Fuzzy 이론과 SVM을 이용한 KOSPI 200 지수 패턴분류기

이 수 용 (연세대학교)
이 일 병 (연세대학교)

FUZZY 이론과 SVM을 이용한 KOSPI 200 지수 패턴분류기

이 수 용 [1] 이 일 병 [2]

2002. 10.12

[1] 연세대학교 문리대학 전산학과, hosu@hosu.yonsei.ac.kr , http://hosu.yonsei.ac.kr/~hosu

[2] 연세대학교 공과대학 컴퓨터과학과, yblee@csai.yonsei.ac.kr , http://csai.yonsei.ac.kr

요 약

데이터마이닝(Data Mining)의 패턴분류 기법은 데이타베이스에서 유용한 정보를 탐색하기 위해 훈련데이터를 기계학습한 후 새로운 입력 데이터를 학습된 패턴에 따라 분류하는 것이다. 최근 패턴분류에 있어서 각광을 받고 있는 SVM (Support Vector Machine : SVM) 패턴분류기 모델은 경험적 위험 최소화 기법인 ERM (Empirical Risk Minimization : ERM) 모형의 단점을 보완하여 구조적으로 오분류률(misclassification rate)을 최소화 시키는 SRM (Structual Risk Minimization : SRM) 모델이다.

SVM 은 통계학자인 V.N.Vapnik 에 의해 제안된 이원분류 모형으로 수학적 이론을 배경으로 오분류률을 최소화 시켜줌으로써 다양한 응용분야의 패턴분류 문제에서 우수한 분류 성능이 입증되고 있다.

본 논문에서는 ERM 모형인 기존의 신경망 (BPN : Backpropagation Neural Net) 학습모델과 SRM 모형인 SVM 모델들의 패턴분류 성능을 비교하였다. 이를 위해 주식시장 KOSPI 200 지수의 주별 데이타에 대한 향후 상승/하락의 이원분류 문제에 적용하였으며 기계학습 시간의 단축과 오분류률의 최소화를 위해 C.F.Lin 이 제안한 FSVM (Fuzzy Support Vector Machine : FSVM) 알고리즘을 KOSPI 200 지수에 적용하여 FSVM 패턴분류기가 BPN 및 SVM 패턴분류기들에 비해 예측률이 우수함을 보였고 컴퓨터 기계학습 시간을 대폭 단축시킬 수 있음을 입증하였다.

KOPI200 지수에 적용한 FSVM 패턴분류기는 파생상품인 선물시장 및 옵션시장에 대한 의사결정 지원시스템으로서 포토폴리오 설계시 유용한 모형이 된다.

Keyword : Fuzzy-SVM, SVM, Neural Net, Fuzzy Neural Net, Pattern

Classification, Kospi200, Data Mining.

1. 서 론

인간의 정보처리 방식을 모방한 신경망의 분류기능이 다양한 분야에 응용되고 있지만 최적화된 신경망 구조의 설계를 위해서는 다양한 실험을 통해 오분류(misclassification) 위험을 최소화 해야 하며 결과에 대한 일관성 유지 및 기계학습 시간의 단축 등을 위한 연구가 필요하다.

한편 인간의 사고 및 추론 과정을 수학적으로 표현하여 전문가의 지식이나 불확실한 데이터를 처리하는데 효과적인 Fuzzy 이론은 소속정도라는 상대적인 기준으로 애매한 데이터의 처리과정에 유용성이 있으므로 보다 향상된 모형의 개발을 위해 다양한 모델들과의 결합을 통해 좋은 성능을 보여 주고 있다.

최근에 패턴분류에 있어서 각광을 받고 있는 SVM (Support Vector Machine : SVM) 모델은 1998 년 V.N.Vapnik[6]에 의해 개발된 통계적 학습이론으로서 학습데이터와 범주 정보의 학습진단을 대상으로 학습과정에서 얻어진 확률분포를 이용하여 의사결정함수를 추정한 후 이 함수에 따라 새로운 데이터를 이원 분류하는 것으로 VC(Vapnik-Chervonenkis) 이론이라고도 한다. 특히 SVM 은 분류문제에 있어서 일반화 기능이 높기 때문에 많은 분야에서 응용되고 있다.

기존의 학습 알고리즘은 학습집단을 이용하여 학습오류(empirical error)를 최소화하는 경험적 위험 최소화 원칙(Empirical Risk Minimization : ERM)을 구현하는 것인데 비해 구조적 위험 최소화 원칙 SRM(Structual Risk Minimizarion : SRM)은 전체집단을 하위 집단으로 세분화한 뒤 이 집단에 대한 경험적 위험도를 최소화하는 의사결정함수를 선택하는 것이다.

SVM 은 신경망처럼 패턴분류나 함수 추정 등을 효과적으로 수행할 능력이 있지만 다음 같은 점에서 개선되는 것이 바람직 하다. 어떤 학습데이타는 패턴분류 문제에 있어서 다른 데이터들보다 더 결정적인 영향력을 가질 수도 있으므로 우선적으로 올바르게 분류되어져야 하며 잡음(noise) 또는 이상치(outlier)들은 기계학습 과정에서 영향력이 작아지길 원한다.

예를들어 순차 데이터들이 추세를 형성할 때 가장 최근 데이터들의 패턴에 영향을 많이 받는다면, 순차적 성질을 갖는 Fuzzy 소속함수를 정의하여 각 훈련 데이타에 적용하면 학습할 때 모든 훈련벡타들이 획일적으로 취급되지 않고 순차 데이터들의 패턴에 영향을 받도록 학습시킬 수 있다. 따라서 SVM 기법은 신경망, 퍼지이론, 유전자 알고리즘, 혼돈이론 등의 인공지능 기법과 통합한 패턴분류기의 설계가 바람직하다.

이에 C.F.Lin[6]은 Fuzzy 소속함수를 SVM 의 완화 변수(slack variable)에 적용하는 FSVM (Fuzzy SVM : FSVM) 모델을 제안했다. FSVM 의 특징은 SVM 이 비선형 분류의 문제를 해결하기 위해 Fuzzy 소속함수를 적용한 훈련 데이타를 사용함으로 오분류량의 단위가 되는 완화 변수들이 Fuzzy 소속함수의 영향을 받아 결정곡면의 기울기를 조정할 때 Fuzzy 소속함수의 성질이 반영되도록 학습시키는 것이다.

본 연구에서는 학습시간의 단축과 오분류률의 최소화를 위해 C.F.Lin 이 제안한 FSVM (Fuzzy SVM : FSVM) 알고리즘을 KOSPI 200 지수의 상승/하락에 대한 패턴분류 문제에 응용하였다. 특히 2 차원으로 정의된 소속함수를 n 차원 소속함수로 확장한 FSVM 패턴분류기가 BPN 및 SVM 패턴분류기들에 비해 예측률이 우수함을 보이고 컴퓨터 기계학습 시간을 대폭 단축시킬 수 있음을 입증하였다.

KOSPI 200 지수에 적용한 FSVM 패턴분류기는 파생상품인 선물시장 및 옵션시장에 대한 의사결정 지원시스템으로서 포토폴리오 설계시 유용한 모형이 된다.

2. 기존 연구

국내에서 금융지수에 대한 기존 연구는 주로 기업도산, 주가, 이자율, 환율의 예측 및 분류 등이다. 특히 현물시장과 선물시장 그리고 옵션시장을 통합한 투자 포토폴리오를 구축하려는 시도가 많았으며 이를 위해 현물시장 지

수의 예측 및 분류의 연구가 주된 내용이다. 기존 연구들이 주로 사용한 모델은 신경망, 유전자 알고리즘, 사례기반추론, Fuzzy 추론시스템, 카오스, 자기회귀모형, 다변량 판별분석, 다중회귀분석, 이동평균법 등이며 SVM 이 국내 금융지표에 응용된 사례는 아직까지 찾을 수 없었다. 반면에 국내에서 SVM 을 응용한 연구가 문자인식, 영상인식, 홍채인식, 그리고 문서분류 등의 분야에서 매우 활발히 진행되고 있다.

외국의 경우 Alan[1] 은 SVM 을 오스트리아 주식시장의 주식종목을 선택하는데 응용하여 포토폴리오 설계시 좋은 결과를 보여 주었으며, Chang[4] 은 SVM 을 이용한 시계열 분석을 하였고, Lin[7]은 순차 데이터에 Fuzzy 소속함수를 도입하여 학습의 유연성을 제안했으며, Tay[9]는 미국 금융지표의 예측을 위해 BP 와 SVM 을 적용한 결과 SVM 이 우수함을 입증했다.

그 외 응용 분야에서는 Yang[13]이 SVM 의 분류 성능을 이용하여 남녀 사진을 분류하였고, Hadzic[14]은 영상인식에 적용했다. Ana[14]는 SVM 처리결과를 기하적으로 해석하는 방법을 제시했으며, Takuya[10]는 polyhedral pyramid 를 Fuzzy 소속함수로 정의하여 Fuzzy SVM 분류기의 성능을 benchmark 데이터들과 비교 분석하여 FSVM 이 SVM 보다 우수함을 입증하였으며, Jeng[6]은 Fuzzy 신경망을 위한 Fuzzy 추론시스템을 개선하는 SVM 을 제안하여 소속함수가 Gausian function 일 때 SVM 이 Fuzzy 추론시스템의 규칙수를 결정할 수 있음을 보였다.

Suykens[14]는 이원분류문제의 해결을 위해 Least Squre SVM 을 제안했으며, Roobaert[8]는 간단한 학습 알고리즘으로 Direct SVM 을 제안하였고, Mangasarian[14]은 데이타마이닝에서 SVM 의 기능에 대해 연구하였다. 또한 SVM 의 빠른 학습시간을 위한 연구가 Anguita[2]와 Yang[13] 에 의해 이루어 졌다.

SVM 의 다양한 연구결과 및 연구과제들은 [14]에서 찾을 수 있다.

3. 관련된 내용

본 연구를 위해 필요한신경망, Fuzzy 이론, SVM, 그리고 FSVM 등의 개념과 이론 그리고 관련된 알고리즘들을 소개한다.

3-1. 신경망 (Neural Net)

선형분류를 위해 반복된 기계학습 이론은 1956 년 F.Rosenblatt 의 퍼셉트론 연구로부터 시작된 단층 신경망이다. 신경망은 병렬로 작동되는 간단한 성분들로 구성되고 이들 성분은 생물학적인 신경체계를 반영하며 네트워크 함수는 성분간에 연결을 나타낸다. 인간의 뇌를 구성하는 가장 기본적인 단위는 뉴런(neuron)이라는 것으로 뇌는 이러한 수많은 뉴런들이 서로 연결되어 있는 것이다. 새로운 정보는 뉴런의 수상돌기를 통해 입력된다. 뉴런은 입력신호를 간단한 변환과정을 거쳐 축색을 통해 이를 다른 뉴런에 전달한다. 여기서 시냅스는 신호의 강약을 조정하는 역할을 한다. 이러한 뇌의 움직임을 응용한 것이 신경망이다. 신경망은 패턴인식, 식별, 분류, 음성, 비젼, 제어와 같이 다양한 응용 분야에서 복잡한 함수를 실행할 수 있도록 훈련된다.

인공 뉴런은 뉴런이 갖는 입력과 가중치 벡터를 내적 연산해서 전달함수에 적용하는 방식으로 계산되는 처리 요소이다. 전달함수는 설계자에 의해 선택되고 인자들은 뉴런 입출력 관계가 어떤 특정 목표에 도달될 수 있도록 하는 학습 규칙들에 의해 조정된다. 가중치는 생물학적인 뉴런에서 시냅스의 강도에 해당되며 세포 몸체는 덧셈과 전달함수로 표현되어 뉴런 출력은 축색에서 신호를 나타낸다. 전달함수는 net 에 관한 선형 또는 비선형함수일 수 있다. 전달함수는 뉴런이 해결하고자 하는 문제에 관해 어떠한 조건을 만족하는 방식으로 선택된다. BP(back propagation)의 학습에 이용될 수 있는 미분 가능한 비선형 전달함수로 쌍곡선탄젠트 함수와 로지스틱 함수를 들 수 있다.

신경망의 BP(Back-Propagation) 알고리즘은 다층 퍼셉트론에 관련된 가중치 및 임계값들에 관한 해를 반복적으로 구하는 일반적인 방법을 나타낸다. 낮은 학습률이 사용되는 경우에 아주 안정적인 최속 강하법에 속하지만 수렴이 느린 단점을 가지고 있다. BP 의 성능을 높이기 위한 방법으로 모멘트(moment) 항을 추가하는 학습율의 적용 등이 고려될 수 있다. 시그모이드 전달함수에서 기울기가 가장 큰 값은 $x=0$ 에서 나타난다. 즉 학습의 속도는 부분적으로 기울기의 크기에 종속되기 때문에 모든 뉴런의 내적인 전달활동은 학습의 촉진을 위해 작게 유지되어야 한다. 이것이 입력의 척도를 변환해야 하고 가중치들을 작은 난수 값들로 초기화 시키는 이유이다.

3.2 Fuzzy 이론

복잡한 현상의 문제를 해결할 때 정확히 표현할 수 없는 애매한 부분들을 단순화하여 처리하는 경향이 있지만 단순화하는 과정에서 필연적으로 관련된 정보가 손실되게 마련이다. Fuzzy 이론은 데이터구조에 대해 0 또는 1 중에 어느 하나만을 선택하는 이분법과는 달리, Fuzzy 분류는 0 과 1 사이의 어떤 실수값과 대응시킬 수 있다. 즉 애매함을 처리할 수 있는 이론적 바탕을 제공하는 것이 Fuzzy 이론이다. Fuzzy 이론은 1965 년 Zadeh 에 의해 제안되어 많은 응용 분야에서 다양한 모델들과의 결합을 통해 좋은 연구 결과들을 보여준다.

자연에서는 단어의 의미가 잘 정의된 것처럼 보이지만 하나의 집합에 관한 라벨로서 단어가 사용될 때 객체의 집합에 대한 포함여부의 경계가 애매하거나 또는 막연한 경우가 많다. 예컨대 ' 빨간 장미' , ' 키가 큰 사람' , ' 아름다운 여자' , ' 신용이 있는 고객' 등은 경계가 애매한 Fuzzy 성질을 나타낸다. Fuzzy 성질은 고유적인 것과 정보적인 것의 두 가지로 구분될 수 있다. 먼저 고유적인 것의 예로서 ' 키가 큰 사람' 을 들 수 있는 데 키가 크다는

의미는 애매한 것으로 관측자의 키와 지역적 특성 등에 영향을 받게 된다. Fuzzy 논리는 의미적으로 막연한 개념들을 취급하는 Fuzzy 집합론과 막연한 성질을 판단 및 전개할 수 있는 Fuzzy 측도론으로 구성된다. Fuzzy 논리의 주요 개념은 어떠한 부정확성에 대한 허용 가능성이라 할 수 있다. Fuzzy 논리는 기호 및 수치처리를 한다는 의미에서 기존의 논리와 차이가 있다. 기존의 논리는 기호적인 조작만을 행하며 넓은 의미에서 Fuzzy 논리는 기존 논리의 확장이라 할 수 있다. Fuzzy 논리의 주된 목표는 정확한 것이라기보다는 근사적인 추론형태를 취급하기 위한 체계적이고 계산적인 기법과 개념을 전개하는 데 있다. Fuzzy 논리에서 정확한 추론은 근사 추론의 극한 개념이라 할 수 있으며 Fuzzy 논리에서 모든 것은 정도의 문제가 된다. 기존의 보통 집합은 유한 또는 무한 가산 집합 X의 원소 또는 객체 집합으로 각 원소는 집합 $A(\subset X)$에 포함되거나 또는 포함되지 않을 수 있다. 즉 포함되는 경우에 명제 'x는 A에 속한다.' 는 참이 되고 포함되지 않는 경우에 명제는 거짓이 된다. 사상은 집합론 형태와 함수적인 형태를 관련시키는 중요한 개념에 속한다. 하나의 사상으로서 특성함수는 유용하게 사용되며, 다음과 같이 정의된다.

$$\mu_A(x)=\begin{bmatrix}1, & x\in A\\ 0, & x\notin A\end{bmatrix}$$

여기서 μ_A는 전체집합의 원소 x에 관해 집합 A에서 소속함수이다. 이러한 소속함수의 개념은 전체집합 X 의 원소 x에서 전체집합 Y 의 두 원소중의 하나로의 사영(projection)이다. X 에서 정의된 집합 A에 관해, 함수 x를 사용한 경우 하나의 값 집합 $\mu_A(x)$가 존재한다. (공집합은 0 의 소속정도, 전체집합은 1 의 소속정도).

보통집합에서 전체집합 내에 하나의 원소에 관해 주어진 집합에 대한 소속함수와 비소속함수 간에 변화는 명백한 것이며 잘 정의되는 명백한 개념이라 할 수 있다. 그렇지만 전체집합 내에 하나의 원소에 관해 Fuzzy 집합에 관한 소속함수와 비소속함수 간의 차이는 명백하게 구분되는 것이 아니다. 여러 소속정도 간에 이러한 차이는 Fuzzy 집합의 경계들이 애매하고 막연하다는 사실에 비추어 보았을 때, 애매성 또는 막연함을 상술하는 함수를 통해

원소에 관한 소속정도가 측정되어야 한다. 요컨대 Fuzzy 집합은 다양한 소속함수정도의 원소를 포함하는 집합이라 할 수 있다.

보통집합의 원소들은 소속정도가 정확히 1 이 배정되지 않은 경우에는 포함될 수 없는 것이 되기 때문에 Fuzzy 집합은 보통집합과 대조를 이루게 된다. 하나의 Fuzzy 집합 원소들은 그들의 소속정도가 완전한 것이 아닌 다른 값이 될 수 있기 때문에 같은 전체집합에서 다른 Fuzzy 집합의 원소가 될 수도 있다. Fuzzy 집합의 원소는 함수적인 형태를 사용하여 멤버쉽 값들의 전체집합으로 사영(projection) 되어 진다 .

전체집합 내에 하나의 원소 x 가 Fuzzy 집합 A의 원소인 경우에 구간 [0, 1]에 포함되는 실수값으로의 사상인 소속함수가 정의될 수 있으며, 이 때 Fuzzy 집합은

$$A = \{(x, \mu_A(x)) \mid x \in X\},\ \mu_A(x) \in [0,1]$$

와 같이 나타낼 수 있다. 전체집합 X가 이산형이고, 유한인 경우에 Fuzzy 집합은 다음과 같은 합의 식으로 표현되며

$$A = \frac{\mu_A(x_1)}{x_1} + \frac{\mu_A(x_2)}{x_2} + \cdots = \sum_i \frac{\mu_A(x_i)}{x_i}$$

전체집합 X가 연속적이고 무한인 경우에 Fuzzy 집합은

$$A = \int \frac{\mu_A(x)}{x}$$

와 같이 나타낸다. 분수에 사용된 바 표기는 나눗셈의 의미가 아니라 하나의 구분을 나타내는 기호이고, + 기호는 덧셈을 나타내는 것이 아니고 함수적인 합집합이다. 또한 적분기호는 대수적인 의미의 적분이 아니고 연속형의 변수들에 관한 합집합이다.

3.3. SVM (Support Vector Machine) 와 FSVM (Fuzzy Support Vector Machine)

Vapnik[3,11]이 제안한 기계학습 알고리즘, 즉 SVM 의 기본 원리는 훈련 데이타들을 고차원의 특징공간으로 사상(mapping) 시킨 후 두 분류 사이의 여백(margine)을 최대화 시키는 결정함수(hyperplane)를 찾는 것이다.

사상(mapping)에 대한 정보가 없더라도 SVM 은 특징공간에서 커널(kernel) 이라는 내적함수를 활용하여 원하는 최적의 결정함수를 찾는다. 최적의 결정함수는 지지벡터(support vector)라는 몇 개의 입력 벡터들의 결합으로 나타낸다. 다음에 설명되는 부호화된 학습집합 $S=\{(y_i,x_i)\colon i=1,...l\}$ 가 주어졌을 때, 각 훈련데이타 $x_i \in R^N$ 는 두개로 부호화된 부분 중 반드시 한 곳에 속하게 되며 이 때 부호는 $y_i \in \{-1,1\}$ $i=1,...l$ 이다. 입력공간에서 적절한 결정함수의 탐색이 쉽지 않으므로 입력공간의 차원보다 더 높은 차원의 특징공간으로 입력공간을 사상(map)시키면 최적의 결정함수를 탐색할 수 있게 된다.

$z=\varphi(X)$ 이 R^N 에서 특징공간 Z 로의 사상일 때, $W\circ Z+b=0$ 를 만족하는 (W,b) 를 결정함수(hyperplane)라 한다. 이 때 $W \in Z, b \in R$ 이고, X_i 는 다음 함수에 의해 분리된다.

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

선형분리가 되지 않는 데이터들을 처리하기 위하여, 유화변수(slack variable) $\xi_i \geq 0$ 를 오분류 척도로 정의하면 결정함수는 다음같이 수정된다.

$y_i(\mathrm{w}\cdot \mathrm{z}_i+b)\geq 1-\xi_i,\ i=1,.....,l$ 여기서 ξ_i 는 결정함수를 만족하지 않는 X_i 에 대한 오분류(misclassification) 척도이므로 $\sum_{i=1}^{l}\xi_i$ 는 훈련데이타 집합 S 에 대한 오분류 척도가 된다. 따라서 최적의 결정함수(hyperplane)는 다음과 같이 표현 된다.

$$\text{minimize}\ \frac{1}{2}\mathrm{w}\cdot\mathrm{w}+\mathrm{C}\sum_{i=1}^{l}\xi_i$$

$$\text{subject to}\ y_i(\mathrm{w}\cdot\mathrm{z}_i+b)\geq 1-\xi_i,\ \xi_i\geq 0, i=1,.....,\ l$$

라그랑지 승수 $\alpha=(\alpha_1,...,\alpha_l)$을 도입하고, $\varphi: R^N \rightarrow Z$, $\varphi(X)=Z$ 일때, 다항식 커널함수 $K(x_i,x_j)=(1+x_i\cdot x_j)^d$가 $z_i \cdot z_j = \varphi(x_i)\cdot\varphi(x_j) = K(x_i,x_j)$. 을 만족하면 최적의 결정함수(hyperplane)은

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

$$\text{subject to } \sum_{i=1}^{l} y_i\alpha_i=0 \quad 0\le\alpha_i\le C, \quad i=1,.....,l$$

을 만족하는 $f(x)=sign(w\cdot z+b)=sign\left(\sum_{i=0}^{l}\alpha_i y_i K(x_i,x)+b\right)$ 이다.

SVM 의 개념은 훈련 데이터들을 서로 다른 두개의 Class 로 분류할 때 분류의 기준이 되는 결정함수를 학습 알고리즘을 이용하여 찾는 것이다

예를 들어 순차 데이터들이 추세를 형성할 때 가장 최근 데이터들의 패턴에 영향을 많이 받는다면 순차적 성질을 갖는 Fuzzy 소속함수를 정의하여 각 훈련 데이타에 적용하면 학습할 때 모든 훈련 데이타들이 획일적으로 취급되지 않고 순차 데이터들의 패턴에 영향을 받도록 학습시킬 수 있다. 따라서 SVM 기법은 신경망, 퍼지이론, 유전자 알고리즘, 혼돈이론 등의 인공지능 기법과 통합하여 확장된 패턴분류기 모델을 설계하는 것이 바람직하다.

이에 C.F.Lin[7]은 Fuzzy 소속함수를 SVM 의 유화변수(slack variable)에 적용하는 FSVM 모델을 제안했다. FSVM 의 특징은 SVM 이 비선형 분류문제를 해결할 때 Fuzzy 소속함수와 결합한 훈련데이타를 사용함으로 오분류량의 단위인 유화변수(snack variable)들이 Fuzzy 소속함수의 영향을 받아 결정곡면의 기울기를 조정할 때 유연성을 지니도록 조정하는 것이다.

즉, $S=\{(y_i,x_i,s_i) : i=1,..l,\ \sigma\le s_i\le 1,\ \sigma: small\ number,\ y=\{-1,1\}\}$ 라 하자 단, $X_i\in R^N$는 훈련 데이터이고, $y_i\in\{-1,1\}$는 부호화된 목표 데이터이며, $S_i=\{s_i real\ number, \sigma\le s_i\le 1\}$ 는 Fuzzy 소속정도 이다. Fuzzy 소속정도 s_i는 벡터 X_i가 한 *Class* 에 속하는 정도를 표시한 속성이고 ξ_i 는 SVM 에서 오분류에 대한 오차의 척도이므로 $s_i\xi_i$ 는 서로 다른 가중치를 갖는 오차의 척도이다. 따라서 FSVM 에서 최적의 결정함수(hyperplane)는 다음의 해이다.

$$\text{minimize} \quad \frac{1}{2} W \cdot W + C \sum_{i=1}^{l} s_i \xi_i$$

$$\text{subject to} \quad y_i (W \cdot Z_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots\ldots, \ l$$

C.F.Lin 은 순차데이타를 위한 Fuzzy 소속함수로 다음 같이 제안했다. Fuzzy 소속정도의 하한(lower bound)으로 $\sigma > 0$ 을 선택하며, 시간 $t_1 \leq \cdots \leq t_l$ 에 관한 소속정도의 함수 $s_i = f(t_i)$ 를 정의한다. 순차 데이타의 마지막 X_l 가 가장 최근의 추세를 반영함으로 소속정도를 $s_l = f(t_l) = 1$ 이라 하고, X_1 의 소속정도는 $s_1 = f(t_1) = \alpha$ 라 하자.

C.F.Lin 이 제안한 2 차원의 Fuzzy 소속함수는 다음과 같다.

$$s_i = f(t_i) = (1 - \sigma)\left(\frac{t_i - t_1}{t_l - t_1}\right)^2 + \sigma$$

본 논문에서는 C.F.Lin 의 2 차원 소속함수를 $n(n \geq 3)$ 차원으로 확대하여 KOSPI 200 지수의 분류 문제에 적용하였다. 실험에 의하면 $n = 1,2$ 일 때 보다 $n(n \geq 3)$ 차원에서 오분류률이 최소화 되었고 기계학습 시간도 크게 단축시킬 수 있는 유용한 결과를 얻었다.

4. 실험

4-1. 실험 환경

데이터베이스는 2000 년 1 월부터 2002 년 8 월까지 24 개 변수의 일별 데이터(국내금융지수:14, 해외금융지수:10) 들로 구성하였고, 일별 데이터를 단순이동평균법에 의해 주별 데이타 137 개로 변환하여 표준 정규분포를 따르도록 전처리 했다. 24 개 변수의 종류는 KOSPI(시가, 종가, 고가, 저가, 거래량), KOSPI 200(시가, 종가, 고가, 저가, 거래량, 거래대금), CBY(3 년만기), 원/달러환율 등의 국내 지표와 Dow-Jones 지수(시가, 종가, 고가, 저가, 거래

량, 거래대금), Nasdaq 지수 (시가, 종가, 고가, 저가, 거래량, 거래대금) 등의 미국 지표들로 구성 되었다. 실험에서는 로짓회귀 분석을 이용하여 KOSPI 200 지수의 등락에 영향이 있는 변수들을 선택했는데 KOSPI200(고가, 저가), KOSPI(시가, 고가), 원달러환율, DOW(시가, 저가), NASDAQ(고가, 저가), 회사채(3 년만기) 등 10 개 변수이다.

실험을 위한 시스템의 운영체제는 Windows 2000/sever 이고, CPU 는 Pentium IV (Dual CPU : 1+1 GHz)이며, RAM 은 1,024 Mb 이다. 이런 사양에서 C++로 구현된 SVM 및 FSVM 을 실험에 사용했다.

4-2. 실험 내용

(1). 순차 데이타(sequential data)에 대한 Fuzzy 소속함수의 검증

FSVM 에서 정의한 n차원 Fuzzy 소속함수의 유용성을 검증하기 위해 병원의 순차 데이터로(인디안 당뇨병) benchmark 한 실험 결과(표-1)는 SVM 이 83.33% 일 때 오히려 C.F.Lin 이 제안한 2 차원 소속함수에서는 70.00%이었지만, 3 차원 이상의 소속함수를 선택하면 SVM 과 동일한 결과를 보였다. 비록 SVM 과 FSVM 의 결과가 동일하지만 실험 내용에서는 다양한 파라메터들의 변화에 대해 FSVM 이 일관성있는 결과를 제시 하였다.

반면에 순차적 경제지표인 KOSPI200 지표에 대한 검증(표-2)에서는 FSVM 의 소속함수가 $n \geq 3$일 때 80.00% 로 기존의 SVM 의 70.00%보다 우수함을 입증하였으며, 병원의 심장(heart)에 대한 임상적 순차 데이터의 검증(표-3)에서도 SVM 이 60.00% 이고 신경망의 BP 알고리즘이 65.00% 인데 비해 제안하는 FSVM 은 $n \geq 3$일 때 63.00% 를 나타냈다. 또한 FSVM 의 경우 Kernel 함수는 RBF 가 Linear 와 Polynomial 일 때 보다 약간 우수하였다.

따라서 순차데이타들에 대한 실험결과(표-1, 표-2, 표-3)에 의해 SVM 과 Fuzzy 소속함수가 결합된 FSVM ($n \geq 3$)이 기존의 SVM 과 신경망의 BP 알고리즘 패턴분류기보다 오분류률을 최소화하는 성능이 있음을 입증 하였다.

실험에서 사용한 FSVM의 Kernel 함수는 RBF를 이용 하였다.

표-1 Diabetes Data (sequential data)

Classifier (Kernel = RBF)		Hit Rate (%)
	SVM	83.33
	n=1	70.00
FSVM	n=2	70.00
	n>2	83.33

표-2. KOSPI 200 Data (sequential data)

Classifier (Kernel = RBF)		Hit Rate (%)
	SVM	70.00
	n=1	70.00
FSVM	n=2	60.00
	n>2	80.00

표-3. Heart Data (sequential data)

Classifier (Kernel)		Hit Rate (%)
SVM	RBF	60.00
	BP	65.00
	Linear	85.00
FSVM	Polynomial	85.00
	RBF	63.00

(2). 컴퓨터 기계 학습시간 비교

표-4 에서 d 는 Polynomial 의 차수, c 는 타협점(trade-off), 그리고 sv 는 support vector 수를 의미한다. 표-4 에 의해 Kernel 함수가 Polynomial 일 때 SVM 의 학습 시간이 가장 많이 소요됨을 보였다.

Polynomial 에 대한 SVM 및 FSVM 의 파라미터들에 관한 실험(표-4)의 결과에 의하면 SVM 과 FSVM 이 동일한 적중률을 나타내는 (d=2, c=1,500 ; sv=98:59)의 경우에 6.8 배, SVM 이 좋은 적중률을 나타내는 (d=2, c=1,000 ; sv=98:60)의 경우에 8.5 배, 그리고 FSVM 이 좋은 적중률을 나타내는 (d=2, c=500 ; sv=98:66)의 경우 5.3 배의 훈련 시간 차이가 있음을 알 수 있다. 윗 결과를 통해 Fuzzy 소속함수가 완화변수(slack variable)와 결합된 FSVM 에서는 support vector 수가 적게 됨으로 기계학습을 위한 시간이 SVM 보다 상대적으로 단축됨을 기대하게 된다.

데이터 크기별로 학습시간을 조사한 실험(표-5)에서는 데이터들은 1997 년 1 월 2 일부터 2002 년 8 월 31 일까지 일별 데이터수 중에서 임의로 92, 127,158, 200, 249, 293, 629, 1,810 로 구분하여 학습시간에 대한 실험을 하였다.

가장 학습시간이 빠른 Kernel 함수는 RBF 이며 신경망의 BP 알고리즘보다 SVM 은 12% 를 FSVM 은 40% 를 단축하였다. 특히 Polynomial 함수의 경우 SVM(49.988 시간=179,957 초)이고 FSVM(4.1525 시간=14,949 초)이 소요됨으로 FSVM 을 이용할 경우 45.8355 시간을 단축하는 효과를 얻을 수 있다.

SVM 의 빠른 학습시간을 위한 연구에서 Anguita[2]은 gradient 를 빠르게 계산하기 위해 block-Toeplitz 행렬을 이용했으며, Yang[13]은 학습 수에 따르는 학습시간의 상계(upper bound)를 제시하는 알고리즘을 발표 하였다.

표-4. Polynomial Kernel 의 학습시간 (단위:초)

d	c	Hit Rate (%)		Time (sec)		SV	
		SVM	FSVM	SVM	FSVM	SVM	FSVM
1	100	73.33	73.33	0.4	0.2	111	111
	500	73.33	73.33	2.5	0.6	113	111
	1000	73.33	73.33	4.9	1.2	113	111
	1500	73.33	73.33	7.3	1.8	113	111
	2000	73.33	73.33	6.6	2.3	113	111
2	100	66.67	80.00	3.5	2.6	94	82
	500	66.67	70.00	23.9	4.5	98	66
	1000	66.67	56.70	43.3	5.1	98	60
	1500	66.67	66.67	70.3	10.3	98	59
	2000	66.67	60.00	89.7	10.9	98	58
Average		70.00	70.00	25.2	4.0	104.90	88.00

표-5. Data 의 크기별 학습시간 (단위:초)

#(data)	Linear		RBF		Polynomial	
	SVM	FSVM	SVM	FSVM	SVM	FSVM
92	5	5	1	1	6	6
127	20	15	1	1	20	15
158	17	11	1	1	13	11
200	23	21	1	1	22	20
249	25	20	1	1	30	20
293	32	7	1	1	31	6
629	145	3	2	1	117	2
–	–	–	–	–	–	–
1,810	6,853	31	2,632	2,118	179,957	14,949
			BP=2,973			

(3). 유용성을 검증을 위한 실험

유용성 검증을 위해 학습기간 및 테스트 데이터의 기간 이동방법을 적용하였다. 즉 2000 년 1 월 첫주부터 2002 년 6 월 둘째주까지 학습한 후 6 월 셋째주를 테스트하고, 2000 년 1 월 둘째주 부터 2002 년 6 월 셋째주까지 학습한 후 2002 년 넷째주를 테스트하는 방법으로, 2002 년 9 월 둘째주까지 시뮬레이션한 결과 Kernel 함수를 RBF 로 사용했을 때 적중률을 비교하면 신경망의 BP 알고리즘은 70%, SVM 은 77.78%, 그리고 FSVM 은 88.89% 를 보였다.

또한 위 기간에서 추세 전환점에 대해 옳게 분류하는지 여부를 실험하였는데 신경망의 BP 알고리즘은 64%, SVM 은 66.67%, 그리고 FSVM 은 73.33% 의 적중률을 보였다.

표-6. 추세 전환점에서 분류 성능 비교

Classifier	BPN	SVM	FSVM
Hit Rate (%)	64.00	66.67	73.33

표-7. 2002.6.셋째주-2002.9.둘째주 (시뮬레이션)

Classifier	NN	SVM	FSVM
(Kernel)	BP	RBF	RBF
Hit Rate (%)	70	77.78	88.89

5. 결론

FSVM 패턴분류기의 성능을 검증하기 위해 KOSPI 200 지수에 적용한 연구 결과는 다음 같이 요약 된다.

첫째, SVM (Support Vector Machine : SVM)에 Fuzzy 소속함수를 결합한 C.F.Lin 의 FSVM 패턴분류기는 소속함수의 차수가 2 차원 이하일 때보다 3 차원 이상일 때 오분류률이 최소화 됨을 입증 하였다.

둘째, FSVM 패턴분류기는 기존의 SVM 분류기와 신경망 분류기들에 비해 분류 성능이 우수함을 제공할 뿐만 아니라 기계학습 시간을 단축시킬 수 있었다. 특히 Kernel 함수가 Polynomial 인 경우 학습시간이 큰 폭으로 단축되었다.

셋째, FSVM 패턴분류기는 각 파라메터들의 값에 항상 일관된 결과를 유지하여 기존의 신경망에 비해 안정성을 유지하였다. 따라서 KOSPI200 지수의 파생상품인 선물시장 및 옵션시장에 대한 포토폴리오 설계시 FSVM 을 적용하는 의사결정지원 시스템의 구축이 바람직 하다.

References

[1]. F.Alan, " Stock Selection using Support Vector Machines" , www.kernel-machines.org

[2]. D.Anguita, " Fast Training of Support Vector Machines for Regression" , www.kernel-machines.org

[3]. N.Cristianini, " An Introduction to Support Vector Machines" , Cambridge University Press, 2000.

[4]. M.W.Chang, " Analsis of nonstaionary time series using support vector machines, www.kernel-machines.org

[5]. C.W.Hsu, " A simple Decomposition Method for Support Vector Machines" , Machine Learning, 46, 291-314, 2002.

[6]. J.T.Jeng, " Support Vector Machines for the Fuzzy Neural Networks" , www.kernel-machines.org.

[7]. C.F.Lin, " Fuzzy Support Vector Machines, IEEE Transactions on Neural Networks, Vol.13, No.2, March 2002.

[8]. D.Roobaert, " DirectSVM:A Simple Vector Machine Perceptron" , J. of VLSI Signal Processing 32, 147-156, 2002.

[9]. F.E.H.Tay, " Application of support vector machines in financial time series forecasting" , Omega 29, 309-317, 2002.

[10]. I.Takuya, " Fuzzy Support Vector Machines for Pattern Classification" , www.kernel-machine.org

[11]. Z.Weida, " Linear programming support vector machines" , J. Pattern Recognition Society, pp:1-10,2001.

[12]. V.N.Vapnik, " Statistical Learning Theory" , Wiley-Interscience Pub., 1998.

[13]. D.Yang, " Provably Fast Training Algirithms for Support Vector Machines" , www.kernel-machines.org.

[14]. Http://www.kernal-machines.org.

[15]. 김대수, " 신경망 이론과 응용(I)" , 하이테크정보, 1989.

[16]. 김대수, " 신경망 이론과 응용(II)" , 하이테정보, 1993.

[17]. 이수영, " 신경망, 유전자알고리즘, 퍼지이론, 카오스, 통계이론의 통합형 모델 개발" ,Kaist, 1995.

Pattern Classifier utilizing Fuzzy Theory and SVM

Lee, Sooyong
Dept. of Computer Science
Yonsei University

The method of pattern classification in data mining is to classify the pattern of new input data, after machine learning classifying training data into patterns, in order to find out useful information from database of a large scale. A SVM (Support Vector Machine) model which has been spot-lighted recently is a SRM (Structural Risk Minimization) model that minimizes the misclassification rate by complementing the defects of an ERM (Empirical Risk Minimization) model. Consequently, minimizing the misclassification rate structurally, a SVM model has the excellent classification ability in pattern classification.

The purpose of this paper is to present Pattern Classifier, an integrated model which can go over the limits of individual models by complementing mutually the merits and faults of ERM-type and SRM-type pattern classification, to find out useful information with the pattern classifier. Especially, it is pointed out that Fuzzy concept combined with SVM, a SRM model, a hitting ratio of pattern classification is raised and the machine learning hours is reduced.

To verify the presented model, we made Financial Index Pattern Classifier and show that applying it to dual classification problems with KOSPI 200 index, a hitting ratio of pattern classification is raised. Therefore, it proves that decision support system using an FSVM model is very useful in planning the portfolio of future market and option market, which are Financial Derivatives of stock market.