(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2017년04월28일
(11) 등록번호 10-1731626
(24) 등록일자 2017년04월24일

(51) 국제특허분류(Int. Cl.)
G06F 17/30 (2006.01)
(52) CPC특허분류
G06F 17/30327 (2013.01)
G06F 17/30961 (2013.01)
(21) 출원번호 10-2016-0111407
(22) 출원일자 2016년08월31일
심사청구일자 2016년08월31일
(56) 선행기술조사문헌
US07467116 B2*
US20160155069 A1*
KR1020160143512 A*
KR101469136 B1*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
백준걸
서울특별시 강동구 고덕로 131, 122동 1503호(암사동, 강동롯데캐슬퍼스트아파트)
김동환
서울특별시 광진구 광나루로56길 32, 215동 2303호(구의동, 구의현대2단지아파트)
(74) 대리인
김등용, 김홍석

전체 청구항 수 : 총 6 항 심사관 : 김경완

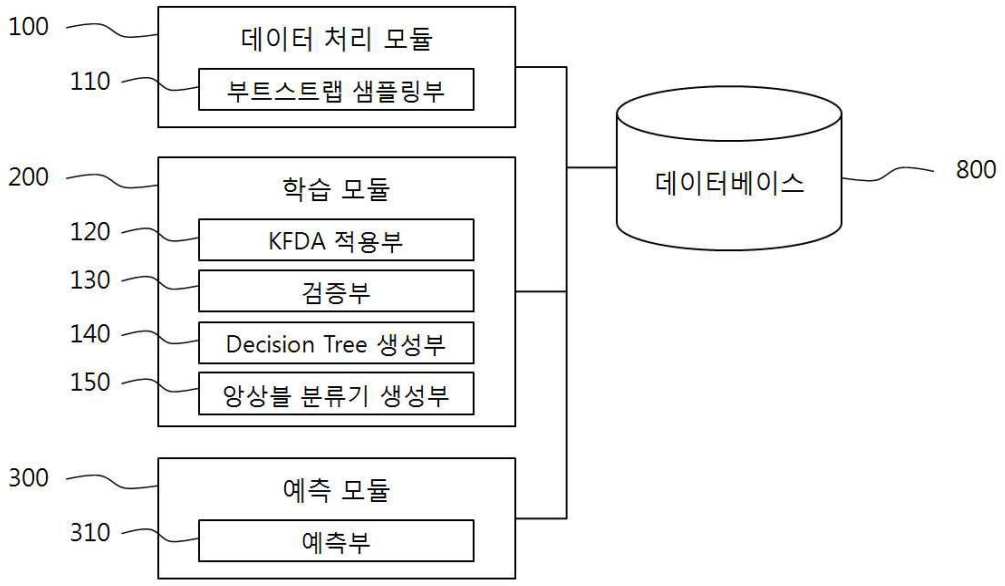
(54) 발명의 명칭 **트리 기반 앙상블 분류기를 이용한 정보 예측 방법 및 시스템**

(57) 요 약

데이터 처리 모듈, 학습 모듈, 예측 모듈, 및 데이터베이스를 포함하는 트리 기반 앙상블 분류기를 이용한 정보 예측 시스템 및 이를 이용한 트리 기반 앙상블 분류기를 이용한 정보 예측 방법이 개시된다. 트리 기반 앙상블 분류기를 이용한 정보 예측 방법은 학습을 위한 데이터를 수집하는 데이터 수집 단계, 상기 데이터를 부트스트랩

(뒷면에 계속)

대 표 도 - 도1



10

샘플링하여 훈련 데이터와 샘플링되지 않은 검증 데이터로 구분하는 부트스트랩 샘플링 단계, 상기 훈련 데이터에 KFDA를 적용하는 KFDA 적용 단계, 샘플링되지 않은 상기 검증 데이터를 이용하여 검증을 수행하고 최적의 커널 파라미터를 추출하는 최적 커널 파라미터 추출 단계, 상기 최적의 커널 파라미터에 따른 의사결정 트리를 생성하는 의사결정 트리 생성 단계, 융합 규칙을 이용하여 적어도 둘 이상의 의사결정 트리를 병합하여, 트리 기반 앙상블 분류기를 생성하는 앙상블 분류기 생성 단계, 및 신규 데이터를 입력 데이터로 하고 상기 트리 기반 앙상블 분류기를 이용하여 상기 신규 데이터의 클래스 라벨(class label)을 예측하는 단계를 포함한다.

이 발명을 지원한 국가연구개발사업
과제고유번호 F16SN26T4601
부처명 교육부
연구관리전문기관 한국연구재단
연구사업명 BK21플러스사업
연구과제명 제조 물류분야에서의 빅데이터 운용사업팀
기 여 율 1/2
주관기관 고려대학교 산학협력단
연구기간 2016.03.01 ~ 2017.02.28
이 발명을 지원한 국가연구개발사업
과제고유번호 1711040094
부처명 미래창조과학부
연구관리전문기관 한국연구재단
연구사업명 개인연구지원
연구과제명 빅 데이터 기반 스마트팩토리 구현을 위한 프리딕티브 에널리틱스 기법 개발
기 여 율 2/2
주관기관 고려대학교 산학협력단
연구기간 2016.06.01 ~ 2017.05.31
공지예외적용 : 있음

명 세 서

청구범위

청구항 1

트리 기반 앙상블 분류기를 이용한 정보 예측 시스템에서 트리 기반 앙상블 분류기를 이용한 정보 예측 방법에 있어서,

학습을 위한 데이터를 수집하는 데이터 수집 단계;

상기 데이터를 부트스트랩 샘플링(Bootstrap sampling)하여 훈련 데이터와 샘플링되지 않은 검증 데이터로 구분하는 부트스트랩 샘플링 단계;

상기 훈련 데이터에 KFDA(Kernel Fisher Discriminant Analysis)를 적용하는 KFDA 적용 단계;

샘플링되지 않은 상기 검증 데이터를 이용하여 검증을 수행하고 최적의 커널 파라미터를 추출하는 최적 커널 파라미터 추출 단계;

상기 최적의 커널 파라미터에 따른 의사결정 트리(Decisoon Tree, D_i)를 생성하는 의사결정 트리 생성 단계;

융합 규칙을 이용하여 적어도 둘 이상의 의사결정 트리를 병합하여, 트리 기반 앙상블 분류기를 생성하는 앙상블 분류기 생성 단계; 및

신규 데이터를 입력 데이터로 하고 상기 트리 기반 앙상블 분류기를 이용하여 상기 신규 데이터의 클래스 라벨(class label)을 예측하는 단계;를 포함하는 트리 기반 앙상블 분류기를 이용한 정보 예측 방법.

청구항 2

제1항에 있어서,

부트스트랩 샘플링 단계, KFDA 적용 단계, 최적 커널 파라미터 추출 단계, 및 의사결정 트리 생성 단계를 반복하여 L 개의 의사결정 트리(D_1, D_2, ... , D_L)를 생성하는 단계를 더 포함하는 트리 기반 앙상블 분류기를 이용한 정보 예측 방법

청구항 3

제1항에 있어서,

상기 융합 규칙은 다수결(majority voting)인 트리 기반 앙상블 분류기를 이용한 정보 예측 방법

청구항 4

학습을 위한 데이터를 수집하고, 상기 데이터를 부트스트랩 샘플링(Bootstrap sampling)하여 훈련 데이터와 샘플링되지 않은 검증 데이터로 구분하는 데이터 처리 모듈,

상기 훈련 데이터에 KFDA(Kernel Fisher Discriminant Analysis)를 적용하고, 샘플링되지 않은 상기 검증 데이터를 이용하여 검증을 수행하고 최적의 커널 파라미터를 추출하여, 상기 최적의 커널 파라미터에 따른 의사결정 트리(Decisoon Tree, D_i)를 생성하고, 융합 규칙을 이용하여 상기 의사결정 트리를 병합하여, 트리 기반 앙상블 분류기를 생성하는 학습 모듈, 및

상기 트리 기반 앙상블 분류기를 이용하여 신규 데이터의 클래스 라벨(class label)을 예측하는 예측 모듈을 포함하는 트리 기반 앙상블 분류기를 이용한 정보 예측 시스템.

청구항 5

제4항에 있어서,

상기 학습 모듈은 L 개의 의사결정 트리(D_1, D_2, ... , D_L)를 생성하며,

상기 데이터 처리 모듈은 각각의 의사결정 트리를 형성할 때마다 새롭게 부트스트랩 샘플링을 수행하여 상기 훈련 데이터 및 상기 검증 데이터를 생성하는 트리 기반 앙상블 분류기를 이용한 정보 예측 시스템.

청구항 6

제4항에 있어서,

상기 융합 규칙은 다수결(majority voting)인 트리 기반 앙상블 분류기를 이용한 정보 예측 시스템.

발명의 설명

기 술 분 야

[0001] 본 발명의 개념에 따른 실시 예는 트리 기반 앙상블 분류기를 이용한 정보 예측 방법 및 시스템에 관한 것으로, 더욱 상세하게는, 트레이닝 데이터를 부트스트랩 샘플링하고 변수들을 랜덤하게 쪼개어 서브셋을 만든 후, 서브셋에 KFDA를 적용하고 검증을 통해 최적의 커널 파라미터를 추출하여 의사결정트리를 생성하고 각각의 의사결정트리를 병합하여 앙상블 분류기를 생성하고, 이를 이용하여 정보를 분류하고 예측하는 방법 및 시스템에 관한 것이다.

배 경 기 술

[0002] 본 발명은 시뮬레이션 및 실제 현실에서 발생되는 데이터(정보)의 분류 기법에 관한 것으로, 다양한 사례(instance)와 특성(feature)을 갖는 데이터의 클래스 라벨(class label)을 정확하게 예측하는 기법에 관한 것이다.

[0003] 과거의 데이터들은 대부분 적은 수의 변수(variable)와 선형의 데이터 구조를 갖는 경우가 많았기 때문에 기존의 알고리즘으로도 충분한 예측/분류 결과를 획득할 수 있었다. 그러나 ICT 및 센서(sensor) 기술의 발달로 인하여 제조공정이나 유전공학 분야에서는 수백 수천에 달하는 변수를 가진 데이터들이 생성되기 시작하였다.

[0004] 다양한 사례(instance)와 특성(feature)을 갖는 데이터의 클래스 라벨(class label)을 정확하게 예측하고 분류하는 다양한 트리(tree) 기반의 알고리즘들이 있지만, 변수가 증가할수록 예측 정확도가 떨어지는 경우가 많다. 이는 변수가 증가할수록 차원의 증가로 인한 문제와 데이터의 노이즈 등 데이터 분석에 어려움이 발생하고, 복잡한 데이터 구조 때문에 입력 공간(input data space)에서 기존의 알고리즘을 적용하는데 어려움이 있기 때문에, 데이터 사례(instance)의 클래스 라벨(Class label)을 정확하게 예측하기 어렵다. 따라서 변수의 수를 줄이지 않으면서도 정확하게 데이터 사례(instance)의 클래스 라벨(class label)를 예측하는 알고리즘이 필요하다.

발명의 내용

해결하려는 과제

[0005] 본 발명이 이루고자 하는 기술적인 과제는 많은 수의 변수를 가진 복잡한 구조의 데이터에 있어서 각 데이터 사례(instance)의 클래스 라벨(class label)을 정확하게 예측하는 것이다.

과제의 해결 수단

[0006] 본 발명의 일 실시 예에 따른 트리 기반 앙상블 분류기를 이용한 정보 예측 방법은 학습을 위한 데이터를 수집하는 데이터 수집 단계, 상기 데이터를 부트스트랩 샘플링(Bootstrap sampling)하여 훈련 데이터와 샘플링되지 않은 검증 데이터로 구분하는 부트스트랩 샘플링 단계, 상기 훈련 데이터에 KFDA(Kernel Fisher Discriminant Analysis)를 적용하는 KFDA 적용 단계, 샘플링되지 않은 상기 검증 데이터를 이용하여 검증을 수행하고 최적의 커널 파라미터를 추출하는 최적 커널 파라미터 추출 단계, 상기 최적의 커널 파라미터에 따른 의사결정 트리(Decisoon Tree, D_i)를 생성하는 의사결정 트리 생성 단계, 융합 규칙을 이용하여 적어도 둘 이상의 의사결정 트리를 병합하여, 트리 기반 앙상블 분류기를 생성하는 앙상블 분류기 생성 단계, 및 신규 데이터를 입력 데이터로 하고 상기 트리 기반 앙상블 분류기를 이용하여 상기 신규 데이터의 클래스 라벨(class label)을 예측하는 단계를 포함한다.

[0007] 본 발명의 일 실시 예에 따른 트리 기반 앙상블 분류기를 이용한 정보 예측 시스템은 학습을 위한 데이터를 수집하고, 상기 데이터를 부트스트랩 샘플링(Bootstrap sampling)하여 훈련 데이터와 샘플링되지 않은 검증 데이

터로 구분하는 데이터 처리 모듈, 상기 훈련 데이터에 KFDA(Kernel Fisher Discriminant Analysis)를 적용하고, 샘플링되지 않은 상기 검증 데이터를 이용하여 검증을 수행하고 최적의 커널 파라미터를 추출하여, 상기 최적의 커널 파라미터에 따른 의사결정 트리(Decisoon Tree, D_i)를 생성하고, 융합 규칙을 이용하여 상기 의사결정 트리를 병합하여, 트리 기반 앙상블 분류기를 생성하는 학습 모듈, 및 상기 트리 기반 앙상블 분류기를 이용하여 신규 데이터의 클래스 라벨(class label)을 예측하는 예측 모듈을 포함한다.

발명의 효과

[0008] 본 발명의 실시 예에 따른 트리 기반 앙상블 분류기를 이용한 정보 예측 방법 및 시스템은 변수가 많은 고차원의 입력 데이터에 대하여 기존의 기법보다 정확하게 분류예측할 수 있는 효과가 있다.

[0009] 또한, 본 발명의 실시 예에 따른 트리 기반 앙상블 분류기를 이용한 정보 예측 방법 및 시스템은 KPCA와 LDA를 이용하여 KFDA를 구현하여 앙상블의 다양성(diversity)을 향상시키는 효과가 있다.

도면의 간단한 설명

[0010] 본 발명의 상세한 설명에서 인용되는 도면을 보다 충분히 이해하기 위하여 각 도면의 상세한 설명이 제공된다.

도 1은 본 발명의 일 실시 예에 따른 트리 기반 앙상블 분류기를 이용한 정보 예측 시스템의 기능 블럭도이다.

도 2는 본 발명의 일 실시 예에 따른 트리 기반 앙상블 분류기를 이용한 정보 예측 방법의 프레임워크를 도시한다.

도 3은 본 발명의 일 실시 예에 따른 트리 기반 앙상블 분류기를 이용한 정보 예측 방법을 설명하기 위한 흐름도이다.

도 4는 본 발명의 일 실시 예에 따른 트리 기반 앙상블 분류기를 이용하여 시뮬레이션을 수행하기 위한 입력 데이터의 예시적인 도면이다.

도 5는 본 발명의 일 실시 예에 따른 트리 기반 앙상블 분류기를 이용하여 도 4의 입력 데이터를 시뮬레이션한 결과를 도시한 표이다.

도 6은 본 발명의 일 실시 예에 따른 트리 기반 앙상블 분류기를 생성하기 위한 예시적인 수도 코드(Pseudo code)이다.

발명을 실시하기 위한 구체적인 내용

[0011] 본 명세서에 개시되어 있는 본 발명의 개념에 따른 실시 예들에 대해서 특정한 구조적 또는 기능적 설명은 단지 본 발명의 개념에 따른 실시 예들을 설명하기 위한 목적으로 예시된 것으로서, 본 발명의 개념에 따른 실시 예들은 다양한 형태들로 실시될 수 있으며 본 명세서에 설명된 실시 예들에 한정되지 않는다.

[0012] 본 발명의 개념에 따른 실시 예들은 다양한 변경들을 가할 수 있고 여러 가지 형태들을 가질 수 있으므로 실시 예들을 도면에 예시하고 본 명세서에서 상세하게 설명하고자 한다. 그러나 이는 본 발명의 개념에 따른 실시 예들을 특정한 개시 형태들에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물, 또는 대체물을 포함한다.

[0013] 본 명세서에서 사용한 용어는 단지 특정한 실시 예를 설명하기 위해 사용된 것으로서, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 본 명세서에 기재된 특징, 숫자, 단계, 동작, 구성 요소, 부분품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성 요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0014] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0015] KPCA는 Kernel 기반의 PCA(Principal Component Analysis)를 수행하는 알고리즘으로, 다양한 커널을 사용하여 비선형 데이터 구조(non-linear data structure)를 이해하는데 도움을 준다. PCA는 기본적으로 데이터의 분산을 최대한 보존하는 방향으로 서로 수직인 새로운 기저(basis)를 찾는 방법이다. 이를 통해서 데이터의 차원을 축소하기도 하고, 새로운 기저(basis)를 특질로 사용하기도 한다. KPCA는 명시적인 데이터의 커널 매핑(kernel mapping) 없이 커널 트릭(kernel trick)을 이용하여 입력 데이터(input data)를 커널 특징 공간(kernel feature space)으로 매핑(mapping) 한다. 이때 입력 공간(input space)에서는 비선형적(non-linear)이고, 비분리적(non-separable)인 구조가, 커널 특징 공간(kernel feature space)과 같이 고차원(high-dimensional)일 경우에는 분리될 가능성이 조금 더 높다.

[0016] LDA(Linear Discriminant Analysis)는 각 클래스(Class)에 속하는 데이터들 간의 분산(within-class scatter)은 최소화하면서 각 클래스간의 분산(between-class scatter)은 최대화하는 프로젝션(projection)을 찾는 알고리즘이다. LDA는 트레이닝 데이터(Training data)를 통하여 최적의 프로젝션(projection)을 찾고, 테스트 데이터(Test data)에 적용하여 테스트 데이터(Test data)의 클래스(Class)를 예측하게 된다.

[0017] 본 발명에서 데이터의 특징을 추출하고 각 데이터 사례(instance)의 클래스 라벨(Class label)을 정확하게 예측하기 위하여, KFDA(Kernel Fisher Discriminant Analysis) 기법(Mika, S. (2003). Kernel fisher discriminants. PhD thesis, University of Technology, Berlin.)을 이용한다. KFDA는 KPCA와 마찬가지로 커널(kernel) 기반의 LDA를 수행하는 알고리즘이다. 이때 KFDA는 KPCA와 LDA의 조합과 정확하게 일치한다. 즉, KFDA는 데이터의 클래스(Class) 정보를 이용하여, 커널 특징 공간(kernel feature space) 상에서 분리 가능한 구조를 갖는 프로젝션(projection)을 찾는 기법이다.

[0018] 먼저, 도 1을 참조하여, 본 발명의 일 실시예에 따른 트리 기반 앙상블 분류기를 이용한 정보 예측 시스템에 대해 상술한다. 도 1은 본 발명의 일 실시 예에 따른 트리 기반 앙상블 분류기를 이용한 정보 예측 시스템의 기능 블럭도이다. 도 1을 참조하면, 트리 기반 앙상블 분류기를 이용한 정보 예측 시스템(10)은 데이터를 수집하고 샘플링하는 데이터 처리 모듈(100), 수집된 데이터를 학습하여 트리 기반 앙상블 분류기(모델)을 생성하는 학습 모듈(200), 신규의 데이터에 대하여 앙상블 분류기를 적용하여 데이터 사례(instance)의 클래스 라벨(Class label)을 예측하는 예측 모듈(300), 및 데이터베이스(800)를 포함한다.

[0019] 본 명세서에서 사용되는 '-부' 또는 '모듈'이라 함은 본 발명의 기술적 사상을 수행하기 위한 하드웨어 및 상기 하드웨어를 구동하기 위한 소프트웨어의 기능적, 구조적 결합을 의미할 수 있다. 예컨대, 상기 '-부' 또는 '-모듈'은 소정의 코드와 상기 소정의 코드가 수행되기 위한 하드웨어 리소스의 논리적인 단위를 의미할 수 있으며, 반드시 물리적으로 연결된 코드를 의미하거나 한 종류의 하드웨어를 의미하는 것은 아니다.

[0020] 트리 기반 앙상블 분류기를 이용한 정보 예측 시스템(10)의 데이터 처리 모듈(100)은 부트스트랩 샘플링부(110)를 포함한다.

[0021] 부트스트랩 샘플링부(110)는 원데이터(original data)를 부트스트랩 샘플링(Bootstrap sampling)하여 훈련 데이터(training data, TD)와 샘플링(sampling)되지 않은 검증 데이터(validation data, VD)로 구분한다. 바람직하게는 원데이터(original data)에 대하여 훈련 데이터(TD)를 75%로 부트스트랩 샘플링(bootstrap sampling)을 한다.

[0022] 부트스트랩 샘플링부(110)는 KFDA를 적용하고 의사결정 트리(Decision Tree)를 형성할 때마다 새로이 부트스트랩 샘플링을 수행하고 훈련 데이터 및 검증 데이터를 생성한다. 부트스트랩 샘플링(bootstrap sampling) 후, 변수들을 랜덤하게 쪼개어 K개의 서브셋(subset)으로 만든다.

[0023] 모델을 학습시킬 때 사용되는 부트스트랩 샘플링 및 변수들을 랜덤하게 쪼개어 만드는 서브셋들을 통해서 앙상블 모델의 다양성(diversity)을 크게 향상시킬 수 있다.

[0024] 트리 기반 앙상블 분류기를 이용한 정보 예측 시스템(10)의 학습 모듈(200)은 KFDA 적용부(210), 검증부(220), 의사결정트리 생성부(230), 및 앙상블 분류기 생성부(240)를 포함한다.

[0025] KFDA 적용부(210)는 부트스트랩 샘플링(bootstrap sampling) 결과 생성된 훈련 데이터(training data)에 KFDA를 적용한다. 방사 기저 함수(Radial Basis Function, RBF) 커널을 사용하여 KFDA를 적용할 수 있다. KFDA를 적용하여 얻어진 프로젝션 패턴(projection pattern)은 새롭게 얻어진 LDs(linear discriminants)와 평행하기 때문에 트리(Tree) 기반의 앙상블 분류기를 사용하기에 아주 적합하다.

[0026] 검증부(220)는 샘플링되지 않은 검증 데이터(validation data)를 이용하여 검증을 수행하고 최적의 커널 파라미

터를 추출한다.

[0027] 의사결정트리 생성부(230)는 최적의 커널 파라미터에 따른 의사결정 트리(Decisoon Tree, D_i)를 L(L은 2 이상의 자연수) 개를 생성한다.

[0028] 앙상블 분류기 생성부(240)는 L 개의 의사결정 트리(D_1, D_2, ... , D_L)를 병합하여 트리 기반 앙상블 분류기를 생성한다. 의사결정 트리를 병합하는데 사용하는 융합 규칙(fusion rule)은 다수결(majority voting)일 수 있다.

[0029] 트리 기반 앙상블 분류기를 이용한 정보 예측 시스템(10)의 예측 모듈(300)은 예측부(310)를 포함한다.

[0030] 트리 기반 앙상블 분류기를 이용한 정보 예측 시스템(10)의 예측 모듈(300)은 클래스 라벨(class label) 예측이 요구되는 신규 데이터를 입력받아 학습 모듈에서 생성된 트리 기반 앙상블 분류기를 이용하여 각 데이터의 클래스 라벨(class label)을 예측한다. 예측 결과를 데이터베이스(800)에 저장하고 출력할 수 있다.

[0031] 데이터베이스(800)는 학습을 위하여 수집된 데이터를 저장한다. 또한, 수집된 데이터를 부트스트랩 샘플링한 트레이닝 데이터 및 검증 데이터를 저장할 수 있다. 또한, 트레이닝 데이터에 KFDA를 적용한 결과, 검증 데이터를 이용하여 추출된 커널 파라미터, 및 앙상블 분류기를 저장할 수 있다. 또한, 신규 데이터 및 이에 대한 예측 결과를 저장할 수 있다. 본 명세서에서 데이터베이스라 함은, 각각의 데이터베이스에 대응되는 정보를 저장하는 소프트웨어 및 하드웨어의 기능적 구조적 결합을 의미할 수도 있다.

[0032] 제어모듈(미도시)은 트리 기반 앙상블 분류기를 이용한 정보 예측 시스템(10) 전반적인 동작을 제어한다. 즉, 데이터 처리 모듈(100), 학습 모듈(200), 예측 모듈(300), 및 데이터베이스(800)의 동작을 제어할 수 있다. 이와는 달리, 트리 기반 앙상블 분류기를 이용한 정보 예측 시스템(10)의 각 모듈은 별도의 장치로 구성될 수도 있다. 이때, 각각의 장치별로 제어모듈를 각각 구비할 수 있다.

[0033] 이하, 도 2 내지 도 6을 참조하여, 본 발명의 일 실시예에 따른 트리 기반 앙상블 분류기를 이용한 정보 예측 시스템을 이용한 트리 기반 앙상블 분류기를 이용한 정보 예측 방법에 대하여 자세히 살펴보도록 한다.

[0034] 도 2는 본 발명에 따른 트리 기반 앙상블 분류기를 이용한 정보 예측 방법의 프레임워크를 도시한 도면이고, 도 3은 도 1에 도시한 트리 기반 앙상블 분류기를 이용한 정보 예측 시스템을 이용한 트리 기반 앙상블 분류기를 이용한 정보 예측 방법을 설명하기 위한 흐름도이다.

[0035] 도 2 및 도 3을 참조하면, 트리 기반 앙상블 분류기를 이용한 정보 예측 방법은 데이터를 수집하고 샘플링하는 데이터 처리 단계(S100), 수집된 데이터를 학습하여 트리 기반 앙상블 분류기(모델)을 생성하는 학습 단계(S200), 및 신규의 데이터에 대하여 앙상블 분류기를 적용하여 데이터 사례(instance)의 클래스 라벨(Class label)을 예측하는 예측 단계(S300)를 포함한다.

[0036] 먼저, 트리 기반 앙상블 분류기를 이용한 정보 예측 시스템의 데이터 처리 모듈(100)은 학습을 위한 데이터(original data)를 수집한다(S110).

[0037] 다음, 트리 기반 앙상블 분류기를 이용한 정보 예측 시스템의 데이터 처리 모듈(100)은 원데이터(original data)를 부트스트랩 샘플링(Bootstrap sampling)하여 트레이닝 데이터(training data, TD)와 샘플링(sampling)되지 않은 검증 데이터(validation data, VD)로 구분한다(S120). 바람직하게는 원데이터(original data)에 대하여 트레이닝 데이터(training data, TD)를 75%로 부트스트랩 샘플링(bootstrap sampling)을 한다. 이때, 트레이닝 데이터와 검증 데이터는 KFDA를 적용하고 의사결정 트리(Decision Tree)를 형성할 때마다 새롭게 생성된다. 부트스트랩 샘플링(bootstrap sampling) 후, 변수들을 랜덤하게 쪼개어 K 개의 서브셋(subset)으로 만든다. 모델을 학습시킬 때 사용되는 부트스트랩 샘플링 및 변수들을 랜덤하게 쪼개어 만드는 서브셋들을 통해서 앙상블 모델의 다양성(diversity)이 크게 향상된다.

[0038] 부트스트랩 샘플링(bootstrap sampling) 결과 생성된 훈련 데이터(training data)에 KFDA를 적용한다(S210). 바람직하게는 방사 기저 함수(Radial Basis Function, RBF) 커널을 사용하여 KFDA를 적용한다. 구현된 KFDA를 통해서 얻어진 프로젝션 패턴(projection pattern)은 새롭게 얻어진 LDs(linear discriminants)와 평행하기 때문에 트리(Tree) 기반의 앙상블 분류기를 사용하기에 아주 적합하며, 기존의 분류기보다 더 나은 분류예측 성능을 제공할 수 있다.

[0039] 다음, 샘플링되지 않은 검증 데이터(validation data)를 이용하여 검증 단계를 수행하고 최적의 커널 파라미터를 추출한다(S220).

[0040] 예를 들어, 목적함수 $J(w)$가 아래의 식과 같을 때, 목적함수 $J(w)$를 최대화하는 변환행렬(w), 즉, 클래스 내 분산을 최대화하면서 클래스 내 분산을 최소화하는 변환행렬(w)를 추출한다. 이때, $\Phi(x)$는 매핑 함수를 의미한다.

[0041] $$J(w) = \frac{w^T S_B^{\Phi} w}{w^T S_W^{\Phi} w}$$

[0042] 이때,

[0043] $$S_B^{\Phi} = (m_1^{\Phi} - m_2^{\Phi})(m_1^{\Phi} - m_2^{\Phi})^{\mathrm{T}},$$

[0044] $$S_W^{\Phi} = \sum_{i=1,2} \sum_{w \in x_i} (\Phi(x) - m_i^{\Phi})(\Phi(x) - m_i^{\Phi})^{\mathrm{T}}$$

[0045] $$m_i = \frac{1}{l_i} \sum_{j=1}^{l_i} \Phi(x_j^i)$$

[0046] 다음, 최적의 커널 파라미터에 따른 의사결정 트리(Decisoon Tree, D_i)를 생성한다(S230). 즉, 의사결정 트리(Decisoon Tree, D_i)를 통해 모델에 학습된다.

[0047] 부트스트랩 샘플링을 통하여 새롭게 생성된 트레이닝 데이터(training data)에 KPCA와 LDA를 결합한 KFDA가 적용되고, 검증 데이터(validation data)를 기반으로 최적의 커널 파라미터를 추출하고, 이에 따른 의사결정 나무(Decisoon Tree, D_i)를 생성하는 단계를 반복하여(S120-S230 또는 S210-S230), L(L은 2 이상의 자연수) 개의 의사결정 트리(D_1, D_2, ... , D_L)를 생성한다.

[0048] 다음, 융합 규칙(fusion rule)을 통하여 L 개의 의사결정 트리(Decisoon Tree, D_i)를 병합하여(S240), 트리 기반 앙상블 모델을 생성한다(S250). 바람직하게는, 의사결정 트리 병합을 위하여 사용하는 융합 규칙(fusion rule)은 다수결(majority voting)일 수 있다.

[0049] 다음, 클래스 라벨(class label) 예측이 요구되는 신규 데이터를 입력받는다(S310).

[0050] 신규 데이터에 대하여 학습 단계(S200)에서 생성된 트리 기반 앙상블 모델을 기반으로 각 데이터의 클래스 라벨(class label)을 예측한다(S320).

[0051] 도 4는 트리 기반 앙상블 분류기를 이용하여 시뮬레이션을 수행하기 위한 입력 데이터의 예시적인 도면이고, 도 5는 도 4의 입력 데이터를 시뮬레이션한 결과를 도시한 표이다.

[0052] 도 4의 (a)는 두 개의 나선형(two spiral) 구조의 시뮬레이션 데이터이고, 도 4의 (b)는 circle in a square 구조의 시뮬레이션 데이터이다. 도 5는 도4의 시뮬레이션 데이터를 5-묶음 교차 검증법(5-fold cross validation)을 기반으로, 앙상블 사이즈를 50으로 하여 Bagging, Adaboost, Random Forest, Rotation Forest 및 KFDA 기법을 이용하여 시뮬레이션 결과표이다. 도 5를 참조하면, 본 발명에 ㄸ따른 KFDA 기법을 적용하는 경우, 기존의 기법들(Bagging, Adaboost, Random Forest, Rotation Forest)에 비하여 정확도(accuracy)가 높고, 표준 오차(standard error)가 낮은 것을 확인할 수 있다. 특히, 나선형(two spiral) 구조에서 보다 정확한 분류 예측 성능을 보인다.

[0053] 도 6은 본 발명의 일 실시 예에 따른 트리 기반 앙상블 분류기를 생성하기 위한 예시적인 수도 코드(Pseudo code)이다. X는 훈련 데이터(training data), Y는 클래스 라벨, L은 분류기의 개수(즉, 앙상블 사이즈), K는 서브셋(subset)의 개수라고 할 때 앙상블 분류기(Ensemble Classifier)를 만들기 위해서, 75%로 부트스트랩 샘플링(bootstrap sampling)하여 훈련 데이터($X_{i,j}$)를 생성하고 변수들을 랜덤하게 쪼개어 K개의 서브셋(subset)으로 만든다. 이때, $X_{i,j}$는 i번째 의사결정 트리를 위하여 샘플링한 j번째 서브셋의 훈련 데이터를 의미한다. 이렇

게 만들어진 서브셋(subset)에 KFDA(Kernel Fisher Discriminant Analysis)를 적용하고 샘플링되지 않은 데이터(X_{val})를 이용한 검증(validation)을 통해서 최적의 커널 파라미터($C_{i,j}$)를 찾은 뒤, L 개의 의사결정 나무(D_i)를 생성하고 결합하여 앙상블 분류기를 생성한다. 다음, 신규의 데이터 사례(instance)에 대하여 앙상블 분류기를 이용하여 클래스 라벨 예측을 수행한다.

[0054] 본 발명은 도면에 도시된 실시 예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시 예가 가능하다는 점을 이해할 것이다. 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 등록청구범위의 기술적 사상에 의해 정해져야 할 것이다.

부호의 설명

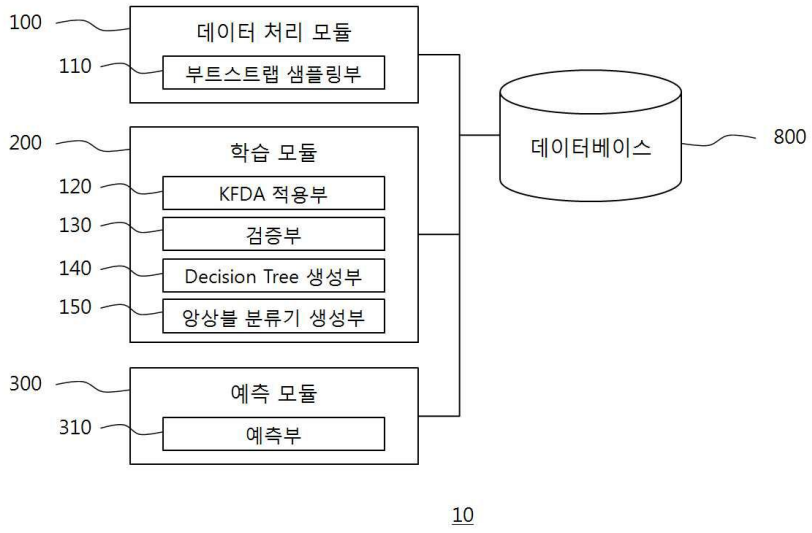
[0055] 100 : 트리 기반 앙상블 분류기를 이용한 정보 예측 시스템

100 : 데이터 처리 모듈 200 : 학습 모듈

300 : 예측 모듈 800 : 데이터베이스

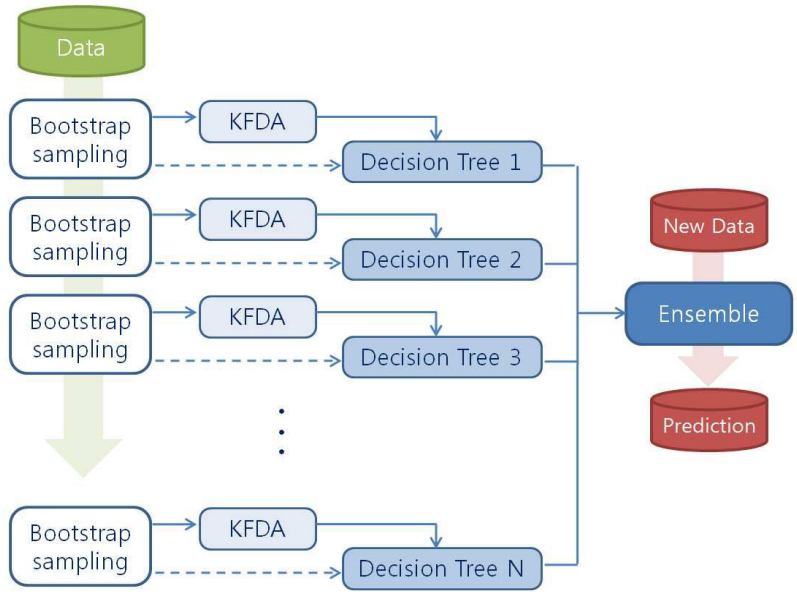
도면

도면1

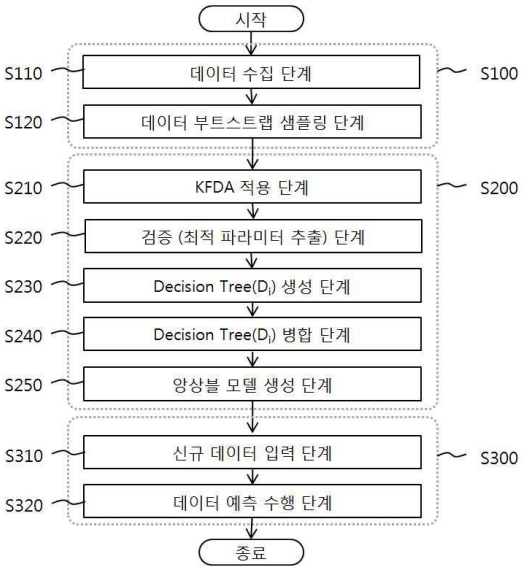


10

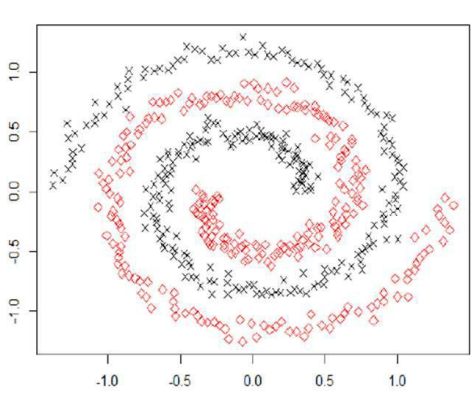
도면2


Data
Bootstrap sampling
KFDA
Decision Tree 1
Bootstrap sampling
KFDA
Decision Tree 2
Bootstrap sampling
KFDA
Decision Tree 3
Bootstrap sampling
KFDA
Decision Tree N
New Data
Ensemble
Prediction

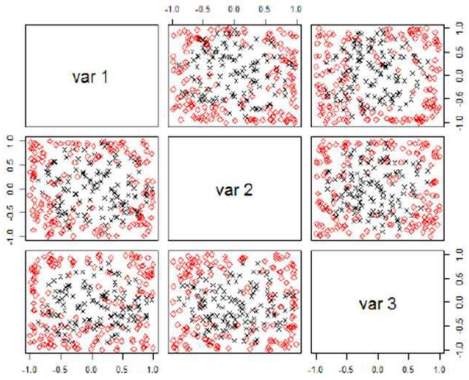
도면3


시작
S110
데이터 수집 단계
S100
S120
데이터 부트스트랩 샘플링 단계
S210
KFDA 적용 단계
S200
S220
검증 (최적 파라미터 추출) 단계
S230
Decision Tree(Di) 생성 단계
S240
Decision Tree(Di) 병합 단계
S250
앙상블 모델 생성 단계
S310
신규 데이터 입력 단계
S300
S320
데이터 예측 수행 단계
종료

도면4



(a) Two spiral



(b) Circle in a Square

도면5

Dataset	Instances	Dimension	Class	Method	Bagging	AdaBoost	Random Forest	Rotation Forest	KFDA Forest
Two spiral	500	2	2	Accuracy (Standard error)	0.9676 (0.0068)	0.9708 (0.0041)	0.9712 (0.0074)	0.9244 (0.1602)	**0.9984* (0.0008)**
Circle in a Square	500	3	2	Accuracy (Standard error)	0.8913 (0.0112)	0.9167 (0.0108)	0.8780 (0.0117)	0.8287 (0.0084)	**0.9433* (0.0058)**

* with bold denote the highest accuracy

도면6

Training Phase

Given:

- X : Training Data
- Y : the labels of the Training Data
- L : the number of classifiers, i.e. ensemble size
- K : the number of subsets

Output: ensemble of decision tress $D_1, D_2, \ldots, D_L$

- For i = 1 ,..., L
- In Train dataset X, data randomly split F (the feature training set) into K subsets : $F_{i,j}$ (for j = 1 ... K)
  - Prepare the rotation matrix R_i^a :
  - For j = 1 ,..., K do
- Make a bootstrap sample $X_{i,j}$ from in $F_{i,j}$ (with sample size of 75% of the number of instances in $X_{i,j}$).
- Remaining data is being automatically validation set X_{val}

Validation phase

- For sigma = 1e-7 ,..., 10 do
- Apply KPCA plus LDA on the sample $X_{i,j}$ to obtain a coefficient matrix $C_{i,j}$
- Fit the model of the coefficient matrix and compare the accuracy of each models by using Validation set X_{val}
- Get the best model and sigma by grid search.

- Arrange the best coefficient matrix of $C_{i,j}$ into a single rotation matrix
- Store information of features of each subset and construct the rotation matrix R_i^a by rearranging by the rows of best $C_{i,j}$ so that they match to the original order of feature set in F

- Build a tree-based classifier D_i using (XR_i^a, Y)

Classification Phase

Given: data instance x, ensemble $D_1, D_2, \ldots, D_L$

Output:

For a given data instance x, the predicted class label from each classifier D is:

$$D = \frac{1}{L}\sum_{i=1}^{L} D_i(XR_i^a)$$