ISBN 978-89-294-***-*-93300

2015년도 기술 및 정책보고서

중소기업 R&D 정보지원 수요 영향 요인 예측 모형: 다변량 분석을 통한 영향 요인 분석

전 승 표

ISBN 978-89-294-***-*-93300

머 | 리 | 글

한국과학기술정보연구원(KISTI), 산업정보분석실에서는 2015년도 한 해를 마무리 지으며 연구원들이 그동안 관심 분야에 대해 정보들을 분석한 '정보분석 보고서'를 발간합니다. 개별 연구자의 관심분야가 다른 만큼 다양한 과학기술과 산업시장에 대한 분석 정보를 담고 있으며, 특히 중소기업 R&D 정보(기획) 정보가 필요한 곳에 조금이라도 도움이 되었으면 하는 바램입니다.

본 보고서는 중소기업을 위한 R&D 정보(또는 기획) 지원 사업의 사용자인 중소기업 이해에 일조할 수 있는 새로운 요인들과 방안을 KISTI 전승표 책임연구원의 관점에서 분석, 연구하였습니다. R&D 정보지원을 필요하거나 이미 경험한 기업의 프로파일을 분석함으로써 특정 중소기업이 R&D 정보지원을 필요할지 예측할 수 있는 방향을 제시하기 위해, 전통적인 회귀분석에서 벗어나 다변량 분석을 통해 영향 요인을 도출하는데 그 주안점을 두고 있습니다.

끝으로 본 보고서는 전승표 책임연구원이 집필한 것으로 노고에 깊이 감사드리며, 본 보고서에 수록된 내용은 연구자 개인의 의견으로서 한국과학기술정보연구원의 공식의견이 아님을 밝혀두고자 합니다.

2015년 11월

한국과학기술정보연구원

원 장 한 선 화

ISBN 978-89-294-***-*-93300

목 차

Abstract ····· 1

제1장 서론 ····· 2

1.1 연구 배경 및 필요성 ····· 2
1.2 중소기업 기술혁신 지원 정책 ····· 3
1.3 다변량 분석과 프로파일링 ····· 5

제2장 판별분석과 의사결정나무 분석 방법론 ····· 7

2.1 판별분석 ····· 7
2.1.1 거리 측정 ····· 7
2.1.2 분류 성과 ····· 9
2.1.3 판별분석의 장단점 ····· 11
2.2 의사결정나무 분석 방법론 ····· 12
2.2.1 분류나무 ····· 13
2.2.2 회귀나무 ····· 19
2.2.3 방법론의 장단점과 추가 고려사항 ····· 21

제3장 선행연구 분석 ····· 24
3.1 중소기업 기술혁신 지원 정책과 효과 연구 ····· 24
3.2 중소기업과 기술기획에 관한 연구 ····· 26
3.3 중소기업 국내 기술기획 현황 ····· 28
3.4 중소기업 기획(또는 정보) 지원 관련 사업 현황 ····· 31
3.5 다변량 분석(판별분석 및 나무분석) 활용 선행 연구 ····· 33

제4장 다변형 분석을 통한 영향 요인 도출 ····· 37

4.1 연구 사례 ····· 37
4.2 변수 설정 ····· 39
4.3 기술 통계 ····· 42

4.4 영향 요인 분석(의사결정나무 분석) ····· 43
4.4.1 개발기획 단계 지원 선호 ····· 43
4.4.2 기획지원 필요성 ····· 45
4.4.3 정보지원 필요성 ····· 47
4.4.4 기획지원 경험 ····· 49
4.4.5 정보지원 경험 ····· 51

제5장 수요 기업 예측 모델 ····· 54

5.1 개발기획 단계 지원 선호군 예측 모델(프로파일링) ····· 54
5.2 개발기획 지원 고수요군 예측 모델(프로파일링) ····· 60
5.3 기술정보 지원 고수요군 예측 모델(프로파일링) ····· 65

제6장 결론 및 시사점 ····· 70

6.1 결론 ····· 70
6.2 연구의 시사점 ····· 72

참고문헌 ····· 74

표 목 차

<표 1> 다변량 분석 방법의 장단점 비교 ····· 22
<표 2> 변수명과 변수 설명 ····· 41
<표 3> 범주형 변수의 기술통계량 ····· 42
<표 4> 연속형 변수의 기술통계량 ····· 43
<표 5> 모형의 분류 정확도(개발기획 단계 지원 선호) ····· 45
<표 6> 모형의 분류 정확도(기획지원 필요성) ····· 47
<표 7> 모형의 분류 정확도(정보지원 필요성) ····· 49
<표 8> 모형의 분류 정확도(기획지원 경험) ····· 51
<표 9> 모형의 분류 정확도(정보지원 경험) ····· 53
<표 10> Wilks's lamda(개발기획 단계 지원 선호) ····· 56
<표 11> 집단평균의 동질성에 대한 검정(개발기획 단계 지원 선호) ····· 56
<표 12> 단계별 판별분석 결과(개발기획 단계 지원 선호) ····· 57
<표 13> 정준 상관 계수(개발기획 단계 지원 선호) ····· 57
<표 14> 표준화 정준판별함수 계수와 정준판별함수 계수(개발기획 단계 지원 선호) ····· 57
<표 15> 구조행렬(개발기획 단계 지원 선호) ····· 58
<표 16> 분류 함수 계수(개발기획 단계 지원 선호) ····· 59
<표 17> 모형의 분류 적중률(개발기획 단계 지원 선호) ····· 60
<표 18> Wilks's lamda(개발기획 지원 고수요) ····· 61
<표 19> 집단평균의 동질성에 대한 검정(개발기획 지원 고수요) ····· 61
<표 20> 단계별 판별분석 결과(개발기획 지원 고수요) ····· 62
<표 21> 정준 상관 계수(개발기획 지원 고수요) ····· 62
<표 22> 표준화 정준판별함수 계수와 정준판별함수 계수(개발기획 지원 고수요) ····· 62
<표 23> 구조행렬(개발기획 지원 고수요) ····· 63
<표 24> 분류 함수 계수(개발기획 지원 고수요) ····· 64
<표 25> 모형의 분류 적중률(개발기획 지원 고수요) ····· 64
<표 26> Wilks's lamda(기술정보 지원 고수요) ····· 65
<표 27> 집단평균의 동질성에 대한 검정(기술정보 지원 고수요) ····· 66
<표 28> 단계별 판별분석 결과(기술정보 지원 고수요) ····· 66
<표 29> 정준 상관 계수(기술정보 지원 고수요) ····· 66
<표 30> 표준화 정준판별함수 계수와 정준판별함수 계수(기술정보 지원 고

수요) ····· 67
<표 31> 구조행렬(기술정보 지원 고수요) ····· 67
<표 32> 분류 함수 계수(기술정보 지원 고수요) ····· 68
<표 33> 모형의 분류 적중률(기술정보 지원 고수요) ····· 69
<표 34> 예측모형별 판별함수의 중심점 ····· 71

그 림 목 차

<그림 1> 기술요소별 세계대비 기술역량 자가 진단 결과 ····· 29
<그림 2> R&D 기획단계 정부 지원 활용 여부 비교 ····· 30
<그림 3> R&D 기획 지원 필요성(5점 매우 높음) 응답 결과 비교 ····· 30
<그림 4> R&D 개발 단계별 지원 선호도 응답 결과 비교 ····· 31
<그림 5> 의사결정나무 분석법을 활용한 우울 노인의 특성 분석 사례 ···· 35
<그림 6> 연구사례의 업종별 분포 ····· 38
<그림 7> 연구사례의 종사자 수(좌) 및 기업 업력(우) 분포 ····· 38
<그림 8> 연구사례의 매출액 규모별 분포 ····· 39
<그림 9> 개발기획 단계 지원 선호군에 대한 의사결정나무 분석결과 ····· 44
<그림 10> 기획지원 고수요군에 대한 의사결정나무 분석결과 ····· 46
<그림 11> 정보지원 고수요군에 대한 의사결정나무 분석결과 ····· 48
<그림 12> 개발기획 정부지원 경험군에 대한 의사결정나무 분석결과 ····· 50
<그림 13> (기술)정보 정부지원 경험군에 대한 의사결정나무 분석결과 ···· 52
<그림 14> 개발기획 단계 지원 선호군에 대한 의사결정나무 분석결과(변환 변수 사용) ····· 55

본 연구의 결과는 2015년 당해 게재된 하기 논문의 이론적 배경이 일부 활용되었음을 밝힌다.

1. *전승표(단독저자), "중소기업 R&D 정보지원과 성과의 관계에 대한 연구: ICT 기업을 중심으로"*, 2015 한국기술혁신학회 추계학술대회
2. *서주환, 성태응, 김유일, 전승표,* "잠재성장모형을 활용한 중소기업 R&D 기획 지원 효과 분석", 2015 한국기술혁신학회 춘계학술대회.

Abstract

우리나라를 비롯한 많은 선진국이나 개발도상국에서는 중소기업 지원을 위해서 규제, 유도, 정보 지원 등 다양한 정책을 통해 중소기업의 기술혁신을 지원하고 있으며, R&D 투자 효율성 제고를 위해서 다양한 연구를 진행하고 있다. 본 연구보고서는 직접 지원 중심의 R&D 투자 효율성을 높이는 제고하기 위한 간접 지원의 한 가지 정책인 R&D 정보지원 정책에 주목했다.

본 연구에서는 기존 이론 중심의 연구에서 벗어나 R&D 정보지원의 수요자에 보다 집중했다. 이론 중심의 연구들에서 미처 밝히지 못하거나 간과했던 요인들이 무엇인지 밝히기 위해 최근 각광받고 있는 데이터 마이닝을 통한 분석 방법을 적용해 수요의 특징을 분석(프로파일링)했으며, 새로운 기업에 대해 R&D 정보제공의 수요가 얼마 높을지 예측하는 모형을 제시했다. 제시된 모형에 따르면 기존에 주목했던 기업의 경제적 특성(매출액 등)보다는 연구개발 기간이나 기술적 특성이 R&D 정보제공 수요를 결정하는데 중요한 요인으로 제시되었다. 또한 R&D 기획 지원 수요 집단과 (기술)정보제공 지원 수요 집단의 프로파일링 결과가 다르게 나타나기도 했다. 따라서 본 연구의 결과에 따르면 기존의 정보 지원 사업에 신청한 기업의 선정 방법에 변화가 필요한 것으로 제안할 수 있었다.

본 연구에서는 이상의 연구결과 도출을 위해서 제1장에서는 먼저 중소기업 기술혁신 지원 정책에 대해 살펴보고서 본 연구 활용하는 다변량 분석과 프로파일링에 대해 살펴본다. 제2장에서는 분석 방법론인 판별분석과 의사결정나무 분석방법에 대해 설명하고, 제3장에서는 중소기업 기술혁신, 기술기획, 정보지원 활동, 다변량 분석(판별분석과 나무분석)과 관련한 선행연구 분석을 살펴본다. 제4장에서는 의사결정나무 분석을 활용해서 중소기업 R&D 정보지원 영향 요인을 도출한다. 제5장에서는 판별분석을 활용해서 R&D 정보지원 수요 기업을 예측할 수 있는 모형을 제시하고 제6장에서는 결과와 정책적 시사점을 제시한다.

제 1 장 서 론

1.1 연구 배경 및 필요성

우리나라에서 중소기업은 2012년 말 현재 한국 기업체 수의 99.9%, 취업자 수의 87.7%를 차지하고 있어 국가 경제에서 중심적인 역할을 하고 있다(중소기업청 2014). 이와 같이 중소기업은 한국 경제에서 차지하는 비중이 높음에도 불구하고 재원 조달, 판로 개척, 기술혁신 등을 하는데 있어 대기업보다 열위에 있는 것이 사실이다. 특히, 세계적인 경기침체와 환율전쟁의 여파로 우리나라에서는 많은 중소기업이 대기업보다 큰 고통을 겪고 있으며, 이러한 중소기업의 애로를 지원하기 위한 정부의 노력도 더욱 중요해졌다. 현재 한국 정부는 조세감면, 금융지원 등의 다양한 중소기업 지원 시책을 마련·시행하고 있지만, 일부 정책은 실효성 논란도 제기되고 있다. 이런 배경에서 기술 중심의 중소기업 경쟁력을 근본적으로 높일 수 있는 기술혁신 지원 정책은 한국 중소기업청의 핵심적 지원 정책으로 자리 잡아왔다. 그러나 확대되는 중소기업 기술혁신 정책에도 불구하고 효과적인 정부 지원정책을 수립하기 위한 연구들은 부족했다(박찬수 외 2012).

중소기업의 기술혁신을 지원하기 위한 정책은 크게 세 가지로 구분해 볼 수 있는데, 1) 법/규제 및 세제 지원과 관련된 규제 정책, 2) 기술개발 자금 지원 및 판로 지원과 같은 유인 정책, 3) 기술정보나 기술 인력을 지원하는 정보제공 정책이다(류숙원 외 2010). 본 연구에서는 이런 정책적 도구 중에서 중소기업의 기술혁신 지원을 위한 정보 지원 노력에 주목한다. 정보 지원 정책과 관련해서 최근 발표된 '제3차 중소기업 기술혁신 촉진계획(안)'을 살펴보면, 현 정부는 개별기업의 기술경쟁력 제고를 위해서 중소기업 신기술(6T)의 투자를 늘리고, R&D 기획역량을 강화시키며, 다양한 수요에 대응할 수 있는 정보제공 사업을 진행할 예정이다(국가과학기술심의회 2014). 최근 정부의 중소기업 기술혁신 지원 정책은 연구개발 또는 사업화를 위한 자금 지원이나 세제지원과 같은 직접 지원 정책을 지양하고, 출연연과 같은 다양한 기관을 활용한 정보 제공과 같은 간접 지원 정책을 확대하고 있다. R&D

정보 지원은 R&D를 위한 정보를 제공해 기술개발 투자를 늘리고, 성공률을 제고하려는 노력의 일환으로 수행되어 왔으며, 대표적인 사업으로 기획단계의 정보지원 사업인 중소기업청의 '중소기업 R&D 기획지원 사업' 이나 개별 지자체들의 기술정보 지원 사업 등이 있다(중소기업중앙회 2014). 2014년 중소기업 기술통계 조사 결과에 따르면, 2200개 설문 대상 기업 중에서 2013년에 124개(5.6%)의 기업이 정보조사 지원 사업을 경험한바 있고, 298개(13.5%)의 기업이 R&D 기획을 위한 자금 지원을 경험한바 있다(중소기업청·중소기업 중앙회 2014). 그러나 R&D 정보 지원의 효과나 관련 기관의 역할에 대해 거의 연구된 바가 없다.

본 보고서에서는 이런 R&D 정보 지원과 중소기업의 여러 가지 조건들이 어떤 관계가 있는지 탐색적으로 연구하고자 한다. 정보 지원과 성과를 면밀히 분석하여 설명적 연구나 인과관계 분석하기 위해서는 수년간 지원 결과는 추적조사 하거나(윤요한 외 2015) 선택 편의를 극복하기 위한 다양한 장치가 필요하다(이종민 외 2013). 그러나 본 보고서는 이런 추적 조사 연구에 앞서 정책에 활용가능한 시사점을 제공하고자 했다. 먼저 R&D 정보 지원을 경험한 기업들(실험집단)의 특징을 살펴보아 프로파일링하고 나아가 주요 수요 집단으로 예상되는 R&D 기획 지원 고수요 집단의 특징도 프로파일링(영향 요인)했다. 여기서 도출된 결과를 통해 향후 중소기업 기술혁신을 위한 기획지원이나 정보 제공이 목표로 해야 할 고객은 어떤 특징을 가지고 있는지 찾고, 도출된 영향 요인을 통해 예측하는 방법을 제시해 중소기업 R&D 지원 정책개발이나 평가에 일조하고자 한다.

1.2 중소기업 기술혁신 지원 정책

우리나라의 중소기업정책은 1960년대 도입되어 1970~80년대 대기업 위주의 경제발전 전략에서 따라 대기업-중소기업간의 계열화 구도를 통해 보호 및 육성 중심의 정책이었다. 1990년대 들어 WTO 출범에 따른 경쟁도입과 구조고도화 정책 중심으로 추진되다가 IMF 위기 이후 벤처창업 촉진정책 및 중소기업 경쟁력 강화를 위한 기반조성 정책으로 변화하는 과정을 거쳤다. 국내외 경쟁 환경과 정책적 목표에 따라 중소기업 정책도 발맞춰 변모

해 왔다. 최근의 중소기업 정책은 대기업과 중소기업의 공정한 협력관계 형성을 통해 기업간 상생관계를 구축하여 상생발전을 유도하고 있으며, 중소기업을 지역과 연계된 혁신 주체로 인식하고 있다는 것이 과거 정책과의 가장 큰 차이점이다. 중소기업을 창조와 혁신의 주체로 인식하고 미래 성장 동력이라는 점이 강조되면서, 국가 성장 동력을 확보하기 위해 국제 경쟁력을 갖춘 혁신형 중소기업 육성 방향으로 가닥이 잡혀있다(서주환 외 2015).

혁신형 중소기업의 육성을 위해 우리나라의 중소기업 R&D 지원 정책은 직접 지원(R&D 보조금, 공공구매 등)과 간접 지원(조세 감면 등)이 적절히 균형을 이루고 있다. 이러한 지원 유형은 국가별로 상이한데 캐나다, 일본, 네덜란드는 간접 지원 비중이 높은 반면에 스웨덴, 핀란드, 독일 등의 국가는 조세 시스템을 통한 기업 R&D 간접 지원 정책 수단이 거의 없는 것으로 알려졌다. 2013년 우리나라 중소기업 지원정책은 13개 중앙부처와 16개 지방자치단체에서 개별적으로 기획·집행되고 있고, 관련 사업만도 1,123개에 다다르는 등 그 지원 범위와 내역사업의 수적인 측면에서는 포괄적이고 광범위하게 운영되고 있다(서주환 외 2015).

우리나라 중앙부처가 집행한 203개 사업을 유형별로 분류하면 1) 자금지원분야 41개 사업(5조2,956억원), 2) 기술 39개 사업(1조7,056억원),3) 인력 24개 사업(6,372억원), 4) 창업벤처 10개 사업(4,228억원), 5) 소상공인 6개(3,031억원), 6) 수출판로 26개 사업(1,559억원)등으로 구분할 수 있다. 기술지원 정책의 경우 정부 보조금 지원이 다수를 이룬다는 점에서 중소기업의 내부 역량강화보다는 중소기업의 부족한 파이낸싱을 보완하는 경향이 지배적이다. 이러한 자금 '투입' 중심의 지원정책이 창업→중소기업→중견기업→글로벌 전문기업 혹은 대기업으로 이어지는 자생적인 기업 생태계 조성에는 어떠한 영향을 미치고 있는지에 대한 효과성 분석이 필요한 상황이다(서주환 외 2015).

중소기업정책은 경제 환경 및 정책의 목표에 따라 다르게 시행되어 왔으며, 크게 자금지원, 기술지원, 인력지원, 판로지원, 정보화지원, 창업지원, 세제지원 등의 영역에서 시행되었다. 이런 지원들 중에서 본 보고서는 간접지원에 관심을 가지고 있으며, 특히 R&D 기획이나 개발을 위한 정보지원 정책에 초점을 맞추고 있다.

1.3 다변량 분석과 프로파일링

어떤 대상의 성격을 규명하기 위하여 한 가지 측면(한 개의 종속변수)에서 그 대상을 관찰하고 분석하는 것을 단변량 분석(univariate analysis)이라고 할 수 있다. 그러나 실제 우리 주변에는 인간의 행동과 같이 어떤 하나의 현상이나 자료를 가지고 설명하기 힘든 경우가 많다. 그러므로 인간의 행동을 좀 더 심층적이며 적절하게 설명하기 위해서는 여러 각도와 측면에서 수집한 자료를 동시에 그 관계성을 분석해 보아야 가능한 것이다.

통계적으로는 종속변수의 관계성(relationships)을 고려한 상태에서 여러 개의 단변량분석을 동시에 수행하는 것을 의미하며, 여기에서 사용되는 자료는 다차원적으로 정상성(multivariate normal distribution)이 가정되어야 한다. 이와 같이 여러 개의 독립변수에 대한 여러 개의 종속변수를 동시에 분석해 보는 통계적 방법을 다변량분석이라고 한다.

이와 같은 통계적 방법으로서 대표적인 것으로서는, 변수들 사이의 유사성을 찾아낼 수 있는 요인분석(factor analysis), 이미 설정되어 있는 피험자 집단의 적절성을 확인하는 판별분석(discriminant analysis), 독립변수의 수준에 따라 피험자를 집단으로 구분하는 군집분석(cluster analysis), 여러 개의 종속변수에 대한 분석을 위한 다변량분산분석(MANOVA), 요인분석과 회귀분석의 혼합형인 구조방정식모형(structural equation model), 여러 개의 변수집단들 사이의 상관계수를 추정하는 정준상관계수(canonical correlation), 의사결정규칙(decision rule)을 도표화하여 관심대상이 되는 집단을 몇 개의 소집단으로 분류(classification)하거나 예측(prediction)을 수행하는 분석을 위한 의사결정나무 분석(decision tree analysis) 등이 있다(이학식 외 2015).

이런 다변량 분석이 자주 활용되는 분야가 데이터 마이닝이다. 대용량 데이터로부터 유용한 정보를 추출하는 기술인 데이터마이닝은 오늘날 그 중요성이 점점 더 커지고 있다. 예를 들어 이메일 스팸 필터기의 일부는 데이터마이닝으로 생성된 규칙을 사용하고 있다. 여기서 데이터마이닝은 수백만 개

의 전자메일 메시지를 조사하여 스팸인지 여부를 판별하는 규칙을 생성시킨다(shmueli 외 2009).

본 보고서에서는 중소기업청에서 수행한 2014년도 '제7차 중소기업 기술통계조사 결과를 활용해서 R&D 기획이나 정보조사를 위해서 정부지원을 받는 기업과 받지 않는 기업을 분류하는 방법을 찾고자 한다. 또한 기획 지원에 정부지원이 필요하다고 강력하게 생각하는 그룹과 그렇지 않은 그룹을 분류하는 방법도 찾고자 한다. 이 결과는 보다 집중된 사업(정책)의 개선 방향 탐색이 가능하게 할 것이다. R&D 기획지원이나 정보지원이 필요한 기업과 그렇지 않은 기업을 구분할 수 있는 요인을 도출하면 예측 모델 수립이 가능하다. 즉 기획(정보) 지원 수요에 미치는 요인을 찾고 수요 기업을 예측하기 위한 프로파일링을 진행하게 된다.

여기서 프로파일링은 예측변수의 관점에서 각 집단 내 관찰치 간의 유사성을 찾거나 집단을 구별시켜 주는 요인을 찾는 활동을 말한다. 예를 들어, 회사가 효과적인 광고를 위해서 구매자 대상이 될 만한 집단에 관한 정보수집 활동도 여기에 해당된다. 주로 로지스틱 회귀분석이나 판별분석 등의 통계적 분류기법을 이용하여 이러한 작업을 수행한다(shmueli 외 2009). 본 보고서에서는 프로파일링을 위해서 다음 장에서 소개할 판별분석과 의사결정나무 분석을 활용한다.

제 2 장 판별분석과 의사결정나무 분석 방법론

2.1 판별분석[1)]

판별분석(discriminant analysis)은 분류기업 중의 하나로서 로지스틱 회귀분석과 같이 분류(classification)와 프로파일링(profiling)에 사용되는 전통적인 통계기법이다. 판별분석은 해당 항목들이 속해 있는 각 집단들을 위해 연속형 변수를 사용하며, 새로운 항목에 대해서는 이 집단들 중의 한 집단에 속하는 것으로 분류한다. 대표적인 적용사례를 살펴보면 대출, 신용카드, 보험가입 신청자들에 대해서 저/고 위험군으로 분류하는 경우 또는 신제품의 고객을 초기수용자, 조기다수자, 후기다수자, 지각수용자 등으로 분류하거나 채권을 채권등급별로 분류하는 경우 등이 있다. 그밖에도 판별분석은 인간 화석에 대한 두개골 분류, 연구에 대한 저자들의 논쟁, 대학입학관련 의사결정, 알코올 중독자와 비중독자에 대한 의학적 분류, 사람 지문의 확인 방법 등의 다양한 분류 의사결정에 주로 사용된다. 또한 판별분석은 집단을 구별시켜 주는 요인들을 찾아내기 위한 활동인 프로파일링에 사용되기도 한다(규모가 크면 관계가 복잡해 잘 나타나지 않을 수도 있음).

2.1.1 거리 측정

항목들을 가장 잘 분할시키는 분리점을 찾기 위해서는 해당 집단과 항목간의 거리를 측정해야 한다. 이에 대한 일반적인 개념은 각 항목을 해당 항목으로부터 가장 가까운 집단으로 분류하는 것이다.

이 경우 유클리드 거리 법칙(Euclidean distance rule)을 이용하여 소득을 대출 제안 수락의 예측변수로 사용할 수 있다. 만일 x가 대출제안을 거절한 집단의 평균소득보다 대출제안을 수락한 집단의 평균소득에 더 가깝다면, 이 고객은 대출제안을 수락한 고객으로 분류되며, 그 반대의 경우에는 대출제안을 거절한 고객으로 분류된다.

단일 변수(소득)에서 2개 이상의 변수로 증가하면, 집단의 평균을 집단의

1) 2장에서 설명하는 판별분석과 의사결정나무 분석 방법에 대한 설명은 Shmueli, G., Patel, N. R., & Bruce, P. C. (2009)의 Data mining for business intelligence의 내용을 인용하여 재정리한 것 임.

중심점(centroid)으로 사용한다. 이는 간단하게 평균벡터를 이용하여 구할 수 있다. 항목(관찰치)과 중심점 사이의 거리인 유클리드 거리는 각 변수 값과 해당 집단의 변수 평균값 간의 차이의 제곱을 합한 후 제곱근을 취한 값으로 정의한다.

유클리드 거리를 이용한 방법은 3가지 단점을 가지고 있다. 첫째, 거리측정이 선택변수의 측정단위에 따라 달라진다는 점이다. 소득을 천 달러 기준이 아닌 일 달러 기준으로 측정할 경우, 소득에 의한 거리는 달라진다. 둘째, 유클리드 거리는 변수의 변동성을 고려하지 못한다. 예를 들어 두 집단의 소득의 변동을 비교할 때, 대출제안 수락자들의 표준편차가 대출제안 거절자들에 비해 낮게 나타났다면, 대출제안 거절자임에도 불구하고 대출제안 수락자의 달러기준 평균소득에 가까울 수도 있다. 따라서 거리측정시 다른 변수들의 변동성을 고려할 필요가 있으며, 원래 측정 단위(척도)보다 표준편파로 나눈 표준단위값을 이용하여 거리를 측정해야 한다. 이것이 z값(z-score)과 같다. 셋째, 유클리드 거리는 변수들 사이의 상관관계를 무시한다. 이는 특히 많은 변수들을 이용하여 다수의 집단을 분리하려고 할 때 특히 중요한 고려사항이다. 변수들 간의 상관관계가 있는 경우 각 집단 사이를 판별하는 데 유용한 변수들이라 할지라도 이들 변수들을 함께 사용할 경우 동일한 변수의 효과가 중복해서 나타나게 된다.

이러한 문제점들을 해결하기 위해서는 통계적 거리(statistical distance) 혹은 마할라노비스 거리(Mahalanobis distance)로 불리는 측정도구를 사용해야 한다. 통계적 거리는 예측변수의 평균값을 고려할 뿐만 아니라 예측 변수 값의 퍼짐(spread) 정도와 다른 예측변수 간의 상관관계를 고려한다. 관찰치와 집단 사이의 통계적 거리를 계산하기 위해서는 반드시 예측변수의 평균(중심점)과 각 두 변수 간의 공분산을 계산해야 한다. 판별분석의 접근방법은 분리선(예측변수가 3개 이상인 경우에는 분리 초평면(hyperplane)을 찾기 위해 각 집단의 평균으로부터 같은 거리에 있는 통계적 거리를 사용한다. 또한 각 집단별로 관찰치의 통계적 거리를 측정하여 가장 근접한 집단으로 관찰치를 할당한다.

피셔의 선형 분류함수

1936년 통계학자인 피셔(Fisher)는 관찰치를 여러 집단으로 분리하는 방법을 개선시키기 위해 선형 분류함수(linear classification function)를 제안하였다. 선형 분류함수의 기본개념은 집단 간의 변동성 대 집단 내의 변동성의 비율을 극대화시키는 측정변수들의 선형 함수를 찾는 것이다. 즉 개별 집단 내에서는 매우 동질적(homogeneous)이지만, 다른 집단과 비교할 때는 가장 큰 차이가 나타나도록 집단을 찾아내는 것이다.

이 분류함수는 개별 관찰치와 각 집단 사이의 근접성을 점수로 계산하기 위해서 사용된다. 각 관찰치는 가장 높은 분류점수(가장 가까운 통계적 거리)를 가지고 있는 집단에 속한 것으로 분류된다. 이 분류함수는 통계 소프트웨어를 이용하여 측정할 수 있다. 분류함수의 개수는 집단의 수와 동일하다.

관찰치를 하나의 집단으로 분류하는 또 다른 방법은 각 집단에 속할 확률을 계산하고, 이 중에서 가장 확률이 높은 집단에 관찰치를 할당하는 것이다. 두 집단의 경우에는, 각 관찰치별로 하나의 확률만 계사하면 된다. 분류기준값 0.5를 사용하는 것은 관찰치를 가장 높은 분류점수를 갖는 집단으로 할당하는 방식과 동일하다. 이 방법의 장점 중 하나는 확률을 이용하여 레코드들을 내림차순으로 정렬하여 리프트 곡선을 그릴 수 있는 것이다.

2.1.2 분류 성과

판별분석은 분류점수에 대해서 2가지 중요한 가정을 하고 있다. 첫 번째 가정은 모든 집단의 측정치는 다변량 정규분포를 따른다는 것이다. 이 가정이 무리 없이 충족될 때, 판별분석은 로지스틱 회귀분석 등의 다른 분류기법보다 더 나은 분석도구가 될 수 있다. 요컨대 판별분석은 데이터가 다변량 정규분포를 따를 경우 로지스틱 회귀분석보다 30%만큼 더 효율적인 결과를 보이며, 동일한 결과를 얻는데 필요한 데이터의 30%정보만 요구된다, 실제로 예측변수들이 비정규분포를 따르고 심지어 더미변수를 포함하는 등 정규화되지 않아도 가장 크기가 작은 집단이 최소 20개보다 큰 데이터를 가진다면 판별분석은 상대적으로 강건한(robust) 것으로 알려져 있다. 또한 판별분석은 단일 예측변수를 이용한 단변량 공간과 다변량 공간 모두에서 극단치에 민감하게 반응한다. 따라서 탐색적 분석을 이용하여 극단치를 갖는 사례를 찾아내어 이를 제거할지 여부를 결정해야 한다.

판별분석의 두 번째 가정은 한 집단 내에서의 측정변수들 간의 상관관계는 다른 집단에서의 측정변수들 사이에서도 이와 동일한 상관관계를 갖는다는 것이다. 이러한 가정은 각 집단별로 변수들 간의 상관계수 행렬을 추정하여 비교해 봄으로써 어느 정도 확인될 수 있다. 만일 집단 사이에 상관관계가 서로 상당히 다르다면 이 판별분석은 사례를 판별할 때 분류결과의 변동성이 상당히 커지는 경향이 있다. 상관구조가 유의적으로 차이가 나고 데이터 집합이 매우 클 경우의 해결방안은 2차 판별분석(quadratic discriminant analysis)을 이용하는 것이다. 실제로 2차 판별분석은 상관관계 행렬에서 차이가 크고 학습용 및 평가용 데이터가 매우 많은 경우에만 유용하다. 이러한 이유는 2차 판별분석이 많은 파라미터를 추정해야하기 때문이다.

분류의 정확성을 평가할 때에는 분류모형의 성과를 판단하기 위해 일반적인 성과측정치를 사용한다. 여기에는 정오분류표(정확도 혹은 비용과 결합된 정확도)와 리프트 도표(lift chart)가 있다. 일반적으로 동일한 데이터에서 오차율을 산정하면 산정과정에는 추정상의 편의가 존재한다. 즉 지나치게 높은 성과를 나타내게 되는데, 그 이유는 분류모형의 파라미터를 추정하고, 모형의 오차를 측정하기 위해 모두 동일한 데이터를 사용하였기 때문이다. 따라서 다른 모형의 경우와 마찬가지로 분류함수를 추정하는 데 사용되지 않은 데이터(평가용 집합)를 이용하여 모형의 성과를 평가해야 한다.

판별분석의 정오분류표를 얻기 위해서는 분류점수를 바로 이용하거나, 분류점수로부터 계산된 소속집단의 확률을 사용한다. 이 경우에는 각 관찰치는 전체 집단 중 가장 높은 점수 또는 가장 높은 확률을 갖는 집단을 그 관찰치의 소속집단으로 할당한다. 그 다음 이 분류 결과를 관찰치의 실제 분류집단과 비교한다. 이와 같은 방식으로 정오분류표가 도출된다.

사전확률

지금까지 판별분석의 목적은 오분류율을 최소화하는 것으로 가정하였다. 앞에서 제시된 방법은 각 집단의 분류를 위해 사용되는 항목(관찰치)의 빈도가 집단별로 동일하다고 가정하였다. 미래의 분류를 위해 해당 항목이 나타날 확률이 집단별로 동일하지 않다면, 기대(평균)오차율을 줄이기 위해 분류함수를 수정해야 한다. 예를 들어 구입할 확률이 구입하지 않을 확률보다

확실히 적을 것으로 예상된다면, 미리 확률을 가정하는 편이 정확한 예측에 도움이 된다(동일한 집단 크기로 가정하고 판별하는 수고를 감소).

비대칭 오분류비용

집단별로 오분류비용이 대칭적이지(동일하지) 않은 경우, 사실상 추가적인 수정이 요구된다, 집단 1의 항목에 대한 오분류비용이 집단 2의 항목에 대한 오분류비용과 차이가 크게 난다면, 단순오차율보다는 기대 오분류비용을 최소화시키는 것이 필요하다(단순오차율은 이러한 비대칭 오분류비용을 인식하지 못함). 두 집단의 경우 오분류비용을 다르게 반영시키기 위해 분류함수를 수정하는 작업은 간단하다(사전확률을 추가).

균형화(balancing) over/under sampling

다변량 분석중에서 로지스틱 분석이나 신경망 분석에서는 상대적으로 크기가 너무 작은 범주가 있으면, 작은 범주의 레코드를 일정비율 배가(boosting)하거나 큰 범주의 레코드를 일정비율 감축(reducing)함으로써 종속변수 분포를 균등하게 하는 전처리(균형화)를 한다. 그러나 판별분석이나 나무분석은 오분류비용을 비대칭적으로 설정하는 것이 가능하기 때문에 균형화(balancing) 작업이 반드시 요구되지 않는다. 이 점이 판별분석이안 나무분석 분류가 신경망과 다른 중요한 한 가지 부분이다(허명회 외 2008). 물론 균형화를 할 수도 있다.

2.1.3 판별분석의 장단점

판별분석은 데이터마이닝 기법보다는 통계적 분류방법론으로 분류되는 경향이 있다. 이에 따라 판별분석은 데이터마이닝 문헌에서 빠져 있거나 짧게 언급되고 있다. 그러나 판별분석은 사회과학에서 매우 일반적인 분석방법이고 좋은 결과를 보여주고 있다. 판별분석은 모형의 적용과 결과측면에서 다중 선형 회귀분석과 유사하다. 따라서 이 두 기법은 여러 가지 측면에서 비슷한 장단점을 가지고 있다. 선형 회귀분석처럼 판별분석은 설명(예측) 변수의 가중치를 최적화시킨다. 선형 회귀분석에서 설명변수의 가중치는 종속변수와의 관계서 가중치가 결정되는 반면, 판별분석에서는 종속변수의 집단들

을 분리시키는 관계에서 가중치가 결정된다. 두 분석 모두 동일한 추정방법으로 최소자승법(least square)을 사용하며, 추정 결과치는 지역적 최적화(local optima)의 문제점도 거의 없다.

두 기법은 정규분포에 대한 가정을 하고 있다. 판별분석의 경우 예측변수들은 다변량 정규분포를 따른다고 가정한다. 그럼에도 불구하고 이 가정은 이진분류형 예측변수가 사용되는 경우와 같이 많은 실제 상황에서 지켜지지 않고 있지만, 판별분석은 이러한 문제에 대해서 강건하다. Hastie et al.(2001)에 의하면, 그 이유는 데이터가 보통의 경우 선형 경계와 같은 단순한 분류 경계를 찾는 데에만 사용되기 때문이다. 그러나 히스토그램을 통해서 분포가 매우 비대칭적인 연속형 변수들에 대해서는 로그변환을 적용함으로써 그 성과를 향상시킬 수 있다. 또한 이 방법을 통해서 극단치에 대한 민감도는 줄어들게 되는데, 이는 데이터의 극단치를 찾아서 분석에서 이러한 극단치를 없애주는 역할을 한다.

분류기법으로서 판별분석의 장점은 로지스틱 회귀분석과 마찬가지로 단일 예측변수의 기여도를 측정한다는 것이다. 이에 따라 예측변수의 중요도 순위를 정하고 변수를 선정할 때 유용하다.

마지막으로 판별분석으로 계산과정이 간단하고, 간명하며, 특히 작은 데이터 집합에서 유용하게 사용될 수 있다. 판별분석은 데이터를 최대한으로 이용하여 추정산식을 만들기 때문에 특히 데이터가 적을 때 유용하다.

2.2 의사결정나무 분석 방법론

데이터 분석가에게는 많은 노력이 요구되지 않으면서 사용자가 쉽게 분석결과를 이해할 수 있는 범용 분류기법으로서 가장 대표적인 것은 Breiman et al.(1984)이 개발한 나무방법론(tree methodology)이 있다. 이 장에서는 먼저 나무방법론(의사결정 나무)에 의한 분류절차에 대해서 논의하고, 이어서 연속형 종속변수를 예측하는 데 이 절차를 어떻게 확장할 수 있는지를 설명한다. CART(classification and regression tree, 분류와 회귀나무)는 Breiman 등이 이러한 절차를 구현하기 위해 개발한 프로그램(프로세스)을 지칭하는 용어이다.

이러한 의사결정 나무의 대표적인 적용사례는 대출심사가 있다. 은행고객들의 소득, 교육수준, 월별 신용카드 평균사용액 등을 이용하여 은행이 고객들의 대출제안을 수락할지 여부를 분류하는 것이다. 의사결정 나무모형이 흔히 사용되는 이유는 나무가 매우 큰 경우에도 나무규칙을 이해하기가 쉽기 때문이다. 종점마디(terminal node)는 대출제안 수락(1) 또는 거부(0)를 숫자로 부여받으며, 원 마디(circle node)에 있는 값은 예측변수에 대한 분할값(splitting vale)을 제공한다.

2.2.1 분류나무

분류나무에서 기본이 되는 2가지 핵심내용은 독립변수의 차원 공간을 반복적으로 분할하는 것과 평가용 데이터를 사용하여 가지치기를 하는 것이다.

주요 구성요소는 다음과 같다.

- 뿌리마디(Root node): 시작되는 마디(제일 상단 1개)
- 자식마디(Child node): 하나의 마디로 부터 분리된 2개 이상의 마디들
- 보모마디(Parent node): 자식마디의 상위마디
- 끝마디(Terminal node): 각 나무줄기의 끝에 위치하고 있는 마디, 잎(leaf)이라고도 하며, 결국 끝마디 개수만큼 분류규칙이 생성됨
- 중간마디(internal node): 나무구조 중간에 있는 끝마디가 아닌 마디
- 가지(branch): 하나의 마디로부터 끝마디까지 연결된 일련의 마디들, 이때 가지를 이루고 있는 다미의 개수를 깊이(depth)라고 함

반복적 분할

먼저 반복적 분할을 살펴보자. 종속(반응) 변수는 y로, 독립(예측)변수는 x1, x2, x3, ..., xp로 표시된다고 하자. 분류문제에서 반응변수를 범주형 변수이어야 한다. 반복적 분할은 변수 x의 p 차원 공간을 겹치지 않는 다차원 직사각형으로 나눈다. 여기서 변수 x는 연속형, 이진 혹은 서열 척도이다. 먼저, 변수 중 하나인 xi가 선택되고 xi의 값, 즉 si가 p차원 공간을 2개의 부분으로 나눈다. 한 부분은 xi≤si 인 모든 점들을 포함하고, 나머지 점들은 xi>si 가 된다. 그리고 나서 다시 변수를 선정해서(xi가 될 수도 있고 또는

다른 변수가 될 수도 있음) 이 두 부분 중 한 부분을 같은 방식으로 나눈다. 이 결과는 3개의 (다차원) 직사각형 영역이 된다. 이 과정이 계속되면 직사각형 영역은 더욱 작아진다. 이 개념은 모든 x-공간을 직사각형으로 나누어서 각 직사각형이 가능한 한 '순수'하게 또는 동질적(homogeneous)이 되도록 하는 것이다. 여기서 순수(pure)의 의미는 각 점들이 단 하나의 집단에 속하는 것을 말한다(물론 이 경우가 항상 가능한 것은 아니며, 이 점들이 다른 집단에 속할 수 있지만 모든 독립변수에 대해 정확히 같은 값을 가짐).

불순도 측정

여기서 분할 값들은 직사각형 안에 불순도(이질성)를 얼마나 많이 제거하는가에 따라서 순위가 매겨진다. 순수한 직사각형은 단일 집단으로 구성된다. 불순도의 감소는 분할 전의 전체 불순도에서 분할 후의 두 직사각형의 불순도의 합을 뺀 값으로 정의한다. 불순도를 측정하는 방법에는 여러 가지 방식이 있다. 가장 대표적인 2가지 측정치로는 지니 지수(Gini index)와 엔트로피 지수(entropy measure)가 있다.

이 방법론이 분류나무 알고리즘(classification tree algorithm)이라고 불리는 이유는 각 분할을 한 마디가 2개의 마디로 분할되는 것으로 묘사할 수 있기 때문이다. 첫 번째 분할은 나무뿌리 마디(root)의 가지로 묘사된다. 여기서 자식들을 갖는 마디는 원으로 표시된다(SPSS는 다르며, 프로그램마다 표현은 다소 상이함). 원 안의 숫자는 분할 값이며 분할되는 마디 밑에 분할에 선택된 변수의 이름이 표시된다. 의사결정 마디의 왼쪽갈래의 숫자는 분할 값보다 작서나 같은 값을 가진 의사결정 다미에 있는 점들의 개수이며, 오른쪽 갈래의 숫자는 분할 값보다 큰 점들의 개수를 의미한다. 이러한 마디들은 의사결정 마디(decision node)라고 부르는데, 그 이유는 만역 독립변수의 값들만 알고 있는 새로운 관찰치를 분류하기 위해 나무를 사용한다면, 각 의사결정 마디에서 관찰치를 적합한 나뭇가지로 할당하는 과정을 마디가 더 이상 자식을 갖지 않을 때까지 지속해야하기 때문이다. 이렇게 더 이상 자식 마디가 없는 마디는 나무의 잎(leaves) 또는 끝마디(Leaf node), 종점 마디(terminal node)라고 부른다, 각 끝마디는 원보다는 직사각형으로 표시하며,

x-공산에서 분할된 마지막 직사각형 중 하나에 대응된다. 관찰치가 모든 과정을 거쳐 끝마디로 종료될 때, 성장을 마친 나무의 끝마디에 포함된 모든 학습용 데이터를 대상으로 단순히 '다수결(voting)'에 의해 소속집단을 예측할 수 있다. 분류기준값(cutoff)을 0.5로 사용할 경우, 이 기준값을 기준으로 가장 많은 선택을 받은 집단이 새로운 관찰치의 예측집단이 된다. 이 집단의 이름은 끝마디에 표시한다. 따라서 소유자(owner)라고 이름을 붙인다.

이진분류 상황에서(일반적으로, 상대적으로 드물지만 특별한 관심의 대상인 성공집단을 가진 경우), 분류기준값을 낮게 설정함으로써 드물게 나타나는 성공집단들을 또한 더 잘 찾아낼 수 있다(그러나 이 경우 실패한 것을 성공한 것으로 잘못 분류하는 비용이 증가함). 분류기준값을 낮게 하면 성공집단에 대한 표결은 전체 끝마디에 대해 성공집단으로 분류하기 위한 낮은 분류기준값으로 인해서 성공집단으로 분류되기가 쉽다. 따라서 분류기준값은 최종 마디의 집단을 결정하는 데 필요한 비율을 결정한다. CART(이진나무라고 불림)에 의해 생성된 나무의 유형은 끝마디의 개수가 의사결정 마다보다 정확히 하나 더 많은 특성을 갖는다.

범주형 예측변수를 다루기 위해서는, 범주형 예측변수에 대해 선택된 분할값들이 범주형 값들을 2개의 하위집합으로 나눌 수 있는 값들 안에 있어야 한다. 4개의 범주를 가진 범주형 변수를 2개의 하위집단으로 나누는 7가지 방법이 있을 수 있다. 범주의 수가 많으면 분할개수도 매우 많아진다(따라서 소프트웨어에 따라서는 이진범주형 변수만 지원하는 경우도 있음). 2개 이상의 값을 가진 범주형 예측변수에 대해서는 회귀분석의 경우처럼 더미변수를 사용하는 방식으로 각 변수들이 이진변수가 되도록 이러한 범주형 변수를 여러 개의 더미변수로 바꾸어야 한다(역시 SW별로 상이).

성과평가(평가용 데이터와 가지치기)

새로운 사례를 분류함에 있어서 나무의 정확도를 평가하기 위해 우선 데이터를 학습용과 평가용 데이터 집합으로 나누도록 한다. 학습용 집합은 나무를 학습하는 데 사용되고 평가용 집합은 나무의 성과를 측정하는 데 사용된다. 평가용 집합을 사용하여 나무를 구축하는 방법도 있는데, 이 경우에 최종 나무의 정확도를 평가하는 데에는 세 번째 검증용 데이터 집합을 사용한

다(학습용, 평가용, 검증용).

평가용(혹은 검증용) 데이터의 각 관찰치는 나무에 적용하여 나무에서 연결된 끝마디에 따라 분류된다. 이렇게 예측된 집단은 정오분류표를 통해 실제 소속값과 비교된다. 특정 집단에 관심이 있을 때 리프트 도표(lift chart)는 모형이 각 소속값을 잘 분류할 수 있는지를 평가하는 데 유용하게 사용된다.

완전히 성장한(full-grown) 나무는 완전히 순수한 끝마디로 종료되기 때문에 학습용 데이터에서는 100%의 분류예측률을 갖는다. 반대로, 평가용과 검증용 데이터(완전히 성장한 나무를 구축하는데 사용되지 않은 데이터)에 대한 정오분류표는 학습용보다는 낮은 분류예측률을 보인다. 이러한 결과가 나타난 주된 이유는 완전히 성장한 나무는 학습용 데이터를 과적합시키기(완벽하게 예측하기) 때문이다. 과적합(overfitting)을 피하기 위한 방법으로 나무가 완전히 성장하기 전에 성장을 중지시키거나 완전히 성장한 나무를 가지치는(pruning) 방법이 있다.

자료는 꼭 임의로 분할 하는게 좋을까?

훈련자료와 검정자료로의 자료 분할에 대하여 50:50으로 분리한 경우 이 중 한부분에 기초하여 모형을 만들고 다른 한 부분으로 모형을 평가한다. 그 역할을 바꾸어 동일 작업을 반복해 재현성을 보는 것도 의미 있는 일이다. 그러면 2개의 모형을 얻었을 텐데 그 중에서 어느 모형을 쓸 것인가? 두 모형을 결합하여 쓰는 방법을 강구하는 것이 좋다. 예컨대 양쪽 모형으로부터 도출된 점수의 평균을 사용한다든지 하는 방법이다(허명회 외 2008).

과적합 해결 방안

과적합은 새로운 데이터에서 낮은 성과를 초래하게 만든다. 나무의 다양한 분할수준에서 전체오차를 볼 때, 과적합을 하게 되면 분할수준의 수가 증가하면서 전체적인 오차가 감소한다. 물론, 학습용 데이터의 경우는 전체 오차가 점점 감소하여 나무의 최대 마디수준에서 0이 된다. 그러나 새로운 데이터에 대해서는 나무가 집단과 예측변수 사이의 관례를 모형화하는 학습이 종료되는 시점까지는 전체 오차가 일반적으로 감소하지만, 이 점을 지나면

나무는 학습용 집합에서 잡음(noise)을 모형화하기 시작하고 평가용 집합에서 전체 오차는 일반적으로 증가한다. 나무의 높은 마디 분리 수준에서 과적합이 발생하는 직관적인 이유는 이 분리된 마디가 매우 작은 관찰치를 사용하기 때문이다. 이러한 경우에는 집단 간의 차이가 예측변수 정보보다는 잡음에 기반하여 발생하였을 가능성이 높다.

과적합을 제한하는 2가지 방법 중 첫 번째 방법은 나무의 성장을 멈추는 규칙을 정하는 것이고, 두 번째 방법은 완전히 성장한 나무를 과적합화되지 않는 수준으로 다시 가지치기를 하는 것이다.

분류나무 학습(분기, 성장)의 중단: CHAID

나무가 데이터를 과적합시키기 전에 나무의 성장을 중단시키기 위한 다양한 기준이 제시될 수 있다. 예를 들면, 나무의 깊이(즉, 마디 분할의 횟수), 마디 안에서의 최소 데이터 수, 불순도의 최호한의 감소 등이다. 그러나 문제는 이러한 규칙을 사용하여 적절한 학습종료시점을 결정하는 것이 그렇게 간단하지 않다는 것이다.

반복적 분할(recursive partition)에 기초하여 개발된 과거의 방법들은 나무가 학습용 데이터에 대해서 과도하게 분기(성장)하여 과적합하는 것을 방지하는 규칙을 사용한다. CHAID(chi-squared automatic interaction detection)는 분류와 회귀나무(CART)의 절차보다 수년 이상 먼저 개발된 반복적 분할방법으로서 오늘날 데이터베이스 마케팅에서 폭넓게 사용되고 있다. 이 방법은 순도를 통계적으로 유의한 양만큼 개선시키기 위해 마디를 더 분할할지 여부에 대한 판단을 잘 알려진 통계검정(독립변수에 대한 카이제곱)을 이용하여 결정한다. 특히 각 마디에서 반응변수와 가장 강하게 연관된 예측변수를 분할한다. 연결강도는 각 독립변수의 카이제곱 검정에 의한 p값으로 측정한다. 만약 가장 좋은 예측변수를 이용한 마디 분할의 검정결과가 유의적으로 향상되지 않는다면, 분할은 수행되지 않고 나무는 그대로 종료된다. 이 방법은 범주형 예측변수에 더 적합하지만, 연속형 변수는 범주형으로 변환되어 이를 적용할 수 있다.

분류나무의 가지치기

과적합에 대한 또 다른 해결방안으로는 나무 성장을 멈추게 하는 것보다 더 성공적인 방법으로서 완전히 성장한 나무에 가지치기를 적용하는 것이다. CART(SAS 등)와 C5.0(SPSS 클레멘타인, 현) 모델러) 등이 이러한 방법을 사용한다. C5.0에서는 학습용 데이터가 나무의 성장과 가지치기에 모두 사용한다. 그러나 CART에서는 학습용 데이터를 이용하여 나무를 성장시키고, 평가용 데이터를 이용하여 의도적으로 나무를 과도하게 성장시킨 다음, 다시 평가용 데이터를 이용하여 가지치기를 수행한다.

가지치기에 대한 기본적인 아이디어는 매우 큰 나무는 학습용 데이터에 대해 과적합화될 여지가 많으며, 오차를 거의 줄일 수 없는 가장 성능이 낮은 가치는 제거되어야 한다는 것이다. 완전히 성장한 마지막 단계의 사각형은 1개의 점만 포함하는 경우도 있는데, 이 마지막 분할들은 평가용 데이터 집합의 미래의 데이터에서 발생할 패턴을 반영한 것이라기보다는 학습용 데이터 집합의 잡음을 반영한 것으로 직관적으로 해설할 수 있다. 가지치기는 계속해서 의사결정 마디를 선택한 후, 이를 끝마디로 재설정하는 것으로 이루어진다. 즉, 이 과정은 의사결정 마디와 하위지식나무(subtree)로 향한 나뭇가지들을 잘라서 나무의 규모를 줄이는 과정이다. 학습용 데이터에서 잡음이 아닌 패턴을 파악하기 위해서 가지치기 과정은 평가용 데이터 집합의 오분류 오차가 최소로 되도록 가지친 나무의 의사결정 마디의 수를 정한다.

이 점을 찾기 위해 CART 알고리즘은 나무의 비용 복잡성(cost complexity)이라 불리는 기준을 사용으로써, 뿌리만을 가진 나무가 될 때까지 연속적으로 더욱 작은 크기의 일련의 나무를 생성한다. 여기서 뿌리 마디만을 갖는 분류규칙이란 각 규모(1,2,3, ...)의 최종 하위자식나무를 찾는 첫 번째 단계를 의미한다. 그리고 나서 평가용 집합의 오차율이 최소가 되도록 이 하위자식나무들 중에서 하나를 선택한다. 결국 최적의 선택나무는 평가용 데이터에서 가장 작은 오분류 오차를 제공하는 나무이다.

각 나무의 크기에 맞는 최적의 나무를 구축하는 것은 비용 복잡성(CC) 기준에 따르면, 이 기준은 나무의 오분류 오차(학습용 데이터에 기반)와 나무의 크기에 대한 벌점요인을 합하여 계산된다. 이 개념은 완전히 성장한 나무로부터 시작하여 전체 나무의 비용 복잡성이 하위 자식나무의 비용 복잡성을 초과하기 전까지 벌점요인 α를 점차적으로 증가시키는 것이다. 그리고

나서 하위자식나무에 대해 같은 절차를 반복한다. 이러한 방식을 계속 수행하면, 여러 개의 마디를 가진 일련의 나무가 생성되는 동안 마디 수는 감소하여 결국에는 단 한 개의 마디로 구성된 나무를 얻게 된다.

이러한 일련의 나무들로부터 평가용 데이터 집합에서 가장 작은 오분류 값을 갖는 나무를 쉽게 선택할 수 있다. 이렇게 선택된 나무를 최소 오차 나무(minimum error tree)라고 부른다.

한편, 더 좋은 성과를 얻기 위해서 다른 표본에서 이 최소점이 변하게 되는 원인으로서 표본오차(sampling error)를 추가적으로 살펴보도록 한다. 나무를 더 가지치기 위해, 즉 최소점보다 큰 표준오차를 갖는 평가용 데이터의 오차율에 대해서 오차의 추정 표준오차를 사용한다. 즉, '최적의 가지친 나무'는 가치치기과정에서 최소 오차 나무의 표준오차 안에 있는 오차를 가진 가장 작은 크기의 나무가 된다. 하지만, 평가용 집합에 대한 가지친 나무의 성과는 평가용 집합이 가지치기에 실제로 사용되었기 때문에 완전히 새로운 데이터에 대한 성과를 전부 반영한 것은 아니다. 이러한 이유로 인해서 선택된 모형의 유형에 상관없이 전혀 사용된 적이 없는 세 번째 데이터 집합인 검증용 집합을 이용하여 선택된 모형에 대한 성과를 평가하는 것이 특히 유용하다.

2.2.2 회귀나무

CART 알고리즘은 출력변수가 연속형인 경우에도 사용될 수 있다. 예측을 위한 회귀나무 역시 분류나무와 동일한 방식으로 작동한다. 회귀나무에서는 출력변수 Y가 연속형 변수이고, 원리와 절차 모두 분류나무와 동일하다. 많은 분할(split)이 이루어지고, 각 분류별로 나무의 각 가지에서의 '불순도'를 측정한다. 그리고 나서 나무절차는 불순도의 합이 최소가 되는 분할을 선택한다. 회귀나무를 통해 연속형 변수의 예측에 활용되는 것이다(사양에 따른 가격 예측). 회귀나무와 분류나무는 예측, 불순도 측정치, 성과평가 등 3가지 측면에서 서로 차이점이 있다.

예측

관찰치를 대상으로 반응변수 Y의 값을 예측하는 것은 분류나무에서와 비

슷하게 수행된다. 예측변수에 대한 정보는 끝마디에 도달할 때까지 나무를 학습시키는 데 사용된다. 예를 들면, 사용기간 = 55개월이고 마력 = 86인 도요타 코롤라의 가격을 예측하기 위해, 이 입력정보를 나무에 적용하여 최종적으로 9,876.5 달러의 값을 갖는 마디에 도달한다면, 이렇게 나무에 의하여 이 차의 가격을 예측한다. 분류나무의 경우에는 끝마디의 값(집단들 중의 하나)은 이 마디에 속한 학습용 데이터 중에서 집단의 수가 가장 큰 집단으로 '다수결'에 의해서 결정된다. 그러나 회귀나무에서는 끝마디의 값은 끝마디의 값은 끝마디의 학습용 데이터의 평균값으로 결정된다. 위의 예처럼, 9,876.5 달러는 사용기간 >50 and 마력 ≤90인 범주에 해당하는 학습용 집합의 자동차 120대의 평균값이다.

불순도 측정

분류나무의 마디에 대한 2가지 불순도(impurity) 측정치로는 지니 지수와 엔트로피 지수가 있다. 두 지수는 모두 해당 마디의 관찰치들의 범주 사이의 비율함수로 정의된다. 끝마디의 평균과의 제곱편차의 합은 회귀나무에서의 전형적인 불순도 측정치이다. 이 제곱편차의 합은 끝마디의 평균값이 정확히 예측값에 해당하기 때문에 제곱오차와 같다. 위의 예에서 9,876.5 달러의 값을 갖는 다미의 불순도는 해당 마디의 학습용 집합의 자동차 120대의 각 가격으로부터 9,876.5 달러의 값을 뺀 후, 이 편차의 제곱을 모두 더한 값이다. 마디의 모든 값들이 같을 때, 불순도는 가장 작은 값인 0을 갖는다.

성과평가

앞에서 언급했듯이 예측은 마디의 출력변수값의 평균을 통해서 얻어진다. 일반적으로 회귀나무의 예측성과는 다른 예측기법과 마찬가지로 평균제곱오차의 제곱근(RMSE, root mean squared error)과 같은 요약측정치와 리프트 도표(lift chart)와 같은 그래프를 이용하여 평가된다.

변수의 선택(feature selection)

거르기(screening): 중요하지 않거나 문제가 많은 예측변수들을 제거한다. 즉 결측치가 많은 예측변수와 한 범주에 너무 많은 관측치가 있는 범주형

변수, 그리고 변화량(표준편차 등)이 너무 작거나 지나치게 큰 연속형 변수들을 제거한다.

순위화(ranking): 나머지 변수들을 중요도에 따라 순위를 부여한다.

선택(selection): 순서화된 예측변수들 중에서 모델링과정에서 이용할 일부 변수들을 선택한다(허명회 외 2008).

2.2.3 방법론의 장단점과 추가 고려사항

분류와 회귀나무기법은 패키지화된 좋은 분류 및 예측도구이다. 이 방법은 또한 나무의 상위 분리수준에서 일반적으로 나타나는 가장 중요한 입력변수들을 사용하는 등의 입력변수를 선정하는 데에도 매우 유용하다. 이 나무는 다음과 같은 측면에서 사용자들에게 상대적으로 적은 노력을 필요로 한다. 첫째, 변수를 변환(transformation)할 필요가 없다(변수의 수서를 바꾸지 않는 어떤 단조변화(monotone transformation)도 동일한 나무를 생성함). 둘째, 변수군의 선택은 분할의 일부이므로 자동으로 수행된다. 여러 개의 독립변수를 제시해도 일부 변수만 선택적으로 활용하는 것이다.

분류와 회귀나무에서 마디 분할의 선택은 관찰값들의 절대크기가 아닌, 관찰값들의 실제크기 순서에 따라 결정되기 때문에 또한 기본적으로 극단치의 효과가 상대적으로 적다. 그러나 이 모형은 데이터의 변화에 민감하고 심지어 약간의 데이터 변경에도 다른 분할을 만들어낸다.

입력변수와 출력변수 사이의 특정한 관계를 가정하는 모형과는 달리(예를 들어 선형회귀식, 선형 판별분석과 같은 선형관계), 분류와 회귀나무는 비선형이고 비모수방법이기 때문에 입력변수와 출력변수 사이의 다양한 유형의 관계가 가능하다. 하지만, 이 점이 약점이 될 수도 있다. 분할이 입력변수들 간의 조합보다 단일 입력변수들을 바탕으로 행해지기 때문에, 이 나무기법은 선형 또는 로지스틱 회귀모형에서 특정 선형 구조하에서 입력변수들 간의 관계를 반영하기 못하는 경우가 종종 있다. 분류나무는 입력변수의 공간을 수평적, 수직적 분할을 통해 충분히 집단을 나눌 수 있는 사례에서 유용한 분류도구이다. 그러나 다음의 예를 가정해 보자. 즉 2개의 입력변수와 2개의 집단을 가진 데이터 집합이 있고, 2개의 집단 사이의 분리는 대각선을 이용하여 가장 확연하게 구분되는 예가 있다고 할 때, 궁극적으로 분류나무는 판

별분석과 같은 방법보다 도 나쁜 성과를 보이게 된다. 성과를 높이는 한 가지 방법은 기존의 입력들을 이용하여 입력변수들 간의 가설관계(회귀모형에서의 상호작용과 유사함)를 파악할 수 있도록 새로운 입력변수를 생성하는 것이다.

분류나무의 성과와 관련된 또 다른 문제는 좋은 분류모형을 만들기 위해서는 많은 데이터 집합이 필요하다는 것이다. 최근에 Breiman과 Cutler는 random forest를 제시하여 이러한 문제를 분류나무로 확장하여 다루었다. 기본 아이디어는 데이터로부터 다수의 분류나무를 생성하고(그러면 'forest'를 얻음) 그 결과를 더 나은 분류기준을 얻기 위해 결합하는 것이다.

분류와 회귀나무의 장점은 결측치를 대체하거나 결측값을 가진 관찰치를 삭제하지 않아도 결측데이터를 처리할 수 있다는 것이다. 따라서 이 방법은 해당 변수가 분류성능에 미치는 영향도의 관점에서 변수들의 중요도 순위를 평가하는 목적으로 확장될 수 있다. 한편, 계산측면에서 나무모형은 모든 변수에 대해 모든 가능한 분리를 계산하기 위해서 수많은 정렬계산과정이 필요하기 때문에 나무를 성장(학습)시키는 데 상대적으로 많은 시간이 소요된다, 또한 평가용 집합을 이용하여 데이터를 가지치기할 경우 더 많은 계산시간이 요구된다. 결론적으로, 분류와 회귀나무에서 가장 실제적으로 중요한 이점은 나무들이 생성하는, 이해하기 쉽게 표현된 규칙에 있다. 이러한 명쾌성으로 인해서 경영분야에서 많이 활용되고 있다.

앞서 설명한 판별분석과 나무분석 그리고 대표적인 다변량분석(데이터마이님) 분석 방법인 신경망의 비교를 정리하면 다음과 같다.

<표 2> 다변량 분석 방법의 장단점 비교

	장점	단점
판별분석 (로지스틱)	생성모형에 대한 해석이 쉽다.	복잡한 상황에 유연하게 대처하지 못한다.
신경망	복잡한 상황에 유연하게 대처한다.	생성모형에 대한 해석이 어렵다
나무 분석	생성모형이 단순하고 해석이 쉽다	설명(입력) 변수의 작은 변화가 출력을 크게 바꾸기도 한다

자료: 허명회 외 2008

Random Forests

Random forests는 Breiman(2001)에 의해 개발된 분류 알고리즘으로서 다양한 붓스트랩 표본으로부터 얻은 분류나무들을 하나의 분류나무로 만드는 알고리즘이다. 이 알고리즘의 목적은 단일 나무를 만드는 대신 여러 개의 나무로 확장시킴으로써 분류의 정확도를 높이고자 한 것이다. Random forests의 가장 큰 장점은 데이터를 과적합시키지 않는다는 점이다. 이는 대수의 법칙에 의해 forests(나무숲)가 커질수록, 즉 나무의 수가 많아질수록 일반화의 오류는 특정값으로 수렴한다는 데서 얻어진 결과이다. 따라서 다른 분류기법에 비해서 낮은 일반화 오류를 갖는다. 일반화 오분류율의 상한값(PE^2)은 random forests에서 각 나무의 정확도로 해석될 수 있는 강도(strength)와 분류나무의 상관관계(correlation)에 의해 계산된다(식은 생략하는데 분류의 정확도를 높이기 위해서는 강도(s)를 유지하면서 상관관계를 최소로 해야 함). Random forests는 위의 조건을 만족시키기 위해서 2가지 임의성(randomness)의 개념을 도입한다. 첫째, 앙상블 나무(ensemble trees)를 확장하기 위해 분석용 데이터에서 붓스트랩 표본을 뽑아낸다. 이 과정에서 Beiman이 제안한 Bagging 기법을 사용한다. 각 나무는 붓스트랩 분석용 데이터에 기초하여 확정된다. 둘째, 나무가 확장될 때 분리에 사용되는 독립변수는 무작위로 선택되거나 이들의 조합으로 구성된다. 그리고 나무는 가지치기하지 않는다.

제 3 장 선행연구 분석

3.1 중소기업 기술혁신 지원 정책과 효과 연구

일찍이 Robson(1993)의 연구에 따르면, 정부 R&D 지원제도는 민간기업의 R&D 투자를 증가시켜 기술혁신활동을 보완하는 관계에 있다. 이 외에도 많은 연구에서 정부의 중소기업 지원은 R&D 투자를 증가시키고 사업화 노력을 촉진시키는 효과가 나타난다(Audretsch et al. 2002; 이병헌 외 2013). 반면 정부의 재정적 지원이 기술 사업화나 경제적 성과에 부정적인 효과를 나타낼 수 있다는 주장도 있었다(Akcomak et al. 2007; Svensson 2007; 송종국 외 2009).

이런 중소기업 기술혁신을 위한 지원 정책은 다양한 방법으로 분류가 가능한데, 많은 연구들은 자금과 인력을 정책 도구 유형으로 구분하여 분석하고 있다(Lin et al. 2006; 신진교 외 2008). 그밖에도 다양한 분류가 있었지만 그러나 잘 알려진 정책도구 유형은 Bemelmans-Videc et al.(2011)이 수단의 강제력의 정도에 따라 규제(sticks), 유인(carrots), 정보제공(sermon)의 세 가지 정책 유형으로 분류한 것이다. Bemelmans-Videc et al.(2011)의 유형론은 정책도구의 상호배제성과 망라성의 조건을 충족하며 단순명료하여 기술혁신의 정책을 분류하는 기준으로 널리 활용되고 있다. 이런 세 가지 분류 중에서 규제 정책은 법, 각종 규제, 표준, 세재 등이 해당되며, 유인 정책은 각종 개발 자금 지원, 기술개발 조세 감면, 공공 구매, 판로 지원 등이 해당되고, 정보 제공 정책에는 기술제공이나 지도, 각종 인력 지원이 포함된다(류숙원 외 2010).

이들 정책의 분류 중에서 규제와 유인정책의 효과에 대한 연구는 상대적으로 활발하다. 송종국 외(2009)는 R&D 촉진을 위한 재정지원 정책의 효과를 분석했는데, 중소기업 기술혁신을 위한 정책의 효율성이라는 측면에서 국내에서 가장 보편적으로 활용되고 있는 R&D 보조금 지원제도와 조세 지원제도가 과연 효과가 있는지를 실증 분석하기 위해서 과학기술 연구개발활동 조사 결과를 활용해 분석했다. 연구결과에 따르면, 정부의 R&D 직접 보조금은 오히려 중소기업의 R&D를 줄이는(구축) 효과를 보였으며, R&D 조세 지

원제도는 중소기업의 R&D 투자를 유인하는 것으로 나타났다. 저자들은 연구결과에 따라 R&D 지원제도를 효율성 제고를 위해서 중소기업에 대해서 R&D 인프라 구축은 물론 기술정보지원과 같은 R&D 서비스 지원이 필요하다고 주장했다. Lichtenberg(1984)의 연구에서도 정부 R&D지원제도가 민간 기업들의 R&D 투자액을 감소시키는 효과를 초래한다고 주장된 바 있다(전승표 2015).

특히 유인정책과 관련해서 중소기업의 경우 정부의 자금 지원이나 벤처캐피탈 투자 등이 혁신을 향상시키기 때문에(Shan & Song 1997), R&D 초기 단계에 있는 중소기업은 초창기에 정부의 다양한 금융, 인력, 판로지원에 대한 활용이 매우 효과적이라는 주장도 있다(김왕동 외 2002). 그러나 정부의 보조금 지원이 기업의 성과에 미치는 영향이 일관된 경향을 나타내고 있는 것은 아니다. 일반적으로는 Busom(2000)의 연구처럼 정부의 직접적인 지원정책이 기업의 생산성이나 혁신을 향상시킨다는 연구가 주류를 이루고 있다. 그러나 Guellec et al.(2004)의 연구를 보면 OECD 국가를 대상으로 한 연구에서 정부의 지원비중이 높을수록 기업의 R&D에 대한 생산성 탄력성이 낮아지는 것으로 나타나기도 했다. 이렇게 기술혁신을 촉진시키기 위한 규제나 유인(자금 지원 등)에 대한 연구는 활발히 진행되어 왔다.

상대적으로 기술혁신을 촉진시키기 위한 정보제공에 대한 연구는 최근에 들어 시도되고 있는데, 그중에서도 인적자원에 대한 연구가 활발히 진행되어 왔다. Vrakking(1990)은 중소기업 혁신을 위한 인적자원 변수는 중요한 요인으로 분석하면서 중요한 인재가 없는 조직에서 창의적 기술이나 혁신을 기대하기 어렵다고 주장하였다. 특히, 인적자원은 정보나 지식과 같은 경제적 부가가치를 창출하는 작업으로, 기술 혁신의 역량확대와 이를 통한 조직경쟁력에 핵심적 역할을 담당하고 있기 때문에 주목받아 왔다(정선양 외, 2009). 또한 Romijn과 Albaladejo(2002)나 이병헌 외(2008)등도 R&D 기술혁신을 높이기 위해서는 인력에 대한 투자를 확대해야 한다고 주장했다. 신진교 외(2008)는 중소기업이 R&D를 통해 혁신성과를 창출하는 과정에 있어서 정부정책의 조절효과에 주목하고, IT분야의 중소기업을 대상으로 정책 조절효과를 정책자금지원, 기술지원, 인력지원의 3개 영역에서 살펴보았으며, 이들 정책수단 중 정책자금 지원과 인력지원 정책이 혁신성과 창출면

에서 유의미한 조절효과를 보이고 있음을 밝혔다(서주환 외 2015 재구성). 이상에서와 같이 정보제공 정책 중에서도 인력지원은 상대적으로 연구가 활발했다. 반면 기술정보 지원은 정보 제공정책 중에서 상대적으로 연구가 활성화되진 못한 영역이다(전승표 2015).

최근 박문수 외(2012)는 중소기업 기술지원 정책을 기술자금·세제지원, 정보지원, 기술인프라 지원으로 구분하고, 중소기업의 기술경쟁력에 영향을 미치는 변수로서 기업의 특성과 기술특성으로 구분하여 연구하였다. 그 결과 정보지원과 같은 간접적인 지원 정책보다는 기술자금·세제 지원, 기술인프라 지원 등 중소기업에 대한 직접적인 지원 정책이 기업 기술경쟁력 강화에 높은 조절효과를 나타낸 것으로 나타나기도 했다. 또한 류숙원 외(2010)은 제조기업을 중심으로 정부 정책이 중소기업의 기술혁신에 미치는 영향을 연구했는데, Bemelmans-Videc et al.(2011)의 세 가지 분류 정책 모두를 분석 대상으로 했다. 그 연구에 따르면 정부의 중소기업정책은 기업의 부가가치 창출을 고취시키고, 장기적인 경쟁력을 확보하게 하는 등 중소기업 혁신을 위해 중요한 수단이지만, 자금, 인력, 정보와 같은 정책도구의 선택에 따라 그 성과에는 차이가 있는 것으로 나타났다. 연구에 따르면, 경제적 유인도구 중에서 기술개발조세감면, 정보제공 도구 중에서 기술정보제공과 정부연구개발 사업 참여는 중소기업혁신에 긍정적인 영향을 끼치는 것으로 밝혀졌다. 반면에 정보제공 도구 중 인력에 의한 기술지원의 경우는 부정적 영향을 끼치는 것으로 나타났다(전승표 2015).

3.2 중소기업과 기술기획에 관한 연구

최근 이종민 외(2013)의 연구에 따르면, 중소기업의 기술경영에서 기술기획(technology planning)의 역할이 잘 정리되어 있는데, 기업이 추구하는 기술전략은 세부적인 연구개발 과정에 있어 방향성을 제시하며 그 실행을 위한 구체적 가이드라인을 제시하는 역할을 수행한다. 여기서 기술기획은 기술전략에 따른 세부적인 실행계획, 연구개발 프로젝트 선정, 자원배분 등 과 같은 활동을 의미하는 것으로 기술전략을 구체화하는 과정으로 파악할 수 있다(정선양 2011). 오해영 외(2007)는 기술기획을 기술자원의 동원가능성

과 과학기술제도와의 부합성 등의 전제조건을 고려하여 실행타당성과 효과성의 관점에서 대안을 검토, 평가하는 활동으로 비가시적인 성격을 지니는 것으로 파악하였다. 정근하(2010)는 기술기획에 대한 국가차원의 접근을 강조하며 경쟁력 강화를 위한 관련 기술의 분석과 예측을 통해 개발기술의 기간별 목표설정과 전략수립 그리고 그 실현 과정에 참여하는 일련의 활동을 기술기획으로 정의하였다.

상기와 같이 기술기획에 대한 개념은 연구자들마다 미래에 대한 합리적인 선택, 실천적 역할 등 해당 연구에서 강조하는 바에 따라 다양한 정의를 내리고 있다. 기업들 또한 기술개발과 관련된 기획, 계획수립, 프로세스 개발 등 다양한 경영활동을 수행하는데, 세부 활동으로는 기술기획(technology planning), 기술전략(technology strategy), 연구기획(R&D planning), 전략적 기획(strategic planning), 신제품개발기획(new product development planning) 등을 들 수 있다(이종민 외 2013).

기획이란 계획을 수립하는 과정을 의미하며, 계획은 기획을 통하여 산출되는 결과를 지칭한다(최영훈 외 2005). 기획은 특별한 목표와 성취해야할 수단을 규명하는 활동으로 어느 특정한 시기에 성취해야할 결정행위를 담고 있으며, 지속적이고 미래지향적인 활동이다(오해영 외 2007). 아울러 초기 계획은 신제품개발 프로젝트 성공에 긍정적인 영향을 미치는 것으로 조사되었다(Verworn et al. 2008). 이러한 점을 고려할 때, 기술기획 역량은 기술개발과 사업화 성공을 위한 계획수립, 조정, 실행 등의 다양한 활동능력을 의미한다. 이와 더불어 기술자원의 동원가능성과 실행 타당성, 그리고 미래지향성이 기술기획 역량에 영향을 미친다고 할 수 있다(이종민 외 2013).

기업이 생산하는 제품은 기업의 수익을 창출하기 위한 주요한 수단이다. 기업들은 신제품으로부터 수익의 1/3을 창출하는 것으로 나타났다(Cooper 2001), 기술혁신 변화가 급격한 산업의 경우 신제품으로부터의 수익 창출은 더 많은 비중을 차지하게 되며, 그 결과 신제품 개발은 비즈니스에 있어 가장 중요한 과업의 하나로 인식되어지고 있지만, 그러나 이러한 신제품 개발의 중요성에도 불구하고 사업화 실패율이 높게 나타나는 것은 연구개발 초기단계인 기술기획과의 관련성이 높다(이종민 외 2013). 전략적인 측면에서 기획에 대한 중요성에 대해서는 상반된 연구결과들이 존재하는데, 기술기획

에 있어서도 프로젝트 및 혁신의 성공에 있어 기술기획이 중요한 역할을 수행하며 초기 계획은 성공에 있어서 긍정적인 영향을 미친다는 견해(Dvir et al, 2003; Stockstrom & Herstatt 2008)와 정교한 초기 계획의 효과성에 의문을 제기하고 변화에 대한 빠른 대응 능력과 성공을 이끌기 위해 어떻게든 업무를 처리하려는 노력의 중요성을 강조하는 의견(Dvir & Lechler 2004)이 공존하고 있다.

기술혁신활동은 중소기업들의 생존, 성장 그리고 개발을 위한 중요한 요인으로 간주되어져 왔으며, 특히 R&D는 연구자들로부터 혁신의 결정요인으로 가장 많은 관심을 받아왔다. 일반적으로 중소기업들은 R&D 집중도와 특허활동이 부족함에도 불구하고 대기업에 비해 보다 혁신적이며, R&D 효율성이 높은 것으로 분석되었다(이종민 외 2013). 중소기업들은 환경변화에 유연하고 빠르게 대응을 할 수 있는 강점을 지니고 있다. 하지만 극심한 경쟁환경과 다양한 분야에 있어 기술적 역량의 필요가 증대됨에 따라 모든 측면에서 경쟁력을 확보하기에는 어려움이 있으며, 대기업에 비해 절대적인 규모의 제약 때문에 경쟁열위에 놓여있다(Narula 2004). 혁신활동 수행을 위한 중소기업의 가용자원은 대기업에 비해 상대적으로 열악하며 한정적이기 때문에 이에 대한 투자 및 실행을 위해서는 철저한 계획수립과 보다 많은 고민이 필요하다. 또한, 연구개발 활동은 본연의 특성상 R&D성과가 실제 사업에 유용하게 사용되었다는 사실을 확인하기까지 많은 시간이 걸리고, 매우 복합적이며 타 성과에 비해 비가시적인 특성을 지니기 때문에 보다 효율적인 관리가 필요하다(이종옥 외 2005). 중소기업의 경우 잘못 추진된 기술개발 프로젝트는 해당 프로젝트의 실패뿐만 아니라, 기업의 존립을 위협할 수도 있다. 그렇기 때문에 기술기획은 매우 중요하며 필수적인 프로세스인 것이다(이종민 외 2013).

3.3 중소기업 국내 기술기획 현황

국내에서는 지속적인 R&D비용 확대에도 불구하고 연구개발 효율성 및 성과 측면에서의 문제점은 여전히 개선되고 있지 않은 문제점으로 지적되고 있는데, 이러한 기술수준의 답보상태는 사전기획을 중요하게 생각하는 반면

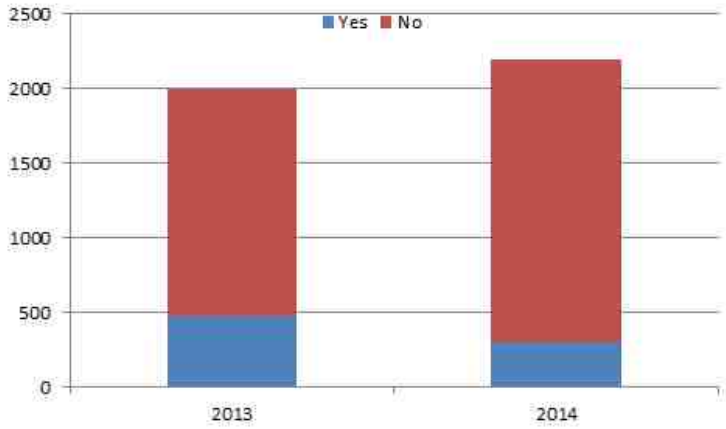
실질적인 기술기획 수준이 낮은데서 그 원인을 찾을 수 있을 것이다(이종민 외 2013). 중소기업연구원(2006)에 따르면 중소기업들은 기술개발 전략 수립을 위한 가장 효율적인 수단으로 '사전기획 조사'라고 응답하였으며, 과제기획 단계에 가장 애로를 느끼는 것으로 조사되었다. 중소기업들은 기술개발전략 수립 시 '기술개발의 시장성 및 수익성(64.1%)'과 '국내외 기술개발 추세(25.6%)'를 우선순위로 고려하는 것으로 나타났으며, 기술개발전략의 효율화를 위한 수단의 1순위로 '사전기획조사 강화를 통한 성공가능성 높은 과제를 발굴 (36.2%)'이라고 응답하였다. 그리고 기술사업화의 주요한 성공요인으로 '잠재적인 기술력 우수(19.7%)', '기술 및 시장조사가 충분히 이루어짐(18.8%)'인 것으로 분석되었다(한국산업기술진흥협회, 2009).

이는 기술기획의 중요성이 점차 증대된 것으로 파악할 수 있을 것이나, 그럼에도 불구하고 본 보고서가 분석한 '중소기업 기술통계조사'에 따르면, <그림 1>에 나타난 것과 같이 우리나라 중소기업의 기술기획 수준은 세계 최고수준 대비 약 70% 수준으로 여전히 취약실정이며 향후 개선이 필요한 것으로 분석되었다.

자료: 중소기업청 및 중소기업중앙회 2013 및 2014

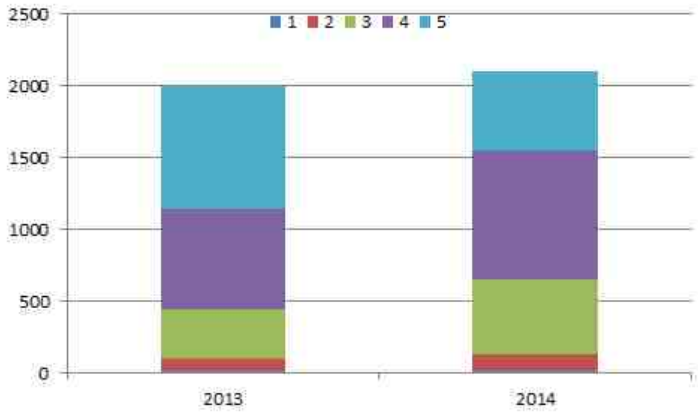
<그림 1> 기술요소별 세계대비 기술역량 자가 진단 결과

동일한 조사에서 2013년에는 R&D 기획단계에서 정부자금 지원을 경험한 집단이 23.8%에 이르렀으며, 2014년 샘플이 바뀌면서 13.5%로 나타났다(<그림 2> 참조). <그림 3>에는 기획지원 제도 필요성이 나타나 있는데, 기획단계의 정부지원 필요성이 매우 높은 것으로 나타났다. 2013년에는 77.7%가 2014년에는 66.0%의 중소기업이 약간(4점) 또는 매우(5점) 필요한 것으로 응답했다. 특히 기획단계 정부지원 경험이 상대적으로 많은 집단으로 조사된 2013년에서 매우 필요하다고 응답한 집단이 더 많았음을 주목할 필요가 있다.



자료: 중소기업청 및 중소기업중앙회 2013 및 2014

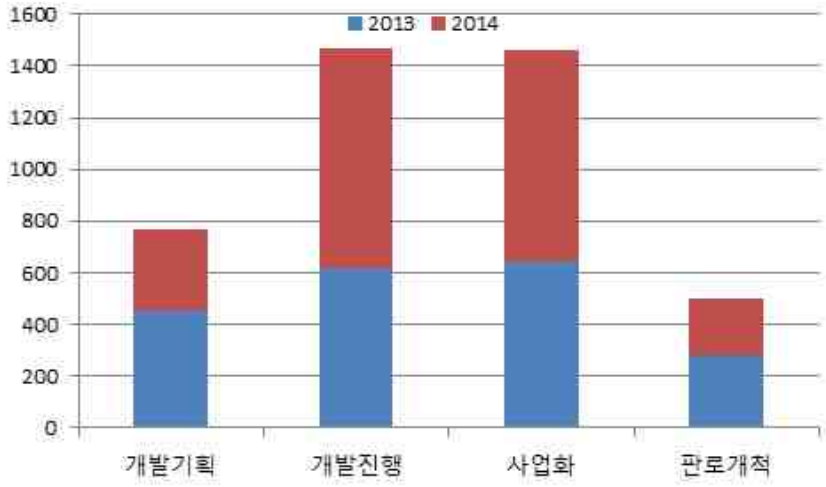
<그림 2> R&D 기획단계 정부 지원 활용 여부 비교



자료: 중소기업청 및 중소기업중앙회 2013 및 2014

<그림 3> R&D 기획 지원 필요성(5점 매우 높음) 응답 결과 비교

<그림 4>에서 보면, 기술개발 단계상 정부의 정책적인 지원이 가장 필요한 단계를 중소기업이 응답한 결과가 나타나 있다. 결과에 따르면 이미 지원사업이 활발한 개발단계나 사업화단계에 대한 지원 필요성이 높게 나타났지만, 국내 관련 지원 사업이 활성화되어 있지 않은 개발기획단계 지원필요성을 선택한 응답자는 상대적으로 많은 것을 확인할 수 있다. 다른 단계보다 기획단계의 필요성을 높게 응답한 비중이 2014년 14.3%, 2013년 22.8%에 이르고 있었다. 중소기업 R&D 지원 정책에서 기획 단계 지원의 강화 필요성을 확인할 수 있는 결과라고 할 수 있다.



자료: 중소기업청 및 중소기업중앙회 2013 및 2014

<그림 4> R&D 개발 단계별 지원 선호도 응답 결과 비교

3.4 중소기업 기획(또는 정보) 지원 관련 사업 현황

정보제공 정책 중에서 정보 지원 사업은 크게 정보를 제공하거나 정보화를 지원하는 사업과 같이 직접적으로 지원하거나 다른 기관을 활용하여 간접적으로 정보를 지원하는 사업으로 나누어 볼 수 있다. 먼저 간접적으로 중소기업에 정보를 제공하는 사업으로 R&D 기획 지원과 같은 정책 사업이 있다.

중소기업청에서는 중소기업의 R&D 혁신활동을 도모하기 위한 다양한 정책적 시도를 하고 있는데, 이러한 지원 프로그램은 R&D 단계의 정책자금

지원이외에도 非R&D 분야에서도 정책적 지원을 하고 있다. 대표적인 사업으로 '중소기업 R&D 기획지원' 사업이 있다. 중소기업 R&D 기획지원 사업은 중소기업의 R&D 혁신 활동을 촉진하고 이를 통한 혁신 성과의 효율적 창출을 위해, R&D 기획단계에서 중소기업 보유 기술에 대한 기술성, 시장성, 사업성에 대한 면밀한 검토를 통해 효율적인 기술개발 및 사업화 전략 수립을 할 수 있도록 지원하고 있다. 매년 약 30~40억 원의 예산 지원을 통해 140개 내외의 중소기업을 지원하고 있다. R&D 기획지원을 통해 기획지원 받은 기업들은 중소기업청의 R&D과제에 연계되게 되는데, 연계 대상 R&D과제는 R&D기획 결과물과 기업유형 및 기술특성에 따라 창업성장기술개발사업 혹은 중소기업 기술혁신개발사업을 통해 1-2년 내의 R&D자금을 지원 받게 된다(서주환 외 2015).

이런 R&D 기획 지원 사업과 유사한 사업으로 R&D 로드맵 지원이 있었는데 중소기업청에서는 중소기업의 기술개발 전략수립을 지원하기 위하여 2008년부터 개별중소기업의 기술로드맵 작성지원을 추진하게 되었다. 개별기업을 위한 기술로드맵 지원사업의 목적은 R&D를 기반으로 하고 있는 기술혁신형 중소기업을 대상으로, 미래 신규시장 개척을 위한 R&D 기획에 활용할 수 있는 개별 중소기업의 기술로드맵 작성을 지원함으로써, 중소기업 R&D 자원의 효율적 이용과 사업화 성공률을 높이고, 중소기업의 기획 능력을 배양하기 위한 목적으로 추진되었다(Jun et al. 2012). 2008년에 1차 사업을 수행한 이후 2013년까지 사업이 진행되었으며, 2014년부터 R&D 기획 지원사업과 통합되었다.

이상의 사업들은 모두 정부에서 중소기업의 R&D 기획 단계에서 지원한다는 점에서는 일반적인 기술 개발 단계의 자금 지원과 유사하지만, 지원되는 자금이 기획이나 로드맵을 제공하는 정보 분석 기관에 모두 제공된다는 측면에서 일반적인 R&D 자금 지원 정책과 큰 차이점이 있다. 지원되는 자금의 규모에서도 큰 차이가 있는데, 기획지원 사업은 기업당 2천만 원 이하 수준으로 상대적으로 작다. 이런 점 때문에 본 보고서에서는 R&D 기획을 위한 자금 지원 사업도 간접적인 정보 제공 정책으로 간주하여 연구를 진행했다.

직접적인 정보지원 산업은 기술이나 시장정보를 제공하는 사업과 정보화를 지원 받는 사업으로 다시 나누어 볼 수 있다. 대표적인 정보화 사업으로

‘기술유출방지시스템구축 지원 사업’ 이 있다(강지헌 2014). 이 사업은 중소기업의 보안인프라에 대한 정밀진단 및 설계를 통해 기업환경에 적합한 보안시스템을 구축 지원하는 사업으로 지원규모는 15억 수준이며, 기업당 최대 6개월간 4천만 원을 지원받을 수 있다(중소기업기술정보진흥원 2015). 기술정보 지원 사업은 지자체 단위로 많이 진행되고 있는데, 국내외 산업기술정보를 지원하는 사업으로 전체 사업규모 수천만 원 수준이며 기업당 수백만 원 규모로 정보조사 비용을 지원하고 있다(한국과학기술정보연구원 2015).

본 보고서에서는 기술개발관련 기획단계 자금 지원 경험이나 정보 지원 사업 참여 경험과 각각에 대한 정부지원 필요성을 높게 응답한 기업을 중심으로 영향 요인을 찾고 프로파일링을 하게 되는 것이다.

3.5 다변량 분석(판별분석 및 나무분석) 활용 선행 연구

판별모형(discriminant model)은 경영분야에서 활발하게 활용되어 왔으며, 대표적인 사례가 기업 부실이나 파산 예측 모형이다. 기업의 재무상태, 거래신뢰도, 일반사항, 신용등급 등 다양한 정보를 종합적으로 고려해 예측 모형을 만드는데 판별분석이 활용된 것이다. 성웅현(2006)은 기술지향적 중소벤처기업의 기술력평가 정보를 바탕으로 벤처중소기업의 파산예측모형으로 선형 판별모형과 로지스틱 판별모형을 활용해 파산을 예측할 수 있는 모형을 제시했다. 모집단에 대하여 다변량 정규분포를 가정하여 분석한 결과에 따르면 선형판별모형이 로지스틱 판별모형보다 예측력이 다소 우월한 것으로 나타났다. 성웅현(2007)에서도 역시 판별분석을 활용해서 중소기업의 부실을 예측하는 최적 모형을 연구하였는데, 기술신용보증기금의 기술력 평가자료를 활용했으며, 평가항목도 자원, 기술성, 시장성, 사업성 등 지표를 확대했다. 특히 확실범주와 유보범주로 구분할 수 있는 경계값을 설정하여 예측력을 높이기도 했다.

성웅현 외(2009)는 단순 이분형 판별분석을 떠나 여러 범주에 대한 판별분석을 활용했는데, 지식재산 투자와 관리가 기업의 무형자산 가치 범주들을 구분하는데 유의미한지 분석하기 위해서 다항로지스틱 판별분석을 활용했다.

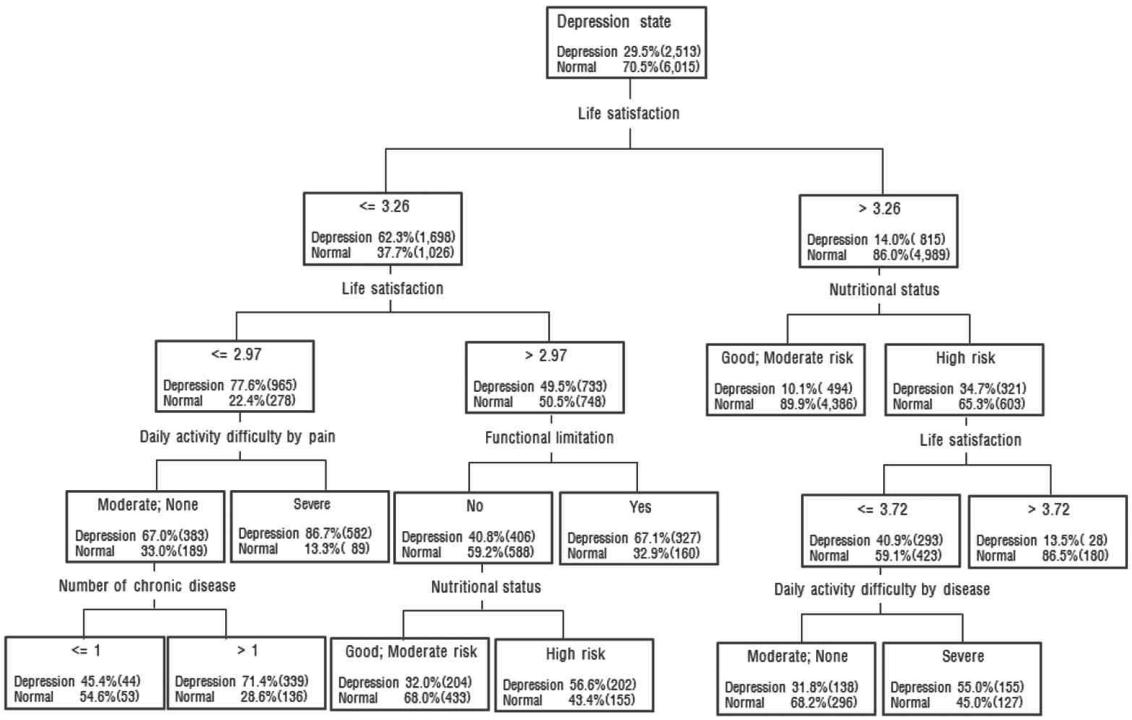
이 연구는 단순 무작위 추출된 90개 기업을 대상으로 매출액 대비 연구개발비 비율이 높은 집단, 중간 집단, 낮은 집단 3가지로 범주화 했다. 이 연구에서 무형자산으로 브랜드도 고려하여 광고선전비도 연구개발비와 같이 기업의 장기적인 가치 향상을 위한 노력으로 간주했다. 선정된 90개 기업에 지식재산 관리 설문조사를 실시했다. 그 결과를 다항 로지스틱 분석을 통해 연구개발비 비율과 특허 관리 역량이 무형자산가치 범주 구분에 유의미한 것으로 나타났으며, 정준변량분석 결과 연구개발비 비율, 특허 관리 역량, 디자인 및 브랜드 관리 역량, 광고선전비 비율 순으로 영향이 높은 것으로 나타났다.

최근 이선영 외(2013)는 대·중소기업협력재단의 '구매조건부 신제품개발사업'을 대상으로 구매 이행 여부를 판별하는 요인이 무엇인가 분석 하였다. 구체적으로 기업(기관)간 협력, 경영자원·역량, 시장환경, 정부지원 등의 요인들 중 사업성과에 영향을 미치는 것들이 무엇인지 연구했다. 역시 피셔의 선형판별함수를 통해 6개 독립변수의 영향을 판별했고, 정준판별함수를 활용해서 영향을 분석했다.

이상에서 살펴볼 수 있듯이 다변량분석 중 판별분석은 범주형 피예측변수에 대한 영향요인의 탐색적 연구나 예측모형 개발에 활발하게 활용되어 왔다. 의사결정나무 분석도 다변량분석으로 활용되지만, 다른 방법과 같이 활용되는 경우가 많고, 단독으로 활용되는 경우가 판별분석(선형 또는 로지스틱) 만큼 많지는 않다.

일반적으로 의사결정나무 분석은 의학(또는 심리학)에서 탐색적 연구를 위해 활용되는데, 박명화 외(2013)는 의사결정나무 분석법을 활용해 우울 노인의 특성을 분석했다. 노인 우울증 유병률이 증가하는데, 기존의 단편적이고 적은 케이스에 대한 연구에서 벗어나 대규모 데이터를 통해 노인의 일반적 특성과 건강, 사회활동, 가족생활 등 전반적인 특성을 바탕으로 노인의 우울과 관련한 세분화된 특성 파악 및 새로운 지식을 얻을 수 있는 연구를 시도한 것이다. 연구는 2008년도 노인실태조사 자료를 활용하였는데, 응답자 중 14,970명의 데이터를 분석했고 분석용과 검증용을 6:4로 나누어 활용했다. 총 15문항으로 구성된 설문 문항은 예, 아니오로 응답하도록 되어 있으며, 우울증상의 절단점으로 8점을 사용하여 총점이 0 –7점인 경우는

정상군으로, 8-15점은 우울군으로 구분하였다. 예측변수(독립변수)는 문헌 고찰을 통해 전체 변수 중 노인의 우울과 관련된 변수들을 선별하였다. 선별된 변수는 일반적 특성 관련 변수 6개, 가족 및 사회적 관계 관련 변수 4개, 경제 상태 관련 변수 3개, 건강상태 관련 변수 9개, 건강행태 관련 변수 12개, 기능상태 관련 변수 9개, 여가 및 사회활동 관련 변수 3개, 노후생활 및 삶의 질 관련 변수 3개, 생활환경 관련 변수 4개로 총 53개의 변수가 초기 입력변수(initial input variables)로 선택되었다(박명화 외 2013). 분석은 SPSS Clementine 12.0 프로그램(현 모델러)을 활용했다. 분석 결과는 <그림 5>와 같았으며, 삶의 만족도가 가장 큰 영향을 주고 있었다.



자료: 박명화 외 2013

<그림 5> 의사결정나무 분석법을 활용한 우울 노인의 특성 분석 사례

의사결정나무 분석은 다른 분석과 같이 활용되는 경우도 많은데, 조광현 외(2011)는 의사결정나무 생성 시, 입력 변수에 대한 매개 관계를 파악하여 나무 생성에 불필요한 입력 변수를 제거하는 방법을 제시하고 그 효율성을 파악하기 위하여 실제 자료에 적용하기도 했다. 연구 결과에 따르면 논문에서 제시하는 다중매개연관성규칙을 이용한 의사결정나무모형 생성의 방법이

모형의 정확도는 거의 동일(실제로는 다소 감소)하면서 불필요한 가지를 생성하지 않으므로 효율적이라고 할 수 있다고 주장했다. 또한 김수진 외(2013)는 로지스틱 회귀분석과 의사결정나무 분석을 동시에 활용하여 대도시 주민의 우울에 영향을 주는 요인을 예측하고 비교하는 연구를 진행했다. 연구대상은 462명이었으며, 통계 프로그램은 SPSS 18.0을 활용했다. 연구결과, 로지스틱 회귀분석과 의사결정나무 분석에서 공통적으로 나타난 우울 예측요인은 사회부적응, 주관적 신체증상 및 가족 지지이었으며, 전체 분류 정확도는 로지스틱 회귀분석은 82.0%, 의사결정나무 분석은 80.5%이었다. 저자들은 연구 결과 민감성과 분류 정확도와 더 높게 나타난 로지스틱 회귀분석 방법이 지역 주민의 우울 예측 모형을 구축하는데 더 유용한 자료로 사용될 수 있으리라 주장했다.

이와 같이 의사결정나무 분석은 데이터마이닝 기법 중의 하나인 의사결정나무 분석법은 각 자료 내에 존재하는 관계와 규칙을 탐색하고 찾아내어 모형화하는 분석기법으로, 선형성(linearity), 정규성(normality) 및 등분산성(equal variance) 가정 등이 필요 없는 비모수적인 방법이다(Choi et al., 2002). 또한 의사결정나무 분석법은 분류와 예측을 위한 효과적인 방법으로 분석대상을 몇 개의 소집단으로 분류하는 규칙을 나무구조로 만들어 이를 새로운 대상에 적용함으로써 해당 특성을 예측하는 분석 방법으로 이해하기 쉬운 규칙을 생성시켜 주고 예측이 쉬워 보건의료관련 데이터의 분석에 유용 한 도구로 사용되고 있다(Bae et al., 2004, 박명화 외 2013).

제 4 장 다변량 분석을 통한 영향 요인 도출

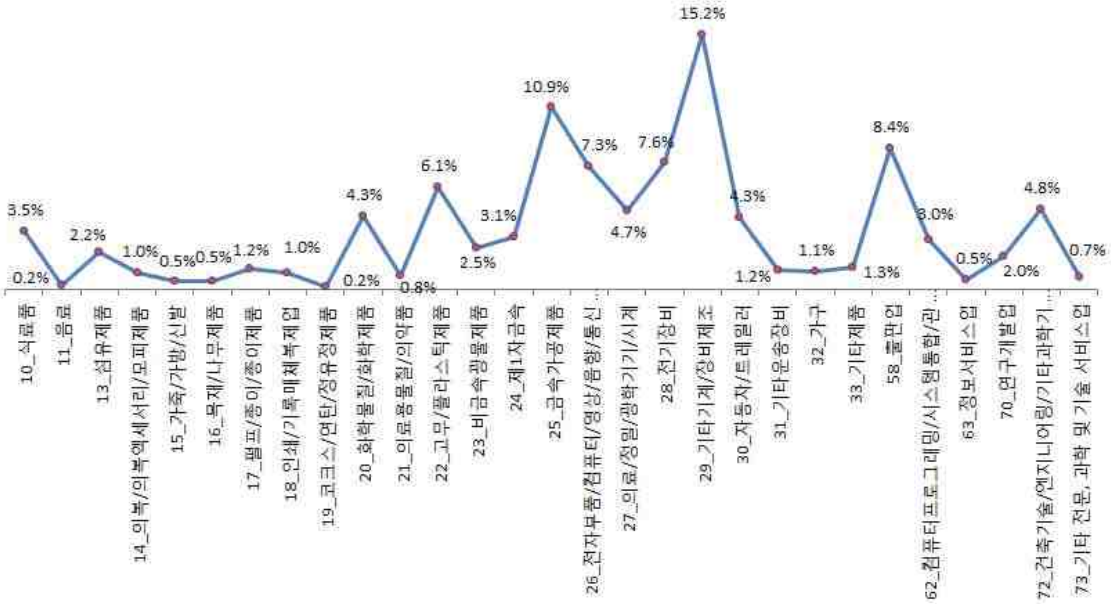
4.1 연구 사례

본 연구에서 분석에 사용된 자료는 2013년도 '제6차 중소기업 기술통계조사' 결과다. 2014년 7차 조사 자료 대신 6차 자료를 활용한 것은 <그림 2~4>에서 살펴본바와 같이 2014년 샘플보다 오히려 2013년 샘플에서 R&D 기획 지원이나 정보 지원에 대한 경험과 수요가 높았기 때문이다.

상기 조사는 중소기업기술혁신촉진법 제8조(중소기업 기술통계의 작성)에 근거하여 중소기업청과 중소기업중앙회에서 공동으로 격년마다 실시하고 있다. 설문 조사의 모집단은 종사자수 5인 이상 300인 미만인 제조업 및 제조업 외 기업 중에서 기술개발을 수행하고 있는 중소기업 38,288개사(제조업 30,864개사, 제조업 이외 업종 7,424개사)이다. 이 중에서 2012년 12월 31일 현재 기준으로 기술개발을 수행하고 있는 2,000개 중소기업(제조업 1,628개사, 제조업 이외 372개사)을 표본 추출하여 방문조사를 통해 기술혁신활동, 투자현황 및 기술수준, 기업성과 등을 조사한 설문 결과가 본 연구가 분석 대상으로 한 원시자료이다[2]. 모집단에서 표본의 추출은 층화 추출법을 활용했는데, 29개 산업중분류와 4개 종사자규모에 따라 1차 층화 추출된 후 16개 시도별로 2차 층화 추출했다. 조사 대상기간은 2011년 1월 1일부터 2012년 12월 31일까지 2년간이며, 조사는 2013년 9월 2일부터 10월 25일까지 진행되었다(중소기업청 및 중소기업중앙회 2013).

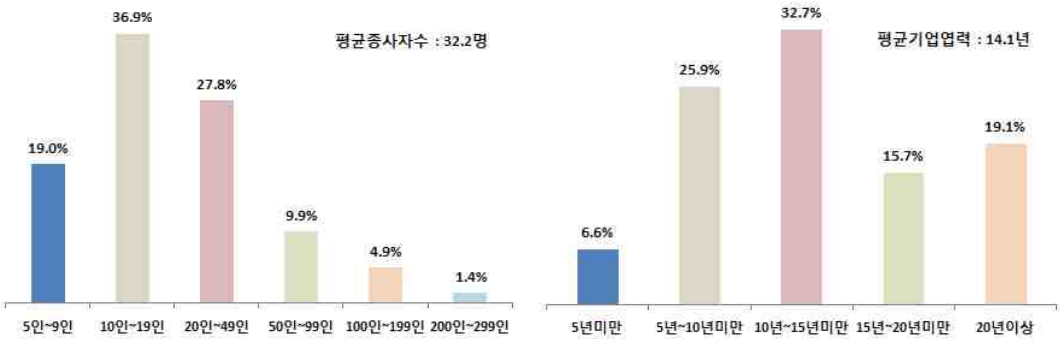
연구 사례의 업종별 분포는 <그림 6>과 같은데 기타 기계 및 장비제조가 가장 높은 분포(15.2%)를 보이고 있었다. <그림 7>을 보면, 기술개발 수행 중소기업의 평균 종사자수는 32.2명이며, 평균 종사자수 10~19명 보유 중소기업이 36.9%로 가장 많고, 그 다음으로 20~49명이 27.%로 많았다. 평균 기업업력은 14.1년으로 업력별 분포는 10~15년 미만 기업이 32.7%로 가장 많고, 5~10년 미만이 25.9%로 다음으로 많았다.

2) 연구에서 활용한 설문결과는 통계청의 2차 자료가 아니며, 중소기업중앙회에서 본 연구를 위해 제공 받은 원자료(1차 자료) 즉 2200건의 설문 원시자료를 분석한 결과이다.



자료: 중소기업청 및 중소기업중앙회 2013

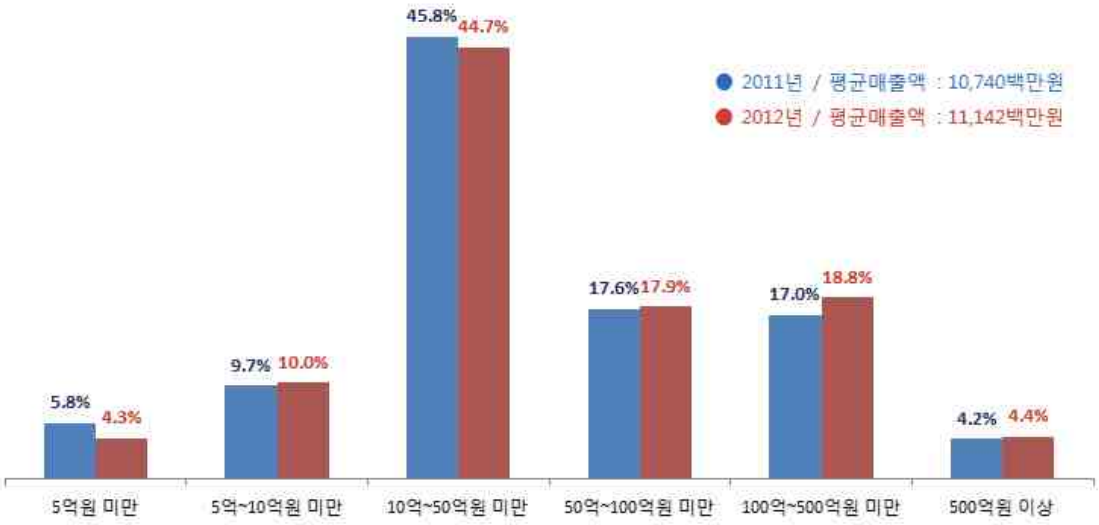
<그림 6> 연구사례의 업종별 분포



자료: 중소기업청 및 중소기업중앙회 2013

<그림 7> 연구사례의 종사자 수(좌) 및 기업 업력(우) 분포

<그림 8>에는 기술개발 수행 중소기업의 평균 매출액이 나타나 있는데 2011년도 107억 4000만원에서 2012년도 111억 4200억원으로 3.17% 증가했으며, 2012년 기준, 종사자수 50인이상 중기업의 평균 매출액(399억원 6600만원)은 소기업(55억 8300만원) 보다 7배이상 큰 것으로 나타났다. 2012년도 기준, 평균 매출액 규모별 분포를 살펴보면, 10~50억원 미만이 44.7%로 가장 많고, 100억~500억원 미만이 18.8%로 뒤를 이었다.



자료: 중소기업청 및 중소기업중앙회 2013

<그림 8> 연구사례의 매출액 규모별 분포

4.2 변수 설정

중소기업이 지향하는 기술혁신성과(기술적 성과)로는 기술개발 성공률 제고나 지재권 확보, 기술경쟁력 강화 등이 있으며, 경제적 성과로는 선획득된 기술적 성과를 바탕으로 수혜기업의 성장성(매출액), 수익성(영업이익), 생산성 지표 등이 자주 활용되고 있다(홍지승 외 2011; 유태욱 외 2009; 장선미 외 2009; 전승표 2015).

본 보고서에서도 R&D 기획지원이나 정보지원에 영향을 줄 수 있는 요인들로 이들 성과에 주목했다. 먼저 투입 요소나 기술능력으로 연구개발비는 물론 정부 지원에 대한 경험과 필요성이라는 측면에서 개발비에서 정부지원 자금이나 자체조달 자금이 차지하는 비중도 영향요인 될 수 있다고 판단했다(안치수 외 2011; 전승표 2015). 또한 기술적 성과 또는 역량이라는 측면에서 특허보유(국내, 해외)는 물론 최근의 특허출원 건수, R&D 시도/진행/성공 건수, 제품화 진행/성공 건수 등이 영향요인이 될 수 있는 것으로 보았다. 마지막으로 경제적 성과나 역량의 측면에서는 매출액, 기술에 의한 매

출액 그리고 영업이익을 잠재적 영향요인으로 주목했다.

중소기업의 성과나 역량분석에서는 다양한 통제변수들이 고려되는데, 기술기획(로드맵)의 성과에는 기술분야나 선진대비 기술수준도 영향을 끼칠 수 있다(박상문 외 2007; 김선영 외 2014), 따라서 산업분류(신기술분야, 생산기반 분야)를 통제변수로 고려하며, 세계대비 기술기획능력도 역시 통제변수로 고려하게 된다. 또한 중소기업의 혁신활동이나 성과 분석에 공통적으로 고려되는 통제변수들로 업력(설립연도), 기업규모(종업원수)가 고려되는데(안치수외 2011; 김선영 외 2014), 본 보고서에서도 업력, 종업원수는 물론 연구원수도 고려하게 된다. 그 밖에 성장단계나 연구소 보유와 같은 인증도 통제변수가 되기도 한다(김선영 외 2014). 따라서 본 보고서에서는 기업 성장단계, 이노비즈 여부, 벤처 여부도 통제 변수로 고려하게 된다. 그밖에 본 연구가 R&D 기획단계에 주목한 만큼 전체 R&D 소요기간과 그중에서 R&D 기획이 차지하는 기간을 추가로 고려했다.

앞선 제3장에서 살펴본 바와 같이 R&D 기획 지원이나 정보 지원은 필요성이나 경험은 중소기업 혁신에서 중요한 역할을 하고 있다(Bemelmans-Videc et al.(2011); 류숙원 외 2010; 이종민 외 2013). 본 보고서에는 이 변수들을 집단변수(또는 종속변수)로 분석하게 된다. 이상의 변수 설정과 관련 문헌을 정리하면 <표 2>와 같다.

<표 2> 변수명과 변수 설명

변수명			변수 설명	관련 문헌
종속변수 (집단변수)		기획단계 강화	4개 단계 중 개발기획 단계 강화를 선택한 경우(기획, 진행, 사업화, 판로개척)	Bemelmans-Videc et al.(2011); 류숙원 외 (2010); 이종민 외 (2013)
		기획지원 필요성	기획지원 필요성을 매우 높게 선택한 경우(5점만점에 5점)	
		정보지원 필요성	정보지원 필요성을 매우 높게 선택한 경우(5점만점에 5점)	
		기획지원 경험	2011~2012년 기획단계 정부지원을 경험한 경우	
		정보지원 경험	2011~2012년 (기술)정보 지원을 경험한 경우	
독립 변수	투자	연구개발비	2012년 기술개발 투자비	Hadjimanolis (2000); 장선미 외 (2009)
		자체개발비 비율	2012년 기술개발 투자비 중 자체조달 자금 비중	
		정부자금 비율	2012년 기술개발 투자비 중 정부지원 자금 비중	
	기술	지재권 출원 건수	2011~12년 지식재산권 출원 건수	류숙원 외 (2010); Audretsch et al.(2002); 이병헌 외 (2013)
		R&D 시도 건수	2011~12년 연구개발 시도 건수	
		R&D 진행 건수	2011~12년 연구개발 진행 건수	
		R&D 성공 건수	2011~12년 연구개발 성공 건수	
		제품화 진행 건수	2011~12년 신기술 제품화 진행 건수	
		제품화 성공 건수	2011~12년 신기술 제품화 성공 건수	
		(국내)특허보유 건수	조사 당시 국내 특허보유건수	
		(해외)특허보유 건수	조사 당시 해외 특허보유건수	
	경영	2012 매출액	2012년 총 매출액	홍지승 외, (2011); 유태욱 외 (2009); 장선미 외 (2009)
		2012 기술매출액	2012년 기술에 의한 매출액	
		2012 영업이익	2012년 영업이익	
통제 변수	기업 일반	업력	2015년 기준 설립 후 경과 기간	박상문 외 (2007); 안치수 외 (2011); 박문수 외(2012); 김선영 외 (2014);
		신기술분야	신기술 6T 분야(기타 포함)	
		생산기반 분야	생산기반 기술 8개 분야(기타 포함)	
		상근 총계	상근 인력 총원	
		연구원 합계	연구원 인력 총원	
		이노비즈 여부	이노비즈 기업 등록 여부	
		벤처 여부	벤처기업 등록 여부	
		성장단계	기업 성장 단계(시장 진입기, 성장기, 성숙기, 쇠퇴기)	
	기획 관련	기획능력	세계일류 기업대비 기획 능력(100점 만점)	김선영 외 (2014); 이종민 외 (2013)
		총 소요기간	기술개발 총 소요기간(개발기획, 개발진행, 사업화)	
		개발기획 소요기간	기술개발 소요기간 중 개발기획 단계 소요기간	

4.3 기술 통계

분석 대상으로 삼은 변수들 중에서 범주형 척도로 측정된 변수들의 기술통계(빈도 분석)는 <표 3>과 같다. 독립변수는 모두 연속형 변수라 <표 3>에는 나타나 있지 않다.

<표 3> 범주형 변수의 기술통계량

변수 명		범주	빈도(개)	해당비율(%)	누적비율(%)
종속변수	기획단계 강화	기타 단계 우선	1545	77.3	77.3
		기획단계 우선	455	22.8	100.0
	기획지원 필요성	약간필요함이하	1140	57.0	57.0
		매우필요함(5점)	860	43.0	100.0
	정보지원 필요성	약간필요함이하	1270	63.5	63.5
		매우필요함(5점)	730	36.5	100.0
	기획지원 경험	경험있음	475	23.8	23.8
		경험없음	1525	76.3	100.0
	정보지원 경험	경험있음	435	21.8	21.8
		경험없음	1565	78.3	100.0
통제변수	이노비즈 여부	이노비즈 해당됨	656	32.8	32.8
		이노비즈 해당없음	1344	67.2	100.0
	벤처 여부	벤처기업 해당됨	547	27.4	27.4
		벤처기업 해당없음	1453	72.7	100.0
	성장단계	시장 진입기	84	4.2	4.2
		성장기	1000	50.0	54.2
		성숙기	855	42.8	97.0
		쇠퇴기	61	3.1	100.0
	신기술분야	IT(정보기술)	310	15.5	15.5
		BT(생명공학기술)	94	4.7	20.2
		ST(우주항공기술)	51	2.6	22.8
		NT(나노기술)	10	.5	23.3
		ET(환경기술)	161	8.1	31.3
		CT(문화기술)	18	.9	32.2
		해당없음	1356	67.8	100.0
	생산기반 기술분야	주물	37	1.8	1.8
		금형	116	5.8	7.6
		용접	59	2.9	10.6
		표면처리(도금)	26	1.3	11.9
		소성가공(단조)	59	2.9	14.9
		열처리	26	1.3	16.2
		화학소재공정	198	9.9	26.1
		해당없음	1479	74.0	100.0
소계			2000	100.0	

본 연구에서 판별(영향요인)을 위한 독립변수는 모두 연속형 변수이며, 일부 통제변수도 연속형 변수로 측정되었다. 이들 연속형 변수의 기술통계량은 <표 3>와 같다.

<표 4> 연속형 변수의 기술통계량

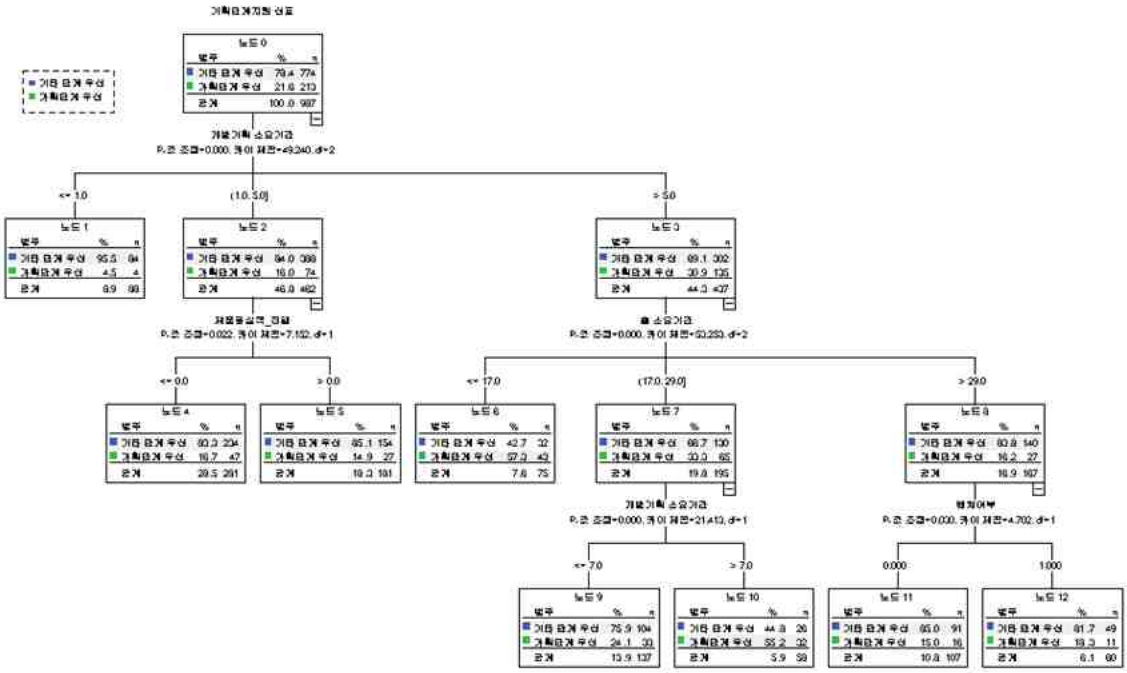
변수명	N	최소값	최대값	평균	표준편차	왜도		첨도	
	통계량	통계량	통계량	통계량	통계량	통계량	표준오차	통계량	표준오차
연구 개발비	2000	0	35000	607.08	1612.211	11.318	0.055	176.582	0.109
자체개발비 비율	1999	0.00	1.00	0.8218	.30824	-1.654	0.055	1.455	0.109
정부자금 비율	1999	0.00	1.00	0.1318	.25467	2.030	0.055	3.270	0.109
지재권 출원	2000	0	50	0.69	2.359	10.685	0.055	168.057	0.109
기술개발시도 건수	2000	0	745	7.73	27.494	15.409	0.055	327.951	0.109
기술개발진행 건수	2200	0	100	2.42	5.998	9.670	0.055	121.666	0.109
기술개발성공 건수	2000	0	600	4.57	22.342	17.407	0.055	378.746	0.109
제품화실적 진행	2000	0	150	1.05	4.412	22.517	0.055	683.910	0.109
제품화실적 성공	2000	0	500	3.34	19.525	18.074	0.055	399.041	0.109
국내 특허 보유	2000	0	230	4.65	11.337	9.555	0.055	140.796	0.109
해외 특허 보유	2000	0	45	0.31	2.104	13.275	0.055	211.393	0.109
2012 매출액	2000	1	543954	19040.95	43557.527	6.643	0.055	58.632	0.109
2012 기술매출액	2000	0	461121	5235.32	18907.682	12.075	0.055	220.347	0.109
2012 영업이익	2000	-24295	76528	1037.68	4050.333	10.876	0.055	168.169	0.109
업력	2000	3.00	73.00	17.61	10.181	1.606	0.055	3.283	0.109
상근 총계	2000	5	299	50.64	55.985	1.968	0.055	3.840	0.109
연구원 합계	2000	0	197	7.35	11.045	6.973	0.055	78.245	0.109
세계대비 기획능력	2000	0	100	71.77	21.585	-1.834	0.055	3.542	0.109
총 소요기간	2000	3	120	20.30	13.581	1.903	0.055	6.061	0.109
개발기획 소요기간	2000	0	60	5.44	4.505	4.024	0.055	35.540	0.109
유효수 (목록별)	1999								

<표 4>에서 보면 연속형 척도로 측정된 대부분 변수들은 왜도와 첨도가 매우 높게 나타났다. 따라서 변수의 분포가 정규성을 기대하기 어려웠다. 의사결정나무 분석은 앞서 설명된 바와 같이 비모수 통계 분석이기 때문에 별다른 변환 없이 원 변수값을 그대로 활용하지만, 제5장에서 예측 모델 도출을 위해서 활용한 판별분석은 독립변수의 정규성이 확보되어야하기 때문에 제5장의 판별분석에서는 변수를 변환했는데, 자연로그로 변환할 경우 대부분의 변수에서 변환 전 값에 0이 존재하기 때문에 변환이 불가능한 문제가 발생되어 제곱근으로 변환했다(단, 영업이익은 음수 값이 존재하기 때문에 변환하지 않았으며, 왜도와 첨도가 이미 안정적인 자체연구비 비율, 정부 자금 비율, 업력, 상근 총계, 기획 능력은 변환 없이 판별분석에 활용했다. 그리고 단위가 상대적으로 큰 매출액과 기술매출액은 자연로그로 변환했다).

4.4 영향 요인 분석(의사결정나무 분석)

4.4.1 개발기획 단계 지원 선호

본 연구에서는 2013년도 중소기업 기술통계조사 자료를 활용하였는데, 응답자 중 2,000명의 데이터를 분석했고 타당성 분석을 위해서 훈련용과 검정용을 5:5로 나누어 활용했다. <표 2>에서 설명된 총 25개의 설문 문항을 독립 및 통제변수들로 활용하였으며, 4가지 개발단계 중에서 기획단계의 지원 필요성이 가장 높다고 선택한 실험군(기획단계 강화 선호군)과 다른 단계 지원을 선호한 대조군으로 구분하였다. 분석은 SPSS 20.0 버전의 분류분석을 활용했다. 분석 결과는 <그림 9>와 같았으며, 개발기획 소요기간, 제품화실적 진행, 총 소요기간, 벤처여부가 25개의 변수 중에서 분류하는데 유의미한 영향 요인으로 나타났으며, 특히 총 소요기간과 기획단계 소요기간이 중요해서 기획단계 소요기간이 5개월 이상인 경우에서 총 소요기간이 17개월보다 짧은 경우 개발기획 단계 필요성을 특히 크게 보는 것으로 나타났다. 전체집단에서 21.6%가 기획단계 지원이 가장 필요하다고 생각했다면, 상기 조건을 만족하는 경우 57.3%까지 기획단계 지원을 필요로 하는 기업이 증가했다. 전반적으로 총 소요기간 대비 기획단계 소요기간이 중요한 영향으로 도출된 것이다.



<그림 9> 개발기획 단계 지원 선호군에 대한 의사결정나무 분석결과

상기 모델의 예측 정확도는 <표 5>에 나타나 있는데, 훈련 자료에서 위험도는 20.4%, 검정 자료에서는 19.9%로 나타났다. 제시된 모형은 검정 자료에서 다른 단계를 선택한 774개 기업중 716개를 정확히 분류했고(92.5%), 기획단계를 선택한 213개 기업중 75개를 정확히 분류(35.2%)를 했다. 검정 자료를 기준으로 전체 분류 정확도는 80.1%로 나타났으며, 훈련 자료의 분류 정확도는 79.6%로 나타났는데, 정확도도 80%에 가까워 높았고, 두 정확도 차이가 크지 않아서 분류가 안정적임을 확인할 수 있었다.

<표 5> 모형의 분류 정확도(개발기획 단계 지원 선호)

분류				
표본	감시됨	예측		
		기타 단계 우선	기획단계 우선	정확도(%)
훈련	기타 단계 우선	722	49	93.6%
	기획단계 우선	158	84	34.7%
	전체 퍼센트	86.9%	13.1%	79.6%
검정	기타 단계 우선	716	58	92.5%
	기획단계 우선	138	75	35.2%
	전체 퍼센트	86.5%	13.5%	80.1%

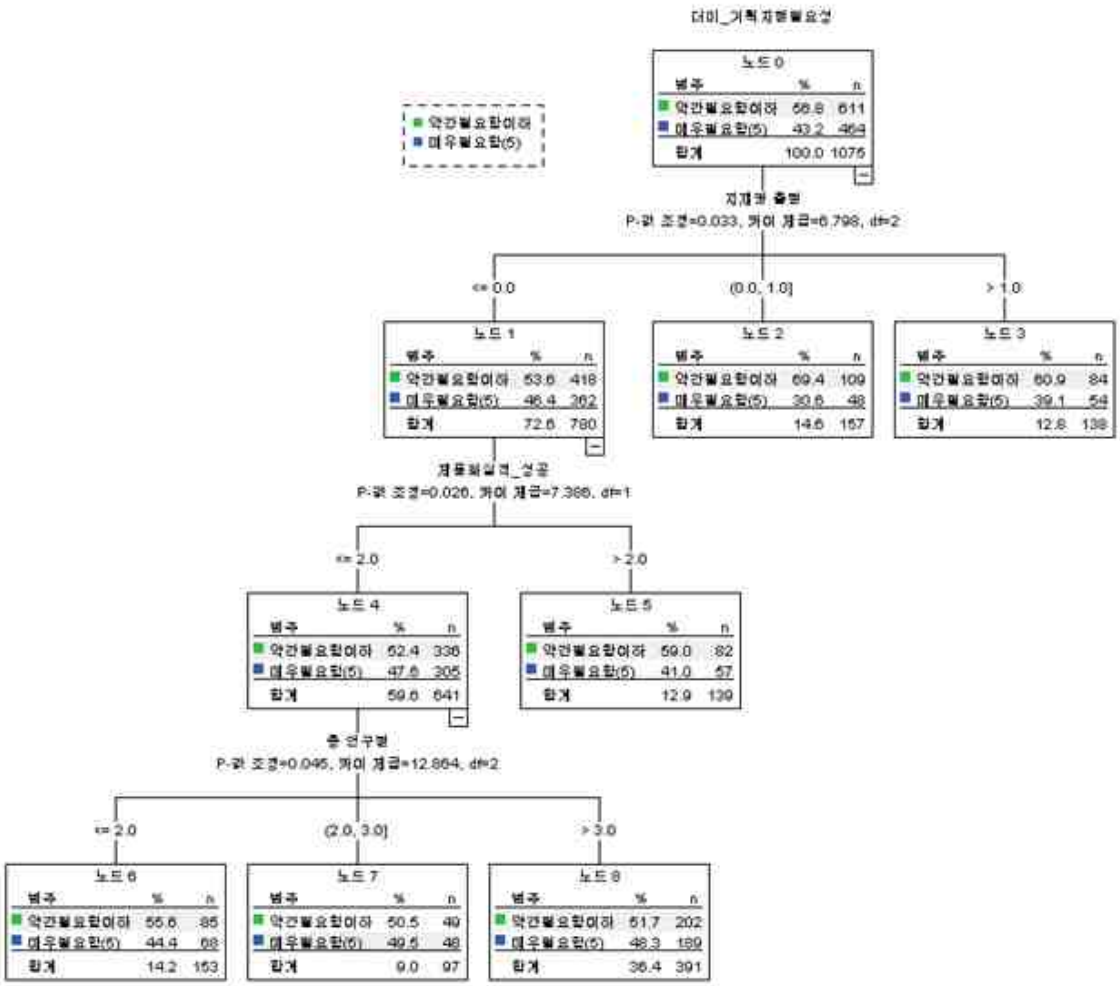
성장방법: CHAID
종속변수: 기획단계 강화

상기 분석결과는 <표 5>에 나타난 바와 같이 CHAID 방법으로 성장했으며, 최대나무 깊이는 3으로 했고, 상위 노드는 100개를 하위 노드는 50개를 최소 케이스로 설정했다. 또한 확장 기준은 유의 수준 0.05로 했고, 오분류 비용은 별도로 가중하지 않았다(추후 선호군 중심의 분류분석을 원한다면 가중하여 진행해 볼 수 있을 것이다). 이상의 분석 방법은 본 본고서 전체에서 공통적으로 적용되었다,

4.4.2 기획지원 필요성

기획지원 필요성에 대한 분석 역시 동일 데이터를 활용했으며, 역시 응답자 중 2,000명의 데이터를 훈련용과 검정용(5:5)으로 나누어 분석했다. 동일하게 총 25개의 설문 문항을 독립 및 통제변수들로 활용하였으며, 기획단계에 대한 지원 필요성이 매우 높다고 응답한(5점 만점에 5점) 실험군(기획지원 고수요군)과 약간 필요함 이하(4점이하)를 선택한 대조군으로 구분

하였다. 분석 결과는 <그림 10>과 같았으며, 지재권 출원 건수, 제품화실적 성공 건수, 연구원 합계가 25개의 변수 중에서 분류하는데 유의미한 영향 요인으로 나타났으며, 특히 연구원 합계가 중요했는데, 지재권 출원 실적이 없고, 제품화 실적도 2건이하인 그룹 중에서 총 연구원이 많지 않은 경우 기획지원의 필요성을 특히 높게 느끼는 것으로 나타났다. 전체집단에서 43.2%가 기획단계 지원이 매우 필요하다고 생각했다면, 상기 조건을 만족하는 경우 49.5%까지 해당 집단이 다소 상승했다. 전반적으로 기술 역량이 부족한 경우 기획 지원의 필요성을 높게 생각하는 것으로 나타났다.



<그림 10> 기획지원 고수요군에 대한 의사결정나무 분석결과

상기 모델의 예측 정확도는 <표 6>에 나타나 있는데, 훈련 자료에서 위험도는 40.5%, 검정 자료에서는 43.3%로 나타났다. 제시된 모형은 검정 자료에서 약간 필요함 이하를 선택한 611개 기업중 562개를 정확히 분류했고(92.0%), 매우 필요함을 선택한 464개 기업중 48개를 정확히 분류(10.3%)를 했다. 검정 자료를 기준으로 전체 분류 정확도는 56.7%로 나타났으며, 훈련 자료의 분류 정확도는 59.5%로 나타났는데, 정확도도 상대적으로 60%에 가까워 낮았고, 두 정확도 차이가 다소 존재해서 분류가 불안정적임을 확인할 수 있었다.

<표 6> 모형의 분류 정확도(기획지원 필요성)

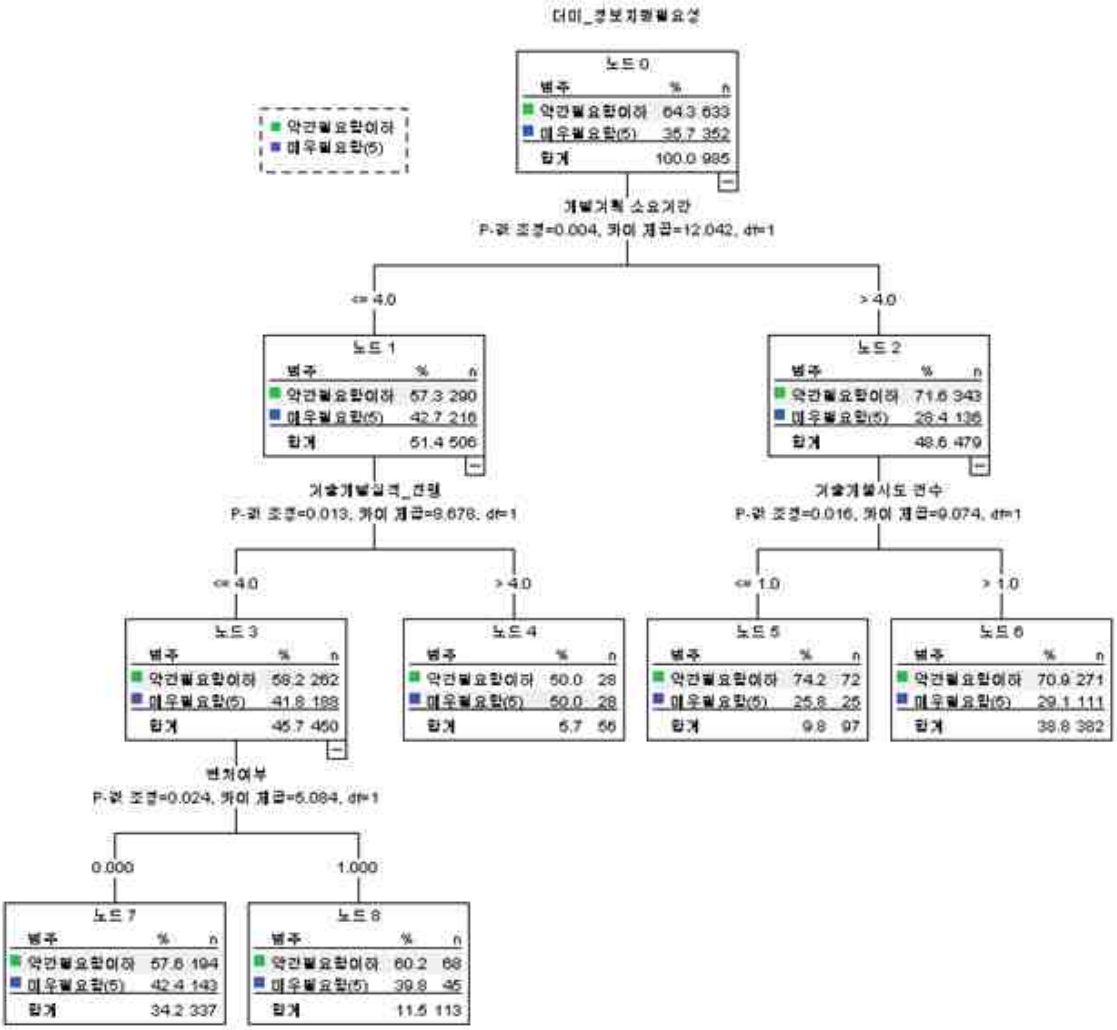
분류				
표본	감시됨	예측		
		기획지원 약간필요이하	기획지원 매우필요(5점)	정확도(%)
훈련	기획지원 약간필요이하	501	28	94.7%
	기획지원 매우필요(5점)	347	49	12.4%
	전체 퍼센트	91.7%	8.3%	59.5%
검정	기획지원 약간필요이하	562	49	92.0%
	기획지원 매우필요(5점)	416	48	10.3%
	전체 퍼센트	91.0%	9.0%	56.7%

성장방법: CHAID
종속변수: 기획지원 필요성

4.4.3 정보지원 필요성

정보지원 필요성에 대한 분석 역시 동일 데이터를 활용했으며, 역시 응답자 중 2,000명의 데이터를 훈련용과 검정용(5:5)으로 나누어 분석했다. 동일하게 총 25개의 설문 문항을 독립 및 통제변수들로 활용하였으며, (기술)정보 지원에 대한 필요성이 매우 높다고 응답한(5점 만점에 5점) 실험군(정보지원 고수요군)과 약간 필요함 이하(4점이하)를 선택한 대조군으로 구분하였다. 분석 결과는 <그림 11>과 같았으며, 개발기획 소요기간, 기술개발 성공 건수, 벤처여부, 기술개발 시도 건수가 25개의 변수 중에서 분류하는데 유의미한 영향 요인으로 나타났으며, 특히 기술개발 진행 건수가 중요했는데, 개발기획 기간이 4개월 이하인 기업 중에서 현재 기술개발을 4건이상 진행하고 있는 경우 (기술)정보지원의 필요성을 특히 높게 느끼는 것으로

나타났다. 전체집단에서 35.7%가 (기술)정보제공이 매우 필요하다고 생각했다면, 상기 조건을 만족하는 경우 50.0%까지 고수요 집단이 상승했다. 전반적으로 개발기획 기간이 짧은 분야에서 현재 진행중 기술개발 건수가 많은 경우 정보지원의 필요성을 높게 생각하는 것으로 나타났다.



<그림 11> 정보지원 고수요군에 대한 의사결정나무 분석결과

상기 모델의 예측 정확도는 <표 7>에 나타나 있는데, 훈련 자료에서 위험도는 36.1%, 검정 자료에서는 35.7%로 나타났다. 제시된 모형은 검정 자료에서 약간 필요함 이하를 선택한 633개 기업중 605개를 정확히 분류했고

(95.6%), 매우 필요함을 선택한 352개 기업중 28개를 정확히 분류(8.0%)를 했다. 검정 자료를 기준으로 전체 분류 정확도는 64.3%로 나타났으며, 훈련 자료의 분류 정확도는 63.9%로 나타났는데, 정확도도 상대적으로 65%에 가까워 낮았지만, 두 정확도 차이가 크지 않아서 분류가 안정적임을 확인할 수 있었다.

<표 7> 모형의 분류 정확도(정보지원 필요성)

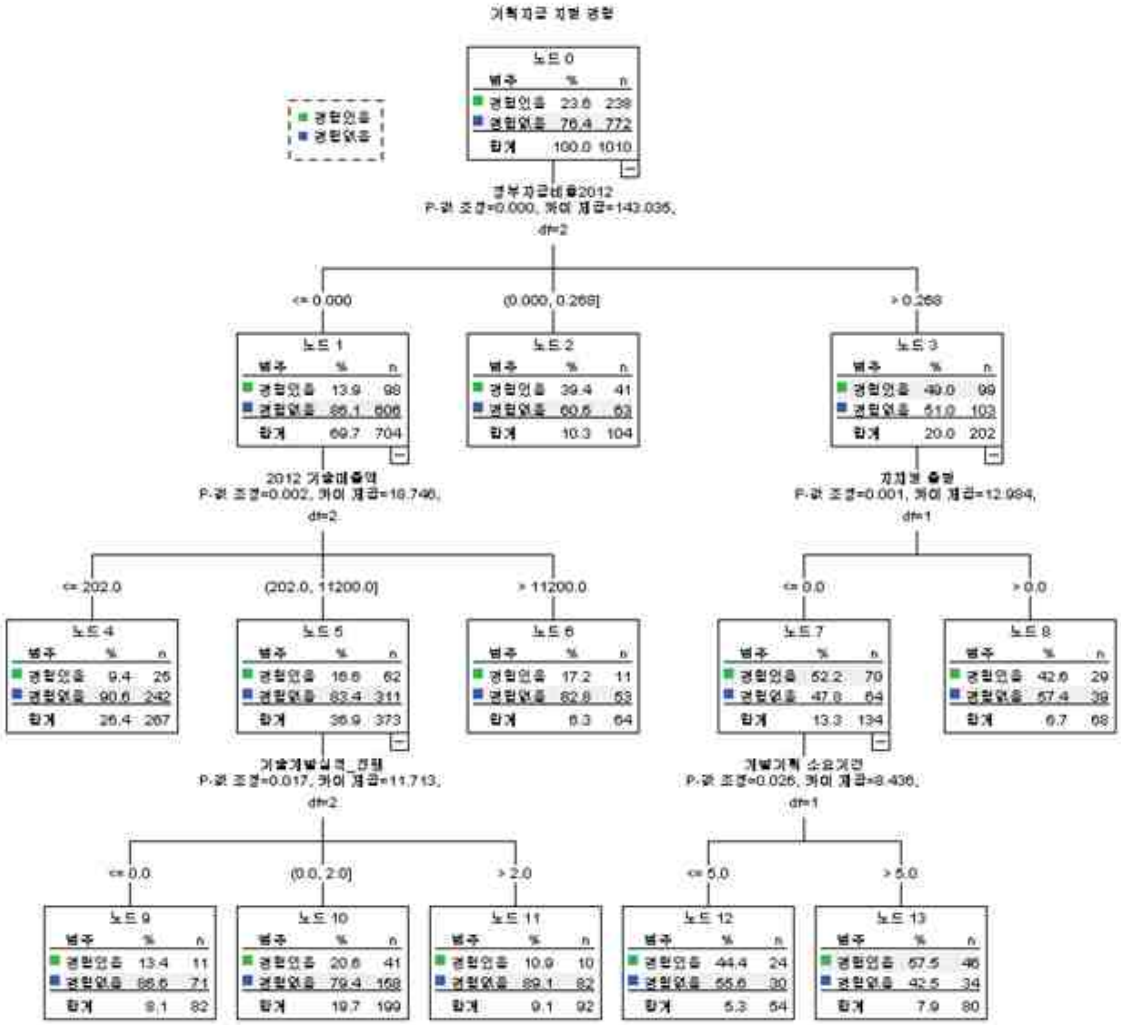
분류				
표본	감시됨	예측		
		정보지원 약간필요이하	정보지원 매우필요(5점)	정확도(%)
훈련	정보지원 약간필요이하	617	20	96.9%
	정보지원 매우필요(5점)	346	32	8.5%
	전체 퍼센트	94.9%	5.1%	63.9%
검정	정보지원 약간필요이하	605	28	95.6%
	정보지원 매우필요(5점)	324	28	8.0%
	전체 퍼센트	94.3%	5.7%	64.3%

성장방법: CHAID
종속변수: 정보지원 필요성

4.4.4 기획지원 경험

개발기획 정부지원 경험 여부에 대한 분석 역시 동일 데이터를 활용했으며, 응답자 중 2,000명의 데이터를 훈련용과 검정용(5:5)으로 나누어 분석했다. 동일하게 총 25개의 설문 문항을 독립 및 통제변수들로 활용하였으며, 개발기획 정부지원 경험이 있다고 응답한 실험군(기획지원 경험군)과 경험이 없다고 응답한 대조군으로 구분하였다. 분석 결과는 <그림 12>과 같았으며, 정부자금 비율, 2012 기술매출액, 기술개발 진행 건수, 지재권 출원 건수, 개발기획 소요기간이 25개의 변수 중에서 분류하는데 유의미한 영향 요인으로 나타났으며, 특히 정부자금비율이 중요했는데, 정부자금 비율이 26.8%이상인 기업들이 경험이 많았으며, 그 중에서도 최근 지재권 출원이 없었고, 개발기획 소요기간이 5개월 이상인 경우 개발기획 정부지원 경험이 많은 것으로 나타났다. 전체집단에서 23.6%가 기획단계 정부지원 경험이 있었는데, 상기 조건을 만족하는 경우 57.5%까지 경험 집단이 크게 상승했다. 전반적으로 이미 기획지원을 경험한 집단은 연구개발에 정부지원 자금을 많

이 받은 기업으로 나타났고, 기획기간이 긴 기업으로 나타났다. 연구개발에 정부지원 경험이 많으면 기획지원도 받은 것이다.



<그림 12> 개발기획 정부지원 경험군에 대한 의사결정나무 분석결과

상기 모델의 예측 정확도는 <표 8>에 나타나 있는데, 훈련 자료에서 위험도는 20.5%, 검정 자료에서는 22.4%로 나타났다. 제시된 모형은 검정 자료에서 경험있음을 선택한 238개 기업중 46개를 정확히 분류했고(19.3%), 경험없음을 선택한 772개 기업중 738개를 정확히 분류(95.6%)를 했다. 검정 자료를 기준으로 전체 분류 정확도는 77.6%로 나타났으며, 훈련 자료의

분류 정확도는 79.5%로 나타났는데, 정확도도 상대적으로 80%에 가까워 보통이었고, 두 정확도 차이가 크지 않아서 분류가 안정적임을 확인할 수 있었다.

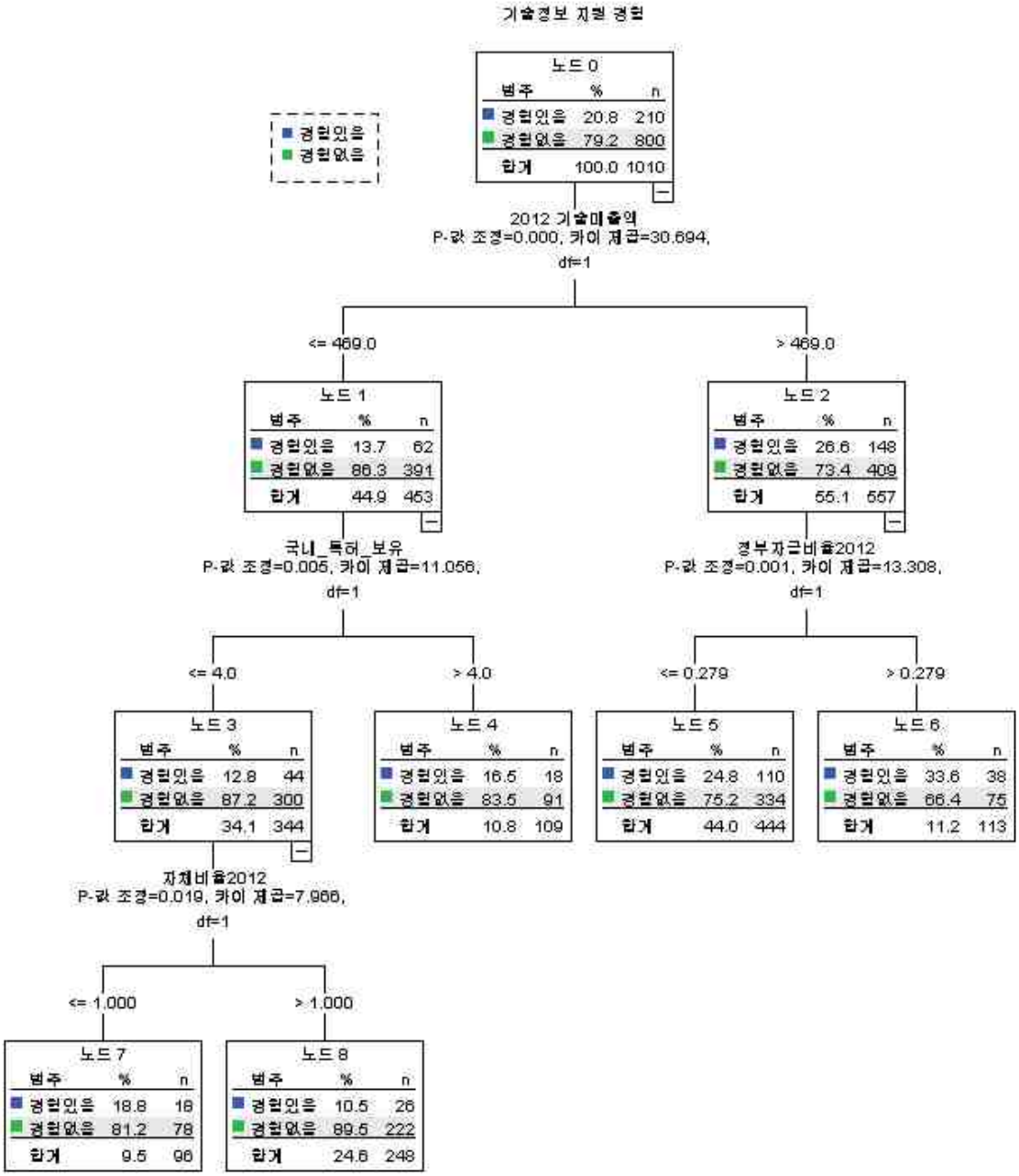
<표 8> 모형의 분류 정확도(기획지원 경험)

분류				
표본	감시됨	예측		
		기획지원 경험있음	기획지원 경험없음	정확도(%)
훈련	기획지원 경험있음	53	184	22.40%
	기획지원 경험없음	19	734	97.50%
	전체 퍼센트	7.30%	92.70%	79.50%
검정	기획지원 경험있음	46	192	19.30%
	기획지원 경험없음	34	738	95.60%
	전체 퍼센트	7.90%	92.10%	77.60%

성장방법: CHAID
종속변수: 정보지원 필요성

4.4.5 기획지원 경험

기술정보 정부지원 경험 여부에 대한 분석 역시 동일 데이터를 활용했으며, 응답자 중 2,000명의 데이터를 훈련용과 검정용(5:5)으로 나누어 분석했다. 동일하게 총 25개의 설문 문항을 독립 및 통제변수들로 활용하였으며, 정보지원 사업 경험이 있다고 응답한 실험군(정보지원 경험군)과 경험이 없다고 응답한 대조군으로 구분하였다. 분석 결과는 <그림 13>과 같았으며, 2012 기술매출액, 국내 특허 보유 건수, 자체연구비 비율, 정부자금 비율이 25개의 변수 중에서 분류하는데 유의미한 영향 요인으로 나타났으며, 특히 기술매출액이 중요했는데, 기술매출액이 469백만원 이상이고, 정부자금 비율이 27.9%이상인 기업들이 경험이 많았다. 전체집단에서 20.8%가 기획단계 정부지원 경험이 있었는데, 상기 조건을 만족하는 경우 33.6%까지 경험집단이 다소 상승했다. 전반적으로 이미 정보지원을 경험한 집단은 역시 연구개발에 정부지원 자금을 많이 받은 기업으로 나타났고, 기술매출액이 많은 기술 중심 기업으로 나타났다. 기획지원이나 정보지원 모두 연구개발에 정부지원 경험이 많은 기업이 역시 수혜 받은 것이다.



<그림 13> (기술)정보 정부지원 경험군에 대한 의사결정나무 분석결과

상기 모델의 예측 정확도는 <표 9>에 나타나 있는데, 훈련 자료에서 위험도는 22.7%, 검정 자료에서는 20.8%로 나타났다. 제시된 모형은 검정 자료에서 경험있음을 선택한 210개 기업중 0개를 정확히 분류했고(0.0%), 경험

없음을 선택한 800개 기업중 800개를 정확히 분류(100.0%)를 했다. 이런 결과는 모든 자료를 정보지원 경험이 없는 것으로 예상했기 때문으로 검정 자료를 기준으로 나타난 79.2%의 전체 분류 정확도는 큰 의미가 없었다.

<표 9> 모형의 분류 정확도(정보지원 경험)

분류				
표본	감시됨	예측		
		정보지원 경험있음	정보지원 경험없음	정확도(%)
훈련	정보지원 경험있음	0	225	0.0%
	정보지원 경험없음	0	765	100.0%
	전체 퍼센트	0.0%	100.0%	77.3%
검정	정보지원 경험있음	0	210	0.0%
	정보지원 경험없음	0	800	100.0%
	전체 퍼센트	0.0%	100.0%	79.2%

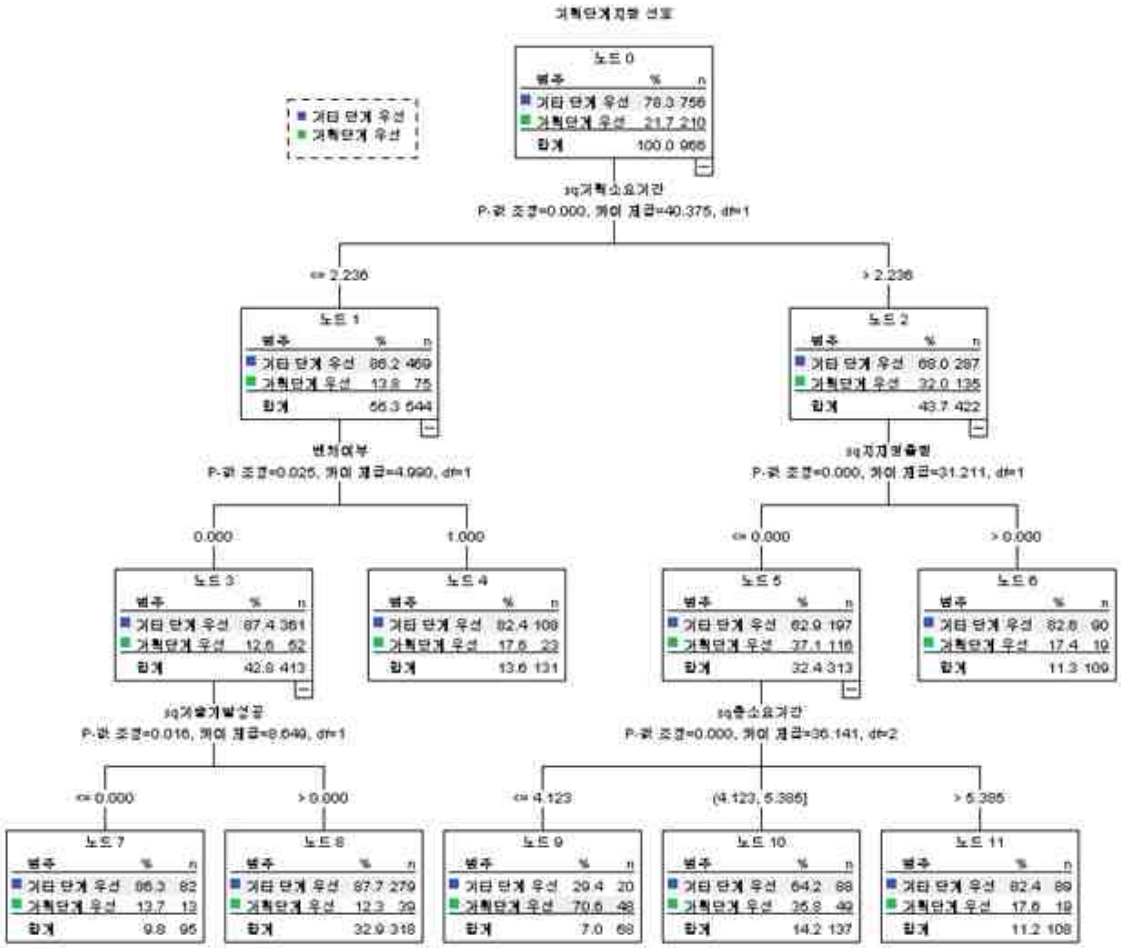
성장방법: CHAID
종속변수: 정보지원 필요성

제 5 장 수요 기업 예측 모델

5.1 개발기획 단계 지원 선호군 예측 모델(프로파일링)

제4장에서 설명한 바와 같이 의사결정나무 분석은 비모수 통계 분석으로 변수의 변환은 불필요 했으며, <그림 9~13>의 나무분석 결과를 활용해서 예측 모형 또는 프로파일링이 가능하다. 그러나 모수 통계분석인 판별분석으로 보다 강건한(robust)한 예측 모델을 제시할 수도 있기 때문에, 본 장에서는 <표 4>에서 설명된바와 같이 정규성이 크게 의심되는 변수들을 제곱근이나 자연로그로 변환해 분석한다. 판별분석은 강건한 결과를 주지만, 모수 분석에 적합한 변수를 직접 선정해야 하는 노력이 필요하다. 일단 판별분석에 적용할 독립 또는 통제 변수를 선정하기 위해서 의사결정나무 분석을 다시 활용하는데, 변환된 변수를 적용해 다시 분석을 진행했다.

역시 응답자 중 2,000명의 데이터를 훈련용과 검정용(5:5)으로 나누어 분석했다. 총 25개의 설문 문항을 독립 및 통제변수들로 활용하였으며, 4가지 개발단계 중에서 기획단계의 지원 필요성이 가장 높다고 선택한 실험군(기획단계 강화 선호군)과 다른 단계 지원을 선호한 대조군으로 구분하였다. 분석 결과는 <그림 14>와 같았으며, 기획단계 소요기간(제곱근), 벤처여부, 기술개발 성공 건수(제곱근), 지재권 출원 건수(제곱근), 총 소요기간(제곱근)이 25개의 변수 중에서 분류하는데 유의미한 영향 요인으로 나타났다. <그림 14>의 결과가 <그림 9>의 영향요인과 다르게 나타난 것은 변수변환에 따른 효과보다는 임의로 구성된 훈련용과 검정용 데이터 구성 차이 때문으로 설명할 수 있다. <그림 14> 분류의 분류 정확도는 훈련 자료에서 78.9%, 검정 자료에서 81.2% 나타나서, 80%가까운 높은 정확도와 안정적 분류 결과를 보여주었다(분류 정확도 도표는 생략함).



<그림 14> 개발기획 단계 지원 선호군에 대한 의사결정나무 분석결과(변환 변수 사용)

응답 기업 2000개사 중 개발기획 단계 지원을 선호한 기업은 455개이며, 다른 단계를 선호한 기업은 1545개 기업이다. 각 집단의 공분산행렬의 동일성에 대한 검증(Box' M)을 실시한 결과 p-value가 0.000으로서 "집단내 독립변수들 사이의 공분산 행렬이 동일하다." 는 가정은 유의수준 5%내에서 기각되어 그대로 판별분석을 할 수 없게 되었다. 즉, Box-M 통계량은 집단의 등분산성을 보는 것인데 등분산성이 인정되면 개별-집단의 공분산행렬을 사용해 적합시키면 되는데, 그렇지 않았기 때문에 집단-내 공분산행렬을 사용해 적합시켰다. 판별분석은 Fisher의 선형판별함수로 분석했고, 단계선택은 Wilks' Lamda를 기준으로 했다. 사전 확률은 집단표본 크기로 계산했다.

의사결정나무 분석에서 유의미하게 제시된 5개의 독립변수를 사용하여 개발기획 단계 지원 선호 기업과 비 선호 기업의 판별식의 적합도를 Wilks' Lamda를 통해 살펴보면, Wilks' Lamda의 수치가 작아 판별식이 통계적으로 유의미함(두 집단의 차이가 있음)을 알 수 있다(<표 10> 참조). 집단 평균의 동질성 검증결과인 <표 11>에 따르면 기획 소요기간, 기술개발 성공 건수, 지재권 출원 건수가 개발기획 단계 지원 선호 기업을 판별하는데 판별력이 통계적으로 유의미함($p<0.05$)을 보여주고 있다. 반면, 총 소요기간과 벤처 기업 여부에 있어서 집단 간 차이는 비유의적으로 나타났다. 5개의 변수 중 기획 소요기간이 가장 작은 Wilks's lamda값과 가장 큰 F값을 나타내어 대체로 판별력이 가장 높음을 알 수 있다($p<0.05$).

<표 10> Wilks's lamda(개발기획 단계 지원 선호)

함수의 검정	Wilks의 람다	카이제곱	자유도	유의확률
1	0.921	163.180	4	0.000

<표 11> 집단평균의 동질성에 대한 검정(개발기획 단계 지원 선호)

	Wilks 람다	F	자유도1	자유도2	유의확률
총 소요기간	1.000	0.064	1	1998	0.800
기획 소요기간	0.969	64.529	1	1998	0.000
기술개발 성공 건수	0.993	14.734	1	1998	0.000
지재권 출원 건수	0.990	21.089	1	1998	0.000
벤처여부	0.999	1.596	1	1998	0.207

판별분석도 회귀분석과 같이 변수의 진입 방법을 결정할 수 있는데, 본 연구에서는 단계적 진입으로 변수 진입을 결정했다(Wilks's lamda 기준). <표 12>를 살펴보면 벤처기업 여부를 제외한 4개 변수가 선택되었다(변수의 정규성이라는 측면에서 볼 때 범주형 변수의 활용은 바람직하지 않음). <표 13>은 판별 점수와 집단 간의 관계를 나타내는 정준상관계수가 제시되어 있다. 정준상관계수가 0.280으로 이를 제곱하면 0.0784이다. 즉, 종속변수의 판별 점수 분산의 7.84%가 4개의 독립변수들에 의해 설명됨을 알 수 있다.

<표 12> 단계별 판별분석 결과(개발기획 단계 지원 선호)

단계	진입된 변수	Wilks 람다							
		통계량	자유도1	자유도2	자유도3	정확한 F			
						통계량	자유도1	자유도2	유의확률
1	기획 소요기간	0.969	1	1	1998.0	64.529	1	1998.0	0.000
2	총 소요기간	0.928	2	1	1998.0	77.366	2	1997.0	0.000
3	지재권 출원 건수	0.924	3	1	1998.0	54.882	3	1996.0	0.000
4	기술개발 성공 건수	0.921	4	1	1998.0	42.488	4	1995.0	0.000

각 단계에서 전체 Wilks의 람다를 최소화하는 변수가 입력됩니다.
a. 최대 단계 수는 10입니다.
b. 입력할 최소 부분 F는 3.84입니다.
c. 제거할 최대 부분 F는 2.71입니다.
d. F 수준, 공차한계 또는 VIN 부족으로 계산을 더 수행할 수 없습니다.

<표 13> 정준 상관 계수(개발기획 단계 지원 선호)

함수	고유값	분산의 %	누적 %	정준 상관
1	0.085[a]	100.0	100.0	0.280

a. 첫 번째 1 정준 판별함수가 분석에 사용되었습니다.

개발기획 단계 지원 선호 여부에 따른 집단을 설명하는데 상대적 중요도를 나타내는 변수에 대한 표준화 정준판별함수 계수가 <표 14>에 제시되어 있는데, 기획 소요기간의 계수 절대값이 1.461로서 선호 여부의 차이를 가장 잘 설명하는 변수이며 따라서 가장 판별력(discriminant power)이 컸으며 다음은 총 소요기간이었다.

<표 14> 표준화 정준판별함수 계수와 정준판별함수 계수(개발기획 단계 지원 선호)

	표준화 정준 판별함수 계수	정준 판별함수 계수
총 소요기간	1.136	0.823
기획 소요기간	-1.461	-1.830
기술개발 성공 건수	0.189	0.117
지재권 출원 건수	0.190	0.259
(상수)		0.207

판별분석의 구조행렬에서 판별적재 값의 크기는 개발기획 단계 지원 선호 여부를 구분하는데 판별력의 크기를 말해주고 있다. <표 15>는 판별적재 값이 크기순으로 제시되었는데 이 값이 클수록 판별함수에 큰 영향을 미치는 변수이다. 또한, 판별적재 값이 ±0.30(또는 0.40)이상인 경우 유의적으로 받아들여진다(이학식 외, 2015; 이선영 외 2013). 구조행렬의 판별적재 값

이 0.3 이상인 요인은 기획 소요기간과 지재권 출원 건수이며 기술개발 성공 건수도 거의 0.3에 가까워서 판별이 높았고(구조행렬에서도 기획 소요기간이 가장 판별력이 큰 것으로 나타남). 반면, 벤처기업 여부나 총 소요기간은 판별력이 별로 없는 것으로 해석될 수 있다.

<표 15> 구조행렬(개발기획 단계 지원 선호)

	함수
	1
기획 소요기간	-0.616
지재권 출원 건수	0.352
기술개발 성공 건수	0.294
벤처여부[a]	0.024
총 소요기간	-0.019

판별변수와 표준화 정준 판별함수 간의 집단-내 통합 상관행렬.
변수는 함수내 상관행렬의 절대값 크기순으로 정렬되어 있습니다.
a. 이 변수는 분석에 사용되지 않습니다.

피셔의 선형판별함수는 개발기획 단계 지원 선호 기업과 다른 단계 지원 선호 기업을 분류하는데 4개의 독립변수들이 어떠한 영향력을 발휘하는지 말해주고 있다. <표 14>의 정준판별함수 계수에 의한 판별함수(예측 모형)는 다음과 같다.

정준판별식 = 0.823(총 소요기간(제곱근)) −1.830(기획 소요기간(제곱근)) + 0.117(기술개발 성공 건수(제곱근)) + 0.259(지재권 출원 건수(제곱근)) − 0.019

이 함수에 질문항목에 대한 수치를 대입하였을 때 다른 단계 지원을 선호한 기업의 중심점은 0.158이였으며, 기획단계 지원을 선호한 기업의 중심점은 −0.538이었다. 이는 향후에 새로운 개체(기업)가 어떤 집단으로 분류될지를 보여 주는 것으로 어떤 개체의 값이 집단의 중심점에 가까운 개체일 때는 그 집단의 구성원으로 판단할 수 있다.

다음의 <표 16>의 분류함수표는 새로운 분류대상을 어느 집단에 분류할 것인가를 결정하데 사용된다. 새로운 분류대상의 독립변수 값들을 다음의 분류함수에 각각 대입하여 결과 값이 크면 큰 값을 나타낸 집단에 분류하고

결과 값이 작으면 작은 값을 나타낸 집단에 분류한다. 이는 향후 어떤 기업이 기획단계 정부 지원을 선호할 집단인가 아닌가를 예측하는데 도움이 될 것이다. 기획단계 지원 선호 기업인 경우 상대적으로 기획 소요기간의 계수가 크고 총 소요기간의 계수는 적다. 기술개발 성공 건수와 지재권 출원 건수의 차이는 적었지만, 상대적으로 기획단계 지원 선호 기업이 낮았다. 이러한 결과는 기획단계 지원 선호에 총 소요기간 대비 투입되는 기간이 많은 기술군이 기획단계 지원 사업의 대상이 될 수 있는 가능성이 커짐을 보여주고 있다. 또한 기술개발 성공이 오히려 별 영향이 없거나 부의 영향이 있을 수도 있음을 보여주고 있다.

<표 16> 분류 함수 계수(개발기획 단계 지원 선호)

	개발기획 단계 지원 선호	
	기타 단계 우선	가획 단계 우선
총 소요기간	2.001	1.429
기획 소요기간	0.898	2.172
기술개발 성공 건수	0.880	0.799
지재권 출원 건수	0.111	-0.069
(상수)	-6.599	-6.875

Fisher의 선형 판별함수

분류함수가 표본 기업의 분류를 얼마나 잘 예측하는가를 나타내는 정오분류결과가 <표 17>에 제시되어 있는데, 의사결정나무 분석의 데이터분할과 같은 검정비교를 위해서 순차제거복원 분류를 통해 교차유효값을 산출했다. <표 17>을 보면 선호 기업 455개 중 301개(원래)와 298개(교차유효)를 각각 정확히 예측해 판별함수는 각각 66.2%와 65.5%의 적중률(hit ratio)을 보였다. 의사결정나무 분석 결과의 정확도인 <표 5>와 비교하면 전반적으로 낮아졌지만, 선호 기업군의 적중률은 오히려 2배 가까이 상승되었는데, 분산-공분산 행력이 동일하다는 가정이 충족되지 못하는 경우 보다 큰 분산-공분산 행렬을 갖는 그룹에 많은 관측치가 분류되는 문제점이 의사결정나무 분석에서 발생되었기 때문으로 설명할 수 있다. 이런 현상은 선행연구인 김선영 외(2013)에서도 나타났는데, 나부분석에서 80%에 가까운 정확도를 보였지만, 민감도가 높은 위험군은 20.8% 밖에 적중시키지 못했었다(로지스틱 회귀분석에서는 42.5%). 그런데 상기 판별분석은 실험군과 대조군의 적

중률 차이가 거의 없는 우수한 분류를 보여 주었다(<표 17> 참조).

<표 17> 모형의 분류 적중률(개발기획 단계 지원 선호)

		기획단계지원 선호	예측 소속집단		전체
			기타 단계 우선	기획 단계 우선	
원래값	빈도	기타 단계 우선	1025	520	1545
		기획 단계 우선	154	301	455
	%	기타 단계 우선	66.3	33.7	100.0
		기획 단계 우선	33.8	66.2	100.0
교차 유효값[b]	빈도	기타 단계 우선	1025	520	1545
		기획 단계 우선	157	298	455
	%	기타 단계 우선	66.3	33.7	100.0
		기획 단계 우선	34.5	65.5	100.0

a. 원래의 집단 케이스 중 66.3%이(가) 올바로 분류되었습니다.
b. 분석시 해당 케이스에 대해서만 교차유효화가 수행됩니다. 교차유효화시 각 케이스는 해당 케이스를 제외한 모든 케이스로부터 파생된 함수별로 분류됩니다.
c. 교차유효화 집단 케이스 중 66.2%이(가) 올바로 분류되었습니다.

5.2 개발기획 지원 고수요군 예측 모델(프로파일링)

역시 응답자 중 2,000명의 데이터를 훈련용과 검정용(5:5)으로 나누어 분석했다. 총 25개의 설문 문항을 독립 및 통제변수들로 활용하였으며, 기획 지원 필요성이 매우 높다고 선택한 실험군(기획 지원 고수요군)과 약간 이하의 필요성을 선택한 대조군으로 구분하였다. 분석 결과는 기획단계 소요기간(제곱근), 지재권 출원 건수(제곱근), 총 소요기간(제곱근)은ㄷ 5.1절(기획단계 선호군)과 동일하게 유의미했고, 국내 보유 특허 건수(제곱근), 제품화 성고 건수(제곱근), 연구원 합계(제곱근), 정부자금 비율도 25개의 변수 중에서 분류하는데 유의미한 영향을 줄 수 있는 후보요인으로 나타났다(의사결정나무 분석 결과는 생략함).

응답 기업 2000개사 중 기획단계 정부지원 필요성이 매우 높은 기업(고수요군)은 860개이며, 약간 이하의 필요성을 보인 기업(수요 유보군)은 1139개 기업이다. 각 집단의 공분산행렬의 동일성에 대한 검증(Box' M)을 실시한 결과 p-value가 0.096으로서 "집단내 독립변수들 사이의 공분산 행렬이 동일하다."는 가정은 유의수준 5%내에서 기각되지 않아 판별분석을 할

수 있게 되었다(즉, 개별-집단의 공분산행렬을 사용해 적합시킴).

의사결정나무 분석에서 유의미하게 제시된 7개의 독립변수를 사용하여 개발기획 지원 고수요군 기업과 수요 유보군 기업의 판별식의 적합도를 Wilks' Lamda를 통해 살펴보면, Wilks' Lamda의 수치가 작아 판별식이 통계적으로 유의미함(두 집단의 차이가 있음)을 알 수 있다(<표 18> 참조). 집단 평균의 동질성 검증결과인 <표 18>에 따르면 지재권 출원 건수와 제품화 성공 건수가 개발기획 단계 지원 선호 기업을 판별하는데 판별력이 통계적으로 유의미함($p<0.05$)을 보여주고 있었으며, 국내 특허 보유 건수나 정부자금 비율도 다소 유의($p<0.1$)하게 나타났다. 반면, 총 소요기간과 기획 소요기간, 연구원 합계에 있어서 집단 간 차이는 비유의적으로 나타났다. 7개의 변수 중 지재권 출원 건수가 가장 작은 Wilks's lamda값과 가장 큰 F값을 나타내어 대체로 판별력이 가장 높음을 알 수 있다($p<0.05$).

<표 18> Wilks's lamda(개발기획 지원 고수요)

함수의 검정	Wilks의 람다	카이제곱	자유도	유의확률
1	0.993	14.882	2	0.001

<표 19> 집단평균의 동질성에 대한 검정(개발기획 지원 고수요)

	Wilks 람다	F	자유도1	자유도2	유의확률
기획 소요기간	1.000	0.000	1	1997	0.997
총 소요기간	1.000	0.020	1	1997	0.889
지재권 출원 건수	0.996	7.346	1	1997	0.007
국내 특허 보유 건수	0.999	2.902	1	1997	0.089
제품화 성공 건수	0.998	4.990	1	1997	0.026
연구원 합계	1.000	0.079	1	1997	0.779
정부자금 비율	0.998	3.603	1	1997	0.058

판별분석도 회귀분석과 같이 변수의 진입 방법을 결정할 수 있는데, 본 연구에서는 단계적 진입으로 변수 진입을 결정했다(Wilks's lamda 기준). <표 20>을 살펴보면 지재권 출원 건수와 국내 특허 보유 건수만 선택되었고 다른 5개 변수는 제외되었다. <표 21>은 판별 점수와 집단 간의 관계를 나타내는 정준상관계수가 제시되어 있다. 정준상관계수가 0.086으로 이를 제곱하면 0.0074이다. 즉, 종속변수의 판별 점수 분산의 0.74%가 2개의 독립변

수들에 의해 설명됨을 알 수 있다.

<표 20> 단계별 판별분석 결과(개발기획 지원 고수요)

단계	진입된 변수	Wilks 람다							
		통계량	자유도1	자유도2	자유도3	정확한 F			
						통계량	자유도1	자유도2	유의확률
1	지재권 출원 건수	0.996	1	1	1997.0	7.346	1	1997.0	0.007
2	국내 특허 보유 건수	0.993	2	1	1997.0	7.469	2	1996.0	0.001

각 단계에서 전체 Wilks의 람다를 최소화하는 변수가 입력됩니다.
a. 최대 단계 수는 14입니다.
b. 입력할 최소 부분 F는 3.84입니다.
c. 제거할 최대 부분 F는 2.71입니다.
d. F 수준, 공차한계 또는 VIN 부족으로 계산을 더 수행할 수 없습니다.

<표 21> 정준 상관 계수(개발기획 지원 고수요)

함수	고유값	분산의 %	누적 %	정준 상관
1	0.007[a]	100.0	100.0	0.086

a. 첫 번째 1 정준 판별함수가 분석에 사용되었습니다.

개발기획 지원 고수요군 여부에 따른 집단을 설명하는데 상대적 중요도를 나타내는 변수에 대한 표준화 정준판별함수 계수가 <표 22>에 제시되어 있는데, 지재권 출원 건수의 계수 절대값이 0.951로서 선호 여부의 차이를 가장 잘 설명하는 변수이며 따라서 가장 판별력(discriminant power)이 컸으며 다음은 국내 특허 보유 건수였다.

<표 22> 표준화 정준판별함수 계수와 정준판별함수 계수(개발기획 지원 고수요)

	표준화 정준 판별함수 계수	정준 판별함수 계수
지재권 출원 건수	0.951	1.293
국내 특허 보유 건수	-0.756	-0.488
(상수)		0.234

판별분석의 구조행렬에서 판별적재 값의 크기는 개발기획 지원 고수요군 여부를 구분하는데 판별력의 크기를 말해주고 있다. <표 23>은 판별적재 값이 크기순으로 제시되었는데 이 값이 클수록 판별함수에 큰 영향을 미치는 변수이다. 또한, 판별적재 값이 ±0.30(또는 0.40)이상인 경우 유의적으로 받아들여진다(이학식 외, 2015; 이선영 외 2013). 구조행렬의 판별적재 값

이 0.4 이상인 요인은 역시 지재권 출원 건수와 국내 특허 보유 건수였다. 반면, 다른 5개 변수는 판별력이 별로 없는 것으로 해석될 수 있다.

<표 23> 구조행렬(개발기획 지원 고수요)

	함수
	1
지재권 출원 건수	0.701
국내 특허 보유 건수	-0.441
제품화 성공 건수	0.148
기획 소요 기간a	-0.105
총 소요기간a	-0.071
연구원 합계a	-0.053
정부자금비율2012a	-0.005

판별변수와 표준화 정준 판별함수 간의 집단-내 통합 상관행렬.
변수는 함수내 상관행렬의 절대값 크기순으로 정렬되어 있습니다.
a. 이 변수는 분석에 사용되지 않습니다.

피셔의 선형판별함수는 개발기획 지원 고수요군 기업과 수요 유보군 기업을 분류하는데 2개의 독립변수들이 어떠한 영향력을 발휘하는지 말해주고 있다. <표 22>의 정준판별함수 계수에 의한 판별함수(예측 모형)는 다음과 같다.

정준판별식 = 1.293(지재권 출원 건수(제곱근)) - 0.488(국내 특허 보유 건수(제곱근)) + 0.234

이 함수에 질문항목에 대한 수치를 대입하였을 때 개발기획 지원 고수요군 기업의 중심점은 -0.100이였으며, 수요 유보군 기업의 중심점은 0.075이었다. 이는 향후에 새로운 개체(기업)가 어떤 집단으로 분류될지를 보여 주는 것으로 어떤 개체의 값이 집단의 중심점에 가까운 개체일 때는 그 집단의 구성원으로 판단할 수 있다.

다음의 <표 24>의 분류함수표는 새로운 분류대상을 어느 집단에 분류할 것인가를 결정하데 사용된다. 새로운 분류대상의 독립변수 값들을 다음의 분류함수에 각각 대입하여 결과 값이 크면 큰 값을 나타낸 집단에 분류하고 결과 값이 작으면 작은 값을 나타낸 집단에 분류한다. 이는 향후 어떤 기업이 개발기획 지원 고수요군 집단인가 아닌가를 예측하는데 도움이 될 것이

다. 개발기획 지원 고수요군 기업인 경우 상대적으로 국내 특허 보유 건수의 계수가 크고 지재권 출원 건수의 계수는 적다. 이러한 결과는 개발기획 지원 고수요는 국내 특허를 다수 보유하고 있지만, 최근 지재권을 출원이 부진한 기술군이 커서 이들 기업들이 기획단계 지원 사업의 대상이 될 수 있는 가능성이 커짐을 보여주고 있다.

<표 24> 분류 함수 계수(개발기획 지원 고수요)

	개발기획 지원 필요성(고수요)	
	수요 유보군	고수요군
지재권 출원 건수	0.406	0.180
국내 특허 출원 건수	0.540	0.625
(상수)	-1.170	-1.213

Fisher의 선형 판별함수

분류함수가 표본 기업의 분류를 얼마나 잘 예측하는가를 나타내는 정오분류결과가 <표 25>에 제시되어 있는데, 의사결정나무 분석의 데이터분할과 같은 검정비교를 위해서 순차제거복원 분류를 통해 교차유효값을 산출했다. <표 25>를 보면 고수요군 기업 860개 중 445개(원래)와 443개(교차유효)를 각각 정확히 예측해 판별함수는 각각 51.7%와 51.5%의 낮은 적중률(hit ratio)을 보였다. 의사결정나무 분석 결과의 정확도인 <표 6>과 비교하면 전반적으로 낮아졌지만, 선호 기업군의 적중률은 오히려 4배 가까이 상승되어 고수요기업을 판별하고자할 경우 오히려 보다 적절한 판별력을 보였다.

<표 25> 모형의 분류 적중률(개발기획 지원 고수요)

		기획지원 필요성	예측 소속집단		전체
			수요 유보군	고수요군	
원래값	빈도	수요 유보군	667	473	1140
		고수요군	415	445	860
	%	수요 유보군	58.5	41.5	100.0
		고수요군	48.3	51.7	100.0
교차 유효값[b]	빈도	수요 유보군	666	474	1140
		고수요군	417	443	860
	%	수요 유보군	58.4	41.6	100.0
		고수요군	48.5	51.5	100.0

a. 원래의 집단 케이스 중 55.6%이(가) 올바로 분류되었습니다.
b. 분석시 해당 케이스에 대해서만 교차유효화가 수행됩니다. 교차유효화시 각 케이스는 해당 케이스를 제외한 모든 케이스로부터 파생된 함수별로 분류됩니다.
c. 교차유효화 집단 케이스 중 55.5%이(가) 올바로 분류되었습니다.

5.3 기술정보 지원 고수요군 예측 모델(프로파일링)

역시 응답자 중 2,000명의 데이터를 훈련용과 검정용(5:5)으로 나누어 분석했다. 총 25개의 설문 문항을 독립 및 통제변수들로 활용하였으며, 정보 지원 필요성이 매우 높다고 선택한 실험군(정보 지원 고수요군)과 약간 이하의 필요성을 선택한 대조군(수요 유보군)으로 구분하였다. 분석 결과는 기획단계 소요기간(제곱근), 총 소요기간(제곱근), 벤처 기업 여부는 5.1절(기획단계 선호군)과 동일하게 유의미했고, 국내 보유 특허 건수(제곱근), 영업이익, 업력이 25개의 변수 중에서 분류하는데 유의미한 영향을 줄 수 있는 후보요인으로 나타났다(의사결정나무 분석 결과는 생략함).

응답 기업 2000개사 중 기술정보 지원 필요성이 매우 높다고 응답한 기업은 730개이며, 약간 이하의 필요성을 응답한 기업은 1270개 기업이다. 각 집단의 공분산행렬의 동일성에 대한 검증(Box' M)을 실시한 결과 p-value가 0.000으로서 "집단내 독립변수들 사이의 공분산 행렬이 동일하다."는 가정은 유의수준 5%내에서 기각되어 그대로 판별분석을 할 수 없게 되었다. 즉, Box-M 통계량은 집단의 등분산성을 보는 것인데 등분산성이 인정되면 개별-집단의 공분산행렬을 사용해 적합시키면 되는데, 그렇지 않았기 때문에 집단-내 공분산행렬을 사용해 적합시켰다.

의사결정나무 분석에서 유의미하게 제시된 6개의 독립변수를 사용하여 (기술)정보지원 고수요군 기업과 수요 유보군 기업의 판별식 적합도를 Wilks' Lamda를 통해 살펴보면, Wilks' Lamda의 수치가 작아 판별식이 통계적으로 유의미함(두 집단의 차이가 있음)을 알 수 있다(<표 26> 참조).

<표 26> Wilks's lamda(기술정보 지원 고수요)

함수의 검정	Wilks의 람다	카이제곱	자유도	유의확률
1	0.979	41.906	2	0.000

집단 평균의 동질성 검증결과인 <표 27>에 따르면 기획 소요기간만 정보 지원 고수요군 기업을 판별하는데 판별력이 통계적으로 유의미함($p<0.05$)을

보여주고 있다. 반면, 다른 5개 변수에 있어서 집단 간 차이는 비유의적으로 나타났다. 6개의 변수 중 기획 소요기간이 가장 작은 Wilks's lamda값과 가장 큰 F값을 나타내어 대체로 판별력이 가장 높음을 알 수 있다($p<0.05$).

<표 27> 집단평균의 동질성에 대한 검정(기술정보 지원 고수요)

	Wilks 람다	F	자유도1	자유도2	유의확률
기획 소요기간	0.991	18.870	1	1998	0.000
총 소요기간	1.000	0.056	1	1998	0.813
벤처 기업 여부	1.000	0.235	1	1998	0.628
국내 특허 보유 건수	0.999	2.410	1	1998	0.121
업력	1.000	0.620	1	1998	0.431
영업이익(2012)	1.000	0.005	1	1998	0.941

판별분석도 회귀분석과 같이 변수의 진입 방법을 결정할 수 있는데, 본 연구에서는 단계적 진입으로 변수 진입을 결정했다(Wilks's lamda 기준). <표 28>을 살펴보면 기획 소요기간과 총 소요기간만 선택되었고 다른 4개 변수는 제외되었다. <표 29>는 판별 점수와 집단 간의 관계를 나타내는 정준상관계수가 제시되어 있다. 정준상관계수가 0.144로 이를 제곱하면 0.0207이다. 즉, 종속변수의 판별 점수 분산의 2.07%가 2개의 독립변수들에 의해 설명됨을 알 수 있다.

<표 28> 단계별 판별분석 결과(기술정보 지원 고수요)

단계	진입된 변수	Wilks 람다							
		통계량	자유도1	자유도2	자유도3	정확한 F			
						통계량	자유도1	자유도2	유의확률
1	기획 소요기간	0.991	1	1	1998.0	18.870	1	1998.0	0.000
2	총 소요기간	0.979	2	1	1998.0	21.175	2	1997.0	0.000

각 단계에서 전체 Wilks의 람다를 최소화하는 변수가 입력됩니다.
a. 최대 단계 수는 12입니다.
b. 입력할 최소 부분 F는 3.84입니다.
c. 제거할 최대 부분 F는 2.71입니다.
d. F 수준, 공차한계 또는 VIN 부족으로 계산을 더 수행할 수 없습니다.

<표 29> 정준 상관 계수(기술정보 지원 고수요)

함수	고유값	분산의 %	누적 %	정준 상관
1	0.021[a]	100.0	100.0	0.144

a. 첫 번째 1 정준 판별함수가 분석에 사용되었습니다.

기술정보 지원 고수요군 여부에 따른 집단을 설명하는데 상대적 중요도를 나타내는 변수에 대한 표준화 정준판별함수 계수가 <표 30>에 제시되어 있는데, 기획 소요기간의 계수 절대값이 1.562로서 선호 여부의 차이를 가장 잘 설명하는 변수이며 따라서 가장 판별력(discriminant power)이 컸으며 다음은 총 소요기간이었다.

<표 30> 표준화 정준판별함수 계수와 정준판별함수 계수(기술정보 지원 고수요)

	표준화 정준 판별함수 계수	정준 판별함수 계수
기획 소요기간	1.562	1.934
총 소요기간	-1.164	-0.844
(상수)		-0.610

판별분석의 구조행렬에서 판별적재 값의 크기는 개발기획 지원 고수요군 여부를 구분하는데 판별력의 크기를 말해주고 있다. <표 31>은 판별적재 값이 크기순으로 제시되었는데 이 값이 클수록 판별함수에 큰 영향을 미치는 변수이다. 또한, 판별적재 값이 ±0.30(또는 0.40)이상인 경우 유의적으로 받아들여진다(이학식 외, 2015; 이선영 외 2013). 구조행렬의 판별적재 값이 0.4 이상인 요인은 기획 소요기간 뿐이었다. 반면, 다른 5개 변수는 판별력이 별로 없는 것으로 해석될 수 있다.

<표 31> 구조행렬(기술정보 지원 고수요)

	함수
	1
기획 소요기간	0.667
업력a	-0.077
국내 특허 보유 건수a	-0.062
총 소요기간	0.036
영업이익(2012)a	0.007
벤처 기업 여부a	-0.005

판별변수와 표준화 정준 판별함수 간의 집단-내 통합 상관행렬.
변수는 함수내 상관행렬의 절대값 크기순으로 정렬되어 있습니다.
a. 이 변수는 분석에 사용되지 않습니다.

피셔의 선형판별함수는 기술정보 지원 고수요군 기업과 수요 유보군 기업을 분류하는데 2개의 독립변수들이 어떠한 영향력을 발휘하는지 말해주고

있다. <표 30>의 정준판별함수 계수에 의한 판별함수(예측 모형)는 다음과 같다.

정준판별식 = 1.934(기획 소요기간(제곱근)) - 0.844(총 소요기간(제곱근)) - 0.610

이 함수에 질문항목에 대한 수치를 대입하였을 때 정보 지원 고수요군 기업의 중심점은 -0.192이였으며, 수요 유보군 기업의 중심점은 0.110이었다. 이는 향후에 새로운 개체(기업)가 어떤 집단으로 분류될지를 보여 주는 것으로 어떤 개체의 값이 집단의 중심점에 가까운 개체일 때는 그 집단의 구성원으로 판단할 수 있다.

다음의 <표 32>의 분류함수표는 새로운 분류대상을 어느 집단에 분류할 것인가를 결정하데 사용된다. 새로운 분류대상의 독립변수 값들을 다음의 분류함수에 각각 대입하여 결과 값이 크면 큰 값을 나타낸 집단에 분류하고 결과 값이 작으면 작은 값을 나타낸 집단에 분류한다. 이는 향후 어떤 기업이 정보 지원 고수요군 집단인가 아닌가를 예측하는데 도움이 될 것이다. 기술정보 지원 고수요군 기업인 경우 상대적으로 총 소요기간의 계수가 크고 기획 소요기간의 계수는 적다. 이러한 결과는 정보 지원 고수요는 기획기간은 짧지만 총 기술개발 소요기간은 장시간인 기술군과 관련된 기업들이 정보 지원 사업의 대상이 될 수 있는 가능성이 커짐을 보여주고 있다. 5.1절에서 개발기획 단계 지원 선호군과 반대되는 경향임에 주목할 필요가 있다.

<표 32> 분류 함수 계수(기술정보 지원 고수요)

	기술정보 지원 필요성(고수요)	
	수요 유보군	고수요군
기획 소요기간	1.176	0.591
총 소요기간	1.728	1.983
(상수)	-5.723	-5.551

Fisher의 선형 판별함수

분류함수가 표본 기업의 분류를 얼마나 잘 예측하는가를 나타내는 정오분류결과가 <표 33>에 제시되어 있는데, 의사결정나무 분석의 데이터분할과

같은 검정비교를 위해서 순차제거복원 분류를 통해 교차유효값을 산출했다. <표 33>을 보면 고수요군 기업 730개 중 433(원래)와 433개(교차유효)를 각각 정확히 예측해 판별함수는 모두 59.3%의 적중률(hit ratio)을 보였다. 의사결정나무 분석 결과의 정확도인 <표 7>과 비교하면 다소 낮아졌지만, 선호 기업군의 적중률은 오히려 7배 가까이 크게 상승되어 고수요기업을 판별하고자할 경우 오히려 보다 적절한 판별력을 보였다.

<표 33> 모형의 분류 적중률(기술정보 지원 고수요)

		정보지원 필요성	예측 소속집단		전체
			수요 유보군	고수요군	
원래값	빈도	수요 유보군	711	559	1270
		고수요군	297	433	730
	%	수요 유보군	56.0	44.0	100.0
		고수요군	40.7	59.3	100.0
교차 유효값[b]	빈도	수요 유보군	711	559	1270
		고수요군	297	433	730
	%	수요 유보군	56.0	44.0	100.0
		고수요군	40.7	59.3	100.0

a. 원래의 집단 케이스 중 57.2%이(가) 올바로 분류되었습니다.
b. 분석시 해당 케이스에 대해서만 교차유효화가 수행됩니다. 교차유효화시 각 케이스는 해당 케이스를 제외한 모든 케이스로부터 파생된 함수별로 분류됩니다.
c. 교차유효화 집단 케이스 중 57.2%이(가) 올바로 분류되었습니다.

제 6 장 결론 및 정책적 시사점

6.1 결론

정부는 재무적 또는 비재무적으로 다양한 중소기업 R&D 지원정책을 수행하고 있는데, 그 중에서 재무적인 지원은 많은 연구와 경험을 통해 상대적으로 정책적 효과 연구가 진행되어 왔다. 그러나 비 재무적인 지원 정책에 대한 연구는 최근 활발해져 왔는데, 대표적인 비재무적 지원 정책중 하나인 R&D 정보 지원의 효과나 관련 기관의 역할에 대해 특히 거의 연구된 바가 없다.

본 보고서에서는 이런 R&D 정보 지원과 중소기업의 여러 가지 조건들이 어떤 관계가 있는지 탐색적으로 연구하고자 했다. R&D 정보 지원을 경험한 기업들(실험집단)의 특징과 대조군의 특징을 비교분석해서 프로파일링 했고, 나아가주요 수요 집단으로 예상되는 R&D 기획 지원 고수요 집단의 특징도 프로파일링(영향 요인)했다. 의사결정나무 분석의 결과에 따르면 R&D 기획 지원이나 정보지원을 경험 기업을 분류하는 공통적으로 큰 영향을 주는 요인 정부자금비율이었다. 즉 최근까지 R&D 정보 제공을 받은 기업은 연구개발 투자에서 전반적으로 정부 지원이 차지하는 비중이 높았다. 또한 기술 매출액과 특허관련 변수(지재권 출원, 국내 특허 보유)이 분류에 영향을 주는 요인으로 밝혀졌다. 기술적 성과가 이미 어느 수준 이상인 기업들이 이미 R&D 정보 지원을 경험한 것으로 나타났다.

미래 수요와 관련해서 R&D 기획 지원이나 정보지원 필요성을 아주 높게 생각하고 있는 기업을 프로파일링한 결과, R&D 기획에는 연구원 합계, 지재권 출원, 제품화 실적 등이 중요한 영향 요인이고, 정보 지원에는 개발이나 기획 소요기간, 기술개발 실적, 벤처 기업 여부 등이 중요한 영향으로 나타났다. 또한 연구개발의 4개 단계 중에서 개발기획 단계의 지원이 가장 중요하다 생각한 기업을 프로파일링한 결과는 개발이나 기획 소요기간, 제품화 실적, 벤처 기업 여부가 중요한 영향 요인으로 밝혀졌다. 여기서 주목해야 하는 것은 그동안 간과했던, 총 개발기간과 기획 소요기간이 수요 대상 기업이 관련 R&D 정보 제공을 필요로 하는데 큰 영향을 주고 있다는 것이다.

또한 R&D 기획을 필요로 하는 기업과 정보 제공을 원하는 기업에서 총 개발 기간과 기간의 특징이 달랐다. 기획의 지원을 필요로 하는 기업은 총 개발 기간에서 기획기간이 상대적 긴 기업이 기획지원을 필요로 했던 반면, 정보지원을 원하는 기업은 오히려 총 개발기간에 비해서 기획기간의 상대적으로 짧으며, 특히 개발기간이 긴 기업이었다.

판별분석을 통해서 구체적인 수요 예측 모형도 개발했다. 개발기획 단계 지원, R&D 기획 지원, 기술정보 지원의 기업 수요를 예측할 수 있는 정준판별함수를 도출했고 그 결과를 정리하면 다음과 같다.

1) 개발기획 단계 지원 선호 (정준)판별식 = 0.823(총 소요기간(제곱근)) −1.830(기획 소요기간(제곱근)) + 0.117(기술개발 성공 건수(제곱근)) + 0.259(지재권 출원 건수(제곱근)) − 0.019
2) R&D 기획 지원 수요 (정준)판별식 = 1.293(지재권 출원 건수(제곱근)) − 0.488(국내 특허 보유 건수(제곱근)) + 0.234
3) 기술정보 지원 (정준)판별식 = 1.934(기획 소요기간(제곱근)) − 0.844(총 소요기간(제곱근)) − 0.610

<표 34> 예측모형별 판별함수의 중심점

예측모형	고수요군	저수요군
개발기획 단계 지원 선호	-0.538	0.158
R&D 기획 지원 수요	-0.100	0.075
기술정보 지원 수요	-0.192	0.110

상기식과 중심점을 활용하면, 새로운 분류대상을 어느 집단에 분류할 것인가를 결정할 수 있다. 새로운 분류대상의 독립변수 값들을 상기 판별식에 대입하여 결과 값이 어느 중심점에 가까운지로 집단을 분류할 수 있다. 이는 향후 어떤 기업이 정보 지원 고수요군 집단인가 아닌가를 예측하는데 도움이 될 것이다.

이상의 도출 과정에서는 다변량 분석을 통한 접근의 장단점을 확인할 수 있었는데, 의사결정나무 분석 방법은 매우 많은 수의 변량을 별다른 조치(변수 변환)등이 없이 활용해서 직관적으로 영향을 요인을 찾고 간편하게 예측

모델을 만들 수 있다는 장점을 확인할 수 있었던 반면, 시뮬레이션 마다 다소 다른 결과가 제시되어(훈련/검정 자료 구분 등) 결과 해석이 어렵고, 상대적으로 작은 값(수요군, 경험군)은 잘 예측하지 못하는 문제점을 확인할 수 있었다. 판별분석은 독립변수의 선정에 노력이 필요하고 변환의 노력까지 필요하며, 상대적으로 낮은 정확도를 보였지만, 다양한 결과 제공과 비교적 균형 잡힌(수요군 예측력) 정확도를 보여주는 확인할 수 있었다.

향후에는 자료를 추가로 확보하고, 독립변수의 대상을 확보할 필요가 있으며, 의사결정나무 분석의 다양한 분석방법이나 조건에 대한 추가 적용이 필요하다, 특히 판별분석을 활용하는 경우 각 변수의 선정에 대한 이론적 근거의 확보도 필요하다고 할 수 있다.

6.2 정책적 시사점

여기서 도출된 결과를 통해 향후 중소기업 기술혁신을 위한 기획지원이나 정보 제공이 목표로 해야 할 고객은 어떤 특징을 가지고 있는지 찾고, 도출된 영향 요인을 통해 예측하는 방법을 제시해 중소기업 R&D 지원 정책개발이나 평가에 일조할 수 있다.

예를 들어 R&D 기획지원이나 정보지원 경험군에 대한 분석 결과에 따르면, 지원을 경험한 기업들은 이미 정부자금을 지원받은 경험이 있는 기업들로 이미 어떻게 하면 정부지원을 잘 수주하는지 알고 있는 기업이며, 이미 연구개발 실적도 어느 정도 있는 기업들이다. 따라서 향후에는 지원 사업의 홍보를 강화할 필요가 있으며, 지원 경험이 있는 기업이 유리한 현재의 지원 대상 선정 기준을 변경할 필요가 있다(연구 실적, 기술 매출액 중심).

본 연구가 제시한 영향 요인들을 활용할 수 도 있다. 이미 상기 지원사업의 지원에서는 기업의 일반적인 특징(지역, 규모, 기술분야, 대표자 성별 등)과 기술적 성과(보유기술 연구개발 실적 등)을 다수 기재하고 있다, 그러나 장작 본 연구에서는 중요한 영향으로 도출된 총 개발 소요기간이나 기획 소요기간에 대한 정보는 누락되어 있다. 따라서 본 연구의 결과를 활용한다면, 기존의 기재 정보는 줄이고, 오히려 필요한 정보를 추가할 수 있을 것이다.

또한 본 연구의 결과를 좀 더 강화한다면(객관화, 일반화), 지원 대상 기업의 선정에도 활용할 수 있다. 기존의 정성적인 분석에 본 연구가 제시한 판별함수를 점수화해서(중심점의 거리를 정량화)평가 보조 자료로 제공함으로써 평가를 보다 객관적으로 진행할 수 있게 될 것이며, 지원 성공률도 제고할 수 있을 것이다.

또한 본 연구결과는 기존의 R&D 정보 제공 정책에 대한 새로운 의미 있는 시사점을 제공하고 있는데, 개발 기간이나 기획기간에 기업들은 제공받고 싶은 정보도 다를 수 있다는 점이다. 따라서 R&D 제공지원 사업을 기획 및 운영할 때 이런 점들이 고려되어야 한다는 시사점을 제공하는 것이다.

ISBN 978-89-294-***-*-93500

참 고 문 헌

1. (2014), “산업기술 유출방지를 위한 정부 지원제도 연구 : 중소기업 기술보호 지원제도를 중심으로,” 석사학위논문, 서울과학종합대학원.
2. 국가과학기술심의회 (2014), “제3차 중소기업 기술혁신 촉진계획(안),” 2014. 7.
3. 김수진 & 김보영 (2013). 로지스틱 회귀분석과 의사결정나무 분석을 이용한 일 대도시 주민의 우울 예측요인 비교 연구. 한국콘텐츠학회논문지, 13(12), 829-839.
4. 김선영, 배국진, 박상문 & 최윤정 (2014), “제품개발공정의 기술능력이 R&D 매출 성과에 미치는 영향 : 성장단계별 정부지원의 조절효과를 중심으로”, 기술혁신연구 제22권 제4호 pp.235-259.
5. 김왕동 & 김인수 (2002), “응용 및 사례: 기술능력의 축적과정 및 영향요인에 대한 연구: 중소 반도체 장비 제조업체를 중심으로,” 지식경영연구, 3(2), pp.49-70.
6. 류숙원 & 김상윤 (2010), “정책도구의 선택이 중소기업혁신에 미치는 영향에 관한 연구,” 「한국정책과학학회보」, 14(2), pp.65-90.
7. 박문수 & 이호형 (2012), “혁신형 중소기업을 위한 기술지원정책 연구,” 「통상정보연구」, 14(1), pp.197-218.
8. 박명화, 최소라, 신아미 & 구철회 (2013), “의사결정나무 분석법을 활용한 우울 노인의 특성 분석,” Korean Acad Nurs, 43(1), 1-10.
9. 박상문, 변도영 & 손석호 (2007) “국가수준 기술로드맵의 활용도 및 개선사항 영향요인,” 기술혁신학회지, 10(1), 143-163.
10. 박찬수 & 손수정 (2012), 중소기업 기술혁신 역량 평가 및 글로벌 정책동향분석(III), STEPI.
11. 서주환, 성태응, 김유일 & 전승표 (2015), “잠재성장모형을 활용한 중소기업 R&D 기획지원 효과 분석,” 한국기술혁신학회, 2015 춘계학술대회, 2015. 7.
12. 성웅현 (2006), “기술력평가 자료를 이용한 중소벤처기업 피산예측 판별모형에 관한 연구,” 기술혁신학회지, 9(2), 304-324.
13. 성웅현 (2007), “기술금융을 위한 부실 가능성 예측 최적 판별모형에 대한 연구,” 기술혁신학회지, 10(4), 183-205.

14. & 조경선 (2009), “지식재산 투자와 관리가 기업의 무형자산가치에 미치는 영향에 대한 연구,” 기술혁신학회지, 12(2), 291-311.
15. 송종국 & 김혁준 (2009), “R&D 투자 촉진을 위한 재정지원정책의 효과분석,” 기술혁신연구, 17(1), pp.1-48.
16. 신진교 & 최영애 (2008), “중소기업의 R&D 와 혁신: 정부정책지원의 조절효과,” 국제· 경영연구, 15(1), pp.119-132.
17. 안치수 & 이영덕 (2011), “우리나라 개방형 혁신활동의 영향요인에 관한 실증분석 연구,” 기술혁신학회지, 14(3), 431-465.
18. 오해영 외 (2007), 「2007년 기술기획・평가 이슈 도출에 관한 연구」, 서울: 한국과학기술기획평가원.
19. 유태욱, 양동우 (2009), “기술혁신 활동, 기술적 성과, 경제적 성과 간의 관계에 관한 실증연구,” 기업가정신과 벤처연구, 12권 4호.
20. 윤요한, 김윤배, 강지석 & 정가섭 (2015), “출연 (연) 의 기술이전・사업화 추적조사를 통한 영향요인 연구,” 대한산업공학회지, 41(1), pp.105-114.
21. 이선영 & 서상혁 (2011), “정부지원 중소기업 기술협력사업의 성과판별 요인에 관한 연구,” 기술혁신학회지, 14(3), 664-688.
22. 이병헌, 강원진 & 박상문 (2008), “학술연구: 혁신형 중소기업과 일반 중소기업간 기술혁신 및 성과 차이와 정책적 시사점,” 기업가정신과 벤처연구 (구 벤처경영연구), 11(1), pp.79-100.
23. 이병헌, 이수욱 & 위세안 (2013), “정부의 기술개발 지원이 중소기업의 기술 혁신 성과에 미치는 영향,” 벤처창업연구, 9권 5호, pp.157-171.
24. 이종민, 노민선 & 정선양 (2013), “중소기업의 기술기획 역량이 기술사업화 성공에 미치는 영향에 관한 연구,” 기술혁신연구 21(1), pp.253-278.
25. 이종옥, 이규현, 정선양, 조성복 & 윤진호 (2005), 「R&D관리」, 서울: 경문사.
26. 이학식, 임지훈, SPSS 22 매뉴얼, 집현재, 2015
27. 장선미, 김한준 & 김갑수 (2007), “국내 제조업 기술혁신기업의 실태분석,” 한국기술혁신학회, 2007 년 추계학술대회, pp.31-48.
28. 정근하 (2010), 「국가 과학기술적 목표의 효율적 달성을 위한 과학기술기획에 관한 연구」, 서울: 한국과학기술기획평가원.
29. 정선양 (2011), 「전략적 기술경영」, 제3판, 서울: 박영사.

30. , 조성복 & 석재진 (2009), “정부출연연구기관의 창의적 인적자원 양성전략: 전주기적 인력관리의 관점에서,” 기술혁신연구, 17(2), pp.187-206.
31. 전승표 (2015), “중소기업 R&D 정보지원과 성과의 관계에 대한 연구: ICT 기업을 중심으로”, 2015 한국기술혁신학회 추계학술대회, 2015. 11.
32. 조광현 & 박희창. (2011), “매개 변수를 이용한 의사결정나무 생성에 관한 연구,” 한국데이터정보과학회지, 22(4), 671-678.
33. 중소기업기술정보진흥원(2015), 홈페이지 - 정보화 지원 사업 소개, http://www.tipa.or.kr/0504.
34. 중소기업연구원 (2006),「중소기업 R&D 지원의 문제점 및 개선방안」, 서울.
35. 중소기업중앙회 (2014), 2014년도 중소기업기술통계 조사표, 12p.
36. 중소기업청 (2014), 중소기업관련통계, 2014. 06., 2p.
37. 중소기업청, 중소기업중앙회 (2013), 2013년 중소기업 기술통계조사 보고서. 2013.12
38. 중소기업청, 중소기업중앙회 (2014), 2014년 중소기업 기술통계조사 보고서, 2014.12.
39. 최영훈 외 (2005),「국가기술혁신체제의 분석 및 평가」, 서울.
40. 한국과학기술정보연구원 (2015), 홈페이지 - 정보지원 사업 소개 http://www.kisti.re.kr/board/bbs_read.jsp?tname=MINBOARD001&bbsid=B301&bbs_seq=629&jkey=&jword=&pg=1&htxt_code=1246324430078131453764717825548 63&bbsYear=&bbsMonth=
41. 한국산업기술진흥협회 (2009),「국내기업의 기술경영 실태연구」, 서울.
42. 허명회, 이용구 (2008), 데이터 마이닝 모델링과 사례, 한나래출판사, 서울
43. 홍지승 & 홍석일 (2011), 중소기업의 기술혁신성과 영향요인 분석 및 정책과제, 산업연구원.
44. Akcomak, İ. S., & Ter Weel, B.(2007), How do social capital and government support affect innovation and growth? Evidence from the EU regional support programmes.
45. Audretsch, D. B., Link, A. N., & Scott, J. T.(2002), "Public/private technology partnerships: evaluating SBIR-supported research," Research policy, 31(1), pp.145-158.
46. Bae, W. S., Cho, D. H., Seok, K. H., Kim, B. S., Choi, K. L., Lee, J. E., et al. (2004). Data mining using SAS enterprise miner. Seoul: Kyowoosa.
47. Bemelmans-Videc, M. L., Rist, R. C., & Vedung, E. O. (Eds.).(2011), Carrots, sticks,

and sermons: Policy instruments and their evaluation (Vol. 1). Transaction Publishers.

48. Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. CRC press.
49. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
50. Busom, I. (2000), "An empirical evaluation of the effects of R&D subsidies," Economics of innovation and new technology, 9(2), pp.111-148.
51. Choi, J. H., Kang, H. C., Kim, E. S., Lee, S. K., Han, S. T., & Kim, M. K. (2002). Prediction and excess of data mining using decision tree analysis. Seoul: SPSS Academy.
52. Cooper, R. G. (2001), Winning at New Products: Accelerating the Process from Idea to Launch. New York: Basic books.
53. Dvir, D. and T. Raz. and A. J. Shenhar, (2003), “An Empirical Analysis of the Relationship between Project Planning and Project Success”, Project Management, Vol. 21, pp. 89-95.
54. Dvir, D. and T. Lechler, (2004), “Plans are nothing, changing plans is everything: The Impact of Changes on Project Success”, Research Policy, Vol. 33, pp. 1-15.
55. Guellec, D., & Van Pottelsberghe de la Potterie, B.(2004), "From R&D to productivity growth: Do the institutional settings and the source of funds of R&D matter?," Oxford Bulletin of Economics and Statistics, 66(3), pp.353-378.
56. Hadjimanolis, A.(2000), "An investigation of innovation antecedents in small firms in the context of a small developing country," R&D Management, 30(3), pp.235-246.
57. Hastie, T., Tibshrani, R., & Friedman, J. (2001). The elements of statistical learning theory. New York: springer-Verlag.
58. Jun, S. P., Seo, J. H., & Son, J. K. (2013), "A study of the SME Technology Roadmapping Program to strengthen the R&D planning capability of Korean SMEs," Technological Forecasting and Social Change, 80(5), pp.1002-1014.
59. Lichtenberg, F. R. (1984), "The relationship between federal contract R&D and company R&D," The American Economic Review, pp.73-78.
60. Lin, J., & Demner-Fushman, D. (2006, August), The role of knowledge in

conceptual retrieval: a study in the domain of clinical medicine. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 99-106). ACM.

61. Narula, R. (2004), “R&D Collaboration by SMEs: New Opportunities an Limitations in the Face of Globalisation”, Technovation, Vol. 24, pp. 153-161.
62. Robson, M. T. (1993), "Federal funding and the level of private expenditure on basic research," Southern Economic Journal, pp.63-71.
63. Romijn, H., & Albaladejo, M. (2002), "Determinants of innovation capability in small electronics and software firms in southeast England," Research policy, 31(7), pp.1053-1067.
64. Shan, W., & Song, J. (1997), "Foreign direct investment and the sourcing of technological advantage: evidence from the biotechnology industry," Journal of International Business Studies, pp.267-284.
65. Shmueli, G., Patel, N. R., & Bruce, P. C. (2009). Data mining for business intelligence: concepts, techniques, and applications in Microsoft Office Excel with XLMiner. John Wiley & Sons.
66. Stockstrom, C. and C. Herstatt, (2008), “Planning and Uncertainty in New Product Development”, R&D Management, Vol. 38, pp. 480-490.
67. Svensson, R.(2007), "Commercialization of patents and external financing during the R&D phase," Research Policy, 36(7), pp.1052-1069.
68. Verworn, B., C. Herstatt, and A. Nagahira, (2008), “The Fuzzy Front End of Japanese New Product Development Projects: Impact on Success and Differences between Incremental and Radical Projects”, R&D Management, Vol. 38, pp. 1-19.
69. Vrakking, W. J.(1990), "The innovative organization," Long Range Planning, 23(2), pp.94-102.