(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2013-0035735
(43) 공개일자 2013년04월09일

(51) 국제특허분류(Int. Cl.)
G06F 17/30 (2006.01) *G06F 17/00* (2006.01)
(21) 출원번호 10-2011-0100236
(22) 출원일자 2011년09월30일
심사청구일자 2011년09월30일

(71) 출원인
경희대학교 산학협력단
경기도 용인시 기흥구 덕영대로 1732, 국제캠퍼스 내 (서천동, 경희대학교)
(72) 발명자
이영구
경기도 용인시 기흥구 덕영대로 1732, 전자정보관 314호 (서천동, 경희대학교)
한용구
경기도 용인시 기흥구 덕영대로 1732, 전자정보관 349호 (서천동, 경희대학교)
박기성
경기도 용인시 기흥구 덕영대로 1732, 전자정보관 349호 (서천동, 경희대학교)
(74) 대리인
특허법인 신지

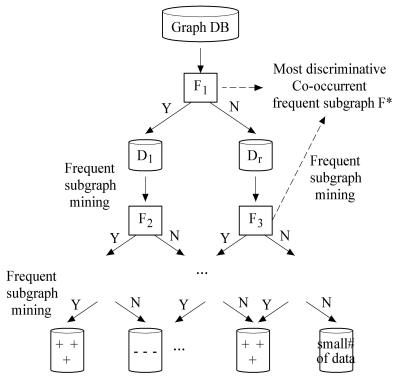
전체 청구항 수 : 총 8 항

(54) 발명의 명칭 **동시 발생 빈발 부분 그래프를 이용한 그래프 분류 장치 및 방법**

(57) 요 약

본 발명은 동시 발생 빈발 부분 그래프를 이용한 그래프 분류 방법으로, 그래프 데이터베이스로부터 빈발 부분 그래프를 마이닝하는 단계와, 상기 빈발 부분 그래프 마이닝에 의해 생성된 빈발 부분 그래프 특징 집합에서 동시 발생 빈발 부분 그래프 특징을 고려하여 가장 변별력이 높은 동시 발생 빈발 부분 그래프 특징을 선택하는 단계와, 상기 선택된 동시 발생 빈발 부분 그래프 특징의 포함 여부에 따라 상기 그래프 데이터 베이스를 분할하는 단계를 포함한다.

대 표 도 - 도2



이 발명을 지원한 국가연구개발사업

과제고유번호	2010-0013689
부처명	교육과학기술부
연구사업명	일반연구자지원사업
연구과제명	그래프 마이닝을 위한 부분감독기반 특징선택 기술 연구
주관기관	경희대학교 산학협력단
연구기간	2010.05.01 ~ 2013.04.30

특허청구의 범위

청구항 1

그래프 데이터베이스로부터 빈발 부분 그래프를 마이닝하는 단계와,

상기 빈발 부분 그래프 마이닝에 의해 생성된 빈발 부분 그래프 특징 집합에서 동시 발생 빈발 부분 그래프 특징을 고려하여 가장 변별력이 높은 동시 발생 빈발 부분 그래프 특징을 선택하는 단계와,

상기 선택된 동시 발생 빈발 부분 그래프 특징의 포함 여부에 따라 상기 그래프 데이터 베이스를 분할하는 단계를 포함함을 특징으로 하는 동시 발생 빈발 부분 그래프를 이용한 그래프 분류 방법.

청구항 2

제 1항에 있어서,

상기 분할된 데이터베이스들에 포함된 그래프가 소정 개수 이하가 아닐 경우, 분할된 데이터베이스들 각각에 대해 상기 빈발 부분 그래프들을 마이닝하는 단계 및 상기 동시 발생 빈발 부분 그래프 특징을 선택하는 단계를 수행함을 특징으로 하는 동시 발생 빈발 부분 그래프를 이용한 그래프 분류 방법.

청구항 3

제 1항에 있어서, 상기 선택하는 단계는

전체 빈발 부분그래프 특징 집합에서 최상위 변별력을 가지는 베스트 빈발 부분 그래프를 선택하는 단계와,

상기 베스트 빈발 부분 그래프와 다른 빈발 부분 그래프의 동시 발생 특징의 변별력을 계산하는 단계와,

가장 높은 동시 발생 특징의 변별력이 상기 베스트 빈발 부분 그래프의 변별력보다 높으면 동시 발생 빈발 부분 그래프의 특징을 베스트 빈발 부분에 추가하는 단계와,

가장 높은 동시 발생 특징의 변별력이 상기 베스트 빈발 부분 그래프의 변별력보다 높지 않으면, 특징 선택 알고리즘을 종료하는 단계를 포함함을 특징으로 하는 동시 발생 빈발 부분 그래프를 이용한 그래프 분류 방법.

청구항 4

제 1항에 있어서,

상기 선택하는 단계의 이전에 높은 지지도의 빈발 부분 그래프 특징들을 제거하는 단계를 더 포함함을 특징으로 하는 동시 발생 빈발 부분 그래프를 이용한 그래프 분류 방법.

청구항 5

그래프 분류 대상인 그래프들 및 그래프들의 정보를 저장하는 그래프 데이터베이스와,

상기 그래프 데이터베이스로부터 빈발 부분 그래프를 마이닝하는 그래프 마이닝부와,

상기 그래프 마이닝부에 의해 출력되는 하나 이상의 빈발 부분 그래프 특징 집합을 저장하는 빈발 부분 그래프 저장부와,

상기 빈발 부분 그래프 특징 집합에서 동시 발생 빈발 부분 그래프 특징을 고려하여 가장 변별력이 높은 동시 발생 빈발 부분 그래프 특징을 선택하는 동시 발생 빈발 부분 그래프 특징 선택부와,

상기 선택된 동시 발생 빈발 부분 그래프 특징의 포함 여부에 따라 상기 그래프 데이터 베이스를 분할하는 데이

터베이스 분할부를 포함함을 특징으로 하는 동시 발생 빈발 부분 그래프를 이용한 그래프 분류 장치.

청구항 6

제 5항에 있어서, 상기 데이터베이스 분할부는

상기 분할된 데이터베이스들에 포함된 그래프가 소정 개수 이하가 아닐 경우, 분할된 데이터베이스들 각각에 대해 상기 빈발 부분 그래프들을 상기 그래프 마이닝부에 출력함을 특징으로 하는 동시 발생 빈발 부분 그래프를 이용한 그래프 분류 장치.

청구항 7

제 5항에 있어서, 상기 동시 발생 빈발 부분 그래프 특징 선택부는

전체 빈발 부분그래프 특징 집합에서 최상위 변별력을 가지는 베스트 빈발 부분 그래프를 선택하고, 상기 베스트 빈발 부분 그래프와 다른 빈발 부분 그래프의 동시 발생 특징의 변별력을 계산하고, 상기 가장 높은 동시 발생 특징의 변별력이 상기 베스트 빈발 부분 그래프의 변별력보다 높으면 동시 발생 빈발 부분 그래프의 특징을 베스트 빈발 부분에 추가하고, 가장 높은 동시 발생 특징의 변별력이 상기 베스트 빈발 부분 그래프의 변별력보다 높지 않으면, 종료함을 특징으로 하는 동시 발생 빈발 부분 그래프를 이용한 그래프 분류 장치.

청구항 8

제 5항에 있어서, 상기 그래프 마이닝부는

상기 높은 지지도의 빈발 부분 그래프 특징들을 제거함을 특징으로 하는 동시 발생 빈발 부분 그래프를 이용한 그래프 분류 장치.

명 세 서

기 술 분 야

[0001] 본 발명은 그래프 분류 기술에 관한 것으로, 특히 빈발 부분 그래프를 이용한 그래프 분류 장치 및 방법에 관한 것이다.

배 경 기 술

[0002] 그래프 분류(graph classification)는 화합물, 소셜 네트워크, XML 문서 등과 같이 그래프로 모델링된 데이터를 분류하는 기술이다. 그래프 분류를 위해 그래프의 구조적 정보를 반영하는 빈발 부분그래프(frequent subgraph)들이 널리 사용되고 있다.

[0003] 빈발 부분그래프 특징 기반 그래프 분류 연구들은 일반적으로 특징 벡터 구성 단계와 분류 모델 학습 단계를 갖는다. 특징 벡터 구성 단계에서는 그래프 DB로부터 빈발 부분그래프를 마이닝한다. 그런데, 그래프 DB에서 일반적으로 수 만개의 빈발 부분그래프가 생성되기 때문에 분류에 변별력 있는 특징 선택(feature selection) 과정이 요구된다. 이 중에 변별력 있는 빈발 부분그래프들을 선택하여 특징 벡터를 구성한다. 분류 모델 학습 단계에서는 특징 벡터로 표현된 그래프 데이터를 이용하여 분류 모델을 학습시킨다.

[0004] 기존 그래프 분류 연구들에서는 개별 빈발 부분그래프의 변별력만을 고려하는 인포메이션 게인(information gain)과 같은 필터 타입(filter-type) 특징 선택 기법을 많이 사용한다. SFS(sequential forward selection)와 같은 래퍼 타입(wrapper-type) 특징 선택 기법은 필터 타입에 비교하여 높은 변별력의 특징들을 선택할 수 있지만 너무 긴 수행시간 때문에 복잡한 구조의 그래프 모델 분류에서는 사용되기 어렵다. 최근에는 그래프 내에서 유사한 분포로 나타나는 빈발 패턴을 특징으로 사용하는 방법, 그래프 커널(graph kernel) 기반 특징 선택 방법들이 제안 되었다. 그런데, 기존의 연구들은 개별 빈발 부분그래프의 변별력만을 고려하여 특징을 선택하고 있

다.

[0005] 도 1은 동시에 발생하는 특징의 변별력을 설명하기 위한 도면이다.

[0006] 도 1을 참조하면, 동시에 발생하는 부분그래프들의 변별력이 개별 부분그래프들보다 높음을 알 있다. '+'와 '-'는 각각 '+'클래스에 속하는 그래프와 '-' 클래스에 속하는 그래프들을 나타낸다. 원은 f_1 부분그래프만을 가지고 있는 그래프를, 삼각형은 f_2 부분그래프만을 가지는 있는 그래프를, 그리고, 사각형은 f_1과 f_2를 동시에 가지고 있는 그래프를 나타낸다. f_1을 가지는 그래프들과 f_2를 가지는 그래프들은 두 가지 클래스에 고르게 분포가 되어 있기 때문에 높은 변별력을 가질 수 없다. 그러나, f_1과 f_2가 동시에 발생하는 그래프들은 '+'클래스에만 속하기 때문에 좋은 변별력을 가질 수 있다.

[0007] 그러나, 종래의 특징 선택 방법은 낮은 변별력을 갖는 개별 빈발 부분그래프들의 변별력만을 고려할 뿐, 본원 발명에서와 같이 동시발생을 고려하여 높은 변별력을 가질 수 있다는 점을 반영하지 못한다.

발명의 내용

해결하려는 과제

[0008] 본 발명은 빈발 부분그래프의 동시 발생을 고려하여 특징을 선택하는 동시발생 빈발 부분그래프를 이용한 그래프 분류 장치 및 방법을 제안한다.

[0009] 또한, 본 발명은 모델 기반 탐색 트리를 적용한 동시 발생 빈발 부분 그래프를 이용한 그래프 분류 장치 및 방법을 제안한다.

과제의 해결 수단

[0010] 본 발명은 동시 발생 빈발 부분 그래프를 이용한 그래프 분류 방법으로, 그래프 데이터베이스로부터 빈발 부분그래프를 마이닝하는 단계와, 상기 빈발 부분 그래프 마이닝에 의해 생성된 빈발 부분 그래프 특징 집합에서 동시 발생 빈발 부분 그래프 특징을 고려하여 가장 변별력이 높은 동시 발생 빈발 부분 그래프 특징을 선택하는 단계와, 상기 선택된 동시 발생 빈발 부분 그래프 특징의 포함 여부에 따라 상기 그래프 데이터 베이스를 분할하는 단계를 포함한다.

[0011] 본 발명은 동시 발생 빈발 부분 그래프를 이용한 그래프 분류 장치로, 그래프 분류 대상인 그래프들 및 그래프들의 정보를 저장하는 그래프 데이터베이스와, 상기 그래프 데이터베이스로부터 빈발 부분 그래프를 마이닝하는 그래프 마이닝부와, 상기 그래프 마이닝부에 의해 출력되는 하나 이상의 빈발 부분 그래프 특징 집합을 저장하는 빈발 부분 그래프 저장부와, 상기 빈발 부분 그래프 특징 집합에서 동시 발생 빈발 부분 그래프 특징을 고려하여 가장 변별력이 높은 동시 발생 빈발 부분 그래프 특징을 선택하는 동시 발생 빈발 부분 그래프 특징 선택부와, 상기 선택된 동시 발생 빈발 부분 그래프 특징의 포함 여부에 따라 상기 그래프 데이터 베이스를 분할하는 데이터베이스 분할부를 포함한다.

발명의 효과

[0012] 본 발명은 마이닝된 빈발 부분그래프들로 구성된 특징 집합들 중에 빈발 부분그래프들의 개별적인 변별력뿐만 아니라 동시 발생 변별력을 함께 고려하여 분류에 더 유용한 특징들을 선택한다. 또한 지수 시간 복잡도를 갖는 동시발생 빈발 부분그래프의 탐색 공간에서 변별력이 높은 특징을 효율적으로 선택하기 위하여, 그리디 탐색 방법을 사용하여 탐색시간을 줄인다.

도면의 간단한 설명

[0013] 도 1은 동시에 발생하는 특징의 변별력을 설명하기 위한 도면이다.

도 2는 본 발명에 따른 모델 기반 탐색 트리를 설명하기 위한 도면이다.

도 3은 본 발명의 바람직한 실시 예에 따른 동시 발생 빈발 부분 그래프를 이용한 그래프 분류 장치를 도시한 도면이다.

도 4는 본 발명의 바람직한 실시 예에 따른 동시 발생 빈발 부분그래프를 이용한 그래프 분류 방법을 설명하기 위한 순서도이다.

도 5는 특징-그래프 행렬을 나타낸 도면이다.

도 6은 본 발명의 바람직한 실시 예에 따른 동시 발생 빈발 부분 그래프 특징 선택 방법을 설명하기 위한 순서도이다.

도 7은 높은 지지도를 갖는 패턴의 개수를 설명하기 위한 도면이다.

도 8은 본 발명에 따른 그래프 분류의 정확도 비교를 위한 도면이다.

도 9는 모델 기반 탐색 트리 구축 시작을 도시한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0014] 이하, 첨부된 도면을 참조하여 기술되는 바람직한 실시 예를 통하여 본 발명을 당업자가 용이하게 이해하고 재현할 수 있도록 상세히 기술하기로 한다.

[0015] 본 발명을 설명함에 있어 관련된 공지 기능 또는 구성에 대한 구체적인 설명이 본 발명 실시 예들의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략할 것이다.

[0016] 그리고, 명세서 전반에 걸쳐 사용되는 용어들은 본 발명 실시 예에서의 기능을 고려하여 정의된 용어들로서, 사용자 또는 운용자의 의도, 관례 등에 따라 충분히 변형될 수 있는 사항이므로, 이 용어들의 정의는 본 발명의 명세서 전반에 걸친 내용을 토대로 내려져야 할 것이다.

[0017] 본 발명은 마이닝된 빈발 부분그래프들 중에 빈발 부분그래프들의 개별적인 변별력뿐만 아니라 동시 발생 변별력을 함께 고려하여 분류에 더 유용한 특징들을 선택하고, 선택된 특징을 이용하여 그래프 분류를 수행하는 장치 및 방법을 제안한다.

[0018] 한편, 최근에는 벡터 구성 단계와 분류 모델 학습 단계를 통합한 형태의 모델 기반 탐색 트리(model based search tree: M^bT)가 제안되었다.

[0019] 도 2는 모델 기반 탐색 트리를 설명하기 위한 도면이다.

[0020] 도 2를 참조하면, 모델 기반 탐색 트리는 가장 높은 변별력을 가지는 빈발 부분그래프를 선택하고, 이 부분그래프를 기준으로 그래프 DB를 두 개의 부분 DB로 반복적으로 분할하면서 탑-다운 방식으로 구축된다. 분할된 부분 DB 중에 선택된 특징을 포함하는 DB의 모든 그래프들은 선택된 특징을 포함하게 된다. 따라서, 다음 분할 단계에서 선택될 특징은 이전에 선택된 특징과 서로 중복되지 않기 때문에 분류 성능이 향상된다.

[0021] 따라서, 본 발명은 이러한 모델 기반 탐색 트리를 기반으로 하는 그래프 분류 방식을 적용한다. 즉, 그래프 DB의 분할 기준을 본 발명의 바람직한 실시 예에 따라 선택된 동시 발생 빈발 부분 그래프 특징(COFSF)의 포함 여부로 한다.

[0022] 도 3은 본 발명의 바람직한 실시 예에 따른 동시 발생 빈발 부분 그래프를 이용한 그래프 분류 장치의 개략적인 블록 구성도이다.

[0023] 도 3을 참조하면, 본 발명의 바람직한 실시 예에 따른 동시 발생 빈발 부분 그래프를 이용한 그래프 분류 장치는 그래프 DB(10), 그래프 마이닝부(20), 빈발 부분 그래프 정보 저장부(30), 동시 발생 빈발 부분 그래프 특징 선택부(40) 및 DB 분할부(50)를 포함한다.

[0024] 그래프 분류 장치는 버텍스(vertex) 및 에지(edge)로 구성되는 그래프로 표현 가능한 다양한 형태의 데이터 예를 들어, 화합물 데이터, 네트워크 분석용 네트워크 데이터를 분류하기 위한 데이터 분석 및 분류 장치일 수 있으며, 컴퓨터, 휴대 단말 등 다양한 형태로 구현될 수 있다.

[0025] 그래프 DB(10)는 그래프 분류 대상인 그래프들 및 그래프들의 정보를 포함한다.

[0026] 그래프 마이닝부(20)는 그래프 DB(10)에 저장된 그래프들로부터 빈발 부분 그래프 집합을 추출하여, 빈발 부분 그래프 정보 저장부(30)에 저장한다.

[0027] 그래프 마이닝부(20)는 그래프 DB(10)의 그래프들에 대하여 gSpan(Graph-Based Substructure Pattern Mining), FSG(Frequent Subgraph) 마이닝 등과 같은 빈발 부분 그래프 마이닝 알고리즘을 사용하여 빈발 부분그래프 집합 $F=\{f_1, f_2, f_3, \ldots, f_n\}$을 추출할 수 있다. 그래프 마이닝부(20)는 최소 지지도(minimum support)를 이용하여 빈발 부분 그래프 마이닝을 수행한다. 빈발 부분 그래프 마이닝에서 최소 지지도란 부분 그래프가 그래프 DB(10) 내에서 최소한으로 발생해야 하는 빈발도(frequency)를 의미한다.

[0028] 만약, 최소 지지도를 낮게 설정한다면, 빈발 부분 그래프를 가지고 있는 그래프 인스턴스의 수가 적은 경우에도, 최소 지지도가 만족되어, 대부분 낮은 분류 성능을 가지고 있는 부분 그래프들이 추출될 수 있다. 또한, 연산량이 증가하여 수행시간이 급격히 증가하게 된다. 반면, 최소 지지도를 높게 설정한다면, 상이한 클래스에 속하는 대부분의 그래프 인스턴스들이 빈발 부분 그래프를 가지게 되므로, 유용한 분류 성능을 갖는 빈발 부분 그래프들을 추출하지 못할 수 있다. 종래에 최소 지지도와 분류 성능과의 관계를 실험을 통하여 확인한 결과에 따르면, 데이터 셋의 특성에 따라 차이는 발생하지만 일반적으로 최소 지지도가 10~15%사이의 값을 가질 때 유용한 분류 성능을 갖는 빈발 부분 그래프 집합을 얻을 수 있었음 알려져 있다.

[0029] 동시 발생 빈발 부분 그래프 특징 선택부(40)는 빈발 부분 그래프 저장부(30)에서 동시 발생을 고려한 특징을 선택하여 유용한 특징을 선택해낸다. 이때, 동시 발생 빈발 부분 그래프 특징의 선택은 그리디 방법을 사용한다. 이에 대한 상세한 설명은 후술하는 동시 발생 빈발 부분 그래프 특징 선택 방법에서 상세히 설명하기로 한다.

[0030] DB 분할부(50)는 동시 발생 빈발 부분 그래프 특징 선택부(40)에 의해 선택된 동시 발생 빈발 부분 그래프 특징의 포함 여부에 따라 그래프 DB(10)를 분할한다. 그리고, 분할된 DB가 소정 개수 이하가 아닐 경우에는 분할된 DB에서 검색을 수행하고, 그렇지 않으면 검색을 중단한다.

[0031] 도 4는 본 발명의 바람직한 실시 예에 따른 동시 발생 빈발 부분그래프를 이용한 그래프 분류 방법을 설명하기 위한 순서도이다.

[0032] 도 4에 도시된 동시 발생 빈발 부분그래프를 이용한 그래프 분류 방법은 다음의 <표 1>와 같은 알고리즘 1로 구현될 수 있다. <표 1>의 알고리즘 1은 COFSF를 이용한 모델 기반 탐색 트리의 알고리즘 COFSF_based_SearchTree이다. 알고리즘 1은 그래프 DB와 최소 지지도 s, 노드 내의 최소 그래프 개수 m을 입력받아, 그래프 분류 트리 모델 T를 반환한다.

[0033] <표 1>

Algorithm COFSF_based_SearchTree(D, s, m)
Input: graph DB $D=\{G_0, G_1, \cdots, G_n\}$ minimum support s minimum node size m
Output: model-based search tree T
1: $F=fp(D,s)$; //frequent subgraph mining
2: Create a feature-graph matrix M from D and F
3: $F^*=COFS(F,M)$; //feature selection
4: Maintain patterns F^* as the testing features in current node of the tree T;
5: D_l= subset of graphs in D containing F^*;
6: $D_r=D-D_l$
7: **FOR** each P in $\{l,r\}$
8: **IF** $\lvert D_p\rvert \le m$ **OR** grpahs in D_l have the same class label, make a leaf node
9: **ELSE**
10: recursively construct T_p with D_p and s;
11:**RETURN** T;

[0034]

[0035] 도 4 및 <표 1>의 알고리즘 1을 참조하면, 그래프 마이닝부(20)는 110 단계에서 빈발 부분그래프 마이닝 $fp(D,s)$ 을 수행하여 전체 빈발 부분그래프 집합 F를 마이닝한다.(<표 1>의 알고리즘의 line 1). 그리고, 각 그래프(G_i)에 대하여 특징(f_j)의 포함 여부를 나타내는 특징-그래프 행렬 M을 하기의 <수학식 1>과 같이 생성한다(<표 1>의 알고리즘의 line 2).

[0036] <수학식 1>

$$M_{i,j}=\begin{cases}1 & if\ f_j \subseteq G_i \\ 0 & otherwise\end{cases}$$

[0037]

[0038] 상기 <수학식 1>에서 M의 (i, j) 엔트리는 i번째 그래프가 j번째 빈발 부분그래프를 포함하고 있음을 의미한다.

[0039] 도 5는 그래프 DB(10)가 $D=\{G_1, G_2, G_3\}$ 이고, 빈발 부분그래프 집합이 $F=\{f_1, f_2, f_3\}$ 일 때, 특징-그래프 행렬을 구성하는 예제이다. 이 행렬로부터 G_1 은 f_1 과 f_2 를 가지고 있고, G_2 는 f_2 와 f_3 를 갖는 정보를 쉽게 찾을 수 있다.

[0040] 그리고, 동시 발생 빈발 부분 그래프 특징 선택부(40)는 120 단계에서 동시발생 특징 선택 함수 COFS(F, M)을 이용하여 가장 변별력이 높은 동시 발생 빈발 부분그래프를 F^* 를 찾는다(<표 1> 알고리즘 1의 line 3).

[0041] 도 6은 본 발명의 바람직한 실시 예에 따른 동시 발생 빈발 부분 그래프 특징 선택 방법을 설명하기 위한 순서

도이다.

[0042] 도 6에 도시된 동시 발생 빈발 부분 그래프 특징 선택 방법은 하기의 <표 2>의 알고리즘 2와 같이 구현될 수 있다.

[0043] <표 2>

```
Algorithm COFS(F, M)
Input: frequent subgraph set F={f1, f2, ..., fn}
Output: selected COFSF F*

1: F*={∅}, MaxIG=0;
2: REPEAT
3:   BestF = null;
4:   FOR each f (f∈F and f∉F*)
5:        F' = F* ∪ {f};
6:        IG = infoGain(F', M);
7:        IF IG > MaxIG THEN
8:             BestF = f;
9:             MaxIG = IG;
10:  IF BestF !=null THEN F*=F* ∪{BestF}
11:UNTIL BestF ==null OR F*==F
12:RETURN F*;
```

[0044]

[0045] 도 6 및 <표 2>의 알고리즘을 참조하면, 동시 발생 빈발 부분 그래프 특징 선택부(40)는 121 단계에서 전체 빈발 부분 그래프 특징 집합에서 가장 변별력이 높은 베스트 빈발 부분 그래프 특징을 선택한다.

[0046] 본 발명에서는 분류 문제에서 특징의 변별력 평가를 위하여 하기의 <수학식 2>의 인포메이션 게인 평가 함수를 사용한다.

[0047] <수학식 2>

$$IG(C|X) = -\sum_{c\in\{0,1,\cdots,k\}} P(c)\log P(c) + \sum_{x\in\{0,1,\cdots,k\}} P(x) \sum_{c\in\{0,1,\cdots,k\}} P(c|x)\log P(c|x)$$

[0048]

[0049] 클래스 $C=\{c_0, c_1, \cdots, c_k\}$ 분류 문제에서 특징 f의 인포메이션 게인을 가정할 때, x는 부분그래프 f를 가지고 있는 그래프이고, P(x)는 그래프들이 f를 가지고 있을 확률이다. P(c)는 전체 그래프들이 각 클래스 c_i에 속할 확률이고, $P(c|x)$는 f를 가지고 있는 그래프가 클래스 c_i에 속할 확률이다.

[0050] 다시 알고리즘 2를 이용하여 살펴보면, 우선 인포메이션 게인 평가 함수 infoGain()을 사용하여 특징 f_i 중에 가장 높은 변별력을 갖는 BestF를 찾아 F^*에 넣는다.

[0051] 그리고, 동시 발생 빈발 부분 그래프 특징 선택부(40)는 122 단계에서 베스트 빈발 부분 그래프 특징과 다른 빈발 부분 그래프 특징과의 동시 발생 특징의 변별력을 계산한다. 즉, F^*와 F^*에 선택되지 않은 각 특징 f_i ($f_i\in F$ and $f_i\notin F^*$)의 동시 발생 특징의 변별력을 계산한다.

[0052] 그리고, 123 단계에서 이들 중에 가장 높은 동시 발생 특징의 변별력이 기존 F^* 의 변별력보다 높은지를 판단한다.

[0053] 123 단계의 판단 결과 가장 높은 동시 발생 특징의 변별력이 기존 F^* 의 변별력보다 클 경우, 동시 발생 빈발 부분 그래프 특징 선택부(40)는 124 단계에서 동시 발생 빈발 부분 그래프를 베스트 빈발 부분 그래프 특징에 추가한다.

[0054] 그러나, 123 단계의 판단 결과, 만약 변별력이 더 높지 않으면 동시 발생 빈발 부분 그래프 특징 선택부(40)는 특징 선택 알고리즘을 종료하고, F^* 를 반환한다. 전술한 121 단계 내지 124 단계의 과정은 전체 특징 집합과 선택 부분 집합이 같아질 때($F^* = F$)까지 반복(line 2-11)된다.

[0055] 다시 도 4를 참조하면, 130 단계에서 DB 분할부(50)는 그래프 DB(10)를 F^* 를 기준으로(알고리즘 1의 line 4) F^* 를 갖는 그래프들로 구성된 데이터베이스(D_i)(알고리즘 1의 line 5)과 갖지 않은 그래프들로 구성된 데이터베이스(D_r)(알고리즘 1의 line 6)로 분할한다.

[0056] 그리고, 140 단계에서 분할된 데이터베이스, 즉, D_i 또는 D_r 을 구성하는 분할 그래프들의 개수가 최소 노드 개수 m 이하의 클래스 레이블을 가지는지 판단한다.

[0057] 상기 140 단계의 판단 결과 분할 그래프들의 개수가 최소 노드 개수 m보다 이하일 경우, DB 분할부(50)는 150 단계에서 분할을 중지한다(알고리즘 1의 line 7-8).

[0058] 그러나, 상기 140 단계의 판단 결과 분할 그래프들의 개수가 최소 노드 개수 m보다 이하가 아닐 경우, 분할 그래프들(D_i 및 D_r)로 110 단계 내지 140 단계를 반복적으로 수행되도록 한다(알고리즘 1의 line 10). 그리고, DB 분할부(50)는 최종적으로 그래프 분류 트리 모델 T를 반환한다.

[0059] 그런데, 가장 변별력이 높은 COFSF를 기준으로 분할된 부분 DB D_i (COFSF 포함)과 (COFSF 미포함) D_r 중에 D_i 은 모든 그래프가 COFSF를 포함하기 때문에 구조적으로 유사할 확률이 크다. 따라서, D_i 내의 그래프들은 많은 수의 부분그래프들을 함께 가지고 있다.

[0060] 도 7은 MCF-7 데이터를 최소 지지도 25%를 사용하여 모델 기반 탐색 트리를 구축할 때, 트리의 깊이와 각 트리의 깊이에서 분할된 부분 DB들에서 높은 지지도(70% 이상)로 발생하는 부분그래프들의 개수의 관계를 보여주는 그래프이다.

[0061] 높은 지지도의 부분그래프들의 개수는 트리의 깊이에 따라 점차 증가하다가 트리의 깊이 4와 같은 특정 지점에서 급격히 감소되는 경향이 있다. 이와 같은 급격한 변화는 높은 지지도를 갖는 부분그래프를 많이 포함하고 있는 D_i 의 분할이 끝난 깊이에서 발생한다.

[0062] 본 발명에서는 높은 지지도로 발생하는 특징들은 분류에 도움이 되지 않는다는 성질을 활용하여, 이와 같은 특징들을 제거하고 COFSF의 변별력을 측정한다. 모델 기반 탐색 트리에서는 마이닝된 빈발 부분그래프 개수가 매우 많은 노드가 있기 때문에, 특징 선택 시 많은 연산량을 줄이는 효과가 있다.

[0063] 실험을 통하여 제안하는 기법이 기존의 개별 특징 선택 기법보다 더 높은 그래프 분류 성능을 보이는 것을 입증하였다.

[0064] <실험 환경 및 실험 데이터>

[0065] 실험을 위하여 암에 대한 효과의 유무 정보가 있는 화합물들로 구성된 MCF-7(유방암), OVCAR(난소암), SW-620(대장암) 데이터 셋을 사용하였다. 각 데이터 셋들에 대하여 암에 효과가 있는 클래스 데이터 100개와 효과가 없는 클래스 데이터 400개를 샘플링하여 훈련용 데이터 셋으로 사용하여 기존의 모델 기반 탐색 트리와 동시 발생 특징 기반 그래프 분류 트리를 각각 구축하였다. 그리고 각 2000개의 테스트 데이터를 이용하여 분류 성능을 평가 및 비교하였다.

[0066] 특징의 변별력 특정은 인포메이션 게인을 사용하였고 빈발 부분그래프는 gSpan 알고리즘을 사용하였다.

[0067] <실험 결과>

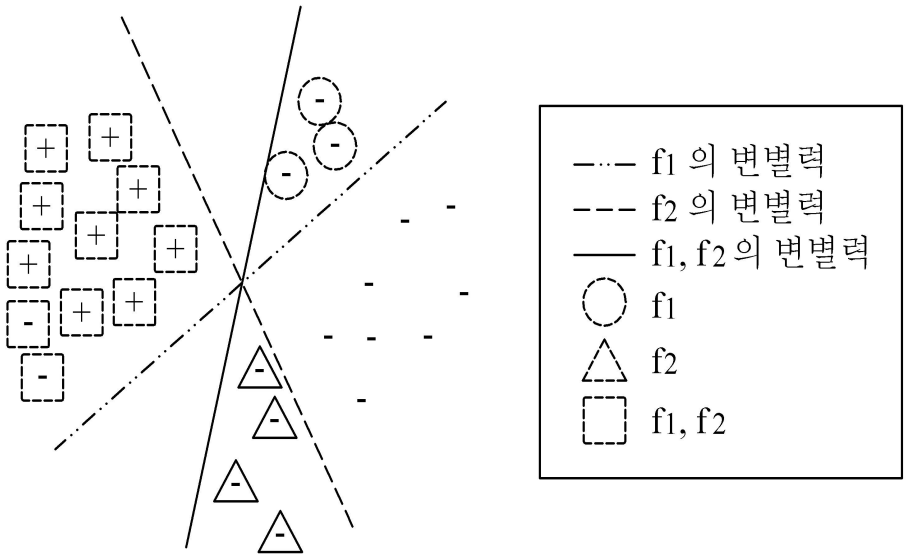
[0068] 도 8은 25%, 30%, 35% 최소지지도 별 그래프 분류의 정확도 실험 결과를 나타낸다. M^bT는 기존의 개별 특징 변별력을 이용한 모델 기반 탐색트리이고, COFS는 제안하는 COFSF를 이용한 모델 기반 탐색 트리이다. 제안하는 그래프 분류 방법이 대부분의 경우에서 더 좋은 분류 정확도를 갖는 것을 알 수 있다.

[0069] 도 9는 제안하는 COFSF를 이용한 모델기반 탐색트리의 구축 시간이 기존의 모델기반 탐색트리보다 더 빠른 것을 보여준다. 제안하는 기법의 가장 높은 변별력의 COFSF 선택은 기존 M^bT의 개별 특징들의 변별력만을 고려한 특징 선택보다 더 많은 시간을 요구한다. 그러나 이와 같은 특징 선택에 관련된 시간은 모델기반 탐색트리의 모든 노드마다 수행되는 빈발 부분그래프 마이닝 시간에 비교하면 그 비중이 작다. 제안하는 기법이 더 적은 수행시간이 소요된 것은 모델기반 탐색 트리의 각 노드에 공통된 부분그래프를 공유하는 그래프의 개수가 기존의 방법보다 적기 때문이다. M^bT에는 공통된 부분그래프들을 많이 존재하는 노드들(주로)이 많다. 일반적으로 그래프 DB에 공통된 부분그래프들이 많이 있는 경우 빈발 부분그래프 마이닝 알고리즘의 수행시간이 많이 소요된다.

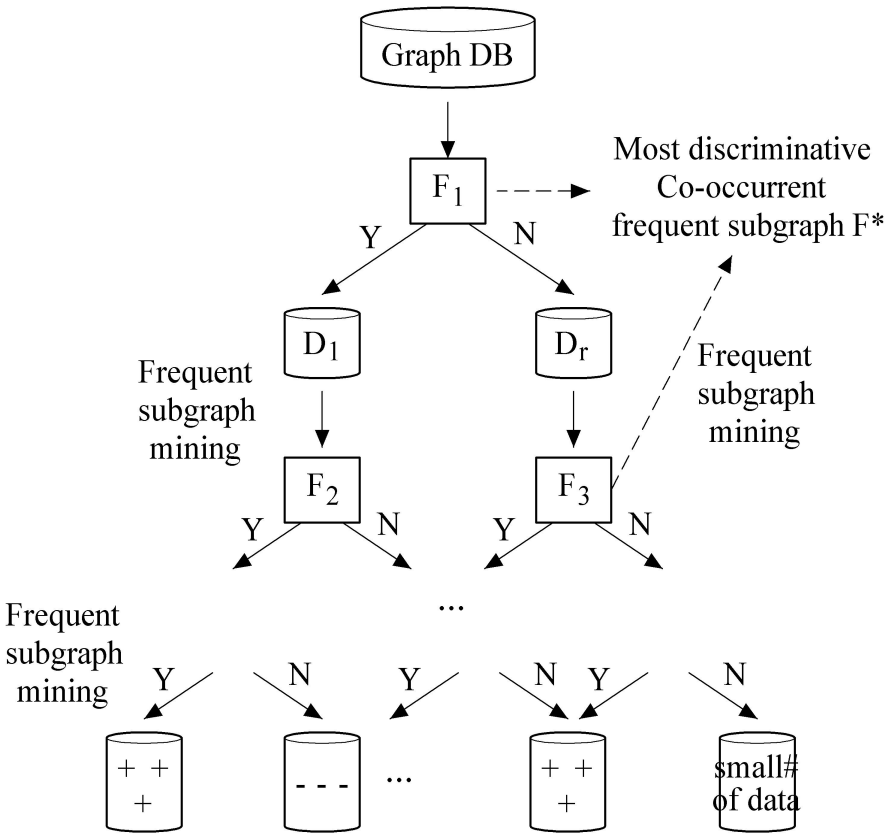
[0070] 본 발명은 최근 세계적으로 커다란 이슈가 되고 있는 소셜 네트워크 분석에 활용이 가능하다. SK 컴즈, NHN와 같은 국내의 대표 포털 업체에서도 미투데이 등과 같은 소셜 네트워크 서비스 제공을 시작했다. 이 외에도 많은 제약 업체의 연구소에서 화합물 분석 및 분류에 활용할 수 있다.

도면

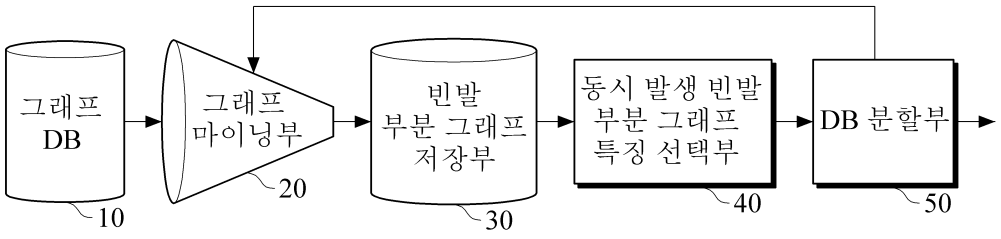
도면1



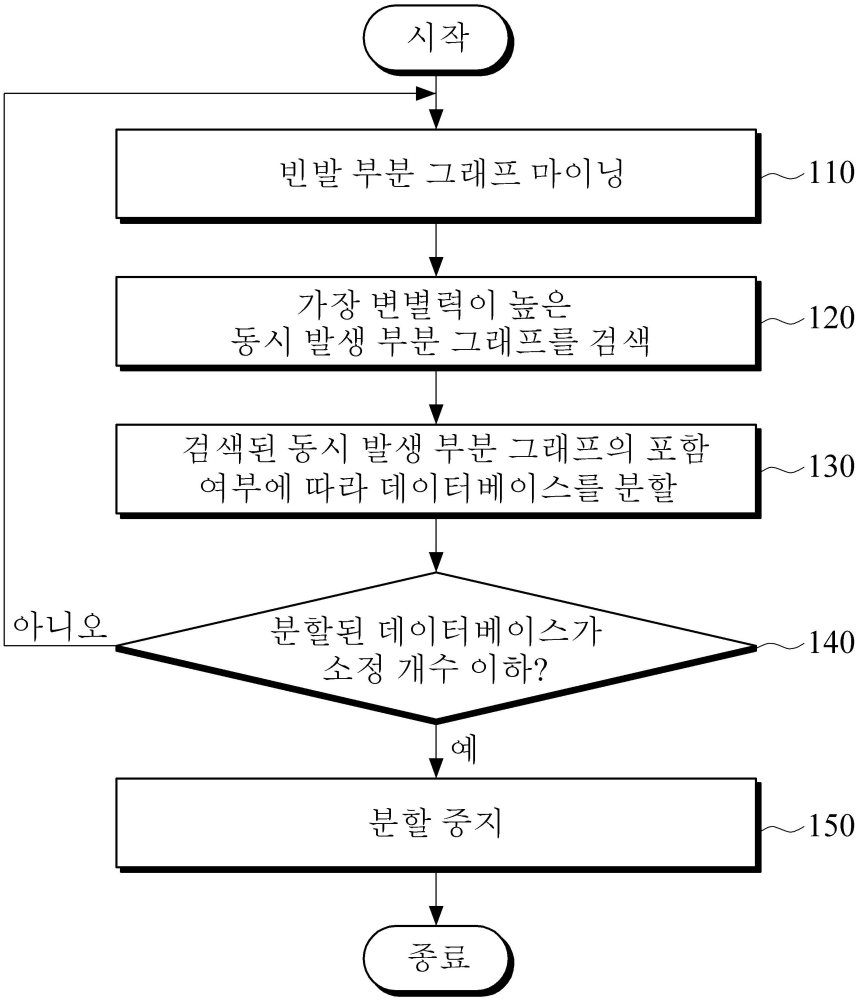
도면2


Graph DB
F1
Most discriminative
Co-occurrent
frequent subgraph F*
Y
N
D1
Dr
Frequent
subgraph
mining
Frequent
subgraph
mining
F2
F3
Y
N
Y
N
...
Frequent
subgraph
mining
Y
N
Y
N
Y
N
+ +
+
- - -
...
+ +
+
small#
of data

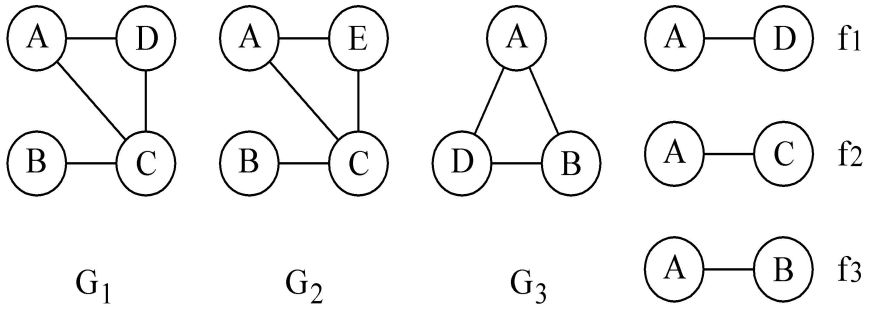
도면3


그래프
DB
10
그래프
마이닝부
20
빈발
부분 그래프
저장부
30
동시 발생 빈발
부분 그래프
특징 선택부
40
DB 분할부
50

도면4

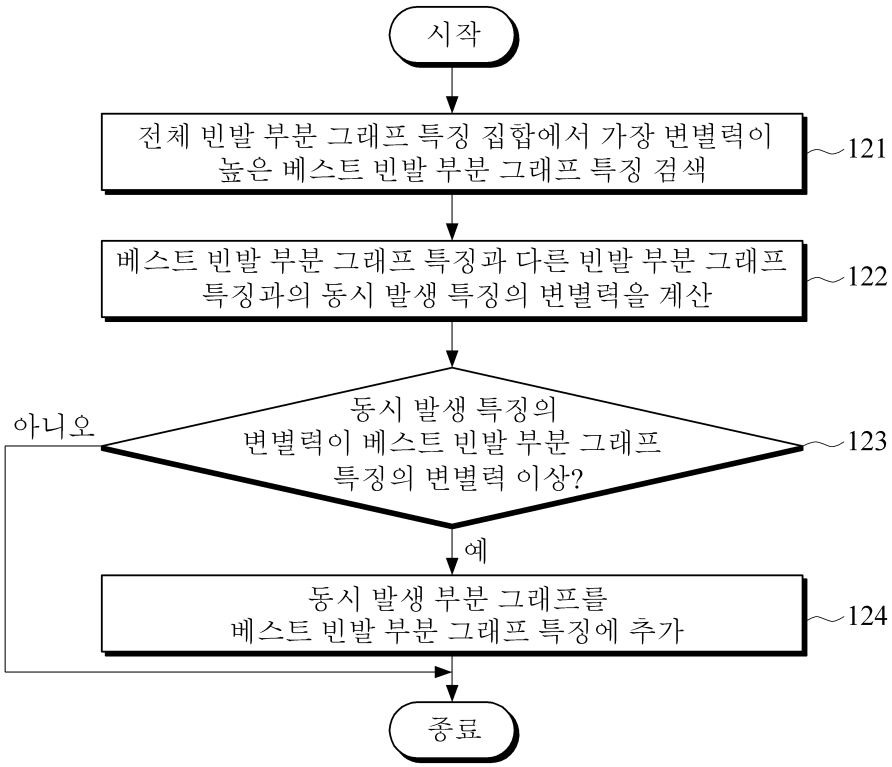

시작
빈발 부분 그래프 마이닝
110
가장 변별력이 높은
동시 발생 부분 그래프를 검색
120
검색된 동시 발생 부분 그래프의 포함
여부에 따라 데이터베이스를 분할
130
분할된 데이터베이스가
소정 개수 이하?
140
아니오
예
분할 중지
150
종료

도면5

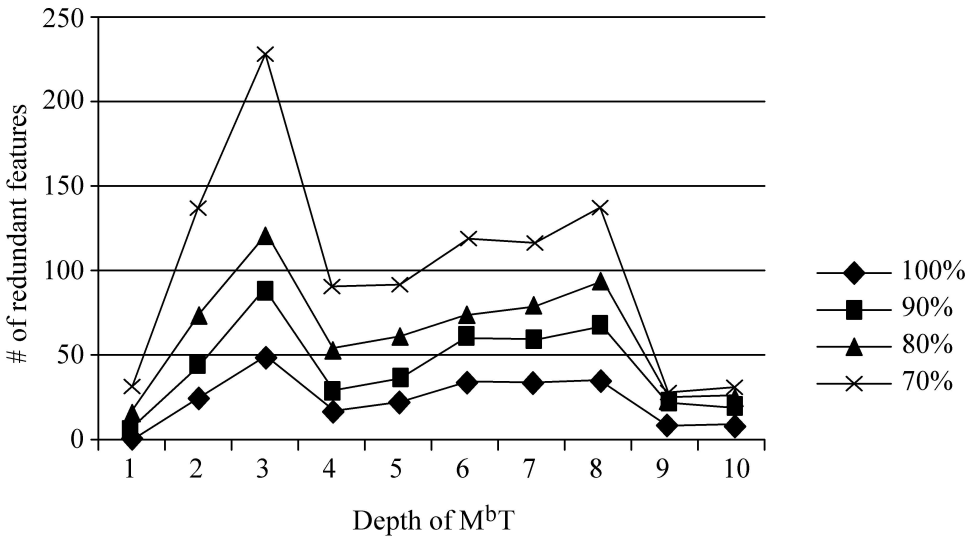


	f1	f2	f3
G_1	1	1	0
G_2	0	1	1
G_3	1	0	1

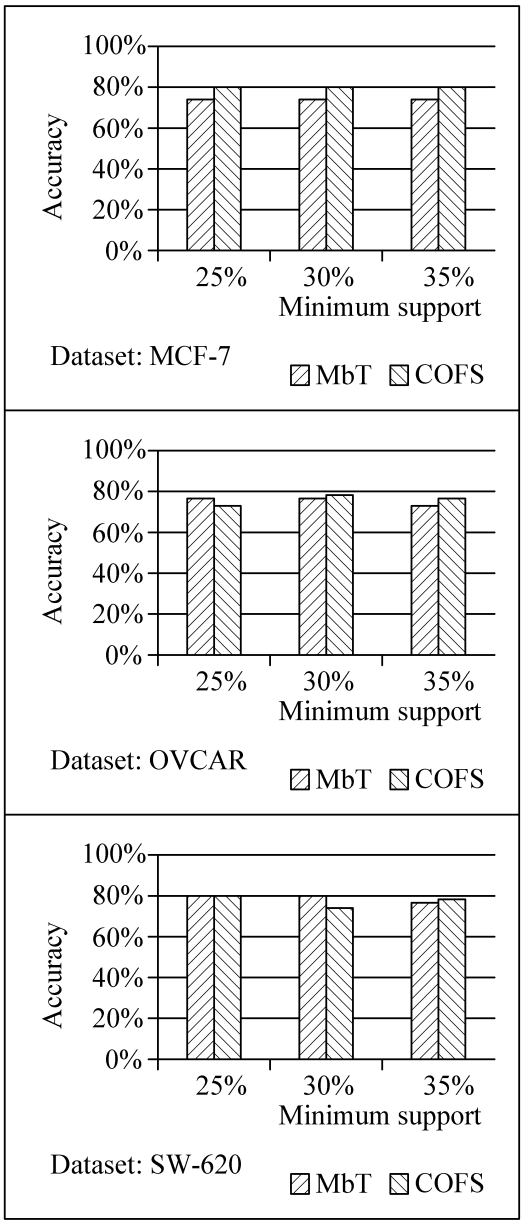
도면6



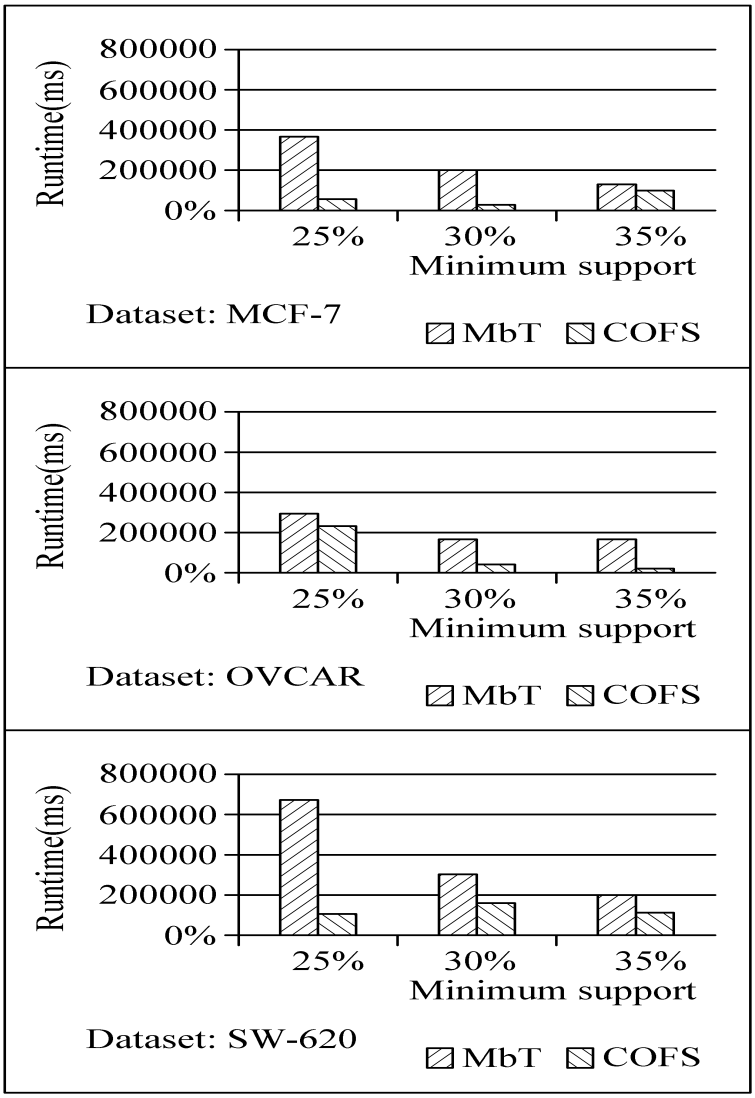
도면7


250
200
150
100
50
0
# of redundant features
1
2
3
4
5
6
7
8
9
10
Depth of MbT
100%
90%
80%
70%

도면8


100%
80%
60%
40%
20%
0%
Accuracy
25%
30%
35%
Minimum support
Dataset: MCF-7
MbT
COFS
100%
80%
60%
40%
20%
0%
Accuracy
25%
30%
35%
Minimum support
Dataset: OVCAR
MbT
COFS
100%
80%
60%
40%
20%
0%
Accuracy
25%
30%
35%
Minimum support
Dataset: SW-620
MbT
COFS

도면9


800000
600000
400000
200000
0%
Runtime(ms)
25%
30%
35%
Minimum support
Dataset: MCF-7
MbT
COFS
800000
600000
400000
200000
0%
Runtime(ms)
25%
30%
35%
Minimum support
Dataset: OVCAR
MbT
COFS
800000
600000
400000
200000
0%
Runtime(ms)
25%
30%
35%
Minimum support
Dataset: SW-620
MbT
COFS