碩士學位論文

의사결정트리 알고리즘을 이용한 학생 취업상황 예측 연구

濟州大學校 大學院

컴퓨터工學科

文 裕 亨

2013年 02月

의사결정트리 알고리즘을 이용한 학생 취업상황 예측 연구

指導敎授 金 度 縣

文 裕 亨

이 論文을 工學 碩士學位 論文으로 提出함

2013年 02月

文裕亨의 工學 碩士學位 論文을 認准함

審査委員長 _____ 印

委 員 _____ 印

委 員 _____ 印

濟州大學校 大學院

2013年 02月

A Study on Context Prediction of Student Employment Using Decision Tree Algorithm

You-Hyeong Moon

(Supervised by professor Do-Hyeun Kim)

A thesis submitted in partial fulfillment of the requirement for the degree of Master of Computer Engineering

2013. 02.

This thesis has been examined and approved.

Thesis director, _____

Thesis director, _____

Thesis director, _____

February 2013

Department of Computer Engineering

Graduate School

Jeju National University

감사의 글

기대와 걱정으로 시작했던 대학원 생활이 어느덧 2년이라는 시간이 흘러 졸업을 앞두고 있습니다. 학업과 사회생활을 병행 하는 일이 쉽지만은 않았지만 지도교수님을 비롯하여 주변의 많은 분들의 도움으로 잘 보냈던 것 같습니다. 많은 분들의 머릿속을 스쳐 지나지만 먼저 지난 2년간 대학원 생활을 함에 있어 아낌없는 가르침과 지도를 해주신 지도교수님이신 김도현 교수님께 감사드립니다. 또한 학업에 정진할 수 있도록 많은 애정과 가르침을 주신 김장형 교수님, 안기중 교수님, 곽호영 교수님, 이상준 교수님, 변상용 교수님, 송왕철 교수님, 변영철 교수님 정말 감사드립니다. 그리고 교수님과 학생들 뒤에서 묵묵히 도와주시는 학과 사무실 김남식, 오은희 선생님 지금은 다른 곳에 계시지만 항상 따뜻한 말 한마디로 힘을 주시던 이정하, 정은경 선생님 모두 감사드립니다.

대학에서 만난 나의 벗 김은철, 나의 후배들 양승철, 김소현, 김현복, 고봉찬, 유충협 등 모두 감사합니다. 지금은 각자의 위치에서 생활하고 있지만 다시 한번 다 같이 모일 수 있는 날이 있었으면 좋겠습니다. 그리고 지나온 세월만큼이나 깊은 우정의 친구들 강영하, 이창익, 박경택, 하건용 모두 고맙다. 그리고 모바일 컴퓨팅 연구실의 후배들 정재훈, 김주영, 홍영기 모두들 많이 도와줘서 고마워. 다들 원하는 곳에 취업하길 바랄게. 그리고 외국인 친구들 진서, 진남, 섭은, Safdar Ali, Rashid 많은 시간을 같이 보내지 못했지만 즐거웠고, 고향에 돌아가서도 이곳에서의 생활의 좋은 추억으로 남기를 기원합니다.

그밖에도 논문 쓴다고 많은 배려와 격려 주신 김선희 팀장님, 문성은 선생님, 나의 친구이자 동료 현진규, 조윤범 이외에도 정보통신원 직원 선생님들께도 감사드립니다.

끝으로 지금까지 끊임없는 지지로 저를 도와주신 아버지, 어머니, 누나, 매형, 형, 형수 그리고 조카 양문혁, 양예람 에게 사랑한다는 말을 전하고 싶습니다.

목 차

그 림 목 차

표 목 차

국문초록

의사결정트리 알고리즘을 이용한 학생 취업상황 예측 연구

컴퓨터공학과 문유형
지 도 교 수 김도현

최근 IT분야에서는 클라우드 컴퓨팅(Cloud Computing), 빅데이터(Big Data), 스마트 미디어 등이 정점화 되고 있다. 이러한 서비스나 시스템은 기본적으로 엄청난 양의 데이터를 처리하고 있으며, 이와 같은 대용량 데이터를 효과적으로 처리하기 위한 서비스 또는 기술에 대해 다양한 연구가 진행되고 있다. 학생정보를 관리하는 학사 시스템에서도 대용량 데이터를 수집, 저장, 조회하는 단순한 처리과정을 하고 있으나, 향후 인공지능이나 기계학습, 통계분석 등을 폭 넓게 사용하여 다양한 대용량 데이터에서 의미 있는 규칙이나 패턴 및 관계를 찾아내고, 실생활에 도움이 되는 데이터를 이용하여 지능적인 학사서비스 제공이 요구 되고 있다.

따라서 본 논문에서는 학사 시스템에서 의사결정트리(Decision Tree)알고리즘을 이용하여 학생들의 기존 취업 정보를 이용하여 취업가능 상태를 예측하는 알고리즘을 제안한다. 더불어 학생의 취업 정보를 기반으로 향후 취업을 위해 준비해야할 보완요소를 식별하여 추천하는 방안을 제시한다. 그리고 시뮬레이션을 통해 예측 알고리즘의 민감도(Sensitivity), 특이도(Specificity), 정밀도(Precision), 정확도(accuracy), 오류율(error ate) 에 대해 성능을 평가하고 제안된 알고리즘의 우수성을 검증한다. 이를 위해 세부적으로 본 논문에서는 취업정보를 정규화(Normalization) 하고 엔트로피(Entropy)를 이용하여 분류화(Classification) 한다. 이때 의사결정트리를 이용하여 기존취업 정보 기반의 트리를 형성하고 트리 기반의 취업 정보와 취업 여부를 결정하는 규칙(Rule)을 제공한다. 이 규칙을 이용하여 새로운 학생의 취업 정보기반의 취업 상황을 예측하는 방안을 제시하고 있다. 그리고 학생의 취업관련 (성별, 나이, 학점, 어학점수, 자격증유무, 어학연수

유무 등) 정보를 바탕으로 취업을 위한 보완요소를 식별하고 준비 하는데 도움되도록 한다. 이 연구를 통해 취업을 위한 기본적인 요건들을 의사결정트리 알고리즘을 통해 분석하여 일정한 규칙을 생성하여 취업상황을 예측함으로써 학생들의 취업을 체계적으로 지원하고, 취업과 채용 활성화에 기여할 것으로 사료된다.

ABSTRACT

A Study on Context Prediction of Student Employment Using Decision Tree Algorithm

Moon, You-Hyeong
Department of Computer Engineering
Graduate School
Jeju National University

Recently, Cloud Computing, Big Data and Smart Media become issues in IT field. We have researched on techniques to process a lot of data efficiently for supporting these service or system basically. Until now, existed university education system that manages student's information like processes, collection and storage of a lot of data simply. But in the future, we is required to find meaningful regularity, pattern or relation to lots of data using artificial intelligent or machine learning, statistical analysis, and provide intelligent education services.

Accordingly, in this paper, we proposes context prediction algorithm that estimate a possibility of finding a job using a decision tree, and student's established data of getting a employment. Also this paper presents complementary method that finds and recommends lack elements for student employment using student's information. And, we evaluate sensitivity and specificity, precision of context prediction algorithm, and verify performance of proposed algorithm through the simulation. In detail, we normalize student's data related employment for preprocessing student's data, and classify these

data using entropy and existed training data of student. So we make decision tree based on entropy information for generating a rule, and provide these tree information and rule for deciding possibility on getting a job. And we proposes predicts method which forecast context based new student's information for student employment using this rule. And, we support to identify student's possibility of employment and prepare lack elements based on these information (gender, age, grade, language score, certification, language study abroad, etc). Through this research, we can support to predict student's employment possibility using decision tree for getting a job effectively. And we can support systematically to get a job of students, and promote activation in recruitment field.

약어표

DT	Decision Tree
ID3	Interactive Dichotomizer ver.3
CHAID	Chisquared Automatic Interaction Detecton
CART	Classification & Regression Tree

Ⅰ. 서 론

1. 연구 배경 및 목적

최근 클라우드 시스템과 더불어 대용량 데이터에 대한 관심과 대용량 데이터를 통한 서비스 개발 등이 큰 화두가 되고 있다. 이에 본 논문에서는 향후 대학 시스템에서 양산 될 빅 데이터를 활용한 의미 있는 정보를 제공하고 동시에 학사 시스템의 지능화를 목적으로 데이터를 이용한 예측 알고리즘을 연구한다. 이를 위해 의사결정트리(Decision Tree) 기법을 이용하여 학생들의 정보를 활용하여 취업 상태를 예측하는 알고리즘을 연구하고, 학생들의 취업을 준비하고 보완할 수 있도록 취업상태에 따른 보완요소를 추천하여 취업률 향상에 기여를 목적으로 한다.

2. 연구 내용 및 방법

본 논문에서는 의사결정트리 알고리즘을 이용하여 학사시스템에서 얻어지는 대용량 데이터를 통한 지능형 서비스 제공을 목적으로 한다. 기존 대학 학사시스템은 데이터를 단순 조회, 저장, 삭제하여 관리하는 시스템 기능을 제공하였지만 향후 학사시스템에서 발생하게 되는 데이터를 통해 새로운 서비스 모델을 제시한다. 이를 위해 기존 취업 정보를 이용하여 데이터를 정규화(Normalization) 하고 엔트로피(entropy), 정보이득(Information Gain)을 계산 하여 의사결정트리(Decision Tree) 통해 데이터를 분류화(Classification)한다. 이렇게 분류화 된 데이터는 규칙(Rule)이 되며, 데이터베이스에 저장한다. 규칙을 이용하여 학생의 취업관련정보에 따른 취업 상태를 비교하여 예측 결과를 제공한다. 예측 결과에 따라서 취업을 위해 보완해야 할 요소를 제공하여 학생들이 취업을 대비 할 수 있도록 한다.

3. 논문 구성

서론에 이어 2장 의사결정트리와 알고리즘에서는 의사결정트리(Decision Tree)와 사용자 패턴 분류과 예측 관련 연구에 대해 알아본다. 3장에서는 취업 예측 처리과정(데이터 전처리 , 정규화, 분류화, 예측, 보완요소 추천)처리 구조와 구조에 대한 프로세스 흐름도와 예를 살펴보며, 데이터베이스 설계, 의사결정트리를 형성하는 트리생성 과정 등을 설명한다. 4장에서는 취업예측 알고리즘을 검증하기 위해 알고리즘의 민감도(Sensitivity), 특이도(Specificity), 정밀도(Precision)에 대해 성능을 평가하고 제안된 알고리즘의 우수성을 확인 한다. 마지막으로 5장에서 결론을 맺는다.

Ⅱ. 관련 연구

본 장에서는 대용량 데이터기반의 예측을 위하여 의사결정트리에 관하여 알아보며, 의사결정 트리 알고리즘(Decision Tree)의 종류와 방법들에 관하여 알아본다. 또한 기존의 예측관련 분야에서 적용된 예측 관련 연구를 살펴보고 알고리즘 또는 특성에 따라 연구된 관련연구에 대하여 살펴보며, 더불어 대표적인 예측알고리즘인 의사결정트리 알고리즘의 개념과 분리 방식 등에 대해서 고찰한다.

1. 예측 관련 연구 동향 분석

본 절에서는 사용자 패턴이 다양한 상황인식과 광범위하게 연구되어지고 있는 여러 가지 영역에서의 기술들을 고려하고 기술한다.

- 위치 예측
- 활동 예측
- 에너지 관리
- 건강 관리
- 비즈니스 관리

1) 위치 예측

사람의 다음 위치를 예측하는 것은 사람의 미래 상황을 예측할 수 있도록 하며, 상황을 준비하고 결과적으로 반응할 추가적인 시간을 제공할 수 있게 해준다. 위치예측은 위치 기반 서비스(Location Based Services)와 같은 지능형 컴퓨팅 어플리케이션을 위한 흥미로운 분야이다. 사용자의 다음 위치 예측은 그들의 현재 위치뿐만 아니라 그들의 미래 목적지와 관련된 서비스를 제공하도록 할 것이다. 이러한 방식으로 각 사용자는 특정 장소(식당, 박물관)와 관련된 정보를 인식할 수 있으며 그 곳에 도착한 직후 그 장소에서의 행동에 대해서 결정할 수 있다. 모바일 기기 자체로도 사용자의 미래 위치를 인식할 수 있으며, 사용자가 어딘가(컴퓨터, 전등, 히터)에 도착했을 때 그 장소와 상호작용하여 미래 위치에 대

한 정보를 준비한다.

Vintan 외 [1]은 사용자의 이동 예측을 위한 신경망 기반의 접근 방식을 제안했다. 이 논문에서는 사람의 다음 움직임을 예측하기 위한 신경 예측 기술을 제안한다. 저자는 선행학습을 유무에 따른 신경 예측장치들(역전파 학습을 가진 다층 퍼셉트론)에 초점을 맞춘다. 신경망의 최적의 구성은 건물 내에서 실제 사람들의 움직임의 순서들을 평가함으로써 결정된다.

Petzold 외 [2]은 스마트 홈에서의 실내 움직임을 예측하는 베이지안 네트워크를 제안했다. 이 논문에서 저자는 사람의 미래 위치와 그 위치에 머무를 기간, 새로운 장소로 이동할 때의 시간을 예측한다. 그림 2-1는 동적 베이지안 네트워크에 의해 형상화된 어플리케이션 시나리오를 보여주고 있다.

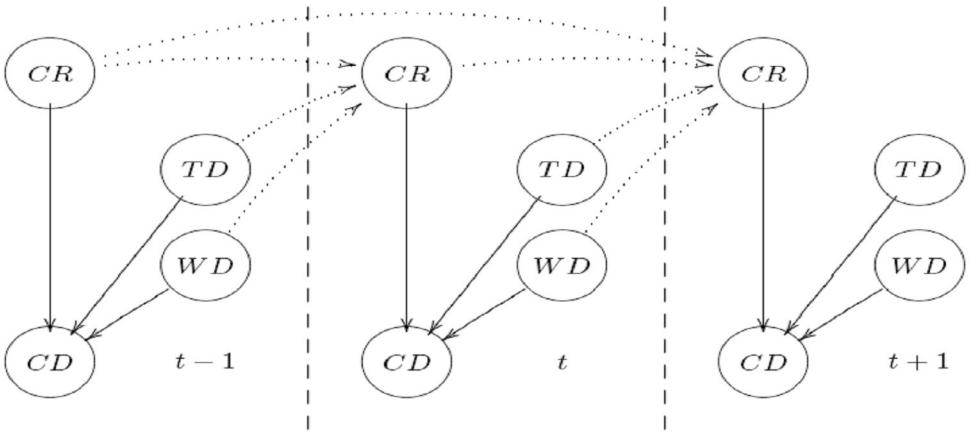


그림 2-1. 시분할 시간차 사이의 의존성이 있는 동적 베이지안 네트워크

이 네트워크는 t-1, t , t+1의 시분할을 잘 보여주지만, 실질적으로 과거나 미래의 시분할의 제한은 없다. 베이지안 네트워크는 시스템 안에서 각각의 사람들에게 할당되기 때문에, 사람들은 네트워크 변수로 나타나지 않는다. 각 시분할에서 현재 기간은 기본적으로 사람의 현재 공간에 달려있다. 현재 공간은 근본적으로 사람이 방문한 마지막 공간의 순서이다. 그래서 이전 시분할로 부터의 CR(Current Room)은 현재 시분할에서의 CR과 연결된다. 하루와 일주일간의 시간은 사람의 특정 행동 예측에 중요하다. 이러한 이유로 CD(Current Duration)는

현재 TD(Time of day)와 현재 WD(Weekday)에 밀접한 관계가 있다. 현재공간은 이 두 요인뿐만 아니라 이전 시분할에도 의존한다.

Christian Voigtmann외 [3]은 위치기반 상황예측 접근방법들에 대한 설문 조사를 제시한다. 이 논문에서 저자는 기존의 실내와 실외 위치 기반 상황예측 접근방법들에 대한 소개함. 저자는 사용자들 혹은 다른 객체들의 현재 위치는 상황예측이 시작된 이래로 연구들에 의해 대부분 빈번하게 사용 되어온 상황들이기 때문에 다음 위치예측 접근 방법에 초점을 두었다. 제시된 접근 방법들은 각각 다른 관점으로 평가되었다. 적용된 관점들의 차이는 저자가 예측 접근 방법의 평가를 위해 사용했던 일련의 데이터와 관련이 있다. 수집된 데이터의 결과로 만들어진 관점들에 대한 분석과 제안된 접근 방법들은 대부분의 경우에 공개적으로 이용 가능하도록 만들어지지 않는다. 그러므로 위치기반 상황 예측에 관심이 있는 새로운 연구자들이 제시된 결과를 평가하고 그와 필적할 만한 그들 자신의 결과물들을 만드는 것은 쉬운 것이 아니다. 그 후에 저자는 4가지 서로 다른 일련의 데이터를 이용하여 상황 예측 접근 방법들(Active Le Zi, Alignment, CCP)의 세 가지를 잘 알려진 기계학습 알고리즘(Bayes Net, Decission Table, J48 Tree) 세 가지와 비교한다.

2) 활동 예측

Nazefard 외[4]은 활동에 상응하는 특성들을 예측하기 위해서 2단계 프로세스에 베이지안 네트워크를 사용한 순서 기반의 활동 예측 접근방법을 제시한다. 제안된 모델뿐만 아니라 저자는 여러 검색 및 점수(S&S)와 제약 기반(CB) 베이지안 구조 학습 알고리즘의 결과를 제시한다. 이 논문에서 활동 예측 성능은 naive Bayes와 앞서 언급한 S&S, CB알고리즘과 비교된다.

실험 결과는 스마트 홈으로부터 다섯달 동안 수집된 실제 데이터에 대하여 수행되었다. 결과는 언급되었던 베이지안 네트워크 구조학습 알고리즘의 우수한 활동 예측 정확성을 말해준다. 그림 2-2는 제안된 베이지안 네트워크를 보여준다.

그림 2-2에서 색칠된 X_{t+1} 노드는 모델에 의해 예측될 클래스 변수(다음활동)를 의미한다. 이와 마찬가지로 노드 X_t 현재 발생 중이거나 완료된 현재 활동을

의미한다. 또한 노드 Y^i_t와Y^i_{t+1} (i=1,2and3) 는 현재와 다음 시간 상태를 나타낸다.

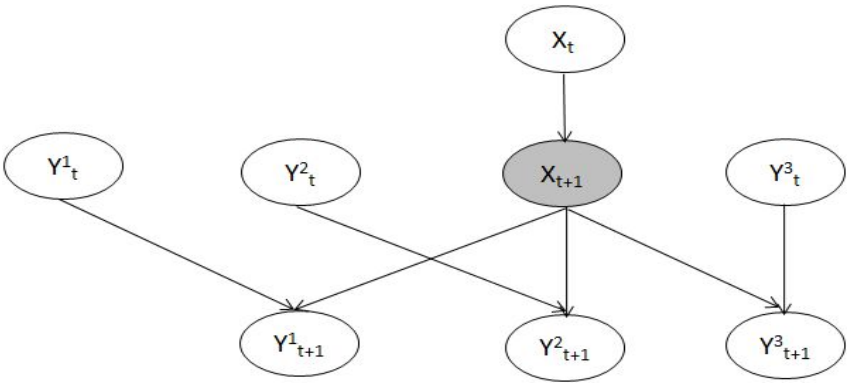


그림 2-2. 활동예측을 위해 제안된 베이지안 네트워크

활동 예측 문제에서 우리는 다음 활동에 대하여 알지 못하기 때문에, t+1시간에 대한 변수들은 처음 위치에서 모두 알려지지 않는다.

Parkka J 외[5]은 활동 분류에 대한 연구를 보여준다. 이 논문에서는 걷기, 뛰기, 자전거 타기와 같은 일상적인 활동의 분류를 위해 사용된 방법들이 설명되어 있다. 이 연구의 목적은 활동들을 어떻게 인식할 것인지, 어떤 감지장치들이 유용할지, 신호 처리와 분류의 종류는 무엇인지 알아내고 실제적인 센서의 데이터 라이브러리는 수집한다. 16명의 테스터들은 일상적인 환경에 기록된 35-채널 데이터와 약 31시간 추가 설명된 결과를 만든 데이터 수집에 참여한다. 테스터들은 2시간 측정 세션 동안 몇몇의 활동을 수행 하면서 착용 가능한 센서들을 착용한다.

Jaeyoung Yang 외[6]은 RFID를 기반으로 한 활동 인식 시스템을 제안한다. 이 논문에서 저자는 시간단위의 긴 시간부터 분단위의 짧은 시간 단위의 범위 동안 보여줄 수 있는 활동이론을 기반으로 한 간단한 접근 방법을 제안한다. 이 접근 방법은 계산 복잡성을 제한하면서 활동에 관련된 객체들에 대한 정보를 이용하여 활동들을 정확하게 인식할 수 있다. 저자는 또한 정확성을 증명하기 위하여 접근 방법들과 현재 실험적인 결과들을 설명한다. 이 논문에서는 활동이론은 적절한 계산적인 구조의 생성에 의해 운영된다.(활동이론은 좋은 계산적인 속성들을 가

지고 있음에도 불구하고 일반적으로 계산에 바로 사용되지 않는다.) 원시센서 데이터는 위치, 시간, 사람 행동에 의해 영향을 받은 객체들, 정보의 손실 없이 감지된 데이터의 표현을 단순화하기 위하여 사용할 준비가 된 활동이론과 같은 요소들과 밀접하게 연관된 의미 있는 구문으로 변환된다. 패널티가 주어진 naive Bayes 분류자가 소개된다. 분류자의 성능은 비분류의 가능성을 줄이기 위한 측면에서 패널티 함수 사용에 대한 보상이다. 운영되는 활동이론과 함께 패널티가 주어진 naive Bayes 분류자를 사용하여 높은 정확성을 가진 가벼운 분류자를 얻는다.

Emmanuel Tapia 외[7]은 집안에서 간단하고 지능적인 센서들 기반의 활동인식을 제안한다. 이 논문에서 설명된 작업은 지능적이고 간단한 감지 장치들이 실제 가정에서 일상생활의 활동을 인식하기 위하여 사용될 수 있다는 것을 보여준다. 활동인식을 위한 센서시스템들과 달리 이 작업에서 개발된 시스템은 실제 거주자와 함께 여러 주거환경에 배치되었다. 거주자들은 연구원이 아니거나 관련 경험이 없다. 게다가 제안된 감지 시스템은 카메라와 마이크로폰과 같은 많은 사람들에게 급속도로 퍼진 센서에 대안을 제시한다. 마지막으로 시스템은 주요 수정이나 손상 없이 기존 홈 환경에서 쉽게 새로 장착될 수 있도록 한다.

표 2-1. 특징계산과 평가

Feature description	Example
exist(sensorA, start, end)	Sensor A fires within time interval
before(sensorA, sensorB, end)	Sensor A fires before sensor B within time interval
before(sensorTypeA, sensorTypeB, start, end)	Sensor in a drawer fires before a sensor in the fridge within time interval
before(sensorLocationA, sensorLocationB, start, end)	Sensor in kitchen fires before sensor in bathroom within time interval

표 2-1는 특성 설명과 그에 상응하는 예들을 보여준다. 예를 들어 “exist(sensor A, start, end)” 규칙은 센서 A가 start와 end 사이의 시간에 경고음을 울리는 것

을 의미한다. 두 번째 규칙에서는 상응하는 기능성에 따라 센서들의 순서를 따르는 프로시저를 나타낸다.

3) 에너지 관리

에너지 관리는 스마트 홈과 스마트 빌딩, 스마트 공장과 스마트 오피스 등을 포함한 여러 환경에서 많은 연구자들에게 흥미로운 분야로 남아있다. 이에 논문에서는 위에 환경들을 포함하는 몇 가지 관련연구들을 기술하였으며, 목적은 어떤 종류의 사용자 패턴들이 에너지 효율 환경에서 고려될 수 있는지를 기술 한다.

Chao Chen 외[8]는 사용자의 행동들을 이용하여 스마트 홈에서의 에너지 예측 매커니즘을 제시 했다. 이 논문에서 저자는 집 에너지 소비의 분배를 분석하여 에너지 사용을 예측하기 위해 인간행동과 시간, 척도등의 특징들이라고 알려진 선형과 비선형의 회귀 학습모형을 제시한다. 제안된 방법들의 타당성을 보장하기 위하여 3달 넘게 수집된 2개의 실세계 정보데이터는 모형을 만드는데 적용 되었다. 학습 모형들을 기반으로 된 웹기반 최종소비자 시스템은 행동변화를 통해 에너지 효율성과 지속가능성을 촉진시키기 위하여 행동기반의 에너지 사용에 대한 피드백을 사용자에게 주기 위하여 개발된다. 저자는 에너지 사용 예측을 위해 선형 회귀와 SVM회귀를 사용했다. 표 2-2은 이 논문에서 사용된 특성과 그것들의 설명이 나타난다.

표 2-2. 예측과 회귀를 위한 데이터 특징

Feature Name	Description
하루의 길이	하나의 인스턴스(초)에서 발생 하는 때 자정 이후 시간 길이
주중 요일	주중(월, 화, 수, 목, 금, 토 일요일)의 현재 날짜
평일/주말	현재 날이 평일 인지 주말 인지 결정 하는 이진 변수
하루 중 시간	서로 다른 타임슬롯(아침, 정오, 오후, 저녁, 밤, 늦은 밤)
개인 센서의 시간	시간 창 동안 다른 활성화 된 모션 센서의 시간
모션 센서 종류	다양한 객실에서 발생된 모션 센서의 수
모션 센서의 총 수	시간 창 동안 실행된 모션 센서 이벤트의 총 수

Christian Beckel 외[9]는 전기 소비 데이터를 기반으로 가전제품의 자동 분류를 제시한다. 이 논문에서 저자는 디지털 전력량계에 의해 수집된 정보를 사용하여 개인 가전제품들의 자동분류를 시행하는 것의 문제점을 역설했다. 특히 찾기에 적절하고 유망할 수 있는 일련의 가전제품들을 인식하고, 우선 세 가지 다른 에너지 공급업체의 노동자들과의 심오한 인터뷰를 함으로써 일련의 적절한 특징들을 이끌어 낸 다음 자아 조정 맵을 사용하여 많은 양의 전기 소비 흔적 데이터를 분석한다. 분석은 일반적인 분류 방식을 이용하여 얻은 전기 소비 데이터로부터 추론하기 쉬운 일련의 가전제품 특징들이 존재함을 보여준다. 그 결과들은 가전제품의 크기와 거주자의 수입과 같은 특징들을 전기 소비 데이터로부터 매우 쉽게 추론할 수 있을 뿐만 아니라 에너지 공급업체에게 유용할 수 있다는 것을 보여준다. 그러한 특징들의 인식은 전기 소비 데이터를 이용하여 개인의 가전제품들의 자동분류 잠재성의 분석을 향한 필수적인 첫 번째 걸음이 된다는 것을 나타낸다.

S.A. Pour mousavi 외[10]은 매우 단기적인 바람 속도 예측을 위한 혼합 모형을 제시한다. 이 연구의 목적으로서 인공 신경망(ANN)과 마르코프 연쇄(MC)는 매우 단기적인 시간 척도에 바람 속도를 예측하기 위하여 새로운 ANN-MC 모형을 개발하기 위해 사용된다. 미래에 짧은 시간에 매우 단기적인 바람 속도의 예측을 위해 현재 시간에 앞서 기록된 단기적인(약 1시간) 것과 매우 단기적인(약

몇 분 또는 몇 초) 것에 대한 데이터 패턴들이 고려된다. 이 연구에서 바람 속도 데이터에서의 단기적인 패턴은 ANN에 의해 기록되고 장기적인 패턴은 MC 접근 방법과 4가지 이웃한 지수들을 사용하여 고려된다. 그 결과들은 입증되고 새로운 ANN-MC 모형의 효율성이 보모형의 효율성이 보여진다. 예측의 불확실성과 계산시간이 감소되는 동안 예측오류들이 감소될 수 있다는 것이 발견된다. 그림 2-3은 제안된 혼합 모형을 보여준다.

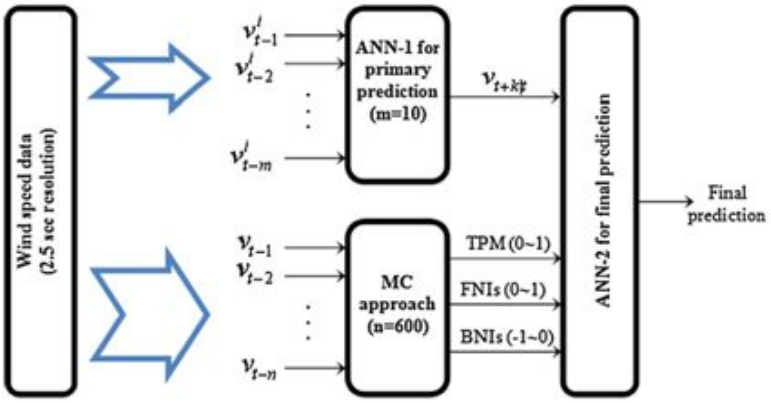


그림 2-3. 제안된 하이브리드 모델

Vladimir Cherkassky 외[11]은 상업용 건물에서 전력 소비의 예측을 제시한다. 이 논문은 전력 소비의 예측을 위하여 컴퓨터를 사용한 지능기술의 적용을 나타낸다. 제안된 접근 방법은 여러 가지 상업용 건물과 공공건물로부터 얻은 실제 생활 데이터를 이용해 시간(하루)과 온도의 함수로써 동력 소비의 예측 정확성을 향상시키기 위하여 회귀와 클러스터링 방법을 결합한다.

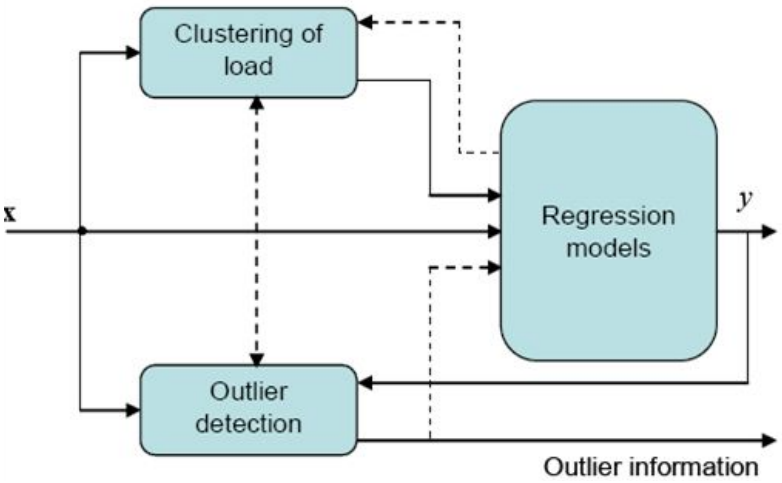


그림 2-4. 세 가지 데이터 모델링 작업 간의 관계

그림 2-4에서 보여지는 데이터 분석 시스템에서 회귀모델화는 사용유무(하루 중의) 기간 동안 회귀모형을 추정하기 위하여 클러스터링 모듈로부터 얻어진 나눠진 훈련 데이터를 사용하고, 분리물 감지는 이러한 추정된 회귀모형을 필요로 한다. 클러스터링(데이터 분할)과 회귀모델화의 작업이 훈련 데이터를 사용하여 수행되는 반면 분리물 감지 작업은 시험 데이터를 사용하여 작동(예측)하는 동안 수행된다. 그림 2-4에서 파선은 클러스터링회귀 추정과 분리물 감지 클러스터링이 상호간에 의존적인 방식으로 수행되는 미래에 더 발전된 시스템에서의 정보흐름을 나타낸다.

A. Badri 외[12]는 단기 부하예측을 위해 인공 신경망과 퍼지논리방법들의 적용을 제안한다. 저자에 따르면 정확한 부하예보는 전기 회사들이 장치운용, 발전과 유지계획의 관점에서 가장 좋은 결정을 내리는 데 큰 도움이 된다. 전기 발전회사들은 큰 정확성을 가지고 미래수요에 대한 사전 지식을 가지는 것이 필요하다. 몇몇 데이터 마이닝 알고리즘은 부하예보를 예측하는 것에 커다란 역할을 수행한다. 이 논문은 단기 범주의 load수요를 예측하기 위한 예보도구로서 인공 신경망과 퍼지논리의 적용을 검사한다. 이러한 경우에 예보는 하루 앞서게 되고 ANN은 FL과 비교해서 좀 더 정확한 결과를 나타낸다.

이 논문은 단기부하예보를 위한 신경망과 퍼지논리 두 개의 현대적 방법 사이

에서 수행되어온 비교를 통한 연구를 제공한다. 저자는 시뮬레이션 결과로부터 ANN 모형이 FL모형보다 상당히 나은 결과를 생산한다는 것이 발견된다고 설명한다. 이것은 같은 한도를 가지고서 FL 방법이 정확하고 포괄적인 선형회귀곡선을 찾는 것이 어렵다는 사실 때문이다. 저자는 제시된 모형이 체계적으로 검사되지 않을 것이며 검사의 결과는 완전히 만족스러울 수준으로 제공되지 않을 것이라고 설명한다. 다른 연구처럼 중단기의 load 예보를 위한 ANN의 적용이 제시되었다. 단기적인 load 데이터의 정확성과 load 발생시간에 가까움 때문에 단기적인 Load예보는 중단기의 LF와 비교하여 좀 더 정확한 결과를 제공한다. 저자는 훨씬 더 나은 결과를 얻기 위해서는 시작날과 다른 날들을 식별할 수 있는 신경망을 위해 더 정교한 토폴로지를 소유할 필요가 있을지도 모른다고 결론을 내린다. 이러한 작업을 위하여 저자는 다른 날씨정보들 중에 오직 온도만을 사용한다.

Ren'e Schumann 외[13]은 스마트 그리드에 블록을 만드는 것으로서 수요 예측와 스마트 장치를 제시한다. 사용자 수요의 예측 도입이 스마트 그리드에서의 에너지 비용을 감소시키는 데 도움을 준다는 것이다. 저자는 마이크로 그리드에서의 스마트 장치의 삽입의 영향을 연구해 왔다. 마이크로 스마트 그리드는 에너지 비용을 최소화하기 위해 예측, 계획, 조정과 같은 기술들을 응용할 수 있다. 마이크로 그리드 소비 관리의 몇 가지 근본적인 형태는 사물 인터넷(M2M)의 상황에서 현재 연구되는 기술들이다. 장치들은 지역 컴퓨팅과의 의사소통 능력을 더 함으로써 스마트해진다. 저자는 a) 마이크로 그리드에서의 스마트 장치의 양을 달라지게 하는 것과 b) 에너지 발전계획을 위하여 예측 능력을 더함으로써 에너지 비용을 감소시킬 가능성을 연구한다. 이러한 스마트 마이크로 그리드의 배열형태와 에너지 비용의 최소화를 분명히 소비자에게 가져다 줄 것 이라고 한다. 실험에 근거하여 지역 에너지 생산 계획에 예측 정보를 더하는 것은 그리드의 나머지가 다른 스마트 장치들을 소유하지 않는다 할지라도 상당한 비용 감소를 이끌 수 있다고 한다. 또한 더 많은 스마트 장치 / 소비자를 마이크로 그리드에 더함으로써 에너지와 비용이 더욱 감소될 수 있다고 한다.

4) 건강 관리

Smitha T 외[14]은 질병 예측에 기반을 둔 의사 결정트리를 제안한다. 사용 중인 의사 결정트리는 데이터 마이닝 응용에 관한 규칙을 발견하기에 적합하도록 해주는 특정한 이점들을 소유한 학습 알고리즘들 중의 하나이다. 주로 어떤 지역, 특히 슬럼에서의 질병의 발생 가능성을 예측하기 위해 의사 결정트리를 사용하여 예측 모형을 만드는 목적을 가졌던 데이터 마이닝 기술에 초점을 둔다. 이 모형은 또한 모형 형성에 도움이 될 수도 있는 다른 중요한 인자(파라미터)를 인식한다. 이 논문에서 의사 결정트리는 질병이 발생할 가능성에 근거한 지역의 거주자들을 분류하는데 적용되고 있다. 이 논문은 의사 결정트리알고리즘을 이용하여 질병 발생에 관한 규칙을 발견하고자 한다. 논문은 또한 미래 예측을 위해 이 지역에서 어떤 규칙들이 발생되는지 연구한다.

표 2-3. 가능성 있는 값들에 대한 몇 가지 특성

No	Attribute	Possible Values
1	Income	Low/Average/High
2	Employment	Employed/Un-Employed
3	Environmental Condition	Good/Average/Poor
4	No of Members in a Family	Below 5/Between 6-10/Above 10
5	Sanitation Facility	Good/Fair/Poor
6	Education	Educated/Non-Educated

Osmar R. Zaiane 외[15]는 분류 시스템을 만들기 위해 새로운 방법을 제안한다. 연관 규칙 마이닝에 기반을 두고 의학적 이미지들을 분류하기 위한 응용에 실제의 데이터를 시험한다. 이 논문에 제안된 방법은 디지털 유방암 검사를 3가지 범주로 분류한다. 보통 ,양성, 악성 보통의 유형은 건강한 환자를 특징으로 하는 것이며, 양성유형은 종양을 보여주는 유방암 검사를 대표하지만 그 종양은 암세포에 의해 형성되지 않으며 악성유형은 암적인 종양을 가진 환자로부터 만들어진 유방암 검사이다. 이 연관 규칙에 기반을 둔 분급기는 실제의 데이터모음에서

실험된다. 이 논문에서 제시된 시스템은 다음과 같은 단계로 구성 되어있다. 전처리 단계, 처리된 업무의 데이터베이스를 마이닝 하는 단계, 그리고 분류모형에서의 처리된 연관 규칙을 정리하기 위한 마지막 단계로 구성된다.

Mantzaris D.H. 외[16]은 의학적인 질병 예측을 위해 인공적인 신경망을 제안한다. 이 연구는 다층 퍼셉트론과 확률론적 신경망뿐만 아니라 골다공증 위험 요소 예측의 문제에 이러한 모형들의 적용을 기반으로 한 인공 신경망 패턴 인식 모형들의 발전과 평가에 초점을 둔다. 더 정확히 말하자면 어떤 사람이 높은 수준의 골다공증에 있으며 그러므로 더 많은 골다공증 검사를 받아야 하는지를 인식하는 임상의들을 도와주기 위함이다. 이 응용 분야는 초기 골다공증의 발견이 골다공증으로 인한 골절의 예방에 중요하기 때문에 매우 중요하다고 여겨지는데, 이것은 증가된 질병률과 사망자 수, 높은 사회경제적 비용과 관련이 있다. 제안된 인공 신경망 구조들과 임상 데이터에서의 그것들의 활용이 이 논문에 제시된다.

Hani Neuvirth 외[17]은 만성 질환을 가진 환자에 대해 데이터 구동 평가 시스템을 위한 프로토타입과 환자에게 최적의 관리를 할 수 있는 내과 의사를 지정하기 위한 새로운 응용을 제시한다. 프로토타입은 당뇨병의 경우 사용하기 위해 고안되고, 대략 4500명의 당뇨병 환자들로부터 얻은 일상적인 수술 데이터를 통해 평가된다. 이 연구에서 환자의 결과를 위해 두 가지 측정법을 탐구했다. 측정법은 당뇨병 환자를 위한 치료 질의 평가에 필수적이고 분석적인 체계 중 두 가지 유형에 영향을 받는다. 분류와 생존 분석 프로토타입 시스템은 최종소비자가 상호적으로 바라던 결과를 선택할 수 있게 해준다. Cox 비례 위험 모형, 종적인 데이터를 분석하는데 일반적으로 사용되는 통계적 방법, 그리고 여러 가지의 최신식 기계 학습 방법론 등을 포함한 다양한 학습 접근 방식을 시험하고 비교한다.

5) 비즈니스 관리

Hyunchul Ahn 외[18]은 기업의 파산 예측을 위하여 CBR(Case based reasoning)의 예측 수행을 향상시키기 위한 새로운 접근 방식을 제안한다. 저자는 유전자 알고리즘을 사용함으로써 사례 기반의 추론(CBR)을 위한 동시적인 특징 가중치의 최적화와 사례 선택을 제안한다. 저자에 따르면 금융상에서 가장 중요

한 연구 주제 중 한 가지가 효과적인 회사의 파산 예측모형들을 만드는 것이다. 왜냐하면 그것은 금융기관의 위기 관리를 위해서 필수적이기 때문이다. 연구자들은 통계적인 지능기술과 인공적인 지능기술을 포함한 예측 수행을 향상시키기 위해 다양한 데이터 구동접근 방식을 적용해 왔고, 그것들 중에 상당수가 유용하다고 증명되어 왔다. 사례 기반의 추론은 가장 인기 있는 데이터 구동접근 방식 중에 하나이다. 왜냐하면 적용하는 것이 쉽고 과잉적합(over fitting)의 가능성도 없으며, 산출된 것에 적절한 설명을 제공하기 때문이다. 그러나 그것은 중요한 한계를 가지고 있으며 예측 수행이 일반적으로 낮다. 저자는 그들의 모형이 더 관계있는 사례를 참고하고 필요 없는 것을 제거함으로써 예측 수행을 향상시킬 수 있다고 주장한다. 제안된 모형을 실세계의 사례에 적용. 실험의 결과는 전통적인사례 기반의 추론의 예측 정확성이 제안된 모형을 사용함으로써 상당히 향상될지도 모른다는 것을 보여준다. 저자는 금융기관들이 이러한 결과물들에 적절한 설명뿐만 아니라 정확한 결과를 만들어내는 파산 예측모형을 만드는 방법을 제안 한다

2. 의사결정트리 알고리즘

본 논문에서 예측을 위한 알고리즘으로 대표적인 방법 중 하나인 의사결정트리를 사용되고 있다. 의사결정트리 모형은 1964년 Sonquist와 Morgan에 의해 처음 시도되었고, 1973년 Morgn 과 Messaenger에 의해 THAID(THeta Automatic Interaction Detection)라는 알고리즘으로 일반인도 많이 이용하게 되었다. 1980년 Kass는 카이제곱 적합성검정에 근거한 CHAID(Chisquared Automatic Interaction Detecton)라는 알고리즘을 소개하였는데 현재까지 많이 사용되고 있다. 1982년 컴퓨터 학자 Quinlan은 ID3(Interactive Dichotomizer Ver.3)라고 하는 의사결정트리 알고리즘을 소개하였고, 후에 이를 발전시켜 C4.5라는 알고리즘을 구현 한다. 1984년 Breiman등은 CART(Classification & Regression Tree)를 통해 의사결정트리의 성장 및 가지치기 등의 이론을 정립한다. C4.5와 CART는 비모수적인 분류 방법인데 1988년 Loh와 Vanichetake는 모수적인 접근 방식의 FACT를 소개한다. 최근에는 하나의 의사결정트리로서 데이터를 분류하는 방법

보다 붓스트랩(bootstrap) 방법으로 표본을 여러 개 추출한 후 이 표본에 근거한 여러 개의 의사 결정트리 분류를 통합하는 앙상블(ensemble)모형을 많이 이용한다.[19]

의사결정트리는 데이터를 분류(Classification)하기 위해 사용하며, 목표 변수가 범주형인 경우 사용되며 수치형인 경우 결정트리 알고리즘 종류가 달라져야 한다.

의사결정트리 알고리즘 종류는 ID3 알고리즘, C4.5 알고리즘, C5.0 알고리즘, CART , CHAID 알고리즘이 있으며, 데이터 마이닝에서 가장 많이 사용되는 알고리즘은 C4.5 또는 C5.0이다. ID3알고리즘의 수치형 데이터를 분류 할 수 없어 이를 보완한 알고리즘이 C4.5가 개발되었으며 다시 이를 보완한 알고리즘이 C5.0이다. 이들 알고리즘들은 크게 인공지능, 기계학습 분야에서 발전된 ID3, C4.5, C5.0과 통계학 분야에서 개발된 CART, CHAID 알고리즘으로 분류 된다.

또한 서로 유사한 방식을 갖지만 평가지수 즉, 선택방법에 다른 접근 방식을 갖게 되며, 그 분류는 인공지능 계열의 알고리즘들은 엔트로피, 정보이득 개념을 사용하여 분리기준을 결정하고, 통계학에 기초한 CART 및 CHAID 알고리즘들은 카이스퀘어, T검정, F검정 등의 통계분석법을 사용한다.

결정트리 알고리즘들은 기본적인 생성 방식은 유사하며 가지를 분리하는 방식(분리에 사용될 변수 및 기준을 선택하는 방식)에서의 약간의 차이를 갖는다. 분리 방식의 차이점을 아래의 표 2-4로 설명한다. [20]

표 2-4. 의사결정트리 알고리즘 분리방식

알고리즘	평가지수(선택방법)	비고
ID3	엔트로피	다지분리(범주)
C4.5	정보이득	다지분리(범주) 및 이진분리(수치)
C5.0	정보이득	C4.5와 거의 유사
CHAID	카이제곱(범주), F검정(수치)	통계적 접근 방식
CART	Gini Index(범주), 분산의 차이(수치)	통계적 접근 방식, 항상 2진 분리

ID3 알고리즘은 의사결정트리 기반 분류 알고리즘의 대표적인 알고리즘으로서, 다양한 의사결정트리 기반의 분류 알고리즘인 C4.5, CART, CHAID 들도 ID3의 기초하고 있다. ID3알고리즘을 이해하기 위해 의사결정트리를 설명할 때 가장 많이 사용되는 예제로 날씨에 따라 경기를 했는지 안했는지를 기록한 데이터를 갖고 설명한다.

표 2-5. 날씨 및 기온기반의 경기 예측 예

Day	Outlook	Temp	Humidity	Windy	Play(Y/N)
D1	Sunny	Hot	High	Weak	N
D2	Sunny	Hot	High	Strong	N
D3	Overcast	Hot	High	Weak	Y
D4	Rain	Mild	High	Weak	Y
D5	Rain	Cool	Normal	Weak	Y
D6	Rain	Cool	Normal	Strong	N
D7	Overcast	Cool	Normal	Strong	Y
D8	Sunny	Mild	High	Weak	N
D9	Sunny	Cool	Normal	Weak	Y
D10	Rain	Mild	Normal	Weak	Y
D11	Sunny	Mild	Normal	Strong	Y
D12	Overcast	Mild	High	Strong	Y
D13	Overcast	Hot	Normal	Weak	Y
D14	Rain	Mild	High	Strong	N

표 2-5는 날씨에 따라 경기를 했는지 안했는지 기록된 데이터이며, 맨 위 데이터는 각 속성필드 이름이고 14개의 레코드가 존재한다.

ID3알고리즘의 구현을 위한 각 단계는 :

1. 전체 데이터를 포함하는 루트 노드를 생성한다.

2. 만약 샘플들이 모두 같은 클래스라면, 노드는 잎이 되고, 해당 클래스로 레이블을 부여한다.

3. 그렇지 않으면 정보이득이 높은(즉, 데이터를 가장 잘 구분할 수 있는) 속성을 선택한다. (이때 정보이득은 엔트로피의 변화를 가지고 계산한다.)

4. 선택된 속성으로 가지(Branch)를 뻗어 하위 노드들을 생성한다.

(각 하위 노드들은 가지의 조건을 만족하는 레코드들이다.)

5. 각 노드에 대해여 2단계로 이동한다.

위 5단계를 통해 구현되, 구현을 위해서는 정보이득을 구하며, 정보이득을 얻기 위해서는 엔트로피 값 계산이 필수적이며 값을 구하는 공식은 다음과 같다.

$$p_i = \frac{freq(C_i, S)}{|S|}, \quad Entropy(S) = -\sum_{i=1}^{m} p_i \log_2(p_i) \qquad \text{(식 1)}$$

S : 주어진 데이터들의 집합

C = {C1, C2, ... , Ck} : 클래스 값들의 집합

freq(Ci,S) : S에서 class Ci에 속하는 레코드의 수

|S| : 주어진 데이터들의 집합의 데이터 개수

위 식에서 는 i번째 클래스 값에 대하여 해당 데이터집합 중에서 차지하는 비율(Probability)을 의미한다.

엔트로피값을 구한 후 정보이득을 계산할 수 있다.

정보이득이란, 어떤 속성을 선택함으로 인해서 데이터를 더 잘 구분되게 하는 것을 의미하며 정보이득을 구하는 공식은 다음과 같다.

$$Gain(A) = I(s_1, s_2, s_3 \cdots s_m) - E(A) \qquad \text{(식 2)}$$

I : 상위 노드의 엔트로피

S : 데이터 속성값

E : A선택시 파생되는 엔트로피 가중치 평균값

I는 상위 노드의 엔트로피이며, E(A)는 A라는 속성을 선택 시 파생되는 하위노드의 엔트로피 가중치 평균값을 의미한다.

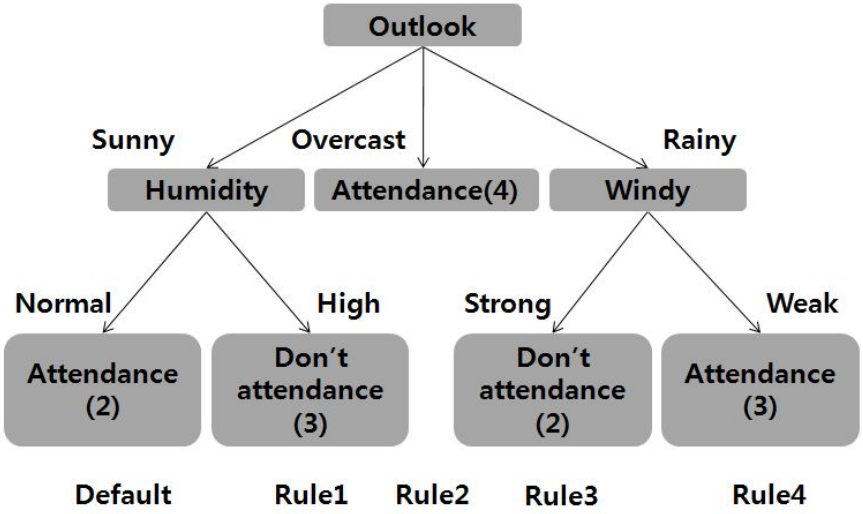


그림 2-5. 의사결정트리 구성

C4.5알고리즘은 ID3알고리즘을 보완하여 알고리즘이라 할 수 있으며, ID3알고리즘의 문제점은 다음과 같다.

(a) 수치형 속성 취급 (handling continuous attributes)
(b) 무의미한 속성을 제외하는 문제
(b) 트리의 깊이 문제 (how deeply to grow the decision tree)
(c) 결측치 처리 (Handling missing attributes values)
(d) 비용고려 (handling attributes with different costs)
(e) 효율성 (Improving computational efficiency)

ID3 알고리즘은 범주형 속성에 대해서만 트리를 생성하는 방법을 제시하고 있다. 따라서 수치형 속성은 모델 생성에 활용할 수 없는 한계가 있다. C4.5에서는 수치형 속성까지 사용하는 방법에 대해서 제안한다.

스마트 홈 환경에서 데이터 마이닝 기법을 이용하여 사용자에게 상황에 적합한 서비스를 추론하는 모델을 제안한다. 의사결정트리 알고리즘들 중에 하나인 C4.5 알고리즘을 기반으로 서비스 추론에 쓰이는 서비스 트리를 생성하고, 정량적 특성 규칙과 정량적 판별 규칙을 이용하는 정량적 가중치 산정 알고리즘을 통해 사

용자에게 제공될 서비스를 추론 한다. [21]

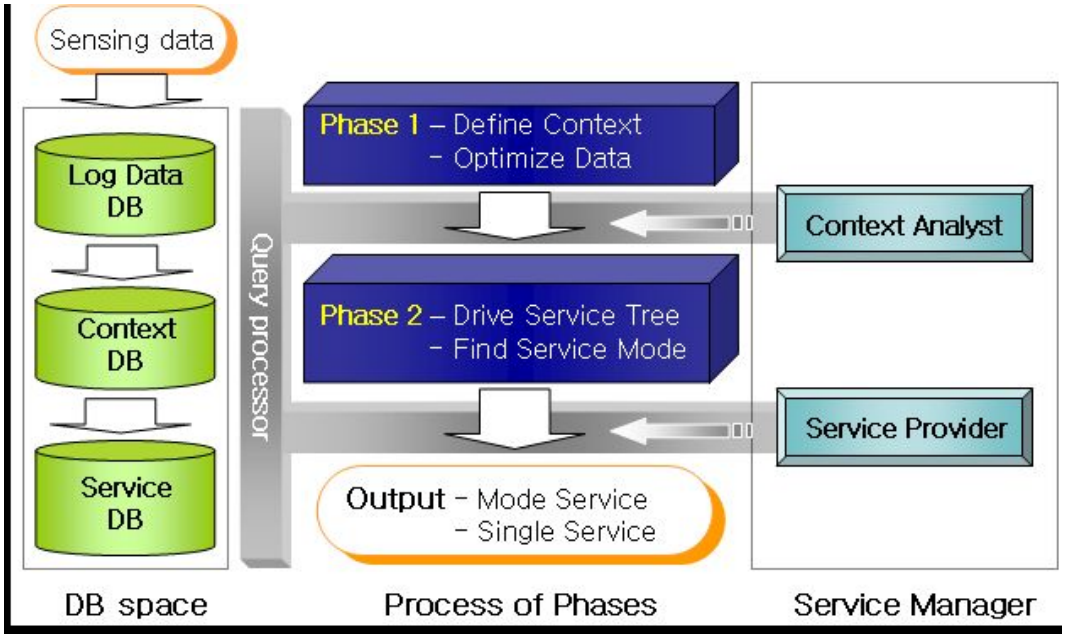


그림 2-6. ISR 모델

Ⅲ. 의사결정트리을 이용한 취업상황 예측알고리즘

1. 개요

본 장에서는 의사결정트리 알고리즘을 이용하여 규칙을 생성하여 취업상황을 예측하고 보완요소를 추천하기 위한 전체 프로세스 구조와 각 프로세스의 주요기능 및 각 프로세스의 특성들에 대해 설명한다. 또한 각 프로세스의 세부 흐름도(Flow Chart)를 통한 상세한 구조 설명과 흐름도이용한 실제 데이터의 흐름 예를 나타낸다. 구현된 프로세스 기능과 특성들의 설명에 이어서 알고리즘을 구현하기 위해 뿌리(Root)노드 생성과정, 하위노드 생성과정, 의사결정트리 생성 마지막으로 규칙생성 방법을 설명하고 알고리즘을 이용하여 생성된 규칙을 통해 예측결과를 제시함과 동시에 예측 결과에 따른 취업을 위한 보완요소 제공방안을 설명한다.

2. 취업상황 분류 및 예측알고리즘

1) 전체 시스템 구성 및 Process 구조

학생취업상황 예측 및 추천 전체 처리과정은 그림 3-1과 같다. 처리과정은 6가지 분류의 기능으로 나누어져 있으며, 각 기능은 첫 번째 데이터를 입력받고저장처리 해주는 데이터 처리(Data Process), 두 번째 데이터를 변환하여 예측 가능한 범주형 데이터로 변환시켜주는 정규화(Normalization Process), 세 번째는 범주형 데이터로 구성된 데이터를 기반으로 엔트로피, 정보이득 값을 구한 후 트리를 구성하여 데이터의 결과를 분류시켜주는 분류화(Classification Process), 그리고 분류되어진 데이터를 기반으로 예측을 위한 규칙생성 과정과 규칙을 기반으로 사용자 요청 데이터에 대한 예측을 처리하는 예측(예측 처리 과정) 끝으로 예측된 결과를 바탕으로 결과에 따라 사용자에게 보완요소를 추천해주는 보완요소추천프로세스(Complementary element Suggest Process) 로 구성 되어 있다.

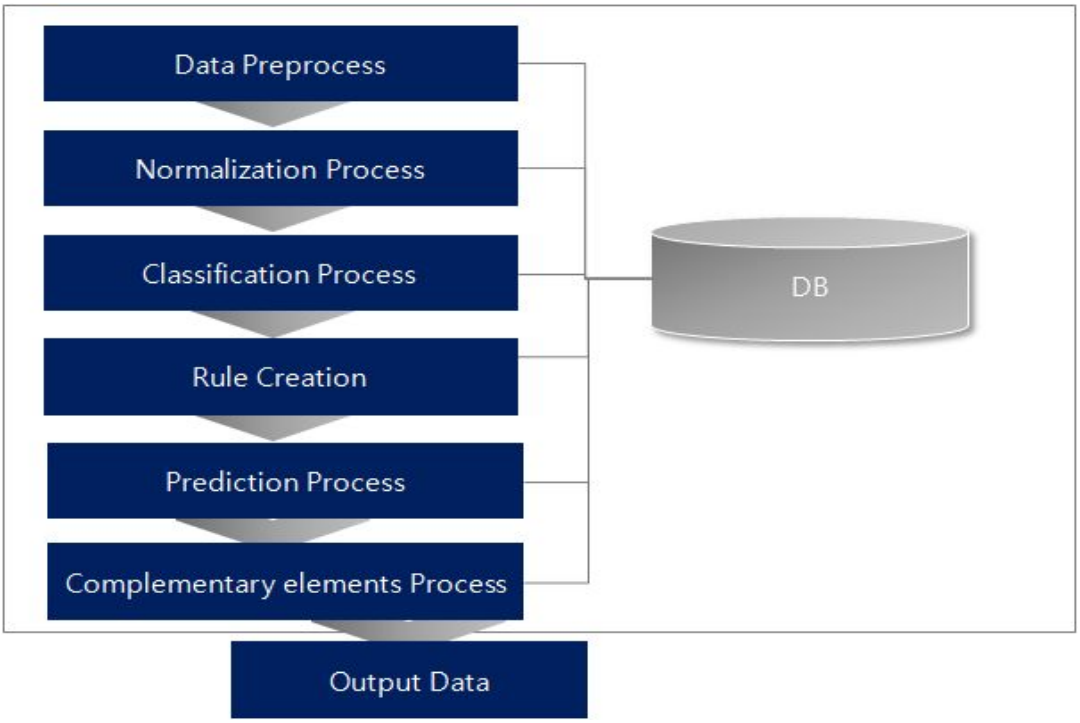


그림 3-1. 학생취업상황 예측 및 추천 전체 처리과정

제안하는 의사결정트리 알고리즘 적용을 통한 학생 취업예측을 위한 전체 처리과정은 그림 3-2와 같다. 학생의 취업 요청데이터를 받으면 시스템은 그림에서 Input Data가 되며, 요청된 데이터는 저장이 이루어진다. 데이터는 저장 과정 중에 분류화(Classification)을 위해 범주형 데이터로 전환되어지며 전환된 데이터를 기반으로 분류화(Classification)가 이루어진다. 분류된 데이터는 예측의 결과를 위한 규칙이 되며, 결과에 따른 보완요소의 추가적인 기능을 위해 사용된다. 데이터는 의사결정트리 C4.5의 평가지수 계산방법에 의해 엔트로피값을 우선 구하고 그 다음 정보이득값을 구하여 가장 높은 값을 트리노드로 생성하고 차례로 하위노드를 생성하여 leaf노드를 구하여 정지규칙이 만족 될 때까지 반복적으로 수행되어 구성된 트리는 Root노드에서 부터 Leaf노드까지의 하나의 줄기가 하나

의 규칙이 되며 이 데이터들은 규칙(Rule)데이터 테이블에 저장된다. 이러한 일련의 과정이 끝나면 사용자가 요청한 데이터를 규칙데이터와 비교하여 취업 가능유무를 확인할 수 있게 된다.

Request Data
성별 : 여 , 나이 : 24 , 학점 4.0, 어학 : 800 , 자격증 : N

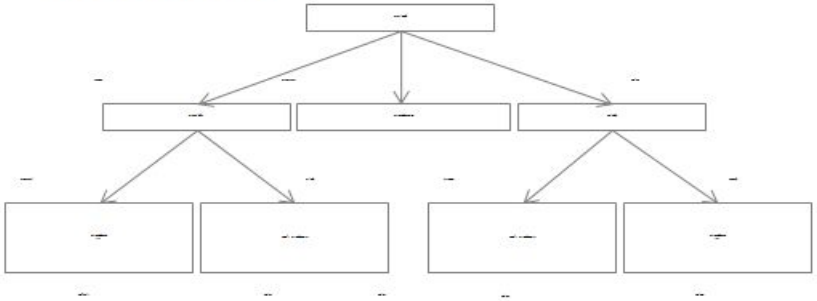
Input Data Table

성별	나이	학점	어학	자격증	취업
남	23	3.5	900	Y	Y
여	24	4.0	800	N	Y
남	25	4.2	650	N	Y
여	25	2.8	700	Y	N
. . 남	. . 26	. . 3.9	. . 450	. . Y	. . N

Normalization Data

성별	나이	학점	어학	자격증	취업
남	23	3	A	Y	Y
여	24	4	B	N	Y
남	25	4	D	N	Y
여	25	2	C	Y	N
. . 남	. . 26	. . 3	. . E	. . Y	. . N

Classification
& Rule Creation



Rule Data

Rule	성별	나이	학점	어학	자격증	취업
1	남	23	3	A	Y	Y
2	여	24	4	B	N	Y
3	남	25	4	D	N	Y

Response Data
취업 가능함.

그림 3-2. 취업상황 예측을 위한 전체 처리 과정

2) 데이터 전처리 프로세스(Preprocess Process) 설계

규칙을 이용하여 상황을 예측하기 위해서는 데이터 전처리 과정을 통한 분류화와 규칙생성의 준비 단계가 요구된다. 우선 Input Data(나이, 성별, 학점, 어학점수, 자격증 유무 등)를 입력받게 되며 데이터 전처리 과정(Preprocess)에서 이러한 데이터들은 저장, 수정 처리 할 수 있으며 입력 데이터는 범주형 데이터 형식의 아닌 실제 학점 (cf,. 4.5. 3.8) 또는 어학점수(900, 700)와 같은 형태로 입력받게 될 것이다. 하지만 본 논문에서 알고리즘 구현을 위해 입력받은 값을 범주형 데이터로 변환(cf, 학점 4.5 -> 4, 어학 900-> A)처리 하여 정규화 프로세스(Normalization Process)가 이루어진다. 데이터 정규화 프로세스 에서 입력받은 데이터를 데이터베이스에 저장된 Threshold 값과 일치하는 데이터를 조회하여 입력한 토익점수(Language_Score)나 학점(Credit)과 일치하는 데이터를 정규화 프로세스(Normalization Process)가 변환 후 정규화 데이터(Normalization Data)로 변환 하는 작업을 거치게 된다. 이렇게 변환 된 데이터를 PredictMst 테이블에 저장함으로서 규칙을 생성하기 위한 기본적인 데이터를 구성 할 수 있다.

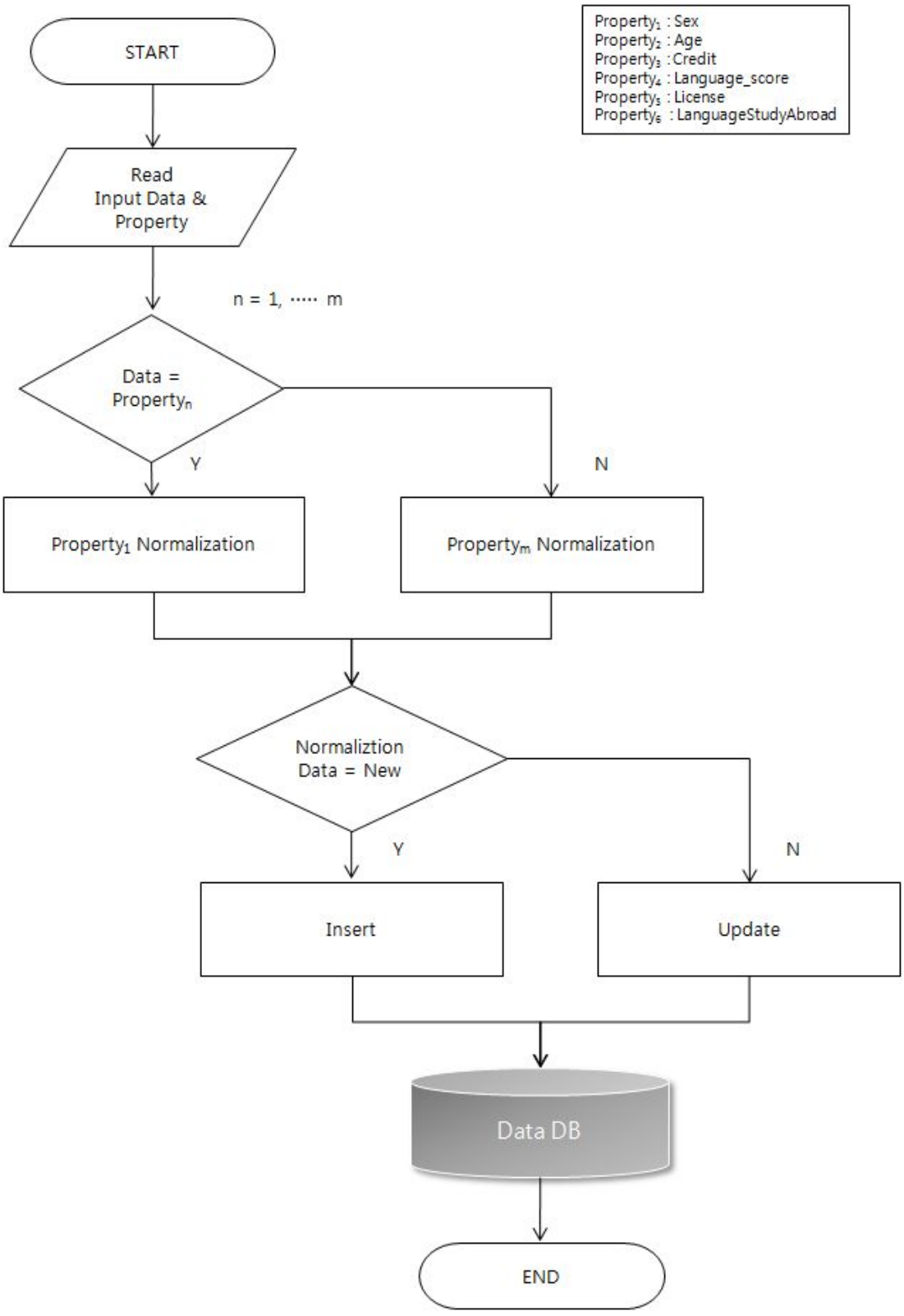


그림 3-3. 데이터 전처리 프로세스 흐름도

그림 3-3는 데이터 전처리 처리 흐름도이며, 데이터 저장과 정규화과정을 흐름도를 통해 나타내고 있다. 흐름도 상단의 흐름에서 사용자를 통해 데이터 입력이 이루어지면 Property 데이터는 범주형 데이터로 정규화 되어야 하는 컬럼인지 아닌지 판단이 이루어진다. $Property_1$은 정규화가 필요한 경우이므로 그림 3-6에서 와 같이 데이터의 Normailization과정을 거치게 된다. 정규화 과정이 필요 하지 않는 경우의 변수$Property_n$ ($_n$, 주어진 변수 종류)는 기존의 저장된 데이터의 유무의 따라 새로운 데이터라면 저장(Insert)을 통해 저장되며 기존에 존재하는 데이터의 경우라면 수정(update)되어 Preprocess 과정을 마친다. 그림 3-4는 위에 언급된 데이터 Preprocess과정의 예를 나타낸다. 그림에서 사용자가 학점(Credit)을 입력하였고 학점(Credit)은 정규화가 필요한 데이터이므로 'Y'로 정규화 과정을 거치게 된다. 데이터의 기존 존재 유무에 따라 데이터는 추가 또는 수정 로직을 걸치게 될 것이며, License_YN은 정규화 과정이 필요 없는 변수이므로 판단에서 'N'으로 판단되고 데이터 존재 유무에 따라 저장 되는 흐름을 보여주고 있다.

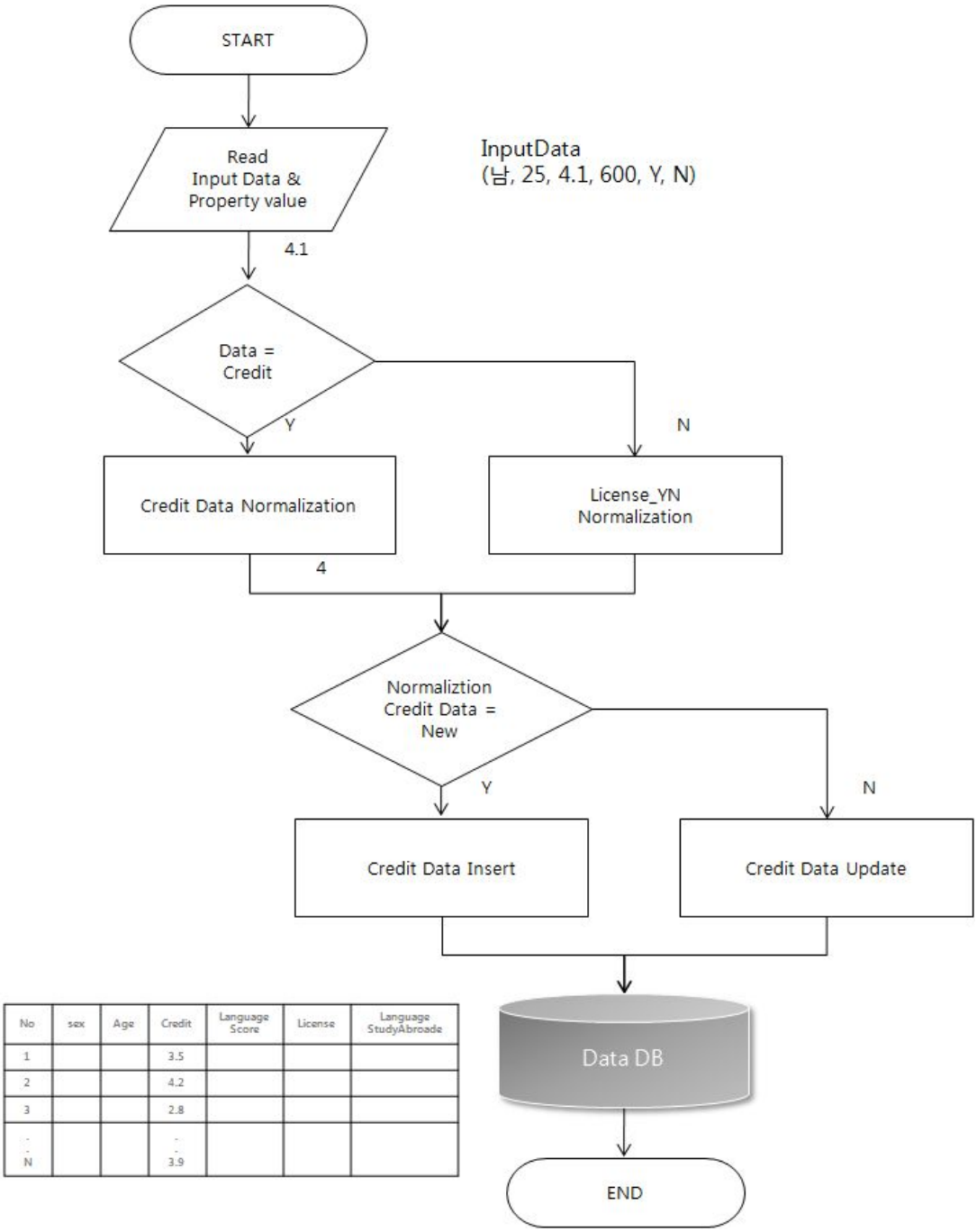


그림 3-4. 데이터 전처리 프로세스 흐름도(예)

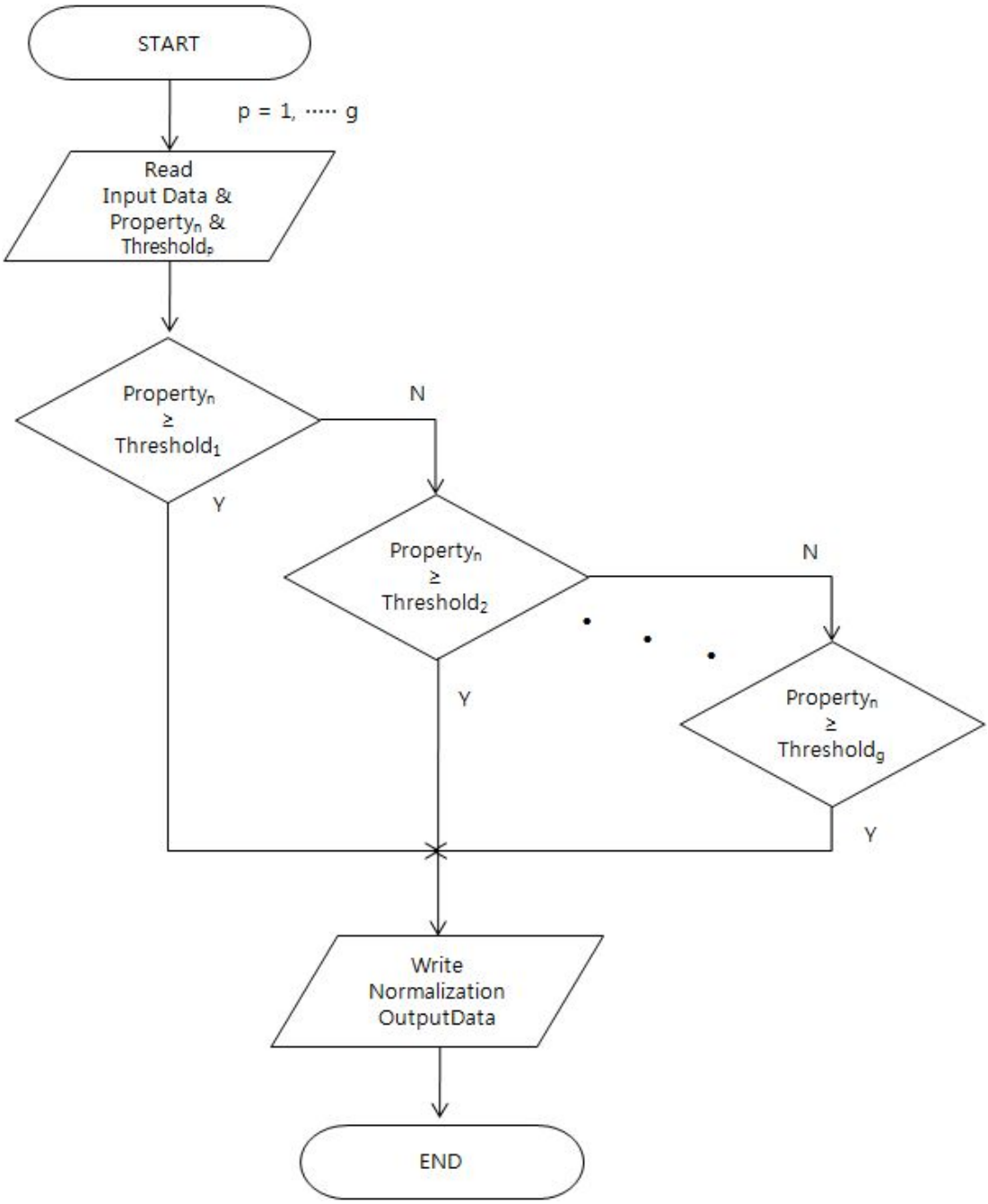
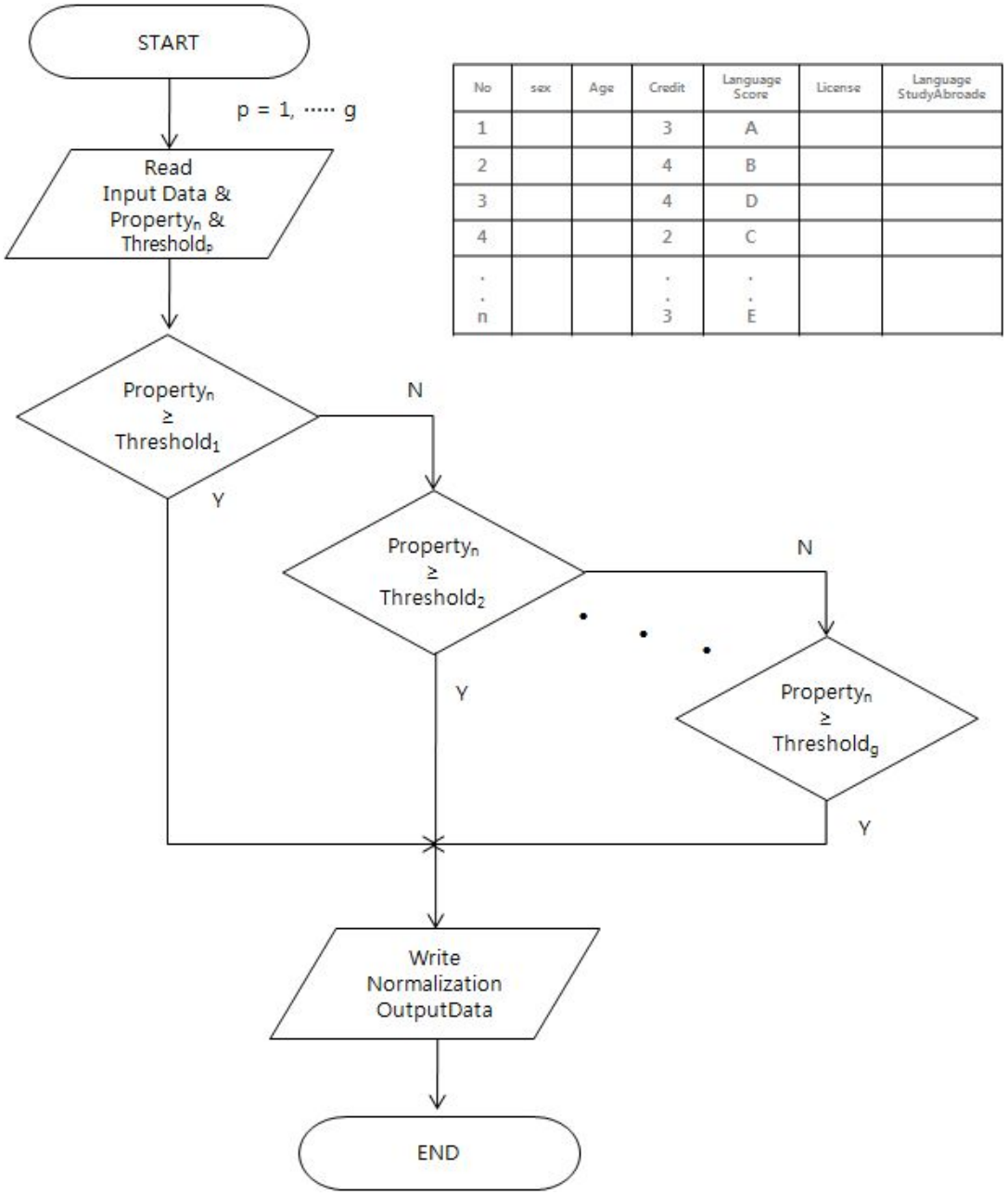


그림 3-5. 정규화 프로세스 흐름도

그림 3-5은 데이터 정규화(Normalization) 과정을 흐름도(Flow Chart)로 나타내고 있다. 입력된 데이터는 기존의 데이터 테이블에 저장되어 있는 해당 변수값과 비교하여 기준값(Threshold)에 있는 데이터 값을 해당 범주형 데이터로 변

화하는 과정을 나타낸다. p는 각 변수가 가질 수 있는 경계치의 수를 나타낸다. 그림 3-6은 정규화의 예를 보여준다. 그림에서 사용된 변수 항목은 Credit(학점)이다. 입력된 학점에 따라 그림에서 4.0보다 큰 경우, 3.0.....1.0보다 큰 경우 순으로 대상값의 경계값에 따라 변환 되는 예를 보이고 있다.



No	sex	Age	Credit	Language Score	License	Language StudyAbroade
1			3	A		
2			4	B		
3			4	D		
4			2	C		
. . n			. . 3	. . E		

그림 3-6. 정규화 프로세스 흐름도(예)

3) 분류화 프로세스(Classification Process) 설계

분류화 프로세스(Classification Process) 구조는 그림 3-7과 같다. 정규화 프로세스에서 처리된 데이터는 의사결정트리 알고리즘 생성을 위한 범주형 데이터가 된다. 이를 활용하여 분류화 프로세스는 각 필드의 엔트로피 값을 계산하고 엔트로피 계수로 변수에 대한 기대정보와 정보이득을 구하고 정보이득이 가장 큰 변수를 뿌리(Root)노드로 트리를 생성하게 되며, 최상위 노드를 구한 후 Leaf노드를 구하고 모든 Leaf노드가 구해지면 데이터의 분류가 종료되어 의사결정트리를 이루게 되며 각 Leaf노드를 끝으로 뿌리노드를 시작점을 갖는 가지는 각각의 규칙이 되어 저장된다.

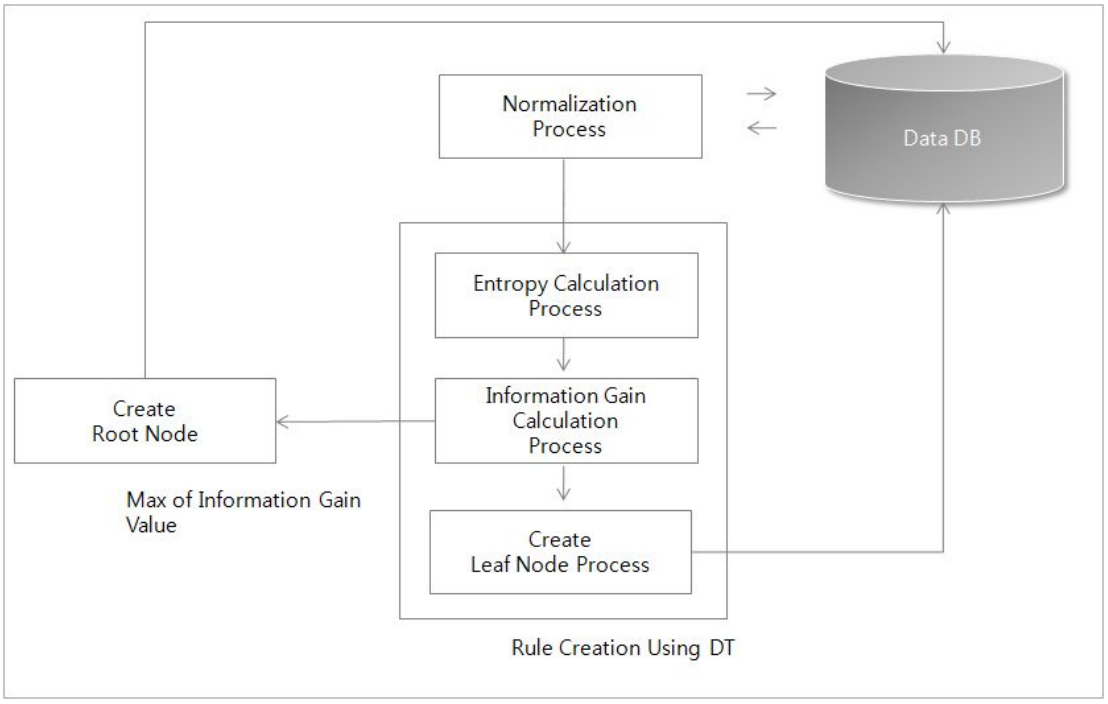


그림 3-7. 분류화 프로세스 구조

이렇게 생성된 가지는 정지 규칙이 적용되어 만족될 때까지 알고리즘이 반복되어 적용된다. 이렇게 생성된 의사결정트리의 가지분할이 완료되면 최종적으로 Leaf 노드의 값으로 결과를 예측할 수 있으며 , 뿌리노드로부터 Leaf 노드에 이

르는 가지가 규칙을 형성하게 되며, 분류화 프로세스(Classification Process)에서는 완성된 규칙을 데이터베이스에 저장하게 된다. 따라서 규칙 데이터베이스에 생성된 데이터를 기준으로 다음에 설명될 예측 처리 과정에서 규칙 데이터베이스를 활용하여 사용자요청에 따른 결과를 출력 해주게 된다.

그림 3-8 분류화 흐름도(Classification Flow Chart)는 알고리즘 구현을 위한 의사결정트리의 알고리즘 구현 흐름도를 보여준다. 흐름도에서 상단의 전처리 과정이 완료 된 후 정규화 된 데이터를 이용하여 각 변수의 엔트로피를 계산하게 된다. 엔트로피 값은 계산식에 의해 구하며, 이렇게 얻어진 각 변수의 엔트로피 값을 이용하여 정보이득을 계산식을 이용하여 구한다. 엔트로피와 정보이득 값을 이용하여 정치 규칙이 적용되어 만족 될 때까지 반복된다.

$$p_i = \frac{freq(C_i, S)}{|S|} \quad \text{여기서,} \quad Entropy(S) = -\sum_{i=1}^{m} p_i \log_2(p_i) \qquad \text{(식 3)}$$

S : 주어진 데이터들의 집합

C = {C1, C2, ... , Ck } : 클래스 값들의 집합

freq(Ci,S) : S에서 class Ci에 속하는 레코드의 수

|S| : 주어진 데이터들의 집합의 데이터 개수

$$Gain(A) = I(s_1, s_2, s_3 \cdots s_m) - E(A) \qquad \text{(식 4)}$$

I : 상위 노드의 엔트로피

S : 데이터 속성값

E : A선택시 파생되는 엔트로피 가중치 평균값

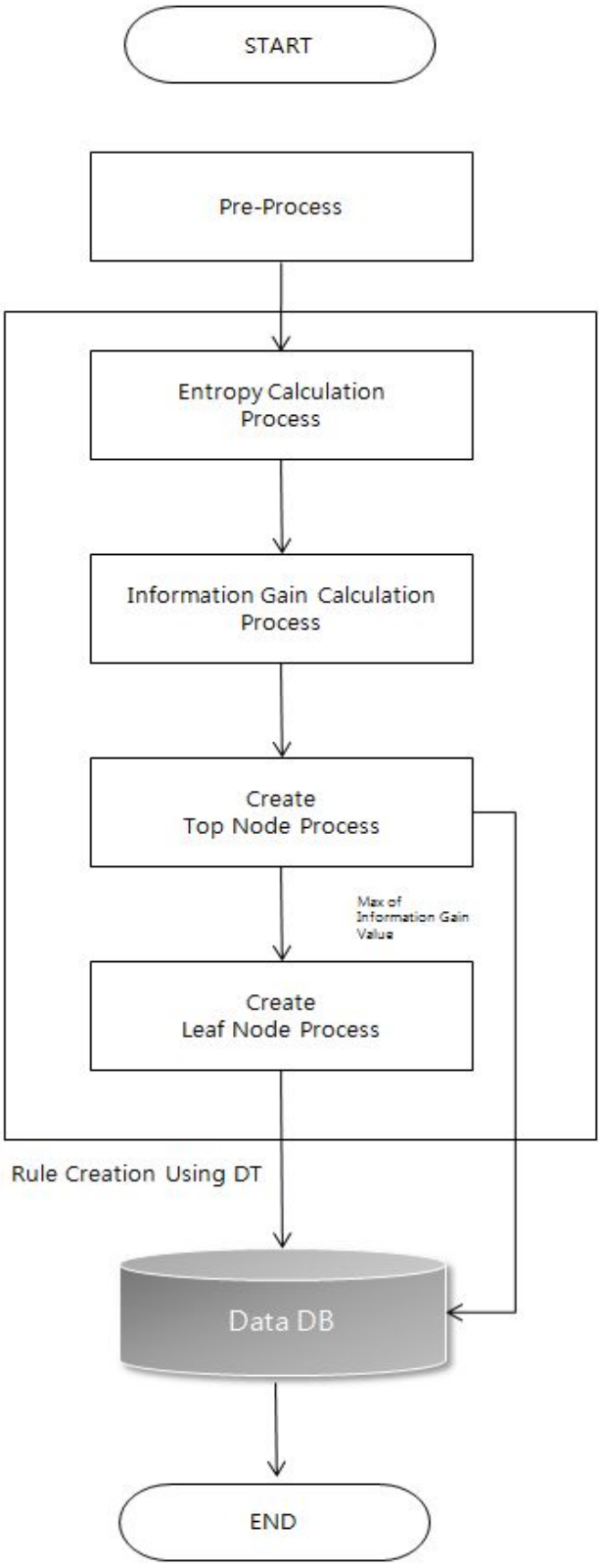


그림 3-8. 분류화 프로세스 흐름도

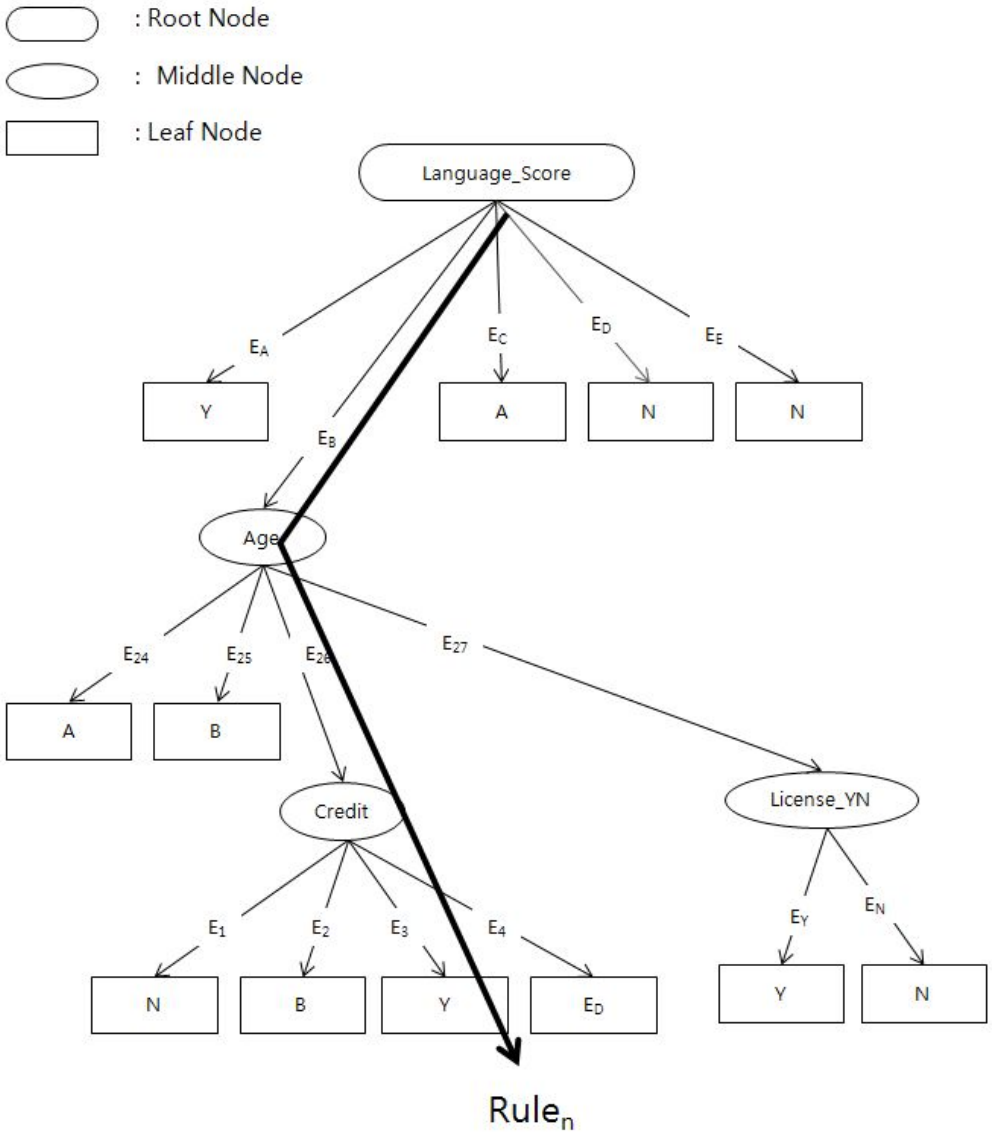


그림 3-9. 분류화 프로세스 흐름도(예)

그림 3-9는 위의 분류화 흐름도 예로 의사결정트리의 구성과정을 나타낸다. 우선 상단 최상위 노드를 구하기 위해 각 변수마다의 엔트로피 계산되고 엔트로피 값을 이용하여 정보이득값을 구하고, 각 변수의 정보이득 값이 가장 높은 변수 Language_Score가 최상위 노드로 선정된다. 최상위 노드를 구하였으므로 그 다음은 Language_Score를 제외한 각 변수의 엔트로피와 정보이득 값을 반복하여

구한다. Language_Score변수에서 각 속성값은 5가지로 나눠지며 이중 B항목을 제외한 나머지 값은 정지규칙이 만족되어 더 이상 분할되지 않으며 B항목은 다시 나머지 변수들중 가장 정보이득이 높은 Age로 분할되어 다시 Credit 와 License_YN으로 분할되어 Leaf노드까지 완료 된다. 완성된 트리는 총 12개의 가지를 이루며 규칙 12개를 생성한다.

4) 상황예측 처리과정

예측처리 과정에 관련된 구조는 그림 3-10과 같다. 위 분류화 프로세스(Classification Process)에서 저장한 규칙 데이터베이스데이터를 활용하여 예측은 사용자가 요청한 데이터를 규칙 데이터베이스와 비교하여 존재유무를 가리고 결과를 사용자에게 출력하게 된다.

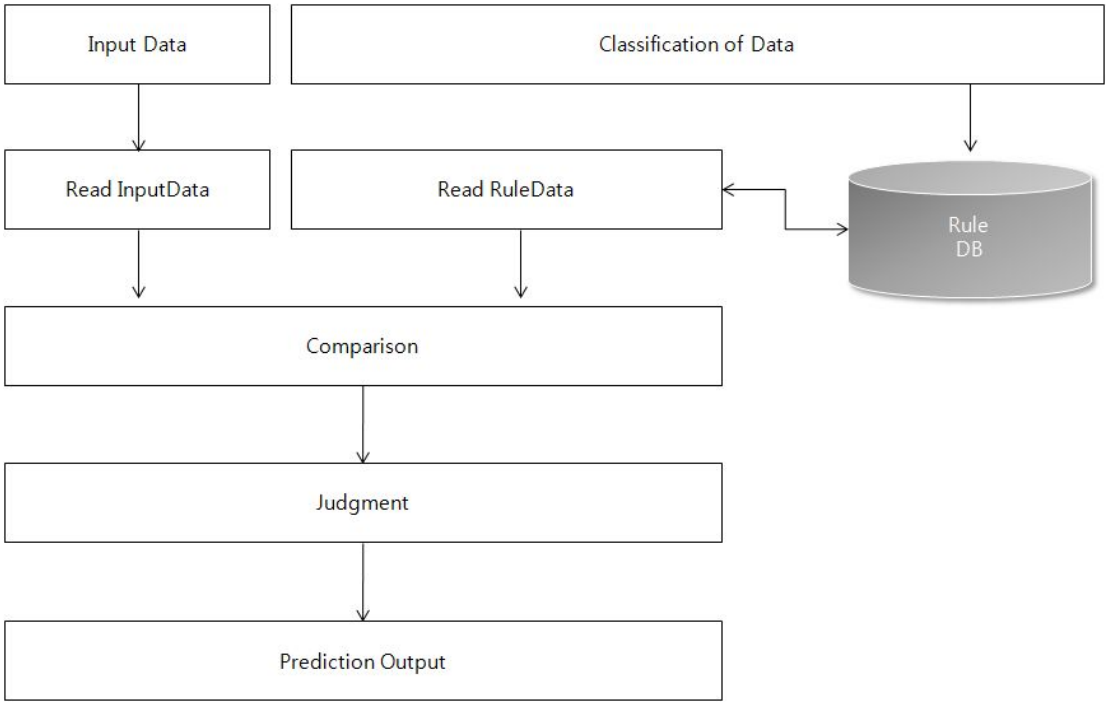


그림 3-10. 예측 프로세스 구조

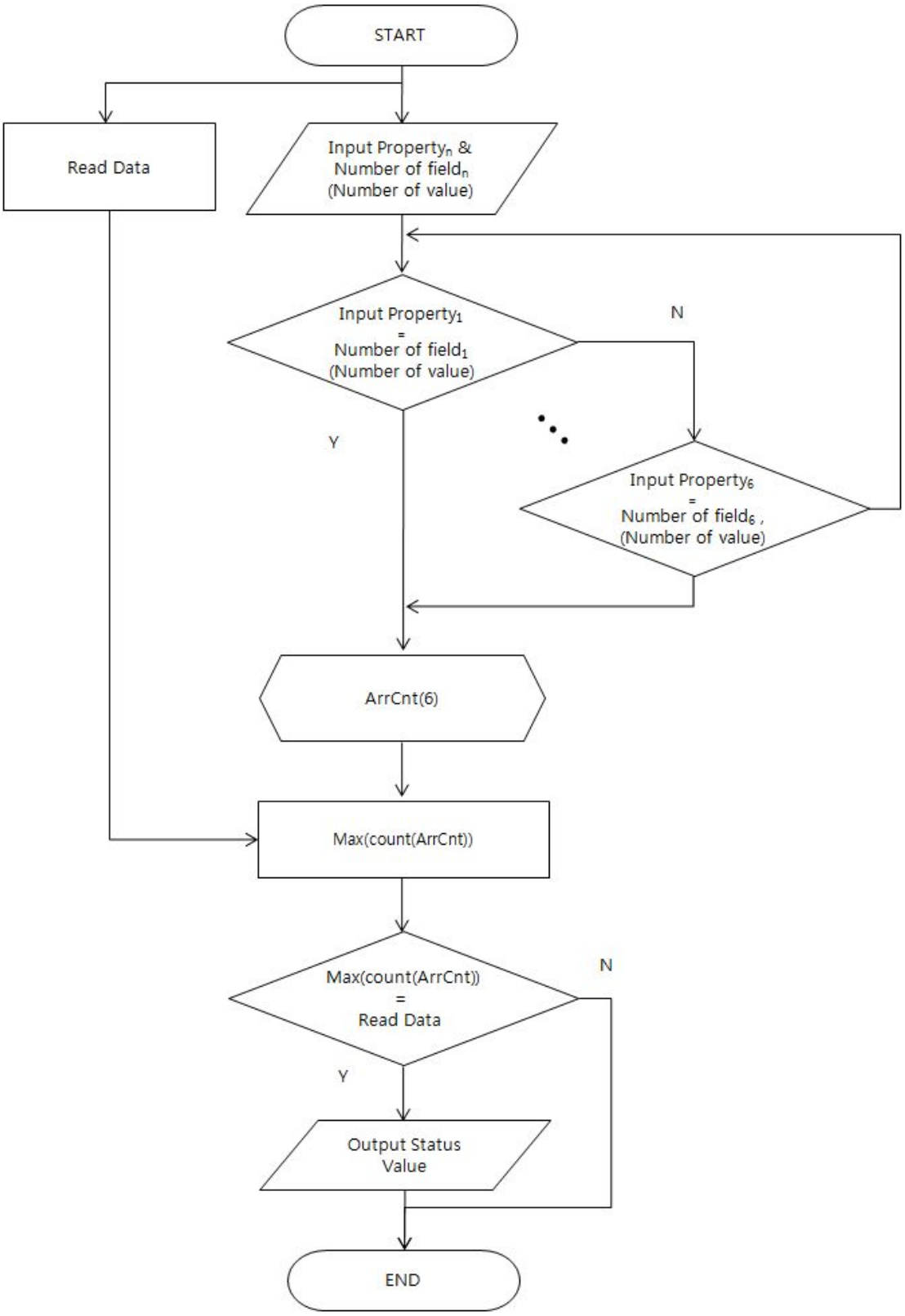


그림 3-11. 예측 프로세스 흐름도

그림 3-11에서는 규칙 데이터베이스데이터와 요청 데이터를 예측 프로세스(예측 처리 과정)가 비교 처리하여 결과를 출력하는 과정을 보여주고 있다. 흐름도 상단에서 사용자의 입력데이터와 데이터베이스에 저장되어있는 규칙(Rule) 데이터 값을 입력 받고 입력된 해당 데이터와 데이터베이스에 저장된 각 항목에 들어있는 레코드 값과 비교하여 일치하는 항목들을 검색한다. 이중에서 일치하는 변수가 존재하는 경우 배열 ArrCnt라는 곳에 해당 차례 순으로 Y를 저장하며 , 일치하는 값이 없는 경우 null값을 저장하여준다 이렇게 저장된 배열은 데이터베이스에서 검색한 항목 중 가장 많은 수의 항목의 일치하는 WORK_STATUS값을 검색하게 되고, 결과로 출력하게 된다. 그림 3-12은 예측 프로세스 흐름도의 예제를 나타낸다. 흐름도 에서 입력 값 (25, 남, 3, B, N,, N) 사용자로부터 요청되었고, 예측 프로세스는 Read한 데이터베이스에 데이터 값과 비교 한다. 그림 3-12에서는 남자, 학점 3, 어학점수 B인 변수를 비교하였으며 해당 데이터 값과 일치하므로 해당 배열 ArrCnt에 해당 필드에 해당되는 곳에 'Y'를 할당 한다. 할당된 배열 규칙 테이블에서 가장 많은 값과 일치하는 해당 변수 Max값을 선택하여 해당 WORK_STATUS(N)값을 출력하여 완료된다.

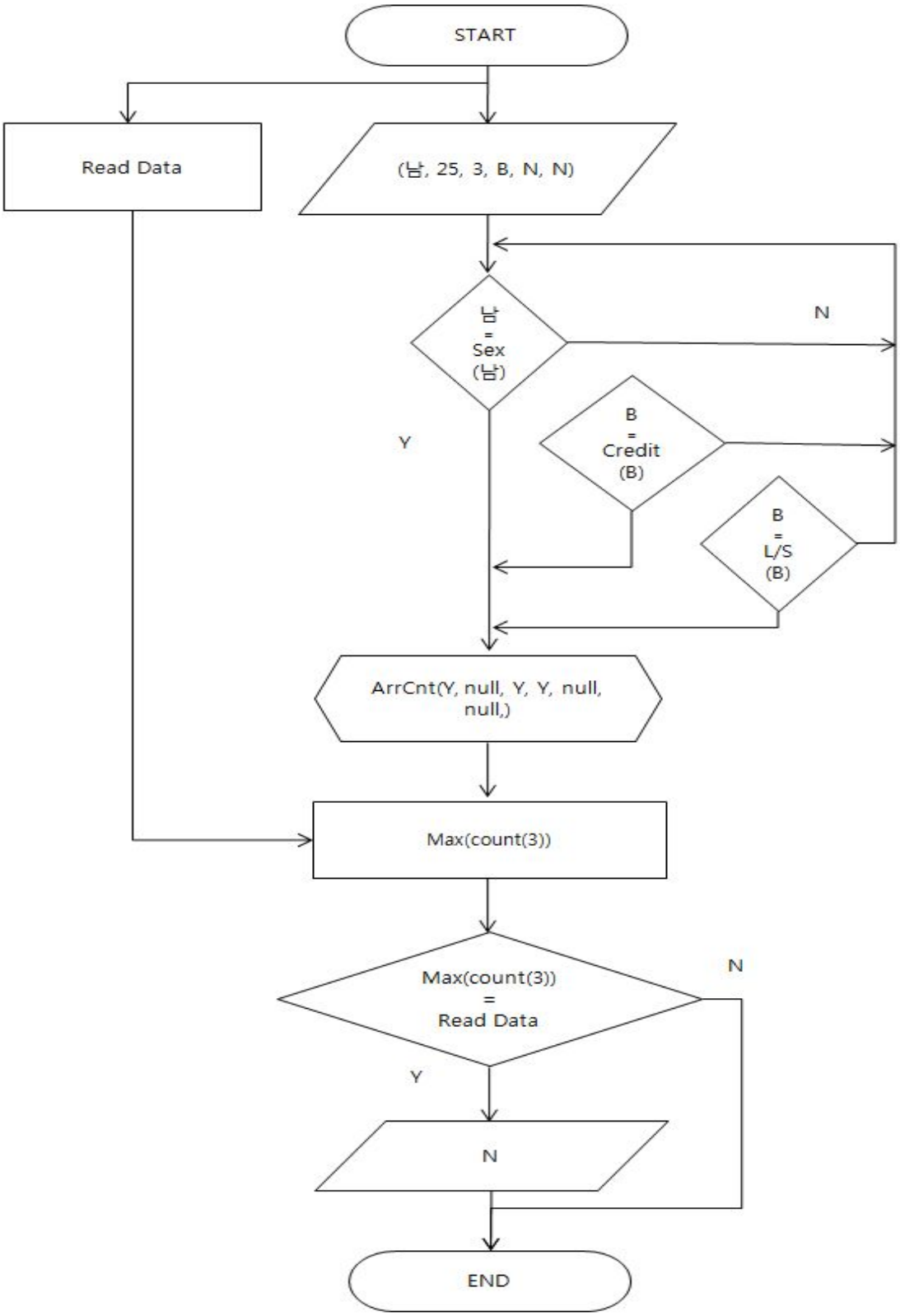


그림 3-12. 예측 프로세스 흐름도(예)

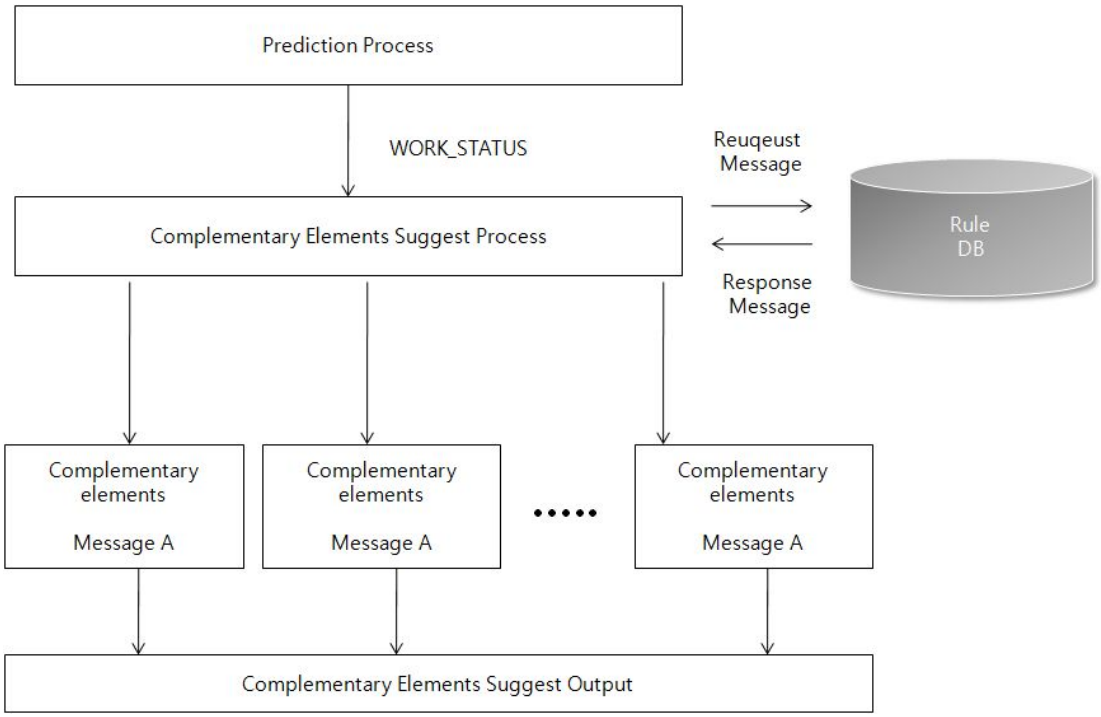


그림 3-13. 보완요소 추천 프로세스 구조

5) 취업 보완 추천 처리과정

취업 여부를 예측하고 더불어 취업하기 위한 보완요소를 추천(Complementary Element Suggest Process)한다. 그림 3-13는 예측결과에 따라 취업을 위한 보완요소를 추천해주는 프로세스(Complementary Elements Suggest Process)구조를 나타내고 이TEk.. 보완요소 추천은 취업상태가 'Y'인 경우 즉, 취업상태인 경우는 제외되며 나머지 미취업(N), 서류통과(A), 면접통과(B)의 결과에 경우에만 보완요소를 추천하여준다. 보완요소는 데이터베이스에서 해당 결과에 따라 출력 한다.

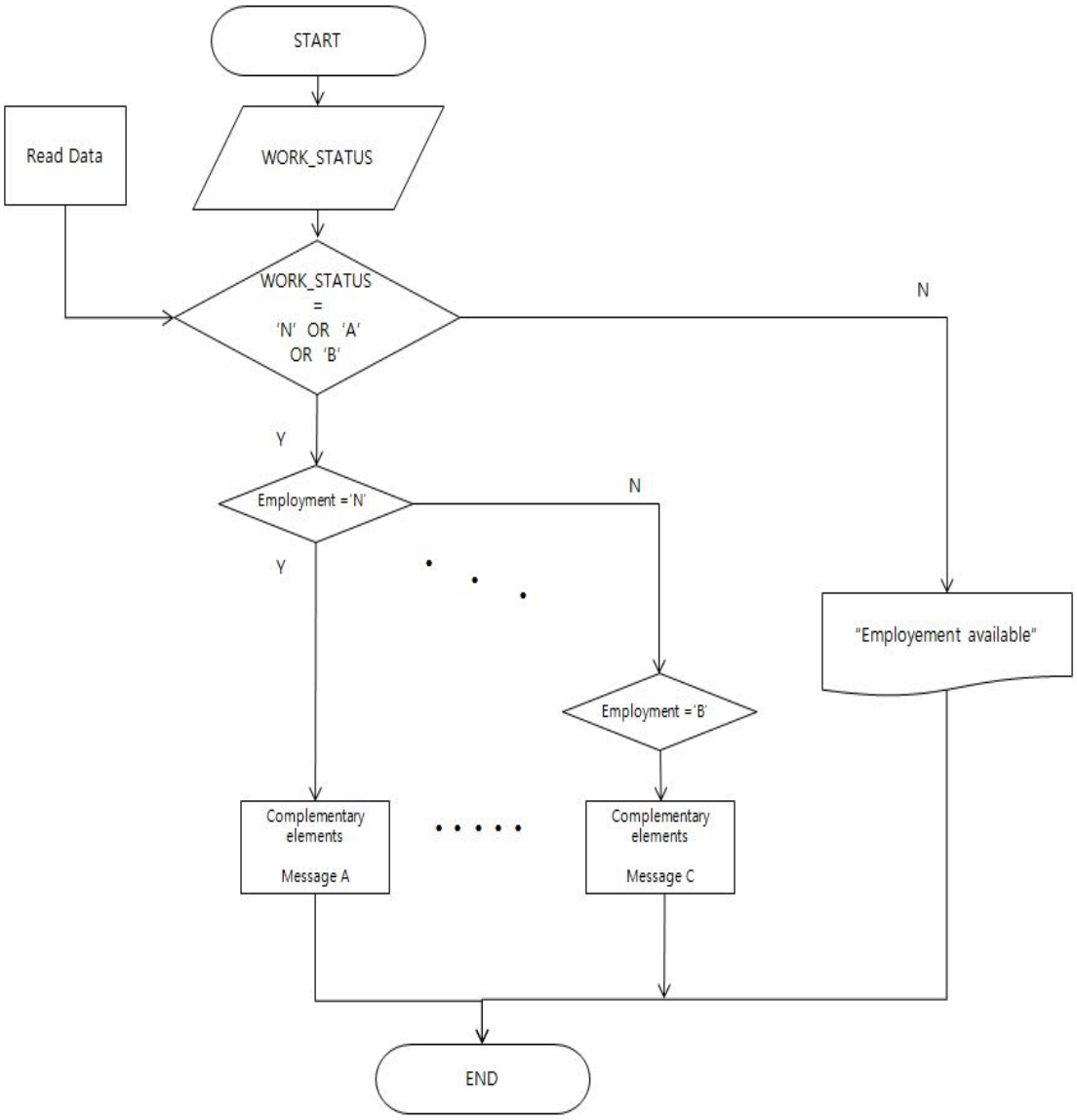


그림 3-14. 보완요소 추천 프로세스 흐름도

그림 3-14은 보완요소 추천의 흐름도를 나타낸다. WORK_STATUS값이 'N', 'A' , 'B'인 경우 메시지를 출력하여 주며 'Y'인 경우는 Employment Available로 취업가능 메시지를 출력하여 완료 된다.

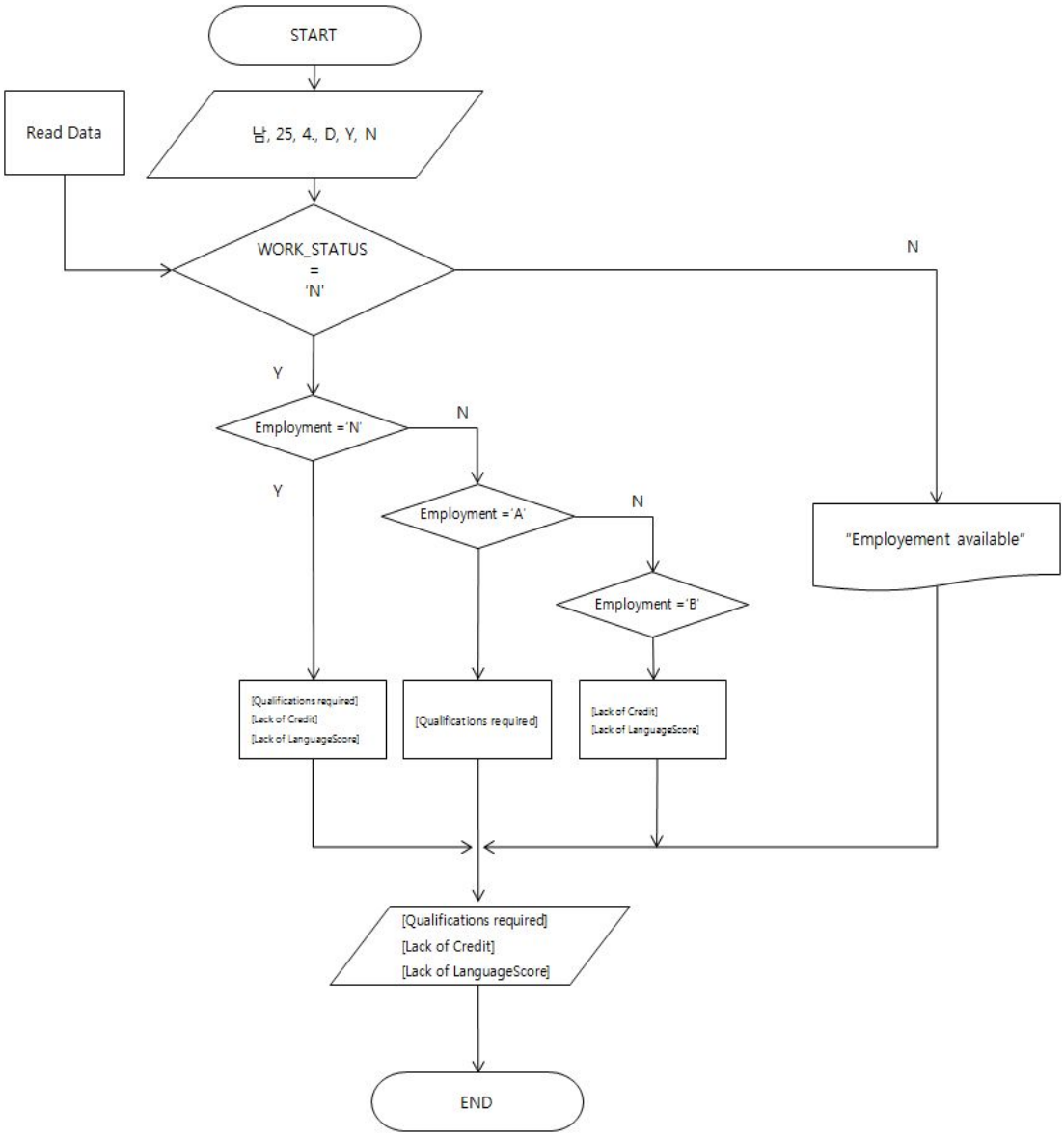


그림 3-15. 보완요소 추천 프로세스 흐름도(예)

그림 3-15은 보완요소 추천 프로세스 흐름도 예를 나타낸다. 취업상태가 'N' 인 경우로 해당 ReadData의 메시지 자격증 요구됨(Qualifications required), 학점 부족(Lack of Credit), 어학점수부족(Lack of Language_Score)로 출력 하여 주고 있다. 만약 데이터가 'Y' 라면 "Employment available" 메시지를 출력하여 완료 된다.

6) 취업예측데이터베이스 Table 설계

취업예측시스템 테이블은 모두 4가지로 구성 되며, 각 테이블은 사용자가 입력한 데이터를 저장하는 테이블 (성별, 나이, 학점, 어학점수, 자격증유무, 어학연수유무, 학적상태, 취업상태)인 PredictMst테이블, 트리생성이후 규칙이 된 데이터를 저장하는 테이블 (성별, 나이, 학점, 어학점수, 자격증유무, 어학연수유무, 취업결과)인 RuleData테이블, 사용자가 입력한 어학점수를 범주형 데이터로 변환하기 위해 기준정보를 갖고 있는 테이블 (어학구분, 어학명, 시작점수, 종료점수, 변환점수)인 Language테이블, 사용자가 입력한 학점점수를 범주형 데이터로 변환하기 위해 기준정보를 갖고 있는 테이블 (학점구분, 학점명, 시작점수, 종료점수, 변환점수)인 Credit테이블로 구성된다.

표 3-1. 데이터 베이스설계

PredictMst

열 이름	데이터 형식	Null 허용
student_no	varchar(50)	☐
sex	varchar(50)	☑
age	varchar(50)	☑
credit	varchar(50)	☑
language	varchar(50)	☑
licnese_yn	varchar(1)	☑
exp_language_yn	varchar(1)	☑
status	varchar(1)	☑
work_status	varchar(50)	☑

RuleData

열 이름	데이터 형식	Null 허용
seq	int	☐
sex	varchar(50)	☑
age	varchar(50)	☑
credit	varchar(50)	☑
language	varchar(50)	☑
license_yn	varchar(1)	☑
exp_language_yn	varchar(1)	☑
result	varchar(50)	☑

Language

열 이름	데이터 형식	Null 허용
language_gb	varchar(50)	☑
language_nm	varchar(50)	☑
from_score	varchar(50)	☑
to_score	varchar(50)	☑
con_score	varchar(50)	☑
		☐

Credit

열 이름	데이터 형식	Null 허용
credit_gb	varchar(1)	☑
crdeit_nm	varchar(50)	☑
from_score	varchar(50)	☑
to_score	varchar(50)	☑
con_score	varchar(50)	☑

Ⅳ. 시뮬레이션 및 성능 분석

본 논문에서는 취업상황 예측을 위해 알고리즘을 이용하여 학생들의 정보를 통해 취업을 예측한다. 학생들의 데이터는 신뢰성 있는 데이터로 성별, 나이, 학점, 어학점수, 어학연수경험 유무, 자격증 유무에 관한 정보를 얻어 데이터화 시켰다. 인공지능 예측에 사용되는 의사결정트리 C4.5알고리즘을 잉용하여 절차에 따라 입력된 정보를 사용하여 각 변수별 엔트로피 값을 계산하고 각 변수별 정보이득 값을 얻어낸 후 Root노드를 정하게 된다. 뿌리노드가 지정되고 난후에는 사용된 변수는 제외하여 하위노드를 생성하게 되며 각 변수별 정보이득값을 구하여 노드를 생성하여 의사결정트리를 만들며 각 노드가 leaf노드로 모두 완료되면 각 Leaf노드는 규칙으로 만들어져 데이터베이스에 저장하게 된다. 이렇게 저장된 규칙을 바탕으로 3장에서 의사결정트리 기반으로 설계한 취업예측 알고리즘을 작성된 시나리오를 바탕으로 시뮬레이션 성능 평가를 수행하여 결과를 분석한다. 그림 4-1은 시뮬레이션을 위해 데이터베이스에 입력된 데이터 값 화면이다.

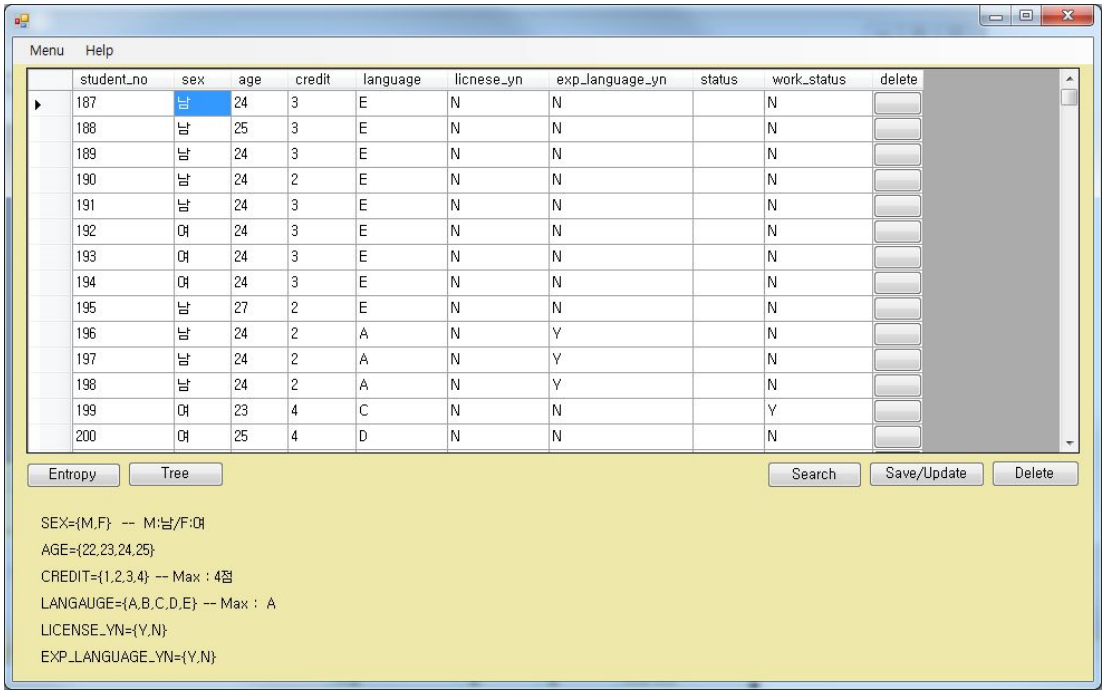


그림 4-1. 신뢰성 있는 취업정보 예

1. 시뮬레이션 환경

본 논문에서 제안한 취업예측알고리즘은 아래와 같은 환경에서 실행한다.

표 4-1. 구현환경

구분	하드웨어	소프트웨어	비고
의사결정트리 C4.5알고리즘 구현 및 실행	Intel(R) Core(TM) i-5-2430M CPU @ 2.40Ghz 8GB RAM	Microsoft Visual Studio	C#
사용자 인터페이스	Intel(R) Core(TM) i-5-2430M CPU @ 2.40Ghz 8GB RAM	Microsoft Visual Studio	C#
데이터베이스	Intel(R) Core(TM) i-5-2430M CPU @ 2.40Ghz 8GB RAM	Microsoft SQL2008 R2	

의사결정트리 C4.5를 이용한 취업예측시스템은 Microsoft Visual Studio 이용하여 C#언어로 구현하였고 , 간단한 실험 사용자 인터페이스 또한 Microsoft Visual Studio에서 C#으로 구현한다. 사용되는 학생정보 데이터는 MS-SQL을 이용하여 Predict 데이터베이스에 저장 한다.

2. 취업상황 예측을 위한 트리 생성결과

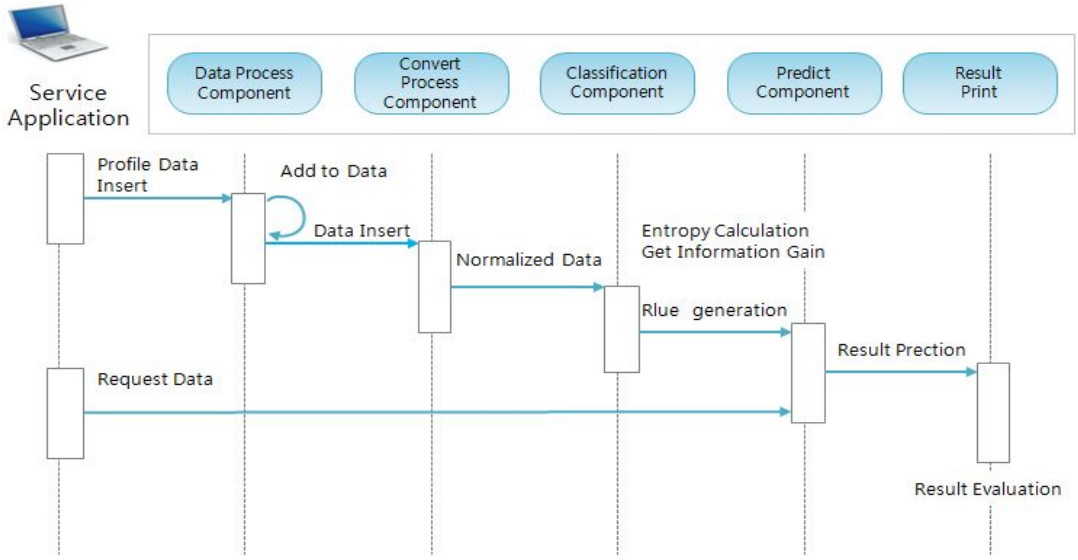


그림 4-2. 취업예측 시퀀스 다이어그램

그림 4-2은 취업예측과정을 시퀀스 다이어그램을 통해 나타내고 있다. 사용자가 데이터를 입력하면 데이터 전처리 프로세스는 데이터가 없는 경우라면 정규화 프로세스를 통해 데이터를 알고리즘에서 사용가능하도록 범주형 데이터로 변환시켜 프로그램 로직에서 각 항목별 엔트로피와 정보이득을 구하고 적절한 트리를 구성하여 분류화 프로세스(Classification Process)에서 규칙을 생성하여 규칙 테이블에 저장하고 저장된 데이터를 활용하여 예측 처리 과정은 사용자가 검색 요청한 데이터에 대해 적절한 예측 응답을 출력하면 본 시스템의 절차가 완료된다. 이러한 일련의 과정을 위해서는 시스템의 실질적인 핵심로직인 의사결정트리의 알고리즘을 구현한다. 그림 4-3은 의사결정트리 알고리즘을 보여주고 있다.

```
TreeGrowth(E,F)

 1 : if stopping_condition(E,F) = true then
 2 :         leaf = createNode()
 3 :         leaf.label = Classify(E).
 4 :         return leaf.
 5 : else
 6 :         root = creatNode().
 7 :         root.test_condition = find_best_spilt(E,F)
 8 :         let V = {v|v is a possible outcome of root.test_condition}.
 9 :         for each v  ∈ V do
10 :                 Ev = { {root.test_condtion(e) = v} ∩ {e ∈ E} }
11 :                 child = TreeGrowth(Ev,F))
12 :                 add child as descendent of root and label the edge(root -> child) as v
13 :         end for
14 : end if
15 : return root
```

그림 4-3. 의사결정트리 알고리즘

1) 뿌리노드생성

알고리즘의 구현을 위해 32개로 구성된 데이터를 가지고 트리생성의 과정을 위한 엔트로피 계산과 정보이득 값을 구하는 절차를 설명하고 트리를 생성하는 과정을 설명한다.

표 4-2. 취업여부에 따른 학생 프로파일(Profile) 정보

sex	age	credit	language	licnese_yn	exp_language_yn	work_status
F	24	1	E	N	N	N
M	24	2	D	N	N	N
F	24	3	B	N	Y	B
M	24	3	B	Y	Y	A
F	24	3	C	Y	N	B
M	24	4	C	Y	Y	Y
M	25	1	B	N	N	N
F	25	2	A	Y	Y	A
F	25	2	B	N	N	N
F	25	3	B	N	Y	A
M	25	3	C	N	N	B
M	25	4	C	N	N	B
F	26	2	D	N	N	N
F	26	3	A	Y	Y	Y
M	26	3	A	Y	Y	Y
M	26	3	A	Y	Y	Y
F	26	3	B	N	N	N
M	26	3	B	Y	N	A
F	26	3	B	Y	N	Y
M	26	4	E	Y	Y	A
F	27	2	E	N	Y	N
F	27	3	B	Y	N	B
F	27	3	E	N	N	N
M	27	4	B	N	N	A
M	27	4	B	Y	N	Y
M	27	4	D	N	Y	B
F	28	2	A	N	Y	B
M	28	3	A	N	N	A
F	28	3	C	Y	Y	B
M	28	4	A	N	Y	Y
M	28	4	A	Y	Y	Y
F	28	4	D	Y	Y	A

표 4-2는 취업 여부에 따른 학생들의 프로필정보 나타낸다. 표의 데이터 집합 E의 데이터 수는 32개 이므로 그림 4-3의[19] 단계1의 stopping, condition(E,F)은 false 이므로 6단계로 간다. 뿌리 노드 T의 createNode()를 위해 취업집단(G_1)과

미취업 집단(G_2)분포, 면접경험(G3)분포, 서류통과(G4)분표 (8/32, 8/32, 8/32 ,8/32)에 대한 엔트로피 계수 I(T)는 다음과 같다.

$$I(T) = -\ 8/32 \times \log_2 8/32 - 8/32 \times \log_2 8/32 - 8/32 \times \log_2 8/32 - 8/32 \times \log_2 8/32 = 2$$

7단계의 최적가치분할 find_best_split()을 찾기 위해 각 변수별 × 집단별 교차표를 구하고 엔트로피 계수로 변수에 대한 기대정보와 정보이득(Δ = I(T) - 기대정보)구하면 표 4-3와 같다.

표 4-3. 각 변수별 정보이득

Attribute	M	F				SUM	엔트로피	0.1171
SEX	16	16				32	1	
AttributeCnt	6,2,5,3	2,6,3,5						
엔트로피								
Attribute	24	25	26	27	28	SUM	엔트로피	1.2806
AGE	6	6	8	6	6	32	2.3113	
AttributeCnt	1,2,1,2	0,2,2,2	4,2,2,0	1,2,1,2	2,0,2,2			
엔트로피	1.9184	1.5849	0	1.9184	1.0566			
Attribute	1	2	3	4		SUM	엔트로피	1.093953125
CREDIT	2	6	15	9		32	1.7299	
AttributeCnt	0,2,0,0	0,4,1,1	4,2,4,5	4,0,3,2				
엔트로피								
Attribute	A	B	C	D	E	SUM	엔트로피	1.33443125
language	8	11	5	4	4	32	2.1981	
AttributeCnt	5,0,2,1	2,3,4,2	1,0,0,4	0,2,1,1	0,3,1,0			
엔트로피	0.875	1.9362	0.2575	1.5	0			
Attribute	Y	N				SUM	엔트로피	1.08996875
licnese_yn	15	17				32	0.9971	
AttributeCnt	7,0,5,3	1,8,3,5						
엔트로피								
Attribute	Y	N				SUM	엔트로피	0.1727
exp_language_yn	16	16				32	1	
AttributeCnt	6,1,5,4	2,7,3,4						
엔트로피		1.8496						

2) 중간노드 및 Leaf노드 생성

Language_Score인 어학점수가 정보이득이 가장 크므로 뿌리노드는 어학점수가 되고 단계 8의 어학점수 변수 값 집합은 V={A, B, C, D, E} 이 된다. 어학점수에 따른 단계 10의 E_A, E_B, E_C, E_D, E_E 을 의사결정트리형태로 그리면 그림 4-4와 같다.

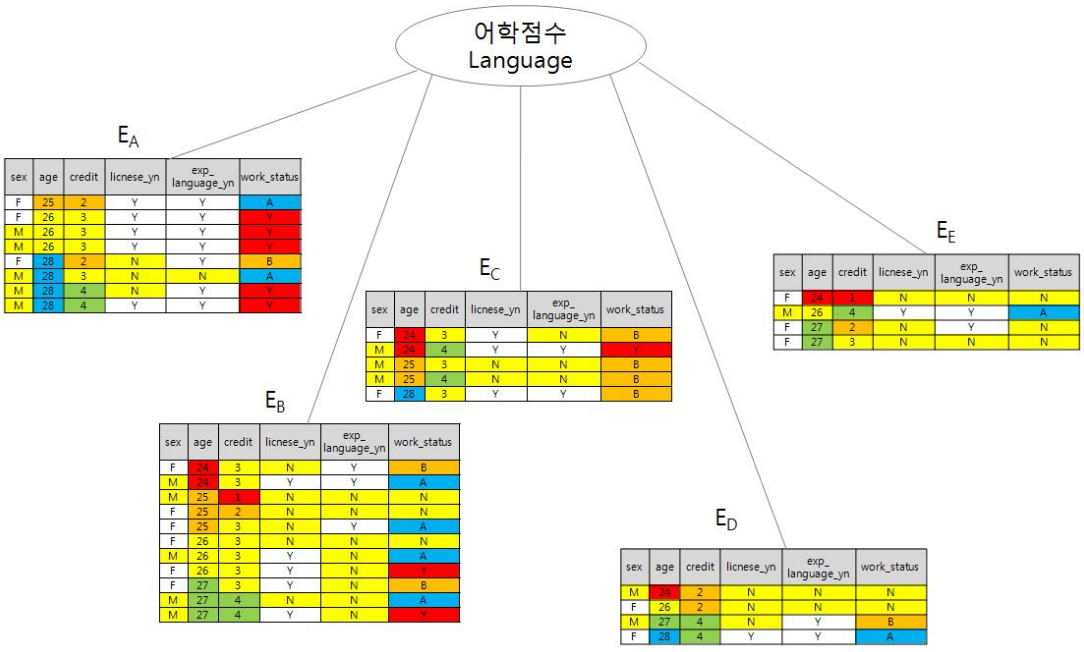


그림 4-4. 어학점수에 따라 가지분할 된 의사결정트리

각각의 데이터 집합은 정지규칙이 만족될 때까지 TreeGrowth() 알고리즘이 반복되어 적용 된다(단계11). 이중에서 어학점수 A, C, D, E 데이터 집합 E_A, E_C, E_D, E_E 더 이상 분할되지 않고 정지규칙을 만족시키므로 다수결에 의해 각 Record 별 결과가 정해진다. 잎은 다수결에 의해 E_A는 취업, E_C는 서류통과 , E_D, E_E는 미취업으로 결정된다.

어학점수 (B) 11명의 데이터 집합 E_B 에 대해 취업집단(G_1)과 미취업 집단(G_2) 분포, 면접경험(G3)분포, 서류통과(G4)분포 (2/11, 3/11, 4/11, 2/11)에 대한 엔트로피계수는 다음과 같다.

$-(2/11) \times \log_2(2/11) - (3/11) \times \log_2(3/11) - (4/11) \times \log_2(4/11) - (2/11) \times \log_2(2/11) = 1.9362$

11명의 E_B에 대한 최적가치분할 find_best_split()을 찾기 위해 각 변수별 × 집단별 교차표를 구하고, 엔트로피계수로 변수에 대한 기대정보와 정보이득을 구하면 표 4-4과 같다.

표 4-4. 어학점수 범주가 E_B에 대한 각 변수별 정보이득

Attribute	M	F			0.688854545
SEX	5	6			
AttributeCnt	1,1,3,0	1,2,1,2			
엔트로피	0.4422	1.9183			
Attribute	24	25	26	27	1.124881818
AGE	2	3	3	3	
AttributeCnt	0,0,1,1	0,2,1,0	1,1,1,0	1,0,1,1	
엔트로피	0.5	0.5283	1.0566	1.0566	
Attribute	1	2	3	4	0.672790909
CREDIT	1	1	7	2	
AttributeCnt	0,1,0,0	0,1,0,0	1,1,3,2	1,0,1,0	
엔트로피	0	0	1.8425	0.5	
Attribute	Y	N			-7.270454545
licnese_yn	5	6			
AttributeCnt	2,0,2,1	0,3,2,1			
엔트로피	0.9932	1.4592			
Attribute	Y	N			0.406227273
exp_language_yn	3	8			
AttributeCnt	0,0,2,1	2,3,2,1			
엔트로피	0.5283	1.9056			

어학점수가 B범주인 경우 정보이득은 AGE가 가장 높다. 그림 4-5과 같은 의사결정트리가 형성된다.

정보이득은 어학점수가 B범주인 11명 전체 취업상황 분포(2/11, 3/11, 4/11, 2/11)의 엔트로피에서 기대정보를 빼어 계산한다.

성별 정보이득(Δ = I(T) - 기대정보)

기대정보 = 각속성수/전체속성수 × 각 변수의 속성 엔트로피값

Δ = 1.9362 -(5/11 × 0.4422 + 6/11 ×1.9183)

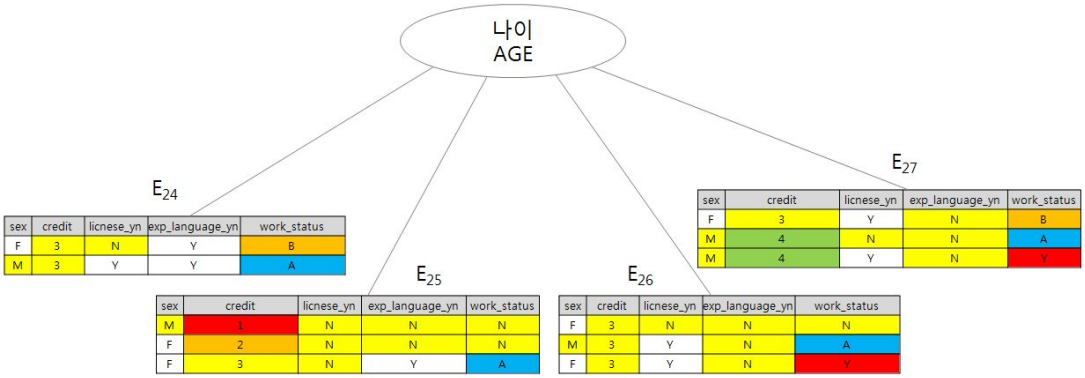


그림 4-5. 어학점수 범주가 B에 대해 나이로 분할된 의사결정트리

나이까지 분할된 트리는 E_{24}, E_{25}, E_{26}, $E_{,27}$로 분할되며, 각 트리는 다수결에 의해 규칙(Rule)으로 저장되거나 또는 같은 수의 결과를 갖는 경우 각각의 Row를 규칙으로 저장하게 된다.

3) 의사결정트리 완성

이상을 종합하면 그림 4-6과 같은 의사결정 트리가 완성된다.

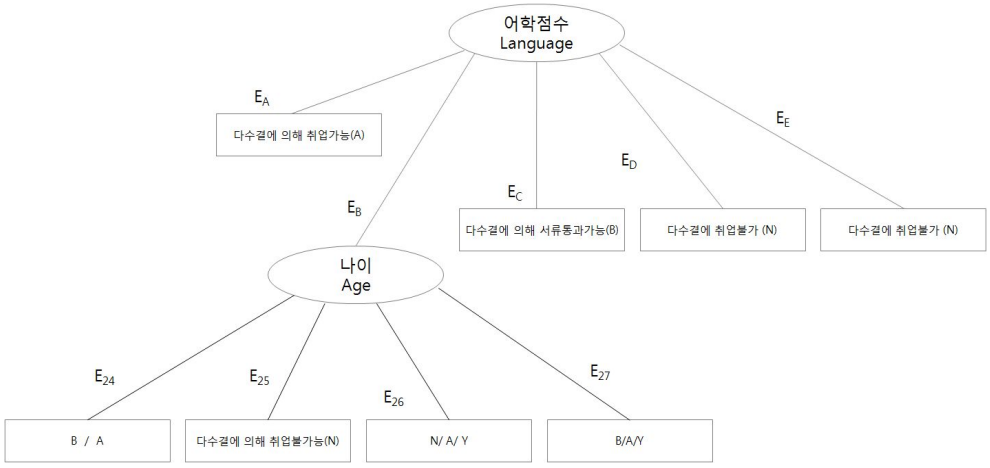


그림 4-6. 학생이 취업이 가능여부를 분류하는 의사결정트리

4) 규칙(Rule)생성

표 4-5. 규칙생성 결과

seq	sex	age	credit	language	licnese_yn	exp_language_yn	result
8675	NULL	NULL	NULL	A	NULL	NULL	Y
8676	NULL	24	NULL	B	NULL	NULL	A
8678	NULL	25	NULL	B	NULL	NULL	N
8679	NULL	26	NULL	B	NULL	NULL	Y
8682	NULL	27	NULL	B	NULL	NULL	Y
8685	NULL	NULL	NULL	C	NULL	NULL	B
8686	NULL	NULL	NULL	D	NULL	NULL	N
8687	NULL	NULL	NULL	E	NULL	NULL	N

cf)

```
IF LANGUAGE=A THEN WORK_STATUS =Y
IF AGE= 24 AND LANGUAGE=B THEN WORK_STATUS = A
IF AGE= 254 AND LANGUAGE=B THEN WORK_STATUS = N
IF AGE= 26 AND LANGUAGE=B THEN WORK_STATUS = Y
IF AGE= 27 AND LANGUAGE=B THEN WORK_STATUS = Y
IF LANGUAGE=C THEN WORK_STATUS = B
IF LANGUAGE=D THEN WORK_STATUS = N
IF LANGUAGE=E THEN WORK_STATUS = N
```

표 4-5는 의사결정트리의 모든 과정이 끝난 후 생성된 규칙을 나타내고 있다. 표를 이용하여 cf와 같이 코드로 표현 할 수 있다.

표 4-6. 시뮬레이션을 위한 훈련 데이터

구분	속성값	기타
성별 (SEX)	M	M : 남 , F : 여
나이 (AGE)	27	23 , 24 , 25 , 26
학점 (CREDIT)	3	4 : 4.5 ~ 4.0 3 : 3.9 ~ 3.0 2 : 2.9 ~ 2.0 1 : 1.9 미만
어학점수 (LANGUAGE)	D	A : 990-900 B : 899-800 C : 799-700 D : 699-600 E : 599 미만
자격증유무 (LICENSE)	N	Y : 있다 N : 없다
어학연수경험유무 (EXP_LANGUAGE_YN)	Y	Y : 있다 N : 없다

본 시뮬레이션 에서는 입력되어진 데이터를 범주형 데이터로 정규화 시키고 정규화 된 데이터를 바탕으로 규칙을 만들어 준다. 표 4-6은 시뮬레이션을 위한 훈련(Training Data)데이터 이며 각 구분값에 따른 속성 값을 표시하고 있다. 기타의 내용은 각 필드가 갖는 속성 변수 값을 의미한다.

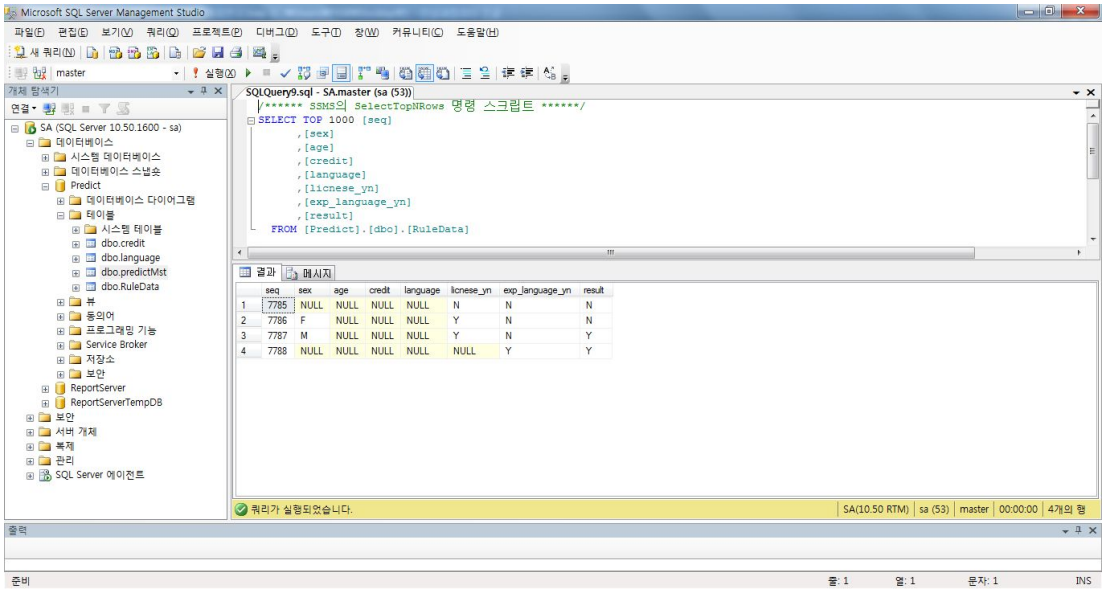


그림 4-7. 데이터베이스에 생성된 규칙생성결과

그림 4-7은 훈련용 데이터를 이용하여 규칙을 생성하여 데이터베이스 테이블(RuleData)에 저장된 데이터를 나타낸다.

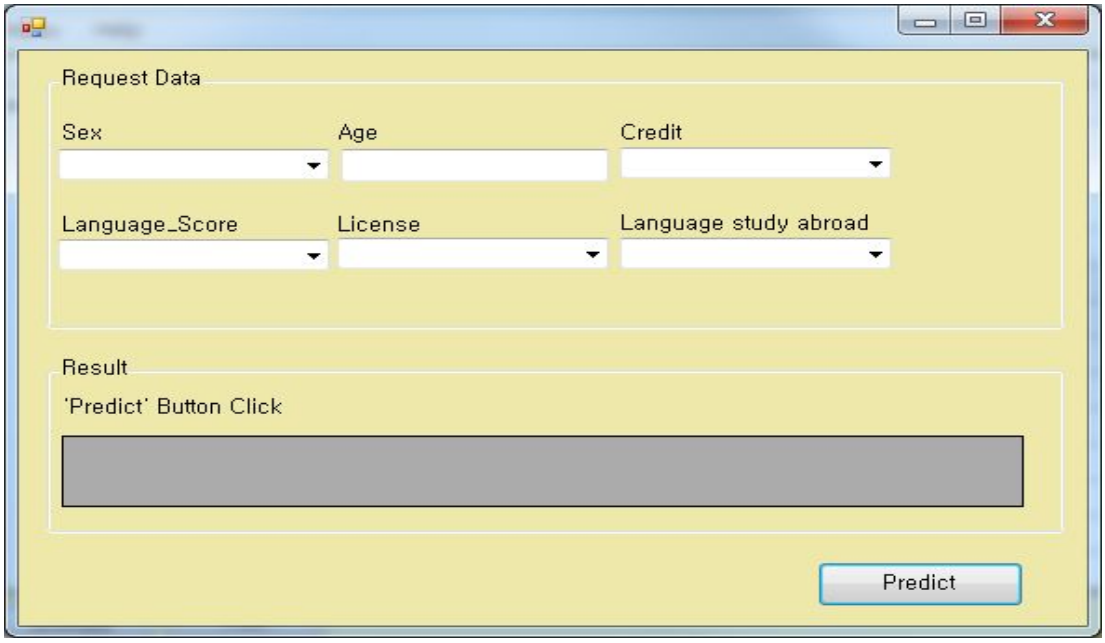


그림 4-8. 사용자 예측요청 데이터 입력화면

그림 4-8은 사용자 예측요청 데이터를 입력하는 화면이다. 예측을 위해 개인의 정보를 입력하면 이미 트리생성을 통해 구성된 규칙을 조회 한 후 가장 부합하는 결과를 그림 4-9와 같이 보여준다.

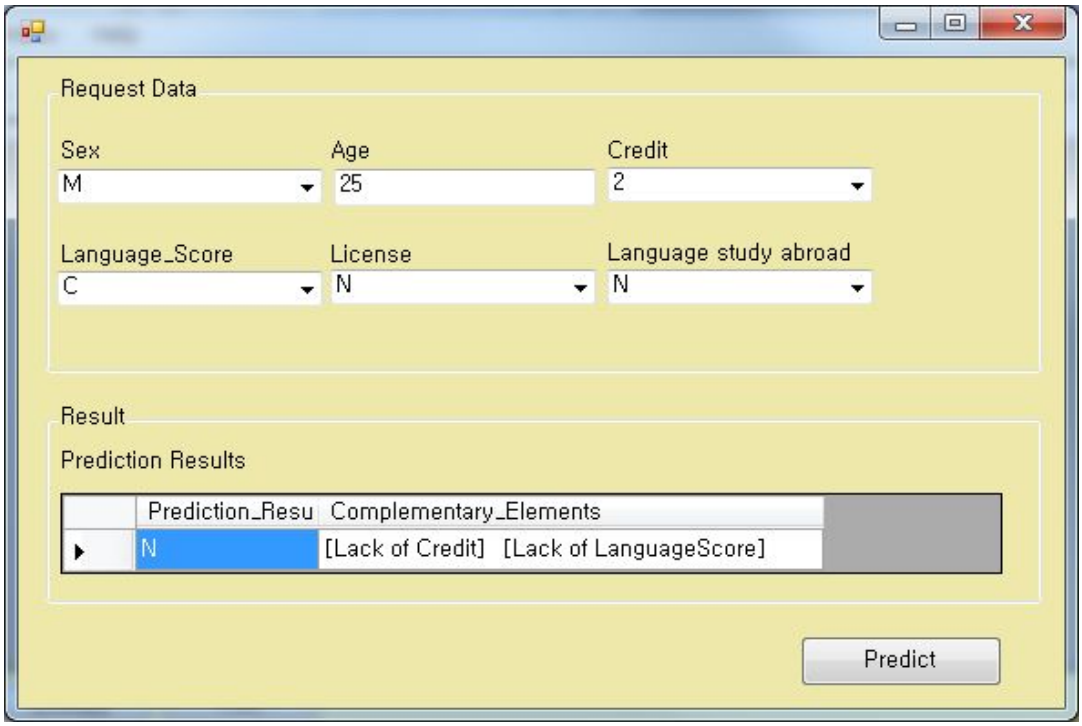


그림 4-9. 사용자 예측요청 데이터 결과화면

위와 같은 조건으로 규칙에서 학점과 어학점수가 부족하여 결과에 취업이 불가능하다는 결과를 출력 해주고 있다.

3. 취업 상황 예측결과

그림 4-10, 4-11는 생성된 규칙을 적용하여 취업상황을 예측한 화면이다. 그림 4-10은 입력된 데이터 성별 : M, 나이 : 25, 학점 : 2, 어학점수 : E, 자격증유무 : N, 어학연수경험 : Y에 따라 결과 'N' 예측 하였으며, 그림 4-11는 입력된 데이터 성별 : M, 나이 : 26, 학점 : 2, 어학점수 : A, 자격증유무 : Y, 어학연수경험 : Y에 따라 결과 'Y' 예측 하였다.

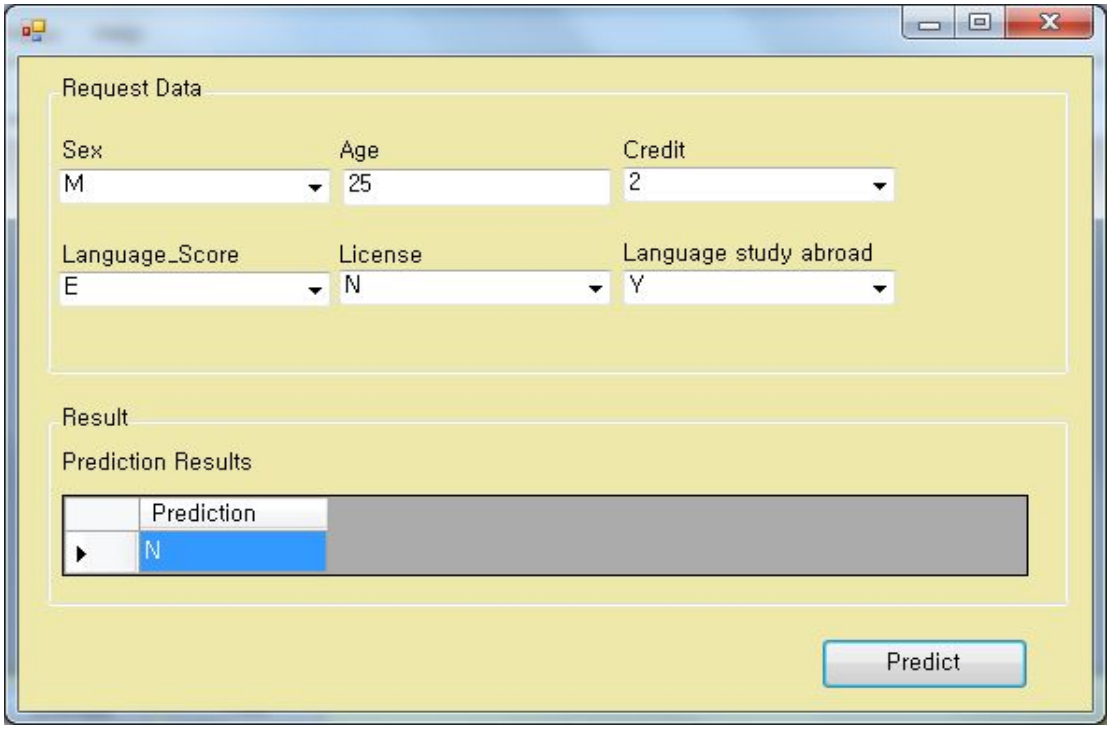


그림 4-10. 취업 예측결과(예1)

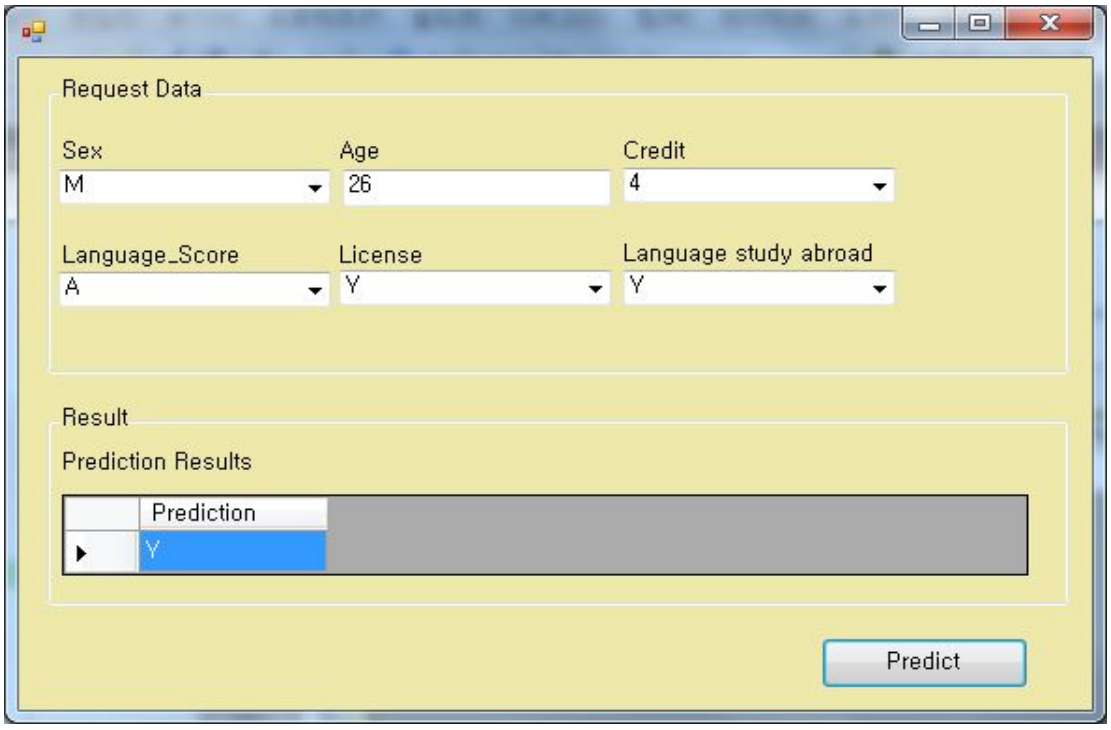


그림 4-11. 취업 예측결과(예2)

4. 취업상황 예측 기반의 보완요소 추천결과

트리생성 과정을 통해 규칙이 생성되고, 취업상황에 대한 결과를 예측 할 수 있다. 하지만 만약 취업이 불가능한 결과가 도출 될 경우, 어떠한 이유에서 취업이 불가능한지 부족한 부분이 무엇인지를 제공하는 기능이 필요하다. 그림 4-12는 요청데이터에 의해 취업이 가능성을 예측하고, 시뮬레이션한 예측 결과 화면이다. 만약 취업이 불가능한 경우 또는 면접이나 서류전형을 통과하는 경우라면 취업을 위해 학생들은 어떠한 부분이 부족한지를 알려주기 위해 규칙에 생성된 정보에 의해 부족한 부분을 그림 4-13과 4-14에서 보여주고 있다.

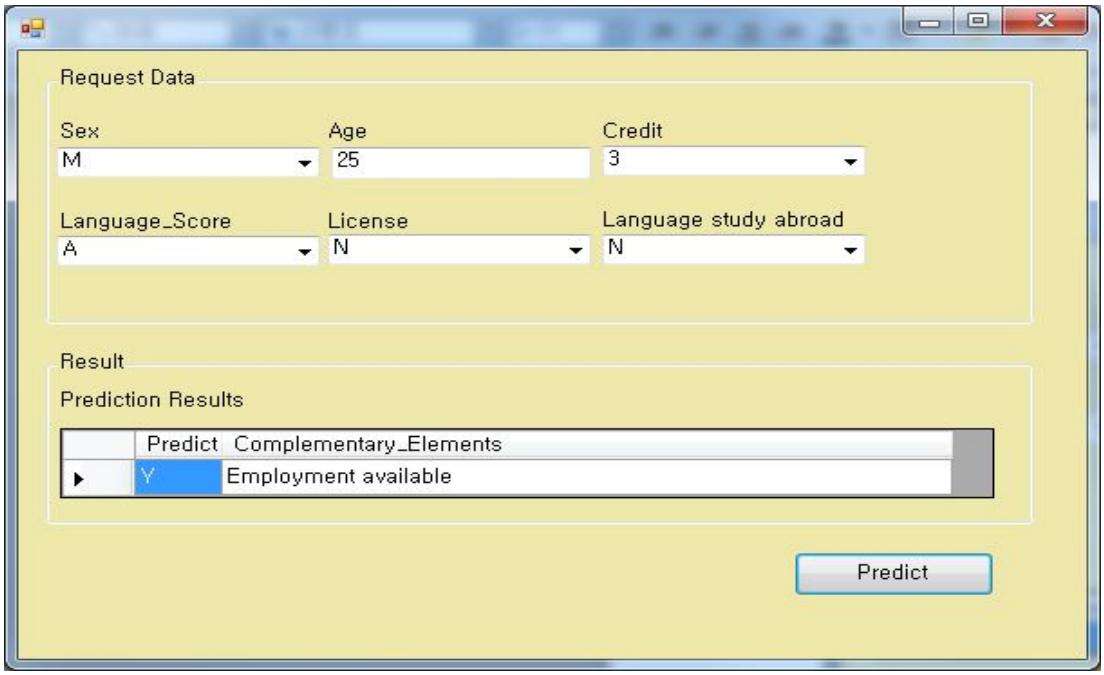


그림 4-12. 요청 데이터에 의해 취업 결과를 예측한 화면

그림 4-13은 취업 불가능인 취업을 위해 학점과 어학점수가 부족하다고 보충하여 주고 있으며, 그림 4-14는 A(면접통과)가 결과로 나왔으며 면접을 통과하는 조건은 되지만 취업을 위해 자격증을 보완할 것을 추천하고 있다.

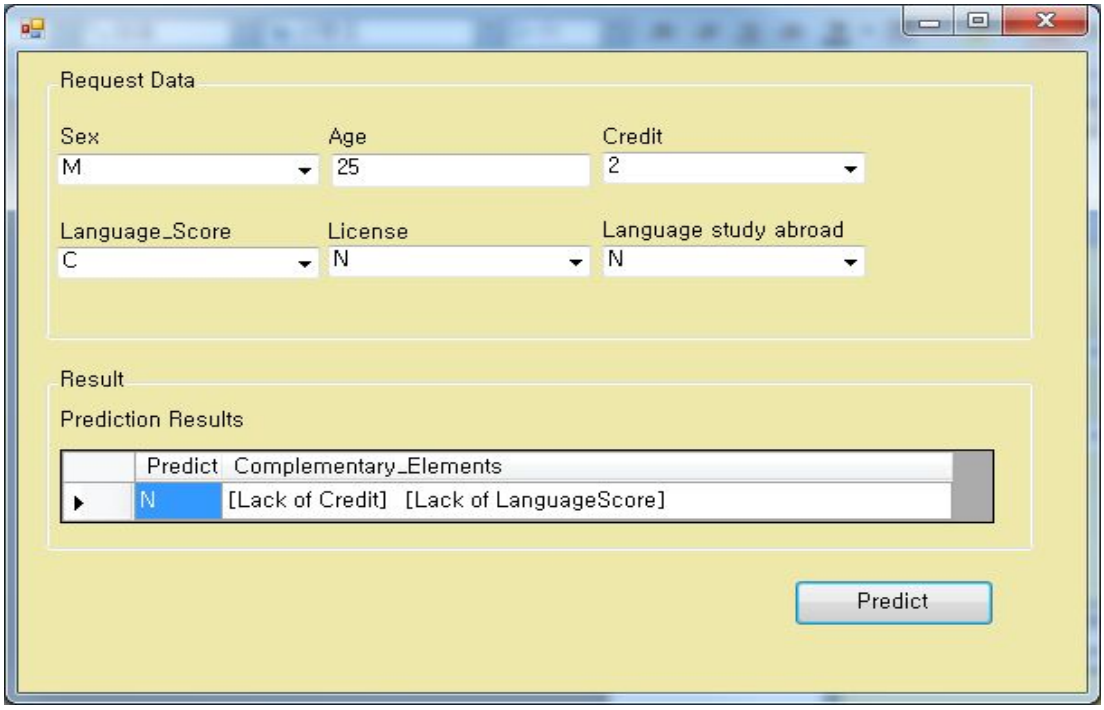


그림 4-13. 요청 데이터에 취업불가인 경우 보완요소를 제공하는 화면

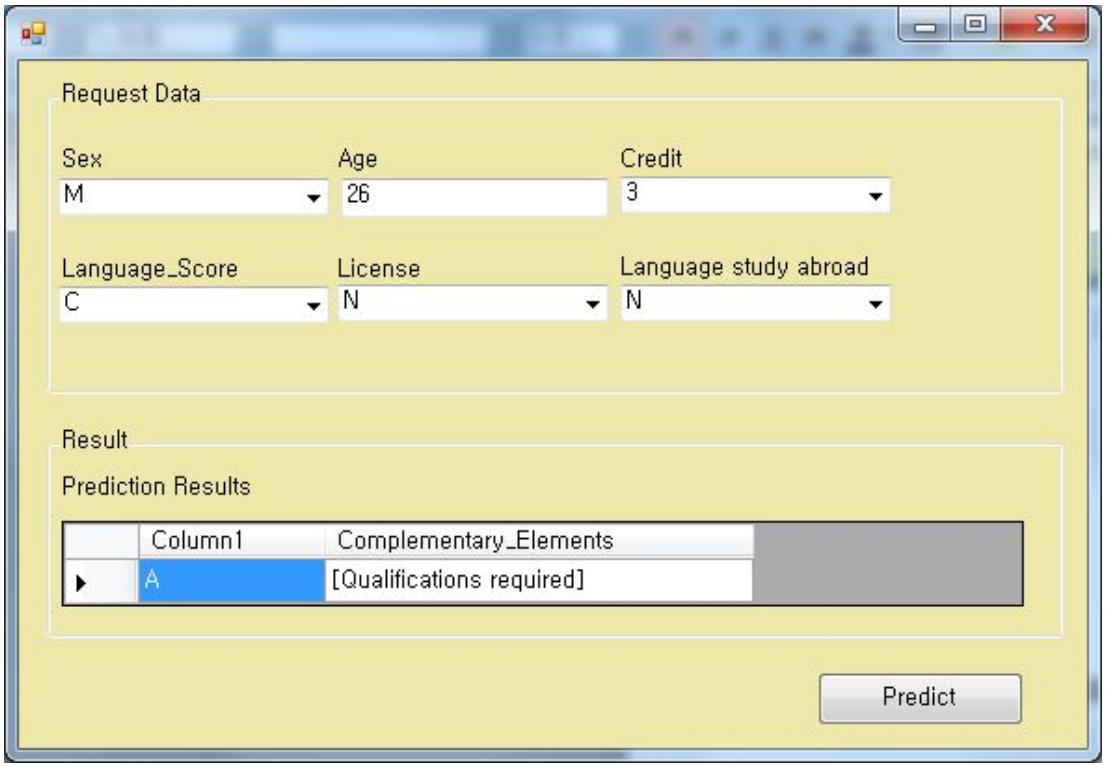


그림 4-14. 요청 데이터에 면접가능인 경우 보완요소를 제공하는 화면

5. 성능평가 및 분석

분류모형이 평가는 일반적으로 훈련용 데이터에 의해 만들어진 모형함수를 시험용 데이터에 적용하였을 때 나타나는 분류의 정확도를 이용하게 된다. 평가를 위해 1000개의 데이터를 이용하여 훈련용 데이터(500)개와 시험용 데이터(500)를 이용한다.

표 4-7. 시험용 데이터의 실제집단과 분류된 집단의 결과 분류

		분류된 집단	
		G_1	G_2
실제집단	G_1	f_{11}	f_{12}
	G_2	f_{21}	f_{22}

여기서 G1와 G2는 분류된 집단을 의미하고,

f_{11} 은 실제집단 G_1 와 분류집단 G_1 으로 정상 분류한 수이고,
f_{12} 은 실제집단 G_1 와 분류집단 G_2 으로 비정상 분류된 수이고,
f_{21} 은 실제집단 G_2 와 분류집단 G_1 으로 비정상 분류된 수이고,
f_{22} 은 실제집단 G_2 와 분류집단 G_2 으로 정상 분류한 수이다.

정의된 분류모형의 정확도(accuracy)는 전체 데이터의 수 중 올바르게 분류된 수의 비율이고, 오류율(error rate)은 오분류 된 수의 비율이다.

표 4-8. 시험용 데이터의 실제집단과 분류된 집단의 결과 분류 결과표

		분류된집단	
		G1	G2
실제집단	G1	14	39
	G2	6	441

$$\text{정확도(accuracy)} = \frac{f_{11}+f_{22}}{n} = \frac{14+441}{500} = 0.91 \quad \text{(식 5)}$$

$$\text{오류률(error rate)} = \frac{f_{12}+f_{21}}{n} = \frac{39+6}{500} = 0.09 \quad \text{(식 6)}$$

표 4-8을 이용하여 식5과 식6를 이용하여 정확도와 오류율을 구할 수 있다. 정확도(accuracy)는 0.91 이며, 오류율(error rate)는 0.09 로 정확도와 오류율은 집단 G_1과 G_2의 오분류(f12, f21)에 대한 위험성이 동일하다는 가정에서 합리적인 측도라 볼 수 있지만 현실문제에서 오분류에 대한 집단별 위험성이 다를 수 있으므로 오분류의 위험성이 서로 다른 경우 민감도(sensitivity), 특이도(specificity) 및 정밀도(precision)라는 측도를 이용한다.

$$\text{민감도(sensitivity)} = \frac{f_{11}}{f_{11}+f_{12}} = \frac{14}{14+39} = 0.264151 \quad \text{(식 7)}$$

$$\text{특이도(specificity)} = \frac{f_{22}}{f_{21}+f_{22}} = \frac{441}{6+441} = 0.986577 \quad \text{(식 8)}$$

$$\text{정밀도(precision)} = \frac{f_{11}}{f_{11}+f_{12}} = \frac{14}{14+6} = 0.7 \quad \text{(식 9)}$$

민감도는 실제 취업가능자를 취업가능으로 분류하는 비율이고, 특이도는 취업 불가자를 취업불가로 분류하는 비율, 정밀도는 취업가능자로 분류된 사람 중에서 실제 취업가능의 비율을 의미한다. 정확도(accuracy)는 민감도와 특이도의 가중합으로 표시 할 수 있다.

$$\text{정확도(accuracy)} = \frac{f_{11}+f_{12}}{n} \times (\text{sensitivity}) + \frac{f_{12}+f_{21}}{n} \times (\text{specificity}) \quad \text{(식 10)}$$

$$= \frac{14+39}{500} \times 0.264151 + \frac{6+441}{500} \times 0.986577 = 91\% \quad \text{(식 11)}$$

훈련용 데이터를 통해 생성된 분류모형을 시험용 데이터를 비교함으로서 정확도를 구하였고 그 결과는 정확도 91%로 비교적 정확한 분류가 되었음을 보여주고

있다. 위의 결과와 함께 평가를 위해 시험용 데이터의 변화에 따른 정확도를 살펴보고자 500개의 데이터를 100개단위로 나누어 정확성을 평가하였고, 그림 4-15은 예측정확도 변화를 나타내고 있다.

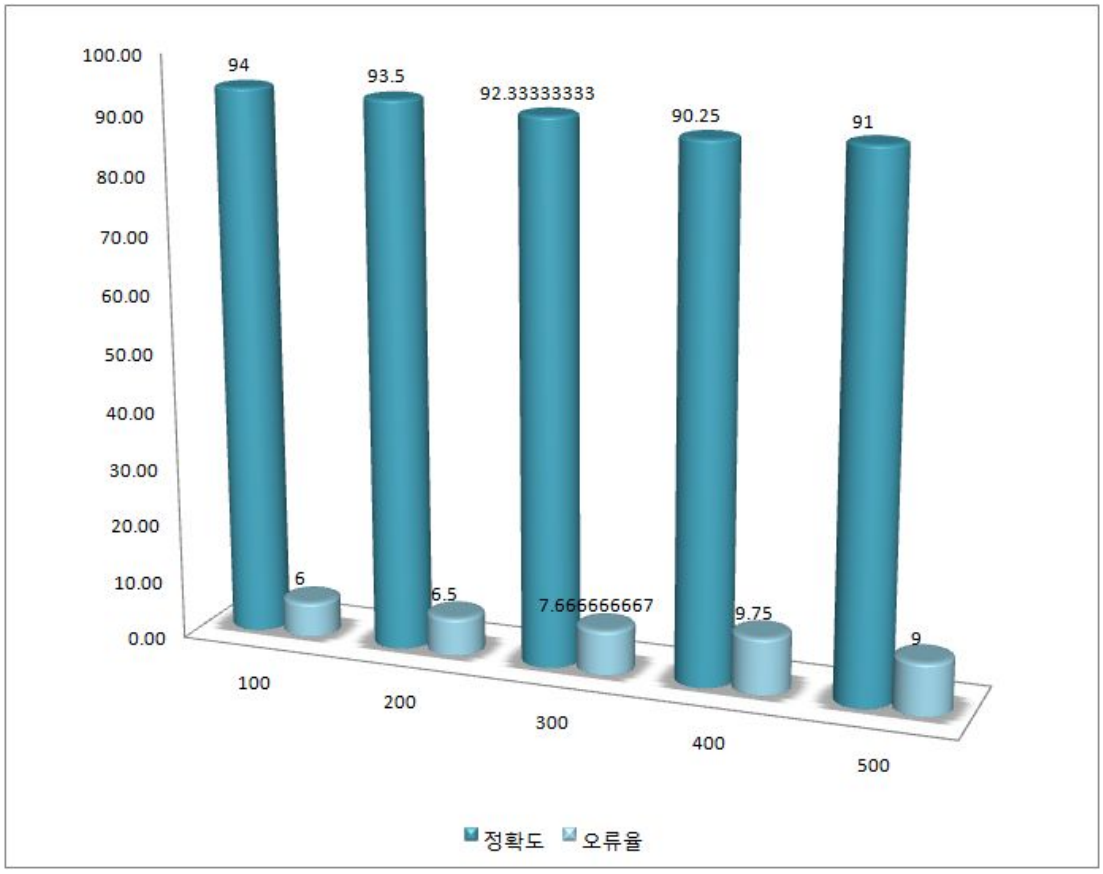


그림 4-15. 데이터변화에 따른 정확도와 오류율

표 4-9. 시험용 데이터를 통한 정확도와 오류율 변화

	100	200	300	400	500
정확도	94	93.5	92.33333	90.25	91
오류율	6	6.5	7.666667	9.75	9

시험용 데이터를 100개 단위로 변화를 주어 데이터가 100개 일 때 정확도는

94%이고, 500개 일 때 91%로, 데이터가 늘어날수록 정확도가 조금씩은 떨어졌지만 90%이상을 유지하며 비교적 높은 정확도를 나타냈으며 반대로 오류율은 적은수가 상승하는 것을 볼 수 있었다.

Ⅶ. 결 론

앞으로의 미래사회에서는 많은 분야에 걸쳐 수년간에 걸쳐 축적된 대용량 데이터를 효과적으로 처리되고 이 데이터를 기반으로 한 다양한 지능적인 서비스가 개발 될 것이다. 이와 같은 데이터처리 패러다임의 변화는 교육이나 학사시스템에서도 예외가 아닐 것이며, 각 대학에서 큰 이슈가 되는 졸업이후 취업에 대한 정보는 중요성이 증대 할 것으로 생각 된다.

이에 본 논문에서는 이러한 대용량 데이터처리 이슈와 문제들을 학생취업 관점에서 의미 있는 데이터 및 서비스를 제공하기위해 취업 상황 예측 알고리즘을 제안하고, 취업을 위한 학생의 보완 요소를 제공하는 방안을 연구하고, 제안 해 볼 수 있는 방안을 연구하였고, 기존의 무의미한 데이터에서 의미 있는 데이터로 그리고 단순히 데이터를 저장, 조회, 처리하는 시스템에서 데이터를 기반으로 결과를 예측 및 보완 알고리즘을 제안 한다.

이를 위해 취업 관련 학생 데이터를 범주에 맞는 데이터로 정규화 하였고, 정규화 된 데이터를 기반으로 알고리즘을 이용하여 규칙을 생성하고, 생성되어진 규칙을 기반으로 예측하고, 취업 보완요소를 식별하고 제시 하였다.

규칙을 생성하기 위한 기초 데이터에 대한 논의가 필요하고 좀 더 유용한 시스템 서비스로 발전하기 위해서는 전문가의 고견이 더 필요할 것으로 사료되나 향후 학사정보시스템에서 나아가야 할 시스템 방향과 알고리즘을 활용한 서비스 모델의 방향을 제시 한다고 사료 된다.

참고문헌

[1] Lucian Vintan , Arpad Gellert , Jan Petzold , Theo Ungerer, Person Movement Prediction Using Neural Networks, In First Workshop on Modeling and Retrieval of Context, 2004

[2] Jan Petzold, Andreas Pietzowski, Faruk Bagci, Wolfgang Trumler, and Theo Ungerer. Prediction of indoor movements using bayesian networks. In Proceedings of the First international conference on Location- and Context-Awareness (LoCA'05), 2005.

[3] Christian Voigtmann and Klaus David, A Survey To Location-Based Context Prediction, In Proceedings of the First Workshop on recent advances in behavior prediction and pro-active pervasive computing (AwareCast), June 2012.

[4] Nazerfard, Ehsan Cook, Diane J. Bayesian Networks Structure Learning for Activity Prediction in Smart Homes, 8th International Conference on Intelligent Environments (IE), 2012.

[5] Parkka, J, Ermes, M.; Korpipaa, P.; Mantyjarvi, J.; Peltola, J.; Korhonen, I.; , "Activity classification using realistic data from wearable sensors," Information Technology in Biomedicine, IEEE Transactions on , vol.10, no.1, pp.119-128, Jan. 2006.

[6] Jaeyoung Yang, Joonwhan Lee, and Joongmin Choi, Activity recognition based on RFID object usage for smart mobile devices. Journal of Computer Science and Technology, March 2011.

[7] Emmanuel Tapia, Stephen Intille, and Kent Larson. Activity recognition in the home using simple and ubiquitous sensors pervasive computing, Lecture Notes in Computer Science, Berlin, Heidelberg, 2004.

[8] Chao Chen and Diane J. Cook, Behavior-based Home Energy Prediction, Eighth International Conference on Intelligent Environments, 2012.

[9] Christian Beckel, Leyna Sadamori, Silvia Santini, Towards Automatic

Classification of Private Households Using Electricity Consumption Data. 4th ACM Workshop On Embedded Sensing Systems For Energy-Efficiency In Buildings (BuildSys'12). Toronto, Canada. November 2012

[10] S.A. Pourmousavi Kani, M.M. Ardehali, Very short-term wind speed prediction: A new artificial neural network–Markov chain model, Energy Conversion and Management, Volume 52, Issue 1, January 2011, Pages 738-745,

[11] Vladimir Cherkassky, Sohini Roy Chowdhury, Volker Landenberger, Saurabh Tewari, and Paul Bursch, Prediction of Electric Power Consumption for Commercial Buildings, Proceedings of International Joint Conference on Neural Networks, 2011.

[12] A. Badri, Z. Ameli, A.Motie Birjandi, Application of Artificial Neural Networks and Fuzzy logic Methods for Short Term Load Forecasting, Energy Procedia, Volume 14, 2012, Pages 1883-1888, 2011.

[13] Ren´e Schumann, Dominique Genoud, Demand forecasting and smart devices as building blocks of smart micro grids, Sixth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, 2012.

[14] Smitha.T, V.Sundaram, Classification 규칙s by Decision Tree for Disease Prediction, International Journal of Computer Applications, Volume 43 – No.8, April 2012

[15] Osmar R. Zaiane, Maria L. Antonie, and Alexandru Coman. Mammography classification by an association 규칙-based classifier, International Workshop on Multimedia Data Mining (with ACM SIGKDD) 2002.

[16] Mantzaris, D.H, Anastassopoulos, G.C, Lymberopoulos, D.K, "Medical disease prediction using Artificial Neural Networks," , 8th IEEE International Conference on BioInformatics and BioEngineering 2008.

[17] Hani Neuvirth, Michal Ozery-Flato, Jianying Hu, Jonathan Laserson, Martin S. Kohn, Shahram Ebadollahi, and Michal Rosen-Zvi. Toward personalized care management of patients at risk: the diabetes case study. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data

mining (KDD '11), ACM, New York, NY, USA.

[18] Hyunchul Ahn, Kyoung-jae Kim, Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach, Applied Soft Computing, Volume 9, Issue 2, March 2009.

[19] 이정진, "R, SAS, MS-SQL을 활용한 데이터 마이닝", 2011, 8 p 190-211

[20] http://ai_times.tistory.com

[21] 강명석, 김학배, "스마트 홈 환경에서 데이터마이닝 기법을 이용한 지능형 서비스 추론 모델", 한국통신학회논문지 제32권 제12호 (2007년 12월)-네트워크 및 서비스 pp.767-778 1226-4717 KCI