데이터 마이닝의 능률적인 군집화를 위한 유전자 알고리즘 적용에 관한 연구

최호진*, 홍성표**
*나주대학 정보전산센타
**조선대학교 전자정보공과대학
e-mail : hjchoi@naju.ac.kr
hongsp@chosun.ac.kr

A Study on Gene Algorithm Application for Efficient Clustring of Data Mining

Ho-Jin Choi*, Sung-Pye Hong**
*Dept of Information Computer Center, Naju College
**Dept of Computer Engineering, Chosun University

요 약

데이터 마이닝의 대표적인 기법인 군집화는 군집내의 유사성을 최대화하고, 군집들간의 유사성을 최소화 시키도록 데이터의 집합을 분할하는 것이다. 대용량의 데이터베이스에서 최적의 효율화를 내기 위해서는 원시데이터에 대한 접근 횟수를 줄이고, 이것을 알고리즘 적용 대상이 데이터 구조의 크기를 줄이는 군집화 기법에 많은 관심이 보이고 있다.

본 논문에서는 유전자 알고리즘을 이용하여 자동으로 군집의 개수를 결정하는 군집화 알고리즘을 제안하는 적합도 함수는 보다 양질의 군집을 찾아내는 것으로 평가 되었다. 또한 유전자 알고리즘 중 8가지를 세부 분석하여 평가하였다.

키워드 : Data Mining, Gene Algorithm, Clustering

I. 서론

현재 관심의 대상이 되고 있는 데이터 마이닝은 대용량의 실재 데이터로부터, 미리 알려지지 않았지만 잠재적으로 유용한, 암시적인 정보를 발굴하는 작업이라는 의미이다. 정보를 발굴하는 작업이라는 말은 수 많은 데이터로부터 의미 있는 정보 패턴을 분류해 내어 이것을 지식의 형태로 추출 하는 것이다. 여기서 데이터를 분류해내는 것은 군집화를 이루어 내는 것과도 같다.

데이터 마이닝의 대표적인 기법인 군집화는 군집내의 유사성을 최대화하고, 군집들간의 유사성을 최소화 시키도록 데이터의 집합을 분할하는 것이다.

대용량의 데이터베이스에서 최적의 효율화를 내기 위해서는 원시데이터에 대한 접근 횟수를 줄이고, 이것을 알고리즘 적용 대상이 데이터 구조의 크기를 줄이는 군집화 기법에 많은 관심이 보이고 있다.

본 연구에서는 군집화에 유전자 알고리즘을 적용한다. 특히 유전자 알고리즘 성능에 영향이 많은 여러가지 함수를 적용하여 이에 유전자 알고리즘내에서 최적의 군집화를 이끌어 내는 함수를 찾아내는 연구에 중점을 둔다.

II. 관련 연구

유전자 알고리즘은 자연계의 진화현상으로부터 유도된 기계학습의 한 모델이다. "자연은 자연선택과 적자생존을 통해 적용된 유기체를 골라낸다"는 다윈의 자연선택의 원리를 모방하였다.

유전자 알고리즘의 수행과정은 첫째, 다양한 정보를 다루기 쉽게 부호화 할 수 있도록 일정 길이의 이진 문자열로 개체를 만들고 초기값으로 설정할 세대를 생성한다. 둘째, 가능한 가설 집합인 모집단에서 모든 개체들의 적합도를 계산한다. 셋째, 개체에서 교배 적합도, 비례 재생성, 돌연 변이 등

과 같은 유전적 조작들을 수행함으로써 새로운 모집단을 생성한다. 마지막 단계로, 과거의 모집단은 무시하고 새로운 모집단을 사용하여 위의 과정을 반복한다. 유전자 알고리즘은 적합한 가설을 얻어내는 방법으로서 모델링 하기 힘든 복잡한 문제에도 적용이 가능하고 병렬화가 가능하다[1][2].

본 연구에서는 군집화에 유전자 알고리즘을 적용하며, 특히 유전자 알고리즘의 성능을 좌우한다고 할 수 있는 적합도 함수에 관한 연구에 중점을 두고 있다. 적합도 함수란, 임의의 개체가 문제의 해에 얼마나 적합한지를 나타내는 척도이다. 따라서 문제의 해가 될 가능성이 있는 것들을 평가하는 환경의 역할을 수행 하는 것으로, 일종의 목적함수라고 할 수 있다. 적합도 함수의 중요성은 다양한 군집화 평가 함수로 이용되면서 부각 되었다. 응집성과 분리성을 이용한 군집화의 개념에 이용되면서 이 두 가지 평가함수를 이용한 목적함수가 나오면서 군집화의 여부를 평가하여 분리 할 수 있게 되었다.

현재 데이터 마이닝에 적용되는 유전자 알고리즘은 초기의 유전자 알고리즘, Mcssy Genetic Algorithm, Genetic Programming, Parallel Genetic Algorithm, GABIL, GA-IDE, GIGAR, LONGEPRO 등을 들 수 있다. 각기 다양한 특성을 가지고 있으며 현재 데이터 마이닝에 폭넓게 적용되고 있다.

III. 유전자 알고리즘 분류 기준과 분석

3.1 유전자 알고리즘 분류기준

본 연구에서 세부 알고리즘을 대상으로 분류하기 위하여 각 기준에 대한 척도를 제시하였다. 여기서 사용된 기준은 각 유전자 알고리즘별 여러 특징들을 도출한 것으로 각 알고리즘별 성능면, 사용자 측면, 특징, 그리고 기타 정보 등을 중심으로 기준 척도를 만들었다.

분할적 군집화는 사전에 군집의 개수를 결정해주어야 하며, 초기 군집의 중심 설정과 잡음에 따라 알고리즘의 성능이 민감하게 좌우되는 문제점이 있다. 그래서 최근 통계적 기법이나 유전자 알고리즘을 이용하여 자동으로 군집의 개수를 결정해주고자 하는 연구가 이루어지고 있다[4][5][6][7]. 또한 유전자 알고리즘을 이용하여 지역적 최적해(local minima)에 수렴될 수 있는 문제점을 해결하기 위한 연구도 진행되고 있다.

본 논문에서 적용하고 있는 유전자 알고리즘은 다양한 NP-Complete조합 최적화 문제를 해결하는데 매우 유용한 것으로 알려져 있다[4][5]. 자연도태와 진화의 원리에 기반을 둔 확률적인 탐색 알고리즘이다. 특히 전역적 탐색 및 최적화, 기계학습의 도구로 많이 사용되고 있다[3].

군집에서 두 개의 대표값을 가지고 군집의 내부적 특징인 응집거리와 군집간의 외부적 거리를 나타내는 근접거리를 계산한다. 이를 이용하여 군집간의 연관성과 특징을 고려한 유사도(similarity)[8]값을 측정하고 적합도 함수로 이용한다. 다음의 식에 유사도 개념이 정리되어 있다.

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

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

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

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

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

ra, rb : 대표값 벡터
ni : 소군집 i에 속하는 대표값의 개수
w_{ra} : 대표값 ra가 대표하는 원시데이터 개체 수

• 선택(selection)
룰렛휠(roulette wheel)방법과 엘리트 방법(elitist model)을 함께 사용한다[3].

• 교배(Crossover)
유전인자들이 고정길이가 아닌 가변길이이며 위치에 상관없게 정의되었다.

• 돌연변이(Mutation)
정규적 돌연변이 연산자와 가우시안 함수를 사용하여 선택된 특정 유전인자의 값을 바꿔준다.

3.2 유전자 알고리즘 분석

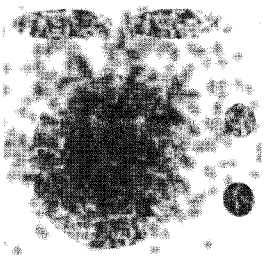
유전자 알고리즘의 입력 데이터 타입은 기본적으로 비트 문자열에 기반하고 있다. 규모 확장성은 작으나 병렬화가 가능해지면서 VL(very large)의 개념이 가능해졌다. 또한, 이 알고리즘은 최적해가 결과값으로 산출됨으로 결과 설명이 가능하고 일반적으로 유전자 알고리즘은 별도의 훈련시간이 필요치 않으나 일부 알고리즘에서는 훈련시간이 필요하다는 것을 알 수 있었다[9].

사용되는 교배 확률, 돌연변이 확률, 모집단의 크기, 그 외 선택, 암호화, 교배, 돌연변이형 등의 여러 가지가 있으나 사용이 용이하지가 않다. 이 알고리즘은 병렬 분석이 가능하고 분류, 예측, 최적화의 영역에 적용 되어 질 수 있다.

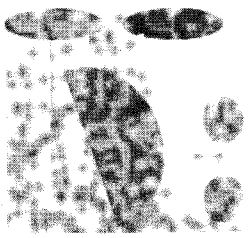
IV. 구현 및 고찰

본 시스템은 Windows 2000 Server 환경에서 C#, MSSQL2000를 이용하여 현재 학생 진로지도 시스템에 응용되어 구현되었으며 기존의 알고리즘은 국소 값에 수렴하여 적절한 중심을 찾지 못하는 반면, 제안한 알고리즘은 자동적으로 군집의 개수를 찾아낼 뿐만 아니라 비교적 정확한 군집을 현성하고 있다.

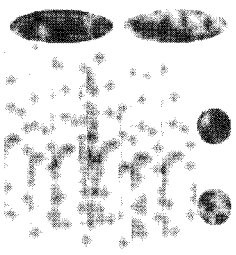
2000개의 데이터와 5개의 군집이 존재하는 이차원 공간 데이터로 실험한 결과, 제안한 알고리즘은 서로 다른 크기의 잡음이 섞여 있는 데이터 집합에서 양질의 군집을 찾아냄을 알 수 있다.



Data Type(1)



Data Type(2)



Data Type(3)

〈그림 1〉 Dimensions Space Data

〈표1〉 Performance of grouping techniques for set of space data(i)

	Hierarchic grouping		
	Shortest	Longest	Average
Data Type(1)	8.063	6.012	5.673
Data Type(2)	9.700	8.042	5.854
Data Type(3)	2.495	2.620	2.577

〈표2〉 Performance of grouping techniques for set of UCI data(Q)

	Hierarchic grouping		
	Shortest	Longest	Average
Australian	1.538	1.334	1.537
Diabetes	0.604	0.637	0.605
Heart	1.675	1.653	1.376
Iris	0.854	0.892	0.915
Soybean	3.092	3.092	3.092
Wine	0.987	0.712	0.985
Zoo	1.356	1.372	1.303

실제 데이터인 UCI Machine Learning Repository의 데이터로 본 알고리즘의 유효성을 검증해 본 결과, 다음과 같이 비교적 좋은 군집화 성능을 보였다. 군집의 성능을 평가하기 위해 사용된 성능 평가함수는 다음의 식의 i로 측정하였는데 D_i는 군집 i에 속하는 모든 데이터 체계들 간의 평균거리이고 Q는 D_i의 평균이다.

$$i=\sum_{i=1}^{k}\frac{D_i}{k}\,D_i=\frac{\sum_{a=1}^{ni}\sum_{b=1}^{ni}\sqrt{(x_a-x_b)^2}}{n_i(n_i-1)}$$

각 알고리즘별로 최적의 알고리즘을 선정하였으며, 그 근거는 각기 다르다. Data Type(1), 즉 데이타 마이닝 지향 알고리즘인 연관규칙과 클러스터링은 대형 데이타베이스를 목표로 하여 끊임없는 성능 개선을 위해 개발되고 있으므로 알고리즘별 특성과 그 성능의 차이를 분석할수 있는 근거들이 존재한다. 그러나 Data Type(2)의 경우에는 기계 학습 알고리즘에 그 근원을 두고 있어 성능면에서의 비교 분석은 불가능하다. 따라서 Data Type(2)의 알고리즘들은 각 알고리즘들의 특성 비교측면에서만 언급하였다.

유전자 알고리즘의 장점으로는 최종세대의 결과값이 곧 최적해를 나타냄으로 결과에 대한 별도의 해석이 필요 없다는 점과 결과값을 적용하기가 쉽다는 것, 그리고 다양한 데이타 형태를 적용할 수 있으며 넓은 영역의 데이타를 핸들링 할 수 있으며, 많은 최적화 문제의 응용에 적용되며 신경망과 잘

통합 될 수 있다.

단점으로는 많은 문제들을 일정한 길이의 유전자로 인코딩하기가 어렵다는 점과 최적화에 대한 보장을 할 수 없고, 계산비용이 높으며 현재까지 적은 수의 부분에서 유용된다는 것이다. 몇 가지 문제점만 해결되면 앞으로 유전자 알고리즘을 향후 데이터 마이닝 분야에서 폭 넓게 응용될 수 있을 것으로 본다.

V. 결론

본 논문에서는 유전자 알고리즘을 이용하여 자동으로 군집의 개수를 결정하는 군집화 알고리즘을 제안하는 적합도 함수는 보다 양질의 군집을 찾아내는 것으로 평가 되었다. 또한 유전자 알고리즘 중 8가지를 세부 분석하여 평가하였다.

각 알고리즘의 실험 환경과 마이닝 목적이 상이한 관계로 알고리즘들이 서로 똑같은 기준과 환경에서 평가 한다는 것은 불가능한 것이다. 다만 앞으로 구축하게 될 시스템내의 최적화 알고리즘을 찾아내어 구현하는 것이 목적이므로 본 연구결과는 데이터 마이닝에서 유전자 알고리즘을 도구로 효과적으로 개발할 수 있는 기반을 제공하고 데이터 마이닝에 맞는 적절한 알고리즘을 선택 할 수 있는 지표를 제시하는데 목적이 있다.

향후 연구 계획은 학생 진로 시스템에서 알고리즘을 최적화하여 성능 및 데이터 처리에 최적의 시스템 구현을 목표로 한다.

참고문헌

[1] Michael J. A Berry, and Gorden Linoff, Data Mining Techniques : For Marketing, Sales, and Customer Support, John Wiley & Sons, Inc., (1997)

[2] Rakesh Agrawal and John C. Shafer, "Parallel Mining of Association Pules," IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6, pp. 962-969, December

[3] Alexander Hinneburg, and Daniel A.Keim, "An Efficient Approach tho Clustering in Large Multimedia Databases with Noise," In proc. of 4th International Conference of Knowledge Discovery and Data Mining, NewYork, pp. 58-65, (1998)

[4] Mihael Ankerst, Markus M. Breuning, Hans-Peter Kriegel, and Jorg Sander, "OPTICS : Ordering Points To Identify the Clustering Structure,"In proc. of ACM SIGMOD International Conference on Management of Data, Philadephia, Pennsylvania, USA, pp. 49-60, June (1999).

[5] J.Bala, J.Huang, H.Vafaie, K.DeJong and H.Wechsler, "Hyvrid Learning Using Genetic Algorithms and Decision Tree for Pattern Classification," In Proc. of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI95), Volume I pp.719-724, August(1995).

[6] James D Kelly, and lawrence Davis," hybridizing the Genetic Algorithms and the K Nearest Neighbors Classification Algorithms," In Proc. of the 4th International Conference on Genetic Algorithms and their Applications Morgan Kaufmann Publishers, (1991).

[7] S.Anadn, D.Patterson, J.G.hughes, and D.A.Bell, Discovering Case Knowledge using Data Mining, Northern Ireland knowledge engineering Laboratoy, School of Information and software Engineering, University of Ulster. (1998).

[8] Xiaowei Xu, Martin Ester, Hans-Peter Kriegel, and Jorg Sander, " A Distribution-Based Clustering Algorithm for Mining in Large Spatial Databases," In proc. of 14th International Conferenceon Data Engineering(ICDE), Orlando, Florida, USA, pp. 324-331, february (1998).

[9] Goldberg David.E, korb Bradley, and Deb K. "Messy Genetic Algorithms : Motivation, Analysis and Results," TCGA Report 90005, May (1995).