.	Regr.	T.S.	ave.
kernel				RBF	POLY	RBF	POLY				
OPEN	85.0	38.0	58.0	89.0	85.5	84.0	90.9	89.1	90.9	52.7	76.3
HIGH	75.0	38.0	53.0	53.0	67.3	49.0	56.4	63.6	56.4	58.2	57.0
LOW	61.0	38.0	53.0	69.0	63.6	58.0	67.3	69.1	72.7	56.4	60.8
CLOSED	59.0	38.0	58.0	55.0	58.0	49.0	54.0	50.9	49.1	43.6	51.5
ave.	70.0	38.0	66.5	66.5	68.6	60.0	67.2	68.2	67.3	52.7	61.4

표 6. 심전도 신호의 부정맥 패턴분류 모델의 성능비교 (단위: 분류 적중률 %)
Table 6. Performance comparison of ECG signal's Arrhythmia pattern classification (unit: hit rate %)

Arrhythmia		SVM			$FSVM_l$			$FSVM_n$				$FSVM_a$			
Fuzzy		-			Linear Type			Polynomial Type			n	Exponetial Type			a
Kernel		LINEAR	POLY	RBF	LINEAR	POLY	RBF	LINEAR	POLY	RBF		LINEAR	POLY	RBF	
Hit Rate (units,%)	NSR	98.67	98.68	99.04	98.72	98.46	88.02	97.01	97.56	99.42	28	98.32	98.32	98.98	400
	VT	97.10	97.12	98.75	95.81	95.75	97.35	89.94	89.75	99.00	50	86.93	86.94	98.06	48
	VF	89.66	89.66	89.65	89.66	89.66	99.74	65.29	65.29	99.79	5	0.00	0.00	0.00	-
	Etc.	98.74	98.75	99.77	89.00	89.00	99.69	99.88	99.88	100.00	4	0.00	0.00	100.00	3100
	AVE.	96.04	96.05	96.80	93.30	93.22	96.20	88.03	88.12	99.55	21.75	46.31	46.32	74.26	1182.67

체 분류적중률 평균이 76.3과 비교하여 FSVM과 Regr.이 90.9%의 높은 분류적중률을 나타내었다. 반면에 종가의 경우 전체 분류적중률 평균이 51.5%에 대하여 Regr.과 T.S.들은 상대적으로 낮은 적중률 50.9%와 49.1%를 나타내었다.

(II). ECG: 심전도 신호와 같이 시계열 속성이 있는 데이터들의 패턴을 학습할 때 심실빈맥과 심실세동과 같은 응급상황의 심전도 패턴을 인식하는데 유용하였다. 특히 심실세동(VF)의 전단계인 심실빈맥(VT)의 패턴을 인식하는데 $FSVM_n$은 99.00%의 적중률을 나타내어 기존의 SVM보다 0.25%를, $FSVM$보다는 1.65%를 향상 시켰다. 따라서 심전도의 신호가 정상리듬 단계에서 심실빈맥 단계로 전환된 후 심실세동의 응급상황으로 전환되는 시계열성 데이터들을 학습할 때 보다 효과가 높다. 획일적인 데이터 학습만으로 구축된 심전도 부정맥 패턴분류기보다는 보다 더 빠르게 신속한 응급상황을 판별함으로서 이전보다 소생성공률의 향상을 기대할 수 있다.

(III). 훈련 데이터가 시계열 속성을 갖는 경우, 가장 최근의 훈련 데이터가 패턴분류기의 학습효율을 높이는 중요한 데이터라면, 학습할 때 가장 순차적으로 최근 데이터부터에서 과거의 데이터로 갈수록 학습에 차등을 주는 가중치가 필요하다. 이는 모든 학습 데이터들을 획일적으로 학습에 영향력을 행사하도록 모델링하는 것보다 오분류률을 크게 낮출 수 있다.

(IV). 본 연구를 통해 복잡계에서 형성되는 패턴을 분류하기 위하여 다양한 모델들을 융합하는 것이 바람하다. 예를 들면, KOSPI 패턴분류에서 퍼지신경망과 FCM알고리즘을 이용한 군집방법과 연동하여 처리하거나, ECG 패턴분류에서 퍼지소속 함수와 SVM을 결합함으로 기존의 모델들보다 분류적중률이 향상됨을 입증하였다. 다양한 분류모델의 장점들을 최대한 이용하고 단점을 보완할 수 있는 융합형 모형의 개발이 필요해 지고 있다.

참 고 문 헌

[1] V.X. Afonso and W. J Tompkins, "Detection Ventricular Fibrillation," *IEEE Engineering in Medicine and Biology Magazine,* vol. 14, no. 2, pp.152-159, 1995.

[2] H. Bae, S. Kim, H Kim, and B.W. Kwang, "A Comparative Study on the Prediction of KOSPI 200 Using Intelligent Approaches," *International Journal of Fuzzy Logic and Intelligent Systems,* vol. 3, no. 1, pp.7-12, June, 2003.

[3] Y.G. Jang, S.J. Jang, S.O. Hwang, and Y.R. Yoon, "Prediction of the Successful Defibrillation using Hilbert-Huang Transform," *Journal of The Institute of Electronics Engineers of Korea - SC,* vol. 44, no. 5, pp. 45-54, 2007.

[4] D.S. Kim, "Stock and News Application of Intelligent Agent System," *International Journal of Fuzzy Logic and Intelligent Systems,* vol. 3, no. 2, pp. 239-243, December, 2003.

[5] M.S. Kim and S.Y. Lee, "Pattern Classification for Biomedical Signal using BP Algorithm and SVM," *Journal of Fuzzy Logic and Intelligent Systems,* vol. 14, no.1, pp. 82-87, 2004.

[6] C.F. Lin, "Fuzzy Support Vector Machine," *IEEE Transactions on Neural Networks,* vol. 13, no, 2, March, 2002.

[7] S.Y. Lee, "Fuzzy Support Vector Machine for Pattern Classification of Time Series Data of KOSPI200 Index," *Journal of Fuzzy Logic and Intelligent Systems,* vol. 14, no. 1, pp. 52-56, 2004.

[8] S.Y. Lee, "On the Fuzzy Membership Function of fuzzy Support Vector Machines for Pattern Classification of Time Series Data," *Journal of Fuzzy Logic and Intelligent Systems,* vol. 17, no. 6, pp. 799-803, 2007.

[9] S.Y.Lee, A Comparative Study of Model Design and Performance Test in Financial Index Pattern Classification System, Doctoral Thesis, Yonsei University, 2004.

[10] S.Y.Lee, D.Y.Ahn, M.H.Song, and K.J.Lee, "The Classification of Electrocardiograph Arrythmia Pattern using Fuzzy Support Vector Machines", *International Jounal of Fuzzy Logic and Intelligen Systems,* Vol.11, no.3, pp.204-21-. 2011.

[11] S.Y.Lee, D.Y.Ahn, M.H.Song, and K.J.Lee, "The Arrythmia Pattern Classifier of Electrocardiograph using the Fuzzy SVM", *12th ISIS, Proceeding of ISIS2011, September*, pp.526-529. 2011.

[12] J. Mill , A. Inoue, "An application of fuzzy support vectors, Fuzzy Information Processing Society," *NAFIPS. 22nd International Conference of the North American, Chicargo, IL,* pp.302-306, 2003.

[13] K.I. Minami, H.Nakajima, "Real-Time Discrimination of Ventricular Tachyarrhythmia with Fourier-Transform Neural Network," *IEEE Trans. Biomedical Engineering,* vol. 46, no. 2, pp. 179-185, 1999.

[14] M.I. Oiws, A.H. Abou-Zied and A.M. Youssef, "Study of features based on Nonlinear Dynamical Modeling in ECG Arrhythmia Detection and Classification," *IEEE Transaction Biomedical Engineering,* vol. 49, no.7, pp.733-736, July, 2002.

[15] L.Y. Shyu, Y.H. Wu, and W. Hu, "Using wavelet transform and fuzzy neural network for VPC detection from the Holter ECG," *IEEE Transactions on Biomedicine and Engineering,* vol. 51, no. 7, pp.1269-1273, 2004.

[16] M.H. Song, J. Lee, S.P. Cho, K.J. Lee, and S.K. Yoo, "Support Vector Machine Based Arrhythmia Classification Using Reduced Features," *International Journal of Control, Automation, and Systems,* vol. 3, no, 4, pp. 571-579, 2005.

[17] M.H. Song, J. Lee, S.P. Cho, and K.J. Lee, "SVM Classifier for the Detection of Ventricular Fibrillation,"

Journal of The Institute of Electronics Engineers of Korea - SC, vol. 42, no. 5, 2005.

[18] E. Spyrou, and others, Fuzzy support vector machines for image classification fusing MPEG-7 visual descriptors, 2nd European Workshop, Integration of Knowledge, Semantics and Digital Media Technology, 2005-Dec, pp.23-30, 2005.

[19] T. Sugiura, H. Hirata, Y. Harada and T. Kazui, "Automatic Discrimination of Arrhythmia waveforms using Fuzzy Logic," *in Proc. Of IEEE Conf. on Engineering in Medicine and Biology Society,* vol. 20, no.1, pp. 108-111, Hong Kong, China, October-November, 1998.

[20] Y. Wang, S. Wang, and K.K. La, "A New Fuzzy Support Vector Machine to Evaluate Credit Risk," *IEEE Transactions on Fuzzy Systems,* vol. 13, no, 6, pp. 820-831, 2005.

저 자 소 개

Lee Soo Yong

Professor of General Education and Teacher Training, Yonsei University.
Major: Computer Science / Mathematics
Ph.D.(Com.Sci. Yonsei Univ., 2004)
Ph.D.(Mathe. KyungHee Univ., 1992)

Research Area : Data Mining, Pattern Classification.
E-mail : 0691@yonsei.ac.kr

Lee Kyoung Joung
(Corresponding author)

Professor of Biomedical Engineering, Yonsei University, Korea.
Ph.D. (Yonsei University, 1988),

Resear Area : Medical Instrumentation & Modeling
E-mail : lkj5809@yonsei.ac.kr