점진적 앙상블 SVM을 이용한 고객 분류 시스템

박상호° 이종인 박선 강윤희 이주홍
인하대학교 컴퓨터공학과
parksangho@datamining.inha.ac.kr°, leejongi7446@hanmail.net,
sunpark@datamining.inha.ac.kr, yjfloo@hanmail.net, juhong@inha.ac.kr

Customer Classification System Using Incrementally Ensemble SVM

Sangho Park° Jongin Lee, Sun Park, Yunhee Kang, Juhong Lee
Dept. of Computer Science and Engineering, Inha University

요 약

소비자의 신용 대출 규모가 점차 증가하면서 기업에서 고객의 신용 등급에 의한 정확한 고객 분류를 필요로 하고 있다. 이를 위해 판별 분석과 신경망의 역전파(BP:BackPropagation)를 이용한 고객 분류 시스템이 연구되었다. 그러나, 판별 분석을 사용한 방법은 불규칙한 신용 거래의 성향을 보이는 비정규 분포의 고객 데이터의 영향으로 여러 개의 판별 함수와 판별점이 존재하여 분류 정확도가 떨어지는 단점이 있다. 신경망을 이용한 방법은 불규칙한 신용 거래의 성향을 보이는 고객 데이터에 의해서, 지역 최소점(Local Minima)에 빠져 최대의 분류 정확률을 보이는 분류자를 얻지 못하는 경우가 발생할 수 있다. 본 논문에서는 이러한 기존 연구의 분류 정확률을 저하시키는 단점을 해결하기 위해 SVM(Support Vector Machine)을 사용하여 고객의 신용 등급을 분류하는 방법을 제안한다. SVM은 SV(Support Vector)의 수에 의해서 학습 성능이 좌우되므로, 불규칙한 거래 성향을 보이는 고객에 대해서도 높은 차원으로의 매핑을 통하여, 효과적으로 학습시킬 수 있어 분류의 정확도를 높일 수 있다 하지만, SVM은 근사화 알고리즘(Approximation Algorithms)을 이용하므로 분류 정확도가 이론적인 성능에 미치지 못한다. 따라서, 본 논문은 점진적 앙상블 SVM을 사용하여, 기존의 고객 분류 시스템의 문제점을 해결하고 실제적으로 SVM의 분류 정확률을 높인다. 실험 결과는 점진적 앙상블 SVM을 이용한 방법의 정확성이 기존의 방법보다 높다는 것을 보여준다.

1. 서 론

현대 사회는 정보화의 발전과 신용 거래의 증가로 개인의 무형 자산인 신용에 근거하여 각종 거래와 평가가 이루어지는 개인 신용 사회이다. 신용을 제공하는 기업으로서는 고객의 신용도를 최대한 정확히 분류 예측하여, 이를 근거로 신용 부여나 거래 중지 등의 고객 신용 관리를 수행 할 수 있어야 한다.

고객 분류란 금융, 보험업, 카드사 등에서 기존 가입자들의 축적된 특성 자료를 바탕으로 가입자들의 신용도를 파악하여 이들을 우량 또는 불량(good or bad)집단으로 혹은 A,B,C,D등의 여러 개의 그룹으로 분류하는 일련의 절차를 말한다.

현재까지 많은 방법들을 이용하여 고객의 신용을 분류하였다. 대표적으로 판별 분석(Discriminant Analysis)을 이용한 방법[1]과 신경망(Neural Network)을 이용한 방법[2]등이 있다.

판별 분석은 입력변수들의 값에 근거하여 유사한 사례들을 몇 가지 부류로 분별하는 기법이다. 판별 분석을 이용하여 고객의 신용을 분류하는 방법은 결과에 대한 수학적 설명(증명)이 가능하여 모니터링(Monitoring)을 할 수 있다는 장점이 있다. 그러나, 고객 데이터가 불규칙한 신용거래 성향을 보이고, 정규 분포를 따르지 않는 경우에는 여러 개의 판별 함수와 판별점이 존재하여 분류 정확도가 낮아지는 단점이 있다.

신경망은 인간 두뇌의 신경망을 모방하여 실제 자신이 가진 데이터로부터 반복적인 학습 과정을 거쳐 데이터에 숨어 있는 패턴을 찾아내는 기법이다. 신경망의 역전파(BP:BackPropagation)를 이용한 방법은 불규칙한 신용거래의 성향을 보이는 고객 데이터에 의해서, 분류 에러율을 최소화하지 못하는 지역 최소점(Local Minima)에 빠지는 경우가 발생할 수 있다.

이와 같이 기존 기법들은 불규칙 거래 성향을 보이는 고객 데이터에 의한 분류기 학습은 분류기의 분류 정확률을 낮아지게 하는 단점을 가지고 있다. 이를 해결하기 위해 본 논문에서는 점진적 앙상블 SVM을 사용하여 불규칙 거래 성향을 보이는 고객 데이터의 효율적인 학습과 높은 분류 정확률을 보이는 고객 신용 분류 방법을 제안한다. SVM(Support Vector Machine)은 최소의 일반화 에러를 가진 최적의 분류 평면을 결정하는 기법이다.[5] SVM은 차원(Dimension)보다는 SV(Support Vector)의 수에 의해서 학습 성능이 좌우된다.[3] 그러므로, SVM은 불규칙한 거래 성향을 보이는 고객 데이터들에 효과적으로 학습시킬 수 있어 고객 분류 시스템에

서 높은 분류 정확률을 보일 수 있다. 하지만, SVM은 근사화된 알고리즘을 이용하기 때문에 이론적인 분류성능을 보이지 않을 수 있다. 따라서, 본 논문은 SVM을 앙상블로 구성하여, 학습단계에서는 전단계의 SV를 점진적으로 분류자의 학습에 포함시켜 높은 분류 정확률을 보인다. 테스트 단계에서는 앙상블 SVM들의 분류 결정들을 다수결의 원칙(Majority Voting)에 의해서 통합하여 최종 분류 결정을 한다.

본 논문은 2장에서는 SVM에 대하여 간단히 알아보고, 3장에서는 점진적 앙상블 SVM을 이용한 고객 분류 방법에 대하여 설명하고, 4장에서는 실험, 5장에서는 결론 및 향후 과제를 살펴본다.

2. SVM

본 논문에서 사용하는 SVM은 Vapnik[5]에 의해서 1995년 이원문제를 해결하기 위해서 제안된 알고리즘이다. 먼저, SVM은 데이터집합의 입력이 X_i 이고, 출력이 y_i인 학습데이터의 집합을 D라고 하면 SVM은 매핑함수(Φ)에 의해서 입력공간의 데이터(X_i)를 특징 공간으로 매핑(Mapping)한다.

$$W^T\Phi(X_i)+b \le 1 \quad (y_i=-1\text{인 경우}) \tag{1}$$

$$W^T\Phi(X_i)+b \ge 1 \quad (y_i=+1\text{인 경우}) \tag{2}$$

특징 공간에서 데이터가 선형적으로 분리 가능하면 학습데이터 집합의 모든 요소들에 대하여 식(1)과 (2)를 만족하는 벡터 W와 스칼라 b가 존재하여 두 클래스를 분리할 수 있는 하이퍼플랜(Hyperplane, $W^T\Phi(X)+b$)들을 만들 수 있다. 이들 중에서 양과 음의 구역에 있는 데이터들의 분리를 가장 명확하게 하는 것을 최적의 하이퍼플랜으로 결정한다. 예를 들면, 그림1의 하이퍼플랜 A, B, C중에서 각각의 하이퍼플랜을 기준으로 양과 음의 두 영역 각각에 가장 근접한 데이터들의 거리(Margin: $\frac{1}{\|W\|^2}$)를 최대화시키는 B가 최적의 하이퍼플랜으로 결정된다.

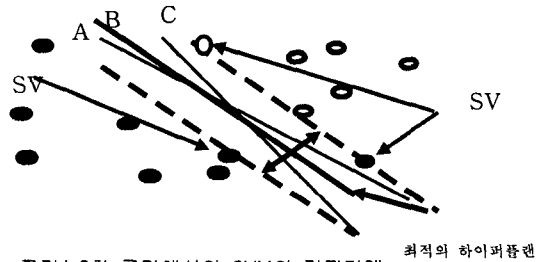

그림1.2차 공간에서의 SVM의 결정경계

즉, 선형으로 분리 가능한 경우에는 최적의 하이퍼플랜을 위한 W를 $\|W\|^2$의 최소화 과정을 통해 결정하고, 식 (3)의 선형결정함수를 이용하여 최적의 선형함수를 결정한다.

$$f(x)=sign\left(\sum_{i=1}^{l} y_i \alpha_i \cdot (X \cdot X_i)+b\right) \tag{3}$$

선형적으로 분리할 수 없는 경우에는 입력공간을 분리하는 비선형 결정면을 이용한다. 하지만 비선형 결정면의 식을 분석적으로 계산해낸다는 것은 어려운 일이므로, 다항식, RBF(Radial Basis Function), 다층 퍼셉트론(Multi-Layer Perceptron)등의 커널함수를 사용하여 입력 벡터X를 고차원 특징공간으로 매핑한 후, 선형의 경계선을 찾는 문제로 전환한다. 이처럼 커널함수를 사용하면은 입력벡터를 특징공간으로 투영시킴으로써 내적에 대한 계산만을 하므로 계산이 간편해진다. 결국, 입력공간에서 식(4)의 비선형 결정함수를 이용하여 최적의 선형 함수를 결정한다.

$$f(x)=sign\left(\sum_{i=1}^{l} y_i \alpha_i \cdot K(X, X_i)+b\right) \tag{4}$$

y_i는 학습데이터의 레이블, α_i는 랑그랑즈 승수, K( )는 커널함수, X는 입력데이터, X_i는 SV(Support Vector), b는 bias이다.

3.점진적 앙상블 SVM을 이용한 고객 분류 방법

앙상블(Ensemble)[4]이란 서로 다른 여러 개의 분류기들의 출력을 통합하여 최종 분류하는 일종의 복수 분류기 시스템(Multiple Classifier System:MSC)을 말한다.

3.1 점진적 앙상블 SVM 학습

본 논문은 고객 분류 프로세스의 2단계에서 SVM을 점진적 앙상블 구조로 구성한다. 점진적 앙상블 SVM은 학습과정에서 점진적으로 여러 개의 분류자를 학습시킨다. 이때에, 여러 개의 분류자 학습을 위해서 전체 학습데이터에서 분류자의 수만큼 랜덤하게 복원 추출하여, 각 분류자를 학습시킨다. 분류자의 학습은 전단계에서 학습되어진 분류자의 SV을 다음 단계의 분류자의 학습에도 포함시켜 점진적으로 이루어진다. 예를 들어,학습데이터가 T라고 했을 경우에, T에서 n개의 데이터를 복원 추출하는 과정을 분류자의 수만큼 반복한다. 여러 개의 분류자들 가운데 SVM1의 학습은 위해서 제일 먼저 복원 추출된 데이터를 가지고 학습되어진다. SVM2에서는 SVM1에서 얻어진 SV1과 다시 복원 추출된 학습데이터를 가지고 학습된다. 이와 같이 전단계에서 얻어진 SV들을 다음단계의 분류자의 학습에 포함시키는 점진적 앙상블 학습을 SVM1이 최종 학습되어질 때까지 수행한다. 즉, 최

종단계에서, SVM1은 SVM5에서 얻어진 SV5와 다시 복원 추출된 학습데이터를 이용하여 SVM을 다시 학습시킨다. 이러한 과정을 통해서 5개의 분류자를 가진 앙상블 구조의 SVM이 학습된다. 그림3.은 학습단계를 검정색으로, 테스트 단계를 빨간색으로 표시한 점진적 앙상블 SVM을 이용한 고객 분류 방법이다.

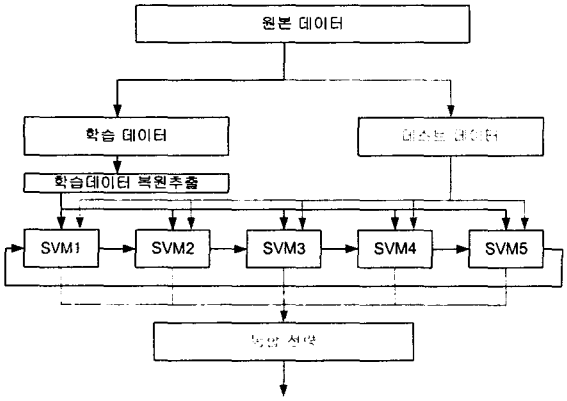

그림3. 점진적 앙상블 SVM을 이용한 고객 분류 방법

3.2 점진적 앙상블 SVM 테스트

테스트 단계에서는, 위에서 학습되어진 5개의 분류기를 가지고 똑같은 테스트 데이터에 대하여 분류작업을 한다. 분류되어진 결과는 N×M의 매트릭스를 만든다. 여기에서 N은 각각의 분류자를 의미하고, M은 분류될 클래스를 의미한다. 이렇게 N×M매트릭스의 M에 분류된 클래스를 '1'로 Voting한다. 각 클래스 마다 Vote의 수를 다른 클래스의 Vote와 비교한다. 비교되어진 Vote를 가장 많이 가진 클래스를 다수결의 원칙에 의해서 최종 고객 분류가 이루어진다.

분류 정확률은 오분류(Misclassification)된 경우의 수를 이용한 식(5)의 분류 에러(Classification Error)식을 이용하여 계산되어진다.

$$\text{분류에러} = \frac{\text{오분류된경우의수}}{\text{모든경우의수}} \quad (5)$$

4. 실험

본 논문은 펜티엄IV 1.5GHz, RAM 256M상의 리눅스 시스템 상에서 C언어로 구현하였다. 실험 자료는 UCI repository중에서 credit-screening의 실험을 위한 690개의 데이터를 가지고 실험하였다. 커널함수는 RBF(Radial Basis Function)을 사용하였고, 복원추출비율은 생략(Missing)될 학습 데이터의 경우의 수를 줄이기 위해서 80%로 하였다.

학습과 테스트의 비율은 7대3으로 하였다. 분류정확도는 식(5)의 분류 에러율로 측정했다. 학습된 분류자를 이용하여 테스트 데이터에 대한 실험결과는 그림4.와 같다. IE_SVM은 점진적 앙상블 SVM을 의미한다.

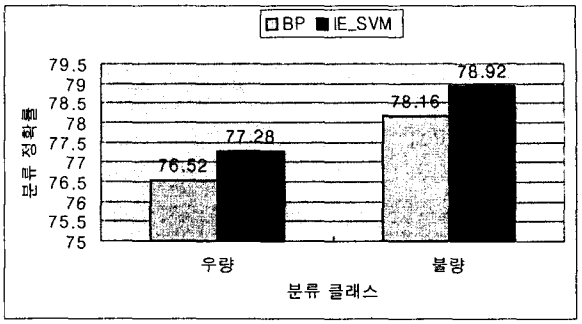

그림4. IE_SVM과 BP를 이용한 고객 분류 성능평가

5. 결론 및 향후 과제

위 실험결과, BP를 사용한 방법의 에러율은 우량 클래스의 경우 23.48%,불량 클래스는 21.84%를 보였다. 이에 비해서 본 논문에서 제시한 방법의 에러율은 우량 클래스의 경우 22.72%, 불량 클래스는 21.08%로 우량과 불량 클래스에 대하여 분류의 정확률이 각각 0.76%씩 높았다.

향후의 연구계획으로 SVM이 3원 분류이상에서 분류 정확도가 현저히 떨어지는 단점을 극복할 수 있는 분류 기법에 대하여 연구할 계획이다.

6. 참고문헌

[1] Altman. E. I., "Financial Ratio, Discriminant Analysis and the Prediction of Corporate Bankruptcy," Journal of Finance, No. 23, pp. 589-609, 1968.

[2] Atiya, Amir, F., "Bankruptcy prediction for credit risk using neural networks : a survey and new results," IEEE transactions on neural networks, Vol.12, No.4, pp.929-935, 2001.

[3] Joachims, T., "Text categorization with support vector machines: learning with many relevant features," In Proc. of the 10th European Conf. On Machine Learning, pp.137-142, 1998.

[4]Thomas G. Dietterich, "Machine Learning Research: Four Current Directions". The AI Magazine, vol 18, no.4, 97-136,1998.

[5] Vapnik, V., The Nature of Statistical Learning Theory. Springer-Verlag, 1995.