2001년 한국정보처리학회 춘계 학술발표논문집 제8권 제1호

진화 알고리즘을 적용한 효율적 군집화 기법

이수정*, 권혜련, 김은주, 이일병
연세대학교 컴퓨터과학과
e-mail : {crystal, comtrue, outframe, yblee}@csai.yonsei.ac.kr

An Efficient Clustering using the Genetic Algorithm

Soo-Jung Lee*, Hye-Ryun Kwon, Eun-Ju Kim, Yill-Byung Lee
Dept. of Computer Science, Yonsei University

요 약

최근 들어 관심의 대상이 되고 있는 CRM, eCRM 은 비즈니스 분야에 중요한 역할을 담당하고 있다. 이를 위해 여러 방법들이 사용되고 있으나, 그 중 데이터 마이닝은 핵심 기술이라 할 수 있다. 다양한 데이터 마이닝 기법가운데 군집화 기법은, 데이터 집합을 유사한 데이터 개체들의 군집들로 분할하여 데이터 속에 존재하는 의미 있는 정보를 얻는 과정이다. 그런데 기존의 군집화 알고리즘들은 사전에 군집의 개수를 미리 결정해줘야 하며, 지역적 최적해(local minima)에 수렴할 수 있다는 문제점을 가지고 있다. 본 논문에서는 진화 알고리즘을 사용하여 자동적으로 적절한 군집의 개수를 결정하여 군집화 될 수 있도록 하고, 병렬 탐색을 통해 지역적 최적해에 수렴되는 문제점을 개선한 알고리즘과 적합도 함수를 제안한다.

1. 서론

최근 들어 CRM, eCRM 이 많은 관심의 대상이 되고 있다. CRM(Customer Relationship Management, 고객관계관리)은 선별된 고객으로부터 수익을 창출하고 장기적인 고객관계를 가능케 하는 솔루션을 말한다. 즉 CRM 은 고객과 관련된 기업의 내외부 자료를 분석, 통합하여 고객특성에 기초한 마케팅 활동을 계획하고 지원하며 평가하는 과정이다.

비즈니스 영역에 중요한 위치를 차지하고 있는 CRM, eCRM 에는 여러 가지 방법이 사용된다. 그 가운데 가장 중요한 기법 중 하나가 데이터 마이닝이다. 데이터 마이닝(Data Mining)이란 대용량의 데이터베이스로부터 아직 알려지지 않았으나 의미 있는 패턴을 지식의 형태로 추출하는 작업이다[5]. 기존에 기계 학습 분야에 KDD(Knowledge Discovery in Database)라는 개념으로 존재하다가 최근 데이터 마이닝이라는 개념으로 발전되었다. 데이터 마이닝은 기계 학습, 인공지능, 데이터베이스, 통계학 등 다른 연구 분야로부터 발전된 최신 데이터 분석 기술로서, 정보 분석에 있어 기존의 Query, OLAP 툴, 통계적 기법들이 제공하는 것 이상의 역할을 수행한다. 기존의 기법들이 분석가에게 부담이 큰 반면, 데이터 마이닝은 많은 부담을 컴퓨터에게 전이한 것이다.

데이터 웨어하우스로부터 의미있는 지식을 찾아내는 데이터 마이닝의 주된 기능은, 네 가지 정도로 요약된다.

먼저, 데이터 마이닝의 대표적 기법인 군집화(Clustering)는, 데이터 분류와 이미지 처리 같은 많은 실용적 문제영역에 적용되고 있는 탐색적 데이터 분석의 중요한 기법 중 하나이다. 군집화는 입력 데이터 집합을 유사한 관찰값들의 군집들로 구분하여 데이터 집합속에 존재하는 의미 있는 정보를 얻는 과정이다[1][2]. 즉, 군집내의 유사성은 최대화하고 군집들간의 유사성은 최소화 시키도록 데이터 집합을 분할하는 것이다[3]. 이러한 군집발견 과정은 우리에게 군집의 데이터 분포가 갖고 있는 특징을 설명하며 다른 분석 기법을 위한 토대를 마련해주는 역할을 수행한다[4]. 따라서 군집화 기법은 공학, 생명과학, 금융, 마케팅 등 다양한 분야에서 응용되고 있다. 예를 들어, 기업에서 고객을 구매패턴에 근거해서 분류하거나, 웹 문서의 범주별 분류에 이용하거나, 유사한 기능을 하는 유전자를 분류하는 등 다양한 응용 분야에 적용이 가능하다. 최근에는 데이터 마이닝의 출현으로 인해 원시데이터에 대한 접근회수를 줄이고 알고

리즘이 다루어야 할 데이터 구조의 크기를 줄이는 군집화 기법에 관한 연구들이 활발하다.

데이터 마이닝에서 많이 사용되는 또 다른 기법으로 분류(classification)가 있다. 이것은 데이터베이스 내의 객체의 집합에 대하여 그 안에 내재하는 공통 특성을 뽑아내어 이 객체들을 서로 다른 클래스로 분류해내는 작업이다. 분류에 주로 사용되는 지능형 알고리즘으로는 오류 역전파 알고리즘(Back propagation)을 이용한 다층 신경회로망(Neural Network), 유전자 알고리즘(Genetic Algorithm), 의사결정트리(Decision Tree), 퍼지 알고리즘(Fuzzy Algorithm), 사례기반추론(Case Based Reasoning) 등을 들 수 있으며, 통계적인 방법론으로는 분류하는데에 주로 사용되는 판별분석(Discriminant Analysis)과 연속된 값의 예측에 주로 사용되는 회귀분석(Regression) 등을 들 수 있다. 이러한 분류 기법은 주로 고객의 신용 평가나 사기 검출(Fraud Detection) 등에 이용된다.

이 외에 예측(Prediction) 기능은 주가 예측, 고객 수요 예측, 고객 이탈 예측, 제품 가격 산출 등에 이용되고, 연관(Association)은 주로 장바구니분석(Market Basket Analysis)로서 이용되고, Cross Selling, Inventory Display 등에 사용된다.

본 논문의 구성은 다음과 같다. 2 장에서는 기존의 알고리즘과 문제점들을 살펴보고, 3 장에서는 군집화 알고리즘을 제안한다. 4 장에서는 실험을 통한 결과를 요약하고, 5 장에서 결론을 맺는다.

2. 관련 연구

1 장에서 설명한 데이터 마이닝의 기능 중에 군집화는 주어진 데이터 혹은 개체를 군집(clusters, subsets, groups, classes)으로 분할하는 것이다. 군집화는 특별한 정보나 배경지식 없이 데이터들 간의 주어진 척도를 이용하여 결과를 이끌어내므로 비교사학습(unsupervised learning)에 속하는 패턴 분류 방법이다. 즉, 주어진 자료들에 숨겨져 있는 다양한 패턴을 발굴하여 다음 분석과정의 토대를 제공하는, 데이터 마이닝의 기초 작업이라고 할 수 있다. 비즈니스 영역에 응용되는 군집화의 예를 보면, 기업의 마케팅 전략수립과 관련하여 소비자 행동의 유사성을 바탕으로 시장 세분화에 이용된다. 마케팅 의사결정 과정에서 소비자의 상품구매 형태나 생활 방식에 따라 몇 개의 소비자 군으로 분할하여 시장 세분화 전략을 수립한다. 또한 기존 상품을 몇 개의 속성으로 구분하여 이를 군집화 함으로써 새로운 시장 진출 가능성을 미리 예측하기도 하며, 특히 군집화를 통해 나타나는 정보를 고객 세분화에 이용하여 Direct Mail 발송 혹은 표적 시장 마케팅(Target Marketing) 등에 활용한다.

군집화 알고리즘은 크게 두 가지로, 분할적 군집화(Partitional Clustering)와 계층적 군집화(Hierarchical Clustering)로 나눌 수 있다[3][9]. 분할적 군집화[3][9]는 주어진 목적함수를 최적화 하도록 데이터 집합을 k 개의 군집으로 나누는 것으로, K-means 등이 이에 속한다. 임의의 초기 분할로부터 시작하여, 데이터 개체에 대한 소속 군집의 재할당 과정과 목적함수의 평가를 반복적으로 수행하여 목적함수를 최적화한다. 계층적 군집화[3][10]는 가장 유사한 두 개체들을 선택하여 병합해가는 병합적 계층 군집화 방법과 가장 먼 개체들을 선택하여 나누어 나가는 분할적 계층 군집화 방법이 있다. 두 군집의 유사도를 측정하는 기준에 따라 최단연결법, 최장 연결법, 중심연결법, 평균연결법 등으로 나뉜다[3].

3. 제안하는 알고리즘

군집화 문제에 있어서 최적의 군집을 찾아내는 것은 NP-complete 문제로 알려져 있다. 또한 어떻게 이루어진 군집화가 최적인가에 대한 수학적 모델이 아직 알려지지 않았다.

분할적 군집화의 대표적인 알고리즘인 K-means 알고리즘[11]과 Fuzzy C-Means(FCM)[12]알고리즘은 모든 데이터로부터 각각의 군집 중심까지의 거리와 제곱의 합으로 정의되는 목적함수(object function)를 최소화하는데 바탕을 둔 알고리즘이다. 이들 알고리즘의 목적함수는 단지 같은 군집내의 유사성을 중심과 입력 데이터간의 거리만으로 고려하고 있기 때문에 유사성이 가장 높은 경우는 각각의 데이터가 군집의 중심이 된다. 그러므로 목적함수의 값은 군집의 개수와 입력데이터의 개수가 같은 경우에 가장 최소의 값을 갖는다. 따라서 사전에 군집의 개수를 결정해주어야 하며, 초기 군집의 중심 설정과 잡음에 따라 알고리즘의 성능이 민감하게 좌우되는 문제점이 있다. 그래서 최근 자동으로 군집의 개수를 결정해주기 위해 통계적 기법이나 진화 알고리즘을 적용하는 연구가 이루어지고 있다[8][11][12][13]. 또한 지역적 최적해(local minima)에 수렴될 수 있는 문제점을 해결하기 위해 진화 알고리즘을 이용하는 연구가 이루어지고 있다[11][12].

본 논문에서 적용하고 있는 진화 알고리즘은 자연계에 있어서 생물의 유전과 진화의 메커니즘을 공학적으로 모델화하여 생물이 환경에서 갖는 적응능력을 모방한 것으로, 전역적 탐색 기법의 하나이다. 이것은 1975 년 John Holland(당시 Michigan 대학교)에 의해 제안된 자연도태와 진화의 원리에 기반을 둔 확률적인 탐색 알고리즘이며, 특히 탐색 및 최적화, 기계학습의 도구로 많이 사용되고 있다[14]. 제안한 알고리즘의 흐름 역시, 일반적인 진화 알고리즘과 유사하다.

3.1. 개체 표현 및 개체군 초기화

일반적으로 진화 알고리즘은 문제에 대한 후보해(candidate solution) 또는 개체(chromosome)들의 집단인 개체군(population)을 유지한다.

본 논문에서는 각 군집의 중심값으로 개체를 표현하고, 각각의 개체는 임의의 값에 의해 가변길이를 갖도록 하였다.

3.2. 적합도 함수(fitness function)를 이용한 개체 평가

진화 알고리즘의 성능은 적합도 함수(fitness function)에 많은 영향을 받으므로, 적절한 적합도 함수를 정의하는 것은 매우 중요한 문제이다. 적합도란 임의의 개체가 문제의 해에 얼마나 적합한지를 나타내는 척도이다. 따라서 문제의 해가 될 가능성이 있는 것들을 평가하는 환경의 역할을 수행하는 것이 적합도 함수이다.

1971년 Cormack은 응집성(compactness)을, 1980년 Gordon은 분리성(separation)을 이용한 군집화를 정의하게 위해 시도하였으며, 최근 이 두 가지 평가 척도를 고려한 목적함수를 제안한 연구도 발표되었다[8]. 이러한 목적함수의 경우 응집성은 작고 분리성은 큰 값을 가질 때 군집화가 잘 이루어지게 된다.

본 논문에서는 군집간의 연관성과 특징을 고려한 유사도 값[15]을 적합도 함수로 이용한다. 즉 군집의 두 개의 대표값을 가지고 군집의 내부적 특징인 응집거리과 군집간의 외부적 거리를 나타내는 근접거리를 계산하고, 이를 이용하여 두 군집간의 유사도를 나타낸다.

$$\text{유사도}_{ij} = \frac{1}{\text{응집거리}_{ij} \times \text{근접거리}_{ij}^2}$$

$$\text{응집거리}_{ij} = \frac{\text{연결거리}_{ij}}{\frac{\text{연결거리}_i + \text{연결거리}_j}{2}}$$

$$\text{근접거리}_{ij} = \frac{\sum_{a}^{ni}\sum_{b}^{nj} w_{ra} \times w_{rb} \times \|r_a - r_b\|^2}{n_i \times n_j}$$

$$\text{연결거리}_{ij} = \sum_{a}^{ni}\sum_{b}^{nj} w_{ra} \times w_{rb} \times \|r_a - r_b\|^2$$

$$\text{연결거리}_i = \frac{\sum_{a}^{ni}\sum_{b}^{ni} w_{ra} \times w_{rb} \times \|r_a - r_b\|^2}{2}$$

r_a, r_b : 대표값 벡터

n_i : 소군집 i에 속하는 대표값의 개수

w_{ra} : 대표값 r_a가 대표하는 원시데이터 개체 수

3.3. 선택 (selection)

부모 개체를 선택하고 다음 세대의 개체 집단을 선택하는 방법으로 룰렛휠(roulette wheel)방법을 사용한다. 룰렛휠 방법은 우수한 성질의 염색체에 보다 많은 선택의 기회를 주는 방법이다. 또한 현재 개체 집단에서 가장 우수한 성질을 갖는 염색체를 다음세대에도 확보하기 위해 엘리트 방법(elitist model)을 병행하여 사용한다[14].

3.4. 교배 연산 (Crossover)

유전인자들이 가변길이이며 위치에 상관없기 때문에, 부모 염색체에서 공통으로 가지고 있는 유전인자는 자식세대에 그대로 전해진다. 반면 서로 다른 유전인자는 생성된 자식 염색체 모두가 가지거나, 모두 갖지 않거나, 또는 자식 염색체 중 한 염색체만 가질 수 있다.

3.5. 돌연변이 (Mutation)

진화 알고리즘에서 돌연변이 연산은 한 개체에서 임의로 선택된 유전인자를 임의의 가능한 다른 값으로 바꾸는 것이다. 이로써, 현재 개체군에 존재하지 않는 새로운 개체를 생성하며 개체군의 다양성을 유지한다.

본 논문에서는 정규적 돌연변이 연산자와 가우시안 분포 랜덤 변수(Gaussian distribution random variable)를 사용한다.

정규적 돌연변이 연산에서는 개체군 초기화 단계에서 고려된 예비후보 중심값들중 하나를 선택하여 돌연변이 확률에 의해 선택된 유전인자의 값과 바꾸는 것이다. 가우시안 분포 랜덤 변수의 경우는, 가우시안 함수를 사용하여 돌연변이 확률에 의해 선택된 특정 유전인자의 값을 바꿔주는 것이다.

4. 실험 결과

제안된 적합도 함수를 적용한 진화 알고리즘으로 군집화가 이루어지는지에 대한 실험을 하였다. 먼저 적합한 군집의 개수를 자동적으로 찾아낼 수 있는지를 알아보기 위해 이차원 데이터를 가지고 실험하였다. 그리고, 가우시안 함수를 이용하여 임의로 생성시킨 500 개의 패턴과 20 개의 군집을 갖는 실험 데이

터를 생성하여 본 알고리즘을 적용해 보았다.

매개 변수	값
진화회수	1000
개체집단의 크기	30
교배 연산자의 확률	0.5
돌연변이 연산자의 확률	0.05

<표 1> 실험 데이터 변수

실험결과 FCM, Isodata 등 기존의 알고리즘은 지역적 최적해에 수렴하여 적절한 중심을 찾지 못한 반면, 제안한 알고리즘은 자동적으로 군집의 개수를 찾아낼 뿐만 아니라 비교적 정확한 군집을 형성하고 있는 것을 보인다.

5. 결론

본 논문에서는 진화 알고리즘을 이용하여 자동으로 군집의 개수를 결정하는 군집화 알고리즘을 제안하였다. 이는 군집의 초기치를 잘못 설정했을 경우 발생할 수 있는 부정확한 군집화 결과를 방지할 수 있다. 또, 제안하는 알고리즘의 적합도 함수는 보다 정확한 군집을 찾아내는 것으로 평가되었다.

향후 계획으로는 알고리즘을 좀 더 최적화하여 데이터 마이닝에서 사용하는 실제 데이터들에 적용해 보고자한다. 특히 본 논문의 적합도 함수 정의에서 사용한 군집의 대표값 개념은 요약 정보만을 이용하기 때문에 계산속도가 향상되어 대용량 데이터를 사용하는 데이터 마이닝에 적합할 것으로 기대된다.

참고문헌

[1] Tian Zhang, Raghu Ramakrishnan, and Miron, " Birch: An efficient data clustering method for very large databases" , the ACM SIGMOD Conference on Management of Data, Montreal, Canada, June 1996.

[2] Tian Zhang, Raghu Ramakrishnan, and Miron, Birch: A New Data Clustering Algorithm and Its Applications," Data Minning and Knowledge Discovery, 1, 141-182, 1997.

[3] Richard O. Duda and Peter E. Hard, Pattern Classification and Scene Analysis, A Wiley-Interscience Publication, New York, 1973.

[4] Berry, Linoff, Data Mining Techniques for Marketing, Sales, and Customer Support, Jone Wiley & Sons, 1997.

[5] Fayyad, Piatetsky-Shapiro, Smyth, " Advances in knowledge discovery and data mining" , 1996.

[6] J. T. Tou, R. C. Gonzalez, "Pattern Recognition Principles", Addison-Wesley Publishing Company, Inc. pp. 75-109, 1974.

[7] George J. Klir, Bo Yuan, "Fuzzy Sets and Fuzzy Logic", Prentice-Hall Inc. 1995.

[8] 김명원, 류정우, "진화 알고리즘을 이용한 클러스터링 알고리즘", 2000 봄 학술발표논문집(B) 제 27 권 1 호 pp. 313-315, 2000.

[9] Kaufman, Leonard and Rousseuw, Peter J., Finding Groups in Data - An Introduction to Cluster Analysis, Wiley Series in Probability and Mathematical Statistics, 1990.

[10] Murtagh, F., " A Survey of Rescent Advances in Hierarchical Clustering Algorithms" , The Computer Journal, 1983.

[11] Susu Yao, "Evolutionary Search Based Fuzzy Self-Organising Clustering", Congress on Evolutionary Computation,pp. 185-188. 1999.

[12] K.Krishna and M. Narasimha Murty, "Genetic K-Means Algorithm", IEEE Trans. Syst., Man, Cybern., VOL. 29, No. 3, pp. 433-439,1999.

[13] G.Phanendra Babu and M. Narasimha Murty, "Clustering with Evolution Strategies", Pattern Recognition VOL 27, No.2 pp. 321-329, 1994.

[14] Z. Michalewicz, "Genetic Algorithm + Data Structures = Evolution Programs", Third, Extended Edition, Springer-Verlag, 1995.

[15] 안병주, 김은주, 이일병, " 데이터 마이닝을 위한 계층적 대표값 군집화 기법" , 정보과학회 학술발표논문집, 제 27 권, 2000.