닫힌 빈발 패턴을 기반으로 한 특징 선택과 분류방법 비교

장뢰, 김성호, 류근호
충북대학교 컴퓨터과학과
e-mail : {zhanglei, kimsungho, khryu}@dblab.chungbuk.ac.kr

A Comparative Study on Feature Selection and Classification Methods Using Closed Frequent Patterns Mining

Lei Zhang, Cheng Hao Jin, Keun Ho Ryu
Dept. of Computer Science, Chungbuk National University South Korea

요 약

분류 기법은 데이터 마이닝 기술 중 가장 잘 알려진 방법으로서, Decision tree, SVM(Support Vector Machine), ANN(Artificial Neural Network) 등 기법을 포함한다. 분류 기법은 이미 알려진 상호 배반적인 몇 개 그룹에 속하는 다변량 관측치로부터 각각의 그룹이 어떤 특징을 가지고 있는지 분류 모델을 만들고, 소속 그룹이 알려지지 않은 새로운 관측치가 어떤 그룹에 분류될 것인가를 결정하는 분석 방법이다. 분류기법을 수행할 때에 기본적으로 특징 공간이 잘 표현되어 있다고 가정한다. 그러나 실제 응용에서는 단일 특징으로 구성된 특징공간이 분명하지 않기 때문에 분류를 잘 수행하지 못하는 문제점이 있다. 본 논문에서는 이 문제에 대한 해결방안으로써 많은 정보를 포함하면서 빈발패턴에 대한 정보의 순실이 없는 닫힌 빈발패턴 기반 분류에 대한 연구를 진행하였다. 본 실험에서는 χ^2(Chi-square)과 정보이득(Information Gain) 속성 선택 척도를 사용하여 의미있는 특징 선택을 수행하였다. 그 결과, 이 연구에서 제시한 척도를 사용하여 특징 선택을 수행한 경우, C4.5, SVM 과 같은 분류기법보다 더 향상된 분류 성능을 보였다.

1. 서론

분류 기법은 기계 학습, 통계 및 데이터 마이닝 중에서도 널리 연구되는 핵심 테마중의 하나로서 C4.5, C5.0, SVM(Support Vector Machine), K-NN(K-Nearest Neighbor)등 분류 기법들이 있다. 그러나 일부 분류에서는 특징공간이 선명하지 않기 때문에 분류를 잘 진행하지 못한다. 그리하여 연관 규칙을 적용하여 빈발패턴을 특징으로 새로운 특징공간으로 만들어 특징공간이 선명하지 못한 문제를 해결하였다. 최근에 빈발 패턴 기반한 분류 알고리즘[1], [2], [3], [4], [5], [6]이 많이 개발되었다. 이 분류기들은 높은 지지도와 높은 신뢰도에 기초하여 특징을 선택한다.

빈발 패턴은 데이터 집합에서 사용자가 임의적으로 지정하는 최소지지도(minsup)임계값을 만족하는 모든 항목집합이다. 빈발패턴은 단일 속성보다 정보량이 더 많기 때문에 빈발패턴으로 구성하는 새로운 특징공간에 대해 분류하게 되면 분류 정확도를 많이 향상시킬 수 있다.

그러나 빈발 패턴 마이닝 과정 중에 생성되는 빈발 항목 집합은 양이 아주 많다. 특히, 최소 지지도(minsup)를 낮은 임계 값으로 설정할 경우이다. 그러나 빈발 패턴 마이닝 과정 중에서 생성된 많은 빈발 집합에는 보통 일정한 양의 중복 빈발 항목집합이 존재하기 때문에 특징으로 사용하면 분류성능에 영향을 많이 미친다.

닫힌 패턴은 패턴의 지지도 정보의 손실 없이 그것들의 최소의 표현을 규정한다. 그리하여 닫힌 빈발 패턴은 마이닝 과정중에서 생성된 패턴의 정보를 완전하게 유지하면서 수량을 대폭 감소 할 수 있고, 빈발 패턴에서 중복된 빈발 패턴을 제거한다. 닫힌 패턴의 이러한 성질 때문에 좋은 특징이 될 수 있다. 본 논문에서는 빈발 패턴 마이닝 과정을 통해 생성된 유용한 패턴의 정보를 완전하게 유지할 수 있는 닫힌 빈발 패턴 마이닝과 특징 선택을 적용한 실험을 통해 일부 데이터 집합에서 높은 정확도를 보였다.

본 논문의 구성은 다음과 같다. 2 장에서는 본연구에 관한 문제를 정의하며 실험 설계한다. 3 장에서는 닫힌 빈발 패턴 마이닝 기법을 이용하여 UCI 데이터에 대한 실험하며 그 실험 결과에 대해 분석한다. 마지막장에서는 결론 및 향후 연구 방향을 제시한다.

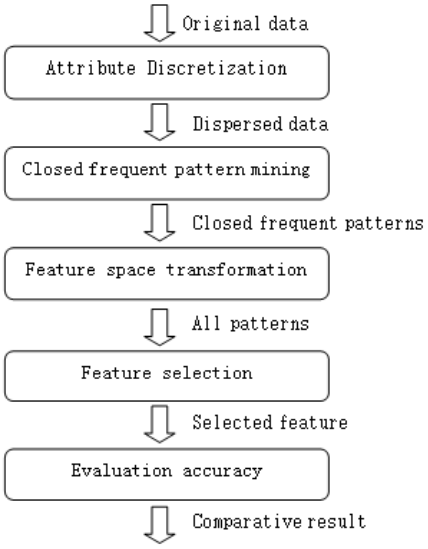
[1] 이 논문은 2010 년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2010-0001732)와 2009 년 교육과학기술부로부터 지원받아 수행된 연구임"(지역거점연구단육성사업 / 충북 BIT 연구중심대학 사업단)

2. 문제 정의 및 실험 설계

I = {i_1, i_2, ..., i_m}은 항목 집합이 되고 P = {T_1, T_2, ..., T_n}은 트랜잭션 데이터베이스가 된다고 가정한다. 각각 트랜잭션 집합 T_q($T_q \in P$)은 I 의 하위 집합의 하나이며 $T_q \subseteq I$ 이다. 만약 항목집합 X 의 포함 집합들이 어느 것도 X 와 정확히 같은 지지도 카운트를 가지지 않은 경우 항목집합 X 는 닫혀있다고 정의한다. 즉 T_q 의 하위집합 중에 T_q 과 같은 지지도 가지고 있는 하위집합이 없다면 T_q 이 닫힌 항목집합이라고 한다. 만약 항목집합이 닫혀있고, 그 지지도가 minsup 보다 크거나 같으면 그 항목집합은 닫힌 빈발 항목집합(closed frequent itemset)이다[7]. CLOSET[8], CHARM[9]등 우수한 닫힌 빈발 패턴 마이닝 알고리즘이 개발되었다.

데이터 집합에서 유용한 특징들을 얻기 위해서 특징선택 기법을 이용한다. 본 논문에서는 특징선택 기법으로 filtering 에 기반한 Chi-Square 및 Information gain 을 선택하여 분류를 진행하였다.

본 연구에서 진행되는 실험단계는 다음과 같다:



(그림 1) 본 논문의 방법의 각 단계

[step 1] 속성 이산화: 속성 이산화는 연속형 속성들을 처리하는 데 가장 평범한 접근법이다. 이 접근법은 연속형 속성의 인접한 값들을 한정된 수의 구간그룹들로 만든다. 원본 데이터를 감소시키고 간략화할 수 있다. 본 논문에서 이산화는 닫힌 빈발 패턴을 기반으로한 특징선택의 전처리과정으로 사용된다. 여기서 모든 특징이 같은 유형을 요구한다.

[step 2] 닫힌 빈발 패턴 마이닝: 속성들이 이상화하기를 통해서 모두 이산 데이터가 되었다. 그 이산 데이터를 대상으로 닫힌 빈발 마이닝한다. 닫힌 빈발 항목 집합은 빈발 패턴 마이닝 과정중에서 생성된 패턴의 정보를 완전하게 유지하면서 빈발 항목 집합 수량을 대폭 감소 할 수 있다.

[step 3] 특징 공간 변환: 더 명확한 특징공간을 얻기 위해서 닫힌 빈발 패턴과 속성의 조합에서 중복된 속성들을 제거한다.

[step 4] 특징 선택 (Chi-Square 및 Information gain): 평가 척도를 사용해서 모든 패턴의 순서를 배열한다. 그 다음에 순위가 높은 패턴 일부를 선택한다.

[step 5] 분류 정확도 평가: 선택된 패턴이 분류에 대해서 효과가 좋은지 나쁜지를 C4.5 및 SVM 를 사용해서 나오는 정확도를 통해 평가한다.

3. 실험 결과 분석

본 연구의 분류성능을 테스트하기 위해서 UCI 데이터베이스 중 10 개 데이터를 이용한다[10]. 그 10 개 데이터에 대한 정보는 표 1 에서 표현한다. 뉴질랜드 Waikato 대학에서 개발된 WEKA 에서 속성 이산화를 진행하였으며 Chi-square 과 Information Gain 을 이용하여 C4.5 와 SVM 분류 성능을 비교했다. 표 2 와 표 3 에서 나오는 Original 분류정확도는 원본 데이터 원래 있는 단일 특징에서 분류하는 정확도이고 표 4 와 표 5 에서 나오는 Closed frequent pattern 분류정확도는 단일 특징과 닫힌빈발패턴의 조합(중복패턴 저거)에서 분류하는 정확도이다. 표 2 와 표 3 는 원본 데이터 및 Chi-square 과 Information Gain 특징선택 기법을 사용해서 C4.5 와 SVM 분류기를 통한 분류 정확도 비교이고, 표 4 와 표 5 는 닫힌 빈발 패턴 및 Chi-square 과 Information Gain 특징선택 기법을 사용해서 C4.5 와 SVM 분류기를 통한 분류 정확도 비교이다. 비교 결과, 표 4 와 표 5 는 닫힌 빈발 패턴의 모든 인스턴스에 대하여 5%의 최소 지지도를 사용했다.

<표 1> 실험에 사용한 데이터들에 대한 요약

데이터	특징 수	샘플 수	클래스 수
hayes-roth	4	132	3
tic-tac-toe	9	958	2
balloons	4	76	2
auto-mpg	7	398	3
breast-cancer	9	286	2
diabetes	8	768	2
heart-h	13	294	2
labor	16	57	2
solar-flare	10	333	8
heart-statlog	13	270	2

본 실험에서 특징선택은 매우 중요한 단계이다. 선택된 하위집합의 분류 정확도는 특징선택 성능을 판단하는 근거중에 하나이다. 선택된 하위집합에 얼마나 좋은 특징을 포함하는지를 고려해야된다. 보다높은 분류 정확도를 유지하면서 특징이 적을 수록 좋은 알고리즘으로 판단된다.본 연구에서는 전체 새로운 특징공간에서의 상위 10% 특징을 선택하여 분류를 진행한다.

<표 2> 원본 데이터 및 Chi-square 과 Information Gain 특징선택 기법을 사용해서 C4.5 분류기를 통한 분류 정확도 비교

Data	Original	Chi-square	Information Gain
hayes-roth	72.73 %	54.55 %	54.55 %
tic-tac-toe	85.07 %	69.94 %	69.94 %
balloons	67.11 %	69.74 %	69.74 %
auto-mpg	80.65 %	74.37 %	74.37 %
breast-cancer	75.52 %	75.17 %	75.17 %
diabetes	73.83 %	74.74 %	74.74 %
heart-h	77.89 %	80.27 %	80.27 %
labor	80.70 %	91.23 %	91.23 %
solar-flare	79.88 %	86.79 %	86.79 %
heart-statlog	76.67 %	76.30 %	76.30 %
average	77.01 %	75.31 %	75.31 %

<표 3> 원본 데이터 및 Chi-square 과 Information Gain 특징선택 기법을 사용해서 SVM 분류기를 통한 분류 정확도 비교

Data	Original	Chi-square	Information Gain
hayes-roth	81.06 %	54.55 %	54.55 %
tic-tac-toe	98.33 %	69.94 %	69.94 %
balloons	72.37 %	69.74 %	69.74 %
auto-mpg	68.84 %	74.87 %	74.87 %
breast-cancer	69.58 %	67.13 %	67.13 %
diabetes	77.34 %	74.74 %	74.74 %
heart-h	82.65 %	81.29 %	81.29 %
labor	91.23 %	91.23 %	91.23 %
solar-flare	80.78 %	86.79 %	86.79 %
heart-statlog	84.07 %	76.30 %	76.30 %
average	80.63 %	74.66 %	74.66 %

표 2 와표 3 에서 나오는 결과를 보면 원본 데이터에서 Chi-square 과 Information Gain 특징 선택을 사용해서 C4.5, SVM 의 분류 정확도가 solar-flare 등 일부 데이터에서 좋지만 대부분은 원본 데이터에서 직접 분류하기보다 더 나쁘다. 특히 hayes-roth, tic-tac-toe 등 같은 데이터에서 분류 정확도가 많이 떨어졌다. 평균 분류정확도도 원본 데이터에서 직접 분류하기에 지나지 않다.

각 데이터 세트, 각 알고리즘의 성능에 따라 표 4 과 표 5 의 분석 결과가 다르게 나타났다. 각 알고리즘은 일부 데이터집합에서 가장 높은 분류 정확도를 얻지만 모든 데이터집합에서 뛰어난 성능을 얻지 못했다. 예를 들어 표 5 에서 보면 Information Gain 알고리즘은 solar-flare 데이터집합에서 가장 높은 정확도를 얻었지만 다른 데이터집합에서는 일반 성능을 얻었다.

<표 4> 닫힌 빈발 패턴 및 Chi-square 과 Information Gain 특징선택 기법을 사용해서 C4.5 분류기를 통한 분류 정확도 비교

Data	Closed frequent pattern	Chi-square	Information Gain
hayes-roth	80.30 %	84.09 %	84.09 %
tic-tac-toe	98.12 %	100 %	100 %
balloons	72.37 %	81.58 %	71.05 %
auto-mpg	78.89 %	76.88 %	76.38 %
breast-cancer	73.08 %	76.57 %	76.57 %
diabetes	77.34 %	77.60 %	77.86 %
heart-h	77.21 %	80.61 %	79.93 %
labor	82.46 %	85.96 %	85.96 %
solar-flare	86.79 %	86.79 %	86.79 %
heart-statlog	82.96 %	84.07 %	81.85 %
average	80.95 %	83.42 %	82.05 %

<표 5> 닫힌 빈발 패턴 및 Chi-square 과 Information Gain 특징선택 기법을 사용해서 SVM 분류기를 통한 분류 정확도 비교

Data	Closed frequent pattern	Chi-square	Information Gain
hayes-roth	81.82 %	84.09 %	84.09 %
tic-tac-toe	99.90 %	100 %	100 %
balloons	78.95 %	78.95 %	71.05 %
auto-mpg	78.89 %	78.14 %	76.88 %
breast-cancer	66.08 %	70.63 %	72.38 %
diabetes	77.08 %	77.08 %	76.95 %
heart-h	81.29 %	80.27 %	79.25 %
labor	92.98 %	87.72 %	89.47 %
solar-flare	82.28 %	85.59 %	86.49 %
heart-statlog	80.37 %	82.22 %	83.33 %
average	81.96 %	82.47 %	81.99 %

각 알고리즘은 항상 장점과 단점을 가지고 있으며 다른 알고리즘보다 완전히 뛰어난 방법은 없다. 그리고 각각의 알고리즘에 의해 선택된 하위 집합은 다른 분류에 따라서 다른 성능을 가지고 있다. Information Gain 알고리즘에 의해 선택된 하위 집합은 SVM 분류기에 대해서는 일반 성능을 가지고 있지만 C4.5 분류기에 대해서는 뛰어난 성능을 가지고 있다. 비록 auto-mpg, diabetes 같은 일부 데이터집합에서 좋지 않은 성능 얻을 수도 있지만 전체적으로 다부분의 데이터집합에서 가장 높은 평균 정확도를 얻었다. 그리고 C4.5, SVM 에 대해 모두 가장 높은 평균 분류 정확도를 얻었다. 특히, hayes-roth, labor 같은 대이터집합에서는 성능 향상이 훨씬 확실하다.

위의 분석은 Chi-square 과 Information Gain 두 개의 특징 선택 결과를 결합했으며 두 개의 분류 알

고리즘을 이용, 그 원본 데이터를 기반으로 한 분류하는 결과 및 닫힌 빈발 패턴을 분류하는 결과의 비교를 보여준다. 이 선택된 하위 집합은 높은 평균 분류 정확도를 얻기 때문에 그 방법이 안정적이고 효율적이다.

4. 결론 및 향후 연구 방향

최근까지 특징 선택에 대해 연구가 활발히 진행되고 있지만, 그에 반해 현존하는 문제점도 많이 발견되어진다. 닫힌 빈발 항목집합은 빈발 항목 집합 마이닝과정에 생성된 패턴의 수를 크게 감소시킬 수 있다. 그리고 완전한 빈발 항목집합을 유지한다. 빈발 패턴을 생성하는 과정에서 중복된 빈발 항목 집합들을 제거한다. 동일한 분류기 사용을 기반으로 닫힌 빈발 패턴을 이용하는 것은 원본 데이터보다 더욱 향상된 분류 정확도를 획득한다. 본 논문에서 특징 선택 방법 2 가지를 실험을 통하여 비교해 보았다. 대부분의 경우 특징 선택을 통해서 패턴의 수를 줄리고 정확도를 향상 시킨 것을 확인할 수 있었다. 그리고 닫힌 빈발 패턴을 기반한 특징 선택을 이용하는 것은 훨씬 더 높은 분류 정확도를 얻었다. 이 방법을 통하여 높은 정확도를 얻을 수 있으므로, 안정적이고 효율적인 기법이 된다. 향후에는 분류 정확도를 향상시키기 위해 특징 선택에 대한 세밀한 연구가 이루어져야 한다.

참고문헌

[1] B. Liu, W. Hsu, and Y. Ma, “Integrating classification and association rule mining,” in Proc. of KDD, 1998, pp. 80–86.

[2] W. Li, J. Han, and J. Pei, “CMAR: Accurate and efficient classification based on multiple class-association rules,” in Proc. of ICDM, 2001, pp. 369–376.

[3] X. Yin and J. Han, “CPAR: Classification based on predictive association rules,” in Proc. of SDM, 2003, pp. 331–335.

[4] G. Cong, K. Tan, A. Tung, and X. Xu, “Mining top-k covering rule groups for gene expression data,” in Proc. of SIGMOD, 2005, pp. 670–681.

[5] J. Wang and G. Karypis, “HARMONY: Efficiently mining the best rules for classification,” in Proc. of SDM, 2005, pp.205–216.

[6] A. Veloso, W. M. Jr., and M. Zaki, “Lazy associative classification,” in Proc. of ICDM, 2006, pp. 645–654.

[7] Pangning Tan. Introduction To Data Mining. 355-356, 2006.

[8] J. Pei, J. Han, and R. Mao. CLOSET: An efficient algorithm for mining frequent closed itemsets. In DMKD’2000.

[9] M. J. Zaki and C. J. Hsiao. CHARM: An efficient algorithm for closed itemset mining. In SDM’2002.

[10] UCI Machine Learning Repository, http://www.ics.uci.edu/~mlearn/MLRepository.html