ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2020 December: 26(4): 173~198 http://dx.doi.org/10.13088/jiis.2020.26.4.173

캠페인 효과 제고를 위한 자기 최적화 변수 선택 알고리즘*

서정수
케이티디에스㈜ / 국민대학교 비즈니스IT전문대학원
(se0007@kt.com)

안현철
국민대학교 비즈니스IT전문대학원
(hcahn@kookmin.ac.kr)

최근 온라인의 비약적인 활성화로 캠페인 채널들이 다양하게 확대되면서 과거와는 비교할 수 없을 수준의 다양한 유형들의 캠페인들이 기업에서 수행되고 있다. 하지만, 고객의 입장에서는 중복 노출로 인한 캠페인에 대한 피로감이 커지면서 스팸으로 인식하는 경향이 있고, 기업입장에서도 캠페인에 투자하는 비용은 점점 더 늘어났지만 실제 캠페인 성공률은 오히려 더 낮아지고 있는 등 캠페인 자체의 효용성이 낮아지고 있다는 문제점이 있어 실무적으로 캠페인의 효과를 높이고자 하는 다양한 연구들이 지속되고 있다. 특히 최근에는 기계학습을 이용하여 캠페인의 반응과 관련된 다양한 예측을 해보려는 시도들이 진행되고 있는데, 이 때 캠페인 데이터의 다양한 특징들로 인해 적절한 특징을 선별하는 것은 매우 중요하다. 전통적인 특징 선택 기법으로 탐욕 알고리즘(Greedy Algorithm) 중 SFS(Sequential Forward Selection), SBS(Sequential Backward Selection), SFFS(Sequential Floating Forward Selection) 등이 많이 사용되었지만 최적 특징만을 학습하는 모델을 생성하기 때문에 과적합의 위험이 크고, 특징이 많은 경우 분류 예측 성능 하락 및 학습시간이 많이 소요된다는 한계점이 있다. 이에 본 연구에서는 기존의 캠페인에서의 효과성 제고를 위해 개선된 방식의 특징 선택 알고리즘을 제안한다. 본 연구의 목적은 캠페인 시스템에서 처리해야 하는 데이터의 통계학적 특성을 이용하여 기계 학습 모델 성능 향상의 기반이 되는 특징 부분 집합을 탐색하는 과정에서 기존의 SFFS의 순차방식을 개선하는 것이다. 구체적으로 특징들의 데이터 변형을 통해 성능에 영향을 많이 끼치는 특징들을 먼저 도출하고 부정적인 영향을 미치는 특징들은 제거를 한 후 순차방식을 적용하여 탐색 성능에 대한 효율을 높이고 일반화된 예측이 가능하도록 개선된 알고리즘을 적용하였다. 실제 캠페인 데이터를 이용해 성능을 검증한 결과, 전통적인 탐욕알고리즘은 물론 유전자알고리즘(GA, Genetic Algorithm), RFE (Recursive Feature Elimination) 같은 기존 모형들 보다 제안된 모형이 보다 우수한 탐색 성능과 예측 성능을 보임을 확인할 수 있었다. 또한 제안 특징 선택 알고리즘은 도출된 특징들의 중요도를 제공하여 예측 결과의 분석 및 해석에도 도움을 줄 수 있다. 이를 통해 캠페인 유형별로 중요 특징에 대한 분석과 이해가 가능할 것으로 기대된다.

주제어 : 특징 선택, AI기반 캠페인 시스템, 탐욕 알고리즘, 캠페인 수행 결과 예측, 머신 러닝

논문접수일 : 2020년 11월 19일 논문수정일 : 2020년 12월 26일 게재확정일 : 2020년 12월 28일
원고유형 : 학술대회 Fast-track 교신저자 : 안현철

1. 서론

캠페인 데이터는 다양한 고객 속성 특징들을 가진 대용량 데이터이다. 최근 온라인의 비약적인 활성화로 캠페인 채널들이 다양하게 확대되면서 과거와는 비교할 수 없을 수준의 다양한 유형들의 캠페인들이 기업에서 수행이 되고 있다. 일반적인 고객의 획득과 정보들을 획득하기 위

* 본 연구는 2020 한국지능정보시스템학회 추계학술대회에서 발표되었던 연구를 학술지 논문으로 발전시킨 것입니다. 논문의 개선에 큰 도움을 주신 두 익명의 심사위원께 감사의 말씀을 드립니다.

한 획득 캠페인(acquisition campaign)에서부터 관계 강화(cultivation)를 위한 캠페인들, 고객 유지와 고객 돌봄(care)을 위한 수많은 유형들의 캠페인들이 수행되고 있다. 문제는 이러한 캠페인들이 다양해지고 채널이 많아지면 많아질수록 고객의 입장에서는 중복 노출로 인해 캠페인에 대한 피로감이 커지면서 스팸으로 인식하는 경향이 있으며 기업입장에서는 캠페인에 투자하는 비용은 점점 더 늘어났지만 실제 캠페인 성공률은 오히려 더 낮아지고 있는 등 캠페인 자체의 효용성이 낮아지고 있다는 문제점이 있다. 특히 캠페인 기획 단계에서 캠페인 성격과 목적을 정의하고 그에 맞는 캠페인 대상자 선정을 위해 최소 1~2개월 이상이 소요가 되어도 기획자의 수행 경험 기반으로(성공에 영향을 많이 주는 특징에 대한 경험) 캠페인 대상자를 선정 후 수행하는 경우가 많아 실제 수행 성공률이 떨어지는 등 비효율적이지만 캠페인 비용을 줄이지는 못하는 현실에 직면해 있다. 이에 최근에는 기계학습을 통한 데이터 분석기법 및 딥러닝(deep learning) 등을 활용하여 캠페인에 적용하여 캠페인 성공률을 높이려는 시도들이 일어나고 있다. 기계학습을 통한 데이터 분석 기법은 학습 과정을 통해 도출된 분류 모델을 기반으로 방대한 데이터를 손쉽게 처리할 수 있으므로 캠페인 데이터 분석을 비롯한 다양한 분야에서 사용되고 있다(Oh et al., 2004). 하지만, 데이터의 양이 증가할수록 특징 차원과 학습 데이터의 양이 늘어남에 따라 학습시간 및 분류시간이 비례하여 늘어나는 문제점이 발생하게 된다(Ladha and Deepa, 2011). 대용량의 데이터를 분류하는 과정에서 전체 입력 데이터를 모두 사용한다면 분류 클래스가 확장됨에 따라 학습 시간이 많이 소요된다. 따라서 전체 데이터에서 최소한의 입력데이터 셋을 추출하여 사용하여야 한다(Zhou and Li, 2016). 또한, 특징을 너무 많이 사용하여 학습된 모델을 생성할 경우, 과적합(overfitting) 되거나 특징들 사이의 상관관계로 인해 예측 정확도가 낮아질 수 있다. 따라서, 정확도를 향상하기 위하여 잡음(noise)에 가까운 특징들을 제거하는 특징 선택 기법이 적용되어야 한다(Chandrashekar and Sahin, 2014).

특징선택(feature selection)은 고차원의 데이터 셋을 분석하기 위해서 필요한 과정이다. 분류 분석을 위한 특징 선택은 선택된 특징 조합에 의하여 학습된 분류기의 예측 성능을 최대화하는 조합을 선택하는 문제로 정의할 수 있다(Ohn and Han, 2013). 즉, 특징 선택 과정을 통해 특징 집합의 크기를 줄일 수 있고 이로써 고차원의 특징에서 발생하는 문제(curse of dimension)를 해결할 수 있다. 또한, 분별력 있는 특징 집합을 구성함으로써 분류 성능을 보장할 수 있다. 여러 연구에서 이러한 목적을 달성하기 위해 다양한 특징 선택 방법들을 제안하고 있다(Guyon and Elisseeff, 2003; Hong and Shin, 2003; Kim and Ahn, 2011; Lee and Jeong, 2008). 일반적인 특징 선택 기법은 특징 간의 상관관계를 고려하지 못하기 때문에 선택된 특징이 과적합 된 모델을 생성하게 된다는 단점이 있다. 이러한 모델은 낮은 일반화율 때문에 높은 오탐률을 보이며 이는 실제 기업에서의 캠페인 시스템에는 적용하기 어렵다. 이러한 문제점을 해결하기 위해 본 연구에서는 개선된 방식의 특징 선택 알고리즘을 제안한다. 본 연구의 목적은 캠페인 시스템에서 처리해야 하는 데이터의 통계학적 특성을 이용하여 기계 학습 모델 성능 향상의 기반이 되는 특징 부분 집합을 선택하는 것이다. 제안하는 알고리즘은 특징 부분 집합을 탐색하는 과정에서 개선

된 방식의 SFFS(Sequential Forward Floating Selection)를 사용한다. 기존의 SFFS 기법은 순차적 탐색을 수행하기 때문에 처리해야 하는 특징 평가 작업이 수행되는 동안 다른 작업을 수행하지 못하며 전반적으로 탐색 성능의 효율이 많이 떨어진다. Parallel 형태로 동시 수행을 할 수 있지만(Lee, Park and Lee, 2017) 자원을 많이 소모한다는 한계가 있다. 게다가 목적 함수로 사용되는 특징 선택 기준에 따라 분류 성능의 향상도가 크지 않다는 단점이 존재한다. 이러한 문제점을 해결하기 위하여 특징 탐색을 특징이 분류 성능에 가장 영향을 적게 미치는 것으로 예상되는 경우부터 제거를 하면서 가장 최적의 성능을 내는 특징을 선택할 수 있도록 한다. 또한 랜덤포레스트 분류기를 활용하여 특징 부분 집합을 도출할 수 있도록 한다. 랜덤포레스트는 많은 수의 의사결정트리를 기반으로 분류기를 생성하기 때문에 일반화(generalization)된 분석 모델을 도출할 수 있다. 하지만, 불필요한 특징을 모두 포함하거나 특징 해석이 어렵다는 단점이 있다. 이를 해결하기 위하여 특징 부분 집합을 학습시킬 때 도출된 특징 중요도를 특징 부분집합과 함께 제공함으로써 분석 모델의 결과 해석에 도움이 되도록 한다. 제안하는 SOFS(Self Optimizing Feature Selection) 알고리즘은 특징 선택 기법의 성능상의 단점을 개선하고 일반화의 오류를 보완할 수 있는 형태이다. 이를 통해 최적 특징을 도출할 수 있고, 캠페인 예측 분류 모델의 정확도와 성능향상을 기대할 수 있다.

본 연구의 구성은 다음과 같다. 제2장에서는 기계학습과 특징 선택 기법에 대해 살펴본다. 제3장에서는 특징 부분 집합 최적화를 위한 SOFS 특징 선택 알고리즘에 대하여 설명한다. 제4장에서는 기계 학습 기반 캠페인 시스템에 특징 선택 알고리즘을 적용하고 제5장에서는 특징 선택 알고리즘이 적용된 기계 학습 기반 캠페인 시스템에 대한 비교 모델들과의 우위점에 대하여 기술할 것이다. 마지막으로 제6장에서는 연구의 결론과 함께, 본 연구의 한계점에 대해 논의하고 향후의 방향성에 대해서도 약술할 것이다.

2. 관련 연구

본 연구에서 제안하는 모형은 SOFS(Self-optimizing Feature Selection) 알고리즘 기반의 특징 선택에 대한 개선된 모델이다. 이에 기존 문헌의 검토에서는 우선 기계학습 기법과 특징 선택 기법에 관한 관련 연구를 정리하고 특징 선택을 위한 개선된 기존 연구들에 대해 살펴 보도록 한다.

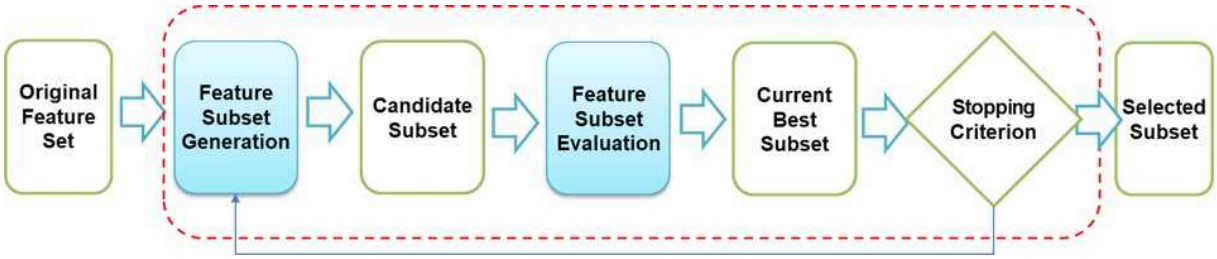
2.1. 기계학습(Machine Learning)

기계 학습(Machine Learning)은 ‘코드로 정의하지 않은 동작을 실행하는 능력’에 대해 연구하는 분야이다(Samuel, 1959). 즉, 기계가 특정 작업에 대해 꾸준한 경험을 통하여 학습하고 그 작업을 얼마나 잘 수행하는지에 대한 실행 성능을 높이는 것을 목적으로 한다. “컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야”를 포괄하여 기계 학습이라 할 수 있다(Mitchell, 1997). 기계학습에서 가장 중요한 것은 경험에 해당하는 데이터이며 좋은 품질의 데이터를 많이 확보할수록 좀 더 나은 성능을 끌어낼 수 있다. 반면에 잡음이 많은 데이터를 사용할 경우 분류 모델의 과적합으로 인해 결과 성능이 저하되는 현상이 발생한다. 잡음(noise)은 분류에 불필요한 정보가 너무 많이 분류기에 입력되는

것을 말한다. 다시 말해 측정된 변수에 무작위의 오류(random error) 또는 분산(variance)가 존재하는 것을 말한다. 잡음이 발생하는 원인은 관측 오류, 시스템에서 발생하는 오차 등으로 파악할 수 있다. 데이터에 잡음이 많아지면 분류 성능이 낮아지고 이를 해결하기 위해서는 많은 시간과 노력이 필요하다. 일반적으로 잡음 데이터의 처리를 위해서는 분류기에 나쁜 영향을 미치는 대상을 파악하여 제거하는 과정을 거친다. 과적합은 데이터의 잡음이 모델에 학습되거나 원하는 결과 예측에 도움이 되지 않는 요소들이 함께 학습되어 발생한다. 많은 특징을 사용하면 분류 모델의 학습률은 올라가지만, 구축에 사용되지 않은 새로운 데이터(테스트 데이터)에 적용할 경우에는 분류 정확도가 낮아지는(오류율이 높아지는) 문제가 발생한다(Lee, 2007). 따라서 분류 모델에 대한 일반화(generalization) 실험을 수행하여 과적합을 배제한다. 일반화는 분류 모델을 실제 시스템 환경에 적용 가능한지를 측정하기 위한 과정을 의미한다. 학습 데이터를 통해 도출된 분류기를 테스트 데이터로 실험하여 그 결과에서 보여주는 성능을 평가 기준으로 사용하는 것을 말한다. 이처럼 기계 학습의 성능은 데이터의 양과 질에 의존적이기 때문에 일반적으로 충분한 데이터를 먼저 수집한 후에 이 중 유용한 데이터만을 도출하여 활용한다. 이 때 사용하는 각 데이터를 특징(feature)이라 하며 어떤 특징이 유용한지 아닌지 확인하는 과정을 거친다. 이 과정을 특징 선택(feature selection)이라 하며 기존 입력을 토대로 줄어든 새로운 입력 데이터를 만들기 때문에 보통 학습 과정 전에 수행되는 핵심적인 전처리 과정 중 하나이다.

2.2. 특징 선택

패턴 인식, 기계학습 분야에서 사용되는 분석 기준, 혹은 데이터를 특징이라 한다. 예를 들어 캠페인 성공 예측에 사용되는 데이터에서는 고객의 성별, 나이, 서비스가입기간, ARPU(Average Revenue Per User), 고객등급, 멤버십가입여부, 멤버십 잔여 포인트, 서비스 만료 일자, 포인트 사용량 등을 특정 캠페인 수행을 위한 특징으로 정의하여 활용한다. 특징 선택은 주어진 문제에서 효과적이고 개선된 해를 얻기 위해 유용한 특징들을 선택하는 처리 과정이다(Cho et al., 2008). 특징 선택 기법은 특징 개수의 감소를 통해 계산량을 줄이고 분류 정확도를 향상하게 시키는 동시에 일반화 능력을 향상하게 시키는 것을 목적으로 한다(Molina et al., 2002). 특징 선택은 선택된 특징 조합을 탐색하는 최적의 조합을 선택하는 문제로 정의할 수 있으며 특징의 효율성과 분석 결과의 정확도 등을 향상하게 시키기 위한 중요한 문제이다(Guyon and Elisseeff, 2003). 특징 선택 방법은 데이터를 분석하는 과정에서 관련성이 높은 특징 혹은 변수, 특징들의 부분 집합을 선택하는 과정이다. 모든 특징에 대하여 특징 중요도를 평가하고 이 순서에 따라 순위를 부여하여 특징 선택에 활용한다. 특징 중요도를 결정하는 기준은 크게 두 가지로 구분할 수 있다. 첫 번째는 분별력(discriminatory power)이다. 좋은 특징은 서로 다른 부류를 잘 구별할 수 있어야 한다. 두 번째는 차원(dimensionality)이다. 특징 차원이 낮을수록 계산 효율이 높고 의미해석이 가능해진다. 일반적으로 정보기반의 예측값에 근거하여 특징 중요도 순위를 매긴 후 어떤 임계 값을 만족하거나 초과하는 특징만 고르거나 단순히 상위 k개의 특징을 고르는 형태



<Figure 1> Feature selection process

로 수행된다

2.2.1. 특징 선택 실행 절차

특징 선택 과정은 경쟁하는 여러 개의 후보 부분 집합 중에서 최적 특징 집합을 구하는 과정이며 총 두 단계로 구성된다. 다음 후보 집합을 생성하기 위한 생성 단계, 대상 후보 집합을 평가하는 단계이다. <Figure 1>은 특징 선택의 각 단계를 도식화한 그림이다.

첫 번째로 생성 단계는 새로운 부분 집합을 탐색하는 절차이다. 기본적으로 평가를 위한 특징 부분 집합을 생성한다. 이를 위한 기법들은 탐색 전략에 따라 크게 전역 탐색(global search), 순차 탐색(sequential search), 임의 탐색(random search)으로 구분된다. 두 번째는 생성한 후보 특징 집합이 특정 평가 기준에 따라 분류에 적합한지 아닌지를 확인하기 위한 평가 단계이다. 평가 방법에 따라 두 가지로 구분할 수 있다. 필터 방법의 경우 특징 집합의 개별 특징 정보만으로 평가하는 방법이며, 래퍼 방법의 경우 대상 특징 집합으로 학습된 분류기가 도출하는 분류 정확도를 평가 기준으로 사용하는 방식이다. 특징 선택 프로세스의 각 단계는 반복적으로 수행되고 미리 정의된 중지 기준이 충족되었을 경우 전체 프로세스를 종료하는 형태로 동작한다. 특징 선택이 종료된 이후, 선택된 특징 집합은 여러 가지 테스트를 통해 유효성을 검사하는 과정을 거친다.

2.2.2. 실행 단계에 따른 분류

특징 선택은 수행하는 단계에 따라 각각 다른 알고리즘이 사용된다. 특징 생성 과정(feature subset generation)에서 주로 사용하는 탐색 기법은 탐욕 알고리즘(greedy algorithm), 유전자 알고리즘(Genetic Algorithm, GA), 파티클 집단 최적화 알고리즘(Particle Swarm Optimization, PSO), 개미 군집 최적화 알고리즘(Ant Colony Optimization, ACO) 등이 있다. 탐욕 알고리즘은 SFS(Sequential Forward Selection), SFFS(Sequential Forward Floating Selection), SBS(Sequential Backward Selection), SFBS(Sequential floating backward selection) 등이 있다. 이 중 SFS방식은 빈 세트로 시작하여 하나의 특징마다 분류 정확도를 측정하고 그 중에서 가장 정확도가 높은 특징을 추가한다. 이후 모든 특징을 비교할 때까지 새로 추가할 특징 중 가장 정확도가 높은 특징을 추가하는 과정을 반복한다(Devijver, 1982). 다음 <Figure 2>는 SFS의 처리 절차를 의사 코드로 나타낸다(Ferri et al., 1994).

1. Start with the empty set $Y_0 = \{\emptyset\}$
2. Select the next best feature $x^+ = \arg\max_{x \notin Y_k} J(Y_k + x)$
3. Update $Y_{k+1} = Y_k + x^+;\ k = k + 1$
4. Go to 2

<Figure 2> SFS pseudo code

Step 1 (Inclusion):

$$x^+ = \arg\max J(x_k + x),\ \text{where } x \in Y - X_k$$
$$X_{k+1} = X_k + x^+$$
$$k = k + 1$$
Go to Step 2

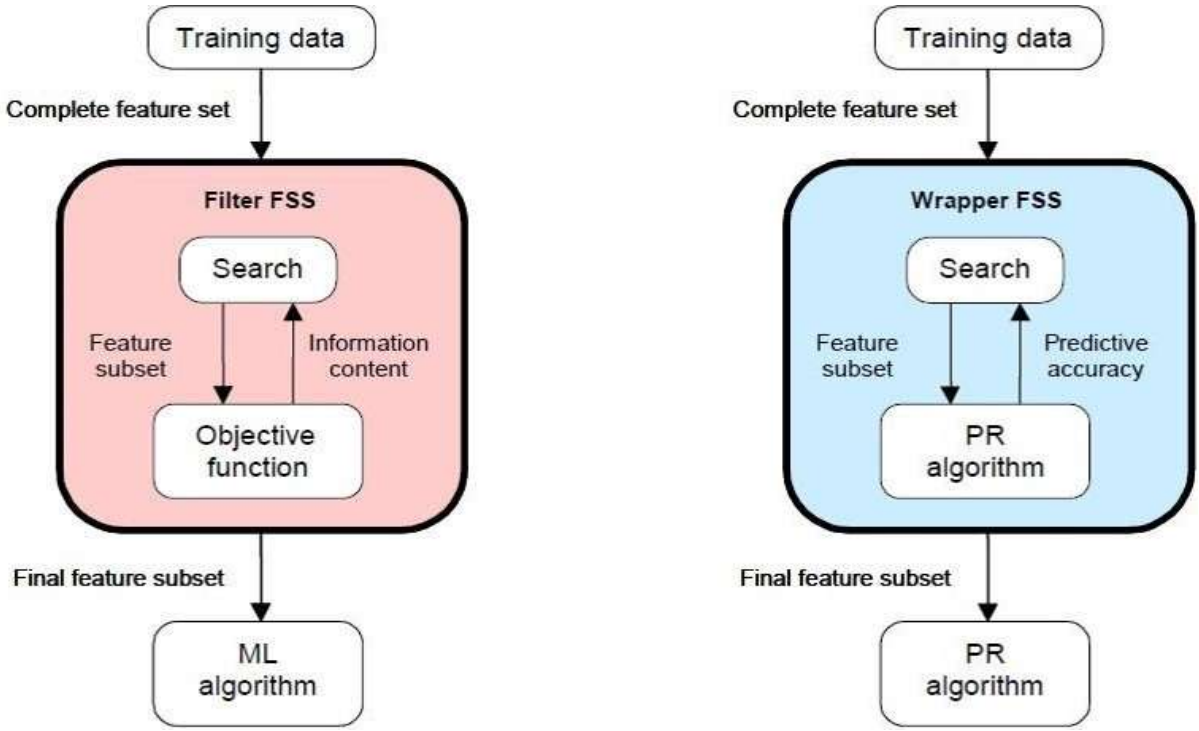
Step 2 (Conditional Exclusion):

$$x^- = \arg\max J(x_k - x),\ \text{where } x \in X_k$$
$$if\ J(x_k - x) > J(x_k - x):$$
$$X_{k-1} = X_k - x^-$$
$$k = k - 1$$
Go to Step 1

<Figure 3> SFFS pseudo code

비교적 단순하게 계산할 수 있고 적은 특징일 경우 최적 성능을 보이는 특징 선택 방법이다. 단점은 많은 수의 특징을 사용할수록 계산 복잡도가 증가하고, 한번 추가된 특징은 다시 제거할 수 없다는 점이다. 이러한 단점을 보완하기 위한 SFFS 방식이 있다. SFFS방식은 공집합으로 시작하여 각 순차 단계에서 최적 특징을 선택하여 특징 부분 집합을 생성한 후 그 집합 중 최악의 특징을 선택하여 제거하는 방식이다(Pudil, 1994). 한번 최적 특징으로 선택된 특징이라 하더라도 분류 성능을 감소시킬 경우 특징 조합에서 제외할 수 있으므로 SFFS 방식을 이용하여 SFS의 단점을 보완할 수 있다. <Figure 3>은 SFFS의 처리 절차를 나타낸 코드이다(Pudil, 1994)

특징 평가(feature subset evaluation) 단계는 모든 부분 집합을 이용하여 분별력을 기준으로 평가한 후 그 중의 가장 높은 점수를 받은 부분 집합을 선택하는 과정이다. 크게 필터 방식(filter method)(Bluma and Langley, 1997)과 래퍼 방식(wrapper method)(Kohavi, Ron and John, 1997) 두 종류로 나눌 수 있다. 필터 방식은 특징 집합을 개별 특징의 독립된 정보만으로 평가하는 방식이며 래퍼 방식은 분류기의 성능을 평가 척도로



<Figure 4> Filter method vs. Wrapper method

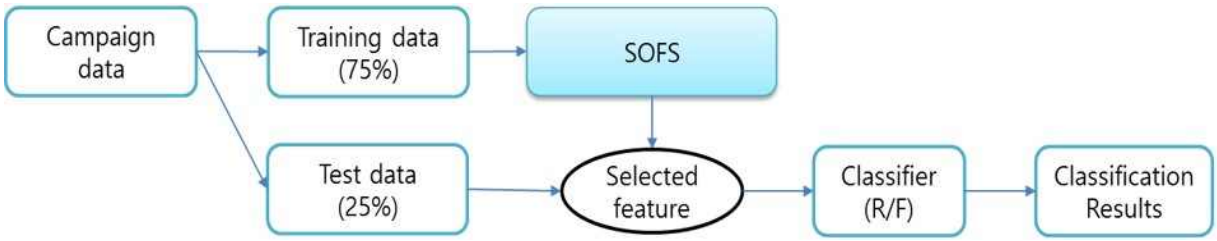
사용하는 방식이다. 다음 <Figure 4>는 필터 방식과 래퍼 방식을 나타낸다.

필터 방식은 학습 알고리즘의 독립성을 가진 전처리 단계로서 다양한 통계적 방법을 이용하여 각 특징의 성능을 평가하여 특징 선택 과정을 수행한다. 선택된 부분 집합의 성능을 부분 집합에 포함된 특징들과 분류 기준 사이의 고유한 특징을 이용하여 평가한다. 높은 순위의 특징이 먼저 선택된다(Yu, Lei and Liu, 2003). 필터 방식에서 독립적인 기준으로 특징을 평가하는 기준으로 일반적으로 쓰이는 방식은 상호 정보량(mutual information), 상관계수(correlation criteria), 데이터들의 거리(distance) 등이 있다(Guyon and Elisseeff, 2003). 필터 방식은 연산 속도가 빠르므로 고차원 데이터 셋을 분석하는 데 적합하다(Lee and Oh, 2016). 필터 방식의 단점은 다음과 같다. 첫째, 어떤 평가 기준을 사용하는가에 따라 선택되는 특징이 달라질 수 있다. 둘째, 하나의 특징의 개별적인 평가는 우수할 수 있으나 선택된 특징이 실제 분류기를 사용했을 때의 분류 성능과 차이를 보일 수 있다. 셋째, 단일 특징들의 가중치만을 기준으로 평가한 독립된 수치를 통해 특징을 선택하기 때문에 특징 간의 상관관계를 고려하지 못한다. 래퍼 방식은 종속적인 특징 부분 집합 평가 기준을 두는 방법이다. SVM(Support Vector Machine), 의사결정 트리(decision tree), KNN(K-Nearest Neighbor) 등과 같은 특정 기계학습 알고리즘을 적용하고 해당 분류 성능을 사용하여 특징 선택을 한다. 이 방법은 직접적으로 분류기를 사용하여 해당 특징 부

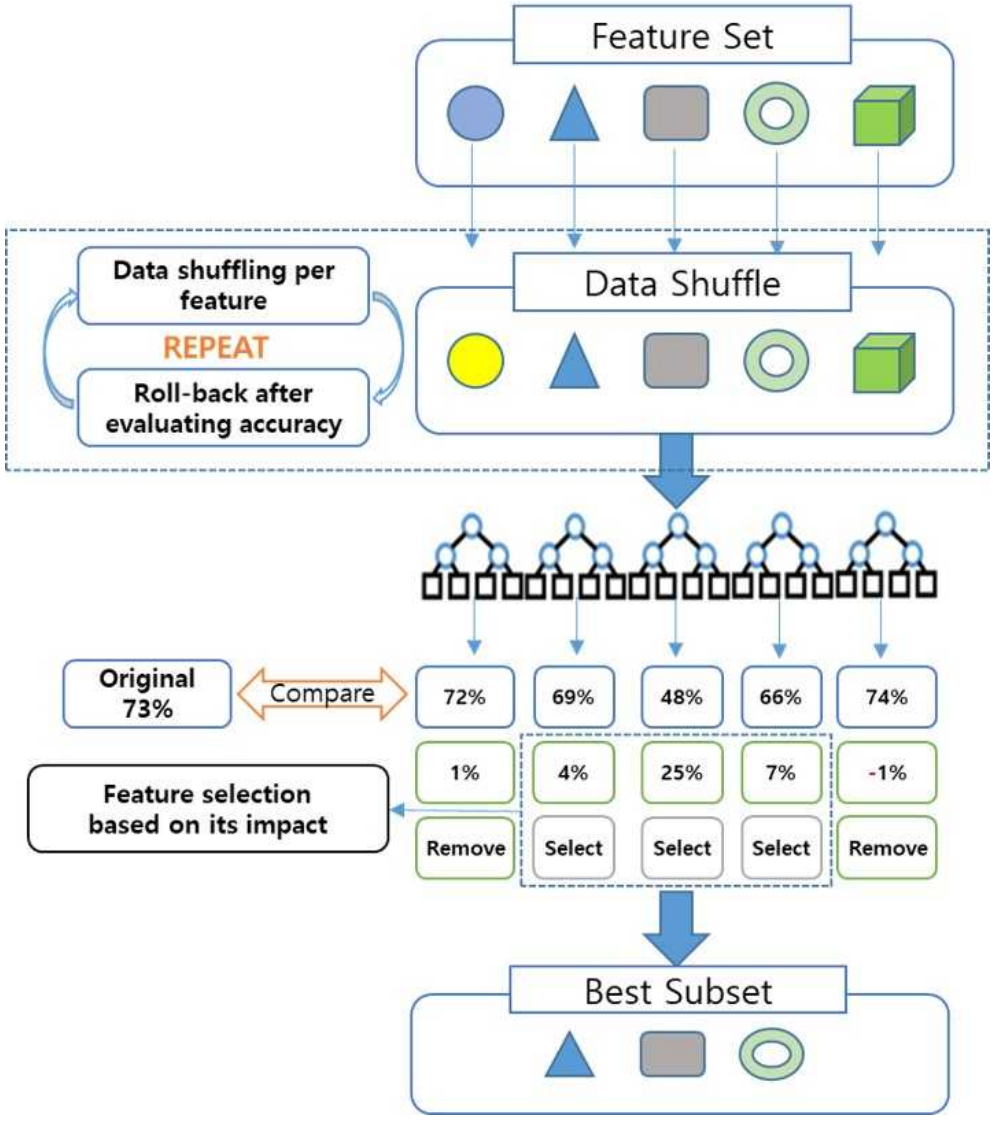
분 집합의 성능을 평가한다. 사용하는 분류기를 사전에 결정한 다음, 매번 특징 부분 집합이 형성될 때마다 기계 학습 알고리즘을 사용해 데이터를 분류하고 그 분류 성능을 기준으로 특징 부분 집합을 평가한다(Kohavi, Ron and John, 1997). 래퍼 기법의 장점은 다음과 같다. 첫째, 분류기의 분류 결과를 평가하여 특징을 선택하기 때문에 직접적으로 분류기에 최적화된 특징을 선택할 수 있어 가장 잘 분류할 수 있는 특징 부분 집합을 찾아낼 수 있다. 둘째, 특징 간의 상관관계가 높은 부분 집합에 선택할 수 있다. 래퍼 기법의 단점은 다음과 같다. 첫째, 매번 새로운 분류 모델을 생성하여 특징 부분을 평가하기 때문에 계산 복잡도가 높다. 따라서 필터 기법보다 연산 시간이 오래 걸린다(Lee and Oh, 2016). 실제 캠페인 시스템에 적용하기 위해서는 이 부분에 대한 고려가 필요하다. 둘째, 과적합의 위험이 높으므로 일반화된 분류 모델을 생성하기 어렵다. 학습 데이터를 정확히 분류할 수 있는 모델이더라도 일반화가 되지 않았다면 실제 데이터를 처리할 경우 오탐이나 미탐을 발생시킬 수 있다. 보통 필터 방식은 다수의 특징을 갖는 고차원 데이터에서 특징을 줄이기 위해 사용되고, 래퍼 방식의 경우 특징 간의 상관관계를 고려한 최적의 특징 조합을 생성하기 위해 사용되기 때문에 필터와 래퍼 방식은 데이터의 특성에 맞게 취사 선택하여 사용이 필요하다. 본 연구에서는 특징 부분 집합을 생성하기 위하여 SFFS방식을 활용한다. SFFS 방식은 관련 없는 특징을 배제함으로써 컴퓨팅 효율과 모델의 일반화 오류를 감소시킬 수 있는 장점을 가지지만, 순차적 실행 방법의 한계를 가지기 때문에 이를 보완하려는 방안이 필요하다. 또한 SFFS의 특징 평가 방식으로 단일 분류기를 사용할 경우 성능 향상이 크지 않기 때문에 다양한 분류 알고리즘을 혼합하여 최적의 분류 예측이 되도록 단점을 보완할 필요가 있다. 이를 위해 본 연구에서는 SOFS 알고리즘을 제안한다. 제안하는 알고리즘은 SFFS의 순차적 특징 평가 방식에 비해 특징이 평가에 미치는 부정적인 특징을 제거하는 방식으로 변경하여 일반성과 성능 개선에 목적을 두는 형태로 목적 함수는 랜덤포레스트 분류기를 사용한다.

3. SOFS 특징선택 알고리즘

본 연구에서 제안하는 모형은 SOFS(Self-optimizing Feature Selection) 알고리즘 기반의 특징 선택에 대한 개선된 모델이다. SOFS 알고리즘은 지도학습(supervised learning)을 위한 특징 선택 알고리즘이며, 제2장 관련 연구에서 언급하였던 래퍼 방식을 기반으로 설계하였다. <Figure 5>는 제안하는 특징 선택 기법에 대한 간단 개념도를 나타낸다. 기계학습 과정은 크게 학습 과정(training phase)과 테스트 과정(test phase)으로 나뉜다. 첫 번째로 학습 과정에서는 데이터를 정제하고 모든 특징에 대해서 특징 선택 과정(feature selection phase)을 수행한다. 전체 특징에서 필요한 최적의 특징 부분 집합을 선택하는 과정이며 이 단계에서 제안하는 알고리즘이 수행된다. 마지막으로 테스트 과정은 모델의 성능을 평가하는 단계이다. SVM, KNN 등등 기계학습 알고리즘을 활용하여 분류기(classifier)를 생성하고 해당 분류기의 분류 결과(성공, 실패)를 테스트하는 단계이며 캠페인 시스템에서는 분류 결과를 반응성공률(response rate) 및 분석정확도(accuracy)로 평가한다. 여기서 반응성공률과 분석 정확도는 캠



<Figure 5> Self-Optimizing Feature Selection (SOFS) for machine learning



<Figure 6> SOFS process

페인성공과 실패를 정확하게 분류하는 능력을 의미한다.

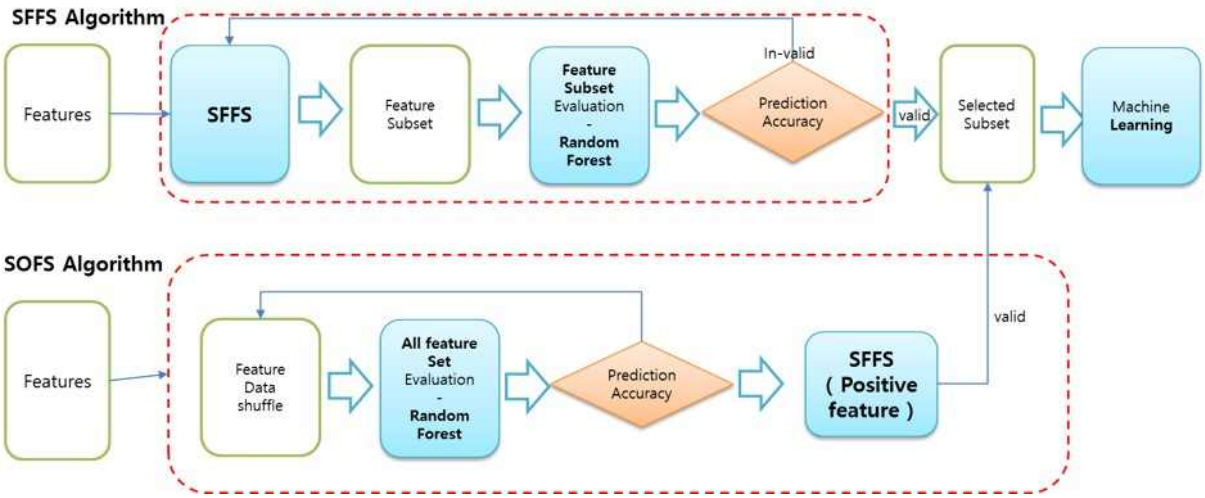
기존 SFFS 기법은 각 특징별로 평가를 수행하여 불필요한 특징을 줄여주는 장점이 있지만 최적 특징만을 학습하는 모델을 생성하기 때문에 과적합에 빠질 가능성이 높다. 또한, 단일 분류기로 사용할 경우 분류 성능이 떨어진다. 따라서, SFFS 기법을 보완하기위해 다중 분류기를 사용하여야 성능 향상을 기대할 수 있다. 이를 해결하기 위해 랜덤포레스트 분류 기를 사용한다. 랜덤포레스트 분류기는 랜덤으로 선택된 특징을 기반으로 자체적으로 다수 개의 분류모델(의사결정트리)을 생성하기 때문에 다중 분류기의 특성이 있다. 하지만, 모든 특징을 기계학습에 사용하기 때문에 중복 특징이나 관련 없는 특징까지도 학습하는 단점이 있다. 두 기법은 단일 기법으로는 사용하는데 한계가 있으며 본 연구에서는 두 기법의 상호보완을 통해 단점을 해결하기 위한 SOFS 특징 최적화 알고리즘을 제안한다

3.1. SOFS 특징 부분 집합 최적화 알고리즘

일반적으로 특징 선택 과정은 전체 데이터에서 유용한 특징들의 부분 집합을 선택하는 과정이다. 새로운 특징 부분 집합을 선택하기 위한 탐색전략(search strategy)과 생성된 특징 부분 집합을 평가하려는 방법(evaluation method) 그리고 특징 선택 과정을 종료하기 위한 정지 기준(stop criterion), 이렇게 세 부분으로 구성된다. <Figure 6>는 SOFS의 특징 선택 과정을 도식화하여 나타내고 있다.

3.2. 탐색전략

특징 탐색 단계에서는 변형한 SFFS 기법을 통해 데이터의 불필요한 특징이나 중복된 특징을 제거하고 중요한 특징을 추출한다. 분류 성능이 높은 특징들의 부분 집합을 선택하는 과정이다. 이 과정에서 목적함수(objective function)는 특징 평가 또는 재학습을 통해 예측 모델을 만들어내는 역할을 한다. 예측 모델은 특징이 추가되었을



<Figure 7> Comparison between SFFS and SOFS algorithms

때 분류 성능이 향상되는지를 파악하는 데 사용된다. 본 연구에서 제안하는 SOFS 알고리즘은 기존 SFFS 알고리즘이 순차적으로 하나씩 특징을 탐색하기 때문에 자원과 성능 측면에서 비 효율적이라는 단점을 개선하는 것을 목적으로 한다. <Figure 7>은 특징 선택 과정에서 기존 SFFS 알고리즘을 사용하는 경우와 개선된 SFFS 알고리즘을 사용하는 경우에 대한 동작 흐름을 도식화하여 나타내고 있다.

기존 SFFS 알고리즘은 순차적 탐색 방식을 사용하여 특징 부분 집합을 생성한다. 후보 특징 부분 집합을 하나씩 순차적으로 생성하는 과정이며 모든 특징에 대한 최초 집합을 생성하고 성능 평가에 따라 최적 부분 집합에 특징을 추가하는 형태로 동작한다. 다음 <Figure 8>은 기존 SFFS의 처리 절차를 나타낸 의사 코드이다.

Input:
All features $\{F\} := (f_1, f_2 \ldots)$
Output:
Optimal features without redundant and irrelevant features $\{F_O\}$

Code:
function ***SFFS()***{
  initialize $\{F_O\} = empty;$

  for $k = 1$ *to number of* $\{F\}$
    //forward selection
    $f_i^+ = \underset{f_i \in F}{arg\,max}[eval(\{F_O\} + f_i)]$
    $\{F_O\} = \{F_O\} + f_i^+$

    // backward elimination
    $f_i^- = \underset{f_i \in F}{arg\,max}[eval(\{F_O\} - f_i)]$
    if $(eval(\{F_O\} - f_i) > eval(\{F_O\})$ *then*
      $\{F_O\} = \{F_O\} - f_i^-$
    end if
  end for
  return $\{F_O\}$
};

<Figure 8> SFFS pseudo code

입력으로 모든 특징이 포함된 집합 F를 사용하여 도출되는 결과로 최종적으로 선택된 최적 특징 집합 FO가 생성된다. 결과로 도출되는 특징 집합 FO는 초기값이 공집합으로 시작한다. 생성되는 특징 집합 FO는 순차적으로 탐색하면서 특징을 하나씩 비교한다. 첫 번째 단계로 전진 선택(forward selection)에서는 하나의 특징이 추가하는 부분이며, 기존 특징 집합 FO에 있는 특징과 비교하여 정확도가 높은 특징만 특징 집합에 저장하고 정확도가 낮은 특징은 저장하지 않는다. 두 번째 단계로 후방 제거(backward elimination)는 특징 집합 FO에서 있는 특징 중에서 정확도가 낮은 특징을 제거하는 단계다. 특징 집합을 평가하는 기준으로는 랜덤포레스트 분류기를 사용한다. 특징 부분 집합에 대한 학습시킨 후 분류 정확도를 도출하여 사용한다. 이처럼 SFFS는 최적의 특징 부분 집합을 도출할 수 있지만 하나의 특징을 한 번에 하나씩 비교하기 때문에 속도가 너무 느리다는 단점을 가지고 있다. 본 연구에서 제안하는 탐색 전략은 한 번에 하나씩 특징을 탐색하여 평가하는 기법을 변형하여 분류 성능에 가장 영향을 많이 미치는 특징들을 선택하여 최소 탐색을 하는 방식을 사용하여 개선하였다.

Input:
All features{F}:=(f1,f2...)
Output:
Optimal features without redundant and irrelevant features {Fo}

Code:
function SOFS(){
  // get Feature Importance list from randomforest
  for i=1 to number of {F}
    {Fi} = shuffle({fi})

    // Impurity of Features
    Importance = randomforest({F},{Fi})

    // negative feature remove
    if Importance > 0.05 then
      {Fj} = {Fi}
    end if
  end for

  initialize{Fo} = {Fj} // features subset;
  k = number of {Fo} ;
  j = number of {F} - k

  for i=1 to j
  //forward selection & get best accuracy with
    ACCURACY = randomforest({Fo} + fi)
    if(ACCURACY > MAX_ACCURACY)
      then MAX_ACCURCY =
      {Fo} = {Fo} + fi
    end if
  end for
  return {Fo}
};

<Figure 9> SOFS pseudo code

<Figure 9>는 SOFS 알고리즘의 실행 절차를 기록한 의사 코드이다. 입력으로 모든 특징이 포함된 집합 F을 사용하고 도출되는 결과는 최종적으로 선택된 최적 특징 집합 F0과 특징 중요도가 생성된다. 결과로 도출되는 특징 집합 F0는 초기값이 공집합으로 시작하지만 F에 포함된 모든 특징을 개별적으로 혼합(shuffle)하여 특징의 중요도를 측정하여 F0에 추가한다. 최종 생성되는 특징 집합 F0는 부정적인 영향을 끼칠 수 있는 특징은 제거를 한 특징의 수를 기준으로 이후 제거된 특징을 순차적으로 하나씩 추가하면서 평가 과정을 수행하여 분류 정확도가 더 이상 증가하지 않으면 종료하게 된다. 특징 집합을 평가하는 기준으로는 랜덤포레스트 분류기를 사용한다. SFFS의 특징을 하나씩 추가하면서 진행하는 순차적인 탐색에 대비하여 분류 성능에 영향을 미치는 특징들을 우선 선별하고 이후 제외된 특징들을 순차적으로 추가하는 형태로 동작하므로 순차적 탐색 기법보다 탐색 횟수가 훨씬 줄고 컴퓨팅 자원을 효율적으로 활용할 수 있는 장점을 가진다.

3.3. 평가 기법

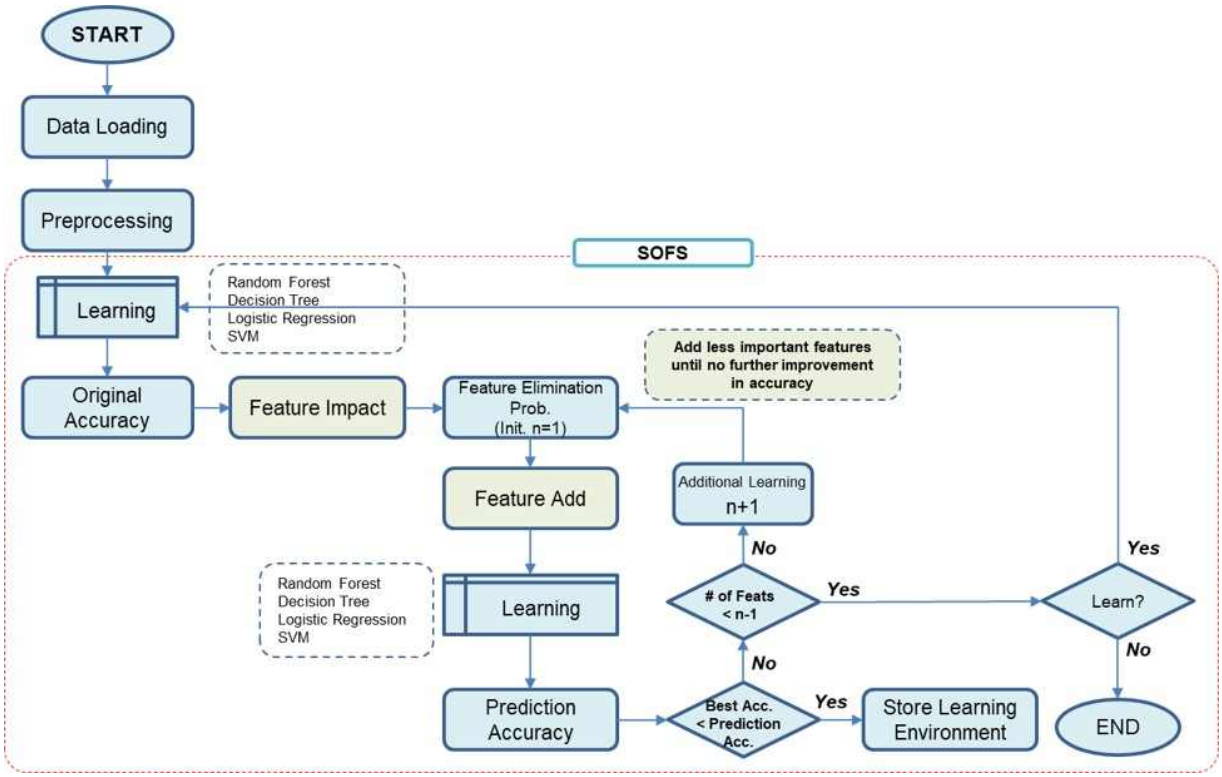
본 연구에서는 랜덤포레스트 분류기를 목적함수로 사용하여 분류 모델을 생성하고 최적 특징 집합을 도출한다. SOFS를 통해 특징 부분 집합이 생성될 때마다 랜덤포레스트를 통해 분류 모델을 도출하고 이 모델의 분류 정확도를 특징 부분 집합을 평가하는 수치로 활용하는 형태이다. 우수성을 평가하기 위해 정확도를 사용하여 중요도를 평가한다. 정확도는 다음과 같이 정의된다.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

특징 선택 과정이 완료되면 랜덤포레스트를 수행한다. 최종 특징 집합에 특징이 저장되어 있으면 여러 개의 특징 부분 집합으로 나눈다. 또한, 전체 데이터도 여러 개의 부트스트랩 부분 집합으로 나눈다. 최종 특징 집합과 부트스트랩 부분 집합을 이용하여 의사결정트리를 만들게 되며 여러 개의 트리는 앙상블을 통해서 하나의 분류 모델을 만들어 낸다. 랜덤포레스트 분류기는 랜덤으로 선택된 특징을 기반으로 많은 수의 분류기를 학습시키는 방식으로 동작하며 학습된 많은 수의 분류 모델을 앙상블 기법을 통해 결합하여 예측 결과를 도출하기 때문에 단일 분류기 사용보다 더 좋은 예측 성능을 보인다.

3.4. 정지 기준

제안하는 SOFS 알고리즘에서는 SFFS와 같이 모든 특징에 대한 평가를 수행하고 더 이상 특징 공간에 추가할 특징이 존재하지 않은 경우를 정지 기준으로 설정하지 않고 분류 정확도를 기준으로 더 이상 정확도가 늘지 않으면 최종 특징 선택 과정을 종료하고 최적 성능을 내는 특징 부분 집합이 도출된다. 또한, 특징 중요도에 따른 순위 정보를 함께 제공한다. 이는 전체 특징들을 분류와 높은 연관이 있는 특징을 순위로 표기한 것을 의미한다. 제안하는 SOFS 특징 최적화 알고리즘을 이용하여 분류 정확도에 기여하는 특징은 남기고 필요 없는 특징은 제외할 수 있는 기대효과가 있으며 결과에 영향을 미치는 특징을 명확히 확인할 수 있으므로 랜덤포레스트의 단점으로 지적됐던 분류 결과에 대한 해석이 어



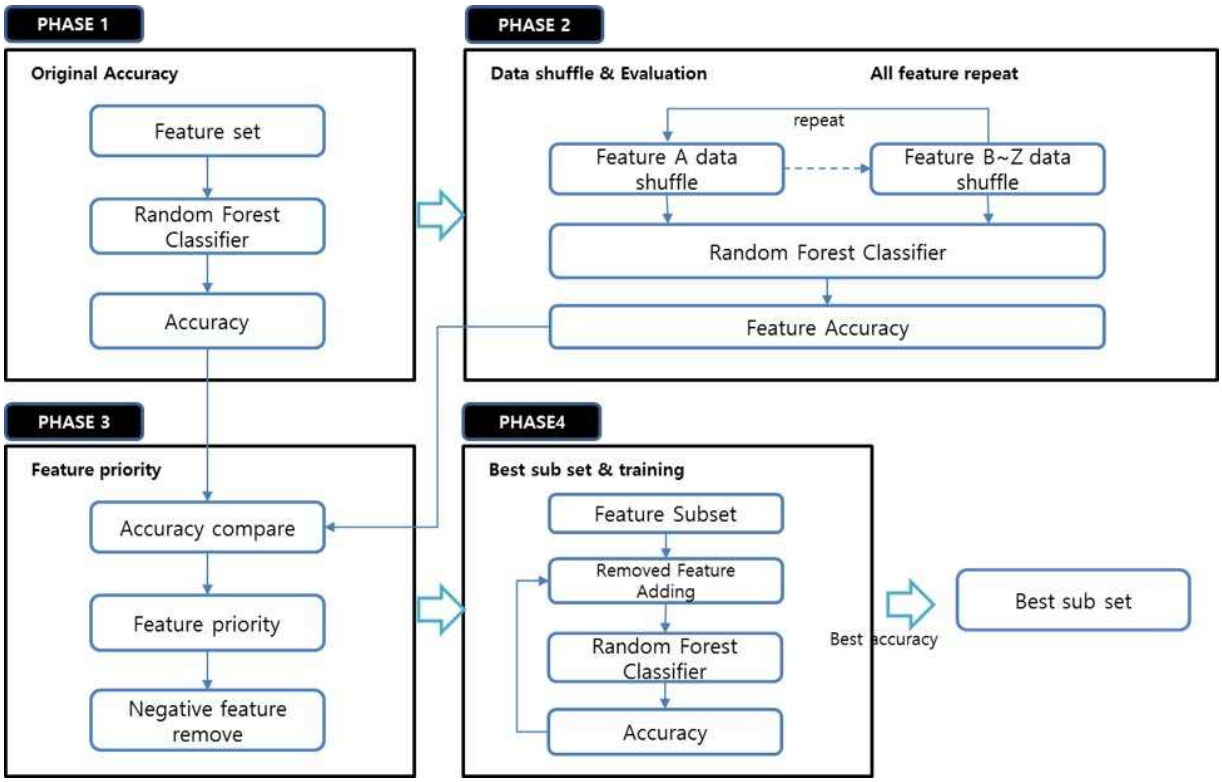
<Figure 10> Campaign targeting system based SOFS algorithm

려웠던 부분을 보완할 수 있다. 특히 많은 수의 데이터와 특징들이 있는 경우 좋은 성능을 나타내고 있어 캠페인 수행을 위한 성공 예측 시스템에 매우 적합하다.

4. SOFS 기반 캠페인 타겟팅 시스템 구성

앞 장에서 살펴본 기존 연구들에 대한 분석결과를 토대로 본 연구에서는 캠페인 성공 여부에 대한 예측을 위해 기존 SFFS의 성능을 개선한 SOFS 알고리즘을 새롭게 제안한다. 특히 본 연구에서는 그간 탐욕 알고리즘인 SBS, SBS, SFFS, SFBS 과의 비교와 더불어 GA(Genetic Algorithm), RFE(Recursive Feature Elimination) 과 같이 그 효과가 입증된 알고리즘 과 분류성능을 비교하여 어느 정도 최적화가 개선되는지를 확인하고자 한다. 본 연구에서는 편의상 제안 모형을 SOFS(Self Optimizing Feature Selection) 알고리즘으로 명명하였다.

<Figure 10>은 캠페인 시스템에서의 SOFS 알고리즘을 활용한 캠페인 성공 예측을 위한 흐름도이다. 학습을 위한 데이터를 로딩하고 전처리를 한 후 학습 모델을 먼저 수행한다. SOFS 알고리즘은 컬럼(column)의 데이터를 무작위로 재정렬하면 결과 데이터가 더 이상 실제 데이터와 일치하지 않으므로 예측 정확도가 떨어진다는 점을 이용한다. 모델이 예측에 크게 의존하는 특징



<Figure 11> Detailed process of SOFS

을 섞으면 모델 정확도가 특히 떨어지게 되는데 거의 정확도가 떨어지지 않는 특징들을 제거한 하위 데이터셋(sub dataset)을 기준으로 제거한 특징을 순차적으로 추가해 가면서 정확도가 증가하지 않을 때까지 수행하여 분류 정확도가 가장 높은 특징들을 선택을 한다.

해당 시스템은 특징 선택 과정에서 전체탐색을 하지 않고 중요도가 낮은 특징들 위주로 순차탐색을 함으로써 기계 학습 기법을 적용한 캠페인 대상자 선정을 위한 분류 효율을 향상시킬 수 있다.

다음의 <Figure 11>은 본 연구의 제안 알고리즘인 SOFS이 동작하는 단계를 프로세스 흐름 형식으로 보여주고 있다. 그림에서 보듯이, SOFS은 다음과 같이 크게 4단계에 의해 구현되도록 설계되었다.

Phase 1. 단계1은 입력된 특징 데이터에 대한 본래 분류 성능을 측정하는 단계이다. 이 단계에서는 이미 전처리 된 특징 데이터에 대한 기본적인 분류 성능을 측정하며 분류 성능 측정을 위해 랜덤포레스트를 사용한다. 이는 여러 개의 의사결정트리들을 학습하는 앙상블 (ensemble) 방식의 분류기이다(Breiman, 2001). 앙상블은 여러 분류 모델을 만들고 이 모델의 결과를 조합하여 예측 결과를 도출하는 것이다. 이것은 많은 약한 분류기(weak classifier)를 결합하여 하나의 강력한 분류기(strong classifier)를 생성하는 기법이다. 하나의 약한 분류기는 지역 최적화(local optimization)에 빠지거나 특정 학습 데이터 셋에

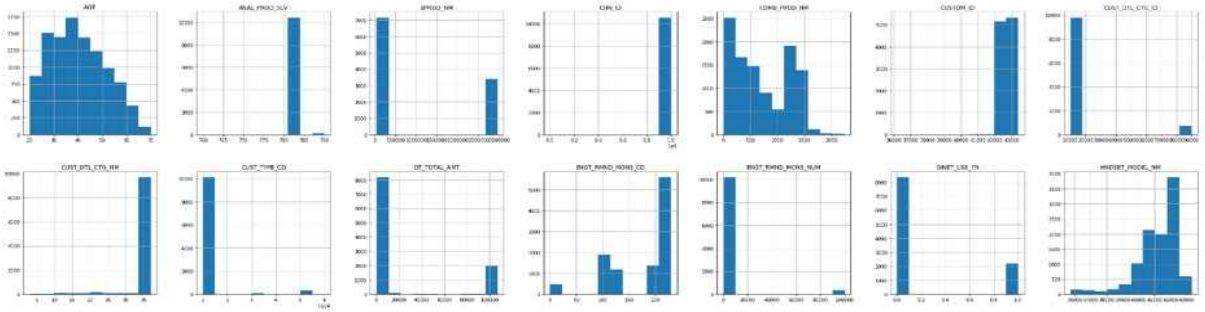
과적합 될 수 있으므로 여러 개의 약한 분류기를 결합하면 정확도를 개선할 수 있다.

앙상블 과정은 분류 모델 생성과 예측값 결합 과정으로 구분할 수 있다. 예측값을 결합하는 마지막 단계는 보통 투표(voting)를 통해 수행된다. 약한 분류기 중 가장 많은 득표를 한 예측 클래스가 결합한 앙상블 모델이 출력된다. 하나의 약한 분류기는 각각 다른 정확도로 결과를 예측한다. 따라서 약한 분류기의 정확도에 따라 투표에 가중치를 줄 수도 있다. 즉, 높은 정확도를 가지는 약한 분류기는 마지막 통합과정에서 낮은 정확도의 분류기보다 높은 대표성을 가진다. 여러 개의 분류 모델을 결합하는 과정을 거치기 때문에 과적합을 평균화시켜 오류를 줄일 수 있다. 이를 통해 분류 모델의 일반화가 가능하다는 장점이 있다. 이처럼 앙상블 방법은 오류율을 개선하고 개별 모델의 편향(bias)을 극복할 수 있으므로 분류 문제에 있어서 가장 중요한 기법이다. 또한 랜덤포레스트는 배깅(bagging) 기법을 활용하여 앙상블을 수행한다. 배깅은 조금씩 다른 임의의 표본으로 여러 모델을 학습시켜 그 결과를 결합하는 앙상블 방식을 의미한다(Breiman and Leo, 1996). 결국 배깅과 의사결정 트리를 조합해서 사용하는 방식으로 동작한다. 전체 데이터에서 추출된 각 부트스트랩 샘플을 기반으로 의사결정트리를 생성하고 그 결과를 조합하는 과정을 거치게 된다.

Phase 2. 단계2에서는 특징별로 데이터 변이를 통해서 특징이 분류 성능에 얼마만큼의 영향을 미치는지를 확인하는 단계이다. 먼저 단일 컬럼 값을 섞고 결과 데이터 집합을 사용하여 예측을 수행한다. 데이터를 원래대로 되돌린 후 각 열의 영향도를 계산할 때까지 데이터 집합의 다음 열에 대해 위 단계를 반복한다. 컬럼을 무작위로 재정렬하면 결과 데이터가 더 이상 실제 세계에서 관찰되는 것과 더 이상 일치하지 않으므로 예측 정확도가 떨어지게 되며 모델이 예측에 크게 의존하는 열을 섞으면 모델 정확도가 특히 떨어지는 점을 활용한 알고리즘이다. 예를 들어 특징 중 LTE평균매출액을 셔플(shuffle)하면 예측 정확도가 많이 떨어지며 LTE최종가입일자나 사용자ID를 셔플하면 결과 예측에 거의 영향을 미치지 않는다. 이 특징 데이터의 변이를 하는 방법은 첫째 데이터를 무작위로 재정렬 하는 방법이 있으며 실제 세계에서 관찰되는 데이터를 그대로 활용하여 특징 영향도를 더 잘 판단할 수 있다는 장점이 있다. 두번째는 데이터의 평균값 등으로 일괄 업데이트를 하는 방법이 있으며 실제 데이터와는 약간 다른 형태이지만 아주 대량의 데이터를 처리하는데 성능상의 장점이 있는 방법이다. 이러한 특징 데이터의 셔플 및 변이를 활용하면 SFFS 에 비해 탐색 횟수가 훨씬 줄어 처리 속도가 빠르고 결과에 대해 이해하기가 쉬우며 일관된 특징의 중요도를 측정할 수 있다는 장점이 있다.

Phase 3. 이 단계는 단계 2에서 특징들의 데이터 셔플 후 정확도를 본래(original) 정확도와 비교하여 특징별 영향도를 정렬시키고 영향이 거의 없거나 부정적인 영향을 미치는 특징들은 제거를 한 하위 집합(subset)을 만드는 단계이다. 정확도가 많이 떨어진 특징들은 실제 더 중요하고 정확도에 영향을 많이 끼치는 특징이므로 우선순위를 상위로 하고 영향이 거의 없거나 오히려 정확도가 올라간 특징들은 제거(remove) 한다(기본값 0.05).

	USER_ID	INET_COMB_YN	IPTV_COMB_YN	PSTN_COMB_YN	MOBL_COMB_YN	MPHON_ENGT_EXP_RPERD_ITG	MPHON_ENGT_EXP_RPERD_ITG_CD	AGE
0	239682827	1	1	1	1	216	-16	57
1	239837710	1	1	1	1	210	-10	29
2	240720694	1	0	0	1	224	-24	49
3	241177915	1	1	0	1	116	16	44
4	242517942	1	1	0	1	124	24	37



<Figure 12> Customer cultivation campaign data set & feature histogram

<Table 1> Summary of experimental data set

Dataset	Campaign response		Total
	Success	Fail	
Training data	2,952	4,954	7,906
Test data	904	1,732	2,636

Phase 4. 단계 3에서 선택된 특징들에 대해 정확도 측정하여 저장 후 삭제된 특징들을 영향도가 높은 순으로 추가하면서 정확도를 측정하고 종료 조건(이전정확도 > 측정정확도) 충족될 때까지 반복하여 작업을 수행하게 된다. 이러한 과정을 통하여, 최적의 특징 데이터 셋이 결정이 되면 제안모형이 새로운 데이터(test data)의 예측에 있어서도 탁월한 성과를 보여주는지 확인하여, 제안모형의 일반화 가능성을 검증하게 된다.

5. 실증 분석

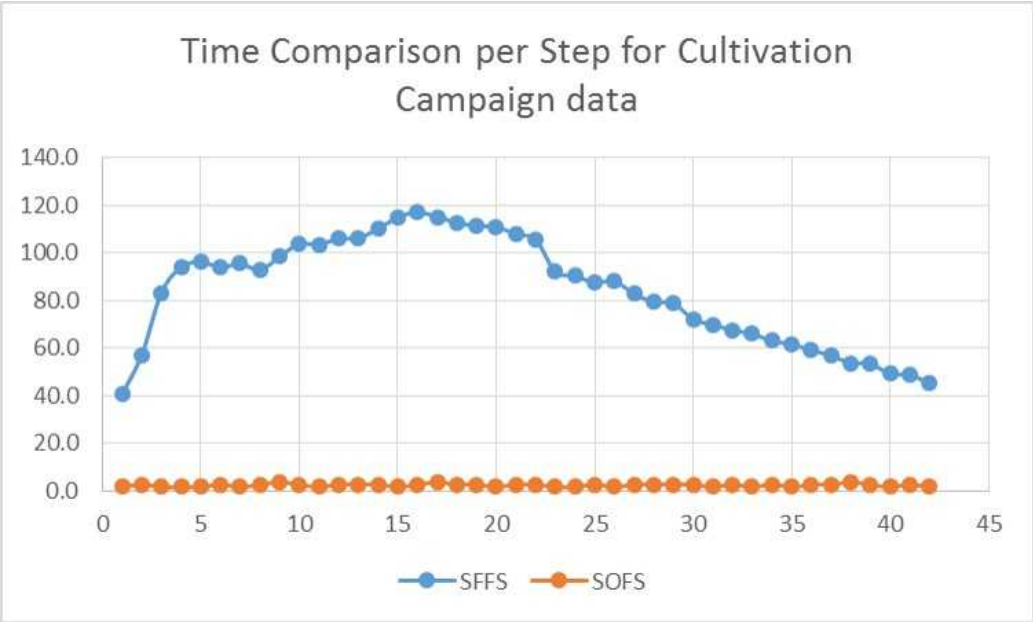
5.1. 실험 설계

본 연구에서는 SOFS 알고리즘을 검증하기 위해 2018년 4월 12일부터 2018년 5월 11일 일까지의 통신사의 고객 관계 강화 캠페인(customer cultivation campaign) 수행 이력 데이터 및 이와 관련된 통신사의 고객속성 정보, 캠페인 수행 성공결과가 기록된 캠페인 데이터를 기반으로 실험용 데이터셋을 수집하였다. <Table 1>에 제시한 바와 같이 10,542건의 관측치를 확보하였으며, 이 중 학습용 데이터 셋(training data set)으로 전체 데이터의 75%, 검증용 데이터 셋(test data set)으로 25%를 사용하였다.

특징변수들은 연령, 가입상품, 고객등급, 인터넷결합여부, 멤버십 잔여점수, 평균 ARPU 등 고객 속성 특징 40개와 캠페인 수행 결과(응답) 1개 특징으로 구성되어 있다. 종속변수로 활용한 캠페인 수행 결과는 실패(0) 6,686건, 성공(1)

<Table 2> List of the features

ID	Feature Name	Description
1	USER_ID	User's ID
2	INET_COMB_YN	Combination with Internet (1:Yes / 0:No)
3	IPTV_COMB_YN	Combination with IPTV (1:Yes / 0:No)
4	PSTN_COMB_YN	Combination with PSTN (1:Yes / 0:No)
5	MOBL_COMB_YN	Combination with Mobile (1:Yes / 0:No)
6	MPHON_ENGT_EXP_RPERD_ITG_CD	Remaining period of mobile phone contract expiration (Level Code)
7	MPHON_ENGT_EXP_RPERD_ITG	Remaining period of mobile phone contract expiration (Days)
8	AGE	Age
9	R3M_IPTV_AVG_ARPU_AMT	Average ARPU amount of IPTV in the last 3 months
10	R3M_MPHON_AVG_ARPU_AMT	Average ARPU amount of Mobile in the last 3 months
11	R3M_INET_AVG_ARPU_AMT	Average ARPU amount of Internet in the last 3 months
12	VIP_ADM_TAG_ITG_CD	VIP status code
13	GINET_USE_YN	Use of Giga Internet (1:Yes / 0:No)
14	CUST_DTL_CTG_NM	Customer's detailed category
15	SRVC_USE_CD	Period of total service use
16	ENGT_RMND_MONS_CD	Remaining period of contract code
17	ENGT_RMND_MONS_NUM	Remaining period of contract (Months)
18	OTS_SBSC_YN	Use of OTS(Olleh TV Skylife)
19	BPROD_NM	Number of basic product use
20	R3M_MPHON_AVG_ARPU	Average ARPU of Mobile in the last 3 months
21	CUSTOM_ID	Customer's ID
22	CUST_TYPE_CD	Customer's type code
23	R3M_LTE_AVG_AMT	Mobile Internet usage in the last 3 months
24	RM_LTE_AVG_AMT	Mobile data usage in the last month
25	SVC_USE_MONS_NUM	Period of mobile service use
26	RMONTH_TOT_BILL_AMT	Total billing amount in the last month
27	CUST_DTL_CTG_ID	Consent to consignment of customer's information
28	HNDSET_MODEL_NM	Handset model code
29	R6M_AVG_ARPU_AMT	Average ARPU of Internet in the last 6 months
30	COMB_PROD_NM	Combination product name
31	SPN_START_DATE	Start date of sponsorship application
32	SPN_TERM_DATE	End date of sponsorship application
33	MOBILE_PRE_CD	Ex-mobile service provider's code
34	SRVC_USE_DAY	Period of total service use
35	MIX_TYPE_TOTAL_CD	Code for service combination discount
36	ANAL_PROD_5LV	5-level analytic products
37	CHN_ID	Channel code
38	MEMBERSHIP_POINT	Remaining membership points
39	DT_TOTAL_AMT	Total amount of data provided
40	LAST_DAY_LTE_AMT	Data usage amount at the end of the month
41	LAST_DAY_LTE_RATE	Data usage rate at the end of the month
42	M2A_LTE_RATE	Average data usage rate in the last 3 months
43	RESPONSE	Campaign response (1:Success / 0:Fail)



<Figure 13> Processing time comparison between SFFS and SOFS

3,856건으로 구성되었다. 보다 상세한 특징변수 목록은 다음의 <Table 2>에 정리되어 있다. 독립변수로 사용된 고객 속성 특징들은 실제 캠페인 데이터 기반으로 Null값 처리 및 결측값 제거, 이상치 제거 및 연속형 변수, 범주형 변수 등의 전처리를 수행한 정제된 데이터로 알고리즘을 수행한다.

5.2. 실험 결과

제안하는 SOFS 알고리즘의 우수함을 검증하기 위해 먼저 특징 선택 과정 중 특징 탐색 과정과 특징 평가 방법을 구분하여 실험하였다.

5.2.1 특징 탐색 시간 비교

첫 번째 실험은 SFFS 알고리즘과 SOFS 알고리즘의 특징 탐색 시간을 비교하는 실험이다. 특징 탐색 방식의 성능 개선 여부를 파악하기 위해 기존 SFFS 알고리즘 방식과 개선된 방식으로 특징 선택을 수행하고 랜덤포레스트를 이용하여 평가를 하는 과정에서 소요되는 시간을 측정하였다.

<Figure 13>은 특징 선택을 진행하는 단계별로 걸리는 시간을 나타낸 그래프이다. 기존 방식은 단계당 평균 84.4초의 시간이 소요되는데 반하여 SOFS 알고리즘의 경우 평균 2.19초의 시간을 나타내고 있다. 또한 모든 단계에서 95% 이상 감소된 소요시간을 보이며 STEP 15의 경우 최대 98.7%의 시간 감소량을 나타내고 있다. 전체 소요된 시간은 SFFS는 3544.1초 SOFS는 92초가 소요되어 특징 선택에 걸리는 시간을 큰 폭으로 감소 시킬 수 있다는 것을 의미한다.

5.2.2 특징 평가 정확도 비교

두 번째 실험은 특징 선택하는 과정에서 부분집합을 평가하는 두 가지 방법을 비교하는 실험

<Table 3> Comparison of selected features for cultivation dataset

Method		No. of features	Selected Features
Search	Evaluator		
SFFS	R/F	24	1, 2, 3, 4, 5, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 23, 34, 36, 37, 39, 40, 42
SOFS	R/F	20	7, 8, 10, 16, 19, 20, 21, 23, 24, 26, 29, 30, 31, 32, 37, 38, 39, 40, 41, 42
GA	R/F	21	2, 3, 4, 8, 9, 10, 12, 14, 16, 20, 22, 23, 26, 27, 33, 34, 35, 36, 39, 40, 42
RFE	R/F	20	1, 8, 10 ,20, 21, 23, 24, 26, 28, 29, 30, 31, 32, 34, 37, 38, 39, 40, 41, 42

<Table 4> Comparison of prediction accuracy for cultivation dataset

Algorithm	Training dataset			Test dataset		
	DR(%)	FPR(%)	Time(sec)	DR(%)	FPR(%)	Time(sec)
SFFS	99.8	0.2	0.1	93.7	6.3	0.1
SOFS	99.9	0.1	0.1	94.08	5.92	0.1
GA	99.8	0.2	0.1	93.93	6.07	0.1
RFE	99.8	0.2	0.2	93.77	6.23	0.2
All feature	99.9	0.1	0.1	94	6	0.1

이다. 특징 부분 집합을 실험 성능을 검증하기 위해 SFFS, SOFS, GA, RFE 를 각각 특징부분집합을 탐색을 하였다.

<Table 3> 은 특징 선택 후 알고리즘 별 실험 결과이며 총 42개의 특징 중에서 서로 다른 특징들을 탐색 알고리즘에 따라 선택을 하였으며 특징선택의 개수도 서로 달랐다.

분류 성능 검증을 위하여 특징선택 기법을 통해 도출된 특징 셋을 기반으로 학습을 시켜 그 결과를 평가하는 실험을 수행하였다. <Table 4>는 학습 데이터 셋에 따른 분류 정확도와 테스트 데이터 셋에 대한 분류 성능 비교 결과를 나타내고 있으며 각각 성공률(DR, Detection Rate), 오탐율(FPR, False Positive Rate) 결과로 구성된다. 학습 데이터에서는 모든 알고리즘이 과적합되었기 때문에 탐지율에서 높은 정확도를 보이지만 테스트 데이터로 실험한 결과를 보면 성공율과 오탐율에서 차이를 보인다. SOFS를 기반으로 실험한 결과가 다른 알고리즘에 비해서 더 낮은 오탐율을 보여주고 있어 더 좋은 일반화를 보여주고 있다라고 해석된다.

추가로 SOFS의 경우 특징 별 영향도를 먼저 확인 후 영향도가 큰 특징을 선택하는 과정을 거쳤기 때문에 특징에 대한 중요도를 도출할 수 있다. <Table 5>는 각 특징이 분류 예측에 얼마만큼의 영향을 미치는지의 나타낸 표이다.

총 42개의 특징 중에서 분류기의 정확도 향상에 영향을 미치는 특징 20개가 결과로 도출되었고 특이한 점은 기존 캠페인 기획자들은 캠페인 대상자 선정에 사실 전혀 사용하지 않았던 결합상품명(COMB_PROD_NM), 평균3개월데이터소진율(M2A_LTE_RATE), 최근3개월무선데이터사용량(R3M_LTE_AVG_AMT)과 같은 특징들이 의외로 캠페인 반응에 중요한 특징으로 선택이

<Table 5> Impact rate of the selected feature subset from cultivation dataset

Rank	Feature ID	Feature name	Impact rate
1	39	DT_TOTAL_AMT	0.0443±0.0031
2	21	CUSTOM_ID	0.0386±0.0047
3	8	AGE	0.0191±0.0022
4	30	COMB_PROD_NM	0.0134±0.0012
5	41	LAST_DAY_LTE_RATE	0.0099±0.0029
6	31	SPN_START_DATE	0.0073±0.0021
7	42	M2A_LTE_RATE	0.007±0.0016
8	40	LAST_DAY_LTE_AMT	0.0062±0.001
9	23	R3M_LTE_AVG_AMT	0.0062±0.0027
10	26	RMONTH_TOT_BILL_AMT	0.0061±0.001
11	24	RM_LTE_AVG_AMT	0.0039±0.002
12	32	SPN_TERM_DATE	0.0034±0.0005
13	37	CHN_ID	0.0031±0.0027
14	20	R3M_MPHON_AVG_ARPU_AMT	0.0029±0.002
15	38	MEMBERSHIP_POINT	0.0029±0.0029
16	19	BPROD_NM	0.0024±0.0011
17	7	MPHON_ENGT_EXP_RPERD_ITG_CD	0.002±0.0018
18	16	ENGT_RMND_MONS_CD	0.0019±0.0013
19	29	R6M_AVG_ARPU_AMT	0.0016±0.0016
20	10	R3M_MPHON_AVG_ARPU	0.0014±0.0009

되었다는 점이다. 이처럼 특징 별로 예측 정확도에 얼마나 많은 영향을 미치는지 특징의 중요도를 구할 수 있어 향후 캠페인 데이터의 분석에 매우 용이하다.

6. 결 론

본 연구에서는 기계학습 기반으로 캠페인 시스템의 효과 제고를 위한 SOFS 특징 선택 알고리즘을 제안하였다. SOFS 알고리즘은 기존 SFFS 알고리즘이 순차적인 특징 평가로 인한 과도한 시간이 소요된다는 문제와 단일 분류기를 사용할 경우 발생하는 특징 선택의 정확도 향상의 한계점을 보완한 알고리즘이다. 또한, 특징 선택 결과로 특징 중요도를 함께 제공하기 때문에 분류 모델 생성 과정에서 추가로 활용할 수 있다는 장점을 가진다. 제안하는 알고리즘의 효용성을 검증하기 위하여 캠페인 시스템의 실제 수행 데이터를 대상으로 실험을 수행하였다. 그 결과 기존 방식과 비교했을 때 95% 이상의 탐색 시간 감소와 0.38%의 분류 정확도 향상을 확인할 수 있었다. 추가로 각기 다른 3개 분야(Keep & Care, Acquisition, Retention)의 캠페인 데이터 셋을 대상으로 특징 개수, 분류 정확도, 학습시간, 메모리 사용량 등을 측정한 결과 SOFS 알고리즘을 통해 최적화된 특징을 적용한 분류모델이 그렇지 않은 분류모델보다 더욱 나은 성능을 나타냄을 확인하였다.

결론적으로 제안하는 SOFS 특징 선택 알고리즘은 기계학습 기반 캠페인 시스템의 성능 및 분류예측의 활용에 긍정적인 영향을 미칠 수 있는 특징을 구분할 수 있는 기반이 되고 나아가 기계학습 기반 분류모델의 정확도를 향상하는데도 가능한 방식임을 확인하였다. 또한 캠페인 실무 관점에서는 대부분 기업이 캠페인 수행 시 활용하던 통계와 경험 기반 캠페인 수행으로 성공률이 높지 않던 부분을 기계학습을 활용하여 더 높은 캠페인 성공에 도움이 될 수 있어 기업에는 매출향상과 고객 관리(care) 활동과 유지(retention) 활동 등에 충분한 적용하여 활용할 만한 가치를 확인하였다. 더불어 실제 고객들이 캠페인에 반응하는 기준이 전통적으로 알고 있던 나이나 등급, 상품, 월 매출과 같은 일반적인 속성정보로 고정된 것이 아니라 캠페인의 유형이나 성격에 따라 다양한 속성들의 중요도는 서로 상호적으로 달라진다는 점도 확인할 수 있어 기획자들은 캠페인 유형에 맞는 중요 속성을 확인할 수 있게 된다. 궁극적으로 기업에서는 고객에 관리가 생존을 위한 핵심이며 많은 캠페인 비용을 들이지만 그 효과가 점점 더 낮아지고 있는 상황에서 반복적인 캠페인 기획 및 수행 비용을 낮추는데 도움이 될 수 있고 다양한 캠페인 성공률을 높일 수 있으며 고객입장에서는 여러 채널로 받던 캠페인들로 인한 피로를 줄일 수 있을 것으로 기대한다.

그러나 본 연구는 이런 실무적 의의에도 불구하고 다음과 같은 몇 가지 한계점을 또한 가지고 있다. 첫째, 캠페인의 예측을 위해 특히 영향도가 크다고 한 특징들이 왜 중요한지에 대한 설명을 하기에는 아직 부족하다는 점이다. 둘째, 지금은 전체 캠페인 유형들 중 적은 특징 및 고객 속성을 활용을 하였기 때문에 훨씬 많은 데이터 및 특징들을 기준으로 더 다양하게 테스트를 해 볼 필요가 있을 것이다. 특히 본 연구에서 사용된 설명변수들은 개인정보로 인해 실제 개인의 성향과 관련된 특징들은 제외하였는데 향후에는 포함한 테스트를 통해 캠페인에 실제 영향을 미치는 특징들에 대한 좀 더 많은 분석이 필요할 것으로 보인다.

참고문헌(References)

Bluma, A. L. and P. Langley, “Selection of relevant features and examples in machine,” *Artificial Intelligence*, Vol. 97, Nos. 1-2(1997), 245~271.

Breiman, L., "Bagging predictors," *Machine Learning*, Vol. 24, No. 2(1996), 123~140.

Breiman, L., "Random forests," *Machine Learning*, Vol. 45, No. 1(2001), 5~32.

Chandrashekar, G. and F. Sahin, “A survey on feature selection methods,” *Computers & Electrical Engineering*, Vol. 40, No. 1(2014), 16~28.

Cho, J., D. Lee, C. Song and M. Chun, “Feature Selection by Genetic Algorithm and Information Theory,” *Journal of Korean Institute of Intelligent Systems*, Vol. 18, No. 1(2008), 94~99.

Devijver, P. A. and J. Kittler, *Pattern Recognition : A Statistical Approach*, Vol. 761. London: Prentice-Hall, 1982.

Ferri, F. J., P. Pudil, M. Hatef, and J. Kittler, "Comparative study of techniques for large-scale feature selection," *Machine Intelligence and Pattern Recognition*, Vol. 16(1994), 403~413.

Guyon, I. and A. Elisseeff, "An Introduction to

Variable and Feature Selection," *Journal of Machine Learning Research*, Vol. 3(2003), 1157~1182.

Hong, S.-H., and K.-S. Shin, "Using GA based Input Selection Method for Artificial Neural Network Modeling Application to Bankruptcy Prediction", *Journal of Intelligence and Information Systems*, Vol. 9, No. 1 (2003), 227~249.

Kim, K.-J., and H.-C. Ahn, "Optimization of Support Vector Machines for Financial Forecasting", *Journal of Intelligence and Information Systems*, Vol. 17, No. 4 (2011), 241~254.

Kohavi, R. and G. H. John, "Wrappers for feature subset selection," *Artificial intelligence*, Vol. 97, Nos. 1-2(1997), 273~324.

Ladha, L. and T. Deepa, "Feature selection methods and algorithms," *International Journal on Computer Science and Engineering*, Vol. 1, No. 3(2011), 1787~1797.

Lee, C. H., "A Study on the Important Variable Selection Method by Feature Selection," Doctoral Dissertation, The University of Chung- Ang, Korea, 2007.

Lee, J., D. Park and C. Lee, " Feature Selection Algorithm for Intrusions Detection System using Sequential Forward Search and Random Forest Classifier," *KSII Transactions on Internet and Information Systems (TIIS)*, Vol. 11, No. 10(2017), 5132~5138.

Lee, J. S., and M. K. Jeong, "A Hybrid Feature Selection Method using Univariate Analysis and LVF Algorithm", *Journal of Intelligence and Information Systems*, Vol. 14, No. 4 (2008), 179~200.

Lee, W. and S. Oh, "Efficient Feature Selection Based Near Real-Time Hybrid Intrusion Detection System," *KIPS Transactions on Computer and Communication Systems*, Vol. 5, No. 12(2016), 471~480.

Mitchell, T. M., *Machine Learning*, McGraw Hill, 1997.

Molina, L. C., L. Belanche and A. Nebot, "Feature selection algorithms: A survey and experimental evaluation," *Proceedings of 2002 IEEE International Conference on Data Mining*, (2002), 306~313.

Oh, I. S., J. S. Lee, and B. R. Moon, "Hybrid genetic algorithms for feature selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 26, No. 11(2004), 1424~1437.

Ohn, S. Y., S. D. Chi and M. Y. Han, "Feature Selection for Classification of Mass Spectrometric Proteomic Data Using Random Forest," *Journal of the Korea Society for Simulation*, Vol. 22, No. 4(2013), 139~147.

Pudil, P., J. Novovičová and J. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, Vol. 15, No. 11(1994), 1119~1125.

Samuel, A. L., "Some studies in machine learning using the game of checkers," *IBM Journal of Research and Development*, Vol. 3, No. 3(1959), 210~229.

Yu, L. and H. Liu. "Feature selection for high-dimensional data: A fast correlation-based filter solution," *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, (2003), 856~863.

Zhou, Q., H. Zhou and T. Li, "Cost-sensitive feature selection using random forest: Selecting low-cost subsets of informative

features," *Knowledge-based Systems*, Vol. 95(2016), 1~11.

Abstract

Self-optimizing feature selection algorithm for enhancing campaign effectiveness

Jeoung-soo Seo* · Hyunchul Ahn**

For a long time, many studies have been conducted on predicting the success of campaigns for customers in academia, and prediction models applying various techniques are still being studied. Recently, as campaign channels have been expanded in various ways due to the rapid revitalization of online, various types of campaigns are being carried out by companies at a level that cannot be compared to the past. However, customers tend to perceive it as spam as the fatigue of campaigns due to duplicate exposure increases. Also, from a corporate standpoint, there is a problem that the effectiveness of the campaign itself is decreasing, such as increasing the cost of investing in the campaign, which leads to the low actual campaign success rate. Accordingly, various studies are ongoing to improve the effectiveness of the campaign in practice. This campaign system has the ultimate purpose to increase the success rate of various campaigns by collecting and analyzing various data related to customers and using them for campaigns. In particular, recent attempts to make various predictions related to the response of campaigns using machine learning have been made. It is very important to select appropriate features due to the various features of campaign data. If all of the input data are used in the process of classifying a large amount of data, it takes a lot of learning time as the classification class expands, so the minimum input data set must be extracted and used from the entire data. In addition, when a trained model is generated by using too many features, prediction accuracy may be degraded due to overfitting or correlation between features. Therefore, in order to improve accuracy, a feature selection technique that removes features close to noise should be applied, and feature selection is a necessary process in order to analyze a high-dimensional data set. Among the greedy algorithms, SFS (Sequential Forward Selection), SBS (Sequential Backward Selection), SFFS (Sequential Floating Forward Selection), etc. are widely used as traditional feature selection techniques. It is also true that if there are many risks and many features, there is a limitation

* ktds / Graduate School of Business IT, Kookmin University
** Corresponding author: Hyunchul Ahn
Graduate School of Business IT, Kookmin University
77 Jeongneung-ro, Seongbuk-gu, Seoul 02707, Republic of Korea
Tel: +82-2-910-4577, Fax: +82-2-910-4017, E-mail: hcahn@kookmin.ac.kr

in that the performance for classification prediction is poor and it takes a lot of learning time. Therefore, in this study, we propose an improved feature selection algorithm to enhance the effectiveness of the existing campaign. The purpose of this study is to improve the existing SFFS sequential method in the process of searching for feature subsets that are the basis for improving machine learning model performance using statistical characteristics of the data to be processed in the campaign system. Through this, features that have a lot of influence on performance are first derived, features that have a negative effect are removed, and then the sequential method is applied to increase the efficiency for search performance and to apply an improved algorithm to enable generalized prediction. Through this, it was confirmed that the proposed model showed better search and prediction performance than the traditional greed algorithm. Compared with the original data set, greed algorithm, genetic algorithm (GA), and recursive feature elimination (RFE), the campaign success prediction was higher. In addition, when performing campaign success prediction, the improved feature selection algorithm was found to be helpful in analyzing and interpreting the prediction results by providing the importance of the derived features. This is important features such as age, customer rating, and sales, which were previously known statistically. Unlike the previous campaign planners, features such as the combined product name, average 3-month data consumption rate, and the last 3-month wireless data usage were unexpectedly selected as important features for the campaign response, which they rarely used to select campaign targets. It was confirmed that base attributes can also be very important features depending on the type of campaign. Through this, it is possible to analyze and understand the important characteristics of each campaign type.

Key Words : Feature selection, Artificial Intelligence-based Campaign system, Greedy Algorithm, Campaign prediction, Machine learning

Received : November 19, 2020 Revised : December 26, 2020 Accepted : December 28, 2020
Corresponding Author : Hyunchul Ahn

저 자 소 개

서 정 수
현재 ktds CRM사업팀장으로 재직 중이다. 창원대학교 미생물학과를 졸업하고, 국민대학교 경영대학원에서 경영학석사 학위를 취득하였다. 국민대학교 비즈니스IT전문대학원 박사과정을 수료했고, 관심분야는 고객관계관리 분야의 AICC(AI Contact Center), Data preprocessing, machine vision 등이다.

안 현 철
현재 국민대학교 비즈니스IT전문대학원 교수로 재직 중이다. KAIST에서 산업경영학사를 취득하고, KAIST 테크노경영대학원에서 경영정보시스템을 전공하여 공학석사와 박사를 취득하였다. 주요 관심 분야는 금융 및 고객관계관리 분야의 인공지능 응용, 정보시스템 수용과 관련한 행동 모형 등이다.