ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2013 June: 19(2): 125~140
http://dx.doi.org/10.13088/jiis.2013.19.2.125

불균형 데이터 집합의 분류를 위한 하이브리드 SVM 모델

이재식
아주대학교 경영대학 e-비즈니스학과
(leejsk@ajou.ac.kr)

권종구
아주대학교 일반대학원 경영정보학과
(karacocha@hanmail.net)

어떤 클래스에 속한 레코드의 개수가 다른 클래스들에 속한 레코드의 개수보다 매우 많은 경우에, 이 데이터 집합을 '불균형 데이터 집합'이라고 한다. 데이터 분류에 사용되는 많은 기법들은 이러한 불균형 데이터에 대해서 저조한 성능을 보인다. 어떤 기법의 성능을 평가할 때에 적중률뿐만 아니라, 민감도와 특이도도 함께 측정하여야 한다. 고객의 이탈을 예측하는 문제에서 '유지' 레코드가 다수 클래스를 차지하고, '이탈' 레코드는 소수 클래스를 차지한다. 민감도는 실제로 '유지'인 레코드를 '유지'로 예측하는 비율이고, 특이도는 실제로 '이탈'인 레코드를 '이탈'로 예측하는 비율이다. 많은 데이터 마이닝 기법들이 불균형 데이터에 대해서 저조한 성능을 보이는 것은 바로 소수 클래스의 적중률인 특이도가 낮기 때문이다.

불균형 데이터 집합에 대처하는 과거 연구 중에는 소수 클래스를 Oversampling하여 균형 데이터 집합을 생성한 후에 데이터 마이닝 기법을 적용한 연구들이 있다. 이렇게 균형 데이터 집합을 생성하여 예측을 수행하면, 특이도는 다소 향상시킬 수 있으나 그 대신 민감도가 하락하게 된다. 본 연구에서는 민감도는 유지하면서 특이도를 향상시키는 모델을 개발하였다. 개발된 모델은 Support Vector Machine (SVM), 인공신경망(ANN) 그리고 의사결정나무 기법 등으로 구성된 하이브리드 모델로서, Hybrid SVM Model이라고 명명하였다. 구축과정 및 예측과정은 다음과 같다.

원래의 불균형 데이터 집합으로 SVM_I Model과 ANN_I Model을 구축한다. 불균형 데이터 집합으로부터 Oversampling을 하여 균형 데이터 집합을 생성하고, 이것으로 SVM_B Model을 구축한다. SVM_I Model은 민감도에서 우수하고, SVM_B Model은 특이도에서 우수하다. 입력 레코드에 대해서 SVM_I와 SVM_B가 동일한 예측치를 도출하면 그것을 최종 해로 결정한다. SVM_I와 SVM_B가 상이한 예측치를 도출한 레코드에 대해서는 ANN과 의사결정나무의 도움으로 판별 과정을 거쳐서 최종 해를 결정한다. 상이한 예측치를 도출한 레코드에 대해서는, ANN_I의 출력값을 입력속성으로, 실제 이탈 여부를 목표 속성으로 설정하여 의사결정나무 모델을 구축한다. 그 결과 다음과 같은 2개의 판별규칙을 얻었다. 'IF ANN_I output value < 0.285, THEN Final Solution = Retention' 그리고 'IF ANN_I output value ≥0.285, THEN Final Solution = Churn'이다. 제시되어 있는 규칙의 Threshold 값인 0.285는 본 연구에서 사용한 데이터에 최적화되어 도출된 값이다. 본 연구에서 제시하는 것은 Hybrid SVM Model의 구조이지 특정한 Threshold 값이 아니기 때문에 이 Threshold 값은 대상 데이터에 따라서 얼마든지 변할 수 있다.

Hybrid SVM Model의 성능을 UCI Machine Learning Repository에서 제공하는 Churn 데이터 집합을 사용하여 평가하였다. Hybrid SVM Model의 적중률은 91.08%로서 SVM_I Model이나 SVM_B Model의 적중률보다 높았다. Hybrid SVM Model의 민감도는 95.02%이었고, 특이도는 69.24%이었다. SVM_I Model의 민감도는 94.65%이었고, SVM_B Model의 특이도는 67.00%이었다. 그러므로 본 연구에서 개발한 Hybrid SVM Model이 SVM_I Model의 민감도 수준은 유지하면서 SVM_B Model의 특이도보다는 향상된 성능을 보였다.

논문접수일 : 2013년 05월 21일 게재확정일 : 2013년 06월 21일
투고유형 : 학술대회 우수논문 교신저자 : 이재식

* 본 연구는 2012~2013학년도 아주대학교 일반연구비 지원에 의하여 연구되었음.

1. 서론

하나의 데이터는 다수의 레코드를 포함하고 있으며, 레코드는 여러 개의 속성으로 이루어져 있다. 속성은 레코드를 구분하는 목표속성과 목표속성에 영향을 주는 설명속성으로 구분할 수 있다. 목표속성 값에 따라 레코드들은 여러 개의 클래스로 군집화 될 수 있다. 각 클래스들이 비교적 균등한 개수의 레코드들을 포함하고 있을 때, 이 데이터 집합을 균형데이터 집합(Balanced Data Set)이라고 부르고, 어떤 특정 클래스가 다른 클래스들보다 현저히 많은 레코드들을 포함하고 있을 때 이를 불균형 데이터 집합(Imbalanced Data Set) 혹은 비대칭(Skewed) 데이터 집합이라고 부른다. 이러한 불균형 데이터 집합의 예로는 의료 분야에서 희귀한 질병을 가진 환자, 이동통신에서의 이탈 고객, 금융 분야에서의 신용불량자 등이 있다(McNamee et al., 2002).

불균형 데이터 집합을 분류하고자 할 때 우리가 알고자 하는 클래스는 소수 클래스이다. 즉, 희귀한 질병의 환자, 이탈 고객, 신용불량자 등이다. 하지만, 소수 클래스에 속한 레코드의 개수가 적기 때문에 소수 클래스를 분류하는 중요한 특징을 추출하기가 어렵다. 그러므로 데이터 마이닝 기법을 적용하여 불균형 데이터 집합을 분류하면 소수 클래스의 적중률이 낮아지게 된다.

이와 같은 불균형 데이터 집합의 분류 문제를 해결하기 위하여 다음 두 가지의 방법이 주로 사용되었다(Barandela et al., 2003). 첫째, 소수 클래스에 속한 레코드를 중복해서 추출하는 Oversampling이나 다수 클래스에 속한 레코드를 적게 추출하는 Under-sampling을 통하여 균형 데이터 집합으로 구성한 후 모델에 적용하여 학습을 시킨다. 둘째, 소수 클래스와 다수 클래스의 균형을 맞추기 위하여 소수 클래스에 의도적으로 학습을 기울이도록 분류 모델을 구성하는 것이다. 하지만 이러한 방법을 사용하면 소수 클래스의 적중률은 다소 올라가나 다수 클래스의 적중률은 낮아지게 되어 전체적으로 적중률이 낮아지는 문제점이 발생한다.

따라서 본 연구에서는 불균형 데이터 집합 분류에서 나타나는 문제점을 해결하기 위한 Hybrid Model을 제시하고자 한다. 즉, 불균형 데이터 집합을 분류할 때 상대적으로 낮게 나오는 소수 클래스의 적중률을 높이고 상대적으로 높게 나오는 다수 클래스의 적중률은 유지하고자 한다.

본 논문은 총 8장으로 구성되어 있다. 제 1장에서는 연구의 배경 및 목적과 논문의 구성에 대해서 기술하고, 제 2장에서는 불균형 데이터 집합에 대한 기존 연구들을 살펴본다. 제 3장에서는 본 연구에서 사용된 데이터 마이닝 기법들인 SVM(Support Vector Machine), 인공신경망, 의사결정나무에 대해서 간략하게 소개한다. 제 4장에서는 본 연구에서 사용된 데이터 및 연구를 위해 Sampling한 방법에 대해서 기술한다. 제 5장에서는 Hybrid Model의 유형에 대해서 기술한다. 제 6장에서는 본 연구에서 개발된 Hybrid SVM Model의 구축 과정에 대해서 설명하고, 제 7장에서 그 성능을 평가한다. 제 8장에서는 결론 및 향후 연구과제에 대해 제시한다.

2. 불균형 데이터 집합

불균형 데이터 집합을 분류할 때 각 클래스의 상대적인 차이점을 고려하지 않고 설계된 인공신경망이나 의사결정나무 모델들은 불균형 데이터 집합의 분류 문제를 해결하기에는 부적합 하다. 왜냐하면 이러한 모델들은 다수 클래스는 정확하게 분류하면서 소수 클래스는 무시하는 경향이 있기 때문이다(Jo and Japkowicz, 2004). Breiman et al.(1984)은 불균형 데이터 집합을 분류하면 소수 클래스가 다

수 클래스로 오・분류되는 비율이 높다고 하였다. 전체 오・분류를 낮추기 위해서 다수 클래스의 분류 적중률을 높이고자 학습을 많이 하게 되고, 그 영향을 받은 학습 패턴에 의해 소수 클래스가 다수 클래스로 오・분류되기 때문이다.

불균형 데이터 집합의 분류 문제점을 극복하기 위하여 다양한 방법들이 연구되어 왔다(Ganganwar, 2012). 예를 들어, 소수 클래스를 중복해서 추출하는 Oversampling(Chen et al., 2005), 다수 클래스를 적게 추출하는 Undersampling(Kubat and Matwin, 1997) 방법들이 있다. 또한 소수 클래스에 가중치를 주어 학습하는 방법(Cardie and Howe, 1997), 학습 규칙 자체에 가중치를 주는 방법(Grzymala-Busse et al., 2000), 소수 클래스를 부스트(Boost) 하는 방법도 있다(Joshi et al., 2001).

SVM으로 불균형 데이터 집합의 분류 문제점을 해결하기 위해서 많은 연구가 있었다. Veropoulos et al.(1999)은 클래스마다 일정한 페널티를 부여하는 것을 제안하였다. Wu and Chang(2003)은 불균형 데이터 집합을 균형 데이터 집합으로 인식하기 위하여 SVM의 주요함수를 바꾸는 알고리즘을 개발하였으며, Akbani et al.(2004)과 Calleja et al.(2011)은 SMOTE (Synthetic Minority Oversampling Technique) 기법(Chawla et al., 2002)을 기본으로 사용하여 향상된 성능의 기법을 개발하였다. Kotsiantis and Pintelas (2003)는 여러 분류기법들의 결과를 Voting하는 Hybrid Model로 적중률을 향상시켰으며, Chen et al.(2005)은 일단 구축된 SVM을 Pruning하는 방법으로 성능을 향상시켰다.

3. 사용된 데이터 마이닝 기법들

본 연구에서는 SVM, 인공신경망 그리고 의사결정나무의 세 가지 기법이 사용되었다.

SVM은 Vapnik(1995)에 의해 제안된 통계학적 학습 이론에 기반을 둔 기법이다. 이 기법은 1979년 Vapnik(1979)에 의하여 발표된 바 있으나, 최근에 와서야 그 성능을 인정받아 각광을 받게 되었으며 분류 문제에서 좋은 성능을 보이고 있다. 기존의 분류 기법들 대부분이 경험적 위험(Empirical Risk)을 최소화한다는 아이디어에 기초하는 반면, SVM은 일반화 에러의 상한(Upper Bound)을 최소화하는, 구조적 위험(Structural Risk)의 최소화에 기반을 둔 기법이다. SVM은 두 집단으로 구분된 입력값을 가지는 학습 데이터를 분류할 때, 기준이 되는 분리 경계면(Separating Hyperplane)을 특수한 학습 알고리즘을 사용하여 찾는다. SVM이 주목받는 이유는 첫째, 명백한 이론적 근거에 기반을 두므로 결과 해석이 용이하고(Cristianini and Shawe-Taylor, 2000), 둘째, 실제 응용에 있어서 인공신경망 수준 또는 그 이상의 성과를 내고, 셋째, 적은 학습 데이터로 신속하게 분류 학습을 수행할 수 있고, 특히 불균형 데이터 집합에 대해서 우수한 성능을 보이기 때문이다(Kim, 2012; Lee and Ahn, 2011; Min and Lee, 2005).

인공신경망(ANN: Artificial Neural Network)은 인간의 두뇌와 신경세포에 대한 연구에 기초를 두고 있는데 이는 인간 뇌의 가장 기본적인 단위인 뉴런(Neuron)을 모방한 처리요소(Processing Element)라 불리는 노드(Node)들 간의 연결로 이루어져 있다. 각 노드는 다수의 입력을 받아 하나의 결과값을 만들어낸다. 각 노드들은 서로 연결되어 앞 노드의 결과값이 후속 노드의 입력값으로 사용된다. 노드 간의 흐름은 입력값에서 결과값 방향으로만 진행되며, 역방향 진행이나 순환은 허용되지 않는 구조가 일반적이다(Linoff and Berry, 2011). 인공 신경망은 병렬분산 처리 시스템으로서 여러 개의 처리요소가 동시에 작동하며 학습을

할 수 있다는 큰 장점이 있다. 하지만 패턴을 추정하기 위해 다량의 학습 자료가 필요하고, 과잉적합(Overfitting)으로 인해 일반화의 어려움이 있을 뿐만 아니라 결과에 대한 해석이 어렵다는 한계점이 있다.

의사결정나무(Decision Tree)는 의사결정 진행 과정을 나무 형태로 표현한 것으로서 분류와 예측을 하는데 있어서 효과적인 데이터마이닝 기법이다. 인공신경망과 달리 분석 결과로 규칙이 도출되고, 규칙은 곧바로 글로 표현되어 사람이 쉽게 이해할 수 있다는 장점이 있다. 따라서 의사결정나무는 분류 또는 예측을 목적으로 하는 어떠한 경우에도 사용될 수 있으나, 분석의 정확도 보다는 분석 과정의 설명이 필요한 경우에 더 유용하게 사용된다(Linoff and Berry, 2011). 의사결정나무는 분류나 예측의 근거를 알려줌으로써 인공신경망이나 전통적인 통계학적 기법에 비해 모델의 구조를 이해하기 쉽다는 장점이 있으나 인공신경망에 비해 예측력이 다소 떨어질 수 있으며, 모델을 구축하는데 사용되는 레코드의 크기에 민감하다. 또한 연속형 속성을 범주형 속성으로 취급하기 때문에 분류의 경계점 부근에서 예측오류가 클 가능성이 존재한다.

4. 사용된 데이터

본 연구에서는 UCI Machine Learning Repository에서 제공하는 Churn Data Set을 사용하였다(Bache and Lichman, 2013). 이 데이터는 21개의 속성으로 구성되어 있으며 이동통신 고객의 유지(Retention)와 이탈(Churn)에 대한 총 5000개의 레코드가 있으며, 다수 클래스인 유지고객이 85%이고 소수 클래스인 이탈고객이 15%인 불균형 데이터 집합이다. Churn Data Set의 속성은 <Table 1>과 같다.

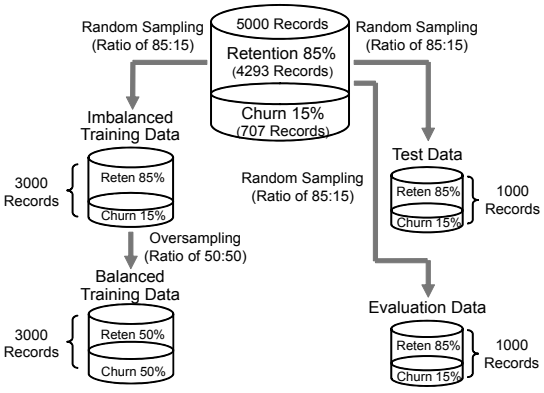
<Table 1> Attributes of Churn Data Set

Input Attributes	
Name	Description
STA	The 50 states and Washington D.C.
ACL	Account length : integer-valued variable for how long account has been active.
AC	Area code : categorical variable.
PN	Phone number.
IP	International Plan : dichotomous categorical having yes or no value.
VMP	Voice Mail Plan : dichotomous categorical variable having yes or no value.
NVM	Number of voice mail messages : integer-valued variable.
TDM	Total day minutes : continuous variable for number of minutes customer has used the service during the day.
TDC	Total day calls : integer-valued variable
TDG	Total day charge : continuous variable based on previous two variables.
TEM	Total evening minutes : continuous variable for minutes customer has used the service during the evening.
TEC	Total evening calls : integer-valued variable.
TEG	Total evening charge : continuous variable based on previous two variables.
TNM	Total night minutes : continuous variable for storing minutes the customer has used the service during the night.
TNC	Total night calls : integer-valued variable.
TNG	Total night charge : continuous variable based on previous two variables.
TIM	Total international minutes : continuous variable for minutes customer has used service to make international calls.
TIC	Total international calls : integer-valued variable.
TIG	Total international charge : continuous variable based on previous two variables.
NCSC	Number of calls to customer service : integer-valued variable.
Target Attribute	
Name	Description
OFF	An indication of whether or not that customer churned.

예측 시스템의 설계를 위해서는 적절한 Sampling 작업이 필요하다. 본 연구에서는 두 단계에 걸쳐 데이터 Sampling 작업을 하였다. 첫 번째는 불균형 데이터 집합을 만드는 단계로서 유지고객 85%(4293개), 이탈고객 15%(707개)로 구성된 전체 5000개의 레코드 중에서 3000개를 Sampling하여 Training (학습용) Data로, 1000개를 Sampling하여 Test(테스트용) Data로 사용하고, 나머지 1000개의 레코드는 모델의 최종 적중률을 측정하기 위한 Evaluation (평가용) Data로 사용하였다.

두 번째 단계에서는 불균형 데이터 집합으로 구성된 Training Data로부터 이탈고객을 Oversampling하여 유지고객 50%(1500개), 이탈고객 50%(1500개)로 구성된 클래스 간에 균형을 이룬 Training Data를 만들었다. Sampling 과정은 <Figure 1>과 같다.

이러한 Sampling 작업을 10번 하여 10-fold Cross Validation을 수행하였다.



<Figure 1> Sampling Process of this Research

5. Hybrid Model의 유형

Hybrid Model 사용의 목적은 하나의 문제를 해결하기 위해 여러 다양한 모델들을 사용함으로써 하나의 모델을 사용할 때보다 더 좋은 예측 성능을 얻고자 하는 것이다. Hybrid Model은 입력 데이터의 사용 방법, 하위 모델들의 역할, 하위 모델들로부터 얻어진 해의 결합 방법 등에 따라 여러 유형으로 분류될 수 있는데, Lee and Lee는 Hybrid Model을 구축하는 방법으로서 Whole Data Approach와 Segmented Data Approach의 두 가지 접근법을 제시하였다(Lee and Lee, 2006).

5.1 Whole Data Approach

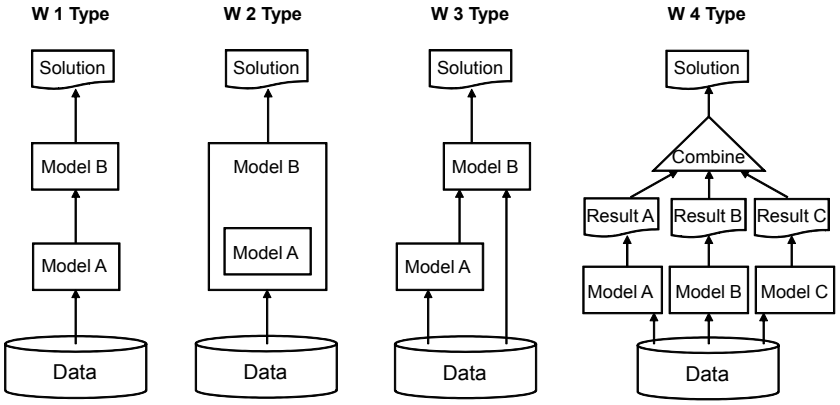
이 접근법은 입력 데이터 전체를 여러 다른 모델들에서 동등하게 사용하는 방법으로서 <Figure 2>와 같이 네 가지 유형이 있다.

W1 Type은 모델 A가 모델 B의 전처리 기능을 담당하는 유형이다. 예를 들어, 의사결정나무 기법으로 속성 선정을 수행한 후에, 그 선정된 속성들만으로 인공신경망 모델을 구축하는 경우이다.

W2 Type은 W1 Type처럼 두 개의 모델이 직렬적으로 수행되는 것이 아니고, 모델 A가 모델 B의 내부에 포함되어 있는 유형이다. 예를 들어, 인공신경망을 구축할 때에 속성 선정을 포함한 인공신경망의 구조를 유전적 알고리즘(Genetic Algorithm)을 이용하여 탐색하는 경우이다.

W3 Type은 모델 A를 수행하여 얻은 해(Solution)를 모델 B의 입력으로 사용하는 유형이다. 즉, 입력 데이터 전체가 모델 A에 사용되어 어떤 해를 (또는 해들을) 산출하고, 모델 B는 입력 데이터 전체와 그 해를(또는 해들을) 사용하여 최종 해를 도출하는 것이다. 예를 들어, 의사결정나무의 해가 인공신경망에서 하나의 입력노드를 차지하여 사용되는 경우이다.

마지막 유형인 W4 Type은 동일한 입력 데이터에 대해서 모델 A, 모델 B, 모델 C 등을 적용하여 해들을 얻은 후에 이들을 적절한 방법으로 통합하



<Figure 2> Whole Data Approach of Hybrid Model

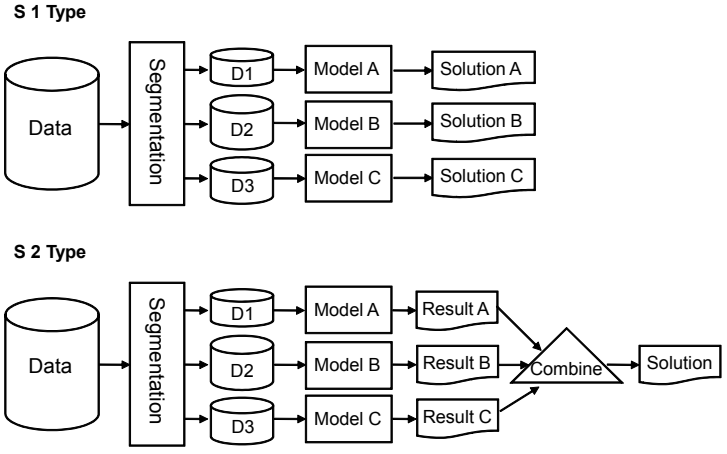
여 최종의 해를 도출하는 것이다

5.2 Segmented Data Approach

이 접근법은 입력 데이터를 분할한 후에 각 부분마다 다른 모델을 구축하는 방법이다. 최종 해를 구할 때에 각 모델이 도출한 해를 통합하느냐 하지 않느냐에 따라 <Figure 3>와 같이 크게 두 유형으로 나눌 수 있다.

입력 데이터를 분할하는 방법은 크게 두 종류로 나눌 수 있는데, 첫 번째는 입력 데이터를 속성의 특성에 따라 분할하는 것이고, 두 번째는 군집화 기법(Clustering Technique)과 같은 모델링 기법을 사용하여 분할하는 것이다. 첫 번째의 경우는 주로 해당 문제 영역에 대한 분석가의 영역 지식(Domain Knowledge)에 의해 수행된다.

분할된 데이터를 사용하는 대부분의 Hybrid Model에서는 하위모델로부터 얻어진 해들을 통합



<Figure 3> Segmented Data Approach of Hybrid Model

하는 단계가 없다. 왜냐하면, 데이터가 어느 부분에 속하느냐에 따라서 그에 해당하는 하나의 모델만 수행되기 때문이다.

하지만, 특수한 형태의 앙상블 모델에서는 S2 Type의 Hybrid Model이 사용된다. 예를 들어, 속성의 부분집합을 여러 개 만들고, 각 부분집합에 대해서 구축된 하위모델로부터 얻어진 해들을 결합하여 하나의 최종 해를 도출하는 것이다.

6. 이탈고객 예측 모델의 설계

6.1 SVM Model의 설계

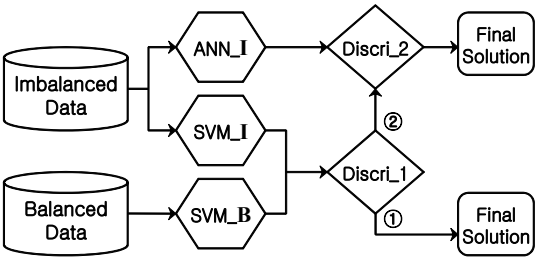
SVM Model의 설계에서는 격자탐색 알고리즘을 사용하여 SVM Model의 마진폭에 대응하는 분류오차의 페널티 C와 커널 파라메타 γ를 선정하였다. 5-fold Cross Validation을 이용하여 최적의 C, γ를 구하였으며 이를 이용하여 SVM 분석을 수행하였다. C와 γ의 값이 각각 2^3, 2^{-1}일 때 예측성과가 가장 우수한 것으로 나타났으며, 이때의 Training Data에 대한 적중률은 94.4%이었다.

6.2 ANN Model의 설계

ANN Model은 은닉층의 노드의 개수를 처음 10개부터 시작하여 점차 2개씩 증가시키면서 학습을 시켰다. 은닉노드의 개수가 14일 때 Training Data의 적중률이 제일 높았으며 14개를 전후로 적중률이 떨어지는 현상이 나타났다. 따라서 은닉노드의 개수를 14로 설정하였다. Training Data에 대한 적중률은 94.0%이었다.

6.3 Hybrid SVM Model의 구조

본 연구에서 개발하는 Hybrid SVM Model은 아래의 두 가지 목적을 달성하여야 한다. 첫째, Hybrid SVM Model의 민감도(Sensitivity : '유지'를 '유지'로 예측하는 비율)는 불균형 데이터 집합으로 학습한 SVM Model의 민감도의 수준이 되어야 하고, 둘째, Hybrid SVM Model의 특이도(Specificity : '이탈'을 '이탈'로 예측하는 비율)는 균형 데이터 집합으로 학습한 SVM Model의 특이도의 수준이 되어야 한다. 이 두 가지 목적이 달성된다면 자연스럽게 전체 적중률은 높아질 것이다. 본 연구에서는 SVM Model과 ANN Model의 결합을 통하여 <Figure 4>와 같은 예측모델을 제시 한다. 이 모델은 제 5장에서 소개한 Hybrid Model의 유형 중에서 S1 Type에 속하는 Hybrid Model이다.



<Figure 4> Structure of Hybrid SVM Model

<Figure 4>의 왼쪽에 있는 두 개의 데이터 집합 'Imbalanced Data'와 'Balanced Data'는 <Figure 1>의 Sampling 작업에 의하여 생성된 것이다. <Figure 4>에 제시된 Hybrid Model의 구축과정 및 예측과정은 다음과 같다.

불균형 데이터 집합으로 학습하여 ANN_I Model과 SVM_I Model을 만들고, 균형 데이터 집합으로 학습하여 SVM_B Model을 만든다. 입력 레코드에 대해서 SVM_I이 '유지'로, SVM_B도 '유지'로 예측하거나, 혹은 SVM_I이 '이탈'로, SVM_B도 '이탈'로 예측한 레코드는 Discri_1 Model에서 ①번 방향으로 예측과정이 진행되어 최종 해가 결정된다.

그 외의 레코드는 ②번 방향으로 예측과정이 진행된다. Discri_1 Model에서 ②번 방향으로 진행된 레코드는 ANN_I Model의 예측치를 이용해서 구축된 의사결정나무의 규칙을 이용하여 Discri_2 Model에서 최종 해가 결정된다. 본 연구의 Hybrid SVM Model에 포함된 각각의 Model별 역할은 <Table 2>와 같다.

<Table 2> The Roles of Individual Models

Basic Models	
ANN_I	ANN Model trained by imbalanced data.
SVM_I	SVM Model trained by imbalanced data.
SVM_B	SVM Model trained by balanced data.
Discrimination Models	
Discri_1	It discriminates the path of prediction process, i.e., ① for final solution and ② for further analysis, by comparing the results of SVM_I and SVM_B.
Discri_2	On the records coming from the path ②, it determines the final solution by applying the discrimination rules obtained from decision tree model.

6.4 Basic Models

본 연구에서 사용하는 3개의 Basic Model과 ANN_B Model의 Test Data에 대한 성능은 <Table 3>과 같다. ANN_B Model은 다른 모델과의 성능 비교를 위하여 참고적으로 포함시켰다.

<Table 3> Performance of Basic Models on Test Data (%)

	Sensitivity	Specificity	Accuracy
ANN_I	97.4	65.5	92.8
ANN_B	91.3	75.6	89.0
SVM_I	97.0	73.0	93.6
SVM_B	93.2	77.6	90.9

전반적으로 SVM Model의 적중률이ANN Model의 적중률보다 뛰어났다. SVM_I가 SVM_B보다 적중률은 높았으나, 민감도와 특이도에서는 성능이 엇갈렸다. 민감도는 SVM_I가 높았고 특이도는 SVM_B가 높았다. 이러한 결과를 토대로 제 6.3절에서 제시한 Hybrid SVM Model의 두 가지 목적을 다음과 같이 설정하였다. 즉, Hybrid SVM Model의 민감도는 SVM_I의 수준을, 특이도는 SVM_B의 수준을 유지하거나 또는 그 이상으로 향상 시키고자 한다.

6.5 Discrimination Models

Hybrid SVM Model에 있어서 Discrimination (판별) Model의 역할은 SVM_I와 SVM_B의 예측치를 통합하고, SVM_I와 ANN_I의 예측치를 비교하여 최종 해를 결정해주는 것이다. 본 연구에서는 다음과 같은 과정으로 Discrimination Model을 설계하였다.

Discri_1 Model은 SVM_I의 SVM_B의 예측치를 통합하는 모델로서, <Table 4>와 같은 규칙에 의해 수행된다. <Figure 4>의 Hybrid SVM Model에서 SVM_I와 SVM_B가 동일하게 예측한 경우에는 ①번 방향으로 진행되어 그 값이 최종 해로 결정된다.

<Table 4> Rules of Discri_1 Model

IF SVM_I result = Retention AND SVM_B result = Retention, THEN Final Solution = Retention
IF SVM_I result = Churn AND SVM_B result = Churn, THEN Final Solution = Churn
IF SVM_I result ≠ SVM_B result, THEN Proceed to the path ②

두 모델이 서로 다르게 예측한 경우에는 ②번 방

향으로 진행되어 Discri_2 Model의 판별 과정을 더 거치게 된다. ②번 방향으로 진행된 레코드들에 대해서는, 의사결정나무 모델을 사용하여 최종 해를 결정하였다. Training Data에 대한 ANN_I의 예측치를 입력값으로, 각 레코드의 이탈 여부를 목표값으로 설정하여 의사결정나무 모델을 구축하였다. Discri_2 Model의 규칙은 <Table 5>와 같다.

<Table 5> Rules of Discri_2 Model

IF ANN_I output value < 0.285, THEN Final Solution = Retention
IF ANN_I output value ≥ 0.285, THEN Final Solution = Churn

<Table 5>에 제시되어 있는 규칙의 Threshold 값인 0.285는 본 연구에서 사용한 데이터에 최적화되어 도출된 값이다. 본 연구에서 제시하는 것은 Hybrid SVM Model의 구조이지 <Table 5>에서 제시된 것과 같은 특정 Threshold 값이 아니기 때문에 이 Threshold 값은 대상 데이터에 따라 얼마든지 변할 수 있다.

6.6 Hybrid SVM Model의 Test Data에 대한 성능

완성된 Hybrid SVM Model과 SVM_I, SVM_B Model의 Test Data에 대한 성능은 <Table 6>와 같다. 본 연구에서는 Sampling을 10번 하여 10-fold Cross Validation을 수행하였으므로, 10개 Fold의 결과가 제시되어 있다.

<Table 6>에서 보듯이, Hybrid SVM Model의 적중률이 SVM_I나 SVM_B Model의 적중률보다 높다. 그리고 우리가 제 6.4절에서 밝힌 두 가지 목적, 즉 민감도는 SVM_I의 수준으로, 특이도는 SVM_B의 수준으로 얻고자 하는 목적이 달성되었다. Hybrid SVM Model이 다른 두 모델보다 우수한 것은 ROC(Receiver Operating Characteristic) 곡선의 분석으로도 알 수 있다. ROC 곡선은 '1-특이도'

<Table 6> Performance on Test Data (%)

	SVM_I			SVM_B			Hybrid SVM		
Fold	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
1	98.36	66.66	93.80	93.45	73.61	90.60	98.00	71.50	94.20
2	91.13	71.32	88.30	91.83	74.82	89.40	95.00	72.00	91.70
3	98.12	74.82	94.70	93.20	78.23	91.00	97.00	77.60	94.10
4	97.79	78.10	95.10	94.20	83.21	92.70	98.20	81.00	95.80
5	98.26	77.37	95.40	95.48	82.48	93.70	98.40	77.40	95.50
6	96.92	73.77	94.10	93.28	82.78	92.00	96.60	81.20	94.70
7	96.26	73.42	93.00	92.99	73.42	90.20	97.70	72.00	94.00
8	97.75	75.48	94.30	92.54	76.12	90.00	97.40	78.70	94.50
9	98.23	71.52	94.20	91.75	73.50	89.00	97.50	74.80	94.10
10	97.26	67.72	92.60	92.99	77.80	90.60	96.80	74.70	93.30
Avg	**97.01**	**73.02**	**93.55**	**93.17**	**77.60**	**90.92**	**97.26**	**76.09**	**94.19**

를 X 축으로, '민감도'를 Y축으로 찍은 점들을 연결한 곡선이다(Egan, 1975). 이 곡선의 아래 면적인 AUROC(Area Under ROC)가 넓을수록 민감도와 특이도가 적절하게 배분되어 있음을 나타낸다. Test Data에 대한 결과에서 도출한 세 Model의 AUROC는 <Table 7>과 같다.

<Table 7> AUROC of Models(Test Data)

Fold	SVM_I	SVM_B	Hybrid SVM
1	0.8252	0.8353	0.8477
2	0.8123	0.8333	0.8351
3	0.8648	0.8572	0.8725
4	0.8795	0.8871	0.8958
5	0.8782	0.8898	0.8788
6	0.8535	0.8803	0.8887
7	0.8485	0.8321	0.8485
8	0.8662	0.8434	0.8805
9	0.8489	0.8263	0.8618
10	0.8250	0.8542	0.8574
Avg	**0.8502**	**0.8539**	**0.8667**

<Table 7>에서 보듯이, Hybrid SVM Model의 AUROC가 가장 크다.

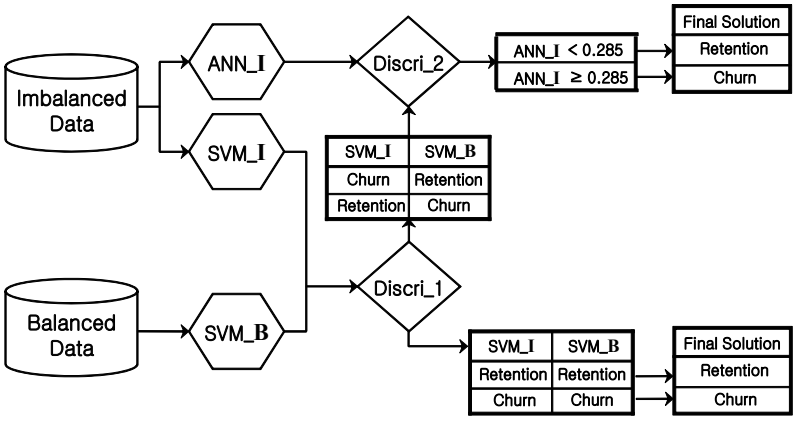
7. 하이브리드 모델의 성능 평가

이동통신 고객의 이탈예측을 위해서 완성된 Hybrid SVM Model은 <Figure 5>와 같다. 성능 평가를 위해 마련해 놓은 1000건의 Evaluation Data에 적용하여 얻은 예측 결과는 <Table 8>와 같다.

<Table 8> Performance on Evaluation Data (%)

			Prediction		Accuracy
			Retention	Churn	
Actual	SVM_I	Retention	94.65	5.35	90.76
		Churn	33.80	66.20	
	SVM_B	Retention	91.62	8.38	88.25
		Churn	33.00	67.00	
	Hybrid SVM	Retention	**95.02**	4.98	**91.08**
		Churn	30.76	**69.24**	

<Table 8>에서 보듯이, Hybrid SVM Model의 적중률이 SVM_I나 SVM_B의 적중률보다 높다. 비록 Hybrid SVM Model의 적중률이 SVM_I의 적중률보다는 미세하게 높지만, 우리가 주목할 점은 민감도와 특이도이다. Hybrid SVM Model의 민감도



<Figure 5> Hybrid SVM Model for Customer Churn Prediction

는 95.02%이었고, 특이도는 69.24%이었다. 동일 데이터에 대한 SVM_I의 민감도는 94.65%이었고, SVM_B의 특이도는 67.00%이었다. 그러므로 SVM 단일 기법만을 사용한 Model에 비해서, 본 연구에서 개발된 Hybrid SVM Model이 다수 클래스의 적중률은 유지하면서 소수 클래스의 적중률은 향상시키는 성능을 보였다.

Evaluation Data에 대한 결과에서 도출한 세 Model의 AUROC는 <Table 9>과 같다.

<Table 9> AUROC of Models(Evaluation Data)

SVM_I	SVM_B	Hybrid SVM
0.8043	0.7931	0.8213

<Table 9>에서 보듯이, Hybrid SVM Model의 AUROC가 가장 크다. 즉, Hybrid SVM Model은 적중률이 가장 높을 뿐만 아니라, 민감도와 특이도도 가장 적절하게 배분되어 있다.

8. 결론 및 향후 과제

본 연구에서는 불균형 데이터 집합의 분류에 있어서 두각을 나타내고 있는 SVM을 기본으로, 적중률을 높일 뿐만 아니라 민감도와 특이도에 있어서도 만족할만한 성능을 보이는 Hybrid SVM Model을 구축하였다.

연구를 수행하는 과정에서, 불균형 데이터 집합으로 학습한 SVM인 SVM_I Model, Oversampling하여 생성한 균형 데이터 집합으로 학습한 SVM인 SVM_B Model도 구축되었다. Hybrid SVM Model의 적중률이 이 두 SVM Model 보다 높게 나왔다.

여기서 주목할 점은 단순한 적중률의 증가가 아니라, 민감도와 특이도의 적절한 조화이다. SVM_I와 SVM_B의 성능을 비교해 보면, 민감도는 SVM_I가 높고, 특이도는 SVM_B가 높다. Hybrid SVM Model의 민감도는 SVM_I의 수준을 유지하였으며, 특이도는 SVM_B 보다 높게 나왔다. 즉, 불균형 데이터 집합에 대한 예측을 수행하고자 할 때에, 불균형 데이터에 대한 학습결과 및 그 데이터 집합으로부터 생성한 균형 데이터 집합에 대한 학습결과를 동시에 사용하고, 본 연구에서 제시한 판별규칙의 Threshold를 구하여 사용하는 것이 예측 적중률을 높이는데 효과적이라고 할 수 있다.

본 연구의 수행 과정에서 얻은 한계점 및 그에 따른 향후 연구과제는 다음과 같다.

첫째, Discrimination Model인 Discri_2의 규칙을 도출할 때에, ANN 이외의 다른 기법의 예측치도 사용하여 Hybrid Model의 유연성을 확장할 필요가 있다.

둘째, 본 연구에서 제시한 Hybrid SVM Model을 다양한 불균형 데이터 집합에 적용해 봐서 성능의 우수성을 일반화시킬 필요가 있다.

참고문헌

Akbani R., K. Wek, and S. J. Apkwicz, "Applying Support Vector Machines to Imbalanced Data Sets," *Proc. 15th European Conf. on Machine Learning*, (2004), 39~50.

Barandela, J., S. Sanchez, V. Garcaa, and E. Rangel, "Strategies for Learning in Class Imbalance Problems," *Pattern Recognition*, Vol.36(2003), 849-851.

Bache, K. and M. Lichman, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA : University of California, School of Information and Computer Science, 2013.

Breiman, L., J. H. Friedman, J. A. Olshen, and C.

J. Stone, *Classification and Regression Trees*, Wadsworth, 1984.

Calleja, J., A. Benitez, M. A. Medina, and O. Fuentes, "Machine Learning from Imbalanced Data Sets for Astronomical Object Classification," *Proc. Int'l Conf. on Soft Computing and Pattern Recognition*, (2011), 435~439.

Cardie, C. and N. Howe, "Improving Minority Class Prediction Using Case-Specific Feature Weights," *Proc. 14th Int'l Conf. on Machine Learning*, (1997), 57~65.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE : Synthetic Minority Over Sampling Technique," *Journal of Artificial Intelligence Research*, Vol.16(2002), 321~357.

Chen, X., B. Gerlach, and D. Casasent, "Pruning Support Vectors for Imbalanced Data Classification," *Proc. Int'l Joint Conf. on Neural Networks*, (2005), 1883~1888.

Cristianini, N. and J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, MA, 2000.

Egan, J. P., *Signal Detection Theory and Roc Analysis*. New York : Academic Press, 1975.

Ganganwar, V., "An Overview of Classification Algorithms for Imbalanced Datasets," *Int'l Journal of Emerging Technology and Advanced Engineering*, Vol.2, No.4(2012), 42~47.

Grzymala-Busse, J., X. Zheng, L. Goodwin, and W. Grzymala-Busse, "An Approach to Imbalanced Data Sets Based on Changing Rule Strength," *Proc. AAAI Workshop*, (2000), 69~74.

Jang, Y. S., J. W. Kim, and J. Hur, "Combined Application of Data Imbalance Reduction Techniques Using Genetic Algorithm," *Journal of Intelligence and Information Systems*, Vol.14, No.3 (2008), 133~154.

Jo, T. and N. Japkowicz, "Class Imbalances versus Small Disjuncts," *ACM SIGKDD Exploration*, Vol.6(2004), 40~49.

Joshi, M., V. Kumar, and R. Agarwal, "Evaluating Boosting Algorithms to Classify Rare Classes : Comparison and Improvements," *Proc. 1st IEEE Int'l Conf. on Data Mining*, (2001), 257~264.

Kim, M.-J., "Ensemble Learning with Support Vector Machines for Bond Rating," *Journal of Intelligence and Information Systems*, Vol.18, No.2(2012), 29~45.

Kotsiantis, S. B. and P. E. Pintelas, "Mixture of Expert Agents for Handling Imbalanced Data Sets," *Ann. Math. Computer Teleinformatics*, (2003), 46~55.

Kubat, M. and S. Matwin, "Addressing the Curse of Imbalanced Data Sets : One-sided Sampling," *Proc. 14th Int'l Conf. on Machine Learning*, (1997), 179~186.

Lee, H.-U. and H. Ahn, "An Intelligent Intrusion Detection Model Based on Support Vector Machines and the Classification Threshold Optimization for Considering the Asymmetric Error Cost," *Journal of Intelligence and Information Systems*, Vol.17, No.4(2011), 157~173.

Lee, J. S. and J. C. Lee, "Customer Churn Prediction by Hybrid Model," *Advanced Data Mining and Applications*, Lecture Note on Artificial Intelligence Vol.4093(2006), 959~966.

Ling, C. and C. Li, "Data Mining for Direct Marketing Problems and Solutions," *Proc. 4th Int'l Conf. on Knowledge Discovery and Data Mining (KDD-98)*, New York, 1998.

Linoff, G. and M. Berry, *Data Mining Techniques*, 3rd Ed., Wiley Pub. Inc., 2011.

McNamee, B., P. Cunningham, S. Byrne, and O. Corrigan, "The Problem of Bias in Training

Data in Regression Problems in Medical Decision Support," *Artificial Intelligence in Medicine*, Vol.24(2002), 51~70.

Min, J. H. and Y. C. Lee, "Bankruptcy Prediction Using Support Vector Machine with Optimal Choice of Kernel Function Parameters," *Expert Systems with Applications*, Vol.28(2005), 603~614.

Vapnik, V., *Estimation of Dependences Based on Empirical Data*, Nauka, Moscow, 1979.

Vapnik, V., *The Nature of Statistical Learning Theory*, Chapter 5. Springer-Verlag, New York, 1995.

Veropoulos, K., C. Campbell, and N. Cristianini, "Controlling the Sensitivity of Support Vector Machines," *Proc. Int'l Joint Conf. on AI*, (1999), 55~60.

Wu, G. and E. Chang, "Class-Boundary Alignment for Imbalanced Dataset Learning," *Proc. Int'l Conf. on Machine Learning : 2003 Workshop on Learning from Imbalanced Data Sets*, Washington, D.C., 2003.

Abstract

A Hybrid SVM Classifier for Imbalanced Data Sets

Jae Sik Lee* · Jong Gu Kwon**

We call a data set in which the number of records belonging to a certain class far outnumbers the number of records belonging to the other class, 'imbalanced data set'. Most of the classification techniques perform poorly on imbalanced data sets. When we evaluate the performance of a certain classification technique, we need to measure not only 'accuracy' but also 'sensitivity' and 'specificity'. In a customer churn prediction problem, 'retention' records account for the majority class, and 'churn' records account for the minority class. Sensitivity measures the proportion of actual retentions which are correctly identified as such. Specificity measures the proportion of churns which are correctly identified as such. The poor performance of the classification techniques on imbalanced data sets is due to the low value of specificity.

Many previous researches on imbalanced data sets employed 'oversampling' technique where members of the minority class are sampled more than those of the majority class in order to make a relatively balanced data set. When a classification model is constructed using this oversampled balanced data set, specificity can be improved but sensitivity will be decreased.

In this research, we developed a hybrid model of support vector machine (SVM), artificial neural network (ANN) and decision tree, that improves specificity while maintaining sensitivity. We named this hybrid model 'hybrid SVM model.' The process of construction and prediction of our hybrid SVM model is as follows.

By oversampling from the original imbalanced data set, a balanced data set is prepared. SVM_I model and ANN_I model are constructed using the imbalanced data set, and SVM_B model is constructed using the balanced data set. SVM_I model is superior in sensitivity and SVM_B model is superior in specificity. For a record on which both SVM_I model and SVM_B model make the same prediction, that prediction becomes the final solution. If they make different prediction, the final solution is determined by the discrimination rules obtained by ANN and decision tree. For a record on which SVM_I model and SVM_B model make different predictions, a decision tree model is

* Corresponding Author: Jae Sik Lee
Dept. of e-Business, School of Business Administration, Ajou University
San 5, Wonchun-Dong, Youngtong-Gu, Suwon 443-749, Korea
Tel: +82-31-219-2719, Fax: +82-31-219-1616, E-mail: leejsk@ajou.ac.kr
** Dept. of Management Information Systems, Graduate School, Ajou University

constructed using ANN_I output value as input and actual retention or churn as target. We obtained the following two discrimination rules: 'IF ANN_I output value <0.285, THEN Final Solution = Retention' and 'IF ANN_I output value ≥0.285, THEN Final Solution = Churn.' The threshold 0.285 is the value optimized for the data used in this research. The result we present in this research is the structure or framework of our hybrid SVM model, not a specific threshold value such as 0.285. Therefore, the threshold value in the above discrimination rules can be changed to any value depending on the data.

In order to evaluate the performance of our hybrid SVM model, we used the 'churn data set' in UCI Machine Learning Repository, that consists of 85% retention customers and 15% churn customers. Accuracy of the hybrid SVM model is 91.08% that is better than that of SVM_I model or SVM_B model. The points worth noticing here are its sensitivity, 95.02%, and specificity, 69.24%. The sensitivity of SVM_I model is 94.65%, and the specificity of SVM_B model is 67.00%. Therefore the hybrid SVM model developed in this research improves the specificity of SVM_B model while maintaining the sensitivity of SVM_I model.

Key Words : Data Mining, Imbalanced Data Set, SVM, Hybrid Model

저 자 소 개

이재식

현재 아주대학교 경영대학 e-비즈니스학과의 교수로 재직 중이다. 서울대학교 경영학과에서 경영학사, KAIST 산업공학과에서 공학석사, 그리고 미국 University of Pennsylvania, Wharton School에서 경영정보시스템 전공으로 경영학 박사학위를 취득하였다. Management Science, Decision Support Systems, Expert Systems with Applications, Annals of OR 등의 외국저널에 Heuristic Algorithm, Model Management, Case-based Reasoning, Context Reasoning, Recommender Systems 등을 주제로 한 논문들을 게재하였고, 국내저널에도 다수의 논문을 게재하였다. 주요 관심분야는 Data Mining, CRM, Recommender Systems, Ubiquitous Computing, Intelligent Information Systems, AI Application to Business Problem Solving 등이다.

권종구

아주대학교 공과대학 정보컴퓨터공학과에서 공학사, 아주대학교 일반대학원 경영정보학과에서 석사학위를 취득한 후 KT에서 근무하였고, 현재는 유학 준비 중이다. 주요 관심분야는 Data Mining, CRM, Recommender Systems 등이다.