경영정보학연구
제14권 제3호
2004년 9월

# 새로운 추론 기법 소개: 코드배열기반 추론

강 민 철*, 임 호 윤**

# Introduction to a New Reasoning Technique: Code Arrangement-Based Reasoning

Mincheol Kang, Hoyoun Im

When humans make decisions, they differentiate classifications of individual attribute variables that affect the decisions according to the importance and pattern of each attribute variables. The present study examines the practicality of the proposed Code Arrangement-Based Reasoning (CABR), which resembles the human's way of reasoning. To this end, we developed a CABR technique that classifies each attribute variable affecting significant impacts on the target variable into a cluster and assigns a code to the cluster. For verifying the proposed technique, both case-based reasoning and CABR were used for the customer continuance judgment problem of an automobile insurance company. Results indicated that the performance of CABR is close to the one of the case-based reasoning. The CABR also shows the possibility of using bio-informatics techniques for organizational data analysis in the future.

Keywords : Code Arrangement-based Reasoning, Case-based, Reasoning

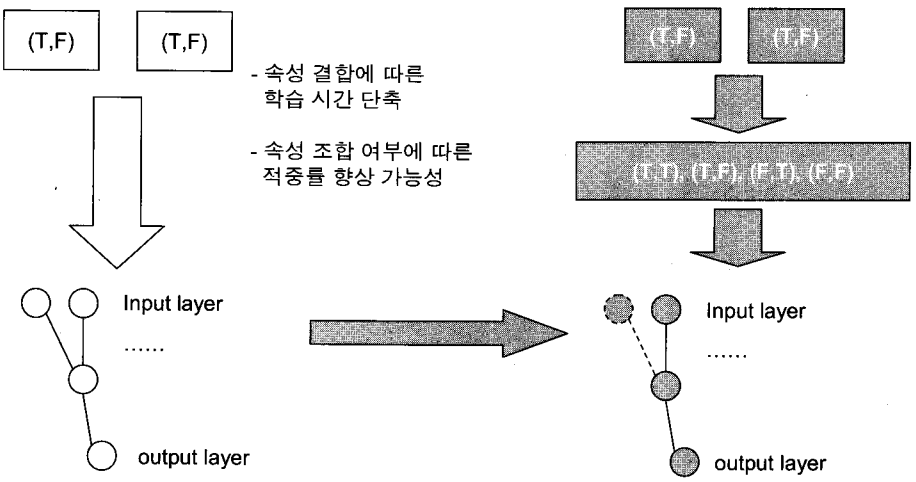
* 아주대학교 경영대학 e-비즈니스학부
** 아주대학교 일반대학원 경영정보학과

# Ⅰ. 서 론

현대 사회에서 정보 기술의 사용이 보편화되고, 축적되는 정보의 양이 늘어남에 따라 각 조직에서는 데이터를 처리하기 위하여 방대한 데이터 베이스를 필요로 하게 되었으며, 아울러 이러한 방대한 데이터를 분석하고 처리하기 위해서 다양하고 복잡한 프로세스가 필요하게 되었다. 이에 따라, 지식의 발견과 분석을 위한 수단으로서 통계적 기법, 인공 신경망, 사례기반 추론 등의 데이터 마이닝 기법들이 그 해결책으로 등장하였다.

그러나, 이들 기법들은 인간의 의사 결정 과정 중 문제 해결을 위해 사용되는 여러 고려 사항들, 즉, 목적 변수를 설명하고 있는 각 속성 변수들을 정리하는 방식을 내포하고 있지 않다. 인간은 의사결정 과정에서 자신의 결정에 영향을 미치는 여러 변수들에 대해서 세분화하고 분류하여, 중요한 영향을 미치는 변수와 그렇지 않은 변수에 대해서 그 패턴의 정확도를 달리하여 기억해 두었다가 이를 결정에 반영한다. 예를 들어, 꽃을 판단하기 위한 결정에 있어, 목적변수가 되는 꽃 이름을 판별하기 위해서 꽃의 크기, 모양, 색, 냄새 등과 같은 속성 변수를 세분화하여 판별하게 되는데, 이 때 이들 각 속성의 중요도와 패턴에 따라서 속성별 분류 체계를 달리함으로써 보다 정확한 판별이 가능해진다[Duda et al., 2001; Gose et al., 1996]. 본 연구에서는 목적변수를 설명하고 있는 각 속성 변수들을 정리하는 인간의 추론 방식과 유사한 방식의 추론을 하는 새로운 방식의 추론 기법인 코드배열기반 추론 기법을 개발하여 이를 소개하고자 한다.

속성 변수의 조절을 통한 분석 기법은 이미 인공 신경망에서 사용되고 있는데[Doak, 1992], 모델 설계 시에 속성 변수의 결합, 또는 조절을 통해서 인공 신경망을 구축하는 경우, <그림 1>과 같이 입력 변수들의 조합을 통해 새로운 인공신경망 모델을 만들어 낼 수 있다.




<그림 1> 인공 신경망에서의 입력 변수의 조합

그러나 속성 변수의 조절을 통한 인공 신경망의 분석 기법은 불연속적인 값을 가지는 이산형 데이터(discrete data)에 대해서는 속성 결합이 가능하나, 연속적인 값을 가지는 연속형 데이터(continuous data)는 속성 결합을 할 수가 없다. 본 연구에서 제시하는 코드배열기반 추론은 사례의 코드화를 통해 사례기반 추론을 변형한 것으로서 이러한 데이터 형태에 무관하게 속성 결

합을 할 수 있다.

방대한 데이터베이스에는 수치적 자료뿐만 아니라, 텍스트와, 기호와 같은 상징적인 데이터 등을 포함하고 있으며, 이러한 데이터들은 중복되고, 많은 에러를 포함하고 있다, 따라서, 이러한 데이터를 저장하고 있는 데이터베이스에서 새로운 지식을 발견하고 의미를 부여하기 위해서는 패턴 인식이 이러한 분석 과정에서 중요한 법칙이 된다[Pal, 2004]. 본 연구에서 소개하는 코드배열기반 추론은 데이터의 패턴을 이용한 추론 방식이다.

본 논문에서는 원시 데이터를 코드화하여 유사 패턴을 추론하는 새로운 방식의 추론 기법인 코드배열기반 추론을 소개한 다음, 이의 유용성을 검증하기 위하여 해당 추론 기법을 사용하여 자동차 보험사 고객의 보험의 유지 및 해지에 대한 적중률을 조사하고, 이 결과를 기존의 사례기반 추론 기법을 적용하여 얻은 결과와 비교하고자 한다.

## Ⅱ. 문헌 연구

### 2.1 인공 신경망

인공 신경망은 인간 두뇌의 학습 능력을 모방하여 문제 해결에 적용하고자 하는 기법으로써, 1957년 Rosenblatt에 의해 발표된 Perceptron이 최초의 인공 신경망 모델이다. 인공 신경망은 기존의 인공지능 기법이나 학습 방법과는 다른 여러 가지 장점들을 가지고 있다. 인공 신경망의 특징을 정리하면 다음과 같다[Nelson and Illingworth, 1991].

첫째, 인공 신경망은 분산 처리의 특성을 가지고 있다. 인공 신경망은 연결선의 연결 강도를 조정함으로써 학습을 하기 때문에 전체 연결선 자체가 지식을 표현하고, 저장하고 있는 것을 의미한다. 또한, 인공 신경망은 인간의 뇌가 입력 자극을 뇌 속의 기억과 매칭함으로써 기억 내용을 찾는 것처럼 저장된 기억을 찾아낸다. 인간은 데이터가 변형, 파괴되거나 생략되더라도 정확한 패턴을 재생하는데 능숙하며 이러한 특징이 인공신경망에도 적용되어, 일반적으로 인공신경망이 오류극복성을 가지게 한다.

둘째, 인공 신경망은 구조적으로 병렬일 뿐만 아니라, 처리 순서도 병렬적이며 동시적이다. 따라서 데이터 수가 증가함에 따라 처리속도가 늦어지는 특성을 가지는 기존의 폰노이만 방식의 직렬 처리 시스템이 가지는 단점을 극복할 수 있다. 뇌가 병렬 분산 처리를 통해 다량의 정보를 안정적으로 처리할 뿐만 아니라 컴퓨터로 처리하기 힘든 뉴런의 인식을 짧은 시간에 수행할 수 있듯이, 뇌를 모방한 인공 신경망도 여러 개의 처리 요소들이 서로 영향을 주며, 동시에 서로 다른 연산을 수행하여 짧은 시간에 방대한 양의 데이터들의 연산을 처리할 수 있다.

셋째, 인공 신경망은 적응 능력을 가진다. 이는 자기 조정 능력을 말하는 것으로, 학습 및 자기 조직화와 훈련 등을 조정 능력이라 할 수 있다. 학습은 각 노드의 연결 강도를 계속 변화시킴으로써 이루어지게 되며, 한번에 여러 연결 강도의 조정이 일어나는데 반복적인 훈련을 통해 목표 출력 값과 출력 값 간의 오차가 더 이상 커지지 않는 안정적인 상태가 되도록 한다.

위와 같은 장점들을 가진 인공 신경망은 비교적 높은 예측률을 보이는 것으로 알려져 있으나, 데이터 분석 과정에 블랙 박스와 같은 숨겨진 층(Hidden Layer)이 존재하기 때문에 어떻게 그러한 출력 결과가 나오게 되었는지를 설명하지는 못한다. 따라서 출력된 결과에 대해서 수용할 지 거부할 지에 대한 논란이 있다. 또한 인공신경망은 데이터 정제 과정 방식의 종류가 많으며, 사용된 정제 과정 방식에 따라서 결과에 많은 영향을 미친다. 그리고 인공 신경망은 인간의 두뇌의 학습 능력을 모방하여 개발되기는 하였으나, 의사결정을 하기 위해 필요한 각 변수들을 분류, 패턴화하는 과정을 모방하지는 못한다. 또한, 실제

적으로 인간이 목적 변수를 추론하기 위하여 필요한 수 많은 변수 데이터를 각각 0~1 사이 값으로 변환하여야 하기 때문에, 각 변수가 가지는 의미를 놓치게 되는 문제가 있다.

본 연구에서 소개하는 코드배열기반 추론은 그 과정이 추후 Ⅲ장에서 보다 자세히 설명이 되겠지만, 코드화된 데이터들간을 비교하여 유사 배열을 추론하는 방식을 채택하고 있기 때문에 추론 과정 및 결과가 논리적으로 명확하게 서로 연결이 된다. 또한 코드배열기반 추론은 인공 신경망과 달리 인간의 추론 방식과 유사하게 데이터의 중요도에 따라 상세화 또는 간략화를 하는 차별적인 패턴화 방식을 채택하고 있다. 인간은 매우 복잡한 정보 시스템으로서 데이터를 간략화 시키는 능력과 아주 우수한 패턴 인식 능력을 보유하고 있다[이성환, 1994]. 인간은 중요하다고 판단되는 데이터에 대해서는 매우 상세하게 기억 장치에 저장하고, 중요하지 않다고 판단되는 데이터에 대해서는 간략화하여 기억한 후 이를 추론에 사용하는데, 코드배열기반 추론은 차별적인 패턴화를 통해 이를 모방하고 있다.

## 2.2 사례기반 추론

사례기반 추론은 인간이 과거 경험에 비추어 사물을 인식한다는 점에 착안하여 만들어진 것으로, 기억을 통해 현재 자신이 직면한 문제와 가장 유사한 과거 문제의 해결을 조회하여 이를 수정함으로써 새로운 문제의 해를 찾는다는 점에서 사례기반 추론의 정당성을 찾을 수가 있다[Riesbeck and Schank, 1989]. 다시 말해, 사례기반 추론은 과거에 문제를 풀었던 경험을 새로운 문제에 적용하여 해결책을 제시하는 시스템으로 정의할 수 있다. 사례기반 추론이 가지는 일반적인 특징은 다음과 같다[Watson, 1997].

첫째, 사례기반 추론은 새로운 문제가 입력되어야만 학습을 할 수 있는 사후 학습 기법이다. 즉, 학습과 추론이 문제가 입력되는 시점에 일어난다.

둘째, 과거의 사례를 기반으로 새로운 사례의 해를 구하기 때문에 설명력이 강하다.

셋째, 인공 신경망은 연속형 속성, 범주형 속성들 중에서 연속형 속성만을 사용하기 때문에 원본 데이터에 대한 변환 과정이 필요하다. 그러나 사례기반 추론은 이러한 데이터 속성 유형에 상관없이 추론과 학습이 가능하다.

넷째, 학습과정이 쉽다. 사례기반 추론은 해결한 문제의 사례를 사례베이스에 저장하고, 경우에 따라서는 실패한 문제의 사례까지 저장을 함으로써 다음에 해결해야 할 사례에 대한 해결책을 더욱더 풍부히 가지게 된다.

사례기반 추론은 의사결정에 사용된 데이터를 다시 저장하여 경험으로 축적하고 있다는 점에서 의사결정 시 인간이 과거의 경험을 바탕으로 결정을 내리는 것을 모방한 기법이다[Riesbeck and Schank, 1989]. 그러나 목적 변수 추론 시 각 속성 변수에 대한 적정 가중치를 탐색하는 것은 쉽지 않은 일이고, 많은 시간이 소요되기 때문에 전체적인 학습에 상당히 오랜 시간이 걸리게 된다. 또한, 근접한 범위 내에 있는 값들을 개략적으로 동일한 것으로 취급하는 인간과 달리, 사례기반 추론은 속성 변수가 가질 수 있는 모든 값의 경우의 수를 구별하여 추론하기 때문에, 인간의 의사결정 과정을 모방하는데 한계가 있다. 예를 들어, 꽃의 색깔을 구별하려고 할 때 인간은 빨간색, 녹색, 파란색 등과 같이 몇 가지 색상에 대한 정보만을 필요로 하나, 사례기반 추론의 경우는 RGB에 속하는 모든 경우의 수(= 256 × 256 × 256)에 대한 정보가 필요하다는 것이다.

본 연구에서 소개하는 코드배열기반 추론은 코드화된 레코드를 사용한다는 점 등에서 사례기반 추론과 확연한 차이가 있으나, 레코드 간의 비교를 통하여 유사한 배열을 추론한다는 점에서 사례기반 추론과 방법상의 유사성이 존재한

다. 그러나 코드배열기반 추론은 사례기반 추론과 달리 정규화된 데이터를 사용하기 때문에 적정 가중치의 탐색이 매우 용이하다. 또한 코드배열기반 추론은 앞서 인공신경망 부분에서 설명한 바와 같이 데이터의 중요도에 따라 차별적인 패턴화를 하는 방식을 채택하고 있기 때문에 사례기반 추론의 경우처럼 속성 변수가 가질 수 있는 모든 값의 경우의 수를 구별하여 추론을 하는 것이 아니라 유사한 경우들간의 간략화를 통하여 추론을 하기 때문에 매우 효율적이며 인간의 추론 과정에 보다 근접해 있다고 볼 수 있다. 예를 들어, 코드배열기반 추론에 있어서 녹색 계통의 값을 가지는 모든 RGB 값은 단순히 녹색으로 간략화하여 추론할 수 있다.

## 2.3 K-Means 클러스터링

K-Means 클러스터링(Clustering) 기법은 각 데이터와 클러스터 중심 값과의 거리 차이를 최소화 시킬 수 있는 그룹화를 통해서 클러스터링을 하는 것으로, 이러한 작업을 반복적으로 시행하여 각 클러스터에 대한 데이터를 재배정함으로써 클러스터를 나누게 된다. 클러스터링의 장점과 단점을 살펴 보면 다음과 같다[Berry and Linoff, 1997].

클러스터링의 장점으로, 먼저 클러스터링은 대용량 데이터에 대한 탐색적인 기법으로서, 주어진 데이터의 도메인에 대한 사전 지식이 없이 의미 있는 데이터 구조를 찾아낼 수 있는 방법이라는 것을 들 수 있다. 두 번째 장점은 다양한 거리 측정 단위를 사용함으로써 거의 모든 형태의 데이터에 적용이 가능하다는 것이다. 또한 클러스터링은 적용하기가 쉽다. 사전에 특정 속성들을 입력이나 출력과 같은 역할로 정의하는 것이 필요하지 않고, 다만 데이터들 사이의 거리만이 분석에 필요한 입력 데이터가 된다.

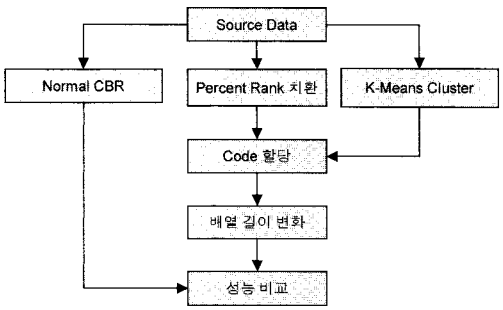
클러스터링의 단점은, 우선 클러스터 분석의 결과가 데이터들 사이의 거리 또는 유사성을 어떻게 정의하는가에 크게 좌우된다는 것이다. 특히 여러 다른 유형을 포함하는 데이터의 경우, 데이터들 사이의 거리 측정 단위를 정의하고 각 속성에 대한 가중치를 결정하는 것은 쉽지 않은 문제이다. 두 번째 단점으로, 클러스터의 수인 K를 정하는데 있어서, 만일 K가 원래의 데이터구조에 적합하지 않으면 좋은 결과를 얻을 수 없게 된다. 그러므로 여러 번의 탐색이 필요하다. 또 다른 단점은, 사전에 주어진 목적이 없기 때문에 클러스터링 후에 발견된 지식이 사용자가 찾는 목적이 아닐 수 있다는 것이다. 따라서 사용자가 발견해 낸 클러스터의 의미를 충분히 이해하고 실제적으로 활용하기가 쉽지 않다.

본 연구는 앞서 살펴 본 바와 같이, 인공 신경망과 사례기반 추론의 문제점을 보완하고자, 데이터를 상세화 또는 간략화하는 차별적인 패턴화 방식을 사용하는 인간의 추론 방법을 모방한 코드배열기반 추론을 개발하였다. 한편 본 코드배열기반 추론 기법은 데이터 정규화를 위해서 Percent Rank 또는 K-Means 클러스터링 알고리즘을 사용하였다.

# Ⅲ. 연구 방법

## 3.1 연구 모델

본 연구에서 사용된 연구 모델은 <그림 2>와 같다.




<그림 2> 코드배열기반 추론 방법

본 연구에서는 동일한 데이터에 대해서 사례기반 추론과 코드배열기반 추론을 사용하여 그 성능적 차이를 분석하고, 코드배열기반 추론의 현실적 적용 가능성 여부를 살펴보고자 하였다. 본 연구에서 특별히 코드배열기반 추론과 사례기반 추론의 결과를 비교한 이유는 코드배열기반 추론이 레코드간의 비교를 통하여 유사한 배열을 추론하는 점에 있어서 사례기반 추론과 방법상의 유사성이 있기 때문이다. 코드배열기반 추론에 있어서 코드배열을 통해 저장되는 각 레코드를 사례기반 추론에 있어서 하나의 사례와 유사한 개념으로 볼 수 있다. 한편, 본 연구에서는 코드배열기반 추론의 경우, 데이터 정규화를 하는 방식에 따라 Percent Rank를 이용한 코드배열기반 추론과 K-Means를 이용한 코드배열기반 추론으로 각각 나누어 그 차이를 비교 조사함으로써 코드배열기반 추론에 있어서 코드 할당 방법 변경에 따른 성능적 변화도 알아보고자 하였다.

Percent Rank를 이용한 코드배열기반 추론은 어느 속성에 해당하는 값들이 있는 필드(예: <표 1>의 X1열)에서 특정 변수 값(예: <표 1>의 X1열의 첫 번째 값인 25)이 해당 필드 전 구간에서 차지하는 퍼센트를 나타내는 Percent Rank를 구한 다음, 이 값이 전체 구간을 사분위로 나누었을 때 어디에 속하는 지에 따라 코드를 할당하여 목적 변수를 추론하는 방식이며, K-means를 이용한 코드배열기반 추론은 각 속성에 해당하는 필드를 K-means를 이용하여 그룹화 하고 나누어진 각 그룹에 코드를 할당하여 목적 변수를 추론하는 방식이다.

Percent Rank를 이용하여 코드를 할당하는 데이터 정규 과정은 다음과 같다.

자동차별 속성을 표기한 원시 데이터를 나타낸 <표 1>에서 Y 변수는 자동차의 종류를 나타내는 변수이며, X 변수는 자동차의 속성을 나타내는 변수이다. <표 1>에 나타난 각 X 변수의 값들에 대한 Percent Rank[1]값을 구하면 <표 2>와 같다. 이 변환 과정을 설명하기 위해, 차종 Y4의 X1 속성에 해당하는 값인 19가 변환되는 과정을 살펴 보면, 19라는 값은 X1 속성에 해당하는 총 9개의 값들 중 제 4순위이며, 상대적으로 19보다 작은 순위를 가지는 값들은 4개, 19와 같거나 큰 순위를 가지는 값들은 4개가 있다. 따라서 Percent Rank 값은 4/(4+4) = 0.5가 된다.

<표 1> 원시 데이터 - 자동차 Spec 데이터

| | X1 | X2 | X3 | X4 | X5 | X6 |
|---|---|---|---|---|---|---|
| Y1 | 25 | 30 | 2,390 | 16 | 5 | 118 |
| Y2 | 18 | 22 | 3,077 | 16.6 | 5 | 161 |
| Y3 | 23 | 27 | 2,568 | 10.2 | 5 | 108 |
| Y4 | 19 | 22 | 2,932 | 16.8 | 5 | 130 |
| Y5 | 18 | 22 | 2,855 | 14.3 | 5 | 168 |
| Y6 | 18 | 23 | 3,395 | 16.2 | 5 | 168 |
| Y7 | 19 | 28 | 3,289 | 16.4 | 6 | 165 |
| Y8 | 17 | 24 | 4,209 | 87.9 | 8 | 140 |
| Y9 | 19 | 28 | 3,267 | 16.4 | 6 | 165 |

<표 2> Percent Rank 값으로 변환한 데이터

| | X1 | X2 | X3 | X4 | X5 | X6 |
|---|---|---|---|---|---|---|
| Y1 | 1 | 1 | 0 | 0.25 | 0 | 0.125 |
| Y2 | 0.125 | 0 | 0.5 | 0.75 | 0 | 0.5 |
| Y3 | 0.875 | 0.625 | 0.125 | 0 | 0 | 0 |
| Y4 | 0.5 | 0 | 0.375 | 0.875 | 0 | 0.25 |
| Y5 | 0.125 | 0 | 0.25 | 0.125 | 0 | 0.875 |
| Y6 | 0.125 | 0.375 | 0.875 | 0.375 | 0 | 0.875 |
| Y7 | 0.5 | 0.75 | 0.75 | 0.5 | 0.75 | 0.625 |
| Y8 | 0 | 0.5 | 1 | 1 | 1 | 0.375 |
| Y9 | 0.5 | 0.75 | 0.625 | 0.5 | 0.75 | 0.625 |

데이터 정규 과정의 다음 단계로, <표 2>에서 변환된 값에 대해서 각 X 변수의 필드 값의 범위를 4등분하여 네 개 구간으로 나눈 후, 4개의 코드 (A, C, G, T)를 각 구간별로 하나씩 할당하여

1) Percen Ran: 순위를 전체 관측치로 나눈 분수형 순위에 100을 곱한 백분율 순위를 계산한다.

<표 3>을 생성한다. 이 때, 각 코드는 A, C, G, T의 순서대로 0~0.25, 0.25~0.50, 0.50~0.75, 0.75~1의 구간에 할당된다. 예를 들어, 차종 Y4의 X1 속성에 해당하는 Percent Rank 값인 0.5는 0.25보다 크고 0.5보다 작거나 같은 값들을 나타내는 구간에 속하기 때문에 C 코드 값으로 변환된다. 이러한 변환 과정을 통해, 원시 데이터가 벡터 값의 의미를 가지는 데이터로 변환됨으로써 패턴 분석이 가능하게 되는 것이다. 한편, 특별히 본 연구에서 Percent Rank 값을 코드 값으로 전환할 때 특별히 4개의 코드를 사용한 이유는 염기 서열을 나타내는 기본 단위인 A, C, G, T를 사용함으로써, 데이터 분석 시 바이오 인포메틱스(Bioinformatics) 분야에서 사용되는 기법을 적용시킬 수 있는 가능성을 열어 놓기 위함이다.

<표 3> Percent Rank에 의해 코드를 할당한 데이터

| | X1 | X2 | X3 | X4 | X5 | X6 |
|---|---|---|---|---|---|---|
| Y1 | T | T | A | AT | A | A |
| Y2 | A | A | C | GT | A | C |
| Y3 | T | G | A | AA | A | A |
| Y4 | C | A | C | TC | A | A |
| Y5 | A | A | A | AC | A | T |
| Y6 | A | C | T | CC | A | T |
| Y7 | C | G | G | CT | G | G |
| Y8 | A | C | T | TT | T | C |
| Y9 | C | G | G | CT | G | G |

본 연구에서는 X 변수에 해당하는 값들을 기본적으로 4개의 구간에 나누어 배정한 후 각 구간별로 A, C, G, T 등 네 종류의 문자를 할당하는 것으로 처리를 하고 있으나, 목적변수에 대한 적중률에 중요한 영향을 미치는 주요 변수에 대해서는 보다 높은 해상도를 부여함으로써 적중률에 변화를 주고자 하였다. 여기서 해상도를 높인다는 뜻은 <표 3>에서 X4 변수 값 필드에 나타난 바와 같이 보다 자세한 분류를 위해 코드 문자의 자릿수, 즉 코드 배열을 늘려서 할당하는 것을 말한다. 인간은 의사결정 과정에서, 특정 변수가 결정에 중요한 영향을 미치는 속성 변수이거나 또는 이 속성에 다양한 패턴이 존재하는 경우, 그 속성에 대한 분류 즉, 패턴의 복잡성을 인간 스스로가 결정하고 인식하기 때문에[김상운, 2003], 코드배열기반 추론에서는 이러한 특성을 반영하기 위해 해상도를 조절하는 것이다. 예를 들어, 수 십만 건의 레코드를 포함하고 있는 데이터의 경우, 단순히 각 속성 필드값으로 4가지의 코드 중의 하나를 할당하는 것, 즉 1자리 코드 수로 나타내는 것보다 AA, AC, AG, AT, CA, … TG, TT 등과 같이 두 자리 코드 수로 나타내어 총 16개로 구분하는 것이 해당 속성을 통해 목적 변수를 추정함에 있어서 더 정교한 설명을 할 수 있기 때문이다. 참고로 <표 3>에서 X4 변수 값 필드의 경우처럼 두 자리 코드 수로 나타내기 위해서는 X 변수의 필드 값의 범위를 16등분하여 총 16개 구간으로 나눈 후, AA에서부터 TT까지 16개의 코드를 각 구간별로 하나씩 할당한다. 이 때, 각 코드는 AA, AC, AG, AT, CA, … TG, TT의 순서대로 0~0.0625, 0.0625~0.125, 0.125~0.1875, 0.1875~0.25, 0.25~0.3125, … 0.875~0.9375, 0.9375~1의 구간에 할당된다. 예를 들어, <표 2>에 있어서 차종 Y1의 X4 속성에 해당하는 Percent Rank 값인 0.25는 0.1875보다 크고 0.25보다 작거나 같은 값들을 나타내는 구간에 속하기 때문에 <표 3>에서 AT라는 코드 값으로 변환된다.

한편 K-means 클러스터링을 이용하여 코드를 할당하는 데이터 정규 과정을 살펴 보면, 먼저 원시 데이터에 각 속성 변수별로 K-means 알고리즘을 적용하여 해당 속성 변수에 속하는 값들을 그 값이 제일 작은 그룹부터 큰 그룹까지 4개 그룹으로 자동 그룹핑(Grouping)을 한다. K-means 알고리즘은 데이터 마이닝 분야에서 이미 널리 사용되고 있는 것이기 때문에 본 연구에서는 별도의 설명을 생략하기로 하며, 혹 자세한 설명을

필요로 하는 경우에는 Berry and Linoff(1997) 등 데이터 마이닝 관련 문헌을 참고하면 될 것이다. 데이터 정규 과정의 다음 단계는 앞서 K-means 알고리즘을 통해 그룹핑한 것들 중에서 제일 작은 그룹에 속하는 속성 변수 값은 A 코드로, 그 다음 그룹에 속하는 값은 C 코드로, 그 다음 그룹에 속하는 값은 G 코드로, 그리고 제일 큰 그룹에 속하는 속성 변수 값은 T 코드로 변환하는 것이다. 해상도를 조절하는 방식은 앞서 Percent Rank를 이용하여 코드를 할당하는 경우에서 설명한 내용과 동일하다.

본 연구에서는 Percent Rank 또는 코드배열기반 K-Means를 이용한 데이터 정규화 과정을 통해 속성 변수값들을 코드화한 후, 이 코드 패턴들에 대한 코드배열기반 추론을 실시하였는데, 각 레코드의 속성 변수에 할당된 코드들의 차이를 측정하고자 유사도 산정 방식 중 Distance 측정을 이용하였으며, Distance는 일괄적으로 1을 적용하여 저장된 전체 레코드와 비교하는 추론 방식을 사용하였다.

## 3.2 실험 데이터

본 연구에서는 Percent Rank를 이용한 코드배열기반 추론과 K-Means를 이용한 코드배열기반 추론 기법의 실용성을 검증하고자 이들 기법들을 사용하여 자동차 보험 회사 고객의 보험 유지 및 해지에 대한 적중률을 조사하였으며, 그 결과를 동일 데이터에 사례기반 추론 기법을 사용한 결과와 비교하였다. 이 실험을 위해 A 자동차 보험 회사의 실제 데이터를 사용하였으며, 총 254,052개의 Case 중 1,000개의 분석용 데이터를 Random Sampling 방식으로 추출하였다. 그리고 추출된 분석용 데이터를 구성하고 있는 총 50개의 필드 중, 해당 필드의 값이 100% Null 값을 가지는 경우, 그리고 Code 할당이 불가능한 값인 경우, 즉 이름 또는 주소와 같이 정규화가 불가능한 필드인 경우를 제외한 32개의 필드로 데이터 셋(Set)을 구성하였다. 실험에 사용된 데이터 셋의 각 필드에 해당하는 변수는 <표 4>와 같다.

<표 4> 데이터 필드 속성표

| 변수 | 변수 설명 | Type |
|---|---|---|
| A1 | 납 기 | Integer |
| A2 | 년 만기 | Integer |
| A3 | 보험기간 | Integer |
| A4 | 현 계약상태 | Code |
| A5 | 전 계약상태 | Code |
| A6 | 변경CODE | Code |
| A7 | 계약내용 변경횟수 | Integer |
| A8 | 계약자 거주지역 | Code |
| A9 | 계약자 생년 | Integer |
| A10 | 계약자 성별 | Code |
| A11 | 사망보험금 수익자관계 | Code |
| A12 | 피보험자 거주지역 | Code |
| A13 | 증권발행횟수 | Integer |
| A14 | 납입방법 | Code |
| A15 | 집금방법 | Code |
| A16 | 합계보험료 | Integer |
| A17 | 모집점포 CODE | Code |
| A18 | 집금 점포 CODE | Code |
| A19 | 입금횟수 | Integer |
| A20 | 피보험자 연령 | Integer |
| A21 | 대부 CODE | Code |
| A22 | 대부잔액 | Integer |
| A23 | 만기환급 지급액 | Integer |
| A24 | 영위직종(직업) | Code |
| A25 | 이재 CODE | Code |
| A26 | 이재횟수 | Integer |
| A27 | 청약일 | Integer |
| A28 | 보험시기 | Integer |
| A29 | 보험종기 | Integer |
| A30 | 최종 월도 | Integer |
| A31 | 생존기간 | Integer |
| **T1** | **정상/해지 여부** | **Code** |

<표 4>에 나타난 총 32개의 변수 중 T1으로 분류된 목적 변수를 뺀 나머지 A1부터 A31까지의 31개의 변수를 속성 변수로 정의하였으며, 각 속성 변수의 해당 값들은 데이터 정규화 시 Percent Rank와 K-means를 사용하기 위하여 수치 자료로 변환하였다. 예를 들어, 성별의 경우는 '여성'을 1로 '남성'을 0으로 변환하였다. 한편 추출된 1,000개의 실험용 데이터 셋 중에서 사용이 불가능한 레코드(Record) 8개를 제외한 992개의 레코드에 대한 Training Set과 Evaluation Set의 비율은 9 : 1로 정하였다.

## Ⅳ. 실험 결과

본 연구에서는 사례기반 추론, Percent Rank를 이용한 코드배열기반 추론, K-Means를 이용한 코드배열기반 추론 등 3개의 다른 추론 기법들을 실험 데이터 셋에 적용하여 목적 변수를 추론한 결과에 대해서 적중률을 비교, 분석하는 방식으로 실험을 진행하였으며, 그 결과는 다음과 같다.

### 4.1 사례기반 추론

사례기반 추론 시스템의 속성 가중치는 0~1로 하였으며, 유사도 산정 방식은 Distance 측정을 사용하였다. 또한 테스트 방법으로는 Leave-One-Out / 10 Fold Test 방식을 사용하였으며, 적응 방법은 Voting 방법을 사용하였다[Riesbeck and Schank, 1989]. 사례기반 추론에서 가장 좋은 적중률을 보인 5개의 weight set은 <표 5>와 같고, Training Set의 적중 평균 개수는 635개, 적중률 평균은 71.1%이다.

<표 5> 사례기반 추론의 가중치 및 적중률 - Training Set

| | w1 | W2 | w3 | w4 | w5 | w6 | w7 | w8 | w9 | w10 | w11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.80 | 0.42 | 0.62 | 0.76 | 0.41 | 0.19 | 0.83 | 0.08 | 0.69 | 0.88 | 0.55 |
| 2 | 0.77 | 0.54 | 0.69 | 0.85 | 0.40 | 0.17 | 0.13 | 0.68 | 0.85 | 0.56 | 0.32 |
| 3 | 0.82 | 0.05 | 0.13 | 0.82 | 0.41 | 0.93 | 0.60 | 0.77 | 0.02 | 0.30 | 0.32 |
| 4 | 0.45 | 0.69 | 0.45 | 0.90 | 0.26 | 0.94 | 0.80 | 0.26 | 0.94 | 0.19 | 0.11 |
| 5 | 0.37 | 0.40 | 0.43 | 0.99 | 0.34 | 0.07 | 0.68 | 0.21 | 0.42 | 0.58 | 0.54 |
| | w12 | W13 | w14 | w15 | w16 | w17 | w18 | w19 | w20 | w21 | w22 |
| 1 | 0.95 | 0.45 | 0.06 | 0.56 | 0.10 | 0.67 | 0.12 | 0.71 | 0.93 | 0.75 | 0.26 |
| 2 | 0.74 | 0.35 | 0.02 | 0.34 | 0.69 | 0.86 | 0.76 | 0.23 | 0.98 | 0.78 | 0.80 |
| 3 | 0.64 | 0.82 | 0.41 | 0.95 | 0.45 | 0.03 | 0.81 | 0.41 | 0.63 | 0.37 | 0.65 |
| 4 | 0.17 | 0.67 | 0.07 | 0.02 | 0.23 | 0.58 | 0.28 | 0.03 | 0.82 | 0.13 | 0.38 |
| 5 | 0.30 | 0.19 | 0.12 | 0.52 | 0.59 | 0.11 | 0.76 | 0.08 | 0.67 | 0.73 | 0.22 |
| | w23 | W24 | w25 | w26 | w27 | w28 | w29 | w30 | w31 | correct | % |
| 1 | 0.20 | 0.43 | 0.03 | 0.32 | 0.43 | 0.28 | 0.67 | 0.13 | 0.28 | 632 | 70.77 |
| 2 | 0.83 | 0.77 | 0.42 | 0.32 | 0.01 | 0.69 | 0.76 | 0.37 | 0.71 | 652 | 73.01 |
| 3 | 0.31 | 0.65 | 0.20 | 0.08 | 0.50 | 0.89 | 0.38 | 0.45 | 0.38 | 635 | 71.10 |
| 4 | 0.73 | 0.91 | 0.18 | 0.89 | 0.36 | 0.39 | 0.26 | 0.15 | 0.86 | 638 | 71.44 |
| 5 | 0.14 | 0.41 | 0.13 | 0.24 | 0.92 | 0.70 | 0.79 | 0.75 | 0.71 | 618 | 69.20 |

위 Training Set을 통해 얻어진 가중치를 가지고 Evaluation Set에 대해 실험한 결과는 <표 6>과 같다. Evaluation Set 99개 레코드 중 적중 평균 개수는 66.2개이며, 적중률 평균은 66.87%로 Training Set의 적중률과 비교하여 평균 4.24% 낮은 수치를 보였다.

<표 6> 사례기반 추론의 적중률 및 에러율 편차 - Evaluation Set

| | Correct | % | Training | Err |
|---|---|---|---|---|
| 1 | 67 | 67.677 | 70.773 | 3.09591 |
| 2 | 68 | 68.687 | 73.012 | 4.32545 |
| 3 | 64 | 64.646 | 71.109 | 6.46216 |
| 4 | 67 | 67.677 | 71.445 | 3.7678 |
| 5 | 65 | 65.657 | 69.205 | 3.54836 |

## 4.2 Percent Rank를 적용한 코드배열기반 추론

Percent Rank를 이용한 코드배열기반 추론에서는 해상도를 변화시킴으로써, 즉 변수에 할당되는 코드 자리수인 배열의 길이를 달리하여 적용함으로써 사례기반 추론과의 평균 적중률 차이를 분석하고자 하였다. 코드배열기반 추론에 있어서 Training Set의 적중률은 <표 7>과 같고, 적중 평균 개수는 632개, 적중률 평균은 70.77%로 사례기반 추론에 비해서 약간 낮은 수치를 보였다.

다음 <표 7>에서 각 셀(Cell)의 값은 각 필드에 해당하는 코드의 자리수, 즉 코드 배열의 길이를 나타내고 있다. 예를 들어, 1번 Set에서 각 필드에 할당한 코드는 모두 한 자리 수이며, 각 셀의 실제 코드 값은 A, C, G, T 중의 하나로 이루어져 있다. <표 7>에 나타난 결과를 분석해 보면, 1번 Set에서는 83%의 적중률을 보이고 있으며, 각 속성 변수에 해당하는 코드 자리수를 일괄적으로 모두 두 자리수로 변경함으로써 해상도를 높인 2번 Set에서는 적중률이 상당히 떨어지는 것을 볼 수 있다. 또한 일부 속성 변수들만의 코드 자리수를 두 자리 수~네 자리 수로 변경함으로써 해상도를 조절한 3, 4, 5번 Set의 경우에도 적중률의 수치가 코드 자리수로 한 자리 수를 사용한 1번 Set보다 낮은 적중률을 보이는 것으로 나타났다. 따라서 일반적으로 해상도를 낮추게 되면 적중률은 높아진다는 것을 알 수 있다. 그러나 해상도를 낮춘다는 것은 해당 속성 변수가 나타낼 수 있는 구간의 세분화를 희생함으로써 얻어지는 것으로서, 이는 해당 속성이 나타낼 수 있는 값의 범주를 줄여 필요 이상으로 너무 요약하여 나타낸 것이 될 수 있으며, 이에 따라 해당 속성의 목적 변수에 대한 설명력을 저하시키게 된다. 예를 들어, 꽃의 색깔이라는 속성의 경우, 해상도를 낮추어 청색 계열, 홍색 계열 등으로 나타낼 수 있으나, 해상도를 높이면 청색 계열은 하늘색, 파란색 등으로 홍색 계열은 붉은색, 분홍색 등으로 세분화하여 나타낼 수 있다. 따라서 적중률만을 고려하여 무조건 해상도를 낮추는 것은 바람직하지 않으며, 문제에 따라 적정 해상도를 유지하여야 할 것이다.

Evaluation Set을 가지고 코드배열기반 추론을 실시한 결과, 99개 레코드 중 적중 평균 개수는 63.6개이며, 적중률 평균은 64.24%로 66.87%의 적중률 평균을 보인 사례기반 추론과 비교하여 평균 2.63% 낮은 수치를 보였다. 한편 Training Set을 가지고 추론한 경우의 적중률과 Evaluation Set을 가지고 추론한 경우의 적중률 간의 차이인 에러율 편차는 평균 6.53%로서 Evaluation Set의 경우의 적중률이 Training Set 경우에 비해 낮은 수치를 보였고, 이 결과는 <표 8>에 나타나 있다. Evaluation Set을 가지고 실험한 결과, 앞서 <표 7>에 나타난 Training Set의 경우와 마찬가지로 코드 배열의 길이, 즉 속성 변수에 할당된 코드의 자리수가 짧을수록 높은 적중률을 보였고, 코드 배열의 길이를 길게함으로써 해상도를 높일수록 낮은 적중률을 보였다. 그러나 배열의 길이를 짧게하여 해상도를 낮춘 경우,

Training Set에서의 적중률과 Evaluation Set에서의 적중률 간의 차이를 나타내는 에러율 편차는 증가됨을 알 수 있고, 배열의 길이를 길게하여 해상도를 높인 경우에는 에러율 편차가 감소됨을 알 수 있다.

종합적으로 Percent Rank를 적용하여 코드를 할당한 추론의 경우, 다른 Set보다 코드 배열의 총 길이(각 필드에 할당된 코드 자리 수를 모두 합산한 것)가 짧은 1번 Set을 사용하였을 때, Test Set과 Evaluation Set에 있어서 모두 가장 높은 적중률을 보였다. 그러나 해상도를 낮추는 것은 속성 변수의 적정 세분화가 이루어 지지 못하는 문제와 Test Set 경우의 적중률과 Evaluation Set 경우의 적중률의 차이가 증대되는 문제가 발생한다.

<표 7> Percent Rank를 적용한 코드배열기반 추론의 해상도 및 적중률 - Training Set

| | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 | C9 | C10 | C11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | 1 |
| 4 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 3 | 1 | 1 |
| 5 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 4 | 1 | 1 |
| | C12 | C13 | C14 | C15 | C16 | C17 | C18 | C19 | C20 | C21 | C22 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 3 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 1 | 2 | 1 | 1 |
| 4 | 2 | 1 | 1 | 1 | 3 | 3 | 3 | 1 | 3 | 1 | 2 |
| 5 | 1 | 1 | 1 | 1 | 3 | 3 | 3 | 1 | 4 | 1 | 1 |
| | C23 | C24 | C25 | C26 | C27 | C28 | C29 | C30 | C31 | correct | % |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 745 | 83.427 |
| 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 546 | 61.142 |
| 3 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 647 | 72.452 |
| 4 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 3 | 628 | 70.325 |
| 5 | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 4 | 594 | 66.517 |

<표 8> Percent Rank를 적용한 코드배열기반 추론의 적중률 및 에러율 편차 - Evaluation Set

| | Correct | % | Training | Err |
|---|---|---|---|---|
| 1 | 74 | 74.747 | 83.427 | 8.67918 |
| 2 | 55 | 55.556 | 61.142 | 5.58666 |
| 3 | 67 | 67.677 | 72.452 | 4.77564 |
| 4 | 64 | 64.646 | 70.325 | 5.67828 |
| 5 | 58 | 58.586 | 66.517 | 7.9315 |

## 4.3 K-Means를 적용한 코드배열기반 추론

Training Set의 적중률은 <표 9>와 같고, 적중 평균 개수는 604개, 적중률 평균은 67.65%로 사례기반 추론에 비해 평균 3.45% 낮은 수치를 보였으며, Percent Rank를 이용한 코드배열기반 추론보다 평균 3.12% 낮은 수치를 보였다.

데이터 정규화로 K-Means를 이용한 코드배열기반 추론의 경우, 각 필드별로 클러스터링을 하여 분류된 그룹에 대하여 코드를 할당하였기 때

문에 아래의 <표 9>와 같이 변수와 데이터 Set에 할당된 값은 클러스터의 갯수로 표현하였다. Percent Rank를 적용한 추론 결과와 비교하기 위하여, K-Means를 이용한 추론에 있어서 클러스터 4개가 Percent Rank 경우의 코드 자리수 하나에 해당하도록 하였다. Percent Rank 실험 결과를 나타낸 <표 7>을 보면, 각 셀에 나타난 값은 코드 자리수를 나타내는 것으로서, Percent Rank를 적용한 추론에서는 A, C, G, T 네 개중 한 개를 하나의 코드 자리 수에 할당하였기 때문에, 예를 들어, <표 7>에 나타난 1 이라는 값은 <표 9>에 있어서 4에 대응된다고 볼 수 있다. 그런데, K-Means를 적용한 실험의 경우, 각 필드의 전체 데이터가 표현하고 있는 그룹의 범위를 넘어서서 클러스터를 지정할 수는 없도록 되어 있기 때문에 약간의 조정이 필요하다. 예를 들어, 성별의 경우에는 남과 여라는 두 가지 값만이 존재하기 때문에, 클러스터를 두 개 이상 분리 할 수 없다. 따라서 이러한 경우는 ACGT 중 A와 C만을 코드 배정에 사용하였다. 마찬가지로, <표 4>의 A1, A2, A3에 대응되는 변수들인 <표 9>의 CK1(납기), CK2(년 만기), CK3(보험기간) 변수들도 원시 데이터의 값이 두 가지만으로 구성되어 있기 때문에 A와 C만을 코드 배정에 사용하였다. 예를 들어, 보험기간의 경우, 5년과 10년 중 하나의 값을 가지고 있다.

<표 9> K-Means를 적용한 코드배열기반 추론의 해상도 및 적중률 - Training Set

| | CK1 | CK2 | CK3 | CK4 | CK5 | CK6 | CK7 | CK8 | CK9 | CK10 | CK11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 2 | 2 | 4 | 4 | 4 | 4 | 4 | 4 | 2 | 4 |
| 2 | 2 | 2 | 2 | 4 | 4 | 4 | 4 | 16 | 16 | 2 | 4 |
| 3 | 2 | 2 | 2 | 4 | 4 | 4 | 4 | 16 | 16 | 2 | 4 |
| 4 | 2 | 2 | 2 | 4 | 4 | 4 | 4 | 16 | 64 | 2 | 4 |
| 5 | 2 | 2 | 2 | 4 | 4 | 4 | 4 | 4 | 64 | 2 | 4 |
| | CK12 | CK13 | CK14 | CK15 | CK16 | CK17 | CK18 | CK19 | CK20 | CK21 | CK22 |
| 1 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 3 | 4 |
| 2 | 16 | 4 | 4 | 4 | 16 | 16 | 16 | 16 | 16 | 3 | 4 |
| 3 | 16 | 4 | 4 | 4 | 16 | 16 | 16 | 4 | 16 | 3 | 4 |
| 4 | 16 | 4 | 4 | 4 | 64 | 64 | 64 | 4 | 64 | 3 | 4 |
| 5 | 16 | 4 | 4 | 4 | 64 | 256 | 256 | 16 | 64 | 3 | 4 |
| | CK23 | CK24 | CK25 | CK26 | CK27 | CK28 | CK29 | CK30 | CK31 | correct | % |
| 1 | 4 | 4 | 3 | 4 | 4 | 4 | 4 | 4 | 4 | 706 | 79.05 |
| 2 | 16 | 16 | 3 | 4 | 16 | 16 | 16 | 16 | 16 | 546 | 61.14 |
| 3 | 16 | 16 | 3 | 4 | 4 | 4 | 4 | 4 | 16 | 657 | 73.57 |
| 4 | 64 | 64 | 3 | 4 | 4 | 4 | 4 | 4 | 16 | 598 | 66.96 |
| 5 | 64 | 64 | 3 | 4 | 4 | 4 | 4 | 4 | 16 | 514 | 57.55 |

다음 <표 10>에 나타난 바와 같이, Evaluation Set을 가지고 K-Means를 이용한 코드배열기반 추론을 실시한 결과, 99개 레코드 중 적중 평균 개수는 58.6개이며, 적중률 평균은 59.19%로 66.87%의 적중률 평균을 보인 사례기반 추론과 비교하여 평균 7.68% 낮은 수치를 보였으며, 64.24%의 적중률 평균을 보인 Percent Rank를 이용한 코드배열기반 추론보다 평균 5.05% 낮은 수치를 보였

다. 한편 Training Set 경우의 적중률과 Evaluation Set 경우의 적중률 간의 차이를 나타내는 에러율 편차는 평균 8.46%로서 Evaluation Set의 경우의 적중률이 Training Set 경우에 비해 낮은 수치를 보였다. K-Means를 이용한 코드배열기반 추론의 에러율 편차는 사례기반 추론이나 Percent Rank를 적용한 추론에 비해 더 큰 것이다.

<표 10> K-Means를 적용한 코드배열기반 추론의 적중률 및 에러율 편차 - Evaluation Set

| | Correct | % | Training | Err |
|---|---|---|---|---|
| 1 | 67 | 67.677 | 79.059 | 11.3826 |
| 2 | 51 | 51.515 | 61.142 | 9.62707 |
| 3 | 64 | 64.646 | 73.572 | 8.92576 |
| 4 | 63 | 63.636 | 66.965 | 3.32892 |
| 5 | 48 | 48.485 | 57.559 | 9.07394 |

3가지 추론 기법의 성능을 비교한 결과를 요약하면 <표 11>과 같다. 먼저, Percent Rank를 적용한 코드배열기반 추론의 적중률은 Evaluation Set의 경우를 기준으로 사례기반 추론의 적중률에 비해 평균 2.63% 정도 약간 낮은 것으로 나타났으며, 에러율 편차는 조금 더 높은 것으로 나타났다. 한편 K-means를 사용하는 코드배열기반 추론은 Evaluation Set의 경우를 기준으로 적중률에서 사례기반 추론은 물론, Percent Rank를 적용한 추론에 비해 더 낮은 것으로 나타났으며, 에러율 편차는 더 높은 것으로 나타나 비교 대상 추론 기법 중 제일 성능이 미흡한 것으로 나타났다.

<표 11> 추론 결과 요약

| | Training set 평균 적중률(%) | Evaluation set 평균 적중률(%) | 에러율 편차(%) |
|---|---|---|---|
| 사례기반 추론 | 71.10 | 66.87 | 4.23 |
| Percent Rank 적용 코드배열기반 추론 | 70.77 | 64.24 | 6.53 |
| K-Means 적용 코드배열기반 추론 | 67.65 | 59.19 | 8.46 |

# V. 결 론

본 연구에서는 목적 변수를 설명하는 각 속성 변수들을 조절하는 방식의 추론을 하는 인간의 의사결정 과정을 모방하는 코드배열기반 추론이라는 새로운 추론 기법을 제시하였으며, 코드배열기반 추론의 현실적 적용 가능성을 검증하기 위해 원시 데이터에 대해 사례기반 추론을 실시한 결과와 데이터 정규화를 통해 원시 데이터를 가공 처리한 데이터에 대해 코드배열기반 추론을 실시하고 그 결과를 비교하였다. 비교 결과, Percent Rank를 적용한 코드배열기반 추론이 사례기반 추론에 비하여 적중률에 있어서 2.63% 정도 약간 낮게 측정 되는 것으로 나타났다. 그러나 해상도의 조정을 통해서 결정을 최적화하는 인간의 의사결정 처리 과정과 더 유사한 특성을 가지고 있는 코드배열기반 추론이 기존의 사례기반 추론에 비해 성능적인 면에서 크게 손색이 없다는 것은 이 새로운 추론 기법의 현실적 적용 가능성을 확인하여 준 것이라 할 수 있다.

본 연구에서 제시한 코드배열기반 추론은 다음과 같은 특징을 가지고 있다. 먼저, 추론을 실행함에 있어서 해상도 조정은 적중률의 변화를 가져오기 때문에, 이를 통해 데이터 셋의 적정 해상도를 찾아 낼 수 있다. 또한, 본 코드배열기반 추론은 데이터 베이스에 직접적으로 기계적 학습법을 적용시켜서 별도의 정제과정을 거치지 않고 추론을 실행 할 수 있다는 장점을 지니고 있다. 일반적으로 데이터 베이스에 저장된 데이터는 범용적인 목적에 사용될 수 있는 형태로 저장되어 있기 때문에, 데이터를 사용하려는 데이터 마이닝 기법에 맞도록 가공하지 않고 있는 그대로 처리하는 것에 한계가 있다. 예를 들어, 사례기반 추론이나 인공신경망의 경우, 데이터를 추출한 뒤, 수작업 등과 같은 인간의 개입을 통해 별도의 정제 과정을 거친 다음에야 비로서 실행될 수 있다는 단점이 있다. 그러나 본 연구에서 소개하는 코드배열기반 추론에 사용되는 데이터의 경우는

데이터 정제의 자동화가 가능할 것으로 판단된다. 그리고 인공신경망은 Trial and Error 프로세스 방식이기 때문에 문제 해결을 위한 최상의 시스템을 구축하기 위해서는 모델을 구성하고 결과를 이끌어내는 과정까지 느리게 진행될 수 밖에 없고, CBR의 경우 데이터의 중복 문제와 에러 의 교정, 그리고 데이터 정제가 주로 수작업을 통해 이루어지기 때문에 이 부분에서의 속도 저하가 존재할 수 밖에 없다[Kim et al., 2004]. 그러나 코드배열기반 추론은 데이터 정제 및 정규화 과정의 자동화가 가능하기 때문에 이러한 부분의 자동화가 이루어 진다면 추론 소요 시간이 기존의 인공신경망, 사례기반 추론에 비해 크게 감소될 수 있을 것으로 예측된다. 또한 현재의 수작업을 통해 적정 해상도를 찾는 방식을 대체할 수 있는 알고리즘이 개발된다면 추론 전 과정의 자동화가 가능해지기 때문에 추론 시간이 더욱 획기적으로 개선될 수 있을 것이다. 아울러 본 코드배열기반 추론은 해상도의 개념을 적용시켰기 때문에 인간이 패턴을 인지하는 방식이 데이터 자체에 담겨져 있는 것으로 볼 수 있다. 그리고, 코드배열 기반 추론은 원시 데이터를 변환한 후 원시 데이터와 분리된 데이터인 변환 데이터를 사용함으로써 원시 데이터에 대한 보안성을 향상 시킬 수 있다. 특히 본 연구는 인간의 유전 정보를 담고 있는 DNA의 기본 정보 단위인 ACGT(Adenine, Cytosine, Guanine, Thymine)와 같이, 기업의 기본 정보를 처리하는 단위로 AGCT를 사용함으로써 생물체와 같이 역동성을 지닌 조직의 정보를 분석함에 있어 인간의 유전자 정보를 분석하는 분야인 바이오 인포메틱스 분야의 기술을 사용할 수 있는 가능성을 열어 보고자 하였다. 바이오 인포메틱스 분야에서는 단백질 구조의 정보학적인 표현 및 단백질 정보를 제공하고 있는 데이터베이스, 그리고 단백질 구조를 비교하기 위한 알고리즘 등을 연구하고 있다[Mohammed and Wang, 2003; Gary and Corne, 2003].

한편 본 연구에서 제시한 코드배열기반 추론의 활용을 위해서는 최상의 해상도를 결정할 수 있는 알고리즘의 개발이 필요한 것으로 나타났다. 본 연구에서는 수작업을 통하여 각 해상도 별 데이터 Set을 비교하는 것에서 그쳤으나, 가능한 모든 해상도를 고려하기 위해서는 코드화한 모든 레코드를 고려해야만 하는데, 예를 들어 본 연구에서 사용된 실험 데이터에 대해 모든 속성을 코드화하면 총 31개의 변수에 대하여 각각 네 개의 코드(A, C, G, T)를 할당하게 되기 때문에 생성되는 레코드의 가지 수는 $4^{31}$개가 존재 하게 되며, 이를 모두 고려해야 한다. 따라서 유전적 알고리즘이나 기계적 학습 방법을 이용하여 적정 해상도를 찾아내기 위한 알고리즘의 개발이 선행되어야 할 것이다. 그리고 이러한 알고리즘을 개발함에 있어서, Numeric 데이터보다 처리가 속도가 느린 Symbolic 데이터를 처리해야 한다는 것을 시스템의 성능적 측면에서 고려해야 할 것이다. 또한 본 연구는 데이터 정규화에 있어서 Percent Rank와 K-Means 두 가지 방식을 적용하였으나, 향후 연구에서는 이들을 대체할 새로운 데이터 정규화 방식들에 대한 제시가 필요할 것이다.

끝으로, 방대한 데이터베이스 및 데이터웨어하우스 내에 있는 데이터를 처리, 분석하여 새로운 지식을 발견하는 수단으로 사용되는 데이터마이닝은 산업체 및 학계의 관심이 증대되고 있는 경영정보학 분야의 중요한 주제이며 특히 고객관계관리 (CRM) 분야에 있어서 그 활용도가 크게 증대하고 있는 바, 본 연구에서 소개한 코드배열기반 추론 기법은 바로 이러한 데이터마이닝의 기반 기술로 널리 활용되는 인공신경망, 사례 기반 추론 등의 새로운 대안이 될 수 있을 것으로 사료된다.

# 〈참 고 문 헌〉


[1] 김상운, *패턴 인식 및 학습*, 홍릉과학출판사, 2003.

[2] 이성환, *패턴 인식의 원리*, 홍릉과학출판사, 1994.

[3] Berry, M.J.A. and Linoff, G., *Data Mining Techniques for Marketing, Sales, and Customer Support*, John Wiley & Sons Inc., 1997.

[4] Berry, M.J.A. and Linoff, G., *Mastering Data Mining*, John Wiley & Sons Inc., 2000.

[5] Doak, J., *An Evaluation of Feature Selection Methods and Their Application to Computer Security*, Technical Report CES-92-18, Department of Computer Science, Davis, CA., University of California, 1992.

[6] Duda, R.O., Hart, P.E. and Stock D.G., *Pattern Classification*, 2nd ed, John Wiley & Sons, 2001.

[7] Gary, B.F. and Corne, D.W., "Computational intelligence in bioinformatics," *Biosystems*, Vol. 72, Issues 1-2, November 2003, pp. 1-4.

[8] Gose, E., Johnsonbaugh R. and Jost, S., *Pattern Recognition and Image Analysis*, Prentice Hall, 1996.

[9] Kim, G.H, An, S.H. and Kang, K.I., "Comparison of Construction Cost Estimating Models Based on Regression Analysis, Neural Networks, and Case-Based Reasoning," *Building and Environment*, in press, 2004.

[10] Zaki, M.J. and Wang, J.T.L., "Special issue on data management in bioinformatics," Information Systems, Vol. 28, Issue 4, June 2003, pp. 241-242.

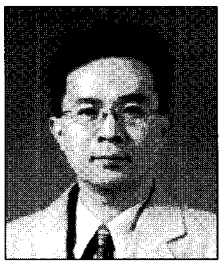
[11] Nelson, M.M. and Illingworth, W.T., *A Practical Guide to Neural Nets*, Addison-Wesley Inc., 1991.

[12] Pal, S.K., "Soft Data Mining, Computational Theory of Perceptions, and Rough-Fuzzy Approach," *Information Sciences*, Vol. 163, Issues 1-3, June 2004, pp. 5-12.

[13] Riesbeck, C.K. and Schank, R.L., *Inside Case- Based Reasoning*, Lawrence Erlbaum Associates, 1989.

[14] Watson, I., *Applying Case-Based Reasoning: Techniques for Enterprise System*, Morgan Kaufmann, 1997.

## ◆ 저자소개 ◆



강민철 (Kang, Mincheol)
현재 아주대학교 경영대학 e-비즈니스학부에 조교수로 재직 중이다. 한국항공대학교 항공전자공학과에서 학사(1984), 미국 뉴욕주립대(SUNY at Albany)에서 전산학 석사(1989), 미국 Rensselaer Polytechnic Institute(RPI)에서 공학박사(1996) 학위를 취득하였다. 삼보컴퓨터에서 연구원, 삼성SDS에서 경영컨설턴트, 그리고 계명대학교 경영학부 경영정보학 전공의 조교수로 근무한 바 있으며, 주요 연구분야는 e-Business, Multi-Agents, Computational Organization Theory 등 이다.

임호윤 (Im, Hoyoun)
현재 아주대학교 경영정보학 박사과정에 있으며, 협성대학교 e-Business학부에서 강의를 담당하였다. 아주대학교 경영학과에서 학사(2002), 아주대학교 경영정보학과에서 석사(2004) 학위를 취득하였다. 주요 연구분야는 인공지능(AI), 인터넷 마케팅 등이다.

◈ 이 논문은 2004년 4월 6일 접수하여 1차 수정을 거쳐 2004년 6월 8일 게재확정되었습니다.