한국전산구조공학회
2005년 봄 학술발표회 논문집

진화연산에 의한 공학 데이터의 활용

Practical Utilization of Engineering Data based on Evolutionary Computation Method

이 경 호* 연 윤 석** 양 영 순***
Lee, Kyung-Ho Yeon, Yun-Seog Yang, Young-Soon

ABSTRACT

Korean shipyards have accumulated a great amount of data. But they do not have appropriate tools to utilize the data in practical works. Engineering data contains experts' experience and know-how in its own. It is very useful to extract knowledge or information from the accumulated existing data by using datamining technique. This paper treats an evolutionary computation method based on genetic programming (GP), which can be one of the components to realize datamining.

1. 서론

지금까지 공학설계 분야에서 컴퓨터 기술의 활용은 컴퓨터 원용 설계(CAD)와 컴퓨터 원용 해석(CAE)이 대부분을 차지하였다. 그러나 최근 들어 정보기술의 발전에 힘입어 다각적인 컴퓨터의 활용이 이루어지고 있다. 즉 정보통신 기술의 발전은 산업 환경을 급격히 분산화, 글로벌화로 변화시키고 있으며, 세계를 단일시장으로 하고 있는 조선산업에서 정보기술은 단순한 자동화의 도구가 아니라 고도로 정보화된 21 세기 산업 환경에서 국제경쟁력의 확보를 위한 전략적 도구로 활용되고 있다. 정보통신 기술의 발전과 함께 국가나 기업의 경쟁력은, 지금까지의 토지, 공장, 설비 등 눈에 보이는 자산으로부터, 지식의 힘이나 서비스 능력으로 옮겨가고 있다. 즉, 제품이나 서비스 가치의 대부분은 눈에 보이는 하드웨어가 아닌 기술적인 Know-How 와 고객의 요구를 파악한 제품설계 방식과 마케팅 방법, 기업의 선진성 등 지식을 기반으로 하는 눈에 안 보이는 소프트웨어가 되고, 지식이 경쟁력의 근원이 되는 시대를 맞고 있다.(1) 특히, 조선산업은 지금까지 우리나라 경제성장을 실질적으로 주도해 왔으며, 이에 그치지 않고 차세대 성장동력으로 재도약하기 위해서는 어떻게 그 방향을 설정해야 하는가 하는 것이 매우 중요한 시점이 되었다. 그 돌파구중의 하나는 분명히 지식관리를 통한 기술지식(공학지식)의 활용일 것이다.

그러나 기술지식의 활용 측면에서 보면, 우선 기술 지식을 어떻게 정의할 것인가에서부터 그렇게 정의된 기술지식을 어떻게 활용할 수 있는 형태로 표현할 것인가, 또한 어떻게 이를 체계적으로

* 정회원 · 인하대학교 선박해양공학과 교수
** 대진대학교 컴퓨터응용기계설계공학과 교수
*** 서울대학교 조선해양공학과 교수

공유/활용할 수 있는 관리 시스템 환경을 구축할 것인가 하는 문제가 풀려야 할 과제로 남는다.

본 논문에서는 이러한 과제들 중에서 먼저 지식관리 관점에서의 기술지식에 대해 정의하고, 이의 활용 방안 중에서 특히, 공학 데이터의 활용 측면에서 이의 중요성과 활용 방안으로서 진화연산에 의한 데이터 활용에 대해 기술하였다.

2. 지식관리 관점에서의 공학데이터

지식은 어떠한 관점에서 바라보느냐에 따라 여러 가지로 분류되지만 형태에 따른 분류와 생성과정에 따른 분류로 나뉘는 것이 일반적이다. 먼저 형태에 따른 분류를 보면, 명시적인 지식(형식지: Explicit Knowledge)과 암시적 또는 묵시적인 지식(암묵지: Tacit Knowledge)으로 분류할 수 있다.[2] 형식지는 말 그대로 명백하게 드러내 보이는 형체와 형식이 있는 지식을 의미하는 것이라면 암묵지는 형상화되어 있지는 않지만 은연중에 그 뜻을 나타내 보일 수 있는 지식을 의미한다. 예를 들어, 형식지는 문서 또는 파일의 형태로 된 보고서, 영업실적 분석자료, 업무처리절차, 지침서 등을 들 수 있으며, 암묵지는 개인의 축적된 기술, 경험 등에 내재되어 있는 지식으로 엔지니어들이 보유하고 있는 기술, 문서화되지는 않았으나 고객의 문의에 유연하고 효율적으로 대처하는 프로세스 등을 들 수 있다. 다음으로 생성과정에 따라 지식을 분류해 보면 경험적 지식(Experimental Knowledge)과 분석적 지식(Analytical Knowledge)으로 나눌 수 있다. 경험적 지식은 기업 내 업무수행 중 동일하게 반복되는 과정에서 겪게 되는 경험과 시행착오를 통해 지속적으로 누적시켜 온 지식을 말한다. 여기서 경험적 지식은 형태에 따른 분류 방식 중 형식지와 일맥상통한다고 볼 수 있다. 반면, 분석적 지식은 업무 수행을 하기 위해 기업이 기존부터 보유하고 있던 데이터나 정보를 활용 및 분석하여 얻어낸 지식이라 할 수 있다. <Table 1>은 이상의 지식분류 방법을 간단히 정리한 것이다. 본 논문에서는 지식관리 관점에서의 기술지식(여기서는 공학지식이라고 생각함)에 대해 다루고 있는데 이를 한마디로 정의하기는 어렵다. 물론 기술지식은 형식지와 암묵지, 경험적 지식과 분석적 지식을 모두 다 포함하고 있다. 그러나 여기서는 형식화된 지식보다는 명시적으로 나타나 있지 않는 암묵지와 기업이 보유는 하고 있으나 이의 가공을 통해 유용한 지식을 얻어낼 수 있는 분석적 지식의 활용 측면에서의 기술지식의 활용을 다루고자 한다.

본 연구에서 대상으로 하고 있는 기술지식을 다시 한번 정의하면 다음과 같다.

<Table 1> 지식의 분류

분류 방식	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

"기술지식은 지식의 분류 측면에서 형식지와 암묵지, 경험적 지식과 분석적 지식을 모두 다 포함하고 있다. 그러나 여기서는 형식화된 지식보다는 명시적으로 나타나 있지 않는 암묵지(사람의 머릿속에 존재)와 기업에 산재해 있는 구조화되지 못한 지식요소, 기업이 보유는 하고 있으나 이의 가공을 통해 유용한 지식을 얻어낼 수 있는 분석적 지식을 의미한다."

<Table 2>는 이러한 관점에서의 기술지식의 형태 및 이를 구현하기 위한 기술을 정리한 것이다. 본 논문에서는 이러한 지식 중에서 가공을 통해 지식을 얻어낼 수 있는 분석적 지식에 초점을 맞추고 있다. 이는 기술지식은 현실적으로 추출 및 표현하기가 쉽지 않다. 이에 반해 공학 데이터는 현장에서 쉽게 구할 수 있으며, 최근 들어 ERP (Enterprise Resource Planning) 시스템 구축을 통해 상당 부분 체계화되어 축적되어 있다. 이러한 축적된 공학 데이터는 그 데이터 자체에 전문가의 경험과 노하우, Yard Practice 가 녹아 들어 있다고 할 수 있다. 데이터마이닝(Data Mining)은 이러한 데이터로부터 유용한 정보나 지식을 추출하는 기술로서 본 논문에서는 진화연산 (Evolutionary Computation)에 의한 데이터마이닝 방법에 대해 소개하고자 한다.

<Table 2> 기술지식의 분류 및 관련 구현 기술

암묵지에 대한 접근	구조화되지 못한 지식요소	가공을 통해 지식을 얻어낼 수 있는 분석적 지식
Expert System Case-Based System . 기존의 KMS EDM 기업의 전략 및 의지	XRML (eXtensible Rule Markup Language) Ontology	데이터마이닝 기계학습

3. 데이터마이닝에 의한 활용 방안

데이터마이닝이라는 것은 방대한 양의 데이터 속에서 쉽게 드러나지 않는 유용한 정보를 찾아내는 과정, 데이터간의 숨겨진 관계, 혹은 겉으로 드러나지 않거나 또는 기존의 통계학적 방법을 통해 뽑아내기에는 너무나 복잡한 관계를 찾아내고, 이 관계를 바탕으로 앞날을 예측하는 기술이라고 정의할 수 있다. 즉, 대규모 데이터베이스 내에 존재하지만 숨겨져 있는 패턴이나 상호 관련성에 대한 탐색 및 추출을 통하여 지식의 형태로 변환하는 과정으로서 기계학습, 인공지능, 데이터베이스, 통계학 등 다른 연구 분야로부터 발전된 데이터 분석 기술이라 할 수 있다. [3]

본 논문에서는 축적된 데이터로부터 반복적인 학습 과정을 거쳐 데이터의 패턴을 찾아내고 이를 일반화하여 향후를 예측(Prediction)하도록 하는 데이터마이닝의 기법으로 진화연산 방법을 소개하고자 하며, 특히 비선형 데이터의 예측이 탁월한 유전적 프로그래밍(Genetic Programming : 이하 GP) 방법에 대해 소개하고자 한다.

4. 진화연산에 의한 공학 데이터 예측 모델의 구현

4.1 유전적 프로그래밍(GP) 개요

GP는 유전적 알고리즘(GA)의 확장으로써 그 개체(Individual)가 트리(Tree) 형태의 컴퓨터 프로그램이 된다.[4] 여기서의 컴퓨터 프로그램은 터미널 집합(Terminal set)과 함수 집합(Function set)의 조합으로 생성된 문법적으로 올바른 GP 트리를 뜻한다. 진화과정을 통하여 GP 트리는 적합도(Fitness)를 최적화하기 위해서 그 구조 자체가 동적으로 변화하는데, 적합도 계산을 위해서 트리의

학습오차(Learning error)를 계산할 수 있는 함수가 사용된다. 기저함수 바탕의 근사화 기법은 그 함수의 형태가 이미 결정되어 있는 반면, GP에서는 함수 즉 GP 트리의 구조 차제가 적합도를 최적화하기 위하여 변화된다. 이러한 특징을 고려할 때, GP는 함수 근사화 및 데이터마이닝의 유용한 도구로 활용될 가능성이 크다.[5] 그러나 GP의 가장 큰 단점 중의 하나가 진화의 과정을 통하여 학습을 하는 과정에서 많은 양의 데이터와 복잡한 트리 구조를 최적화해 나감으로 인해 계산시간이 많이 소요된다는 것이다. 또한 대부분의 공학 문제에서 GP의 학습에 이용할 만큼의 일관성 있는 많은 양의 실적 데이터를 얻는다는 것도 쉽지는 않다. 이러한 문제를 해결하기 위한 방법으로 GP의 복잡한 트리 구조 대신 선형 모델을 도입하고, 이를 통해 학습 오류가 적으면서 간단한 근사 모델로서 좋은 일반화 성능을 보이는 모델을 찾음으로써 계산시간을 줄이고 적은 양의 학습 데이터를 가지더라도 우수한 성능을 나타내는 데이터 예측 모델을 제시하고자 한다.

4.2 유전적 프로그래밍에서의 선형모델 구현

일반적으로 회귀분석이나 함수근사의 문제는 주어진 샘플 데이터를 바탕으로 일반적 성능을 가진 우수한 모델을 찾는 것이다. 이를 위해서 가장 중요한 문제는 생성할 모델에 사용될 적절한 기저 함수 의 형식을 선택하는 것이다. 기저함수가 선택되면 이들의 조합을 통해서 적절한 모델을 생성해 간다. 여기에서는 근사 모델의 구조가 고정되어 있다. 이와는 달리 GP를 이용하게 되면 GP트리의 구조 자체가 점진적 진화연산에 의해 개선되고 최적화되어 더 적절하고 정교한 근사모델을 얻을 수 있는 확률을 높일 수 있다. 이러한 이유로 GP는 회귀분석이나 시스템 인식(System Identification) 분야에서 활발하게 적용되고 있다.[6]

4.2.1 MDL에 의한 최적 선형모델 생성

GP가 유전적 진화연산에 의해 생성되는 트리 구조를 사용함으로써 매우 잠재능력을 가진 도구로서 평가되지만 또한 이것이 GP의 큰 단점이 될 수도 있다. 예를 들어,

$\theta_1(+\,\theta_2 x_1\ \theta_3(*\ \theta_4 x_2\ \ \theta_5(\sin\theta_6(+\,\theta_7 x_3\ \theta_8 x_4))))$ 와 같은 트리를 생각하면, 여기서 θ_i 는 각 노드에 붙어진 가중치(Weight)이고 x_i 는 변수(Variable) 또는 상수(Constant)임, 이 트리는 비선형 함수로서 θ_i 를 구하기 위해서는 복잡하고 많은 양의 계산을 요하는 비선형 최적화 방법을 사용해야 한다.

일반적으로 좋은 모델을 생성하는 데는 모델의 일반화 성능이 우수해야 한다. 또한 이것은 모델의 복잡도와 연관되어 있다. 보통 모델이 복잡하면 학습 성능은 뛰어나지만(학습 오차가 적음) 테스트 오차가 매우 크게 되는 Overfitting 경향을 나타낸다. 이러한 현상은 학습 데이터의 수가 적을 때 발생하기 쉽다. 이러한 문제를 해결하기 위한 방법으로 본 논문에서는 일반적인 GP 트리 구조로부터 Symbolic Processing 알고리즘을 활용하여 선형 모델을 생성하였으며, 가장 성능이 뛰어난 모델을 선정하기 위하여 GP의 적합도 함수로서MDL 방법을 도입하였다. 여기서, MDL은 Ockham' s Razor[7]와 밀접하게 관련되어 있는데, 즉 학습 데이터를 잘 근사하면서 가장 단순한 모델이 가장 좋은 일반화 성능을 나타내는 모델이라는 것이다.

일반적인 선형 모델은 식 (1)과 같이 표현된다.

$$y = \sum_i \theta_i x_i = \underline{\theta}\,\underline{x}^T \qquad (1)$$

여기서, $\underline{\theta}$는 구하고자 하는 파라메터 θ_i의 벡터이고, $\underline{x}$는 d 차원의 설계변수이다. y는 θ_i의 선형 함수로서 선형 모델이라 한다. 식 (1)은 식 (2)와 같이 확장될 수 있다.

$$y = \sum_{i=1}^{\kappa-1} \theta_i b_i = \underline{\theta}\underline{b}^T \qquad (2)$$

여기서, $\kappa-1$은 기저함수(Base Function)의 수이고, $\underline{b} = (b_i)_{i=1}^{\kappa-1}$는 d차원 벡터 $\underline{x}$로부터 선택된 변수들의 임의의 연속 함수의 벡터이다. 식 (2)도 여전히 선형 모델이고, b_i는 표준 기저함수가 아니며, b_i와 b_j ($i \neq j$)는 같은 함수 형태가 아니다. 만일 학습 데이터 세트가 $L = \{(\underline{z}_i, t_i)\}_{i=1}^{n}$이라고 할 때, 여기서 $\underline{z}_i$는 d 차원 벡터 $(z_1^i, \ldots, z_d^i)$이고, t_i는 목표치(Target Value)임, GP의 임무는 L에 담겨진 정보로부터 $\underline{b}$를 찾아내는 것이다.

간단한 예를 들어 설명하면 다음과 같다.

전형적인 GP 트리로부터 선형 모델을 만들기 위해서는 먼저 식 (2)에서의 기저함수 b_i를 먼저 추출해야 한다. 만일 다음과 같이 생성된 GP 트리 하나를 생각해보자.

(- (* 0.7 (* 1.5 (sin (+ x_1 (* 0.3 (exp x_2)))))) (* (* 0.1 (* x_1 x_3)) (* x_1 (cos (- x_2 1)))))

이렇게 GP에서 생성된 트리 구조를 표준 수학함수 형태로 고쳐보면 다음과 같다.

$1.05\sin(x_1 + 0.3\exp(x_2)) + (-0.1)x_1^2 x_3 \cos(x_2 - 1)$ 이 식으로부터 우리는 다음과 같은 3개의 기저함수를 찾아낼 수 있다.

$$b_1 = 1, \quad b_2 = \sin(x_1 + 0.3\exp(x_2)),$$

$$b_3 = x_1^2 x_3 \cos(x_2 - 1)$$

그러면 θ_i는 OLS (Ordinary Least Square Method)에 의해 쉽게 구해진다. 이러한 변환은 궁극적으로 GP 트리로부터 가능한 모든 기저함수를 모으기 위한 것이다.

GP에서 사용되는 터미널 노드와 함수 세트는 다음과 같다.

$T_{GP} = \{x_1, \ldots, x_d, R, one\}$

$F_{GP} = \{g_1, g_2 \ldots, +, -, *\}$

여기서, R은 $|R| < 1$인 난수(Random Number)이고, 'one' 은 1이며, g_i는 임의의 연속 함수이다.

g_i로서 본 논문에서는 <Table 3>과 같은 다양한 수학 함수를 사용하고 있다. 또한 g_i로서 수학함수 대신 <Table 4>에 정의한 함수에 해당하는 Low order Taylor Series를 사용하게 되면 이경호[4]가 사용한 다항식 기반의 GP가 된다. 본 논문에서는 이러한 다항식을 이용한 GP 트리로부터의 선형모델 생성(이를 LM-GP라 함) 을 다루고 있는데, 이것을 특별히 PLM-GP (LM-GP with Polynomial)이라 한다.

4.2.2 가상 데이터 생성

일반적으로 MDL에 의한 선형모델 생성은 계산양을 줄여준다는 장점을 가지고 있지만 일반적으로 데이터의 양이 많은 때 좋은 결과를 보여준다. 본 논문에서와 같이 데이터의 수가 제한되어 있는 경우에 대해서는 좋은 모델을 만들어 준다는 보장을 할 수가 없게 된다. 더구나 이렇게 데이터의 수가 적고, 선형 모델의 기저함수 자체가 진화연산 과정에서 비선형성을 보일 때, 이렇게 생성된 모델은 과도한 Overfitting 경향을 보이게 된다. 이러한 문제를 해결하기 위하여 MDL과 병행하여 방향성을 가지고 가상 데이터를 생성 시킬 수 있는 DDBS(Directional Derivative based Smoothing) 방법의 도입이

<Table 3> Mathematic functions used for GP functions

cos, acos, sec, asec, sin, asin, csc, acsc, tan, atan, cot, acot, cosh, acosh, sech, asech, sinh, asinh, csch, tanh, atanh, coth, acoth, sqrt, exp, log(ln), iexp(1/exp)

<Table 4> Taylor series used for GP functions

Symbol	Math. function	Taylor series	Symbol	Math. function	Taylor series
tcos	$\cos(x)$	$1-1/2x^2$	tlsqrt	$(1+x)^{1/2}$	$1+1/2x-1/8x^2+1/16x^3$
tsec	$\sec(x)$	$1+1/2x^2$	tilsqrt	$(1+x)^{-1/2}$	$1-1/2x+3/8x^2-5/16x^3$
tsin	$\sin(x)$	$x-1/6x^3$	texp	$\exp(x)$	$1+x+1/2x^2+1/6x^3$
ttan	$\tan(x)$	$x+1/3x^3$	tllog	$\log(1+x)$	$x-1/2x^2+1/3x^3$
tcosh	$\cosh(x)$	$1+1/2x^2$	tilpx	$(1+x)^{-1}$	$1-x+x^2-x^3$
tsinh	$\sinh(x)$	$x+1/6x^3$	ti2px	$(1+x)^{-2}$	$1-2x+3x^2-4x^3$
ttanh	$\tanh(x)$	$x-1/3x^3$	texpsin	$\exp(\sin(x))$	$1+x+1/2x^2$
tlogcos	$\log(\cos(x))$	$-1/2x^2$	texptan	$\exp(\tan(x))$	$1+x+1/2x^2+1/2x^3$

요구된다.

Fig.1에서와 같이 선형 모델 y 가 주어지면, y 의 거동은 가장 인접한 두 샘플 포인트인 z_i 과 z_j 로부터 이 두 점을 연결하는 $^{i,j}l$ 선을 따라 탐색을 하면서 원치 않는 급격한 Peak나 Valley가 발생한 곳에서 y 의 방향 도함수를 사용하여 새로운 점들을 찾게 된다. DDBS를 사용하면 y 의 부드러운 결과를 효과적으로 얻을 수 있을 뿐만 아니라 샘플 데이터의 추가 없이도 효과적인 학습을 수행할 수 있게 된다.

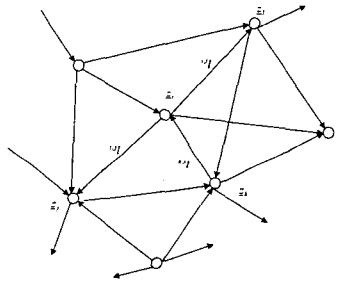


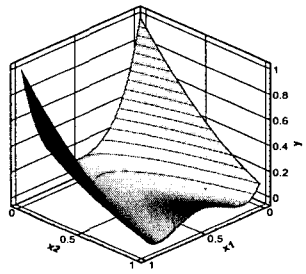
Fig. 1 Generation of Virtual Data by DDBS

4.3 공학 데이터의 적용

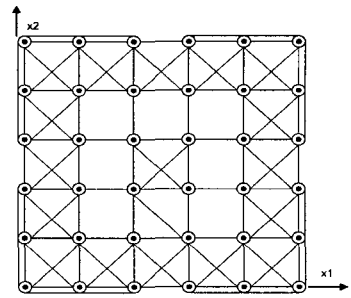
본 논문에서 개발된 GP의 선형모델을 검증하기 위해 수학적 함수의 근사를 수행하였다. 이것은 일반적인 표준GP가 학습 데이터의 수가 적을 경우 비정상적인 현상을 나타낼 수 있으며, LM-GP 또는 PLM-GP의 도입을 통해 이 문제를 해결 할 수 있음을 보이기 위한 것이다. 검증을 위해 사용한 함수는 Rosen Brock 함수이며, 단지 6x6개의 학습 데이터와 25x25개의 테스트 데이터를 그리드로 생성하여 사용하였다.

Rosen Brock 함수는 다음과 같으며, GP에 사용된 파라메터 값들은 <Table 5>와 같다. 또한 Fig.2는 그 결과이다. 예상대로 적은 수의 데이터에서도 LM-GP와 PLM-GP는 우수한 학습 성능을 나타내고 있음을 알 수 있다. Fig.3은 PLM-GP의 선형 모델로부터 변환된 다항식의 결과이다.

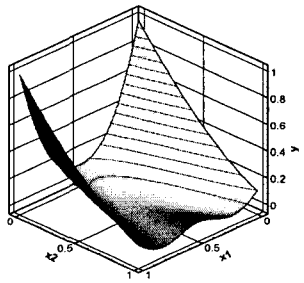
$$y = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$$
$$-2 \le x_i \le 2, i = 1,2$$



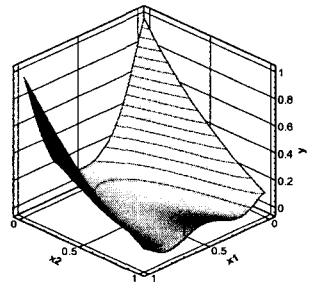
a. The original function.



b. Generated lines for creating virtual samples.



c. The best result of LM-GP.



d. The best result of PLM-GP.

Fig.2 Fitting results of the Rosen Brock's function with noiseless samples.

0.995 - 1.404x2 - 2.859x1 + 6.883E-1x2^ 2 + 3.217x1x2 - 1.407E1x1^ 2 - 4.718E-1x1x2^ 2 - 1.667E-1x2^ 3 + 1.182E2x1^ 3 - 1.481x1^ 2x2 + 6.665E-3x1x2^ 3 - 7.959E-1x1^ 3x2 - 1.957E 1x1^ 2x2^ 2 - 4.123E2x1^ 4 + 2.651E-1x1x2^ 4 - 9.611E-1x1^ 4x2 + 9.560E2x1^ 5 + 7.688E-1x1^ 2x2^ 3 - 2.171x1^ 3x2^ 2 - 1.663E3x1^ 6 - 1.050x1^ 4x2^ 2 - 2.941x1^ 5x2 - 1.750Ee1x1^ 2x2^ 4

.....

-1.017E-1x1^ 7x2^ 7 - 1.533E1x1^ 14 + 1.186E-1x1^ 12x2^ 2 - 2.209E-1x1^ 8x2^ 7 + 6.746E 2x1^ 11x2^ 4 - 2.791E-1x1^ 10x2^ 5 - 3.374E-2x1^ 12x2^ 3 - 5.438E-1x1^ 9x2^ 6 + 3.389E-2x1^ 9x2^ 7 + 2.283E-1x1^ 10x2^ 6 - 3.858E-2x1^ 12x2^ 4 - 7.216E-2x1^ 11x2^ 5 + 3.624E-2x1^ 12x2^ 5 + 2.455E-2x1^ 10x2^ 7 + 6.042E-2x1^ 11x2^ 6 - 3.008E-2x1^ 12x2^ 6 - 3.765E 3x1^ 11x2^ 7

<Table 5> The default parameters used for GP

Population size	300
Max. generation	100
Selection method	Tournament with 20 trees
Reproduction probability	0.15
Crossover probability	0.7
Mutation probability	0.15

5. 향후 연구

4절에서의 Fig.3의 결과를 보면 GP를 통해 얻어지는 데이터 근사 모델은 좋은 예측결과를 줄 수는 있으나 공학적으로 의미를 찾기 위해서는 생성된 데이터 모델이 물리적으로 의미를 가져야 하며, readability를 가져야 한다. 또한 이러함에도 불구하고 알려지지 않은 새로운 관계를 찾아낼 수 있어야 한다. 이를 위하여 Fourier knowledge를 이용한 파라메터 간의 관계를 미리 설정할 수 있어야 하고, Bayesian Statics를 이용하여 공학적으로 의미 있는 Term만을 가지고 갈 수 있어야 한다. 이러한 연구를 통하여 Evolutionary Computation, 특히 GP를 이용한 공학분야의 데이터 근사 모델의 생성이 가능하게 된다.

6. 결론

본 논문은 공학 분야의 축적된 데이터를 효율적으로 활용할 수 있도록 데이터 분석 및 성능 예측을 위한 공학 분야에 적합한 도구를 개발하는데 그 목적이 있다. 지금까지 공학 데이터의 축적을 위한 도구에 대한 노력과 연구는 많이 되어왔다. 그러나 이것의 활용 측면에서는 많은 연구가 없었던 것이 사실이다. 본 연구를 통하여 지식관리 측면에서의 기술지식을 정의하고, 의미 있고 활용 가능한 기술지식으로서의 축적된 공학 데이터 활용을 위한 진화 연산, 특히 유전적 프로그래밍 방법의 접근에 대해 소개하였으며, 유전적 프로그래밍 방법의 진화적 성능 향상을 도모하면서 간단하고 효율적인 선형(Linear) 모델의 개발을 통해 데이터의 일반화된 학습 성능을 높이고, 학습 데이터의 수가 적은 경우에도 뛰어난 학습 성능을 발휘하는 시스템을 개발하여 이를 검증하였다.

이러한 방법론은 조선 현장의 데이터를 효과적으로 활용할 수 있는 도구로 사용될 것으로 기대하며, 궁극적으로는 데이터로부터 유용한 정보 및 지식을 추출해 내는 데이터마이닝의 도구로 사용될 것이다

Acknowledgement

본 논문은 한국과학재단 첨단조선공학연구센터의 지원으로 수행된 연구 결과의 일부임을 밝힌다.

참고문헌

1. 이경호, 손미애, "표준화와 기술지식관리" , 대한조선학회지 제41권 3호, pp.15-26, 2004. 9.
2. 노나카 이쿠지로, "지식창조의 경영" , 21세기북스, 2001.
3. 오경주, "Data Mining with Case Studies" , 온톨로지 기반 지식 시스템 및 Digital Meister (한국 CAD/CAM 학회 응용연구회 발표집, pp.57-86, 2004
4. 이경호, 연윤석, 양영순 "개선된 유전적 프로그래밍 기법을 이용한 설계 파라메터 추정" , 대한조선학회 설계연구회 하계발표회, 1998. 8.
5. 이경호, 연윤석, "데이터마이닝을 위한 다항식기반의 유전적프로그래밍 기법과 조선분야 응용" 대한조선학회 춘계학술대회 논문집, 2004. 4.
6. Gray G.J., Murray D.J. and Sharman K.C., 1996, "Structural System Identification using Genetic Programming and a Block Diagram oriented Simulation Tool" Electronics Letters, Vol.32, pp. 1422-1424.
7. Barron A., Rissanen J. and Yu B., 1998, "The Minimum Description Length Principle in Coding and Modeling" IEEE Trans. Information Theory, Vol.44, No.6, pp. 2743-2760.