조선기술지식 활용을 위한 유전적 프로그래밍 기반의 데이터 마이닝 도구개발

Development of Data Mining Tool for the Utilization of Shipbuilding Knowledge based on Genetic Programming

이 경 호* · 오 준** · 박 종 현*** · 박 종 훈****
Lee, Kyung-Ho · Oh, June · Park, Jong-Hyun · Park, Jong-Hoon

Abstract

As development of information technology, companies stress the need of knowledge management. Companies construct ERP system including knowledge management. But, it is not easy to formalize knowledge in organization. They experience that constructing information system help knowledge management. Now, we focus on engineering knowledge. Because engineering data contains experts' experience and know-how in its own, engineering knowledge is a treasure house of knowledge. Korean shipyards are leader of world shipbuilding industry. They have accumulated a store of knowledges and data. But, they don't have data minning tool to utilize accumulated data. This paper treats development of data minning tools for the utilization of shipbuilding knowledge based on genetic programming(GP).

Keywords: knowledge management, data minning, engineering knowledge, shipbuilding, genetic programming

1. 서 론

요즘 사회는 산업사회에서 정보화 사회로 변해가고 있다. 정보화 사회는 흔히들 지식사회라고도 일컫는다. 지식사회의 특징은 여러 가지가 있지만, 그 중에서 가장 중요한 것은 지식이나 정보의 가치가 예전에 비교할 수 없을 만큼 그 중요하다는 것이다. 최근 산업 전 분야에서 지식의 중요성이 매우 강조되고 있다. 한 기업의 경쟁력은 곧 기업의 생존과 연결되기 때문이다. 따라서 IT기술의 발전함에 따라 기업의 경쟁력을 위하여 산업 환경을 급격히 분산화, 글로벌화로 변화시키고 있는 것이다. 또한, 기업이 어떠한 지식을 지니고 있으며, 이 지식을 어떻게 공유하고 활용을 극대화 할 수 있느냐의 여부에 따라서 그 기업의 경쟁력을 결정짓는다고 할 수 있다. 특히, 세계를 단일시장으로 하고 있는 조선 산업은 고도로 정보화된 21세기 지식 산업 환

* 정회원 · 인하대학교 선박해양공학과 교수 E-mail: kyungho@inha, ac.kr
** 인하대학교 선박해양공학과 석사과정 E-mail: shipman98@lycos.co.kr
*** 인하대학교 선박해양공학과 석사과정 E-mail: dairy5@nate.com
**** 인하대학교 선박해양공학과 석사과정 E-mail: 15knight@nate.com

경에서 국제 경쟁력을 높이기 위해서는 이러한 노력이 더욱 필요하다고 할 수 있을 것이다. 이를 위하여 기업들은 지식관리(Knowledge Management)도 이야기하고 있으며, 전사적인 ERP(Enterprise Resource Planning) 구축에 온힘을 다 하고 있다. ERP 구축에 있어서 핵심 요소 중의 하나는 지식관리 시스템과 정보의 공유이다. ERP 시스템을 통한 지식의 공유 및 활용 환경의 구축이 요구되고 있다. 그러나 현실적으로 조직 내에 스며들어 있는 지식을 형식화 한다는 것은 그렇게 쉬운 일이 아니며, 이를 정보시스템으로 지원하기 위해서 많은 어려움이 따른다(박우창 등, 2004). 여기서 우리의 관심은 기술지식(Engineering Knowledge)이다. 기술지식 관점에서 보면 축적된 공학 데이터의 활용 측면을 간과해서는 안 된다. 왜냐하면, 공학데이터에는 데이터 그 자체에 전문가의 경험과 노하우가 녹아들어있는 있는 정보의 보고이기 때문이다(이경호등, 2005-1). 현재 세계시장에서 선두에 있는 우리나라의 조선 산업은 지금까지 많은 배들을 건조하며 축적된 많은 데이터를 가지고 있다. 하지만 이러한 데이터들을 활용하기 위한 도구를 보유하고 있지 못한 것이 현실이다. 본 논문에서는 이러한 조선기술지식의 활용을 위하여 기술지식이 녹아있는 데이터를 활용한 유전적 프로그래밍을 이용한 데이터 마이닝 툴의 개발에 대해서 이야기 하고자 한다.

2. 기술 지식의 분류 및 정의

지식은 어느 관점에서 바라보느냐에 따라 여러 가지로 분류되지만 표1에서와 같이 형태에 따른 분류와 생성과정에 따른 분류로 나눌 수 있다.(이경호와 손미애, 2004). 본 논문에서 다루는 데이터들은 지식관리 관점의 기술 지식에 대해 다루고 있는데 이를 한마디로 정의하기는 힘들다. 물론 기술지식이나 기술지식이 녹아있는 데이터에 대하여 한마디로 정의하기란 불가능하다. 따라서 본 논문에서 대상으로 하고 있는 기술지식과 데이터를 다시 정의하면 다음과 같다. "기술지식은 지식의 분류 측면에서 형식적 지식과 암묵적 지식, 경험적 지식과 분석적 지식을 모두 다 포함하고 있다. 그러나 여기서는 형식화된 지식보다는 명시적으로 나타나 있지 않는 암묵지와 기술지식이 녹아있는 데이터, 구조화되지 못한 지식요소 등, 데이터 마이닝을 통하여 지식을 얻어낼 수 있는 분석적 지식을 의미한다. (이경호 등, 2005-2)

표1. 일반적인 지식의 분류

분류방식	지식분류	정의	실례
형태	명시적 지식 (형식지)	언어, 코드, 구조성을 지닌 형태로 표현된 지식	영업실적에 대한 분석자료
	암시적 지식 (암묵지)	언어, 코드, 구조성을 지닌 형태로 표현하기 힘든 지식	기술자가 보유한 기술, 비즈니스 감각
생성과정	경험적 지식	업무수행 중 동일하게 반복되는 과정에서 겪게 되는 경험과 시행착오를 통해 지속적으로 누적시켜 온 지식	시스템운영 지침서, 작업 방법론
	분석적 지식	업무를 수행하기 위해 기업이 기존부터 보유하고 있던 데이터나 정보를 활용 및 분석하여 얻어낸 지식	특정제품의 시장점유율, 판매전략 변화에 따른 매출액 증가비율

3. 데이터 마이닝 기술

앞에서 언급하였듯이 본문에서 다루고자 하는 지식은 형식적으로 나타난 기술지식이 아닌 명시적으로 나

타나지 않는 지식과 그러한 지식이 녹아있는 데이터, 즉 분석적 지식에 초점을 맞추고 있다. 암묵지에 대한 접근은 전문가 시스템이나 Case-based System과 기존의 지식관리 시스템의 접근방법으로 가능하며, 구조화 되지 못한 지식요소는 XRML(eXtensible Rule Markup Language)을 통하여 접근할 수 있다. 하지만 여기서 초점을 맞출 대상은 위에서 언급한 분석적 지식이다. 즉, 데이터의 가공을 통해 지식을 얻어낼 수 있는 지식에 초점을 맞추고 있으며, 이러한 데이터들은 데이터 마이닝과 기계학습을 통하여 지식을 얻어낼 수 있다.

3.1 데이터 마이닝의 정의

데이터 마이닝의 일반적인 정의는 다음과 같다.

데이터 마이닝은 잠재적으로 유효하고, 새롭고 타당성 있으면서 궁극적으로 데이터에서 이해할 수 있는 패턴을 찾아내는 단순하지 않은 프로세스이다.

위의 정의에서 사용한 주요 용어들의 의미를 살펴보면 다음과 같다.

- '데이터'는 데이터베이스내의 사례를 나타내는 사실의 집합이다.
- '패턴'은 사실의 부분집합으로 사실을 묘사할 수 있는 언어의 표현이다.
- '프로세스'란 용어는 데이터 마이닝이 여러 단계로 구성되어 있고 각 단계는 데이터 준비, 탐색, 지식 평가, 정제, 변경 후 반복을 하는 작업을 포함한다.
- '단순하지 않다'는 것은 탐색이나 추론이 포함된다는 뜻이다.
- 발견된 패턴은 '새로운' 이어야 하며 데이터(현재의 가치를 이전의 가치나 기대했던 가치와의 비교)와 지식(새로운 발견이 과거의 발견과의 어떤 관련)의 변화된 관계를 비교하여 측정된다.
- 또한 발견된 패턴은 어느 정도의 확실성을 가지고 새로운 데이터에 대해 '타당'해야 한다.
- '잠재적으로 유용' 하다는 것은 발견된 패턴은 유용한 함수를 통해서 측정되어진 것처럼 잠재적으로 유용한 행동을 이끌어 낼 수 있어야 한다는 의미를 내포한다.
- '궁극적으로 이해할 수 있는' 이라는 것은 데이터마이닝의 목표는 패턴을 기초로 하는 데이터의 더 나은 이해를 돕기 위하여 인간이 이해할 수 있도록 패턴을 만드는 것이다.

3.2 데이터 마이닝 기술 분류

3.2.1 예측(Prediction)

데이터의 학습을 통하여 만들어진 예측모형이나 패턴을 이용하여 특정한 속성이나 값을 예측하는 것으로서, 데이터의 해석 및 일반화 모형의 구현을 통해 새로운 패턴을 찾아낸다. 이를 위하여 인공신경망이나 진화연산방법 등이 활용된다.

3.2.2 연관성 추정(Association)

연관성측정(Associations)은 어떤 특정 문제에 대한 예측(Prediction)이나 고객들을 특정목적에 따라 분류(Segmentation)하는 문제가 아니라, 상품 혹은 서비스의 거래기록(Historical)데이터로부터 상품간의 연관성 정도를 측정하여 연관성이 많은 상품들을 그룹화 하는 clustering의 일종으로서, 동시에 구매될 가능성이 큰 상품들을 찾아냄으로써 시장바구니분석(Market Basket Analysis)에서 다루는 문제들에 적용될 수 있다. 연관성측정에서 얻어지는 결과물인 연관규칙 Association rule 은 If A, then B (A→B) 와 같은 형식으로 표현된다. 연관성측정에서의 연관규칙은 '상품 A가 구매되어진 경우는 상품 B 도 구매된다.'라고 해석된다. 여하튼 연관규칙은 구체적인 상품들이 언급되어지므로 이해가 쉽고 명쾌하며 실제 업무로의 적용이 용이하다.

3.2.3 클러스터링(Clustering)

의사결정나무(Decision Trees) 또는 신경망 모형(Neural Networks)을 이용한 데이터마이닝 작업의 목적은 대체로 목적변수가(Target) 있어서, 이 목적변수에 영향을 주는 다른 변수들을 찾아내어 그들의 상호관계를 파악한 후, 이 상호관계를 근거로 추후에 일어나는 사건들에게 어떤 결과가 있을지를 예측하는 것이다. 즉, Supervised 데이터에 적합한 분석기법들인 것이다. 이에 반면 클러스터링은 어떤 목적변수(Target)를 예측하기 보다는 고객수입, 고객연령과 같은 속성이 비슷한 고객들을 묶어서 몇 개의 의미 있는 군집으로 나누는 것을 목적으로 한다. 숲이 너무 복잡해서 전체를 파악할 수 없을 때, 나무들부터 살펴보아야 하듯이, 대용량의 데이터가 너무 복잡할 때는 이를 구성하고 있는 몇 개의 군집을 우선 살펴봄으로써 전체에 대한 윤곽을 잡을 수 있을 것이다. 클러스터링은 이런 상황에 유용하게 쓰일 수 있다.

3.3 데이터 마이닝 도구 개발의 초점

본 논문에서 개발할 데이터 마이닝 도구는 위에서 언급한 데이터 마이닝의 기술 중 예측(Prediction)과 연관성(Association)추정에 초점을 맞추고 있다. 즉, 조선분야의 데이터를 가공하여 나오는 지식은 결과를 예측하는 것이나 수식이 될 것이다. 하지만 조선분야의 데이터의 경우 같은 선종의 데이터가 그리 많지 않은 점과 데이터의 비 선형성과 불연속성으로 인한 함수의 비정확성이 데이터 마이닝 도구개발의 가장 큰 어려움으로 꼽힌다.

4. 데이터 마이닝 도구 개발

본 논문에서는 위에서 언급한 데이터 마이닝의 분류에서, 축적된 데이터를 바탕으로 이들의 학습을 통하여 데이터의 패턴을 찾아내고 이것을 일반화 하는 모형으로 근사시킬 수 있는 예측모델이나 패턴을 찾는 데이터마이닝 기술로 유전적 프로그래밍(Genetic Programming : 이하 GP라 지칭)을 사용하였다. GP는 비선형성과 불연속 특성을 가진 공학 데이터의 학습에 탁월한 능력을 가지고 있다고 생각되어 GP를 활용하여 데이터 마이닝 도구를 개발하였다. 본 논문에서는 조선기술지식 활용을 위한 유전적 프로그래밍 기반의 데이터 마이닝 도구개발을 다루고 있다.(이경호 등, 2004)

4.1 유전적 프로그래밍(GP)의 도입

GP는 유전적 알고리즘(GA:Genetic Algorithm)의 확장으로 그 개체(Individual)가 트리(Tree) 형태의 컴퓨터 프로그램이 된다.(이경호등, 1998). 여기서의 컴퓨터 프로그램은 터미널 집합(Terminal set)과 함수 집합(Function Set)의 조합으로 생성된 문법적으로 올바른 GP 트리를 뜻한다. 진화과정을 통하여 GP트리는 적합도를 최적화하기 위하여 그 구조 자체가 동적으로 변화하는데, 적합도 계산을 위해서 트리의 학습오차를 계산할 수 있는 함수가 사용된다. 기저함수 바탕의 근사화 기법은 그 함수의 형태가 이미 결정되어 있는 반면, GP에서는 함수 즉 GP 트리구조 자체가 적합도를 최적화하기 위하여 변화한다(Koza, 1992). 이러한 특성 때문에 GP는 새로운 데이터 영역의 학습에 있어 전문가(Domain Expert)가 예측하지 못한 새로운 모델을 찾을 가능성을 가지고 있다. 이러한 특징을 고려할 때 GP는 함수 근사화 및 데이터 마이닝의 유용한 도구로 활용될 가능성이 크다.(이경호 등, 2004; Gray et al,1996)

4.2 조선기술지식 활용을 위한 GP기반의 데이터 마이닝 도구 개발

4.2.1 도구 개발

조선분야의 데이터는 다음과 같은 특징을 지녔다. 조선 산업은 주문자 생산방식으로 제작될 뿐만 아니라

같은 선종의 배의 숫자가 GP의 학습에 이용할 수 있을 정도로 그렇게 많은 양의 데이터를 구하기도 힘들다. 또한 기존의 여러 가지 경험식 또한 과거의 선박(Tanker, Bulk Carrier)등의 중심으로 만들어진 식이기 때문에 현재 많이 건조하고 있는 컨테이너선이나 LNG Carrier와 같은 선박에 적용하기 힘들기 때문이다. 또한, 조선분야의 데이터는 비선형적이고 불연속성이라는 특성을 가진 데이터가 많기 때문에 일반적인 회귀식이나 기존의 상용 툴로는 예측모형을 만들어 내는 것이 어렵다고 생각되었다(이경호 등, 2005-2). 따라서 조선전용 데이터 마이닝 도구 개발이 필요하였다.

4.2.2 개발 과정

조선기술 지식활용을 위한 데이터 마이닝 도구로 Microsoft Visual Studio .Net C#을 이용하여 데이터 마이닝 프로그램을 작성하였다. 개발 프로그램은 다음 그림2와 같은 인터페이스를 지니도록 프로그램을 하였고, 인터페이스에서 먼저 GP의 옵션을 살펴보면 다음 표2 와 같다.

표 2 개발된 데이터 마이닝 도구에서 사용되는 GP의 옵션

GP 옵션	옵션의 기능설명
High Order Polynomial	예측모형의 수식 표현을 차수 다항식, 즉 다항식정리로 나타낸다.
Linear Model With Polynomial(PLM-GP)	예측모형의 수식 표현을 Polynominal로 나타냄. 함수자체가 비선형적일때 사용
Linear Model With Math (LM-GP)	예측 모형의 수식표현을 수학식(Sin,Cos)등과 같은 다항식으로 표현

GP를 학습시킬 트레이닝 데이터와 테스트 데이터를 나누지 않고 하나의 데이터 파일로 입력을 받아 테스트할 데이터의 건수만 입력하면 자동으로 나누어 질수 있도록 프로그래밍 했다. 프로그램 상에서 GP의 특징 중 하나인 교배, 재생산, 돌연변이의 확률을 사용자가 직접 수정할 수 있도록 하였다.

4.3 GP를 이용한 데이터 마이닝 도구 적용 예

현재 우리나라 각 조선소는 그동안 배를 건조하며 쌓은 많은 데이터와 모형선을 실험한 데이터가 있다. 하지만, 이러한 데이터들은 대외비에 붙여진다. 따라서 이런 데이터를 외부에서 구할 수 없다. 따라서 본 논문에서는 본교에서 행하여진 Trimaran 선박의 모형선 데이터 55개를 이용하여 테스트를 진행하였다. 옵션의 설정은 PLM-GP를 사용하였고, 모형선 실험데이터 파라미터에서 검증을 위해 사용된 입력과 출력에 사용된 정보는 다음과 같다. 선속, 시간, 풀루드 넘버(Fn), 트림(선박의 선수와 선미부의 기울기), Ct(전 저항계수-선박의 전체저항에 대한 계수) 등을 입력 파라메터로 하고, RTS(모형선 실험의 저항 측정값) 값을 출력변수로 설정하였다. 학습을 통하여 나온 모형에서의 테스트 값은 RTS값이 나오도록 하였다. 그중 35건의 데이터는 GP를 학습시키는데 사용하였고 나머지 데이터는 GP를 이용하여 학습결과를 테스트 하였다. 그림2는 35건의 학습을 통하여 얻어진 학습모델을 바탕으로 테스트를 한 값과의 오차(실제값-추정된 결과값)에 대한 그래프이다.

GP의 단점중의 하나는 진화 과정을 통하여 학습을 하기 때문에 진화 과정 중에 생성된 복잡한 트리구조로 인하여 그림3과 같은 사람이 알아보기 어려운 수식 또는 사용하기 힘든 식으로 나타난다는 점이다. 따라

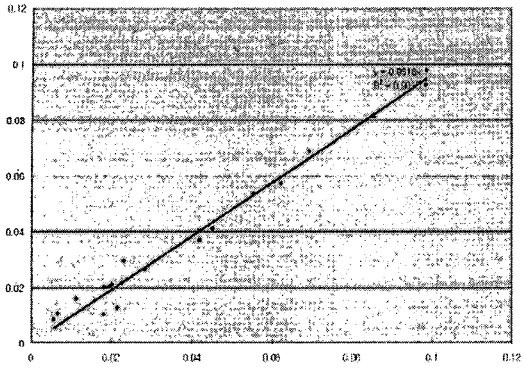

그림 2 테스트 데이터의 오차에 대한 그래프

서 본 논문에서는 데이터 마이닝을 통한 조선기술 지식활용에 목적이 있기 때문에 다른 프로그램이나 사람이 알아보도록 하기 위하여 , 그림4와 같이 C 코드로 수식을 생성하여 표현할 수 있도록 하였다.

```
0.7041403765 + 5.4176944816e-002x1 + 7.8629378184e-001x5 - 9.459042539 4e-003x4 + 3.8714236782e-001x2
+ 4.7791487031e-001x2^2 + 1.8841593368e-001x4^2 + 1.5261995540e-001x5^2 + 1.4854091430e-001x2^3
+ 2.5739344715e-001x4^3 + 0.0000000000e+000x4x5^2 - 3.5650157521e-001x4^2x5^2 + 1.2719246766e-001x4^4
+ 1.1883385846e-001x4^3x5^2 + 5.2681866989e-002x4^5 + 1.3130922272e-002x4^6 + 3.4601623476e-001x4^4x5^2
+ 1.1865666655e-003x4^7 - 2.3867748984e-001x4^5x5^2 - 8.7377837862e-002x4^6x5^2
+ 1.2582408537e-001x4^7x5^2 - 4.1941361790e-002x4^8x5^2 + 4.6601513180e-003x4^9x5^2
```

그림 3 데이터 마이닝 도구를 이용하여 추정된 수식의 결과

4.3 앞으로의 개선 사항

본 논문에서는 GP 자체의 단점인 학습과정을 통하여 나온 결과 수식이 사용하기 어려운 식으로 나온 것을 C 코드로 바꾸도록 개발 하였다. 하지만, 결과 자체가 알아보기 쉬운 그래프나 수식으로 나온 것이 아니므로 그 결과를 일일이 도표로 바꾸는 작업을 하여야 하였고, GP가 학습을 통하여 도출한 결과가 얼마나 정확한 것인가 기존에 존재하는 식과 바로 비교할 수 있도록 하는 작업이 더 필요하다고 생각된다. 또한 프로그램 내에서 간단한 회귀분석의 결과와 GP의 결과를 편하게 비교할 수 있도록 도표로 바꾸어 주는 작업등도 더 필요하다고 생각된다. 또한 정확성의 향상과 좋은 결과를 위하여 다른 상용 데이터 마이닝 툴과의 비교를 하여 조금 더 조선분야 환경에 맞도록 개발을 할 필요가 생각된다.

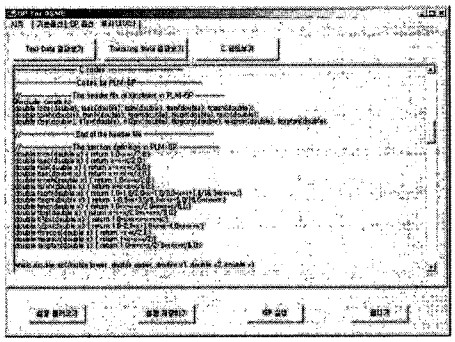

그림 4 추정된 경험식의 C 코드 생성

5. 결 론

세계 1등 산업중의 하나인 조선분야의 데이터가 축적되면서 이것을 활용하고자 하는 요구가 증대되고 있는 시점에서 본 논문에서는 지식관리의 관점에서 기술지식을 정의하였다. 또한 조선기술 지식의 활용을 위하여 GP를 이용한 데이터 마이닝 도구를 개발하여 예측 모델 생성을 제시하였다. GP는 모든 학습 데이터를 한꺼번에 근사하는 글로벌 모델로서 비선형성이나 불연속성의 특징을 가진 조선분야의 데이터의 근사에 아주 좋은 결과를 보여주었고 이를 통하여 조선분야의 기술지식 활용을 위한 도구개발에 알맞다는 것을 보여주었다. 본 논문에서 언급한 데이터 마이닝 툴을 통하여 조선 분야의 축적된 많은 데이터로부터 유용하고 좋은 지식의 생성을 기대한다.

감사의 글

본 논문은 본 논문은 한국과학재단 첨단조선공학연구센터 지원과제(R11-2002-104-08002-0)로 수행된 연구결과의 일부로서, 위 기관의 지원에 감사드립니다.

참고문헌

박우창, 송현우, 용환승, 최기현 (2004) 데이터 마이닝, 자유 아카데미, 서울

이경호, 손미애 (2004) 차세대 성장동력과 조선산업(어떻게 해야 하나? How-to-do) ; 표준화와 기술지식관리, 대한조선학회 학회지 Vol.41 No.3 pp.15-26

이경호, 연윤석 (2005) 데이터 마이닝 개념에 의한 조선 분야 데이터의 해석 및 활용 기법 연구, CAD/CAM 학회 학술발표회 논문집 Knowledge Engineering I pp.110-115

이경호, 연윤석, 양영순 (1998), 개선된 유전적 프로그래밍 기법을 이용한 설계 파라메터 추정, 대한조선학회 설계연구회 하계발표회

이경호, 연윤석, 양영순 (2004) 데이터 마이닝을 위한 다항식기반의 유전적 프로그래밍 기법과 조선분야 응용, 대한조선학회 춘계학술대회 논문집 pp.845-850

이경호, 연윤석, 양영순, (2005)-1 조선 기술지식 활용을 위한 데이터 마이닝 기법의 적용, 한국해양과학기술협의회 공동학술대회 Vol.2005, No.0 pp.375-380

이경호, 연윤석, 양영순, (2005)-2 조선분야의 축적된 데이터 활용을 위한 유전적 프로그래밍에서의 선형 모델개발, 대한조선학회 논문집 Vol.42 No.5 pp.309-405

Gray G.J. , Murray D.J. and Sharman K.C., (1996) Structural System Identification using Genetic Programming and a Mlock Diagram oriented Simulation Tool, Electronics Letters, Vol.32,pp1422-1424

Koza, J.R (1992), Genetic Programming: On the Programming of Computers by Means of Natural Selection, The MIT Press