(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2012-0024074
(43) 공개일자 2012년03월14일

(51) 국제특허분류(Int. Cl.)
G06F 17/15 (2006.01) *G06F 17/10* (2006.01)
(21) 출원번호 10-2010-0086705
(22) 출원일자 2010년09월03일
심사청구일자 2010년09월03일

(71) 출원인
경희대학교 산학협력단
경기도 용인시 기흥구 덕영대로 1732, 국제캠퍼스 내 (서천동, 경희대학교)
(72) 발명자
이승룡
경기도 성남시 분당구 구미동 삼환빌라 102-102
이영구
경기도 수원시 영통구 덕영대로 1709, 706호 (영통동, 경희유니빌)
라더빈
경기도 용인시 기흥구 덕영대로 1732, 전자정보대학 351호 (서천동, 경희대학교)
(74) 대리인
특허법인 신지

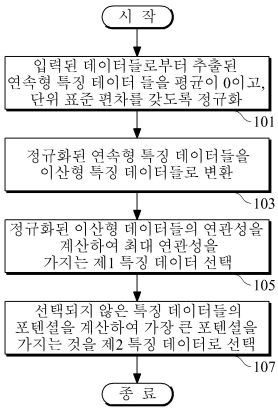
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 **패턴 인식을 위한 특징 데이터 선택 방법 및 그 방법을 이용한 특징 데이터 선택 장치**

(57) 요 약

패턴 인식을 위해 특징 데이터의 선택을 위한 상호 정보 계산시 특징 데이터의 연관성을 나타내는 클래스-특징 상호 정보와 특징 데이터의 중복성을 나타내는 특징-특징 상호 정보 간의 불균형을 제거하기 위한 것으로, 일 실시예에 따른 패턴 인식을 위한 특징 데이터 선택 방법은 입력된 데이터로부터 추출된 연속형 특징 데이터들을 평균이 0이고, 단위 표준 편차를 갖도록 정규화하는 단계와, 정규화된 연속형 특징 데이터들을 이산형 특징 데이터들로 변환하는 단계와, 정규화된 이산형 특징 데이터들의 연관성을 계산하여 최대 연관성을 가지는 제1 특징 데이터를 선택하는 단계와, 제1 특징 데이터 선택 단계에서 선택되지 않은 특징 데이터들의 포텐셜을 계산하여 가장 큰 포텐셜을 가지는 것을 제2 특징 데이터로 선택하는 단계를 포함함으로써, 특징 데이터의 연관성과 중복성 간에 불균형을 제거할 수 있다.

대 표 도 - 도1



이 발명을 지원한 국가연구개발사업

과제고유번호	NIPA-2010-(C1090-1021-003)
부처명	지식경제부
연구사업명	대학 IT연구센터 지원 사업
연구과제명	동서신의학 u-라이프케어 기술 연구
주관기관	경희대학교 산학협력단
연구기간	2006.11.01 ~ 2014.12.31

특허청구의 범위

청구항 1

입력된 데이터로부터 추출된 연속형 특징 데이터들을 평균이 0이고, 단위 표준 편차를 갖도록 정규화하는 단계와;

상기 정규화된 연속형 특징 데이터들을 이산형 특징 데이터들로 변환하는 단계와;

상기 정규화된 이산형 특징 데이터들의 연관성을 계산하여 최대 연관성을 가지는 제1 특징 데이터를 선택하는 단계와;

상기 제1 특징 데이터 선택 단계에서 선택되지 않은 특징 데이터들의 포텐셜을 계산하여 가장 큰 포텐셜을 가지는 것을 제2 특징 데이터로 선택하는 단계;를 포함하는 패턴 인식을 위한 특징 데이터 선택 방법.

청구항 2

제1항에 있어서,

상기 정규화 단계는,

특징 데이터의 연관성을 나타내는 클래스-특징 상호 정보를 계산하기 위하여

$$MI(C;X_i)=\frac{I(C;X_i)}{H(C)}$$

의 식을 사용하는 패턴 인식을 위한 특징 데이터 선택 방법.

청구항 3

제1항에 있어서,

상기 정규화 단계는,

특징 데이터의 중복성을 나타내는 특징-특징 상호 정보를 계산하기 위하여

$$MI(X_s;X_j)=\frac{I(X_s;X_j)}{B}$$

의 식을 사용하는 패턴 인식을 위한 특징 데이터 선택 방법.

청구항 4

제1항에 있어서,

상기 변환 단계는,

상기 연속형 특징 데이터들을 8비트, 16비트 및 32비트 중 어느 하나로 양자화하는 패턴 인식을 위한 특징 데이터 선택 방법.

청구항 5

제1항에 있어서,

상기 제2 특징 선택 단계는,

특징 데이터의 포텐셜을 계산하기 위하여

$$f(X_j) = MI(C; X_j) - \frac{1}{|S|}\sum_{s \in S} MI(X_s; X_j)$$

의 식을 사용하는 패턴 인식을 위한 특징 데이터 선택 방법.

청구항 6

입력된 데이터로부터 추출된 연속형 특징 데이터들을 평균이 0이고, 단위 표준 편차를 갖도록 정규화하는 데이터 정규화부;

상기 정규화된 연속형 특징 데이터들을 이산형 특징 데이터들로 변환하는 데이터 변환부;

상기 정규화된 이산형 특징 데이터들의 연관성을 계산하여 최대 연관성을 가지는 제1 특징 데이터를 선택하는 제1 특징 데이터 선택부와;

상기 제1 특징 데이터 선택 단계에서 선택되지 않은 특징 데이터들의 포텐셜을 계산하여 가장 큰 포텐셜을 가지는 것을 제2 특징 데이터로 선택하는 제2 특징 데이터 선택부;를 포함하는 패턴 인식을 위한 특징 데이터 선택 장치.

청구항 7

제6항에 있어서,

상기 데이터 정규화부는,

특징 데이터의 연관성을 나타내는 클래스-특징 상호 정보를 계산하기 위하여

$$MI(C; X_i) = \frac{I(C; X_i)}{H(C)}$$

의 식을 사용하는 패턴 인식을 위한 특징 데이터 선택 장치.

청구항 8

제6항에 있어서,

상기 데이터 정규화부는,

특징 데이터의 중복성을 나타내는 특징-특징 상호 정보를 계산하기 위하여

$$MI(X_s; X_j) = \frac{I(X_s; X_j)}{B}$$

의 식을 사용하는 패턴 인식을 위한 특징 데이터 선택 장치.

청구항 9

제6항에 있어서,

상기 데이터 변환부는,

상기 연속형 특징 데이터들을 8비트, 16비트 및 32비트 중 어느 하나로 양자화하는 패턴 인식을 위한 특징 데이터 선택 장치.

청구항 10

제6항에 있어서,

상기 제2 특징 선택부는,

특징 데이터의 포텐셜을 계산하기 위하여

$$f(X_j) = MI(C; X_j) - \frac{1}{|S|}\sum_{s \in S} MI(X_s; X_j)$$

의 식을 사용하는 패턴 인식을 위한 특징 데이터 선택 장치.

명 세 서

기 술 분 야

[0001] 입력 데이터의 특징 데이터를 선택하는 것으로, 보다 구체적으로는 패턴 인식을 위해 데이터의 상호 정보 정규화하여 특징 데이터를 선택하는 기술과 관련된다.

[0002] 본 발명은 동서신의학 U-라이프케이연구센터 사업의 일환으로 수행한 연구로부터 도출된 것이다.[과제관리번호 : NIPA-2010-(C1090-1021-003), 과제명 : 동서신의학 U-라이프케어 기술 연구]

배 경 기 술

[0003] 특징선택이란 입력 데이터로부터 추출된 많은 수의 특징들로부터 다른 클래스와 구분할 수 있는 정보를 가진 적절한 특징들의 집단을 선택하는 기술이다. 패턴 인식에서 가장 차별된 특징들을 식별하는 것이 가장 중요한 단계이다. 이는 많은 수의 특징들의 집합을 인식모델에 반영하는 것이 계산 증가에 따른 부담뿐만 아니라 일반적으로 차원수의 저주(curse of dimensionality)로 알려진 문제를 야기시키기 때문이다.

[0004] 특히, 차원수의 저주는 추가차원(extra dimensions)을 공간에 더하는 것과 관련된 것으로 부피에서의 지수적 증가에 의해 야기되는 문제이다. 이는 상태 변수의 차원이 높을 때 수치적 후진귀납법(numerical backwards induction)에 의해 동적 최적화 문제를 해결하는데 중요한 장애물이다. 또한 고차원의 유한 자료 표본들과 고차원 공간에서 최근접 이웃 검색(nearest neighbor search)으로부터 무한분포(state of nature)의 학습을 포함하는 기계학습 문제를 복잡하게 만든다.

[0005] 종래에는 상호정보척도(mutual information measurements)에 기초하여 강력히 연관된 특징을 선택하는 방법이 소개되었다. 그러나, 이는 파라미터 종속적인 방법이었다. 따라서, 파라미터가 제대로 설정되지 않는 경우에는 올바른 결과값을 얻지 못하였다. 이러한 제한을 극복하기 위해 소개된 방법은 특징들의 중복과 연관성에 기초한 개념인 파라미터로부터 자유로운 특징 선택방법이다. 그러나, 이러한 방법도 클래스-특징과 특징-특징 측정 사이에서의 불균형으로 인해 좋은 특징들을 누락할 수 있다.

[0006] 비록 특징-특징 상호 정보를 정규화하여 위와 같은 문제를 해결하고 있지만 완전히 해결하지는 못하고 있는 실정이다. 게다가 이는 특징 연관된 측정을 정규화할 때, 가중치의 불균형이란 문제를 가져온다.

[0007] 따라서, 본 발명자는 정규화 과정에서의 문제와 특징 종속적인 가중치들의 문제를 해결하기 위해 연구하게 되었다.

발명의 내용

해결하려는 과제

[0008] 패턴 인식을 위해 특징 데이터 선택을 위한 상호 정보 계산시 특징 데이터의 연관성을 나타내는 클래스-특징 상호 정보와 특징 데이터의 중복성을 나타내는 특징-특징 상호 정보 간의 불균형을 제거할 수 있다. 또한, 특징에 독립적인 정규화 팩터를 이용하여 모든 특징 데이터들을 공평하게 고려할 수 있다.

과제의 해결 수단

[0009] 일 실시예에 따른 패턴 인식을 위한 특징 데이터 선택 방법은 입력된 데이터로부터 추출된 연속형 특징 데이터들을 평균이 0이고, 단위 표준 편차를 갖도록 정규화하는 단계와, 정규화된 연속형 특징 데이터들을 이산형 특징 데이터들로 변환하는 단계와, 정규화된 이산형 특징 데이터들의 연관성을 계산하여 최대 연관성을 가지는 제1 특징 데이터를 선택하는 단계와, 제1 특징 데이터 선택 단계에서 선택되지 않은 특징 데이터들의 포텐셜을 계

산하여 가장 큰 포텐셜을 가지는 것을 제2 특징 데이터로 선택하는 단계를 포함한다.

[0010] 일 실시예에 따른 패턴 인식을 위한 특징 데이터 선택 장치는 입력된 데이터로부터 추출된 연속형 특징 데이터들을 평균이 0이고, 단위 표준 편차를 갖도록 정규화하는 데이터 정규화부와, 정규화된 연속형 특징 데이터들을 이산형 특징 데이터들로 변환하는 데이터 변환부와, 정규화된 이산형 특징 데이터들의 연관성을 계산하여 최대 연관성을 가지는 제1 특징 데이터를 선택하는 제1 특징 데이터 선택부와, 제1 특징 데이터 선택 단계에서 선택되지 않은 특징 데이터들의 포텐셜을 계산하여 가장 큰 포텐셜을 가지는 것을 제2 특징 데이터로 선택하는 제2 특징 데이터 선택부를 포함한다.

발명의 효과

[0011] 특징에 독립적으로 정규화함으로써 특징 데이터의 연관성과 중복성 간에 불균형을 제거할 수 있다. 또한, 특징과 관련된 척도를 정규화할 때 가중치의 불균형을 제거할 수 있다.

도면의 간단한 설명

[0012] 도 1은 일 실시예에 따른 패턴 인식을 위한 특징 데이터 선택 방법의 흐름도,

도 2는 일 실시예에 따른 패턴 인식을 위한 특징 데이터 선택 방법 중 상호 정보와 엔트로피 간의 관계도,

도 3은 일 실시예에 따른 패턴 인식을 위한 특징 데이터 선택 장치의 구성도이다.

발명을 실시하기 위한 구체적인 내용

[0013] 이하, 첨부된 도면들을 참조하여 본 발명의 실시예를 상세하게 설명한다. 사용되는 용어들은 실시예에서의 기능을 고려하여 선택된 용어들로서, 그 용어의 의미는 사용자, 운용자의 의도 또는 판례 등에 따라 달라질 수 있다. 그러므로 후술하는 실시예들에서 사용된 용어의 의미는, 본 명세서에 구체적으로 정의된 경우에는 그 정의에 따르며, 구체적인 정의가 없는 경우는 당업자들이 일반적으로 인식하는 의미로 해석되어야 할 것이다.

[0014] 도 1은 일 실시예에 따른 패턴 인식을 위한 특징 데이터 선택 방법의 흐름도이다.

[0015] 도 1을 참조하면, 패턴 인식을 위한 특징 데이터 선택 방법은 입력된 데이터들로부터 추출된 연속형 특징 데이터들을 평균이 0이고, 단위 표준 편차를 갖도록 정규화한다(101). 입력되는 데이터는 음성 데이터, 얼굴 데이터, 지문 데이터, 홍채 데이터일 수 있으나 이에 한정하는 것은 아니다. 얼굴 인식, 음성 인식의 패턴을 분석하기 위해서는 각 데이터의 특징이 될 만한 부분을 추출하여야 한다. 추출된 특징 데이터만을 이용하여 패턴을 인식하기 때문에, 패턴을 정확하게 인식하면서도 계산과정을 단순화시킬 수 있다.

[0016] 이산형 특징 데이터를 정규화하는 과정은 추출된 특징 데이터를 패턴인식 과정에서 사용하기에 적합하게 설정하는 것이다. 이를 위해서는 상호 정보(Mutual Information)를 계산하여야 한다. 상호 정보란 두 개의 대상이 서로 의존적으로 나타난 정보를 의미한다. 상호 정보는 특징 데이터를 선택을 위한 기준으로 사용된다. 이는 변수들의 상호 연관성(relevance)과 중복성(redundancy)을 특성화하기 위해 사용되는 것이다. 예를 들어, 두개의 이산 랜덤 변수(discrete random variable)인 x와 y의 상호 정보(Mutual Information)는 수학식 1과 같이 정의될 수 있다.

수학식 1

[0017]
$$I(X;Y)=\sum_{x\in\Omega_X}\sum_{y\in\Omega_Y}p(x,y)\log_2\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

[0018] 수학식 1에서, Ω_X와 Ω_Y는 X와 Y의 샘플 공간(sample space)이고, p(x,y)는 x와 y의 결합 확률 분포(joint probability distribution)의 함수이다. p(x)와 p(y)는 x와 y의 주변 확률 분포(marginal probability distribution)이다.

[0019] 한편, 연속적인 경우에는 수학식 1의 상호 정보를 이중 적분을 이용하여 나타낼 수 있다.

수학식 2

$$I(X;Y)=\int_{\Omega_X}\int_{\Omega_Y} p(x,y)\log_2\left(\frac{p(x,y)}{p(x)p(y)}\right)dxdy$$

[0020]

[0021] 수학식 2에서, p(x,y)는 x와 y의 결합 확률 밀도(joint probability density)의 함수이다. p(x)와 p(y)는 x와 y의 주변 확률 밀도(marginal probability density)이다.

[0022] 수학식 1과 수학식 2에서 정의된 상호 정보의 정의에서 로그의 밑수(base)가 정해지지 않으면 로그 함수가 불명확해질 수 있다. 따라서, 상호 정보 함수 I는 I(X,Y,b)와 같이 지정될 수 있다. 이 경우, b는 로그함수의 밑수이다. 대안적으로, 상호 정보의 측정의 가장 일반적인 단위는 비트이므로, 로그 함수의 밑수는 2로 특정할 수 있다.

[0023] 도 2는 일 실시예에 따른 패턴 인식을 위한 특징 데이터 선택 방법 중 상호 정보와 엔트로피(200) 간의 관계도이다.

[0024] 도 2를 참조하면, 왼쪽의 타원은 변수 X의 엔트로피(210)를 나타내며, 오른쪽의 타원은 변수 Y의 엔트로피(230)를 나타낸다. 두 타원체의 겹치는 부분은 변수 X, 변수 Y의 상호 정보(250)를 나타낸다. 이 경우, 타원 중 겹치는 부분을 제외한 영역은 X|Y, Y|X와 같은 조건적 엔트로피를 나타낸다. 이를 통해, 상호 정보(250)는 다음의 수학식 3과 같이 계산될 수 있다.

수학식 3

$$I(X;Y)=H(X)-H(X\mid Y)$$

[0025]

[0026] 수학식 3에서, H(X)와 H(X|Y)는 엔트로피 함수이다.

[0027] 상호 정보(250)는 변수 X와 변수 Y가 공유하는 정보를 측정하는 것이다. 이는 다른 것에 대한 불확실성을 감소시킬 수 있는 변수들 중 하나를 얼마나 많이 알고 있는지를 측정하는 것이다. 예를 들어, 변수 X와 변수 Y가 독립적인 경우, 변수 X를 아는 것만으로는 변수 Y에대한 정보를 얻을 수 없다. 또한, 그 반대의 경우도 마찬가지이다. 따라서, 이들의 상호 정보의 값은 0이 된다. 극단적으로, 변수 X와 변수 Y가 동일한 경우, 상호 정보는 변수 X 또는 변수 Y 자체에 포함된 불확실성과 같다. 즉, 변수 X와 변수 Y가 일치하는 경우, 그들의 엔트로피는 같다.

[0028] 다시 도 1로 돌아가서, 본 발명의 특징 선택 방법에서 포함하는 정규화 단계는 특징 데이터의 연관성을 나타내는 클래스-특징 상호 정보를 계산하기 위하여 다음과 같은 식을 이용할 수 있다.

수학식 4

$$MI(C;X_i)=\frac{I(C;X_i)}{H(C)}$$

[0029]

[0030] 수학식 4에서, C는 클래스의 변수이고, X는 특징 변수이다. $MI(C;X_i)$는 클래스-특징 상호 정보이고, H(C)는 클래스 변수 C의 엔트로피를 나타낸다. $I(c;X_i)$는 수학식 2를 통해 구할 수 있다.

[0031] 또한, 정규화 단계에서 특징 데이터의 중복성을 나타내는 특징-특징 상호 정보를 계산하기 위하여 다음과 같은 식을 이용할 수 있다.

수학식 5

[0032] $$MI(X_s;X_j)=\frac{I(X_s;X_j)}{B}$$

[0033] 수학식 5에서, $I(X_S;X_j)$는 두 특징 변수 X_S, X_j의 상호 정보를 나타낸다. 또한, B는 연속형 특징 데이터를 양자화하는 비트의 수를 나타낸다. 한편, 종래의 상호 정보를 구하기 위해서는 다음과 같은 식을 이용하여야 했다.

수학식 6

[0034] $$MI(X_i;X_j)=\frac{I(X_i;X_j)}{\min(H(X_i),H(X_j))}$$

[0035] 수학식 6에서, 정규화 팩터인 $\min((H(X_i), H(X_j))$은 특징-종속적이므로, 특징 선택 과정에서 특징들이 불공평하게 취급된다. 그러나, 본 발명의 특징 데이터 선택 방법에서는 수학식 5와 같이, 특징-독립적이다. 이에 따라, 특징 데이터들의 연관성과 중복성 사이에 불균형을 제거할 수 있다.

[0036] 다음으로, 정규화된 연속형 특징 데이터들을 이산형 특징 데이터들로 변환한다(103). 데이터 변환 단계는 추출된 연속형 특징 데이터들을 예를 들어 8비트, 16비트 및 32비트 중 어느 하나로 양자화(Quantize)할 수 있으나 반드시 이에 한정하는 것은 아니다. 이는 컴퓨터를 이용하여 수치를 분석하기 위해, 음성, 지문 이미지 등과 같은 아날로그형 데이터를 일정한 비트 단위로 샘플링하는 것이다. 따라서, 연속형 특징 데이터들은 이산형 특징 데이터로 변환된다. 따라서, 사용자는 샘플링되는 비트 수를 조절하여 적절한 이산형 특징 데이터로 변환할 수 있다.

[0037] 다음으로, 정규화된 이산형 특징 데이터들의 연관성을 계산하여 최대 연관성을 가지는 제1 특징 데이터를 선택한다(105). 제1 특징 데이터는 다음과 같은 식을 이용한다.

수학식 7

[0038] $$j=\underset{i=1:N}{\operatorname{argmax}} f(X_i)$$

$$S=S\cup j$$

[0039] 수학식 7에서, j는 최근에 선택된 특징 인덱스이고, S는 선택된 특징 데이터들의 집합을 나타낸다. i는 특징 변수이며, $f(X_i)$는 특징 변수의 함수로 나타낼 수 있다. $f(X_i)$는 본 발명의 특징 선택 방법에 따른 클래스-특징

상호정보를 구하는 수학식 4를 통해 구할 수 있다. 따라서, i가 1에서 N 범위에 있을 때, 최대의 클래스-특징 상호정보를 갖는 특징 인덱스 j가 제1 특징 데이터로 선택될 수 있다.

[0040] 다음으로, 제1 특징 데이터 선택 단계에서 선택되지 않은 특징 데이터들의 포텐셜을 계산하여 가장 큰 포텐셜을 가지는 것을 제2 특징 데이터로 선택한다(107). 이렇게 하여 선택된 특징 데이터는 선택되지 않은 다른 특징 데이터들 중에서 가장 높은 포텐셜을 가지게 된다. 이 경우, 제2 특징 선택 단계에서 특징 데이터의 포텐셜을 계산하기 위하여는 다음과 같은 식을 이용할 수 있다.

수학식 8

$$f(X_j) = MI(C;X_j) - \frac{1}{|S|}\sum_{s\in S} MI(X_s;X_j)$$

$$= \frac{I(C;X_i)}{H(C)} - \frac{1}{|S|}\sum_{s\in S}\frac{I(X_s;X_i)}{B}$$

[0041]

[0042] 수학식 8에서, X_j의 포텐셜은 수학식 4에 나타난 특징 데이터의 연관성 정보 $MI(c;X_j)$와, 수학식 5에 나타난 특징 데이터의 중복성 정보 $MI(X_S;X_j)$를 이용하여 구할 수 있다. 이 경우, 선택되지 않은 특징 데이터 중 최대의 포텐셜을 구하기 위한 조건은 다음과 같은 식으로 나타낼 수 있다. 종래에는 1/|S| 대신에 β 파라미터를 이용하였다. 그러나, 이는 β 파라미터를 잘못 구하면 결과값이 달라지는 문제점이 있었다. 또한, β 파라미터를 1/|S|로 대체한 이후에도, 같은 범위내에서의 값을 구할 수 없어, 특징 값이 최대 연관성 특징에 치우치거나 최소 중복성 특징에 치우치는 문제점이 있었다. 그러나, 수학식 8을 이용함에 따라, 연관성과 중복성간에 균형있는 조화를 이룰 수 있게 되었다.

[0043] 제2 특징 데이터를 선택하기 위해서는 다음과 같은 기준의 식을 이용할 수 있다.

수학식 9

$$j = \underset{i\notin S}{\operatorname{argmax}}\, f(X_i)$$

$$S = S \cup j$$

[0044]

[0045] 수학식 9에서, j는 최근에 선택된 특징 인덱스이고, S는 선택된 특징 데이터들의 집합을 나타낸다. i는 특징 변수이며, $f(X_i)$는 특징 변수의 함수로 나타낼 수 있다. 이 때, i는 S에 포함되지 않는 것이어야 한다. 즉, 앞의 단계에서 선택되지 않은 특징 데이터들의 포텐셜을 구하는 것이다. $f(X_i)$는 본 발명의 특징 선택 방법에 따른 클래스-특징 상호정보를 구하는 수학식 4를 통해 구할 수 있다. 이에 따라, 앞의 단계에서 선택되지 않은 특징 데이터들 중에서 최대의 클래스-특징 상호정보를 갖는 특징 인덱스 j가 제2 특징 데이터로 선택될 수 있다.

[0046] 이러한 과정은 사용자가 원하는 특징 데이터의 수에 이를 때까지 반복될 수 있다. 따라서, 선택된 특징 데이터

들 중 어느 하나는 선택되지 않은 특징 데이터 중에서 최대의 포텐셜을 가질 수 있다.

[0047] 도 3은 일 실시예에 따른 패턴 인식을 위한 특징 데이터 선택 장치의 구성도이다.

[0048] 도 3을 참조하면, 본 발명의 특징 데이터 선택 장치는 데이터 정규화부(11)와, 데이터 변환부(13), 제1 특징 데이터 선택부(15)와, 제2 특징 데이터 선택부(17)와, 특징 데이터 출력부(19)를 포함할 수 있다. 데이터 정규화부(11)는 앞서 설명한 바와 같이 입력된 데이터에서 추출된 연속형 특징 데이터를 평균이 0이고, 단위 표준 편차를 갖도록 정규화할 수 있다. 데이터 정규화부(11)에서 정규화된 연속형 특징 데이터는 데이터 변화부로 출력된다.

[0049] 또한, 데이터 정규화부(11)는 특징 데이터의 연관성을 나타내는 클래스-특징 상호 정보를 계산하기 위해 수학식 4를 구현한 논리회로를 구현할 수 있다. 또한, 데이터 정규화부(11)는 특징 데이터의 중복성을 나타내는 특징-특징 상호 정보를 계산하기 위해 수학식 5를 구현한 논리회로를 구현할 수 있다. 이에 따라, 특징 데이터를 특징에 독립적으로 정규화함으로써 특징 데이터의 연관성과 중복성 간에 불균형을 제거할 수 있다.

[0050] 데이터 변환부(13)는 데이터 정규화부(11)로부터 정규화된 연속형 특징 데이터를 이산형 특징 데이터로 변환한다. 이 경우 이산형의 특징 데이터는 예를 들어 8비트, 16비트 및 32비트 중 어느 하나의 비트로 양자화할 수 있으나 이에 한정하는 것은 아니다. 데이터 변환부(13)에서 변환된 특징 데이터는 데이터 변환DB(14)에 저장된다. 데이터 변환DB(14)에는 특정의 비트로 샘플링된 이산형 특징 데이터들이 저장, 관리될 수 있다.

[0051] 제1 특징 데이터 선택부(15)는 데이터 변환DB(14)에 저장된 특징 데이터들 중에서 가장 연관성이 큰 특징 데이터를 제1 특징 데이터로 선택한다. 이를 위해, 수학식 7과 같은 기준을 논리회로로 구현할 수 있다. 제1 특징 데이터 선택부(15)는 수학식 7의 기준을 만족하는 특징 데이터를 선택하면, 이를 특징 데이터 DB에 저장하고, 데이터 변환DB(14)에서는 해당 특징 데이터를 삭제할 수 있다.

[0052] 제2 특징 데이터 선택부(17)는 데이터 변환DB(14)에 저장된 특징 데이터들 중에서 포텐셜이 가장 큰 특징 데이터를 제2 특징 데이터로 선택한다. 앞서 설명한 바와 같이, 데이터 변환DB(14)에는 제1 특징 데이터로 선택된 특징 데이터들이 저장된다. 제2 특징 데이터 선택부(17)는 수학식 8을 구현한 논리회로로 구현될 수 있다. 따라서, 데이터 변환DB(14)의 특징 데이터들 중 수학식 9의 기준을 만족하는 제2 특징 데이터를 선택할 수 있다. 이 경우, 데이터 변환DB(14)에서는 제2 특징 데이터를 삭제하고, 제2 특징 데이터는 특징 데이터DB(18)에 저장될 수 있다.

[0053] 또한, 제2 특징 데이터 선택부(17)는 설정된 특징 데이터의 수를 만족할 때까지 복수의 제2 특징 데이터를 선택하여 특징 데이터DB(18)에 저장할 수 있다. 이에 따라, 본 발명의 특징 데이터 선택 장치는 포텐셜이 가장 높은 복수의 특징 데이터들을 선택함으로써, 패턴 인식을 보다 편리하게 진행할 수 있도록 할 수 있다.

[0054] 특징 데이터 출력부(19)는 특징 데이터DB(18)와 연결된다. 이는 특징 데이터DB(18)에 저장된 특징 데이터를 추출하여, 이를 외부로 출력한다. 특징 선택 장치(10)와 연결된 외부 장치는 패턴 인식을 위해 특징 데이터를 필요로 하는 장치일 수 있으나, 이에 한정하는 것은 아니다.

[0055] 이상에서 본 발명은 도면을 참조하면서 기술되는 바람직한 실시예를 중심으로 설명되었지만 이에 한정되는 것은 아니다. 따라서 본 발명은 기재된 실시예로부터 도출 가능한 자명한 변형예를 포괄하도록 의도된 특허청구범위의 기재에 의해 해석되어져야 한다.

부호의 설명

[0056] 10 : 특징 선택 장치(10)

11 : 데이터 정규화부

13 : 데이터 변환부

14 : 데이터 변환DB

15 : 제1 특징 데이터 선택부

17 : 제2 특징 데이터 선택부

18 : 특징 데이터DB

19 : 특징 데이터 출력부

200 : 엔트로피

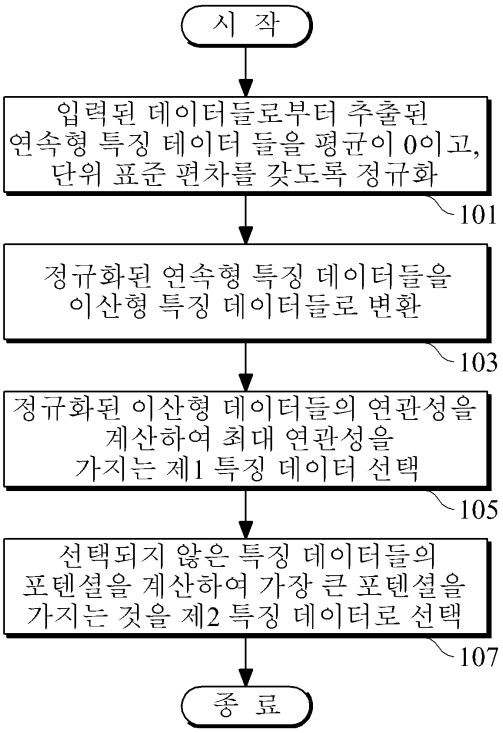
210 : X의 조건적 엔트로피

230 : Y의 조건적 엔트로피

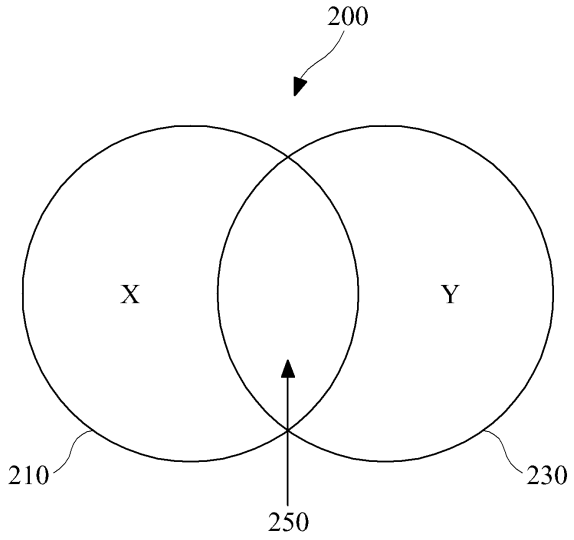
250 : 상호 정보

도면

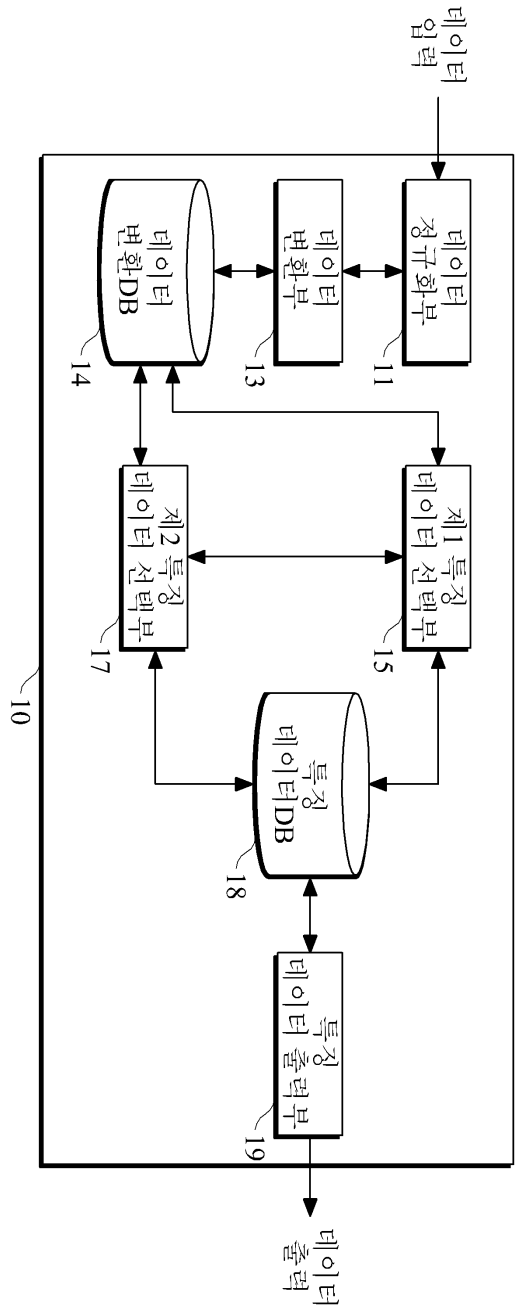
도면1



도면2



도면3


데이터 입력
데이터 정규화부
11
데이터 변환부
13
데이터 변환DB
14
제1 특징 데이터 선택부
15
제2 특징 데이터 선택부
17
특징 데이터DB
18
특징 데이터 출력부
19
10
데이터 출력