ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2016 March: 22(1): 139~157 http://dx.doi.org/10.13088/jiis.2016.22.1.139

부도예측을 위한 KNN 앙상블 모형의 동시 최적화

민성환
한림대학교 경영학부
(shmin@hallym.ac.kr)

앙상블 분류기란 개별 분류기보다 더 좋은 성과를 내기 위해 다수의 분류기를 결합하는 것을 의미한다. 이와 같은 앙상블 분류기는 단일 분류기의 일반화 성능을 향상시키는데 매우 유용한 것으로 알려져 있다. 랜덤 서브스페이스 앙상블 기법은 각각의 기저 분류기들을 위해 원 입력 변수 집합으로부터 랜덤하게 입력 변수 집합을 선택하며 이를 통해 기저 분류기들을 다양화 시키는 기법이다. k-최근접 이웃(KNN: k nearest neighbor)을 기저 분류기로 하는 랜덤 서브스페이스 앙상블 모형의 성과는 단일 모형의 성과를 개선시키는 데 효과적인 것으로 알려져 있으며, 이와 같은 랜덤 서브스페이스 앙상블의 성과는 각 기저 분류기를 위해 랜덤하게 선택된 입력 변수 집합과 KNN의 파라미터 k의 값이 중요한 영향을 미친다. 하지만, 단일 모형을 위한 k의 최적 선택이나 단일 모형을 위한 입력 변수 집합의 최적 선택에 관한 연구는 있었지만 KNN을 기저 분류기로 하는 앙상블 모형에서 이들의 최적화와 관련된 연구는 없는 것이 현실이다. 이에 본 연구에서는 KNN을 기저 분류기로 하는 앙상블 모형의 성과 개선을 위해 각 기저 분류기들의 k 파라미터 값과 입력 변수 집합을 동시에 최적화하는 새로운 형태의 앙상블 모형을 제안하였다. 본 논문에서 제안한 방법은 앙상블을 구성하게 될 각각의 KNN 기저 분류기들에 대해 최적의 앙상블 성과가 나올 수 있도록 각각의 기저 분류기가 사용할 파라미터 k의 값과 입력 변수를 유전자 알고리즘을 이용해 탐색하였다. 제안한 모형의 검증을 위해 국내 기업의 부도 예측 관련 데이터를 가지고 다양한 실험을 하였으며, 실험 결과 제안한 모형이 기존의 앙상블 모형보다 기저 분류기의 다양화와 예측 성과 개선에 효과적임을 알 수 있었다.

주제어 : 랜덤 서브스페이스, 부도 예측, 앙상블, 유전자 알고리즘

논문접수일 : 2016년 3월 2일 논문수정일 : 2016년 3월 14일 게재확정일 : 2016년 3월 14일
원고유형 : 일반논문 교신저자 : 민성환

1. 서론

앙상블 분류기란 개별 분류기보다 더 좋은 성과를 내기 위해 다수의 분류기를 결합한 분류기를 의미한다. 이와 같은 앙상블 분류기는 단일 분류기의 일반화 성능을 향상시키는데 매우 유용한 것으로 알려져 있다(Dietterich, 1997). 그러나, 앙상블 모형의 일반화 성능을 향상시키기 위해서는 앙상블을 구성하고 있는 기저 분류기들의 예측 성과가 좋아야 할 뿐만 아니라 다양성이 존재해야 한다. 특히 기저 분류기의 다양성은 앙상블의 성과에 중요한 영향을 미친다 (Kuncheva and Whitaker, 2003; Bian and Wang, 2007). 만약 기저 분류기의 다양성이 존재하지 않는 다면 아무리 많은 기저 분류기들을 결합하더라도 단일 모형보다 성과 개선을 기대하기 어렵다. 예를 들어, 동일한 기저 분류기를 결합할 경우에는 아무리 많은 수의 분류기를 결합하더라도 앙상블 모형의 예측 성과는 단일 분류기와 차이가 없을 것이다. 이와 같은 이유로 대다수의 앙상블 기법은 기저 분류기를 다양화 시키는 데에 초점이 맞춰져 있다. 다양성이 있는 기저분류기를 발생시키

기 위한 방법으로는 학습 데이터의 다양화, 학습 알고리즘의 다양화, 학습 알고리즘 파라미터의 다양화 등의 방법이 있다. 가장 대표적인 앙상블 기법으로 많은 응용 분야에서 성공적으로 활용되고 있는 배깅, 부스팅, 랜덤 서브스페이스 기법은 학습 데이터의 다양화를 통해서 기저 분류기를 다양화 시키는 대표적인 방법이다. 배깅은 bootstrap aggregating의 약자로 학습 데이터의 다양화를 통해 기저분류기를 다양화시키는 대표적인 앙상블 기법 중의 하나이다 (Breiman, 1996). 배깅 기법은 원 학습데이터로부터 부트스트랩 샘플링(bootstrap sampling) 방식에 의해 다수의 서로 다른 학습 데이터 셋을 발생시키고, 이들 각각의 서로 다른 학습 데이터를 가지고 기저 분류기를 학습시키게 된다. 이와 같은 방식에 의해 다양성이 존재하는 다수의 기저 분류기가 생성되게 된다. 랜덤 서브스페이스 앙상블 기법은 각각의 기저 분류기들을 위해 원 입력 변수 집합(feature set)으로부터 랜덤하게 입력 변수 집합을 선택하며 이를 통해 기저 분류기들을 다양화 시키는 기법이다(Ho, 1998).

k-최근접 이웃(KNN:k nearest neighbor) 모형은 데이터의 변화에 강건한 성격이 있지만 입력 변수의 변화에는 매우 민감한 것으로 알려져 있다. 이 때문에 KNN은 랜덤 샘플링을 통해 데이터에 변화를 주는 앙상블 기법인 배깅 기법의 기저 분류기로는 부적합하지만, 입력 변수의 변화를 통해 기저 분류기를 다양화시키는 기법인 랜덤 서브스페이스 앙상블 모형의 좋은 기저 분류기가 될 수 가 있다. 이로 인해 KNN 랜덤 서브스페이스 앙상블 모형은 단일 모형보다 성과 개선의 효과가 큰 것으로 알려져 있다. 이와 같은 KNN 랜덤 서브스페이스 앙상블 모형의 성과는 각 기저 분류기들을 위해 어떠한 입력 변수 집합이 선택되었는지에 크게 영향을 받는다. 또한, KNN의 파라미터인 k의 값도 앙상블 모형의 성과에 중요한 영향을 준다고 볼 수 있다. 하지만, KNN 단일 모형을 위한 k의 최적 선택이나 단일 모형을 위한 입력 변수 집합의 최적 선택에 관한 연구는 있었지만 KNN을 기저 분류기로 하는 앙상블 모형에서 이들의 최적화와 관련된 연구는 없는 것이 현실이다. 이에 본 연구에서는 KNN을 기저 분류기로 하는 앙상블 모형의 성과 개선을 위해 각 기저 분류기들의 k 파라미터 값과 입력 변수 집합을 동시에 최적화하는 새로운 형태의 앙상블 모형을 제안하였다. 또한 제안한 모형의 검증을 위해 국내 부도 기업 관련 문제에 적용해 다양한 분석을 하였다.

본 논문은 다음과 같이 구성된다. 제 2장에서는 부도 예측 관련 연구에 대한 설명을 하고 제 3장에서는 본 논문에서 제안한 앙상블 모형에 대해 설명을 하였다. 제 4장에서는 연구 데이터와 실험 설계에 대한 설명을 하였으며, 제 5장에서는 실험 결과와 이에 대한 분석 및 해석을 하였다. 끝으로 제 6장에서는 연구의 결론에 대해 설명하였다.

2. 부도 예측 관련 연구

부도 예측은 오랜 기간 동안 연구되어 온 중요한 주제로 부도와 관련된 비용은 매우 크기 때문에 국가 경제에 중요한 영향을 주는 것으로 알려져 있다. 과거 수 십 년 동안 많은 학자들이 부도 예측과 관련된 주제를 연구해오고 있으며, 초창기의 부도 예측 관련 연구는 주로 단일 변량 분석, 다변량 판별분석, 다중 회귀 분석, 로지스틱 회귀 분석과 같은 전통적인 통계를 활용하는 연구가 주를 이루었다 (Beaver, 1966; Altman, 1968; Meyer and Pifer, 1970; Ohlson, 1980). 그 뒤로,

사례기반추론, 귀납적 학습, 인공 신경망 등과 같은 인공지능 기법들이 부도 예측 문제에 적용되기 시작하였다 (Buta, 1994; Messier and Hansen, 1998; Tam and Kiang, 1992; Zhang et al., 1999).

최근에는 부도 예측 모형의 성능을 개선시키기 위해 앙상블 모형을 활용하려고 하는 연구가 시작되고 있다. 앙상블 분류기란 개별 분류기보다 더 좋은 성과를 내기 위해 다수의 분류기를 결합한 분류기를 의미한다. 이와 같은 앙상블 분류기는 단일 분류기의 일반화 성능을 향상시키는데 매우 유용한 것으로 알려져 있다. 이와 같은 앙상블 기법은 최근에 부도 문제에 성공적으로 적용되고 있다.

Kim and Kim (2007)은 SOHO의 부도 예측 문제에 선택적 배깅 앙상블 모형을 적용해 보았다. 기존의 일반 배깅 모형은 발생한 모든 기저 분류기들을 결합하여 앙상블 모형을 구성하지만 본 논문에서 제안한 모형에서는 성능 평가용 데이터에서 평균 이상의 예측 정확도를 가지는 분류기만을 선택하여 앙상블 모형을 구성하였다. 이와 같이 구성한 선택적 배깅 앙상블 모형은 기존의 배깅 앙상블 모형보다 높은 예측 정확도를 보였다. Tsai and Wu (2008)는 부도 예측 모형에서 다수의 신경망 모형을 사용할 경우의 효과성에 대한 연구를 수행하였다. Kim (2009)은 기업 부도 데이터의 불균형 문제 해결을 위해 기하 평균 개념을 부스팅 알고리즘에 적용한 GM-Boost 알고리즘을 적용해 보았으며, 실험 결과 제안한 모형이 기존의 다른 기법에 비해 데이터 불균형 정도에 관계없이 우수한 예측 성과를 보였다. Nanni and Lumini (2009) 는 다양한 형태의 앙상블 모형을 부도 예측과 신용 평가 문제에 적용해 보았으며 실험 결과 Levenberg-Marquardt 신경망을 기저 분류기로 사용하는 랜덤 서브스페이스 앙상블 모형이 가장 좋은 성과를 보였다. Hung and Chen (2009)은 기대확률(expected probability)에 기반을 둔 선택적 앙상블 모형을 부도 예측 문제에 적용해 보았으며 제안한 모형이 기존의 앙상블 모형보다 우수한 예측 성과를 보였다. Wang and Ma (2009)는 대표적인 앙상블 기법인 배깅과 랜덤 서브스페이스 앙상블 기법을 결합하는 혼합 앙상블 모형을 재무 부실화 문제에 적용해 보았으며, 제안한 모형이 기존의 단일 앙상블 모형보다 좋은 성과를 낸다는 것을 실증적으로 보였다. Kim (2010)은 기업의 부도 예측 문제에 배깅과 부스팅 앙상블 모형을 적용해 보았다. 또한, 이들 앙상블 모형의 성과 개선을 위해 유전자 알고리즘을 적용한 모형을 제안하였으며 제안한 모형이 기존의 앙상블 모형보다 예측 성과가 개선되었다는 것을 보여주었다. Louzada et al. (2011)는 앙상블을 구성하는 기저 분류기들을 연속적인 샘플링 과정의 반복을 통해 결합하는 배깅 변형 모형을 제안하였으며 제안한 모형의 효과성을 검증하기 위해 재무 부실화 문제에 적용해 보았다. Li et al. (2011)는 로지스틱 회귀 모형을 기저 분류기로 하는 랜덤 서브스페이스 앙상블 모형을 부도 예측 문제에 적용하여 전통적인 개별 분류 모형과 예측 성과를 비교하여 앙상블 모형의 우수성을 보였다. Min (2014)은 기존의 배깅 앙상블 모형의 성능을 개선하기 위해 유전자 알고리즘을 이용한 사례 선택 기법과 배깅의 결합 모형을 제안하였다. 본 논문에서는 제안한 모형을 국내 부도 예측 문제에 적용해 보았으며 제안한 모형이 기존의 모형보다 우수한 성과를 보였다. Abellan and Mantas (2014) 는 다양한 앙상블 모형을 부도 예측과 신용 평가 문제에 적용해 보았으며 실험 결과 크리덜 의사결정 트리(creedal decision tree)를 기저 분류기로 사용하는 배깅 앙상블 모형이 가장 좋은 예측 성과를 보였다.

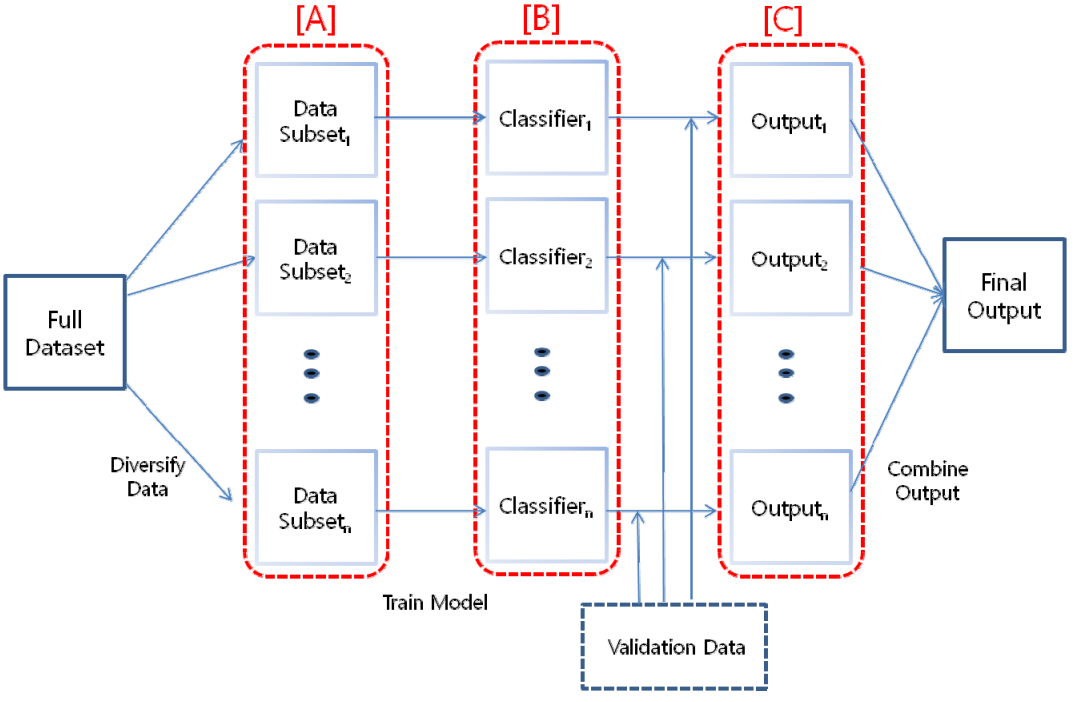
3. 연구 모형

3.1 기존 연구와의 차별성

앙상블 모형은 개별 분류기를 사용한 것보다 좋은 예측 성과를 얻기 위해 다수의 분류기를 결합하여 사용하는 것으로, 이와 같은 앙상블 기법은 단일 분류기의 일반화 성능을 향상시키는 데에 매우 유용한 것으로 알려져 있다. 앙상블 분류기의 일반화 성능을 향상시키기 위해서는 앙상블을 구성하고 있는 기저 분류기들의 예측 정확도가 가능하면 좋아야 하며, 이들 기저 분류기들이 가능하면 다양성을 가져야 한다.

앙상블 분류기는 다수의 기저 분류기들로 구성되어 있으며, 이들 기저 분류기간의 다양성이 앙상블 분류기의 성과에 중요한 역할은 한다. 앙상블을 구성하게 될 기저 분류기들을 다양화하는 방법으로는 학습데이터의 다양화, 학습 알고리즘의 다양화, 학습 알고리즘 파라미터의 다양화 방법 등이 있다. <Figure 1>은 학습 데이터의 다양화를 통한 앙상블 모형의 전반적인 모습을 보여주고 있다. <Figure 1>에서 (A)부분은 원 학습데이터로부터 일부만을 선택하여 서로 다른 다수의 학습 데이터 집합을 구성하는 데이터 다양화를 나타내고 있다. 대표적인 앙상블 기법인 배깅, 부스팅, 랜덤 서브스페이스 앙상블 기법은 이와 같이 데이터의 다양화를 통해 기저 분류기를 다양화시키는 앙상블 기법들이다.

이와 같은 방법에 의해 데이터 셋이 다양화되면 이들 각각의 서로 다른 데이터 셋을 이용해 각각의 분류기를 학습시키고 이를 통해 (B)와 같이 다양성이 존재하는 기저 분류기들의 집합 $\{C_1,\cdots,C_n\}$이 생성되게 된다. 또한, 이들 기저 분



〈Figure 1〉Ensemble Model using Diversified Data Set

류기들 각각의 예측 결과 값 $\{O_1,\cdots,O_n\}$는 다수결 투표와 같은 결합 방식에 의해 합쳐지게 된다. 이와 같이 하나의 문제를 해결하기 위해 다수의 분류기를 이용하는 앙상블 기법은 분류기의 예측 성능을 개선시킬 수 있다는 장점이 있는 것으로 알려져 있으며, 이로 인해 최근에 다양한 연구가 활발하게 진행되고 있다. 초기의 앙상블 모형 관련 연구는 기본 앙상블 모형을 다양한 문제에 적용하여 기존의 단일 모형과 비교해 보는 연구가 대부분이었으며, 많은 응용 분야에서 앙상블 모형이 기존의 단일 모형보다 좋은 예측 결과를 보였다. 그 뒤로 앙상블 모형의 예측 성과를 개선시키기 위한 다양한 연구가 시도되었다. Marques et al. (2012)과 Min(2012)과 같은 연구자들은 기존의 데이터 다양화 기법 중에 두 개의 기법을 결합하여 앙상블 성능을 개선하려는 연구를 수행하였다. 또한, (B)부분에서 생성된 모든 기저 분류기들을 결합하여 앙상블 모형을 구성하는 것이 아니고, 일부만을 선택하여 앙상블을 구성하는 선택적 앙상블 모형에 관한 연구도 최근에 활발하게 진행되고 있다 (Hung and Chen, 2009; Li et al., 2012). 또한 (C) 부분의 각 기저 분류기의 예측 결과값들을 어떻게 결합할 지에 관한 많은 연구도 있었다(Alexandre et al., 2001; Kuncheva et al., 2001; Mandler and Schurmann, 1998). Ho(2002)에 따르면 앙상블 분류기의 성과를 최적화하는 방법은 기저 분류기 풀(classifier pool) $\{C_1,\cdots,C_n\}$로 부터 최적의 분류기를 선택하는 범위 최적화(coverage optimization)와 각 분류기들의 출력정보(output) $\{O_1,\cdots,O_n\}$를 결합하기 위한 규칙을 최적화 하는 결정 최적화(decision optimization)로 분류될 수 있으며 그림에서 살펴보면 범위 최적화는 [B] 부분과 관련된 것이고 결정 최적화는 [C]부분과 관련된 것으로 앙상블 모형의 성과 개선을 위해 이와 관련된 많은 연구가 있었다. 하지만, [A]부분의 최적화와 관련된 연구는 거의 없는 것이 현실이다. 이에 본 논문에서는 [A]부분의 최적화를 통해 앙상블 모형의 성능을 개선하는 방법을 제안하고자 한다. 즉, 본 논문에서 제안한 앙상블 모형 개선 방법은 주어진 다양화 방법을 통해서 결정된 기저 분류기들을 최적화하거나 이들 기저 분류기 결과값들을 최적화하는 것이 아닌 앙상블 모형의 성과를 개선시킬 수 있는 다양화 방법을 제안하였으며, 이는 기존의 연구와 차별성이 있다고 볼 수 있다.

일반적인 랜덤 서브스페이스 모형의 경우 입력 변수 집합 $\{feat_1,\cdots,feat_m\}$에서 임의로 f개 만큼 선택하여 새로운 입력 변수 집합을 생성하고, 이를 이용해 기저 분류기를 학습시킴으로써 기저 분류기를 다양화시키게 된다. 이때 입력 변수를 임의로 선택하여 입력 변수 집합을 구성하기 때문에 서로 다른 입력 변수 집합이 생성되고 이들로 인해 다양성이 존재하는 기저 분류기들이 생성되고 앙상블 모형의 성과 개선이 이뤄지게 된다. 본 논문에서는 이와 같은 일반 랜덤 서브스페이스 앙상블 모형을 수정하여 앙상블 모형의 성과 측면에서 최적인 각 기저 분류기별 입력 변수 집합을 유전자 알고리즘을 이용하여 구하는 방법을 제안하였다. 뿐만 아니라 각 기저 분류기의 KNN 파라미터인 k 또한 유전자 알고리즘을 이용하여 최적화하게 된다.

3.2 제안 모형

본 연구는 부도 예측을 위한 KNN 앙상블 모형의 성능 개선에 관한 연구이다. 본 논문에서는 KNN 모형의 성능 개선을 위해 앙상블 모형을 사용하였으며, KNN 앙상블 모형의 추가 성능 개선을 위해 KNN 기저 분류기의 파라미터 k와 입력변수 집합을 최적화하는 방법을 제안하였다.

The Proposed Model

1. Setup:
   -Partition Data Set: T_1, T_2, V
   -Define the chromosome
   -Determine parameters of GA
2. Generate the initial population
3. Select k parameters and the feature subsets for each chromosome
4. Generate base classifiers using the k parameters and the feature subsets generated in step 3 (Different classifiers are generated ➡ $\{C_1, .., C_n\}$
5. Apply the base classifiers to T_2 ➡ $\{O_1, \ldots, O_n\}$
6. Combine the output data by a combining method and Calculate the fitness values
7. Perform GA operations and create a new generation
8. Repeat from step 3 to 7 until the termination criteria are satisfied
9. Select the optimal k and optimal feature subsets
10. Generate base classifiers using the optimal values ➡ Optimal Classifiers Set: $\{C^*_1, \ldots, C^*_n\}$
11. Apply the base classifiers to the validation data set(V) ➡ Output Set :$\{O^*_1, \ldots, O^*_n\}$
12. Combine the output data by a combining method

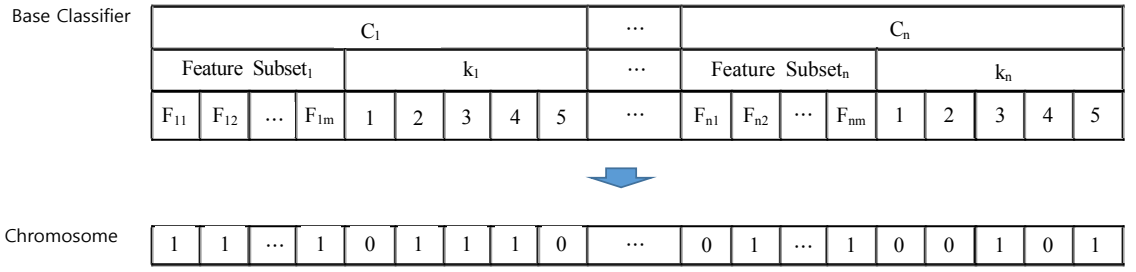
〈Figure 2〉 Steps of the Proposed Model

본 논문에서 제안한 방법은 앙상블을 구성하게 될 각각의 KNN 기저 분류기들에 대해 최적의 앙상블 모형 성과가 나올 수 있도록 각각의 k의 값과 입력 변수를 유전자 알고리즘을 이용해 탐색하였다. 본 논문에서 제안한 방법은 앙상블을 구성하게 될 기저 분류기 발생을 최적화하는 효과를 가져올 것으로 기대되며, 이를 통해 KNN 앙상블 모형의 성과가 개선될 것으로 기대된다. 본 논문에서 제안한 모형의 절차는 <Figure 2>와 같으며 자세한 설명은 다음과 같다.

제안한 모형을 실행하기 위한 준비 단계에서는 먼저 데이터를 분할하고 유전자 알고리즘을 위한 염색체(chromosomes)를 정의하고, 유전자 알고리즘을 위한 파라미터를 설정한다. 본 연구에서는 전체 데이터를 학습용 데이터(T)와 검증용 데이터(V)로 분할하여 사용하였으며 학습용 데이터는 다시 모형 구축을 위한 데이터(T_1)와 유전자 알고리즘에서 적합도 함수 계산을 위해 사용한 데이터(T_2)로 구분하여 사용하여 과적합을 피하려고 노력하였다.

본 논문에서는 KNN 앙상블 모형에서의 최적의 k 값과 최적의 입력 변수 집합을 탐색하기 위해 유전자 알고리즘을 사용하였다. 유전자 알고리즘은 자연 선택과 진화 이론에 기초를 둔 효율적인 최적화 절차를 제공하는 기법이다. 유전자 알고리즘은 적자 생존의 기본 개념을 바탕으로 두고 있어서 단계적으로 계속해서 문제에 대해 더 좋은 해를 선택하는 과정을 거치게 된다(Goldberg, 1989). 유전자 알고리즘은 문제에 대한 하나의 해에 점진적인 변화를 줌으로써 최적의 해를 찾아나가기 보다는 모집단이라는 다수의 해 집합을 유지하면서 더 좋은 해들을 계속해서 생성시키는 방식으로 최적의 해를 탐색하는 방식을 이용한다. 유전자 알고리즘에서 처음에는 다수의 초기해 집합인 모집단을 랜덤하게 발생시키고 선택, 교배, 돌연변이 등과 같은 유전자 연산을 통해 초기의 해집합은 계속해서 다음 세대로 진화를 하게 된다.

본 논문에서는 앙상블 모형의 성과 측면에서 최적의 k 값과 최적의 입력 변수 집합을 선택하기 위해서 유전자 알고리즘을 활용하였으며, 이와 같은 최적화 문제에 유전자 알고리즘을 적용



〈Figure 3〉 An Example of Chromosome

하기 위해서는 문제에 맞게 염색체를 암호화 필요가 있다.

본 논문에서 염색체는 각 기저 분류기에서 사용하게 될 각각의 k 파라미터와 입력 변수 집합을 나타내도록 이진열(binary string) 형태로 설계되었으며 <Figure 3>은 이와 같은 염색체 구조의 예를 보여주고 있다. 그림에서와 같이 각각의 기저 분류기(C_i)를 생성시키기 위해서는 각각의 기저 분류기별로 입력변수 집합(Feature $Subset_i$)과 k의 값(k_i)이 필요하게 된다. Feature Subset에 해당하는 이진열의 각각의 비트는 해당되는 입력변수의 선택유무를 나타내게 된다. 즉, F_{ij}에 해당하는 값이 의미하는 것은 i 번째 기저 분류기에서 j번째 입력 변수의 선택 유무를 의미한다. 예를 들어 그림에서 F_{12}에 해당하는 비트의 값은 0이며 이것이 의미하는 것은 첫 번째 기저 분류기에서는 입력 변수 집합에서 두 번째 입력변수를 사용하지 않는다는 것을 의미한다. k_i에 해당하는 비트는 i번째 기저 분류기로 사용하게 될 KNN 분류기에서의 파라미터 k의 값을 의미한다. 그림에서와 같이 각 분류기별로 k의 값은 5비트로 설계하였으며 이것이 나타내는 수는 0부터 31(2^5-1)까지의 수를 나타내며 k의 값은 1부터 32까지 대응하도록 하였다. 예를 들면 <Figure 3>에서 k_1에 해당하는 비트는 (01110)로 이 값은 이진수 값을 의미하며 10진수로 변환할 경우 $0\times2^0+1\times2^1+1\times2^2+1\times2^3+0\times2^4=14$로 변환할 수 있으며 해당되는 k의 값은 여기에 1을 더해 15가 되도록 설계하였다. 원 학습데이터가 m개의 입력변수로 구성되어 있고, 앙상블을 구성하게 될 기저 분류기의 수를 n개로 한다면 총 $(n\times m+5\times n)$의 비트로 구성된 염색체가 필요하게 된다.

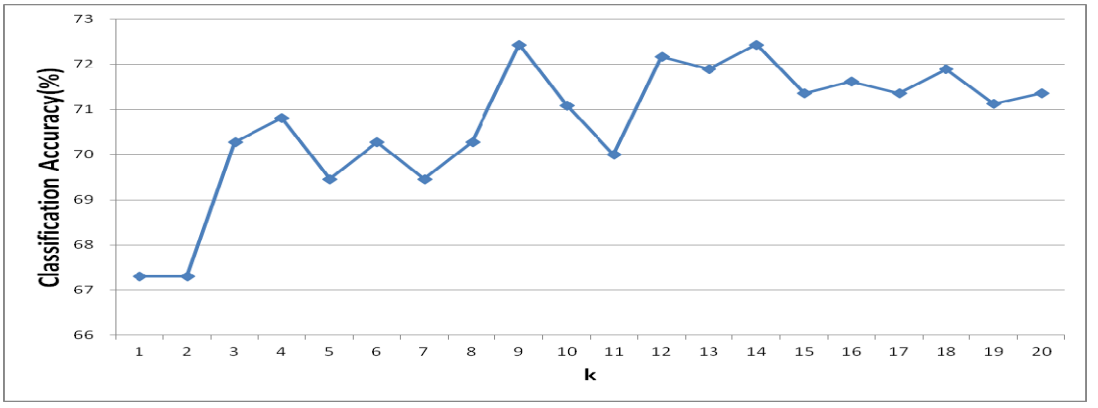
유전자 알고리즘은 랜덤하게 모집단을 발생시킴으로써 시작하게 된다. 모집단은 다수의 염색체로 구성되어 있으며 랜덤하게 발생한 각각의 염색체는 살펴본 바와 같이 앙상블 모형을 구성하게 될 각 KNN 기저분류기의 파라미터 k 값과 입력 변수 집합을 의미하게 되며 이들 값들을 이용해 각각의 기저 분류기 $\{C_1,\cdots,C_n\}$를 발생시키고, 이들 각각에 대해 T_2데이터 셋에서의 예측 결과값 $\{O_1,\cdots,O_n\}$을 구하고 이들을 다수결투표 방식을 이용해 결합하여 앙상블 모형의 예측 결과값을 구한다. 이와 같이 구한 앙상블 모형의 예측 정확도는 해당되는 염색체에 대한 적합도 값으로 사용하였으며, 이 값은 재생, 교배, 돌연변이 등의 유전자 연산을 통한 다음 세대로 진화시 중요한 정보로 활용되게 된다.

4. 실험 설계

본 논문에서는 KNN 모형의 성능 개선을 위해 새로운 형태의 앙상블 모형을 제안하였으며 제

〈Table 1〉 Input Variables

Category	Description
Profitability	EBITDA to Sales
	Financial Expenses to Sales
	Financial Expenses to Debt
	Ordinary Income Rate
	Ordinary Income to Sales
	Net Income to Sales
	Interest Expenses to Net income
	Ordinary Income to Capital
	Ordinary Income to Total Asset
Stability	Fixed Asset to Owner's Equity
	Quick Asset to Current Liability
	Debt Ratio
	Current Liability to Total Asset
	Current Asset to Current Liability
	(Capital Surplus + Retained Earnings- Dividend)/Total Assets
	Borrowings to EBITDA
	Borrowings to Sales
	Cash Ratio
Growth	Coefficient of Variation of Sales
Cash Flow	Cash Flow after Interest Payment to Sales
	Cash Flow to Financial Expenses
Activity	Sales to Net Change in Working Capital
	Total Assets Turnover Period
	Sales to Net Change in Account Receivable



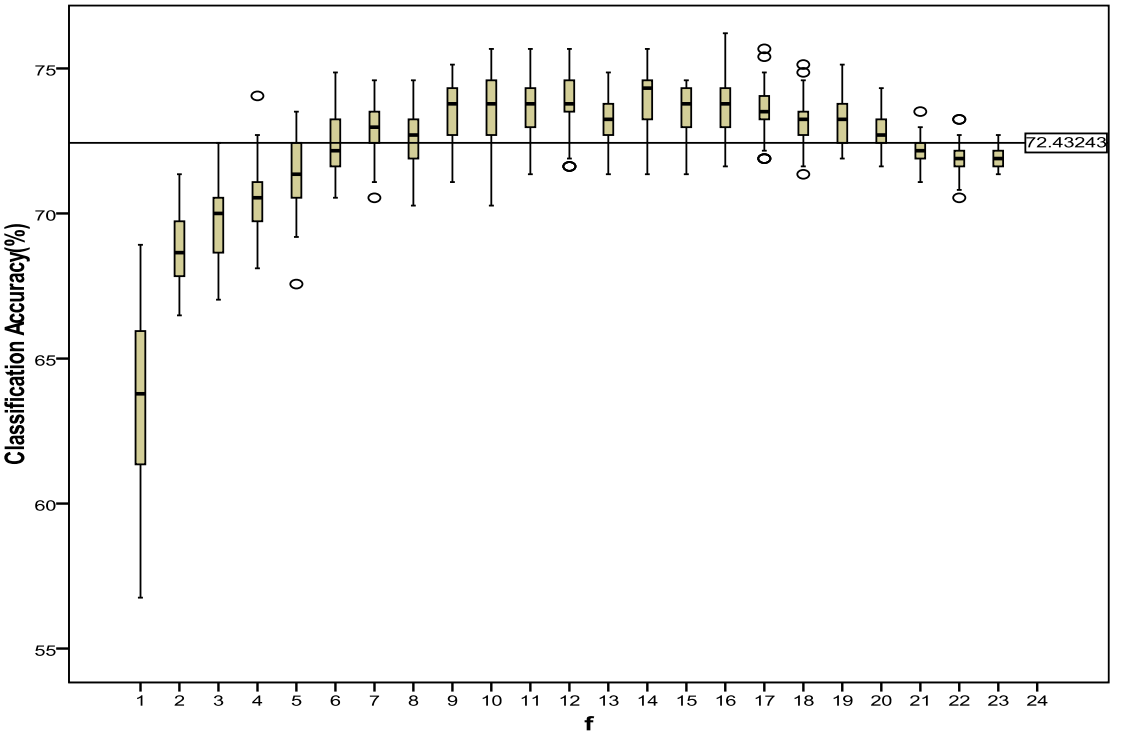
〈Figure 4〉 Classification Accuracy of Single KNN Model at Different k values (%)

안한 모형의 효과성을 검증하기 위해 국내 기업의 부도 예측 문제에 적용해 보았다. 모형의 검증을 위해 사용한 데이터는 총 1800개의 국내 비외감 기업의 데이터로 구성되어 있으며 부도 기업과 비부도 기업이 각각 900개로 구성되어 있다. 기업의 부도를 예측하기 위한 입력 변수로는 수익성, 안정성, 성장성, 활동성 및 현금 흐름으로 분류된 재무비율을 사용하였다. 실험에 사용한 최종 입력 변수는 단일 표본 t검정(independent-samples t-test)과 후진 선택법(backward selection)을 이용한 로지스틱 회귀분석을 이용하여 최종 선정 하였으며 그 결과는 <Table 1>과 같다.

데이터는 학습용 데이터와(T)와 모형 검증을 위한 검증용 데이터(V)로 분할하여 실험을 수행하였으며 학습용 데이터는 다시 모형 구축을 위한 데이터(T_1)와 유전자 알고리즘에서 적합도 계산을 위한 데이터(T_2)로 분류하였다. 10-겹 검증(10-fold cross validation) 방법으로 실험을 수행하였으며 앙상블을 구성하고 있는 기저 분류기의 수는 25로 고정하고 실험을 하였다. 앙상블 모형의 경우 10회 반복하여 실험한 결과의 평균값을 대푯값으로 사용하였다.

5. 실험 결과

본 논문은 KNN을 기저 분류기로 하는 앙상블 모형의 성능 개선에 관한 연구이다. 단일 KNN 분류기의 성과는 파라미터 k에 큰 영향을 받으며 KNN 랜덤 서브스페이스 앙상블 모형의 성과는 파라미터 k와 선택된 입력 변수 집합에 큰 영향을 받는다. <Figure 4>는 k에 따른 단일 KNN 분류기의 예측 성과 변화를 보여주고 있다. 그림

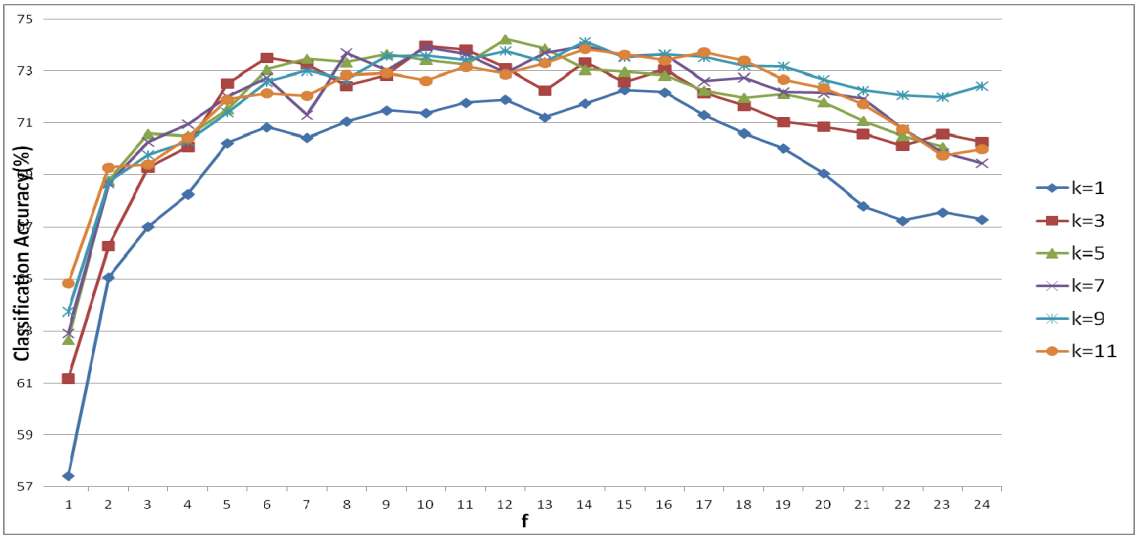


〈Figure 5〉 Box plots of the classification accuracy of KNN ensemble model (%)

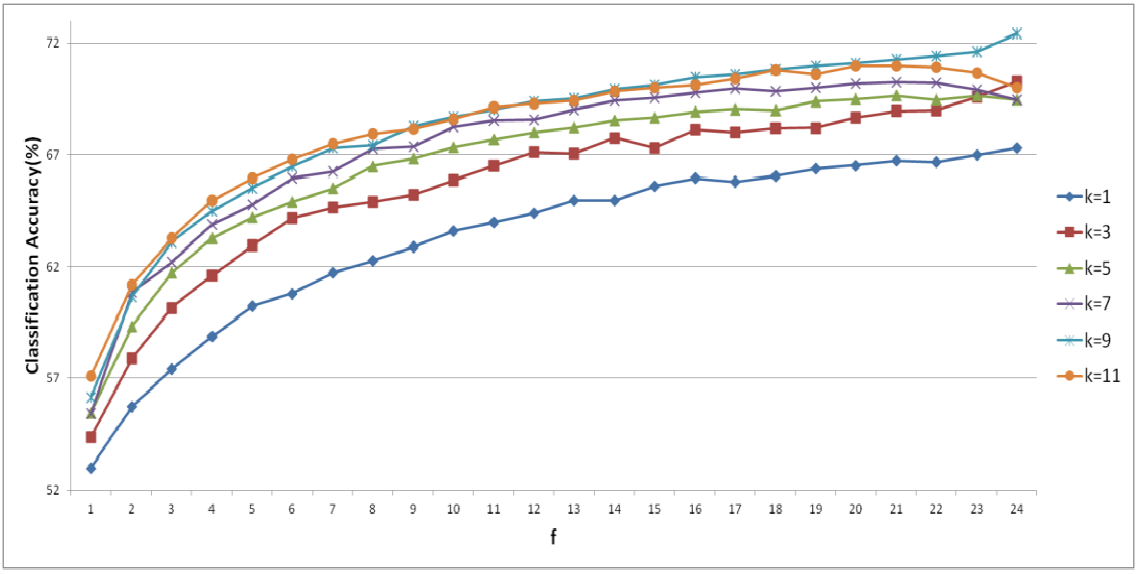
에서 보는 바와 같이 KNN 분류기의 성과는 파라미터 k에 크게 영향을 받는 것을 알 수 있다. 본 실험에서는 k=9 일 때 가장 좋은 예측 성과를 보였다.

<Figure 5>는 단일 모형에서 가장 좋은 성과를 보인 k의 값을 선택하여 기저 분류기로 사용한 랜덤 서브스페이스 앙상블 모형의 성과를 상자그림(box plot)으로 보여주고 있다. 그림에서 f는 전체 입력 변수 중에서 랜덤하게 선택된 입력 변수의 수를 의미한다. 그림에서 f=12가 의미하는 것은 24개의 전체 입력 변수 중에서 랜덤하게 12개의 입력 변수를 선택하여 입력 변수 집합 Feature $Subset_i$를 생성하고 이것을 25번 반복하여 25개의 서로 다른 입력 변수 집합{Feature $Subset_1$, ..., Feature $Subset_{25}$}를 생성시킨 후 이를 이용하여 서로 다른 KNN 분류기를 생성하여 앙상블 모형을 구성한다는 것을 의미한다. 그림에서 수평선은 단일 KNN 분류기의 예측률을 의미한다. 그림에서 보는 바와 같이 KNN 앙상블의 성과는 f의 값이 작을 때는 단일 모형보다 좋지 않다가 f가 증가함에 따라 단일 모형보다 좋은 예측률을 보이는 것을 알 수 있다. 앙상블 모형의 예측률은 f의 값이 전체 입력 변수의 0.5에 해당하는 12 부근에서 가장 좋은 값을 보이다가 점차로 감소하는 것을 알 수 있으며 이는 선행 연구 결과와 일관된 결과이다(Ho, 2002). <Figure 5>의 실험 결과에서 알 수 있듯이 랜덤 서브스페이스 앙상블 모형의 성과는 f의 값에 따라 차이가 날 뿐만 아니라 f의 값이 같더라도 어떤 입력 변수 군이 선택되었냐에 따라 편차를 보임을 알 수 있다. 이런 점에 착안하여 본 연구에서는 앙상블 모형의 성과를 최대로 할 수 있는 입력 변수 군을 유전자 알고리즘을 이용해 탐색하여 사용한 새로운 형태의 앙상블 모형을 제안하였으며 이는 기존의 앙상블 모형의 성과 개선에 큰 기여를 할 것으로 기대 된다.

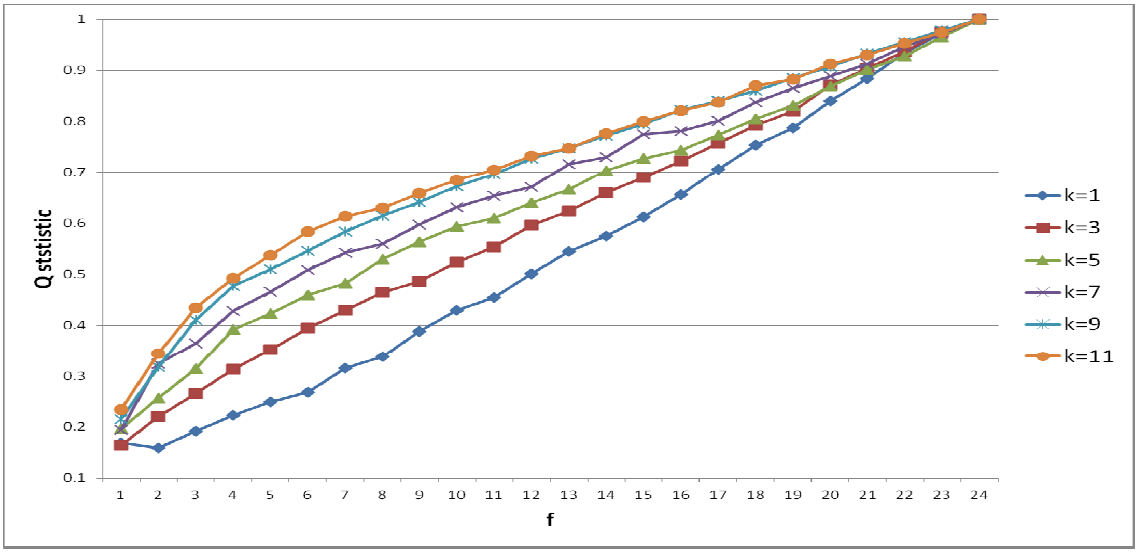
<Figure 6>, <Figure 7>, <Figure 8>은 다양한 k 값에 따른 KNN 앙상블 모형의 여러 결과값을 보여주고 있다. 우선 <Figure 6>은 KNN 앙상블 모형의 예측 성과를 나타내고 있다. <Figure 6>에서 보는 바와 같이 KNN 앙상블 모형은 f의 값에 영향을 받을 뿐만아니라 k의 값에도 영향을



〈Figure 6〉 Average Classification accuracy of KNN ensemble model



〈Figure 7〉 Average Classification Accuracy of Base Classifiers in KNN ensemble model (%)



〈Figure 8〉 Average Q-Statistic Value of Base Classifiers in KNN ensemble model (%)

받고 있다는 것을 알 수 있다. 우선 전반적인 KNN 앙상블 모형의 성과는 f의 값이 12 부근에서 대체로 가장 좋은 값을 보이고 있다. 단일 모형의 경우 k=9일 때 가장 좋은 예측 성과를 보였지만 앙상블 모형의 성과의 경우 k와 f에 따라 서로 다른 예측 성과를 보이고 있으며 그림에서는 k=5일 때 그리고 f=12일 때 가장 좋은 예측 성과를 보였다.

앞에서 살펴본 바와 같이 앙상블 모형의 성과는 그것을 구성하고 있는 기저 분류기들의 평균

예측률과 다양성 정도가 매우 중요하다. 이에 본 논문에서는 앙상블 모형의 최종 성과뿐만 아니라 그것을 구성하고 있는 기저 분류기들의 평균 예측률과 다양성 지수도 함께 살펴 보았다. <Figure 7>과 <Figure 8>은 각 앙상블 모형을 구성하고 있는 기저 분류기들의 평균 예측률과 Q-통계량을 보여주고 있다. 각 기저 분류기들의 평균 예측률은 f가 커짐에 따라 증가함을 알 수 있으며 대체로 k가 클 수록 좋은 값을 보이는 것을 알 수 있다. 하지만 f의 값이 24일 때는 <Figure 4>에 나와 있는 단일 KNN의 성과와 앙상블 모형의 성과가 같다는 것을 알 수 있다. 이는 랜덤 서브스페이스 기법에서 랜덤하게 입력 변수를 선택할 때 비복원 추출 방식을 이용했기 때문이다. 즉, f=24인 앙상블 모형은 KNN 단일 모형과 똑같은 기저 분류기 25개로 구성되어 있으며, 이들의 다양성 지수는 0으로(Q-통계량 값은 1) 앙상블 모형의 개선 효과는 전혀 없다고 볼 수 있다.

앙상블 모형의 성과는 기저 분류들의 예측 성과가 좋을 수록 그리고 기저 분류기들 사이의 다양성 지수가 높을 수록 좋은 것으로 알려져 있다. <Figure 7>에서 보면 앙상블 모형의 성과 개선 측면에서는 f의 값과 k의 값이 클 수록 대체로 좋다는 것을 알 수 있다.

<Figure 8>은 기저 분류기들 사이의 평균 Q-통계량 값을 나타내고 있다. Q-통계량은 분류기 사이의 다양성을 측정하기 위한 대표적인 척도로 두 분류기 C_i와 C_j가 있다고 가정할 때 Q-통계량은 다음과 같이 계산될 수 있다(Kuncheva and Whitaker, 2003).

$$Q-statistic(C_i, C_j) = \frac{(ad-bc)}{(ad+bc)} \qquad (1)$$

여기서, a는 양쪽 분류기가 옳게 분류하는 경우, b는 첫 번째 분류기는 옳게 분류하고 두 번째 분류기는 오분류하는 경우, c는 두 번째 분류기가 옳게 분류하고, 첫 번째 분류기가 오분류하는 경우, d는 모두 오분류하는 경우를 의미한다. Q 값은 -1에서 1사이의 값을 갖게 되며, 통계적으로 독립적인 분류기에 대해서는 0의 값을 갖게 된다. 낮은 오분류율을 보이는 통계적으로 독립적인 기저 분류기들을 결합할 경우 앙상블 모형의 성과는 크게 좋아지는 것으로 알려져 있다.

그림에서 보는 바와 같이 Q-통계량의 값은 f의 값이 커질수록 증가하는 것을 알 수 있으며 대체로 k의 값이 큰 경우가 그렇지 않은 경우보다 Q-통계량의 값이 크다는 것을 알 수 있다. 앙상블 모형의 성과 측면에서는 다양성 지수가 높을수록 좋으므로 Q 통계량이 0에 가까울 수록 좋다는 의미이다. 본 실험 결과에서 볼 때 앙상블 모형의 성과 측면에는 f의 값이 작을수록 그리고 k의 값이 작을 수록 좋다는 것을 알 수 있으며 이는 평균 예측률 관점에서의 경우와 상충하는 것을 알 수 있다. 그러므로, 앙상블 성과 측면에서 최적의 k의 값을 찾는 것은 매우 복잡한 문제라고 볼 수 있으며, 이런 이유로 대부분의 KNN 앙상블 모형에서는 최적의 k와 최적의 입력 변수집합을 고려하지 못하고 있다. 이에 본 논문에서는 앞에서 살펴본 바와 같이 KNN 앙상블 모형의 성과를 향상시키기 위해 각 기저 분류기별 최적의 k값과 각 기저 분류기별 최적의 입력 변수 집합을 유전자 알고리즘을 이용하여 탐색하고 이 값을 이용해 최적의 KNN 앙상블 모형을 구축하였으며 제안한 모형의 최종 성과는 <Table 2>에 나와 있다.

여기서 SORS는 각 기저분류기별로 k와 입력 변수 집합을 동시에 최적화한 KNN 랜덤서브스페이스 앙상블 모형을 의미하며, Single은 가장

〈Table 2 〉 Final Experimental Results

	Model (%)	Base Classifiers (%)	Q-Statistic
Single	72.43	-	-
RS(S)	73.77	69.39	0.73
RS(B)	74.23	68.00	0.64
SORS	76.43	68.07	0.61

〈Table 3 〉 Wilcoxon Signed-Rank Test (p-value)

	RS(S)	RS(B)	SORS
Single	0.007	0.005	0.005
RS(S)		0.093	0.005
RS(B)			0.005

좋은 예측 결과값을 보인 k=9일 때의 KNN 단일 모형을 의미한다. RS(S)는 표준 랜덤 서브스페이스 앙상블 모형을 의미하며, 여기서는 k=9, f=12일 때의 앙상블 모형을 의미한다. RS(B)은 <Figure 6>에서 가장 좋은 결과를 보인 앙상블 모형(k=5, f=12)을 의미한다. Model은 각 모형의 예측 정확도를 의미하고 Base Classifiers는 앙상블을 구성하고 있는 기저 분류기들의 평균 예측률을 의미한다. Q-ststistic는 각 앙상블 모형을 구성하고 있는 기저 분류기들 사이의 Q-통계량의 평균값을 의미한다. 표에서 보는 바와 같이 각 앙상블 모형이 단일 모형보다는 좋은 성과가 나왔으며, 본 논문에서 제안한 모형이 기존의 앙상블 모형보다 예측 성과 면에서 우수한 결과를 보였다. 본 논문에서 제안한 모형은 특히 Q-통계량 값이 가장 작은 값을 보여, 제안한 모형이 기존의 방법보다 기저 분류기의 다양화 측면에서 좋은 결과를 보였으며 이를 통해 전체 앙상블 모형의 성과가 크게 개선되었음을 알 수 있다.

본 논문에서 제안한 모형의 우수성을 통계적으로 검증하기 위해 윌콕슨 부호 순위 검정(Wilcoxon Signed-Rank Test)을 실시하였으며 결과는 <Table 3>과 같다. <Table 3>에서 보는 바와 같이 본 논문에서 제안한 모형의 예측 성과 개선은 통계적으로 유의미하다는 것을 알 수 있으며 이는 본 논문에서 제안한 모형이 예측 성과 개선 측면에서 매우 효과적이었다고 해석할 수 있을 것이다.

6. 결론

앙상블 분류기란 개별 분류기보다 더 좋은 성과를 내기 위해 다수의 분류기를 결합하는 것을 의미한다. 이와 같은 앙상블 분류기는 단일 분류기의 일반화 성능을 향상시키는데 매우 유용한 것으로 알려져 있다. 랜덤 서브스페이스 앙상블 기법은 각각의 기저 분류기들을 위해 원 입력 변수 집합으로부터 랜덤하게 입력 변수를 선택하며 이를 통해 기저 분류기들을 다양화 시키는 기법이다.

본 연구는 부도 예측을 위한 KNN 앙상블 모

형의 성능 개선에 관한 연구이다. KNN을 기저 분류기로 하는 랜덤 서브스페이스 앙상블 모형의 성과는 단일 모형의 성과를 개선시키는 데 효과적인 것으로 알려져 있으며, 이와 같은 랜덤 서브스페이스 앙상블 모형의 성과는 각 기저 분류기를 위해 랜덤하게 선택된 입력 변수와 KNN의 파라미터 k의 값이 중요한 영향을 미친다. 하지만, 단일 모형을 위한 k의 최적 선택이나 단일 모형을 위한 입력 변수 집합의 최적 선택에 관한 연구는 있었지만 KNN을 기저 분류기로 하는 앙상블 모형에서 이들의 최적화와 관련된 연구는 없는 것이 현실이다. 이에 본 연구에서는 KNN을 기저 분류기로 하는 앙상블 모형의 성과 개선을 위해 각 기저 분류기들의 k 파라미터 값과 입력 변수 집합을 동시에 최적화하는 새로운 형태의 앙상블 모형을 제안하였다. 본 논문에서 제안한 방법은 앙상블을 구성하게 될 각각의 KNN 기저 분류기들에 대해 최적의 앙상블 성과가 나올 수 있도록 각각의 기저 분류기가 사용할 파라미터 k의 값과 입력 변수를 유전자 알고리즘을 이용해 탐색하였다.

본 연구에서는 제안한 모형의 성과를 검증하기 위해 국내 기업의 부도 예측관련 데이터를 사용하여 다양한 실험을 수행하였으며 제안한 모형의 예측 성과뿐만 아니라 앙상블을 구성하고 있는 기저 분류기들의 평균 예측률 및 다양성 지수도 함께 살펴 보았다. 실험 결과 제안한 모형은 기존의 앙상블 모형보다 기저 분류기들을 다양화시키는 데 효과적이었다는 것을 알 수 있었다. 또한, 제안한 앙상블 모형은 기존의 앙상블 모형보다 예측 성과 면에서 우수한 결과를 보였다.

본 연구의 한계와 향후 연구 방향을 정리하면 다음과 같다. 첫째, 본 논문에서는 부도 예측을 위해 KNN 모형만을 살펴 보았으나 향후 다양한 분류 모형에 관한 비교 분석과 본 논문에서 제안한 접근 방법을 다양한 분류 모형에 적용해 보는 연구가 필요할 것이다. 둘째, 본 논문에서 제안한 모형은 기존의 단일 모형보다 예측 성과 측면에서 개선이 있었지만 모형의 복잡성이 증가하여 많은 연산이 필요하다는 단점이 있다고 볼 수 있다. 끝으로, 본 논문에서 제안한 모형을 부도 예측 문제에만 적용해 보았지만, 향후 다른 분류 문제에서의 제안 모형의 유용성에 대한 검증이 필요할 것이다.

참고문헌(References)

Abellan, J. and C. J. Mantas, “Improving Experimental Studies about Ensembles of Classifiers for Bankruptcy Prediction and Credit Scoring,” *Expert Systems with Applications*, Vol.41, No.8(2014), 3825~3830.

Alexandre, L., A. Campihlo, and M. Kamel, “On combining classifiers using sum and product rules,” *Pattern Recognition Letter*, Vol.22, No.12(2001), 1283~1289.

Altman, E. I., “Financial ratios, discriminant analysis and the prediction of corporate bankruptcy,” *The Journal of Finance*, Vol.23, No.4(1968), 589~609.

Beaver, W. H., “Financial ratios as predictors of failure,” *Journal of Accounting Research*, Vol.4(1966), 71~111.

Bian, S. and W. Wang, “On diversity and accuracy of homogeneous and heterogeneous ensembles,” *International Journal of Hybrid Intelligent Systems*, Vol.4, No.2(2007), 103~128.

Breiman, L., “Bagging predictors,” *Machine*

Learning, Vol.24, No.2(1996), 123~140.

Buta, P., "Mining for financial knowledge with CBR," *AI Expert*, Vol.9, No.10(1994), 34~41.

Dietterich, T. G., "Machine-learning research: Four current directions," *AI Magazine*, Vol.18, No.4(1997), 97~136.

Goldberg, D. E., *Genetic algorithms in search, optimization and machine learning*, Addison-Wesley, New York, 1989.

Ho, T., "The random subspace method for construction decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.20, No.8(1998), 832~844.

Ho, T., "Multiple classifier combination: Lessons and next steps," *Series in Machine Perception and Artificial Intelligence,* Vol.47(2002), 171~198.

Hung, C. and J-H. Chen, "A Selective Ensemble Based on Expected Probabilities for Bankruptcy Prediction," *Expert Systems with Applications*, Vol.36, No.3(2009), 5297~5303.

Kuncheva, L., J. Bezdek, and R. Duin, "Decision templates for multiple classifier fusion: an experimental comparison," *Pattern Recognition,* Vol.34, No.2(2001), 299~314.

Kuncheva, L. I. and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Machine Learning*, Vol.51, No.2(2003), 181~207.

Kim, M., "Ensemble Learning for Solving Data Imbalance in Bankruptcy Prediction," *Journal of Intelligence and Information Systems*, Vol.15, No.3(2009), 1~15.

Kim, M., "Optimal Selection of Classifier Ensemble Using Genetic Algorithms," *Journal of Intelligence and Information Systems*, Vol.16, No.4(2010), 99~112.

Kim, S. H. and J. W. Kim, "SOHO Bankruptcy Prediction Using Modified Bagging Predictors," *Journal of Intelligence and Information Systems*, Vol.13, No.2(2007), 15~26.

Li, H., Y.-C. Lee, Y.-C. Zhou, and J. Sun, "The random subspace binary logit (RSBL) model for bankruptcy prediction," *Knowledge-Based Systems*, Vol.24, No.8(2011), 1380~1388.

Li, K., Z. Liu, and Y. Han, "Study of Selective Ensemble Learning Methods Based on Support Vector Machine," *Physics Procedia*, Vol.33(2012), 1518~1525.

Louzada, F., O. Anacleto-Junior, C. Candolo, and J. Mazucheli, "Poly-bagging predictors for classification modelling for credit scoring," *Expert Systems with Applications*, Vol.38, No.10(2011), 2717~12720.

Mandler, E. and J. Schurmann, "Combining the classification results of independent classifiers based on the Dempster-Shafer theory of evidence," In E.S. Geselma and L.N. Kanal (eds.), *Pattern Recognition and Artificial Intelligence*, North Holland, Amsterdam, (1988), 381~393.

Marques, A. I., V. Garcia, and J. S. Sanchez, "Two-Level Classifier Ensembles for Credit Risk Assessment," *Expert Systems with Applications,* Vol.39, No.12(2012), 10916~10922.

Meyer, P. A. and H. W. Pifer, "Prediction of bank failures," *The Journal of Finance*, Vol.25, No.4(1970), 853~868.

Messier, W. F. Jr. and J. V. Hansen, "Inducing

rules for expert system development: an example using default and bankruptcy data," *Management Science*, Vol.34, No.12(1998), 1403~1415.

Min, S., "Developing an Ensemble Classifier for Bankruptcy Prediction," *Journal of the Korea Society Industrial Information System,* Vol.17, No.7(2012), 139~148.

Min, S., "Bankruptcy Prediction Using an Improved Bagging Ensemble," *Journal of Intelligence and Information Systems,* Vol.20, No.4(2014), 121~139.

Nanni, L. and A. Lumini, "An Experimental Comparison of Ensemble of Classifiers for Bankruptcy Prediction and Credit Scoring," *Expert Systems with Applications*, Vol.36, No.2(2009), 3028~3033.

Ohlson, J. A., "Financial ratios and the probabilistic prediction of bankruptcy," *Journal of Accounting Research,* Vol.18, No.1(1980), 109~131.

Tam, K. Y. and M. Y. Kiang, "Managerial applications of neural networks: the case of bank failure predictions," *Management Science*, Vol.38, No.7(1992), 926~947.

Tsai, C. and J. Wu, "Using Neural Network Ensembles for Bankruptcy Prediction and Credit Scoring," *Expert Systems with Applications*, Vol.34, No.4(2008), 2639~2649.

Wang, G. and J. Ma, "A hybrid ensemble approach for enterprise credit risk assessment based on Support Vector Machine," *Expert Systems with Applications*, Vol.39, No.5 (2009), 5325~5331.

Zhang, G., Y. M. Hu, E. B. Patuwo, and C. D. Indro, "Artificial neural networks in bankruptcy prediction: general framework and cross-validation analysis," *European Journal of Operational Research*, Vol.116, No.1(1999), 16~32.

Abstract

Investigating Dynamic Mutation Process of Issues Using Unstructured Text Analysis

Sung-Hwan Min*

Bankruptcy involves considerable costs, so it can have significant effects on a country's economy. Thus, bankruptcy prediction is an important issue. Over the past several decades, many researchers have addressed topics associated with bankruptcy prediction. Early research on bankruptcy prediction employed conventional statistical methods such as univariate analysis, discriminant analysis, multiple regression, and logistic regression. Later on, many studies began utilizing artificial intelligence techniques such as inductive learning, neural networks, and case-based reasoning. Currently, ensemble models are being utilized to enhance the accuracy of bankruptcy prediction. Ensemble classification involves combining multiple classifiers to obtain more accurate predictions than those obtained using individual models. Ensemble learning techniques are known to be very useful for improving the generalization ability of the classifier.

Base classifiers in the ensemble must be as accurate and diverse as possible in order to enhance the generalization ability of an ensemble model. Commonly used methods for constructing ensemble classifiers include bagging, boosting, and random subspace. The random subspace method selects a random feature subset for each classifier from the original feature space to diversify the base classifiers of an ensemble. Each ensemble member is trained by a randomly chosen feature subspace from the original feature set, and predictions from each ensemble member are combined by an aggregation method.

The k-nearest neighbors (KNN) classifier is robust with respect to variations in the dataset but is very sensitive to changes in the feature space. For this reason, KNN is a good classifier for the random subspace method. The KNN random subspace ensemble model has been shown to be very effective for improving an individual KNN model. The k parameter of KNN base classifiers and selected feature subsets for base classifiers play an important role in determining the performance of the KNN ensemble model. However, few studies have focused on optimizing the k parameter and feature subsets of base classifiers in the ensemble. This study proposed a new ensemble method that improves upon the performance KNN ensemble model by optimizing both k parameters and feature subsets of base classifiers. A genetic algorithm was used to optimize the KNN ensemble model and improve the prediction accuracy of the

* Corresponding Author: Sung-Hwan Min
Department of Business Administration, Hallym University
39 Hallymdaehak-gil,Chuncheon Gangwon-do, 200-702, Korea
Tel: +82-33-248-1841, Fax: +82-33-256-3424 ,E-mail: shmin@hallym.ac.kr

ensemble model.

The proposed model was applied to a bankruptcy prediction problem by using a real dataset from Korean companies. The research data included 1800 externally non-audited firms that filed for bankruptcy (900 cases) or non-bankruptcy (900 cases). Initially, the dataset consisted of 134 financial ratios. Prior to the experiments, 75 financial ratios were selected based on an independent sample t-test of each financial ratio as an input variable and bankruptcy or non-bankruptcy as an output variable. Of these, 24 financial ratios were selected by using a logistic regression backward feature selection method. The complete dataset was separated into two parts: training and validation. The training dataset was further divided into two portions: one for the training model and the other to avoid overfitting. The prediction accuracy against this dataset was used to determine the fitness value in order to avoid overfitting. The validation dataset was used to evaluate the effectiveness of the final model. A 10-fold cross-validation was implemented to compare the performances of the proposed model and other models.

To evaluate the effectiveness of the proposed model, the classification accuracy of the proposed model was compared with that of other models. The Q-statistic values and average classification accuracies of base classifiers were investigated. The experimental results showed that the proposed model outperformed other models, such as the single model and random subspace ensemble model.

Key Words : Random Subspace, Bankruptcy Prediction, Ensemble, Genetic Algorithms

Received : March 2, 2016 Revised : March 14, 2016 Accepted : March 14, 2016
Publication Type : Regular Paper Corresponding Author : Sung-Hwan Min

저 자 소 개

민 성 환
KAIST 테크노 경영대학원에서 경영정보시스템을 전공하여 박사를 취득하고 현재 한림대학교 경영학과에 재직 중이다. 주요 관심 분야는 데이터 마이닝, 재무예측 모형 개발, 고객관계관리 등이다.