Journal of Biosystems Engineering Vol. 29. No. 1, pp. 79~96 (2004)

자 료

데이터 마이닝의 농업적 활용

농업공학연구소 정밀농업기계연구실장 / **이용범** 연구관

1. 서 론

정보와 시간은 21세기에 가장 가치있는 두 가지 자원이다. 흔히들 요즘을 정보의 홍수시대속에 살고 있다[1]고 한다. 그만큼 많은 정보를 접하며 살고 있다. 이러한 정보의 홍수속에서 살아남기 위해서는 몇 가지 기술이 필요하다. 첫째, 정보의 옥석을 구별할 수 있어야 한다. 정보의 가치를 구별해서 선별하는 일은 정보의 홍수속에서 살아남는 첫번째 열쇠라고 할 수 있다. 둘째, 정보가 담고 있는 의미를 이해해야 한다. 그러기 위해선 큰 그림을 볼 줄 알아야 한다. 셋째, 얻은 정보를 가지고 무슨 일을 해야 하는지를 알아야 한다. 무작위로 쏟아지는 정보의 홍수, 정보의 홍수 속에서 꼭 필요한 정보만을 처리하고 활용할 수 있는 선택과 집중의 지혜가 필요하다.

이러한 정보 수집이나 수집된 정보 중 꼭 필요한 또는 의미있는 정보를 가려내는 기술 중 하나가 데이터 마이닝[2](data mining)이다. 마이닝이라는 것은 쉽게 이해할 수 있듯이 방대한 양의 부스러기를 제거해야만 제대로 된 다이아몬드나 금을 찾아낼 수 있다는 의미이다. 즉, 데이터로부터 정보를 찾아내는 작업이 마치 금이나 다이아몬드를 발견하기 전에 수 많은 양의 흙과 잡석들을 파헤치고 제거하는 것과 유사하다는 것이다. 이와 같은 논리를 데이터[3]에 적용해보면 대용량의 데이터로부터 의미있는 지식을 발견하는 것이 데이터를 마이닝 한다고 할 수 있는 것이다.

데이터 마이닝을 현재 정보기술의 핵심으로 보는 이유는 의사결정을 위해 필요한 정보를 추출하고 업무에 도움을 주는 추가적인 내용을 유추해 내는 단계를 가능케 하고 효과 또한 무궁무진하기 때문이다.

근래 전 세계적인 친환경 농업에 대한 요구와 주변 첨단기술의 발달에 힘입어 새로운 농업기술로 각광받는 기술이 정밀농업[4]이다. 정밀농업에서는 전년도의 기상자료나 수확량 자료를 기반으로 현재의 작물 생육상태, 토양상태, 기상상태 등을 종합적으로 판단하여 수행해야 할 농작업의 종류와 수준을 결정하는 농작업 의사결정 기술이 필요하다. 이는 마치 숙련된 농업전문가가 수행해야 할 농작업의 종류나 시기, 수준을 결정하는 것처럼 수많은 정보 속에서 필요한 정보만을 가려내서 전문가시스템에 의한 농작업 의사결정을 내려야 한다. 바로 이러한 과정에서 데이터 마이닝 기술이 활용된다.

본고에서는 데이터 마이닝의 정의, 기법 등을 알아보고, 이러한 기법을 이용하여 실제 농업공학 분야에 이용된 사례를 통해 금후 농업시스템공학 연구의 기초자료를 제공하고자 한다.

2. 데이터 마이닝

가. 출현 배경

데이터 마이닝의 개념은 정보기술의 발달과 비즈니

1) http://www.ncs.gov/N5_HP/Customer_Service/XAffairs/Speech Service/SS98-037.htm

2) 데이터 마이닝이란 말은 대부분 통계학자들과 데이터베이스 학자들이 최근 MIS 분야와 경영분야에서 사용하는 말이다. 관련 용어로 데이터웨어하우징, 의사결정지원시스템, OLAP, 지식관리 등을 더 살펴보면 이해하는데 도움이 될 것이다.

3) 데이터란 컴퓨터로 처리될 수 있는 숫자, 사실, 혹은 텍스트라고 볼 수 있다. 오늘날 다양한 형태와 다양한 데이터베이스 안에 대용량의 데이터가 축적되어 있다. 데이터의 종류는 논리적인 데이터, 설계나 데이터 자체에 대한 데이터인 메타데이터도 될 수 있고, 농업분야에서는 포장의 속성정보나 기상정보 같은 예측 데이터들도 포함된다. 이와 같은 모든 데이터들 사이의 패턴이나, 조합, 혹은 관련성을 발견하면 이는 다시 정보로 이어지며 이러한 정보가 패턴과 미래에 대한 경향을 예측할 수 있다면 이는 다시 지식으로 전환된다.

4) 정밀농업이란, "농산물의 생산에 영향을 미치는 변이 정보를 탐색하여, 그 정보를 바탕으로 한 의사결정 및 처리과정을 거쳐, 생산물의 공간적 변이를 최소화하는 농업기술"이다.

스적 요구에 의해 시장에서부터 등장하게 되었다고 볼 수 있다. 1980년대 이후 정보기술의 급속한 발달에 근거하여 기업들은 방대한 양의 데이터를 저장하고 관리하기 위한 데이터베이스 구축에 많은 투자와 노력을 기울여 왔고 이러한 대용량 데이터베이스의 활용도를 높이기 위한 방편이 데이터웨어하우스[5]의 구축이다. 이러한 정보기술과 시스템의 발전으로 기업은 고객, 상품, 경쟁사 관련 데이터, 매일 발생하는 거래 데이터 등을 보다 손쉽고 효과적으로 관리할 수 있게 되었다.

더불어서 다양화, 개성화 되는 고객의 요구에 대한 적절하고 빠른 대응이 기업간 경쟁력의 척도가 되고 있는 가운데, 지속적인 경쟁우위를 확보하기 위해서는 신속한 전략과 효과적인 의사결정이 중요하게 되었다. 기업 운영상 생성, 수집, 관리되는 데이타 양의 증가에도 불구하고 유용한 정보의 부족으로 인한 의사결정의 어려움을 해결하고, 급변하고 세분화되는 시장의 변화 속에서 소비자의 구매 패턴 및 욕구를 분석하고 예측할 정보, 즉 데이타 웨어하우스 활성화에 따른 데이타 마이닝 구축 필요성에 대한 인식의 확산이 데이터 마이닝의 출현 배경이라 할 수 있다. 더욱이 대규모 트랜잭션[6] 시스템의 폭 넓은 활용으로 인하여 대용량 데이터를 다루는데 있어서 시간과 노력을 절약할 수 있는 것도 데이터 마이닝의 출현 배경에 한 몫을 한다고 할 수 있다.

경영층의 의사결정에 도움을 주는 고급정보를 가공하고 축적하는데 관심을 가졌던 의사결정지원시스템(DSS : Decision Support System)은, 정보화의 개념을 조직의 하부계층의 반복업무를 지원하는 자동화 업무에서 전사적인 개념으로 확장시키는 역할을 하였다. 그런데 이를 구축하기 위한 선결과제가 발견되었는데 바로 전사적인 시스템을 통합관리하는 통합데이타베이스의 구축이었다. 각 부서별로 독립적으로 운영되는 시스템으로는 경영층이 의사결정을 내리는데 별로 도움이 되지 않았던 것이다. 통합 데이터베이스 구축이 어느 정도 이루어졌을 때 발생한 또 다른 문제점은, 방대한 데이터와 정보들 가운데서 찾고자 하는 정보를 정확하고 빠르게 찾는다는 것이 아주 힘들다는 점이다. 결국 이를 해결하고자 하는 노력에 의해 등장한 개념이 바로 데이터 마이닝, 데이터 웨어하우징[7] 등의 개념이다.

따라서 각 기업들은 최적의 전략이나 의사결정을 뒷받침해 줄 수 있는 희귀성 있는 고급정보를 필요로 하게 되었으며, 데이터 마이닝은 이러한 요구를 만족시킬 수 있는 중요한 정보기술로 등장하고 있다. 이미 알려져 있고 기대했던 정보 뿐만 아니라 전혀 예상하지 못했고, 쉽게 드러나지도 않는 정보까지를 데이터베이스나 데이터 웨어하우스로부터 찾아내고자 하는 목적하에 개념적인 정보 추출 방법론인 데이터 마이닝이 시장에 등장하게 되었다. 데이터 마이닝의 활용분야는 매우 다양하며, 현재도 그 활용분야가 점차 늘어나고 있다.

농업 기술적인 측면에서 보면, 데이터의 양적 증가, 자동화된 자료 수집 및 전자도구의 일상화, 시스템구조, 프로세서 속도, 자료저장 구조의 기술적 발전 등을 배경으로 들 수 있다.

나. 정 의

마인(mine)이란 채광하다 라는 의미이며, 데이터 마이닝이란 용어는 비교적 새로운 용어이지만 그에 관련된 기술은 이미 여러 분야에서 활용되어 왔다. 데이터 마이닝은 자동화되고 지능을 갖춘 데이터베이스 분석기법으로서 1990년대 초반부터 기업 마케팅을 포함한 다양한 분야에서 소개되고 활용되어 졌다. 일반적인 데이터 마이닝은 데이터내에 존재하는 자료들간의 관계, 패턴, 규칙을 발견해내기 위해 대용량의 데이터를 탐색하고 분석하여 모형화함으로써 유용한 정보[8]를 추출하는 일련의 자동화되거

5) 데이터 웨어하우스(data wearhouse)는 모든 데이터에 관한 중앙창고라 할 수 있다. 즉, 다양한 경로를 통해 모아진 외부데이터를 주제별로 통합하여 즉시 여러 각도에서 분석하기 용이하도록 만들어진 통합 데이터베이스이다. 이 용어는 W.H. Inmon에 의해 처음 사용되었다.

6) 컴퓨터 프로그램에서 트랜잭션의 일반적인 의미는 정보의 교환이나 데이터베이스 갱신 등 연관되는 작업들에 대한 일련의 연속을 의미하는데, 데이터베이스의 무결성이 보장되는 상태에서 요청된 작업을 완수하기 위한 작업의 기본 단위로 간주된다.

7) 데이터 웨어하우징(data wearhousing)이란 실제업무에서 대용량의 데이터베이스의 활용도를 높이기 위해 데이타를 좀더 정제되고 일관성있게 통합된 형태로 쌓아두고자 하는 시도이다.

8) 여기서 의미하는 정보는 묵시적이고 잘 알려져 있지는 않지만 잠재적으로 활용가치가 있는 정보를 말한다. 농업분야의 정보를 찾아내는 방법은 어떤 특정 기법과 그 기술 자체만을 의미하는 것이 아니고, 농업 전반의 문제를 이해하고 이러한 문제를 해결하기 위하여 정보기술을 적용하는 포괄적인 과정을 의미한다. 즉 유용한 정보의 추출을 위한 방법론이라고 할 수 있다. 따라서 데이터 마이닝을 효율적으로 수행하기 위하여 시계열분석 등 각종 통계기법과 데이터베이스 기술뿐만 아니라 산업공학, 신경망, 인공지능, 전문가시스템, 퍼지논리, 패턴인식, 기계적 학습(Machine Learning), 불확실성 추론(Reasonging with Incertainty), 정보검색에 이르기까지 각종 정보기술과 기법들을 이해하고 있어야 한다.

나 반자동화된 과정이라고 정의할 수 있다. 즉, 거대하고 복잡한 데이터의 분석을 통한 새로운 지식을 창출하는 과정을 수행하기 위한 핵심적인 과정중의 하나인 관찰된 자료로부터 의미있는 패턴이나 모형을 추출하는 과정이라고 할 수 있다. 한편으로는 대량의 데이터로부터 새롭고 의미있는 정보를 추출하여 의사결정에 활용하는 작업으로 축소 정의할 수도 있다. 물론 이 과정에서 통계학과 데이터 마이닝은 구별된다[9]. 좀더 확대해서 정의하면 막대한 양이 축적된 데이타로부터 유용한 정보를 뽑아내기 위해 사용되는 모든 기술로 정의할 수도 있고, 정의나 기준이 모호한 데이타로부터 유용한 정보를 뽑아내기 위해 사용되는 모든 기술로도 정의할 수 있다. 데이터 마이닝은 하나의 분석기법을 의미하는 것이 아니라 여러 기법과 방법들의 적절한 조합으로 이루어진 일련의 과정이므로 통계적인 관점에서 간략하게 표현한다면 일련의 데이터에 대한 탐색적 데이터 분석이라고 할 수도 있을 것이다.

이러한 데이터 마이닝을 통하여 거대한 데이터베이스에 숨어있는 전략적인 정보를 발견할 수 있다. 데이터 마이닝은 흔히 정보발견(Knowledge Discoverty in Database), 지식추출(Knowledge Extraction), 정보수확(Information Havesting), 정보고고학(Data Archeology), 자료패턴처리(Data Pattern Processing) 등 으로도 불린다[10].

농업분야에 한정하여 적용해 보면, 위치별 토양 이화학성, 수확량 자료, 기상정보, 생육정보 등 농업경영자가 가진 모든 자료와 실시간으로 입력되는 정보를 포함하여 사용가능한 모든 데이타를 기반으로 드러나 있지 않은 경향, 기대하지 못했던 패턴, 새로운 법칙과 관계를 발견하고 이를 실제 농작업 경영의 의사결정을 위한 정보 및 농작업의 프로세서를 개선하는데 활용하자는 것이다.

그러나 데이터 마이닝은 다소 새로운 분야이고 지금도 발전되어 가는 과정에 있기 때문에 그 정의를 한마디로 요약하기는 매우 어렵다.

다. 관심이 집중되는 이유

현재의 데이터 마이닝에 대한 폭발적인 관심은 두 가지 요인 때문으로 볼 수 있다. 첫째는, 정보의 홍수라고 하는 방대한 양의 자료이다. 정밀농업에서 위치별 의사결정을 위해서는 전년도의 수확량, 기상상태 등의 자료를 기반으로 현재의 작물 생육상태, 토양상태, 기상상태 등을 종합적으로 판단하여 수행해야 할 농작업의 종류와 수준을 결정하는 농작업 의사결정 기술이 필요하다. 좀더 정확한 예측을 위해서는 그만큼 많은 정보가 필요하다. 근래들어 과학기술의 발전과 함께 고성능의 계측장비가 나오면서 더욱 방대한 양의 자료를 만들어 내고 있고 농업분야에도 유비쿼터스의 개념[11]과 그리드 컴퓨팅 개념[12]의 도입이 요구되고 있다. 또한 근래 도입되기 시작한 농산물생산이력시스템[13]을 위해서는 농업활동을 통해 생성

9) 통계학은 통계적 이론에 근거하여 추정, 가설 및 검증을 수행하고 예측 등의 결과에 대한 이론적인 해석을 한다. 반면 데이터 마이닝은 대용량의 데이터로부터 숨겨진 지식을 발굴해 내는 것으로 명확하게 정형화된 가설이 없는 데이터 중심의 분석 방법이다.

10) Data Mining은 크게 Computer Science 관점, MIS 관점, Statistics 관점에 의한 정의로 나누어 살펴볼 수 있다.
- Computer Science 관점 : 패턴 인식 기술, 통계적, 수학적 분석방법을 이용하여 저장된 거대한 자료로부터 우리에게 유익하고 흥미있는 새로운 관계, 성향, 패턴 등 다양한 가치있는 정보를 찾아내는 일련의 과정
- MIS 관점 : 거대한 데이터베이스 혹은 자료에서 유용한 정보를 유출하는 일련의 과정 뿐 아니라 값진 정보를 사용자가 전문적 지식 없이 사용할 수 있는 의사결정지원 시스템의 개발과정을 통틀어 지칭
- Statistics 관점 : 올바른 의사결정을 지원하기 위한 자료분석(Data Analysis) 및 모델선택(Model Selection)

11) 유비쿼터스(Ubiquitous) 개념은 지난 88년 제록스 팰러앨토연구소(PARC)의 마크 와이저(Mark Weiser)가 처음 제시한 '유비쿼터스컴퓨팅'이 그 효시다. 라틴어로 '언제 어디서나', '동시에 존재한다'라는 뜻으로 일반적으로 물이나 공기처럼 도처에 있는 자원이 언제 어디서나 시공을 초월하여 존재한다는 것을 의미한다. 보통 신분야에서는 이것을 'ubiquitous computing'이나 'ubiquitous network' 처럼 유비쿼터스화되고 있는 새로운 IT 환경 또는 IT 패러다임의 의미로 받아들여지고 있는 것이다. 즉 유비쿼터스 통신 또는 유비쿼터스 컴퓨팅이란 쉽게 말해 현재의 컴퓨터에 어떠한 기능을 추가한다든가 컴퓨터 속에 무엇을 집어넣는 것이 아니라 역으로 컵이나, 자동차, 안경, 신발과 같은 일상적인 사물에 제 각각의 역할에 부합되는 컴퓨터를 집어넣어 사물끼리도 서로 커뮤니케이션을 하도록 해주는 것이다.

12) 그리드 컴퓨팅(grid computing)이란, 인터넷에 연결된 PC나 다른 장비를 구동시켰을 때 단순히 데스크톱에 들어있는 자원에만 연결되는 것이 아니라, 훨씬 강력한 가상 컴퓨터의 자원을 공유하는 것을 의미한다. 가상 컴퓨터에는 컴퓨팅 파워, 스토리지, 애플리케이션, 데이터, I/O 디바이스 등이 포함되며 인터넷을 통한 곳이면 어디나 분산될 수 있다. 다시 말해, 그리드 컴퓨팅이란 지리적으로 분산된 고성능 컴퓨터, 대용량 저장장치, 첨단 장비 등의 자원을 고속 네트워크로 연결해 상호 공유하고 이용할 수 있도록 하는 차세대 디지털 신경망 서비스라 정의할 수 있다.

되는 모든 정보가 연결되어 있어야 한다. 더군다나 최근에는 농업그리드시스템[14] 구축까지 논의 되고 있다. 그만큼 정보의 량이 늘어나는 것이다.

데이터 마이닝의 관심을 고조시키는 두 번째 이유는, 급속히 발달한 컴퓨터와 더불어 성장한 데이터 마이닝 도구의 발전을 들 수 있다. 모든 자료의 디지털화에 따른 거대자료의 출현은 새로운 종류의 또다른 거대한 자료의 출현을 이끌었다. 새로운 종류의 자료와 더불어 발전한 기계학습에 대한 이론, 데이터베이스, 그리고 여러 가지 시각적 자료분석방법은 새로운 정보환경에서 요구되는 문제들을 거대자료를 바탕으로 효과적으로 풀 수 있는 방법을 제공한다. 뿐만아니라 컴퓨터의 용량, 자료 처리 속도, 통계 소프트웨어의 지속적인 발전으로 분석의 정확성은 획기적으로 높아지고 하드웨어의 비용 또한 크게 하락하였다. 이러한 정보환경의 변화는 컴퓨터 분야, 경영학 분야, 통계학[15] 분야 그리고 문헌정보학 분야에서도 데이터 마이닝 기법[16] 에 더욱 관심을 갖게된 배경이다.

이러한 데이터 마이닝이라는 개념과 실제 도구가 국내에 소개된 것은 그리 오래전 얘기가 아니다. 국내에서의 데이터 마이닝은 통신/금융산업의 이탈방지, 카드 혹은 휴대폰 도난/도용방지, 유통시장의 상품진열분석 등을 위해 사용되다가 현재는 크게 고객관계관리(CRM[17])의 비즈니스 업무 분석, 전략 도출, 고객 데이터 웨어하우스 구축, 고객데이터 가치 분석, 고객별 전략 실행, 고객별 전략 실행 결과 분석에 데이터 마이닝 기법이 사용되고 있다.

아직 농업분야에는 많은 연구가 보고되어 있지 않다.

라. 데이터 마이닝의 진행단계

데이터 마이닝은 데이터로부터 유용한 정보를 추출하는 프로세스의 전 과정이다. 데이터 마이닝 과정은 크게 여섯 단계로 구성되는데, 각 단계는 여러 속성으로 분할할 수 있으며, 각각 다른 단계를 반복적으로 수행하여 최적의 결과를 만들어 낸다. 최종적으로 필요로 하는 정보를 정의하고 이를 위해 요구되는 원천 데이터의 성격과 소재, 그리고 충실도를 파악하는 준비작업을 거친후 시작한다.

1) **문제정의** : 적용하고자 하는 문제의 정의와 목표결정

2) **적절한 데이터 준비** : 정의된 문제에 따라 필요한 데이터를 선정하고 준비

3) **데이터 마이닝 과정** : 준비된 데이터를 샘플링하고, 사전분석을 통해 탐색과 변형과정을 거친 후 적절한 데이터 마이닝 기법을 이용하여 정보의 패턴을 발견, 평가

4) **리포트** : 데이터 마이닝 과정에서 얻어진 결과물에 대해 사용자가 쉽게 이해할 수 있도록 문제 목적에 맞게 재표현 하는 단계

5) **의사결정** : 데이터 마이닝으로부터 얻어진 정보를 기반으로 전략이나 의사를 결정

6) **Feed-Back** : 적용 결과나 효과를 토대로 향상된 정보를 얻기 위해 데이터 마이닝의 초기 단계로 회기

---

13) 농산물생산이력(traceability)이란, 작물의 재배 또는 가축의 사육에서부터 가공, 유통(운송 및 저장까지도 포함), 판매에 이르기까지의 모든 과정(farm to table)을 소비자가 역으로 거슬러 올라가 확인할 수 있도록 각 단계별 기록을 작성하고 기록된 내용을 바코드 또는 IC 카드, 인터넷 등을 통하여 검색할 수 있는 시스템을 말한다.

14) 농업용 그리드 시스템(grid system)은 분산협조시스템이라고도 불리는데, 이론적으로 서로 묶을 수 없는 분산된 데이터를 연결하여 복수데이터로 조합하는 시스템으로 농업용 의사결정에 이용된다. 한편, 그리드라는 게 분산된 모든 자원(인적자원까지도)을 초고속 네트워크환경하에서 최대한 활용하는 협업체계라고 할 수 있기 때문에 농업분야에서 활용가능한 그리드기술은 다른 분야와 마찬가지로, 컴퓨팅그리드, 데이터그리드, 그리고 억세스그리드 기술 등이다. 현재는 바이오텍의 유전자정보탐색, 가공, 표출, 정보공유 등이 농업분야에서는 가장 유망한 분야라 할 수 있고, 기계분야에서는 첨단시스템구조 설계, 정밀농업시스템, 기상재해 대응체계, 기후변화영향평가 등 고도, 고용량 자료처리 기능이 필요한 분야는 어디나 적용가능한 차세대 인터넷기술이라 할 수 있다. 우리나라에서는 국가그리드프로젝트로 국내 지구환경시스템 구축을 위한 그리드기반의 uMeteo- Korea 프로젝트가 수행중에 있다.

15) 관심대상이 되는 자연현상이나 사회 현상으로부터 숫자화 정보인 자료를 수집, 정리, 분석, 표현하여 원하는 정보를 얻는 과정에 관련된 이론과 방법에 대한 학문을 통계학이라 한다.

16) 문헌정보학 분야에서는 도서를 찾는 검색 엔진의 효율성을 높이는데 데이터 마이닝 기법을 이용한다. 또한 근래는 사서의 개념이 바뀌고 있는데, 도서목록작업과 외부자료복사 서비스와 같은 단순한 업무는 외부로 아웃소싱 시키고 대신에 사서들은 정보공학자, 전자출판, 데이터 마이닝, 지식관리에서 주도적인 업무를 수행하고 있다. 사서들은 웹페이지의 표준화와 기술적인 분야의 표준화, 어떠한 정보를 인트라넷에 올릴 것인가에 대하여 간여하고 있다.

17) Customer Relationship Management: 고객의 관점에서 영업 및 마케팅 활동의 효율성과 효과성을 제고하여 기업 경쟁력 및 수익성을 향상시키는 고객관리방법.

마. 데이터 마이닝의 특징

데이터 마이닝 기법들은 통계학, 컴퓨터과학, 인공지능, 공학과 같은 분야에서 개발된 특징을 가지고 있으며, 대용량의 관측된 자료를 다룬다. 실험자료의 경우는 목적에 따라 여러 요인들이 통제되고 조작되어 얻어지지만, 관측자료는 통제되지 않은 상태에서 시간이 흐름에 따라 순차적으로 축적되며, 자료분석을 염두에 두지 않고 수집된 자료이다.

데이터 마이닝은 이론보다는 실무위주의 컴퓨터 중심적인 방법이다. 따라서 기존의 이론으로 해결되지 않는 문제를 강력한 컴퓨터의 처리속도와 능력을 활용하여 해결해 가는 방법을 택하고 있다. 데이터 마이닝은 경험적방법에 근거하고 있다. 많은 데이터 마이닝 기법들이 이론에 기초하여 개발되었다기 보다는 경험에 기초하여 개발되었다. 따라서 이러한 기법들은 그 특성이 수리적으로 밝혀지지 않은 것이 많다.

데이터 마이닝은 일반화된 결과를 도출하는데 초점을 두고 있다. 여기서 일반화는 만들어진 예측모형이 새로운 자료에 얼마나 잘 적용되도록 하는 것인가를 의미한다. 따라서 일반화는 데이터 마이닝 기법의 비정형성을 어느 정도 해결 또는 보완하여 주는데 도움을 주고 있다.

데이터 마이닝은 다양한 상황하에서의 의사결정을 지원하기 위해서 활용될 수 있다.

바. 데이터 마이닝 지식의 종류 및 사용 기법

1) 지식의 종류

- **연관 규칙(association Rule) 기법 :** 연관 규칙은 레코드[18]의 셋에 대하여 아이템의 집합 중에 존재하는 친화도나 패턴을 찾아내는 규칙이다. 예를 들어 수확량과 토양 특성간의 관계를 살펴보고 이로부터 유용한 규칙을 찾아내고자 할 때 이용될 수 있는 기법이 연관 규칙기법이다. 연관 규칙은 포장 속성 데이터베이스로부터 수확량 간의 관련성을 발견한다. 연관 규칙탐사는 포장의 속성데이터로부터 수확량의 연관 정도를 측정하여 연관성이 많은 속성들을 그룹화 하는 클러스터링의 일종이다.

- **일반화/요약(generalization/Summarization) 기법 :** 데이터 일반화란 데이터베이스에서 많은 관련된 데이터를 낮은 개념 레벨에서 높은 개념 레벨로 추상화시키는 작업이다. 즉, 데이터베이스 요약은 방대한 양의 데이터베이스 레코드들을 적은 양의 일반화된 대표적 표현으로 축약시키는 것을 의미하며, 자료 이해도의 증진, 자료 전달 및 저장의 효율화와 같은 목적을 위해 수행된다.

수십 만 개의 레코드를 수 개의 일반화된 형태의 데이터로 요약하는 상향식 요약기법과 가장 일반적인 가정에서 출발하여 사용자가 관심있는 세부적인 내용으로 전개해 나가는 하향식 요약기법이 있다.

- **분류(classification) 기법 :** 데이터베이스 내의 객체의 셋에 대하여 그 안에 내재하는 공통 특성을 뽑아내어 이 객체들을 서로 다른 클래스로 그룹화하는 작업을 말한다. 이와 같은 차별적인 특성은 소속 클래스를 알 수 없는 미지의 객체가 있을 때, 그 소속 클래스를 결정하는데 활용된다. 자료의 분류는 귀납적 학습문제에서 가장 많이 연구되어진 분야로서 각 클래스가 갖는 특징에 근거하여 분류하는 것이다. 분류의 목적은 입력 데이터를 분석하여 각각 클래스에 대해 정확한 표현이나 모델을 개발하는 것이다. 여기서 입력 데이터는 속성이나 특성에 대한 레코드들로 구성된다. 분류화를 표현하기 위한 방법으로는 의사결정 트리방법을 가장 많이 사용한다.

- **군집(clustering) 기법 :** 클러스터링이란 물리적 혹은 추상적 객체를 비슷한 객체군으로 그룹화하는 과정이다. 이 때 유사성에 따라 함께 모여진 개체의 셋을 클러스터라 한다. 클러스터링 작업은 먼저 필수 객체들이 셋으로 모여지고 이로부터 일련의 규칙이 유도된다. 클러스터링 기법은 어떤 목적 변수를 예측하기 보다는 토양물리성, 수확량과 같은 속성이 비슷한 정보들을 묶어서 몇 개의 의미있는 군집으로 나타내는 것을 목적으로 한다. 숲이 너무 복잡해서 전체를 파악할 수 없을 때 나무들을 살펴보아야 하듯이, 대용량의 데이터가 너무 복잡할 때는 이를 구성하고 있는 몇 개의 군집을 우선 살펴봄으로써 전체에 대한 윤곽을 잡을 수 있을 것이다. 유사한 특성을 갖는 클래스를 함께 그룹화하고 분할하는 방법으로, 어떤 그룹이 사전에 정의되어 있지 않다는 점에서 분류방법과 차이가 있다. 이 방법은 유사한 속성을 가진 객체들끼리 군집화 한다.

- **유사성 탐색(similarity search) 기법 :** 공간 데이터베이스에서 비슷한 패턴을 찾는 일은 포장정보를

18) 컴퓨터 데이터 처리에서 레코드란 프로그램에 의해 처리되기 위해 정렬된 데이터 항목의 집합이다. 하나의 파일이나 데이터 셋에는 통상 여러 개의 레코드가 들어 있다. 레코드 내의 데이터 구조는 보통 레코드를 정의하는 프로그램 언어나, 또는 그 데이터를 처리할 응용프로그램에 의해 미리 기술된다. 대개 레코드는 고정된 길이이지만 가변길이도 될 수 있다.

분석하는 데 필수적인 작업이다. 예를 들어 매년 비슷한 경향으로 변해가는 토양 화학성을 알아내거나, 비슷한 변화 추세를 갖는 토양 속성을 알아내는 일들이 이 범주에 속한다.

- **순서 패턴(sequential patterns)** : 일정한 경향을 식별해 내기 위해 일정시간 동안의 레코드를 분석하여 순서 패턴을 찾아낸다. 순서 패턴은 이와 같이 일련의 연관된 레코드 집합을 분석하여 시간에 따른 발생 패턴을 발견해 내는 작업이다. 이러한 지식추출 작업에 자주 사용되는 기법에는 신경망, 결정트리 등의 기법들이 많이 쓰이고 있다.

2) 데이터 마이닝 기법

- **신경망(neural network)** : 훈련 셋을 통하여 학습하는 비선형 예측모델로서 생물학적 신경 네트워크와 같은 구조를 갖는다.

- **결정트리(decision trees)** : 결정트리는 트리 모양의 구조로 결정의 셋을 나타낸다. 이 결정트리는 데이터 셋의 분류를 위한 규칙을 생성해 낸다.

- **규칙 귀납(rule induction)** : 통계적 중요도에 근거해서 데이터로부터 유용한 if-then 규칙을 추출해 낸다.

- **데이터 가시화(data visualization)** : 다차원 데이터 내의 복잡한 관계의 비쥬얼 해석을 제공한다.

- **발생 알고리즘(genetic algorithms)** : 유전 조합, 변이 등과 같은 과정을 사용하는 최적화 기법을 말한다.

- **클러스터 분석(cluster analysis)** : 관련된 객체의 서브셋을 발견하고 이 서브셋의 각각을 기술하는 묘사를 발견해 나가는 기법이다.

- **OLAP(on-line analytical processing)** : 대규모 데이터에 대한 동적 합성, 분석, 합병 기술이다. 여러 비즈니스 요소들의 관계를 분석하여 중요한 의사결정 정보를 제공할 수 있다.

지금까지의 이러한 기법들은 독자적으로도 쓰일 수도 있으나 성능향상 등 경우에 따라서 복합되어 사용될 수도 있다.

3. 데이터 마이닝 연구

지난 8월 미국 워싱턴 DC에서 열린 데이터 마이닝 학술대회[19](KDD 2003)에서는 월드컵 축구대회나 로보컵처럼 데이터 마이닝 알고리즘을 놓고 경쟁하는 데이터 마이닝컵(KDD cup) 대회가 열렸었다. 주어진 문제는 한 전자문서보관소에서 지난 12년 동안 수집된 29,555개의 과학 관련 논문 가운데서 "흥미있는 패턴"을 뽑아내는 것이었다. 재미있는 것은 흥미있는 패턴이 무엇인지를 정해주지 않고 참가자가 스스로 정의하도록 했다는 점이다[20].

이처럼 이제는 학술대회에서 데이터 마이닝 대회까지 열리고 있다. 본고에서는 사회과학분야, 농업분야로 나누어 데이터 마이닝 관련 연구 경향을 알아보고, 관련 논문 리뷰를 통해 데이터 마이닝 기술을 농업적으로 활용하는 기초자료를 제공코자 한다.

가. 사회과학 분야

1) 미국의 데이터 마이닝 연구 동향

정보화 혁명 이후 매일 쏟아지고 있는 정보의 양은 도저히 한 사람의 능력만으로는 소화할 수 없을 정도로 방대하다. 게다가 기존 개념의 자연 언어 소프트웨어로는 처리가 불가능한 멀티미디어 정보의 비율도 계속 늘고 있다. 이에 따라 컴퓨터과학 분야에서는 동영상 등 멀티미디어 데이터 마이닝에 대한 연구가 많은 관심을 끌며 이루어지고 있는 추세이다. 이들이 특히 주목하고 있는 분야는 엄청난 양의 멀티미디어 데이터를 하나하나 어딘가에 저장해둘 필요 없이, 스트리밍[21] 형태로 흘러가도록 놓아두면서 마이닝 알고리즘을 수행시키는 것이다.

볼티모어 소재 메릴랜드 대학교의 Hillol Kargupta 교수는 분산 환경에서의 데이터 마이닝을 전공하는 소장 학자로, 스트리밍 동영상 데이터에 대한 마이닝 분야를 집중적으로 연구하고 있다. Kargupta 교수는 최근 무선 핸드

19) www.acm.org/sigkdd/kdd2003 이번 학술대회에서는 관련 행사의 하나로 데이터 마이닝 기술의 미래와 응용 방안에 대한 패널 토의가 이루어졌다. 이 패널 토의에서는 그동안 주로 이론개발 분야에 집중해 온 데이터 마이닝 기술의 개발방향을 좀 더 실용적인 응용이 가능한 분야로 돌려야 한다는 의견이 주로 제시됐다. 예로든 분야는 생물정보공학, 전자상거래, 신용카드 등의 사기범죄 적발 등이다.

20) http://www.dailycollegian.com/vnews/display.v/ART/2003/09/23/3f6fa4c1c8aa5?in_archive=1

21) streaming : 전송되는 데이터를 마치 끊임없고 지속적인 물흐름처럼 처리할 수 있는 기술을 의미한다. 스트리밍 기술은 인터넷의 성장과 함께 더욱더 중요해지고 있는데, 그 이유는 대부분의 사용자들이 대용량 멀티미디어 파일들을 즉시 다운로드할 만큼 빠른 접속회선을 가지고 있지 못하기 때문이다. 스트리밍 기술을 이용하면, 파일이 모두 전송되기 전에라도 클라이언트 브라우저 또는 플러그인이 데이터의 표현을 시작할 수 있다.

헬드 컴퓨터를 통해 주가 정보 등 재테크 관련 자료들을 지속적으로 모니터링 하다가 미리 지정해둔 종목에서 주가 변동의 폭이 크다든지 하는 특별한 사건이 발생하면 사용자에게 곧바로 알려주는 시스템을 개발하기도 했다. 그는 화물 트럭 추적 시스템 등 무선 통신을 통해 전송되는 스트리밍 데이터에 특히 관심을 가지고 있다.

미국 정부는 9.11 테러 사건 이후 이민 관련 서류, 범죄 자료, 주가 자료 등의 데이터베이스 마이닝 기능과 함께 스트리밍 동영상과 음성 자료의 마이닝에 큰 관심을 보이고 있다. 미국 국방성은 테러 예방 관련 데이터 마이닝 연구 분야에 새 프로그램을 신설해 예산을 배정하기도 했다. 이 가운데 스트리밍 데이터 마이닝이 유망한 이유는 이처럼 많은 양의 정보를 모두 저장해 둘 만한 저장 공간이 부족하기 때문이다.

세인트루이스에 위치한 워싱턴 대학교 전자공학과 교수인 Ronald Indeck은 컴퓨터 정보의 속성이 사람과 비슷하다면서, 빈집에 이사간 사람이 며칠 사이 차고까지 짐으로 꽉 채우듯 정보도 공간이 허락하는 만큼 계속 쏟아진다고 비유하고 있다. 그에 따르면 하루에 미국 정보부가 모아들이는 정보의 양은 미 의회 도서관에 있는 장서에 들어있는 양만큼이나 많다. 웹만 해도 하루에 백오십만 페이지씩 늘어나고 있어 검색 엔진이 따라잡기가 버거울 정도이다. 이 같은 정보의 증가 속도는 18개월마다 두 배로 뛰어오르는 컴퓨터 처리 속도의 증가세[22]를 이미 추월한 상태이다. Indeck 교수는 하드웨어 측면에서 데이터베이스 기술의 속도를 높이는데 중점을 두어, 프로세싱 능력을 겸비한 신형 하드디스크 개발에 나서고 있다.

한편 정보량이 증가하는 동안 저장 장치의 용량도 계속 증가하고 있다. 4년 전의 일반적인 노트북 컴퓨터에는 신용카드 크기의 하드디스크가 포함되어 6기가 바이트 가량의 데이터를 저장할 수 있었던 반면, 요즈음의 노트북 하드디스크의 크기는 40~60기가 바이트가 보통이며 120기가 바이트도 볼 수 있다. 게다가 저장 용량은 증가하면서도 가격은 계속 떨어지고 있는 추세이다. 저장 장치의 대용량화와 저가화 추세에 이동 무선 컴퓨팅 환경이 결합하면서 "입는 컴퓨터[23]"에 대한 관심도 차세대 데이터 마이닝의 한 분야로 간주되고 있다.

데이터 마이닝 분야가 강하고 국방성이 위치한 워싱턴과 가깝기 때문에 인공 지능 기술의 군사적 응용에도 관심이 높은 볼티모어 소재 매릴랜드 대학교의 Anupam Joshi 교수도 이 분야에서 손꼽히는 학자이다. Joshi 교수는 슈퍼마켓에서 판매 증진을 위해서 어떤 제품을 어떻게 배치해야 하는지 사용자들의 구매 패턴을 분석한 예를 들면서, 데이터 마이닝의 마케팅 분야에 대한 응용 가능성에 대해 설명했다. 가장 유명한 예가 언뜻 보기에는 전혀 관계가 없을 것 같은 1회용 기저귀와 맥주 판매의 상호 연관성이다. 휴일 저녁 미식 축구 중계를 보며 맥주를 마시고 싶은 남성들이 아내가 부탁한 기저귀를 함께 사가는 경우가 많다는 사실이 데이터 마이닝 알고리즘을 이용해본 결과 드러났다[24].

그 외에 데이터 마이닝 분야에서 최근 1~2년 사이에 떠오른 주제로는 프라이버시 보호 문제가 있다. 마케팅 목적에 따라 상거래 활동을 감시당하고 있는 사용자 입장에서는 개인 정보가 노출되는 것을 꺼릴 수 밖에 없기 때문이다. 따라서 마이닝 알고리즘의 적용 과정에서 개인 데이터의 프라이버시가 침해될 가능성이 있는지, 또 그렇다면 어떻게 프라이버시 보호 대책을 기술적으로 구축할 수 있는지 여부도 중대한 관심사 중 하나이다. 이 문제는 정교한 수학적 기술을 이용해 데이터베이스를 설계함으로써, 전반적인 흐름에 대한 질의에는 답변하되

22) 이를 무어의 법칙(Moore's Law)이라고 한다 무어의 법칙은 마이크로집 기술의 발전속도에 관한 것으로, 마이크로칩에 저장할 수 있는 데이터의 량이 매년 또는 적어도 매 18개월마다 두 배씩 증가한다는 법칙이다. 이에 관한 강연을 준비할 당시인 1965년에, 고든 무어는 마이크로칩의 용량이 매년 두 배가 될 것으로 보인다고 예고했었다. 그러나, 변화의 속도는 지난 수년간 다소 느려져서, 매 18개월마다 두 배가 되는 것을 반영하기 위해 고든 무어의 승인 하에 이 법칙의 정의를 수정하였다.

23) 지난 2002년 6월, 미국 버지니아의 한 회사에서는 '방탄기능을 겸한 입는 컴퓨터'를 시판한다고 발표하였다. 자이버노트(Xybernaut)사는 오래 전부터 입는 컴퓨터를 개발하여 왔으며 이미 200여 고객에게 자그마한 입는 컴퓨터를 판매하여 왔다. 그런데 2002년 6월에는 '방탄조끼를 겸한 입는 컴퓨터'라는 보다 업그레이드 된 발명품을 개발한다고 발표한 것이었다. 관계자의 말을 인용하면, 이 방탄조끼를 겸하는 입는 컴퓨터는 방탄 기능 외에도 데이터를 수집할 수 있고 통신 기능을 겸할 수 있어 군인이나 경찰이 현장에서 긴급상황을 완벽하게 장악할 수 있도록 도와준다고 한다. 이 회사는 2002년 12월 입는 컴퓨터에 관한 특허를 획득함으로써 명실상부하게 이 분야에 대한 기술력을 인정받게 되었다.

24) http://www.hanbitbook.co.kr/developers/columns/datamining_5.html

자세한 정보 자체는 겉으로 드러나지 않도록 하는 방법으로 해결하는 방안이 연구되고 있다[25].

2) 한국의 데이터 마이닝 적용 사례 및 응용분야

다양한 산업 분야에 속한 많은 기업들은 그들이 구축한 세세한 거래정보 데이터베이스에 데이터 마이닝 기법을 적용하여 얻은 유용한 정보를 활용하므로 써 경쟁우위를 확보하기 위해 노력하고 있다.

〈유통업(Retail)〉

유통업자들은 자사가 발행한 신용카드와 컴퓨터화 된 결제시스템을 통하여 고객들의 매일 매일의 자세한 구매정보를 보유할 수 있게 되었다. 이러한 정보는 유통업자들로 하여금 여러 다른 성격의 고객 집단을 보다 잘 이해하는데 도움을 주고 있다. 이에 대한 예를 들면 아래와 같다.

- **바구니 분석(basket analysis) 수행 :** 바구니 분석은 일명 친화성 분석이라고도 하는데 고객들의 구매행위시 어떤 상품들이 같이 구매되는가를 밝혀낸다. 이와 같은 지식은 상점의 진열 전략이나 재고 전략, 판매촉진 등의 성과 제고에 활용할 수 있다.

- **시계열 패턴 조사(temporal pattern/sequences) :** 시간에 따른 구매행위에 대한 지식은 유통업자들의 재고관리 의사결정에 많은 도움을 준다. 예를 들어 "오늘 한 고객이 캠코더를 구매하였다면 이 고객은 언제쯤 별도의 건전지와 추가적인 테이프를 구매할 것인가?"와 같은 질문의 해답을 구하는데 많은 도움을 줄 수 있다.

- **예측모델의 개발 :** 유통업자들은 고객의 구매행위, 예를 들어 어떤 상품의 구매행위나 할인 행사에 참여하는 행위 등을 통하여 특성을 파악할 수 있으며 이러한 지식을 통하여 특정 고객집단을 겨냥한 효과적이고 경제적인 판매촉진책을 구사할 수 있다.

〈은행업(Banking)〉

은행은 사기행위 색출, 고객집단 분류, 라이프 싸이클에 따른 고객가치 관리(predictive life cycle management)등 다양한 분야에 데이터 마이닝을 이용하고 있다.

- **사기행위 색출(fraud detection) :** 신용카드회사의 골치거리중 하나가 남의 카드를 훔치거나 주워서 몰래 사용하는 경우다. 한 카드사는 이러한 일이 발생하는 것을 미리 방지하기 위해 데이터 마이닝 기법을 도입했다. 예를 들어 신용카드를 사용하는 패턴이 평소 사용하는 특성에서 벗어나게 되면 위험도 점수가 높아지며 이 점수가 일정 수준에 이르면 자동으로 거래가 거절된다. 만일 이 때 신용 카드 가맹점에서 카드를 사용하게 되면 그 카드사로 연락을 요청하는 메시지가 승인 단말기에 해당 전화번호와 함께 자동으로 나타난다. 이렇게 되면 카드사용자는 카드사 직원과의 전화통화로 본인 여부를 확인하는 절차를 밟아야 하고 아무 이상이 없다고 판단되면 정상적으로 카드거래가 이뤄지게 되는 것이다. 이 카드사의 분실/도난 카드에 대한 조기검색시스템은 SAS사[26]의 데이터 마이닝 기법 중의 하나인 인공신경망(Neural Network Application) 방법을 이용하여 개발됐다.

- **고객집단 분류(customer segmentation) :** 특정 고객집단을 찾아내고 이 집단만을 겨냥한 차별화된 서비스를 제공한다. 예를 들어, 어떤 상품은 자주 여행을 다니는 고객집단에게, 어떤 상품은 언제나 결제일을 잘 지키는 고객들에게 중점적으로 판매할 수 있다. 또한 은행은 고객집단 편성에 관한 지식을 이용하여 특정 판촉활동에 의하여 가장 많은 효과와 혜택을 얻게 될 지점을 찾는데도 사용할 수 있다. 예를 들면 한 고객집단이 여행과 관련하여 숙박업소나 교통수단을 자주 이용하는 특성이 나온다면 여행상품 할인권을 발송한다든지 스카이패스 등과 같은 마일리지 적립이 가능한 카드사용을 유도하는 것이다.

- **라이프 싸이클 예측 관리(predictive life-cycle management) :** 데이터 마이닝은 은행이 고객의 시간에 따른 가치를 예측하고 이에 따라 개개의 고객집단에 알맞은 서비스를 제공하는데 도움을 준다. 은행은 현재의 수익성이 높은 고객집단을 정의하고 데이터 마이닝을 이용하여 이들의 몇 년전의 공통된 특성을 발굴한다. 그 다음 이러한 특성을 지닌 현재의 고객들을 찾아낼 수 있는데 이들은 가까운 장래에 수익성이 높은 고객이 될 가능성이 매우 높은 고객들이다. 은행은 이들에게 특별한 상품 거래를 제안하거나 수수료를

25) http://www1.kisti.re.kr/~trend/Content548/computer05.html

26) 전세계 37,000개 이상의 기업, 정부 및 대학에서 사용하고 있는 소프트웨어로, Fortune 500의 100대 기업 중 98개의 기업과 Fortune 500 전체 기업 중 90%가 SAS의 고객이다.

면제해 주는 것과 같은 고객 이탈방지 프로그램 같은 것을 실시하고 있다.

〈통신산업(Telecommunications)〉

전세계적으로 점점 치열해져가는 경쟁에 직면하고 있는 통신회사들은 기존 고객을 유지하고 새로운 고객을 끌어들이기 위해 적극적인 마케팅 정책과 가격 정책을 실시하고 있다. 이러한 통신산업 분야에 데이터 마이닝이 적용된 예는 다음과 같다.

- **통화 기록 분석 :** 통신사업자들은 고객의 자세한 통화 기록을 가지고 있다. 비슷한 통화 사용패턴을 가진 집단을 찾아내어 그들에게 유리한 가격정책이나 기능 등을 개발할 수 있다.

- **고객 충성도(customer loyalty) :** 어떤 고객은 계속 통신서비스 제공자를 바꾸면서 각 통신회사가 제공하는 인센티브를 이용한다. 통신회사는 데이터 마이닝 기술을 이용하여 한 번 고객이 되면 오랫동안 지속적인 거래를 하게될 고객과 그들의 특성을 찾아내고 이들을 중심으로 가장 이익이 많은 곳에 마케팅 투자를 할 수 있다.

〈보험업(Insurance)〉

보험회사는 오랜기간에 걸쳐 집적된 방대한 데이터를 가지고 있는데 이것은 효과적인 계획을 세우는데 참고자료로 활용될 수 있다. 이러한 보험산업 분야에 데이터 마이닝 기술이 적용된 예는 다음과 같다.

- **사기 색출 :** 보험회사는 예를 들어 허리부상과 같은 높은 보험 청구율을 가진 분야의 청구자, 의사, 변호사들 사이의 관련성 또는 보험청구 패턴을 찾아냄으로써 보험사기를 줄일 수 있다.

- **상품 설계(product design) :** 보험업자는 가장 수익성이 좋은 상품 구성 즉, 보험가입 신청자의 특성, 보험증권의 보장범위 및 보험증권 특약의 최적 결합을 알고 싶어한다. 보험업자들은 이 정보를 이용하여 새로운 상품을 설계하고 장래의 판매를 위하여 기존의 상품을 고부가가치화 하는데 이용한다.

- **위험 분석(risk analysis) :** 보험업자는 보험지급액과 관련된 여러 요인들을 찾아내므로써 지급부담 위험을 줄이고 있다. 예를 들어 한 대형 보험회사는 최근 지난 2년간의 중요한 보험청구건을 검토한 결과 기혼자의 청구금액이 미혼자의 청구금액의 두 배에 달한다는 사실을 발견하였다. 이 지식을 바탕으로 이 회사는 기혼자에게 일률적으로 적용, 할인하여 주는 정책을 조정하였다.

〈그 밖의 데이터 마이닝 응용 사례〉

데이터 마이닝 기술의 응용은 다양한 여러 산업 분야에서 활발하게 이루어지고 있다.

- **고객집단 편성 :** 거의 모든 산업분야는 데이터 마이닝을 이용하여 뚜렷한 고객집단을 편성하는데 활용하므로써 많은 이익을 얻을 수 있다. 이들 조직은 데이터 마이닝을 이용하여 전통적인 데이터 웨어하우징에서 다루는 것보다 훨씬 많은 변수들을 고려할 수 있다.

- **자동차 제조 :** 자동차 제조업자들은 고객들을 위하여 맞춤 자동차를 생산하기 시작했고 따라서 어떤 특징들이 선호될 것인지 그리고 이러한 특징과 함께 어떤 점들이 요구될지를 예측해야 할 필요가 있다.

- **보증계약(warranties) :** 제조업자는 보상 청구를 해 올 고객의 수를 예측하고 이에 따른 비용을 예측하여야 한다.

- **탑승객 인센티브(frequent flier incentives) :** 항공사는 자사 비행기를 더 자주 이용할 수 있도록 인센티브를 제공할 고객 집단을 찾아내는데 데이터 마이닝을 이용할 수 있다. 예를 들면 한 항공사는 비행거리 누적 혜택을 받는데는 별 도움이 되지 않을 정도의 짧을 거리를 매우 자주 여행하는 고객집단이 존재함을 발견하였다. 따라서 이 항공사는 비행거리 뿐만 아니라 비행횟수에 의해서도 항공사가 제공하는 혜택을 받을 수 있도록 규칙을 변경하였다.

나. 농업분야

데이터 마이닝이 과학기술분야에 활용되고 있는 예는 많다. 과학기술분야의 데이터는 그 양이 클 뿐만 아니라 동시에 매우 복잡함을 가지고 있다[27]. 그러나 데이터 마이닝 기술을 활용하여 많은 부분 단순화시킬 수 있고 해석도 가능해지고 있다. 근래들어서는 천문학[28], 원

27) 멀티센서, 멀티 스펙트럴, 멀티 해상도, 시간/공간 복합자료, 다차원자료, 모의시험에 의한 격자데이터, 소음 등에 의한 데이터의 오염 등 과학기술분야의 데이터가 복잡할 수 밖에 없는 이유는 많다.

28) 은하/성단 분류(fayyad, 1996), 금성의 화산폭발 감지(Burl, 1998), 다이아몬드 눈의 발견/분석(Burl, 2001), 사파이어의 발견(Kamath, 2001) 등을 들 수 있다.

격탐사[29], 생명공학[30], 생물학과 의료영상[31], 화학분야[32], 비파괴 검사[33], 보안/감시분야[34], 전산유체역학[35], 구조역학[36], 연소공학[37], 기상[38], 등에서 가시적인 효과를 나타내고 있다[39].

이 중 본 자료에서는 농업분야에 대해서만 알아보고자 한다.

1) 인간 게놈(Human Genome) 해석

인간 게놈은 인체 세포에 존재하는 23쌍의 염색체, 이들 염색체를 이루는 DNA, 다시 이들 DNA를 구성하는 30억 개 염기쌍들의 나선형 조합으로 구성되어 있다. 대부분의 과학자들은 30억 개에 이르는 인간의 염기 서열을 다 밝히는 것은 불가능하며 그럴 필요도 없다면서 겨우 3%만 인간에게 유용한 정보이며, 나머지 97%는 쓰레기라고 주장한다. 그러나 문제는 30억 개 중 어느 부분이 바로 3%에 해당하는 것이며, 그 속에 담겨있는 유전자 코드 정보를 찾아 어떤 변이가 어떠한 질병을 발생시키는지 알아내는 것이 중요하다. 여기에서 과학자들은 대용량의 데이터에서 숨겨진 패턴을 찾아주는 데이터 마이닝을 이용했다[40].

바이오 인포메틱스[41](Bio-informatics) 분야에서도 데이터 마이닝이 많이 활용된다. 피부나 목소리 같은 생체 신호를 분류하는데도 활용되고, 유전자 분석시 단백질 분석 등 분류, 그룹화, 가시화 등에 활용된다.

29) 해양천연자원지도 작성, 유전발견, 토지이용도 작성, 수자원관리, 환경, 농업과 산림분야의 데이터 가시화.
30) 생물학과 정보공학의 가교로서 신경망, 마코브 모델 등을 이용하여 염기서열 분석, 단백질구조 분석 등의 효과를 내고 있다.
31) 잡음이 타기 쉽고 형상이 선명하지 않은 MRI, PET, 초음파, X선사진 등의 영상분석을 통해 종양발견, 변화감지, 단백질구조, 유전자 등의 차이를 구별해 낸다.
32) 분자패턴 분석, 신약 개발, 분자간의 관계 분석.
33) 교량검사, 광산탐색, 재질검사, 결함 분석, 위험요소 조사 등.
34) 지문, 홍채, 얼굴, 인장 등을 실시간으로 분석.
35) 항공기 등의 유체 분석, 교란검사.
36) 자동차 충돌 시험, 교량 등의 안전성 검사.
37) 연소실내 와류와 화학적 작용의 교호작용 분석.
38) 엘리뇨, 지구온난화 모델링.
39) http://www.ipam.ucla.edu/publications/sdm2002/sdm2002_ckamath.pdf
40) 생명과학/ SAS NEWS 2000. 08. 19.
41) 생명 현상 관련 연구에서 나오는 다양한 정보를 수집, 관리, 분석하는 데 필요한 제반 분야. 생명공학 기술과 정보기술의 융합.

2) 기상정보 분석

일기예보의 과정은 일기자료관측 - 자료의 수집 - 일기분석 및 일기도 작성 - 일기 예보의 순서로 진행된다. 우리나라의 일기자료 관측은 73개의 유인지상기상관측소에서 총 15종의 기상요소를 3시간마다 관측하고, 460개 지점의 무인자동기상관측장비를 이용하여 매분 관측하는 지상관측자료와, 항공기상종합자동관측장비를 이용한 항공기상관측, 지상으로부터 30km 상공까지의 고층기상관측, 해양기상영상감시시스템을 이용한 해양기상관측, 인공위성을 이용한 기상위성관측, 지진자동관측망을 이용한 지진 및 해일 관측, 기상레이터를 이용한 기상레이터관측, 낙뢰관측시스템을 이용한 낙뢰관측 등으로 구성된다.

이러한 관측장치를 통해 수집된 방대한 양의 데이터를 이용하여 일기 분석 및 일기도를 작성하게 되는데, 현재는 수퍼컴퓨터(SX-5)[42]에서 수치예보모델을 이용하여 24, 36, 48시간 예상 일기도를 생산한다. 이러한 과정에서 관측정보의 종합, 다중/시계열 분석을 통한 예보 등에 데이터 마이닝 기법을 활용하는 연구가 국내외적으로 수행중이다.

우리나라의 한 벤처기업에서는 전국 4백50여 개소의 기상청 무인자동기상관측시스템 모니터링 자료를 데이터 마이닝 기법으로 분석하여 10분간격으로 전국 기상정보를 제공해 주는 인터넷 사이트를 개설해 관심을 끌고 있다[43].

태풍은 일상생활에 큰 영향을 끼치는 기상현상 중의 하나다. 따라서 태풍의 정확한 해석과 신속한 예보는 매우 중요한 의미가 있으며, 그동안 이를 위해 수 많은 예보기술과 기상학적인 연구가 진행되어 왔다. 태풍의 보다 정확한 해석을 위해서는 태풍운 패턴의 시계열적 해석 부분을 필요로 하나, 이는 기상학적 접근만으로는 해결되지 않는 부분으로서 정보학적인 접근을 요하는 부분이기도 하다. 이러한 분야에 대량의 태풍영상을 모은 태풍 콜렉션(collection)을 대상으로 기상학적 지식과 정보학적 접근을 데이터 마이닝으로 융합한 태풍 패턴의 시계열 해석법을 개발하여, 예보분석 전문가로 하여

42) 슈퍼컴퓨터는 초당 연산능력이 약 10~15GFlops(1GFlops는 초당 10억 회의 연산) 이상 되며 주기억장치(RAM) 용량은 4GB 이상, 메모리간 차선개수(Bandwidth)가 약 10GB/시스템 정도 되는 것이라고 볼 수 있다. 기상용 슈퍼컴퓨터 Model: NEC SX-5/16A & SX-4/2A(세계50위) 도입: 99.6.1 초당 1,280억 번 연산 2000.6 : SX-5/8B를 추가(초당 1,920억 번 연산 :세계 30위권).
43) http://chumsungdae.com/korean/press/press.html

금 태풍자료를 분석하는 데에 보다 효율적이고 유용한 자료로 활용토록 하고자 하는 연구도 수행중이다. 동시에 패턴인식방법이나 데이터 마이닝 방법 등에 근거하여 드보락 방법[44]에 잠재된 기상학적 지식과 이것들의 기상학적 지식 체계를 시스템에 대입하여 정보학적 접근과 융합시키는 연구도 진행중이다[45].

3) 관련 논문 review

김석중 등(2000)은 의사결정 트리분석기법[46]을 이용하여 우유의 소비행태 변화에 관한 연구를 수행했다. 일반적으로 우유의 소비행태에 관한 연구는 크게 시계열[47] 자료나 횡단면자료[48], 설문조사방법을 이용하여 분석되었다. 시계열자료는 가격과 소득이 소비에 미치는 영향분석을 통하여 미래의 수요를 예측하는데 유용하게 사용되고 있으며, 횡단면자료는 미래의 수요를 예측하지는 못하지만 경제적·비경제적 요인을 모두 포함하여 계량화 할 수 있는 장점이 있다. 그러나 이러한 분석은 이러한 우유소비에 영향을 미치는 요인들의 교호작용을 설명하기 어려운 단점을 내포하고 있다. 따라서 우유의 소비행태에 영향을 미칠 것으로 판단되는 경제적 요인뿐만 아니라 비경제적 요인들을 포함하며 모든 가능한 설명변수의 교호 효과를 분석함으로써 기존의 시계열자료나 횡단면자료, 설문조사방법에서 다룰 수 없었던 소비행태의 요인들을 구체적으로 제시하고자 의사결정트리 분석을 이용하여 소비행태를 분석하였다. 연구결과 우유의 소비확대를 위하여 가구주의 연령이 50대 이후에서 우유 주소비 계층이 크게 감소하고 있는 바 노년층을 대상으로 한 마케팅 내지는 제품 개발이 필요하다고 판단되며 가구주의 학력이 저학력으로 갈수록 우유의 주소비 계층이 감소하고 있으므로 이들을 대상으로 우유의 영양학적인 효과와 필요성을 부각시킬 필요가 있다고 보고했다.

Fileto 등(2002)은 XML[49]과 데이터 마이닝을 결합하여 브라질의 청과물 유통체인에 의사결정 지원시스템을 적용했다. 이 과제의 주요 목표는 데이터 웨어하우스의 구축과 웹상에서 이러한 자료를 관리하고 필요한 의사를 결정하는 방법을 개발하는 것이다.

Harms 등은 미국농무성 위험관리서비스의 수준을 높이기 위해 데이터 마이닝 기술을 이용하여 가뭄 위험 관리를 위한 지리적 의사결정지원시스템을 개발했다[50].

Lazarevic 등은 공간데이터를 분석하고 이 데이터를 이용하여 모델링할 수 있는 소프트웨어를 개발했다. 근래 GIS기술의 비약적인 발전으로 다양한 데이터를 대량으로 수집 관리할 수 있다. 그러나, 기존의 방법으로는 다차원적이고 시간적/공간적 변이를 가진 데이터 속에서 의미있는 지식을 발견하기란 어렵다. 이러한 문제를 해결하고자 Lazarevic 등은 공간데이터 분석과 이를 통한 모델링 작업이 가능한 소프트웨어 시스템을 개발했다. 이 소프트웨어는 기계학습을 통해 통계에 대한 전문적인 기술이나 분석 없이 공간도메인 정보를 추출할 수 있다[51].

근래 그리드 컴퓨팅 분야에서 연구되고 있는 분야 중 하나가 향후 반세기 동안의 기후 변화를 예측하는

44) 드보락(Dvorak) 방법은 전 세계적으로 널리 이용되는 태풍 강도 분석법으로 인공위성 영상 등을 이용하여 여러 단계의 주관적인 분석과정을 필요로 하는 기상분석법 중 하나이다.

45) http://www.kma.go.kr/kma15/2003/contents/200307_04.htm

46) 대용량의 데이터로부터 이들 데이터 내에 존재하는 관계, 패턴, 규칙 등을 탐색하고 찾아내어 모형화하는 데이터 마이닝 기법중의 하나

47) 시계열(time series) : 시간의 경과에 따라 연속적으로 관측된 관측값의 계열을 말하는 것으로써 동일한 시간간격으로 측정되는 것인데 측정한 시간에서만 취하는 시계열을 이산시계열(discrete time series)이라 하며, 같은 시간 구간에 걸쳐 동시적으로, 또한 순서적으로 배열된 이러한 시계열의 특징은 연속적인 관측값이 대개 독립적이 아니며, 반드시 시간순서에 따라 관측값을 분석해야 한다. 이 시계열은 대개 4개의 성분(추세, 계절, 순환, 불규칙성분)으로 분해한다.

48) 횡단면 분석(cross-sectional analysis) : 동일시점에 또는 동일기간에 걸쳐 여러 변수에 대하여 관찰된 횡단면자료를 이용하는 통계적 분석.

49) XML(eXtensible Makeup Language : 확장성표기언어) : 인터넷정보는 일반적으로 HTML(Hyper Text Markup Language)이라는 인터넷 언어를 사용하여 표시하고 있으나 HTML은 다양한 신수요에 적절히 대응하기에는 표현과 운용상의 한계가 있다. XML은 이러한 문제점을 보완하기 위해서 개발된 언어로 W3C SGML 워킹그룹의 멤버인 Tim Berners-Lee가 '96년 4월 웹에서 SGML을 사용하여 문서데이터기술 및 교환을 위해 SGML을 경량화해 사용하기 쉬우며 응용프로그램을 쉽게 구현하도록 설계되었다. XML은 기존의 HTML의 기능을 대폭 향상시켜 e-비즈니스, 전자정부 등 다양한 분야에 적용시킬수 있어 차세대 인터넷용 언어로 널리 활용될 전망이다.

50) http://www.isi.edu/dgrc/dgo2001/papers/session-1/harms.pdf

51) http://www.computer.org/proceedings/hicss/0493/04932/04932006.pdf

프로젝트이다. 이 프로젝트는 영국에 있는 리딩, 옥스퍼드, 그리고 오픈 대학교들과 메트 오피스, 그리고 여러 곳의 기업들이 함께 추진하는 공동 프로젝트이다[52]

유전자 발현 패턴을 효과적으로 동정하는데 활용할 수 있는 새로운 유전자 데이터 마이닝 기수이 미국 펜실베니아 주립대와 뉴욕대, 버팔로의 과학자들에 의해 보고되었다[53].

Motonaga 등(2002)은 품질관정기능을 갖춘 농업정보 경영 시스템을 개발했다. 고품질 농산물의 안정적인 공급을 위해서는 품질관리, 재배 관리, 포장관리 등의 정보를 효율적으로 관리하는 것이 필요하다. 이 연구팀에서는 인터넷을 이용하여 농산물 품질의 디지털화와 품질관정시스템을 구현했다. 이 시스템은 데이터베이스, 데이터 마이닝 같은 분석툴로 구성된 웹 서버로 통합관리된다. 먼저 포도를 대상으로 재배, 수확과 관련된 다양한 정보를 다루는 농업정보 경영 시스템을 구축했다. 이 시스템은 작물의 생육정보와 관련된 영상과 스펙트럼 정보를 통합 관리하기 위한 데이터베이스 시스템이다. 시스템을 적용해 본 결과, 계층적 모델로 이루어진 데이터베이스를 통해서 포도의 경작 정보를 관리하는 것이 가능하다고 보고했다. 데이터 마이닝을 통한 효과적인 분석, 평가기능 때문에 기존에 이용하지 않고 있던 데이터를 이용하여 효과적으로 농업정보를 관리할 수 있음을 알 수 있었다. 이러한 모든 정보는 인터넷을 위해 쉽게 접근이 가능하다. 이러한 과정이 농산물생산이력과 접목될 수 있다.

Bertis 등은 미국 농무성 위기관리소(Risk Management Agency) 국가작물보험 프로그램의 보험판매인, 배상조정자, 보험구입자 등의 행동을 데이터 마이닝 기술을 이용하여 분석했다. 기존 연방법에 있는 보상에 관한 지표변수는 6개의 잭나이프[54] 변수를 분석하여 얻은 값의 150%에 달한다. Berties 등은 순차적이고 비순환적인 노드의 분석에 로그 모델을 이용했다. 여기에서 보험료의 남용이나 부정사용을 적발하는데 데이터 마이닝기술을 이용했다. 이와같이 농업 보험의 적절한 설계와 보험사고의 확률을 낮추는데도 데이터 마이닝이 이용된다[55].

데이터 마이닝 기법을 이용하여 반자동 토지분할시스템을 개발한 보고도 있다. 기존의 순차적 클러스터링 방법에 의해 토지를 분할 한 후 사용자의 요구에 따라 지리적인 흐름을 분석하고, 공간집합을 분석하였다[56].

산림자원을 분류하는데 데이터 마이닝 기법이 이용된 연구도 있다. 분류시스템에서 일반적으로 많이 사용되는 Wilson XCS시스템에 신경망, 의사결정 트리구조 등을 접목하여 예측정밀도를 알아보았다. 연구결과 일반적인 방법보다 데이터 마이닝 기법을 이용하면 예측정밀도가 10% 정도 높아졌다[57].

Robert 등은 농업과 원예분야의 문제를 해결하는데 적용할 수 있는 기계학습에 대해 보고했다[58].

Glenn 등(2002)은 토마토의 숙성유전자를 검색하는데 데이터 마이닝 기법을 이용했다. 연속된 유전자 정보 데이터베이스 중에서 과일의 숙성과 관련된 생물학적, 생화학적 정보를 이용하여 토마토의 영양을 개선하는 연구를 수행했다.

Lukman 등 52개국 83개 도시의 공기오염 데이터를 데이터 마이닝 기법으로 분석하여 공기오염 평가를 위한 모델, 패턴, 규칙, 경향 등을 분석했다[59].

다. 정밀농업 분야

근래들어 다양한 작물 매개변수를 측정하는 센서가 개발되고, 하드웨어의 속도가 빨라짐에 따라 연속/비연속 정보의 수집, 관리, 분석에 대한 요구가 늘어나고 있다. 그 방법의 하나로 웨이블릿 신호처리[60]와 신경망 컴퓨팅을 이용하여 농업분야 측정 데이터의 보완에 관한 연구가 많이 수행되고 있다. 즉, 농업자료의 분석에 관한 방법론적인 접근이 많이 활발해진 것이다. 특히

52) http://zdnet.com.com/2100-1103_2-5075485.html?tag=zdnnfd.main
53) http://www.psu.edu/ur/2003/dataclustering.html
54) Jackknife 분석 : 비모수 검증법의 하나로 주로 통계치의 편기를 추정하기 위해서 사용된다.
55) http://www.cse.unsw.edu.au/~qzhang/papers/p28.pdf
56) http://www.geocomputation.org/2000/GC059/Gc059.htm
57) http://www.sys.uea.ac.uk/~ajb/PDF/IJCNN2003.pdf
58) http://craig.nevill-manning.com/~nevill/publications/COMPAG.pdf
59) http://www.iwr.uni-heidelberg.de/HPSCHanoi2003/abstracts/lukman_i.pdf
60) 웨이블릿은 1983년 Morlet에 의해 소개된 이후 신호를 분석하고 해석하는데 효과적인 수학적 도구로 알려져, 순수수학분야(조화해석학, 선형대수)부터 여러 응용분야(전자공학, 컴퓨터공학, 지구과학)에 이르기까지 폭넓게 연구되어 왔다. 웨이블릿 변환은 푸리에(Fourier) 변환에 기반을 둔 기존의 신호처리 알고리즘에 비해 속도가 빠르고 시간과 주파수 영역에서 신호의 국소화를 효율적으로 구현하기 때문에 최근 신호 및 영상처리 분야에 많이 응용되고 있다.

정밀농업에서는 원격탐사 정보를 이용하여 곡물 정보, 수확량, 생육상태, 병해충 등의 다양한 정보를 해석할 수 있다. 이러한 다양한 데이터들의 신호의 처리와 특징 추출 등을 통해 의미있는 정보를 추출하는데 데이터 마이닝이 활용된다.

정밀농업에서의 데이터 마이닝은 데이터 속에 있는 보이지는 않지만 잠재적으로 활용가치가 있고 유용한 정보를 찾아내는 탐색과 분석에 대한 기술이다.

대체적으로 정밀농업과 관련된 연구는 자료의 수집(저비용 자료수집, 자료형태, 자료종류 등)과 수확량 분석(수집된 데이터에 근거한 수확량 증대)에 대한 내용이 주를 이룬다. 특히 근래는 자료의 수집에 대한 연구가 많다. 한 예로 2000년에 있었던 제5차 정밀농업학회에서 발표된 자료를 보면, 총 238편의 논문 중 139편이 자료수집에 대한 내용이고, 단지 26편만이 수확량과의 관계를 해석하는 논문이다[61].

이렇게 수확량 관련 논문이 적은 이유는 지난 수년간 연구자들이 노력한 수확량과 측정결과간의 수학적인 관계를 찾고자 하는 연구의 실패에 일부분 기인한다. 이러한 실패는 데이터 셀의 제어가 없는 단순관계 분석에 의한 당연한 결과라고 볼 수 있다. 수확량은 생각보다 복잡한 함수관계를 이용해야만 분석이 가능하다. 즉, 이는 수확량과 관련된 다중변수의 수학적 모델을 이해해야만 해석할 수가 있다(Kastens, 2000). 일본에서는 이미 1999년부터 정밀농업 포장정보관리에 데이터 마이닝 기술을 활용하고자 적응적 데이터 마이닝에 근거하는 포장정보관리 데이터베이스 시스템 개발에 관한 연구를 수행했다[62].

1) 원격탐사 정보 분석

어떤 밴드에서의 반사특성이 농업현상과 깊은 관련이 있는지를 찾고자 하는 시도가 원격탐사에서의 데이터 마이닝이다. 이러한 과정은 원격으로 측정된 데이터로부터 지식이나 정보를 추출하는데 필수적인 과정이다. 정밀농업에서 원격탐사분야에 데이터 마이닝을 적용했을 때의 잇점은, 파장별 정보와 농업 현상과의 관계를 2차원 평면으로 나타낼 수 있다는 것과, 사용자의 관심별로 농업현상을 구분할 수 있고, 비교적 고 정밀도의 분류 규칙을 만들 수 있다는 점이다. 반면 단점은 데이터 처리에 손이 많이 가고 필요에 따라 일정한 선처리가 필요하다는 점이다.

원격탐사 기술과 데이터 마이닝 기술이 만나 보다 나은 경작기술을 개발할 수 있게 되었다. 오늘날의 농부는 재배하고 있는 작물에 최적인 기계를 포함한 다양한 정밀농업 도구들을 선택할 수 있다. 이러한 장비들의 도움으로 비료나 병충해 문제를 필요한 곳에만 처방할 수 있어 시간과 비용을 줄일 수 있다. 또한 정밀농업을 통해 농약의 사용량도 줄일 수 있다.

그러나, 농부들이 긴 세월에 걸쳐 그들의 농지에 관한 많은 자료를 축적하는 동안, 수집 방법들과 자료의 품질은 계속해서 변한다. 농부들은 엄청나게 많은 양의 정보를 필요로 하면서 동시에 유용한 정보를 필요로 한다.

Lei Tian(일리노이 대학) 교수는 농민들에게 정보농업에 필요한 정보를 주기 위한 연구를 수행중이다. 농사를 짓는 농민들에게 정밀농업의 실제 데이터가 없기 때문에 실제 포장에는 20에서 30% 정도의 잡초밖에 없지만 실제로는 전체 포장에 제초제를 뿌리게 된다. 그는 원격탐사를 이용해서 작물, 병해충 등에 관한 고품질의 자료를 얻고 있다. 이로 인한 환경오염을 방지하기 위해 이들 자료에 대해 데이터 마이닝 기법을 이용하여 새로운 작업을 실천하는 의미있는 정보를 추출하고자 하는 연구다[63].

2) 공간 정보 데이터 마이닝

통계적인 공간 분석의 배경인 공간 데이터 마이닝은 공간자료를 분석하는 일반적인 방법이다. 통계분석에 관해서 수 많은 연구가 이루어 졌으며 문제해결을 위한 다양한 최적화 기법 등의 알고리즘이 존재한다. 이러한 알고리즘은 대량의 수치적 자료를 취급하고, 공간적인 현상을 현실적으로 나타내 준다. 그러나, 이러한 접근은 서로 상호 연관이 있는 공간적으로 분포된 데이터간의 통계적인 독립을 가정하는 전제를 둬야하는

61) 나머지 73편은 정밀농업 교육, 사업, 환경, 정부정책에 대한 내용임. 139편의 데이터 수집에 대한 논문을 살펴보면, 1) 토성이나 원격탐사 등에 활용할 수 있는 저가의 변이 측정 센서, 2) 최적 경영 구역, 3) 최적 격자 크기, 4) 개선된 정보수집 센서, 5) 개선된 통계분석기술 등으로 분류할 수 있다. 26편의 수확량 관련 논문은 1) 토양화학성, 비옥도 등의 변이와 수확량과의 관계 2) 변량살포와 구역크기간의 관계로 분류할 수 있다.

62) 북해도 정밀농업연구센터.

63) http://www.ncsa.uiuc.edu/News/Access/Stories/agsensing/

문제가 있다. 크리깅[64]이나 회기 모델을 이용하여 어느 정도 이 문제를 해결할 수 있지만 궁극적인 해결은 아니다. 불행하게도, 이러한 방법은 문제를 더욱 복잡하게 만들고, 더욱 고도의 통계적인 지식을 요구하게 된다. 더군다나 비선형 규칙들은 모델화 할 수 없다. 다시말해, 공간 자료를 분석하기 위해 실제 사용자가 접근해야 할 방법이 아니라는 것이다. 이런 상황에서 필연적으로 나온게 기존의 기계 학습, 데이터베이스, 통계 등을 결합한 공간 데이터 마이닝이다[65].

공간 데이터는 공간을 점유하고 있는 객체와 관련된 데이터이다. 공간 데이터베이스는 공간 데이터 형태와 객체간의 공간적 관계에 의한 표현 등을 저장한다. 공간 데이터의 데이터 마이닝은 공간 데이터베이스에 있는 명백한 패턴이 없는 공간 관계에서 내재된 데이터를 추출한다. 기계학습, 데이터베이스 시스템, 통계학 등이 데이터 마이닝 연구에 관한 근거를 만들어 주었다. 동시에 공간 데이터 구조, 공간 추리, 계산 기하학 등 진보된 공간 데이터베이스 등이 데이터 마이닝에 날개를 달아주었다. 이러한 데이터 마이닝 기술은 GIS, 원격탐사, 영상 데이터베이스 개발, 의료 영상, 로보트 네비게이션 등 다양한 영역에 적용할 수 있다[66].

캐나다 Simon Fraser 대학 컴퓨터 과학과의 Koperski 등은 공간 데이터 마이닝과 관련된 최근의 연구결과를 요약한 바 있다. 데이터베이스의 기술적 진보와 바코드, 원격탐사[67], 위성기술 등 데이터 수집 기술의 진보로 큰 데이터베이스에서 방대한 양의 데이터를 수집하고 있다. 이러한 폭발적인 데이터 량의 증가는 데이터 마이닝이라는 데이터에서 지식과 정보를 추출하는 과정이 필요하게 되었다. 데이터베이스에서 데이터 마이닝에 관한 많은 연구가 있긴 했지만, 아직도 공간 자료 데이터베이스, 시간 자료 데이터베이스, 실용적인 데이터 마이닝, 객체지향형 데이터베이스, 멀티미디어 데이터베이스 등 연구해야 할 분야가 많다.

일리노이 대학에서는 일반적으로 이용하는 정밀농업 데이터를 이용하여 포장내 위치별 수확량을 예측하는데 데이터 마이닝 기법을 이용하였다. 연구에 사용된 데이터는 정밀농업을 실천 중인 농가에서 구했는데, 4년간의 수확량, 매월 강우, 1에이커 단위별로 수집한 13가지 토양 이화학성, 토성, 지형 등의 자료인데 분석에 사용된 모든 데이터가 포장내 위치별로 수집된 것이 아니였다. 평균, 역거리반비례평균법(Inverse Distance Weighting)과 크리깅을 이용하여 위치별 수확량을 예측하였는데 결과는 실망적이였다. 세 가지 방법 모두 다 수확량과 년도별, 포장 위치별 일정한 상관을 나타내지 못한 것이다. 이러한 결과의 원인은 시간적/공간적 데이터의 부적합, 자료의 부족(겨우 2년간의 수확량, 너무 적은 토양 샘플), 기타 요인(질소시비량, 파종 품종)과 데이터 오류(위치나 다른 경영상의 오류)를 들 수 있다. 이러한 상태에서 전통적인 통계모델을 이용하기 위해서는 사전에 공식화된 가설이 필요하다. 그러나 데이터 마이닝은 이러한 상태에서 의미있는 결과를 얻을 수 있는 툴을 제공한다[68].

일반적으로 정밀농업과 변량 작업에서는 대단히 많은 조사와 자료수집이 행해진다. 그런 일에 비하면 상대적으로 적은 데이터 분석이 이루어지고 있다. 데이터 분석이 보다 풍부해지기 위해서는 비료나 파종 품종, 토양의 이화학성 기상자료 등의 시간적 공간적 교호작용에 대한 자료가 필요하다. 일리노이 대학에서는 작물 수확량을 극대화할 수 있도록 포장내 공간변이를 줄이는 농업 실천을 위한 공간변이 분석 툴을 개발했다. 초기작업으로 토양 이화학성, 기상정보, 파종일시, 농작업 이력 등의 자료에 근거한 수확량의 공간변이 예측에 대한 가능성을 평가했다. 다음 작업은 농업에서의 공간정보를 분석할 수 있는 인터넷 기반 데이터 저장 및 분석툴 이다. 이러한 결합은 데이터 저장에서부터 동적 분석과 모델까지를 다 수행할 수 있도록 확장될 것이다[69].

64) Kriging은 지질통계학 분야(hybrid discipline of mining engineering, geology, mathematics and statistics)에 근거를 두고 있다. 일반적으로 공간적으로 상호 연관된 데이터를 예측하는데 유용하게 사용되고 있다. 즉, 관측된 많은 지점으로부터 관심있는 지역 안에 있는 많은 위치(격자점에서)에서 공간적으로 분포된 변수를 예측한 후에 이 변수의 지도를 생성할 수 있다. 예를 들면 등고선, 다르게 등급화된 지역 혹은 삼차원 공간에서의 곡면과 같은 기법이다. 이러한 행위가 "interpolation"이다. Kriging은 1950년대에 샘플링된 광물질 등급에 기초한 분포로부터 최적의 광물질 등급 분포를 결정하기 위한 경험적 방법을 개발한 남아프리카의 광산 기술자였던 D. G. Krige의 이름에서 유래한다.

65) http://db.cs.sfu.ca/GeoMiner/survey/html/node2.html

66) http://db.cs.sfu.ca/GeoMiner/survey/html/node1.html

67) 초창기 단색 영상에서부터 출발한 원격탐사 영상은 칼라, 근적외 영상, 멀티밴드를 넘어서서 근래에는 하이퍼스팩트럴 데이터로 발전하고 있다. 그만큼 정보의 양이 많아진 것이다.

68) http://web.aces.uiuc.edu/sriit/view_report.asp?ID=206

69) http://www.gis.uiuc.edu/cfardatamining/default.htm

3) 농작업 의사결정을 위한 데이터 마이닝

농작업 의사결정을 위한 전략적 모듈의 통합 모델인 "HighQ"라는 인터넷기반의 혁식적인 쌍방향 프로그램이 데이터 마이닝 기법을 이용하여 미국의 한 컨설팅 회사에서 개발되었다. HighQ는 전략적 생산 의사결정을 지원하는 데이터를 생성하는 다양한 프로그램 모듈로 구성되어 있다. 이 온라인 프로그램은 사용자의 수동적인 질문에 대해 사용자 포장 개개의 최적 농작업 투입량을 자동으로 계산해 준다. HighQ 성공의 비결은 포장작업을 지원하는 다양한 작업기의 개발, 웹 기반 프로그램의 꾸준한 갱신, 포장 데이터의 신속한 분석 및 농민을 지원하는 시기별 지원 등을 들 수 있다[70].

"AgFleet."라는 데이터 마이닝 기법을 이용한 농작업 의사결정 모듈도 있다. 이 프로그램은 온라인으로 웹상에서 작동하는 다양한 프로그램 모듈로 구성되어 있으며, 상담자와 재배자가 계절별, 포장별 농작업 의사결정을 위한 자료를 수집, 분석, 관리하는데 도움을 준다. 이러한 프로그램은 정밀농업 장비, PDA 등과 호환성을 갖추고 있어서 자료를 주고받을 수 있을 뿐만 아니라 농민들에게 시각화된 정보를 제공해 준다[71].

4) 관련 논문 review

정밀농업에서 생성된 데이터는 그 크기와 복잡성 때문에 기존의 통계적인 방법으로 분석하기는 곤란한 점이 많다. Canteri 등(2002)은 포장내에서 수확량이 낮은 부분의 토양 이화학성을 분석하여 수확량과 포장 이화학성간의 관계를 구명하고자 데이터 마이닝 기법을 이용하였다. 특정위치에서의 토양이화학성 정보가 데이터베이스로 구축되었다. 이러한 데이터베이스와 정밀농업 장비를 이용하여 측정된 수확량 정보를 이용하여 메타데이터[72]를 구축하였다. 이러한 정보는 의사결정트리구조를 이용하여 두 개 부분으로 구분(수확량, 토양이화학성)하는 규칙을 생성했고, 이 규칙들에 대한 신뢰도를 평가했다. 평가결과 농업전문가가 평가한 것 처럼 수확량과 토양 이화학성간의 관계를 설명할 수 있었고 이러한 결과는 고도의 별도 처리 없이 데이터 마이닝 기법을 이용하여 분류가 가능함을 알 수 있다고 보고했다.

Odhiambo 등(2002)은 지표탐사레이더 영상을 이용한 토양 분류에 데이터 마이닝을 이용했다. 넓은 영역의 토양 표면을 촬영한 레이터 영상을 해석할 때 종종 오류가 발생할 수 있다. 이러한 해석을 자동화하기 위해 토양 단면의 특성에 따라 몇 개 그룹으로 분류하는데 퍼지 - 뉴럴 네트워크를 이용했다. 토양의 물리적특성(깊이, 토성 등)에 따라 일차적인 분류가 이루어지고, 레이더 영상을 이용하여 이차적인 분류가 이루어진다. 분류결과 시각적인 분류와 데이터 마이닝에 의한 분류가 높은 상관이 있음을 알 수 있었다. 이 결과는 또한 데이터 마이닝을 이용하여 토양조사시 실시간으로 토양 특성을 분류할 수 있음을 의미한다고 보고했다.

Yang 등(2001)은 정밀농업에서 포장내 위치별 변량살포를 위해 포장을 분류하는데 하이퍼스팩트럴 영상에 대해 데이터 마이닝 기술을 이용하였다. 42개 포장에 대해 400 nm 부터 950 nm 사이의 파장대 중 72개 파장대의 반사도를 이용하여 비료량을 5단계로 구분하는데 데이터 마이닝 트리구조를 이용하였다. 재배 초기에는 91 %의 정확도로 분류할 수 있었고, 재배 중반기에는 99 %, 수확전에는 95 %의 정확도로 분류할 수 있었다. Yang 등은 이 결과를 통해 데이터 마이닝 기술이 원격탐사 영상을 분류하는데 유용함을 보고했다.

Liu 등(2001)은 옥수수의 목표수확량을 정하는데 데이터 마이닝 기법을 이용했다. 정밀농업에서 포장내 위

70) http://www.farmresearch.com/infoag/ia_PresentationInfo.asp?PRID=189&SSID=34

71) http://www.farmresearch.com/infoag/ia_PresentationInfo.asp?PRID=237&SSID=34

72) 메타데이터(metadata) : 실제로 저장하고자 하는 데이터(예를 들면, 토양이화학성, 수확량, 생육상태 등)는 아니지만, 이 데이터와 직접적으로 혹은 간접적으로 연관된 정보를 제공하는 데이터를 나타내는 말이다. 메타데이터라는 말은 최근에 생성된 말이 아니며, 이미 수십 년 전부터 정보과학분야에서 다루어져 왔고, 최근에는 더욱 다양한 분야에서 사용되고 있는 용어이다. 농업분야 메타데이터의 예를 토양이화성을 예로들어 살펴보면, 데이터베이스에 있는 토양이화학성에는 작토층깊이, 토성, 배수등급, 공극량, 질소, 인산, 칼륨 량 등 해당 위치에 대한 정보들이 수록되어 있는데, 이 정보들이 바로 토양이화학성 데이터에 대한 메타데이터가 되는 것이다. 질소, 인산과 같은 성분량 뿐만 아니라, 배수등급처럼 양호, 불량과 같은 등급 데이터처럼 데이터의 형태가 다양해짐에 따라 메타데이터도 간단한 형태가 아닌 저장 포맷, 격자크기, 생산년도, 샘플링 수 등과 같이 다양해지고 있다. 따라서 메타데이터의 활용도가 높아질수록 밀접하게 데이터와 메타데이터를 연결시키고, 공통의 메타데이터 집합과 표기방식을 제정하는 것은 필수적이다.

치별 목표수확량을 현실적으로 설정하는 것은 매우 중요하다. 목표 수확량을 계산하는 데는 토양, 기상, 경영방법 등 매우 다양한 요인이 고려되어야 하기 때문에 기존의 통계적인 방법으로는 정밀한 예측이 불가능할 정도다. 따라서 이 연구에서는 데이터 마이닝 기법의 하나인 역전파 신경망을 이용한 자동학습툴을 개발하였다. 60가지 품종에 대해 실험한 결과 RMS가 약 20 % 였다고 보고했다.

Cunningham 등은 데이터 마이닝 기술을 이용하여 지식분석 시스템을 개발했으며 이의 적용성을 확인하기 위해 식용 버섯을 선별하는데 이 시스템을 이용했다. 적용시험 결과 기계학습모델을 이용하여 버섯의 품질을 분류하고 시장 가격을 매기는 객관적인 기준으로 활용할 수 있음으로 보고했다[73].

Dong 등은 원격탐사 자료를 정밀농업에 응용하기 위해 데이터 마이닝 기술을 이용했다. 원격탐사 자료를 특성별로 분류하고 비연속적인 패턴에서 속성별 동일한 자료와 동일하지 않는 자료 군으로 구분하여 기존의 알고리즘에 의한 분류방법과 비교하였다[74].

Lee 등은 데이터 량이 방대하고 서로 이질적인 정밀농업에서의 데이터 마이닝과 지식발견을 위한 방법에 대해 논했다. 그들은 복잡하고 이질적인 데이터베이스에서 의미있는 패턴을 추출하는 방법으로 귀납적 방법에 의한 학습, 통계분류 등 다양한 학습전략을 통합해야 한다고 주장했다. 이러한 주장을 뒷받침 하기 위해 다이다호주의 한 농장에서 수확량을 예측하는 사례연구를 수행했다[75].

Soh 등(1999)은 데이터 마이닝 기술을 이용하여 인공위성 영상에서 천연자원을 분할해 내는 연구를 수행했다. 그들은 전통적인 영상처리의 한계를 극복하기 위해 인공위성 영상 분류에 데이터 마이닝 기술을 이용하였다.

4. 결 론

발달된 계측기의 등장과 정보산업의 발달로 농업관련 데이터가 점점 복잡해지고 그 양도 커지고 있다. 이 많은 데이터 중에서 의미있고 가치있는 데이터를 찾기 위한 노력은 필수적이다. 또한 축적된 데이터베이스를 효과적으로 활용하여 보다 나은 의사결정을 위한 전략정보를 추출하는 것이 중요하다. 지금까지는 쿼리(Query) 혹은 리포팅(Reporting) 도구들을 이용하여 정보들을 도출하였으나, 이것은 고급정보의 창출이라기보다는 현황 분석 또는 보고서용 등으로 정보 제공범위가 국한되어 있다.

이에 반해 데이터 마이닝은 이 보다 한층 진보된 숨어있는 정보들을 얻는데 사용되어 진다. 예를 들면 금년도 포장정보 중 어떤 토양 화학성이 작년에 비해 줄어들었는가? 라는 질문에는 쿼리나 레포팅 도구로 해결이 가능하지만, 왜? 그런 현상이 일어났는지? 그러한 현상이 수확량과 어떠한 관련이 있는지 등의 질문에는 데이터 마이닝을 이용해 해결할 수 있는 것이다.

또한 본고에서 다루지는 않았지만 분산 마이닝이나 웹 환경하에서의 마이닝 등은 아직 극히 초보단계이므로 많은 연구가 이루어져야 할 것이다. 분명한 것은 데이터 마이닝 기술은 지식추출도구로서 각광을 받게 될 것이므로 우리도 이에 준비를 해야 한다는 것이다.

본 자료에서는 이러한 데이터 마이닝에 관하여 그 개념, 기법에 대해 알아보고 사회과학분야 및 농업분야에 활용된 예를 살펴보았다. 이러한 예를 통해 데이터 마이닝에 관한 실질적인 접근에 대해 생각해보고 성공적인 데이터 마이닝 작업을 위해 필요한 요소들을 점검해봐야 할 것이다. 데이터 마이닝 툴만 이용하면 모든 정보해석에 대한 문제가 해결될 수 있을 거라는 생각은 갖지 않아야 한다.

앞에서 언급한 대로 과학기술의 발달과 첨단장비의 개발로 엄청난 양의 데이터가 우리들 앞에 쏟아지고 있다. 이 방대한 량의 데이터 중 의미있는 데이터만을 추려내서, 필요한 정보만을 처리하고 활용할 수 있는 선택과 집중의 지혜가 필요하다. 이제는 데이터를 수집하는 문제보다는 수집된 데이터를 어떻게 저장하고, 수집/저장된 데이터에서 어떻게 의미있는 정보를 추출해 낼 것인가를 고민할 단계이다. 데이터 수집에 앞서서 데이터 분석에 대한 설계를 할 줄 아는 혜안이 필요하다.

73) http://www.cs.waikato.ac.nz/~ml/publications/1999/99SJC-GH-Innovative-apps.pdf
74) http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf
75) http://cs.gmu.edu/~swlee/Paper/smc-98-cr.pdf

참 고 문 헌

1. 강현철, 한상태, 최종후, 김은석, 김미경. 2001. 데이터 마이닝 방법론 및 활용. 자유아카데미.
2. 고은지. 2001. 생명공학과 정보기술의 만남, 바이오인포매틱스. LG주간경제 12월호.
3. 김석중, 신인자, 이병오. 2000. 우유의 소비행태 변화에 관한 연구 - 의사결정나무분석기법을 이용하여. 농업경영정책연구 27:148-161.
4. 미국의 데이터 마이닝 연구동향. Available at:http://www1.kisti.re.kr/~trend/Content548/computer05.html Accessed on 14 Oct. 2003.
5. 박경우, 이대영. 2002. 데이터 마이닝 알고리즘 분석 및 동향. 光州保健大學 論文集 27.
6. 오승준 외. 2001. 데이터 마이닝 기법의 현황 및 추세. 韓國OA學會誌 8(2).
7. 우철웅, 장병욱, 원정윤. 2003. Fuzzy C-means 클러스터링 기법을 이용한 콘 관인 데이터의 해석. 한국농공학회지 45(3):73-83.
8. 장남식, 홍성완, 장재호. 1999. 데이터 마이닝. 대청.
9. 진휘철. 2000. Data Mining 은 우리에게 어떤 이득을 주는가?. 삼성 SDS IT Review.
10. A semi-automatic method to build territorial partitions. Available at: http://www.geocomputation.org/2000/GC059/Gc059.htm Accessed on 14 Oct. 2003.
11. An introduction to scientific data mining. Available at: http://www.ipam.ucla.edu/publications/sdm2002/sdm2002_ckamath.pdf Accessed on 14 Oct. 2003.
12. Applying Machine Learning to Agricultural Data. Available at:http://craig.nevill-manning.com/~nevill/publications/COMPAG.pdf Accessed on 14 Oct. 2003.
13. Bruton J. M., G. Hoogenboom. and R. W. McClendon. 2000. A Comparison of Automatically and manually Collected Pan Evaporation Data. Trans. ASAE 43(5):1097-1101.
14. Cabena P., P. Hadjinian, R. Stadler, J. Verhees. and A. Zanasi. 1997. Discovering Data Mining-from concept to implementation. Prentice Hall.
15. Canteri M. G., B. C. Avila, E. L. dos Santos, M. K. Sanches, D. Kovalechyn, J. P. Molin. and L. M. Gimenez. 2002. Application of Data Mining in Automatic Description of Yield Behavior in Agricultural Areas. Proceedings of the World Congress of Computers in Agriculture and Natural Resources. pp. 183-189.
16. Crisler M. T., R. M. Strickland, D. R. Ess. and S. D. Parsons. 2002. Data Mining Methods for Use with Geo-Referenced Field Crop Data. Proceedings of the World Congress of Computers in Agriculture and Natural Resources. pp. 265-271.
17. Data Mining for Site-specific Agriculture. Available at: http://web.aces.uiuc.edu/sriit/view_report.asp?ID=206 Accessed on 14 Oct. 2003.
18. Data Mining for Site-specific Agriculture. Available at: http://www.gis.uiuc.edu/cfardatamining/default.htm Accessed on 14 Oct. 2003.
19. Data Mining Resources. Available at:http://www.scd.ucar.edu/hps/GROUPS/dm/dm.html Accessed on 14 Oct. 2003.
20. Data Mining. Available at:http://human21.new21.org/dataMining/dm_4.htm Accessed on 14 Oct. 2003.
21. Data Mining. Available at:http://www.data-science.net/mining_adapt.htm Accessed on 14 Oct. 2003.
22. Data Mining. Available at:http://www.soinet.net/~water/main3.html Accessed on 14 Oct. 2003.
23. Demmel M., R. M. Ehr, M. Rothmund, A. Spangler, and H. Auernhammer. 2002. Automated Process Data Acquisition with GPS and Standardized Communication - The Basis for Agricultural Production Traceability. ASAE Paper No. 023013. Chicago, Illinois, USA : ASAE.
24. Developments gain attention. Available at:http://www.kdnuggets.com/press/wt97 Accessed on 14 Oct. 2003.
25. Digital Information Management and Data Mining. Available at:http://library.smsu.edu/LIS/workshops/datamining.shtml Accessed on 14 Oct. 2003.
26. Fileto R. C., A. A. Meira, J. Naka, A. S. Neto. and C. B. Medeiros. 2002. An XML-Centered Warehouse to Manage Information of the Fruit Supply Chain. Proceedings of the World Congress of Computers in Agriculture and Natural Resources. pp. 540-547.
27. Glenn E. B. and K. I. Betty. 2002. Digital Fruit Ripening: Data Mining in the TIGR Tomato Gene Index. International Society for Plant Molecular Biology, Plant

Molecular Biology Reporter 20:115-130.

28. Government Data Mining Systems Defy Definition. Available at:http://www.washingtontechnology.com/news/13_22/tech_features/393-1.html Accessed on 14 Oct. 2003.
29. Guo L. S. and Q. Zhang. 2002. A Wireless LAN for Collaborative Off-road Vehicle Automation. Proceedings of Automation Technology for Off-Road Equipment 2002 Conference. pp. 51-58.
30. Kaleita A. L. and L. Tian. 2002. Remote Sensing Of Site-Specific Soil Characteristics for Precision Farming. ASAE Paper No. 021078. Chicago, Illinois, USA : ASAE.
31. Kastanek F. 2001. Traditional and New Methods to Describe Multimodal Soil Water Characteristics. Preferential Flow Water: Movement and Chemical Transport in the Environment, Proc. 2nd Intl. Symp. pp. 181-184.
32. Kastens T. L. 2000. Precision Ag update. Proceedings of Risk and Profot 2000 Conference. Manhattan, Kansas.
33. Liu J., C. E. Goering. and L. Tian. 2001. A Neural Network for setting Target Corn Yields. Trans. ASAE 44(3):705-713.
34. Mesterhazi P. A., M. Nemenyi, K. Kacz. and Z. Stepan. 2002. Data Transfer among Precision Farming Systems. ASAE Paper No. 021047. St. Joseph, Mich.: ASAE.
35. Motonaga Y., H. Semba, H. Kondou, H. Kitamura, K. Nakanishi, A. Hashimoto. and T. Kameoka. 2002. Agroinformatic Management System with Quality Analysis Function. Proceedings of the World Congress of Computers in Agriculture and Natural Resources. pp. 580-587.
36. Odhiambo L. O., R. S. Freeland, R. E. Yoder. and J. W. Hines. 2002. Application of Fuzzy-Neural Network in Classification of Soils using Ground-penetrating Radar Imagery. ASAE Paper No. 023097. St. Joseph, Mich.: ASAE.
37. Peschel J. M., G. R. Assistant, P. K. Haan. and R. E. Lacey. 2003. A SSURGO Pre-Processing Extension for the ArcView Soil and Water Assessment Tool. ASAE Paper No. 032123. Las Vegas, Nevada, USA : ASAE.
38. Soh L. K. and T. Costas. 1999. Segmentation of satellite imagery of natural science using data mining. IEEE Trans. On Gerscience and Remote Sensing. 37(2):1086-1098.
39. The Data Mine. Available at:http://www.the-data-mine.com/Accessed on 14 Oct. 2003.
40. Wang Z., F. Xiong. and X. Hang. 2002. A New Algorithm for Mining Sequential Patterns and the Application in Agriculture. Proceedings of the World Congress of Computers in Agriculture and Natural Resources. pp. 622-628.
41. Yang C. C., O. S. O. Prasher, J. Whalen. and P. K. Goel. 2001. Application of data mining technology for hyperspectral imagery classification in agricultural fields. ASAE Paper No. 013116. Sacramento, California: ASAE.
42. Zhenyu Wang, Xiong Fanlun. and Hang Xiaoshu. 2002. A New Algorithm for Mining Sequential Patterns and the Application in Agriculture. Proceedings of the World Congress of Computers in Agriculture and Natural Resources. pp. 622-628.