Journal of Korean Society of Industrial and Systems Engineering
https://doi.org/10.11627/jksie.2022.45.4.233

Vol. 45, No. 4 : 233-239, December 2022
ISSN : 2005-0461 / EISSN : 2287-7975

Simulated Annealing for Overcoming Data Imbalance in Mold Injection Process

Dongju Lee†

Department of Industrial & Systems Engineering, Kongju National University

사출성형공정에서 데이터의 불균형 해소를 위한 담금질모사

이 동 주†

공주대학교 산업시스템공학과

The injection molding process is a process in which thermoplastic resin is heated and made into a fluid state, injected under pressure into the cavity of a mold, and then cooled in the mold to produce a product identical to the shape of the cavity of the mold. It is a process that enables mass production and complex shapes, and various factors such as resin temperature, mold temperature, injection speed, and pressure affect product quality. In the data collected at the manufacturing site, there is a lot of data related to good products, but there is little data related to defective products, resulting in serious data imbalance. In order to efficiently solve this data imbalance, undersampling, oversampling, and composite sampling are usally applied. In this study, oversampling techniques such as random oversampling (ROS), minority class oversampling (SMOTE), ADASYN(Adaptive Synthetic Sampling), etc., which amplify data of the minority class by the majority class, and complex sampling using both undersampling and oversampling, are applied. For composite sampling, SMOTE+ENN and SMOTE+Tomek were used. Artificial neural network techniques is used to predict product quality. Especially, MLP and RNN are applied as artificial neural network techniques, and optimization of various parameters for MLP and RNN is required. In this study, we proposed an SA technique that optimizes the choice of the sampling method, the ratio of minority classes for sampling method, the batch size and the number of hidden layer units for parameters of MLP and RNN. The existing sampling methods and the proposed SA method were compared using accuracy, precision, recall, and F1 Score to prove the superiority of the proposed method.

Keywords : Data Imbalance, Mold Injection, Data Mining, Simulated Annealing

1. 서 론

사출성형공정은 열가소성 수지를 가열하여 유동 상태로 만들어 금형의 공동부에 가압 주입한 후에 금형 내에서 냉각시키는 공정으로, 금형의 공동모양과 동일한 형상의 제품을 만드는 방법이다. 대부분의 제조현장과 마찬가지로 사출공정은 불량데이터의 확보가 어려워, 데이터 불균형이 심하다.

데이터 불균형이란 한 클래스에 속하는 데이터의 수가 다른 클래스에 속하는 데이터의 수보다 매우 많거나 작은 경우를 의미한다. 의사결정모형이나 신경망 모형은 훈련 데이터 등이 클래스 간에 균일하게 분포한다고 가정하기에[2], 데이터 불균형 문제는 분류 예측을 할 때 문제가 될 수 있다. 클래스(Class)란 학습할 데이터에서 데이터를 분류하는 기준이다. 데이터의 수가 많은 클래스를 다수 클래스(Major Class)라 하고, 데이터의 수가 적은 클래스를

Received 3 December 2022; Finally Revised 19 December 2022; Accepted 20 December 2022
† Corresponding Author : djlee@kongju.ac.kr

소수 클래스 (Minor Class)라 칭한다. 데이터불균형 문제는 텍스트 분류, 지진이나 폭발, 희귀병 탐지, 침입탐지시스템[3]에서 자주 발생한다.

데이터 불균형으로 일어나는 오류를 해결하기 위한 방법 중 데이터를 균형 있게 맞추어 학습시키는 샘플링 방법이 있다. 즉, 소수 클래스의 데이터를 증가시켜 다수 클래스의 데이터 수에 가깝도록 하는 오버샘플링(Oversampling) 방법이 있고, 반대로 다수 클래스의 데이터를 소수 클래스의 데이터 수에 가깝도록 감소시키는 언더샘플링(Undersampling)방법이 있다. 또한, 오버샘플링과 언더샘플링을 모두 적용시키는 복합샘플링 방법도 있다.

Lee et al.[10]은 다수클래스의 모집단 분포를 잘 추출하도록 검증하고, 다수 클래스의 데이터를 학습을 통해 제거하는 언더샘플링 기법을 제안하였다.

Jung et al.[6]은 불균형 데이터를 이용한 분류 예측의 정확도를 향상시키기 위해 오버샘플링, 군집분석, 부스팅을 이용한 방법을 제안하였다.

Lee and Kwon[9]은 불균형 데이터를 이용한 분류예측에서민감도는 유지하면서 특이도를 향상시키기 위해 Support Vector Machine, 인공신경망, 의사결정나무 기법으로 구성된 하이브리드 모델을 제안하였다.

Son et al.[13]은 조건부 적대적 생성 신경망(CGAN, Conditional Generative Adversarial Networks)을 이용하여 데이터의 특징을 학습하여 실제 데이터와 유사한 데이터를 생성하여 데이터 수의 균형을 맞추는 기법을 제안하였다.

제조공정에서의 데이터불균형 문제를 해결하기 위한 연구들이 행해졌다. Kim and Lee[7]은 생성적 적대 신경망(GAN)을 이용하여 실제와 가까운 데이터 생성을 위해 유도항을 추가하여 가상데이터를 합성하는 방법을 제안하고, 강판 품질 데이터에 적용하였다.

Jang et al.[5]은 데이터 불균형이 존재하는 문제에서 소수 클래스의 F1 Score를 최소화하는 불균형 문제 해소 기법들의 매개변수들에 대한 최적값을 구하기 위해 유전자알고리즘을 이용하였다. Shin et al.[12]은 불균형 문제가 심한 신용카드 사기 탐지 문제에 F1 Score를 최소화하도록 오버샘플링 기법들의 최적 비율을 구하는 유전자알고리즘을 제안하였다.

데이터 불균형을 처리하기 위해 본 연구에서는 오버샘플링과 복합샘플링 기법을 적용하였다. 대표적인 오버샘플링 기법으로는 SMOTE[1]와 Borderline-SMOTE[4] 등이 있고, 복합샘플링으로는 SMOTE+ENN, SMOTE+Tomek이 있다.

기존의 연구들은 새로운 샘플링방법을 개발하거나, 데이터 불균형이 심한 문제들을 다양한 샘플링 방법을 적용하여 해당 문제들에 적합한 샘플링방법을 탐색하였다. 하지만, 샘플링방법들, 샘플링방법의 매개변수, 인공신경망 방법들의 매개변수간에 최적 조합이 있을 수 있는데, 이들 모두를 고려한 최적화기법에 대한 연구가 없다. 그러므로, 본 연구에서는 샘플링 방법, 샘플링기법별 매개변수, 인공신경망(ANN, Artificial Neural Network)의 매개변수를 동시에 최적화하는 담금질모사(SA, Simulated Annealing) 기법을 제안하고, 데이터 불균형이 심한 사출성형 데이터에 적용하였다. 인공신경망으로는 다층퍼셉트론(MLP, Multilayer Perceptron)과 순환신경망(RNN, Recurrent Neural Network) 기법인 장단기메모리(LSTM, Long Short Term Memory)를 적용하였다.

본 논문의 구성은 다음과 같다. 이어지는 2장에서는 오버샘플링과 복합샘플링 기법에 대해 살펴본다. 3장에서는 SA에 기반한 해법을 제안한다. 4장에서는 사출성형 데이터에 제안한 해법과 샘플링 기법들을 적용하고 장단점에 대해 살펴본다. 마지막으로, 5장에서는 결론과 미래연구방향에 대해 논하고자 한다.

2. 오버샘플링, 복합샘플링 기법

이번 장에서는 4가지의 오버샘플링 기법과 2가지의 복합샘플링 방법에 대해 살펴보고자 한다. 오버샘플링 기법으로는 ROS(Random Over Sampling), SMOTE(Synthetic Minority Over-Sampling Technique), Borderline‐SMOTE, ADASYN(Adaptive Synthetic Sampling)이 있다.

ROS는 기존에 존재하는 소수 클래스(Minority)를 단순 복제하여 비율을 맞춰주는 방법으로 단순히 복제하기에 분포는 변화하지 않지만 소수의 클래스의 데이터 수가 증가하기에 더 많은 가중치를 받게 된다.

SMOTE는 KNN (K-Nearest Neighbor)에 기반하며, 가장 많이 쓰이는 방법이다. 임의의 소수 클래스에 해당하는 관측치 X와 X에 가장 가까운 K개의 이웃 관측치 X(nn)를 탐색한다. X와 K개의X(nn)중 임의의 1개의 관측치 사이에 새로운 데이터 X'를 생성한다.

$$X' = X + U(X(nn) - X)$$

여기서 U는 Uniform Distribution를 의미하며, (0,1)사이의 임의의 값을 생성한다.

Borderline‐SMOTE는 다수 클래스(Majority)와 소수 클래스(Minority)를 나누는 경계선(Borderline)이 다수 클래스와 소수 클래스를 구분하는데 중요하므로, 경계선에 있는 소수 클래스의 데이터에 SMOTE를 적용하는 방법이다. 즉, 임의의 소수 클래스 X를 정하고, X에 가장 근접한 K개의 데이터(소수 클래스와 다수 클래스가 섞여 있다.)를 찾는다. 이들 K개의 데이터 중에서 소수 클래스에 속한

개수와 다수 클래스에 속한 개수에 따라 Noise, Safe, Danger로 나누고 경계에 있다고 판단되는 Danger에만 SMOTE를 적용하여 소수 클래스의 수를 증가시킨다. K'을 다수 클래스에 속한 데이터의 수라고 할 때 Noise, Safe, Danger를 나누는 기준과 그에 대한 설명은 다음과 같다.

- Noise: $K = K'$일 때. X에 가장 근접한 K개의 데이터가 모두 다수 클래스에 속할 때, X는 잘못된 데이터(Noise)로 판단하고, SMOTE를 적용하지 않는다.
- Safe: $0 \le K' \le K/2$일 때. 다수 클래스에 속한 데이터 수가 절반 이하, 즉, 절반 이상의 데이터가 소수 클래스에 속하므로 안전하다고 판단하고 SMOTE를 적용하지 않는다.
- Danger: $K/2 < K' < K$일 때. 절반이상이 다수 클래스에 속하므로 위험하다고 판단하고, SMOTE를 적용한다.

마지막으로, ADASYN은 각 소수 클래스 주변의 다수 클래스 관측치의 비율(Ratio, r_i)을 이용해 SMOTE를 적용시키는 방법. 총 m개의 소수 클래스에 속한 데이터가 있다고 할 때, 소수 클래스의 데이터에 대해 r_i를 구하면,

$$r_i = \frac{\triangle_i}{K},\ i = 1, ..., m$$

이다. 여기서, $\triangle_i$는 소수 클래스에 속한 데이터 x_i에 가장 가까운 K개의 데이터 중 다수 클래스에 속하는 데이터의 수이다.

소수 클래스의 모든 데이터들에 대해 K개의 이웃 관측치를 탐색하고, r_i를 구한다. 계산한 r_i를 정규화한 후, 각 소수 클래스의 데이터에 대해 생성하고자 하는 개수(다수 클래스 개수 - 소수 클래스 개수)를 계산하여, 계산된 수만큼 각각 데이터를 생성한다.

복합샘플링으로는 SMOTE+ENN과 SMOTE+Tomek이 있다.

SMOTE+ENN은 SMOTE와 ENN(Edited Nearest Neighbor)을 혼용하여 오버샘플링과 언더샘플링을하는 기법이다. ENN은 다수 클래스의 데이터가 모두 혹은 대부분 다수클래스가 아니면 이들 다수 데이터를 삭제하여 소수 클래스 주변의 다수 데이터를 삭제하는 언더샘플링 기법이다.

SMOTE+Tomek은 SMOTE와 Tomek을 혼용하여 오버샘플링과 언더샘플링을 하는 기법이다. Tomek Link는 클래스가 다른 두 데이터가 가까이 붙어 있고, 그 사이에는 다른 데이터가 없는 경우를 의미한다. 이러한 Tomek Link를 찾고, 다수 클래스의 데이터를 제거하는 언더샘플링 방법이다.

3. 제안하는 방법

기계학습(Machine Learning)알고리즘을 적용하여 분류(Classification)를 한 경우, 얼마나 잘 분류되었는지 확인하는 척도로는 Accuracy(정확도), Precision(정밀도), Recall(재현율), F1 Score가 있다. 먼저 혼동행렬(Confusion Matrix)는 <Table 1>에 주어져 있다.

<Table 1> Confusion Matrix

		Predictive Values	
		Positive	Negative
Actual Values	Positive	TP (True Positive)	FN (False Negative)
	Negative	FP (False Positive)	TN (True Negative)

Accuracy, Precision, Recall, F1 Score는 아래와 같다.

$$\text{Accuracy} = \frac{TP+FN}{TP+TN+FP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F1 Score} = 2 \times \frac{Recall \times Precision}{Recall + Precision}$$

F1 Score는 정밀도와 재현율의 조화평균으로 정밀도와 재현율을 동시에 고려한 지표로 데이터불균형이 존재하는 경우 자주 쓰이는 지표이다. 본 연구에서는 소수클래스를 중심으로 한 F1 Score를 계산하고, 이를 최대화하고자 한다.

인공신경망 중 MLP와 RNN을 분류를 위한 알고리즘으로 선택하였다. 단층 퍼셉트론은 XOR과 같은 비선형적으로 분리가 되는 데이터는 학습이 불가능하기에, 고안된 방법이 MLP(다층퍼셉트론)이다. MLP는 입력층, 은닉층, 출력층으로 구성되어 있는데, 은닉층이 여러 개인 인공신경망을 심층신경망(Deep Neural Network)이라고 한다. 또한, RNN은 자기 계층의 출력정보를 입력신호로 다시 사용하여 계층의 출력이 순환하는 구조를 지닌다. 하지만, RNN은 역전파 중에 기울기 값이 줄어드는 기울기 소멸 문제가 생기는데, 이를 보완한 것이 LSTM이다. LSTM은 RNN계열로 RNN의 숨겨진 상태에 셀 상태를 추가한 것이다.

Kirkpatrick et al.[8]이 제안한 담금질모사(SA, Simulated Annealing)기법은 지역최적점(Local Optimum)이 다수 존재하는 최적화문제에서 전역최적점(Global Optimum)에

대한 근사치를 구하기 위해 적용하는 메타휴리스틱 기법이다. 본 연구에서 사용한 SA 알고리즘은 다음과 같다.

Step 1: 초기조건 설정. 무작위로 초기해 X를 구하고, 초기온도(T_{max})를 설정한다.

Step 2: 이웃해 생성. 현재해(X)에서 변화를 주어 이웃해(X')를 생성한다. 4종류(샘플링 방법, 다수클래스와 소수클래스의 비율, 배치크기, 은닉층의 유닛 수)의 변화를 줄 수 있는데 랜덤하게 4종류 중 하나의 방법으로 이웃해를 생성한다. 4종류 변화에서 고려되는 값들은 <Table 4>에 주어져 있다.

Step 3: 목적식값의 차이 계산. 현재해와 이웃해의 목적식값의 차이(Δ)를 구한다.

$$\Delta = f(X') - f(X)$$

목적식값으로는 F1 Score가 사용되었다. Step 2에서 생성한 이웃해에 대해 샘플링을 수행하고, 이를 인공신경망(MLP 혹은 RNN)에 적용하여 이웃해의 F1 Score를 계산한다.

Setp 4: Metropolis 기준을 적용한 수락여부 판정. 목적식이 F1 Score의 최대화이므로 이웃해가 현재해보다 F1 Score가 크면 현재해를 이웃해로 변경하고(수락), 이웃해의 F1 Score가 작더라도 식 (1)을 만족하면 현재해를 이웃해로 변경한다(수락). 아니면, 현재해를 이웃해로 변경하지 않는다(거절). 수락의 경우 Step 2로 이동하고, 거절의 경우 Step 5로 이동한다.

$$e^{\left(\frac{-\Delta}{T}\right)} \geq U(0,1) \quad (1)$$

여기서 U(0,1)는 Uniform Distribution을 따르는 (0,1) 사이의 임의의 값을 생성한다.

Step 5: 냉각스케줄(Cooling Schedule) 갱신. 식 (2)를 이용하여 온도를 계산한다.

$$T_k = \alpha \times T_{k-1} \quad (2)$$

여기서 k는 반복횟수를 의미한다. T_k는 반복횟수 k에서의 온도를 의미하며, α는 0.5~0.99 사이의 값으로 정해지며, 냉각속도를 조절하는 역할을 한다.

Step 6: 종료조건. 미리 정해둔 최대 반복회수 (Maximum Number of Iteration)에 도달하거나, 미리 정해둔 최저온도(T_{min})에 도달하면 종료한다. 아니면, Step 2로 돌아간다.

4. 실 험

분석에 사용한 데이터는 KAMP의 사출성형기 AI 데이터 셋[11]을 활용하였는데 자동차 앞 유리 사이드 몰딩 사출 공정 데이터이다. 사용된 독립변수는 온도(스크류/실린더, 수지, 금형, 건조, 유압, 주변환경), 압력(충진, 보압, 배압, 이형, 형개, 형체), 시간(충진, 보압, 냉각, 건조), 속도(사출, 스크류 회전, 형개, 이형(이젝팅)), 양(계량, 이형량, 쿠션량) 관련 24개이다. 종속변수로는 불량여부 (0: 양품, 1: 불량품)가 사용되었다.

제품명은 Cn7으로, 양품 관측치 수 1194개, 불량품 관측치 수는 17개로 데이터 불균형이 심하다. 분류를 위한 인공신경망 기법으로 MLP와 RNN이 사용되었는데, 이들의 매개변수는 <Table 2>, <Table 3>과 같다.

<Table 2> Parameters for MLP

Parameter	Value
Activation Function	ReLU
Optimizer	Adam
Loss Function	Binary Crossentropy
Learning Rate	0.001
Epochs for Training	100

<Table 3> Parameters for RNN

Parameter	Value
Activation Function	ReLU
Optimizer	Adam
Loss Function	Binary Crossentropy
Learning Rate	0.001
Epochs for Training	70

인공신경망은 파이썬으로 코딩되었는데, Keras Package의 Sequential 모델을 이용해 구축하였다.

MLP는 4개의 은닉층이 있는데, 활성화함수로는 은닉층에는ReLU를, 마지막 신경망은 Sigmoid 함수를 사용하였다. 과대적합을 방지하기 위해 완전연결층 사이에는 Drop Out=0.3을 적용하였다.

RNN은 2개의 은닉층이 있는데, 활성화함수로 은닉층에는ReLU를, 출력층은 Sigmoid 함수를 사용하였다.

SA을 이용하여 구하고자 하는 최적 조합은 샘플링 방법, 샘플링 시 매개변수인 소수클래스의 비율, 인공신경망의 매개변수인 배치크기와 은닉층의 유닛 수이다. 이들에 대한 값들과 범위는 <Table 4>에 주어져 있다.

MLP의 경우에는 6×5×11×17=6732개의 가능한 조합이 있으며, RNN의 경우에는 6×5×11×6=2376개의 조합이 가능하다.

<Table 4> Considering Ranges for Combinations

Considering Combinations	Range
Sampling Method	6 Methods
Minority Ratio	(0.1, 0.2, 0.3, 0,4, 0.5)
Batch Size	(20, 30,⋯,120)
# of Unis in Hidden Layer for MLP (nh)	(10, 12, 14, ⋯, 42)
# of Units in Hidden Layer for RNN	(2^3, 2^4, ..., 2^8)

은닉층 별로 유닛 수를 다르게 하는 경우 더욱 다양한 조합이 가능하나, 1개의 변수 nh를 이용하여 은닉층의 유닛 수를 조절하였다. 즉, MLP는 4개의 은닉층으로 구성되어 있다. 1번째 은닉층의 유닛의 개수를 nh라고 할 때, 각 은닉층의 유닛의 개수는 다음과 같이 결정하였다.

1번째 은닉층: nh
2번째 은닉층: 2 × nh
3번째 은닉층: nh
4번째 은닉층: nh / 2

RNN의 은닉층은 1개의 LSTM 층과 1개의 완전연결층(Fully Connected Layer)을 사용하였는데, 각 은닉층의 유닛의 개수는 다음과 같이 결정하였다.

LSTM: 2 × nh
완전연결층: nh

다양한 조합들 중 소수 클래스의 F1 Score를 최대화하도록 SA를 적용하였다. SA에 사용된 매개변수들은 <Table 5>에 주어져 있다.

<Table 5> Parameters for Simulated Annealing

Parameter	Value
Maximum # of Iteration	200
α	0.95
T_{max}	100
T_{min}	0.01

70%의 데이터는 Training용으로, 나머지 30%의 데이터는 Test용으로 사용되었는데, 양품, 불량품 각각에 대해 해당 비율로 랜덤하게 선택되었다.

각 샘플링 방법 별 혼동행렬은 <Table 6>에 주어져 있다. 샘플링을 적용하지 않은 원데이터(Original)와 6개의 샘플링 방법은 소수클래스의 비율은 10%로 하였으며, MLP는 에포크(Epochs)=200, 배치 크기=30을 적용하였다. RNN은 에포크(Epochs)=70, 배치 크기=30을 적용하였다. 원데이터와 6개의 샘플링 방법 별 결과들은 5번의 시행 중 최고의 F1 Score를 주는 경우의 혼동행렬이다.

제안하는 SA기법은 1번의 시행의 결과로, 이때의 혼동행렬은 제일 마지막 행에 주어져 있다.

MLP와 RNN을 적용한 경우의Accuracy, Precision, Recall, F1 Score는 <Table 7>과 <Table 8>에 각각 주어져 있다.

제안하는 SA기법에서 <Table 6>, <Table 7>, <Table 8>의 결과를 보여준 조합은 <Table 9>에 주어져 있다.

<Table 6> Confusion Matrix for MLP and RNN by Sampling Method

Sampling Method		MLP			RNN		
			Predictive Values			Predictive Values	
			Positive	Negative		Positive	Negative
Original	Actual Values	Positive	359	0	Positive	346	13
		Negative	5	0	Negative	5	0
ROS	Actual Values	Positive	357	2	Positive	357	2
		Negative	3	2	Negative	3	2
SMOTE	Actual Values	Positive	359	0	Positive	357	2
		Negative	3	2	Negative	3	2
Borderline-SMOTE	Actual Values	Positive	350	9	Positive	359	0
		Negative	2	3	Negative	3	2
ADASYN	Actual Values	Positive	355	4	Positive	358	1
		Negative	3	2	Negative	3	2
SMOTE +ENN	Actual Values	Positive	355	4	Positive	357	2
		Negative	3	2	Negative	3	2
SMOTE +Tomek	Actual Values	Positive	350	9	Positive	358	1
		Negative	3	2	Negative	3	2
Proposed SA	Actual Values	Positive	359	0	Positive	359	0
		Negative	3	2	Negative	3	2

<Table 7> Accuracy, Precision, Recall, and F1 Score for MLP

	Accuracy	Precision	Recall	F1 Score
Original	0.951	0.000	0.000	0
ROS	0.986	0.500	0.400	0.444
SMOTE	0.992	1.000	0.400	0.571
Borderline-SMOTE	0.970	0.250	0.600	0.353
ADASYN	0.981	0.333	0.400	0.364
SMOTE+ENN	0.981	0.333	0.400	0.364
SMOTE+Tomek	0.967	0.182	0.400	0.250
Proposed SA	0.992	1.000	0.400	0.571

<Table 8> Accuracy, Precision, Recall, and F1 Score for RNN

	Accuracy	Precision	Recall	F1 Score
Original	0.951	0.000	0.000	0
ROS	0.986	0.500	0.400	0.444
SMOTE	0.986	0.500	0.400	0.444
Borderline-SMOTE	0.992	1.000	0.400	0.571
ADASYN	0.989	0.667	0.400	0.500
SMOTE+ENN	0.986	0.500	0.400	0.444
SMOTE+Tomek	0.989	0.667	0.400	0.500
Proposed SA	0.992	1.000	0.400	0.571

<Table 9> Best Combinations for MLP and RNN

	MLP	RNN
Sampling Method	SMOTE+Tomek	ADASYN
Minority Ratio	0.1	0.4
Batch Size	70	30
# of Units in Hidden Layer	38	128

실험결과를 요약하면 다음과 같다.

- 원데이터는 MLP, RNN 모두에서 실제불량을 전혀 예측하지 못했다. 이는 샘플링방법을 통하여 데이터 불균형 문제를 어느 정도 해결할 필요가 있음을 의미한다.
- 샘플링 방법과 제안하는 SA 기법들은 일부의 실제불량을 예측할 수 있었다. 하지만, 불량갯수 5개 중 최대 2개의 실제불량만 불량으로 예측할 수 있었다.
- 샘플링 기법들 중 다른 기법들보다 항상 우수한 기법은 없어 우열을 논할 수 없었다. 다만, F1 Score로 볼 때, MLP에서는 SMOTE가 가장 좋은 결과인 0.571이었고, RNN에서는 Borderline-SMOTE가 가장 좋은 결과인 0.571이었다.
- 제안하는 SA는 MLP, RNN 모두에서 F1 Score=0.571로 가장 좋은 결과를 보여 주었다. 또한, Accuracy = 0.992, Precision = 1.000, Recall = 0.400으로 가장 좋은 결과를 보여주었다.
- F1 Score를 포함한 지표들로 볼 때, 언더샘플링과 오버샘플링을 모두 사용하는 복합샘플링이 오버샘플링보다 우수한 결과를 보여준다고 할 수 없으며, 우열을 논할 수 없었다.

5. 결론 및 연구과제

사출공정은 대부분의 제조현장에서 그러하듯 불량데이터의 확보가 어렵기에, 양품데이터의 수가 다수를 차지하는 데이터 불균형이 심하다.

한 클래스에 속하는 데이터의 수가 다른 클래스에 속하는 데이터의 수보다 매우 많거나 적은 데이터 불균형이 존재하는 경우, 의사결정모형이나 신경망 모형을 이용하여 분류 예측을 할 때 문제가 될 수 있다.

데이터 불균형으로 일어나는 오류를 해결하기 위해 소수 클래스의 데이터를 증가시키는 오버샘플링 방법과 다수 클래스의 데이터를 소수 클래스의 데이터 수에 가깝도록 감소시키는 언더샘플링 방법, 오버샘플링과 언더샘플링을 모두 적용시키는 복합샘플링 방법이 있다. 본 연구에서는 오버샘플링방법과 복합샘플링 방법이 고려되었으며, 분류 예측을 위해서 인공신경망 기법인 MLP와 RNN이 사용되었다.

다수의 샘플링 방법, 샘플링 방법의 매개변수, 인공신경

망 기법들의 매개변수들의 조합을 최적화하는 SA기법을 제안하고, 데이터불균형이 심한 사출공정 데이터에 적용하였다.

제안하는 SA기법을 원데이터, 샘플링기법등과Accuracy, Precision, Recall, F1 Score로 비교해 볼 때 우수한 결과를 보여주었다.

미래의 연구과제로는 좀 더 다양한 데이터 셋에 적용하여 결과를 비교해 볼 필요가 있다. 또한, 인공신경망 기법들의 매개변수들을 좀 더 다양하게 변경하여 결과를 확인하고, 장단점을 논할 필요가 있다.

References

[1] Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P., SMOTE: Synthetic Minority Over-sampling Technique, *Journal of Artificial Intelligence Research*, 2002, Vol. 16, pp. 321-357.

[2] Chen, Z., Yan, Q., Han, H., Wang, S., Peng, L., Wang, L., and Yang, B., Machine learning based mobile malware detection using highly imbalanced network traffic, *Information Science*, 2018, Vol. 433, pp. 346-364.

[3] Cheong, Y.-G., Park, K., Kim, H., Kim, J., and Huun, S., Machine Learning Based Intrusion Detection Systems for Class Imbalanced Datasets, *Journal of The Korea Institute of Information Security & Cryptology*, 2017, Vol. 27, No. 6, pp. 1385-1395.

[4] Han, H., Wang, W.-Y., and Mao B.-H., Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning, *Proceedings of ICIC 2005: Advances in Intelligent Computing*, 2005, pp. 878-887.

[5] Jang, Y.S., Kim, J.W., and Hur, J., Combined Application of Data Imbalance Reduction Technique Using Genetic Algorithm, *Journal of Intelligence and Information Systems*, 2008, Vol. 14, No. 3, pp. 133-154.

[6] Jung, H.N., Lee, J.-H., and Jun, C.-H., A Data Mining Procedure for Unbalanced Binary Classification, *Journal of the Korean Institute of Industrial Engineers*, 2010, Vol. 36, No. 1, pp. 13-21.

[7] Kim, H.S. and Lee, H.S., Generative Adversarial Networks based Data Generation Framework for Overcoming Imbalanced Manufacturing Process Data, *J. of Korean Ins. of Intell. Syst.*, 2019, Vol. 29, No. 1, pp. 1-8.

[8] Kirkpatrick, S., Gelatt, C.D., and Vecchi, M.P., Optimization by Simulated Annealing, *Science*, 1983, Vol. 220, No. 4598, pp. 671-680.

[9] Lee, J.S. and Kwon, J.G., A Hybrid SVM Classifier for Imbalanced Data Sets, *J. Intell. Inform. Syst.*, 2013, Vol. 19, No. 2, pp. 125-140.

[10] Lee, K.N., Lim, J., Bok, K., and Yoo, J., Handling Method of Imbalance Data for Machine Learning : Focused on Sampling, *The Journal of the Korea Contents Association*, 2019, Vol. 19, No. 11, pp. 567-577.

[11] Ministry of SMEs and Startups, Korea AI Manufacturing Platform(KAMP), CNC Machine AI Dataset, KAIST(UNIST, EPM Solutions), 2020.12.14., https://kamp-ai.kr.

[12] Shin, S.S., Cho, H.Y., and Kim, Y.H., Optimal Ratio of Data Oversampling Based on a Genetic Algorithm for Overcoming Data Imbalance, *J. of the Korea Convergence Society*, 2021, Vol. 12, No. 1, pp. 49-55.

[13] Son, M.J., Jung, S.W., and Hwang, E.J., A Deep Learning Based Over-Sampling Scheme for Imbalanced Data Classification, *KIPS Trans. Softw. And Data Eng.*, 2019, Vol. 8, No. 7, pp. 311-316.

ORCID

Dongju Lee | http://orcid.org/0000-0001-6650-9270