유전자 알고리즘을 이용한 군집화 기법의 적합도 함수에 관한 연구

이수정[0] 권혜련 김은주 이일병

연 세 대 학 교 컴 퓨 터 과 학 과

{crystal, comtrue, outframe, yblee}@csai.yonsei.ac.kr

A Study on Fitness Function of Clustering Algorithm based on Genetic Algorithm

Soo-Jung Lee[0] Hye-Ryun Kwon Eun-Ju Kim Yill-Byung Lee

Dept. of Computer Science, Yonsei University

요 약

최근 관심의 대상이 되고 있는 CRM, eCRM에는 데이터 마이닝 기법이 핵심 기술로 이용되고 있다. 이러한 데이터 마이닝 기법가운데 가장 널리 사용되고 있는 군집화는, 데이터 집합을 유사한 데이터의 군집들로 분할하여 데이터 속에 존재하는 의미 있는 정보를 얻는 것이다. 그런데 기존의 군집화 알고리즘은 사전에 군집의 개수를 미리 결정해줘야 하고 잡음에 민감하며 지역적 최적해(local minima)에 수렴할 수 있다는 문제점을 가지고 있다. 이러한 문제점의 개선을 위해, 본 논문에서는 유사도 개념을 적합도 함수로 사용하는 유전자 알고리즘을 적용한 군집화 기법을 제안한다. 특히 적합도 함수에 사용된 군집의 대표값 개념은 요약 정보만을 이용하여 계산속도가 향상되기 때문에 대용량 데이터를 다루는 데이터 마이닝에 적합할 것으로 기대된다.

1. 서론

최근 많은 관심의 대상이 되고 있는 CRM(Customer Relationship Management, 고객관계관리)은 고객과 관련된 기업의 내외부 자료를 분석하고 통합하여 고객의 특성에 기초한 마케팅 활동을 계획하고 지원하며 평가하는 과정이다. 비즈니스 영역에 중요한 위치를 차지하고 있는 CRM, eCRM에는 사용되는 가장 중요한 기법중 하나가 데이터 마이닝(Data Mining)이다. 데이터 마이닝이란 대용량의 데이터베이스로부터 아직 알려지지 않았으나 의미 있는 패턴을 지식의 형태로 추출하는 작업이다[4]. 데이터 마이닝은 기계 학습, 인공지능, 데이터베이스, 통계학 등 다른 연구 분야로부터 발전된 최신 데이터 분석 기술이며, 정보 분석에 있어 기존의 Query, OLAP 툴, 통계적 기법들이 제공하는 것 이상의 역할을 수행한다.

데이터 마이닝의 주된 기능은, 네 가지 정도로 요약된다. 대표적 기법인 군집화(Clustering)는, 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]. 따라서 군집화 기법은 공학, 생명과학, 금융, 마케팅 등 다양한 분야에서 응용되고 있다. 예를 들어, 기업에서 구매패턴에 근거한 고객의 분류, 웹 문서의 범주별 분류, 유사한 기능을 하는 유전자의 분류 등 다양한 응용 분야에 적용이 가능하다. 최근에는 대용량 데이터를 다루는 데이터 마이닝의 출현으로, 원시데이터에 대한 접근회수를 줄이고 알고리즘이 다루어야 할 데이터 구조의 크기를 줄이는 군집화 기법에 관한 연구들이 활발하다. 이외에 데이터베이스 내의 개체의 집합에 대하여 그 안에 내재하는 공통 특성을 뽑아내어 이 개체들을 서로 다른 클래스로 분류해내는 작업인 분류(classification)가 있다. 또, 예측(Prediction) 기능은 주가 예측, 고객 수요 예측, 고객 이탈 예측, 제품 가

격 산출 등에 이용되고, 연관(Association)은 주로 장바구니분석(Market Basket Analysis), Cross Selling, Inventory Display 등에 사용된다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 군집화 알고리즘과 적합도 함수에 대한 관련연구를 살펴보고, 3장에서는 유전자 알고리즘을 적용한 군집화 알고리즘을 제안한다. 4장에서는 실험 결과를 요약하고, 5장에서 결론을 맺는다.

2. 관련연구

데이터 마이닝의 기능 중에 군집화는 주어진 데이터를 군집으로 분할하는 것이다.

군집화 알고리즘은 분할적 군집화와 계층적 군집화로 나눌 수 있다. 분할적 군집화(Partitional Clustering)는 주어진 목적함수를 최적화 하도록 데이터 집합을 k개의 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(Hierarchical Clustering)는 가장 유사한 두 개체들을 선택하여 합병해가는 병합적 계층 군집화 방법과 가장 먼 개체들을 선택하여 나누어가는 분할적 계층 군집화 방법이 있다[3].

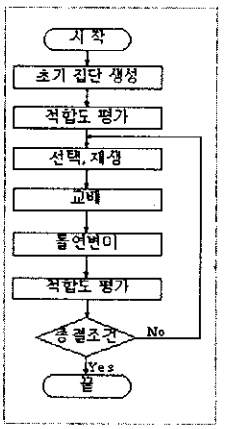
본 연구에서는 유전자 알고리즘을 군집화에 적용한다. 특히 유전자 알고리즘의 성능을 좌우한다고 할 수 있는 적합도 함수(fitness function)에 중점을 두고 있다. 적합도란, 임의의 개체가 문제의 해에 얼마나 적합한지를 나타내는 척도이다. 따라서 문제의 해가 될 가능성이 있는 것들을 평가하는 환경의 역할을 수행하는 것이 적합도 함수이다. 이것은 일종의 목적함수(object function)로서, 각 개체의 적합도를 평가하는 기반이 된다.

Cormack은 1971년 응집성(compactness)을 이용한 군집화를 정의했고, Gordon은 1980년 분리성(separation) 개념을 군집화의 정의에 적용하기 위해 시도하였다. 최근에는 이 두 가지 평가 척도를 고려한 목적함수를 제안한 연구가 발표되었다[9]. 이러한 목적함수의 경우 응집성은 작고 분리성은 큰 값을 가질 때 군집화가 잘 이루어지게 된다.

3. 제안하는 알고리즘

군집화 문제에 있어서 최적의 군집을 찾아내는 것은 NP-complete문제로 알려져 있다. 또한 어떻게 이루어진 군집화가 최적인가에 대한 수학적 모델이 아직 알려져 있지 않다.

분할적 군집화의 대표적 알고리즘인 K-means 알고리즘과 Fuzzy C-Means(FCM) 알고리즘은 모든 데이터로부터 각각의 군집 중심까지의 거리와 제곱의 합으로 정의되는 목적함수를 최소화하는데 바탕을 둔 알고리즘이다. 이들 목적함수는 단지 같은 군집내의 유사성을 중심과 입력 데이터간의 거리만으로 고려하기 때문에 각각의 데이터가 군집의 중심이 되는 경우 유사성이 가장 높게 된다. 즉, 목적함수의 값은 군집의 개수와 입력데이터의 개수가 같을 때 가장 최소값을 갖는다. 따라서 사전에 군집의 개수를 결정해주어야 하며, 초기 군집의 중심 설정과 잡음에 따라 알고리즘의 성능이 민감하게 좌우되는 문제점이 있다. 그래서 최근 통계적 기법이나 유전자 알고리즘을 적용하여 자동으로 군집의 개수를 결정해주고자 하는 연구가 이루어지고 있다[6][7][8][9]. 또한 지역적 최적해(local minima)에 수렴될 수 있는 문제점을 해결하기 위해 유전자 알고리즘을 이용하는 연구도 진행되고 있다[6][7].



본 논문에서 적용하고 있는 유전자 알고리즘은 자연계에 있어서 생물의 유전과 진화의 메커니즘을 공학적으로 모델화하여 생물이 환경에서 갖는 적응능력을 모방한 것으로, 전역적 탐색 기법의 하나이다. 자연도태와 진화의 원리에 기반을 둔 확률적인 탐색 알고리즘이며, 특히 탐색 및 최적화, 기계학습의 도구로 많이 사용되고 있다[5]. 제안한 알고리즘 역시, 왼쪽 그림에 보이는 일반적인 유전자 알고리즘의 흐름과 유사하다.

(1) 개체군 초기화

일반적으로 유전자 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.

(2) 적합도 함수(fitness function)

유전자 알고리즘의 성능이 적합도 함수에 의해 좌우될 수 있는 만큼, 적절한 적합도 함수를 정의하는 것은 매우 중요한 문제이다.

본 논문에서는 군집간의 연관성과 특징을 고려한 유사도(similarity) 값[10]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 측정하고, 그 값이 적합도 함수로서 이용된다. 유사도 개념은 다음 페이지의 박스 안에 정리되었다.

(3) 선택 (selection)

우수한 성질의 염색체에 보다 많은 선택의 기회를 주는 룰렛휠(roulette wheel)방법을 사용한다. 또한 현재 개체 집단에서 가장 우수한 성질을 갖는 염색체를 다음 세대에도 확보하기 위해 엘리트 방법(elitist model)을 함께 사용한다[5].

(4) 교배 (Crossover)

본 연구에서는 유전인자들이 고정길이가 아닌 가변길이이며 위치에 상관없게 정의되었다. 따라서 부모 염색체에서 공통으로 가지고 있는 유전인자는 자식세대에 그대로 전해진다. 반면 부모 염색체의 서로 다른 유전인자는 생성된 자식 염색체 모두가 가지거나, 모두 갖지 않거나, 또는 자식 염색체 중 한 염색체만 가질 수 있게 정의하였다.

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

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

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

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

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

r_a, r_b : 대표값 벡터
n_i : 소군집 i에 속하는 대표값의 개수
w_{ra} : 대표값 r_a가 대표하는 원시데이터 개체 수

(5) 돌연변이 (Mutation)

돌연변이 연산은 한 개체에서 임의로 선택된 유전인자를 임의의 가능한 다른 값으로 바꾸는 것이다. 이는 현재 개체군에 존재하지 않는 새로운 개체를 생성하여 개체군의 다양성을 유지하기 위함이다.

본 논문에서는 정규적 돌연변이 연산자와 가우시안 함수를 사용하여 돌연변이 확률에 의해 선택된 특정 유전인자의 값을 바꿔주는 방법을 함께 사용한다.

4. 실험 결과

제안된 적합도 함수를 적용한 유전자 알고리즘으로 군집화에 대한 실험을 해 보았다. 먼저 적합한 군집의 개수를 자동적으로 찾아낼 수 있는지를 알아보기 위해 이차원 데이터를 가지고 실험하였다. 그리고, 가우시안 함수를 이용하여 임의로 생성시킨 500개의 패턴과 20개의 군집을 갖는 실험 데이터를 생성하여 본 알고리즘을 적용해 보았다. 사용된 실험 데이터 변수는 아래 표와 같다.

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

실험 결과 FCM, Isodata 등 기존의 알고리즘은 지역적 최적해에 수렴하여 적절한 중심을 찾지 못하는 경향을 보이는 반면, 제안한 알고리즘은 자동적으로 군집의 개수를 찾아낼 뿐만 아니라 비교적 정확한 군집을 형성하고 있다.

5. 결론

본 연구에서는 유전자 알고리즘을 이용하여 자동으로 군집의 개수를 결정하는 군집화 알고리즘을 제안하였다. 제안된 기법은 기존 알고리즘의 단점인 군집의 초기치를 잘못 설정했을 경우 발생할 수 있는 부정확한 군집화 결과를 방지할 수 있다. 또한 제안하는 적합도 함수는 보다 적합한 군집의 중심을 찾아내는 것으로 평가되었다.

앞으로의 연구 계획은 알고리즘을 좀 더 최적화하여 데이터 마이닝에 적용해 보고자한다. 특히 본 논문의 적합도 함수 정의에서 사용한 군집의 대표값 개념은, 요약 정보만을 이용하기 때문에 계산속도의 향상을 보인다. 따라서 대용량의 실제 데이터를 많이 사용하는 데이터 마이닝에의 적용효과가 기대된다.

6. 참고 문헌

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

Logic", Prentice-Hall Inc. 1995.

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

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

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

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

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

[10] 안병주, 김은주, 이일병, " 데이터 마이닝을 위한 계층적 대표값 군집화 기법" , 정보과학회 학술발표논문집, 제27권, 2000.