유사도와 유클리디안 계산패턴을 이용한 CBR 패턴연구

윤종찬* · 김학철** · 김종진*** · 윤성대****

A Study on the CBR Pattern using Similarity and the Euclidean Calculation Pattern

Jong-Chan Yun* · Hak-Chul Kim** · Jong-Jin Kim*** · Sung-Dae Youn****

요 약

사례기반추론(CBR:Case-Based Reasoning)은 기존 데이터와 사례 데이터들의 관계성을 추론하는 기법으로 유사도(Similarity)와 유클리디안(Euclidean) 거리 계산 방법이 가장 많이 사용되고 있다. 그러나 이 방법들은 기존 데이터와 사례 데이터를 모두 비교하기 때문에 데이터 검색과 필터링에 많은 시간이 소요되는 단점이 있다. 따라서 이를 해결하기 위한 다양한 연구들이 진행되고 있다. 본 논문에서는 기존의 유사도와 유클리디안 계산과정에서 발견된 패턴을 활용한 SE(Speed Euclidean-distance) 계산방법을 제안한다. SE 계산방법은 새로운 사례입력에 발견된 패턴과 가중치를 적용하여 빠른 데이터 추출과 수행시간 단축으로 시간적 · 공간적 제약사항에 대한 연산 속도를 향상시키고 불필요한 연산 수행을 배제하는 것이다. 실험을 통해 유사도나 유클리디안 방법으로 데이터를 추출하는 기존의 방법보다 제안하는 방법이 다양한 컴퓨터 환경과 처리 속도에서 성능이 향상됨을 확인할 수 있었다.

ABSTRACT

CBR (Case-Based Reasoning) is a technique to infer the relationships between existing data and case data, and the method to calculate similarity and Euclidean distance is mostly frequently being used. However, since those methods compare all the existing and case data, it also has a demerit that it takes much time for data search and filtering. Therefore, to solve this problem, various researches have been conducted. This paper suggests the method of SE(Speed Euclidean-distance) calculation that utilizes the patterns discovered in the existing process of computing similarity and Euclidean distance. Because SE calculation applies the patterns and weight found during inputting new cases and enables fast data extraction and short operation time, it can enhance computing speed for temporal or spatial restrictions and eliminate unnecessary computing operation. Through this experiment, it has been found that the proposed method improves performance in various computer environments or processing rate more efficiently than the existing method that extracts data using similarity or Euclidean method does.

키워드

사례기반추론, 유사도, 유클리디안, 데이터마이닝

Key word

Case-Based Reasoning, Similarity, Euclidean, Data Mining

* 부경대학교 병렬운영체제 및 데이터마이닝연구실
** 부경대학교 전자공학과 박사과정
*** 부경대학교 전자공학과 교수
**** 부경대학교 컴퓨터공학과 교수(교신저자)

접수일자 : 2009. 10. 13
심사완료일자 : 2010. 01. 12

Ⅰ. 서 론

사례기반추론(CBR:Case-Based Reasoning)은 기존 데이터와 사례 데이터의 관계성을 추론하는 기법으로 유사도(Similarity)와 유클리디안(Euclidean) 거리 계산 방법이 가장 많이 사용되고 있다[1,2]. 이 방법들은 정보 시스템의 활용의 증가로 발생한 수많은 데이터들로부터 유용한 데이터를 추출하고 데이터를 사례기반추론에 적용하기 위해 사용된다[3-9]. 그러나 이 방법들은 기존 데이터와 사례 데이터를 모두 비교하기 때문에 데이터 검색과 필터링에 많은 시간이 소요되는 단점이 있다. 따라서 이를 해결하기 위한 다양한 연구들이 진행되고 있다.

본 논문은 많은 사례 데이터와 기존 데이터를 비교 및 검증하는 데 좀 더 빠르고 정확한 결과를 도출하기 위해 SE계산방법을 제안하고자 한다. SE 계산방법은 새로운 사례입력에 발견된 패턴과 가중치를 적용하고 발생가능성이 있는 사례 데이터에 대한 연산만 수행하여, 빠른 데이터 추출과 수행시간 단축으로 시간적·공간적 제약사항에 대한 연산 속도를 향상시키고 불필요한 연산 수행을 배제하는 것이다.

실험데이터는 랜덤으로 사례데이터를 100개 생성하여 전산 모의실험을 실시하였다. 사용한 데이터의 정확도는 기존의 계산방식인 유클리디안 계산방법으로 확인하였다. 전산 모의실험은 컴퓨터 환경에 따른 데이터 처리 속도변화를 평가하기 위해서 제안한 SE 계산 알고리즘을 적용하는 방법이 기존 계산방법보다 연산 속도면에서 성능이 향상됨을 증명하기 위해 제안하는 알고리즘과 기존의 알고리즘을 비교·분석하고자한다.

본 논문의 구성은 2장에서는 기존 관련연구에 대해 기술하고, 3장에서는 제안하는 알고리즘에 대해 기술하고, 4장에서는 각 알고리즘에 따른 전산모의실험과 평가결과를 마지막으로 5장에서 결론을 맺는다.

Ⅱ. 관련연구

2.1 유사도(Similarity) 계산

현재의 새로운 사례와 유사한 사례를 검색하기 위해서, 유사도계산은 새로운 사례와 사례 베이스에 있는 과거의 모든 사례와의 유사도를 측정함으로써 유사 사례를 찾는 방법이다. 일반적으로 가장 많이 사용되는 방법은 새로운 사례와 가장 유사한 k개의 과거 사례를 검색해 주는 k-NN(k Nearest Neighbors)방법이다. 일반적으로 유사도 측정의 식은 다음 식 (1)과 같다[6].

$$\text{Similarity(N, C)} = \frac{\sum_{i=1}^{n} f(N_i, C_i) \times W_i}{\sum_{i=1}^{n} W_i} \quad (1)$$

N : 새로운 사례.

C : 사례베이스에 저장된 과거 사례.

n ; 사례가 가지는 속성의 개수.

N_i : 새로운 사례의 i번째의 속성 값.

C_i : 과거 사례의 i번째 속성 값.

$f(N_i, C_i)$: N_i와 C_i 사이의 거리 측정 함수.

W_i ; i번째 속성에 대한 가중치.

사례간의 유사도는 일반적으로'0'에서'1'사이의 정규화 된 실수 값으로 표현되는데,'0'에 가까울수록 두 사례의 유사성이 낮다는 것을 의미하고,'1'에 가까울수록 유사성이 높다는 것을 의미한다.

유사도 계산은 사례기반 시스템이 새로운 문제의 해를 위해 과거의 문제를 적용할 경우, 어떤 방법으로 유사한 사례를 인식할 것인지가 문제이다. 따라서 유사 매트릭스 또는 여러 차원으로 유사도를 판단할 수 있는 방법이 있어야 한다.

다음은 유사도 계산방법의 한 예이다[5].

$$V_i = SIM_i(\text{고객특성}) + SUU_i(\text{사례선정})$$

$$SIM_i(\text{고객특성}) = SIM_i(\text{나이}) + SIM_i(\text{학력}) + SIM_i(\text{분야}) + SIM_i(\text{직업})$$

$$\text{SUCCi(사례선정)} = \frac{T_i(\text{성공횟수})}{\text{성공횟수의최대값}} \times C$$

유사도 계산을 위해 추출한 고객의 특성과 선정된 사례의 특성은 나이, 학력, 분야, 직업으로 제한하였으며,

그 유사도는 다음과 같이 계산한다.

$$IF|S(나이) - T_i(나이)| \le 50$$

$$SIM_i(나이) = W_{나이} - \frac{|S(나이) - T_i(나이)|}{C_3}$$

$$IF|S(나이) - T_i(나이)| > 50$$

$$SIM_i(나이) = 0$$

$$SIM_i(학력) =$$

$$W_{학력} \times (1 - \frac{|S(학력) - T_i(학력)|}{C_2})$$

$$SIM_i(분야) = \begin{cases} W_{분야} & IF일치하면 \\ 0 & Otherwise \end{cases}$$

$$SIM_i(직업) = \begin{cases} W_{직업} & IF일치하면 \\ 0 & Otherwise \end{cases}$$

여기서, 학력의 도메인 지식은 편의상 대학은 1, 고등학교는 2, 중학교는 3으로 하였으며, 가중치 및 상수를 경험 값으로 다음과 같이 설정하였다.

$W_{학력} = 10$, $W_{나이} = 10$, $W_{분야} = 3$, $W_{직업} = 3$, $C_1 = 10$, $C_2 = 3$, $C_3 = 5$ 그리고 성공회수의 최대값=1,000이다.

다음 과정은 사례베이스를 검색하는 단계로, 고객의 등록정보로부터 고객의 정보를 검색하여, 검색된 고객의 개인 정보로부터 색인을 구성한다. 표 1은 검색하려는 고객의 예이고, 표 2는 사례베이스의 검색 예이다.

표 1. 검색 고객
Table 1. Search of customers

학력	나이	분야	직업
대졸	26	컴퓨터	학생

표 2. 사례베이스 검색
Table 2. Search of case database

키워드	학력	나이	분야	직업	서적 코드	성공 회수
VB	고졸	32	전기	?	B104	200
VB	고재	18	?	학생	B155	300
VB	대재	19	컴퓨터	학생	B123	800

표 3은 유사도 계산에 대한 결과를 나타낸 것이고, 표 4는 표 3의 결과를 가지고 선정된 사례를 나타낸 것이다.

표 4와 같이 유사도 계산에서는 새로운 데이터를 사례 데이터와 비교하여 유사도 값이 가장 큰 값을 가진 사례 데이터를 선정하여 사례 고객이 구입했던 서적 분야를 새로운 고객에게 추천하는 것이다.

표 3. 유사도 계산 결과
Table 3. Result of similarity computation

서적 코드	학력	나이	분야	직업	성공값	유사도
B104	6.67	8.80	0.00	0.00	2.00	17.47
B155	6.67	8.40	0.00	3.00	3.00	21.07
B123	10.0	8.60	3.00	3.00	8.00	28.60

표 4. 사례선정의 예
Table 4. Example of case decision

키워드	학력	나이	분야	직업	서적 코드	성공 회수
VB	대재	19	컴퓨터	학생	B123	800

본 논문에서 제안하는 방식은 기존 유사도 계산방법인 모든 변수들을 적용하여 기존 데이터와 사례 데이터를 분석하는 것이 아니라, 각 열에 있는 변수들을 비교하여 동일한 조건을 만족하는 변수들이 많은 사례를 선택하는 방법을 채택하고 있다. 이 계산방법은 사례변수들의 가중치들을 곱하여 나올 수 있는 최대의 값을 구하고, 그 값을 가지고 추출할 최소의 데이터를 검출하는 계산방법이다.

2.2 유클리디안 거리 계산(Euclidean distance)

일반적으로 유클리디안 거리 계산 개념은 다음과 같이 설명되고 있다. 우리가 일반적으로 생각할 수 있는 공간이 3차원 공간이기 때문에 여기서 n = 3이 된다. 따라서 3차원 공간상에서 x로 표현되는 한 점과 y로 표현되는 한 점의 기하학적 거리를 나타나게 된다.

관측 값에서 각 변수는 공간상에 있는 점을 나타내는 벡터의 한 성분이다. 두 점 사이의 거리가 연관성에 있는 측도로 사용된다. 두 점이 거리상 가깝다면, 대응하는 관측 값들은 유사하다고 여겨진다. 두 점간의 관계를 하나의 숫자로 표현하는 함수는 연관성의 측도로 사용될 수 있으나, 제대로 된 거리 측도가 되기 위해서는 다음을 만족해야한다[2].

Distance(X,Y) = 0 ↔ X = Y; 객체에서 자신으로의 거리는 0이다.
Distance(X,Y) ≥ 0; 거리는 음이 아닌 수이다.
Distance(X,Y) = Distance(Y,X); 거리는 대칭함수이다.
Distance(X,Y) ≤ Distance(X,Z) + Distance(Z,Y); 공간에서, X에서 Y로 직접 가는 것은 다른 어떤 객체 Z를 우회하여 가는 것보다 크지 않다(삼각부등식).

n개의 행 벡터

$x_i = \{x_1, x_2, \cdots, x_n\}$과 $y_i = \{y_1, y_2, \cdots, y_n\}$이 있다고 할 때, 다음과 같이 유클리디안 거리(Euclidean distance)에 대한 식은 다음 식 (2)와 같이 정의된다 [1,2].

$$d = \left[(x_i - y_i)^T (x_i - y_i)\right]^{1/2} = \left[\sum_{i=1}^{n} (x_i - y_i)^2\right]^{1/2}, \quad i = 1,2,\cdots,n$$

$$d(X, Y) = \sqrt{\sum_{i=1}^{D} (x_i - y_i)^2} \qquad (2)$$

다음은 유클리디안 거리의 계산방법의 한 예이다. 표 5는 기존 고객의 사례데이터베이스이고, 사례데이터베이스의 구성은 기존 고객들의 사례에 대한 데이터로 결과값에서 1은 구매를 0은 비구매를 나타낸다.

표 5. 사례 데이터베이스
Table 5. Databased of Cases

사례 번호	변수1	변수2	변수3	결과값
1	12	2	32	1
2	31	2	45	0
3	23	5	34	0
4	43	4	45	1
5	34	8	12	1

새로운 고객이 있을 때, 이를 검증하기위한 작업은 다음과 같다. 새로운 데이터의 사례(사례고객 6번)는 변수1의 값은 23, 변수2의 값은 3 그리고 변수3의 값은 33이다.

다음 단계에서 학습 조건 결정 k를 3으로 한다.

표 6은 학습 조건으로 결정된 k(3)를 가지고 기존 데이터와 사례 데이터를 학습 조건 결정을 하는 단계이다. 학습 조건 결정방법은 다음과 같다.

유사도 추출을 위해 유클리디안 공식에 사례 데이터와 새로운 데이터를 대입한다.

표 6. 학습 조건 결정 단계
Table 6. Step of deciding learning conditions

사례번호	변수1	변수2	변수3	결과값
1	12	2	32	1
2	31	2	45	0 k=3
3	23	5	34	0
4	43	4	45	1
5	34	8	12	1

d(1,6) = 11.09, d(2,6) = 14.46, d(3,6) = 2.24의 해답을 도출한다. 사례 3, 1, 2의 결과는 각각 0, 1, 0으로 유클리디안 거리값이 작게 나온 d(3,6)을 선택하므로 사례 6의 결과를 0으로 결정하여 새로 들어온 고객은 상품을 비구매할 경우가 높은 것으로 보는 것이 유클리디안 거리를 통해 구하는 방법이다.

기존의 유클리디안 알고리즘은 각 변수에 입력한 값에 각각의 값을 대입해 거리를 구하는 방법이다. 유클리디안 거리 계산방법은 유사도 알고리즘과 달리 학습조건이 있고, 계산된 결과가 가장 작은 값을 가진 데이터를 추출하는 것이다. 유클리디안은 학습조건 외에는 모든 데이터를 추출하기 때문에 데이터 추출하는 시간과 필터링이 늦어지는 단점이 있다.

본 논문에서는 유클리디안 계산패턴에서 발견된 각 열에 있는 변수들을 비교하여 동일한 조건을 만족하는 변수들이 많은 사례를 선택하는 방법을 채택하고 있다. 유사도 계산방법과 동일하게 적용된 계산방법은 변수들의 가중치들을 곱하여 나올 수 있는 최대의 값을 구하고, 그 값을 가지고 추출할 최소의 데이터들을 검출하는 계산방법을 사용한다.

Ⅲ. 제안하는 알고리즘

본 논문에서는 상품 추출 데이터베이스가 대용량화 됨에 따라 연산시간과 메모리 사용면에서 효율적인 데이터 처리가 필요함에 따라 기존의 사례기반추론기법 중에서 많이 알려져 있는 유사도와 유클리디안 거리의

계산에서 발생하는 패턴을 이용하여 상품추출 방법을 제안하게 되었다.

사례번호	변수1	변수2	변수3	성공횟수	유사도값	유클리디안 결과값
1	30	1	0	24	6.4	6.08
2	37	2	0	18	6.8	1.00
3	29	3	0	20	5.5	7.07
4	35	4	0	17	8.2	2.27
5	36	2	0			

그림 1. 유사도 값에서의 패턴분석
Fig 1. Pattern analysis of similarity values

그림 1은 사례 데이터와 새로운 데이터를 유사도 계산방법에서 사용되는 패턴분석으로 나타낸 것이다. 각 열에 있는 변수1, 변수2, 변수3을 비교하여 동일한 조건을 만족하는 변수들이 많은 사례를 선택하는 방법의 결과를 나타낸 것이다.

	항목	유클리디안 계산	결과	
k=3	d(1,5)	= √((36-30)²+(2-1)²+(0-0)²) = 6.08	2	1
	d(2,5)	= √((36-37)²+(2-2)²+(0-0)²) = 1.00	1	0
	d(3,5)	= √((36-29)²+(2-3)²+(0-0)²) = 7.07	3	0
	d(4,5)	= √((36-35)²+(2-4)²+(0-0)²) = 2.27		

그림 2. 유클리디안에서의 패턴분석
Fig 2. Pattern analysis of Euclidean

그림 2는 유클리디안에서의 사례 데이터와 새로운 데이터의 계산 방법 패턴 분석을 나타낸 것이다. 점선은 새로 입력된 데이터와 사례 데이터 중 유클리디안 거리 계산 방식의 키인 거리가 짧은 것인 d(2,5)와 d(4,5)가 추출되고, 이중 변수가 여러 개 일치하는 d(2,5)가 추출되는 것이다.

표 7은 기존의 방법에서 발견된 패턴을 가지고 새로운 사례 데이터의 계산 패턴 방법을 나타내는 것을 수식화한 것이다.

아래의 과정은 입력 데이터 선별과정을 나타내고 있다.

사례 $X=(x_1,x_2,x_3,...x_n)$와 입력 $Y=(y_1,y_2,y_3,...y_n)$는 입력된 값을 계산하기 위한 패턴이며, 두 패턴의 유사도 △(X,Y)는 식 (3)과 같이 정의된다.

표 7. 데이터의 수식화 모델
Table 7. Mathematization model of data

UID \ 변수	1	2	3	4	5	6	n
$1(x_1)$	x_1^1	x_2^1	...	...	...	...	x_n^1
$2(x_2)$	x_1^2	x_2^2	...	...	...	...	x_n^2
$3(x_3)$	x_1^3	x_2^3	...	...	...	...	x_n^3
$4(x_4)$	x_1^4	x_2^4	...	...	...	...	x_n^4
5(Y=y)	y_1	y_2	...	...	...	...	y_n

$$\triangle(X,Y) = \triangle(x_i^m,y_i) = |x_i^m,y_i| = \triangle(x^m,y) \quad (i=1,2,.....n) \tag{3}$$

$\exists x_{n1},x_{n2},x_{n3} \geq 0, n=1,2,3,...,k$이고 x_1,x_2,x_3에 대한 각각의 데이터는 0이상인 양의 실수로 입력되어야 하고, $x_1(n)$은 x_1변수의 n번째 색인을 가진 사례 변수들의 각 변수값을 명시한다. 또한, 비교가 끝난 데이터는 새로운 사례DB에 저장되도록 한다.

$$\min|x_i^m - y_i|, m=1,2,3,...,n$$

즉, $|x_1^1-y_1|,|x_1^2-y_2|,|x_1^3-y_3|,|x_1^4-y_4|$ (4)

여기서 $\min|x_i^m-y_i|$는 각각의 i번째 입력되는 값의 최소값을 의미한다.

식 (4)는 각 첫 번째 변수에 대해 값을 비교한 후, 최소값을 가진 변수로 저장한다. 식 (4)를 변수에 대입하여 최소값을 구한다. 또한 같은 변수 항목을 비교하여 같은 항목을 가진 것에 대해 카운터를 준다.

각 변수에 대한 항목에서 항목의 차가 0으로 나온 결과와 차가 작게 나온 항목을 가진 변수를 해 값으로 선정하도록 한다. 비교하는 각 항목이 같은 것이 많거나 만약 같을 때에는 두 항목에 대한 차가 각 레코드를 비교했을 때 가장 작게 나온 레코드를 가진 레코드를 해로 선택한다.

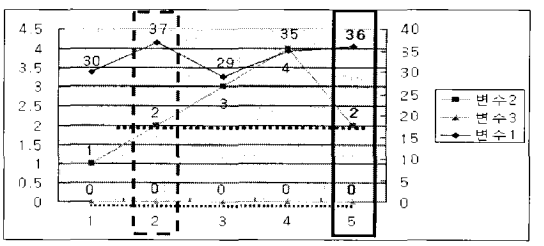
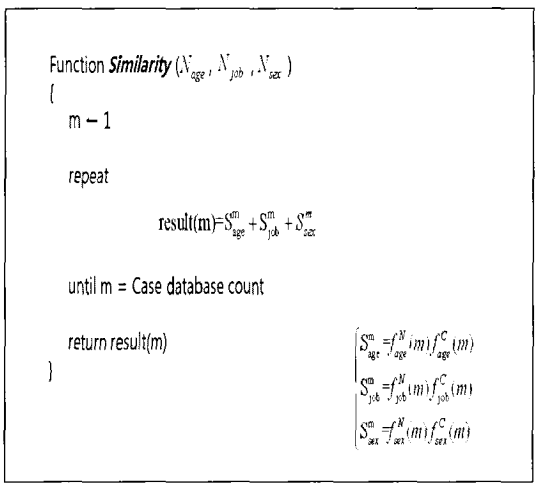


그림 3. 차트로 보는 SE 알고리즘
Fig 3. SE algorithm depicted by chart

그림 3은 기존 패턴방식과 제안하는 새로운 패턴 방식을 전체적으로 비교한 것을 나타낸 것으로, 사례 데이터와 새 데이터 중에서 유사도와 유클리디안 거리를 혼합한 SE계산방식을 이용하면 사례 데이터와 새 데이터의 값이 서로 가까운 데이터를 추출하고, 변수 중에서 일치하는 변수가 많은 사례 데이터를 추출하는 결과 화면을 나타낸 것이다.

그림 4는 기존의 계산 알고리즘과 제안하는 SE계산방법에 대한 알고리즘을 나타낸 것이다. 첫 번째의 유사도 알고리즘은 이전에 언급되었던 유사도 검색 방법의 새로운 사례와 사례베이스에 있는 과거의 모든 사례와의 유사도를 측정하여 유사 사례를 찾는 방법이다. 두 번째 유클리디안 알고리즘은 나이, 직업, 성별이라는 스칼라 입력 값에 대하여 각각의 값들에 대한 거리를 구하는 방법이다. 마지막에 기술한 알고리즘은 전체적인 사례들을 모두 계산하는 기존의 계산방법을 벗어나 새로운 사례가 들어왔을 경우 발생가능성이 있는 사례 데이터에 대한 연산만 수행(distance= job_count*sex_count)하여 계산수행에 영향을 줄 수 있는 불필요한 연산을 최소화하여 관련된 사례를 추출하는 것이다.

제안하는 마지막 SE알고리즘은 사례변수들의 가중치들을 곱하여 나올 수 있는 최대의 값을 구하고, 그 값을 가지고 추출할 최소의 데이터들을 검출한다. 검출된 데이터를 가지고 기존의 유클리디안 거리 공식을 이용해 데이터를 마지막으로 추출한다.



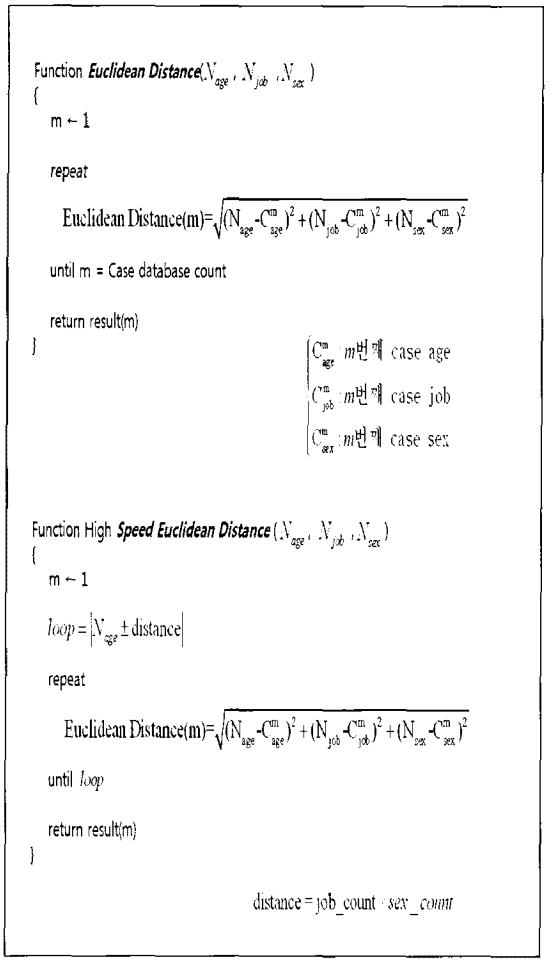


그림 4. 기존의 계산방법과 SE계산방법에 대한 알고리즘
Fig 4. Algorithm for existing computation method and SE computation method.

Ⅳ. 전산 모의 실험 및 평가결과

4.1 시스템 평가방법

시스템 평가방법은 무작위 데이터 100개를 30회, 50회씩 컴퓨터 사양의 변화(Delay Time)를 주어 사용하였다. 또한, 알고리즘의 성능을 평가하기 위해서 듀얼코어 CPU 2.5G와 3GB의 메인 메모리를 장착한 개인용 컴퓨터에서 실험을 하였다.

실험 컴퓨터의 운영체제는 Windows XP이고, 사용된 언어는 Micro_ soft Visual C++6.0으로 구현하였다. 사용한 데이터의 정확도는 기존의 계산방식인 유클리디안 계산방법으로 확인하였고, 모든 데이터를 계산하는 방법을 보완하여 새로운 사례가 들어왔을 경우 발생가능성이 있는 사례 데이터에 대한 연산만 수행하였다. 또한, 컴퓨터 환경에 따른 데이터 처리 속도변화를 보기위해 기존 계산방법과 제안하는 방법에 대해 성능평가를 하였다.

4.2 시스템 평가결과

다음은 기존의 계산방법과 제안한 계산방법을 평가한 결과를 나타낸 것이다.

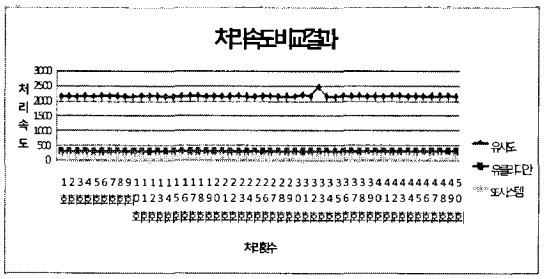


그림 5. 제안한 시스템의 결과화면
Fig 5. Screen shot of suggested system result

그림 5는 본 논문에서 사용한 무작위 데이터 100개를 50회 처리한 경우, 제안하는 패턴 분석 방식을 작성된 프로그램으로 실행했을 경우 나타나는 결과 차트이다. 제안한 SE시스템이 기존의 계산방법보다 빠르다는 것을 알 수 있다.

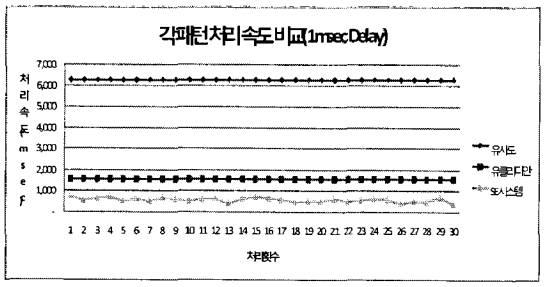


그림 6. 컴퓨터 환경에 따른 Delay 속도 비교(1msec)
Fig 6. Delay speed comparison (1msec) according to computer specifications

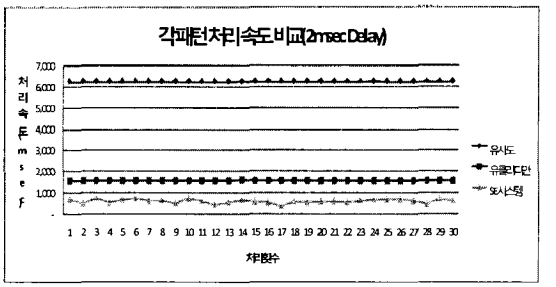


그림 7. 컴퓨터 환경에 따른 Delay 속도 비교(2msec)
Fig 7. Delay speed comparison (2msec) according to computer specifications

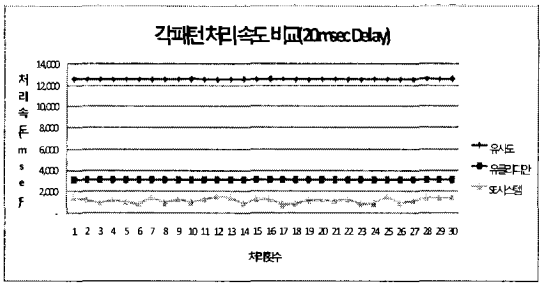


그림 8. 컴퓨터 환경에 따른 Delay 속도 비교 (20msec)
Fig 8. Delay speed comparison (20msec) according to computer specifications

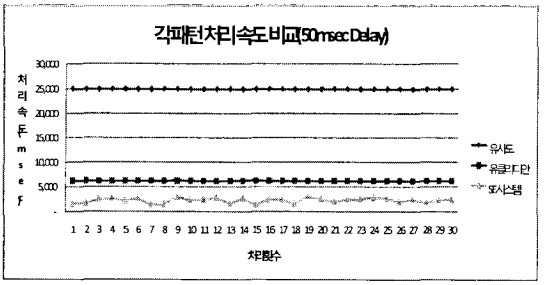


그림 9. 컴퓨터 환경에 따른 Delay 속도 비교 (50msec)
Fig 9. Delay speed comparison (50msec) according to computer specifications

그림 6, 7, 8, 9는 기존 알고리즘과 제안한 SE알고리즘을 발생할 수 있는 경우의 사항 등을 고려하여 무작

위 데이터 100개를 30회 처리하는 동안 처리 속도의 Delay Time을 각각 1msec, 2msec, 10msec, 50msec 등을 주어 알고리즘을 비교한 결과를 차트로 나타낸 것이다.

표 8. 기존 패턴방식과 제안한 패턴방식에 대한 시간과 공간 복잡도
Table 8. Time and Space Complexity for existing and suggested patterns

알고리즘	시간 복잡도	공간 복잡도
유사도	O(n)	O(n)
유클리디안	O(n)	O(n)
제안시스템	O(n)	O(n)

표 8은 기존의 계산방법과 제안하는 SE방식에 대한 계산알고리즘에 대한 시간복잡도와 공간복잡도에 대한 결과표이다. 표에서 보는바와 같이 시간복잡도와 공간복잡도는 기존의 알고리즘과 제안한 SE알고리즘은 같게 나타나지만, 제안한 SE알고리즘이 모든 사례 데이터를 비교·검색하지 않고 지정한 가중치에 속한 사례 데이터만을 검색하기 때문에 연산 속도가 가장 빠르게 나타나고 있다.

표 9는 기존 패턴방식과 제안한 SE계산방식에 사용한 계산 알고리즘을 동일 컴퓨터에서 무작위 데이터 100개를 50회 처리한 처리 속도의 결과 데이터를 표로 나타낸 것이다.

표 10과 그림 10은 기존 패턴방식과 제안한 SE계산방식에 사용한 계산 알고리즘을 동일 컴퓨터에서 무작위 데이터 100개를 50회 처리한 처리 속도의 결과 데이터중에서 각 알고리즘들의 처리속도 중 최고·최저의 속도 값을 표와 차트로 나타낸 것으로 기존의 알고리즘보다 제안한 SE알고리즘의 처리속도가 빠르게 나타났다.

그림 11은 각 알고리즘을 컴퓨터 환경(286컴퓨터에서 686컴퓨터)에 따른 처리 속도 변화를 보기 위해 Delay Time(각 10msec에서 400msce)을 주어 평가한 결과 차트이다. 이 실험은 무작위 데이터 100개를 각 Delay Time을 준 데이터 중 시간 당 2개 표본 데이터를 80번 처리한 것으로 제안한 SE알고리즘의 연산 처리 속도가 빠른 것으로 나타났다.

표 9. 기존 패턴방식과 제안한 패턴방식에 대한 처리 속도
Table 9. Processing time for existing and suggested patterns

	유사도	유클리디안	SE시스템
1회	2125	297	125
2회	2141	297	125
3회	2125	297	109
4회	2141	281	141
5회	2125	297	78
6회	2141	281	110
7회	2141	296	110
8회	2141	297	125
9회	2125	297	109
10회	2125	297	94
11회	2141	297	125
12회	2141	297	47
13회	2141	297	125
14회	2125	297	62
15회	2125	296	110
30회	2125	297	125
31회	2172	297	46
32회	2141	296	110
33회	2438	296	125
34회	2140	297	125
35회	2125	297	109
36회	2141	297	125
37회	2140	297	94
38회	2157	297	62
39회	2140	297	125
40회	2141	281	125
41회	2125	297	125
42회	2141	281	94
43회	2141	281	94
44회	2125	297	125
45회	2141	297	109
46회	2140	297	110
47회	2140	297	110
48회	2140	297	125
49회	2141	297	93
50회	2125	296	110

표 10. 각 알고리즘의 처리속도의 최고·최저 속도
Table 10. Max and Min speed for each algorithm processing time

(단위 : msec)

처리속도	유사도	유클리디안	SE시스템
최고	2,438	297	141
최저	2,125	281	46

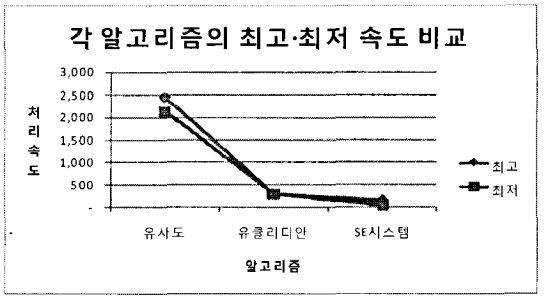


그림 10. 각 알고리즘의 처리속도의 최고·최저 속도 비교
Fig 10. MAX and MIN speed comparison for each algorithm processing time

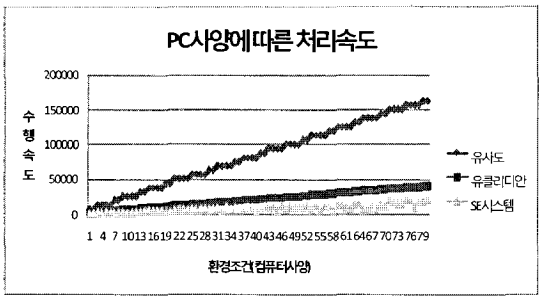


그림 11. PC사양에 따른 Delay변화 추이
Fig 11. Progress of changes in Delay according to PC Specifications

표 11. 컴퓨터 사양별 Delay Time한 처리속도의 최고·최저 속도
Table 11. MAX and MIN speed for processing Delay time according to PC Specifications

처리속도	유사도	유클리디안	SE시스템
최고	162,579	40,625	19,500
최저	6,250	1,562	672

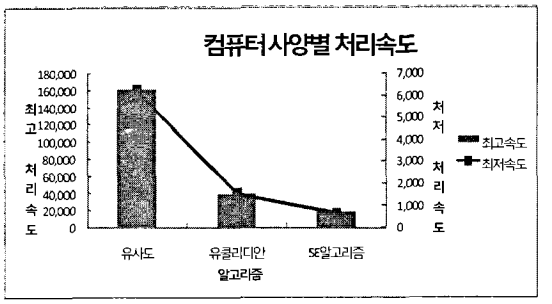


그림 12. PC사양에 따른 Delay변화 속도결과
Fig 12. Result for Speed of Delay change according to PC Specifications

표 11과 그림 12는 컴퓨터 사양별 Delay Time을 10msec에서 400msec으로 변화를 준 실험으로 기존의 알고리즘보다 제안한 SE알고리즘의 처리속도의 최고·최저 처리 속도가 2배에서 9배정도로 빠르게 처리되는 결과를 나타낸 것이다.

V. 결 론

사례기반추론(CBR:Case-Based Reasoning)은 데이터베이스가 대용량화됨에 따라 기존 데이터와 사례 데이터의 관계성을 추론하는 기법으로 유사도(Similarity)와 유클리디안(Euclidean) 거리 계산 방법을 이용해 효율적인 데이터 추출 알고리즘의 개발이 중요해지고 있다. 따라서 본 논문에서는 새 데이터를 모든 사례데이터에서 비교하는 것이 아니라, 발생가능성이 있는 범위를 두어 그 영역 안에 들어있는 사례데이터들에 대한 연산 패턴을 찾아 불필요한 후보 집합 수를 배제시켜 연산처리를 하는 컴퓨터 사양에 크게 영향을 받지 않고, 효율적으로 연산 속도를 높이는 알고리즘을 제안한다.

제안한 SE(Speed Euclidean-distance)계산 알고리즘은 데이터마이닝 기법 중 하나인 사례기반 추론 중 두 개체 사이의 유사성을 측정하는 유클리디안 거리(Euclidean Distance)와 유사도(Similarity)를 계산할 때 발견된 패턴을 이용한 것이다.

제안하는 SE알고리즘의 성능 평가를 하기 위해 사용한 데이터는 매 회마다 무작위 사례 데이터 100개를 생성하여 유사도, 유클리디안, SE계산 알고리즘에 대해 동시성을 가지기 위해 서로 비교 및 검토에 대한 실험을 하

였으며, 이러한 작업을 50회 반복하여 일관성을 가질 수 있도록 시도하였다.

그 결과 발생가능성이 있는 범위에 의해 추출된 사례 데이터에서 새로운 데이터를 비교 탐색하는 SE알고리즘이 기존 알고리즘보다 연산 속도 면에서 향상된 성능을 보임을 확인할 수 있었다.

본 논문에서 실험한 결과는 기존의 알고리즘인 유클리디안과 유사도 계산 알고리즘보다 각각 최저 2배에서 15배와 최고 6배에서 47배 연산속도가 빠르게 처리되는 것으로 나왔다. 환경조건을 최대한 악화(Delay Time을 더 많이 주었을 때의 환경)시켜 시도하여 실험할 필요가 있겠지만, 결과에 대한 자료만 보더라도 그 이상에 대한 실험을 시도할 경우 현재 제시된 자료보다 더 우수한 결과가 나올 수 있다.

제안하는 SE알고리즘이 기존 알고리즘보다 연산속도가 빠르게 나타나는 것은 기존 알고리즘은 모든 사례 데이터를 스캔하거나 계산식이 복잡하기 때문인 것이다. 알고리즘의 성능분석에 있어서 사용하는 시간복잡도와 공간복잡도에서 각각의 결과는 동일하게 나왔지만, 제안한 SE알고리즘이 기존의 알고리즘보다 연산속도가 빠르게 나왔다. 이유는 가중치를 두어 모든 사례 데이터를 스캔하는 것이 아닌 가중치에 포함한 사례 데이터에서 새로운 데이터를 비교하기 때문에 연산 속도가 빠르게 나타난 것이다. 또한, 각 알고리즘을 컴퓨터 환경(286컴퓨터에서 686컴퓨터)에 따른 처리 속도에 Delay Time(각 10msec에서 400msec)을 주어 처리속도를 평가한 실험에서도 제안한 SE알고리즘이 최저 2배에서 최고 9배 정도로 빠른 것으로 나타났다.

참고문헌

[1] 이태림 외 4인 공저, 데이터마이닝, pp.107-138, 한국방송통신대학교출판부, 서울, 2005.

[2] 김연형 외 3인 공저, 고객관계관리와 데이터마이닝, pp. 278-288, 교우사(출), 서울, 2006.

[3] 이형용, "자기조직화지도와 사례기반추론을 결합한 추천 모형:온라인 커뮤니티 추천 시스템의 사례", e-비지니스연구, 제9권 제1호, pp. 309-327, 2008.

[4] 김우생, "XML 문서의 구조와 내용을 고려한 유사도 측정", Journal of korea Multimedia Society Vol. 11, No. 8, pp. 1043-1050, 2008.

[5] 成百均, "판매지원 에이전트에서의 사례기반추론 방법", 産業科學技術硏究所 論文集 第8輯 Vol. 8, pp.221-230, 2000.

[6] 이재식, 이진천, "유사도 임계치에 근거한 최근접 이웃 집합의 구성", 한국지능정보시스템학회논문지 제13권 제2호, pp. 1-14, 2007.

[7] 정석훈, 서용무, "사례기반 추론을 이용한 암 환자 진료비 예측 모형의 개발", 경영정보학 연구(The Journal of MIS research), Vol.16, No.2, 2006.

[8] 김경재, 김병국, "데이터 마이닝을 이용한 인터넷 쇼핑몰 상품추천시스템", 한국지능정보시스템학회논문지 제11권 제1호, pp. 191-205, 2005.

[9] Han, J. and M. Kamber, Datamining : Concepts and Techniques, Morgan Kaufmann Publishers, San Francisco, CA, 2001.

[10] 이정원, 김호숙, 최지영, 김현희, 용환승, 이상호, 박승수, "데이터마이닝 알고리즘의 분류 및 분석", 정보과학회논문지 : 데이터베이스 제28권 제3호, pp. 279-300, 2001.

[11] B. Ralph, J. Kolodner and E, "Representation in case-based reasoning", Knowledge Engi- neering Review, vol. 20, pp. 209-213, 2005.

[12] YAN LI, XI-ZHAO WANG, MING-HU HA, "On-Line Multi-CBR Agent Dispatching", Proceedings the second International Con_ ference on Machine Learning and Cyber_ netics, Xi'an, pp. 2-5, 2003.

[13] Chun, S. H. and Y. J. Park, "Dynamic Adaptive Ensemble Case-based Reasoning : Application to Stock Market Prediction", Expert System with Applications, Vol. 28, pp. 435-443, 2005.

[14] Chanchien, S. W. and M. Lin, "Design and Implementation of a Case-based Reasoning System for Marketing Plans", Expert Systems with Applications, Vol. 28, pp. 43-53, 2005.

[15] Roh, T.H., K.J. Oh, and I. Han, "The Collaborative Filtering Recommendation Based on SOM Cluster-Indexing CBR", Expert Systems with Applications, Vol. 25, No. 3, pp. 413-423, 2003.

[16] 안현철, 김경재, 한이구. "효과적인 고객관계관리를 위한 사례기반추론 동시 최적화 모형", 한국지능정보시스템학회논문지, 제11권 제2호, pp. 175-195, 2005.

[17] Park, C. S. and I, Han, "A case-based reasoning with the feature weights derived bt analytic hierarchy process for bankruptcy prediction", Expert Systems with App_ lications, Vol. 23, pp. 255-264, 2002.

[18] Aamodt, A. and E. Plaza, "Case-based Rea_ soning : Fundamental Issues, Methodological Variations, and System Approaches", Artificial Intelligence Communication, Vol. 7, No. 1, pp. 39-59, 1994.

[19] Finnje, G., Sun, Z., "Similarity and Metrics in Case-Based Reasoning", International Journal of Intelligent Systems, Vol. 17, pp. 273-287, 2002.

저자소개

윤종찬(Jong-Chan Yun)

2003. 2 동명정보대학교 경영정보학과 경영학사
2005. 2 부경대학교 대학원 전산정보학과 공학석사
2008. 2 부경대학교 대학원 전자상거래시스템학과 공학박사
※ 관심분야 : 전자상거래, 데이타마이닝, 유비쿼터스, e-CRM 등

김학철(Hak-Chul Kim)

2000. 2 부경대학교 전자공학
2005. 2 부경대학교 전자공학과 공학석사
2010. 현재 부경대학교 대학원 전자공학과 박사과정
※ 관심분야 : 병렬분산처리, 패턴인식, 알고리즘 등

김종진(Jong-Jin Kim)

1983. 2 경북대학교 공학사
1985. 2 한국과학기술원 전기 및 전자공학과 공학석사
1995. 2 경북대학교 대학원 전자공학과 공학박사
1987~현재 부경대학교 전자공학과 교수
※ 관심분야 : 병렬분산처리, 컴퓨터구조, 상호접속망 등

윤성대(Sung-Dae Youn)

1980. 2 경북대학교 컴퓨터공학과 공학사
1984. 2 영남대학교 대학원 전자계산학과 공학석사
1997. 2 부산대학교 대학원 전자계산학과 이학박사
1981~1986 경남정보대학 전산과 조교수
1991~1992 MIT 방문교수
1992~1995 부산공업대학교 전산소장
1989~현재 부경대학교 컴퓨터공학과 교수
※ 관심분야 : 병렬처리, 멀티캐스트통신, 데이터마이닝 등