



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2019-0136969
(43) 공개일자 2019년12월10일

(51) 국제특허분류(Int. Cl.)
G06F 16/2457 (2019.01)
(52) CPC특허분류
G06F 16/24578 (2019.01)
(21) 출원번호 10-2019-0062882
(22) 출원일자 2019년05월29일
심사청구일자 2019년05월29일
(30) 우선권주장
1020180061513 2018년05월30일 대한민국(KR)

(71) 출원인
경희대학교 산학협력단
경기도 용인시 기흥구 덕영대로 1732 (서천동, 경희대학교 국제캠퍼스내)
(72) 발명자
이승룡
경기도 성남시 분당구 구미로144번길 25,102동 102호 (구미동, 삼환빌라)
박불알리
경기도 용인시 기흥구 덕영대로 1732 경희대학교 국제캠퍼스 전자정보대학 351
강병호
경기도 용인시 기흥구 덕영대로 1732(서천동)
(74) 대리인
특허법인도담

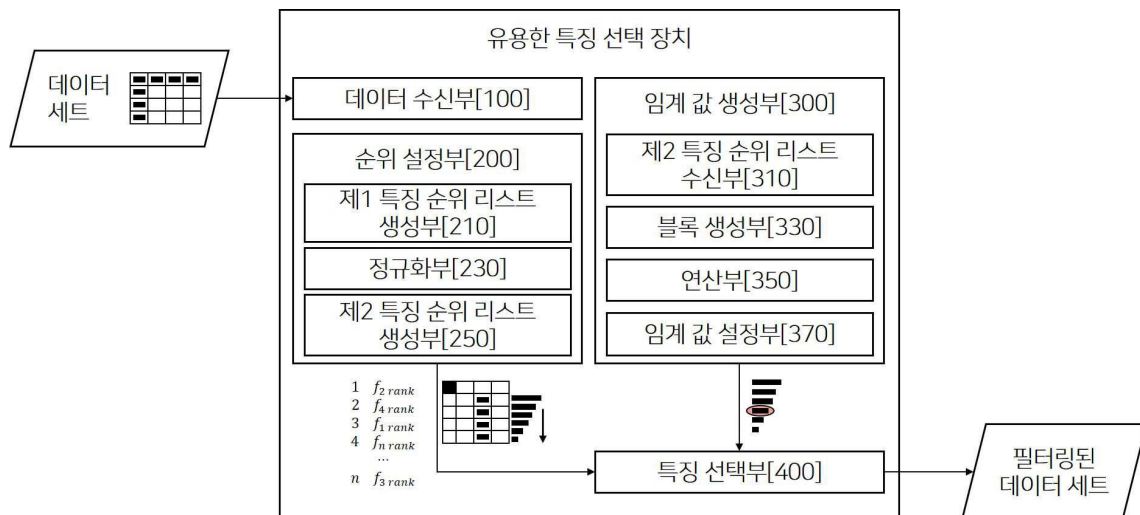
전체 청구항 수 : 총 8 항

(54) 발명의 명칭 유용한 특징 선택 방법 및 그 장치

(57) 요약

본 발명은 유용한 특징을 선택하는 방법에 관한 것으로, 데이터베이스에서 데이터 세트를 로드하고, 상기 데이터 세트에 포함된 하나 이상의 제1 특징을 추출하는 단계, 상기 제1 특징에 n개의 특징 평가 모델을 각각 적용하여 제2 특징 순위 리스트를 생성하는 단계, 상기 제2 특징 순위 리스트에 기반하여 상기 제1 특징의 비율에 따라 적어도 하나 이상의 특징 블록을 생성하는 단계, 복수 개의 분류기를 이용하여 상기 특징 블록에 포함된 제2 특징을 각각 분류하고, 상기 분류 결과의 정확도를 연산하는 단계, 상기 정확도를 이용하여 임계 값을 생성하는 단계, 및 상기 제2 특징 순위 리스트에 포함된 상기 제1 특징 중 상기 임계 값 이상인 제3 특징(상기 유용한 특징)을 식별하는 단계를 포함하는 것을 특징으로 한다.

대표도 - 도1



이 발명을 지원한 국가연구개발사업

과제고유번호 1711070401
 부처명 과학기술정보통신부
 연구관리전문기관 정보통신기술진흥센터
 연구사업명 연구개발고급인력지원
 연구과제명 지능형 의료 플랫폼 개발
 기 여 율 40/100
 주관기관 경희대학교산학협력단
 연구기간 2019.01.01 ~ 2019.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호 1711065120
 부처명 과학기술정보통신부
 연구관리전문기관 정보통신기술진흥센터
 연구사업명 디지털콘텐츠원천기술개발
 연구과제명 다양한 디지털 제품의 UX 평가를 위한 Lean UX 핵심기술 및 플랫폼
 기 여 율 30/100
 주관기관 경희대학교산학협력단
 연구기간 2019.01.01 ~ 2019.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호 1415146278
 부처명 산업통상자원부
 연구관리전문기관 한국산업기술평가관리원
 연구사업명 지식서비스산업핵심기술개발
 연구과제명 퍼스널 빅데이터를 활용한 마이닝마인즈 핵심기술 개발
 기 여 율 30/100
 주관기관 경희대학교산학협력단
 연구기간 2017.06.01 ~ 2018.05.31

명세서

청구범위

청구항 1

서버가 유용한 특징을 선택하는 방법에 있어서,

데이터베이스에서 데이터 세트를 로드하고, 상기 데이터 세트에 포함된 하나 이상의 제1 특징을 추출하는 단계;

상기 제1 특징에 n 개의 특징 평가 모델을 각각 적용하여 제2 특징 순위 리스트를 생성하는 단계;

상기 제2 특징 순위 리스트에 기반하여 상기 제1 특징의 비율에 따라 적어도 하나 이상의 특징 블록을 생성하는 단계;

복수 개의 분류기를 이용하여 상기 특징 블록에 포함된 제2 특징을 각각 분류하고, 상기 분류 결과의 정확도를 연산하는 단계;

상기 정확도를 이용하여 임계 값을 생성하는 단계; 및

상기 제2 특징 순위 리스트에 포함된 상기 제1 특징 중 상기 임계 값 이상인 제3 특징(상기 유용한 특징)을 식별하는 단계를 포함하는 것을 특징으로 하는 유용한 특징 선택 방법.

청구항 2

제1항에 있어서,

상기 특징 순위 리스트를 생성하는 단계는,

상기 제1 특징에 n 개의 특징 순위 측정 모델을 각각 적용하여 n 개의 제1 특징 순위 리스트를 생성하는 단계;

상기 제1 특징 순위 리스트를 정규화하는 단계;

상기 정규화된 제1 특징 순위 리스트를 병합하여 제2 특징 순위 리스트를 생성하는 단계를 포함하는 것을 특징으로 하는 유용한 특징 선택 방법.

청구항 3

제1항에 있어서,

상기 임계 값을 생성하는 단계는,

상기 특징 블록을 기반으로 하나 이상의 분류기를 이용하여 측정된 정확도의 평균 값을 연산하는 단계;

상기 평균 값을 이용하여 상기 임계 값을 설정하는 단계를 더 포함하는 것을 특징으로 하는 유용한 특징 선택 방법.

청구항 4

제1항에 있어서,

상기 제1 특징 중 상기 임계 값 미만인 제4 특징을 배제하는 단계를 포함하는 것을 특징으로 하는 유용한 특징 선택 방법.

청구항 5

유용한 특징을 선택하는 장치에 있어서,

데이터베이스에서 데이터 세트를 로드하는 데이터 수신부;

상기 데이터 세트에 포함된 하나 이상의 제1 특징을 추출하고, 상기 제1 특징에 n개의 특징 평가 모델을 각각 적용하여 제2 특징 순위 리스트를 생성하는 순위 설정부;

상기 제2 특징 순위 리스트에 기반하여 상기 제1 특징의 비율에 따라 적어도 하나 이상의 특징 블록을 생성하며, 복수 개의 분류기를 이용하여 상기 특징 블록에 포함된 제2 특징을 각각 분류하고, 상기 분류 결과의 정확도를 연산하며, 상기 정확도를 이용하여 임계 값을 생성하는 임계 값 생성부; 및

상기 제2 특징 순위 리스트에 포함된 상기 제1 특징 중 상기 임계 값 이상인 제3 특징(상기 유용한 특징)을 식별하는 특징 선택부를 포함하는 것을 특징으로 하는 유용한 특징 선택 장치.

청구항 6

제5항에 있어서,

상기 순위 설정부는,

상기 제1 특징에 n개의 특징 순위 측정 모델을 각각 적용하여 제1 특징 순위 리스트를 생성하고, 상기 제1 특징 순위 리스트를 정규화하며, 상기 정규화된 n개의 제1 특징 순위 리스트를 병합하여 제2 특징 순위 리스트를 생성하는 것을 특징으로 하는 유용한 특징 선택 장치.

청구항 7

제5항에 있어서,

상기 임계 값 생성부는,

상기 특징 블록을 기반으로 하나 이상의 분류기를 이용하여 측정된 정확도의 평균 값을 연산하고, 상기 평균 값을 이용하여 상기 임계 값을 설정하는 것을 특징으로 하는 유용한 특징 선택 장치.

청구항 8

제5항에 있어서,

상기 특징 선택부는 상기 제1 특징 중 상기 임계 값 미만인 제4 특징을 식별하고, 상기 제4 특징을 배제하는 것을 특징으로 하는 유용한 특징 선택 장치.

발명의 설명

기술 분야

[0001] 본 발명은 데이터 세트에서 유용한 특징을 선택하는 방법 및 그 장치에 관한 것으로, 보다 자세하게는 데이터 세트에 포함된 특징의 순위를 설정하여 특정 순위 이상인 특징을 유용한 특징으로 선택하는 방법에 관한 것이다.

배경 기술

[0002] 기계 학습과 데이터 마이닝 분야는 방대한 양의 데이터를 이용하기 때문에 데이터 처리에 있어서 많은 연구가 진행되고 있다. 특히 데이터의 전처리 과정에서 핵심적인 목표는 데이터에 포함된 특징 중 유의미한 특징을 선택하는 것으로, 기계 학습과 데이터 마이닝은 데이터에서 추출된 특징을 기반으로 데이터 처리를 하기 때문에 데이터에서 유효한 특징을 추출하는 기술은 특히 중요하다. 데이터의 양이 많아질수록 방대한 데이터에서 추출된 특징에서 유의미한 특징을 선택하는 것은 적은 데이터를 기반으로 하는 특징 선택과 대비하여 복잡도가 증가할 수 있다.

[0003] 일반적으로 유의미한 특징을 선택하기 위하여 특징에 대하여 정보 이론 측정, 상관 관계 측정, 일관성 측정, 거리 기반 측정, 분류/예측 정확도 측정 등의 기법을 적용할 수 있다. 각각의 기법은 고유의 장점과 한계점을 가질 수 있다.

발명의 내용

해결하려는 과제

[0005] 본 발명은 전술한 문제점을 해결하기 위한 것으로서, 하나 이상의 특징 선택 알고리즘을 이용함으로써 보다 유의미한 특징을 선택할 수 있게 하는 것을 일 목적으로 한다.

[0006] 또한 본 발명은 특징이 내포하는 의미에 대한 순위를 설정함으로써 특징을 선택할 수 있게 하는 것을 일 목적으로 한다.

[0007] 또한 본 발명은 특징의 순위 리스트를 기반으로 특징 선택의 임계 값을 설정하는 것을 일 목적으로 한다.

[0008] 또한 본 발명은 하나 이상의 분류기를 이용하여 특징 선택의 임계 값을 설정함으로써 특징 선택의 정확도를 향상시키는 것을 일 목적으로 한다.

과제의 해결 수단

[0009] 이러한 목적을 달성하기 위한 본 발명은 서버가 유용한 특징을 선택하는 방법에 있어서, 데이터베이스에서 데이터 세트를 로드하고, 상기 데이터 세트에 포함된 하나 이상의 제1 특징을 추출하는 단계, 상기 제1 특징에 n 개의 특징 평가 모델을 각각 적용하여 제2 특징 순위 리스트를 생성하는 단계, 상기 제2 특징 순위 리스트에 기반하여 상기 제1 특징의 비율에 따라 적어도 하나 이상의 특징 블록을 생성하는 단계, 복수 개의 분류기를 이용하여 상기 특징 블록에 포함된 제2 특징을 각각 분류하고, 상기 분류 결과의 정확도를 연산하는 단계, 상기 정확도를 이용하여 임계 값을 생성하는 단계, 및 상기 제2 특징 순위 리스트에 포함된 상기 제1 특징 중 상기 임계 값 이상인 제3 특징(상기 유용한 특징)을 식별하는 단계를 포함하는 것을 일 특징으로 한다.

[0010] 또한 상기 특징 순위 리스트를 생성하는 단계는, 상기 제1 특징에 n 개의 특징 순위 측정 모델을 각각 적용하여 n 개의 제1 특징 순위 리스트를 생성하는 단계, 상기 제1 특징 순위 리스트를 정규화하는 단계, 상기 정규화된 제1 특징 순위 리스트를 병합하여 제2 특징 순위 리스트를 생성하는 단계를 포함하는 것을 일 특징으로 한다.

[0011] 나아가 상기 임계 값을 생성하는 단계는, 상기 특징 블록을 기반으로 하나 이상의 분류기를 이용하여 측정된 정확도의 평균 값을 연산하는 단계, 상기 평균 값을 이용하여 상기 임계 값을 설정하는 단계를 더 포함하는 것을 일 특징으로 한다.

[0012] 또한 상기 제1 특징 중 상기 임계 값 미만인 제4 특징을 배제하는 단계를 포함하는 것을 일 특징으로 한다.

[0013] 나아가 본 발명은 유용한 특징을 선택하는 장치에 있어서, 데이터베이스에서 데이터 세트를 로드하는 데이터 수신부, 상기 데이터 세트에 포함된 하나 이상의 제1 특징을 추출하고, 상기 제1 특징에 n 개의 특징 평가 모델을 각각 적용하여 제2 특징 순위 리스트를 생성하는 순위 설정부, 상기 제2 특징 순위 리스트에 기반하여 상기 제1 특징의 비율에 따라 적어도 하나 이상의 특징 블록을 생성하며, 복수 개의 분류기를 이용하여 상기 특징 블록에 포함된 제2 특징을 각각 분류하고, 상기 분류 결과의 정확도를 연산하며, 상기 정확도를 이용하여 임계 값을 생성하는 임계 값 생성부, 및 상기 제2 특징 순위 리스트에 포함된 상기 제1 특징 중 상기 임계 값 이상인 제3 특징(상기 유용한 특징)을 식별하는 특징 선택부를 포함하는 것을 일 특징으로 한다.

[0014] 또한 상기 순위 설정부는, 상기 제1 특징에 n 개의 특징 순위 측정 모델을 각각 적용하여 제1 특징 순위 리스트를 생성하고, 상기 제1 특징 순위 리스트를 정규화하며, 상기 정규화된 n 개의 제1 특징 순위 리스트를 병합하여 제2 특징 순위 리스트를 생성하는 것을 일 특징으로 한다.

[0015] 나아가 상기 임계 값 생성부는, 상기 특징 블록을 기반으로 하나 이상의 분류기를 이용하여 측정된 정확도의 평균 값을 연산하고, 상기 평균 값을 이용하여 상기 임계 값을 설정하는 것을 일 특징으로 한다.

[0016] 또한 상기 특징 선택부는 상기 제1 특징 중 상기 임계 값 미만인 제4 특징을 식별하고, 상기 제4 특징을 배제하는 것을 일 특징으로 한다.

발명의 효과

- [0017] 전술한 바와 같은 본 발명에 의하면, 하나 이상의 특징 선택 알고리즘을 이용함으로써 보다 유의미한 특징을 선택할 수 있다.
- [0018] 또한 본 발명은 특징이 내포하는 의미에 대한 순위를 설정함으로써 특징을 선택할 수 있다.
- [0019] 또한 본 발명은 특징의 순위 리스트를 기반으로 특징 선택의 임계 값을 설정할 수 있다.
- [0020] 또한 본 발명은 하나 이상의 분류기를 이용하여 특징 선택의 임계 값을 설정함으로써 특징 선택의 정확도를 향상시킬 수 있다.

도면의 간단한 설명

- [0021] 도 1은 본 발명의 일 실시 예에 의한 유용한 특징 선택 장치의 구성을 도시한 도면이다.
- 도 2는 본 발명의 일 실시 예에 의한 유용한 특징 선택 방법을 설명하기 위한 도면이다.
- 도 3은 본 발명의 일 실시 예에 의한 특징 순위 리스트를 생성하는 방법을 설명하기 위한 도면이다.
- 도 4은 본 발명의 일 실시 예에 의한 특징의 순위를 설정하기 위한 알고리즘이다.
- 도 5는 본 발명의 일 실시 예에 의한 특징 순위 리스트를 정규화하기 위한 알고리즘이다.
- 도 6은 본 발명의 일 실시 예에 의한 유용한 특징을 선택하기 위한 임계 값을 연산하기 위한 알고리즘이다.

발명을 실시하기 위한 구체적인 내용

- [0022] 전술한 목적, 특징 및 장점은 첨부된 도면을 참조하여 상세하게 후술되며, 이에 따라 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 본 발명의 기술적 사상을 용이하게 실시할 수 있을 것이다. 본 발명을 설명함에 있어서 본 발명과 관련된 공지 기술에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 상세한 설명을 생략한다.
- [0023] 도면에서 동일한 참조부호는 동일 또는 유사한 구성요소를 가리키는 것으로 사용되며, 명세서 및 특허청구의 범위 기재된 모든 조합은 임의의 방식으로 조합될 수 있다. 그리고 다른 식으로 규정하지 않는 한, 단수에 대한 언급은 하나 이상을 포함할 수 있고, 단수 표현에 대한 언급은 또한 복수 표현을 포함할 수 있음이 이해되어야 한다.
- [0024] 본 명세서에서 사용되는 용어는 단지 특정 예시적 실시 예들을 설명할 목적을 가지고 있으며 한정할 의도로 사용되는 것이 아니다. 본 명세서에서 사용된 바와 같은 단수적 표현들은 또한, 해당 문장에서 명확하게 달리 표시하지 않는 한, 복수의 의미를 포함하도록 의도될 수 있다. 용어 "및/또는," "그리고/또는"은 그 관련되어 나열되는 항목들의 모든 조합들 및 어느 하나를 포함한다. 용어 "포함한다", "포함하는", "포함하고 있는", "구비하는", "갖는", "가지고 있는" 등은 내포적 의미를 갖는 바, 이에 따라 이러한 용어들은 그 기재된 특징, 정수, 단계, 동작, 요소, 및/또는 컴포넌트를 특정하며, 하나 이상의 다른 특징, 정수, 단계, 동작, 요소, 컴포넌트, 및/또는 이들의 그룹의 존재 혹은 추가를 배제하지 않는다. 본 명세서에서 설명되는 방법의 단계들, 프로세스들, 동작들은, 구체적으로 그 수행 순서가 확정되는 경우가 아니라면, 이들의 수행을 논의된 혹은 예시된 그러한 특정 순서로 반드시 해야 하는 것으로 해석해서는 안 된다. 추가적인 혹은 대안적인 단계들이 사용될 수 있음을 또한 이해해야 한다.
- [0025] 또한, 각각의 구성요소는 각각 하드웨어 프로세서로 구현될 수 있고, 위 구성요소들이 통합되어 하나의 하드웨어 프로세서로 구현될 수 있으며, 또는 위 구성요소들이 서로 조합되어 복수 개의 하드웨어 프로세서로 구현될 수도 있다.
- [0026] 이하, 첨부된 도면을 참조하여 본 발명에 따른 바람직한 실시 예를 상세히 설명하기로 한다.
- [0027] 본 발명은 데이터 세트에서 추출된 하나 이상의 특징에서 불필요한 특징을 제외시켜 유의미한 특징만을 추출하는 특징 선택 방법을 개시하고 있다. 특히 본 발명은 데이터 세트에서 추출된 특징에 대해 하나 이상의 특징 평가 모델을 적용하여 특징에 순위를 부여하여 유의미한 특징을 식별할 수 있다. 이를 보다 구체적으로 설명하기 위하여 본 발명의 일 실시 예에 의한 유용한 특징 선택 장치의 구성을 도시한 도 1을 참조할 수 있다.
- [0028] 도 1을 참조하면, 유용한 특징 선택 장치는 데이터 세트를 입력 변수로 하여 유의미한 특징만을 포함하는 필터

링된 데이터 세트를 출력할 수 있다. 이를 수행하기 위한 유용한 특징 선택 장치는 데이터 수신부(100), 순위 설정부(200), 임계 값 생성부(300), 그리고 특징 선택부(400)를 포함할 수 있다.

[0029] 데이터 수신부(100)는 데이터베이스에서 데이터 세트를 로드할 수 있다. 데이터 세트는 이미지 데이터, 문장 데이터, 센서 데이터 등 다양한 데이터를 포함할 수 있으며, 특히 본 발명의 일 실시 예에 의한 데이터 세트는 의료 데이터일 수 있다.

[0030] 순위 설정부(200)는 데이터 수신부(100)에서 수신한 데이터 세트에 포함된 특징을 평가하여 특징의 순위를 설정할 수 있다. 보다 구체적으로 순위 설정부(200)는 제1 특징 리스트 생성부(210), 정규화부(230), 그리고 제2 특징 리스트 생성부(250)를 포함할 수 있다. 순위 설정부(200)는 도 4에 도시된 알고리즘을 통해 동작할 수 있다.

[0031] 제1 특징 순위 리스트 생성부(210)는 데이터 세트에 포함된 하나 이상의 제1 특징을 추출할 수 있다. 여기서 사용되는 특징은 데이터의 특성을 표현하기 위한 개별적이고 경험적인 속성을 의미하는 것으로, 특징을 추출하기 위하여 SIFT(Scale Invariant Feature Transform), HOG(Histogram of Oriented Gradient), Haar Feature, Ferns, LBP(Local Binary Pattern), MCT(Modified Census Transform) 과 같은 다양한 알고리즘을 사용할 수 있다. 본 발명에서 사용되는 제1 특징은 데이터 세트에 미리 추출된 상태로 저장되어 있을 수 있다.

[0032] 제1 특징 순위 리스트 생성부(210)는 제1 특징에 n 개의 특징 평가 기법($n \geq 1$)을 각각 적용하여 제1 특징 순위 리스트를 생성할 수 있다. 제1 특징 순위 리스트 생성부(210)는 제1 특징에 n 개의 특징 평가 기법 각각 적용하기 때문에 n 개의 제1 특징 순위 리스트를 생성할 수 있다. 보다 구체적으로 특징 평가 기법은 특징의 순위를 연산하여 특징을 선택하는 것을 목적으로 하여, IG(Information Gain), Gain Ratio, Chi-Squared, Symmetric Uncertainty, Significance 등과 같은 단변량 평가 기법을 포함할 수 있다.

[0033] IG(Information Gain, 정보 이득) 평가 기법은 특징 선택 분야에서 가장 보편적으로 사용되는 기법으로, 특정 속성이 데이터를 얼마나 잘 구분하는 지를 판단할 수 있다. IG 평가 기법은 전체 데이터 세트, 제1 특징 및 제1 특징에 대한 속성의 엔트로피를 이용할 수 있다. 엔트로피(entropy)는 데이터 집합의 혼잡도를 나타내는 정보 측정 방법으로, 구분된 데이터의 비율이 균일하면 낮은 값을 가질 수 있다. 본 발명의 제1 특징 순위 리스트 생성부(210)는 전체 데이터 세트(D)의 엔트로피를 $Info(D)$, D 에 속하는 제1 특징(A)에 대한 엔트로피를 $Info_A(D)$ 라고 하면 제1 특징(A)에 대한 IG 값을 $IG(A) = Info(D) - Info_A(D)$ 와 같이 연산할 수 있다. 예를 들어 알러지(D)에 사과 알러지(A), 복숭아 알러지(B)가 포함된다고 가정하면, 사과 알러지에 대한 IG 값은 $IG(\text{사과 알러지}) = Info(\text{알러지}) - Info_{\text{사과}}(\text{알러지})$ 와 같이 연산할 수 있다.

[0034] Gain Ratio 평가 기법은 제1 특징(A)에 대한 IG 값이 보다 의미있는 값을 가질 수 있도록 정규화하는 기법으로, 수학적 1과 같이 제1 특징(A)의 IG 값을 정규화할 수 있다. 수학적 1에 개시된 바와 같이 Gain Ratio는 IG 값을 정규화하기 위하여 분할된 정보($SplitInfo_A(D)$)를 이용할 수 있다.

수학적 1

$$SplitInfo_A(D) = - \sum_{j=1}^v \frac{|D_j|}{|D|} * \log_2 \frac{|D_j|}{|D|}$$

[0035] $GainRatio(A) = IG(A)/SplitInfo(A)$

[0036] Chi-Squared(카이 제곱) 평가 기법은 제1 특징(A)과 제1 특징(A)가 속하는 클래스 또는 카테고리인 C_i 사이의 연관성을 연산하는 기법으로, 일 클래스에 포함된 제1 특징(A)의 독립성을 측정할 수 있다. Chi-Squared 기법은 수학적 2를 이용할 수 있는데, 수학적 2에서 F_1 는 A 와 C_i 모두가 발생하는 빈도 수, F_2 는 A 만 발생하고 C_i 는 발생하지 않는 빈도 수, F_3 은 A 는 발생하지 않고 C_i 는 발생하는 빈도 수, F_4 는 A 와 C_i 모두가 발생하지 않는 빈도 수를 의미하고, N 는 데이터에 포함된 속성(속성 A)의 총 개수를 의미할 수 있다. 이 때 $CHI(A, C_i)$ 가 0의

값을 가지면 A와 C_i 는 상호 독립적인 관계에 있다고 판단할 수 있다.

수학식 2

$$CHI(A, C_i) = \frac{N * (F_1 F_4 - F_2 F_3)^2}{(F_1 + F_3) * (F_2 + F_4) * (F_1 + F_2) * (F_3 + F_4)}$$

$$CHI_{\max}(A) = \max_i(CHI(A, C_i))$$

Symmetric uncertainty(SU, 대칭 불확실성) 평가 기법은 구축된 솔루션을 평가하기 위한 정보 분석 기법으로, 특히 불균형한 특징 세트에서 사용될 수 있다. SU 평가 기법은 수학식 3을 이용하여 솔루션의 등급을 평가할 수 있다. 수학식 3에서 $IG(A|B)$ 는 독립적인 제1 특징(A)과 클래스 특징(B)를 통해 연산된 IG 값을, $H(A)$ 는 제1 특징(A)에 대한 엔트로피를, $H(B)$ 는 클래스 특징(B)에 대한 엔트로피를 의미할 수 있다.

수학식 3

$$SU(A, B) = \frac{2 * IG(A|B)}{H(A) + H(B)}$$

Significance(유의성) 평가 기법은 클래스 특징과 관련된 제1 특징(A_i)의 가치를 평가하는 기법으로, 제1 특징(A_i)에 대한 유의성 값은 $\sigma(A_i)$ 로 표시할 수 있으며 이는 수학식 4와 같은 방법으로 연산될 수 있다. 수학식 4에서 $AE(A_i)$ 는 제1 특징(A_i)이 연결 가능한 모든 클래스 연결에 있어서 발생하는 누적 효과를 의미하고, k 는 제1 특징(A_i)의 다른 값을 나타낼 수 있다. $AE(A_i)$ 와 유사하게 $CE(A_i)$ 는 클래스의 변경에 따른 제1 특징의 변경으로 인한 효과를 연산하고, 이를 이용하여 제1 특징(A_i)과 다양한 클래스의 결정에 있어서 그 연관성을 측정할 수 있다. $CE(A_i)$ 를 측정함에 있어서 사용되는 m 은 클래스의 수를, $+(A_i)$ 는 제1 특징(A_i)에 대한 클래스-제1 특징 간의 연관성을 의미할 수 있다.

수학식 4

$$\sigma(A_i) = \frac{AE(A_i) + CE(A_i)}{2}$$

$$AE(A_i) = \left(\frac{1}{k} \sum_{r=1,2,\dots,k} \vartheta_i^r \right) - 1.0$$

$$CE + (A_i) = \frac{1}{m} * \left(\sum_{j=1,2,\dots,m} A_i^j \right) - 1.0$$

제1 특징 순위 리스트 생성부(210)는 특징 평가 기법으로 상술한 5가지 기법 외에 다른 기법 또한 포함할 수 있다. 다른 기법의 예로는 OneR, Relief, ReliefF, DRB-FS, $GR - \chi^2$, Borda method, EMFFS 평가 기법이 있을 수 있다.

제1 특징 순위 리스트 생성부(210)는 n 개의 특징 평가 기법을 이용함으로써 각 특징 평가 기법이 갖는 단점을 극복할 수 있다. 보다 구체적으로 IG 평가 기법은 데이터 세트에서 가장 많은 수가 포함된 제1 특징을 선택하는

것을 추구하고, Chi-Squared 평가 기법은 제1 특징의 크기에 민감하게 반응하며, Gain Ratio와 Symmetric Uncertainty 평가 기법은 데이터 세트에서 가장 적은 수를 갖는 제1 특징에 치우치는 경향이 있다는 단점을 각각 내포하고 있다. 따라서 본 발명의 제1 특징 순위 리스트 생성부(210)는 이러한 각각의 특징 평가 기법의 단점을 극복하기 위하여 다양한 특징 평가 기법을 이용하여 통계적인 편견 없이 특징을 평가할 수 있게 한다.

[0044] 제1 특징 순위 리스트 생성부(210)는 상술한 특징 평가 기법을 이용하여 제1 특징을 평가하여 하나 이상의 제1 특징 순위 리스트를 생성할 수 있다.

[0045] 정규화부(230)는 제1 특징 순위 리스트 생성부(210)에서 생성한 n개의 제1 특징 순위 리스트를 정규화할 수 있다. 정규화부(230)는 도 5에 도시된 알고리즘을 통해 동작할 수 있다. 정규화부(230)는 제1 특징 순위 리스트에 포함된 제1 특징의 평가 수치를 정규화할 수 있다. 보다 구체적으로 정규화부(230)는 제1 특징에 대하여 가장 높은 평가 수치를 1로, 가장 낮은 평가 수치를 0으로 설정하여 제1 특징 평가 수치를 정규화하여 제1 특징 순위 리스트를 정규화할 수 있다. 예를 들어 가장 높은 평가 수치를 갖는 제1 특징의 평가 수치가 0.007384이고 가장 낮은 평가 수치를 갖는 제1 특징의 평가 수치가 0.001361이면, 0.007384를 1에 대응시키고 0.001361을 0에 대응시킬 수 있다.

[0046] 정규화부(230)는 IG 평가 기법을 통해 생성된 제1 특징 순위 리스트에 포함된 제1 특징의 평가 수치를 정규화하여 $IGSR_i$ (Information Gain Scaled Rank)를, Gain Ratio 평가 기법을 통해 생성된 제1 특징의 평가 수치를 정규화하여 $GRSR_i$ (Gain Ratio Scaled Rank)를, Chi-square 평가 기법을 통해 생성된 제1 특징의 평가 수치를 정규화하여 $CSSR_i$ (Chi Square Scaled Rank)를, Symmetrical Uncertainty 평가 기법을 통해 생성된 제1 특징의 평가 수치를 정규화하여 $SUSR_i$ (Symmetrical Uncertainty Scaled Rank)를, Significance 평가 기법을 통해 생성된 제1 특징의 평가 수치를 정규화하여 SSR_i (Significance Scaled Rank)를 생성할 수 있다.

[0047] 제2 특징 순위 리스트 생성부(250)는 정규화된 n개의 제1 특징 순위 리스트를 병합하여 하나의 제2 특징 순위 리스트를 생성할 수 있다. 제2 특징 순위 리스트 생성부(250)는 제1 특징 순위 리스트에 포함된 제1 특징^(j) 각각에 대한 평가 수치를 이용하여 제2 특징 순위 리스트의 제1 특징^(j)으로 설정할 수 있다.

[0048] 먼저 제2 특징 순위 리스트 생성부(250)는 제1 특징 순위 리스트에 포함된 제1 특징^(j) 각각에 대한 평가 수치의 평균 값을 연산하여 제1 특징^(j)에 대한 최종 평가 수치를 생성할 수 있다

$$(\text{final score}_i = \frac{IGSR_i + GRSR_i + CSSR_i + SUSR_i + SSR_i}{n})$$
 예를 들어 $IGSR_1$ 의 평가 수치가 a이고, $GRSR_1$ 의 평가 수치가 b, $CSSR_1$ 의 평가 수치가 c, $SUSR_1$ 의 평가 수치가 d, SSR_1 의 평가 수치가 e이면, 제2 특징 순위 리스트에 포함된 제1 특징^(f1)의 최종 평가 수치는 $\frac{a+b+c+d+e}{5}$ 일 수 있다.

[0049] 제2 특징 순위 리스트 생성부(250)는 제1 특징에 대한 최종 평가 수치를 오름차순 또는 내림차순으로 정렬할 수 있다. 이는 제1 특징의 최종 평가 수치를 보다 직관적으로 확인하기 위함이다.

[0050] 나아가 제2 특징 순위 리스트 생성부(250)는 제1 특징^(j)에 대한 가중치를 더 연산할 수 있다. 제2 특징 순위 리스트 생성부(250)는 수학적 식 5와 같이 n개의 제1 특징 순위 리스트에 포함된 제1 특징^(j) 각각에 대한 평가 수치의 합인 평가 수치 합계를 n개의 제1 특징 순위 리스트에 포함된 모든 제1 특징에 대한 평가 수치의 합으로 나누어 가중치를 연산할 수 있다.

수학식 5

$$\text{Weight} = \frac{\text{combineRank}_i}{\sum_{i=1}^l \text{combineRank}_i}$$

$$\text{combineRank}_i = \sum_{j=1}^n (IGSR_j + GRSR_j + CSSR_j + SUSR_j + SSR_j)$$

[0051]

[0052]

제2 특징 순위 리스트 생성부(250)는 제1 특징의 최종 평가 수치와 가중치를 이용하여 제2 특징 순위 리스트를 생성할 수 있다. 보다 구체적으로 제2 특징 순위 리스트 생성부(250)는 제1 특징에 대한 최종 순위를 제1 특징에 대한 최종 평가 수치에 가중치를 곱하여 제2 특징 순위 리스트를 생성할 수 있다.

[0053]

임계 값 생성부(300)는 제2 특징 순위 리스트 생성부(250)에서 생성된 제2 특징 순위 리스트에 기반하여 유의미한 특징과 무의미한 특징을 구분하기 위한 임계 값을 생성할 수 있다. 임계 값 생성부(300)는 제2 특징 순위 리스트 생성부(250)로부터 제2 특징 순위 리스트를 수신하는 제2 특징 순위 리스트 수신부(310), 제2 특징 순위 리스트를 기반으로 하나 이상의 특징 블록을 생성하는 블록 생성부(330), 특징 블록에 포함된 특징을 이용하여 예측 정확도를 연산하는 연산부(350), 그리고 임계 값을 설정하는 임계 값 설정부(370)를 포함할 수 있다.

[0054]

블록 생성부(330)는 제2 특징 순위 리스트 수신부(310)에서 수신한 제2 특징 순위 리스트를 기반으로 하나 이상의 특징 블록을 생성할 수 있다. 블록 생성부(330)는 제2 특징 순위 리스트에 포함된 제1 특징의 비율의 단위에 따라 하나 이상의 특징 블록을 생성할 수 있다. 예를 들어 블록 생성부(330)는 제2 특징 순위 리스트에 포함된 제1 특징의 수의 5%를 포함하는 특징 블록(FB_1), 10%를 포함하는 특징 블록(FB_2), ..., 100%를 포함하는 특징 블록(FB_{20})을 생성할 수 있다. 위의 예시에는 제1 특징의 수에 대한 비율을 5% 단위로 설정하였으나, 이는 설정된 값으로 다른 비율을 이용할 수도 있다.

[0055]

연산부(350)는 하나 이상의 분류기를 이용하여 특징 블록에 포함된 제2 특징을 각각 분류할 수 있다. 연산부(350)는 제2 특징을 분류하기 위하여 Naive Bayes, J48, k-NN, JRip, 또는 SVM 분류기 중 적어도 하나를 이용할 수 있다. 연산부(350)는 하나 이상의 특징 블록에 포함된 제2 특징을 하나 이상의 분류기를 각각 이용하여 분류할 수 있다. 즉 연산부(350)는 하나의 특징 블록에서 분류기의 수만큼의 분류 결과를 얻을 수 있다.

[0056]

나아가 연산부(350)는 각 분류기를 통해 제2 특징을 분류한 결과에 대한 정확도를 측정할 수 있는데, 이를 위해 k-fold 교차 검증 기법을 이용할 수 있다. 연산부(350)는 특징 블록 모두에 대하여 k-fold 교차 검증을 수행하여 특징 블록 각각에 대한 정확도를 연산할 수 있다. K-fold 교차 검증법은 특징 블록을 k개의 fold로 구분하여 k-1개의 fold를 학습 데이터로, 하나의 fold를 검증 데이터로 하여 특징 블록을 k번 검증하는 기법으로, 특징 블록에 포함된 모든 데이터를 학습 및 검증 데이터로 활용할 수 있어 통계적 신뢰도가 높다.

[0057]

연산부(350)는 k-fold 교차 검증 기법을 이용하여 분류기의 정확도를 측정함에 있어서 오차 행렬(Confusion Matrix)를 더 이용할 수 있다. 오차 행렬은 분류기의 성능을 시각화할 수 있는 표로, TP(True Positives), TN(True Negatives), FP(False Positives), FN(False Negatives)를 이용할 수 있다. 오차 행렬을 보다 쉽게 설명하기 위하여 분류 결과가 1이나 0인 이진 분류기를 이용한다고 가정하면, TP(True Positives)는 분류 결과가 1이고 실제 결과가 1인 것을, TN(True Negatives)는 분류 결과가 0이고 실제 결과가 0인 것을, FP(False Positives)는 분류 결과가 1이고 실제 결과가 0인 것을, FN(False Negatives)는 분류 결과가 0이고 실제 결과가 1인 것을 의미한다. 연산부(350)는 오차 행렬을 이용하여 정확도를 연산할 수 있다 ($\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN}$). 또한 연산부(350)는 분류기의 정확도를 평가하기 위하여 정밀도(Precision)와 민감도(Recall), 그리고 정밀도와 민감도를 모두 고려하는 F-measure를 더 이용할 수 있다.

수학식 6

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F1 - measure} = \frac{2 * (\text{Recall} * \text{Precision})}{\text{Recall} + \text{Precision}}$$

[0058]

[0059]

연산부(350)는 각 특징 블록 별로, 측정된 분류 결과의 정확도에 대한 평균 값을 연산할 수 있다. 연산부(350)는 각 특징 블록 별로 측정된, 분류기의 수만큼의 정확도의 평균 값을 연산할 수 있다.

[0060]

임계 값 설정부(370)는 연산부(350)에서 연산된 평균 값을 그래프에 도식할 수 있다. 임계 값 설정부(370)는 하나 이상의 특징 블록에 포함된 제2 특징의 수를 기준으로 하여 평균 값을 정렬(내림차순)하여 그래프에 도식할 수 있다. 임계 값 설정부(370)는 평균 값이 도식된 그래프에서 기울기의 부호가 변화하지 않는 구간을 식별하고, 식별된 구간의 시작점에 대응되는 제1 특징의 수에 대한 비율을 식별할 수 있다. 예를 들어 임계 값 설정부(370)는 평균 값 a, 0.8a, 0.74a, 0.71a, 0.68a, 0.65a에 대한 그래프를 도식하여 그래프에 도식할 수 있다. 임계 값 생성부는 그래프에 표시된 기울기의 부호가 더 이상 변화하지 않는 0.74a, 0.71a, 0.68a, 0.65a에 대한 구간을 식별하고, 식별된 구간의 시작점이 되는 0.74a를 인식할 수 있다.

[0061]

임계 값 설정부(370)는 제1 특징의 수에 대한 비율을 데이터 세트에 포함된 유용한 특징을 구분하기 위한 임계 값으로 설정할 수 있다.

[0062]

특징 선택부(400)는 제2 특징 순위 리스트에 임계 값(비율)을 적용하여 유용한 특징을 추출할 수 있다. 예를 들어 설명하면, 임계 값으로 설정된 값이 45%이면 특징 선택부(400)는 제2 특징 순위 리스트에 포함된 제1 특징 중 상위 55%(100-45)의 특징을 추출하여 제3 특징으로 식별할 수 있다. 나아가 특징 선택부(400)는 제2 특징 순위 리스트에 포함된 제1 특징 중 상위 55%에 속하지 않는 제1 특징을 제4 특징으로 구분하여 데이터 세트에서 배제할 수 있다.

[0063]

이하에서는 도 2를 이용하여 본 발명의 일 실시 예에 의한 유용한 특징 선택 방법을 설명한다. 유용한 특징 선택 방법에 관한 설명에 있어서 전술한 유용한 특징 선택 장치와 중복되는 세부 실시 예는 생략될 수 있다. 또한 유용한 특징 선택 방법의 주체인 유용한 특징 선택 장치는 서버로 구현될 수 있는 바, 이하에서는 설명의 편의를 위하여 서버로 명명한다.

[0064]

도 2를 참조하면, 서버는 데이터베이스에서 데이터 세트를 로드할 수 있다(S100). 데이터 세트는 이미지 데이터, 문장 데이터, 센서 데이터 등 다양한 데이터를 포함할 수 있으며, 특히 본 발명의 일 실시 예에 의한 데이터 세트는 의료 데이터일 수 있다.

[0065]

서버는 데이터 세트에 포함된 하나 이상의 제1 특징을 추출할 수 있다(S200). 이 때, 데이터 세트는 의료 데이터에 대한 제1 특징이 이미 추출된 데이터 세트일 수 있어서 S200은 생략될 수 있다.

[0066]

서버는 제1 특징에 n개의 특징 평가 기법($n \geq 1$)을 각각 적용하여 n개의 제2 특징 순위 리스트를 생성할 수 있다(S300). 이 때 사용되는 특징 평가 기법은 Information Gain, Gain Ratio, hi-Squared, Symmetric Uncertainty, Significance 등과 같은 단변량 평가 기법을 포함할 수 있고, 또한 OneR, Relief, ReliefF, DRBFS, GR- χ^2 , Borda method, EMFFS 평가 기법을 더 이용할 수 있다.

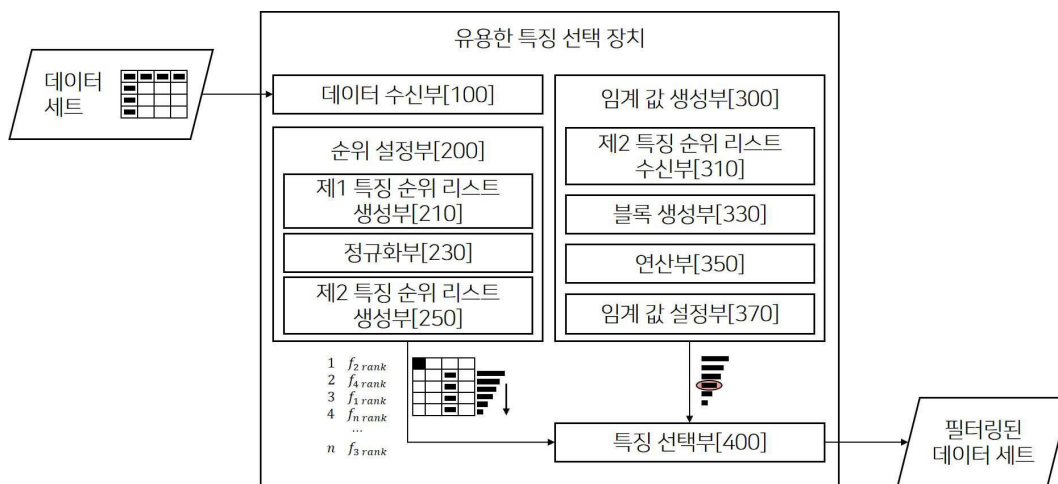
[0067]

서버는 제2 특징 순위 리스트를 생성하기 위하여 먼저, 제1 특징에 n개의 특징 평가 기법을 각각 적용하여 제1 특징의 평가 결과에 따라 n개의 제1 특징 순위 리스트를 생성할 수 있다. 나아가 서버는 생성된 n개의 제1 특징 순위 리스트를 정규화할 수 있다. 서버는 제1 특징 순위 리스트에 포함된 제1 특징의 평가 수치 중 최대 값을 1로, 최소 값을 0으로 하여 제1 특징 순위 리스트에 포함된 제1 특징의 평가 수치를 정규화할 수 있다. 이는 서버의 연산 속도를 향상시키기 위함이다. 서버는 정규화된 n개의 제1 특징 순위 리스트를 병합하여 제2 특징 순위 리스트를 생성할 수 있다(S300).

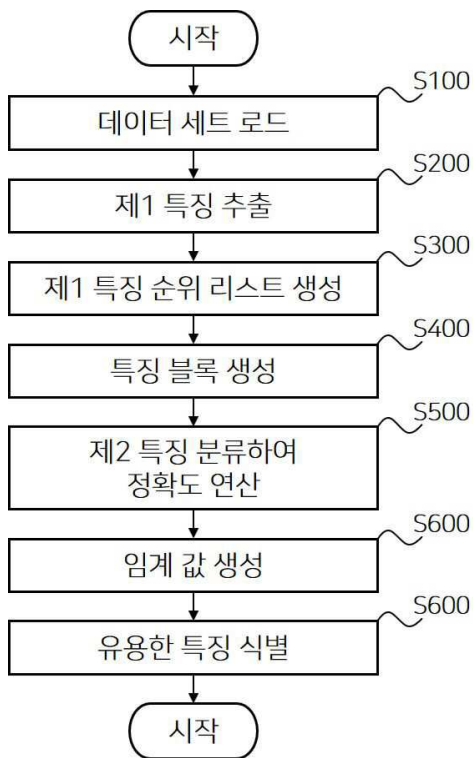
- [0068] 서버는 생성된 제2 특징 순위 리스트를 기반으로 특징 블록을 생성(S400)할 수 있는데, 이 때 제2 특징 순위 리스트에 포함된 제1 특징의 비율에 따라 하나 이상의 특징 블록을 생성할 수 있다. 예를 들어 서버는 제2 특징 순위 리스트에 포함된 제1 특징의 5%를 포함하는 특징 블록, 10%를 포함하는 특징 블록, ..., 100%를 포함하는 특징 블록을 생성할 수 있다.
- [0069] 서버는 특징 블록에 하나 이상의 분류기를 적용하여 특징 블록에 포함된 제2 특징을 분류하고, 이를 검증하기 위하여 정확도를 연산할 수 있다(S500). 서버는 분류 정확도를 연산하기 위하여 k-fold 교차 검증 기법과 오차 행렬(Confusion Matrix)을 이용할 수 있다. 서버는 분류 정확도를 연산하기 위하여, 특징 블록에 대하여 하나 이상의 분류기를 적용하였을 때의 정확도에 대한 평균 값을 연산할 수 있다. 이 때 연산되는 평균 값은 특징 블록 각각에 대하여 연산될 수 있다.
- [0070] 서버는 특징 블록 각각에 대하여 연산된 평균 값을 이용하여 임계 값을 생성할 수 있다(S600). 서버는 특징 블록에 대하여 연산된 평균 값을 그래프에 도시하여, 그래프의 기울기의 부호가 변화하지 않는 구간을 식별할 수 있다. 서버는 해당 구간이 시작하는 지점에 대응되는 특징 블록을 식별하여 해당 특징 블록이 생성되기 위하여 사용된 제1 특징의 비율을 임계 값으로 설정할 수 있다.
- [0071] 서버는 제2 특징 순위 리스트에서 임계 값(비율)을 적용하여 제3 특징(유용한 특징)을 추출할 수 있다. 예를 들어 설명하면, 임계 값으로 설정된 값이 45%이면 서버는 제2 특징 순위 리스트에 포함된 제1 특징 중 상위 55%(100-45)의 특징을 추출하여 제3 특징으로 식별할 수 있다. 서버는 제3 특징으로 식별되지 않은 제1 특징을 제4 특징으로 구분하여 데이터 세트에서 배제할 수 있다.
- [0072] 본 명세서와 도면에 개시된 본 발명의 실시 예들은 본 발명의 기술 내용을 쉽게 설명하고 본 발명의 이해를 돕기 위해 특정 예를 제시한 것뿐이며, 본 발명의 범위를 한정하고자 하는 것은 아니다. 여기에 개시된 실시 예들 이외에도 본 발명의 기술적 사상에 바탕을 둔 다른 변형 예들이 실시 가능하다는 것은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에게 자명한 것이다.

도면

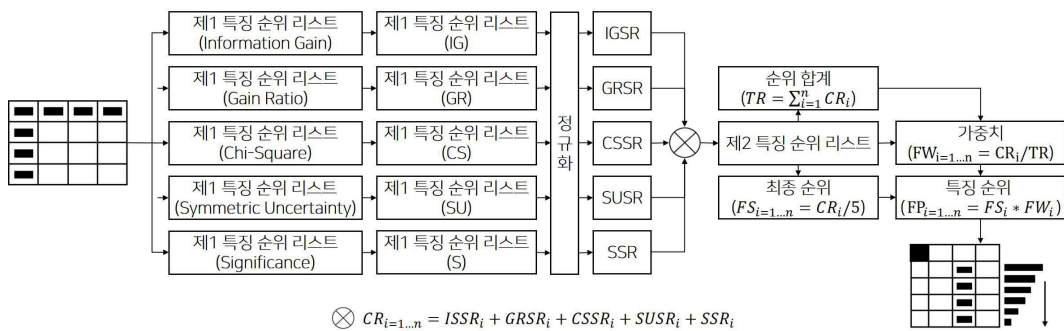
도면1



도면2



도면3



도면4

Algorithm 1: UFS (D)
Input: D: Input data set (data)
Output: FR- Features Ranks

```

1 noOfAttrs ← numAttributes(data) // compute the number of
  attributes;
2 /* Consider n attribute evaluation measures, also called univari-
  ate filter measures (AttrEv1, AttrEv2, AttrEv3,..., and AttrEvn) */;
3 /* Compute the ranks using each selected measure */;
4 CR1[] ← computeRanks(data, AttrEv1) //where CR represents computed
  ranks;
5 CR2[] ← computeRanks(data, AttrEv2);
6 CR3[] ← computeRanks(data, AttrEv3);
7 CRn[] ← computeRanks(data, AttrEvn);
8 /* Compute the scaled ranks of each computed ranks using Algorithm
  2 */;
9 scaledRanks1[] ← scaleRanks(CR1) // invoke Algorithm 2;
10 scaledRanks2[] ← scaleRanks(CR2) // invoke Algorithm 2;
11 scaledRanks3[] ← scaleRanks(CR3) // invoke Algorithm 2;
12 scaledRanksn[] ← scaleRanks(CRn) // invoke Algorithm 2;
13 /* Compute the combined sum of all computed ranks */;
14 combinedRanksSum ← 0;
15 combinedRanks[];
16 for ∀ noOfAttrs ∈ D do
17 /* For each attribute, compute the combined rank by adding all
  computed scaled ranks */;
18 combinedRanksi ←  $\sum_{j=1}^n \text{scaledRanks}_{ji}$  //where n represents the number of
  filter measures;
19 combinedRanksSum = combinedRanksSum + combinedRanksi;
20 end
21 /* Rank the list in ascending order */;
22 sortedRanks[] ← sort(combinedRanks);
23 /* Compute the score, weight, and priority of each attribute */;
24 for ∀ noOfAttrs ∈ D do
25 attrScoresi ← combinedRanksi/n //where n represents number of fil-
  ter measures;
26 attrWeightsi ← combinedRanksi/combinedRanksSum;
27 attrPrioritiesi ← attributesScoresi * attributesWeightsi;
28 /* Assign an index (Rank ID) on ascending order to each attri-
  bute based on its priority value */;
29 FR[] ← assignRank(attrPrioritiesi);
30 end
31 return FR: features ranks

```

도면5

Algorithm 2: Scaling the Computed Ranks (CR)
Input: CR: Input computed ranks (ranks)
Output: SR- Scaled Ranks

```

1 smallest ← ranks0;
2 largest ← ranks0;
3 for ∀ noOfAttrs ∈ CR do
4   if ranki > largest then
5     largest ← ranki;
6   else
7     if ranki < smallest then
8       smallest ← ranki
9     end
10  end
11 end
12 min ← smallest;
13 max ← largest;
14 SR[] ← (ranks - min) / (max - min);
15 return SR: scaled ranks

```

도면6

Algorithm 3: TVS (D, C)

Input: $D = (d_1, d_2, \dots, d_n)$ // set of n datasets with varying complexities
 $C = (c_1, c_2, \dots, c_m)$ // set of m machine learning classifiers
Output: V - cutoff value

```

1 initialization;
2 for  $d_i \leftarrow$  in  $D$  do
3    $d_i \leftarrow \text{computeFeatureRank}(d_i)$  // rank each feature;
4    $d_i \leftarrow \text{sortByRankASC}(d_i)$  // sort features by rank in ASC;
5 end
6  $P \leftarrow 100$ ;
7 for  $d_i \leftarrow$  in  $D$  do
8   while  $P \geq 5$  do
9      $k \leftarrow \text{sizeof}(d_i) * (p/100)$  // compute partition size;
10     $\text{Acc} \leftarrow \text{newSet}()$  // initialize empty set;
11    for  $c_i \leftarrow$  in  $C$  do;
12       $P_{\text{acc}} \leftarrow \text{predictiveAccuracy}(c_i, \text{topKFeatures}(d_i, k))$ ;
13       $\text{Acc.add}(P_{\text{acc}})$  // add accuracy to set;
14    end
15     $\text{AVG}_{\text{acc}} \leftarrow \text{computeAVG}(\text{Acc})$  // compute average accuracy;
16     $G \leftarrow \text{Plot}(\text{AVG}_{\text{acc}}, k)$  // plot the average point;
17     $P \leftarrow P - 5$  // decrease the partition size by 5;
18  end
19 end
20  $V \leftarrow \text{getCutoffValue}(G)$ ;

```