기계학습 기반 바이오의료분야 과학기술정보데이터 분석·활용 모형 개발

김한해 외

KISTEP 한국과학기술기획평가원
Korea Institute of S&T Evaluation and Planning

제 출 문

한국과학기술기획평가원 원장 귀하

본 보고서를 “기계학습 기반 바이오의료분야 과학기술정보데이터 분석활용 모형개발”의 최종보고서로 제출합니다.

2018. 1.

연구기관명 : 한국과학기술기획평가원
연구책임자 : 김 한 해 부연구위원
연 구 원 : 김 은 정 연구위원
연 구 원 : 홍 미 영 연구위원
연 구 원 : 김 주 원 부연구위원
연 구 원 : 유 거 송 부연구위원
연 구 원 : 정 지 연 연구원
연 구 원 : 황 은 혜 연구원
연 구 원 : 김 진 희 연구원
연 구 원 : 고 기 오 연구원

요 약 문

1. 제목 : 기계학습 기반 과학기술지식정보 분석·활용 모형개발

2. 연구목적 및 필요성

가. 연구목적

- 기계학습 기반 바이오의료분야 과학기술정보 분석·활용 모형 구축
  - 딥러닝 등 고도화된 기계학습 방법을 이용하여 바이오의료분야 1) 과학기술지식정보 자연어처리, 2) 연구과제 간 유사·중복(관계성) 분석, 3)의약과제 분류 모형 개발

- 2016년도 신약개발 정부 R&D 투자포트폴리오 분석
  - 국가과학기술지식정보를 바탕으로 신약개발단계, 의약품종류, 대상질환별 신약개발 정부 R&D 투자포트폴리오를 분석하고 의약과제분류모형 검증 데이터로 활용

나. 연구필요성

- 과학기술정보통신부(舊 미래부)는 최근 국가과학기술지식정보* 공개 범위를 28%~70%로 대폭 확대하며 과학기술지식정보 분석 및 활용의 장을 마련

  * 과학기술지식정보(NTIS)는 세계 최초의 국가R&D정보 지식 포털로 2008~2016년까지 정부 예산으로 지원된 모든 과학기술과제정보를 제공(약 500만건)

  - 개인정보, 보안과제 등 민감 정보를 제외하고 누구나 제공된 정보의 직접적 가공·분석 가능

- 2016년 구글 딥마인드 챌린지 매치를 통해 빅데이터와 인공지능 접목 결과물의 우수성이 전 세계에 알려지며 인공지능에 대한 사회적 관심도 증가

- 2010년대에 이르러 컴퓨터 하드웨어의 비약적 발전, 인터넷 보급·확산에 따른 방대한 데이터 인프라 축적, 딥러닝과 같은 혁신기술의 등장으로 인공지능 연구가 급속히 발전

■ 동 연구는 과학기술지식정보라는 빅데이터와 인공지능의 기반이 되는 기계학습을 접목하여 바이오의료분야 과학기술정보데이터 분석·활용 모형 개발을 목표로 함

- 동 연구는 데이터과학에 근거, 바이오의료분야 과학기술지식정보의 고도화된 분석·활용방안을 제시하여 업무효율성을 높일 수 있을 것으로 기대

3. 연구내용

가. 기계학습 기반 바이오의료분야 과학기술정보 분석·활용 모형 구축

■ (1단계) 기계학습 기반 과학기술지식정보 자연어처리 모형 개발

- 과학기술지식정보를 훈련데이터로 사용하여 과학기술 특이적 단어와 문맥을 이해할 수 있는 자연어 처리 모형 개발
- 동 자연어 처리 모형은 연구과제 간 유사·중복 분석 모형 및 의약과제 분류 모형의 기반 기술로 활용
- 동 자연어 처리 시스템은 보건의료분야 외 타 기술분야에도 적용이 가능하도록 최대한 범용적 설계를 지향

■ (2단계) 바이오의료분야 연구과제 간 유사·중복(관계성) 분석 모형 개발

- 주요 용어의 노출빈도 및 중복여부와 같은 단순 유사·중복 관계성 개념을 탈피하여 핵심키워드간 주변 문맥 흐름을 중심으로 연구과제 유사·중복 분석모형 개발
  - 연구과제명, 연구목표 및 연구내용(초록)을 바탕으로 연구과제간 유사도를 정량화(수치제공)하여 사용자가 유사정도를 판단할 수 있도록 의사결정 지원
- 정량화된 유사도를 기반으로 연구과제간 관계성을 네트워크 방식으로 시각화

– 특정연구과제 혹은 문헌정보(연구내용, 신문기사 등)를 입력받아 이와 관계가 있는 연구과제를 추출하고 관계성 기반 네트워크 분석 수행

■ (3단계) 의약과제 과학기술지식정보 분류 모형 개발

● KISTEP 생명기초사업센터가 구축한 정부신약개발과제DB를 훈련데이터로 이용하여 신약개발단계·의약품종류·질환별 신약개발과제(의약과제) 분류 모형 개발

● 유사·중복 분석 모형과 같이 핵심키워드간 과학기술적 문맥을 고려하여 분류 모형을 개발

– 분류모형은 인공신경망 네트워크(artificial neural network) 구축을 통해 딥러닝 방식 적용

● 구축과정에서 2016년도 국가과학기술지식정보 조사·분석데이터를 활용하여 신약개발 정부 R&D 투자포트폴리오를 분석하고 분석결과를 모형 검증에 사용

– 검증단계의 신뢰성 확보를 위해 분류모형 개발에 사용되지 않은 독립된 데이터 이용

나. 2016년도 신약개발 정부 R&D 투자포트폴리오 분석

■ 신약개발 분야 정부 R&D 투자현황을 파악하기 위해 국가과학기술정보서비스(NTIS)에서 제공하는 국가연구개발사업 조사분석데이터('16)를 활용하여 투자 추이(2014년~2016년), 투자 포트폴리오(2016년)를 분석

● 생명보건의료분야 예산심의대상 사업('16년 기준 93개)을 대상으로, 전문가를 통해 신약개발을 목적으로 하는 과제 선별, 신약개발단계, 의약품 종류, 타겟 질환 등의 분류기준*에 따라 과제 분류

* 「신약개발 R&D 투자 효율화 방안(2012)」에서 제안된 분류기준으로, 생명의료전문위 등 관련 전문가 의견을 반영하여 수립

※ 개별 연구자가 수행하는 기초 및 기전연구는 최종 목표가 설정되기 전에 수행하는 과제로 간주하여 제외하여 분석

- 신약개발 단계: 타겟발굴 및 검증, 후보물질도출 및 최적화, 비임상, 임상1~3상, 인프라 및 기타
- 의약품 종류: 합성신약, 바이오신약, 천연물신약, 개량신약(합성), 바이오베터, 바이오시밀러
- 질환: 혈관질환, 천식, 종양, 감염증, 정신질환, 퇴행성뇌질환, 골다공증, 당뇨, 비만, 관절염

- 신약개발분야 정부 R&D 투자는 2016년 신약개발분야 정부 R&D 투자 규모는 3,059억원으로, 연구비 기준 3년간('14~'16) 연평균 약 4.4% 증가한 반면 신약개발과제 수는 '14년 1,047건에서 '16년 967건으로 연평균 약 -3.9% 감소

<표 2-2> 신약개발 분야 정부 R&D 투자 규모

구분	2014년	2015년	2016년	연평균 증가율
과제 수(개)	1,047	1,201	967	-3.9%
정부연구비 (백만원)	280,473	335,738	305,890	4.4%

- 신약개발단계별 정부 R&D 투자는 '16년 기준 신약플랫폼기술(871억원, 26.4%), 후보물질 도출 및 최적화(674억원, 20.4%), 타겟발굴 및 검증(373억원, 12.2%) 순으로 이루어지고 있음
  - **인프라 중 인력양성, 제도·정책, 인·허가 부분의 투자 비중이 낮아** 신약개발 및 제약산업의 성장을 위해서는 신약개발을 촉진시킬 수 있는 제도·정책 및 인프라적인 부분에 대한 정부 차원의 지원 필요
    - 특히 신약개발 분야는 신약 개발 각 단계에 인력이 필요할 뿐만 아니라 임상 후의 각 단계의 리스크를 관리할 수 있는 전문 인력의 양성이 중요함
    - 제도·정책 및 인·허가 단계에 대한 지원이 각 연평균 71.8%, 26.7%로 증가한 부분은 적절하나, 인력 양성을 위한 지원을 확대하여야 할 것으로 판단됨

- 의약품 종류별로는 바이오신약에 대한 투자가 연평균 약 25.3%로 가장 많이 증가한 반면 바이오베터에 대한 투자는 연평균 약 -25.9%로 감소
  - **바이오신약 중 백신에 대한 투자가 '14년 79억원에서 '16년 259억원으로 가장 많이 확대됨(연평균 약 80.7%)**
    - '16년 기준 투자 규모의 절반 이상이 신약개발(54.8%)에 집중되어 투자되고 있으며, 공통기반기술(852억원, 27.9%), 바이오신약(788억원, 25.8%), 합성신약(736억원, 24.1%) 순으로 투자

- 합성 신약의 성공 빈도가 낮아지고, 합성의약품에 비해 부작용이 적은 바이오의약품의 시장이 활발해지고 있음
  - 바이오의약품에 대한 투자가 크게 확대되고 있어 시장의 흐름이 적절히 반영되는 것으로 나타남

■ 질환별 투자현황을 살펴보면, 종양질환(22.0%)의 투자비중이 가장 높으나 투자 규모가 연평균 -1.8%로 감소하는 반면 천식, 감염병 및 퇴행성뇌질환 등에 대한 투자는 증가(75.4%, 33.0%, 23.5%)

- 환경적 변화 및 치매, 신종 감염병 등에 대한 대응을 위한 공공적 보건의료 R&D 지원을 강화한 것으로 보임

■ 과기부 및 복지부는 후보물질 도출 및 최적화 단계에 가장 큰 투자 비중을 보였으며, 범부처 사업은 타겟발굴부터 임상 2상 단계까지, 산업부는 후보물질도출 및 최적화부터 신약플랫폼기술 단계까지 비교적 고르게 지원하였음

- 부처간 투자 효율성을 제고하기 위하여 각 부처의 역할 분담 필요

■ 신약개발분야 정부 R&D 투자포트폴리오를 분석한 결과 대기업 및 중견기업에의 투자는 연평균 -39.8%, -13.4%로 크게 축소하고 있었으며, 중소기업에의 투자는 11.1%로 확대하고 있었음

- 민간기업 중심의 신약개발 투자를 유도하기 위해서는 중소기업에의 투자 확대가 필요하며 대기업 및 중견기업의 경우에는 비교적 연구개발 역량을 갖추었다고 판단되므로 이를 감안한 전략적 지원이 요구됨

4. 결론 및 시사점

- 정부 R&D 예산배분조정 시 보건의료분야 신규과제 유사중복 검토 업무 효율성 제고 기대
  - 키워드 빈도수에 의한 유사중복 탐색보다는 핵심 키워드를 바탕으로 연구과제의 내용간 문맥적 유사성을 검토하는 것이 필요하므로 동 과제 모형을 예산배분조정업무 과정 중 유용하게 활용 할 수 있을 것으로 예상
  - 향후 정부 연구개발 사업간 과제 수행내용을 바탕으로 연구개발 사업간 관계성을 분석하는 기능 등이 추가될 경우 모형의 활용성이 보다 제고될 수 있을 것

- 신약개발 연구과제DB 지속 구축 및 관리에 소모되는 비용 및 노력 절감
  - 대상질환의 분류결과는 최우선예측(80.9%), 우선예측(92.5%)로서 실무적 활용이 가능하다고 판단됨
    - 의약품종류 및 신약개발단계 분류모형의 경우는 최우선예측은 50% 수준이나 우선예측이 70% 이상의 예측율을 보이는 점을 감안하여 업무 개선효과 방안을 마련할 필요

- 타 기술분야 응용연구 등 동 연구 방법론을 바탕으로 후속 연구성과 창출
  - 현재 우리나라 과학기술은 10대 기술분야를 중심으로 투자방향 및 예산배분조정이 진행 중이며 해당 기술분야는 모두 과학기술지식정보를 근간으로 하고 있기 때문에 동 연구방법론 적용이 가능한 상황

- 동 과제모형은 전문가 의사결정지원모형으로 판단되며, 따라서 일정 수준의 한계점이 있음을 숙지할 필요
  - 모든 잠긴 문을 열 수 있는 만능열쇠가 존재하지 않듯이, 동 과제 모형도 최신 기계학습 방법을 적용한 모형일지라도 한계점이 존재
  - 이는 실질적으로 최신 기계학습 방법론을 적용한 모형이 인간과 같이 문맥이나 핵심 키워드, 행간에서 정보를 이해한다기 보다 인간이 입력한 질문에 대한 답을 과거보다 정확성 높게 찾아준다는 시각으로 보는 것이 타당

- 훈련 데이터에 국한되어 있는 해결과제*의 경우 분석활용 모형성능 향상관점에서 긴 호흡을 갖고 접근 및 해결 필요

  * 훈련데이터 부족 혹은 일부 내용에 데이터가 편향된 경우

  - 바이오의료분야 연구과제간 유사과제(관계성) 분석 모형은 과학기술지식정보의 공개 범주에 따라 연구책임자 성명 등을 관계성 분석에 활용할 수 없었음
  - 의약과제 분류모형 훈련정보의 경우 소분류를 기준으로 일부 코드에 데이터가 집중되거나 상대적으로 과제 정보가 매우 부족한 경우가 있음
  - 현재로서는 학습파라미터를 조정하는 기술적 방법을 적용하여 분석·활용 모형의 성능을 제고하는 것이 최선의 방법으로 판단됨

- 궁극적 동 과제 모형은 일회성 사용에 그치지 않고 사용자의 의지에 따라 지속적 업데이트 및 성능향상이 가능한 점을 감안, 추가 모형 발굴, 타 빅데이터 융합, UI/UX 강화, 플랫폼화 등을 통해 활용성을 극대화 할 필요

  - 연구과제간 관계성 분석 이외에 정부 연구개발사업 수준에서의 관계성 분석 모형이 개발될 경우 예산배분조정 업무에 활용성이 강화될 것으로 전망
  - 과학기술지식정보 외에 공개적으로 접근이 가능한 타 바이오의료 관련 빅데이터를 통합할 경우 모형 활용 방안 및 도출할 수 있는 분석정보도 다변화 될 것으로 기대
  - 범용적 활용 위해 UI의 친밀성을 강화하고 타 기술분야 적용이 용이하도록 플랫폼화 필요

목 차

표 목 차

그 림 목 차

제 1 장
서 론

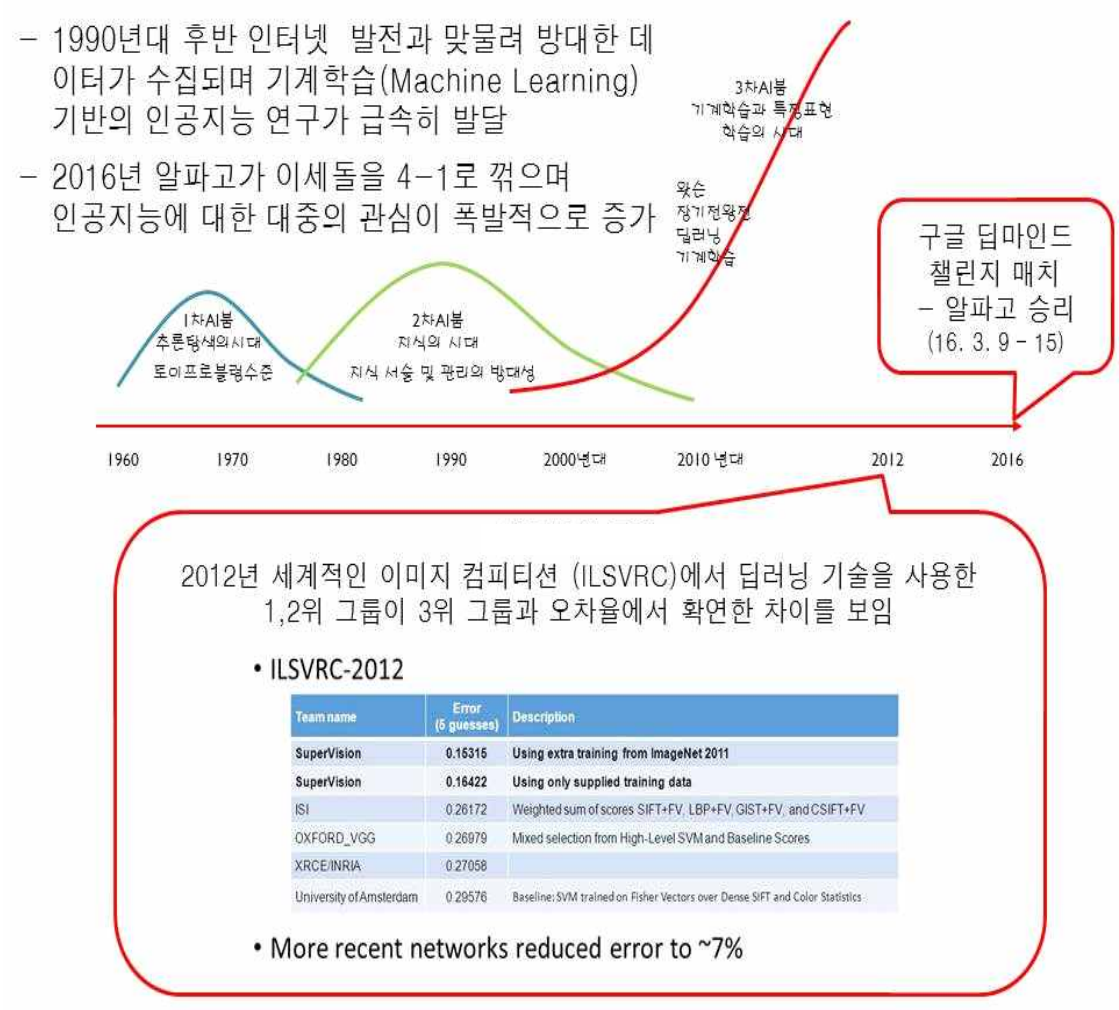
제1장 서론

제1절 연구배경 및 필요성

가. 연구배경 및 필요성

- 과학기술정보통신부(舊 미래부)는 2015년 국가과학기술지식정보서비스(NTIS)를 과학기술계의 구글, 네이버로 만들겠다는 포부를 밝힘
  - 이에 발맞추어 과학기술지식정보서비스 개방을 공표하고 정보 공개 범위를 28%~70%로 대폭 확대하며 과학기술지식정보 분석 및 활용의 장을 마련
    - 개인정보, 보안과제 등 민감 정보를 제외하고 누구나 제공된 정보의 직접적 가공·분석 가능
      ※ NTIS는 세계 최초의 국가R&D정보 지식 포털로 2008~2016년까지 정부 예산으로 지원된 모든 과학기술과제정보를 제공(약 500만건)

- 2016년 구글 딥마인드 챌린지 매치를 통해 빅데이터와 인공지능 접목 결과물의 우수성이 전 세계에 알려지며 인공지능에 대한 사회적 관심도 증가
  - 인공지능은 1950년 최초 등장한 용어로 인간의 두뇌 모방(Artificial Intelligence, AI)에 관한 연구를 의미
  - 1960년, 1990년대 인공지능 붐이 일었으나 컴퓨팅 파워 및 충분한 데이터 확보·활용 측면에서 연구의 한계를 드러냄
  - 2010년대에 이르러 컴퓨터 하드웨어의 비약적 발전, 인터넷 보급·확산에 따른 방대한 데이터 인프라 축적, 딥러닝과 같은 혁신기술의 등장으로 인공지능 연구가 급속히 발전
  - 2016년 구글 딥마인드에서 개발한 알파고와 한국 프로기사 이세돌 9단간 대국에서 알파고가 4:1로 승리하며 인공지능의 인간대체 가능성 논의 활성화



[그림 1-1] 인공지능 기술 발전 동향(마쓰오유타카, 2015 재구성)

- 과학기술지식정보 분석·활용에 관한 선도적 방안 마련 및 업무 효율성 개선을 위해 최근 과학기술 트렌드를 주도*하는 빅데이터와 인공지능(Artificial Intelligence, AI) 연구에 주목

  * 2016년 '4차 산업혁명의 이해(Mastering the Fourth Industrial Revolution)'라는 주제로 개최된 제46회 다보스포럼에서는 ICT, 데이터기반의 기술융합이 미래를 선도할 것임을 공표

■ 동 연구는 과학기술지식정보라는 빅데이터와 인공지능의 기반이 되는 기계학습을 접목하여 바이오의료분야 과학기술정보데이터 분석·활용 모형 개발을 목표로 함

● KISTEP 사업조정본부는 과학기술지식정보를 활용, 정부 연구개발 투자방향 및 기준(안) 마련 및 연구개발사업 예산배분조정 업무를 수행

● KISTEP 생명기초사업센터는 해당 업무 수행하며 바이오의료분야 투자포트폴리오 구축, 신규·기 R&D사업 과제간 유사·중복 검토, 신약개발과제 통계브리프 발간 등을 추진

– 과학기술지식정보 분석·활용 목적으로 전문가풀을 장기 유지·관리하고 균일한 결과를 얻기 위해 노력하는 과정*에서 상당한 업무시간 및 비용 소모

* 전문가 개별 주관적 판단 최소화를 위한 데이터 정규화 등

● 동 연구는 데이터과학에 근거, 바이오의료분야 과학기술지식정보의 고도화된 분석·활용방안을 제시하여 업무효율성을 높일 수 있을 것으로 기대

– 과학기술지식정보 조사분석데이터(과제정보)는 과제명, 연구목표 및 내용을 비롯하여 인력, 성과, 수행주체, 지역, 연구비 규모 등 수십 가지 특이적 지표로 구성되어 기계학습 적용에 최적의 데이터 포맷을 지님

– 연차별 5~6만 여건 내외의 연구과제정보가 10년 이상 축적되어 있기 때문에 그간 정부 R&D 투자현황을 반영하는 빅데이터로서 높은 가치와 활용성 내재

※ 동 연구과제는 최근 10년간(2007~2016) 바이오의료분야 과학기술지식정보(약 14만건) 사용

제 2 절 연구 목표 및 방법

가. 연구 목표

- 기계학습 기반 바이오의료분야 과학기술정보 분석·활용 모형 구축
  - 딥러닝 등 고도화된 기계학습 방법을 이용하여 바이오의료분야 과학기술지식정보 자연어처리, 유사·중복연구과제 분석(관계성분석) 및 의약과제 분류 모형 개발
    ※ 데이터 분석 및 활용의 중요성, 인공지능에 대한 이해도 및 가능성에 대한 직접 체감도 향상

- 2016년도 신약개발 정부 R&D 투자포트폴리오 분석
  - 국가과학기술지식정보를 바탕으로 신약개발연구 전문가를 활용하여 신약개발단계, 의약품 종류, 대상질환별 정부 신약개발 연구과제 심층 분석
    ※ 의약과제 과학기술지식정보 분류모형 예측력 검증을 위한 교차분석 데이터로 활용

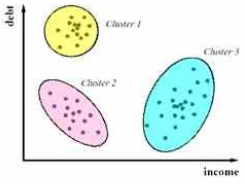
나. 연구 내용

- 연구 목표 달성을 위해 그간 축적된 정부 바이오의료분야 과학기술정보에 기계학습 방법론을 적용
  - 인공지능 기술을 실현하는 기계학습은 지도학습(supervised learning)과 비지도학습(unsupervised learning)으로 구분 [그림 1-2 참조]
    ※ 최근 자율주행, 알파고 등의 등장으로 이에 기반이 되는 강화학습 또한 주목
  - 지도학습은 준비된 훈련데이터(training data)를 기계에 학습시켜 분류모형(분류기, classifier)을 구축하고 훈련데이터에 근거하여 입력데이터를 분류
  - 비지도학습은 라벨링이 되어있지 않은 데이터를 기계에 제공하여 데이터의 내제된 구조를 파악

기계학습 (Machine Learning)

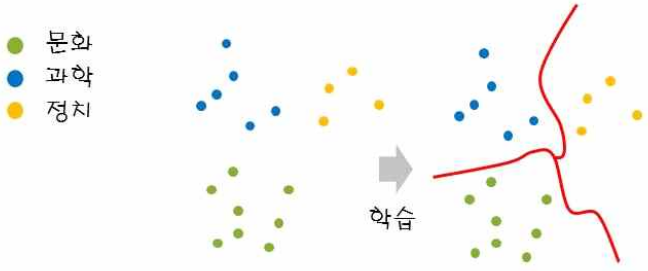
– Unsupervised Learning (비지도학습, 간단예시: Clustering, 군집화)

ㅇ 입력용 데이터만 제공하고 라벨링 없이 데이터에 내재하는 구조 파악



– Supervised Learning (지도학습, 간단예시: Classification, 분류)

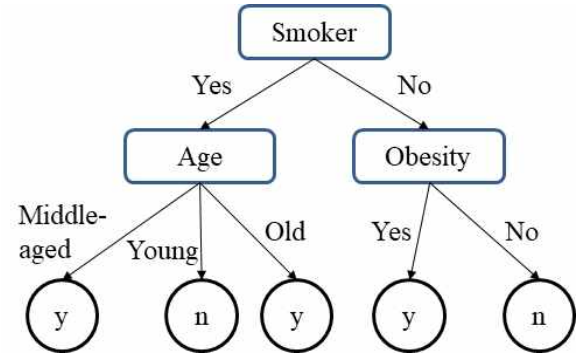
ㅇ 입력과 올바른 출력(분류결과)이 세트가 된 훈련데이터(골드스탠다드)를 미리 준비하여 학습시키고 어떤 입력이 주어졌을 때 올바른 출력(분류)이 가능토록 함



[그림 1-2] 기계학습 방법 예시(마쓰오유타카, 2015 재구성)

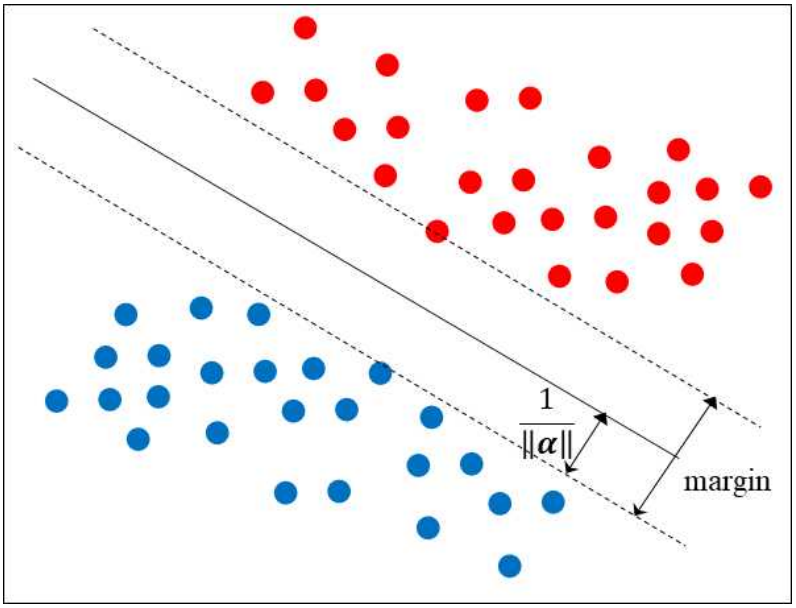
기계학습(지도학습) 방법 예시

- 의사결정트리(Decision Tree): 트리구조의 그래프 모델을 구축, 어떤 입력 값의 해당 속성 만족여부를 '예/아니오'로 분기를 태우는 로직을 통해 정답 도출



[그림 1-3] 의사결정트리 모델 예시

- 서포트백터머신(Support Vector Machine): 데이터를 구분 짓는 구분선과 각 데이터 그룹간의 마진(간격)을 최대로 나누는 분류 방법



[그림 1-4] 서포트백터머신 모델 예시

- 베이지안 이론(Bayesian Theory): 이전 사건 또는 무언가 일어날 것이라는 신념의 정도를 관측[1]한다는 개념으로, 다음의 수식(P(A): 사전확률, P(C|A): 사후확률)에 기반 하여 작동

$$P(C|A) = \frac{P(A|C)P(C)}{P(A|C)P(C) + P(A|C^c)P(C^c)}$$

1) 제이슨벨, 곽승주, 머신러닝 워크북, 길벗 2016

(1단계) 기계학습 기반 과학기술지식정보 자연어처리 모형 개발

- 과학기술지식정보를 훈련데이터로 사용하여 과학기술 특이적 단어와 문맥을 이해할 수 있는 자연어 처리 모형 개발
- 동 자연어 처리 모형은 연구과제 간 유사·중복 분석 모형 및 의약과제 분류 모형의 기반 기술로 활용
- 동 자연어 처리 시스템은 보건의료분야 외 타 기술분야에도 적용이 가능하도록 최대한 범용적 설계를 지향

(2단계) 바이오의료분야 연구과제 간 유사·중복(관계성) 분석 모형 개발

- 주요 용어의 노출빈도 및 중복여부와 같은 단순 유사·중복 관계성 개념을 탈피하여 핵심키워드간 주변 문맥 흐름을 중심으로 연구과제 유사·중복 분석모형 개발
  - 연구과제명, 연구목표 및 연구내용(초록)을 바탕으로 연구과제간 유사도를 정량화(수치제공)하여 사용자가 유사정도를 판단할 수 있도록 의사결정 지원
- 정량화된 유사도를 기반으로 연구과제간 관계성을 네트워크 방식으로 시각화
  - 특정연구과제 혹은 문헌정보(연구내용, 신문기사 등)를 입력받아 이와 관계가 있는 연구과제를 추출하고 관계성 기반 네트워크 분석 수행

(3단계) 의약과제 과학기술지식정보 분류 모형 개발

- KISTEP 생명기초사업센터가 구축한 정부신약개발과제DB를 훈련데이터로 이용하여 신약개발단계·의약품종류·질환별 신약개발과제(의약과제) 분류 모형 개발
- 유사·중복 분석 모형과 같이 핵심키워드간 과학기술적 문맥을 고려하여 분류 모형을 개발
  - 분류모형은 인공신경망 네트워크(artificial neural network) 구축을 통해 딥러닝 방식 적용
- 구축과정에서 2016년도 국가과학기술지식정보 조사·분석데이터를 활용하여 신약개발 정부 R&D 투자포트폴리오를 분석하고 분석결과를 모형 검증에 사용
  - 검증단계의 신뢰성 확보를 위해 분류모형 개발에 사용되지 않은 독립된 데이터 이용

다. 추진방법

- KISTEP 연구진은 바이오의료분야 과학기술정보 분석·활용 모형 개발 개념설계를 비롯하여, 연구추진 방향 설정, 분석·활용 모형 콘텐츠 발굴 등 연구 전반을 총괄

  - 동 연구과제 결과물을 직접적으로 활용하게 될 KISTEP 실무진과 대학, 연구소, 산업체 등에 재직하는 기계학습 및 바이오빅데이터 관련 전문가의 종합자문을 바탕으로 분석·활용 모형 개념설계 추진

    – 외부전문가 자문 및 기술용역을 바탕으로 동 연구 과제의 완성도 제고

  - 기계학습 이해 심화 및 적용방법 발굴, 연구내용·성과 발표, 개선·활용방안 도출을 위해 관련된 국·내외 교육 및 학회 참석, 연구기관방문 등 추진

- 2016년도 신약개발 정부 R&D 투자포트폴리오를 분석하고 신약개발과제DB 정보 업데이트 추진

  - 2016년 과학기술지식정보 조사분석데이터 확보 후 정부 신약개발 대상과제에 한하여 의약분야 전문가 지식* 기반 정부 R&D 투자포트폴리오 분석

    * 포트폴리오 분석결과 및 분류모형 검증에 대한 신뢰성 확보

    – KISTEP 정부신약개발과제DB 분류기준에 부합하도록 신약개발단계, 의약품종류, 질환별 연구과제 분류 및 교차검증 수행

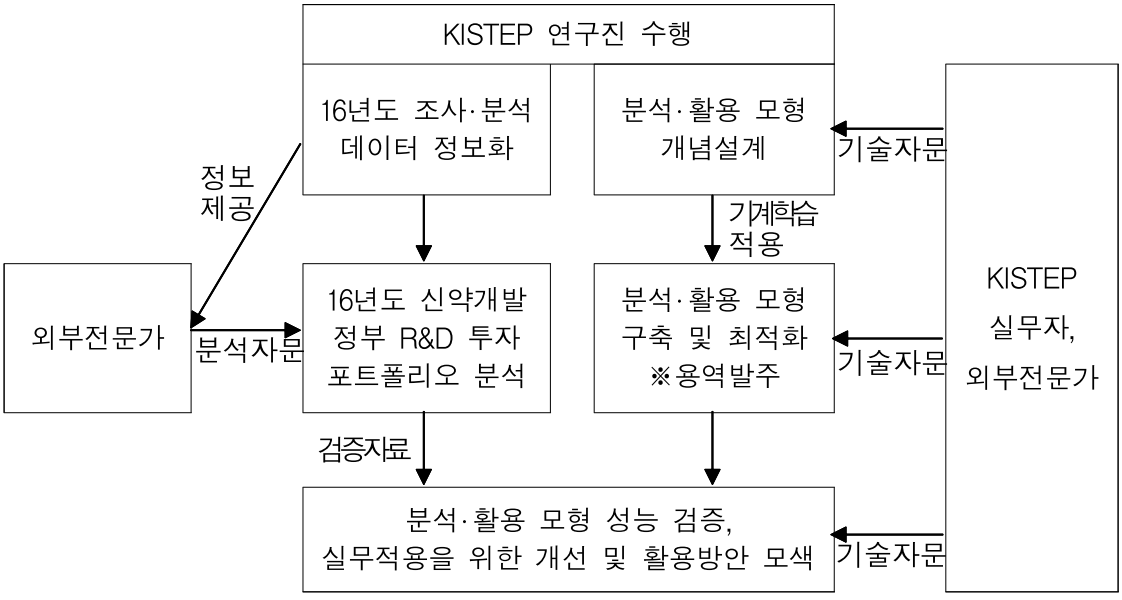
    ※ 관련 상세내용은 '제2장 신약개발 정부 R&D 투자포트폴리오 분석'에 제시

  - 2016년도 신약개발 정부 R&D 투자포트폴리오 분석결과를 기존 신약개발과제 DB('08~'15)에 추가하고 의약과제 과학기술지식정보 분류모형의 검증자료로 사용

    – 모형개발 시 입력되는 훈련데이터는 '08~'15년에 해당하는 신약개발R&D 연구과제로 구성

    – 모형개발에 사용되지 않은 '16년 분석결과(별도데이터)로 모형을 검증함으로써 모형의 분류결과 신뢰성 및 검증과정의 타당성 확보

<표 1-1> 보건의료분야 과학기술정보 분석·활용 모형 개발 추진전략



제2장

신약개발 정부 R&D 투자포트폴리오 분석

제2장 신약개발 정부 R&D 투자포트폴리오 분석

제1절 신약개발 투자포트폴리오 분류기준

- 생명보건의료분야 예산심의대상 사업('16년 기준 93개)을 대상으로, 국가과학기술정보서비스(NTIS)에서 제공하는 국가연구개발사업 조사분석데이터('16)의 과제정보를 활용
  - 신약개발 분야 전문가를 통해 신약개발을 목적으로 하는 과제 선별, 신약개발단계, 의약품 종류, 타겟 질환 등의 분류기준에 따라 과제 분류(<표 2-1> 참고)
    ※ 「신약개발 R&D 투자 효율화 방안(2012)」에서 제안된 분류기준으로, 생명의료전문위 등 관련 전문가 의견을 반영하여 수립
  - 개별 연구자가 수행하는 기초 및 기전연구는 최종 목표가 설정되기 전에 수행하는 과제로 간주하여 제외하여 분석

<표 2-1> 신약개발분야 정부 R&D 투자포트폴리오 분류기준

구분	대분류	중분류	소분류
신약개발 단계	타겟발굴 및 검증	타겟발굴 및 검증	타겟발굴 및 검증
	후보물질도출 및 최적화	후보물질도출 및 최적화	후보물질도출 및 최적화
	비임상	비임상	비임상
	임상	임상1상	임상1상
		임상2상	임상2상
		임상3상	임상3상
	인프라	신약플랫폼기술	타겟발굴 플랫폼
			후보물질 발굴 플랫폼
			비임상 플랫폼
			질환동물 플랫폼
			임상 플랫폼
		인력양성	인력양성
		제도·정책	제도·정책
		인·허가	인·허가
	기타	기타	기타
의약품 종류	신약	합성신약	합성신약
		바이오신약	단백질 치료제
			유전자 치료제
			세포 치료제
			백신
			항체
		천연물신약	천연물신약
	개량신약	개량신약(합성)	개량신약
		바이오베터	단백질 치료제
			유전자 치료제
			세포 치료제
			백신
			항체
	복제약	바이오시밀러	바이오시밀러
	공통기반기술 및 기타	공통기반기술	공통기반기술
		기타	기타
질환	혈관질환, 천식, 종양, 감염증, 정신질환, 퇴행성뇌질환, 골다공증, 당뇨, 비만, 관절염, 기타		

제 2 절 2016년도 신약개발 R&D 투자포트폴리오 분석

가. 신약개발 분야 정부 R&D 투자 현황

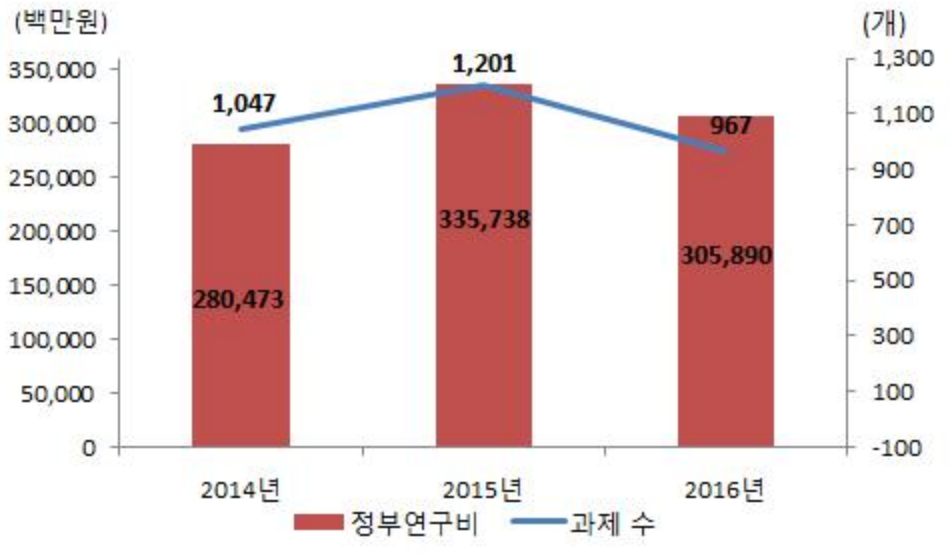
■ 신약개발분야 정부 R&D 투자 현황

● 2016년 신약개발분야 정부 R&D 투자 규모는 3,059억원으로, 연구비 기준 3년간('14~'16) 연평균 약 4.4% 증가

– 반면 신약개발과제 수는 '14년 1,047건에서 '16년 967건으로 연평균 약 -3.9% 감소

<표 2-2> 신약개발 분야 정부 R&D 투자 규모

구분	2014년	2015년	2016년	연평균 증가율
과제 수(개)	1,047	1,201	967	-3.9%
정부연구비 (백만원)	280,473	335,738	305,890	4.4%

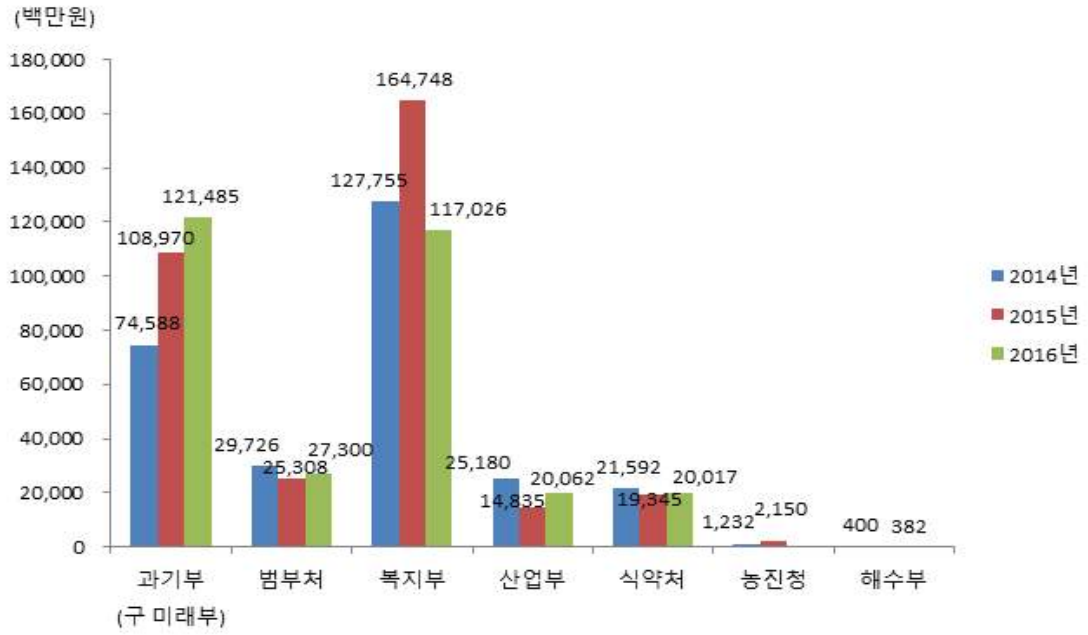


[그림 2-1] 신약개발분야 정부 R&D 투자 현황

부처별 투자 현황

- 모든 부처의 투자 규모가 감소한 반면 과기부의 경우 '14년 746억원에서 '16년 1,215억원으로 연평균 27.6% 증가
- '16년의 경우, 부처별로는 과기부(1,215억원, 39.7%), 복지부(1,170억원, 38.3%), 범부처(273억원, 8.9%) 순으로 신약개발과제 지원

  – 3개 부처가 전체 예산의 86.9%를 차지



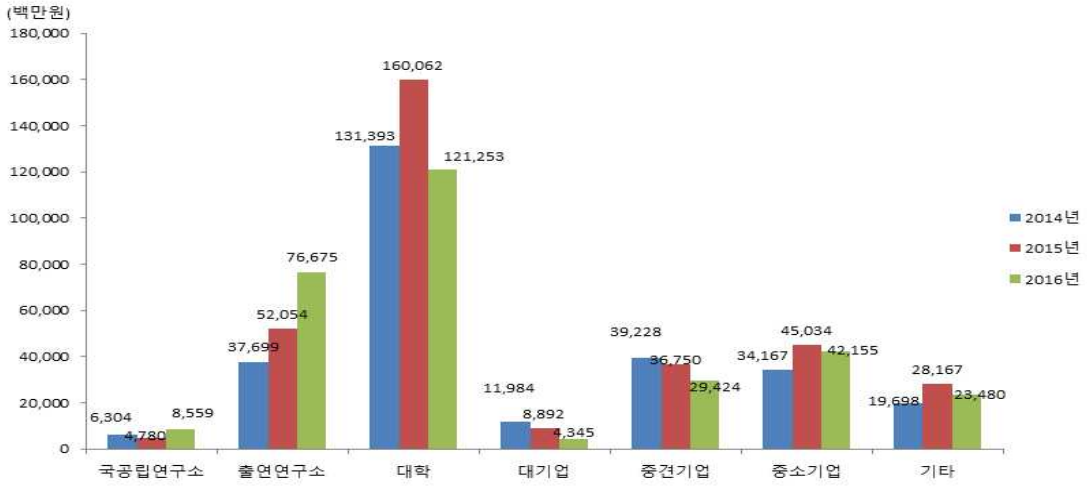
[그림 2-2] 신약개발분야 정부 R&D 부처별 투자 현황

<표 2-3> 신약개발분야 정부 R&D 부처별 투자 현황

구분	2014년		2015년		2016년		연평균 증가율 (%)
	연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)	
과학기술정보통신부	74,588	26.6	108,970	32.5	121,485	39.7	27.6
범부처 사업	29,726	10.6	25,308	7.5	27,300	8.9	−4.2
보건복지부	127,755	45.5	164,748	49.1	117,026	38.3	−4.3
산업통상자원부	25,180	9.0	14,835	4.4	20,062	6.6	−10.7
식품의약품안전처	21,592	7.7	19,345	5.8	20,017	6.5	−3.7
농촌진흥청	1,232	0.4	2,150	0.6	−	−	−100.0
해양수산부	400	0.1	382	0.1	−	−	−100.0
합계	280,473	100.0	335,738	100.0	305,891	100.0	4.4

연구수행주체별 투자 현황

- ’16년 기준 대학에서 1,213억원(39.6%) 규모의 가장 큰 투자 비중을 보임
  - 다음으로 출연연(767억원, 25.1%), 중소기업(422억원, 13.8%), 중견기업(294억원, 9.6%) 순으로 투자
- 대기업 및 중견기업에의 신약개발분야 정부 R&D 투자는 축소한 반면(연평균 -39.8%, −13.4%), 출연연의 경우 연구비 기준 ’14년 377억원에서 ’16년 767억원으로 연평균 약 42.6%로 가장 많이 증가



[그림 2-3] 신약개발분야 정부 R&D 연구수행주체별 투자 현황

<표 2-4> 신약개발분야 정부 R&D 연구수행주체별 투자 현황

구분	2014년		2015년		2016년		연평균 증가율 (%)
	연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)	
국공립연구소	6,304	2.2	4,780	1.4	8,559	2.8	16.5
출연연구소	37,699	13.4	52,054	15.5	76,675	25.1	42.6
대학	131,393	46.8	160,062	47.7	121,253	39.6	-3.9
대기업	11,984	4.3	8,892	2.6	4,345	1.4	-39.8
중견기업	39,228	14.0	36,750	10.9	29,424	9.6	-13.4
중소기업	34,167	12.2	45,034	13.4	42,155	13.8	11.1
기타	19,698	7.0	28,167	8.4	23,480	7.7	9.2
합계	280,473	100.0	335,738	100.0	305,891	100.0	4.4

주요 대상사업별 투자 현황

- 신약개발분야를 지원하는 주요 사업으로는 과기부의 바이오·의료기술개발이 681억원(22.3%)을 지원하고 있었으며, 동 사업 내 신약개발분야에 대한 투자 비중은 35.0% 수준
- 다음으로 범부처전주기신약개발(273억원, 8.9%), 복지부의 첨단의료기술개발(242억원, 7.9%), 질환극복기술개발(237억원, 7.8%) 순으로 신약개발분야 주요 사업으로 나타남

<표 2-5> 신약개발분야 정부 R&D 주요 사업(2016)

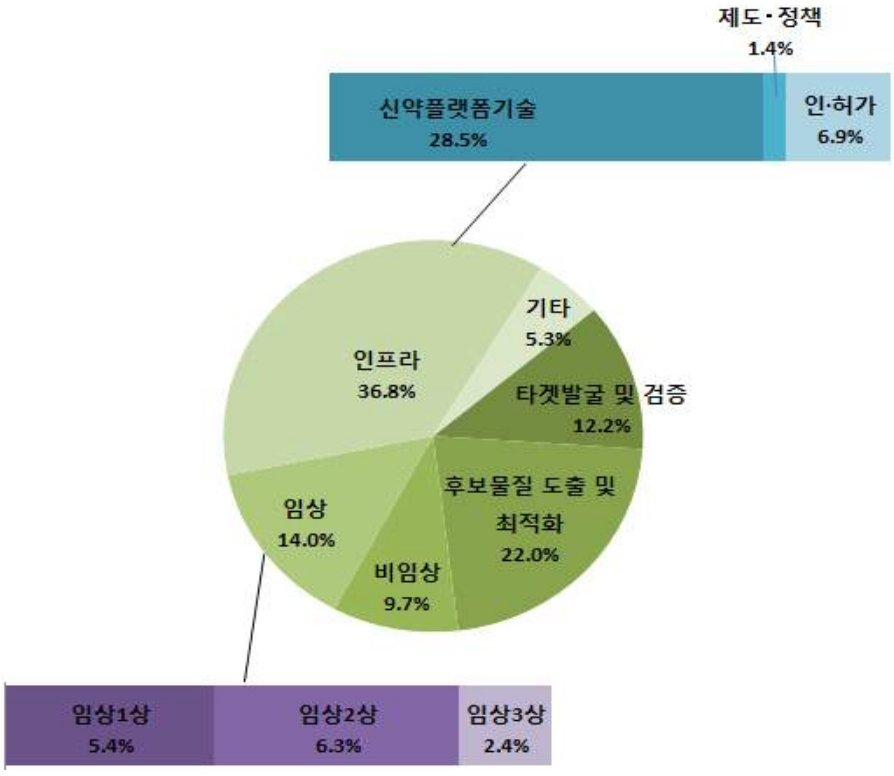
사업명	총 사업비 (백만원) (A)	의약분야 투자액 (백만원)		비중(%) (B/A)
		(B)	비중(%)	
바이오.의료기술개발	194,691	68,073	22.3	35.0
범부처전주기신약개발	30,000	27,300	8.9	91.0
첨단의료기술개발	73,144	24,185	7.9	33.1
질환극복기술개발	80,976	23,743	7.8	29.3
바이오산업핵심기술개발	73,124	20,062	6.6	27.4
안전성평가연구소연구운영비지원	26,777	16,838	5.5	62.9
감염병위기대응기술개발	26,572	15,747	5.1	59.3
의약품등안전관리	22,387	14,367	4.7	64.2
한국화학연구원연구운영비지원	78,401	13,635	4.5	17.4
한국생명공학연구원연구운영비지원	78,486	13,524	4.4	17.2
임상연구인프라조성	52,000	11,588	3.8	22.3
연구중심병원육성	26,250	10,178	3.3	38.8
첨단바이오의약품글로벌진출사업	12,500	9,756	3.2	78.0
한국한의학연구원연구운영비지원	48,806	8,515	2.8	17.4
암연구소및국가암관리사업본부연구운영비지원	58,316	6,597	2.2	11.3
선도형특성화연구사업	11,500	6,418	2.1	55.8
안전성평가기술개발연구	14,553	3,360	1.1	23.1
감염병관리기술개발연구	14,793	3,239	1.1	21.9
한의약선도기술개발	13,206	3,160	1.0	23.9
안전기술선진화	3,833	2,290	0.7	59.7
첨단의료복합단지기반기술구축	2,798	1,780	0.6	63.6
한국원자력의학원연구운영비지원	52,642	900	0.3	1.7
100세사회대응고령친화제품연구개발	4,209	356	0.1	8.5
국가보건의료연구인프라구축	12,417	280	0.1	2.3
합계	1,012,381	305,891	100.0	30.2

나. 신약개발단계별 정부 R&D 투자 현황

신약개발단계별 정부 R&D 투자 현황

- 중분류 기준 인프라 중 제도·정책부분은 '14년 15억원에서 '16년 44억원으로 연평균 71.8%, 타겟발굴 및 검증 단계는 연평균 49.7%(('14) 166억원 → ('16) 373억원)으로 증가한 반면 비임상 단계는 연평균 -22.9%(('14) 499억원 → ('16) 297억원)로 감소
- '16년 기준 신약플랫폼기술(871억원, 26.4%), 후보물질 도출 및 최적화(674억원, 20.4%), 타겟발굴 및 검증(373억원, 12.2%) 순으로 투자가 이루어지고 있음

  – 신약개발단계 후반부로 갈수록 투자 감소하며 인프라에 대한 투자 확대



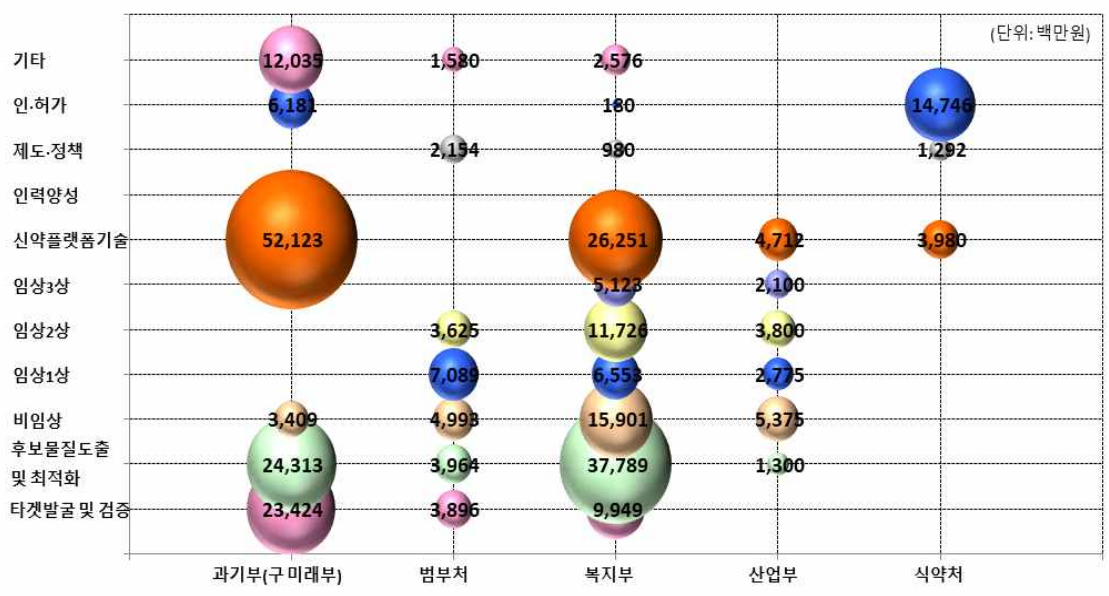
[그림 2-4] 신약개발분야 정부 R&D 신약개발단계별 투자 현황(2016)

<표 2-6> 신약개발분야 정부 R&D 신약개발단계별 투자 현황

구분			2014년		2015년		2016년		연평균 증가율 (%)
			연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)	
타켓발굴 및 검증	타겟발굴 및 검증		16,637	5.9	15,519	4.6	37,268	12.2	49.7
후보물질 도출 및 최적화	후보물질도출 및 최적화		47,422	16.9	42,563	12.7	67,366	22.0	19.2
비임상	비임상		49,887	17.8	63,828	19.0	29,678	9.7	-22.9
임상	임상1상		19,315	6.9	27,122	8.1	16,417	5.4	-7.8
	임상2상		14,619	5.2	16,474	4.9	19,150	6.3	14.5
	임상3상		5,597	2.0	7,356	2.2	7,223	2.4	13.6
인프라	신약 플랫폼 기술	타겟발굴 플랫폼	12,335	4.4	11,942	3.6	7,845	2.6	-20.3
		후보물질 발굴 플랫폼	23,002	8.2	28,271	8.4	26,952	8.8	8.2
		비임상 플랫폼	39,853	14.2	42,978	12.8	24,767	8.1	-21.2
		질환동물 플랫폼	6,351	2.3	6,485	1.9	17,870	5.8	67.7
		임상 플랫폼	15,245	5.4	18,735	5.6	9,632	3.1	-20.5
	인력양성		200	0.1	1,150	0.3	-	-	-
	제도·정책		1,500	0.5	6,308	1.9	4,426	1.4	71.8
	인·허가		13,149	4.7	14,407	4.3	21,107	6.9	26.7
기타	기타		15,360	5.5	32,600	9.7	16,190	5.3	2.7
합계			280,473	100.0	335,738	100.0	305,891	100.0	4.4

부처별 단계별 투자 현황

- '16년 신약개발분야에 가장 많이 지원한 과기부는 임상 이전 단계에 투자를 집중하였으며, 복지부 역시 후보물질 도출 및 최적화 단계(378억원)에 투자를 주력하였음
  - 과기부는 후보물질 도출 및 최적화(243억원), 타겟발굴 및 검증(234억원), 인프라의 신약플랫폼기술 중 전임상 플랫폼(174억원) 순으로 투자
  - 복지부는 후보물질 도출 및 최적화 단계(378억원), 비임상(159억원), 임상2상(117억원) 순으로 투자
  - 범부처 및 산업부의 경우 임상단계에의 투자 비중이 타 부처에 비해 높게 나타남



[그림 2-5] 신약개발분야 정부 R&D 부처별 단계별 투자 현황(2016)

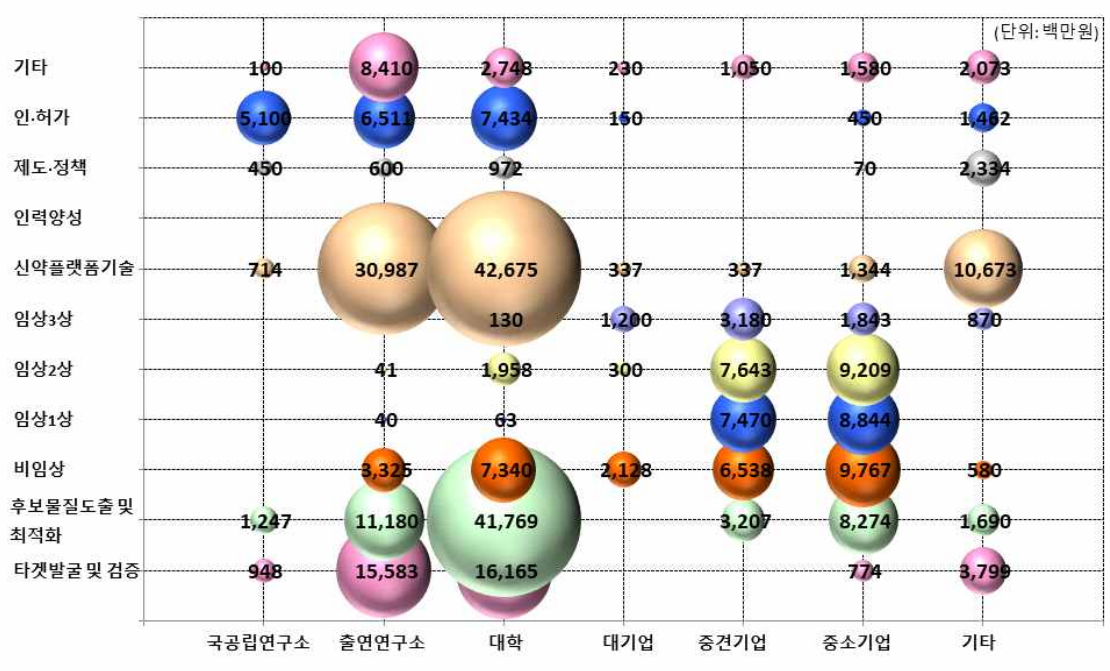
<표 2-7> 신약개발분야 정부 R&D 부처별 단계별 투자 현황(2016)

(단위: 백만원)

구분		과기부 (구 미래부)	범부처	복지부	산업부	식약처	합계
타겟발굴 및 검증		23,424	3,896	9,949	-	-	37,268
후보물질도출 및 최적화		24,313	3,964	37,789	1,300	-	67,366
비임상		3,409	4,993	15,901	5,375	-	29,678
임상1상		-	7,089	6,553	2,775	-	16,417
임상2상		-	3,625	11,726	3,800	-	19,150
임상3상		-	-	5,123	2,100	-	7,223
신약 플랫폼 기술	타겟발굴 플랫폼	6,697	-	1,148	-	-	7,845
	후보물질 발굴 플랫폼	16,115	-	10,337	500	-	26,952
	전임상 플랫폼	17,369	-	3,050	2,538	1,810	24,767
	질환동물 플랫폼	10,100	-	4,096	1,674	2,000	17,870
	임상 플랫폼	1,842	-	7,620	-	170	9,632
인력양성		-	-	-	-	-	-
제도·정책		-	2,154	980	-	1,292	4,426
인·허가		6,181	-	180	-	14,746	21,107
기타		12,035	1,580	2,576	-	-	16,190
합계		121,485	27,300	117,026	20,062	20,017	305,891

연구수행주체별 단계별 투자 현황

- 가장 큰 투자 비중을 보였던 대학에서는 신약플랫폼기술(427억원)과 후보물질도출 및 최적화(418억원) 단계에 투자를 주력
  - 출연연은 신약플랫폼기술(310억원), 타겟발굴 및 검증(156억원), 후보물질도출 및 최적화(112억원) 순으로 투자
  - 국공립연구소의 경우 인·허가에 투자를 집중하고 있었으며 중견기업 및 중소기업은 후보물질도출 및 최적화부터 임상2상단계에 주로 투자하였음



[그림 2-6] 신약개발분야 정부 R&D 주체별 단계별 투자 현황(2016)

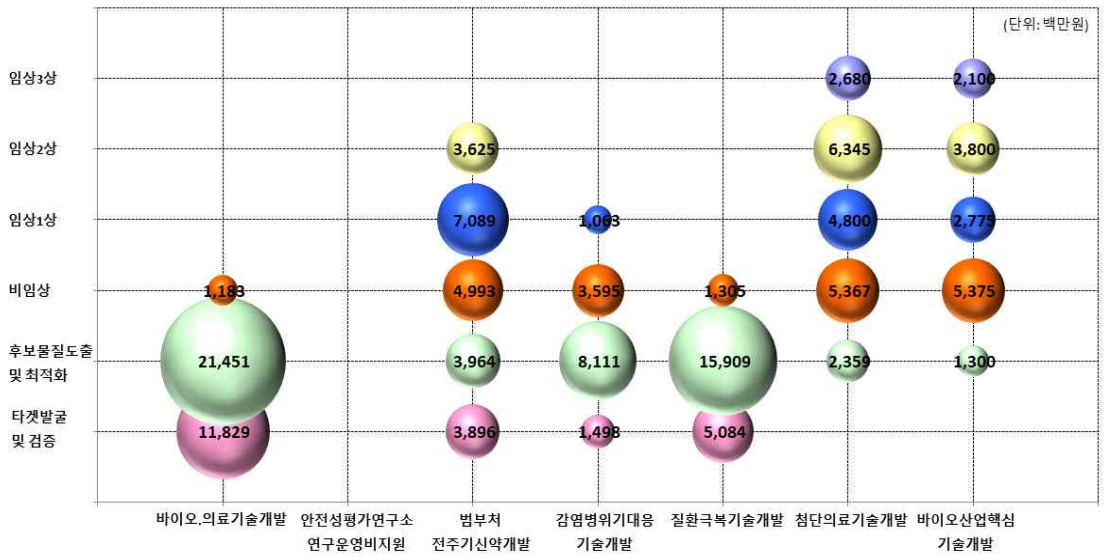
<표 2-8> 신약개발분야 정부 R&D 주체별 단계별 투자 현황(2016)

(단위: 백만원)

구분		국공립 연구소	출연 연구소	대학	대기업	중견 기업	중소 기업	기타	합계
타겟발굴 및 검증		948	15,583	16,165	-	-	774	3,799	37,268
후보물질도출 및 최적화		1,247	11,180	41,769	-	3,207	8,274	1,690	67,366
비임상		-	3,325	7,340	2,128	6,538	9,767	580	29,678
임상1상		-	40	63	-	7,470	8,844	-	16,417
임상2상		-	41	1,958	300	7,643	9,209	-	19,150
임상3상		-	-	130	1,200	3,180	1,843	870	7,223
신약 플랫폼 기술	타겟발굴 플랫폼	260	3,078	3,919	-	-	-	588	7,845
	후보물질 발굴 플랫폼	144	5,271	14,010	-	-	750	6,778	26,952
	전임상 플랫폼	310	18,519	4,468	-	-	-	1,470	24,767
	질환동물 플랫폼	-	3,869	11,650	-	-	594	1,757	17,870
	임상 플랫폼	-	250	8,628	337	337	-	80	9,632
인력양성		-	-	-	-	-	-	-	-
제도·정책		450	600	972	-	-	70	2,334	4,426
인·허가		5,100	6,511	7,434	150	-	450	1,462	21,107
기타		100	8,410	2,748	230	1,050	1,580	2,073	16,190
합계		8,559	76,675	121,253	4,345	29,424	42,155	23,480	305,891

주요사업별 단계별 투자 현황

- 신약개발분야 주요 사업 중 가장 큰 비중을 차지하는 과기부의 바이오·의료기술개발은 신약플랫폼기술(290억원)에 가장 많이 투자하였으며, 후보물질 도출 및 최적화(215억원), 타겟발굴 및 검증(118억원) 순으로 지원
  - 복지부의 질환극복기술개발과 감염병위기대응기술개발 모두 후보물질 도출 및 최적화(159억원, 81억원)에 투자 집중
  - 범부처전주기신약개발은 임상1상(71억원), 비임상(50억원), 후보물질 도출 및 최적화(40억원) 순으로 투자



[그림 2-7] 신약개발분야 정부 R&D 주요사업별 단계별 투자 현황(2016)

<표 2-9> 신약개발분야 정부 R&D 주요사업별 단계별 투자 현황(2016)

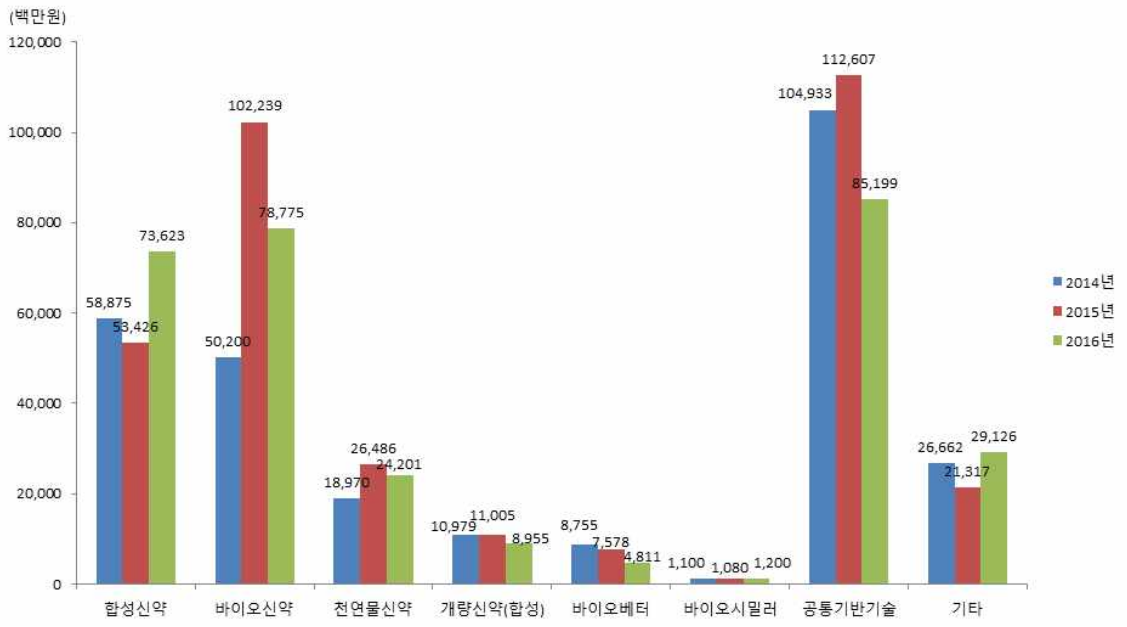
(단위: 백만원)

구분		바이오·의료 기술개발	안전성평가 연구소연구 운영비지원	범부처 전주기 신약개발	감염병 위기대응 기술개발	질환극복 기술개발	첨단의료 기술개발	바이오 산업핵심 기술개발
타겟발굴 및 검증		11,829	-	3,896	1,498	5,084	-	-
후보물질도출 및 최적화		21,451	-	3,964	8,111	15,909	2,359	1,300
비임상		1,183	-	4,993	3,595	1,305	5,367	5,375
임상1상		-	-	7,089	1,063	-	4,800	2,775
임상2상		-	-	3,625	-	-	6,345	3,800
임상3상		-	-	-	-	-	2,680	2,100
신약 플랫폼 기술	타겟발굴 플랫폼	4,492	-	-	-	-	-	-
	후보물질 발굴 플랫폼	11,804	-	-	650	1,005	380	500
	전임상 플랫폼	798	10,657	-	-	-	-	2,538
	질환동물 플랫폼	10,100	-	-	530	260	530	1,674
	임상 플랫폼	1,842	-	-	-	80	674	-
인력양성		-	-	-	-	-	-	-
제도·정책		-	-	2,154	-	100	-	-
인·허가		-	6,181	-	-	-	-	-
기타		4,573	-	1,580	300	-	1,050	-
합계		68,073	16,838	27,300	15,747	23,743	24,185	20,062

다. 의약품 종류별 정부 R&D 투자 현황

의약품 종류별 투자 현황

- 바이오신약에 대한 투자가 연평균 약 25.3%로 가장 많이 증가한 반면 바이오베터에 대한 투자는 연평균 약 -25.9%로 감소
  - 바이오신약 중 백신에 대한 투자가 '14년 79억원에서 '16년 259억원으로 가장 많이 확대됨(연평균 약 80.7%)
- '16년 기준 투자 규모의 절반 이상이 신약개발(54.8%)에 집중되어 투자되고 있으며, 공통기반기술(852억원, 27.9%), 바이오신약(788억원, 25.8%), 합성신약(736억원, 24.1%) 순으로 투자



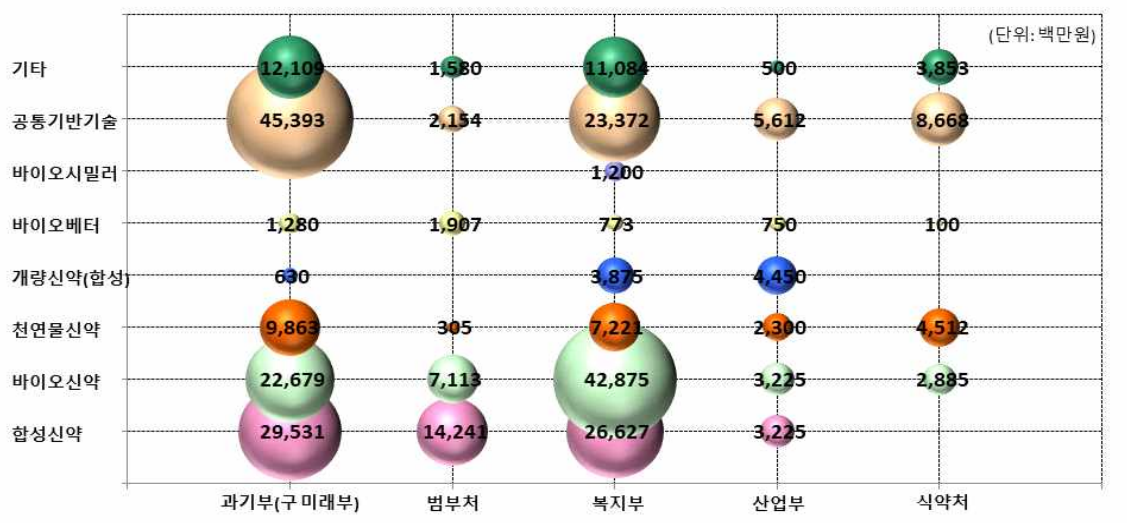
[그림 2-8] 신약개발분야 정부 R&D 의약품 종류별 투자 현황

<표 2-10> 신약개발분야 정부 R&D 의약품 종류별 투자 현황

구분		2014년		2015년		2016년		연평균 증가율 (%)
		연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)	연구비 (백만원)	비중	
합성신약		58,875	21.0	53,426	15.9	73,623	24.1	11.8
바이오 신약	단백질 치료제	12,810	4.6	18,180	5.4	14,384	4.7	6.0
	유전자 치료제	9,362	3.3	11,426	3.4	7,632	2.5	-9.7
	세포 치료제	7,107	2.5	45,793	13.6	17,313	5.7	56.1
	백신	7,923	2.8	13,049	3.9	25,878	8.5	80.7
	항체	12,998	4.6	13,791	4.1	13,568	4.4	2.2
천연물신약		18,970	6.8	26,486	7.9	24,201	7.9	12.9
개량신약(합성)		10,979	3.9	11,005	3.3	8,955	2.9	-9.7
바이오 베터	단백질 치료제	2,665	1.0	2,050	0.6	1,573	0.5	-23.2
	유전자 치료제	200	0.1	250	0.1	-	-	-100.0
	세포 치료제	75	0.0	130	0.0	-	-	-100.0
	백신	3,980	1.4	3,883	1.2	1,523	0.5	-38.1
	항체	1,835	0.7	1,265	0.4	1,714	0.6	-3.4
바이오시밀러		1,100	0.4	1,080	0.3	1,200	0.4	4.4
공통기반기술		104,933	37.4	112,607	33.5	85,199	27.9	-9.9
기타		26,662	9.5	21,317	6.3	29,126	9.5	4.5

부처별 의약품종류별 투자 현황

- 과기부는 공통기반기술(454억원), 합성신약(295억원), 바이오신약(227억원) 순으로 투자
- 복지부는 바이오신약(429억원)에 투자 주력하며, 합성신약(266억원), 공통기반기술(234억원), 천연물신약(75억원) 순으로 지원
- 범부처는 합성신약(142억원), 산업부·식약처는 부처 특성에 맞게 공통기반기술(56억원, 87억원)에 가장 많이 투자



[그림 2-9] 신약개발분야 정부 R&D 부처별 의약품종류별 투자 현황(2016)

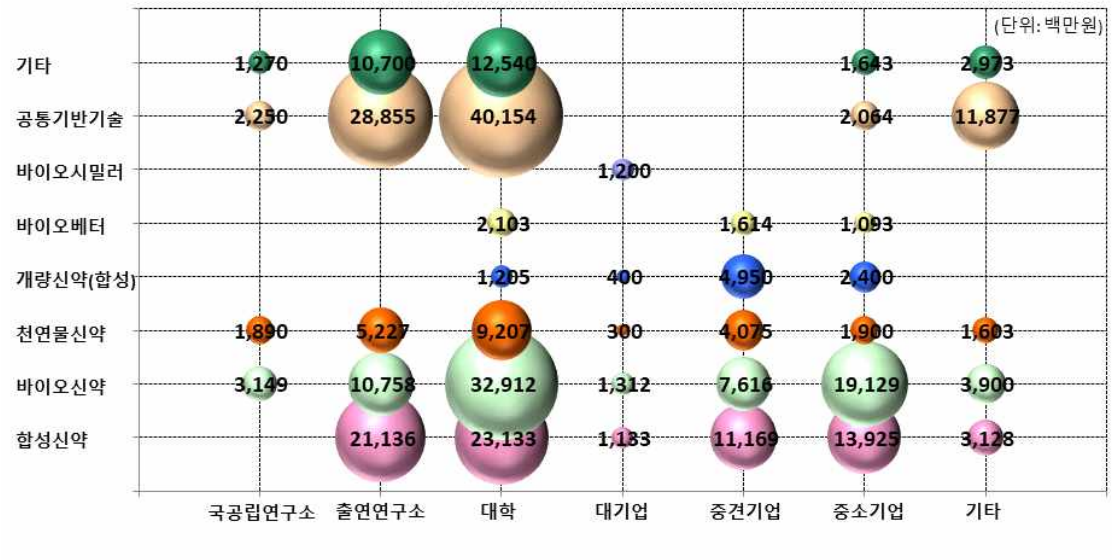
<표 2-11> 신약개발분야 정부 R&D 부처별 의약품종류별 투자 현황(2016)

(단위: 백만원)

구분		과기부 (구 미래부)	범부처	복지부	산업부	식약처	합계
합성신약		29,531	14,241	26,627	3,225	-	73,623
바이오신약	단백질 치료제	2,145	3,472	6,093	2,525	150	14,384
	유전자 치료제	2,200	200	5,023	-	209	7,632
	세포 치료제	9,308	-	7,140	700	165	17,313
	백신	2,425	-	21,092	-	2,361	25,878
	항체	6,601	3,441	3,527	-	-	13,568
천연물신약		9,863	305	7,221	2,300	4,512	24,201
개량신약(합성)		630	-	3,875	4,450	-	8,955
바이오베터	단백질 치료제	1,280	293	-	-	-	1,573
	유전자 치료제	-	-	-	-	-	-
	세포 치료제	-	-	-	-	-	-
	백신	-	-	773	750	-	1,523
	항체	-	1,614	-	-	100	1,714
바이오시밀러		-	-	1,200	-	-	1,200
공통기반기술		45,393	2,154	23,372	5,612	8,668	85,199
기타		12,109	1,580	11,084	500	3,853	29,126
합계		121,485	27,300	117,026	20,062	20,017	305,891

연구수행주체별 의약품종류별 투자 현황

- 가장 많이 투자한 대학의 경우 의약품종류 중에는 바이오신약(329억원)과 합성신약(231억원)에 투자를 집중하였으며, 바이오신약 중 백신(130억원) 개발에 투자 주력
  - 전반적으로 개량신약 보다는 합성신약과 바이오신약에 투자를 집중



[그림 2-10] 신약개발분야 정부 R&D 주체별 의약품종류별 투자 현황(2016)

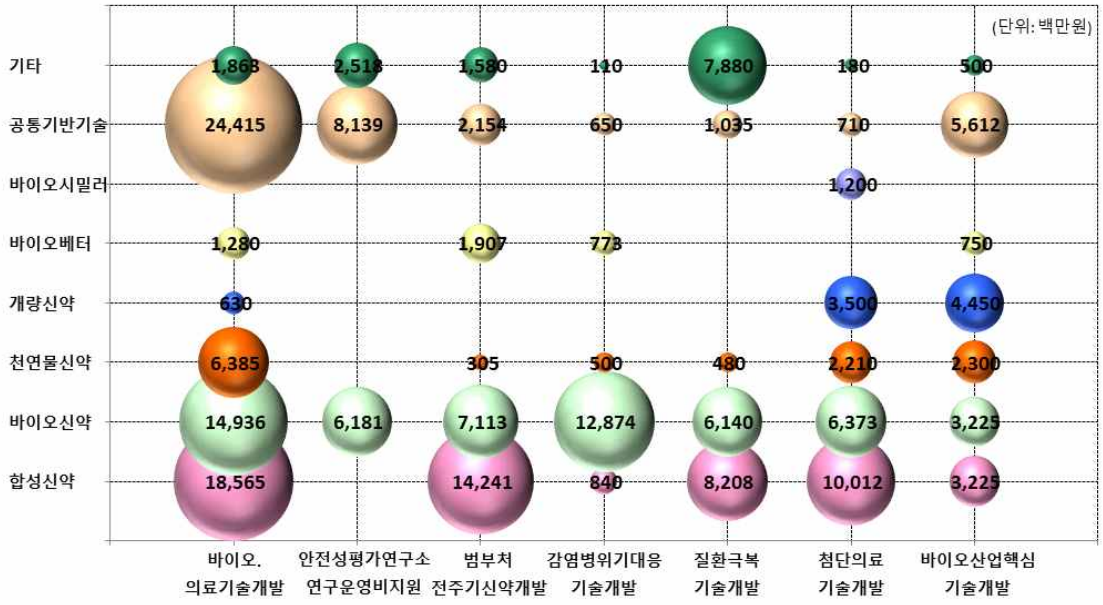
<표 2-12> 신약개발분야 정부 R&D 주체별 의약품종류별 투자 현황(2016)

(단위: 백만원)

구분		국공립 연구소	출연 연구소	대학	대기업	중견 기업	중소 기업	기타	합계
합성신약		-	21,136	23,133	1,133	11,169	13,925	3,128	73,623
바이오 신약	단백질 치료제	150	-	4,539	25	575	8,866	230	14,384
	유전자 치료제	-	620	3,839	-	336	2,537	300	7,632
	세포 치료제	-	7,738	5,385	-	3,220	100	870	17,313
	백신	2,601	1,210	13,014	1,287	1,250	5,836	680	25,878
	항체	398	1,190	6,135		2,235	1,790	1,820	13,568
천연물신약		1,890	5,227	9,207	300	4,075	1,900	1,603	24,201
개량신약(합성)		-	-	1,205	400	4,950	2,400	-	8,955
바이오 베터	단백질 치료제	-	-	480	-	-	1,093	-	1,573
	유전자 치료제	-	-	-	-	-	-	-	-
	세포 치료제	-	-	-	-	-	-	-	-
	백신	-	-	1,523	-	-	-	-	1,523
	항체	-	-	100	-	1,614	-	-	1,714
바이오시밀러		-	-	-	1,200	-	-	-	1,200
공통기반기술		2,250	28,855	40,154	-	-	2,064	11,877	85,199
기타		1,270	10,700	12,540	-	-	1,643	2,973	29,126

주요사업별 의약품종류별 투자 현황

- 주요사업 전반적으로 합성신약, 바이오신약, 천연물신약 순으로 투자하는 양상을 보여 신약 개발의 투자 비중이 높은 양상을 보임
  - 과기부의 바이오·의료기술개발은 공통기반기술(244억원)이 가장 큰 투자 비중을 차지하고 있으며, 복지부의 감염병위기대응기술개발은 바이오신약 중 백신 개발(127억원)에 투자를 집중하고 있음



[그림 2-11] 신약개발분야 정부 R&D 주요사업별 의약품종류별 투자 현황(2016)

<표 2-13> 신약개발분야 정부 R&D 주요사업별 의약품종류별 투자 현황(2016)

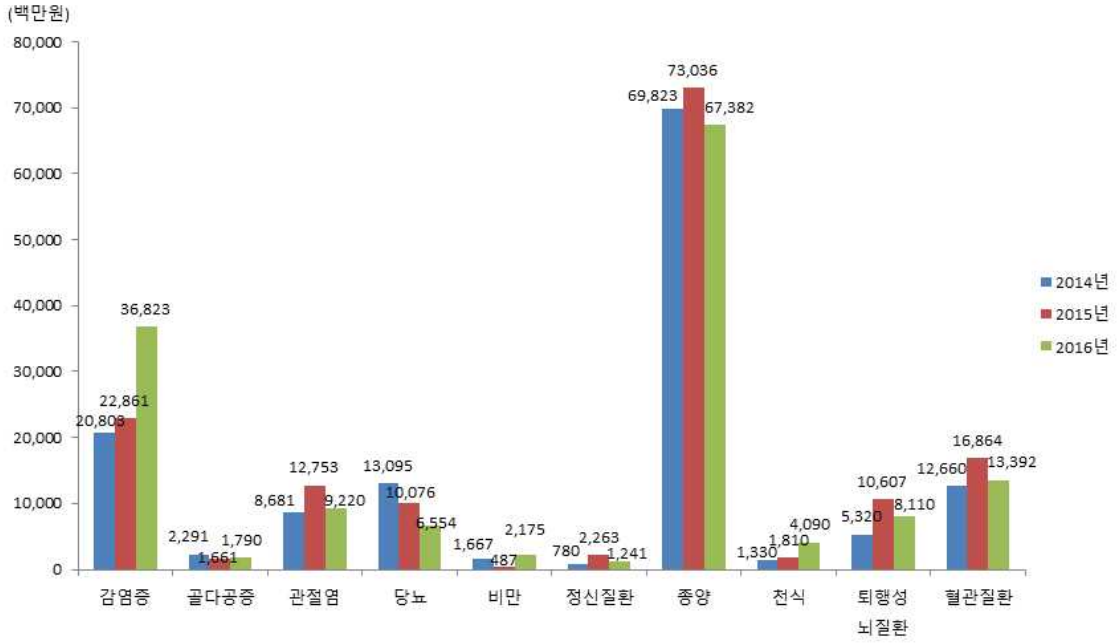
(단위: 백만원)

구분		바이오·의료기술개발	안전성평가연구소 연구운영비 지원	범부처 전주기 신약개발	감염병 위기대응 기술개발	질환극복 기술개발	첨단의료 기술개발	바이오 산업핵심 기술개발
합성신약		18,565	-	14,241	840	8,208	10,012	3,225
바이오신약	단백질 치료제	2,145	-	3,472	50	2,360	2,808	2,525
	유전자 치료제	2,200	-	200	-	1,110	973	-
	세포 치료제	2,370	6,181	-	125	1,200	1,380	700
	백신	1,620	-	-	12,699	1,010	837	-
	항체	6,601	-	3,441		460	375	-
천연물신약		6,385	-	305	500	480	2,210	2,300
개량신약		630	-	-	-	-	3,500	4,450
바이오베터	단백질 치료제	1,280	-	293	-	-	-	-
	유전자 치료제	-	-	-	-	-	-	-
	세포 치료제	-	-	-	-	-	-	-
	백신	-	-	-	773	-	-	750
	항체	-	-	1,614	-	-	-	-
바이오시밀러		-	-	-	-	-	1,200	-
공통기반기술		24,415	8,139	2,154	650	1,035	710	5,612
기타		1,863	2,518	1,580	110	7,880	180	500

라. 질환별 정부 R&D 투자 현황

질환별 투자 현황

- '16년 정부 R&D는 종양(674억원, 22.0%)에 투자를 주력하였으며, 천식이 연평균 75.4%로 크게 증가하였고 감염증(연평균 33.0%), 정신질환(연평균 26.1%), 퇴행성 뇌질환(연평균 23.5%)도 투자가 확대되었음
  - 반면, 당뇨는 연평균 -29.3%, 골다공증은 -11.6%, 종양은 -1.8%로 투자가 감소되고 있음



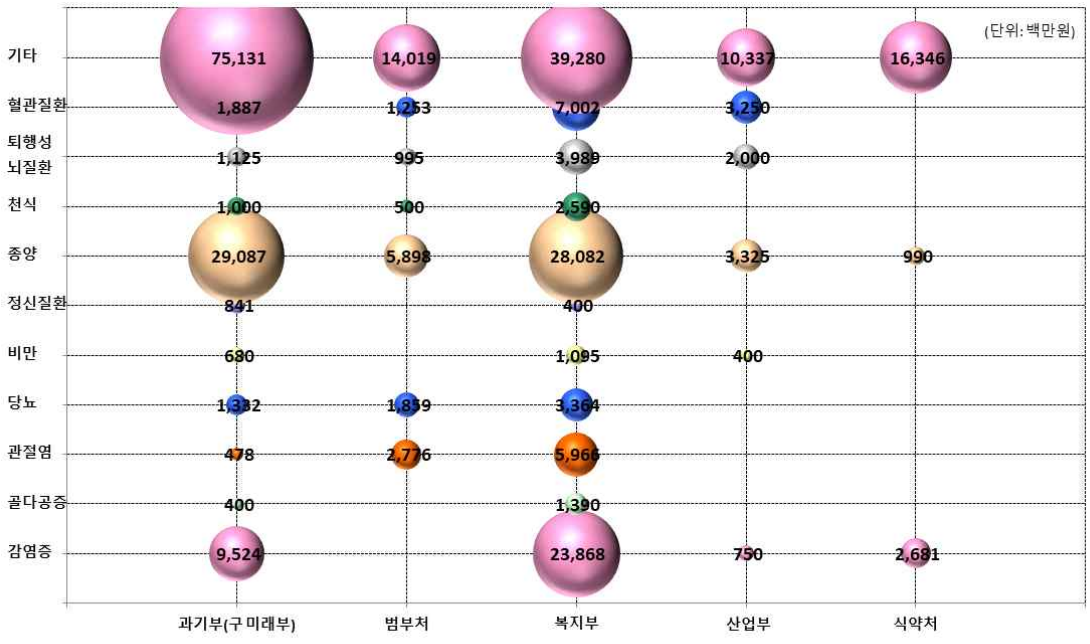
[그림 2-12] 신약개발분야 정부 R&D 질환별 투자 현황

<표 2-14> 신약개발분야 정부 R&D 질환별 투자 현황

구분	2014년		2015년		2016년		연평균 증가율 (%)
	연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)	
감염증	20,803	7.4	22,861	6.8	36,823	12.0	33.0
골다공증	2,291	0.8	1,661	0.5	1,790	0.6	-11.6
관절염	8,681	3.1	12,753	3.8	9,220	3.0	3.1
당뇨	13,095	4.7	10,076	3.0	6,554	2.1	-29.3
비만	1,667	0.6	487	0.1	2,175	0.7	14.2
정신질환	780	0.3	2,263	0.7	1,241	0.4	26.1
종양	69,823	24.9	73,036	21.8	67,382	22.0	-1.8
천식	1,330	0.5	1,810	0.5	4,090	1.3	75.4
퇴행성뇌질환	5,320	1.9	10,607	3.2	8,110	2.7	23.5
혈관질환	12,660	4.5	16,864	5.0	13,392	4.4	2.9
기타	144,025	51.4	183,320	54.6	155,113	50.7	3.8

부처별 질환별 투자 현황

- 부처별로 질환별 투자 현황을 분석한 결과 과기부·범부처·복지부·산업부 모두 종양(291억원, 59억원, 281억원, 33억원)에 가장 많이 투자
  - 과기부와 복지부는 종양 다음으로 감염증(95억원, 239억원)에 많이 투자하였으며, 식약처도 감염증(27억원)에 투자 비중이 가장 큼



[그림 2-13] 신약개발분야 정부 R&D 부처별 질환별 투자 현황(2016)

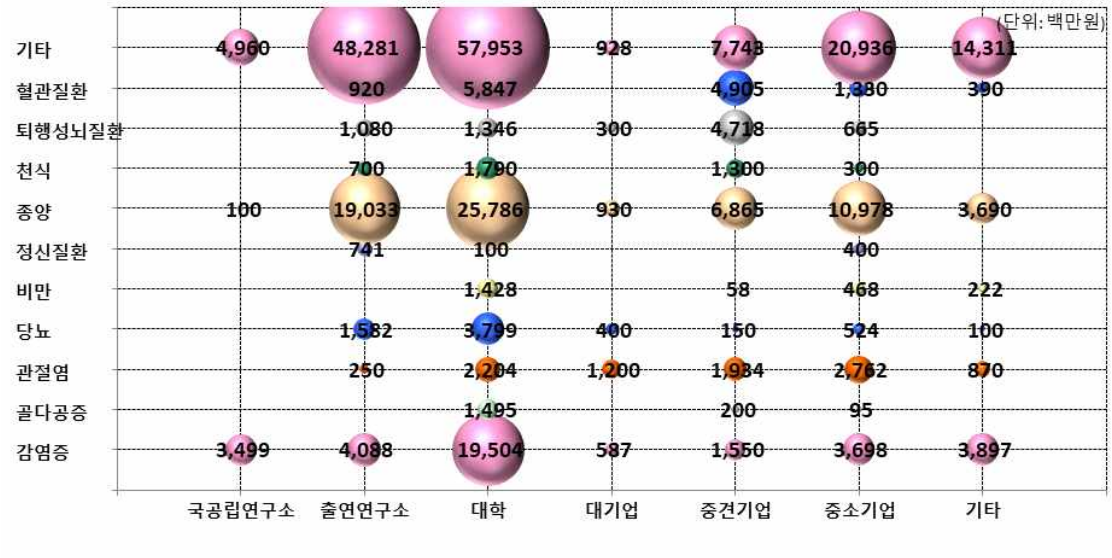
<표 2-15> 신약개발분야 정부 R&D 부처별 질환별 투자 현황(2016)

(단위: 백만원)

구분	과기부 (구 미래부)	범부처	복지부	산업부	식약처	합계
감염증	9,524	-	23,868	750	2,681	36,823
골다공증	400	-	1,390	-	-	1,790
관절염	478	2,776	5,966	-	-	9,220
당뇨	1,332	1,859	3,364	-	-	6,554
비만	680	-	1,095	400	-	2,175
정신질환	841	-	400	-	-	1,241
종양	29,087	5,898	28,082	3,325	990	67,382
천식	1,000	500	2,590	-	-	4,090
퇴행성뇌질환	1,125	995	3,989	2,000	-	8,110
혈관질환	1,887	1,253	7,002	3,250	-	13,392
기타	75,131	14,019	39,280	10,337	16,346	155,113
합계	121,485	27,300	117,026	20,062	20,017	305,891

연구수행주체별 질환별 투자 현황

- 전반적으로 종양과 감염증에 투자가 집중되었으며 특히 대학의 경우 다른 연구수행주체보다 감염증에 대한 투자가 비교적 많이 이루어짐
  - 중견기업의 경우 다른 연구수행주체에 비해 퇴행성뇌질환(47억원)과 혈관질환(49억원)에 높은 투자 비중을 보임



[그림 2-14] 신약개발분야 정부 R&D 주체별 질환별 투자 현황(2016)

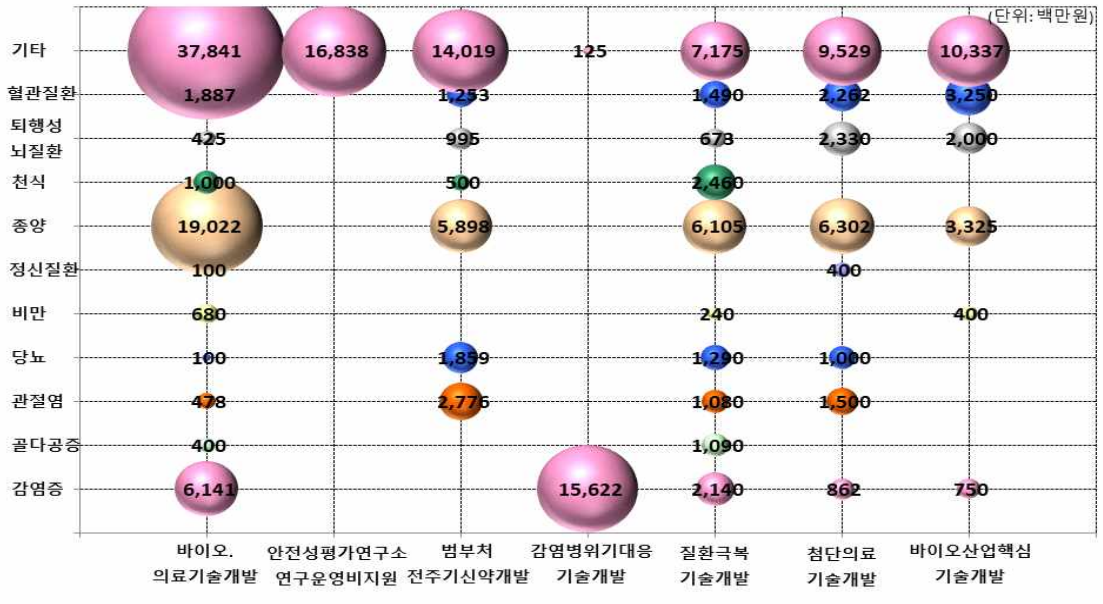
<표 2-16> 신약개발분야 정부 R&D 주체별 질환별 투자 현황(2016)

(단위: 백만원)

구분	국공립 연구소	출연연 구소	대학	대기업	중견 기업	중소 기업	기타	합계
감염증	3,499	4,088	19,504	587	1,550	3,698	3,897	36,823
골다공증	-	-	1,495	-	200	95	-	1,790
관절염	-	250	2,204	1,200	1,934	2,762	870	9,220
당뇨	-	1,582	3,799	400	150	524	100	6,554
비만	-	-	1,428	-	58	468	222	2,175
정신질환	-	741	100	-	-	400	-	1,241
종양	100	19,033	25,786	930	6,865	10,978	3,690	67,382
천식	-	700	1,790	-	1,300	300	-	4,090
퇴행성 뇌질환	-	1,080	1,346	300	4,718	665	-	8,110
혈관질환	-	920	5,847	-	4,905	1,330	390	13,392
기타	4,960	48,281	57,953	928	7,743	20,936	14,311	155,113

주요사업별 질환별 투자 현황

- 복지부의 감염병위기대응기술개발을 제외하고 신약개발분야 주요사업 전반적으로 종양에 가장 많이 투자



[그림 2-15] 신약개발분야 정부 R&D 주요사업별 질환별 투자 현황(2016)

<표 2-17> 신약개발분야 정부 R&D 주요사업별 질환별 투자 현황(2016)

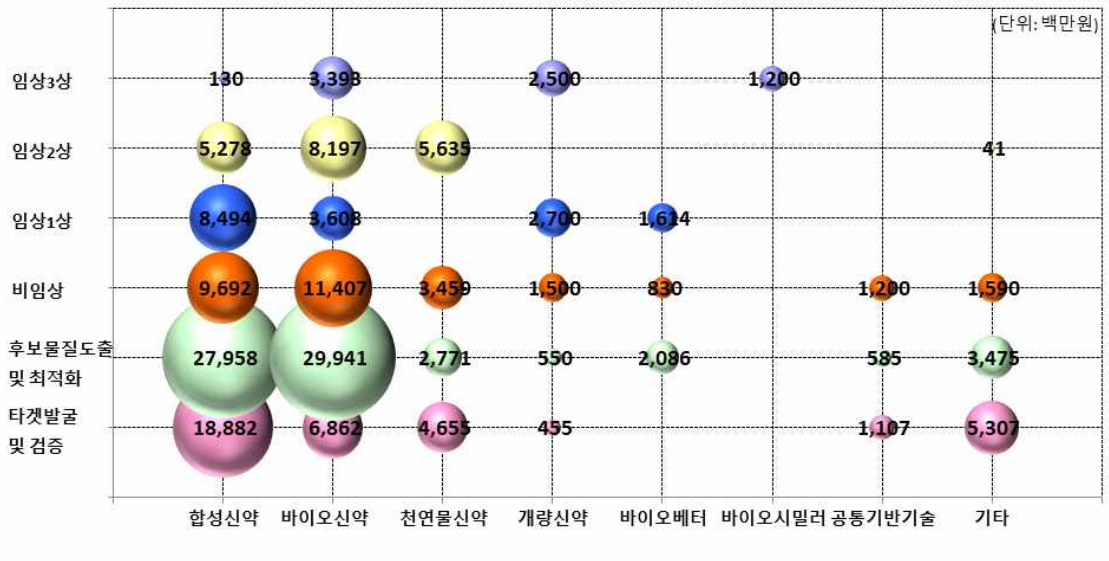
(단위: 백만원)

구분	바이오·의료기술개발	안전성평가연구소연구운영비지원	범부처전주기신약개발	감염병위기대응기술개발	질환극복기술개발	첨단의료기술개발	바이오산업핵심기술개발
감염증	6,141	-	-	15,622	2,140	862	750
골다공증	400	-	-	-	1,090	-	-
관절염	478	-	2,776	-	1,080	1,500	-
당뇨	100	-	1,859	-	1,290	1,000	-
비만	680	-	-	-	240	-	400
정신질환	100	-	-	-	-	400	-
종양	19,022	-	5,898	-	6,105	6,302	3,325
천식	1,000	-	500	-	2,460	-	-
퇴행성뇌질환	425	-	995	-	673	2,330	2,000
혈관질환	1,887	-	1,253	-	1,490	2,262	3,250
기타	37,841	16,838	14,019	125	7,175	9,529	10,337
합계	68,073	16,838	27,300	15,747	23,743	24,185	20,062

마. 교차분석

신약개발단계별 의약품종류별 투자 현황

- 합성신약 및 바이오신약은 후보물질도출 및 최적화 단계(280억원, 299억원)에 투자가 집중되어 있으며, 천연물신약은 임상2상 단계(56억원)에 가장 많이 투자
- 천연물신약은 임상2상(56억원), 타겟발굴 및 검증(47억원), 인·허가(41억원)순으로 투자
- 개량신약은 임상1상(27억원), 임상3상(25억원), 비임상(15억원) 순으로 투자
- 바이오 베터는 후보물질도출 및 최적화(21억원) 단계에 가장 많이 투자되었으며, 바이오시밀러는 임상3상(12억원)에만 투자되었음



[그림 2-16] 신약개발분야 정부 R&D 신약개발단계별 의약품종류별 투자 현황(2016)

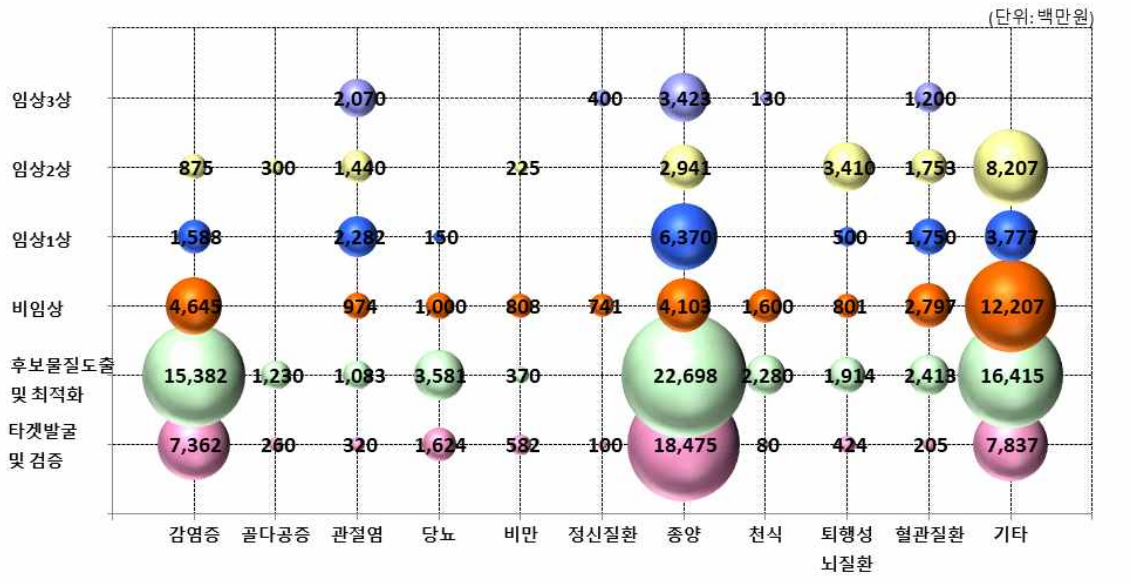
<표 2-18> 신약개발분야 정부 R&D 신약개발단계별 의약품종류별 투자 현황(2016)

(단위: 백만원)

구분		합성신약	바이오신약					천연물신약	개량신약	바이오베터					바이오시밀러
			단백질치료제	유전자치료제	세포치료제	백신	항체			단백질치료제	유전자치료제	세포치료제	백신	항체	
타겟발굴 및 검증		18,882	455	1,880	257	1,888	2,382	4,655	455	-	-	-	-	-	-
후보물질도출 및 최적화		27,958	4,901	2,483	4,520	10,952	7,085	2,771	550	1,393	-	-	693	-	-
비임상		9,692	3,747	600	1,250	5,255	555	3,459	1,500	-	-	-	830		-
임상1상		8,494	525	40	-	1,563	1,480	-	2,700	-	-	-		1,614	-
임상2상		5,278	3,982	1,900	1,440	875	-	5,635		-	-	-			-
임상3상		130	-	-	1,950	1,443	-	-	2,500	-	-	-			1,200
신약플랫폼기술	타겟발굴 플랫폼	1,050	150	240	-	-	300	140	-	-	-	-	-	-	-
	후보물질 발굴 플랫폼	162	-	180	350	794	1,676	1,473	200	180	-	-	-	-	-
	전임상 플랫폼	200	350	80	130	-	90	378	-	-	-	-	-	-	-
	질환동물 플랫폼	-	-	-	-	130	-	-	-	-	-	-	-	-	-
	임상 플랫폼	337	225	100	-	337	-	290	-	-	-	-	-	-	-
인력양성		-	-	-	-	-	-	-	-	-	-	-	-	-	-
제도·정책		-	-	-	435	-	-	380	-	-	-	-	-	-	-
인·허가		-	-	129	6,181	2,541	-	4,132	-	-	-	-	-	100	-
기타		1,440	50	-	800	100	-	888	1,050	-	-	-	-	-	-
합계		73,623	14,384	7,632	17,313	25,878	13,568	24,201	8,955	1,573	-	-	1,523	1,714	1,200

신약개발단계별 질환별 투자 현황

- 가장 많은 투자가 이루어진 종양 및 감염증은 후보물질도출 및 최적화 단계(227억원, 154억원)와 타겟발굴 및 검증 단계(185억원, 74억원)에 가장 많이 투자되었음
  - 반면 타 질환은 비교적 고르게 투자된 양상을 보이며, 혈관질환은 비임상(28억원), 관절염은 임상1상(23억원), 퇴행성뇌질환은 임상2상(34억원)이 가장 큰 투자 비중을 차지



[그림 2-17] 신약개발분야 정부 R&D 신약개발단계별 질환별 투자 현황(2016)

<표 2-19> 신약개발분야 정부 R&D 신약개발단계별 질환별 투자 현황(2016)

(단위: 백만원)

구분		감염증	골다공증	관절염	당뇨	비만	정신질환	종양	천식	퇴행성 뇌질환	혈관질환	기타	합계
타겟발굴 및 검증		7,362	260	320	1,624	582	100	18,475	80	424	205	7,837	37,268
후보물질도출 및 최적화		15,382	1,230	1,083	3,581	370	-	22,698	2,280	1,914	2,413	16,415	67,366
비임상		4,645	-	974	1,000	808	741	4,103	1,600	801	2,797	12,207	29,678
임상1상		1,588	-	2,282	150	-	-	6,370	-	500	1,750	3,777	16,417
임상2상		875	300	1,440	-	225	-	2,941	-	3,410	1,753	8,207	19,150
임상3상		-	-	2,070	-	-	400	3,423	130	-	1,200	-	7,223
신약 플랫폼 기술	타겟발굴 플랫폼	1,479	-	300	-	-	-	1,702	-	-	-	4,364	7,845
	후보물질 발굴 플랫폼	1,264	-	150	-	-	-	1,768	-	-	1,310	22,460	26,952
	전임상 플랫폼	-	-	200	-	-	520		-	940	23,107	24,767	7,845
	질환동물 플랫폼	530	-	180	-	-	-	2,699	-	490	610	13,361	17,870
	임상 플랫폼	337	-	421	-	-	-	230	-	-	220	8,424	9,632
인력양성		-	-	-	-	-	-	-	-	-	-	-	-
제도·정책		-	-	-	190	-	-	-	-	-	4,236	4,426	19,150
인·허가		2,861	-	-	-	-	-	250	-	-	-	17,996	21,107
기타		500	-	-	-	-	-	2,204	-	570	194	12,722	16,190
합계		36,823	1,790	9,220	6,554	2,175	1,241	67,382	4,090	8,110	13,392	155,113	305,891

의약품종류별 질환별 투자 현황

- 가장 많이 투자가 이루어지는 종양은 합성신약(294억원), 감염증은 바이오신약(232억원)에 투자가 가장 크게 이루어지고 있음
  - 전반적으로 개량신약보다는 신약 개발에 투자가 이루어지며, 바이오 시밀러는 관절염을 대상으로 투자가 되고 있음

[그림 2-18] 신약개발분야 정부 R&D 의약품종류별 질환별 투자 현황(2016)

<표 2-20> 신약개발분야 정부 R&D 의약품종류별 질환별 투자 현황(2016)

(단위: 백만원)

구분		감염증	골다공증	관절염	당뇨	비만	정신질환	종양	천식	퇴행성 뇌질환	혈관질환
합성신약		6,465	400	3,336	4,821	1,160	100	29,358	2,830	1,484	4,053
바이오신약	단백질 치료제	430	830	305	-	-	-	1,012	700		730
	유전자 치료제	670	-	-	-	-	-	5,090	-	-	260
	세포 치료제	-	-	2,460	380	-	-	5,112	80	-	1,250
	백신	20,568	-	-	-	-	-	4,760	-	-	-
	항체	1,508	-	803	224	-	-	8,458	-	390	-
천연물신약		1,411	300		100	665	741	919	-	4,535	494
개량신약(합성)		-	-	300	700	-	400	1,950	-	920	2,950
바이오베터	단백질 치료제	-	-	-	-	-	-	500	-	-	-
	유전자 치료제	-	-	-	-	-	-	-	-	-	-
	세포 치료제	-	-	-	-	-	-	-	-	-	-
	백신	1,523	-	-	-	-	-	-	-	-	-
	항체	-	-	-	-	-	-	-	-	-	-
바이오시밀러		-	-	1,200	-	-	-	-	-	-	-
공통기반기술		3,458	-	376	-	-	-	5,784	165	640	2,800
기타		790	260	440	330	350	-	4,439	315	140	855
합계		36,823	1,790	9,220	6,554	2,175	1,241	67,382	4,090	8,110	13,392

바. 분석결과 및 시사점

- 타겟발굴 및 검증부터 인·허가까지의 신약개발단계별 투자 현황을 살펴보면 타겟발굴 및 후보물질 도출 단계(약 34.2%)의 비중이 높은 반면, 인프라 중 인력양성, 제도·정책, 인·허가 부분의 투자 비중이 낮아 조정 필요
  - 신약개발 및 제약산업의 성장을 위해서는 신약개발을 촉진시킬 수 있는 제도·정책 및 인프라적인 부분에 대한 정부 차원의 지원이 필요[2)]
  - 특히 신약개발 분야는 신약 개발 각 단계에 인력이 필요할 뿐만 아니라 임상 후의 각 단계의 리스크를 관리할 수 있는 전문 인력의 양성이 중요함
    - 제도·정책 및 인·허가 단계에 대한 지원이 각 연평균 71.8%, 26.7%로 증가한 부분은 적절하나, 인력 양성을 위한 지원을 확대하여야 할 것으로 판단됨

- 합성 신약의 성공 빈도가 낮아지고, 합성의약품에 비해 부작용이 적은 바이오의약품의 시장이 활발해지고 있음[3)]
  - 바이오 의약품에 대한 투자가 크게 확대되고 있어 시장의 흐름이 적절히 반영되는 것으로 나타남

- 종양질환(22.0%)의 투자비중이 가장 높으나 투자 규모가 연평균 -1.8%로 감소하는 반면 천식, 감염병 및 퇴행성뇌질환 등에 대한 투자는 증가(75.4%, 33.0%, 23.5%)
  - 환경적 변화 및 치매, 신종 감염병 등에 대한 대응을 위한 공공적 보건의료 R&D 지원을 강화한 것으로 보임

2) 신영기 (2010), "우리나라 신약개발의 주요 현안 및 대응방안", 「과학기술정책」 178 : 39-42.
3) 한국수출입은행 (2017), 「세계 의약품 산업 및 국내산업 경쟁력 현황: 바이오의약품 중심」, 한국수출입은행.

■ 과기부 및 복지부는 후보물질 도출 및 최적화 단계에 가장 큰 투자 비중을 보였으며, 범부처 사업은 타겟발굴부터 임상 2상 단계까지, 산업부는 후보물질도출 및 최적화부터 신약플랫폼기술 단계까지 비교적 고르게 지원

● **부처간 투자 효율성을 제고하기 위하여 「바이오 중기('16~'18) 육성전략(안)(2016)」에서는 신약개발 관련 부처의 역할분담(안)*을 제안**

* 과기부는 기초연구부터 후보물질최적화 단계까지, 복지부는 비임상 및 임상단계, 산업부는 IP 사업화, 식약처는 허가 및 컨설팅을 담당하도록 조정

– 각 부처에서는 제안된 신약개발단계별 부처간 역할 분담(안) 보다 폭넓게 지원하고 있어 역할 분담의 조정 필요

■ 민간기업 중심의 신약개발 투자를 유도하기 위해서는 중소기업에의 투자 확대가 필요하며 대기업 및 중견기업의 경우에는 연구개발 역량을 감안하여 전략적인 투자가 필요함[4)]

● **신약개발분야 정부 R&D 투자포트폴리오를 분석한 결과 대기업 및 중견기업에의 투자는 연평균 -39.8%, -13.4%로 크게 축소하고 있었으며, 중소기업에의 투자는 11.1%로 확대하고 있었음**

■ 연구자가 제안한 연구 목표 및 내용을 바탕으로 신약개발단계를 분류하며, 단계가 명확하게 제시되어 있지 않은 경우 분류가 어려움

● **특히 계속과제의 경우 신약개발단계를 전부 다루는 과제가 있음**

– 해당 과제가 어느 단계까지 진행되었는지 확인이 어려움

4) 홍미영 외 (2016), 「신약개발 분야 정부/민간 R&D의 역할조정을 통한 효율화 방안 연구」, 한국과학기술기획평가원.

■ 현재 의약품 종류는 의약품의 제조방식(합성의약품/바이오의약품) 및 신약 여부(오리지널 의약품/복제의약품) 등에 따라 분류

- 기존의 의약품과 새로 개발한 의약품을 복합적으로 투여하는 콤비 형식의 의약품과 같은 경우 구분이 어려우며, 의약품에 대한 기능적 또는 물질적 관점에 따라 분류가 모호해지는 면이 있음
- 고부가가치 창출을 위한 첨단 바이오의약품, 동반진단 등 기술적·치료적 측면 및 기술 동향을 고려한 분류 기준의 개선 필요
  - 분류가 명확하지 않은 부분을 최소화하기 위해 바이오신약, 바이오베터 등 중분류에도 기타 분류코드 추가

■ 기존 질환 분류 기준은 「신약개발 R&D 투자 효율화 방안.(2012)」에서 제안된 것으로, 전문가 의견을 바탕으로 수립

- 현재 분류 기준인 10대 질환에 포함되지 않는 질환을 대상으로 하는 과제가 많음(기타 50.7%)
  - 희귀·난치성 질환 등 대두되고 있는 사회적 문제에 대응하기 위한 R&D가 확대되고 있어 이를 반영하기 위한 분류체계 개선 필요
  - 전문가별로 국내 환경을 감안 10대 질환종류에 시각차가 있을 수 있어 이에 대한 고려가 필요하며, 보편적으로 사용되는 국제질병분류(ICD), 한국표준질병사인분류(KCD) 혹은 WHO의 의약품 성분코드(ATC)와 같은 분류체계의 활용 필요

■ 신약개발분야에 대한 정부 R&D 투자의 효과 분석이 필요

- 신약개발분야 민간 R&D 투자 현황 파악을 통해 민간 영역 투자 효과 분석 필요
- 더불어 정부의 지원으로 수행된 과제가 어느 단계까지 발전이 되었는지 등에 대한 파악 및 추가 분석 필요

제3장
기계학습방법의 이론적 배경

제3장 기계학습방법의 이론적 배경

제1절 기계학습 정의 및 발전동향

가. 기계학습의 정의

통계적 기계학습

- 응용 통계학 관점에서 기계학습은 컴퓨터를 이용하여 복잡한 함수를 통계적으로 추정하는 응용 통계학 기법의 일종[5)]

경험의 강조

- Tom Mitchell은 기계학습을 경험(E), 성능(P), 태스크(T)의 관점에서 정의
  - 태스크(T)와 관련된 경험(E)이 늘어나면 성능(P)이 향상되는 프로그램을 기계학습 프로그램으로 정의[6)]

기계학습과 다른 인공지능과의 차이

- 기계학습은 축적된 데이터를 이용, 태스크 해결 성능 향상이 가능
  - 일반적인 인공지능 프로그램이 문제해결을 위한 전략을 프로그램화 한다면, 기계학습 기반의 프로그램은 수집된 데이터로부터 문제 해결을 위한 모델을 만들어내는 방법으로 구현

5) I. Goodfellow et al. Deep Learning, MIT Press, 2016.
6) T. M. Mitchell. Machine Learning, McGraw-Hill Education, 1997.

나. 기계학습 발전동향

1) 인공지능의 대표성과를 통해 본 기계학습의 발전양상

프로스펙터(Prospector)[7]

- 광물 탐사 과정에서 채집한 데이터를 기반으로 문제 해결책을 제시하는 광물 탐사 전문가 시스템
  - 지질학자가 탐사정보를 시스템에 제공(입력) 시, 제공된 정보를 광석이 매장된 지질의 특성과 비교하고 추가 정보 분석을 통해 광석 매장 가능성을 추정
  - 광석 특성에 대한 데이터베이스를 검색하여 광물자원의 분포에 대한 분석을 제공
- 모델 기반한 추정을 위해 베이즈 정리(Bayes' theorem)을 이용하여 우도값(likelihood)을 계산, 적절한 추론 규칙을 프로스펙터 데이터베이스에 저장
  - 기본적으로 추론 규칙은 인간 지질 전문가의 지식으로부터 수집되며, 수집된 지식을 통해 베이즈 룰을 이용하여 가장 가능성이 높은 가설을 선택
  - 불완전하고 불분명한 지식이 주어진다는 현실에 대응할 수 있도록 가설의 우도(likelihood)를 계산하여 후보 가설을 걸러냄
- 프로스펙터와 같은 전문가 시스템은 목표 도메인에 대한 전문가들의 접근 방법(규칙)을 수집 후 수집된 정보를 데이터베이스에 저장, 관측된 사실에 근거하여 찾고자하는 규칙을 점화하는 방식으로 추론을 진행
  - 일부 도메인에서는 인간 전문가와 유사하거나 능가하는 성과를 달성
  - 그러나 전문가 시스템은 규칙 데이터베이스의 갱신은 가능하나 경험으로부터 학습할 수는 없다는 한계점이 있음

7) P. E. Hart and R. O. Duda, "PROSPECTOR - a computer-based consultation system for mineral exploration," Artificial Intelligence Center, SRI International, Technical Note, No. 155, 1977.

딥블루(Deep Blue)[8)]

- 1987년 체스 그랜드마스터인 게리 카스파로프에게 승리하면서 인공지능의 놀라운 성능을 입증하는 계기가 됨
- 실제 구현된 내용 및 개별 구성요소를 살펴보면, 딥블루는 기계학습 시스템이라기보다는 전문가의 체스 플레이 과정을 규칙으로 표현한 후 빠르게 문제 공간을 탐색하는 전문가 시스템에 가까움
  - 인간 그랜드마스터에 의해 개발된 오프닝전략을 데이터베이스화하여 사용
  - 위치 평가 함수를 계산하는 과정에서 대부분의 특성 값과 가중치가 인간에 의해 결정
  - 위치 검색 함수는 휴리스틱 알고리즘(heuristic algorithm)[9)]을 활용하여 탐색 공간의 크기를 감소
- 즉 딥블루는 한정된 시간 동안 위치를 결정할 수 있도록 빠르게 문제 공간을 탐색할 수 있는 고성능 연산머신에 체스 플레이 규칙을 구현한 전문가 시스템
  - 경험을 통해 문제 해결 성능을 향상시키는 기계학습 시스템에 해당하지는 않음

왓슨(Watson)[10)]

- 2011년 유명 텔레비전 퀴즈 쇼인 'Jeopardy!'에 출연, 인간과 동일한 조건에서 인간에게 승리하는 업적 달성
  - (1) 자연언어로 주어진 질문의 처리, (2) 오픈 도메인에 대한 질문 해결, (3) 인간 경쟁자보다 빠른 시간에 정확한 답변을 탐색함
- 질문을 이해하는 것이 아니라 자연언어로 주어진 질문을 처리하여 그에 맞는 답을 탐색
  - 질문으로 주어진 문장 분석 후 해답 탐색을 위한 실마리를 찾고 답에 대한 가설을 생성

8) M. Campbell, A. J. Hoane Jr., and F.-h. Hsu, "Deep Blue," Artificial Intelligence, Vol. 134, 2002, pp. 57-83.
9) 문제의 답을 알기 어렵거나 많은 시간이 요구될 경우 이러한 목표를 하나 혹은 둘 모두를 포기하고 값을 어림짐작하는 알고리즘
10) D. A. Ferrucci, "Introduction to "This is Watson"," IBM Journal of Research and Development, Vol. 56, 2012, pp. 1:1-1:15.

- 퀴즈 쇼 처리를 위하여 수백 개의 알고리즘을 결합하였으며, 여러 개의 가설이 정확할 가능성을 계산하기 위하여 증거를 결합하는 과정에서 기계학습 방법을 활용

- 증거를 결합하는 과정에서 기계학습을 통해 성능 향상을 기대할 수 있음
- 확정된 지식을 빠른 시간에 결합하는 능력뿐만 아니라 오픈 도메인에 대한 질의를 정확하게 처리함으로써 딥블루에 비해 한층 발전했다고 볼 수 있음

알파고(AlphaGo)

- 구글 딥마인드에서 개발한 바둑 인공지능 시스템 알파고는 2016년 이세돌 9단과의 5번 대국에서 4번을 승리
  - 바둑은 그간 컴퓨터가 처리하기 어려운 크기의 문제 공간을 가졌다고 평가를 받아왔는데, 알파고는 기계학습 관점에서 기존의 규칙 기반 시스템과는 전혀 다른 발전을 보여줌
  - 인간 기보의 데이터를 이용, 포석을 위한 판단 시스템을 일차 훈련시킨 후 알파고 프로세스간의 자체 게임을 통해 판단 시스템을 강화시킴으로써 인간 기보에서는 존재하지 않았던 새로운 포석을 찾을 수 있었음
- 알파고에서 활용한 딥뉴럴넷은 태스크 해결을 위한 문제의 표현 학습에 능한 방법
  - 바둑과 같이 표현 공간이 큰 문제를 대상으로, 간단하지만 비선형의 함수를 이용하여 표현 공간을 전이하면서 태스크 해결에 적합한 표현을 찾음
  - 데이터에 기반한 딥뉴럴넷을 이용하여 인간 전문가가 충분하게 분석할 수 없는 큰 문제 공간에서 문제 해결에 적합한 표현을 찾을 수 있었음
  - 프로세스간의 게임을 통해 포석 결정 정책을 강화시켰다는 점에서 기계학습이 크게 기여

■ 인공지능/기계학습의 발전 양상 요약

- 초창기 인공지능/기계학습 모델은 인간 전문가의 문제 해결 방식을 규칙으로 표현한 후 규칙의 결합을 통한 추론으로 문제를 해결하는 전문가 시스템을 구현
- 대표적으로 체스게임에서 인간에게 승리한 딥블루와 퀴즈 게임에서 인간에게 승리한 왓슨은 초고성능 연산 머신에 기반한 전문가 시스템에 가까운 존재였음
- 전문가 시스템은 해당 분야에서 큰 성공을 거두었으나 경험 기반 학습모델과 차이가 있어 문제 해결 경험이 성능 향상으로 이어지지 못한다는 점에서 한계가 존재
- 이와 달리 알파고는 경험을 통해 스스로의 성능을 향상시킬 수 있다는 특징을 가지고 있으며, 바둑에서 인간에게 승리했을 뿐 아니라 승리 과정에서 사용한 포석이 일반적인 포석과는 다르다는 점에서 학습을 통한 성능 향상 접근법의 가능성을 보여 주었음

2) 기계학습의 향후 전망

■ 장난감 문제(Toy problem)에서 현실세계로의 적용 확대

- 태동기 시점 및 1960년대까지 자동번역 등의 분야에서 인공지능은 큰 기대를 받았으나 이를 충족하지는 못함
- 1990년대 이후 관련 이론이 고도화 되었으나 , 컴퓨팅 파워 부족으로 인공지능은 비전공자의 관심을 받기 어려운 장난감 문제(toy problem)에 주로 적용
- 현재는 비약적으로 발전한 컴퓨팅 파워와 데이터량을 바탕으로 현실세계의 문제해결 및 활용에 적극 활용될 가능성이 높음
  - 알파고가 바둑이라는 게임의 룰을 규칙 데이터베이스의 형태로 제공하지 않고 프로그램이 승리할 수 있는 방법을 데이터로부터 학습했다는 점을 주목할 필요

■ 데이터기반 학습의 제약은 향후 발전의 병목요소로 작용 가능

- 알파고로 대표되는 딥뉴럴넷은 이미지 인식과 음성 인식에서 인간과 동등한 수준의 성능을 달성, 이에 따라 RNN(Recurrent Neural Network)을 이용해 생성된 문장을 현장에 적용하기 위해 노력
  - 그럼에도 데이터에 기반한 학습은 인공지능/기계학습에 대한 일반 대중의 기대를 만족하기에 부족한 면이 있음
  - 많은 병원들이 진료의 정확성과 편의성을 높이고자 IBM 왓슨을 질병 진단에 사용하고 있으나 이는 의사를 대체한다기 보다는 의사결정을 지원하는 시스템으로 보는 것이 적절[11]
- 알파고 또한 단순하게 기보를 학습하였을 경우 인간이 생각지도 못한 수를 생산하지 못하였을 가능성이 있음
  - 알파고는 학습과정에서 또 다른 알파고과 꾸준히 대국하였고 이러한 과정에서 독자적인 수 생성이 가능했을 것으로 판단
- 알파고와 같이 새로운 가설을 생성하고 가설을 확인할 수 있는 상황이 아닐 경우 학습한 데이터를 능가하거나 학습한 데이터에 존재하는 편향성을 극복하기는 쉽지 않음
- 또한, 데이터의 수집이 여의치 않을 경우 다양하고 풍부한 학습 자체가 어려워질 수 있음
  - 특히 개인 프라이버시와 관련된 데이터의 경우 개인 데이터의 처리 및 이동에 관련된 EU 규정[12](2018년 5월 시행 예정)으로 개인 데이터를 축적한 포털 사이트여도 과거와 같이 비교적 자유롭게 개인 데이터를 확보·분석·활용하기 어려울 수 있음

11) http://www.yonhapnews.co.kr/bulletin/2017/01/18/0200000000AKR20170118179700017.HTML
12) http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679

■ 궁극적으로는 새로운 가설을 스스로 생성하고 학습할 수 있는 인공지능/기계학습 모델 필요

- 딥블루씨, 왓슨과 같이 특정 분야에 대한 인간의 문제해결 방식을 규칙화하여 적용하는 전문가 시스템은 접근 방식에서 제약이 있을 수 있음
  - (1) 적용 분야 한정 (2) 전문가 문제해결 방식의 명확화·규칙화 (3) 경험을 통한 학습
- 과거 축적된 방대한 데이터를 분석하여 학습을 진행하는 데이터 기반 기계학습의 경우 다음과 같은 제약이 존재
  - (1) 전분야 방대한 데이터 확보의 어려움 (2) 데이터에 포함된 패턴을 이상을 보여주는데 한계 (3) 편향된 데이터가 주어졌을 경우 데이터의 편향성 극복 필요
- 이러한 문제를 해결하고 현재 생성되고 있는 인공지능/기계학습에 대한 대중의 기대를 만족시키기 위해서는 새로운 가설을 생성하고 테스트 할 수 있는 방법 개발 필요

제 2 절 기계학습 적용 방법 탐색

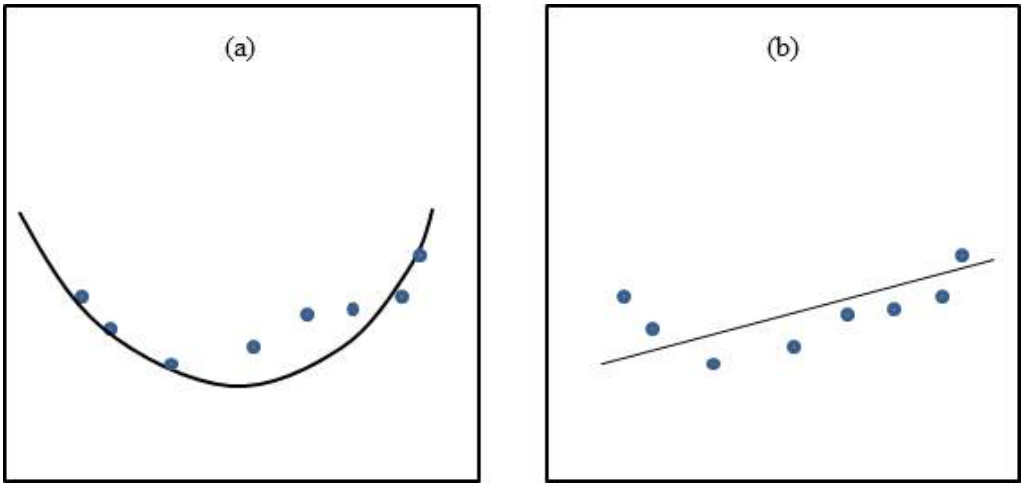
가. 지도학습(Supervised learning)

1) 지도학습의 개념

- 입력 **x**가 주어졌을 때 입력과 연관된 결과 **y** 가 존재하는 상황에서 훈련데이터를 이용하여 입력과 결과를 매칭하는 모델을 학습하는 방법
  - 결과가 명시되지 않은 입력 데이터 **x**가 주어졌을 때 주어진 입력 데이터에 대한 결과를 예측하는 것이 주요 임무이며 분류(classification)와 회귀(regression)로 구분 가능
  - 분류(Classification)
    - 입력 **x**를 $y \in \{1, \ldots, C\}$(여기서 C는 클래스 혹은 카테고리의 수)로 매칭하는 작업

      ※ 고양이에 해당하는 사진들을 모아 고양이 클래스를, 개에 해당하는 사진들을 모아 개 클래스를 생성하여 고양이 사진을 고양이로, 개 사진을 개로 분류하는 모델을 생성한 후 레이블이 부여되지 않은 사진이 입력되었을 때 개 혹은 고양이인지여부를 판단하는 모델을 생성하는 작업이 분류에 해당
  - 회귀(Regression)
    - 입력 **x**와 연관되는 결과 **y**가 실수로 주어지며 학습의 목표는 훈련 데이터로부터 입력과 **y**를 연결하는 모델 $\hat{f}$ $(f: \mathbb{R}^n \rightarrow \mathbb{R})$을 찾는 것

      ※ 과거 입학생들의 학점에 기반하여 신입생의 학점을 예상해보거나, 과거 주가지수 변화를 모델링한 후 주가지수를 예측하는 과정이 회기에 해당
    - [그림 3-1]은 원으로 표시된 훈련 데이터에 대하여 차수가 2인 다항함수와 차수가 1인 선형 함수를 적용 하여 회기분석(regression)를 실시한 결과
    - (a)의 모델과 (b)의 모델 모두 주어진 데이터를 완벽하게 설명하지는 못하지만, 적용 모델의 표현 능력에 따라 예측의 정확도가 변할 수 있다는 것을 알 수 있음



[그림 3-1] (a) 다항회귀(Polynomial regression) (b) 선형회귀(Linear regression)

- 분류와 회귀분석 모두 데이터에 기반한 모델을 형성하기 위하여 모델의 성능을 측정할 수 있는 성능 측도를 요구

  – 평균제곱근에러(Mean Squared Error, MSE)는 m개의 테스트 데이터 $\boldsymbol{y}^{(\text{test})}$와 해당 데이터에 대한 예측값 $\hat{\boldsymbol{y}}^{(\text{test})}$에 대하여 다음과 같이 정의됨:

$$\text{MSE}_{\text{test}} = \frac{1}{m}\sum_{i}\left(\hat{\boldsymbol{y}}^{(\text{test})} - \boldsymbol{y}^{(\text{test})}\right)_i^2$$

  – 잔차제곱합(Residual Sum of Square, RSS)는 테스트 데이터 $\boldsymbol{y}^{(\text{test})}$와 테스트 데이터에 대한 regression 함수 $f_{\boldsymbol{\theta}}$에 대하여 (여기서 $\boldsymbol{\theta}$는 회귀분석 함수를 결정하는 인자 벡터) 다음과 같이 정의함:

$$\text{RSS}(\boldsymbol{\theta}) = \sum_{i}^{m}\left(\boldsymbol{y}^{(\text{test})} - f_{\boldsymbol{\theta}}\left(\mathbf{x}^{(\text{test})}\right)\right)_i^2$$

  – 쿨백-라이블러 발산(Kullback-Leibler divergence, KL divergence)는 두 개의 확률분포 p와 q의 차이를 측정하기 위하여 사용되며 다음과 같이 정의함:

$$\text{D}_{\text{KL}}(p||q) \equiv \sum_{k=1}^{K} p_k \log\frac{p_k}{q_k}$$

2) 지도학습 기반 알고리즘

선형회귀(Linear Regression)

- 주어진 독립 변수 x와 종속 변수 y 사이에 선형 관계 $y \approx \beta_1 x + \beta_0$가 존재한다고 가정하는 방법
  - 훈련 데이터로부터 모델 $\hat{y} = \hat{\beta}_1 + \hat{\beta}_0$를 추정 시, 모델 추정을 위하여 잔차제곱합을 사용
  - n개의 훈련 데이터가 주어질 경우

    $\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$과 같이 계산되며, 잔차제곱합 최소화 위해 미분 시 선형 관계를 설명하는 계수 $\hat{\beta}_1$, $\hat{\beta}_0$는 다음과 같이 구해짐:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

    여기서 $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$, $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$.

로지스틱 회귀(Logistic Regression)

- 로지스틱 시그모이드(Logistic sigmoid) $\sigma(x) = \frac{1}{1+\exp(-x)}$는 입력을 0과 1사이로 매핑시키기 때문에 확률을 도입하여 논리를 전개하고 싶은 상황에서 사용됨
  - 로지스틱 시그모이드 함수를 활용한 회귀기법으로 예를 들어 타겟 카테고리가 0과 1만 존재하는 문제에 대하여 $p(X) = \Pr(Y = 1|X)$와 X의 관계를 모델링 할 때 사용됨
  - 보다 구체적으로 선형 관계를 이용하여 logistic function을 표현하면

    $p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$가 되며 동일 식을 변형하면 $\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$로 표현됨
  - 로그 연산자를 적용하면 $\log\frac{p(X)}{1-p(X)} = \beta_0 + \beta_1 X$로 정리되며, $\log\frac{p(X)}{1-p(X)}$ (logit이라고 표현)가 선형 관계로 표현될 수 있음을 알 수 있음

- 로지스틱 회귀에서는 회귀계수(regression coefficient) β_0와 β_1을 추정하기 위하여 최대우도법(maximum likelihood)을 활용

– $p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$를 이용하여 i번째 훈련 데이터를 확률로 표시할 경우 훈련의 목적은 출력 카테고리 0과 1에 대하여 모든 훈련 데이터의 확률 값을 최대로 하는 회귀계수를 찾는 것이며 이 때 확률 값의 우도(likelihood)는 다음과 같은 우도함수(likelihood function)으로 표현됨:

$$l(\beta_0,\beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1-p(x_{i'}))$$

– 회귀계수를 벡터 $\boldsymbol{\beta}$로 표시하고 최초 우도에 로그 연산자를 씌우면 새로운 우도는

$$l(\boldsymbol{\beta}) = \sum_{i}^{N} \{y_i \log p(x_i;\boldsymbol{\beta}) + (1-y_i)\log(1-p\log p(x_i;\boldsymbol{\beta}))\}$$

로 표시되며 각각의 회귀계수에 대하여 $\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_n} = 0$, $\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_i} = 0$을 계산하여 회귀계수 추정

다항회귀(Polynomial Regression)

- 선형회귀는 차수 1차의 선형함수를 활용하기 때문에 표현에 한계가 있는 반면, 다항 회귀(Polynomial regression)는 $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$를 최고 차수 d의 방정식으로 교체함:

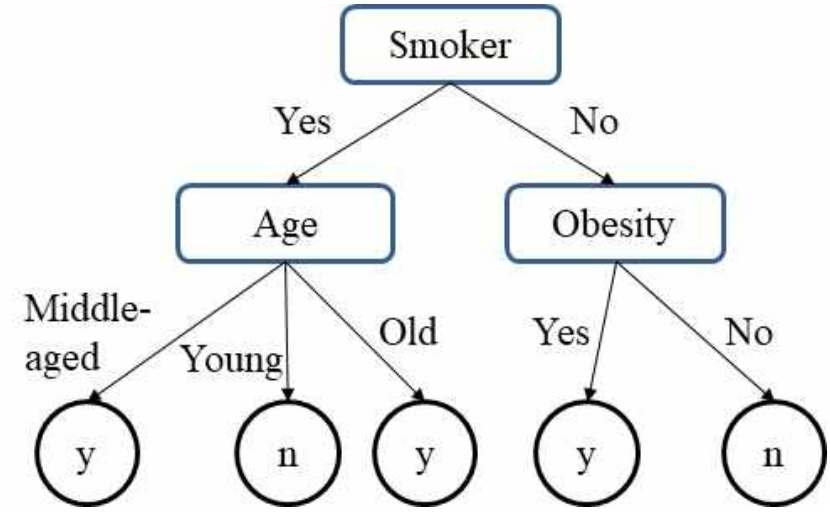
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \cdots + \beta_d x_i^d + \epsilon_i$$

– 여기서 ϵ_i는 오류를 표시하며, 최고 차수 d가 충분히 큰 다항식을 사용하면 극단적인 비선형 관계도 표시할 수 있음

– 그러나 최고 차수 d가 과도하게 클 경우 주어진 훈련 데이터를 과도하게 학습할 수 있기 때문에 지나치게 큰 차수를 사용하는 것은 적합하지 않음

의사결정트리(Decision Tree)

- 의사결정트리(Decision Tree)는 [그림3-2]와 같은 구조를 이용하여 탐색 공간을 중첩 분할한 후 분류나 회귀를 수행
  - 중첩 분할 과정에서 가장 유용한 특성(feature)부터 시작하여 공간을 분할하며 생성된 구조가 트리와 유사하기 때문에 의사결정 트리(decision tree)라고 함
  - 분류를 목표로 하는 의사결정트리는 말단 노드(leaf node)가 카테고리에 해당하며 루트 노드(root node)부터 말단 노드에 도달하는 경로를 구성하는 노드들은 데이터를 구성
  - 카테고리를 모르는 데이터가 들어왔을 때 루트 노드부터 시작하여 해당 특성의 특성값을 확인한 후 다음 노드로 이동하며 이와 같은 과정을 거쳐 도달한 말단 노드의 카테고리 값이 해당 데이터의 카테고리로 부여됨



[그림 3-2] 의사결정트리(Decision tree) 예시

- 의사결정트리에서는 노드에 특성을 할당하는 과정이 매우 중요하며, 특성의 선택에 따른 엔트로피의 변화를 판단의 근거로 사용
  - 의사결정트리의 방법 중 하나인 C4.5의 트리 구성 방법은 다음과 같으며, 훈련 데이터에 존재하는 카테고리의 수가 C라고 할 때 우선 전체집합의 엔트로피(Entropy)는 다음과 같이 정의:

$$\text{Entropy}(S) \equiv \sum_{i=1}^{c} -p_i \log_2 p_i$$

- p_i는 i번째 카테고리에 속한 데이터의 비율을 의미
- 의사결정트리를 구성하는 각 서브트리의 루트 부위에 속한 데이터의 집합을 S라고 간주했을 때 루트를 구성하는 특성으로는 해당 특성을 알게 됨으로써 발생하는 엔트로피 변화 값이 가장 큰 특성이 선택되며, 엔트로피 변화는 정보이득(information gain)*이라고 하며 다음과 같이 정의:

$$\text{Gain}(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entorpy(S_v)$$

* 선택된 특성을 통해 데이터를 보다 잘 분류(구분)하게 되는 것

- 의사결정트리 훈련 과정에서는 루트 노드부터 시작하여 각 노드에 해당하는 특성을 결정한 후 해당 노드의 가지(branch)에 속하는 서브트리의 루트 노드 특성을 정보이득을 이용하여 결정하는 방식으로 트리를 구성

서포트벡터머신(Support Vector Machine, SVM)

- 최대마진분류기(maximal margin classifier)라는 개념을 일반화시킨 분류 방법으로 서포트벡터머신에서 서포트벡터는 초평면(hyperplane)과 가장 가까운 데이터 인스턴스를 지칭
- 2개의 클래스 A와 B가 있을 때 문제 표현이 올바를 경우 같은 클래스에 속한 데이터 인스턴스는 서로 근접해서 모여있을 것이라고 예상할 수 있음
- 데이터가 p차원의 벡터로 표시될 경우 다음의 부등식을 만족시키는 초평면(hyper-plane)*을 이용하여 두 개의 클래스를 구분 가능:

$$\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_p x_p > 0$$

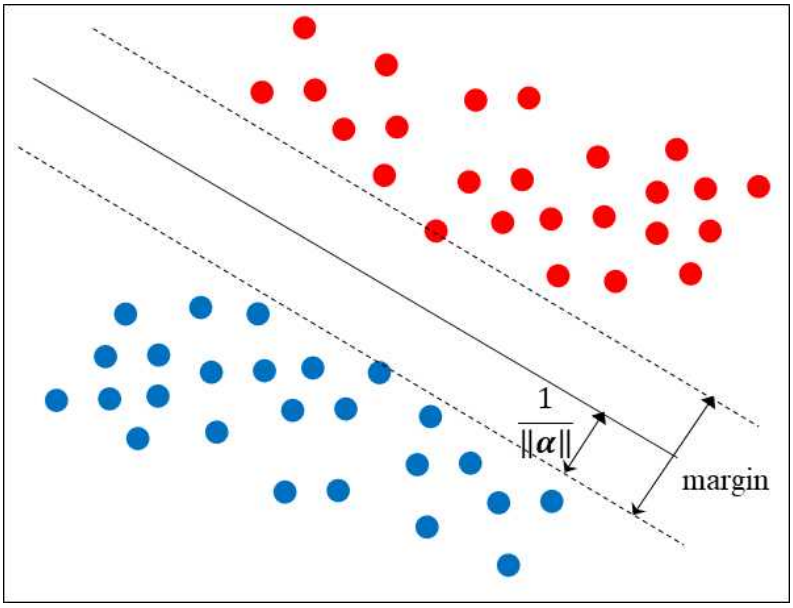
$$\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_p x_p < 0$$

* p차원 공간에서 클래스를 구분하는 $p-1$ 차원의 서브 공간으로 [그림3-3]에서 데이터를 구분하고 있는 직선이 해당

위의 부등식을 약간 수정하여 클래스 A의 레이블을 1, 클래스 B의 레이블을 −1로 표시하면 두 개의 클래스를 분류하는 초평면은 다음의 부등식을 만족시키는 초평면으로 표현할 수 있음:

$$y_i(\alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \cdots + \alpha_p x_{ip}) > 0$$

여기서 $i = 1, \ldots, |\mathcal{D}|$.



[그림 3-3] 서포트벡터머신(SVM) 예시

- 상기 그림과 같이 간단한 예에서도 두 개의 클래스를 분류하는 초평면의 수는 무한대이기 때문에 하나의 초평면을 특정할 방법이 필요
- 최대마진초평면(maximal margin hyperplane)은 클래스에 속한 데이터 인스턴스 중 초평면과 가장 가까운 원소의 거리가 최대가 되는 초평면을 의미
- 최대마진초평면을 선택하면 미지의 데이터에 대한 처리 가능성을 높일 수 있음
- 보다 구체적으로 찾으려는 초평면을 $\{\mathbf{x}: g(\mathbf{x}) = \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}\}$로 표현하였을 때 해당 초평면과 가장 가까운 원소간의 거리(간격) $\frac{y_i g(\mathbf{x}_i)}{\|\boldsymbol{\alpha}\|} = b$를 최대로 하는 $\boldsymbol{\alpha}$를 찾으면 최대마진초평면 설정 가능

- 이 때 발견된 초평면 특정을 위해 $b\|\boldsymbol{\alpha}\| = 1$이라는 제약을 도입, 거리를 $b = \frac{1}{\|\boldsymbol{\alpha}\|}$로 표시
- 원소와 초평면과의 거리 $\frac{1}{\|\boldsymbol{\alpha}\|}$를 최대화하기 위해서는 $\|\boldsymbol{\alpha}\|$를 최소화해야 하는데 이 때 예제와 같이 모든 데이터 인스턴스가 완벽하게 구분되는 상황은 현실적이지 못함
- 이에 잘못된 분류 가능성을 표현하기 위하여 $y_i g(\mathbf{x}_i) \geq 1 - \xi_i$와 같이 에러 텀 ξ_i를 도입하게 되는데 ξ_i의 도입과 함께 학습 목표는 모든 데이터 인스턴스에 대하여 $\min\|\boldsymbol{\alpha}\|$인 $\boldsymbol{\alpha}$를 찾는 것으로 변경됨
- 이 때 $\boldsymbol{\alpha}$는 모든 데이터 인스턴스에 대하여 $y_i g(\mathbf{x}_i) \geq 1 - \xi_i$ 를 만족하고 $\xi_i \geq 0$이며 $\sum \xi_i \leq C$ (C는 상수)라는 제약조건을 만족해야함
- 제약 조건이 존재할 때의 최적화 기법인 라그랑주 승수법(Lagrange multiplier)을 사용하면 학습 목표는 다음 함수 L의 최소화로 변경됨:

$$L(\boldsymbol{\alpha}, \boldsymbol{b}) \equiv \frac{1}{2}\|\boldsymbol{\alpha}\|^2 - \sum_{i=1}^{|\mathcal{D}|} b_i [y_i \boldsymbol{\alpha}^{\mathrm{T}} \mathrm{x}_i - 1]$$

- 이 함수로부터 다음과 같은 라그랑주 듀얼 객체함수(Lagrangian dual objective function) L_D을 획득하며 L_D를 계산하여 원하는 초평면 탐색 가능:

$$L_D = \sum_{i=1}^{|\mathcal{D}|} b_i - \frac{1}{2} \sum_{k,j}^{|\mathcal{D}|} b_k b_j y_k y_j \mathbf{x}_j^{\mathrm{T}} \mathbf{x}_k$$

- 커널함수(Kernel function)
  - 서포트벡터머신은 기본적으로 선형 모델이라는 한계가 존재
  - 선형 모델의 표현력을 확장시키기 위하여 특성 공간을 확장하는 방법을 취할 수 있는데, 서포트벡터머신에 이 아이디어를 적용할 경우 함수 L_D가 다음과 같이 변환되게 되며 이 때 특성 공간 확장에 사용되는 함수를 kernel function이라고 함

$$L_D = \sum_{i=1}^{|\mathcal{D}|} b_i - \frac{1}{2} \sum_{k,j}^{|\mathcal{D}|} b_k b_j y_k y_j \langle h(\mathbf{x}_k), h(\mathbf{x}_j) \rangle$$

- 다양한 커널함수를 이용하여 특성 공간을 확장할 수 있지만 $h(\mathbf{x}_k)$를 실제 계산하는 것은 어려울 수 있으며. 이러한 경우 서포트벡터머신은 커널트릭(kernel trick)을 사용

- 각각의 데이터 인스턴스를 새로운 특성 공간에서 확장하는 것이 아님
- $K(\mathbf{x},\mathbf{x}') = \langle h(\mathbf{x}_k), h(\mathbf{x}_j)\rangle$를 만족하는 kernel function을 이용하여 벡터간의 내적을 계산하는 것만으로 특성 공간에서의 계산 효과를 대체, 대표적인 커널함수는 다음과 같음:
  a. d차 다항 커널(polynomial kernel): $K(\mathbf{x},\mathbf{x}') = (1 + \langle \mathbf{x},\mathbf{x}'\rangle)^d$
  b. 레이디얼 베이시스(Radial basis): $K(\mathbf{x},\mathbf{x}') = \exp(-\|\mathbf{x}-\mathbf{x}'\|^2/c)$
  c. 신경망(Neural network): $K(\mathbf{x},\mathbf{x}') = \tanh(\kappa_1\langle \mathbf{x},\mathbf{x}'\rangle + \kappa_2)$

k-최근접 이웃(k-Nearest Neighbor, kNN)

- 기계학습에서는 관찰 데이터를 생성하는 확률분포의 존재를 가정한 후 해당 확률분포의 특성을 데이터로부터 추정하고자 함
  - 확률 분포의 형태를 사전에 가정한 후 분포를 설명하는 인자를 찾으려고 할 경우, 정확하게 분포의 형태를 추정할 수 있다면 문제가 없으나, 대부분의 경우 연구자가 채택한 분포와 실제 데이터의 분포 형태에는 큰 차이가 있음
  - k-최근접이웃방법에서는 구체적인 분포의 형태를 지정하지 않은 상태에서 n개의 훈련 인스턴스(training instance)로부터 생성 분포를 추정하기 위하여 언노운 인스턴스(unknown instance) $\mathbf{x}'$ 로부터 가장 가까운 k개의 훈련 인스턴스를 포착할 때까지 셀(cell)의 크기를 키워나간 후, 셀에 포함된 훈련 인스턴스의 카테고리 중 가장 많은 카테고리를 $\mathbf{x}'$의 카테고리로 할당하는 방식을 사용

나이브 베이즈 분류기(Naïve Bayes Classifier)

- 베이지안 이론(Bayes theorem)에서는 데이터($\mathcal{D}$)가 주어졌을 때 가설(h)의 확률을 계산할 수 있는 방법을 제공:

$$P(h|\mathcal{D}) = \frac{P(\mathcal{D}|h)P(h)}{P(\mathcal{D})}$$

  - $P(h|\mathcal{D})$는 사후확률(posterior probability), $P(\mathcal{D}|h)$는 우도(likelihood), $P(h)$는 사전확률(prior probability)에서,
  - $P(h)$는 데이터를 관찰하기 전에 가설 h에 주어진 믿음 혹은 신뢰성의 정도라고 해석할 수 있으며, $P(\mathcal{D}|h)$는 데이터를 관찰함으로써 가설 h에 부여되는 가중치
  - 따라서 사후확률 $P(h|\mathcal{D})$는 데이터를 관찰한 결과 가설 h에 가지게 되는 확신/신뢰도를 표시한다고 해석할 수 있음
  - 여러 개의 가설 집합 H가 존재할 때 사후확률이 가장 큰 가설을 선택하는 것이 합리적인 결정일 수 있는데, 이를 최대사후확률(maximum a posteriori, MAP)이라 정의:

$$h_{MAP} \equiv \underset{h \in H}{\operatorname{argmax}} P(h|\mathcal{D}) = \underset{h \in H}{\operatorname{argmax}} \frac{P(\mathcal{D}|h)P(h)}{P(\mathcal{D})}$$

  - 개별 데이터가 d개의 특성으로 설명되는 데이터($\langle a_1, \ldots, a_d \rangle$)인 경우 해당 데이터가 속할 수 있는 카테고리 집합을 V로 표시하면 가장 타당한 카테고리는 v_{MAP}, 즉 최대 사후확률 값을 최대로 하는 카테고리가 되어야 함:

$$v_{MAP} \equiv \underset{v_j \in V}{\operatorname{argmax}} P(v_j|a_1, \ldots, a_d) = \underset{v_j \in V}{\operatorname{argmax}} \frac{P(a_1, \ldots, a_d|v_j)P(v_j)}{P(a_1, \ldots, a_d)}$$

  - 일반적으로 결합확률분포(joint probability) $P(a_1, \ldots, a_d|v_j)$을 계산하기 위해 많은 데이터가 요구되나 나이브 베이즈 분류기에서는 목표 값(target value)이 주어졌을 때 특성 값이 서로 조건부 독립이라고 가정하여 결합확률을 간단하게 계산
  - 그 결과 $P(a_1, \ldots, a_d|v_j) = \prod_i P(a_i|v_j)$로 계산이 단순해지며 타겟 카테고리는 다음의 공식을 이용하여 결정:

$$v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \prod_i P(a_i|v_j)$$

학습 모델 평가 및 개선 방법

교차검증(Cross-validation)

- 학습 과정에서 활용할 수 있는 기법 중 하나로, 기계학습의 훈련 과정에서는 훈련 데이터를 이용하여 학습 커브를 생성하지만 태스크 해결 과정에서 더욱 중요한 것은 훈련 과정에서 접하지 못한 테스트 데이터에 대하여 훈련된 모델이 보이는 성능임
- 테스트 데이터가 충분할 경우 테스트 성능의 측정이 문제되지 않으나, 일반적인 경우에서는 만족할만한 규모의 데이터를 확보하는 것이 쉽지 않으므로 데이터의 부족을 극복하기 위해 교차검증(Cross-validation)을 활용

검증집합(Validation set) 접근법

- 최고 차수가 d인 다항 회귀를 이용하여 주어진 입력에 대응하는 실수 값을 예측하려면 서로 다른 d값의 다항식을 생성 후 훈련 데이터에 기반하여 다항식의 계수를 추정해야 함
- 이 때 검증 집합(validation set) 접근법에서는 우선 가용한 훈련 데이터를 무작위로 훈련 집합(training set) 검증 집합으로 구분 후, 훈련집합을 이용하여 계수를 예측, 검증 집합을 이용하여 서로 다른 d값의 다항식 모델에 대한 회귀 성능을 측정

검증집합(Validation set)과 과적합(Overfitting)

- 기계학습에서는 훈련집단(training set)과 테스트집단(test set)이 서로 동일한 분포에서 독립적으로 샘플링되었다고 간주하나 동일한 데이터 집합을 사용하지 않는 이상 관측 오류 등으로 인하여 훈련집단에 목표 분포에 일반적이지 않은 특이 분포가 섞일 가능성이 있음
- 그 결과, 가능한 가설이 모두 모인 가설 공간 H에 속한 가설 h와 h'에 대하여 훈련 집단에 대해서는 가설 h의 성능이 h' 보다 뛰어나지만 테스트 집단에 대해서는 가설 h의 성능이 h' 보다 저조한 현상(과적합(overfitting)) 발생 가능
- 검증집합은 과적합 문제 해결에도 사용 가능하며, 훈련집단을 이용하여 모델을 훈련하는 과정에서 계속하여 검증집단에 대한 성능을 측정하면서, 검증집단의 성능은 약화되는 반면 훈련집단의 성능이 계속해서 개선되는 현상이 발견되는 시점에서 훈련을 중단

- 리브원아웃 교차검증(Leave-one-out cross validation, LOOCV)
  - n개의 데이터 집합이 주어졌을 때 1개의 데이터를 검증 목적으로 사용하고 나머지 $n-1$개의 데이터를 훈련에 사용
  - 평균제곱근에러(MSE)를 성능 측도로 사용 시 (x_1, y_1)을 제외한 데이터가 훈련에 사용되면 $MSE_1 = (y_1 - \hat{y}_1)^2$로 정의
  - 유사한 방식으로 (x_i, y_i)가 훈련에서 제외되었을 경우에 대한 $MSE_i = (y_i - \hat{y}_i)^2$를 정의하면 주어진 데이터에 대한 성능은 다음과 같이 측정됨:

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{MSE}_i$$

- k배 교차검증(k-fold cross-validation, k-fold CV)
  - 주어진 훈련 데이터를 k개의 그룹으로 분할 후 1개의 그룹은 검증집단으로 활용하고 나머지 k-개의 그룹은 훈련집단으로 사용하여 모델을 훈련
  - 검증 그룹에 대하여 훈련된 모델의 평균제곱근에러(MSE)를 측정하면 총 k개의 평균제곱근에러가 도출되게 되며 k배 교차검증의 성능은 다음 식을 통해 측정됨:

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k} \mathrm{MSE}_i$$

- 부트스트랩(Bootstrap)
  - 추정의 정확성을 높이기 위하여 훈련집단 $\mathcal{D}$로부터 n개의 데이터를 무작위로 복원선택하는 과정을 B번 반복한 후, 그 결과 생성된 B개의 데이터 집합을 이용하여 통계치 θ를 다음과 같은 방식으로 추정:

$$\hat{\theta}^{*(\cdot)} = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}^{*(b)}$$

  - $\hat{\theta}^{*(b)}$는 부트스트랩 샘플(bootstrap sample) b를 이용하여 확보한 추정치를 의미

- 배깅(Bagging; Bootstrap aggregation)
  - 배깅(bagging)에서는 훈련집단(training set) $\mathcal{D}$로부터 n'개의 샘플($n' < |\mathcal{D}|$)을 복원추출하는 과정을 거쳐 B개의 훈련집단을 새롭게 생성한 후 각각의 훈련집단에 대하여

모델을 학습

- 마지막으로 B개의 모델의 예측 성능을 통합, 다음과 같이 최종 결과 획득:

$$\hat{f}_{\text{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x)$$

$\hat{f}^{*b}(x)$는 b번째 Bootstrap data 집합에서 획득한 모델의 예측 성능을 의미

- 부스팅(Boosting)
  - 여러 개의 분류기(classifier)를 학습한 후 개별 분류기의 예측 결과를 종합하여 개별 분류기 보다 우수한 예측 성능을 달성하는 것을 목표로 함
  - 훈련과정에서는 현재의 분류기 관점에서 가장 정보량이 많은 부분집합을 이용하여 새로운 분류기를 훈련함
  - 전체 훈련데이터(training data) $\mathcal{D}$로부터 $n_1 < |\mathcal{D}|$개의 훈련데이터를 무작위로 비복원 추출한 후 ($\mathcal{D}_1$) $\mathcal{D}_1$을 이용하여 첫 번째 classifier $\mathcal{C}_1$을 훈련시킴
  - 이후 훈련집단(training set) $\mathcal{D}_2$는 1/2의 확률의 사건 A가 발생하면 $\mathcal{D} - \mathcal{D}_1$으로부터 샘플을 하나씩 선택한 후 $\mathcal{C}_1$이 분류에 실패한 샘플을 $\mathcal{D}_2$에 추가하며 A^c가 발생하면 다시 $\mathcal{C}_1$이 분류에 성공한 샘플을 추가하는 방식으로 $\mathcal{D}_2$를 구성한 후 $\mathcal{D}_2$를 이용하여 새로운 분류기 $\mathcal{C}_2$를 훈련
  - $\mathcal{D}_1$,$\mathcal{D}_2$에 속하지 않은 잔여 샘플들에 대하여 $\mathcal{C}_1$과 $\mathcal{C}_2$의 예측결과가 틀린 샘플을 모아 $\mathcal{D}_3$를 구성한 후 $\mathcal{D}_3$로 새로운 분류기 $\mathcal{C}_3$를 훈련시킴

나. 비지도학습(Unsupervised learning)

1) 비지도학습의 개념

- 비지도학습(unsupervised learning)은 특성 벡터 $\mathbf{x}$에 대해 클래스 레이블 혹은 실수 값이 제공되는 지도학습(supervised learning)과 달리 특성 벡터 $\mathbf{x}$ 자체만 제공
  - 데이터에 존재하는 구조적인 특성을 학습하는 것을 목표로 하며, 보다 구체적으로 비슷한 데이터를 무리 짓는 클러스터링(군집화, clustering) 혹은 문제 공간 축소 등에서 활용
  - 비지도학습에서 데이터 구조를 학습하기 위해서는 데이터 그룹의 유사도에 대한 측도가 필요한데, 주로 사용되는 측도는 다음과 같음:
    - 오차제곱합기준(Sum-of-Squared-Error criterion): i번째 그룹 D_i의 평균벡터(mean vector) $\mathbf{m}_i = \frac{1}{n_i}\sum_{\mathbf{x}\in D_i}\mathbf{x}$에 대하여 다음과 같이 정의(여기서 c는 클러스터의 수):

$$J_e = \sum_{i=1}^{c}\sum_{\mathbf{x}\in D_i}\|\mathbf{x}-\mathbf{m}_i\|^2$$

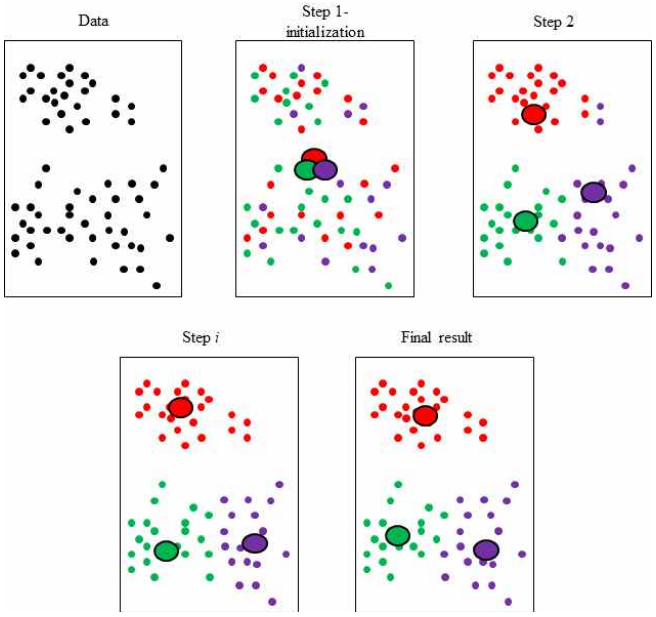
    - 산점도 행렬(Scatter matrix):

$$\boldsymbol{S}_T = \boldsymbol{S}_W + \boldsymbol{S}_B$$

로 정의되는데, $\boldsymbol{S}_W$는 클러스터 내 산점도 행렬(Within-cluster scatter matrix)를 의미하며 $\boldsymbol{S}_W = \sum_i^c \boldsymbol{S}_i$, $\boldsymbol{S}_i = \sum_{\mathbf{x}\in D_i}(\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^t$, $\boldsymbol{S}_B$는 클러스터 간 산점도 행렬(Between-cluster scatter matrix)로 $\boldsymbol{S}_B = \sum_{i=1}^{c}|D_i|(\mathbf{m}-\mathbf{m}_i)(\mathbf{m}-\mathbf{m}_i)^t$, $\mathrm{m} = \frac{1}{|D|}\sum_{i=1}^{c}|D_i|\mathbf{m}_i$로 정의

2) 비지도학습 기반 알고리즘

k-평균 클러스터링(k-means clustering)



[그림 3-4] k-평균 클러스터링의 동작

- k-평균 클러스터링(k-means clustering)은 클러스터의 개수(k)가 주어졌을 때, 클러스터 분산 합(total within cluster variance) 최소화를 목표로 클러스터 센터를 계속하여 이동시키면서 주어진 데이터를 k개의 클러스터로 나눔
  - 무작위로 클러스터를 부여한 후 기댓값 최대화(Expectation-Maximization) 방법을 이용하여 목표 측도를 최소화시키는 그룹을 탐색(상기 그림에 표현된 원은 클러스터 중심을 의미)

- 이 때 $\boldsymbol{S}_W$의 최소화를 목표로 하게 되며. 주어진 수식을 그대로 사용할 경우 규모가 큰 클러스터가 불리해지므로 $\boldsymbol{S}_W = \sum_i^c \frac{1}{|D_i|}\boldsymbol{S}_i$, $\boldsymbol{S}_i = \sum_{\mathbf{x}\in D_i}(\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^t$로 수정한 후 k개의 클러스터는 다음 수식을 만족하는 클러스터로 결정

$$C_1,\dots,C_k = \underset{C_1,\dots,C_k}{\operatorname{argmin}} \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{\mathbf{x}\in C_k} (\mathbf{x}-\mathbf{m}_k)(\mathbf{x}-\mathbf{m}_k)^t$$

기댓값 최대화(Expectation-Maximization, EM)

- m개의 데이터 $\mathbf{X}=\{\mathbf{x}_1,\ldots,\mathbf{x}_m\}$이 관찰되었고 미관찰데이터 $\mathbf{Z}=\{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_m\}$이 존재하면 전체 데이터는 $\mathbf{Y}=\mathbf{X}\cup\mathbf{Z}$로 표현할 수 있음
- $\mathbf{Z}$를 미관찰 인자 $\boldsymbol{\theta}$에 의해 설명되는 확률분포를 따르는 확률변수(random variable)이라 간주하면 $\boldsymbol{\theta}$의 현재 값을 h, 개정 값을 h'로 표현하면서 반복에 의해 주어진 조건을 만족하는 h를 찾을 수 있음
- 기댓값 최대화(EM)에서는 $\mathrm{E}[\ln P(Y|h')]$를 최대화하는 h'을 찾음으로써 최대우도(maximum likelihood) 가설 h'을 찾게 됨
- 전체 데이터 $\mathbf{Y}$는 미관찰 인자 $\boldsymbol{\theta}$에 의해 결정되기 때문에 $\mathbf{Y}$의 분포를 정확하게 알 수 없어, 기댓값 최대화에서는 $\boldsymbol{\theta}$에 대한 현재 추정 값(h)이 데이터를 생성했다고 가정 후 데이터를 생성하는 분포를 찾아가며, 이를 위해 다음과 같이 Q함수를 정의:

$$\mathrm{Q}(h'|h)=E[\ln p(Y|h')\,|h,\mathbf{X}]$$

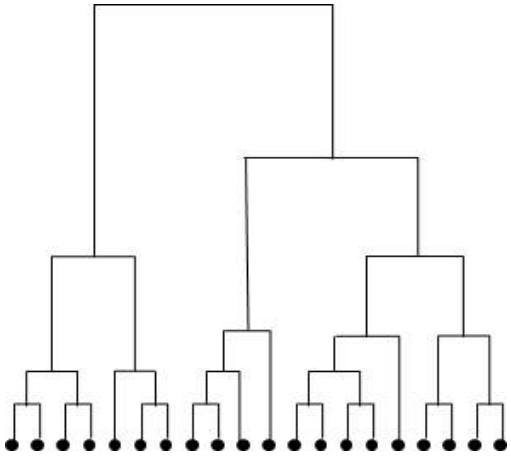
  - Step 1: 기대(expectation) 단계로 현재 가설 h와 관찰된 데이터 $\mathbf{X}$를 이용하여 $\mathrm{Q}(h'|h)$를 계산한 후 전체 데이터 $\mathbf{Y}$에 대한 분포를 추정
  - Step 2: 최대화(maximization) 단계로 Q를 최대화시키는 h'으로 h를 대체:

$$h\leftarrow\underset{h'}{\operatorname{argmax}}\,\mathrm{Q}(h'|h)$$

종료 조건을 만족할 때까지 step1과 step2를 반복하면서 가설을 업데이트

계층적 클러스터링(Hierarchical clustering)

- k-평균 클러스터링은 클러스터의 개수(k)를 사전에 알고 있어야 한다는 제약이 존재하며 일반적인 경우에서는 클러스터의 개수를 사전에 파악하는 것 또한 쉽지 않음
- 계층적 클러스터링(Hierarchical clustering)은 데이터 인스턴스 간의 거리를 비교하여 가장 근접한 데이터 인스턴스 쌍을 하나의 그룹으로 묶음
- 새롭게 생성된 데이터 그룹의 거리를 비교하여 가장 가까운 그룹을 다시 하나로 묶는 과정을 모든 데이터 그룹이 하나의 그룹으로 통합될 때까지 반복
  - 개별 인스턴스의 거리는 비교적 쉽게 계산할 수 있지만, 그룹간의 거리 비교 혹은 그룹과 개별 데이터 인스턴스 간의 거리 비교는 수행하기 쉽지 않음
  - 이에 그룹 혹은 그룹과 인스턴스 간의 거리 비교에는 다음 중 한 가지 측도를 사용:
    - $d_{\min}(\mathcal{D}_i, \mathcal{D}_j) = \min_{\mathbf{x}\in\mathcal{D}_i, \mathbf{x}'\in\mathcal{D}_j} \|\mathbf{x}-\mathbf{x}'\|$
    - $d_{\max}(\mathcal{D}_i, \mathcal{D}_j) = \max_{\mathbf{x}\in\mathcal{D}_i, \mathbf{x}'\in\mathcal{D}_j} \|\mathbf{x}-\mathbf{x}'\|$
    - $d_{\text{avg}}(\mathcal{D}_i, \mathcal{D}_j) = \frac{1}{|\mathcal{D}_i||\mathcal{D}_j|}\sum_{\mathbf{x}\in\mathcal{D}_i}\sum_{\mathbf{x}'\in\mathcal{D}_j}\|\mathbf{x}-\mathbf{x}'\|$
    - $d_{\text{mean}}(\mathcal{D}_i, \mathcal{D}_j) = \|\mathbf{m}_i - \mathbf{m}_j\|$



[그림 3-5] 계층적 클러스터링으로부터 획득한 dendrogram

- 상기 네 가지 거리 측도는 데이터 그룹 $\mathcal{D}_i$와 $\mathcal{D}_j$의 거리 비교에 사용되는데 $d_{\min}$은 $\mathcal{D}_i$의 멤버 인스턴스 $\mathbf{x}$와 $\mathcal{D}_j$의 멤버 인스턴스 $\mathbf{x}'$의 가능한 모든 조합에 대하여 거리를 계산한 후 최소 거리를 $\mathcal{D}_i$와 $\mathcal{D}_j$의 거리로 설정하며, $d_{\max}$에서는 가능한 모든 조합에서 계산한 거리 중 최대 거리를 $\mathcal{D}_i$와 $\mathcal{D}_j$의 거리로 설정
  - $d_{\text{avg}}(\mathcal{D}_i, \mathcal{D}_j)$에서는 멤버 인스턴스 간의 모든 조합에 대하여 거리를 계산한 후 그 평균을 $\mathcal{D}_i$와 $\mathcal{D}_j$의 거리로 설정하며 $d_{\text{mean}}(\mathcal{D}_i, \mathcal{D}_j)$에서는 $\mathcal{D}_i$의 중심점 $\mathbf{m}_i$와 $\mathcal{D}_j$의 중심점 $\mathbf{m}_j$간의 거리를 $\mathcal{D}_i$와 $\mathcal{D}_j$의 거리로 설정

주성분 분석(Principal Component Analysis, PCA)

- 주성분 분석(Principal Component Analysis, PCA)는 실제 입력보다 낮은 차원의 표현 공간을 학습하는 방법으로 고유값(eigenvalue) 및 고유벡터(eigenvector)를 사용
  - 정사각행렬(square matrix) A에 대하여

$$\mathbf{A}\boldsymbol{v} = \lambda \boldsymbol{v}$$

관계를 만족하는 0벡터가 아닌 벡터 $\boldsymbol{v}$를 고유벡터라고 하며 scalar λ를 고유값으로 정의

  - 차원 축소: n차원의 문제 공간에서 표현된 데이터를 $l < n$인 l차원으로 축소하고자 할 때 $\mathbf{x} \in \mathbb{R}^n$인 $\mathbf{x}$를 $\mathbf{c} \in \mathbb{R}^l$인 $\mathbf{c}$로 표현하는 상황을 생각할 수 있으며 인코딩 함수 $f, f(\mathbf{x}) = \mathbf{c}$, 디코딩 함수 g $\left(g\text{의 조건 } g(f(\mathbf{x})) \approx \mathbf{c}\right)$를 도입하면 디코딩 함수 $g(\mathbf{c}) = \boldsymbol{D}\boldsymbol{c}$ $(\boldsymbol{D} \in \mathbb{R}^{n \times l})$를 정의할 수 있음
  - $\boldsymbol{D}$의 유도과정이 용이하도록 $\boldsymbol{D}$의 열이 서로 직교(orthogonal)하게 제약을 가한 후 원래 벡터 $\mathbf{x}$와 $g(\mathbf{c})$의 차이를 최소로 만드는 g를 찾으면 다음과 같음:

$$\boldsymbol{c}^* = \underset{\boldsymbol{c}}{\operatorname{argmin}} \|\mathbf{x} - g(\mathbf{c})\|_2$$

  - L^2 노름(norm)*에 제곱을 하면 최소화되어야 할 함수는

$$(\mathbf{x} - g(\mathbf{c}))^{\mathsf{T}}(\mathbf{x} - g(\mathbf{c})) = \mathbf{x}^{\mathsf{T}}\mathbf{x} - 2\mathbf{x}^{T}g(\mathbf{c}) + g(\mathbf{c})^{\mathsf{T}}g(\mathbf{c})\text{로 단순화}$$

* 유클리디안노름이라고도 하며 n차원의 유클리디안 노름공간에서 원점에서 벡터까지의 직선 거리를 의미

– $g(\mathbf{c})$를 포함하고 있는 부분만 추출하면

$$\begin{aligned} \boldsymbol{c}^* &= \underset{\mathbf{c}}{\operatorname{argmin}} -2\mathbf{x}^{\mathrm{T}} g(\mathbf{c}) + g(\mathbf{c})^{\mathrm{T}} g(\mathbf{c}) \\ &= \underset{\mathbf{c}}{\operatorname{argmin}} -2\mathbf{x}^{\mathrm{T}} \boldsymbol{D}\mathbf{c} + \mathbf{c}^{\mathrm{T}} \boldsymbol{D}^{\mathrm{T}} \boldsymbol{D}\mathbf{c} \\ &= \underset{\mathbf{c}}{\operatorname{argmin}} -2\mathbf{x}^{\mathrm{T}} \boldsymbol{D}\mathbf{c} + \mathbf{c}^{\mathrm{T}} \boldsymbol{I}_l \mathbf{c} \\ &= \underset{\mathbf{c}}{\operatorname{argmin}} -2\mathbf{x}^{\mathrm{T}} \boldsymbol{D}\mathbf{c} + \mathbf{c}^{\mathrm{T}} \mathbf{c} \end{aligned}$$

여기서 $\mathbf{c}$는 $\mathbf{c} = \boldsymbol{D}^{\mathrm{T}}\mathbf{x}$로 유도되며 $f(\mathbf{x}) = \boldsymbol{D}^{\mathrm{T}}\mathbf{x}$, $r(\mathbf{x}) = g\big(f(\mathbf{x})\big) = \boldsymbol{D}\boldsymbol{D}^{\mathrm{T}}\mathbf{x}$가 되며, D의 유도는 다음 과정을 거침

$$\boldsymbol{D}^* = \underset{\boldsymbol{D}}{\operatorname{argmin}} \sqrt{\sum_{i,j} \left(\mathbf{x}_j^{(i)} - r(\mathbf{x}^{(i)})_j\right)^2} \quad (\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D} = \boldsymbol{I}_l)$$

l=1인 상황에서 문제를 풀기 시작하면 D가 아니라 단일 벡터 $\boldsymbol{d}$,

$\boldsymbol{d}^* = \underset{\boldsymbol{d}}{\operatorname{argmin}} \sum_i \left\|\mathbf{x}^{(i)} - \mathbf{x}^{(i)\mathrm{T}} \boldsymbol{d}\boldsymbol{d}^{\mathrm{T}}\right\|_2^2 \ (\|\boldsymbol{d}\|_2 = 1)$를 찾는 문제로 변형할 수 있음. 여기서 주어진 m개의 데이터 인스턴스를 모두 반영하는 행렬 X ($\mathbf{X} \in \mathbb{R}^{m \times n}$, $\mathbf{X}_{i,:} = \mathbf{x}^{(i)\mathrm{T}}$)를 설정하면, 다음 식을 통해 원하는 행렬d를 확보할 수 있음:

$$\boldsymbol{d}^* = \underset{\boldsymbol{d}}{\operatorname{argmin}} \|\mathbf{X} - \mathbf{X}\boldsymbol{d}\boldsymbol{d}^{\mathrm{T}}\|_F^2 \ \ (\boldsymbol{d}^{\mathrm{T}}\boldsymbol{d} = 1)$$

– 주어진 식을 풀어 쓰면

$$\boldsymbol{d}^* = \underset{\boldsymbol{d}}{\operatorname{argmin}} -2\mathrm{Tr}(\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{d}\boldsymbol{d}^{\mathrm{T}}) + \mathrm{Tr}(\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{d}\boldsymbol{d}^{\mathrm{T}}\boldsymbol{d}\boldsymbol{d}^{\mathrm{T}}) \ \ (\boldsymbol{d}\boldsymbol{d}^{\mathrm{T}} = 1)$$

– 와 같고, 이는 고유값 분해(eigen-decomposition)을 통해 고유값 크기에 맞춰 상응하는 $\mathbf{X}^{\mathrm{T}}\mathbf{X}$의 고유벡터 선택 시 차원 축소 가능

- 주성분 분석(PCA)의 해석

– 주성분 분석은 데이터가 가장 많이 변하는 축으로 데이터를 프로젝션하는 방법

– 유도 과정을 거쳐 확보한 상위 k개의 고유값에 대응하는 고유벡터들을 데이터의 변화 정도가 큰 축에서 작아지는 축으로 정렬된 축의 집합이라고 해석 가능

– 혹은 주성분(principal component)을 데이터와의 거리의 합이 가장 작은 축이라고 해석 가능

Independent Component Analysis (ICA)

- PCA는 표현 공간(특성 공간)에서 데이터를 가장 잘 표현하는 방향을 찾으려는 접근법이나, ICA는 가장 독립적인 방향을 찾으려는 접근법임
- ICA에서는 d개의 독립된 소스로부터 생성된 데이터 $\mathbf{x} \in \mathbb{R}^d$가 존재할 때, 관찰된 데이터 $\mathbf{s} = \mathbf{A}\mathbf{x}$는 mixing matrix $\mathbf{A}$에 의해 변화가 발생한 상태라고 가정함
  - $\{\mathbf{s}^{(i)} : i = 1, \ldots, m\}$의 데이터가 관찰되었을 때 ICA에서는 데이터를 생성한 소스 $x^{(i)}$를 복원하고자 함
  - $\mathbf{W} = \boldsymbol{A}^{-1}$를 unmixing matrix라고 하면 $x^{(i)} = \mathbf{W}s^{(i)}$를 계산하여 관찰한 데이터로부터 원래 데이터를 복원할 수 있음
  - 각 소스 x_i의 데이터가 density p_x에 의해 생성되었다고 가정하면 소스 x의 joint distribution은 다음 식과 같이 표현됨:

$$p(\mathbf{x}) = \prod_{i=1}^{d} p_x(x_i)$$

관찰된 데이터 s에 대하여 복원된 데이터를 y라고 표현하면

$$p_{\mathbf{y}}(\mathbf{y}) = \frac{p_{\mathbf{x}}(\mathbf{s})}{|\mathbf{J}|}$$

여기서

$$\mathbf{J} = \begin{pmatrix} \frac{\partial y_1}{\partial s_1} & \cdots & \frac{\partial y_d}{\partial s_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_1}{\partial s_d} & \cdots & \frac{\partial y_d}{\partial s_d} \end{pmatrix}, |\mathbf{J}| = \left| |\mathbf{W}| \prod_{i=1}^{d} \frac{\partial y_i}{\partial s_i} \right|$$

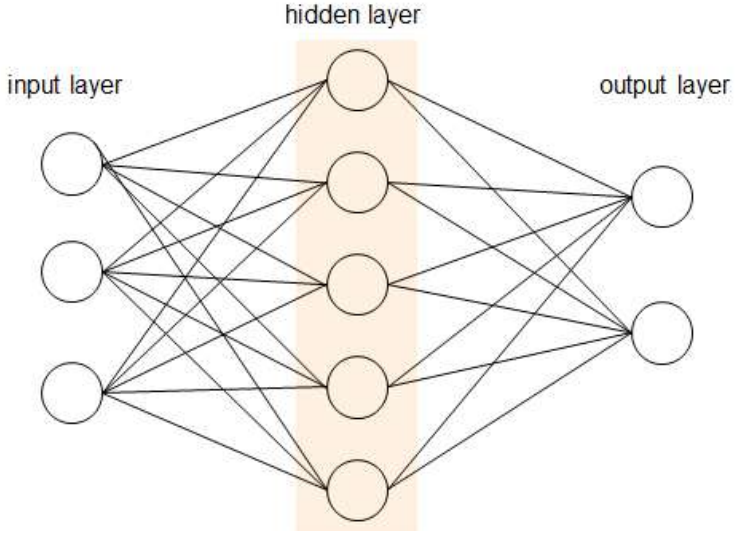
  - 데이터 복원 과정을 source signal의linear transform으로 해석할 경우
    $\mathbf{y} = f[\mathbf{W}\mathbf{s} + \mathbf{w}_0]$ (f는 보통 sigmoid function 사용)로 표현할 수 있는데 W와 $\mathbf{w}_0$를 구하기 위하여 joint entropy를 활용함:

$$\boldsymbol{H}(\mathbf{y}) = -E[\ln p_{\mathbf{y}}(\mathbf{y})] = E[\ln|\mathbf{J}|] - E[\ln p_{\mathbf{s}}(\mathbf{s})]$$

다. 딥 러닝(Deep Learning Network)

1) 딥 러닝의 개념

- 딥 러닝은 기본적으로 인공 신경망(Artifical Neural Netwokr, ANN)에 기반을 둔 기술로 인공신경망 개념 및 알고리즘 자체는 상당한 역사를 가지고 있음
  - 인공신경망은 인간의 뇌가 신경세포(Neuron) 간 조직적으로 연결된 구조임을 착안, 이를 모방한 것으로 다음과 같이 입력계층(input layer), 은닉계층(hidden layer), 출력계층(ouput layer)로 구성



[그림 3-6] 인공 신경망(Artificial Neural Network) 예시

- 딥러닝은 인공신경망의 은닉계층을 심층화하고 학습과정을 통해 각 노드 간 연결된 엣지 할당된 가중치를 탐색하는 과정을 의미
  - 인공신경망은 각 노드는 이 전 노드에서 보내진 데이터 처리 후 다음 노드로 전달하는 방식으로 작동
  - 전달 과정에서 엣지에 할당된 가중치가 데이터에 반영되어 처리되는데 딥러닝을 통해 적절한 가중치를 학습하게 됨

- 딥러닝은 입력된 데이터의 특징을 스스로 학습하는 비지도학습 방법을 기반으로 작동하며 이미지, 음성인식 등에 활용
  - 딥러닝의 기본이 되는 인공신경망 알고리즘은 학습 및 결과값 도출에 방대한 연산이 요구되어 과거 활용에 제약이 있었음
  - 2000년대 이후 GPU(Graphic Process Unit) 등 컴퓨터 하드웨어 파워의 비약적 발전으로 인공신경망 알고리즘의 활용성 강화
    - NVIDIA社는 고효율 연산을 위한 GPU 병렬처리 알고리즘을 손쉽게 구현할 수 있도록 CUDA(Compute Unified Device Architecture)라는 GPGPU 기술을 개발
    - 방법론적 측면에서도 학습 시 요구되는 방대한 연산량을 개선하는 알고리즘이 개발
  - 가중치를 학습하는 과정에서 데이터의 특징을 추출하여 학습함으로써 주어진 데이터 클래스 레이블이나 실수값을 바탕으로 학습하는 지도학습과 차별됨
  - 2012년 세계 최대의 이미지 인식 기술대회(Imagenet Large Scale Visual Recognition Challenge)에서 딥러닝 적용 팀이 압도적 점수차로 우승하며 유용성 입증

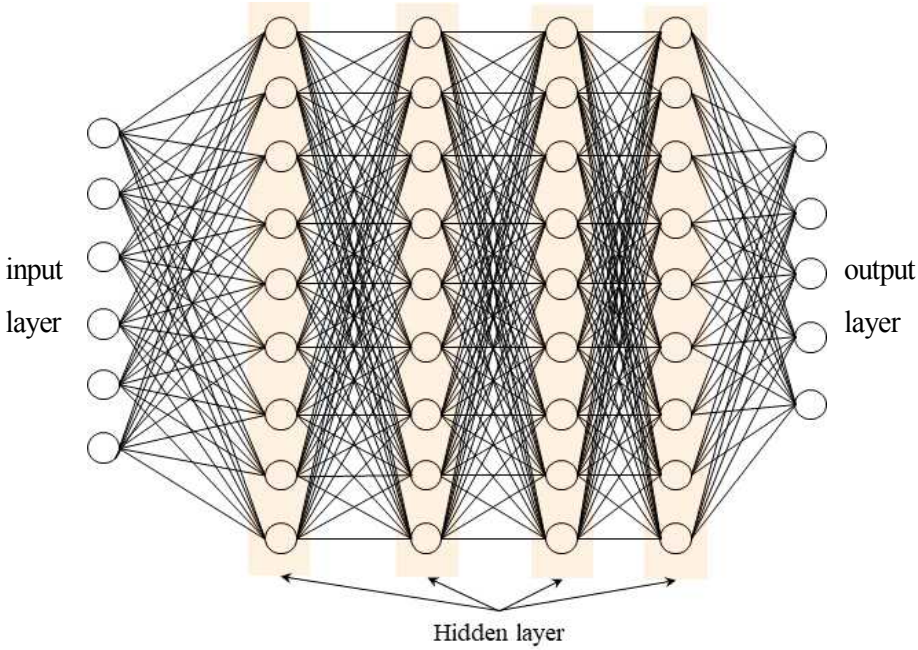
2) 딥러닝 기반 알고리즘[13)]

- 심층 피드포워드 네트워크(Deep feedforward network)
  - 심층 피드포워드 네트워크(Deep feedforward network), 피드포워크 신경망(feedforward neural network), 다층 퍼셉트론(multilayer perceptron)은 모두 같은 개념
  - 기본적으로 인공신경망 구조이므로 여러 개의 노드(node)가 모인 레이어(layer)와, 노드와 노드를 연결하는 엣지(edge)로 구성되며 각 엣지에는 가중치가 부여됨
  - 노드는 하위 계층*의 노드에서 계산한 결과를 입력으로 받아 규정된 연산을 수행한 후 계산 결과를 상위 계층의 노드로 전달

    * 심층 피드포워드 네트워크는 입력계층(input layer)에서 은닉계층(hidden layer)을 거쳐 출력계층(output layer)으로 정보가 흐른다고 간주

13) I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, The MIT Press, 2016. 발췌 인용

- 그 과정에서 목표 함수 $f^*(\mathbf{x})$를 근사하는 것을 목표로 함
- 보다 구체적으로 피드포워드 네트워크에서는 $\mathbf{y} = f(\mathbf{x};\boldsymbol{\theta})$의 매핑 관계를 정의하며 $\boldsymbol{\theta}$는 가장 우수한 성능의 근사 결과를 생성



[그림 3-7] 심층 신경망(Deep Neural Network) 예시

- 피드포워드 네트워크는 비순환 방향 그래프(acyclic directed graph)로 표현되며 여러 개의 함수가 체인구조로 연결
  - 예를 들어 3개의 함수 $f^{(1)}, f^{(2)}, f^{(3)}$가 체인으로 연결될 경우 $f(\mathbf{x}) = f^{(3)}\left(f^{(2)}\left(f^{(1)}(\mathbf{x})\right)\right)$로 표시되며 $f^{(1)}$를 첫 번째 계층, $f^{(2)}$를 두 번째 계층을 의미
  - 체인의 길이를 “깊이(depth)”라고 하며 체인의 최종 레이어를 출력계층(output layer)으로 정의
  - 입력 계층과 출력 계층 사이의 계층들의 경우 훈련 데이터가 적합한 출력 값을 제공해주지 않으며 각 계층에는 여러 개의 노드가 존재할 수 있는데, 동일 계층에 존재하는 노드의 수를 “폭(width)”이라고 함

- 피드포워드 네트워크를 구성하는 개별 노드는 벡터를 입력으로 받아 스칼라(scalar) 값을 출력하는데, 구체적인 함수는 모델 표현력 및 계산 용이성을 고려하여 결정

$$w_i = \text{sign}(w_i^*)\max\left\{|w_i^*| - \frac{\alpha}{H_{i,i}}, 0\right\}$$

  획득된 솔루션에서 일부 파라미터의 경우 최적의 값이 0이 되기 때문에 특성 선택 방법으로도 응용 가능

- 심층 피드포워드 네트워크 및 심층 신경망은 노드에서 비선형 함수를 활용하기 때문에 선형 모델의 최적화 접근법을 사용할 수 없으며, 따라서 그레디언트(gradient)에 기반한 반복 계산을 통해 목표 함수를 개선시키는 방향으로 $f^*(\mathbf{x})$를 근사함

  – 딥러닝에서 모델은 분포 $p(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})$를 표현하게 되며 최대우도(maximum likelihood)를 만족하는 파라미터 $\boldsymbol{\theta}$를 찾음

컨볼루션 신경망(Convoluation Neural Network, CNN)

- 컨볼루션 신경망(CNN)은 이미지와 같이 격자형의 토폴로지를 가지고 있는 데이터 처리를 위한 신경망 구조 네트워크를 지칭

  – 최소 하나의 계층에서 일반적인 행렬 곱셈 대신 컨볼루션(convolution) 연산을 사용

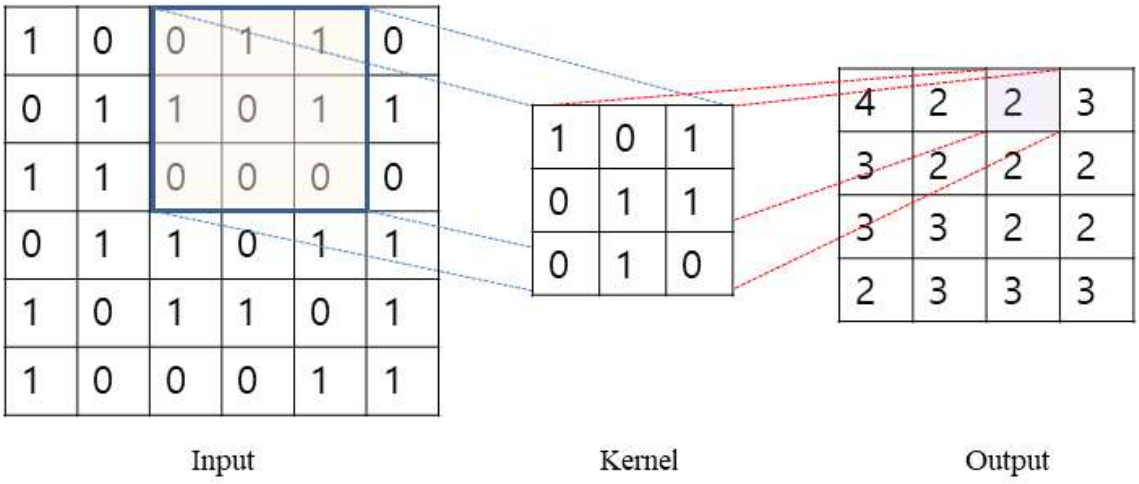
- 컨볼루션 오퍼레이션(convolution operation) 시 입력행렬(input matrix)와 커널행렬(kernel matrix)의 대응 셀의 값을 곱한 후 더한 결과가 출력행렬(output matrix)의 셀 값이 되며 다음과 같은 형태를 지님:

$$s(t) = (x * w)(t)$$

- 컨볼루션 신경망에서는 function x를 input, function w를 kernel이라고 하며, 출력을 feature map이라고 지칭

  – 정수 t에 대하여 이산 컨볼루션(discrete convolution)은 다음과 같이 정의

$$s(t) = (x * w)(t) = \sum_{-\infty}^{\infty} x(a)w(t-a)$$



[그림 3-8] 컨볼루션 오퍼레이션(Convolution operation) 예시

- 컨볼루션 신경망이 2차원 데이터에 적용될 경우 커널 K(kernel K)에 대한 컨볼루션 오퍼레이션은 다음과 같이 정의:

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n)$$
$$= (K * I)(i,j) = \sum_m \sum_n I(i-m, j-n)K(m,n)$$

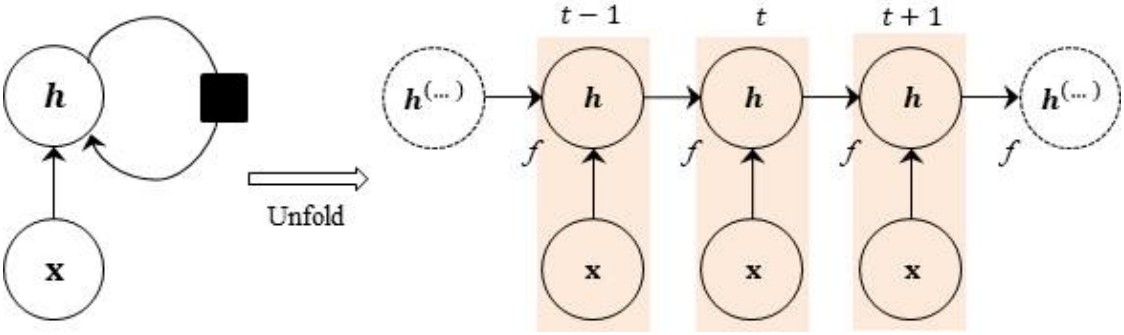
  - 일반적인 신경망에서는 선형 변환(linear transformation)을 위하여 행렬 연산을 사용하는데 이 경우 모든 출력 유닛이 모든 입력 유닛과 관련을 맺게 됨
  - 이에 반하여 컨볼루션 신경망에서는 입력보다 작은 커널을 사용하여 sparse interaction(상호작용) 혹은 sparse connectivity(연결성)를 구현함

순환 신경망(Recurrent Neural Network, RNN)

- 순환 신경망(RNN)은 순차적인 데이터 처리에 특화된 신경망으로, 인자 공유(parameter sharing)이라는 아이디어를 이용, 특정되지 않은 길이의 시퀀스 처리가 가능

$$\boldsymbol{h}^{(t)} = f\left(\boldsymbol{h}^{(t-1)}, \mathbf{x}^{(t)}; \boldsymbol{\theta}\right)$$

- 순환 신경망은 상기와 같은 식을 통해 은닉상태(hidden state)를 계산하는데, 인자 $\boldsymbol{\theta}$에 의하여 동작이 규정되며 현재 타임스텝(time step) t에서의 입력 $\mathbf{x}^{(t)}$와 이전 타임스텝 t−1에서의 은닉상태 $\boldsymbol{h}^{(t-1)}$가 현재 상태(state) $\boldsymbol{h}^{(t)}$의 값을 결정



[그림 3-9] 출력이 없는 순환 신경망(RNN) 예시

- RNN으로 과거 입력 정보에 기반하여 미래 값을 예측하는 상황은 $\mathbf{x}^{(t)}, \mathbf{x}^{(t-1)}, \mathbf{x}^{(t-2)}, \ldots, \mathbf{x}^{(2)}, \mathbf{x}^{(1)}$의 시퀀스를 $\boldsymbol{h}^{(t)}$로 매핑하는 상황으로 표현할 수 있으며 이 상황은 다음과 같은 형태로 표현 가능

$$\boldsymbol{h}^{(t)} = g^{(t)}\left(\mathbf{x}^{(t)}, \mathbf{x}^{(t-1)}, \mathbf{x}^{(t-2)}, \ldots, \mathbf{x}^{(2)}, \mathbf{x}^{(1)}\right) = f\left(\boldsymbol{h}^{(t-1)}, \mathbf{x}^{(t)}; \boldsymbol{\theta}\right)$$

  - $g^{(t)}$를 이용하면 과거 시퀀스를 모두 입력으로 받지만, 순환구조(recurrent structure)에서는 함수 f를 반복 적용하여 (1) 학습된 모델은 항상 동일한 크기의 입력을 받으며 (2) 동일한 인자를 공유하는 동일한 전이 함수 f를 매 타임스텝에서 사용 가능
  - 그 결과 모든 가능한 시퀀스에 대하여 $g^{(t)}$의 학습이 불필요

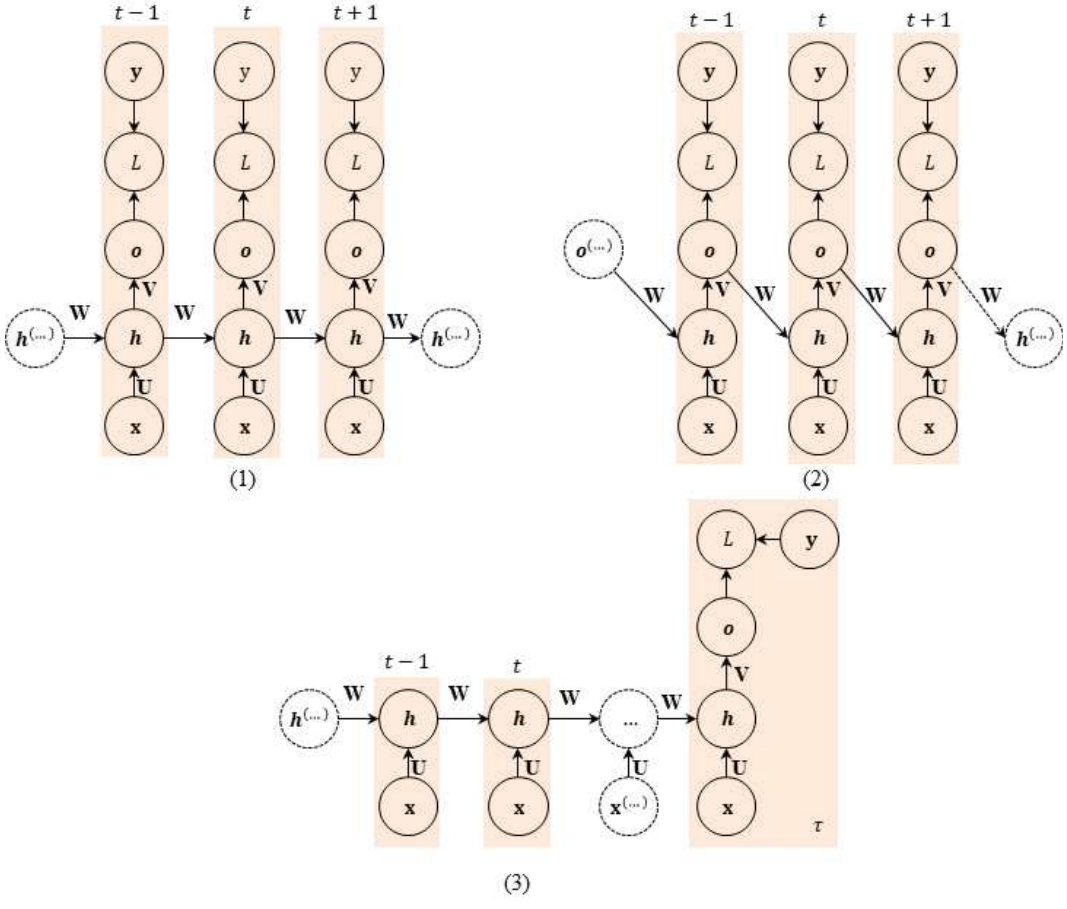
- 순환 신경망의 다음과 같은 주요 유형이 존재
  - (1) 매 타임스텝에서 출력을 생성하고 은닉유닛(hidden unit) 사이에 순환연결(re-

current connection)이 존재하는 경우

- (2) 매 타입스텝에서 출력을 생성하고 한 타임스텝의 출력과 다음 타임스텝의 은닉유닛 사이에 순환연결이 존재하는 경우
- (3) 은닉유닛 간에 순환연결이 존재하며 출력은 하나만 생성되는 경우

● 순환 신경망은 설계구조 상 학습 시 그레디언트 소실(gradient vanishing) 문제 발생

- 신경망이 곱하기 연산으로 이루어져있어 은닉값을 추적하는 과정에서 시간을 거슬러 올라갈수록 그레디언트가 소실되는 구조



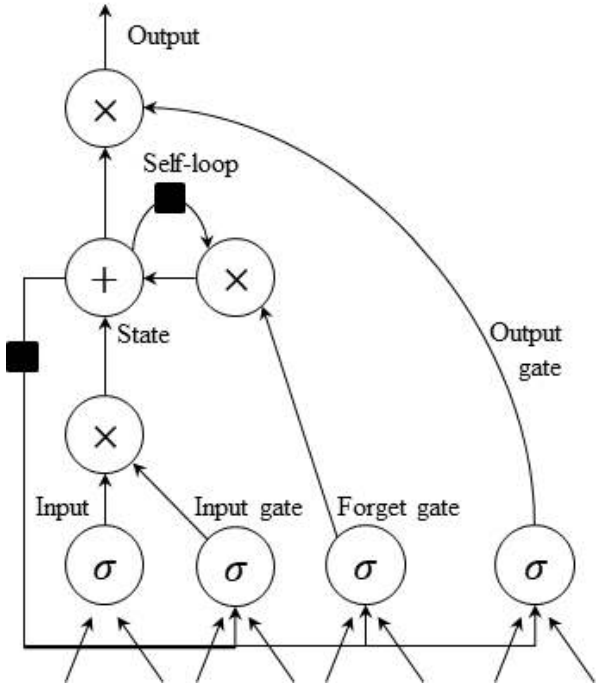
[그림 3-10] 순환 신경망(RNN) 유형

장단기 기억네트워크 (Long Short-Term Memory, LSTM)

- 장단기 기억네트워크(LSTM)은 게이트화된 순환 유닛(gated recurrent unit)을 활용하는 게이트화 순환 신경망(gated RNN)의 일종으로 그레디언트 소실 문제를 해결
  - 미분값이 소멸하지도 너무 커지지도 않는 경로(path)를 활용하는 순환 신경망으로, 연결 가중치가 매 타임스텝마다 변경될 수 있도록 구현
- 장단기 기억네트워크는 그레디언트가 계속해서 전달될 수 있는 셀프루프(self-loop)를 도입, 셀프루프 가중치가 컨텍스트에 의해 제어되도록 설정함으로써 성능을 향상
  - 각 셀은 일반적인 순환 네트워크(recurrent network)와 동일한 입력과 출력을 보유하나 정보의 흐름을 제어하는 게이트 유닛(gate unit)이 추가로 존재
- 장단기 기억네트워크 셀의 가장 중요한 컴퍼넌트는 상태유닛(state unit) $s_i^{(t)}$로 포겟게이트(forget gate) $f_i^{(t)}$의 제어를 받음(i는 i번째 셀, t는 타임스텝):

$$f_i^{(t)} = \sigma\left(b_i^f + \sum_j U_{i,j}^f x_j^{(t)} + \sum_j W_{i,j}^f h_j^{(t-1)}\right)$$

  - 여기서 $\mathbf{x}^{(t)}$는 현재 입력 벡터, $\boldsymbol{h}^{(t)}$는 현재 은닉계층 벡터로 모든 장단기 기억네트워크 셀의 출력을 포함
  - $\mathbf{b}^f, \mathbf{U}^f, \mathbf{W}^f$는 각각 해당 포겟게이트의 바이어스(bias), 입력가중치(input weight), 순환가중치(recurrent weight). σ는 시그모이드 유닛(sigmoid unit)을 의미



[그림 3-11] 장단기 기억네트워크 "셀(cell)"의 블록 다이어그램

- 장단기 기억네트워크 셀의 상태(state)는 다음과 같이 갱신:

$$s_i^{(t)} = f_i^{(t)} s_i^{(t-1)} + g_i^{(t)} \sigma\left(b_i + \sum_j U_{i,j} x_j^{(t)} + \sum_j W_{i,j} h_j^{(t-1)} \right)$$

  - 외부입력게이트(external input gate) $g_i^{(t)}$는 다음과 같이 계산됨:

$$g_i^{(t)} = \sigma\left(b_i^g + \sum_j U_{i,j}^g x_j^{(t)} + \sum_j W_{i,j}^g h_j^{(t-1)} \right)$$

- 장단기 기억네트워크 셀의 출력 $h_i^{(t)}$은 출력게이트(output gate) $q_i^{(t)}$의 값에 따라 제어될 수 있음:

$$h_i^{(t)} = \tanh\left(s_i^{(t)}\right) q_i^{(t)}$$

$$q_i^{(t)} = \sigma\left(b_i^o + \sum_j U_{i,j}^o x_j^{(t)} + \sum_j W_{i,j}^o h_j^{(t-1)} \right)$$

- 장단기 기억네트워크 구조에서는 해당 셀로의 새로운 입력 신호나 에러 신호가 없을 경우 입력게이트 및 출력게이트에 의해 상태 유지
- 게이트가 닫혀 있을 때(엑티베이션이 낮은 상태) 관계없는 신호가 셀 내부로 입력되지 않게 되므로 셀 상태가 다른 게이트 상태를 변경하지 않음
- 포겟 게이트가 없는 장단기 기억네트워크는 정보를 임의 시간 동안 기억할 수 있는데 연속된 입력 스트림이 제공될 경우 셀 상태 값 증가로 출력 값 h의 포화 현상 야기
- 포겟 게이트는 이와 같은 상황을 방지하기 위하여 메모리 블록을 초기화 기능 제공

제 4 장

바이오의료분야 과학기술정보 분석·활용 모형 개발

제 4 장
바이오의료분야 과학기술정보 분석·활용 모형 개발

제 1 절 분석·활용 모형 개발 개요

가. 연구 추진 방향

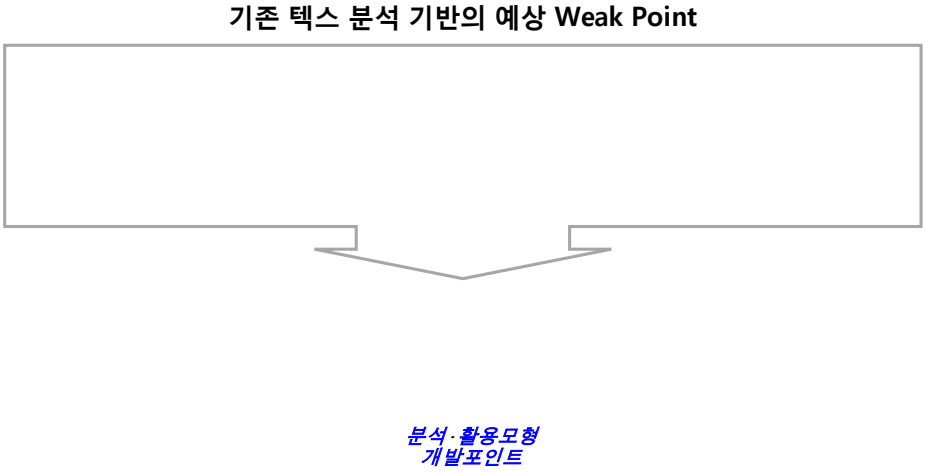
- 동 분석·활용 모형 개발은 사용자가 키워드 기반의 정형 데이터를 구성하여 분석을 수행하던 기존 방식을 개선하는 목표를 설정하고 연구를 추진
  - **실제 업무에 활용 가능하도록 신규 키워드 및 연구과제 내용학습, 키워드·과제간 연관성 분석, 과제분류의 정확도 향상을 추구**

- 고도화된 머신러닝 방법론을 적극 활용하여 업무과정에 자동화 체계를 도입하고 정확도를 높임과 동시에 분석에 대한 원활한 프로세스를 확립하는 것이 중요
  - **개발과정에서 발생하는 문제점을 파악하고 개선방안을 모색하여 분석·활용모형 성능 향상**

- 이를 위해 기존의 텍스트 분석 방식과 더불어 텍스트 임베딩(embedding)[14], 딥러닝 등 최신 기계학습 방법론(알고리즘)을 활용
  - **간단한 기계학습 방법을 적용하기 보다는 고도화된 방법론 적용으로 분석·활용 모형가치 제고**

- 분석·활용 모형이 일회성 사용에 그치지 않도록 지속 학습 방안을 강구하여 모형 성능 향상 기여
  - **과학기술지식정보가 매년 업데이트되는 점을 감안, 분석·활용 모형이 장기 활용될 수 있는 방안 마련 및 적용**

14) 단어, 문장 등 텍스트 데이터를 벡터화(수치정보화) 하는 것



- 머신러닝 알고리즘 적용을 통한 정형화된 분석데이터 생성 자동화
- 정보손실 방지에 의한 예측력 및 분석력 강화

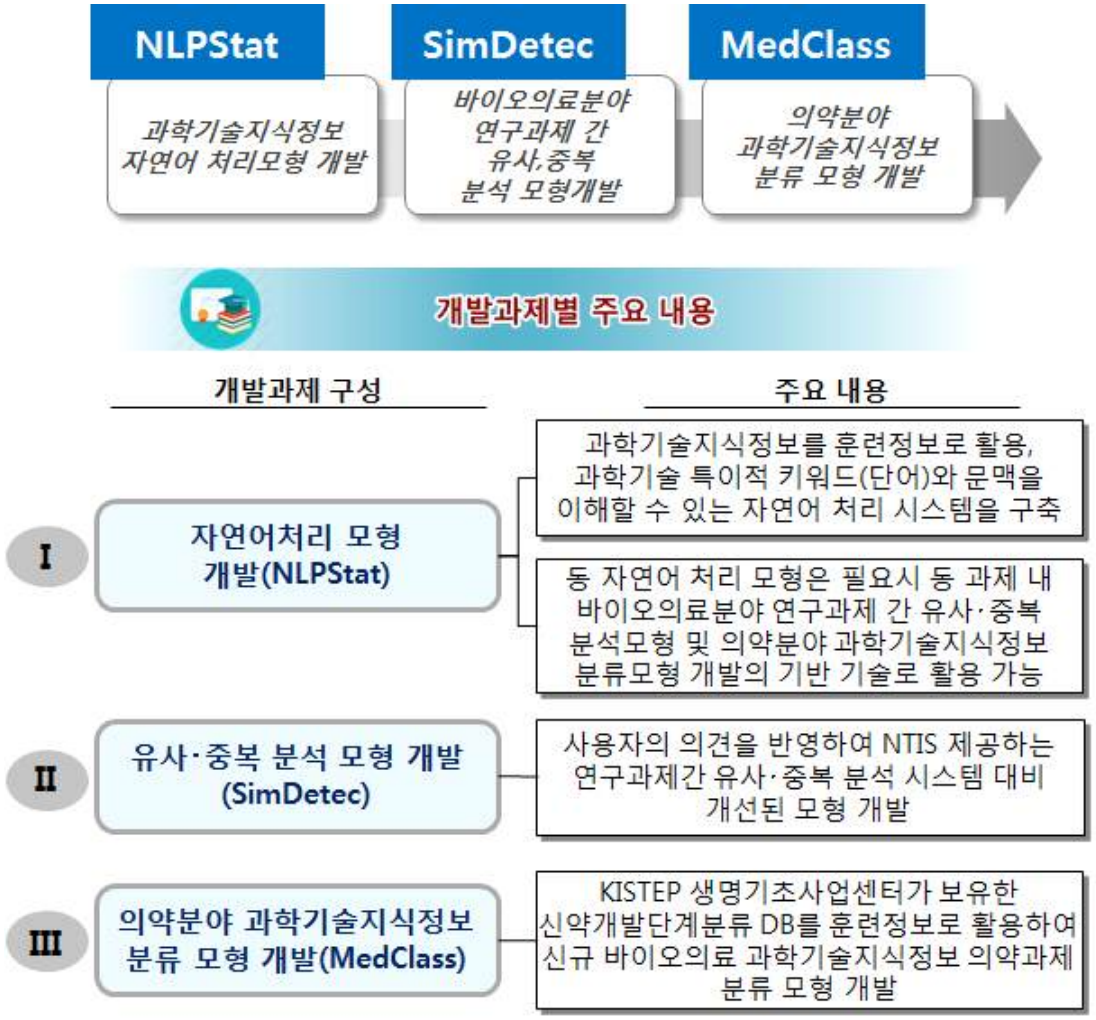
- Doc2vec에 의한 임베딩 벡터의 생성, 딥러닝 알고리즘에 의한 분류모형 개발 등 다양한 방법론 적용 개발

- 다양한 분석 방법 적용 후, 분석 단계별 seamless 연결을 위해 분석 프로세스 정의 필요
- 아울러 데이터 객체 개념 도입 활용

[그림 4-1] 분석·활용 모형 개발 추진방향

나. 분석·활용 모형 개발개요

- 딥러닝 등 고도화된 최신 기계학습 방법을 바탕으로 과학기술지식정보 자연어처리 및 바이오의료 연구과제 분류 모형을 개발
  - 과학기술정보는 연구자가 수행하는 연구내용을 직접 입력하기 때문에 같은 양식(변수)를 사용하더라도 비정형 데이터에 해당
  - 이를 감안하여 비정형 데이터를 탐색하고 정형화하는 첨단 기계학습 분석방법을 적용, 개발된 분석·활용모형을 통해 분석 업무 자동화를 지향하고자 함
  - 바이오의료분야 과학기술정보 분석·활용 모형 개발 내용
    - (NLPStat) 과학기술지식정보 자연어처리 모형 개발
    - (SimDetect) 바이오의료분야 연구과제 간 유사·중복(관계성) 분석 모형 개발
    - (MedClass) 의료분야 과학기술지식정보 분류 모형 개발

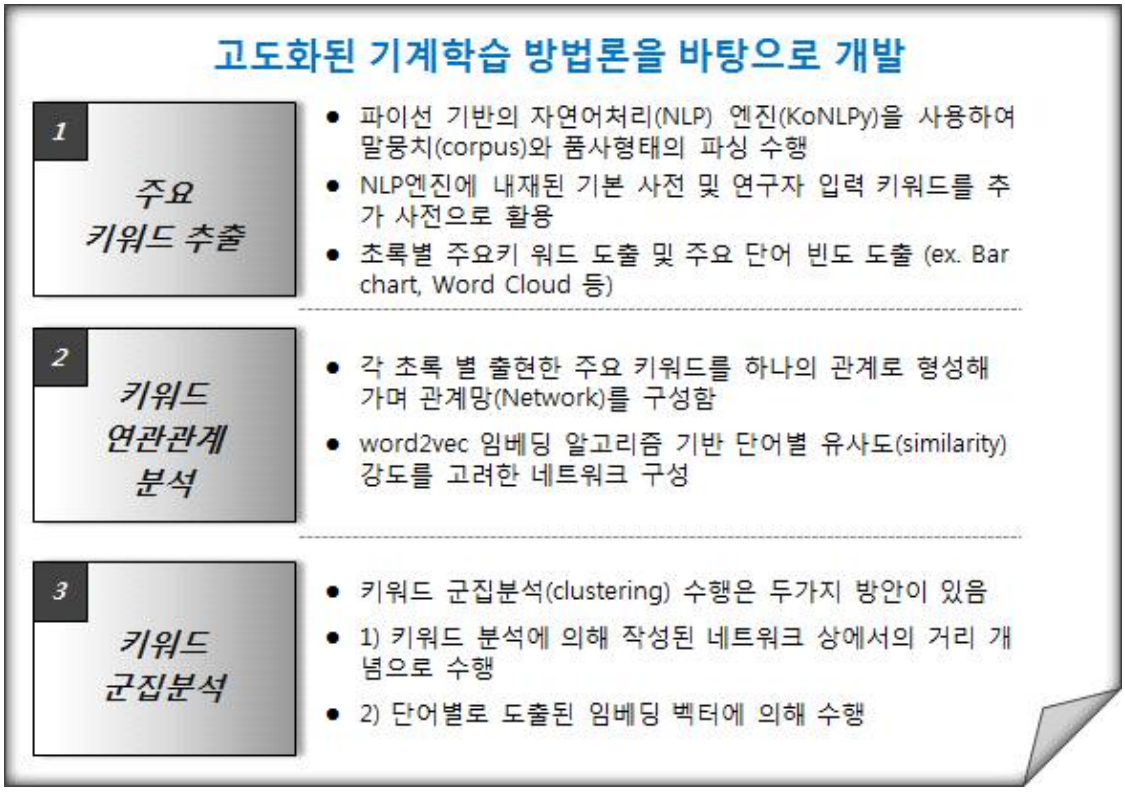


[그림 4-2] 분석·활용 모형 개발 개요

1) 과학기술지식정보 자연어처리 모형(NLPStat)

■ 과학기술지식정보 자연어처리 모형 개발 목표 및 기대효과

- 과학기술지식정보를 훈련데이터로 활용, 과학기술 특이적 용어(키워드)와 문맥을 이해할 수 있는 자연어처리 모형을 개발
- 필요한 정형화 작업을 숫자형 벡터로 표현되는 등의 기계학습 알고리즘을 최대한 반영, 갱신된 훈련데이터를 바탕으로 향후 학습에 의한 성능 향상 및 확장 추구
  - 매년 새로운 과학기술용어가 등장하고 기술분야별 용어가 세분화 되는 점을 감안, 사용자 기반 사전 구성 기능 제공
- 동 모형을 통해 비정형 데이터의 체계적 정형 데이터화 기반을 마련

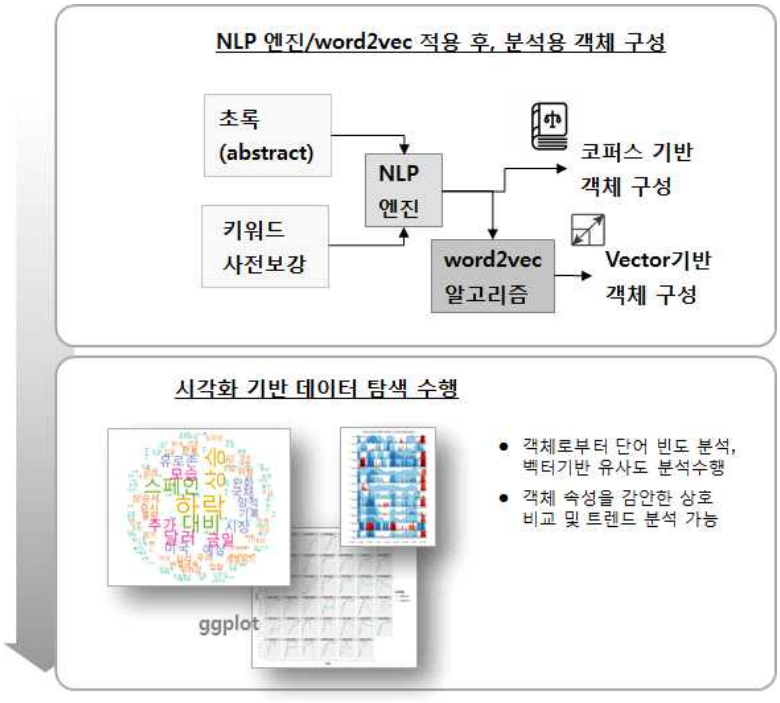


[그림 4-3] 자연어 처리모형(NLPStat) 개발의 주요 목표 및 내용

과학기술지식정보 자연어처리 모형 개발 내용

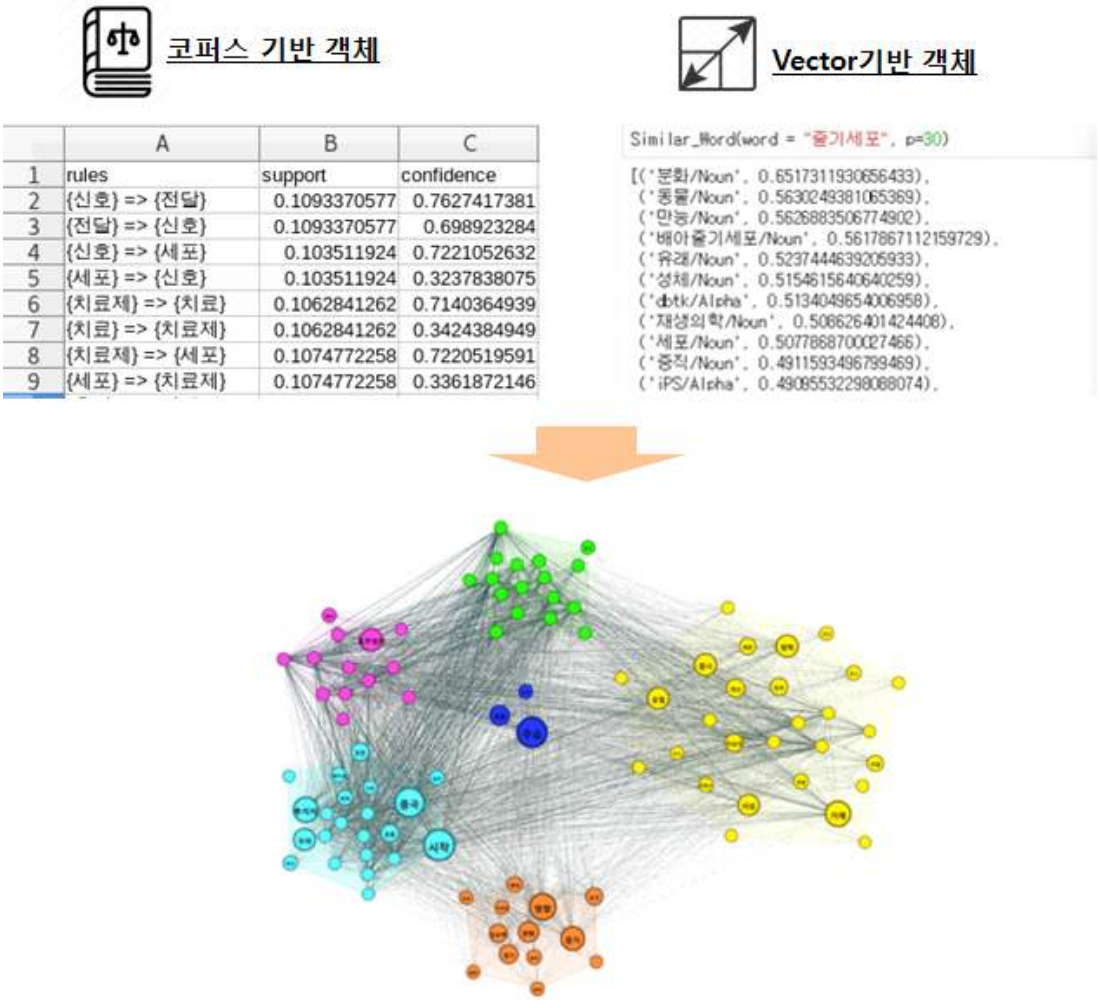
- 동 모형은 비정형 텍스트 데이터를 본격적으로 분석하기 위한 전처리 혹은 기초·탐색단계에 해당
  - 과학기술정보에 내제된 용어들을 사전 기반 형태소 분석을 통해 토큰화하고 주요 키워드 추출, 과학기술정보를 용어(단어) 및 문서 수준에서 벡터화함
- 주요 키워드 추출 단계는 다양한 탐색적 방법을 적용하여 데이터로부터 의미 있는 정보를 발굴하는 단계임
- 동 모형에서는 주요키워드 추출을 다음과 같은 두 가지 단계에서 수행

  1) 사전 기반 단어 빈도수 카운트
  2) word2vec(단어 임베딩 알고리즘) 적용한 단어 벡터 값 활용
- 동 모형에서는 워드클라우드, 벡터화된 단어를 중심으로 탐색적 데이터분석을 통해 연차별 트렌드 비교·분석 등을 가능케 함



[그림 4-4] 주요 키워드 추출 방법 및 데이터 시각화 개요

- 주요 키워드간 연관관계 분석은 1) 코퍼스 기반 객체 2) 벡터(Vector) 기반 객체로 구분되어 수행
  - 코퍼스[15] 기반 객체는 연관관계는 apriori 알고리즘을 기반으로 연관분석(association) 후, 이를 기반으로 관계망 구성
  - 벡터 기반 객체의 경우, word2vec 임베딩 기반으로 주요 키워드 간 유사도 중심 관계망 구성



	A	B	C
1	rules	support	confidence
2	{신호} => {전달}	0.1093370577	0.7627417381
3	{전달} => {신호}	0.1093370577	0.698923284
4	{신호} => {세포}	0.103511924	0.7221052632
5	{세포} => {신호}	0.103511924	0.3237838075
6	{치료제} => {치료}	0.1062841262	0.7140364939
7	{치료} => {치료제}	0.1062841262	0.3424384949
8	{치료제} => {세포}	0.1074772258	0.7220519591
9	{세포} => {치료제}	0.1074772258	0.3361872146

[그림 4-5] 주요 키워드 간 연관관계 분석 적용 방법

15) 말뭉치라는 의미로 특정한 목적을 토대로 추출된 언어 표본 집단

2) 바이오의료분야 연구과제 간 유사·중복(관계성) 분석 모형(SimDetect)

■ 바이오의료분야 연구과제 간 유사·중복(관계성) 분석 모형 개발 목표 및 기대효과

- 키워드 중복 및 발생 빈도에 의한 유사·중복 분석이 아닌, 문맥적 흐름, 문장의 구성 등에 근거한 과학기술정보 연구과제간 유사도* 파악하고 점수를 통해 정량화

  * 공간상에서 벡터화된 과제간 거리를 측정하는 것으로 두 과제간 연관성(상관관계)로 해석 가능

- 국가연구개발사업 예산배분조정 시 바이오의료 세부기술분야 기·신규과제 간 유사성 여부를 판단, 예산배분조정업무 추진효율성 제고 기대

■ 바이오의료분야 연구과제 간 유사·중복(관계성) 분석 모형 개발 내용

- 동 모형은 연구자가 입력한 초록(연구내용) 간 유사성(관계성)을 정량적 수치[16]로 표현
  - 과제 정보(과제명, 연구내용, 연구목표 등)에 적용한 doc2vec 알고리즘에 의해 산출된 코사인 유사도에 근거하여 유사·중복과제 탐색 및 수준 측정
- 자연어처리 모형(NLPStat)을 활용하여 연구과제별 용어에 원핫 인코딩(one-hot encoding)을 수행
  - 문장의 단락 내에 총 1,000개의 단어가 있다면, 길이 1,000개 길이의 벡터 중 해당 단어의 특정 위치를 1로 표기

16) 코사인유사도(cosine similarity): 내적공간에서 두 벡터간 각도의 코사인 값을 활용하여 두 벡터간의 유사성을 측정하는 방법으로, 동 모형에서는 과제를 벡터화하여 유사과제간 코사인 유사도를 측정

(단락) 서울 인천 대구 대전 부산 ....

서울 = [1, 0, 0, 0, 0.... , 0]
인천 = [0, 1, 0, 0, 0.... , 0]
대구 = [0, 0, 1, 0, 0.... , 0]
대전 = [0, 0, 0, 1, 0.... , 0]
부산 = [0, 0, 0, 1, 0.... , 0]

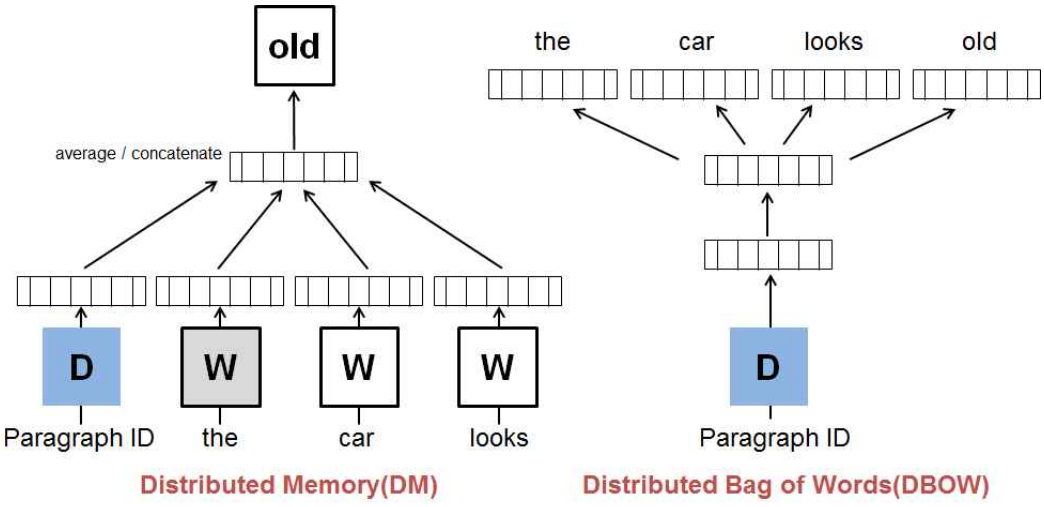
[그림 4-6] one-hot encoding 예시

- Doc2vec 알고리즘에는 분산메모리방식(distributed memory)과 DBOW(distirbuted bag of words)방식이 존재하는데 동 모형에서는 DBOW 방식을 적용

  ※ word2vec 알고리즘도 doc2vec 알고리즘과 유사한 방식으로 수행됨

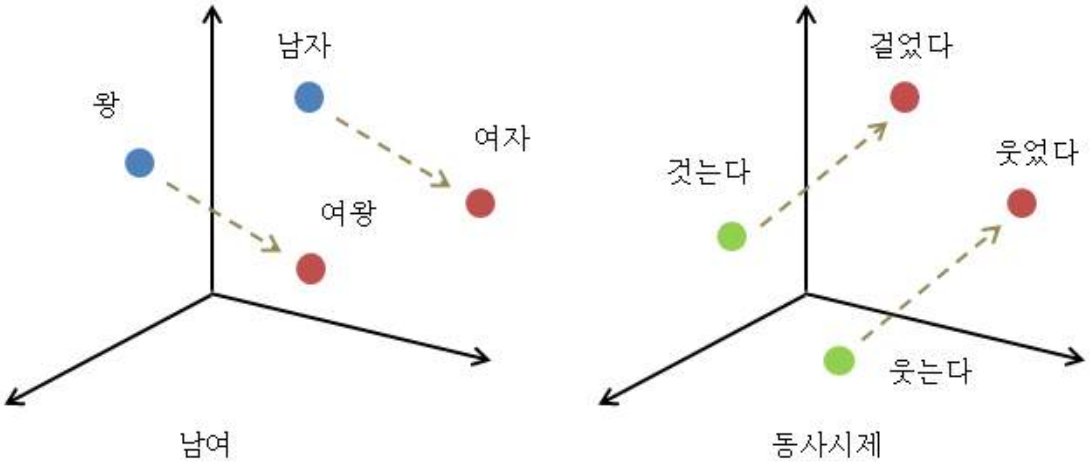
  - 분산메모리방식은 원핫 인코딩을 기반으로 윈도우사이즈(window size)를 설정한 후, 단락 ID와 입력 단어에 대한 타겟 단어(target word) 예측의 방식으로 doc2vec을 수행

  - DBOW 방식은 분산메모리방식과는 반대로 특정 단락ID로부터 문맥을 구성하는 단어를 예측



[그림 4-7] doc2vec 알고리즘 개념도

- 텍스트 데이터가 임베딩 과정을 거쳐 벡터화가 될 경우 아래 그림과 같이 공간상에서 관계성을 보이는 단어들이 나타내는 방향성 및 거리에 따라 관계의 정도 파악 가능



[그림 4-8] 단어 임베딩 결과 예시

- doc2vec에 의해 연구과제별 벡터값이 계산된 결과물*을 바탕으로 연구과제별 코사인 유사도를 측정하여 유사정도를 정량화

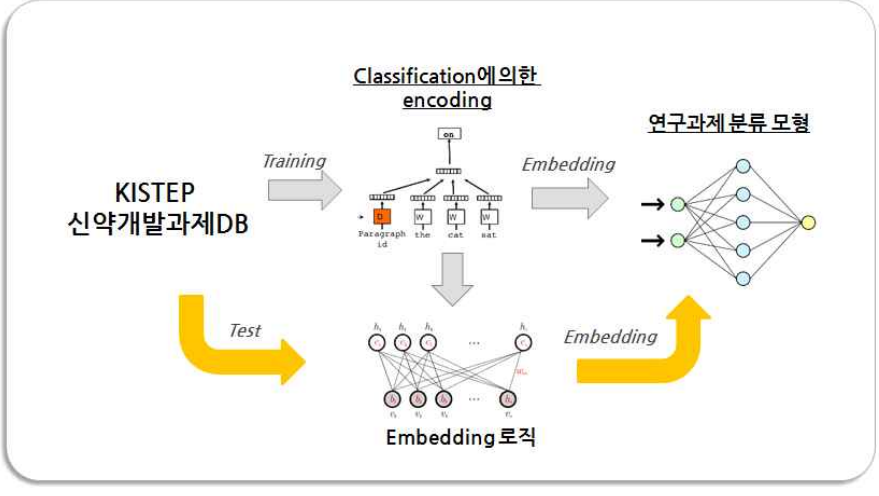
  * doc2vec에 의해 연구과제별 벡터값이 계산된 결과물

- 측정된 연구과제간 유사도 값을 바탕으로 관계망(네트워크 시각화) 기반 연관분석을 수행하여 유사과제간 중요성 여부 등 판단

3) 의약분야 과학기술지식정보 분류 모형(MedClass)

■ 의약분야 과학기술지식정보 분류 모형 개발 목표 및 기대효과

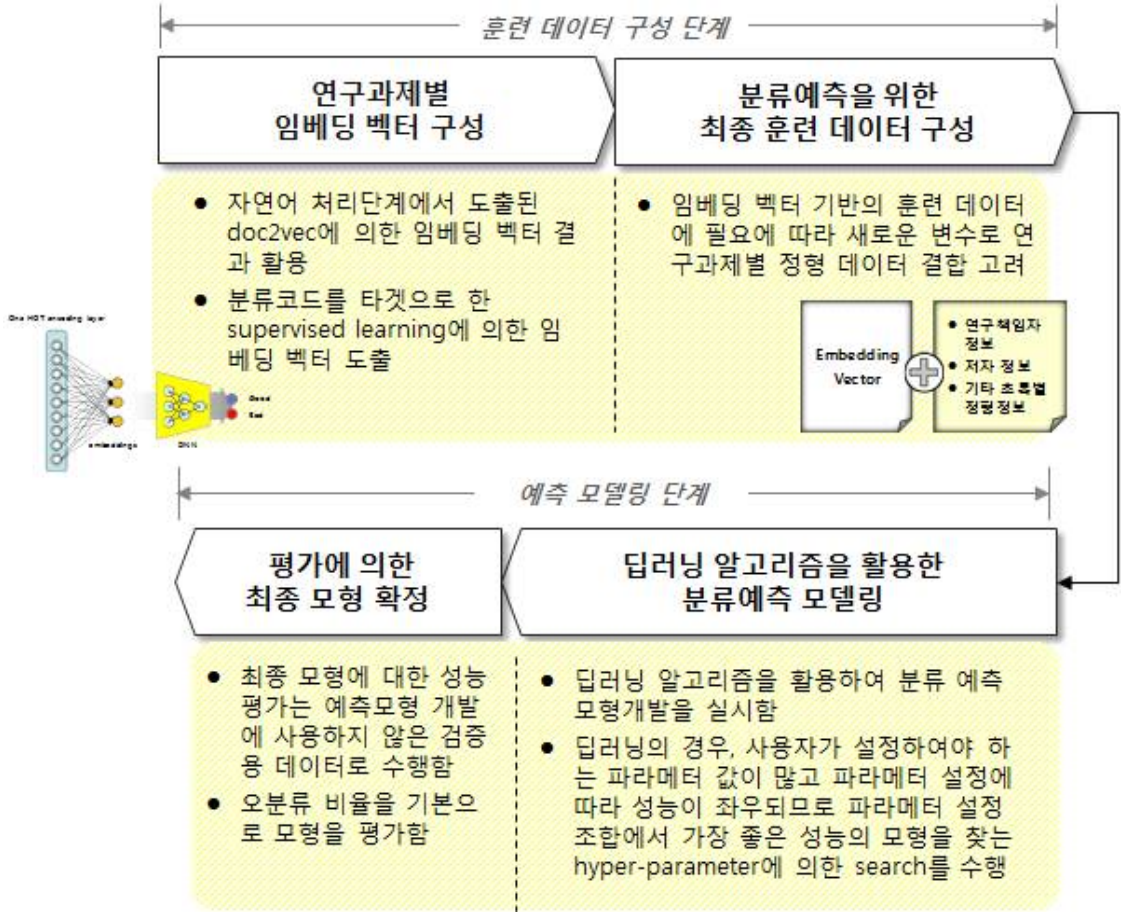
- KISTEP 생명기초사업센터가 보유한 신약개발단계분류 DB를 훈련데이터로 활용하여 신규 바이오의료 과학기술지식정보 의약과제 분류 모형을 개발
  - 입력된 의약분야 연구과제의 사전 분류 정보를 바탕으로 향후 연구과제에 대한 자동분류를 가능하게 함
  - 매년 전문가의견을 바탕으로 신약개발단계분류 DB를 업데이트하고 있으므로, 모형은 신규정보를 수용하고 이를 학습하여 자체성능향상이 가능해야 함
- 최신기계학습 방법론 중 주목을 받는 딥러닝 기반 인공신경망 모델을 구축하고 우수성 여부를 검증
  - 전통적으로 활용되었던 단순 기계학습 방법론 대비 딥러닝 기반 인공신경망 모델을 구축하여 분류 성능을 고도화
  - 훈련데이터(정부신약개발과제DB)를 doc2vec을 통해 임베딩(embedding) 하여 심층 신경망(Deep Neural Netowork) 분류 모형을 구축
- 동 연구 모형를 통해 연구결과 자동분류 및 기 분류결과에 대한 타당성검증 활용이 가능할 것으로 사료됨



[그림 4-9] 의약과제분류모형(MedClass) 구현 및 작동방안 예시

의약분야 과학기술지식정보 분류 모형 개발 내용

- 훈련데이터(KISTEP 정부 신약개발과제DB)를 활용하여 신약개발단계·의약품종류·대상질환별 딥러닝 기반 분류모형을 개발
  - 정부 신약개발과제DB는 2008-2015년도에 해당하는 6,351건의 연구과제*로 구성
    * KISTEP 통계브리프(신약개발 정부 R&D 투자 포트폴리오 분석) 기준에 따라 취합
  - 연구과제는 KISTEP 통계브리프에 제시된 분류기준에 따라 신약개발단계·의약품종류·대상질환별로 전문가에 의해 분석(분류)되어 있음
- 동 모형 개발은 예측모형 도출에 필요한 훈련·검증데이터 구성단계와 실제 예측 모델을 개발하는 단계로 나뉘어 진행됨
  - 신약개발단계·의약품종류·대상질환별 3종류의 분류모형(classifier)를 각각 구축하기 위해 신약개발과제DB 구성별 훈련데이터를 정리·활용하여 학습 및 모형 개발
  - 모형개발 과정에서 2008-2015년 신약개발연구과제DB를 훈련집단(training set) 및 테스트집단(test set)으로 구분하여 분류 성능을 검증
    ※ 성능 검증 시 지도학습 방법을 이용한 분류기(classifier)와 분류성능을 비교 분석
  - 최종적으로 동 과제 수행과정에서 새로이 구축한 16년도 신약개발과제DB*를 얼마만큼 잘 분류해 내는지 전문가의 분류결과와 비교분석 수행
    * 총 967개 과제 중 계속과제 500건, 신규과제 467건



[그림 4-10] 의약분야 과학기술지식정보 분류모형 개발과정 모식도

제 2 절 분석·활용 모형 개발결과

1) 과학기술지식정보 자연어처리 모형(NLPStat)

- 사용자가 기존 사전에 없는 과학기술용어 및 신조어를 자유롭게 입력하는 기능 지원(사전 보강)을 통해 분석 정확도 향상

  - 사전에 제시되어있지 않은 과학기술용어 중 합성어와 유사한 구조를 가지고 있는 용어는 토큰화하여 임베딩되기 때문에 내재공간에서 해당단어의 의미를 내포한 정확한 벡터값 유추가 어려움

    – 가령, '바이오시스템'이라는 용어는 자연어처리 프로그램이 제공하는 사전을 기준으로 '바이오', '시스템'으로 토큰화하여 각각 벡터값을 지님

    – 따라서, 사전에 '바이오시스템'이라는 단어를 추가하지 않을 경우 '바이오시스템'은 '바이오'와 '시스템' 두 단어로 처리되고 해당 벡터값도 '바이오시스템' 자체를 가리키는 것으로 보기 어려움

  - T세포 같은 기존 사전에 부재한 단어는 사용자가 직접 추가하여 모형 성능 제고 가능

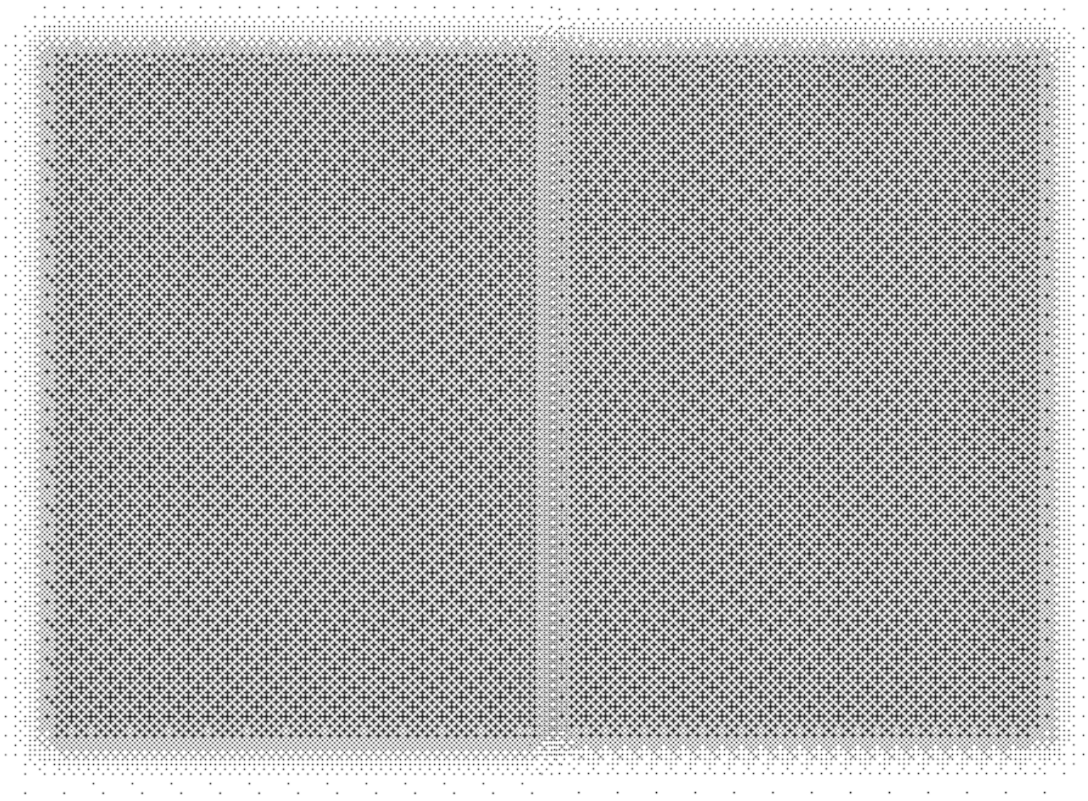
[그림 4-11] 모형성능 향상 및 최적화를 위해 사전에 새로운 단어 추가 가능

- 워드클라우드(WordCloud)를 활용, 중요 키워드 발굴 및 연도별 트렌드 감지 시 과학기술표준분류 중분류 기준으로 탐색이 가능하고 빈도가 높은 단어를 제거 노이즈필터링 기능 제공
  - 대분류 기준으로 빈도수 측정 시 탐색범위가 광범위하기 때문에 기술수준분류가 상대적으로 명확해지는 '중분류'를 기준점으로 설정
  - 동 방법은 단어의 출현빈도를 바탕으로 구현되기 때문에 핵심적인 키워드보다는 일반적인 단어 노출비중이 높아 중요 키워드 발굴 시 용이하지 않은 측면이 있음
  - 따라서, 상위 빈도 단어를 사용자가 임의로 제거할 수 있도록 노이즈필터링 기능을 추가

[그림 4-12] 단어출현 빈도에 의한 워드클라우드 결과

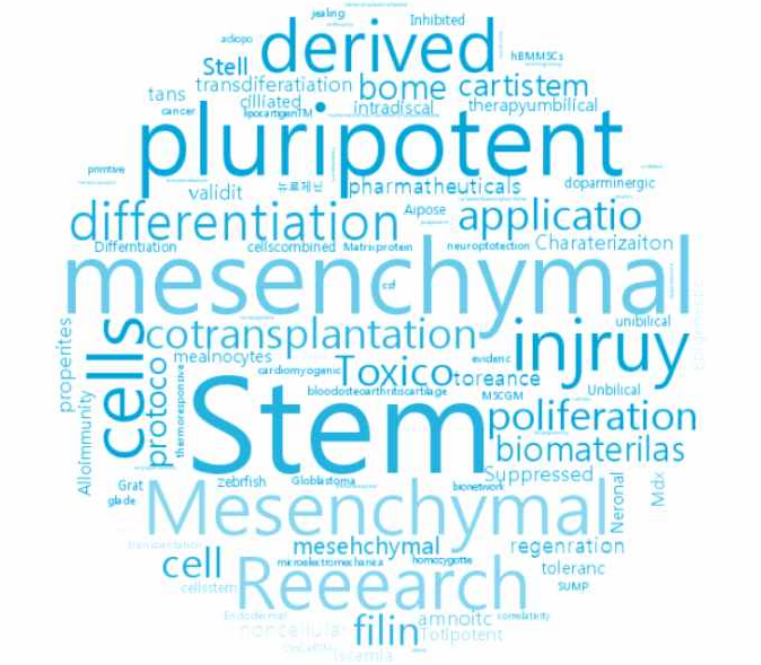
- 단어 출현 빈도수가 아닌 벡터값 기반 연관 키워드 추출 시 WordCloud를 통해 시각화함으로서 빈도수 카운트 결과와 유의미성 비교 가능
  - **단어를 입력하면 연관성이 높은 단어를 내림차순으로 출력, 제공되는 코사인유사도를 바탕으로 유사정도* 파악 가능**

    * 동 과정에서는 연구목적을 감안하여 유사정도를 연관정도로 정의



[그림 4-13] 줄기세포와 stem cell 연관단어(벡터값 기반) 출력

- 해당 결과 역시 워드클라우드를 통해 시각화가 가능하며 연관성이 높은 단어일수록 글자 크기가 크게 표현

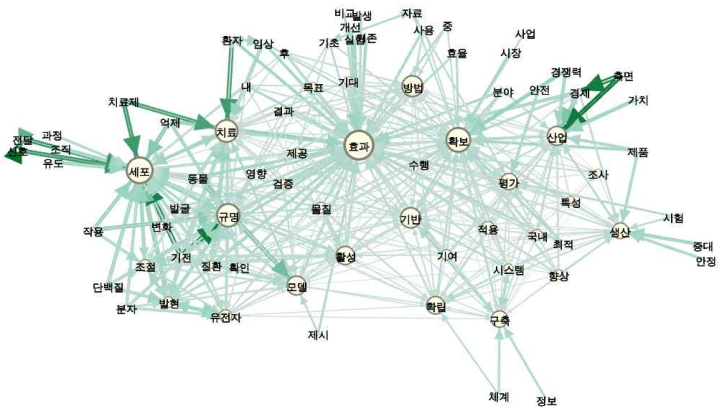


[그림 4-14] 벡터값 기반 워드클라우드 결과

- 앞서 제시된 단어출현빈도 기반의 방법과 비교할 때 벡터값 기반 방법은 입력된 키워드(stemcell)와 유의미하게 연관된 키워드를 제시
- 중요 키워드간 연관분석 수행 후에는 Gephi(오픈소스 네트워크분석 및 시각화 소프트웨어)를 이용하여 연관관계에 의한 관계망 시각화
  - 연관분석(association) 결과는 엑셀파일형식(csv파일)로 다운로드가 가능하며, 사용자는 이를 이용하여 목적에 따라 추가적인 분석수행 및 활용 가능

[그림 4-15] 주요 키워드간 연관분석 결과 제공형태

- 연관분석(association) 후 제공되는 결과파일(csv파일)를 활용하여 관계망(네트워크) 분석 가능



[그림 4-16] 연관분석(association) 네트워크 시각화

2) 바이오의료분야 연구과제 간 유사·중복(관계성) 분석 모형(SimDetect)

- 과학기술표준분류 코드 중 대분류를 기준으로 바이오분야에 해당하는 뇌과학, 생명과학, 보건의료, 농림수산식품을 선택하여 doc2Vec 모델을 학습함
  - 동 모형구축에 사용된 바이오의료분야 과학기술지식정보(연구과제) 수는 약 14만건으로(2007~2016)에 달하기 때문에 학습에 많은 시간이 소요됨
    - 14여만건의 과제정보에는 계속과제가 포함되어있으며 연구과제별 과학기술표준분류 가중치 정도에 따라 타 기술분야 성격의 과제도 일부 포함될 수 있음
  - 학습에 들어가는 변수는 과제명-국문, 요약문_연구내용, 요약문_연구목표, 요약문_기대효과, 요약문_한글키워드, 요약문_영문키워드를 기본적으로 사용
    - 사용자 임의로 연구목적에 따라 학습요소 재구성 및 조합 가능
    - 가령, 연구과제명, 영문키워드 등을 제외하고 학습을 시키는 것이 가능함
  - 훈련데이터양은 사용자 임의설정 및 재구성이 가능하나 훈련데이터 재구성 시 분석모형의 재학습이 필수적으로 요구됨
    - 사용자에 따라 바이오의료분야 연구과제의 설정범위 조정, 계속과제 정리 등을 통해 훈련데이터 재구성 가능
  - 키워드나 연구내용을 입력하면 2007년부터 2016년까지의 자료 중에 관계성이 높은 과제(유사연구)를 분석하여 출력하도록 개발
    - 벡터화된 연구과제 간 유사과제를 탐색은 과제간의 코사인유사도를 측정으로 이루어지는데, 이는 벡터공간에서 관계의 정도를 나타낸다고 볼 수 있음
  - 유사과제 분석 수행을 위해서 사용자 키워드(용어), 문장, 단락 등 연구와 관련된 정보를 입력해야 하며 입력정보의 수준에 따라 결과가 달라질 수 있음
  - 다음의 수행 예시(그림 4-17)는 '동물의 체내에서 사람의 장기를 생산하는 연구'를 입력하였을 때 출력되는 결과물임
    - 줄기세포(stem cell), 유도만능줄기세포(iPSC), 배아줄기세포(ESC) 등의 키워드를 포함하지 않았으나 줄기세포를 이용한 동물 기반 인공장기 연구의 탐색이 가능했음

	유사도	과제수행년도	부처명	사업명	과제명-국문
1	0.923986868	2011	농촌진흥청	차세대바이오그린21	형질전환 복제돼지의 이식면역반응 특성규명 기술 확립(차세대바이오그린21)
2	0.923210918	2016	미래창조과학부	집단연구지원	인간화 돼지 연구센터
3	0.923134512	2015	농촌진흥청	차세대바이오그린21	Rag-2 유전자가 결핍된 면역결핍돼지를 활용한 환자 맞춤형 암 동물모델 개발
4	0.92236391	2007	교육과학기술부	우수연구센터육성<SRC,ERC,MRC,NCRC>	형질전환 돼지 생산을 위한 효율적인 초기배의 개발연구
5	0.919401263	2007	보건복지부	(보건의료기술연구개발)보건의료기술연구개발	돼지태아 줄기세포를 이용한 인슐린분비세포 분화와 이용에 관한 연구
6	0.918693227	2007	농촌진흥청	축산생명환경시험연구	형질전환가축 이용 바이오신약 생산기술 개발
7	0.918123561	2007	교육과학기술부	바이오신약장기사업	이종장기 이식 거부반응 억제를 위한 내피 및 상피세포의 조절
8	0.917363985	2009	교육과학기술부	미래기반기술개발	심혈관 질환 동물모델을 이용한 제대혈 및 지방조직 유래 줄기세포의 기능 연구
9	0.917051035	2012	교육부	일반연구자지원	돼지 심근경색모델에서 줄기세포 이식처료의 안전성 평가와 이식기술의 최적화를 통한 세포처료 기술 개발
10	0.917051035	2015	농촌진흥청	차세대바이오그린21	내분비호르몬이 돼지 및 인간 줄기세포의 EMT 조절을 통한 세포사멸에 미치는 영향 평가
11	0.91635837	2014	농림축산식품부	농생명산업기술개발	가축유래 전분화능 줄기세포를 이용한 고효율 형질전환/질환모델 동물 생산기술 개발
12	0.916288318	2010	농촌진흥청	바이오그린21	기세포 유래 면역세포 연구를 위한 바이오마커 및 기능조절 물질개발(바이오그린21)
13	0.916263344	2014	농림축산식품부	농생명산업기술개발	형질전환 돼지생산
14	0.916242148	2016	농림축산식품부	농생명산업기술개발	돼지 줄기세포를 이용한 유전자조작 기술의 확립
15	0.916228178	2010	교육과학기술부	중견연구자지원	인체 간 유래 줄기세포의 이식 기반기술 연구
16	0.916220564	2016	농림축산식품부	농생명산업기술개발	가축유래 전분화능 줄기세포를 이용한 고효율 형질전환/질환모델 동물 생산기술 개발
17	0.916195202	2016	농림축산식품부	농생명산업기술개발	고부가 달걀 생산을 위한 가금생식줄기세포의 유전자 조절체계 확립

[그림 4-17] 유사과제 분석결과 수행 예시

- 유사과제 수행 결과는 사용자의 후속활용 연계를 위해 csv파일로 추출 가능
  - 결과파일은 예산배분조정 업무수행과정에서 요구*되는 과제에 관한 기본적인 변수 항목으로 구성

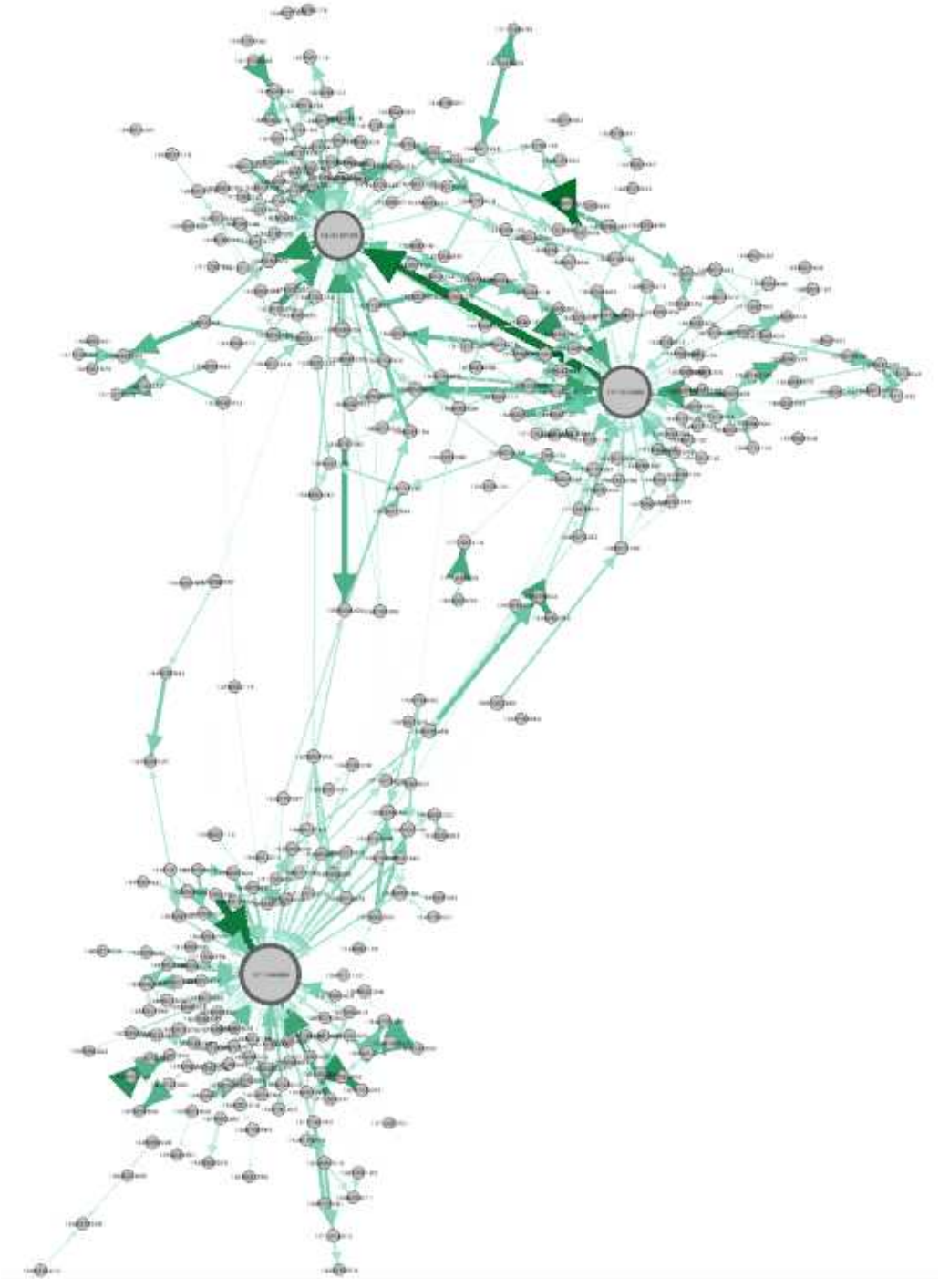
    * 과제수행연도, 처명, 사업명, 내역사업명, 과제명, 연구개발비, 연구목표 및 내용 등
  - 국가과학기술지식정보 웹사이트 링크를 제공하기 때문에 웹사이트 기반 정보 필요시 손쉽게 해당 사이트도 접속 가능

■ 연구과제 간 관계성 분석은 입력받은 과제를 바탕으로 수행되며 해당결과 역시 csv파일 추출 및 관계망(네트워크) 시각화를 통해 분석이 가능

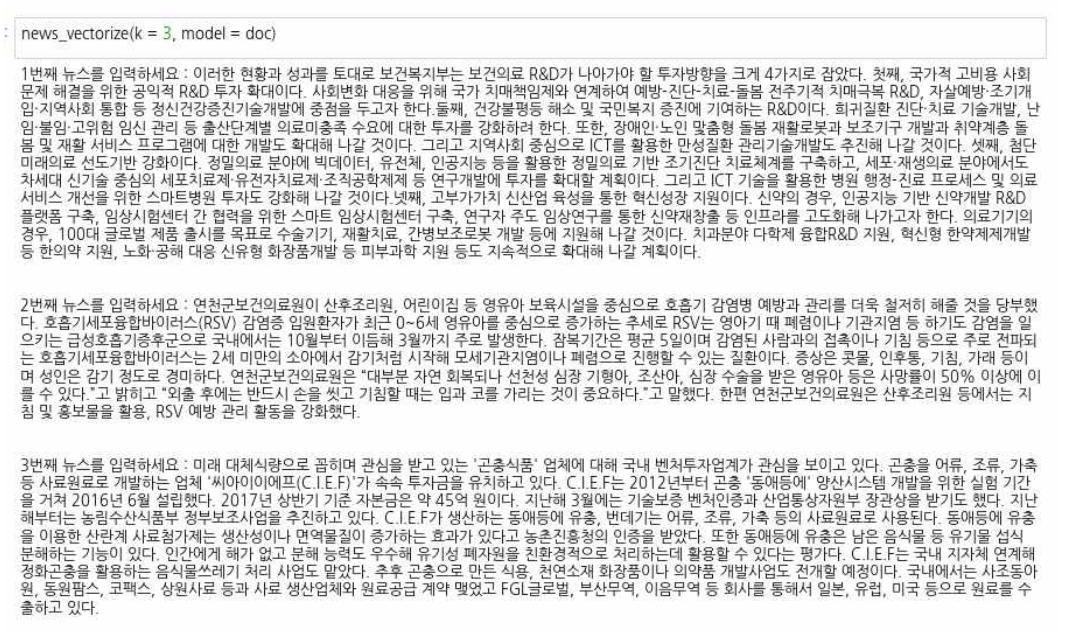
- 사용자는 연구과제 번호를 이용하여 모형에 관심 있는 과제를 입력할 수 있으며 입력된 과제간의 관계성 외에 공간상에서 입력된 과제와 관계가 있는 주변과제간의 관계망 분석을 수행
  - 입력과제수와 분석에 포함되는 주변과제 수도 사용자가 직접 지정할 수 있음

[그림 4-18] 연구과제간 관계성 분석을 위한 과제 입력 예시

- 사용자는 관심 있는 연구과제 번호를 직접 입력할 수 있으며 입력된 과제간의 관계성 외에 공간상에서 입력된 과제와 관계가 있는 주변과제간의 관계망 분석을 수행
  - 벡터값을 바탕으로 관계망을 그리기 때문에 노드(연구과제) 간 방향성 및 관계성 정도에 따라 화살표, 엣지 굵기 등으로 구분되어 표현됨
  - 입력된 연구과제와 주변연구과제 간 관계성 뿐만 아니라 주변노드 간의 관계성도 파악 가능



- 연구과제 외에 사용자가 관심 있는 신문기사와 같은 콘텐츠 입력도 가능하며 이와 연관된 연구과제를 추출, 관계성 분석 수행 가능
  - 콘텐츠 수를 사용자가 직접 지정하고, 문장 및 단락수준으로 지정한 콘텐츠 수만큼 입력 가능



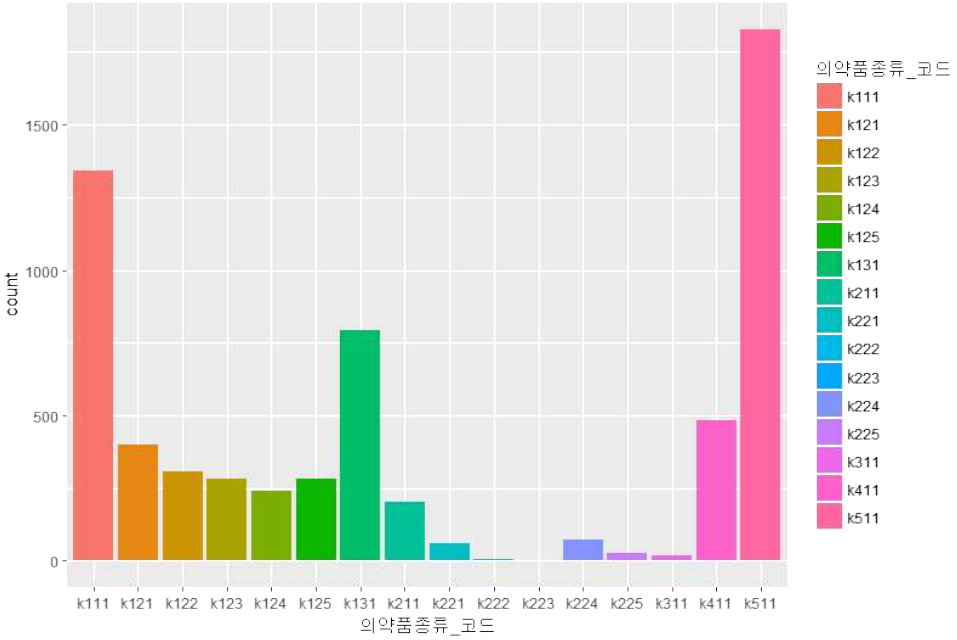
[그림 4-20] 관심 있는 콘텐츠와 연구과제간 관계성 분석 수행 예시

- 유사중복 분석모형 재학습 시 학습 파라미터를 사용자가 직접 설정함으로써 기술적으로도 분류 성능의 향상을 기대할 수 있음
  - 사용자 스스로 학습량, 벡터 및 윈도우 사이즈 등의 학습 파라미터 직접 조정 가능
    - 윈도우 사이즈, 벡터 차원을 확대하거나 학습량을 높일 경우 모형 성능 향상 가능

3) 의약분야 과학기술지식정보 분류모형(MedClass)

- 의약품종류코드, 신약개발단계코드, 질환코드 별로 각각의 doc2Vec 적용 결과를 바탕으로 딥러닝 기반 인공신경망 분류모형(총 3종류)을 구축
  - 신약개발연구단계DB에 저장된 과제의 과제명-국문, 요약문_연구내용, 요약문_연구목표, 변수 내용을 통합하여 모형 학습에 사용
    - 과학기술지식정보의 변수항목 및 중요성, 모형구축의 용이성 등을 검토, 연구과제별 정형 데이터(연구책임자 정보 등)은 훈련데이터에 미포함
    - 연구과제별 국문키워드, 영문키워드 변수는 사용자의 의도에 따라 재학습 시 포함 가능

- 학습에 사용된 과학기술정보(과제기준)는 약 6,351여 건이나, 연구과제는 연속성을 지니고 있기 때문에 연차별 중복되는 과제가 존재
  - 분류기준 내 코드별 연구과제수에 차이가 있고, 계속과제에 의한 학습편향성이 발생할 소지가 있음
    - 가령 의약품종류 훈련 데이터의 경우 k511(공통기반기술), k111(합성신약) 코드에는 많은 양의 연구과제가 존재하는 반면 k222(유전자치료제), k223(세포치료제) 코드는 매우 적은 수준이었음
  - 우선적으로 프로토타입 성격의 모형인 점, 특정 계속과제는 연구책임자가 연도별 과제내용을 갱신하는 사례 등을 감안하여 6,351건의 훈련데이터를 학습에 사용함
    - 동 학습 모형은 사용자에 의해 훈련데이터를 재조정, 재학습이 가능하기 때문에 사용자의 의도에 따라 훈련데이터 조정(계속과제 정리 등) 후 학습을 통해 모형 갱신 가능



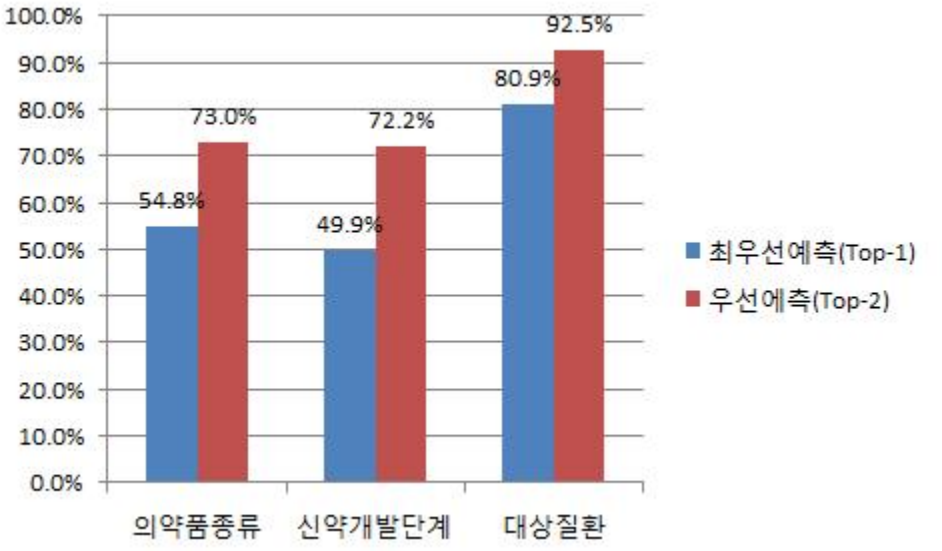
[그림 4-21] 의약품종류코드별 연구과제수

- 일부 지도학습 기반 분류모형과 비교 시 동 분류모형의 예측성능이 가장 우수하게 나타남
  - 훈련데이터를 7:3의 비율로 훈련집단, 테스트집단으로 구분하여 검증을 수행
  - 딥러닝 with Top-1(최우선예측)의 경우는 가장 높은 수준으로 예측된 분류결과(예측 class) 1개만 취하여 테스트셋 간의 정확도를 측정한 것으로, Randome Forest, GBM 분류 모형보다 높은 예측 성능을 보임
  - 추가적으로 2순위로 예측되는 분류결과도 상당한 신뢰성이 있다는 판단하에 예측 Class를 2개까지 취하였을 때(딥러닝 with Top-2, 우선예측)에는 실제 활용이 가능하다고 판단되는 수준의 예측력을 보여줌
  - 그러나 동 분석은 모형구축 시 활용된 훈련데이터를 바탕으로 진행되었기 때문에 실제 성능보다 높게 측정될 수 있음에 유의 필요

		정확도
의약품종류	딥러닝 with Top-1	75%
	딥러닝 with Top-2	87%
	Random Forest	49%
	GBM	66%
산약개발단계	딥러닝 with Top-1	69%
	딥러닝 with Top-2	84%
	Random Forest	40%
	GBM	57%
대상질환	딥러닝 with Top-1	87%
	딥러닝 with Top-2	95%
	Random Forest	64%
	GBM	76%

[그림 4-22] 의약분야 과학기술지식정보 분류모형 테스트 결과

- 의약과제 분류모형 개발 시 사용되지 않은 16년도 신약개발연구과제DB를 바탕으로 분류모형을 검증한 결과 훈련데이터 기반 분류결과에 비해 분류성능이 낮게 측정되었음
  - 16년도 신약개발연구과제DB는 967과제(계속과제 500개, 신규과제 467건)로 구성
  - 계속과제는 15년도 수행되었기 때문에 의약과제 분류모형의 훈련데이터에 포함되어 있으므로 이를 제외하고 신규과제 467건에 대해서만 수행
  - 훈련데이터 기반의 검증결과와 마찬가지로 대상질환, 의약품종류, 신약개발단계 순으로 예측성능이 높게 나타남
    - 최우선예측(Top-1)의 경우 의약품종류 54.8%, 신약개발단계 49.9%, 대상질환 80.9%의 예측성능을 보임
    - 우선예측(Top-2)의 경우 의약품종류 73.0%, 신약개발단계 72.2%, 대상질환 92.5%의 예측 성능을 보임



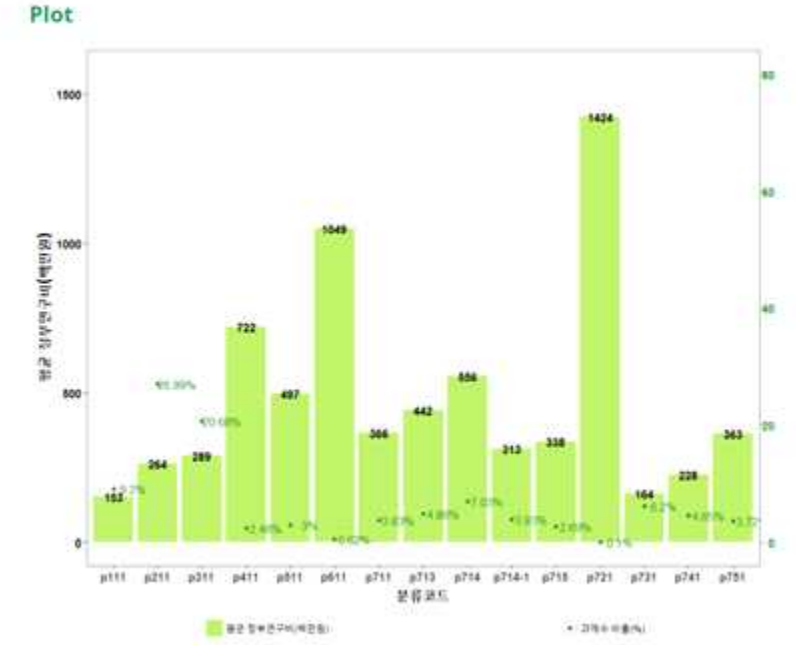
[그림 4-23] 16년 신약개발연구과제 DB 기반 의약분야 과학기술지식정보 분류모형 검증 결과

- 최우선예측 성능이 가장 낮게 측정된 신약개발단계 분류모형은 복수개발단계를 수행하는 과제가 많아 이와 같은 결과가 나온 것으로 판단됨
  - 신약개발연구과제DB의 '소분류' 기준으로 신약개발단계는 15개 코드, 의약품종류 및 질환은 11개 코드로 구성됨
  - 가장 많은 분류코드를 지닌 만큼 코드별로 훈련데이터(연구과제수)가 상대적으로 적게 배분되어 분류성능에 영향을 줄 수 있으나,
  - 우선예측결과가 의약품종류 분류모형과 유사한 수준임을 감안할 때 코드 수 보다는 한 개의 과제 내 에서 연속성 있는 개발단계를 수행하는 과제들로 인해 이와 같은 결과가 나온 것으로 판단됨
  - 16년 신약개발연구과제DB 구축과정에서 500개의 계속과제에 대해 전문가를 통해 재검토한 결과 일부 계속과제들은 분류가 모호*할 수 있다는 의견[17]을 받음

  * 의약품종류 19.2%, 신약개발단계 23.4%, 대상질환 6.0%

17) 분류기준별 전문가 개인의 시각차 존재하고 연구과제가 복수의 내용으로 구성될 수 있음

- 분류결과는 csv파일 형태로 획득 가능하며 연구과제 기반 통계정보 그래프화 기능 제공
  - 연구과제 분류결과는 csv파일로 제공되기 때문에 엑셀이나 기타 통계전문 프로그램을 활용하여 추가 분석이 가능함
  - 분석활용 모형 내에서도 연구과제 기반 통계정보를 바탕으로 기본적인 그래프는 작성 가능



[그림 4-24] 의약분야 과학기술지식정보 분류모형 그래프 시각화 예시

- 의약과제 분류모형은 신규 학습정보를 추가하여 모형의 분류 성능을 향상시키거나 새로운 분류코드 삽입이 가능하도록 설계
  - 신규 훈련데이터가 지속 추가될 경우 모형의 분류 성능이 제고되고 일회성 사용에서 그치지 않고 꾸준한 활용이 가능할 것

    연구트렌드에 따라 신약개발과제분류코드가 향후 수정될 경우 이를 반영할 수 있도록 새로운 분류코드 기반의 학습도 가능

- 의약과제 분류모형 재학습 시 학습 파라미터를 사용자가 직접 설정함으로써 기술적으로도 분류 성능의 향상을 기대할 수 있음
  - 사용자 스스로 학습량, 벡터 및 윈도우 사이즈 등의 학습 파라미터 직접 조정 가능
    - 윈도우 사이즈, 벡터 차원을 확대하거나 학습량을 높일 경우 모형 성능 향상 가능

[그림 4-25] 신약개발연구과제DB를 활용한 분류모형 학습 예시

제5장
결론

제5장 결론

제1절 바이오의료분야 과학기술정보 분석·활용 모형 개발

가. 개선 및 활용방안

■ 정부 R&D 예산배분조정 시 보건의료분야 신규과제 유사중복 검토 업무 효율성 제고 기대

- 정부 R&D 예산배분조정과정에서 신규사업 착수 여부 적절성을 위하여 사업목적, 내용, 수행계획 및 과제 등을 종합 검토
- 이중 투자 효율성 측면에서 신규-기 연구과제간 유사중복 검토요구가 증대되고 있음
- 키워드 빈도수에 의한 유사중복 탐색보다는 핵심 키워드를 바탕으로 연구과제의 내용간 문맥적 유사성을 검토하는 것이 필요하므로 동 과제 모형을 예산배분조정업무 과정 중 유용하게 활용 할 수 있을 것으로 예상
  - 다만, 동 모형에서 구현되는 기능은 실질적으로는 공간상에서 관계성이 높은 과제를 검색하는 것이기 때문에 일정 수준의 코사인유사도값*를 지니는 과제를 모두 유사중복이라고 여기는 것은 적절한 활용방안이 될 수 없음
    * 동 분석활용 모형은 코사인유사도 값을 통해 과제간 관계성 분석도 수행
  - 따라서, 사용자는 다수의 활용을 통해 유사정도를 판단하는 수치를 경험적으로 축적해야하며 경우에 따라서는 가치판단 할 필요가 있음
  - 예를 들어, 동 모형은 구동 방식(벡터 값에 기반 유사도 검색)으로 인하여 질의어 입력 시 항상 결과 값이 도출되고 콘텐츠에 따라 결과로 제시되는 유사도 값이 일반적인 수준보다 높거나 혹은 낮더라도 관련 과제이거나 유사한 과제일 수 있음

  향후 정부 연구개발 사업간 과제 수행내용을 바탕으로 연구개발 사업간 관계성을 분석하는 기능 등이 추가될 경우 모형의 활용성이 보다 제고될 수 있을 것

– 연구개발사업 심층검토 시 동 분석활용 모형에 내제된 과제간의 관계성 분석 기능을 활용하여 정부 연구개발 사업간 추진 중인 과제들의 관계성 혹은 유사한 과제가 수행되는 정도 등을 분석할 경우 활용성이 제고될 것으로 예상

신약개발 연구과제DB 지속 구축 및 관리에 소모되는 비용 및 노력 절감

- 신약개발연구과제DB 구축 및 관리를 위해 매년 일정수준의 연구비와 전문가 네트워크 관리·운영에 노력을 기울이고 있음
- 사용자의 기대정도에 따라 동 모형을 업무 활용에 차이가 있을 수 있으나 지속 검증 및 훈련데이터 추가 확보 필요
  – 학습정보로 활용된 신약개발연구단계DB는 전문가 자문을 중심으로 구축된 것이나 분류기준, 연구과제 정보 등을 감안하였을 때 전문가를 활용하더라도 완벽한 분류가 시행되기 어려움
  – 예를 들어, 복수의 질환을 타깃하거나 신약개발단계에서 여러 단계에 걸쳐 연구를 수행하는 과제는 전문가별 과제를 해석하는 시각차가 존재하여 분류 결과가 상이한 경우가 있었음
  ※ 16년도 신약개발연구과제DB 500개 계속과제 기준 의약품종류 19.2%, 신약개발단계 23.4%, 대상질환 6.0%
- 대상질환의 분류결과는 최우선예측(80.9%), 우선예측(92.5%)로서 실무적 활용이 가능하다고 판단됨
  - 의약품종류 및 신약개발단계 분류모형의 경우는 최우선예측은 50% 수준이나 우선예측이 70% 이상의 예측율을 보이는 점을 감안하여 업무 개선효과 방안을 마련할 필요
- 따라서, 당장의 활용은 동 분류모형으로 우선적 분류를 시행하고 해당 결과를 전문가에서 감수의뢰함으로써 정부 신약개발연구과제DB 구축 업무로드를 줄이는 것이 가능할 것으로 사료됨
  – 전문가의 입장에서도 일정수준의 신뢰성을 확보한 분류가안이 있을 경우 분류 정확도 및 업무 효율이 향상될 수 있을 것으로 판단됨

타 기술분야 응용연구 등 동 연구 방법론을 바탕으로 후속 연구성과 창출

- 과학기술용어 기반 자연어 처리 모형을 바탕으로 향후 타 기술분야로 확대 적용이 가능할 것
- 현재 우리나라 과학기술은 10대 기술분야를 중심으로 투자방향 및 예산배분조정이 진행 중이며 해당 기술분야는 모두 과학기술지식정보를 근간으로 하고 있기 때문에 동 연구방법론 적용이 가능한 상황
  - 자연어처리, 바이오의료분야 연구과제 유사·중복의 연구방법론은 해당 기술분야 과학기술지식정보 학습을 통해 연구영역 확대 가능
  - 의약분야 과학기술지식정보 분류모형의 경우 동 연구진이 구축한 신약개발연구단계 DB를 활용하였기 때문에 다소 특수성이 있으나,
  - 분류결과를 감안할 때 기술분야별 지속 관리가 필요한 핵심 중분야에 학습정보를 구축함으로써 향후 업무에 필요한 정보를 발굴할 수 있을 것으로 기대

동 과제모형은 전문가 의사결정지원모형으로 판단되며, 따라서 일정 수준의 한계점이 있음을 숙지할 필요

- 모든 잠긴 문을 열 수 있는 만능열쇠가 존재하지 않듯이, 동 과제 모형도 최신 기계학습 방법을 적용한 모형일지라도 한계점이 존재
- 사전에 등재되지 않은 단어의 경우 유사성, 연관성 분석에 한계를 보이는 등 추가적인 개선점이 분명히 상존함
  - 사전에 과학기술용어를 추가하는 것도 분명 사용자 입장에서는 실무적 활용에 또 다른 불편을 낳을 수 있음
  - 가령, 과학기술지식정보 등록 시 연구자가 제시하는 과제별 대표 키워드(국문, 영문)를 취합, 모형 학습 시 이와 같은 문제를 최소화 할 수 있을 것으로 판단되나 제시된 키워드에는 오탈자가 포함된 경우가 있어 사전 정제 작업이 요구됨

  당초 유사과제를 판단할 때 흔히 사용하는 유사단어의 출현빈도에 기반을 두지 않기 때문에 적절한 수준의 입력정보가 충족되지 않을 경우 결과가 좋지 않을 수 있음

- 소수의 핵심키워드 중복성만 가지고도 판단이 가능한 유사과제의 경우 간단한 기계학습 방법론으로 분석 수행이 가능
- 내재된 공간상에서 벡터값 기반의 유사도 측정은 해당 입력내용에 유의미하지 않은 내용이 많이 첨가될 경우 공간상에서 적정한 값으로 표현되지 않아 분석결과가 좋지 못할 수 있음

● 이는 실질적으로 최신 기계학습 방법론을 적용한 모형이 인간과 같이 문맥이나 핵심키워드, 행간에서 정보를 이해한다기 보다 인간이 입력한 질문에 대한 답을 과거보다 정확성 높게 찾아준다는 시각으로 보는 것이 타당

- 모형이 도출한 결과를 바탕으로 사용자의 실무적 판단이 수반되어야 함

■ 훈련 데이터에 국한되어 있는 해결과제*의 경우 분석활용 모형성능 향상관점에서 긴 호흡을 갖고 접근 및 해결 필요

* 훈련데이터 부족 혹은 일부 내용에 데이터가 편향된 경우

● 바이오의료분야 연구과제간 유사과제(관계성) 분석 모형은 과학기술지식정보의 공개범주에 따라 연구책임자 성명 등을 관계성 분석에 활용할 수 없었음

● 의약과제 분류모형 훈련정보의 경우 소분류를 기준으로 일부 코드에 데이터가 집중되거나 상대적으로 과제 정보가 매우 부족한 경우가 있음

● 이러한 경우는 분석활용 모형이 아닌 훈련데이터에 국한된 문제로 장기적으로 해결이 가능할 것으로 예상

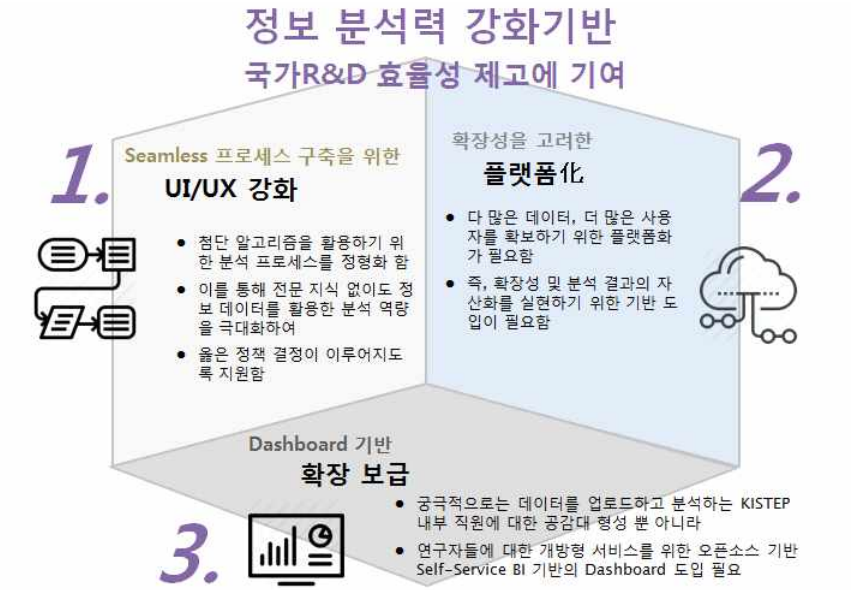
- 개인 정보와 관련되어 공개가 어려운 정보의 경우 유의미성을 가지는 타 변수를 발굴하여 향후 분석활용 모형에 반영할 필요
- 의약품종류 중 유전자치료제, 세포치료제는 바이오기술 발전으로 과거에 비해 최근 개발이 활성화된 분야인 만큼 향후 더 많은 데이터가 축적 될 것으로 기대

현재로서는 학습파라미터를 조정하는 기술적 방법을 적용하여 분석·활용 모형의 성능을 제고하는 것이 최선의 방법으로 판단됨

- 다만, 학습파라미터의 조절 여부에 따라 모형의 학습 및 구동 시 보다 많은 시간이 요구될 수 있음
- 또한, 무한정으로 학습파라미터를 확대한다고 하더라도 기본적으로 훈련데이터의 양이 증가되거나 질이 개선되지 않으면 모형의 성능 제고에는 한계가 있음
- 가령, 현재는 윈도우 사이즈를 10-15, 벡터차원을 300차원 정도로 적용하여 모형을 개발하였는데, 이 이상으로 파라미터를 조정하였을 때 모형의 성능이 크게 향상됨을 체감하지 못하였으며 학습 및 구동 시간만 늘어나는 경향이 있었음
- 학습파라미터 조절 이외에 훈련데이터에서 계속과제를 조정하는 방법으로 학습 시 훈련데이터 쏠림현상을 일정수준 방지할 필요가 있음
- 훈련데이터가 부족한 분류의 경우 현재와 같이 전문가 및 실무자의 판단이 개입되어야함

■ 궁극적 동 과제 모형은 일회성 사용에 그치지 않고 사용자의 의지에 따라 지속적 업데이트 및 성능향상이 가능한 점을 감안, 추가 모형 발굴, 타 빅데이터 융합, UI/UX 강화, 플랫폼화 등을 통해 활용성을 극대화 할 필요

- 연구과제간 관계성 분석 이외에 정부 연구개발사업 수준에서의 관계성 분석 모형이 개발될 경우 예산배분조정 업무에 활용성이 강화될 것으로 전망
- 과학기술지식정보 외에 공개적으로 접근이 가능한 타 바이오의료 관련 빅데이터를 통합할 경우 모형 활용 방안 및 도출할 수 있는 분석정보도 다변화 될 것으로 기대
- 현재는 커멘드 형식의 입력, 특정 사용자 중심으로 모형이 구축되어있으나 향후 범용적 활용을 위해서는 UI의 친밀성을 강화하고 타 기술분야 적용이 용이하도록 플랫폼화가 요구됨



[그림 5-1] 바이오의료분야 과학기술지식정보 분석·활용 모형 개선 방안

참고문헌

마쓰오 유타카, 인공지능과 딥러닝 - 인공지능이 불러올 산업구조의 변화와 혁신, 동아엠엔비, 2015

제이슨벨, 머신 러닝 워크북, 길벗, 2016

김의중, 알고리즘으로 배우는 인공지능, 머신러닝, 딥러닝 입문, 위키북스, 2016

가마타 마사히로, 처음 만나는 파이썬, 제이펍, 2017

장기정, 2015년 신약개발 정부 R&D 투자 포트폴리오 분석, KISTEP, 2017

최근우 외, 딥러닝 기술의 이해와 연구개발 정책과제, KISTEP, 2016

홍세호, 국가연구개발사업 유사·중복 검색 시스템 개발을 위한 실증 연구, 2013

박항식, 제4차 산업혁명시대의 신(新)보건의료서비스 R&D 모델정립 및 프레임웍 연구, KISTEP, 2017

G. James, D. Witten, T. Hastie, andR. Tibshirani, An Introduction to Statistical Learning, Springer, 2013.

I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, The MIT Press, 2016.

K. P. Murphy, Machine Learning: A Probabilistic Perspective, The MIT Press, 2012.

T. M. Mitchell, Machine Learning, McGraw-Hill, 1997.

R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, John Wiley&Sons, Inc., 2001.

C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, Springer, 2001.

부록

제4차 산업혁명시대의 신(新)보건의료서비스R&D 모델 정립 요약 및 정책제언

부록 신(新)보건의료서비스 R&D 모델정립 및 프레임웍 연구[18]

가. 신(新)보건의료서비스 R&D 모델 정립 프레임워크 연구 결과 요약

■ 인공지능을 활용한 분석 기술의 발전과 다양한 의료 정보의 축적 및 활용으로 의료 산업이 급변하고 있음

● 의료서비스는 과거 증상기반의 의료에서 증거기반 의료를 넘어 인공지능을 활용한 정밀의료로 패러다임이 변화하고 있음

– 이러한 의료서비스의 변화과정에서 의료산업은 다양한 데이터를 축적하였고, 인공지능 기술에 활용되면서 의료의 새로운 가능성이 열리고 있음

– 인간 유전체 정보의 해독은 의료의 새로운 가능성을 제시하였으며 개인건강 기록, 식습관정보 등과 결합하여 맞춤·정밀의료가 가능해지고 있음

– 또한, 스마트폰, 스마트워치 등 다양한 모바일 생태계가 열리면서 개인의 건강데이터의 수집이 가능해져, 이를 활용한 다양한 의료서비스가 출현하고 의료 패러다임의 변화가 가속화

■ 4차 산업혁명으로 인해 의료서비스 산업의 범위가 확장되고 의료서비스의 생활화가 가능해질 것으로 예상됨

● 4차 산업혁명으로 인한 다양한 기술적 변화로 의료서비스산업의 경계가 확장될 것으로 예상됨

– 전통적 의미에서 의료서비스는 진단과 치료가 핵심이었으나, 최근에는 항노화, 미용, 건강관리 등과 같은 영역의 중요성이 높아지고 있음

● 사회 인구구조 변화와 과학기술의 급격한 진보로 4차 산업혁명 기반 헬스케어 혁신이 부상하고 있으며 이를 위해 각국 정부 및 민간 기업의 움직임 가속화

– 나노센서, 웨어러블 기술, 장시간 배터리 기술, 인공지능 및 클라우드 컴퓨팅까지 다양한 기술과 방법으로 헬스케어 분야에 새로운 제품 및 서비스 출시 증대

18) 제4차 산업혁명시대의 신(新)보건의료서비스 R&D 모델정립 및 프레임웍 연구, KISTEP, 2017, 결과 요약

- 미국, 영국, 유럽, 일본, 중국 등에서 대규모 프로젝트를 시행하고 있으며 이를 효과적으로 운영하기 위해 정부내 새로운 조직을 출범
- 글로벌 기업 및 벤처기업들이 헬스케어 산업의 성장성을 기대하고 대거 진입

- 4차 산업혁명은 의료서비스산업의 영역을 확장할 뿐만 아니라, 의료서비스의 생활화를 통해 의료서비스의 패러다임을 변화시킬 것으로 예상됨
  - IT 활용 극대화와 웨어러블 기기의 발달로 실시간 건강관리 및 질환 대처가 용이해져, 의료서비스의 핵심이 과거 질병 발생 이후 증상 치료에서 앞으로는 질병 예측과 예방으로 옮겨갈 것임
- 특히, 헬스케어와 ICT의 융합으로 만들어질 정밀의료(Precision Medicine)는 4차 산업혁명의 중심에서 미래의료의 패러다임 변화를 가속화할 것으로 예상됨
- 정밀의료의 확산으로 치료가 아닌 예방 중심의 의료, 표준 치료가 아닌 개인 맞춤형 치료중심의료가 의료서비스의 핵심으로 부상될 것임

■ 제4차 산업형명의 기술적 변화는 현재의 보선산업의 영역을 규정하는 경계를 허물어갈 것임

- 전통적으로 강한 지역 기반의 특징을 지닌 의료서비스 산업도 국경을 넘어 이동하는 환자의 증가 등으로 그간의 산업의 국경이라는 개념도 변화하고 있음
- 원격 로봇수술, 원격 모니터링 등 시간과 공간의 경계를 넘어선 서비스 모델이 등장하고 있음
- 이러한 경계의 소멸은 헬스케어 영역의 전례 없는 확장으로 이어지고 있음

■ 4차 산업혁명으로 인한 미래 의료 기술 발전은 산업적인 가치뿐만 아니라, 고령사회라는 한국의 당면문제에 대한 해결 방안으로서도 매우 중요

- 한국의 의료비 지출액은 평균 OECD 국가 의료비 지출액 대비 빠르게 증가하고 있으며, 향후 고령화가 진행되며 더욱 가속화될 것으로 예상

  ※ 의료비 지출액 연평균 증가율('11~'15): 한국 6.9% / OECD 평균: 3.5%

- 기계학습 기반의 의료는 의료비를 경감시키고, 건강수명(유병기간 제외 기대수명)을 연장함으로써 고령사회의 여러 문제를 해결하고 삶의 질을 높일 수 있음

■ 4차 산업혁명 시대 도래하기 전의 인구고령화는 수명연장, 출산율 감소, 신고령층의 등장 등 많은 사회문제들이 결합되어 단순한 정책으로 해결하기에 어려움이 있음

- 인구의 고령화로 인해 발생될 잉여인력의 감소, 생산성 하락 등

나. 정책제언

1) 한국형 4차 산업혁명 대응전략 및 보건의료서비스 R&D 추진전략 수립

■ 세계 각국은 자국의 강점을 극대화하는 방향으로 4차 산업혁명 대응전략을 전개

- 미국은 인공지능 등 소프트웨어 플랫폼 역량, 독일은 제조시스템 역량, 일본은 로봇, 센서 등 기기·부품 역량을 적극적으로 활용

■ 일본의 4차 산업혁명 개념 이해와 대응방식 등 내용 측면에서 벤치마킹 필요

- 일본은 제4차 산업혁명의 여러 기술적 동인을 고려하되 “데이터의 확보와 활용”이라는 측면을 핵심으로 이해하고 이를 토대로 일관된 대응책을 강구
- 특히 고령과, 자연재해 등 일본 사회가 지닌 문제와 기존 신업의 강점을 분석하여 자국에 맞는 독특한 대응 전략을 수립하려고 노력
  - 일본이 직면한 문제를 해결하는 과정에서 새로운 산업을 창출하고 경제성장을 이루어 사회와 경제의 두 마리 토끼를 잡는다는 발상임
  - 독일의 Industrie 4.0 전략이 제조업 강국이라는 강점을 살리면서 IT가 약한 취약점을 보완하듯이 일본도 산업 발전의 발목을 잡지 않도록 규제 개혁 시기를 미리 설정

- 우리나라의 R&D정책은 선택과 집중이라는 방향성은 옳았지만 지속성 부족이라는 고질적인 문제를 안고 있음
  - 새 정부가 들어설 때마다 새로운 성장동력 산업과 기술을 선정하고 집중 지원하는 방식을 채택
  - 그러나 신성장동력 산업과 기술은 불과 몇 년 단위로 만들어낼 수 있는 것이 아니기 때문에 국가 D&D의 성과가 크지 않았음
  - 미래전략기술 분야를 선정하고 활용하기 위하여 기술 로드맵을 작성하고 R&D프로젝트를 기획하였으나 위원회 중심의 한정된 전문가 풀에 의존하였음
    - 기술분야별 위원장과 위원을 선임하여 '위원회 중심'으로 프로젝트를 기획하는 형태인데, 이 방법은 한정된 전문가의 지식과 경험에 의존도가 높았음
    - 따라서 기술 로드맵 수립 결과가 참여 인원의 전문 분야와 관심사항에 국한되거나 추격형 R&D로 기획되는 경우가 발생하기도 하였음

- 비효율적인 정책 수행도 R&D가 실제 성과로 이어지지 못하는 주요 요인임
  - 우리나라의 GDP대비 R&D 투자 비율은 경제 규모에 비해 기술혁신에 많은 자원을 투자하고 있지만 낮은 R&D 생산성이 약점임
  - 유망 산업에 대해서는 여러 부처가 중복적인 R&D를 진행하고 의사소통이 원활하지 않아 불협화음을 일으키는 경우가 대부분임
  - 또한 인내심이 필요한 기초기술과 원천연구보다 성과에 집착하고 있으며 연구비 지원이나 평가제도가 비효율적으로 운영되는 문제점이 노출됨
    - 유럽의 R&D는 목표가 10일 때 5 정도의 성과를 거둔 상황일 경우, 최종 목표를 위한 추가 지원방안이나 그간의 문제점과 개선책을 고민하는 반면, 우리나라의 평가제도는 성과를 달성하지 못함을 기준으로 R&D 과제 수를 축소하는 경향이 있어 문제해결에 소극적임
    - 또한, 과제 결과가 나온 R&D 프로젝트의 경우, 이미 완결된 것으로 평가하여 제반 환경변화에도 불구하고 같은 주제로 R&D 과제를 진행하기 어려운 상황임

■ R&D 시스템의 연속성, 책임성, 전문성 혁신

- 시스템 설계 역량 및 플랫폼 역량 부족이 심각해지고 있으나 이를 어떻게 확보할 것인지에 대한 국내 논의는 이제 시작 단계가 불과함
- 보건의료서비스 플랫폼 확보 여부가 경쟁의 성패를 좌우하는 핵심요소 연구에 필요한 기간을 단기와 중장기로 나눠 연구를 기획할 필요가 있음
  - 프로젝트에 소요되는 시간을 중·단기(3~5년) 및 장기로(10년 이상) 나눠 설정하여 우수한 생각들이 실현되도록 국가 차원에서 장려할 필요가 있음
- 기업은 연구개발 실패 위험부담이 큰데 비해 학교 혹은 연구기관은 창의성 있는 연구를 진행하여 많은 실패를 거칠 수 있다는 점을 적극 활용할 필요가 있음
- 후불제 연구비 지원과 같은 형태로, 당초에 계획되지 않은 연구개발사업일지라 하더라도, 기업이 성과를 낸다면 연구개발에 소요된 비용을 일부 보전해주는 형태의 지원방안도 기업이 연구개발에 적극적으로 참여하도록 유도하는 방안이 될 수 있음
- 예산삭감 등을 목적으로 하는 정부의 연구개발 과제 감사 활동은 연구 진행의 방해요인이 될 수 있으므로, 연구의 연속성을 보장하는 방안이 병행되어야 함
- 단기에 전략과 집중 분야를 결정하고 자원을 투입하려는 정부주도 마인드 대신 민간이 자생적으로 산업생태계를 키워나갈 수 있도록 환경을 조성하는 정부의 인내심이 필요

2) 4차 산업혁명에 대한 간결하고 입체적인 이해와 접근

■ 4차 산업혁명이 경제·사회에 미치는 영향에 대한 종합적이고 입체적인 전망보다는 고용 감소 여부 등 부정적 효과에 치중

- 인공지능·로봇에 의한 자동화가 고용에 미치는 영향에 대해 많은 연구가 발표되고 있으나 직업의 몇 %가 대체될 것이라는 불안 마케팅 성격의 결과가 대부분이며, 기술과 고용간의 관계, 기술의 발전에 따른 삶의 질의 변화 등에 대한 다각적인 관점의 검토는 미흡
  - 경제·사회적 영향을 객관적으로 진단하는 것이 중요하며, 위협을 과대평가할 경우 사회적으로 4차 산업혁명을 수용하기보다 저항하려는 움직임이 형성될 수 있음

- 일본은 제4차 산업혁명의 핵심 동인을 “데이터”로 이해하고 전 산업에 걸친 일관된 대응을 추구
  - 제4차 산업혁명 개념을 주장한 클라우드 슈밥은 현재의 유망 기술을 모두 모아놓은 23개 기술적 동인을 제시하여 4차 산업혁명의 개념의 명확한 이해를 방해한 측면이 있음
  - 일본은 제4차 산업혁명의 핵심기술을 사물인터넷, 빅데이터, 인공지능, 로봇으로 규정하고 4차 산업혁명을 이 공통 기반 기술들이 부문별 기술(금융, 의약, 생산 등) 및 데이터와 결합하여 새로운 제품과 서비스가 만들어지는 현장으로 간략하게 정리하여 “데이터: 중심의 일관된 전략을 추구”
  - 제4차 산업혁명이 미칠 영향에 대해 논의는 알파고 쇼크를 통해 심화되었는데, 주로 기술 고용과 일자이레 미칠 부정적 영향에 대한 논의에 집중된 경향이 있음
  - 이에 비해 일본은 전체 산업구조에 미칠 영향도 정량적으로 분석함으로써 보다 종합적인 정책적 대응의 토대를 마련하였음

- 4차 산업혁명이 사회에 미치는 영향을 입체적으로 전망할 필요
  - 4차 산업혁명이 일자리를 몇 %를 없앨 것인가를 단편적이고 수동적인 관점으로 막연한 불안감을 조성하는 대신 종합적이고 능동적인 관점에서 대응 방안을 모색할 필요가 있음
  - 고용 및 업무 성격 변화, 신규 사업자와 기존 사업자 간 갈등, 기존 제도와의 모순, 인간의 행태, 사회 인식과 구조 등, 4차 산업혁명의 사회적 영향에 대한 체계적인 조사연구가 필요
  - 인공지능 등 핵심기술을 둘러싼 산업구조를 전망하고 의존도 심화 등 다양한 시나리오에 안정적으로 대응하기 위한 방안 마련

보건의료서비스의 환경변화에 대한 다각적인 예측과 긍정적 비전을 제시하는 접근 필요

- 특히, 보건의료서비스 R&D 분야는 다양한 사회현상과 변화에 따라 수시로 변화하게 되므로, 다양한 관점에서 접근이 필요한 부분임
- 우리나라 국책연구기관의 연구결과물들은 특허, 논문, 세미나 등의 행태를 통해 데이터화 되지만, 연구 과정에서 체화된 지식과 경험은 데이터로 남지 않고 연구원 개개인의 머리에 남아있음
- 일차적으로는 형식적 지식이라도 연구소 간에 자유롭게 공유할 수 있어야만 암묵적 지식과 경험을 공유할 수 있는 문화와 시스템의 구축이 가능함
- 우리나라의 출연연구기관들은 20여개의 독립법인으로 구성되어 융합연구가 어렵고, 융합연구과제를 만들어서 연구기관 간 융합연구를 활성화하기 위한 물리적인 노력에도 불구하고 역동적인 융합혁신을 하기에는 부족한 실정임
- 독일의 연구기관 시스템을 참고한 R&D 과제별 연구 유닛의 활용이 필요함

3) 미래 신기술의 실증의 장 마련을 위한 정부 차원의 규제제도 패러다임 변화

- 현행 규제제도는 새로운 아이디어나 비즈니스 모델이 기존의 규제에 맞는지에 대한 불확실성이 존재하고 비즈니스 성공을 판단하기까지 오랜 기간이 소요됨
  - 우리나라의 규제방식은 포지티브 규제방식으로 새로운 아이디어나 비즈니스 모델이 나타날 경우 기존의 규제방식으로 제어할 수 없기 때문에, 새로운 규제를 정립하는 과정이 오래 걸림
  - 미국, 독일, 일본 등 4차 산업혁명 선도국들은 인공지능, 자율주행 자동차, 사물인터넷과 같은 첨단 산업 분야에서 급속한 기술 변화에 발 빠른 대응을 보이는 반면, 우리나라는 속도전에서 이미 뒤처져있음
    - 과거 정부가 신산업을 육성했기 때문에 융합 신산업 분야의 규제 유형이 법·제도가 기술발전을 따라가지 못하는 '후행성 규제', 융합제품에 대한 인증체계 부재 등 '근거법 미비', 각종 인허가제도 등 '복잡한 행정절차'가 대부분을 차지하고 있음
    - '예비타당성 조사'와 같은 예산 확보에 수년이 소요되며, 새로운 아이디어나 비즈니스 모델에 대한 적용 가능한 규제의 부재 등으로 인해, 혁신과 아이디어가 새로운 산업 동력과 비즈니스로 이어지지 않았기 때문임
    - 즉, 혁신의 조건은 갖춰놓고 느리게 움직여 결국 타이밍을 다른 국가와 기업들에 빼앗기는 실책을 저지르고 있다고 할 수 있음

- 신기술이 창출하는 산업은 그것이 대체하는 기존 산업의 반발을 불러일으키며 정부규제가 중재 역할을 하지 못할 경우 산업 성장이 지연
  - 영국의 적기조례와 미국의 차량공유서비스에 대한 기존 사업자의 반발로 인한 혁신 지연사례를 반변교사로 활용해야 함
    - 영국에서 자동차가 보급되기 시작하던 시기에 철도산업, 마차주들의 반대로 인해 기관차량조례(Locomotive Act) 일명 적기 조례(Red Flag Act)가 1865년 제정
    - 이 조례에 따르면 자동차는 운전수, 기관원, 붉은 기를 가지고 차량의 60야드 전방을 걷는 사람의 3명으로 운용하고, 교외에서는 6km/h, 시가지에서는 3km/h로 차량의

운행속도를 제한하였으며, 영국의 자동차 산업이 독일, 프랑스에 뒤처지는 결정적 원인으로 작용

- 1914년 로스엔젤레스에서는 개인이 소형버스(Jitney)로 승객을 실어 나르는 서비스가 등장하여 인기를 얻었고 175개 도시에서 62,000대가 운행되었으나 전차 사업자들의 반대에 부딪혀 1919년 중단되었음
- 우버(Uber) 서비스로 대표되는 차량을 공유서비스는 등장했다가 사업자의 반발로 사라진지 100년여 만에 다시 등장한 서비스가 되었음

- 기술 발전에 따른 충돌은 늘 있어왔으며, 기술 진보 속도가 빨라지면서 앞으로는 충돌이 더욱 심화될 것으로 예상됨에 따라 정부가 갈등 조율을 얼마나 기민하게 하느냐가 중요한 변수로 작용
  - 인공지능, 자율주행 자동차, 드론 등으로 인해 일시적으로 일자리가 줄어들면서 반발이 더 거세질 확률이 높음

■ 일본은 Society 5.0의 가치를 극대화하는 제도로서 규제 샌드박스 제도를 도입 'Try First'의 신속한 의사결정 체계를 마련

- 일본은 지금까지 세계의 과학기술이나 기술혁신에서 뒤쳐진 원인을 현행 규제정책이라고 판단하고 규제정책의 개혁을 추진하고 있음
  - 일본은 초기 핀테크 정책에 대응하기 위해 규제 샌드박스를 검토하였으나, 2016년 11월 산업구조심의회에서 규체적인 규제 샌드박스 도입을 언급한 이후 국가전략에 포함
  - 국가전략특구의 법 개정(2017.3)에 구제 샌드박스의 내용이 포함되었고, 이는 'Try First'로서 일본이 이루고자 하는 제4차 산업혁명의 혁신을 극대화하는 제도로서 중요하게 인식
- 일본에서 규제개혁을 위해 도입한 '목표역산 로드맵', 지역특구 등의 제도는 한국형 제도 개선 방안의 대ㅎ안으로 검토할 필요가 있음
  - 목표역산 로드맵은 정부가 규제 완화의 목표와 시간을 정하고 이를 통해 각 지자체에 규제 완화를 지시하는 것으로 기업은 이 로드맵으로부터 규제 완화를 예측하고 연구개발이나 시설 투자를 결정할 수 있음

■ 미국은 기존 규제 체계를 적용하기 어려운 신제품이나 서비스, 신산업에 대해서는 별도의 인허가 절차나 심사를 밟도록 하는 '혁신통과(Innovation Pass)'제도를 운영

- 의료기기 개발과 동시에 평가에 필요한 기준을 FDA와 기업이 함께 만드는 방법으로 새로운 의료기기 개발을 촉진하고 향후 발생 가능한 진입장벽을 사전에 막기 위한 취지로 마련되었음
- 사용자가 일상생활에서 해당 신제품을 미리 체험하는 '리빙 랩' 제도 또한 규제적용과 개선을 위한 방법 중의 한가지임

■ 선진입 후보완의 규제전략 도입 검토

- 여러 기술이 융복합하여 산업간 경계가 무너지고 새로운 산업이 태동하고 있기 때문에 하나의 기술을 중심으로 특정 산업이 발전했던 과거와 현재의 변화는 현격하게 다르게 나타남
- 우리나라의 규제프레임은 엄격한 Top-Down 방식으로 4차 산업혁명의 속도와 영향을 신속하게 따라가는데 한계를 지니고 있음
- 구제가 강할수록 경제의 자유와 창의력을 저해하게 되므로 새로운 제품과 서비스가 자유롭게 시장에 진입하기 위해서는 기술혁신을 가로막는 규제를 최소화하고 소비자 관점에서 서비스 경쟁을 시키는 형태로 규제의 패러다임이 변화해야 함
  - 즉, 일단 시장에 인입시킨 뒤에 관리하는 사후 규제의 적용을 적극 검토해야함
- 이를 위해 정부와 국회는 기존 이해당사자들의 거친 저항을 합리적으로 설득하고 고정해야 하며, 이를 위한 전담기구를 운영할 필요가 있음
  - 새로운 제품과 서비스가 시장에 진입하여 일자리가 사라질 처지에 놓인 근로자들을 위해 선제적인 재교육을 활성화하고, 실업자에 대한 사회적 안전망 확대가 필요함

포지티브 규제의 네거티브 규제 전환 검토

- 포지티브 규제란 '법에 규정한 것만 합법이고 나머지는 불법'으로 간주하는 규제형태임
  - 사회·경제 규모가 작고 정부가 컸던 과거에는 정부가 산업 전체를 육성하는 과정에서 관리 및 자원배분을 주도하며 산업을 통제해야 했기 때문에 허가권을 쥐는 규제형태가 적합하였음
  - 그러나, 새로운 산업들이 빠른 속도로 생겨나고 산업구조가 복잡해지면서 신산업 관련 규정과 법을 일일이 만들어야 하고, 이러한 합법화 과정에서 신산업 업체가 기존 업계의 반발에 부딪히는 등의 문제가 발생함
- 즉, 현행법에서 불법이 아닌 것은 모두 합법으로 보는 네거티브 규제로의 전환을 검토할 시점이며, 보건의료서비스와 보건의료서비스 R&D 분야는 변화속도가 빠르기 때문에 우선적으로 검토되어야 할 사항임

우리나라의 R&D 정책은 선택과 집중이라는 방향성은 옳았지만 지속성 부족이라는 고질적인 문제를 안고 있음

- 새 정부가 들어설 때마다 새로운 성장동력 산업과 기술을 선정하고 집중 지원하는 방식을 채택
- 그러나 신성장동력 산업과 기술은 불과 몇 년 단위로 만들어낼 수 있는 것이 아니기 때문에 국가 R&D의 성과가 크지 않았음
- 미래전략기술 분야를 선정하고 활용하기 위하여 기술 로드맵을 작성하고 R&D 프로젝트를 기획하였으나 위원회 중심의 한정된 전문가 풀에 의존하였음
  - 기술분야별 위원장과 위원을 선임하여 '위원회 중심'으로 기획하는 형태인데, 이 방법은 한정된 전문가의 지식과 경험에 의존도가 높았음
  - 따라서 기술 로드맵 수립 결과가 참여 인원의 전문 분야와 관심사항에 국한되거나 추격형 R&D로 기획되는 경우가 발생하기도 하였음

공공 및 민간을 포괄하는 상호협력 기반의 R&D 모델 개발 필요

- 의료법·약사법 등에 따라 산업의 경계가 견고했던 보건산업에서도 최근 새로운 제품과 서비스의 결합, 보건의료와 소비재의 경계영역 제품과 서비스 등장 등으로 그동안 인식해온 산업의 경계가 허물어지고 있는 추세
- 빅데이터·ICT·모바일기기 등과 융합을 통해 시간과 공간의 제약을 넘어선 의료서비스 모델이 등장하고 있으며, 이 같은 보건의료서비스의 확장은 4차 산업의 핵심 기술과 융합되면 더 가속될 수 있을 것으로 보임
- 새로운 시장의 출현 가능성과 기존 보건의료 서비스 시장의 성장 가능성이 증가하는 상황에서 기존의 보건의료 R&D 모델을 체계적으로 분석해야 하며, 동시에 4차 산업 기반의 융·복합 보건의료 신규 사업에 대해 경제성과 적용성이 담보된 최적의 R&D 모델 설계가 필요함