제26회 한국정보처리학회 추계학술발표대회 논문집 제13권 제2호 (2006. 11)

유전자알고리즘의 혼합 초기화법을 이용한 eCRM을 위한 데이터마이닝

강래구*, 임희경*, 정채영*
*조선대학교 전산통계학과
e-mail:kangrg@hanmail.net

Date Mining for eCRM using Mixture Initialization of Genetic Algorithm

Rae-Goo Kang*, Hee-Kyoung Lim*, Chai-Yeoung Jung*
*Dept of Computer Science & Statistics, Chosun University

요 약

고객관리가 기업의 성패를 좌우하는 중요한 화두로 떠오르면서 보다 쉽고 편리하게 고객의 다양한 Pattern을 발견하고 예측하기 위해 많은 기업들이 CRM과 eCRM을 빠르게 도입하고 있고 Data Mining 기법이 대표적으로 이용되고 있다. 본 논문에서는 Data Mining을 적용함에 있어서 Genetic Algorithm의 무작위 초기화법과 유도된 초기화법을 동시에 사용하는 새로운 집단 초기화 방법을 적용하여 A할인점의 2004년과 2005년도 우수고객을 예측하였고 실제 고객 데이터와의 비교를 통해 본 논문에서 제안한 새로운 집단 초기화 방법의 성능을 입증하였다.

1. 서론

70년대 이전까지 기업의 마케팅은 시장 전체 불특정 다수를 겨냥하여 대량으로 상품과 서비스를 제공하는 Mass Marketing이 중심이었다.

2000년대에 들어서면서 기업 마케팅 기법의 화두로 Personal Marketing 이 대두되고 있다.

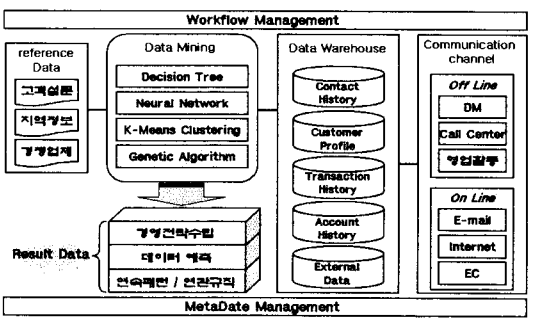
이러한 Personal Marketing의 대표적인 방법이 CRM(Customer Relationship Management)이다.[1][2] 또한, IT산업의 비약적인 발달로 인해 최근엔 CRM을 기반으로 한 eCRM(Electronic Customer Relationship Management)이 새롭게 등장하였다.

eCRM의 기본적인 방법론이나 사상은 CRM과 크게 다르진 않지만 고객정보수집 및 Communication이 기존 CRM과 달리 첨단 IT 기술과 Internet 중심으로 이루어져 있다는게 특징이다.[3][4][5]

본 논문에서는 eCRM과 CRM을 eCRM으로 통일하여 표현하였고 Genetic Algorithm을 이용하여 Data Mining을 진행하였고 집단 초기화 방법으로 무작위 초기화법과 유도된 초기화법을 동시에 사용하는 혼합 초기화법을 제안하였다. 또한 Genetic Algorithm의 혼합 초기화법을 이용한 Data Mining을 통해 A할인점의 우수 고객을 예측하였고 예측한 데이터와 실제 데이터와의 비교를 통해 eCRM에 얼마나 효과적인지 실험을 통해 입증하고자 한다.

2. eCRM을 위한 Data Mining

eCRM은 고객과의 관계를 효과적으로 관리하는 마케팅 기법을 의미하며 기업과 고객간의 상호교류를 관리하는 일종의 Process이다.



(그림 1) eCRM Architecture

그림 1은 eCRM의 전체적인 구성도를 나타내고 있다. 기업의 생존과 지속적인 성장을 위해 Process를 보다 효과적으로 관리하여 통합된 고객중심 마케팅 전략으로 발전시켜 나가기 위해서는 무엇보다도 Process의 자동화가 필요하며 그 중심에 Data Mining이 있다.

갈수록 다양해지는 시장여건에 효과적으로 대응하기 위해 기업의 On/Off Line 채널에서 획득한 대량의 자료로부터 Data Mining을 이용하여 새로운 정보를 발견하고 예측하여 신규고객확보, 기존고객의 유지 및 이탈방지 등에 적극적으로 이용할 수 있게 되었고 보다 공격적인 마케팅 수단으로 활용할 수 있게 되었다.[6] 이러한 Data Mining이 대표적으로 이용되고 있는 분야가 바로 eCRM이다.

2.1 Genetic Algorithm

최근에 Data Mining은 통계와 같은 기존의 고전적 기법에서 Neural Network, Decision Tree, Genetic Algorithm과 같은 차세대 기법으로 점점 대체되고 있다.

이러한 차세대 기법 중 본 논문에서 사용한 Genetic Algorithm은 1975년 John Holland에 의해서 처음 소개되었다. 자연계에서 적자생존의 원리에 따라 세대가 지나면서 우량의 형질을 지닌 객체가 생성되는 과정을 모방한 알고리즘으로 정보의 탐색이나 예측과 같은 모델을 구현하는데 적절한 Algorithm이다. 집단이라는 탐색공간이 초기화되고 집단의 각 객체는 적합도 함수에 의해 평가된다.

이 과정에서 Selection, Crossover, Mutation 연산을 통해 적합도가 우수한 해를 탐색하게 되고 세대가 되풀이되면서 가장 우수한 해를 찾아내게 된다.[7] 본 논문에서 사용한 GA연산자는 표 1과 같다

<표1>. 실험에 사용한 연산자

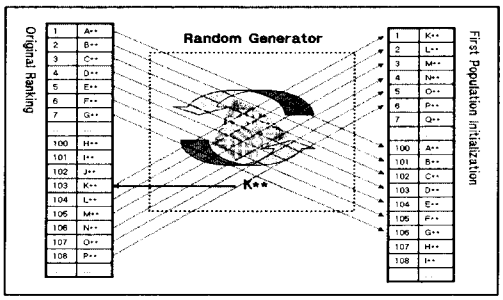
선택(Selection)	Roulette wheel
교배(Crossover)	Edge Recombination
돌연변이(Mutation)	Inversion

3. 제안한 방법

본 논문에서는 Genetic Algorithm을 적용하기 위한 집단 초기화 방법으로 혼합 초기화법을 제안하였다.

3.1 혼합 초기화법

Genetic Algorithm을 이용한 Data Mining을 위해서는 무엇보다도 최초에 생성되는 집단 초기화방법이 중요하다. 집단 초기화방법에는 크게 두 가지 방법이 있다. 하나는 집단 생성에 있어서 특별한 규칙 없이 무작위로 추출해서 집단을 생성하는 무작위 초기화법이 있고 다른 하나는 주어진 값을 통해 어떠한 사전 지식이나 정보를 바탕으로 일관성을 유지하여 집단을 생성하는 유도된 초기화법이 있다. 지금까지 주로 사용해온 Genetic Algorithm의 집단 초기화법으로는 무작위 초기화법이 많이 사용되어 왔다. 본 논문에서는 난수발생기를 이용한 무작위 초기화법과 사전 정보나 경험을 바탕으로 한 유도된 초기화법을 동시에 사용하는 혼합 초기화법을 제안한다.



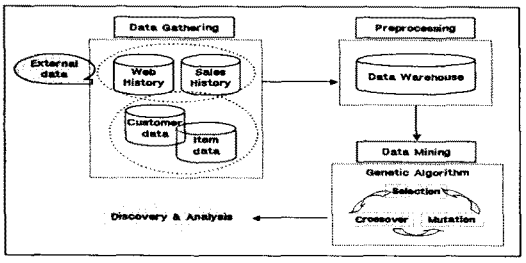
(그림 2) 혼합 초기화법 처리 구성도

그림 2는 본 논문에서 제안한 혼합 초기화법 처리 과정을 나타내는 그림이다. 혼합 초기화법이란 먼저 난수발생기를 이용하여 시작 고객을 랜덤하게 선택한 다음 고객들의 작년도 매출액 정보를 참조하여 선택된 고객보다 매출액이 낮은 고객들을 대상으로 내림차순으로 정렬하고 그 뒤로 선택된 고객보다 매출액이 높은 고객들을 내림차순으로 정렬한다. 이러한 순서로 N개의 모든 고객을 나열하여 최초 집단으로 사용하였다.

4. 시스템 설계

본 논문에서는 A할인점의 2003년과 2004년도 1월부터 12월 까지 On/Off Line 상의 매출 데이터와 고객 데이터를 년도 별로 분리하여 각각 실험을 하였다. 해당 년도 별로 구분한 공산품 관련 매출 데이터와 고객 데이터를 본 논문에서 제안한 혼합 초기화법을 이용한 Data Mining을 적용하여 2004년과 2005년도 우수고객을 각각 순위와 관계없이 50명씩을 예측함과 동시에 각 년도 별 상반기 실제 고객 데이터 중 고객 기여도(매출순위)가 높은 상위 50명

과의 비교를 통해 본 논문에서 제안한 방법이 eCRM을 위해 얼마나 정확한 결과를 예측해 내는지 증명하고자한다. 실험에 사용할 고객 데이터는 각 년도 별 매출 기간 동안 적어도 한 달에 두 번 이상 매장을 방문하였거나 쇼핑몰에 Log-In 한 내역이 있는 고객만을 자동으로 필터링하여 실험하였고 집단의 크기는 500으로 한정하였으며 총 세대수는 600으로 하였다. 그리고 교배 확률(P_c)과 돌연변이 확률(P_m)은 0.7 과 0.2를 사용하였고 각 세대마다 엘리트 전략을 이용하여 우수 개체 2개씩을 보존하였다. 이러한 작업을 거쳐 마지막 세대에 생성된 집단 중 고객기여도가 가장 높은 개체의 상위 50명을 예측 고객으로 선정하였다. 실험에 사용할 Genetic Algorithm의 혼합 초기화법의 구현은 Oracle 기반의 SQL문으로 직접 작성하여 Data Mining 실험을 하였다.

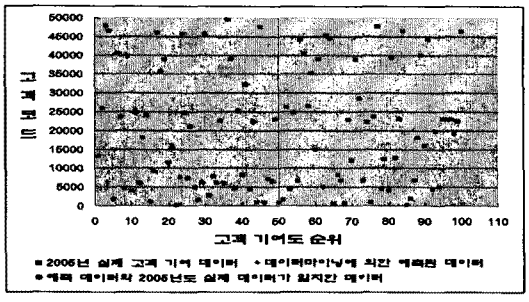


(그림 3) 시스템 구성

그림 3는 Data Mining을 위한 전체적인 시스템 구성을 나타내고 있다.

5. 실험 결과

실험은 P-4 2.4GHz에서 Oracle 9i를 기반으로 Data Mining 작업을 진행하였으며 알고리즘 구현은 SQL문을 이용하여 Procedure와 Trigger를 작성하여 실행하였다.



(그림 4) 2005년 실제 데이터와 예측된 데이터 비교

그림 4은 2005년도 상반기 실제 고객 기여도 순위와 2004년 1월부터 12월까지 고객과 매출 데이터를 본 논문에서 제안한 혼합 초기화법을 이용한 데이터 마이닝으로 예측한 우수고객 데이터와의 비교를 그래프로 나타내고 있다.

2005년도 실제 고객 기여도 데이터는 파랑색 사각형으로 표시하였고 혼합 초기화 법을 통해 예측한 데이터는 빨강색 마름모로 표시하였다.

그림 4에서 알 수 있듯이 논문에서 제안한 Genetic Algorithm의 혼합 초기화법을 이용한 Data Mining을 통해 예측한 고객 50명 중 43명이 2005년도 상반기 실제 고객 기여도 상위 50명 안에 일치함으로써 86%의 예측률을 보였고 50명 안에 포함되지 않은 나머지 7명도 상위 78명 안에 모두 포함되었다. 반면에 무작위 초기화법을 이용한 Data Mining은 80%의 예측률을 보였다.

또한 제안한 방법을 통해 2004년도 실제 고객 기여도 데이터 상위 50명 중 41명을 정확히 예측하여 82%의 예측률을 보였고 예측하지 못한 나머지 9명 중 8명도 상위 80명 안에 포함되는 걸로 예측하였다. 반면에 무작위 초기화법을 이용한 Data Mining은 74%의 예측률을 보였다.

표 2와 표 3은 무작위 초기화법과 혼합 초기화법을 이용하여 2004년, 2005년도 우수고객을 예측한 결과를 표로 나타내어 비교하고 있다.

두 번의 실험에서 2004년과 2005년도 실제 고객 기여도 상위 50명 중 제안한 방법을 통해 예측하지 못한 9명과 7명은 본 실험 조건에 만족하지 않은 고객이 다수 포함 되어 있는 것으로 확인되었다.

실험을 위한 고객 데이터 추출 조건은 2003년과 2004년도에 정상적으로 매출이 일어난 고객 데이터에 한하여 실험하였으나 예측하지 못한 고객 중에는 새로 등록한 고객과 해당년도에 매출이 없었던 고객이 다수 포함되어있었다.

이러한 조건을 감안 한다면 본 논문에서 제안한 Genetic Algorithm의 혼합 초기화법을 이용한 Data Mining 기법에서 보여준 82%와 86%의 예측률은 결코 낮은 수치라 할 수 없을 것이다.

6. 결론

많은 기업들이 성공적인 마케팅을 위해 고객정보의 체계적인 분석과 다양한 Pattern을 발견하고 분석 및 예측을 하기위해 고객관계관리로 불리우는 CRM과 eCRM을 빠르게 도입하고 있다.

	무작위 초기화법을 이용한 2004년도 데이터		혼합 초기화법을 이용한 2004년도 데이터	
실제 고객 순위	상위 1~50위	상위 1~80위	상위 1~50위	상위 1~80위
예측한 고객 50명	37명 일치	44명 일치	41명 일치	49명 일치
예 측 률 (%)	74%	88%	82%	98%
실험 조건에 부적합한 고객 데이터	신규로 등록된 고객 : 4명 매출이 없었던 고객 : 0명		신규로 등록된 고객 : 4명 매출이 없었던 고객 : 0명	

<표 2> 2004년도 예측률 결과를 통한 성능 비교

	무작위 초기화법을 이용한 2005년도 데이터		혼합 초기화법을 이용한 2005년도 데이터	
실제 고객 순위	상위 1~50위	상위 1~80위	상위 1~50위	상위 1~80위
예측한 고객 50명	40명 일치	43명 일치	43명 일치	50명 일치
예 측 률 (%)	80%	86%	86%	100%
실험 조건에 부적합한 고객 데이터	신규로 등록된 고객 : 2명 매출이 없었던 고객 : 1명		신규로 등록된 고객 : 2명 매출이 없었던 고객 : 1명	

<표 3> 2005년도 예측률 결과를 통한 성능 비교

과거엔 고객관리가 통계학자들이나 전문적인 통계 패키지에 의해 관리되어 왔으나 2000년 이후 IT 분야의 급격한 발달을 기반으로 통계적 과정을 자동화하여 통계전문가가 아닌 일반인들도 쉽게 양질의 데이터를 추출하고 예측 할 수 있는 Data Mining으로 점점 대체되고 있는 추세이다.

이러한 Mining 기법의 발달로 인해 SAS나 SPSS, MINITAB과 같은 전문적인 통계 패키지를 습득한다거나 통계전문가가 아니더라도 Data Mining을 통해 비전문가 누구나 원하는 정보에 대한 분석과 예측을 보다 쉽고 편리하게 할 수 있을 것이다.

본 논문에서는 효과적인 Data Mining을 위해 Genetic Algorithm의 혼합 초기화법을 제안하였다. 그리고 제안한 방법을 적용하여 A사의 고객 데이터와 매출 데이터를 기반으로 2004년과 2005년도 우수고객을 예측 해내는 실험을 하였다.

실험 결과 년도 별로 예측한 고객 50명 중 41명과 43명이 2004년과 2005년도 실제 고객 기여도 데이터 상위 50명 안에 포함되는 결과를 나타냄으로써 본 논문에서 제안한 혼합 초기화법을 적용한 Data Mining 기법을 eCRM에 적용하였을 때 성능 향상을 입증하였다. 또한, 본 논문에서는 Data Mining을 위한 Genetic Algorithm을 SQL문으로 직접 작성하여 구현함으로써 추후 본 논문에서 구현한 Algorithm을 이용하여 eCRM System 개발도 가능할 것이다.

이와 같이 효과적인 고객 관리를 위해 eCRM에 Data Mining을 이용함으로써 보다 공격적인 고객 마케팅 수단으로 활용할 수 있을 것이다.

참고문헌

[1] Berson.Alex, Building Data Mining Applications for Crm, McGraw-Hill. 1999

[2] Yim CK, Kannan PK., "Consumer behavioral loyalty: a segmentation model and analysis" Journal of Business Research. Vol.44(2), 1999

[3] Kohli R, Piontek F, "managing customer relationships through e-business decision support applications: a case of hospital-physicion collaboration" Decision Support System, Vol.32(2)

[4] 사와노보리 히데아키, "e-CRM 마케팅" 국일증권경제연구소, 2000

[5] Fayyad, U. M, "Advances in Knowledge Discovery and Data Mining", MIT Press, 1996

[6] 박주석, "성공적인 CRM 구축에 영향을 미치는 요인에 관한 연구", 경영과 컴퓨터, pp.262-265, 2000

[7] Hon, K. K. B, and H. Chi, "A New Approach of Group Technology Part Families Optimization", Annals of the CIRP, 1994