Journal of Korean Institute of Intelligent Systems, Vol. 24, No. 6, December 2014, pp. 646-650
http://dx.doi.org/10.5391/JKIIS.2014.24.6.646

퍼지 매핑을 이용한 퍼지 패턴 분류기의 Feature Selection

Feature Selection of Fuzzy Pattern Classifier by using Fuzzy Mapping

노석범* · 김용수** · 안태천*†
Seok-Beom Roh, Yong Soo Kim, and Tae-Chon Ahn†

*원광대학교 전자융합공학과, **대전대학교 컴퓨터공학과
*Department of Electronics Convergence Engineering, Wonkwang University
**Department of Computer Engineering, Daejeon University

요 약

본 논문에서는 다차원 문제로 인하여 발생하는 패턴 분류 성능의 저하를 방지 하여 퍼지 패턴 분류기의 성능을 개선하기 위하여 다수의 Feature들 중에서 패턴 분류 성능 향상에 기여하는 Feature를 선택하기 위한 새로운 Feature Selection 방법을 제안 한다. 새로운 Feature Selection 방법은 각각의 Feature 들을 퍼지 클러스터링 기법을 이용하여 클러스터링 한 후 각 클러스터가 임의의 class에 속하는 정도를 계산하고 얻어진 값을 이용하여 해당 feature 가 fuzzy pattern classifier에 적용될 경우 패턴 분류 성능 개선 가능성을 평가한다. 평가된 성능 개선 가능성을 기반으로 이미 정해진 개수만큼의 Feature를 선택하는 Feature Selection을 수행한다. 본 논문에서는 제안된 방법의 성능을 평가, 비교하기 위하여 다수의 머신 러닝 데이터 집합에 적용한다.

키워드 : 퍼지 패턴 분류기, Feature Selection, 퍼지 클러스터링, 퍼지 수

Abstract

In this paper, in order to avoid the deterioration of the pattern classification performance which results from the curse of dimensionality, we propose a new feature selection method. The newly proposed feature selection method is based on Fuzzy C-Means clustering algorithm which analyzes the data points to divide them into several clusters and the concept of a function with fuzzy numbers. When it comes to the concept of a function where independent variables are fuzzy numbers and a dependent variable is a label of class, a fuzzy number should be related to the only one class label. Therefore, a good feature is a independent variable of a function with fuzzy numbers. Under this assumption, we calculate the goodness of each feature to pattern classification problem. Finally, in order to evaluate the classification ability of the proposed pattern classifier, the machine learning data sets are used.

Key Words : Fuzzy Pattern Classifier, Feature Selection, Fuzzy Clustering, Fuzzy Number.

1. 서 론

현재는 모바일 기기들의 발달로 인하여 텍스트 형태의 데이터뿐만 아니라 이미지 데이터와 같은 데이터의 차원이 매우 큰 데이터들이 대량으로 생산되고 있다. 이와 같은 데이터를 적절하게 이용하여 올바른 선택을 하고자 하는 시도가 많이 이루어지고 있다. 이와 같은 고차원 대량의 데이터를 분류하기 위하여 다양한 형태의 분류 알고리즘이 개발되어지고 있다. 그러나 획득된 데이터의 차원이 매우 커짐에 따라, 패턴 분류를 위한 계산 시간뿐만 아니라 패턴 분류 성능에 좋지 않은 영향을 주고 있다 [1].

데이터의 다차원 문제를 해결하기 위한 가장 좋은 방법은 데이터의 차원을 낮추어 패턴 분류 알고리즘의 계산량을 줄이고 패턴 분류 성능을 계선 시켜야 한다. 다차원 데이터의 차원을 낮추는 방법들 중에 대표적인 방법으로 전체 입력 변수 들 중 일부분을 선택하여 입력 변수로 사용하는 feature selection [2, 3, 5, 6] 방법이 있다.

Feature Selection을 위한 다양한 방법이 제시되어 왔으며, 특히 Particle Swarm Optimization과 같은 최적화 기법을 이용한 방법도 제시되었다 [4, 7, 8, 9].

본 논문에서는 일반적 함수의 특성에 기반 한 새로운 Feature Selection 알고리즘을 제시하고자 한다.

일반적으로 함수 $f : X \rightarrow Y$가 함수이기 위해서는 정의역

접수일자: 2014년 9월 14일
심사(수정)일자: 2014년 9월 28일
게재확정일자 : 2014년 12월 4일
† Corresponding author
본 논문은 2014학년도 원광대학교의 교비지원에 의해서 수행 됨

X에 속한 모든 원소들이 치역 Y에 속한 단 하나의 원소와 대응되어야 한다. 여기서 치역 Y는 클래스의 레이블을 의미한다.

위 조건을 이용하여 비슷한 특성을 가진 데이터 포인트 $\boldsymbol{x_i}$과 $\boldsymbol{x_j}$ 가 다른 클래스에 속하게 되면, 이와 같은 특성을 가진 데이터 집합은 패턴 분류하기 힘들다고 가정한다. 그래서 위와 같은 특성을 보이는 Feature들을 제거하고 나머지 Feature들로만 구성된 새로운 데이터 집합을 구성한 후 패턴 분류기에 적용한다.

그러나 데이터 포인트 $\boldsymbol{x_i}$과 $\boldsymbol{x_j}$ 가 아주 인접해 있지 않다고 한다면, 두 데이터 포인터가 다른 클래스에 속에 있다고 해도 패턴 분류에는 영향을 미치지 않는다.

데이터 포인트 $\boldsymbol{x_i}$과 $\boldsymbol{x_j}$ 의 상호 유사성을 정의하기 위하여 각 Feature에 퍼지 수들을 정의하고 데이터 포인트 $\boldsymbol{x_i}$과 $\boldsymbol{x_j}$ 가 동일한 퍼지 수에 속하면, $\boldsymbol{x_i}$과 $\boldsymbol{x_j}$ 는 유사성이 있다고 할 수 있겠다.

본 논문에서 각각의 Feature 공간에서 퍼지 수를 정의하기 위하여 Bezdek의 퍼지 클러스터링 기법을 이용한다. 각 퍼지 수의 중심은 Fuzzy C-Means Clustering 기법을 이용하여 얻어진 클러스터의 중심 값을 사용한다. 결정된 퍼지 수의 중심값을 이용하여 삼각형 형태의 퍼지 수의 멤버십 함수를 정의한다.

Feature가 결정된 데이터들을 패턴 분류하기 위하여 Linear Discriminant Analysis를 이용한다.

제안된 패턴 분류를 위한 Feature Selection 기법의 성능을 평가하기 위하여 다양한 형태의 머신 러닝 데이터 집합들을 이용하여 패턴 분류 성능을 비교한다.

2. Feature Selection

본 논문에서는 다차원 데이터의 차원 수를 감소하기 위한 새로운 방법을 제시한다. 일반적인 함수의 정의는 그림 1과 같이 정의역 X에 속한 임의의 원소 $\boldsymbol{x}$ 가 치역의 임의의 원소 $\boldsymbol{y}$ 와 대응 될 때 $f : X \rightarrow Y$를 함수라 할 수 있다.

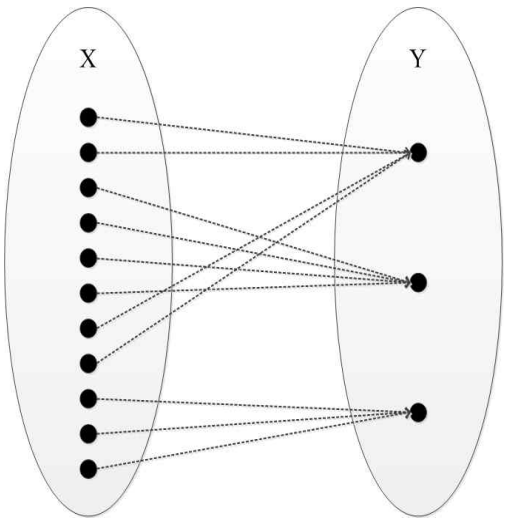

그림 1. 함수의 정의.
Fig. 1. Definition of Function.

만약 정의역 X가 퍼지 수로 정의되었다면 퍼지 함수 $f : \widetilde{X} \rightarrow Y$는 아래 그림 2와 같이 묘사 될 수 있다.

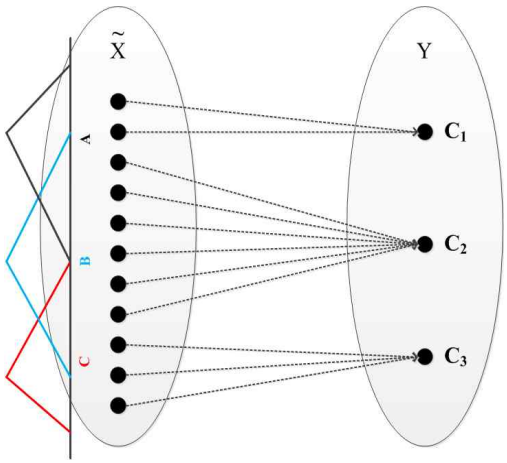

그림 2. 퍼지수 함수의 정의.
Fig. 2. Definition of Function of Fuzzy Number.

그림 2에서 퍼지수 A의 원소는 클래스 1과 클래스 2에 속하지만 퍼지수 B는 클래스 2에 대부분의 원소가 속한다. 일반적으로 비슷한 특성을 가진 원소들이 동일한 클래스에 속하는 특성을 보이는 Feature는 패턴 분류기의 패턴 분류 성능을 향상 시킬 가능성이 크다고 할 수 있다. 그러나 유사한 특성을 가지는 원소들이 다른 클래스에 속하게 되는 Feature는 패턴 분류 성능을 악화 시킬 가능성이 크다.
데이터의 k번째 Feature에 대하여 위에 언급한 패턴 분류 성능향상과 관련된 특성을 파악하기 위한 평가 지수는 식 (1) 과 (2)와 같이 정의 할 수 있다.

$$J_k = \sum_{i=1}^{l} j_{ki} \tag{1}$$

$$j_{ki} = \prod_{j=1}^{m} \sum_{x} (\mu_{ki}(x) \cdot L_j(x)) \tag{2}$$

여기서 $L_j(x) = \begin{cases} 0, & when\ x \notin C_j \\ 1, & when\ x \in C_j \end{cases}$ $\mu_{ki}(x)$는 k번째 Feature에서 정의 된 i번째 퍼지 수의 멤버십 함수를 나타낸다.

식(1)을 통해 얻어진 J값이 크면 클수록 패턴 분류기의 성능을 악화 시킬 가능성이 있는 Feature라 할 수 있겠다.

3. Fuzzy Clustering을 이용한 퍼지 수 정의

본 논문에서 제안된 Feature Selection 기법은 전체 입력 변수들을 하나씩 1차원 공간에서 식(1)을 이용하여 패턴 분류기의 패턴 분류 성능을 개선시킬 수 있는 Feature인지 평

가하는 방법이다. 이를 위해 각각의 1차원 입력변수 공간에서 퍼지 수를 정의해야 하는데 이를 위해 퍼지 클러스터링 알고리즘을 사용한다.

FCM 클러스터링은 n 개의 벡터 $x_i(i=1,..,n)$ 집합을 c 개의 클러스터로 분할하고, 목적함수가 최소가 일 때 생성된 각 클러스터에서 중심 값을 찾는다. FCM 클러스터링의 목적함수는 식(3)과 같다.

$$J(\mu, v) = \sum_{i=1}^{c}\sum_{k=1}^{N}\mu_{ik}^{m}(\| x_k - v_i \|)^2 \quad (3)$$

목적함수를 최소화하는 μ와 v는 식(4), (5)와 같이 얻을 수 있다.

$$u_{ik} = \frac{1}{\sum_{j=1}^{c}(\frac{\|x_k - v_i\|}{\|x_k - v_i\|})^{2/m-1}} \quad (4)$$

$$v_{ij} = \frac{\sum_{k=1}^{N}(u_{ik})^m x_{ki}}{\sum_{j=1}^{N}(u_{ik})^m} \quad (5)$$

여기서, μ는 데이터의 퍼지 집합 소속 멤버십 함수를 의미하며, v는 해당 하는 퍼지 집합의 중심점을 의미한다.

식(5)를 통해 구해진 각 클러스터의 중심 값을 이용해 v_{ij}를 중심으로 하는 퍼지 수를 정의한다. 본 논문에서 사용된 퍼지수는 삼각형 멤버십 함수를 가진다.

그림 3은 Fuzzy C-Means Clustering 알고리즘을 이용하여 분석하여 얻은 멤버십 함수를 보여주고 있다.

여기서 얻어진 각 클러스터의 중심 값은 $[1.57\ 4.85\ 6.57\ 8.0\ 9.70]$ 이다.

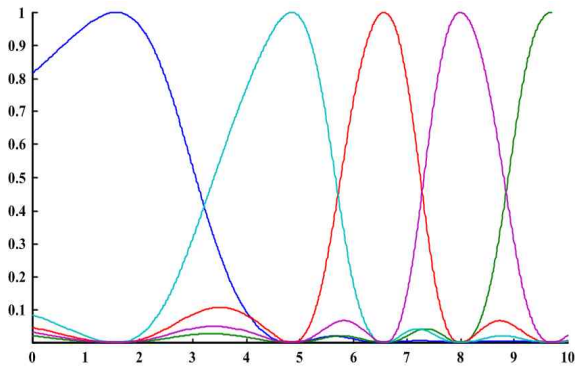

그림 3. 퍼지 클러스터링을 이용하여 정의된 멤버십 함수.
Fig. 3. Membership Function by using Fuzzy Clustering Algorithm.

위에 보인바와 같이 Fuzzy C-Means Clustering 기법을 통해 얻어진 각 클러스터의 중심 값을 기반으로 하는 삼각형 멤버십 함수를 가지는 퍼지 수는 그림 4와 같다.

여기서 각 Feature에 정의 되어진 퍼지 수의 개수는 설계자에 의해 미리 정의 되어져 한다.

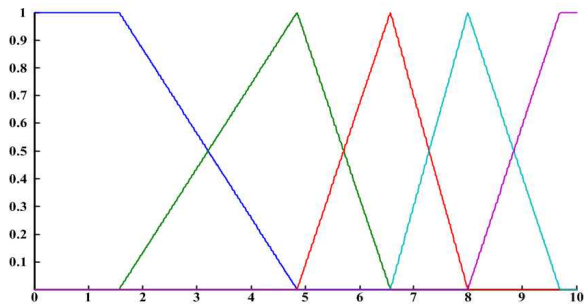

그림 4. 삼각형 멤버십 함수를 가지는 퍼지 수.
Fig. 4. Fuzzy Number with Triangular Membership Function.

식(1)을 통해 평가된 Feature들은 미리 정해진 수만큼의 우수한 Feature들 만을 이용하여 새로운 입력을 구성하고, 이를 Linear Discriminant Analysis를 이용하여 패턴 분류에 이용한다.

4. 선형 판별식

선형 판별식 (Linear Discriminant Analysis; LDA)은 고차원의 데이터 포인트들을 1차원을 가진 임의의 선에 투영하는 방법이다.

이와 같이 고차원에서 1차원의 직선에 투영된 데이터들이 패턴 분류가 잘 되도록 하는 1차원 직선을 정의하는 방법이다.

고차원 데이터를 직선에 투영하여 얻어진 새로운 데이터는 식 (6) 과 같이 정의 한다.

$$z = (\mathbf{w}^T)\mathbf{x} \quad (6)$$

선형 판별식은 투영된 데이터 z가 패턴 분류가 잘 되도록 하는 $\mathbf{w}^t$를 결정한다.

이진 분류 문제에 적용된 선형 판별식의 기준 함수는 (7)과 같다.

$$J(\mathbf{w}) = \frac{|\widetilde{m}_1 - \widetilde{m}_2|^2}{\widetilde{S}_1^2 + \widetilde{S}_2^2} = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}} \quad (7)$$

$$S_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T \quad (8)$$

여기서, $\mathbf{m}_i = \frac{1}{n_i}\sum_{\mathbf{x}\in D_i}\mathbf{x}$.

$$S_i = \sum_{\mathbf{x}\in D_i}(\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T \quad (9)$$

$$S_w = S_1 + S_2 \quad (10)$$

선형 판별식에서는 (7)과 같이 정의된 기준 함수를 최대화 시키는 $\mathbf{w}$를 결정한다.

기준 함수 (7)을 최대화 시키는 $\mathbf{w}$는 (11)을 이용하여 구한다.

$$w = S_w^{-1}(m_1 - m_2) \quad (11)$$

5. 시뮬레이션 및 결과

본 논문에서 제안한 패턴 분류기의 성능을 평가하기 위하여 여러 개의 머신러닝 데이터들을 사용한다. 기계 학습 데이터 집합은 대표적인 Benchmark 데이터 집합인 UCI machine learning repository로부터 얻은 데이터 집합들이다.

Benchmark 데이터 집합을 이용하여 제안된 패턴 분류기의 성능과 특성을 기존 논문에서 이미 제안된 패턴 분류기의 성능과 비교, 평가 한다. 연구되어진 기존 패턴 분류기와 비교하기 위하여, 전체 데이터를 10 fold cross validation 방법에 따라 학습 데이터와 테스트 데이터로 나누어 실험한다.

표 1은 제안된 퍼지 매핑을 이용한 Feature Selection을 위해 미리 설정되어져야 하는 파라미터들을 보인다.

표 1. 설계 파라미터
Table 1. Design Parameters

Parameter	Value
Fuzzifier Coefficient (m)	2
Number of fuzzy number	2, 3, 5
Percentage of the Selected Features	10, 20, 30, 40, 50, 60, 70, 80, 90 (%)

제안된 Feature Selection 기법을 이용하여 선택된 Feature를 가진 패턴 분류기의 성능을 평가하기 위한 기계 학습데이터에 대한 개략적인 정보는 표 2에 열거 하였다.

표 2. 실험에 사용된 기계학습 데이터
Table 2. Machine Learning Data used in the experiments

Datasets	Number of features	Number of Data	Number of Classes
Australian	14	690	2
German	24	1000	2
Ionosphere	34	351	2
Pima	8	768	2
Sonar	60	208	2

표 3은 10 Fold Cross Validation에 의한 데이터 분할 후 실험 결과를 분석, 비교한 결과이다. 패턴 분류 성능을 위하여 패턴 분류기의 오분류율을 평가 지수로 선택 하였다.

표 3. 제안된 패턴 분류기와 기존 패턴 분류기 성능 비교
Table 3. Result of Comparative analysis

	Australian	German	Ionosphere	Pima	Sonar
Bayes Networks	22.14	24.84	N/A	24.25	32.29
kNN (k=3)	15.04	27.79	N/A	26.14	16.24
PART (WEKA)	15.55	29.46	N/A	26.55	22.6
SMO (WEKA)	15.12	24.84	N/A	24.91	23.4
LDA	14.06	30.2	13.36	25.4	24.45
Proposed Classifier	14.35 (Dim=5)	29.8 (Dim=20)	15.09 (Dim=31)	25.64 (Dim=7)	24.6 (Dim=54)

6. 결론 및 향후 연구

본 논문에서는 퍼지수를 이용한 Feature Selection 방법을 제안하였다. 제안된 Feature Selection 기법에 의해 선택된 입력 변수들을 Linear Discriminant Analysis의 입력으로 사용하여 패턴 분류를 수행하였다.

제안된 Feature Selection 알고리즘을 이용한 퍼지 패턴 분류 성능은 전체 입력 변수들을 모두 사용한 경우와 비슷한 성능을 보였다. Feature Selection을 통해 Feature의 차원을 감소시킴으로써 계산 량을 감소시키는데 효과를 얻어다 할 수 있다.

References

[1] Y. J. Lin, J. J. Li, P. R. Lin, G. P. Lin, and J. K. Chen, “Feature Selection via Neighborhood Multi-Granulation Fusion,” *Knowledge Based Systems*, Vol. 67, pp. 162-168, 2014.

[2] Q. Hu, J. Liu, and D. Yu, "Mixed feature selection based on granulation and approximation," *Knowledge Based Systems*, Vol. 21, pp. 294 - 304, 2008.

[3] Q. Hu, W. Pan, L. Zhang, D. Zhang, Y. Song, and D. Yu, "Feature selection for monotonic classification," *IEEE Trans. Fuzzy Systems*, Vol. 20, pp. 69 - 81, 2012.

[4] Y. Zhang, D. W. Gong, Y. Hu, and W. Q. Zhang, “Feature Selection algorithm based on bare bones Particle Swarm Optimization,” *Neurocomputing*, Vol. 148, pp. 150-157, 2015.

[5] G. Pajares, M. Guijarro, A. Ribeiro, “A Hopfield Neural Network for combining classifiers applied to textured images,” *Neural Networks*, Vol.23, pp. 144-153, 2010.

[6] J. C. Bezdek, “*Pattern Recognition With Fuzzy Objective Function Algorithms*”, New York: Plenum, 1981.

[7] A. Bargiela, W. Pedrycz, *"Granular Computing: An Introduction,"* Kluwer Academic Publishers, Dordrecht,

2003.
[8] W. Pedrycz, "Conditional Fuzzy C-Means," *Pattern Recognition Letters*, vol. 17, no. 15, pp 625-631, 1996.
[9] Tae-chon Ahn, Seok-Beom Roh, Yong Soo Kim, "A Design of Fuzzy Classifier with Hierarchical Structure," *Journal of Korean Institute of Intelligent Systems*, vol. 24, no. 4, pp 355-359, 2014.

저 자 소 개

노석범(Seok-Beom Roh)
1994년 원광대학교 제어계측공학과 졸업.
1996년 동 대학원 컴퓨터공학과 석사.
2006년 동 대학원 제어계측공학과 박사.
2007년 ~현재 : 원광대학교 전자융합공학과 연구교수

관심분야 : 퍼지 이론, 신경 회로망, Bio-inspired optimization algorithm, Pattern Recognition
E-mail : nado@wku.ac.kr

김용수(Yong Soo Kim)
1981년 연세대학교 전기공학과 공학사
1983년 KAIST 전기및전자공학과 공학석사
1986년 삼성전자종합연구소 주임연구원
1993년 Texas Tech Univ. 공학박사
1995년~현재 대전대학교 컴퓨터공학과 교수
관심분야 : 신경회로망, 퍼지논리, 패턴인식, 영상처리, 침입탐지 등
Phone : +82-42-280-2547
Fax : +82-42-280-2889
E-mail : kystj@dju.kr

안태천(Tae-Chon Ahn)
1978년 : 연세대학교 전기공학과 공학사
1980년 : 연세대학교 전기공학과 공학석사
1986년 : 연세대학교 전기공학과 공학박사
1981년 ~현재 : 원광대학교 전자융합공학과 교수
2013년~현재 : 한국지능시스템학회 이사

관심분야 : Computational Intelligence, Soft Computing Fuzzy Control, Pattern Recognition
Phone : +82-63-850-6344
E-mail : tcahn@wku.ac.kr