통계적 학습이론을 이용한 최적 군집화

An Optimal Clustering Using Statistical Learning Theory

*최준혁, **전성해, ***오경환

*김포대학 e-비즈니스과

**청주대학교 통계학과

*** 서강대학교 컴퓨터학과

e-mail: jhchoi@kimpo.ac.kr

요 약

모집단의 최적군집 수를 자동으로 결정하고 군집내의 분산은 최소로 하고 군집 간의 분산은 최대로 하는 최적 군집화에 대한 연구는 대부분의 지능형 시스템에서 필요로 하는 모형전략이다. 하지만 아직도 대부분의 군집화 과정에서 분석가의 주관적인 경험에 의존하여 군집수가 결정되어 군집화가 이루어지고 있다. 예를 들어 K-평균 군집화 알고리즘에서도 초기에 K 값을 결정해 주어야 한다. 모집단을 제대로 대표하지 못한 K 값에 의한 군집화 결과는 심각한 오류를 범하게 된다. 본 논문에서는 통계적 학습이론을 이용하여 이러한 문제점을 해결하려고 하였다. VC-차원에 의한 Support Vector를 이용하여 최적의 군집화 기법을 제안하였다. 제안 방법의 성능 평가를 위하여 UCI 기계학습 데이터를 이용하여 객관적인 실험을 수행하였다.

1. 서론

통계적 학습이론은 Vapnik에 의해 1960년대에 처음으로 소개되었다. 하지만 몇 가지의 제약조건들 때문에 그동안 많이 사용되지 않다가 1990년대에 SVM(Support Vector Machine)이라고 불리는 새로운 형태의 학습 알고리즘에 제안되면서 현재는 학습모형들 중에서 이론적 안정성과 성능의 우수함 때문에 많은 분야에서 다양하게 사용되고 있다[13],[14],[15]. 통계적 학습이론은 크게 분류(classification)와 예측(regression)을 위한 SVM과 SVR(support vector regression)이 있다. 최근에는 군집화(clustering)을 위한 SVC(support vector clustering)[1],[2],[3]까지 제안되어 통계적 학습이론은 지도학습(supervised learning)과 자율학습(unsupervised learning)의 학습모형을 모두 구현할 수 있는 알고리즘이 되었다[7],[10],[12]. 본 논문에서는 SVC를 이용하여 군집화를 위한 최적 군집수를 결정하는 방법을 제안하였다. 제안 모형의 성능평가를 위하여 UCI Machine Learning repository의 학습 데이터를 이용하였다[16].

2. 관련 연구

클래스 레이블들을 가진 목표변수 y와 입력벡터(input vector) x로 구성된 데이터 집합 S는 다음 식과 같은 데이터 구조로 표현된다[13],[14].

$$(y_1, x_1), (y_2, x_2), \cdots, (y_l, x_l), \quad x_i \in R^N, \; y_i \in \{-1, 1\} \quad (1)$$

대부분의 분류모형 구축의 경우에 입력공간(input space)에서 서로 다른 클래스 레이블을 분류하는 정확한 초평면(hyperplane)을 찾는 것은 매우 제한적이기 때문에 바로 분류 모형을 사용하기가 어렵다. 이러한 상황에서 해결 방안은 입력공간을 더 높은 차원의 특징 공간(feature space)으로 사상(mapping)시키고, 이 특징 공간에서 최적의 초평면을 찾는 것이다. $z = \phi(x)$를 입력공간 R^N에서 특징

공간 Z로의 사상 ϕ를 갖는 특징 공간 벡터로 표현하면, (w, b)의 쌍으로 이루어진 다음의 초평면을 구해야 한다.

$$w \cdot z + b = 0 \tag{2}$$

식 (2)의 초평면 식을 구하게 되면 다음의 식 (3)의 함수에 의해 개개의 x_i들을 분류할 수 있게 된다.

$$f(x_i) = sign(w \cdot z_i + b) = \begin{cases} 1 & \text{if } y_i = 1 \\ -1 & \text{if } y_i = -1 \end{cases} \tag{3}$$

여기서 $w \in Z$이고 $b \in R$이다. 특히, 집합 S는 (w, b)의 쌍이 존재하면 선형분류 가능(linearly separable)이라고 하고 다음의 부등식이 S의 모든 원소들에 대해 성립한다.

$$\begin{cases} (w \cdot z_i + b) \geq 1, & \text{if } y_i = 1 \\ (w \cdot z_i + b) \leq -1, & \text{if } y_i = -1 \end{cases} \quad i = 1, 2, \cdots, l \tag{4}$$

선형분류 가능한 집합 S는 두 개의 서로 다른 클래스 레이블들의 학습 데이터의 사영(projection)들 사이의 마진(margin)을 최대화 하는 유일한 최적 초평면을 구할 수 있다. 만약 집합 S가 선형 분류 가능이 아니면 분류규칙 위반(classification violations)이 SVM 형식에서 허용되어져야 한다. 선형분류 가능이 아닌 데이터를 다루기 위하여 음이 아닌 변수 ξ_i를 도입하여 아래 식과 같이 식 (4)를 일반화한다.

$$y_i(w \cdot Z_i + b) \geq 1 - \xi_i, \quad i = 1, 2, \cdots, l \tag{5}$$

식 (5)에서 ξ_i는 식 (4)을 만족하지 않는 x_i들이다. 그러므로 $\sum_{i=1}^{l} \xi_i$는 오분류(misclassification)의 양을 나타내는 측도로서 고려된다. 따라서 최적 초평면을 구하는 문제는 아래의 문제에 대한 해(solution)가 된다.

$$\begin{aligned} &minimize \ \frac{1}{2} w \cdot w + C \sum_{i=1}^{l} \xi_i \\ &subject \ to \ y_i(w \cdot z_i + b) \geq 1 - \xi_i \end{aligned} \tag{6}$$

여기서 $\xi_i \geq 0$이고 $i = 1, 2, \cdots, l$이다. C는 상수(constant)이며 조정 모수(regularization parameter)이다. 이 모수의 조정으로 마진 최대화와 분류 규칙 위반 사이의 균형을 맞출 수 있게 된다. 식 (6)에서 최적 초평면을 찾는 것은 다음의 라그랑지 변환(Lagrangian transformation)을 통하여 풀 수 있는 문제가 된다.

$$\begin{aligned} &maximize \ W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j z_i \cdot z_j \\ &subject \ to \ \sum_{i=1}^{l} y_i \alpha_i = 0 \qquad 0 \leq \alpha_i \leq C, \ i = 1, 2, \cdots, l \end{aligned} \tag{7}$$

여기서 $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_l)$는 식 (5)의 제한 조건과 관련된 음이 아닌 라그랑지 승수(multiplier)들의 벡터이다. Kuhn- Tucker 정리는 SVM 이론에서 중요한 역할을 한다. 이 정리에 의하여 식 (7)의 해 $\overline{\alpha_i}$는 다음을 만족한다.

$$\begin{aligned} &\overline{\alpha_i}(y_i(\overline{w} \cdot z_i + \overline{b}) - 1 + \overline{\xi_i}) = 0 \\ &(C - \overline{\alpha_i})\,\overline{\xi_i} = 0, \ i = 1, 2, \cdots, l \end{aligned} \tag{8}$$

식 (8)의 첫 번째 식으로부터 구한 해 $\overline{\alpha_i}$는 식 (5)의 등호를 만족시킨다. $\overline{\alpha_i} > 0$인 x_i를 support vector라고 부른다. 분류가 가능하지 않은(nonseparable) 경우에는 support vector는 두 가지의 형태로 존재한다. $0 < \overline{\alpha_i} < C$인 경우의 support vector x_i는 $y_i(\overline{w} \cdot z_i + \overline{b}) = 1$과 $\xi_i = 0$을 만족하고, $\overline{\alpha_i} = C$인 경우의 ξ_i는 널(null)이 아니고 대응되는 support vector x_i는 식 (4)을 만족하지 않는다. 이 support vector 들은 오차(error)로서 간주된다. $\overline{\alpha_i} = 0$에 대응되는 x_i는 결정 마진(decision margin)과 떨어져서 정확하게 분류된다. 최적 초평면 $\overline{w} \cdot z + \overline{b}$를 구축하기 위하여 다음의 식과 스칼라 $\overline{b}$가 필요하다.

$$\overline{w} = \sum_{i=1}^{l} \overline{\alpha_i} y_i z_i \tag{9}$$

이것은 식 (8)의 첫 번째 식의 Kuhn-Tucker 조건에 의해 결정된다. 결정 함수(decision function)는 식 (3)와 식 (9)에 의해 다음식과 같이 일반화된다.

$$f(x) = sign(w \cdot z + b) = sign\left(\sum_{i=1}^{l} \alpha_i y_i z_i \cdot z + b\right) \tag{10}$$

ϕ에 대한 어떠한 지식(knowledge)도 없기 때문에 식 (7)와 식 (10)의 계산은 불가능하다. 하지만 SVM은 ϕ에 대해서 알 필요가 없다. 단지 커널(kernel)이라 불리우는 $K(\cdot, \cdot)$가 다음과 같은 식에 의해 특징 공간 Z에 데이터의 내적(dot product)을 계산한다.

$$z_i \cdot z_j = \phi(x_i) \cdot \phi(x_j) = K(x_i, x_j) \tag{11}$$

Mercer의 정리를 만족하는 함수들은 내적 계산이 가능하고 따라서 커널로써 사용이 가능하다. SVM 분류기(classifier)를 구축하기 위하여 아래와 같은 차수(degree) d의 다항(polynimial) 커널을 사용한

다.

$$K(x_i, x_j) = (1 + x_i \cdot x_j)^d \quad (12)$$

따라서 비선형 분류 가능 초평면은 다음 식의 해로서 구해진다.

$$maximize\ W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$
$$subject\ to\ \sum_{i=1}^{l} y_i \alpha_i = 0 \quad 0 \le \alpha_i \le C, \quad i = 1, 2, \cdots, l \quad (13)$$

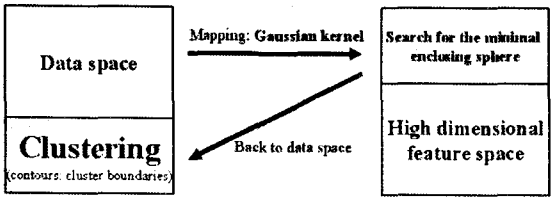
그리고 최종적인 결정 함수는 다음과 같다.

$$f(x) = sign(w \cdot z + b) = sign\left(\sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b \right) \quad (14)$$

제안 모형은 이러한 지지벡터(support vector)를 이용하여 최적의 군집수를 결정하고 이를 통하여 효율적인 군집화 결과를 이끌어 내었다.

3. 통계적 학습이론을 이용한 최적군집화

통계적 학습이론을 이용한 군집화는 SVM과 마찬가지로 지지벡터를 이용한다. 주어진 학습데이터의 데이터 점들(data points)은 가우시안 커널(Gaussian kernel)에 의해 고차원의 특징공간(high dimensional feature space)으로 사상(mapping)된다. 이 공간에서 주어진 데이터 점들을 그룹화 할 수 있는 최소 경계구면(minimal enclosing sphere)을 찾는다. 이 구면은 각 데이터점이 고차원의 특징공간에서 다시 주어진 데이터공간으로 사상될 때 데이터 점들의 분리된 군집을 결정할 수 있는 몇 개의 집단을 구분해 준다. 다음 그림은 이러한 통계적 학습이론에 의한 최적 군집화(SLT-OC) 과정을 요약해서 보여준다.



(그림 1) SLT-OC 절차

$x_i \subseteq X(X \subseteq R^d)$가 N개의 데이터 집합일 때 데이터 공간 X로부터 고차원 특징공간으로의 비선형 변환(nonlinear transformation) Φ를 이용하여 반지름 R의 최소 경계구면을 구한다. 이것은 다음과 같은 제한조건을 갖는다.

$$\|\Phi(x_j) - a\|^2 \le R^2 \ \forall_j \quad (15)$$

여기서 $\|\cdot\|$는 유클리디안 노름(Euclidean norm)이고 a는 구면의 중심이다. Soft 제한조건은 여유변수(slack variable) ξ_j를 추가한 다음 식으로 정의된다.

$$\|\Phi(x_j) - a\|^2 \le R^2 + \xi_j, \quad \xi_j \ge 0 \quad (16)$$

이 문제를 풀기위하여 다음의 랑그랑지 기법(Lagrangian)을 사용한다.

$$R^2(x) = K(x, x) - 2\sum_j \beta_j K(x_j, x) + \sum_{i,j} \beta_i \beta_j K(x_i, x_j) \quad (17)$$

여기서 $\beta_j \ge 0$과 $\mu_j \ge 0$은 라그랑지 승수(Lagrange multipliers)이다. C는 상수이고 $C\sum\xi_j$는 패널티 항(penalty term)이다.

4. 실험 및 결론

본 논문의 실험을 위하여 UCI의 기계학습 데이터베이스로부터 Iris plants 데이터와 Glass identification 데이터를 이용하였다. Iris 데이터는 붓꽃의 종류를 나타내는 1개의 목표변수와 꽃의 외형을 표현하는 4개의 입력변수로 되어 있으며 총 데이터의 개수는 150개이다. Glass 데이터는 총 214개의 데이터 점들로 이루어져 있다. 입력변수들은 굴절률 변수와 성분을 나타내는 8개의 변수들(나트륨, 마그네슘, 알루미늄, 실리콘, 칼륨, 칼슘, 바륨, 철)로 총 9개로 이루어져 있다. 유리의 종류를 나타내는 1개의 목표변수가 있다. 본 논문의 제안 모형을 통하여 Iris와 Glass 데이터 집합들을 위한 최적 군집수는 각각 3개와 7개로 나왔다. 다음은 군집수를 3개와 7개로 했을때의 비교모형들[4],[5],[6],[8],[9][11] 간의 정확도 결과를 보여준다.

<표 1> Iris plants 데이터의 정확도

Method	Accuracy(%)
SOM	88.0
K-means	93.3
Hierarchical	80.7
SLT-OC	94.5

<표 2> Iris plants 데이터의 정확도

Method	Accuracy(%)
SOM	80.8
K-means	89.3
Hierarchical	86.0
SLT-OC	92.3

위의 실험결과를 통하여 본 논문의 제안 모형인 SLT-OC가 가장 정확도가 높음을 알 수 있었다.

5. 결론 및 향후과제

본 논문에서는 Vapnik이 제안한 통계적 학습이론을 군집화에 적용한 SVC를 이용하여 최적 군집수를 결정하고 이를 바탕으로 정확한 군집화를 수행하는 방법을 소개하였다. 기존의 대표적인 군집화 기법들에 비해 제안 기법이 좀 더 향상된 결과를 제공하고 있음을 실험을 통하여 확인하였다. 앞으로 바이오 혹은 유비쿼터스 컴퓨팅 환경에서 발생하는 다양한 데이터의 군집화에 적용해 볼 수 있을 것이다.

Acknowledgement

This research (paper) was performed for the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Ministry of Commerce, Industry and Energy of Korea.

참고문헌

[1] A. Ben-Hur, A. Elisseeff, and I. Guyon. A stability based method for discovering structure in clustered data. in Pacific Symposium on Biocomputing, 2002.

[2] A. Ben-Hur, D. Horn, H.T. Siegelmann, and V. Vapnik. A support vector clustering method. in International Conference on Pattern Recognition, 2000.

[3] A. Ben-Hur, D. Horn, H.T. Siegelmann, and V. Vapnik. A support vector clustering method. in Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference, Todd K. Leen, Thomas G. Dietterich and Volker Tresp eds., 2001.

[4] R.O. Duda, P.E. Hart, and D.G. Stork. Pattern Classification. John Wiley & Sons, New York, 2001.

[5] K. Fukunaga. Introduction to Statistical Pattern Recognition. Academic Press, San Diego, CA, 1990.

[6] A.K. Jain and R.C. Dubes. Algorithms for clustering data. Prentice Hall, Englewood Cliffs, NJ, 1988.

[7] H. Lipson and H.T. Siegelmann. Clustering irregular shapes using high-order neurons. Neural Computation, 12:2331.2353, 2000.

[8] J. MacQueen. Some methods for classification and analysis of multivariate observations. in Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, 1965.

[9] G.W. Milligan and M.C. Cooper. An examination of procedures for determining the number of clusters in a data set. Psychometrika, 50:159.179, 1985.

[10] J. Platt. Fast training of support vector machines using sequential minimal optimization. in Advances in Kernel Methods . Support Vector Learning, B. Sch¨olkopf, C. J. C. Burges, and A. J. Smola, editors, 1999.

[11] B.D. Ripley. Pattern recognition and neural networks. Cambridge University Press, Cambridge,1996.

[12] S.J. Roberts. Non-parametric unsupervised cluster analysis. Pattern Recognition, 30(2): 261.272, 1997.

[13] V. N. Vapnik, The Nature of Statistical Learning Theory, Springer, 1995.

[14] V. N. Vapnik, Statistical Learning

Theory, John Wiley & Sons, Inc., 1998.
[15] V. N. Vapnik, An Overview of Statistical Learning Theory, IEEE Transactions on Neural Networks, Vol. 10, no. 5, 1999.
[16] UCI Machine Learning Repository, http://www.ics.uci.edu/~mlearn