(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2012-0095716
(43) 공개일자 2012년08월29일

(51) 국제특허분류(Int. Cl.)
G06F 17/30 (2006.01)
(21) 출원번호 10-2011-0015206
(22) 출원일자 2011년02월21일
심사청구일자 2011년02월21일

(71) 출원인
경희대학교 산학협력단
경기도 용인시 기흥구 덕영대로 1732, 국제캠퍼스내 (서천동, 경희대학교)
(72) 발명자
이영구
경기도 수원시 영통구 덕영대로 1709, 706호 (영통동, 경희유니빌)
한용구
경기도 용인시 기흥구 덕영대로 1732, 전자정보관 349호 (서천동, 경희대학교)
박기성
경기도 용인시 기흥구 덕영대로 1732, 전자정보관 349호 (서천동, 경희대학교)
(74) 대리인
특허법인 신지

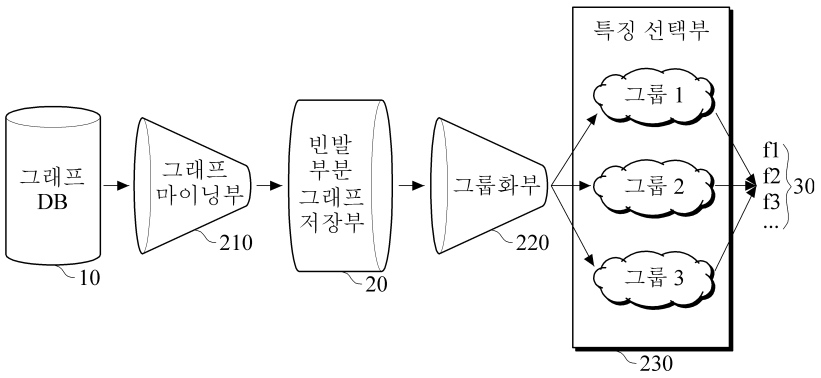
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 **그래프 분류를 위한 유사한 그래프 구조를 이용한 특징 선택 방법 및 장치**

(57) 요 약

유사한 구조의 부분 그래프들은 유사한 분류 성능을 갖는 성질을 이용하여 빈발 부분 그래프간의 구조적 유사성을 이용하여 빈발 부분 그래프들을 그룹화하고, 그룹별로 대표적인 특징을 선택함으로써 특정 클래스들에 편향적이지 않은 특징 서브셋을 구성하는 방법 및 장치를 제공한다. 그래프 분류를 위한 특징 선택 방법은, 복수의 빈발 부분 그래프들의 인포메이션 게인 및 상기 빈발 부분 그래프들의 서로 간의 구조적 유사성을 이용하여 상기 빈발 부분 그래프들을 그룹화하는 단계와, 상기 그룹화된 각 빈발 부분 그래프 그룹별로, 각 빈발 부분 그래프 그룹 내에서 분류 성능이 가장 높은 빈발 부분 그래프를 특징으로 각각 선택하는 단계를 포함한다.

대 표 도 - 도2



이 발명을 지원한 국가연구개발사업

과제고유번호	2010-0013689
부처명	교육과학기술부
연구사업명	일반연구자 지원사업
연구과제명	그래프 마이닝을 위한 부분감독기반 특징 선택 기술 연구
주관기관	경희대학교 산학협력단
연구기간	2010.05.01 ~ 2013.04.30

특허청구의 범위

청구항 1

복수의 빈발 부분 그래프들의 인포메이션 게인 및 상기 빈발 부분 그래프들의 서로 간의 구조적 유사성을 이용하여 상기 빈발 부분 그래프들을 그룹화하는 단계; 및

상기 그룹화된 각 빈발 부분 그래프 그룹별로, 상기 각 빈발 부분 그래프 그룹 내에서 분류 성능이 가장 높은 빈발 부분 그래프를 특징으로 각각 선택하는 단계; 를 포함하는 것을 특징으로 하는 그래프 분류를 위한 특징 선택 방법.

청구항 2

제1항에 있어서,

상기 빈발 부분 그래프들을 그룹화하는 단계는,

상기 빈발 부분 그래프들을 특징들을 인포메이션 게인 값을 기준으로 정렬하는 단계; 및

상기 정렬된 빈발 부분 그래프들 중에서 상기 가장 큰 인포메이션 게인 값을 가지는 빈발 부분 그래프를 기준으로 하여 상기 정렬된 빈발 부분 그래프들에 대하여 유사한 구조를 가지는 빈발 부분 그래프들을 그룹화하는 단계; 를 포함하는 것을 특징으로 하는 그래프 분류를 위한 특징 선택 방법.

청구항 3

제2항에 있어서,

상기 정렬된 빈발 부분 그래프들에 대하여 유사한 구조를 가지는 빈발 부분 그래프들을 그룹화하는 단계에서,

상기 정렬된 빈발 부분 그래프들에 대하여, 동일한 그룹에 속하는 빈발 부분 그래프들의 유사도의 기준이 되는 유사도 임계값(β)을 이용하여 상기 정렬된 빈발 부분 그래프들을 그룹화하는 것을 특징으로 하는 그래프 분류를 위한 특징 선택 방법.

청구항 4

제3항에 있어서,

상기 정렬된 빈발 부분 그래프들에 대하여 유사한 구조를 가지는 빈발 부분 그래프들을 그룹화하는 단계는,

상기 정렬된 빈발 부분 그래프들 중에서 상기 그룹화된 빈발 부분 그래프를 제거하는 단계를 더 포함하고,

상기 그룹화하는 단계는, 상기 정렬된 빈발 부분 그래프들에 대한 그룹화가 종료될 때까지 반복되는 것을 특징으로 하는 그래프 분류를 위한 특징 선택 방법.

청구항 5

제1항에 있어서,

상기 분류 성능이 가장 높은 빈발 부분 그래프를 특징으로 각각 선택하는 단계에서,

그룹화된 각 빈발 부분 그래프 그룹에 대하여, 상기 각 빈발 부분 그래프 그룹에 속하는 빈발 부분 그래프에 대한 인포메이션 게인을 계산하는 단계; 및

상기 각 빈발 부분 그래프 그룹별로, 상기 각 빈발 부분 그래프 그룹에서 가장 높은 인포메이션 게인을 생성한 빈발 부분 그래프를 상기 특징으로 선택하는 단계를 포함하는 것을 특징으로 하는 그래프 분류를 위한 특징 선택 방법.

청구항 6

제1항에 있어서,

분류 대상이 되는 그래프들에 대한 정보를 포함하는 그래프 데이터베이스에 저장된 그래프들에 대하여 그래프

마이닝을 수행하여 상기 복수의 빈발 부분 그래프들을 추출하는 단계를 더 포함하는 것을 특징으로 하는 그래프 분류를 위한 특징 선택 방법.

청구항 7

복수의 빈발 부분 그래프들의 인포메이션 게인 및 상기 빈발 부분 그래프들의 서로 간의 구조적 유사성을 이용하여 상기 빈발 부분 그래프들을 그룹화하는 그룹화부; 및

상기 그룹화된 각 빈발 부분 그래프 그룹별로, 상기 각 빈발 부분 그래프 그룹 내에서 분류 성능이 가장 높은 빈발 부분 그래프를 특징으로 각각 선택하는 특징 선택부; 를 포함하는 것을 특징으로 하는 그래프 분류를 위한 특징 선택 장치.

청구항 8

제7항에 있어서,

상기 그룹화부는, 상기 빈발 부분 그래프들을 인포메이션 게인 값을 기준으로 정렬하고, 상기 정렬된 빈발 부분 그래프들 중에서 상기 가장 큰 인포메이션 게인 값을 가지는 빈발 부분 그래프를 기준으로 하여 상기 정렬된 빈발 부분 그래프들에 대하여 유사한 구조를 가지는 빈발 부분 그래프들을 그룹화하는 것을 특징으로 하는 그래프 분류를 위한 특징 선택 장치.

청구항 9

제7항에 있어서,

상기 특징 선택부는, 그룹화된 각 빈발 부분 그래프 그룹에 대하여, 각 빈발 부분 그래프 그룹에 속하는 빈발 부분 그래프에 대한 인포메이션 게인을 계산하고, 상기 각 빈발 부분 그래프 그룹별로, 상기 각 빈발 부분 그래프 그룹에서 가장 높은 인포메이션 게인을 생성한 빈발 부분 그래프를 상기 특징으로 선택하는 것을 특징으로 하는 그래프 분류를 위한 특징 선택 장치.

청구항 10

제7항에 있어서,

분류 대상이 되는 그래프들에 대한 정보를 포함하는 그래프 데이터베이스에 저장된 그래프들에 대하여 그래프 마이닝을 수행하여 상기 복수의 빈발 부분 그래프들을 추출하는 그래프 마이닝부를 더 포함하는 것을 특징으로 하는 그래프 분류를 위한 특징 선택 장치.

명 세 서

기 술 분 야

[0001] 본 발명은 그래프 분류기 시스템(Graph Classification System)에서의 특징 선택(Feature selection) 방법에 관한 것으로서, 더욱 상세하게는 특징을 선택하는데 있어서, 분류에 효과적인 특징 집합을 구성하는 방법에 관한 것이다.

배 경 기 술

[0002] 그래프 분류는 화합물 분석, 소셜 네트워크 분석 등 다양한 응용 분야에서 활용되고 있다. 일반적으로, 그래프의 분류를 위하여 부분 그래프들을 빈발 부분 그래프들을 특징으로 사용하여 특징 집합으로 구성한다. 논문 [Discriminative frequent pattern analysis for effective classification]에서는 빈발한 패턴이 분류에 유용하다는 것을 보였다. 그러나, 빈발 패턴을 이용한 그래프 분류에서는 너무 많은 빈발 부분 그래프(frequent sub graph)가 존재하기 때문에 분류에 유용한 특징 서브셋(relevant feature subset)을 선택하는 단계가 추가적으로 요구된다.

[0003] 종래의 특징 선택 방식은 인포메이션 게인(information gain)과 같은 필터 기법을 이용한 방식으로서, 그래프 데이터베이스 내에서 빈발 부분 그래프를 추출하고, 추출한 빈발 부분 그래프들 중 인포메이션 게인과 같은 평가 함수의 값이 높은 특징 서브셋을 선택하는 방식이었다. 그러나, 이와 같은 특징 선택 방법은 각 특징들의 분류 성능만을 평가할 뿐 어떤 클래스에 대한 분류 능력을 가지고 있는지를 고려하지 않는다. 따라서, 선

택된 특징 서브 셋이 특정 클래스들에만 한정된 분류 성능을 가져 분류의 성능을 저하시킨다. 예를 들어, C1, C2, C3 세 개의 클래스 분류 문제에서 20개의 특징들 중에 대표적인 필터 기반 기법인 인포메이션 게인을 이용하여 상위 5개의 특징을 선택했다고 가정하자. 이와 같은 필터 기반의 방법에서는 최악의 경우 5 개의 선택된 특징이 모두 클래스 C1과 C2에 대한 높은 분류 능력을 가지고, 클래스 C1과 C3 또는 C2와 C3에 대한 분류 능력이 없을 수 있다.

[0004] 이와 같은 편향적인 특징선택의 문제점을 해결하기 위하여 논문 [Frequent substructure-based approaches for classifying chemical compounds]에서는 특징을 선택할 때 필터 방식을 이용하여 빈발 부분 그래프 구조 중 분류 성능이 제일 좋은 특징을 선택하고, 그 특징을 보유하고 있는 그래프들을 제거한다. 이 과정을 반복함으로써 그래프 데이터베이스 내의 모든 그래프들을 커버할 수 있는 편향적이지 않은 특징 서브셋을 선택할 수 있다. 그러나, 이와 같은 방식의 특징 선택은 인포메이션 게인이 가장 높은 부분 그래프(sub graph) 만의 포함 여부에 따라서 DB를 분할하기 때문에, 같은 클래스에 속하지만 유사한 서브그래프를 포함하는 그래프들이 같은 분할에 속하지 않게 될 수 있다.

발명의 내용

해결하려는 과제

[0005] 본 발명은 그래프 분류를 위한 특징을 선택할 때, 유사한 구조의 부분 그래프들은 유사한 분류 성능을 갖는 성질을 이용하여 빈발 부분 그래프 간의 구조적 유사성을 이용하여 빈발 부분 그래프들을 그룹화하고, 그룹별로 대표적인 특징을 선택함으로써 특정 클래스들에 편향적이지 않은 특징 서브셋을 구성하는 방법 및 장치를 제공하여, 유사한 부분 그래프들을 포함하는 그래프들이 같은 분할된 그룹에 속하도록 해 주어 동일 클래스에 속하는 데이터들이 쪼개지지 않도록 함으로써 분류 시스템의 성능을 높이는데 그 목적이 있다.

과제의 해결 수단

[0006] 일 측면에 따른 그래프 분류를 위한 특징 선택 방법은, 복수의 빈발 부분 그래프들의 인포메이션 게인 및 빈발 부분 그래프들의 서로 간의 구조적 유사성을 이용하여 빈발 부분 그래프들을 그룹화하는 단계와, 그룹화된 각 빈발 부분 그래프 그룹별로, 각 빈발 부분 그래프 그룹 내에서 분류 성능이 가장 높은 빈발 부분 그래프를 특징(feature)으로 각각 선택하는 단계를 포함한다.

[0007] 다른 측면에 따른 그래프 분류를 위한 특징 선택 장치는, 복수의 빈발 부분 그래프들의 인포메이션 게인 및 빈발 부분 그래프들의 서로 간의 구조적 유사성을 이용하여 빈발 부분 그래프들을 그룹화하는 그룹화부와, 그룹화된 각 빈발 부분 그래프 그룹별로, 각 빈발 부분 그래프 그룹 내에서 분류 성능이 가장 높은 빈발 부분 그래프를 특징(feature)으로 각각 선택하는 특징 선택부를 포함한다.

발명의 효과

[0008] 본 발명에 따르면 유사한 부분 그래프들을 포함하는 그래프들이 같은 분할 그룹에 속하도록 해주어 동일 클래스에 속하는 데이터들이 쪼개지지 않도록 하여 그래프 분류 시스템의 성능을 높일 수 있다.

[0009] 본 발명에서 제공하는 빈발 부분 그래프를 이용한 특징 선택은 그래프의 구조적 유사성을 이용한 그룹화 방법을 통하여 모든 클래스에 대하여 높은 분류 성능을 가지고 있는 특징 서브셋으로 선택함으로써 분류의 정확성을 높이고 특징 서브셋이 편향적으로 구성되는 것을 방지하고 분류의 수행 속도를 향상시킬 수 있다.

[0010] 또한, 구조적 유사도 임계값의 조정을 통하여 데이터의 특성에 맞게 그룹화를 할 수 있으며 이에 따라 그룹별 특징 선택에 있어서 최적화된 속도와 정확도를 얻을 수 있다.

도면의 간단한 설명

[0011] 도 1은 본 발명의 일 실시예에 따른 유사한 그래프 구조를 이용한 특징 선택 방법을 이용하는 특징 선택 방법을 나타내는 순서도이다.

도 2는 본 발명의 일 실시예에 따른 유사한 그래프 구조를 이용한 특징 선택 방법을 수행하는 특징 선택 장치의 구성을 나타내는 도면이다.

도 3은 본 발명의 일 실시예에 따른 구조적으로 유사한 빈발 부분 그래프의 그룹화 과정을 나타내는 순서도이다.

도 4는 본 발명의 일 실시예에 따른 구조적으로 유사한 빈발 부분 그래프 그룹내의 빈발 부분 그래프의 특징들을 보여주는 도면이다.

도 5는 본 발명의 일 실시예에 따른 각 빈발 부분 그래프 그룹으로부터 특징을 선택하는 방법을 나타내는 순서도이다.

도 6은 본 발명의 일 실시예에 따른 그래프 데이터베이스를 테이블 형태로 나타내는 도면이다.

도 7은 본 발명의 일 실시예에 따른 빈발 부분 그래프 테이블을 나타내는 도면이다.

도 8은 본 발명의 일 실시예에 따른 구조적으로 유사한 빈발 부분 그래프 그룹 후보 테이블을 나타낸 도면이다.

도 9는 본 발명의 일 실시예에 따른 구조적으로 유사한 빈발 부분 그래프 그룹 테이블을 나타낸 도면이다.

도 10은 본 발명의 일 실시예에 따른 그룹 대표 특징 선택 테이블을 나타낸 도면이다.

발명을 실시하기 위한 구체적인 내용

[0012] 이하, 첨부된 도면을 참조하여 본 발명의 일 실시예를 상세하게 설명한다. 본 발명을 설명함에 있어 관련된 공지 기능 또는 구성에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략할 것이다. 또한, 후술되는 용어들은 본 발명에서의 기능을 고려하여 정의된 용어들로서 이는 사용자, 운용자의 의도 또는 관례 등에 따라 달라질 수 있다. 그러므로 그 정의는 본 명세서 전반에 걸친 내용을 토대로 내려져야 할 것이다.

[0013] 도 1은 본 발명의 일 실시예에 따른 유사한 그래프 구조를 이용한 특징 선택 방법을 이용하는 특징 선택 방법을 나타내는 순서도이다.

[0014] 빈발 부분 그래프들을 각 빈발 부분 그래프들의 인포메이션 게인(information gain) 및 빈발 부분 그래프들 서로 간의 구조적 유사성을 이용하여 그룹화한다(110). 빈발 부분 그래프는, 분류 대상이 되는 그래프들에서 빈번하게 발견되는 부분 그래프로, 그래프들에 대한 분류 기준으로 이용된다. 각 빈발 부분 그래프들의 인포메이션 게인은 인포메이션 게인 특징 평가 함수를 이용하여 계산될 수 있다.

[0015] 각 빈발 부분 그래프 그룹에서 분류 성능이 가장 높은 빈발 부분 그래프를 특징(feature)으로 선택한다(120). 동작 110에서 이용되는 빈발 부분 그래프를 그래프의 특징이라고 할 때, 동작 120에서 선택되는 각 빈발 부분 그래프 그룹에서 분류 성능이 가장 높은 빈발 부분 그래프는, 각 빈발 부분 그래프 그룹의 대표 특징이라고 할 수 있다.

[0016] 도 2는 본 발명의 일 실시예에 따른 유사한 그래프 구조를 이용한 특징 선택 방법을 수행하는 특징 선택 장치의 구성을 나타내는 도면이다.

[0017] 특징 선택 장치(200)는 그래프 데이터베이스(10), 그래프 마이닝부(210), 빈발 부분 그래프 정보 저장부(20), 그룹화부(220) 및 특징 선택부(230)를 포함할 수 있다. 특징 선택 장치(220)는 버텍스(vertex) 및 에지(edge)로 구성되는 그래프로 표현가능한 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.

[0018] 그래프 데이터베이스(10)는 그래프 분류 대상인 그래프들 및 그래프들의 정보를 포함한다. 그래프 데이터베이스(10)의 구성의 일 예는 도 6을 참조하여 후술한다. 그래프들의 정보는, 각 그래프가 속하는 클래스의 클래스 레이블 정보를 포함할 수 있다. 여기에서, 클래스는, 특징 선택 장치(200)가 적용되는 데이터 분석 대상이 되는 실제 데이터, 예를 들어, 화합물 데이터, 소셜 네트워크 분석용 데이터 등에 따라 다양한 형태를 가질 수 있다.

[0019] 그래프 마이닝부(210)는 그래프 데이터베이스(10)에 저장된 그래프들로부터 빈발 부분 그래프 집합을 추출한다. 빈발 부분 그래프 정보 저장부(20)는 그래프 마이닝부(210)에서 추출된 빈발 부분 그래프 집합을 저장한다.

[0020] 그래프 마이닝부(210)는 그래프 데이터베이스(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(210)는 최소 지지도

(minimum support)를 이용하여 빈발 부분 그래프 마이닝을 수행한다. 빈발 부분 그래프 마이닝에서 최소 지지도란 부분 그래프가 그래프 데이터베이스(10) 내에서 최소한으로 발생해야 하는 빈발도(frequency)를 의미한다.

[0021] 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.

[0022] 그룹화부(220)는, 빈발 부분 그래프들을 각 빈발 부분 그래프들의 인포메이션 게인(information gain) 및 빈발 부분 그래프들 서로 간의 구조적 유사성을 이용하여 그룹화하는 도 1의 동작(110)을 수행한다. 그룹화부(220)의 동작의 상세에 대해서는 도 3을 참조하여 후술한다.

[0023] 특징 선택부(230)는, 각 빈발 부분 그래프 그룹별로, 각 빈발 부분 그래프 그래프로부터 분류 성능이 가장 높은 빈발 부분 그래프를 특징(feature)으로 선택하는 도 1의 동작(120)을 수행한다. 특징 선택부(230)의 동작의 상세에 대해서는 도 5를 참조하여 후술한다. 도 3에서는, 예시적인 설명을 위하여, 그룹화부(220)의 빈발 부분 그래프의 그룹화 결과 3개의 그룹이 생성되어, 3개의 그룹 각각으로부터 특징이 선택되는 구성이 도시되어 있으나, 그룹화부(220)의 처리 결과 생성되는 빈발 부분 그래프 그룹의 개수에 제한되지 않는다.

[0024] 특징 서브셋(30)은 특징 선택부(230)의 처리 결과로서, 특징으로 선택된 빈발 부분 그래프의 집합을 나타낸다.

[0025] 도 3은 본 발명의 일 실시예에 따른 구조적으로 유사한 빈발 부분 그래프의 그룹화 과정을 나타내는 순서도이다.

[0026] 도 1의 그룹화부(220)는 그래프 마이닝부(210)에 의해 추출된 빈발 부분 그래프들을 일관성 있게 그룹화하기 위하여 모든 빈발 부분 그래프(f_i)와 모든 클래스들의 그래프들의 집합(Y)에 대한 인포메이션 게인을 계산한다(310).

[0027] 인포메이션 게인은 KLIC(Kullback-Leibler divergence)라고도 한다. 그룹화부(220)에서 이용되는 인포메이션 게인 함수(IG(Y,f_i))는 IG(Y,f_i)=H(Y)-H(Y|f_i)로 나타낼 수 있다. 여기에서, H(Y)는 모든 클래스의 그래프들의 집합(Y)에 대한 엔트로피(entropy)를 나타내고, H(Y|f_i)는 집합(Y) 및 모든 빈발 부분 그래프(f_i)에 대한 크로스 엔트로피(cross entropy)를 나타낸다.

[0028] 그룹화부(220)는 모든 빈발 부분 그래프들에 대해 계산된 각각의 인포메이션 게인 값을 기준으로, 빈발 부분 그래프들을 정렬하고, 구조적 유사한 빈발 부분 그래프 그룹 후보(F={f_1, f_2, f_3,...f_n})를 생성한다(320). 모든 빈발 부분 그래프들에 대해 계산된 각각의 인포메이션 게인 값을 기준으로 빈발 부분 그래프들을 정렬한 결과 생성되는 정보 테이블을 구조적으로 유사한 빈발 부분 그래프 그룹 후보 테이블이라고 한다. 이때, 그룹화부(220)는 인포메이션 게인 값을 기준으로 빈발 부분 그래프들을 정렬하기 때문에, 그래프 데이터베이스(10)의 저장된 순서와는 독립적으로 그룹화를 수행할 수 있다.

[0029] 그룹화부(220)는 구조적으로 유사한 빈발 그래프 그룹 후보 중에서 인포메이션 게인 값이 가장 높은 빈발 부분 그래프를 기준으로 빈발 부분 그래프 후보 내의 모든 빈발 부분 그래프와 구조적 유사도(Topological Similarity)를 측정할 수 있다(330). 이때, 구조적 유사도는 다양한 유사도 측정 방법을 사용할 수 있다. 구조적 유사도 측정 방법의 대표적인 예로 그래프 수정 거리(Graph edit distance)가 있다. 여기에서, 그래프 수정 거리는, 예를 들어, 두 그래프(g_1, g_2)에 대하여 그래프(g_1)를 그래프(g_2)로 변화시키기 위하여 얼마나 많은 수의 에지(edge)와 버텍스(vertex)를 삽입, 삭제 또는 치환 해야 하는지를 나타내는 요소이다. 두 그래프가 서로 같은 경우의 최소 수정 거리는 '0'이고, 두 그래프가 완전히 다를 경우의 최대 수정 거리는 그래프(g_1)의 모든 에지와 버텍스를 삭제한 후에 그래프(g_2)의 모든 에지와 버텍스를 삽입해야 하므로, {(그래프(g_1)의 총 에지 수 + 그래프(g_1)의 총 버텍스 수) + {(그래프(g_2)의 총 에지 수 + 그래프(g_2)의 총 버텍스 수)} 이다.

[0030] 따라서, 그래프 수정 거리를 사용할 경우에 그래프(g_1)와 그래프(g_2)의 구조적 유사도는

$$\left(1-\frac{edit\_distance(g_1,g_2)}{(V_{g1}+E_{g1})+(V_{g2}+E_{g2})}\right)\times 100$$

과 같이 정의할 수 있다. 여기서, edit_distance(g_1, g_2)는 그래프(g_1)과 그래프(g_2)의 그래프 수정 거리이고, V_i 와 E_i 는 각각 그래프 g_i 의 버텍스와 에지의 총 개수이다.

[0031] 그룹화부(220)는, 인포메이션 게인 값이 가장 높은 빈발 부분 그래프와 나머지 빈발 부분 그래프들 사이의 구조적 유사도가 미리 정의된 구조적 유사도 임계값(β)을 만족한다면 그룹화를 하여, 구조적으로 유사한 첫 번째 특징 그룹인 첫 번째 빈발 부분 그래프 그룹(G_1)을 생성할 수 있다(340). 구조적 유사도 임계값(β)은 동일한 그룹에 속하는 빈발 부분 그래프들의 유사도의 기준이 된다.

[0032] 또한, 그룹화부(220)는 인포메이션 게인 값이 가장 큰 빈발 부분 그래프(f_{max})를 기준으로 하여 우선적으로 그룹화함으로써, 구조적 유사도 측정을 위한 비교 연산을 상당히 감소시킬 수 있다. 여기에서는, 인포메이션 게인 값이 가장 큰 빈발 부분 그래프(f_{max})가 첫번째 빈발 부분 그래프(f_1)라고 가정한다.

[0033] 첫 번째 빈발 그래프 그룹(G_1)을 생성한 후에, 그룹화부(220)는 구조적 유사 특징 그룹 후보 집합(F)에서 그룹(G_1)으로 그룹화된 빈발 부분 그래프들을 제거하고(350), 구조적 유사 특징 그룹 후보 집합(F)에 남은 빈발 부분 그래프가 존재하는지 결정하고(360), 빈발 부분 그래프가 존재하면 다시 동작 330으로 진행하여, 다음 그룹(G_2)을 생성하는 과정을 반복한다. 다음 그룹(G_2)을 생성하는 과정에서는, 구조적 유사 특징 그룹 후보 집합(F)에서 그룹(G_1)으로 그룹화된 빈발 부분 그래프들을 제거하고 남은 빈발 부분 그래프들 중에서, 다시 인포메이션 게인 값이 가장 큰 빈발 부분 그래프를 결정하고, 결정된 빈발 부분 그래프를 기준으로, 나머지 빈발 부분 그래프들과의 유사도를 계산하는 방법으로 그룹화가 수행될 수 있다. 이와 같이 각 빈발 그래프 그룹을 생성하는 과정은, 구조적으로 유사한 빈발 부분 그래프 그룹 후보(F)내의 모든 빈발 부분 그래프들이 각 그룹으로 그룹화될 때까지 반복된다.

[0034] 그룹화부(220)는 같은 그룹 내의 모든 빈발 부분 그래프들 간의 구조적 유사도가 유사도 임계값(β)을 만족하도록 빈발 부분 그래프 그룹화를 수행할 수 있다.

[0035] 그룹화부(220)는, 이를 위하여, 첫 번째 단계에서는 인포메이션 게인 값이 가장 큰 빈발 부분 그래프(f_{max})와 빈발 부분 그래프 하나를 비교하여 유사도가 구조적 유사도 임계값(β) 이상이면 그룹(G_i)에 포함시킨다. 다음 단계부터는, 새로운 빈발 부분 그래프가 그룹(G_i)에 포함되기 위해서는 그룹내의 모든 부분 그래프들에 대하여 구조적 유사도를 측정하고, 유사도 임계값(β) 이상이 되어야 한다. 새로운 빈발 부분 그래프와 그룹내의 모든 그래프들과 비교 도중에 유사도 임계값(β)을 만족하지 않는 빈발 부분 그래프가 있을 경우 비교를 중단하고 그룹화를 시키지 않는다. 따라서, 최종적으로 n개의 빈발 부분 그래프로 그룹화가 된 경우에 그룹화를 하기 위한 총 유사도 비교는 평균적으로 n(n+1)/2 회가 발생한다(=$O(n^2)$).

[0036] 구조적 유사도 임계값(β)은 그룹의 수와 비례하므로, 구조적 유사도 임계값(β)이 높을수록 더 많은 그룹을 생성하게 되어 유사도를 비교하는 연산량이 증가하게 되므로 속도는 저하될 수 있다. 그러나, 더 다양한 그룹에서 선택된 특징을 통하여 분류의 정확성을 높일 수 있다. 반면, 구조적 유사도 임계값(β)을 낮추는 경우, 그룹의 수가 줄어들게 되어 유사도를 비교하는 연산량이 감소하여 속도를 향상시킬 수 있다. 하지만, 그룹의 수가 감소하여 더 적은 특징을 선택하게 되어 상대적으로 정확도가 감소하게 된다.

[0037] 도 4는 본 발명의 일 실시예에 따른 구조적 유사한 빈발 부분 그래프 그룹내의 빈발 부분 그래프의 특징들을 보여주는 도면이다.

[0038] 빈발 부분 그래프들은 복수 개의 구조적 유사 특징 그룹으로 분류될 수 있다. 도 4에 도시된 바와 같이, 빈발 부분 그래프들은 그룹 1(G_1), 그룹 2(G_2), 그룹 3(G_3) 및 그룹 4(G_4)로 분류될 수 있다. 예를 들어, 빈발 부분 그래프(f_1, f_4, f_6, f_7)가 그룹화되어 그룹 1(G_1)이 생성될 수 있다.

[0039] 도 5는 본 발명의 일 실시예에 따른 각 빈발 부분 그래프 그룹으로부터 특징을 선택하는 방법을 나타내는 순서도이다.

[0040] 각 빈발 부분 그래프 그룹은 구조적으로 유사한 특징 그룹이라고 할 수 있다. 현재 구조적으로 유사 특징 그룹 내에서의 모든 빈발 그래프들의 인포메이션 게인 값은 그룹화 이전에 측정된 값들이다. 이러한 인포메이션 게인 값은, 그래프 데이터베이스(도 2의 10) 내의 모든 그래프 인스턴스들의 클래스 레이블을 기준으로 인포페이션 게인 평가 함수를 이용하여 측정한 값이기 때문에, 이러한 평가 함수를 통해 계산된 인포메이션 게인 값이 각 그룹 내에서 가장 높다고 해서, 반드시 해당 그룹에 속하는 클래스들에 대하여 분류 성능이 가장 좋은 것은 아니다.

[0041] 예를 들어, 그룹(G_1)은 클래스 레이블이 C1 및 C3인 클래스에 대해서 잘 분류하는 빈발 부분 그래프들의 집합이라고 가정하자. 원래 그룹(G_1)을 그룹화할 때 사용하였던, 빈발 부분 그래프(f_{max})를 가지는 빈발 부분 그래프(f_1)는 클래스 레이블이 C1, C2 및 C3인 클래스를 가장 잘 분류하여 가장 높은 인포메이션 게인 값을 가지고 있었다고 가정한다. 즉, 그룹화를 할 때 계산하였던 인포메이션 게인은 전체 클래스 레이블에 대하여 계산된 값이다.

[0042] 따라서, 구조적으로 유사한 특징 그룹 내의 다른 빈발 부분 그래프(또는 특징) 중에서 클래스 레이블이 C1 및 C3인 클래스를 더 잘 분류하는 특징이 존재할 수 있다. 이와 같은 점을 고려하여, 도 2의 특징 선택부(230)는 구조적으로 유사한 빈발 부분 그래프 그룹 내의 빈발 부분 그래프 특징들을 가지고 있는 그래프 인스턴스들만을 대상으로 하여 인포메이션 게인을 다시 계산할 수 있다(510).

[0043] 여기에서, 특징 선택부(230)가 그룹화된 각 빈발 부분 그래프 그룹별로, 각 그룹에 속하는 빈발 부분 그래프들에 대한 인포메이션 게인을 재측정하기 위하여, 인포메이션 게인 함수($IG(Y_{Gi},f_i)$)를 이용할 수 있다. 이때 이용되는 인포메이션 게인 함수는, $IG(Y_{Gi},f_i)=H(Y_{Gi})-H(Y_{Gi}|f_i)$로 나타낼 수 있다. 여기에서, $H(Y_{Gi})$는 빈발 부분 그래프 그룹(Gi)에 속하는 클래스들의 그래프들의 집합(Y_{Gi})에 대한 엔트로피(entropy)를 나타내고, $H(Y_{Gi}|f_i)$는 빈발 부분 그래프 그룹(Gi)에 속하는 클래스들의 그래프들의 집합(Y_{Gi}) 및 집합(Y_{Gi})에 속하는 빈발 부분 그래프(f_i)에 대한 크로스 엔트로피(cross entropy)를 나타낸다.

[0044] 특징 선택부(230)는 재측정한 인포메이션 게인 값이 가장 높은 빈발 부분 그래프 즉, 분류 성능이 가장 좋은 빈발 부분 그래프를 이 그룹의 대표 특징으로 선택할 수 있다(520). 특징 선택부(230)는 이 과정을 모든 그룹에 대해서 반복하여(530), 대표 특징들을 추출함으로써 최종적인 특징 서브셋을 구성할 수 있다.

[0045] 본 발명의 일 실시예에 따르면, 구조적으로 유사한 빈발 부분 그래프는 비슷한 클래스에 대하여 유사한 분류 성능을 가지는 성질을 이용하여, 각각의 빈발 부분 그래프 그룹 내에서 분류 성능이 가장 좋은 빈발 부분 그래프를 특징으로 선택하였으므로, 편향적으로 특징 서브셋이 구성되는 것을 방지함과 동시에 이와 같은 방법으로 선택한 특징 서브셋을 이용하여 높은 분류 성능을 발휘하는 효과가 있다.

[0046] 도 6은 본 발명의 일 실시예에 따른 그래프 데이터베이스를 테이블 형태로 나타내는 도면이다.

[0047] 도 2의 그래프 데이터베이스(10)는 도 6에 도시된 바와 같은 테이블 형태로 구성될 수 있다. 그래프 데이터베이스(10)는, 분류 대상이 되는 그래프들에 대한 정보로서, 그래프 식별자, 버텍스 정보, 에지 정보, 및 클래스 레이블을 포함할 수 있다.

[0048] 그래프 식별자는, 각 그래프를 식별하는 정보이다. 그래프는 버텍스(vertex)와 에지(edge)의 집합으로 구성되어 있으므로, 각 그래프에 대한 정보로서 버텍스 정보 및 에지 정보가 포함될 수 있다. 버텍스 정보는, 각 버텍스에 대한 버텍스 식별자 및 버텍스 레이블을 포함할 수 있다. 에지는 이러한 버텍스 간의 관계를 표현하기 위하여 사용한다. 에지 정보는, 에지를 형성하는 근원 버텍스 식별자 및 대상 버텍스 식별자와, 에지에 대한 레이블을 포함할 수 있다.

[0049] 예를 들어, 도 6을 참조하면, 그래프(g_1)의 첫 번째 버텍스(V_1)의 레이블은 1이며, 두 번째 버텍스(V_2)의 레이블은 2이다. 그리고, 첫 번째 버텍스(V_1)와 두 번째 버텍스(V_2) 사이에는 에지가 존재하며, 에지의 레이블은 a이다. 그리고 이 그래프(g_1)는 클래스 레이블이 C1인 클래스에 속한다.

[0050] 도 7은 본 발명의 일 실시예에 따른 빈발 부분 그래프 테이블을 나타내는 도면이다.

[0051] 전술한 바와 같이, 빈발 부분 그래프들은, 그래프 데이터베이스(도 2의 10) 내에서 빈발 부분 그래프 마이닝을 통하여 추출될 수 있다. 추출된 빈발 부분 그래프에 대한 정보는 도 7과 같이 테이블 형태로 형성되어 도 2의 빈발 부분 그래프 정보 저장부(20)에 저장될 수 있다. 빈발 부분 그래프에 대한 정보는, 빈발 부분 그래

프 식별자 및 각 빈발 부분 그래프가 속하는 그래프 식별자 정보를 포함할 수 있다.

[0052] 예를 들어, 빈발 부분 그래프(f_1)는 그래프 <g_1, g_2, g_3, g_5, g_7, g_9,...,g_i>에 속하는 빈발 부분 그래프이다. 또한, 도 6에 도시된 바와 같은, 그래프 데이터베이스 테이블의 정보를 이용하면, 빈발 부분 그래프(f_1)가 속하는 그래프 <g_1, g_2, g_3, g_5, g_7, g_9,...,g_i>에 대한 각각의 클래스 레이블을 알 수 있다.

[0053] 도 8은 본 발명의 일 실시예에 따른 구조적으로 유사한 빈발 부분 그래프 그룹 후보 테이블을 나타낸 도면이다.

[0054] 구조적으로 유사한 빈발 부분 그래프 그룹 후보 테이블은 각 빈발 부분 그래프들의 인포메이션 게인 값의 계산 결과에 따라, 인포메이션 게인의 값의 기준으로 각 빈발 부분 그래프들이 내림차순으로 정렬된 테이블이다. 따라서, 첫 번째 빈발 부분 그래프(f_1)가 가장 높은 인포메이션 게인 값을 가지는 빈발 부분 그래프(f_{max})인 경우(즉, $f_{max}=f_1$), 도 1의 그룹화부(220)는 첫 번째 빈발 부분 그래프(f_1)을 기준으로 가장 먼저 그룹화를 실시할 수 있다.

[0055] 그룹화부(220)는 그룹화를 위하여, 빈발 부분 그래프 $f_{max}(=f_1)$와 두 번째로 인포메이션 게인 값이 높은 빈발 부분 그래프(f_3)와의 구조적 유사성을 그래프 수정 거리를 이용하여 측정할 수 있다. 이때, 빈발 부분 그래프 $f_{max}(=f_1)$와 빈발 부분 그래프(f_3)와의 구조적 유사도가 구조적 유사도 임계값(β) 미만의 값으로 측정되는 경우, 빈발 부분 그래프(f_3)는 빈발 부분 그래프(f_1)과 같은 그룹으로 분류되지 않을 수 있다. 반면, 빈발 부분 그래프 $f_{max}(=f_1)$와 빈발 부분 그래프(f_4)와의 유사도를 측정하였을 때, 유사도가 구조적 유사도 임계값(β) 이상의 값으로 측정되는 경우, 빈발 부분 그래프(f_4)를 빈발 부분 그래프(f_1)의 구조적 유사 특징 그룹으로 포함시킨다. 이와 같은 과정을 반복하여 그룹내의 유사도가 그룹내의 유사도 임계값(β)을 만족하는 모든 구조적으로 유사한 빈발 부분 그래프 그룹을 생성할 수 있다.

[0056] 도 9는 본 발명의 일 실시예에 따른 구조적으로 유사한 빈발 부분 그래프 그룹 테이블을 나타낸 도면이다.

[0057] 구조적으로 유사한 빈발 부분 그래프 그룹 테이블은 도 2의 그룹화부(220)에서 도 8과 같은 구조적으로 유사한 빈발 부분 그래프 그룹 후보 테이블을 이용하여 빈발 부분 그래프들을 그룹화한 결과를 나타낸다. 도 9에 도시된 바와 같이, 구조적으로 유사한 빈발 부분 그래프 그룹 테이블은 그룹 식별자, 해당 그룹으로 분류된 빈발 부분 그래프 식별자 정보 및 해당 그룹의 빈발 부분 그래프 중 특징으로 선택된 빈발 부분 그래프 정보를 포함할 수 있다.

[0058] 예를 들어, 그룹 식별자(G_1)은 빈발 부분 그래프 식별자 {f_1, f_4, f_6, f_7,...}을 가지며 대표 빈발 부분 그래프 특징으로 빈발 부분 그래프(f_1)을 사용할 수 있다. 이와 같이, 그룹화된 빈발 부분 그래프들에 대한 예는 도 4에 도시된 바와 같다.

[0059] 도 10은 본 발명의 일 실시예에 따른 그룹 대표 특징 선택 테이블을 나타낸 도면이다.

[0060] 도 10은 도 6, 도 8, 도 9의 테이블들을 결합(join)하여 생성될 수 있다. 그룹 식별자(G_1)는 빈발 부분 그래프 식별자 {f_1, f_4, f_6, f_7,...}을 가지며, 빈발 부분 그래프(f_1)에 대해서, 도 7에 도시된 바와 같은 빈발 부분 그래프 테이블을 참조하여, 빈발 부분 그래프(f_1)가 속하는 그래프 <g_1, g_2, g_3, g_5, g_7, g_9,...,g_i>을 가져올 수 있다. 그리고, 도 6에 도시된 바와 같은, 그래프 데이터베이스 테이블을 참조하여 각 그래프에 대한 각각의 클래스 레이블을 가져올 수 있다.

[0061] 도 1의 특징 선택부(230)는, 이를 통하여 그룹 내의 인포메이션 게인을 재측정할 수 있다. 그 결과, 특징 선택부(230)는 그룹(G_1) 내에서는 빈발 부분 그래프(f_1)의 인포메이션 게인 값이 0.3572로 가장 높으므로, 빈발 부분 그래프(f_1)를 분류 성능이 가장 높은 특징으로 선택할 수 있다. 이와 같은 과정을 모든 그룹에 대하여 반복하여 특징이 선택되어 선택된 특징들의 집합인 특징 서브셋이 생성될 수 있다.

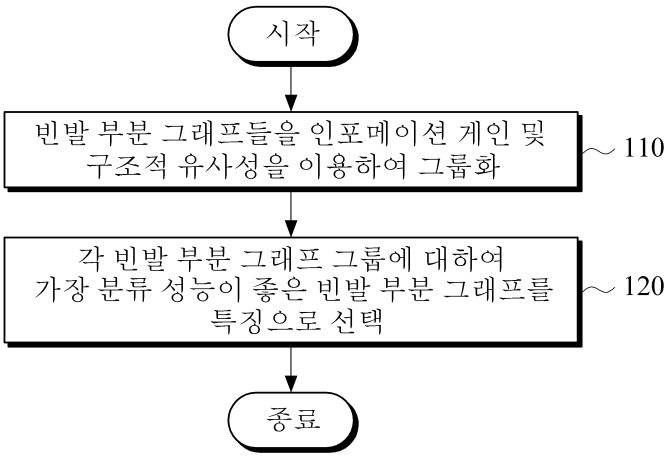
[0062] 본 발명의 일 양상은 컴퓨터로 읽을 수 있는 기록 매체에 컴퓨터가 읽을 수 있는 코드로서 구현될 수 있다. 상기의 프로그램을 구현하는 코드들 및 코드 세그먼트들은 당해 분야의 컴퓨터 프로그래머에 의하여 용이하게 추론될 수 있다. 컴퓨터가 읽을 수 있는 기록매체는 컴퓨터 시스템에 의하여 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록 장치를 포함한다. 컴퓨터가 읽을 수 있는 기록 매체의 예로는 ROM, RAM, CD-ROM, 자

기 테이프, 플로피 디스크, 광 디스크 등을 포함한다. 또한, 컴퓨터가 읽을 수 있는 기록 매체는 네트워크로 연결된 컴퓨터 시스템에 분산되어, 분산 방식으로 컴퓨터가 읽을 수 있는 코드로 저장되고 실행될 수 있다.

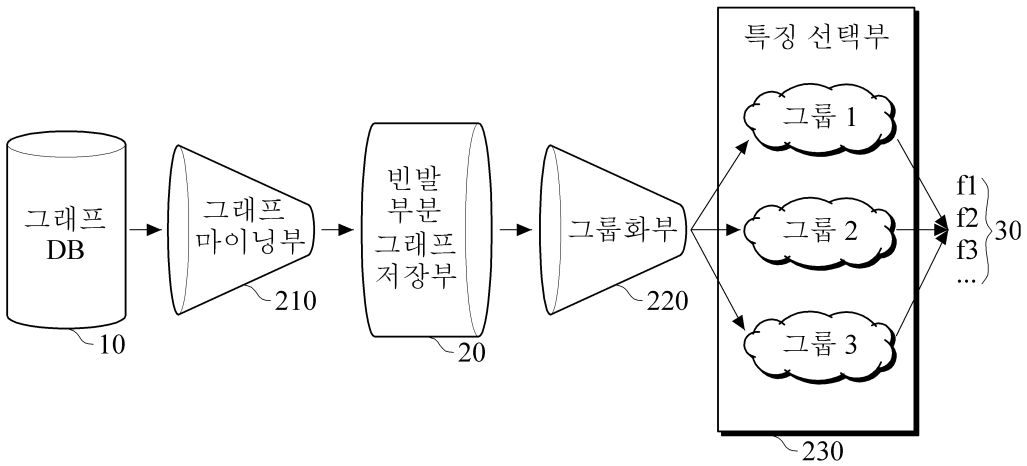
[0063] 이상의 설명은 본 발명의 일 실시예에 불과할 뿐, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자는 본 발명의 본질적 특성에서 벗어나지 않는 범위에서 변형된 형태로 구현할 수 있을 것이다. 따라서, 본 발명의 범위는 전술한 실시예에 한정되지 않고 특허 청구범위에 기재된 내용과 동등한 범위 내에 있는 다양한 실시 형태가 포함되도록 해석되어야 할 것이다.

도면

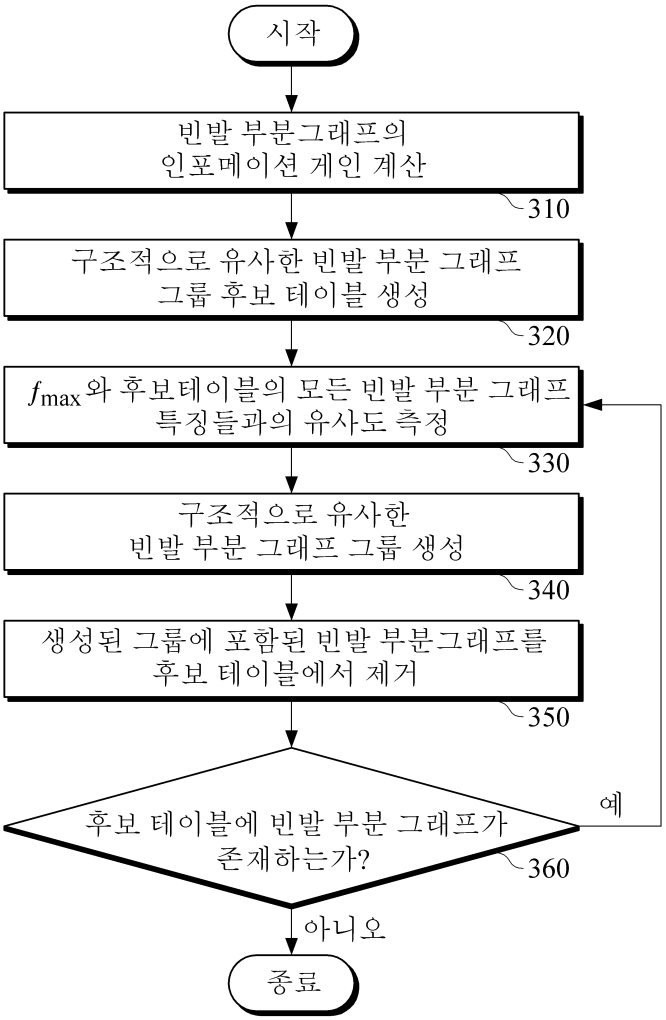
도면1



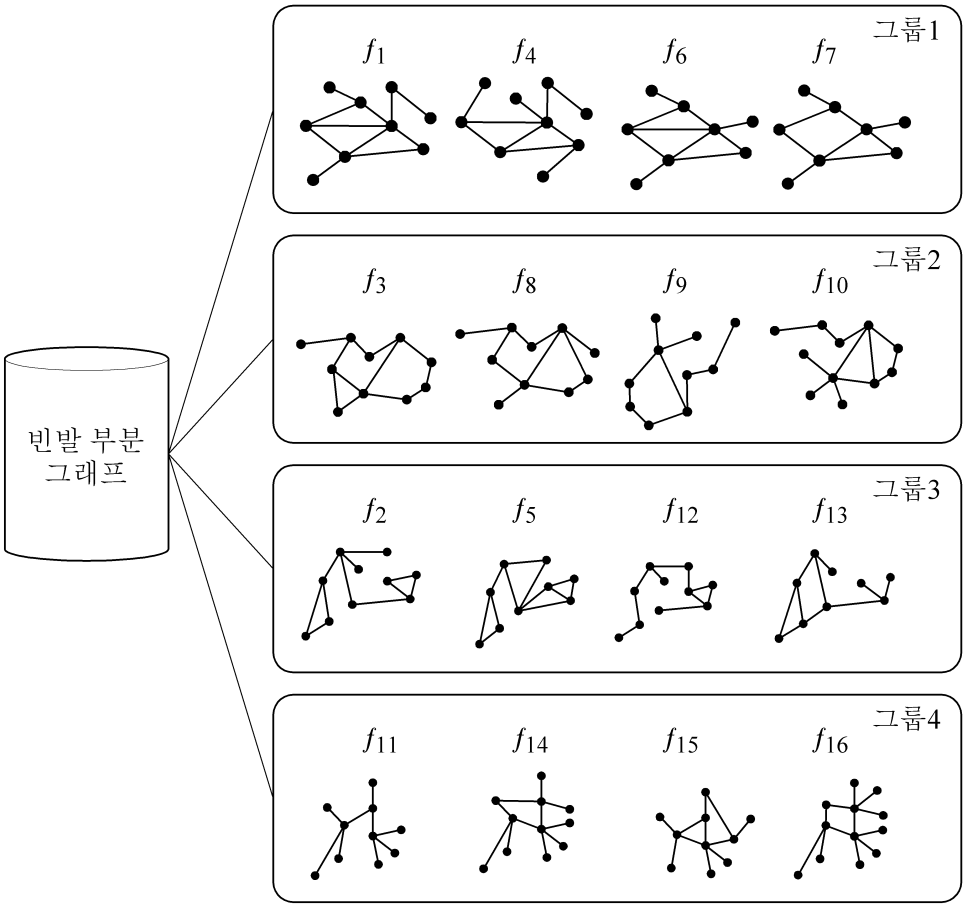
도면2



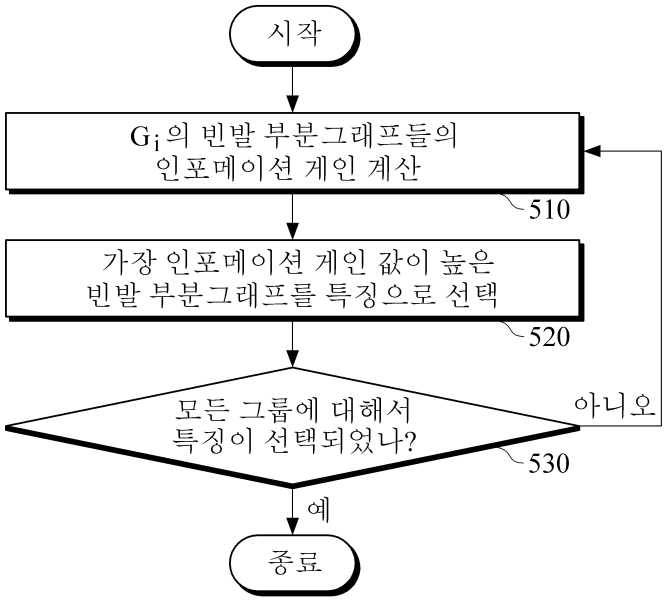
도면3


시작
빈발 부분그래프의
인포메이션 게인 계산
310
구조적으로 유사한 빈발 부분 그래프
그룹 후보 테이블 생성
320
f_{max}와 후보테이블의 모든 빈발 부분 그래프
특징들과의 유사도 측정
330
구조적으로 유사한
빈발 부분 그래프 그룹 생성
340
생성된 그룹에 포함된 빈발 부분그래프를
후보 테이블에서 제거
350
후보 테이블에 빈발 부분 그래프가
존재하는가?
360
예
아니오
종료

도면4


그룹1
f_1
f_4
f_6
f_7
그룹2
f_3
f_8
f_9
f_{10}
그룹3
f_2
f_5
f_{12}
f_{13}
그룹4
f_{11}
f_{14}
f_{15}
f_{16}
빈발 부분
그래프

도면5


시작
G_i의 빈발 부분그래프들의
인포메이션 게인 계산
510
가장 인포메이션 게인 값이 높은
빈발 부분그래프를 특징으로 선택
520
모든 그룹에 대해서
특징이 선택되었나?
530
아니오
예
종료

도면6

그래프 식별자	버텍스 (식별자,레이블)	에지(근원 버텍스, 대상 버텍스, 레이블)	클래스 레이블
g_1	$(V_1,1), (V_2,2), ...$	$(V_1,V_2,a), (V_2,V_3,b), ...$	C1
g_2	$(V_1,3), (V_2,4), ...$	$(V_1,V_2,c), (V_1,V_3,a), ...$	C2
g_3	$(V_1,2), (V_2,1), ...$	$(V_1,V_2,a), (V_1,V_3,c), ...$	C1
g_4	$(V_1,4), (V_2,2), ...$	$(V_1,V_3,d), (V_2,V_3,a), ...$	C3
⋮	⋮	⋮	⋮

도면7

빈발 부분 그래프 식별자	그래프 식별자
f_1	$g_1, g_2, g_3, g_5, g_7, g_9, \cdots$
f_2	$g_1, g_4, g_6, g_8, g_{10}, g_{11}, \cdots$
f_3	$g_4, g_5, g_{11}, g_{13}, g_{14}, g_{16}, \cdots$
f_4	$g_3, g_4, g_5, g_7, g_8, g_{10}, \cdots$
⋮	⋮

도면8

순위	인포메이션 게인 값	빈발 부분 그래프 식별자
1	0.572	f_1
2	0.547	f_3
3	0.512	f_4
4	0.501	f_2
⋮	⋮	⋮

도면9

순위	빈발 부분 그래프 식별자	특징
G_1	$f_1, f_4, f_6, f_7, \ldots$	f_1
G_2	$f_3, f_8, f_{10}, \ldots$	f_3
G_3	$f_2, f_9, \ldots$	f_2
G_4	$f_{11}, f_{14}, \ldots$	f_{11}
⋮	⋮	⋮

도면10

그룹 식별자	빈발 부분 그래프 식별자	그래프 식별자	클래스 레이블	그룹내의 인포메이션 게인
G_1	f_1	g_1	C1	0.3572
		g_3	C1	
		⋮	⋮	
	f_4	g_3	C1	0.3415
		g_7	C2	
		⋮	⋮	
	⋮	⋮	⋮	⋮
		⋮	⋮	
		⋮	⋮	
G_2	f_3	g_2	C2	0.2915
		g_4	C3	
		⋮	⋮	
	f_8	g_2	C2	0.2321
		g_6	C2	
		⋮	⋮	
	⋮	⋮	⋮	⋮
		⋮	⋮	
		⋮	⋮	