한국해양공학회 2006년도 추계학술대회 논문집 pp. 298~302

조선기술지식 관리를 위한 개선된 데이터 마이닝 시스템 개발

이경호*·양영순**·오준*°·박종훈*
*인하대학교 선박해양공학과
**서울대학교 조선해양공학과

Development of Enhanced Data Mining System for the knowledge Management in Shipbuilding

Kyung-Ho Lee*, Young-Soon Yang**, June Oh*° AND Jong-Hoon Park*,

*Department of Naval Architecture and Ocean Engineering , INHA University, Incheon, Korea
**Department of Naval Architecture and Ocean Engineering, Seoul National University, Seoul, Korea

KEY WORDS: Data Mining 데이터 마이닝, Knowledge Management 지식관리, Shipbuilding 조선, Genetic Programming 유전적 프로그래밍, Self Organizing Map(SOM) 자기조직화 지도, Engineering Knowledge 기술지식, Neural Network 인공신경망

ABSTRACT: As the age of information technology is coming, companies stress the need of knowledge management. Companies construct ERP system including knowledge management. But, it is not easy to formalize knowledge in organization. we focused on data mining system by using genetic programming. But, we don't have enough data to perform the learning process of genetic programming. We have to reduce input parameter(s) or increase number of learning or training data. In order to do this, the enhanced data mining system by using GP combined with SOM(Self organizing map) is adopted in this paper. We can reduce the number of learning data by adopting SOM.

1. 서 론

지금 사회를 흔히들 정보화 사회 또는 지식사회라고 한다. 정보화 기술의 발달로 인하여 기업은 컴퓨터에 의해 각 부분에서 들어오는 정보를 실시간으로 중앙에 집중시켜 그 정보로 생산과 판매과정을 제어할 수 있게 되었다. 인간은 실시간으로 모여드는 정보를 판단하여 컴퓨터에 지령을 주기만하면 된다. 즉 정보의 수집과 제품생산과 같은 단순 반복적인 일은 기계와 컴퓨터에게 맡기고, 인간이 가장 잘 할 수 있는 일인 '생각하는 부문'에 집중할 수 있는 것이다. 기업들은 반복으로 인한 효율성 저하를 막을 수 있으며, 인간이 가장 잘할 수 있는 부문인 '생각하는 부문과' 같은 부분에 조직원의 역량을 집중할 수 있게 함으로서 효율성을 높여 경쟁력 강화를 할 수 있다. 이것은 기업의 생존경쟁에 있어 커다란 힘을 발휘 할 수 있을 것이다. 세계를 단일시장으로 하고 있는 조선 산업의 경우에는 더욱 그렇다. 따라서 우리나라 기업들은 경쟁력 강화를 위하여 전사적인 ERP(Enterprise Resource Planning)의 구축과 함께 지식관리에도 힘을 쓰고 있다. 하지만, 형식화 되지 않은 지식을 관리하는 것은 매우 힘든 일이며, 지식관리를 하더라도 지식을 문서화하여 공유하고 있는 수준에 그치는 것이 현실이다(박우창 등, 2004 ; 이경호등, 2005-1). 여기서 우리의 관심은 기술지식(Engineering knowledge)에 있다. 기술지식은 축적된 공학 데이터에서 나올 수 있으며, 공학데이터 속에는 전문가의 경험과 노하우가 녹아들어있다. (이경호 등, 2005-1). 지식은 어느 관점에서 바라보느냐에 따라 여러 가지로 분류되지만 Table.1에서와 같이 형태에 따른 분류와 생성과정에 따른 분류로 나눌 수 있다.(이경호 등, 2004).

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

Table.1 Classification of knowledge

지식관리 관점의 기술지식을 한 마디로 정의하기는 쉽지 않다. 그러나 본 논문에서 대상으로 하고 있는 기술지식과 데

°오준: 인천시 남구 용현동 인하대학교 선박해양공학과 지능형 설계자동화 연구실 shipman98@lycos.co.kr

이터를 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 암묵적 지식과 기술지식이 녹아있는 데이터, 구조화되지 못한 지식요소 등, 데이터 마이닝을 통하여 지식을 얻어낼 수 있는 분석적 지식을 의미한다 (이경호 등, 2005-2).

현재 세계시장에서 선두에 있는 우리나라의 조선 산업은 지금까지 많은 배들을 건조하며 축적된 많은 데이터를 가지고 있다. 하지만, 이러한 데이터들을 활용하여 이로부터 유용한 정보/지식을 추출하기 위한 도구를 보유하고 있지 못한 것이 사실이다. 선행연구로서 "조선설계에서의 데이터 해석 및 활용을 위한 데이터 마이닝 도구 개발"(오준 등, 2006)을 수행하였고, 여기서 이러한 데이터들을 활용한 유전적 프로그래밍을 이용한 데이터 마이닝 도구(Tool)를 개발하였다. 하지만 실제 데이터 마이닝 도구의 학습에 필요한 만큼의 많은 수의 데이터를 구하는 것은 어려운 일이다. 또한 입력 파라미터의 개수가 많으면 많을수록 결과 값과 입력파라미터와의 관계를 수식화하기 힘들기 때문에 더욱 많은 학습데이터가 필요하다. 이러한 조선 분야의 특성 때문에 학습에 필요한 데이터의 개수를 줄이는 일은 매우 중요한 과제이다. 본 논문에서는 기존의 유전적 프로그래밍(Genetic Programming,이하 GP)만으로 학습하던 것을 개선하여 입력 파라미터의 영향도를 인공신경망의 한 종류인 SOM(Self Organizing Map, 이하 SOM)을 통하여 평가하고, 그 영향도가 높은 파라미터만 골라 학습데이터를 만들어 이를 통해 개선된 데이터 마이닝 도구를 통하여 학습한 결과와 영향도 평가 없이 학습데이터를 데이터 마이닝 도구를 사용하여 나온 결과 값과의 비교를 통하여 SOM을 통한 영향도 평가가 데이터 마이닝의 정확도에 얼마나 큰 영향을 미치며, 결과가 비슷하게 나온다면 데이터의 개수를 줄이는데 얼마나 영향을 미치는 가에 대하여 알아보고자 한다.

2. 도입배경과 지식

2.1 지난 데이터 마이닝 도구의 문제점

지난 데이터 마이닝 도구개발(오준 등 , 2006)에서의 문제점은 다음 Table.2과 같다. 첫째, 계산시간의 과다이다. 데이터 마이닝 도구는 GP를 사용하여 연산을 하기 때문에 개체수와 세대수의 증가할 경우 급격한 계산시간 증가를 가져온다. 둘째, 학습데이터가 부족할 경우에 부정확성이 많이 증가한다. 셋째, 결과로 나온 수식이 알아보기 힘들 정도의 많은 수식으로 이루어져 일반적으로 사용하기 힘들다. 위의 3가지 문제점 중 Table.2의 1번,2번 문제의 해결책은 앞에서 말한 바와 같이 SOM을 통한 영향도 평가를 통하여 영향도가 낮은 입력파라미터를 제거하여 입력파라미터의 개수를 줄임으로서 해결이 가능하다. 또한 3번의 경우 데이터 마이닝 도구 안에 수식을 자동으로 컴퓨터 코드(C코드)로 생성하도록 하였고, 이를 통해 다른 설계 프로그램과의 인터페이스를 할 수 있도록 하여 해결하였다.

2.2 SOM(Self organizing map)의 도입배경

SOM은 입력 및 출력 자료가 쌍의 형태로 주어지는 감독 제어형 학습(supervised learning)이 아니라 입력 자료의 규칙성이나 패턴을 찾는 비감독 제어형 학습(unsupervised learning)의 한 예이다.

번호	문제점
1	계산시간 과다
2	학습데이터 부족 → 부정확성 증가
3	수식의 복잡화

Table.2 Problems of data mining tool by using GP

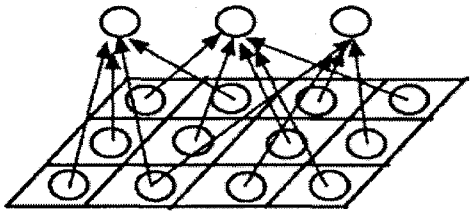

Fig.1 Clustering of self organizing map

즉 입력 벡터는 연속적으로 제시되지만, 요망되는 출력 값은 제시되지 않는다. 또한 SOM은 Fig.1과 같이 입력 자료들을 여러 클러스터로 그룹 짓는 클러스터링 망의 하나이다. 클러스터링 망의 출력 노드들은 각 클러스터를 나타내며 입력 노드들은 입력 벡터의 구성원들을 나타낸다. 각 출력 노드를 위한 가중치 벡터는 망이 각 클러스터를 위해 설정하는 것으로 입력패턴들에 대한 표본 벡터의 역할을 한다. 이는 Kohonen 학습 알고리즘을 사용하는 SOM의 특성중 하나이다. 가중치 벡터가 입력벡터와 가장 유사한 출력 노드(winner) 또는 그와 그의 이웃 노드들도 학습 시킨다. 즉 경쟁학습을 통하여 입력벡터와 가장 유사한 출력노드만이 살아남아 학습에 참여 할 수 있는 것이다. 이 특성을 이용하여 SOM을 통해 데이터마이닝 도구의 학습에 필요한 데이터에서 영향도가 높은 파라미터들을 선택하여 학습을 하게 된다면, 기존의 데이터마이닝도구의 학습에 필요한 입력 파라미터의 개수를 줄이거나 충분한 학습에 필요한 데이터의 수량을 줄일 수 있을 것이다.

3. 도입 및 실험

앞에서 말한바와 같이, SOM을 이용하여 데이터 마이닝 도구의 학습에 필요한 데이터의 입력파라미터의 개수가 줄일 수 있다면 데이터의 개수를 줄이는 효과를 내거나 좀 더 정확한 결과의 예측모델을 만들 수 있을 것이다.

3.1 SOM의 도입과 영향도 평가

먼저 Matlab을 이용하여 SOM Toolbox를 구성하였다. 그리고 데이터 마이닝 도구의 학습에 쓰일 데이터 2종류를 준비하였다. 한 가지는 본교 선형수조에서 실험한 모형선의 Trimaran 의 저항계수(C_t)를 구하는 실험 데이터 50건(Table.2)과 수식을 통하여 생성한 프로펠러의 추력토크를 구하는 데이터 200건(Table.3)을 이용하여 테스트를 진행하고자 하였다. 먼저 Table.2와 Table.3 은 다음과 같다.

	A	B	C	D	E	F	G	H	I	J
1	Fnm(Main)Fn	Rt(N)	Zf(mm)	Za(mm)	EHP(HP)	EHP(KW)	Hs(MM)	Sink	Trim	Ct
2	7.75E+05	0.185	0.9885	0.3404	1.9978	2.134	1.592	1.169	0.049	-0.07
3	1.24E+06	0.295	3.149	-0.8122	5.8636	12.095	9.023	2.526	0.042	-0.083
4	1.69E+06	0.402	7.2895	-3.6719	12.3407	40.582	30.274	4.334	0.038	-0.077
5	2.16E+06	0.514	11.8539	-13.4338	25.2689	85.209	63.566	5.918	0.032	-0.064
6	2.49E+06	0.593	13.9993	-17.6152	25.2035	114.606	85.496	3.794	0.015	-0.031
7	2.84E+06	0.677	16.0983	-19.9528	21.8232	147.944	110.366	0.935	0.003	-0.006
8	3.07E+06	0.732	17.3277	-19.4657	18.3475	170.143	126.927	-0.559	-0.001	0.003
9	3.18E+06	0.756	18.1897	-19.699	16.7942	184.276	137.47	-1.452	-0.004	0.007
10	7.55E+05	0.185	0.9885	0.3404	1.9978	2.208	1.647	1.169	0.049	-0.069
11	1.20E+06	0.295	3.149	-0.8122	5.8636	12.352	9.214	2.526	0.042	-0.083
12	1.64E+06	0.402	7.2895	-3.6719	12.3407	41.198	30.734	4.334	0.038	-0.077
13	2.10E+06	0.514	11.8539	-13.4338	25.2689	86.381	64.44	5.918	0.032	-0.064
14	2.42E+06	0.593	13.9993	-17.6152	25.2035	116.306	86.764	3.794	0.015	-0.031
15	2.76E+06	0.677	16.0983	-19.9528	21.8232	150.348	112.16	0.935	0.003	-0.006
16	2.99E+06	0.732	17.3277	-19.4657	18.3475	173.07	129.11	-0.559	-0.001	0.003
17	3.09E+06	0.756	18.1897	-19.699	16.7942	187.473	139.854	-1.452	-0.004	0.007
18	2.79E+06	0.54	12.6	-32.98	32.24	81.184	60.563	-0.37	-0.002	0.004
19	3.49E+06	0.675	16.61	-36.09	26.2	128.242	95.668	-4.945	-0.016	0.031
20	3.91E+06	0.756	17.74	-37.1	18.84	144.536	107.824	-9.13	-0.023	0.046

Table.2 Trimaran data set for GP learning

	J	P/D	AE/AO	Z		KT	KQ
1	0.163015	1.37341	0.419993		6	0.496533	0.092623
2	0.119045	0.916171	0.969013		4	0.403171	0.057487
3	1.08818	1.05607	0.696387		5	0.018542	0.008631
4	1.33196	1.29028	0.721632		5	0.020089	0.010514
5	0.07309	1.09491	0.775251		6	0.508137	0.081304
6	0.808987	1.35202	0.798638		5	0.321424	0.068983
7	1.38976	1.35379	0.954457		6	0.01504	0.012841
8	0.04966	1.2011	0.789544		5	0.556145	0.097842
9	0.787044	0.833019	0.470046		3	0.056851	0.010704
10	1.13723	1.1759	0.397485		3	0.064139	0.01596
11	1.0654	1.18767	0.427064		6	0.118658	0.029702
12	0.395319	0.812537	0.500755		3	0.203114	0.026705
13	0.869733	1.29985	0.989067		4	0.248424	0.05372
14	0.554828	1.03544	0.367652		4	0.251463	0.040331
15	0.409398	0.669151	0.390143		6	0.162154	0.021588
16	0.904711	1.06174	0.563769		3	0.102169	0.020509
17	0.494797	0.763019	0.958873		4	0.144525	0.021207
18	0.25723	0.87847	0.615503		4	0.30436	0.041279
19	0.836882	1.00925	0.626714		3	0.1056	0.020253
20	0.197465	0.801595	0.705943		4	0.29262	0.037108
21	0.641801	1.08188	0.892555		6	0.276178	0.048714
22	1.12264	1.25657	0.646976		3	0.088725	0.022041
23	0.421574	0.941788	0.371081		5	0.274089	0.039737
24	0.716429	1.12788	0.416899		4	0.231513	0.042068
25	0.393775	1.32573	0.993306		6	0.527713	0.104984

Table.3 KT data set for GP learning

SOM을 통하여 입력파라미터의 영향도를 평가한 결과는 다음 Fig.2, Fig.3과 같다. Fig.2에서 는 Zf, Hs, Sink의 값이 낮게 평가되었고, EHP의 값이 높게 평가 되었다. Fig.3를 보면 Z값이 영향도가 높게 나타나고 KQ의 값이 낮게 평가되었다.

3.2 영향도 평가전의 데이터 마이닝 도구의 학습결과

앞에서 행한 영향도 평가와 비교를 위하여 일단 입력 파라미터의 영향도를 측정하기 전의 학습 데이터를 가지고 데이터 마이닝 도구를 통하여 학습을 시켰다. Trimaran 모형선 실험의 경우 총 50건의 데이터 중 학습 40건과 테스트 10건, PLM-GP를 사용, 세대수 1500 으로 수행하였으며, KT 데이터의 경우 총 200건의 데이터 중 150건 학습, 테스트 50건,

PLM-GP를 사용, 세대수 500으로 실험을 수행하였다. 각각의

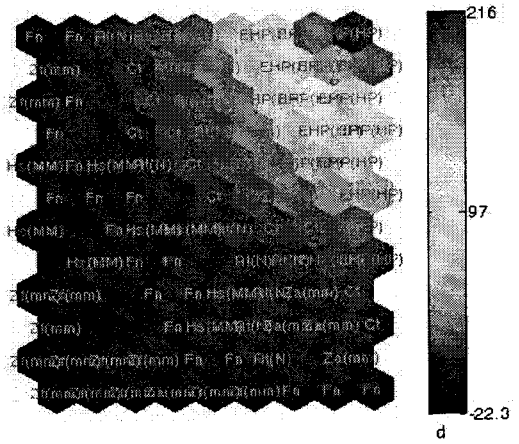

Fig.2 Effect evaluation of Input parameters for Trimaran Data by using SOM

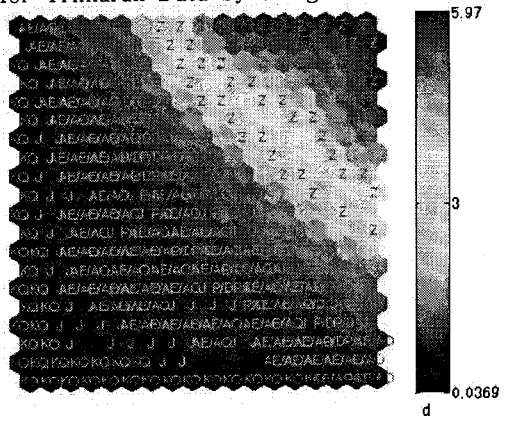

Fig.3 Effect evaluation of Input parameters for KT Data by using SOM

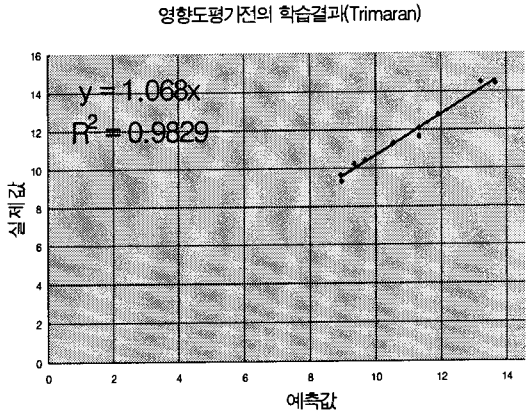

Fig.4 Test result by original GP for Trimaran data

실험을 실행한 후 나온 테스트 데이터의 결과는 다음 Fig.4와 Fig.5와 같다. 그러나 Fig.4의 경우 학습 데이터의 부족으로 상당한 오차를 보일 것으로 예상하였다. 하지만 예측 값과 결과 값의 좌표로 나타낸 기울기에서 1.068 정도의 값을 나타낸 것으로 보아 데이터 마이닝 도구의 함수 근사화가 100%

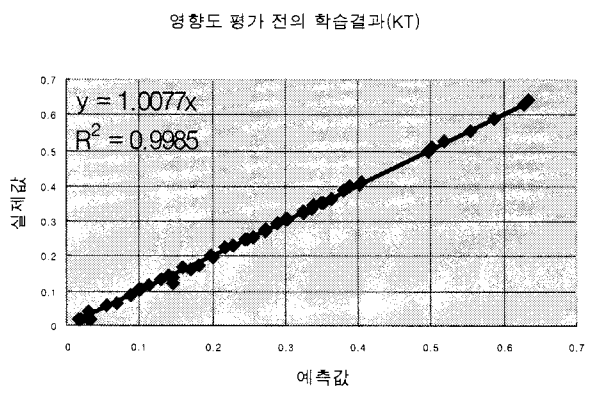

Fig.5 Test result by original GP for KT data

정확했을 경우 y=x의 기울기를 갖는 것과 비교했을 때 예상외로 정확하다고 할 수 있었다. 보통의 경우 데이터 마이닝 도구가 학습 데이터가 모자라 충분히 학습하지 못한 경우 y=x 그래프의 기울기가 Fig.4 보다 크게 나오는 경우(기울기의 크기가 1.2 이상)가 일반적이다. 데이터가 충분한 경우인 Fig.5의 경우에는 기울기는 y=1.0077x를 가지고 있어 만족할 만한기울기와 R^2(실제 값과 예측 값의 차를 제곱한 값) 값은 0.9985로 나타났으며 그래프 상에서 나타나는 점 또한 추세선(y=1.0x)에서 많이 벗어나지 않는 것을 볼 수 있다.

3.3 영향도 평가후의 데이터 마이닝 도구의 학습결과

3.1에서 행한 SOM을 이용한 영향도 평가를 통하여 나온 결과를 토대로 학습데이터를 변경하였다. 수식을 통해 생성한 Data Set의 경우 영향도가 제일 낮은 파라미터 KQ의 값을 삭제하여 파라미터의 개수를 줄였고, 실험을 통해 나온 Trimaran data set의 경우에는 Zf, Hs, Sink의 값을 삭제하고 데이터 마이닝 도구를 통하여 학습 후의 결과 값을 도출하였다(3.2절의 실험방법과동일한 조건으로 각각의 실험 수행). 다음 Fig.6 과 Fig.7은 영향도 평가후의 데이터 마이닝 도구의 학습결과 값을 차트로 나타낸 것이다. Fig.6의 경우는 함수의 값이 영향도 평가를 통하여 학습데이터에서 입력 파라미터의 개수를 줄여 줌으로써 실제 값과 예측 값의 차인 R^2의 값이 Fig.4에 비하여 줄어들었다. 또한 실제 값과 예측 값이 좌표로 찍힌 점이 Fig.4에 비하여 y=x의 추세선에 좀 더 가까운 것을 볼 수 있다. Fig.7의 경우 추세선의 기울기가 y=1.0077에서 y=1.0058로 약간 감소했고 실제 값과 예측값의 차의 제곱의 값인 R^2 또한 줄어들었다. 따라서 데이터의 개수가 부족한 경우 SOM을 이용한 영향도 평가를 통하여 입력 파라미터를 줄여 줌으로써 데이터 마이닝 도구의 학습을 좀 더 정확하게 시킬 수 있다는 결론을 얻을 수 있다.

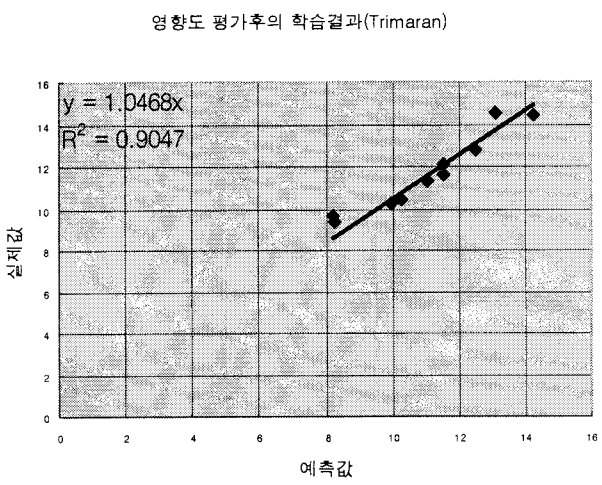

Fig.6 Test result by proposed GP system combined with SOM for Trimaran data

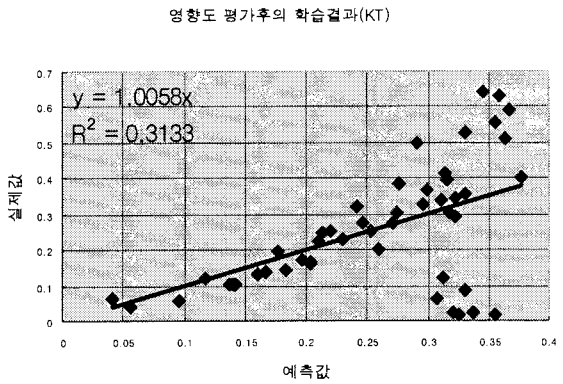

Fig.7 Test result by proposed GP system combined with SOM for KT data

4. 결론 및 향후 과제

SOM을 이용한 영향도 평가는 Fig.4~Fig.7에서 보듯이 데이터 마이닝 도구의 학습에 개선된 결과가 있다는 것을 알 수 있다. 데이터의 개수가 모자랄 경우 SOM을 이용하여 입력 파라미터를 줄여 줌으로써 데이터 마이닝 도구가 좀 더 정확한 학습을 할 수 있었고 좀 더 좋은 결과로 나타난다는 것을 확인 할 수 있었다. 즉, SOM을 활용하여 적은 수의 학습 데이터의 경우 입력 파라미터의 수를 줄임으로서 데이터의 개수를 늘인 효과를 볼 수 있었다.

후 기

본 논문은 대학원 공학석사 학위 논문 중의 일부 내용과 한

국과학재단 첨단 조선 공학 연구 센터 지원과제 (R11-2002-104-08002-0)로 수행된 연구 결과의 일부로서, 위 기관의 지원에 감사드립니다.

참 고 문 헌

- 김대수 (1992), 신경망 이론과 응용, pp169-183, 하이테크 정보사, 서울
- 박우창, 승현우, 용환승, 최기현 (2004) 데이터 마이닝, 자유아카데미, 서울
- 오준, 이경호, 박종현, 박종훈, 최영복, 장영훈 (2006), "조선설계에서의 데이터 핵석 및 활용을 위한 데이터 마이닝 도구개발", "대한조선학회 춘계 학술대회 논문집", pp.478~485
- 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 조선분야 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
- 임종수, (2005), MATLAB GUI PROGRAMMING, 도서출판 아진, pp260-350, 서울
- 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