기계학습을 이용한 불만족 고객의 예측

오세창*, 최민**
*세종사이버대학교 컴퓨터소프트웨어학과
**충북대학교 정보통신공학과
e-mail : scoh713@gmail.com

Prediction of Unsatisfied Customers Using Machine Learning

Se-Chang Oh*, Min Choi**
*Dept. of Computer Software, Sejong Cyber University
**Dept. of Information and Communication Engineering, Chungbuk National University

요 약

많은 기계학습 문제에서 특징 선택 문제는 전체적인 성능을 좌우하는 중요한 부분이다. 이는 불만족 고객의 식별 문제와 같이 수 많은 특징을 사용하는 문제에서 더욱 절실하다. 본 연구에서는 중요한 특징을 찾고 중복성을 제거하기 위한 몇 가지 대표적인 방법들을 불만족 고객의 식별 문제에 적용하였다. 이를 통해 먼저 정보 획득량 지표로 의미 있는 특징들을 선별하고, PCA 를 사용해서 남아있는 중복성을 줄이는 방법이 가장 좋은 결과를 얻었다.

1. 서론

고객 만족은 사업을 성공시키기 위한 핵심요소이다. 따라서 불만족한 고객을 찾아내어 적절히 대응하는 것은 매우 중요하다. 그러나 불만족한 고객들은 떠나가기 전까지에 좀처럼 속마음을 표현하지 않는다. 이러한 문제에 대응하기 위해서는 고객 관계의 초기 단계에서 수집된 데이터를 바탕으로 불만족한 고객을 식별하기 위한 기술이 필요하다. 최근 이를 위해 기계학습 기법을 적용하려는 시도가 진행 중이다. 영국의 쌘텐더 은행이 캐글 사이트에서 진행하는 콘테스트가 바로 그것이다 [1].

그런데 이 콘테스트에서 공개한 데이터를 보면 369 개의 많은 특징들을 사용하고 있어서, 처리 속도의 저하는 물론이고 정확도를 떨어뜨리는 요인이 되고 있다. 이는 이러한 특징들 중에 불만족 고객을 식별하는데 별로 도움이 안 되는 특징들도 포함되어 있고, 의미상 서로 중복된 특징들도 포함되어 있기 때문이다. 따라서 이러한 문제를 해결하고 보다 효율적으로 정확한 결과를 얻기 위한 기계학습 방법이 필요하다.

2. 관련 연구

기계학습 분야에서 특징 선택을 위해 그 동안 다양한 방법들이 연구되어 왔는데, 그 목적은 꼭 필요한 특징을 찾아내는 것과 데이터에 숨어있는 중복성을 찾아내어 특징 공간의 차원을 줄이는 것이다.

먼저 첫 번째로 꼭 필요한 특징을 찾기 위해서는 지도 학습 문제에서 각 특징이 데이터의 부류를 결정하는데 얼마나 기여하는가를 계산할 수 있어야 한다. 이를 위해 다양한 지표들이 개발되었는데, 이 중에는 각 특징이 분류 작업에 얼마나 도움이 되는지를 통계적으로 계산하는 ReliefF 지표 [2], 각 특징을 사용함으로써 엔트로피를 얼마나 줄일 수 있는지를 계산하는 정보 획득량 [3]. 정보 획득량이 특징의 고유 특성에 따라 과대 평가되는 것을 막기 위한 정보 획득 비율 [4], 한 특징에 대해 값에 따라 샘플들을 나누었을 때 각 그룹에 얼마나 다양한 부류가 포함되어 있는지를 나타내는 Gini 불순도 [3] 등이 있다.

두 번째로 데이터에 숨어있는 중복성을 줄이기 위해서 [5]에서는 주어진 데이터에서 거리 척도를 사용해 임의의 두 특징 간의 유사도를 계산한다. 이 때 중복성을 최소화하는 최적의 특징 집합을 구하기 위해 모든 조합을 다 계산하는 방법은 복잡도가 매우 높으므로 그리디 탐색 알고리즘을 사용한다. 그러나 이러한 방법으로는 유사도가 매우 높은 특징들을 배제하는 것은 가능하지만 몇 개의 특징들에 걸쳐서 내재된 중복성의 제거는 어렵다. 반면 PCA (주성분 분석) 방법은 고차원의 데이터를 저차원의 데이터로 환원시킴으로써 다수의 특징들에 걸쳐있는 중복성도 제거할 수 있다 [6].

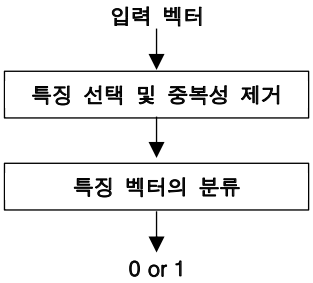
특징의 선택 문제 못지않게 중요한 부분이 바로 분류의 문제이다. 데이터를 분류하기 위한 방법으로 많이 사용되는 방법으로는 나이브 베이즈 [7], SVM [8], 인공신경망 [9] 등이 있다. 이러한 분류 방법들은 모두 데이터로부터 학습을 통해 분류 능력을 갖게 된다. 다만 이러한 방법들은 데이터의 특성에 맞게 선택되어야 최적의 결과를 얻을 수 있다.

본 논문에서는 앞에서 설명한 대표적인 특징 선택 및 분류 방법들을 적용해 봄으로써 불만족 고객의 예측 문제에 적합한 방법들을 찾고자 한다.

3. 예측 시스템

3.1 예측을 위한 처리과정

기계학습을 이용한 분류 시스템은 일반적으로 전처리 과정을 거쳐 입력 샘플을 가공하고, 분류 모듈에 넣어 각 부류 별로 가능성을 계산하고, 이 중 가장 가능성이 높은 부류를 선택하는 처리 과정을 거친다. 물론 분류 모듈은 훈련 데이터 집합에 의해 미리 훈련되어 있어야 한다. 본 논문에서 다루고자 하는 불만족 고객의 예측 문제에 이러한 기계학습 방법을 적용하면 (그림 1)과 같은 처리 과정이 가능하다.



(그림 1) 불만족 고객 예측을 위한 처리 과정

여기에서 예측하고자 하는 부류는 0 (만족) 또는 1 (불만족)로 두 가지이며, 입력 벡터는 (2, 66, 0, 0, ... , 0, 0, 63324.9)과 같이 369 차원 벡터가 된다. 물론 훈련 데이터 샘플은 여기에 부류 정보가 추가된 ((2, 66, 0, 0, ... , 0, 0, 63324.9), 1)와 같은 형태가 된다.

3.2 특징 선택

본 논문에서는 특징 선택을 위해 ReliefF, 정보 획득량, 정보 획득 비율, Gini 불순도 등의 지표를 사용하여 실험하였다. 이 중에서 ReliefF 지표는 식(1)과 같이 훈련 데이터의 각 샘플에 대해 서로 거리가 가까우면서 부류가 같은 샘플과 부류가 다른 샘플의 값의 차이를 각 특징 별로 취합하는 방법으로 각 특징이 분류 작업에 얼마나 도움이 되는지를 계산하는 통계적인 특징 평가 방법이다.

$$W_t(f) = W_{t-1}(f) - diff(f, E_1, H) + diff(f, E_2, F),$$
$$diff(f, E_1, E_2) = \frac{|value(f, E_1) - value(f, E_2)|}{\max(f) - \min(f)} \quad (1)$$

정보 획득량은 각 특징을 사용함으로써 엔트로피를 얼마나 줄일 수 있는지를 식(2)와 같이 계산하며, 이 값이 큰 특징이 분류에 유용하다고 판단한다. 이 식에서 $P(C_x|f_i)$는 특징 f의 값이 f_i인 샘플들 중에서 부류가 x인 샘플들의 비율을 의미한다.

$$InfoGain(f) = -\sum_{v \in values(f)} \left\{ \begin{matrix} P(C_0|v) log_2 P(C_0|v) + \\ P(C_1|v) log_2 P(C_1|v) \end{matrix} \right\} \quad (2)$$

정보 획득량은 샘플들이 특징 f의 값에 따라 고르게 분포할수록 커진다. 따라서 이를 보정하기 위해 고안된 지표가 바로 정보 획득 비율이다. 이는 식(3)과 같이 f의 값에 따라 샘플들을 나눌 때 분포의 특성에 의해 전체적으로 증가하는 정보량의 비율을 나누어 줌으로써 과대 평가되는 것을 방지한다.

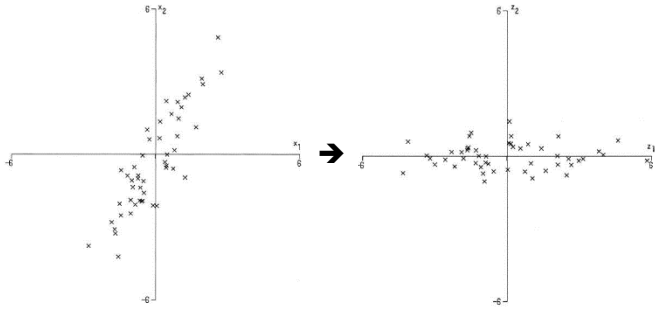
$$GainRatio(f) = \frac{InfoGain(f)}{\sum_{v \in values(f)} \frac{N_v}{N} log_2 \frac{N_v}{N}} \quad (3)$$

Gini 불순도는 주어진 훈련 데이터에서 각 특징 별로 값에 따라 샘플들을 나누었을 때 각 그룹에 얼마나 다양한 부류가 포함되어 있는지를 식(4)와 같이 계산하는 지표로, 이 값이 최소값인 0 에 가까우면 해당 특징은 샘플들을 분류하는데 상당히 유용하다고 볼 수 있다.

$$\text{Gini}(f) = \sum_{v \in values(f)} \{1 - P(C_0|v)^2 - P(C_1|v)^2\} \quad (4)$$

3.3 중복성 제거

많은 특징들을 다루는 기계학습 문제에서 특징들 속에 숨어있는 중복성을 제거하기 위해 일반적으로 사용되는 방법은 바로 PCA 방법이다. PCA 는 데이터를 한 개의 축으로 사상시켰을 때 그 분산이 가장 커지는 축을 첫 번째 주성분으로 선택한다.



(그림 2) 주성분을 기준으로 변환된 특징 공간

(그림 2)는 하나의 주성분을 선택하여 이를 기준으로 특징 공간을 변환한 결과를 보여준다.

두 번째 주성분부터는 앞에서 구한 주성분과 직교되면서 분산이 가장 커지는 축을 선택한다. 이러한 과정을 반복해 서로 직교되는 주성분들을 분산이 큰 순서로 얻게 되는데, 여기에서 분산이 큰 성분이 중요한 성분이다. 따라서 분산이 큰 몇 개의 주성분을 선택하고, 이들을 좌표계로 데이터를 선형 변환함으로써 중복성을 최소화할 수 있다.

3.4 특징 벡터의 분류

본 논문은 데이터에 포함된 많은 특징을 어떻게 효과적으로 줄일 수 있는가가 주안점이므로 분류를 위해서는 대표적인 한 가지 방법을 사용하였다. 즉 여기에서 사용한 방법은 다층 퍼셉트론이라는 신경망 모델이다. 이 모델은 신호가 전달되는 마지막 노드에서 기대하던 결과와 계산된 결과의 차이인 오류를 신호 전달 방향과 반대 방향으로 역전파하는 방식으로 학습을 하게 된다 [9]. 이 모델은 비교적 다양한 성격의 문제에 적용이 가능한 것으로 알려져 있다.

본 논문에서 사용한 다층 퍼셉트론은 입력 층, 은닉 층, 출력 층으로 구성된다. 여기에서 출력 층은 만

족과 불만족을 분류하기 위해 2 개의 노드로 구성되며, 은닉 층은 노드의 개수를 문제의 복잡도에 따라 조정할 수 있다.

4. 실험 결과

4.1 실험 환경

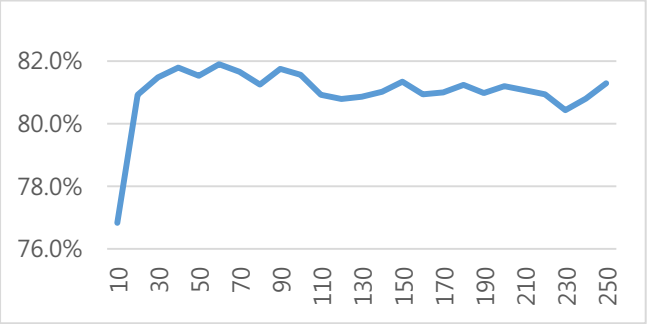
본 논문에서 사용한 실험 데이터는 캐글 사이트에서 얻은 쌘텐더 은행의 불만족 고객 예측에 관한 훈련 데이터이다. 이 데이터는 369 개의 특징과 76019 개의 샘플로 구성되어있으며, 클래스는 0 과 1 로 표기되어 있어 값이 1 인 경우가 불만족 고객의 샘플이다. 본 논문에서는 실험의 편의를 위해 클래스 간의 균형을 맞추어 6015 개의 샘플을 따로 선택하여 사용하였다.

예측 시스템의 성능 평가를 위해 본 논문에서는 주어진 데이터를 5 개의 세트로 나누어 교차 검증을 실시하였다. 그리고 성능평가 지표로 CA (Classification Accuracy)와 AUC (Area Under Curve)를 생각할 수 있는데, CA 에 비해 AUC 가 보다 종합적으로 성능을 판단할 수 있는 지표이므로 실험에서 이를 사용하였다.

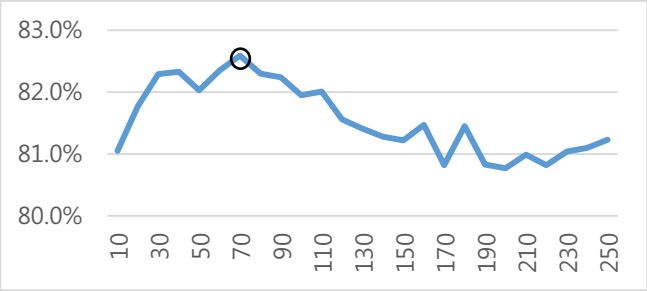
또한 특징 벡터의 분류를 위해서는 3.4 절에서 설명한 다층 퍼셉트론을 사용하였다. 이 모델에서 은닉층은 4 개의 노드로 구성하고, 훈련 횟수는 100 회로 설정하였다.

4.2 특징 선택

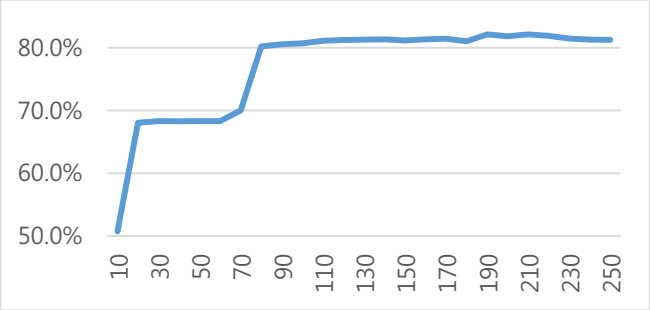
특징 선택을 위해서 ReliefF, 정보 획득량, 정보 획득 비율, Gini 불순도 등의 지표를 실험한 결과는 다음과 같다. 아래 그래프들은 각 방법을 적용하여 모든 특징들을 평가하고 점수에 따라 정렬한 다음 그래프의 수평 축에 표시된 숫자만큼 앞에서부터 차례로 특징들을 선택했을 때 예측 성능을 측정한 결과이다. 여기에서 수직 축은 AUC 값이다.



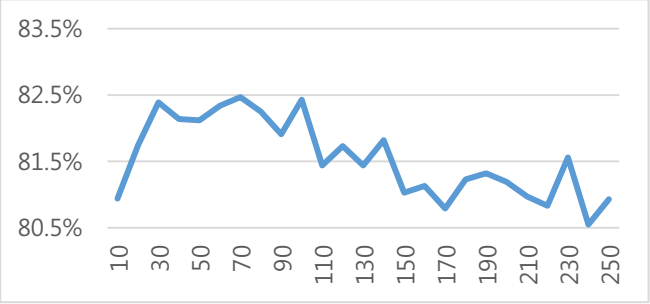
(그림 3) ReliefF 지표를 이용한 특징 선택 결과



(그림 4) 정보 획득량 지표를 이용한 특징 선택 결과



(그림 5) 정보 획득 비율을 이용한 특징 선택 결과

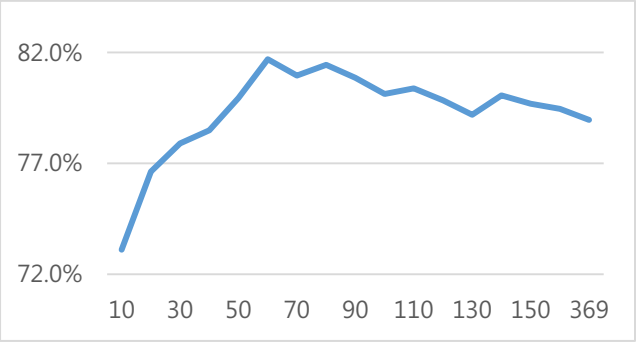


(그림 6) Gini 불순도 지표를 이용한 특징 선택 결과

그래프를 보면 모든 지표가 처음에 특징의 수를 늘리면서 성능이 급격히 상승하다가 어느 수준이 되면 성능이 천천히 떨어지는 결과를 보인다. 따라서 성능이 최대가 되는 지점을 찾아야 하는데, 가장 좋은 결과를 보인 지점은 정보 획득량 지표를 사용해서 70 개의 특징을 선택했을 경우이며, 이 때의 AUC 는 82.6%이다.

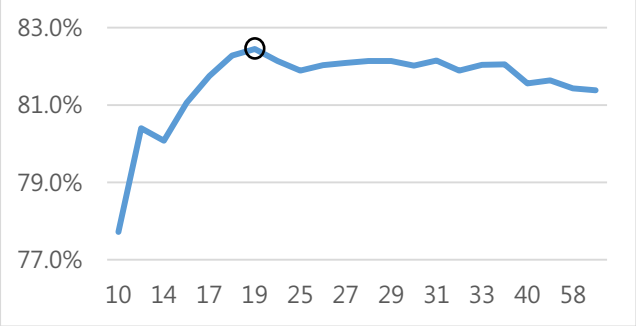
4.3 PCA 에 의한 중복성 제거

별도로 특징 선택을 하지 않고 PCA 만 적용하면 (그림 7)과 같은 결과를 얻을 수 있다. 실제로 PCA 에 의해 변환된 데이터를 살펴보면 59 개의 특징만이 0 이 아닌 값을 가진다. 따라서 PCA 에 의해서 369 개의 특징을 59 개로 줄이는 결과를 얻게 된다.



(그림 7) PCA 를 적용한 결과

다음으로 4.2 절에서 얻은 결과와 같이 정보 획득량 지표를 사용해 70 개의 특징을 선택하고, 여기에 PCA 를 적용해 보면 (그림 8)과 같은 결과를 얻을 수 있다.



(그림 8) 정보 획득량 지표와 PCA 를 결합한 결과

이 결과를 보면 20 개의 특징을 사용해 다층 퍼셉트론을 학습시켰을 때 AUC 가 82.5%로 나타난다. 이는 정보 획득량 지표 만을 사용해 70 개의 특징을 선택했을 경우와 비교해 AUC 가 0.1% 하락하지만 특징의 개수를 71%나 줄이는 결과이다.

5. 결론

본 논문에서는 불만족한 고객을 식별하는 문제에 기계학습 방법을 적용하였다. 이 문제에서 주어진 데이터는 369 개 라는 많은 특징을 포함하고 있어서 이 중 유용한 특징들을 골라내고, 중복성을 제거하기 위한 방법이 이 문제를 해결하는데 있어서 핵심적인 부분이다. 따라서 본 논문에서는 실제로 사용되는 몇 가지 특징 선택 방법과 중복성을 해결하는데 많이 사용되는 PCA 등의 방법을 실험하여 최적의 조합을 찾고자 하였다.

실험 결과 특징 선택 방법 중에서는 정보 획득량 지표가 가장 좋은 결과를 보였다. 또한 특징 선택 방법 만으로 해결하기에는 한계가 있는 중복성의 문제를 PCA 를 사용해 해결할 수 있음을 보였다. 이 두 가지 방법을 결합했을 때 특징을 20 개로 줄이고, AUC 가 82.5%인 결과를 얻었다.

본 연구에서는 앞으로 전체 데이터를 대상으로 실험을 할 예정이며, 또한 분류 방법에 대해서도 다양한 방법들과 이들의 조합에 대해서도 실험해 보고자 한다.

참고문헌

[1] Kaggle contest: Dataset for Santander Customer Satisfaction, https://www.kaggle.com/c/santander-customer-satisfaction/data, 2016.

[2] I. Kononenko, "Estimating attributes: Analysis and extensions of RELIEF," Proc. European Conf. on Machine Learning (ECML-94), pp.171-182. Springer-Verlag, 1994.

[3] Decision tree learning, https://en.wikipedia.org/wiki/Decision_tree_learning, March 2016.

[4] J R Quinlan, "Induction of Decision Trees," Machine Learning, vol.1, pp.81-106, 1986.

[5] Xiubo Geng, Tie-Yan Liu, Tao Qin, Hang Li, "Feature Selection for Ranking," SIGIR'07, Month 1-2, 2007.

[6] Jolliffe I.T., "Principal Component Analysis," Series: Springer Series in Statistics, 2nd ed., Springer, NY, 2002, XXIX, 487 p. 28 illus. ISBN 978-0-387-95442-4.

[7] Domingos, Pedro; Pazzani, Michael "On the optimality of the simple Bayesian classifier under zero-one loss," Journal of Machine Learning, Vol. 29, pp.103-137, 1997.

[8] Cortes, C.; Vapnik, V., "Support-vector networks," Journal of Machine Learning, vol.20, no.3, pp.273, 1995.

[9] Bernard Widrow, Lehr, M.A., "30 years of Adaptive Neural Networks: Perceptron, Madaline, and Backpropagation," Proc. IEEE, vol.78, no.9, pp.1415-1442, 1990.