지능정보연구 제14권 제4호 2008년 12월(pp.179~200)

단변량 분석과 LVF 알고리즘을 결합한 하이브리드 속성선정 방법

이재식
아주대학교 경영대학
(leejsk@ajou.ac.kr)

정미경
(주) ING 생명
(pachen@paran.com)

본 연구에서는 사례기반 추론 기법을 대상으로 효율성과 효과성을 함께 증진시킬 수 있는 속성선정 방법을 개발하였다. 기본적으로, 본 연구에서 개발한 속성선정 방법은 기존에 개발된 단변량 분석 방법과 LVF 알고리즘을 통합하는 것이다. 먼저, 단변량 분석 방법 중 선택효과를 사용하여 전체 속성 중에서 예측력이 우수하다고 판단되는 일부분의 속성들을 추려낸다. 이 속성들로부터 생성해낼 수 있는 모든 가능한 부분집합을 생성해낸 후에, LVF 알고리즘을 이용하여 이 부분집합들이 가지는 불일치 비율을 평가함으로써 최종적으로 속성 부분집합을 선정한다. 본 연구에서 개발한 속성선정 방법을 UCI에서 제공하는 데이터 집합들에 적용하여 성능을 측정한 후, 기존 기법의 성능들과 비교한 결과, 본 연구에서 개발된 속성선정 방법이 선정된 속성의 개수도 만족할만하고 적중률도 향상되어서, 효율성과 효과성 모두의 측면에서 우수함을 보였다.

논문접수일 : 2008년 11월 논문수정일 : 2008년 12월 게재확정일 : 2008년 12월 교신저자 : 이재식

1. 서론

기계학습 알고리즘은 학습방식에 따라 사전학습(Eager Learning) 알고리즘과 사후학습(Lazy Learning) 알고리즘으로 구분될 수 있다. 사전학습 알고리즘은 인공신경망이나 의사결정나무 기법과 같이 질의가 입력되기 이전에, 관련 데이터 분석을 통한 학습을 모두 마쳐놓고 질의가 입력되면 곧바로 응답하는 구조를 가지고 있다. 반면에, 사후학습 알고리즘은 질의가 입력된 후에, 이를 해결하기 위한 관련 데이터들을 모으고 이것들을 분석하여 해를 도출해내는 구조를 가지고 있다. 사례기반 추론(CBR : Case-based Reasoning)은 질의가 입력되는 시점에 비로소 사례베이스로부터 유사한 사례들을 검색하여 해를 도출하기 시작하는 전형적인 사후학습 알고리즘이다.

사후학습 알고리즘의 가장 큰 문제점은, 데이터를 구성하고 있는 속성의 개수가 많으면, 해를 도출하는 시간이 오래 걸린다는 것이다. 이 문제점을 해결하기 위하여 속성선정 방법에 대한 연구가 많이 수행되어 왔다. 물론 속성선정 방법을 적용해야 하는 대상이 사후학습 알고리즘에 국한되는 것은 아니지만, 사후학습 알고리즘이 사전학습 알고리즘 보다는 속성선정의 효과를 많이 볼 수 있다. 왜냐하면, 속성의 개수가 적어지면, 사전학습 알고리즘은 모델의 구축과정에서 효과를 보게 되지만, 사후학습 알고리즘은 문제를 해결할 때마다 시간 단축의 효과를 보게 되기 때문이다. 속성선정 방법의

연구와 더불어 속성에 적절한 가중치를 부여하는 연구도 수행되어 왔는데, 속성선정 방법은 속성가중치 부여 방법의 특수한 형태로 간주할 수 있다. 즉 속성선정이라는 것은 속성가중치를 0 또는 1로 부여하는 것과 동일한 것이다.

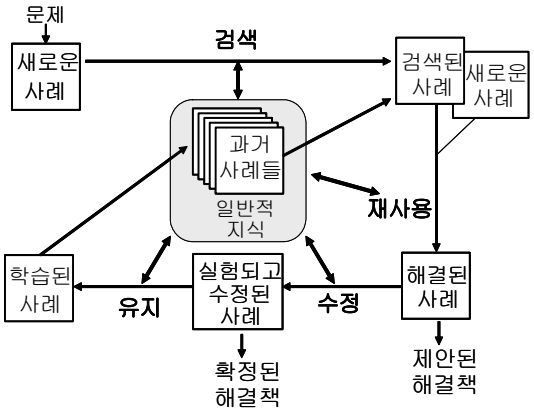
본 연구에서는 기존에 개발된 두 개의 속성선정 알고리즘, 즉 FeaSUA(이재식, 이혁희, 2002)와 LVF (Las Vegas Filter)(Liu and Setiono, 1996)를 결합하여 하이브리드 속성선정 알고리즘을 개발하였다. 우리는 이것을 UV(FeaS**U**A+L**V**F) 알고리즘이라고 명명하였다. 우리는 UV 알고리즘이 효율성과 효과성 측면에서 모두 우수함을 보이고자 하는데, 효율성은 속성개수의 감소를, 효과성은 적중률의 향상을 의미한다. 본 논문은 다음과 같이 구성되어 있다. 제 2장에서는 사례기반 추론, 속성선정 과정, 기존의 속성선정 방법 등의 이론적 배경에 대한 소개를 하고, 제 3장에서는 본 연구에서 개발한 UV 알고리즘의 설계와 구현에 대해서 기술한다. 제 4장에서는 UV 알고리즘의 성능을 UCI Machine Learning Repository의 데이터(Blake et al., 1998)를 사용하여 측정하고 평가한다. 마지막으로, 제 5장에서는 결론과 함께 한계점 및 향후 연구과제에 대해 논의한다.

2. 이론적 배경

2.1 사례기반 추론

사례기반 추론은 이전의 문제 해결 경험을 저장하였다가 새로운 문제를 만나면 가장 유사한 과거 사례를 조회하여 해결책을 제시한다. 사례기반 추론의 기본 순환 과정은 <그림 1>과 같다(Aamodt and Plaza, 1994).

<그림 1>에서 보는 바와 같이 사례기반 추론은 해를 구하고자 하는 새로운 사례가 입력되면, 저장되어 있는 이전 사례들로부터 유사한 사례를 검색하고, 그 유사한 사례를 재사용하고 수정하여 해를 도출한다. 그리고 도출된 해를 다시 저장하여 다음에 입력되는 새로운 사례에 대해서 더욱 풍부해진 과거 사례들로써 대응하도록 한다.



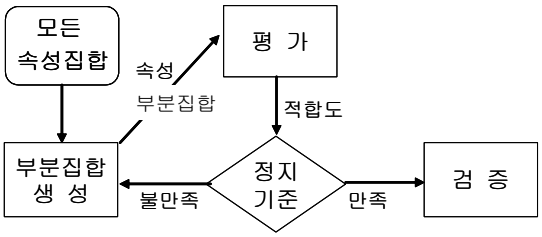
<그림 1> 사례기반 추론의 과정

사례기반 추론은 다양한 현실 문제 해결에 적용되고 있는데, 고객 서비스(Lee and Xon, 1996), 파산예측(Jo. et al., 1997), 고장 진단(Kuo et al., 2005, Varma 1999, Wang and Wang 2005), 헬프 데스크(Goker 1999, Law 1997), 전략 수립(Canchien 2005) 등은 성공적으로 사례기반 추론이 적용되었던 응용 영역이다. 최근에 사례기반 추론은 유비쿼터스 컴퓨팅 시스템에서의 상황인식 기능(Lee and Lee, 2006, 2007) 및 개인화 서비스의 구현(Leake 2006)에도 활용되고 있다.

2.2 속성선정 과정

기존의 많은 연구를 통해서 밝혀진 바와 같이, 사례기반 추론을 포함한 여러 데이터 마이닝 기법

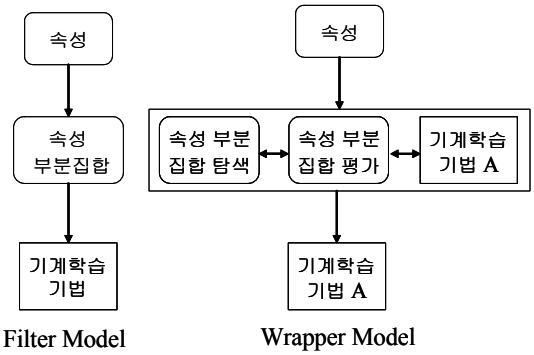
(Data Mining Technique)에서 일부의 속성 부분집합을 사용하는 것이 모든 속성을 이용하여 값을 예측하는 것보다 더 좋은 결과를 보여주는 예가 많았다. 그러한 노력의 일부로서, 많은 연구자들이 속성선정 방법을 연구하였다. 데이터를 구성하고 있는 속성들 중에서 유용한 속성을 선정하는 속성선정은 <그림 2>와 같이 네 단계로 이루어져 있다(Dash and Liu, 1997).



<그림 2> 속성 선정 과정

제 1단계인 '부분집합 생성'에서는 선정될 속성으로 적합한지를 평가하기 위한 속성 후보군을 만든다. 생성 방법에는 완전탐색(Complete Search), 순차탐색(Sequential Search) 그리고 무작위 탐색(Random Search)이 있다(Doak, 1992). 완전탐색은 생성할 수 있는 속성 부분집합의 모든 경우를 고려하는 방법이다. 순차탐색은 일반적으로 전방향 순차 탐색(FSS : Forward Sequential Search)과 역방향 순차 탐색(BSS : Backward Sequential Search)으로 나누어진다. 전방향 순차 탐색은 속성을 하나씩 추가해 가는 것이고, 역방향 순차 탐색은 반대로 속성을 하나씩 제거해 가면서 속성 부분집합을 찾아내는 방식이다. 무작위 탐색은 무작위로 생성된 수치들 또는 유전적 알고리즘(Genetic Algorithms)(Goldberg, 1989) 등을 이용하여 속성 부분집합을 생성해 내는 방법이다.

제 2단계인 '평가'에서는 생성된 속성 부분집합이 평가 기준에 대해서 적합한지를 판정한다. 생성된 속성 부분집합의 평가 방법은 크게 두 가지로 나눌 수 있는데, 그 기준은 속성 부분집합을 사용하여 얻은 분류결과에 대해서 피드백(Feedback)을 받는가의 여부이다. 피드백을 받지 않는 기법을 Filter Model이라 부르고, 기계학습 기법 자신으로부터 피드백을 받는 기법을 Wrapper Model이라고 부른다(John et al., 1994). 이 두 Model을 도식화하면 <그림 3>과 같다.



<그림 3> 속성 부분집합의 두 가지 평가방법

<그림 3>의 왼쪽은 Filter Model인데, 그림에서 보듯이 속성 부분집합을 선정하는 과정이 궁극적으로 사용될 기계학습 기법과는 무관하게 진행된다. 물론 속성 부분집합을 선정하는 방법에 기계학습 기법이 사용될 수도 있지만, 궁극적으로 사용될 기계학습 기법과는 어떠한 피드백도 주고받지 않는다. 반면에, <그림 3>의 오른쪽에 있는 Wrapper Model에서는, 궁극적으로 사용될 기계학습 기법과 속성 부분집합 선정에 사용되는 기계학습 기법은 동일한 것, 즉 "A"라는 동일한 기법이다. 다시 말하면, Wrapper Model에서는 속성선정 과정에서 생성되는 부분집합들을 궁극적으로 사용할 기계학습 기법에 적용하여 그 피드백을 받으면서, 우수

한 속성 부분집합을 선정해 나가는 것이다.

Filter Model의 평가기준으로는 거리, 정보, 의존성 그리고 일치성 등이 있다. 거리 평가기준은, 예를 들어 두 개의 Class가 있는 이진 분류 문제에서 두 속성 F1과 F2가 있다고 할 때에 F1이 F2 보다 두 Class 간의 차별성을 더 부각시키면 F1을 선정하는 것이다. 정보 평가기준은 정보이득(Information Gain)에 기반을 둔 기준으로서, F1의 정보이득이 F2의 정보이득보다 크면 F1을 선정한다. 의존성 평가기준은 속성과 Class 간의 관련성, 즉 상관관계 또는 유사성에 기반을 둔 기준으로서, F1과 Class 간의 관련성이 F2와 Class 간의 관련성보다 크면 F1을 선정한다. 일치성 평가기준은 속성의 값이 동일한 사례들끼리 얼마나 Class의 값들도 일치하는가를 판단하는 것으로서, 자세한 설명은 제 3.2절에서 한다.

Wrapper Model의 평가기준은 사용하는 기법과 그 성능의 평가에 따라 달라진다. 예를 들어 분류문제를 다루는 기계학습 기법에서는 적중률을 평가기준으로 사용할 수 있고, 군집화를 하는 경우에는 각 클러스터를 평가하는 방법들, 즉 클러스터 내의 분산, 클러스터간의 거리 등을 사용할 수 있다.

제 3단계인 '정지기준'은 언제 속성선정 과정을 멈추느냐를 결정하는 것이다. 다양한 기준이 마련될 수 있는데, 예를 들어, 더 이상 속성 부분집합을 생성할 수 없을 때, 미리 설정한 개수의 속성 부분집합을 생성하였을 때 또는 미리 설정한 적중률을 보이는 속성 부분집합을 얻었을 때 등이 있다.

마지막 제 4단계인 '검증'에서는 선정된 속성의 성능을 측정하여 평가하는 것으로서, 일반적으로, 모든 속성을 사용했을 때의 모델의 적중률과 선정된 속성만 사용했을 때의 모델을 적중률을 비교한다.

2.3 기존의 속성선정 방법

본 절에서는, 본 연구에서 개발한 속성선정 방법인 UV 알고리즘에 활용된 FeaSUA 알고리즘(이재식, 이혁희, 2002)과 LVF(Las Vegas Filter) 알고리즘(Liu and Setiono, 1996), 그리고 UV 알고리즘의 성능과 비교할 기법들인 전방향 순차 탐색(Forward Sequential Search) 방법과 역방향 순차 탐색(Backward Sequential Search) 방법(Aha and Bankert, 1994), RC 알고리즘(Domingos, 1997), PSORSFS 알고리즘(Wang et al., 2007)에 대해서 간략하게 설명한다.

FeaSUA 알고리즘(이재식, 이혁희, 2002)은 단변량 분석의 결과를 활용하여 속성을 선정하는 방법이다. 단변량 분석 방법에는 선택효과(Selection Effect)와 제거효과(Exclusion Effect)가 있다. 선택효과는 사례기반추론에 개별 속성 하나씩만을 사용했을 때의 분류적중률로 측정하고, 제거효과는 사례기반추론에 다른 속성을 모두 포함시키되 각 개별 속성 하나씩만을 제거했을 때의 분류적중률로 측정하였다. 선택효과와 제거효과의 순으로 전체 속성 중 상위 30%, 50%, 70% 등의 여러 구성을 속성 부분집합으로 선정하여 분류적중률을 측정한 결과를 제시하였는데, 그 중 전체 속성 중 선택효과 상위 30%의 속성 부분집합의 결과가 비교적 높은 결과를 보여주었다.

LVF 알고리즘(Liu and Setiono, 1996)은 개별 속성에 대해서 선정 여부를 판단하지 않고, 속성 부분집합에 대해서 선정 여부를 판단하는 방법이다. 생성된 속성 부분집합에 대해서, 그 속성들의 값이 동일한 사례들끼리 얼마나 Class의 값들이 불일치하는가를 측정하여 불일치비율라는 것을 계산한다. 그리고 속성 부분집합별 불일치비율이 기준 임계치 보다 작은 경우의 속성 부분집합을

선정하는 것이다. 이 방법은 가능한 모든 경우의 속성 부분집합을 평가하는 방법이 아니며, 범주형 속성을 대상으로 만들어진 속성선정 알고리즘이므로, 그들의 연구에서는 연속형 속성을 포함한 데이터 집합에는 적용하지 않았다.

전방향 순차 탐색(FSS)(Aha and Bankert, 1994)은 모든 속성의 가중치 초기값을 0에서 시작하여 더 이상 개선 될 수 없을 때까지 점진적으로 증가시키는 방법이다. 계속적인 수행을 통해서 속성들 중에는 가중치가 증가되는 속성과 가중치가 증가되지 않는 속성이 생기게 된다. 결과적으로, 속성 가중치가 미리 정한 임계치보다 작은 속성들은 제거하고, 임계치보다 큰 가중치를 가지는 속성들만을 선정하는 것이다.

역방향 순차 탐색(BSS)(Aha and Bankert, 1994)은 FSS와 반대로, 모든 속성의 가중치 초기값을 최대값인 1로 하고 이것을 더 이상 개선 될 수 없을 때까지 점진적으로 감소시키는 방법이다. 그 다음에는 FSS와 마찬가지로 미리 정한 임계치와 속성 가중치를 비교하여, 임계치보다 작은 속성들을 제거하고 임계치보다 큰 속성들만을 선정한다.

RC(Relevance in Context) 알고리즘(Domingos, 1997)은 여러 면에서 BSS와 구동 방식이 비슷하지만 전역적이 아닌 지역적으로, 즉 사례별로 다르게 속성을 선정한다는데 차이가 있다. 어떤 사례 C_i의 속성 F_{ij}의 값이 최근접 이웃 사례의 그 값과 다르고, 또한 그 속성의 제거가 전체적인 분류 오차를 증가시키지 않는다면, F_{ij}를 C_i에서 제거할 수 있다. 모든 사례에 대해서 이와 같은 작업을 수행함으로써 사례별로 상이한 속성 부분집합을 갖게 된다. 사례별로 상이하게 속성들을 제거하다 보면 중복된 사례가 생성될 수 있으나 이를 제거하지는 않는다.

Wang et al.(2007)은 Rough Set과 Particle Swarm Optimization에 기반을 둔 속성선정 방법인 PSO-RSFS 알고리즘을 발표하였다. Rough Set 이란 대상이 되는 집합을 하한 및 상한의 두 개의 집합으로 표현하는 기법인데(Pawlak, 1982), Rough Set 이론에서 속성선정에 사용되는 개념은 Reduct이다. Reduct란 원래 데이터 집합의 특성을 그대

<표 1> 속성선정 방법의 유형

			탐색 전략					
			완 전		순 차 적		무 작 위	
평가기준	Filter	거 리	○		○			
		정 보	○		○	○		
		의존성	○		○	○		
		일치성	○		○		○ LVF	
	Wrapper	분류 적중률 또는 군집의 적합성	○ FeaSUA		○	○	○	
	Hybrid	Filter+Wrapper	**UV**		○	○		
			분류	군집화	분류	군집화	분류	군집화
			Data Mining 작업					

로 나타낼 수 있는 속성의 부분 집합을 말한다. 어떤 속성 부분집합이 Reduct인지를 평가하기 위해서 수많은 Reduct 후보들을 생성해 내야 하는데, 그들은 그 수단으로 Particle Swarm Optimization을 사용하였다. Particle Swarm Optimization은 유전자 알고리즘(GA : Genetic Algorithm)과 흡사한 진화적(Evolutionary) 계산 기법인데, GA의 교배(Crossover)나 변이(Mutation) 같은 복잡한 연산자 없이 단순한 수학적 연산자만으로 작동하는 기법이다.

Liu and Yu(2005)는 기존에 발표된 50여 개의 속성선정 방법들을 그들의 특성들, 즉 평가방법(Filter, Wrapper, Hybrid), 탐색전략(완전, 순차적, 무작위) 그리고 적용되는 Data Mining 작업(분류, 군집화)의 유형들에 따라 분류하여 <표 1>과 같은 결과를 제시하였다.

<표 1>에서 'O'로 표시된 칸은 속성선정 방법이 개발되어 있는 부분이고, 빈 칸은 아직 속성선정 방법이 개발되어 있지 않은 부분이다. FeaSUA는 'Wrapper-완전-분류' 유형이며, LVF는 'Filter-무작위-분류' 유형이다. 본 연구에서 제시하는 UV 알고리즘은 FeaSUA와 LVF를 결합하는 기법으로서, Liu and Yu의 연구에서는 아직까지 연구결과가 없는 부분으로 제시되었던 'Hybrid-완전-분류'의 칸에 새로운 기법을 채워 넣는 것이다.

3. UV 알고리즘의 설계 및 구현

3.1 연구에 사용된 사례기반 추론 시스템

사례기반추론은 그 활용 영역에 따라서 다양한 구조를 가지며, 성능을 높이기 위한 노력 여하에 따라서 매우 많은 변형이 나타날 수 있다. 즉, 속성에 가중치를 부여하는 방법에 따라서, 사례간의 유사도를 계산하는 방법에 따라서, 또한 목표 속성의 값을 산출하는 적응방법에 따라서 사례기반추론의 구조가 달라질 수 있는 것이다. 하지만, 본 연구의 목적이 사례기반 추론의 구조를 변형하여 적중률을 높이고자 하는 것이 아니라, 단순히 속성선정 방법의 개발에 있으므로, 사례기반추론의 구조는 <표 2>과 같이 가장 단순한 형태를 사용하였다.

연속형 속성의 경우에는 단순히 거리를 측정하는 방식을 사용하였으며, 범주형 속성의 경우에는 두 값이 일치하면 1점, 그렇지 않으면 0.5점을 부여하였다. 일치하지 않는 경우에 0.5점을 부여한 이유는 미세한 차이가 있어도 0점을 부여한다면 이로 인해서 유사도 점수 계산에 큰 격차를 나타낼 수 있기 때문이다. 유사도 계산 후, 유사한 사례

<표 2> 본 연구에서 사용된 사례기반추론의 특성

방 법	사 용 방 식
속성 가중치 부여	모두 1
유사도 계산	연속형 속성 : 유사도 = $1-\frac{\lvert\text{새로운 사례의 속성값} - \text{과거 사례의 속성값}\rvert}{\text{해당 속성의 최대값}}$ 범주형 속성 : 유사도 = 1, (새로운 사례의 속성값 = 과거 사례의 속성값) = 0.5, (나머지 경우)
검색(Retrieval)	k개의 최근접이웃 방법(k-NN), k = 5
적응(Adaptation)	Voting 방법 (5개중 3개 이상인 경우)

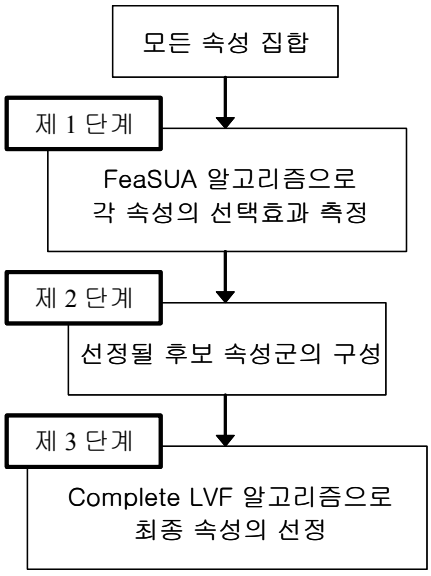
를 검색하는 단계에서는 k개의 최근접이웃(k-NN; k Nearest Neighbors) 방법을 사용하였으며, k값으로 5를 사용하였다. 해를 도출하는 적응(Adaptation)단계에서는 검색된 사례의 해답 중에서 3개 이상이 새로운 사례의 해답과 일치하면 적중한 것으로 정의하는 투표(Voting) 방법을 사용하였다.

본 연구에서 사용된 데이터 집합은 분류 문제 영역으로서 목표 속성은 두 개의 값을 가지는 Two-Class 문제이다. 각 데이터 집합은 학습용 데이터 집합(Training Data Set), 테스트용 데이터 집합(Test Data Set) 그리고 평가용 데이터 집합(Evaluation Data Set)의 세 개의 데이터 집합으로 분류하여 사용된다. 학습용 데이터 집합은 사례베이스로 사용되고, 테스트용 데이터 집합은 입력사례로, 평가용 데이터 집합은 최종적으로 성능을 평가하기 위한 목적으로 사용된다. 데이터 분류의 비율은 학습용, 테스트용, 평가용을 6 : 3 : 1로 하였는데, 데이터 집합을 구성할 때의 Sampling에 따른 영향을 줄이고 결과의 타당성을 높이기 위해서 10-Fold Cross Validation 방법에 따라 데이터 집합 구성을 10번 하고 실험을 하여 결과를 제시한다.

3.2 연구 방법

UV 알고리즘의 수행 과정은 <그림 4>와 같다.

- **제 1단계** : FeaSUA 알고리즘(이재식, 이혁희, 2002)은 개별 속성의 적중률에 대한 영향력을 측정하는 단변량 분석을 통하여 속성선정을 하는 방법이다. 단변량 분석에는 두 가지 종류가 있는데, 첫 째는 단 하나의 속성만을 사용하였을 때의 적중률을 측정하는 선택효과이고, 둘째는 전체 속성에서 단 하나의 속성만을 뺐을 때의 적중률의 하락을 측정하는 제거효과이다. 본 연구에서는 이 중에서 선택효과만을 측정하여 선정 후보 속성군을 구성하는데 사용한다. 각 속성의 선택효과를 측정할 때에도 10-Fold 방식으로 10번 측정하여 평균을 사용한다.



<그림 4> UV 알고리즘의 과정

- **제 2단계** : 선택효과의 측정치가 높은 순서대로 속성을 추가시키면서 사례기반추론 시스템의 분류 적중률을 구하는데, 이 적중률이 더 이상 증가하지 않을 때까지 계속해서 속성을 추가시킨다. 적중률이 더 이상 증가하지 않으면, 그 때까지 추가된 속성들이 속성 선정 후보군이 된다. 이러한 과정을 10개의 상이한 데이터 집합에 대하여 수행하여, 적중률이 가장 높게 나온 실험에 사용된 속성들을 속성 후보군으로 선택한다. 만일 적중률이 가장 높은 실험이 두개 이상이 나오면, 사용된 속성의 개수가 가장 많은 것을 속성 후보군으로 선택한다.

- **제 3단계** : 제 2단계에서 얻어진 속성 후보군을 대상으로 LVF 알고리즘을 적용하여, 부분집합별 불일치비율(Inconsistency Rate)을 계산하고, 불일치비율이 가장 작은 속성 부분집합을 최종적으로 선정된 속성 부분집합으로 결정한다.

이 연구에서 적용한 LVF 알고리즘은 Liu and Setiono(1996)가 개발한 LVF 알고리즘을 변형한 것이다. 기존의 LVF 알고리즘은 <표 1>에 표시되었듯이 무작위로 생성되는 속성 부분집합에 대하여 평가를 하는 것이다. 즉, 이 알고리즘은 모든 부분집합을 고려하지 않기 때문에 최적의 속성 부분집합을 도출한다는 보장이 없다. 하지만, 본 연구에서 적용한 LVF 알고리즘은 가능한 모든 부분집합에 대한 평가를 한다. 그러므로 우리는 이것을 Complete-LVF로 명명하였다. Complete-LVF 알고리즘은 제 2단계에서 얻어진 K개의 속성으로부터 만들어지는 2^K-1개의 모든 부분집합에 대해서 기존의 LVF 알고리즘과 같은 방법으로 불일치비율을 계산한다. 그 다음, 불일치비율이 가장 작은 부분집합을 최종적으로 선정된 속성 부분집합으로 결정한다. 만일, 불일치비율이 가장 작은 부분집합이 두 개 이상인 경우에는 속성의 개수가 적은 집합을 선택한다. Complete-LVF 알고리즘의 과정은 <그림 5>와 같다.

본 연구에서는 사용하는 Complete-LVF 알고리즘이 기존의 LVF 알고리즘과는 다른 점은 다음과 같다. 첫째, 속성 부분집합을 생성할 때에 무작위 방법을 사용하지 않고 가능한 모든 부분집합을 생성하는 것이다. 그러므로 Complete-LVF 알고리즘은 속성의 개수가 증가하면, 계산의 시간이 기하급수적으로 증가할 수 있다. 그러므로 제 1단계에서 FeaSUA 알고리즘을 통하여 적중률 향상에 효과가 큰 속성들만을 발췌함으로써, 고려할 속성

```
Input
    D : Data Set
    A : Set of Attributes: |A| = K
Output
    A_best :Selected Subset of Attributes
Algorithm
    FOR i=1 to 2^K-1
        A_i = Subset of A
        N_i = Number of Attributes in A_i
        δ_i = Inconsistency Rate of A_i
    END FOR
    A_best = A_1
    N_best = N_1
    δ_best = δ_1
    FOR i=2 to 2^K-1
        IF δ_i < δ_best THEN A_best = A_i
        ELSE IF δ_i = δ_best THEN
            IF N_i < N_best THEN A_best = A_i
        END IF
    END FOR
END Algorithm
```

<그림 5> Complete-LVF 알고리즘

의 개수를 줄이는 것이다. 둘째, 기존의 LVF 알고리즘은 속성이 범주형인 경우에만 적용 가능하였지만, Complete-LVF 알고리즘에서는 속성이 연속형인 경우에는 Binning을 통하여 범주형으로 변환하여 적용한다. 속성 값의 변환은 사용하는 방법에 따라 적중률에 차이를 가져올 수 있기 때문에 매우 중요한 작업이다. 속성 값을 변환하는 방법에는 여러 가지가 있는데 그 중 데이터 마이닝 작업에서 대표적으로 이용되는 것이 동일개체수기반 변환방법(Equal-height Bin)과 동일구간기반 변환방법(Equal-width Bin)이다(Berry and Linoff, 2004). 동일개체수기반 변환방법은 각각의 Bin에 거의 동일한 개수의 개체를 할당하는 것이고, 동일구간기반 변환방법은 각 Bin에 포함되는 개체의

개수에 상관없이 일정한 구간으로 나눈 Bin에 개체들을 할당하는 방법이다. 본 연구는 이 중 동일개체수기반 변환방법을 이용한다. 동일개체수기반 변환방법이 실제 프로젝트에서 유용한 것으로 알려져 있고, 빈번하게 사용되는 방법이기 때문이다.

Complete-LVF 알고리즘에서 우리가 구하고자 하는 속성 부분집합의 가장 이상적인 불일치비율은 0이다. 각 속성 부분집합의 불일치비율은 <그림 6>과 같은 방법으로 측정한다(Liu and Setiono, 1996).

$$\text{Inconsistency Rate } \delta = \frac{\text{Inconsistency Count}}{\text{Number of Cases in Patterns}}$$

Inconsistency Count = $\Sigma\,(C_i - M_i)$
C_i = Pattern i에 포함된 Case의 개수, i =1, 2, …., Q
M_i = Pattern i에서 가장 많은 목표 속성 값을 갖는 Case의 개수
Pattern : 목표속성을 제외한 모든 속성의 값이 동일한 Case의 집합
Q : Pattern의 개수
Number of Cases in Patterns = $\Sigma\,C_i$

<그림 6> 불일치비율 계산 방법

간단한 데이터를 가지고 예를 들어서 설명하면, <그림 7>과 같이 F1, F4, F6 세 개의 속성을 갖는 데이터가 있다고 가정하자. 먼저, <그림 7>의 왼쪽의 표를 오른쪽의 표와 같이 동일한 패턴을 가지는 Case들이 한 곳으로 모이도록 정렬시킨다. 그 다음, 각 패턴별로 C_i, M_i, $C_i - M_i$를 계산한다. 그러면, <그림 6>의 계산식에 의해서 불일치비율은(0+1+3)/(4+4+7) = 4/15 = 0.267이 된다.

각 속성 부분집합에 대해서 Complete-LVF 알고리즘으로 불일치비율을 측정하고, 그 값이 가장 작은 속성 부분집합을 최종 속성 부분집합으로 선정한다. 만일 가장 작은 불일치비율이 두 개 이상 나오면, 속성의 개수가 더 적은 것을 선택한다.

4. UV 알고리즘의 성능

4.1 유방암 진단(Breast Cancer)

본 논문에서는 UCI Machine Learning Repos-

F1	F4	F6	목표속성
1	2	8	T
2	7	5	K
1	2	8	T
1	2	5	L
1	2	5	K
1	2	8	T
2	7	5	K
1	2	5	P
2	7	5	L
1	2	5	K
1	2	5	K
1	2	8	T
1	2	5	K
1	2	5	L
2	7	5	K

F1	F4	F6	목표속성	
1	2	8	T	$C_1 = 4$
1	2	8	T	$M_1 = 4$
1	2	8	T	$C_1 - M_1 = 0$
1	2	8	T	
2	7	5	K	$C_2 = 4$
2	7	5	K	$M_2 = 3$
2	7	5	K	$C_2 - M_2 = 1$
2	7	5	L	
1	2	5	K	$C_3 = 7$
1	2	5	K	$M_3 = 4$
1	2	5	K	$C_3 - M_3 = 3$
1	2	5	K	
1	2	5	L	
1	2	5	L	
1	2	5	P	

<그림 7> 불일치비율 계산을 위한 정렬과정

<표 3> 유방암 진단 데이터의 속성 구성

속 성	설 명	유 형
F0	일련번호(Sample Code Number)	사용하지 않음
F1	세균덩어리 두께(Clump Thickness)	범주형 (1-10)
F2	세포 크기의 균일성(Uniformity of Cell Size)	
F3	세포 모양의 균일성(Uniformity of Cell Shape)	
F4	최소 응착력(Marginal Adhesion)	
F5	단일 상피세포 크기(Single Epithelial Cell Size)	
F6	빈 세포핵(Bare Nucleoli)	
F7	차분한 염색질(Bland Chromatin)	
F8	일반 세포핵(Normal Nucleoli)	
F9	유사분열(Mitoses)	
목표 속성	2. 양성(benign); 4. 악성(malignant)	

itory에서 수집한 3개의 데이터 집합을 대상으로 UV 알고리즘의 성능을 측정하였다. 분류 문제의 가장 대표적인 분야의 하나가 진단 문제이다. 유방암 진단 데이터는 유방암의 양성과 악성을 진단하는 문제에 관한 것이다. 데이터의 레코드 개수는 699개인데, 본 연구에서는 그 중 결측치가 있는 레코드 16개를 제외하고 683개를 사용하였다. <표 3>은 유방암 진단 데이터의 속성 값과 유형에 대한 설명이다.

유방암 진단 데이터의 설명 속성은 1부터 10까지의 범주형 숫자 값을 가지고 있다. 속성이 범주형이기는 하지만, 유사도를 계산할 때에 거리 값을 이용할 수 있으며, Complete-LVF 알고리즘을 적용할 때에도 변환(Binning)할 필요가 없다. F0속성을 제외한 모든 속성을 포함하는 사례기반추론 시스템을 구축하여 선택효과를 구한 제 1단계의 결과는 <표 4>와 같다.

<표 4> 속성별 선택효과-제1단계(유방암 진단)

속성	적중률 (%)	순위	속성	적중률 (%)	순위
F6	90.78	1	F1	75.90	6
F4	85.76	2	F7	75.56	7
F8	84.49	3	F5	69.71	8
F2	83.76	4	F9	54.44	9
F3	83.12	5			

유방암 진단 데이터의 제 1단계의 결과는 <표 4>에서 처럼 F6, F4, F8, F2, F3, F1, F7, F5, F9의 순서를 보이고 있다. 순위간의 적중률 차이가 비교적 높게 나타나고 있는데, 1순위 속성과 마지막 9순위의 값의 차이가 36.34% 포인트이다. <표 4>의 결과를 사용하여 제 2단계인 후보 속성군을 구성한 결과는 <표 5>와 같다.

총 10회의 실험 결과가 <표 5>에 보여지고 있는데, 두 번째 열에는 각 실험에서 선정된 '후보 속성군'의 구성을 보여주고 있으며, 세 번째 열에는 후보 속성군만을 사용한 사례기반추론 시스템의 적중률을, 네 번째 열에는 F0를 제외한 모든 속성

<표 5> 후보 속성군-제 2단계(유방암 진단)

실험 번호	후보 속성군	후보 속성군 적중률(%)	모든 속성 적중률 (%)
1	F6, F4, F8, F2	97.38	97.38
2	F6, F4, F8, F2, F3	97.91	97.91
3	F6, F4, F8, F2	97.38	96.86
4	F6, F4, F8	97.91	97.38
5	F6, F4, F8, F2, F3	96.86	97.38
6	F6, F4, F8	95.81	94.76
7	F6, F4, F8, F2	97.38	97.38
8	F6, F4, F8, F2, F3, F1	96.34	95.81
9	F6, F4, F8	97.38	96.86
10	**F6, F4, F8, F2, F3, F1**	97.91	96.86
		97.23	**96.86**

을 사용한 사례기반추론 시스템의 적중률을 보여주고 있다. <표 5>의 결과를 살펴보면, 실험 번호 2, 4, 10에서 후보 속성군에 의한 적중률이 97.91%로 가장 높다. 제 3.2절에서 언급한 바와 같이, 적중률이 동일하면 속성의 개수가 많은 경우를 선택한다. 그러므로 실험번호 10의 후보 속성군인 {F6, F4, F8, F2, F3, F1}을 선택한다. 제 3단계에서는, 후보 속성군 {F6, F4, F8, F2, F3, F1}의 모든 부분집합에 대해서 Complete-LVF 알고리즘으로 불일치비율 δ를 측정하는데, 그 결과는 <표 6>과 같다.

제 3.2절에서 언급하였듯이, 불일치비율이 가장 작은 부분집합이 두 개 이상 나오는 경우에는 속성의 개수가 적은 집합을 선택한다. <표 6>에 보듯이, δ 값이 가장 작은 0을 가지면서 속성의 개수가 가장 적은 부분집합이 모두 7개로 나타났다. 이것들 중에서 어느 것을 선택하는 것에 대한 기준은 마련하지 않았으므로, 가장 먼저 파악된 {F1, F2, F3, F6}을 선택하였다. 이렇게 선정된 속성의 평가용(Evaluation) 데이터 집합에 대한 10-Fold 성능 측정 결과는 <표 7>과 같다.

<표 7>을 보면, UV 알고리즘으로 선정된 속성만을 사용한 사례기반추론의 분류 적중률이 10회 평균 96.72%로서, 전체 속성을 사용한 결과인 96.08%보다 약간 높게 나타나고 있다. Lift는 선정된 속성 적중률을 모든 속성 적중률로 나눈 수치

<표 6> 불일치비율(일부)-제 3단계(유방암 진단)

부분집합 구성속성	δ	부분집합 구성속성	δ	부분집합 구성속성	δ
F1	0.103	F1, F2, F8	0.010	F1, F2, F4, F8	0.002
F2	0.063	F1, F3, F4	0.013	**F1, F2, F6, F8**	**0**
F1, F2	0.036	F1, F3, F8	0.010	F1, F3, F4, F8	0.002
F1, F3	0.033	F1, F4, F6	0.003	**F1, F3, F6, F8**	**0**
F1, F4	0.050	F1, F4, F8	0.027	**F1, F4, F6, F8**	**0**
F2, F3	0.040	F2, F3, F6	0.008	F2, F3, F6, F8	0.002
F2, F4	0.013	**F1, F2, F3, F6**	**0**	**F2, F4, F6, F8**	**0**
F2, F6	0.015	F1, F2, F3, F8	0.006	**F3, F4, F6, F8**	**0**
F6, F8	0.021	**F1, F2, F4, F6**	**0**	F1, F2, F3, F4, F6, F8	0

<표 7> 선정된 속성의 성능(유방암 진단)

선정된 속성 적중률(%)	모든 속성 적중률(%)
97.06	98.53
97.06	95.56
98.53	98.53
94.12	95.56
100.00	98.53
97.06	98.53
92.65	88.24
95.56	98.53
95.12	93.17
100.00	95.56
μ = 96.72	μ = 96.08
σ = 2.41	σ = 3.35
Lift = 1.007	

이다. 적중률의 향상을 많이 보이지는 못했지만 9개 속성 중 4개의 속성만을 사용하여서 적중률이 하락하지 않았다는 것에 의의가 있다고 하겠다.

4.2 피마 인디언의 당뇨병(Pima Indians Diabetes)

피마 인디언의 당뇨병 데이터는 피마 인디언 보호구역에 살고 있는 21세 이상 된 여성 중 당뇨병을 보유하고 있는 여성 268명과 당뇨병을 보유하고 있지 않은 여성 500명의 사례를 포함하고 있으며, 8개의 설명 속성과 1개의 목표 속성이 있다. 결측치는 없으며, 속성 구성은 <표 8>과 같다.

<표 8>에서 보듯이 8개의 설명 속성이 모두 연속형이므로 사례기반추론 시스템의 유사도 측정에서는 거리 값을 이용할 수 있지만, Complete-LVF 알고리즘에서는 그대로 사용할 수가 없다. 즉, 연속형 속성을 범주형으로 변환할 필요가 있다. 속성들의 선택효과를 측정한 결과는 <표 9>와 같다.

<표 9>의 결과를 이용하여 제 2단계인 후보 속성군을 구성한 결과는 <표 10>과 같다. <표 10>에서 보듯이, 후보 속성군은 {F2, F8, F6, F5, F7}로 결정된다. 다음에는 Complete-LVF 알고리즘을 적용하여 불일치비율을 구하기 위해서 연속형

<표 8> 피마 인디언 당뇨병 데이터 속성 구성

속 성	설 명	유 형
F1	임신횟수(Number of Times Pregnant)	연속형
F2	구강내 2시간 동안의 글루코오즈 잔류량 테스트 (Plasma Glucose Concentration of 2 hours in and Oral Glucose Tolerance Test)	
F3	혈압 <mm Hg> (Diastolic Blood Pressure)	
F4	삼두박근 쪽의 피부 두께<mm> (Triceps Skin Fold Thickness)	
F5	2시간 동안의 세룸 인슐린의 양<mu U/ml> (2-Hour Serum Insulin)	
F6	비만도<몸무게 kg/ 키 m^2> (Body Mass Index)	
F7	가계 당뇨병 병력 함수(Diabetes Pedigree Function)	
F8	나이 <년> (Age)	
목표 속성	양성반응(1), 음성반응(0)	범주형

<표 9> 속성별 선택효과-제1단계(피마 인디안 당뇨병)

속성	적중률 (%)	순위	속성	적중률 (%)	순위
F2	72.61	1	F7	61.02	5
F8	65.58	2	F1	60.79	6
F6	63.77	3	F4	59.30	7
F5	61.77	4	F3	58.19	8

<표 10> 후보 속성군-제2단계(피마 인디언 당뇨병)

실험 번호	후보 속성군	후보 속성군 적중률(%)	모든 속성 적중률(%)
1	F2, F8, F6	77.67	74.88
2	F2	73.95	76.28
3	F2, F8, F6, F5, F7	79.07	74.88
4	F2, F8, F6, F5	75.35	73.02
5	F2, F8, F6, F5	78.61	75.81
6	**F2, F8, F6, F5, F7**	80.00	73.49
7	F2, F8	75.81	71.16
8	F2, F8, F6, F5, F7	79.54	77.21
9	F2	71.63	70.70
10	F2, F8, F6, F5, F7	76.74	68.84
		76.84	**73.63**

속성을 범주형으로 변환할 필요가 있다. 속성 값의 변환 기준은 제 3.2절에서 언급한 바와 같이 각 Bin에 거의 동일한 개수의 개체를 포함시키는 동일개체수기반 변환방법을 이용하였다. 각 속성의 변환 분류 기준 및 각 Bin에 포함되는 개체 즉 레코드의 개수는 <그림 8>과 같다.

<그림 8>을 보면, 동일개체수기반 변환방법을 사용하였지만, 각 Bin에 포함되는 레코드의 개수가 다소 차이가 나는 것을 알 수 있다. 그 이유는, 속성 값이 동일한 레코드들이 여러 개 있는 경우에는 이들을 하나의 Bin에 할당할 수밖에 없기 때문이다. 예를 들어, 속성 F5는 0의 값을 갖는 레코드가 255개로 전 데이터의 과반수를 넘지만 이들을 다시 세분할 수 없으므로 하나의 Bin에 할당하였다. <그림 8>의 기준을 적용하여 연속형 속성 값들을 범주형 값으로 변환한 후, 불일치비율을 측정한 결과는 <표 11>과 같다.

<표 11>에서 보듯이, 피마 인디언 당뇨병 데이터에서는 제 2단계에서 얻은 후보 속성군을 모두 사용하였을 때, 즉 {F2, F5, F6, F7, F8}의 5개의 속성을 사용하였을 때에 불일치비율이 가장 작았다. 그러므로 이 집합을 최종적으로 선정된 속성으로 결정하였다. 이렇게 선정된 속성의 10-Fold 성

F2	**≤ 90**	**≤ 100**	**≤ 110**	**≤ 120**	**≤ 130**	**≤ 150**	**≤ 200**	분류기준
	78	72	67	73	75	78	97	레코드개수
	0	**1**	**2**	**3**	**4**	**5**	**6**	Bin 번호

F5	**0**	**≤ 92**	**≤ 168**	**≤ 744**
	255	95	95	93
	0	**1**	**2**	**3**

F6	**≤ 22**	**≤ 27**	**≤ 31**	**≤ 34**	**≤ 38**	**> 38**
	32	101	113	87	102	103
	0	**1**	**2**	**3**	**4**	**5**

<그림 8> 연속형을 범주형으로 변환하는 기준(일부)
(피마 인디언 당뇨병)

능 측정 결과는 <표 12>와 같다.

<표 11> 불일치비율(일부)-제 3단계(피마 인디언 당뇨병)

부분집합 구성 속성	δ	부분집합 구성 속성	δ
F2	0.249	F2, F5, F7	0.186
F5	0.335	F2, F5, F8	0.190
F6	0.335	F2, F6, F7	0.136
F5, F8	0.277	F2, F5, F7, F8	0.112
F6, F7	0.288	F2, F6, F7, F8	0.056
F6, F8	0.260	F5, F6, F7, F8	0.125
F7, F8	0.294	**F2, F5, F6, F7, F8**	**0.028**

<표 12> 선정된 속성의 성능(피마 인디언 당뇨병)

선정된 속성 적중률(%)	모든 속성 적중률(%)
74.78	61.74
80.52	72.73
72.73	66.23
77.92	76.62
72.73	71.43
74.03	74.03
81.82	77.92
80.52	81.82
77.92	75.33
68.83	70.13
μ = 76.18	μ = 72.80
σ = 4.22	σ = 5.82
Lift = 1.046	

<표 12>에서 보듯이, 속성의 개수를 줄이는 효율성 면에서는 8개를 5개로, 즉 62.5%로 줄이는 유용성을 보여주었고, 적중률을 향상시키는 효과성 면에서는 72.80%에서 76.16%로 3.38% 포인트 증가, Lift로는 1.046을 보여서 만족할만한 성능을 보였다.

4.3 신용 평가(Credit Approval Screening)

신용 평가 데이터는 은행이나 신용카드 회사에서 개인 대출이나 카드 발급을 할 때, 은행이나 회사에 주는 위험을 낮추기 위해서 그 한도를 결정하거나 여부를 판정하는 업무에 사용되는 설명 속성들을 포함하고 있고, 목표 속성의 값은 대출여부이다. 총 레코드가 690개인데, 여기에서 결측치를 갖는 레코드를 제거하고 본 연구에서는 651개의 레코드를 사용하였다. 데이터의 속성 구성은 <표 13>과 같다.

<표 13> 신용 평가 데이터의 속성 구성

속 성	설 명	유 형
F1	b, a	범주형
F2	continuous	연속형
F3	continuous	
F4	u, y., l, t.	범주형
F5	g, p, gg	
F6	c, d, cc, i, j, k, m, r, q, w, x, e, aa, ff	
F7	v, h, bb,j, n, z, dd, ff, o	
F8	continuous	연속형
F9	t, f	범주형
F10	t, f	
F11	continuous	연속형
F12	t, f	범주형
F13	g, p, s	
F14	continuous	연속형
F15	continuous	
목표 속성	+ : 적격자, - : 부적격자	이진형

<표 13>에서 보듯이, 신용 평가 데이터에는 각 속성들에 대한 구체적인 설명이 제공되지 않고 암호화된 속성 값들로 제시되어 있는데, 그 이유는 속성 값들이 모두 개인 신용과 관련되는 내용이기 때문이다. 이 데이터에는 연속형 속성과 범주형 속성이 혼합되어 있다. 연속형 속성은 동일개체수기반 변환방법을 통해서 범주형 속성으로 변환하여 사용한다. 속성들의 선택효과를 측정한 결과는 <표 14>와 같다.

<표 14> 속성별 선택효과-제 1단계(신용 평가)

속성	적중률 (%)	순위	속성	적중률 (%)	순위
F9	89.67	1	F4	58.14	9
F11	73.72	2	F5	58.14	10
F10	69.02	3	F3	57.60	11
F15	67.92	4	F2	56.78	12
F8	63.99	5	F12	56.72	13
F7	61.64	6	F1	55.74	14
F6	61.20	7	F13	52.79	15
F14	59.95	8			

<표 14>의 결과를 이용하여, 제 2단계인 후보 속성군을 구성한 결과는 <표 15>와 같다.

<표 15> 후보 속성군-제 2단계(신용 평가)

실험 번호	후보 속성군	후보 속성군 적중률(%)	모든 속성 적중률(%)
1	F9	87.43	83.06
2	F9, F11, F10, F15	87.99	86.34
3	F9, F11, F10, F15	86.89	86.34
4	F9	92.35	87.07
5	F9, F11, F10, F15	91.80	89.07
6	**F9, F11, F10, F15, F8, F7**	91.86	87.43
7	F9, F11, F10	91.80	90.71
8	F9, F11, F10, F15	91.26	92.35
9	F9, F11, F10	87.43	90.71
10	F9, F11, F10	91.80	89.07
		90.00	**88.42**

<표 15>에서 보듯이, 후보 속성군은 {F9, F11, F10, F15, F8, F7}로 결정된다. 다음에는 연속형 속성인 F8, F11, F15를 동일개체수기반 변환방법

F8

0	≤ 0.1	≤ 0.25	≤ 0.6	≤ 1.0	≤ 1.75	≤ 3	≤ 5.5	> 5	분류기준
50	46	65	52	46	63	52	50	34	레코드개수
0	1	2	3	4	5	6	7	8	Bin 번호

F11

0	1	≤ 3	≤ 8	≤ 40
286	44	40	44	44
0	1	2	3	4

F15

0	≤ 10	≤ 100	≤ 350	≤ 1000	> 1000
179	64	53	49	51	62
0	1	2	3	4	5

<그림 9> 연속형 속성을 범주형으로 변환하는 기준(신용 평가)

을 이용하여 범주형 속성으로 변환한다. 각 속성의 변환 분류 기준 및 각 Bin에 포함되는 레코드의 개수는 <그림 9>와 같다. <그림 9>의 기준을 적용해서 연속형 속성을 범주형으로 변환한 후에, 불일치비율을 측정한 결과는 <표 16>과 같다.

<표 16>을 보면, δ값이 가장 작은 것이 0.050인데, 그 중에 속성개수가 가장 적은 집합은{F7, F8, F10, F11, F15}이다. 이렇게 선정된 속성의 10-Fold 성능 측정 결과는 <표 17>과 같다.

이 데이터도 유방암 진단 데이터와 마찬가지로 적중률 증가의 효과성 측면에서는 미흡하다. 하지만 속성개수 감소의 효율성 측면에서는 15개가 5개로 감소하여, 무려 1/3로 감소하였다. 이 정도의 속성들만을 사용하여, 전체 속성을 모두 사용할 때와 비슷한 적중률을 보인다는 것은 의미 있는 결과이다.

<표 17> 선정된 속성의 성능(신용 평가)

선정된 속성 적중률(%)	모든 속성 적중률(%)
84.62	86.15
84.62	83.08
90.77	87.69
83.08	84.62
84.62	83.08
81.54	78.46
87.69	87.69
86.15	87.69
86.14	83.08
76.92	73.85
μ = 84.62	μ = 83.54
σ = 3.70	σ = 4.47
Lift = 1.013	

4.4 성능 종합 평가

UV 알고리즘의 성능을 기존의 기법들과 비교한 결과는 <표 18>과 같다. FSS, BSS와 RC의 결과는 Domingos(1997)에 제시된 것들이다. <표 18>에 기록된 UV 알고리즘의 적중률 96.7 ± 2.4는 10회 실험한 적중률의 평균인 96.7과 표준편차

<표 16> 불일치비율(일부)-제 3단계(신용 평가)

부분집합 구성 속성	δ	부분집합 구성 속성	δ	부분집합 구성 속성	δ
F7	0.100	F7, F8, F9	0.100	F7, F8, F9, F11	0.094
F8	0.245	F7, F8, F10	0.092	F7, F8, F9, F15	0.087
F9	0.288	F7, F8, F11	0.094	F7, F8, F10, F11	0.070
F7, F15	0.098	F8, F9, F10	0.212	F8, F9, F10, F15	0.162
F8, F9	0.245	F8, F9, F11	0.197	F8, F9, F11, F15	0.164
F9, F11	0.234	F9, F10, F11	0.159	F7, F8, F9, F11, F15	0.072
F9, F15	0.260	F9, F10, F15	0.181	**F7, F8, F10, F11, F15**	**0.050**
F10, F15	0.236	F10, F11, F15	0.153	F8, F9, F10, F11, F15	0.092
F11, F15	0.251	F7, F8, F9, F10	0.092	F7, F8, F9, F10, F11, F15	0.050

인 2.4를 의미한다. 다른 기법들의 적중률도 '±'로 표시된 수치들은 평균과 표준편차를 의미한다.

<표 18>에서 보듯이, UV 알고리즘이 다른 기법들보다 우수한 성능을 보이고 있다. 먼저, 전방향 순차 탐색(FSS)과 비교하면 세 가지 데이터 모두에서 적중률이 상승하였음을 볼 수 있다. 속성개수의 감소 측면이 다소 뒤지지만, 유방암 진단 데이터에서는 적중률이 FSS의 66.7%보다 큰 폭으로 상승한 96.7%를 나타내고 있다. 역방향 순차 탐색(BSS)과의 비교에서는 속성개수도 더 적으면서 적중률도 더 높게 나타나고 있다. RC 기법을 보면, RC 기법 자체는 속성개수의 감소 면에서는 별로 유용성이 없는 것을 알 수 있다. 반면에 UV 알고리즘은 속성의 개수를 대폭으로 감소시키면서, 적중률도 신용 평가 데이터에서만 RC 기법과 비슷했을 뿐, 다른 데이터에서는 높은 적중률을 보이고 있다.

FeaSUA 알고리즘(이재식, 이혁희, 2002)의 결과는 선택효과의 상위 순위 30%를 선정했을 때의 적중률과 속성 개수이다. 피마 인디언 당뇨병 데이터에서는 UV 알고리즘보다 적중률이 저조하지만, 신용평가 데이터의 경우에는 UV 알고리즘보다 적중률이 1% 포인트 높게 기록되었다. Wang et al.(2007)은 PSORSFS 알고리즘의 성능을 13개의 UCI 데이터를 사용하여 측정하였다. 하지만 그들의 기법은 속성이 범주형인 경우에만 작동하기 때문에 본 연구 결과와 비교할 수 있는 데이터는 유방암 진단 데이터밖에 없다. 속성의 개수는 UV 알고리즘과 비슷하고, 적중률은 UV 알고리즘보다 다소 낮지만 표준편차가 작아서 안정적이었다고 평가할 수 있다. 하지만, 본 연구에서처럼 적중률이 100%에 이른 경우는 없었다.

<표 18> 타 기법과의 성능비교

알고리즘 \ 데이터		유방암 진단	피마인디언 당뇨병	신용 평가
전체 속성		9개	8개	15개
UV	적중률(%)	96.7 ± 2.4	76.2 ± 4.2	84.6 ± 3.7
	속성선정	4개	5개	5개
FSS	적중률(%)	66.7 ± 6.7	69.6 ± 2.9	80.9 ± 2.3
	속성선정	2.3개	1.9개	5.7개
BSS	적중률(%)	66.9 ± 6.1	69.2 ± 3.3	81.2 ± 2.5
	속성선정	4.8개	6.6개	9.6개
RC	적중률(%)	66.2 ± 5.2	70.5 ± 2.5	83.7 ± 1.9
	속성선정	7.7 ± 1.1개	7.9 ± 0.3개	13.5 ± 1.2개
FeaSUA	적중률(%)	-	71.0 ± 0.0	85.7 ± 0.1
	속성선정	-	2개	4개
PSORSFS	적중률(%)	95.5 ± 0.7	-	-
	속성선정	4.2 ± 0.4개	-	-

5. 결론 및 향후 연구 과제

본 연구에서는 기존의 속성 선정 방법인 FeaSUA와 LVF를 결합하여, 효율성과 효과성을 갖춘 속성선정 방법인 UV 알고리즘을 개발하였다. UV 알고리즘은 Filter Model과 Wrapper Model을 결합하는 Hybrid Model로서, 완전 탐색 전략을 채택하고 분류 문제를 대상으로 하는 속성 선정 방법이다. 효율성은 속성개수의 감소로, 효과성은 전체 속성을 전부 사용했을 때의 적중률에 대한 Lift 값으로 검증하였다. UV 알고리즘의 성능을 평가하기 위하여 UCI Machine Learning Repository에서 수집한 3개의 데이터 집합을 사용하였다. 선정된 속성의 개수는 유방암 진단 데이터의 경우 9개의 속성 중에서 4개(44.4%), 피마 인디언 당뇨병 데이터의 경우 8개의 속성 중에서 5개(62.5%), 신용 평가 데이터의 경우 15개 속성 중에서 5개(33.3%)이었다. 적중률에서는 속성 선정 후에 감소한 경우는 없었고, Lift 값은 각 데이터에 대해서 1.007, 1.046, 1.013 이었다. 본 연구에서는 데이터 마이닝 기법으로 사례기반 추론을 사용하였으나, UV 알고리즘은 기법에 의존적이 아니기 때문에 다른 기법의 경우에도 본 연구에서 제시한 방법론에 따라 적합한 속성 선정 기법을 개발할 수 있을 것이다.

본 연구의 가장 큰 한계점은 제 3.2절에서 언급하였듯이, UV 알고리즘의 제 3단계인 Complete-LVF에서 가능한 모든 부분집합에 대해서 계산이 수행된다는 것이다. UV 알고리즘의 제 2단계에서 전체 속성 중에서 영향력이 큰 속성들이 발췌되기 때문에 전체 속성에 대한 모든 부분집합을 고려해야 하는 부담은 줄일 수 있지만, 전체 속성의 개수가 원래 많다면, Complete-LVF의 부담은 커질 것이다. FeaSUA에서는 전체 속성의 30%만으로도 만족할만한 성능을 얻을 수 있다는 결과를 제시한 바가 있다(이재식, 이혁희, 2002). 그러므로, Complete-LVF의 단계로 가기 전에 속성 후보군의 개수를 좀 더 줄이는 연구는 계속 되어야 할 것이다. 두 번째 한계점은 본 연구에서 사용한 사례기반 추론의 구조를 가장 단순하게 하였다는 것이다. 사례기반 추론에서는 속성들에게 가중치를 상이하게 부여할 수 있는데, 본 연구에서는 모두 1로 부여하였다. 물론 속성 가중치를 상이하게 주는 작업과, 속성 선정을 결정하는 작업을 동시에 수행하는 것이 용이한 작업은 아니지만, 속성의 가중치에 따라 그 속성의 선정 또는 탈락이 영향을 받게 될 것이므로, 향후 연구에서는 사례기반 추론의 구조를 조금 더 복잡하게 할 필요성이 있다.

참고문헌

이재식, 이혁희, “개별 속성의 선택 및 제거효과 순위를 이용한 사례기반 추론의 속성 선정”, *한국지능정보시스템학회논문지*, 8권 2호(2002), 117~137.

Aamodt, A. and E. Plaza, “Case-Based Reasoning : Foundational Issues, Methodological Variations, and System Approaches”, *Artificial Intelligence Communications*, Vol.7, No.1(1994), 39~59.

Aha, D. W. and R. L. Bankert, “Feature Selection for Case-Based Classification of Cloud Types : An Empirical Comparison”, *Proceedings of the AAAI Workshop on CBR*(1994), 106~112.

Berry, M. J. A. and G. S. Linoff, *Data Mining Techniques : For Marketing, Sales, and Customer Relationship Management*, 2nd ed., Wiley Pub. Inc., 2004.

Blake, C., E. Keogh and C. J. Merz, UCI Repository

of Machine Learning Databases, [http://www.ics.uci.edu/~mlearn/MLRepository.html], Dept. of Information and Computer Science, Univ. of California at Irvine, 1998.

Chanchien, S. W. and M. Lin, "Design and Implementation of a Case-based Reasoning System for Marketing Plans", *Expert Systems with Applications*, Vol.28(2005), 43~53.

Dash M. and H. Liu, "Feature Selection for Classification", *Intelligent Data Analysis*, Vol.1(1997), 131~156.

Doak, J., *An Evaluation of Feature Selection Methods and Their Application to Computer Security*, Technical Report CSE-92-18, Dept. of Computer Science, Univ. of California at Davis, 1992.

Domingos, P., "Context-Sensitive Feature Selection for Lazy Learners", *Artificial Intelligence Review*, Vol.11(1997), 227~253.

Goker, M. H. and T. Roth-Berghofer, "The Development and Utilization of the Case-Based Help-Desk Support System HOMER", *Engineering Applications of Artificial Intelligence*, Vol.12(1999), 664~680.

Goldberg, D. E., *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley Pub. Co., Inc., 1989.

Jo, H., I. Han and H. Lee, "Bankruptcy Prediction Using Case-Based Reasoning, Neural Networks, and Discriminant Analysis", *Expert Systems with Applications*, Vol.13 No.2(1997), 97~108.

John, G. H., R. Kohavi and K. Pfleger, "Irrelevant Features and the Subset Selection Problem", *Proceedings of the 11th International Conference on Machine Learning*, (1994), 121~129.

Kuo, R. J., Y. P. Kuo and K. Y. Chen, "Developing a Diagnostic System through Integration of Fuzzy Case-Based Reasoning and Fuzzy Ant Colony System", *Expert Systems with Applications*, Vol.28(2005), 783~797.

Law, Y. F. D., S. B. Foong and S. E. J. Kwan, "An Integrated Case-Based Reasoning Approach for Intelligent Help Desk Fault Management", *Expert Systems with Applications*, Vol.13(1997), 265~274.

Leake, D., A. Maguitman and T. Reichherzer, "Cases, Context, and Comfort : Opportunities for Case-Based Reasoning in Smart Homes", *Lecture Notes in Artificial Intelligence*, Vol.4008(2006), 109~131.

Lee, J. S. and J. C. Lee, "Music for My Mood : A Music Recommendation System based on Context Reasoning", *Lecture Notes in Computer Science*, Vol.4272(2006), 190~203.

Lee, J. S. and J. C. Lee, "Context Awareness by Case-based Reasoning in a Music Recommendation System", *Lecture Notes in Computer Science*, Vol.4836(2007), 45~58.

Lee, J. S. and Y. X. Xon, "A Customer Service Process Innovation using the Integration of Database and Case base", *Expert Systems with Applications*, Vol.11, No.4(1996), 543~552.

Liu, H. and R. Setiono, "A Probabilistic Approach to Feature Selection-a Filter Solution." *Proceedings of the 13th International Conference on Machine Learning* (1996), 319~327.

Liu, H. and L. Yu, "Toward Integrating Feature Selection Algorithms for Classification and Clustering", *IEEE Transactions on Knowledge and Data Engineering*, Vol.17, No.4 (2005), 491~502.

Pawlak, Z., "Rough Sets", *International Journal of Computer and Information Science*, Vol.11,

No.5(1982), 341～356.

Varma, A. and N. Roddy, “ICARUS : Design and Development of a Case-Based Reasoning System for Locomotive Diagnostics”, *Engineering Applications of Artificial Intelligence*, Vol.12(1999), 681～690.

Wang, H. C. and H. S. Wang, “A Hybrid Expert System for Equipment Failure Analysis”, *Expert Systems with Applications*, Vol.28 (2005), 615～622.

Wang, X., J. Yang, X. Teng, W. Xia and R. Jensen, “Feature Selection based on Rough Sets and Particle Swarm Optimization”, *Pattern Recognition Letters*, Vol.28(2007), 459～471.

Abstract

A Hybrid Feature Selection Method using Univariate Analysis and LVF Algorithm

Jae Sik Lee* · Mi Kyoung Jeong**

We develop a feature selection method that can improve both the efficiency and the effectiveness of classification technique. In this research, we employ case-based reasoning as a classification technique. Basically, this research integrates the two existing feature selection methods, i.e., the univariate analysis and the LVF algorithm. First, we sift some predictive features from the whole set of features using the univariate analysis. Then, we generate all possible subsets of features from these predictive features and measure the inconsistency rate of each subset using the LVF algorithm. Finally, the subset having the lowest inconsistency rate is selected as the best subset of features. We measure the performances of our feature selection method using the data obtained from UCI Machine Learning Repository, and compare them with those of existing methods. The number of selected features and the accuracy of our feature selection method are so satisfactory that the improvements both in efficiency and effectiveness are achieved.

Key Words : Feature Selection, Case-based Reasoning, Univariate Analysis, LVF Algorithm

* e-Business Division, School of Business Administration, Ajou University
** ING Life, Co.

저 자 소 개

이재식

현재 아주대학교 경영대학 e-비즈니스학부의 교수로 재직 중이다. 서울대학교 경영학과에서 경영학사(1977), KAIST 산업공학과에서 공학석사(1979), 그리고 University of Pennsylvania, Wharton School에서 경영정보시스템 전공으로 경영학 박사학위(1989)를 취득하였다. Decision Support Systems, Management Science, Expert Systems with Applications, Annals of OR, Lecture Notes in Artificial Intelligence, Lecture Notes in Computer Science 등의 외국저널에 Heuristic Algorithm, Model Management, Case-based Reasoning, Context Reasoning, Recommendation 등을 주제로 한 논문들을 게재하였고, 국내저널에도 다수의 논문을 게재하였다. 주요 관심분야는 Data Mining, Ubiquitous Computing, Intelligent Information Systems, AI Application to Business Problem Solving 등이다.

정미경

현재 ㈜ING생명 코리아에 재정컨설턴트로 재직 중이다. 아주대학교 경영학사(1998), 아주대학교 대학원 경영정보학과에서 석사학위(2002)를 취득하였다. 2002년 이후 현재까지 다수의 Data Mining과 CRM 관련 프로젝트에 참여하였으며, 특히 BI/DW 분야의 실무경험이 많다. ㈜오비씨소프트 소속으로 현대홈쇼핑의 DW 구축, LG전자/LG나라의 Data Mining 프로젝트, 외환은행/예금보험공사 등의 전략컨설팅을 경험하였다. ㈜오토에버시스템즈에서는 DW를 실무에 활용하고 운영, 유지 및 보수 업무를 주로 담당하였다. 주요 관심분야는 Data Mining, BI 기반 시스템인 DW 구축과 BI 정보 요약을 통한 전략컨설팅 등이다.