Journal of the Society of Korea Industrial and Systems Engineering
Vol, 28, No. 4, pp.94－101, December 2005.

데이터마이닝 방법을 응용한 휴리스틱 알고리즘 개발

김 판 수*

*경북대학교 경상대학 경영학부

Development of Heuristic Algorithm Using Data-mining Method

Pan-Soo Kim*

*School of Business Administration, Kyungpook National University

This paper presents a data-mining aided heuristic algorithm development. The developed algorithm includes three steps. The steps are a uniform selection, development of feature functions and clustering, and a decision tree making. The developed algorithm is employed in designing an optimal multi-station fixture layout. The objective is to minimize the sensitivity function subject to geometric constraints. Its benefit is presented by a comparison with currently available optimization methods.

Keywords : data-mining method, fixture layout design, heuristic algorithm

1. 서 론

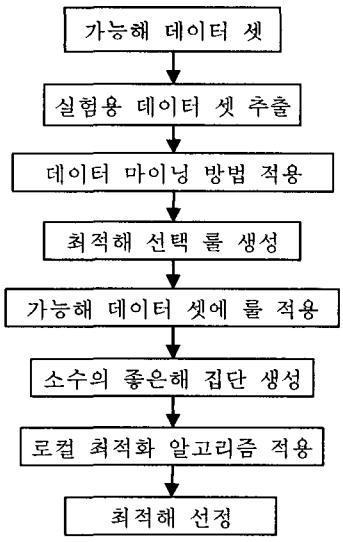
데이터마이닝은 CRM(customer relationship management) 혹은 마케팅 분야에서 주로 쓰이는 데이터 해석 방법이다. 데이터마이닝은 대용량의 데이터 셋에서 통계적 방법이나 모델링 방법과 같은 패턴인식기술을 이용하여 기존에 발견되지 않고 숨겨져 있던 의미 있는 데이터 사이의 관계를 밝혀내는 방법론으로 알려져 있다. 본 연구에서는, 데이터마이닝이 대용량의 데이터에서 데이터 간의 인과관계 혹은 숨어있는 패턴을 해석하는 과정을 통해 소수의 필요로 하는 정보를 찾아내어 가는 과정이라는 특징을 이용하였다. 본 연구는 데이터마이닝이 숨어있는 정보를 찾는 과정을 일반 최적화의 문제에 응용하여 일반 최적화 문제의 최적해(optimal solution) 혹은 글로벌 최적(global optimal)은 아니지만 짧은 시간 내에 의미 있는 해를 찾도록 휴리스틱으로 변형하여 개발하는데 목적이 있다.

목적함수가 비선형(nonlinear)인 수리모형(mathematical programming) 문제의 경우, 일반적인 좋은해를 찾는 방법으로는 비선형 방법혹은 씸플랙스 탐색법(simplex search method) [8] 등이 있다. 그 밖에 주로 사용되는 방법으로는 메타 휴리스틱(meta heuristic)으로 불리우는 유전자 알고리즘(genetic algorithm) 혹은 시뮬레이티드 어닐링(simulated annealing) [1]과 같은 알고리즘이 있고, 문제가 특이하지 않다면 이러한 메타 휴리스틱 알고리즘은 글로벌 최적해(global optimal)를 보장하지는 않지만 납득할 만한 좋은 해를 찾는다. 하지만 이러한 메타 휴리스틱은 해를 추적하는 기본적인 과정이 무작위 추출을 통한 임의해(initial solution) 선정 및 끊임없이 대안해를 찾아내어서 목적함수의 값을 비교하는 방법(random search)을 채택하고 있기 때문에 문제 정의를 위한 모수들(parameters)의 값에 따라 다소 차이는 있을 수 있겠지만, 일반적인 경우에는 총 계산시간(total computation time)이 아주 많이 소요되는 단점이 있다. 특히 제조공정에서 공정디자인을 최적화(design optimization)하는 문제의 경우, 혹은 시뮬레이션을 통해야만 목적함수 값을 알 수 있는 목적함수의 계산시간 자체가 짧지 않은 여러 의사결정 문제의 경우에는 총 계산시간은 아주 길어질 수 밖에 없는 구조적인 한계를 가지고 있다 [10].

* 이 논문은 2005년도 경북대학교 학술진흥연구비에 의하여 연구되었음

2. 관련연구

데이터마이닝을 응용해서 좋은해를 구하기 위한 기존의 연구로는 Schwabacher et al. (2001) [10]와 Igusa, Liu, Schafer and Naiman (2003) [4]의 연구가 있다. 이 연구들의 기본적인 아이디어는 <그림 1>과 같다. 기본 개념은, 알고리즘의 단계를 거쳐 가면서 가능해 데이터 셋의 대상 데이터 개수를 줄여가고, 시간이 많이 걸리는 목적함수를 계산해야 하는 로컬 최적화 기법은 최종 단계에 적용하여 좋은 해를 짧은 시간에 찾는 원리이다. 데이터마이닝의 측면에서 본다면, 가능한 모든 해(candidate solutions)의 집단을 아주 많은 양의 가능해 데이터 셋(data set)으로 보고, 데이터마이닝 방법이 이 데이터 셋으로부터 의미 있는 데이터의 패턴을 발견하여 좋은해를 선택하기 위한 룰을 발견해 낸다. 이 룰을 가능해 데이터 셋에 적용하면 최적해가 될 가능성이 많은, 소수의 좋은해 집단을 찾을 수 있고, 이 소수의 좋은해 집단에 우리가 알고 있는 로컬 최적화 알고리즘을 적용하면 만족할 만한 품질의 로컬 최적해를 찾을 수 있다. 여기서 주로 사용되는 데이터마이닝 방법은 분류법(classification method)이다.



<그림 1> 기존연구의 흐름도

하지만, 이 방법을 사용하기 위해서 우선 해결해야 하는 몇 가지 문제가 있다. 그것는 데이터마이닝 방법 혹은 분류법을 적용해야할 실험용 데이터 셋을 어떻게 결정하는가 하는 문제 이다. 데이터 패턴을 찾기 위해 사용되는 데이터마이닝 방법 특히 분류법은 목적함수의 값을 계산하여 그 값의 유사성으로 진행된다. 그러므로, 우리가 좀더 관심이 있는, 목적함수의 계산시간 자체가 길어 가급적 목적함수 계산횟수를 줄여야 하는 문제의 경우에 가능해 데이터 셋을 대상으로 분류법 혹은 기타 데이터마이닝 방법을 적용 한다면 목적함수 계산에 너무 많은 계산시간이 소요되어 데이터마이닝 방법을 응용하여 좋은해를 찾는 의미가 없어진다.

3. 개발 알고리즘 적용 내용 및 특성

기존의 연구에서는 가능해 데이터 셋으로 부터 패턴을 발견 하기위한 실험용 데이터 셋 추출 방법으로 무작위 추출(random sampling)을 하거나 경험에 의해 목적함수 값을 알고 있는 데이터들을 이용하였다. 예를 들어, Schwabacher et al. (2001) [10]의 연구에서는 경주용 요트나 비행기 설계를 위한 프로토타입 선정을 위한 알고리즘에 데이터마이닝 방법을 사용했는데, 그 때 사용한 실험용 데이터 셋은 경험 데이터를 이용해서 작성하였다. 하지만 만약 많은 수의 경험 데이터를 가지고 있지 못한 경우라면, 이 방법을 통해 찾아진 실험용 데이터 셋이 가능해 데이터 셋의 특성을 잘 반영한다고 보기 힘들다. Igusa et al. (2003) [4]는 무작위로 다수의 데이터를 추출한 다음 몇 가지 과정을 거쳐서 실험용 데이터 셋으로 사용하였다. 이 방법 역시 무작위로 추출한 데이터가 기존의 가능해 데이터 셋의 기하학적인 특징을 잘 반영한다고 보기 힘들고, 얼마만큼의 무작위 데이터가 적절한 양의 실험용 데이터 셋인지 알기 힘든 단점이 있다. 본 연구에서는 가능해 데이터 셋으로부터 실험용 데이터 셋을 생성하는 방법에 대한 새로운 연구가 우선적으로 이루어 졌다.

실험용 데이터 셋은 가능해 데이터 셋의 특징을 잘 반영하면서 데이터의 개수는 적으면 적을수록 총 계산시간을 줄이는데 유리하다. Igusa et al. (2003)는 무작위 추출법 이외에 계산시간이 짧은 특성함수(feature function)를 소개하여 목적함수의 계산횟수를 줄였다. 특성함수는 목적함수의 특성을 같이 가지면서 계산시간은 빠른 특징을 가지도록 정의된 함수이다. 또한 군집법(clustering)이라고 불리는 또 다른 데이터마이닝 방법을 무작위로 추출된 데이터들의 특성함수에 적용하여 실험용 데이터 셋을 생성하였다. 군집법은 데이터 셋을 목적함수의 값이 아닌 데이터 간의 수리적 거리(euclidian distance)로 군집을 형성하는 방법이다 [3].

본 연구에서는 특성함수와 군집법을 이용하기 이전에 대표 데이터 셋을 가능해 데이터 셋에서 추출하는 균일화 알고리즘을 개발하였는데, 개발의 가장 큰 이유는 무작위 추출법으로는 가능해 데이터 셋으로부터 기하학적

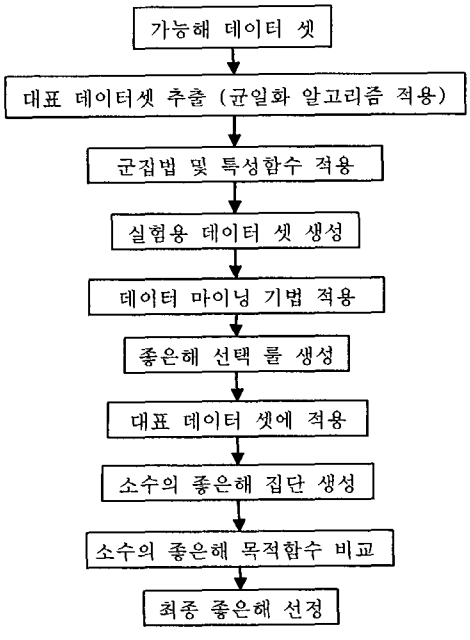
으로 균일한 데이터를 추출하기 어렵기 때문이다. 그것은 통계적인 균일화(uniformity)가 기하학적인 균일화를 보장하지는 못한다는 의미이다. 균일화 알고리즘은 공간충전 디자인(Space Filling Design) [9]의 개념을 이용하여 개발되었다.

공간충전 디자인은 실험계획법에서 많이 쓰이는 디자인 방법인데, 이 방법의 목적은 실험대상 데이터(design point)들을 실험대상 공간(design space)으로부터 기하학적으로 일정한 간격을 가지고 균일하게 추출하여, 실험대상 데이터가 실험대상 공간의 물리적인 특징을 잘 반영하도록 하여 실험결과의 대표성을 높이려는 것이다. 공간충전 디자인에는 라틴격자 샘플링(Latin Hypercube Sampling) [7] 혹은 NTM(Number Theoretic Method) [2]방법 등이 있고, 본 연구는 라틴격자 샘플링의 개념을 이용하였다. 균일화 알고리즘에 의해 생성된 대표 데이터 셋은 가능해 데이터 셋의 기하학적 특징을 잘 반영하는 데이터(design point)들의 최소한의 개수로 이루어진다. 이 단계에서 데이터의 개수는 적을수록 나중 단계에서 목적함수를 계산해야 하는 데이터마이닝 방법의 적용시간이 짧을 것이다. 하지만, 데이터의 개수가 지나치게 적으면 데이터 셋의 물리적 혹은 기하학적 특징을 잘 반영하지 못하게 되어 이후 단계에서 생성되는 좋은해 선택 룰의 품질이 좋기 힘들다. 그러므로 가능해 데이터 셋의 특징을 잘 반영하는 최소수의 데이터로 대표 데이터 셋을 구성하는 것이 이 단계에서의 최우선 목표이다.

다음 단계는 목적함수의 특징에 맞는 데이터를 다시 한번 더 추출하고, 데이터의 개수 역시 최소한으로 다시 한번 조정한 실험용 데이터 셋의 생성이다. 본 연구에서는 균일화 알고리즘에 의해 생성된 대표 데이터 셋에 Igusa et al. (2003)에 의해 소개된 특성함수와 군집법을 적용하여 실험용 데이터 셋을 생성하였다. 특성함수는 어떠한 수학적 이론보다는 좋은해를 찾기위한 대상 시스템에 대한 전문가의 공학적 지식에 의해 결정되므로 좋은 결과를 이끌어내는 특성함수의 선택은 쉽지 않고, 시스템의 특징에 따라 다르게 정의된다. 가능하면 목적함수를 직접 계산하지 않고 목적함수의 최적화 특징을 가진 데이터를 추출하려는 것이 특성함수를 사용하는 가장 큰 이유이다. 군집법은 대표 데이터 셋의 데이터들의 목적함수가 아닌, 특성함수 값들의 상대적 거리가 가까운 데이터 끼리 군집을 형성한다. 몇 개의 군집으로 나누어야 하는 것은 사전에 제시 되어야 하는 모수(parameter)이다. 군집법 적용 후, 각 군집에서 몇 개의 데이터를 뽑아서 실험용 데이터 셋을 생성한다. 여기서 추출되는 데이터의 개수 역시 중요 모수이다. 군집법에서 생성된 각 군집은 로컬해들의 모임이므로 각 군집으로부터 대표되는 몇 개의 데이터들을 추출하면 이 대표되는 소수의 데이터들이 좋은해를 선택하는 좋은 품질의 룰을 제공해 줄 수 있다는 것이 본 연구에서 군집법을 사용하는 중요 개념이다.

그 다음 단계는 데이터마이닝 방법의 적용으로, 본 연구에서는 알려진 데이터마이닝 방법 중 분류법(classification method)을 적용한다. 분류법은 실험용 데이터 셋의 패턴발견을 통해 의사결정나무(decision tree)를 생성한다. 의사결정나무 생성은 일종의 회귀(regression)방법으로 종속변수인 목적함수에 대한 예측에러가 작은 방향으로 각 특성함수 값의 조건을 분화시킴으로써, 분화된 특성함수의 조건들이 트리 형태로 생성되어지게 하는 방법이다. 분화된 조건 트리를 따라가면서 해석을 하면 좋은해 선택 룰(if-then rule)이 생성된다. 이 룰의 조건을 만족하는 데이터들을 대표 데이터 셋에서 추출하면 소수의 좋은 해 집단을 찾을 수 있다.

최종 단계에서는 소수의 좋은해 집단에 알려진 로컬 최적화 알고리즘을 적용하거나 혹은, 개수가 많지 않다면 목적함수를 모두 계산하여 그 중 가장 좋은 해를 찾는다. 이 해가 본 연구가 제시하는 데이터마이닝 방법으로 찾을 수 있는, 목적함수의 계산횟수를 최소화 하는 좋은 품질의 해가 된다. 물론 이 해는 수학적으로 완전한 최적해(global optimal)는 아니다. 하지만, 보다 짧은 시간에 충분히 받아들일 수 있는 수준의 해가 된다.



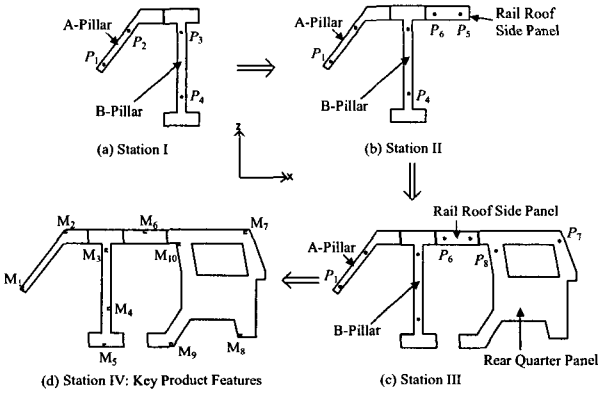
<그림 2> 제안된 최적화 프로세스

전체적으로 이 알고리즘에서 고려해야 하는 총 세 가지의 중요 이슈를 요약하면 다음과 같다. 첫 번째는 대

표 데이터 셋 추출법이다. 이를 위해 본 연구에서는 실험계획법의 공간충전 디자인을 응용한 균일화 알고리즘을 개발하였다. 두 번째 이슈는 특성함수의 정의와 군집법의 적용이다. 특성함수의 정의는 해결하고자 하는 예제의 특성을 파악한 후에 정의되어야 한다. 군집법 적용을 위해 몇 개의 군집으로 나누어야 하는지는 결정해야 하는 연구, 그리고 각 군집으로부터 몇 개의 데이터를 추출해야 하는지에 대한 연구 역시 필요하다. 마지막으로 의사결정나무의 생성에 관한 연구가 필요하다. 최종적으로 분류법에 의해 산출되는 좋은해 선택 룰이 최종 좋은해의 품질을 결정하기 때문에 이 연구 역시 중요하다. 어떻게 의사결정나무를 생성하고 해석 할 것인지에 관한 연구가 필요하다. 제안된 데이터마이닝 방법을 응용한 알고리즘의 흐름은 <그림 2>에 예시되어 있다.

4. 케이스 연구

3장에서 소개된 알고리즘을 적용 할 예제로는 자동차 조립공정에서 판넬(panel)을 고정시키기 위한 고정구의 위치선정 문제를 이용하였다. <그림 3>에 소개된 조립공정은 SUV차량의 측면 조립공정 중 용접공정의 부분이다.



<그림 3> 자동차 측면 판넬 조립공정

4.1 수리적 모델

<그림 3>은 station 1에서 station 4로 공정이 진행되면서 판넬이 공정마다 하나씩 용접되어 추가 되면서 4장의 판넬이 최종 station 4에서 한 장의 SUV의 측면 차체로 완성되는 조립 공정의 흐름을 보여준다. 이 공정의 흐름에서 필요한 총 4개의 판넬을 지지하기 위해서 각 판넬 당 두 개씩 총 8개의 고정구가 필요하다. 고정구의 위치변화에 따라 품질특성이 변화하므로 고정구의 위치는 품질향상에 아주 많은 영향을 미친다. <그림 3>에서 고정구는 $P_1 \cdots P_8$ 로 표현되었고 센서(sensor)의 위치는 $M_1 \ \ldots M_{10}$로 표현되었다. 위 예제에서의 품질특성은 station4에서 완성품의 센서 위치에서의 흔들림(variation)으로 정의 되었다. 본 예제는 이 품질특성을 최소화하는 고정구의 위치를 선정하는 문제로 요약된다. 가능한 고정구의 좌표들은 가능해 집합이 되고, 품질특성은 목적함수가 된다. 목적함수는 고정구의 위치에 의해 결정되는 품질특성함수(sensitivity function)를 사용하였다.

판넬 조립 공정에서의 다 공정 품질특성함수는 기존의 연구에서 이미 개발되었다 [5][6]. 품질특성함수 S는 비 선형 함수로 고정구의 위치 좌표에 의해서 값이 정해지는 함수이다. 고정구는 하나의 판넬 당 2개씩 총 8개의 고정구가 필요하며 전체 문제의 크기를 줄이기 위해 2차원 공간을 가정하면 고정구의 위치 좌표는 16×1의 벡터 즉, $\Theta = [x_1 y_1 \cdots x_8 y_8]^T$로 정의된다. 여기서 x_i과 y_i는 고정구 P_i의 좌표이다. 이 정의를 이용해서 본 연구에서 사용될 수리모델을 정의하면, 목적함수인 품질특성함수 S를 최소화 시키면서 고정구가 기하학적으로 판넬 상에 위치하도록 지역적 제약식 $G(\cdot)$을 만족시키는 고정구 좌표 벡터 Θ를 찾는 문제로 요약할 수 있다. 이를 수리적으로 표현해 보면,

$$\min S(\Theta) \quad subject\ to \quad G(\Theta) \geq 0$$

이다.

여기서 지역적 제약식 $G(\cdot)$는 고정구의 위치는 판넬상에 위치해야 한다는 의미이다. 위 최소화 문제의 해를 효율적으로 찾는 알고리즘의 효율성은 주로 얼마나 자주 목적함수 $S(\cdot)$를 계산하느냐에 많은 영향을 받는다. $S(\cdot)$를 한번 계산하는데 걸리는 시간을 T로 표현한다. 위 예제의 해를 찾기 위해서 판넬의 표면을 10mm 단위로 좌표화 시켜서 가능해 데이터 셋의 개수를 계산하여 보면 $C_2^{239} \times C_2^{707} \times C_2^{200} \times C_2^{3496} \approx 8.6 \times 10^{20}$ 개 이다. 이는 아주 많은 수로 이 많은 수의 가능해 데이터 셋에 목적함수를 계산해야하는 분류법 같은 데이터마이닝 방법을 직접 적용하면 아주 긴 계산시간이 소요 될 것이다.

4.2 대표 데이터 셋 추출

데이터마이닝을 응용한 휴리스틱 알고리즘의 첫 번째 단계는 대표 데이터 셋 추출이다. 3장에서 설명한 공

간충전 디자인을 이용한 균일화 알고리즘을 적용 하는데, 이 방법은 직사각형이 아니거나 좀 더 불규칙적인 모양의 실험 공간에서도 균일한 추출을 할 수 있는 휴리스틱으로 개발되었다. 또한, 사전에 알고 있는 본 예제에 대한 공학적 지식중 하나인 고정구 사이의 거리가 멀면 멀수록 품질특성이 더욱 좋아진다는 지식을 바탕으로 알고리즘이 개발되었다.

대표 데이터 셋이 가능해 데이터 셋의 특징을 잘 반영하면서도 전체 가능해 영역에 걸쳐 골고루 추출된, 그리고 두 고정구 사이의 간격이 가까이 가지 않도록 하는 특징을 반영한 균일화 알고리즘은 다음과 같다.

Step 1. 판넬의 표면을 10mm단위로 좌표화 한다.

Step 2. 각 판넬에서 첫 번째 고정구는 좌표의 순서에 따라 순차적으로 선택한다. 첫 번째 고정구가 선택 되면 두 번째 고정구는 첫 번째 고정구로 부터 각 판넬 크기의 1/2보다 더 멀리 떨어진 좌표들 중에서 무작위로 하나를 선택한다. 결과로 이루어진 고정구 쌍의 집합을 $\Phi_j^{(i)}$라고 표시한다. 여기서 j는 판넬의 번호이고, i는 해당 판넬에서 고정구 쌍의 일련번호, n_j는 해당 판넬에서 추출된 총 고정구 쌍의 개수이다.

Step 3. 본 예제에는 총 4개의 판넬이 있으므로 모든 판넬에 대해서 Step 2를 수행하면, $\Phi_j^{(i)}$, $i = 1 \ldots n_j$, $j = 1..4$가 결정된다.

Step 4. 비 복원 추출로 네 개의 $\Phi_j^{(i)}$에서 고정구 위치한 쌍씩을 추출한다. 추출된 고정구 네 쌍을 하나의 대표 데이터로 저장한다.

Step 5. 네 $\Phi_j^{(i)}$ 중 어느 하나가 소진되면 해당$\Phi_j^{(i)}$를 $\Phi_j^{(0)}$로 재 지정한다.

Step 6. 가장 많은 수의 고정구 쌍($\Phi_4^{(i)}$)이 소진될 때 까지 Step 4와 Step 5를 반복한다.

본 예제에서는 네 번째 판넬의 크기가 가장 크고, 10mm단위로 좌표화 시킨 후에 3,496개의 좌표가 생기므로 $\Phi_4^{(i)}$의 n_j는 3,496이다. 균일화 알고리즘은 가장 큰 판넬의 좌표 수 만큼 고정구 쌍이 생기므로 알고리즘 결과로 생긴 대표 데이터 셋의 데이터의 개수는 3,496개이다. 결국 균일화 알고리즘을 통해 가능해 데이터 셋에서의 8.6×10^{20}개의 데이터에서 3,496개의 데이터를 추출하여 대표 데이터 셋을 생성함으로써 고려해야 하는 대상 시스템의 사이즈를 일차적으로 크게 줄였다.

4.3 실험용 데이터 셋 생성

실험용 데이터 셋 생성을 위해서 가장 먼저 해야 하는 일은 특성함수의 정의이다. 특성함수는 목적함수의 특징을 잘 반영하면서 계산하는데 시간이 많이 소요되지 않는 고정구 레이아웃 벡터 Θ의 함수로 정의 된다. 특성함수를 잘 정의하기 위해서는 해당 문제 시스템에 대한 사전 지식(knowledge)이 필요하다. 본 예제에서는 고정구 간의 거리가 목적함수에 많은 영향을 미친다는 것을 공학적 사전지식으로 알고 있으므로 고정구간의 위치를 다시 한번 이용하여 특성함수를 정의 하였다.

*f1*은 고정구 간의 거리 중 가장 작은값, *f2*는 두 번째로 작은 값, *f3*는 중간값, *f4*는 두 번째로 큰값, *f5*는 가장 큰 값으로 정의 되었다. *D1*과 *D2*는 두 판넬이 조립되기 전 각 판넬의 고정구 간의 거리이고, *D3*는 두 판넬이 조립된 후의 고정구 간의 거리라고 하면, 고정구 거리 변화 비율 r은 다음과 같이 정의된다.

$$r = \frac{D_3}{(D_1 + D_2)/2}$$

계산된 r값에 의해 *f6*는 r의 최소값, *f7*는 중간값, *f8*는 최대값으로 정해진다. 각 값들은 고정구간의 거리 및 고정구 거리 변화 비율의 상대적인 크기이므로 각각의 분포로써 이해될 수 있다. 이를 이용한 특성함수의 정의는 다음과 같다.

$f_1(\Theta) \sim f_5(\Theta)$= 고정구 간의 거리의 분포

$f_6(\Theta) \sim f_8(\Theta)$= 공정을 진행 하면서 변화한 고정구의 거리 변화 비율의 분포

위에서 정의 된 특성함수의 값에 군집법(clustering method)이 적용된다. 데이터마이닝 방법에서 많이 쓰이는 군집법은 *k-means* 방법이다. *k-means* 방법은 모수 k가 주어지면 데이터들의 상대적 거리의 근접성에 따라 데이터 셋을 k개의 군집으로 분류해 준다. 이 방법을 수식으로 표현해 보면 다음과 같다.

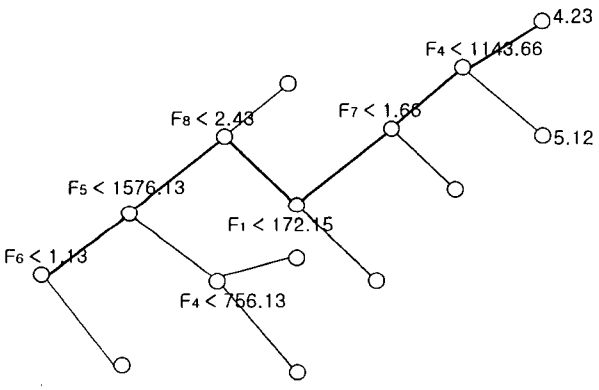
$$\min \sum_{k=1}^{K} N_k \sum_{C(i)=k} | F_i - m_k |^2$$

여기서, N_k는 군집 k 내 데이터들의 개수이고, m_k는 군집의 중앙, F_i는 특성함수의 결과로 나온 값들의 벡터, $| \cdot |$은 2-norm 벡터이다. 이 방법을 사용하기 위해서 미리 정해져야 하는 중요한 모수는 군집의 개수 k이다. 본 예제에 대한 k의 결정은 4.5절 에서 다룬다.

군집법의 결과로 생성된 k개의 군집으로부터 일정량의 데이터들을 무작위로 추출하여 실험용 데이터 셋이 형성되는데, 추출되어야 하는 데이터 개수 n 역시 중요 결정 모수이다. 본 예제에 대한 모수 n의 결정 역시 4.5절에서 설명 한다.

4.4 좋은해 선택 룰 생성

군집법의 결과로 k개의 군집에서 n개의 대안 데이터들을 추출하면 kn개의 실험용 데이터 셋이 형성되고, 이 실험용 데이터 셋에 분류법을 적용하면 의사결정나무가 생성된다. 본 연구에서는 CART(classification and regressin tree) 방법이 one-standard error를을 기반으로한 ten fold cross validation의 트리구조를 결정하기 위해서 분류법으로 사용되었다 [3]. 분류법에 의해 생성된 의사결정나무의 예는 <그림 4>에 예시되었다.



<그림 4> 의사결정나무 생성의 예

의사결정나무의 해석은 다음과 같다. 각 노드의 윗부분의 숫자는 각 노드에서 행해지는 조건이다. 가장 오른쪽의 마지막 노드에 나와 있는 숫자는 해당조건을 만족하는 노드에 해당하는 데이터들의 목적함수 값의 평균치이다. 왼쪽 끝의 시작 노드로부터 오른쪽 노드 중 최소 목적함수 값을 가지는 노드를 연결해주는 선이 *if-then* 룰로 해석되는데, 왼쪽 끝 노드에서 시작하여 만약 조건을 만족하면 위쪽의 선을 따라가고, 조건을 만족하지 않으면 아래쪽 선을 따라가게 된다. 이 룰이 좋은해 선택 룰이고, 하나의 수리적 조건식으로 표현 될 수 있다.

4.5 최종 좋은해 선택 및 중요 모수의 결정

마지막 단계에서는 의사결정나무에 의해 생성된 좋은해 선택 룰을 3,496개의 대표 데이터 셋에 적용한다. 여기서 적용의 의미는 모든 대표 데이터 셋의 데이터를 대상으로 생성된 좋은해 선택 룰에 의해 생성된 조건식으로 조건 검색하여 조건을 만족하는 데이터들을 추출하는 것을 의미한다. 특성함수는 계산에 시간이 소요되지 않는 특징을 지니도록 정의 되었으므로 여기서 소요되는 시간은 아주 짧다. 적용 결과로 소수의 좋은해 집단이 추출된다. 몇 개의 데이터인지는 k와 n에 따라 다르다. 소수의 좋은해 집단은 좋은 품질의 데이터 집단이므로 이 집단에 알려진 일반 휴리스틱 알고리즘을 적용해도 되고, 숫자가 적으면 소수의 좋은해 집단의 목적함수를 모두 계산하여 가장 좋은 목적함수를 가진 데이터를 선택해도 된다. 본 연구에서는 후자의 방법을 통해서 제안 된 데이터마이닝을 응용한 휴리스틱 알고리즘의 해로 선정하였다.

분류법을 적용하기 위해서 결정되어져야 하는 모수인 k와 n의 결정은 많은 의미를 가지고 있다. 전체 휴리스틱의 효율성에 이 두 변수가 많은 영향을 미치기 때문이다. 전체 휴리스틱을 수행하기 위한 시간을 수식으로 나타내면 $T_1 + knT + N_0T$이다. T_1는 목적함수의 계산과 상관없이 공통으로 소요되는, 대표 데이터 셋 추출 등에 사용되는 시간이고, kn는 분류법의 적용대상 데이터 개수, 그리고 N_0는 분류법의 결과로 생성되는 소수의 좋은 해 집단의 데이터 개수, 그리고 T는 목적함수 계산 시간이다. 위 수식을 해석해 보면 T_1은 공통 소요시간, knT는 분류법 적용시간이 되고, N_0T는 최종 좋은해 선택을 위한 목적함수 비교시간이 된다.

물론 k와 n이 크면 클수록 정보가 많아지므로 이용해서 생성되는 의사결정나무가 정교한 좋은해 선택 룰을 제공해서 N_0의 숫자를 줄일 것이다. 하지만, kn이 커지므로 분류법에 많은 시간이 소요되어 전체 휴리스틱의 효율에는 도움이 되지 않는다. 반대로 k와 n이 작다면, knT는 줄어들겠지만, 의사결정나무가 많은 정보를 제공하지 못하여 N_0의 숫자가 늘어나게 되어서 역시 총 계산시간이 길어지게 된다. 이 때문에 가장 효율을 높일 수 있는 적절한 k와 n을 정하는 작업이 중요한 의미를 가진다.

본 연구에서는 k = 3, 6, 9 그리고 n = 5, 10, 15의 레벨로 3^2요인 실험계획법을 통해서, k=9 와 n=12를 결정하였다. 본 예제의 목적함수 인 품질특성함수 계산시간은 0.018초 시간이 2.20GHz P4 프로세스에서 걸렸다.

4.6 결과의 비교

결과의 비교는 첫째, 단순 비선형 최적화 방법인 씸플

랙스 탐색방법, 그리고 본 연구의 예제와 같은 복잡한 프로세스 디자인 최적화 문제에서 일반적으로 좋은 해를 생성한다고 알려진 시뮬레이티드 어니일링 방법과 비교 하였다. 시뮬레이티드 어니일링 방법의 효율에 가장 많은 영향을 주는 어니일링 계수 k_B는 주로 많이 쓰이는 0.95와 0.9 두 가지의 경우 모두와 비교 하였다. 모든 비교실험은 같은 프로세스의 PC에서 MATLAB ver6.5로 수행되었다. 결과의 비교는 <표 1>과 같다.

결과치는 10회의 실험의 평균치를 계산하여 비교하였다. 표준편차는 그 값이 크지 않아 표에서 비교하는 것은 생략하였는데, 데이터마이닝 방법의 경우 목적함수 값의 표준편차는 0.03 미만, 계산시간의 표준편차는 6.3 미만으로 결과의 추론에 영향을 미칠 만큼의 수치는 아니었다.

씸플랙스 추적(Simplex Search)방법은 계산시간은 빠르지만 찾은 해의 품질이 그리 좋지는 않았다. 시뮬레이티드 어니일링 방법은 아주 좋은해를 찾지만 총 계산 시간이 아주 많이 소요되었고, 결과비교를 해석해 보면 데이터마이닝 방법에 비해 총 계산시간, 특히 목적함수 계산 횟수가 많은 것으로 미루어 볼 때, 가능해 데이터 셋의 데이터 개수가 많거나 목적함수 계산시간이 길 경우 특히 효율이 떨어질 것으로 추론 된다.

<표 1> 결과비교

방 법	목적함수 $S(\Theta)$	시간(초)	목적함수 계산 횟수
Simplex Search	6.825	73.8	3,200
Simulated Annealing (k_b=0.9)	3.831	542.8	28,503
Simulated Annealing (k_b=0.95)	3.979	259.5	13,606
데이터마이닝 방법	3.894	54.3	283

5. 결 론

본 연구에서는 데이터마이닝 방법을 응용한 휴리스틱 알고리즘의 개발을 수행하였다. 개발된 방법과 기존의 알려진 가능한 방법들과 비교의 결과 비록 가장 좋은 목적함수의 값은 시뮬레이티드 어니일링 방법에서 찾을 수 있었지만, 데이터마이닝 방법이 시간대비 가장 효율적인 결과를 나타내었다. 특히 데이터마이닝 방법의 경우 목적함수의 계산횟수가 시뮬레이티드 어니일링 방법의 가장 좋은해를 찾는 경우의 1/100 정도밖에 되지 않으므로 목적함수의 계산에 많은 시간이 걸리는 문제에 연구된 방법이 더욱 좋은 효율을 나타낼 수 있다는 결론을 얻을 수 있다.

개발된 방법은 아직까지 많은 부분의 향후 연구 과제를 안고 있다. 우선 특성함수의 정의에 보다 일반적인 접근 방법론의 연구가 필요하다. 그리고 k와 n을 정하는 좀더 이론적인 연구가 필요하다. 이밖에, 개발된 방법이 예제로 사용된 문제 이외에 좀 더 일반적인 문제에 충분히 적용 가능한 알고리즘임을 증명하는 연구가 필요하다. 또한 대표 데이터 셋을 추출하는 균일화 알고리즘과 시뮬레이티드 어니일링 방법 등의 메타 휴리스틱 방법을 함께 사용한 알고리즘을 개발하여 데이터마이닝 방법과 비교해보는 연구도 흥미 있는 향후 과제라고 할 수 있다.

참고문헌

[1] Bertsimas, D., and Tsitsiklis, J.; "Simulated Annealing," Statistical Science, 8 : 10-15, 1993.

[2] Fang, K.T., Lin, D.K.J., Winkle, P., and Zhang, Y.; "Uniform design : theory and application," Technometrics, 42 : 237-248, 2000.

[3] Hastie, T., Tibshirani, R., and Friedman, J.; *The element of statistical learning*, Springer-Verlag, 2001.

[4] Igusa, T., Liu, H., Schafer, B., and Naiman, D. Q.; "Bayesian classification trees and clustering for rapid generation and selection of design alternatives," Proceedings of 2003 NSF DMII Grantees and Research Conference, Birmingham AL, 2003.

[5] Jin, J., and Shi, J.; "State Space Modeling of Sheet Metal Assembly for Dimensional Control," ASME Journal of Manufacturing Science & Engineering, 12 1 : 756-762, 1999.

[6] Kim, P., and Ding, Y.; "Optimal design of fixture layout in multi-station assembly process," IEEE Transactions on Automation Science and Engineering, 1(2) : 133-145, 2004.

[7] McKay, M.D., Bechman, R.J., and Conover, W.J.; "A comparison of three methods for selecting values of input variables in the analysis of output from a computer code." Technometrics, 21 : 239-245, 1979.

[8] Nelder, J.A., and Mead, R.; "A simplex method for function minimization," The Computer Journal, 7 : 308-313, 1965.

[9] Santner, T.J., Williams, B.J., and Notz, W.I.; *Design*

& Analysis of Computer Experiments, Springer Verlag, New York, 2003.

[10] Schwabacher, M., Ellman, T., and Hirsh, H.; "Learning to set up numerical optimizations of engineering designs," *Data Mining for Design and Manufacturing*, Kluwer Academic Publishers, Boston, ed. D. Braha, pp. 87-125, 2001.