KIPS Trans. Softw. and Data Eng.
Vol.9, No.5 pp.161~168 pISSN: 2287-5905

https://doi.org/10.3745/KTSDE.2020.9.5.161

Prediction of English Premier League Game Using an Ensemble Technique

Yi Jae Hyun† · Lee Soo Won††

ABSTRACT

Predicting outcome of the sports enables teams to establish their strategy by analyzing variables that affect overall game flow and wins and losses. Many studies have been conducted on the prediction of the outcome of sports events through statistical techniques and machine learning techniques. Predictive performance is the most important in a game prediction model. However, statistical and machine learning models show different optimal performance depending on the characteristics of the data used for learning. In this paper, we propose a new ensemble model to predict English Premier League soccer games using statistical models and the machine learning models which showed good performance in predicting the results of the soccer games and this model is possible to select a model that performs best when predicting the data even if the data are different. The proposed ensemble model predicts game results by learning the final prediction model with the game prediction results of each single model and the actual game results. Experimental results for the proposed model show higher performance than the single models.

Keywords : Machine Learning, Artificial Intelligence, Sports Game Prediction, Ensemble Technique, Data Analysis

앙상블 기법을 통한 잉글리시 프리미어리그 경기결과 예측

이 재 현† · 이 수 원††

요 약

스포츠 경기 결과예측은 전반적인 경기의 흐름과 승패에 영향을 미치는 변인들의 분석을 통해 팀의 전략 수립을 가능하게 해준다. 이와 같은 스포츠 경기결과 예측에 대한 연구는 주로 통계학적 기법과 기계학습 기법을 활용하여 진행되어 왔다. 승부예측 모델은 무엇보다 예측 성능이 가장 중요시된다. 그러나 최적의 성능을 보이는 예측 모델은 학습에 사용되는 데이터에 따라 다르게 나타나는 경향을 보였다. 본 논문에서는 이러한 문제를 해결하기 위해 데이터가 달라지더라도 해당 데이터에 대한 예측 시 가장 좋은 성능을 보이는 모델의 선택이 가능한 기존의 축구경기결과 예측에서 좋은 성능을 보여온 통계학적 모델과 기계학습 모델을 결합한 새로운 앙상블 모델을 제안한다. 본 논문에서 제안하는 앙상블 모델은 각 단일모델들의 경기 예측결과와 실제 경기결과를 병합한 데이터로부터 최종예측모델을 학습하여 경기 승부예측을 수행한다. 제안 모델에 대한 실험 결과, 기존 단일모델들에 비해 높은 성능을 보였다.

키워드 : 기계학습, 인공지능, 스포츠 승부 예측, 앙상블 기법, 데이터 분석

1. 서 론

스포츠 경기에 대한 결과 예측은 경기 분석가에겐 전반적인 경기들의 흐름을 제공해줄 수 있고 승패에 영향을 미치는 변인들의 분석을 통하여 팀의 전략 수립을 가능하게 해준다. 또한 스포츠 경기의 결과 예측을 통해 수익을 낼 수 있는 국내에서 정식으로 발행되고 있는 스포츠 토토 베팅에 도움이 되기 때문에 스포츠 전문가이외의 스포츠팬들 또한 스포츠 경기들의 결과를 정확히 예측하고자 노력한다. 이에 따라 스포츠 경기 결과 예측에 대한 관심은 계속 증가하고 있고 스포츠 경기 결과 예측에 관련된 연구들 또한 활발히 진행되어 왔다. 이때 스포츠 경기 결과 예측에 활용된 주된 기법은 통계학적 기법과 기계학습 기법이다. 통계학적 기법을 활용한 스포츠 경기 결과 예측은 로지스틱 회귀분석, 푸아송 분포 등을 활용한다[1, 2].

최근 기계학습 기법을 통한 빅데이터 분석의 연구들이 좋은 결과를 보임에 따라 이를 활용한 스포츠 경기결과 예측에 대한 연구가 많이 진행되었다. 기계학습 기법을 활용한 스포츠 경기결과 예측에는 주로 베이지안 네트워크, Multi Layer Perceptron(MLP), Recurrent Neural Networks(RNN), Convolutional Neural Network(CNN) 등의 예측모델들이

※ 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학 ICT연구센터 지원사업의 연구결과로 수행되었음(IITP-2019-2018-0-01419).
† 준 회 원 : 숭실대학교 융합소프트웨어학과 석사과정
†† 정 회 원 : 숭실대학교 소프트웨어학부 교수
Manuscript Received : November 11, 2019
Accepted : January 22, 2020
* Corresponding Author : Lee Soo Won(swlee@ssu.ac.kr)

사용된다[3-6]. 승부예측 모델은 예측 성능이 가장 중요시 된다. 기존의 스포츠 승부예측 연구들을 살펴보면 구승환의 연구[1]에서는 Logistic Regression 모델, Yezus의 연구[9]에서는 Random Forest 모델이 가장 좋은 성능을 보였으며 Pettersson의 연구[4]에서는 LSTM 모델, 김주학의 연구[11]에서는 MLP 모델이 가장 좋은 성능을 보였다. 이와 같이 모델의 학습에 사용하는 데이터에 따라 좋은 성능을 보이는 모델이 다르게 나타나기 때문에 새로운 데이터에 대해 어떠한 단일모델이 가장 좋은 예측 성능을 보일 것인지 결정하기 어렵다. 이러한 문제를 해결하기 위하여 본 논문에서는 스포츠 경기 결과의 예측을 위해 데이터가 달라지더라도 해당 데이터에 대한 예측 시 가장 좋은 성능을 보이는 단일모델의 선택이 가능한 기존의 스포츠 경기결과 예측에 좋은 성능을 보인 여러 기계학습 모델과 통계학적 모델을 결합한 앙상블 모델을 제안한다. 제안하는 앙상블 모델은 각 단일모델의 경기예측결과와 실제 경기결과를 병합한 데이터로부터 최종예측모델을 학습하여 경기에 대한 승부예측을 수행한다. 가장 적합한 최종예측모델의 선정을 위해서 Validation Set을 활용하여 최고의 성능을 보이는 최종예측모델과 해당 모델의 최적의 파라미터를 설정한다. 본 논문에서 제안하는 앙상블 모델에 포함되는 단일모델은 Multi Layer Perceptron(MLP), Support Vector Machine(SVM), Random Forest, Long Short Term Memory(LSTM), Logistic Regression이다. 본 논문에서 제안하는 앙상블 모델은 각 단일모델의 상호보완 효과로 기존의 승부예측 모델에 활용된 단일모델에 비해 보다 정확한 예측 성능을 기대해 볼 수 있다. 본 연구의 의의는 기존의 승부예측에서 좋은 성능을 보인 여러 단일모델을 결합하여 데이터가 달라지더라도 해당 데이터에 대한 최적의 예측을 수행할 수 있는 새로운 앙상블 승부예측 모델의 개발에 있다.

본 논문에서 예측하고자 하는 스포츠 경기는 축구이며 현재 전 세계에서 가장 많은 사람들이 시청하는 스포츠 리그인 잉글리시 프리미어리그의 경기들을 예측 대상으로 선정하였다.

본 논문의 구성은 다음과 같다. 2장에서 통계학적 모델, 기계학습 모델, 앙상블 모델을 활용하여 스포츠 경기 결과 예측을 시도하였던 기존 연구에 대하여 기술한다. 3장에서는 본 논문의 제안방법에 대하여 기술한다. 4장에서는 실험을 통해 학습된 모델의 성능을 비교 평가한다. 5장에서는 본 논문의 결론과 향후연구에 대해 기술한다.

2. 관련 연구

2.1 통계학적 모델 기반 스포츠 경기 결과 예측

구승환[1]은 농구경기의 승패에 영향을 미치는 요인을 팀별, 경기별로 분석하고 Logistic Regression 모델과 인공신경망 모델로 승패예측을 하였다. Groll[2]은 푸아송 모델을 활용하여 UEFA European Football Championship의 이전 3개의 시즌의 분석을 통하여 해당 시즌 경기들의 점수를 예측하였다. 홍종선[7]은 2010년 남아공 월드컵 축구 결과 예측을 위하여 Bradley-Terry모형을 사용하였다. 이 예측모형은 경기력에 영향을 미치는 확률 변수들을 분석하며, 쌍별 비교방법(Paired Comparison Method)을 사용하였다. Prasetio[8]는 축구경기를 이기는데 중요한 변수가 무엇인지 결정하기 위해 Logistic Regression 모델을 구축하고 2015/2016 시즌의 프리미어리그의 경기 결과를 예측하였다. Yezus[9]는 축구경기의 경기결과 예측을 위해 Random Forest 모델을 사용하였다. Ulmer[10]는 잉글리시 프리미어리그의 경기결과를 예측하기 위해 Naive Bayes, Hidden Markov Model, Support Vector Machine(SVM), 그리고 Random Forest 모델을 활용하였다.

2.2 기계학습 모델 기반 스포츠 경기 결과 예측

Owramipur[3]는 베이지안 네트워크를 활용하여 스페인 라리가의 바르셀로나팀의 축구경기 결과를 예측하였다. Pettersson[4]은 축구경기 결과 예측을 위하여 Long Short Term Memory 모델을 활용하였고 이를 통해 기계학습 모델이 축구경기의 결과를 예측하는데 좋은 성능을 보임을 확인하였다. 최형준[5]은 2005년도 영국 윔블던 테니스 대회의 남자 단식 경기 데이터를 활용하여 경기시작 전 다음경기의 승패를 예측하였으며, 실험 결과 인공신경망을 이용하는 방법이 기존의 정준판별분석(Canonical Discriminant Analysis) 또는 Logistic Regression 모델을 이용한 예측기법들에 비해서 더 우수한 성능을 보이는 것을 확인하였다. 이재현[6]은 잉글리시 프리미어리그의 경기결과를 예측하기 위하여 경기정보와 결과에 대한 시계열 데이터를 구축한 후 Long Short Term Memory 모델을 활용하였다. 김주학[11]은 2006 독일월드컵의 경기내용에서 기록한 요인들을 점수화하여 Multi Layer Perceptron기반의 승패 예측모델을 구축하였으며, 이 모델을 통해 다음 경기의 승패를 예측하였다. 오윤학[12]은 2013년 시즌 KBO 프로야구팀과 선수들의 경기데이터로부터 다음 경기의 승패를 예측하기 위해 Decision Tree, Random Forest, Logistic Regression, 신경망 분석, SVM, 판별분석을 이용하였으며, 실험결과 Random Forest 모델이 성능도 우수할 뿐만 아니라, 변수의 중요도 역시 산출해 낼 수 있음을 확인하였다. 최형준[13]은 2002년부터 2018년 사이에 개최된 5번의 월드컵에서 나타난 공식기록을 활용하여 자기구성 지도(Self-Organized Map)를 학습시켜 경기결과를 예측하였다.

2.3 앙상블 모델 기반 스포츠 경기 결과 예측

조수현[14]은 여자부 국제컬링경기대회에서 5엔드 직후 경기의 승패를 예측하기 위하여 기계학습 알고리즘을 이용한 하이브리드 방법을 제안하였다. 하이브리드 기계학습 모델에 포함된 모델은 CNN, LSTM, MLP, Logistic Regression, SVM 등 총 5개의 모델이며 하이브리드 기계학습 모델이 단

일 모델들의 평균 성능보다 10% 정도의 향상을 보였다. Cui[15]는 다수의 Genetic Programming 앙상블 기법을 활용한 잉글리시 프리미어리그 경기 예측모델을 제안하였으며 평가결과 단일 Genetic Programming 시스템을 활용하였을 때 보다 앙상블 기법을 활용하였을 때 더 좋은 예측정확도를 보였다. Saricaoğlu[16]는 터키의 축구리그인 터키쉬 슈퍼리그의 승부예측을 위하여 Logistics Regression, K-nearest Neighbors, Support Vector Machines, Random Forests 등 10개의 모델을 결합한 앙상블 모델을 제안하였다. Hoekstra[17]는 축구의 승부예측을 위하여 첫 번째 단계에서는 여러 개의 모델로 이루어진 앙상블 모델 중 가장 성능이 좋은 모델의 조합을 찾고 두 번째 단계에서 이 모델들의 앙상블 참여시의 최적의 가중치를 찾아주는 방식을 채택한 Evolutionary 앙상블 모델을 제안하였다.

3. 제안 방법

본 연구에서 제안하는 앙상블 모델의 구조도는 Fig. 1과 같다. 제안하는 앙상블 모델은 전처리 모듈, 단일모델 예측 모듈, 데이터 병합 모듈, 최종 예측모듈로 구성되어 있다.

전처리 모듈에서는 입력데이터의 정규화 작업과 라벨링 작업을 수행한다. 본 연구에서 이용한 입력데이터는 Table 1과 같다. 입력 데이터는 예측하고자 하는 경기의 홈팀과 어웨이팀의 리그 내에서의 바로 전 경기 데이터(볼 점유율, 유효슈팅 수, 총 슈팅 횟수, 총 볼터치 횟수, 총 패스 횟수, 총 태클 횟수, 총 수비성공 횟수), 홈, 어웨이팀의 최근 5경기의 경기결과 그리고 경기 직전의 홈, 어웨이팀의 리그 내 순위로 구성되어 있다. Table 1의 9개의 Feature의 경우 홈팀과 어웨이팀 별로 각각 구성되므로 각 단일모델에서 학습되는 Feature수는 경기당 18개이다. 모든 데이터는 정규화 작업을 통해 0과 1사이의 값으로 정규화되고, 각 경기들은 홈팀이 승리했을 경우에는 2, 경기가 무승부로 끝났을 때에는 1, 홈팀이 패배했을 경우에는 0으로 라벨링된다.

각 단일모델 사이의 상호보완을 통한 앙상블 기법의 예측 성능 향상을 위해서 기존의 스포츠 경기결과 예측과 관련된 연구

Table 1. Input Data

Feature	Explanation
Possession	Ratio of Ball Possession of the Latest Game
Shots on Target	Number of Shots on Target of the Latest Game
Shots	Number of Shots of the Latest Game
Touches	Number of Ball Touches of the Latest Game
Passes	Number of Passes of the Latest Game
Tackles	Number of Tackles of the Latest Game
Clearances	Number of Clearances of the Latest Game
Last 5 Match Results	Match Results of Last 5 Games (Win: 1, Draw: 0, Lose:-1; The Sum of the Values of 1, 0, and -1)
Ranking	League Rank

에서 좋은 성능을 보인 단일 모델들이 필요하다. 본 연구에서는 Logistic Regression[1, 5, 8], Long Short Term Memory (LSTM)[4, 6], Random Forest[9, 10, 12], Support Vector Machine(SVM)[10, 12], Multi Layer Perceptron(MLP)[11] 총 5개를 단일모델로 사용한다. 단일 모델 예측 모듈과 최종 예측 모듈의 경우 학습단계와 예측단계로 나뉜다. 단일모델 예측 모듈의 학습단계에서는 Training Set과 Validation Set을 활용하여 각 단일모델의 파라미터 최적화를 진행하고 각 단일모델을 학습한다. 단일모델 예측 모듈의 학습단계가 끝난 후 Validation Set에 대한 각 단일모델의 예측결과와 실제 경기결과를 데이터 병합 모듈을 통해 병합한다. 병합 모듈을 통해 출력되는 최종 예측 모듈의 최종 입력 값의 예시는 Table 2와 같다.

단일모델 예측 모듈과 최종 예측 모듈의 학습단계가 끝난 후 예측하고자하는 새로운 경기에 대한 제안 앙상블 모델의 예측단계는 다음과 같다. 예측하고자하는 새로운 경기들에 대한 각 단일모델들의 예측을 수행하고 병합 모듈을 통해 각 단일모델들의 예측값들을 병합한 데이터를 출력한다. 경기 결과에 대한 각 단일모델의 예측값은 승, 무, 패 각각에 대한 라벨링값(2,1,0)으로 출력된다. 병합 모듈을 통해 병합된 데이터를 입력으로 하여 최종 예측 모듈의 학습단계에서 선정된 최종예측

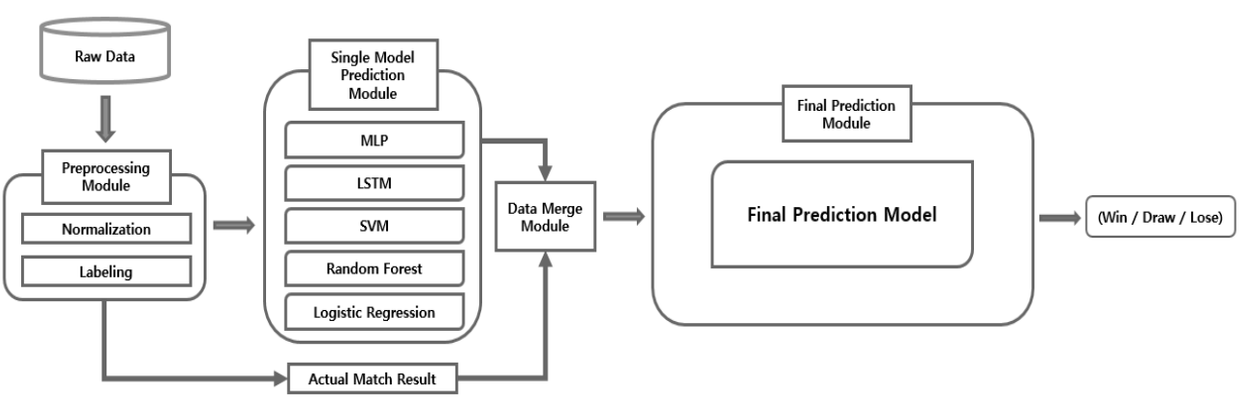

Fig. 1. Structure of the Proposed Ensemble Model

Table 2. Sample Output from the Single Model Prediction Module

Match Number	Predicted Result from Each Single Model					Actual Result
	Logistic Regression	LSTM	MLP	Random Forest	SVM	
1	0	0	0	0	0	0
2	1	1	1	2	1	1
3	2	2	2	2	1	2
...	...					...
350	2	2	2	1	2	-
351	2	2	2	2	2	-
352	0	0	1	1	0	-

모델은 최종적으로 예측하고자하는 경기들에 대한 예측을 수행한다. 이와 같은 앙상블 과정에 의해 단일예측모듈의 각 단일모델의 예측성능이 좋지 않아도 최종예측 모듈을 통해 예측성능이 향상될 수 있다. 최종예측 모델의 선정과 최적의 파라미터 측정과 관련된 실험방법 및 결과는 4장에서 설명한다.

4. 실험 및 결과

본 연구에서는 잉글리시 프리미어리그의 총 13개(2006/2007~2018/2019)시즌에 해당하는 경기에 대한 데이터를 프리미어리그 공식 홈페이지(www.premierleague.com)로부터 수집하였다. 이중 11개의 시즌(2006/2007~2016/2017)은 학습 데이터로 구성하고 2개의 시즌(2017/2018~2018/ 2019)은 테스트 데이터로 구성하였다. Table 3은 수집된 데이터에 대한 Training Set, Validation Set, Test Set의 구성이다.

Table 3. Configuration of Training Set, Validation Set, Test Set

Category	Season	Total Number of Matches
Training Set	(2006/2007 Season) ~ (2014/2015 Season)	3227
Validation Set	(2015/2016 Season) ~ (2016/2017 Season)	728
Test Set	(2017/2018 Season) ~ (2018/2019 Season)	740

제안하는 프리미어리그 경기결과 예측 모델의 실험환경으로는 Ubuntu16.04 운영체제와 Python3.5 프로그래밍 언어가 사용되었으며 이에 사용된 프레임워크는 Tensorflow, Keras, Pandas, Numpy, Scikit-Learn 그리고 BeautifulSoup이다.

제안 모델의 입력데이터에 대한 Feature Selection은 Validation Set을 활용해 수행되었으며 Feature Selection을 위해 사용된 모델은 Logistic Regression 모델과 SVM 모델이다. 별도의 입력데이터 Feature 제거를 하지 않았을 때 Validation Set에 대한 Logistic Regression 모델과 SVM 모델의 정확도가 가장 높았기 때문에 기존의 Feature를 모두 사용하는 방식을 채택하였다.

단일모델 예측 모듈에 포함된 각 단일 모델의 경우 파라미터의 최적화가 필요하다. 파라미터 최적화를 위해 Table 1에서 제시한 9개의 Feature값을 이용해 승부예측을 진행할 때 각 단일모델의 최적의 파라미터를 찾는 작업을 진행하였다. 본 연구에서 사용된 모델의 파라미터 최적화에는 Validation Set을 이용하였다. 이때 사용된 Validation Set은 Validation Set으로 지정해 놓은 2개의 시즌이 포함되어 있는 728개의 데이터 중 2015/2016시즌에 해당하는 356개의 경기 데이터이다.

LSTM모델의 최적 파라미터를 구하기 위해 Hidden Layer Size와 Time-Step을 변경하면서 모델의 성능을 측정하였다. Hidden Layer Size의 경우 2, 4, 8, 16, 32, 64, 128 그리고 Time-Step의 경우 5, 10, 50, 100 으로 변경하면서 모든 경우에 대한 모델의 성능을 측정하였다. LSTM 모델의 파라미터 최적화 실험 결과는 Fig. 2와 같으며 최적의 파라미터가 적용된 LSTM 모델의 실험조건은 Table 4와 같다.

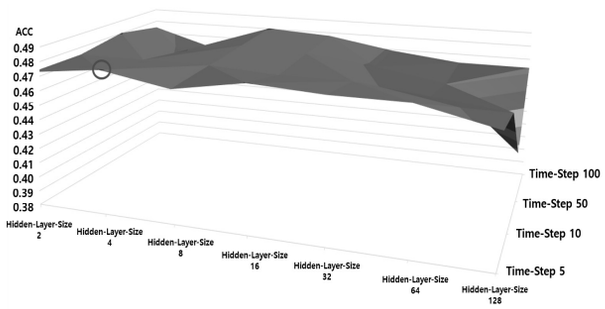

Fig. 2. Results of the Parameter Optimization Experiment for LSTM Model

Table 4. Optimal Parameters of the LSTM Model

	Parameter				Termination Condition	
Model	Input Node	Hidden Layer Size	Output Node	Time Step	Epoch	Learning Rate
LSTM	18	4	3	3	380	0.01

MLP모델에서 사용된 Optimizer는 Adam Optimizer [18]이며, MLP 모델의 최적 파라미터 측정을 위해 2개의 Hidden Layer의 크기와 배치사이즈를 달리하며 모델의 성능을 단계적으로 확인하였다. MLP 모델의 파라미터 최적화 실험 결과는 Table 5와 같으며 MLP 모델의 최적 파라미터는 Table 6과 같다.

SVM 모델의 파라미터 최적화를 위해 SVM상수인 C값을 0.01부터 1까지 0.01의 간격으로 변경해가며 모델의 성능을 측정하였다. 파라미터 최적화 실험의 결과 SVM 상수 C값이 0.11일 때 가장 좋은 성능을 보이는 것을 확인하였다. SVM 모델의 파라미터 최적화 실험 결과는 Fig. 3과 같으며 SVM 모델의 최적 파라미터는 Table 7과 같다.

Table 5. Results of the Parameter Optimization Experiment for MLP Model

Batch_Size	Hidden Layer_1_Size	Hidden Layer_2_Size	Prediction Accuracy (%)
3227	21	16	0.4972
2048	22	23	0.4972
1024	21	18	0.5
512	9	8	0.4972
256	5	28	0.4944
128	23	5	0.4972
64	10	20	0.4944
32	9	17	0.4916
16	8	15	0.4944
8	12	7	0.4888
4	4	4	0.4803
2	4	12	0.4835

Table 6. Optimal Parameters of MLP Model

	Parameter					Termination Condition	
Model	Input Node	Number of Hidden Layers	Each Hidden Layer Size	Output Node	Batch Size	Epoch	Learning Rate
MLP	18	2	21, 18	3	1024	200	0.01

Table 7. Optimal Parameters of SVM Model

	Parameter	
Model	C_Value	Kernel
SVM	0.11	Linear

Random Forest모델의 파라미터 최적화를 위해 n_estimators 값을 1부터 1000까지 변경해가며 모델의 성능을 측정하였다. 이때 n_estimators값이 122일 때 가장 좋은 성능을 보이는 것을 확인하였다. Random Forest모델의 파라미터 최적화 실험 결과는 Fig. 4와 같다.

각 단일모델의 파라미터 최적화를 완료한 후, 최종 예측 모듈에 활용될 최종 예측 모델의 선정과 파라미터 최적화를 위한 실험을 진행하였다. 본 실험은 Validation Set을 활용하여 진행하였으며 각 단일모델의 파라미터 최적화는 단일모델 예측 모듈과 같은 방식으로 진행하였다. 단일모델 예측 모듈을 통해 병합된 데이터에 대한 학습 시 각 단일모델의 최적 파라미터는 Table 8과 같다.

단일모델 예측 모듈을 통해 파라미터 최적화가 완료된 단일모델들의 Validation Set에 대한 예측정확도와 앙상블 기법이 적용된 최종 예측 모듈을 활용했을 때의 Validation Set에 대한 예측정확도는 Table 9와 같다. Table 9에 따르면 모든 단일모델의 예측정확도는 앙상블 기법을 적용한 최종 예측 모듈에서 더 높게 측정되었고 이를 통해 앙상블 기법의 타당성을 확인하였다. 실험결과 단일모델 예측 모듈과 최종 예측 모듈에서 모두 LSTM모델이 가장 좋은 성능을 보였으므로 본 실험에서의 최종 예측 모듈에 사용되는 학습모델은 LSTM모델로 결정하였다. 실험에 사용된 데이터는 경기가 일어난 순서에 따라 수집된 시계열데이터의 형태이기 때문에 시계열데이터 처리에 가장 적합한 모델인 LSTM 모델이 가장 좋은 성능을 보인 것으로 추측된다. 본 실험에서 선정한 데이터를 활용할 경우 최종 앙상블 모델 구조도는 Fig. 5와 같다.

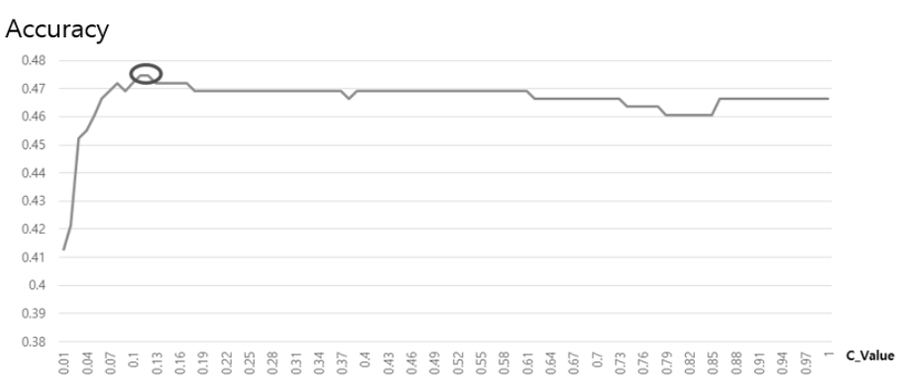

Fig. 3. Results of the Parameter Optimization Experiment for SVM Model

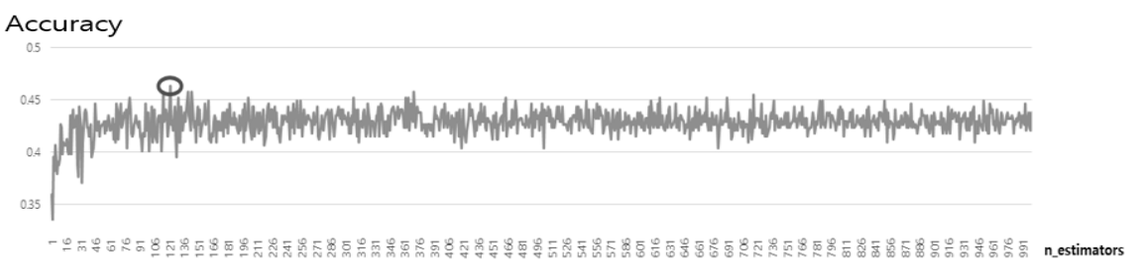

Fig. 4. Results of the Parameter Optimization Experiment for Random-Forest Model

Table 8. Optimal Parameters of each Single Model for the Final Prediction Module

Model	LSTM		MLP		SVM		Random-Forest	
Parameter	Input Node	5	Input Node	5	C_Value	0.01	n_estimators	3
	Hidden Layer Size	2	Number of Hidden Layers	2				
	Output Node	3	Each Hidden Layer Size	4,11	Kernel	Linear		
	Time Step	110	Output Node	3				
			Batch Size	256				
Termination Condition	Epoch	390	Epoch	200				
	Learning Rate	0.01	Learning Rate	0.01				

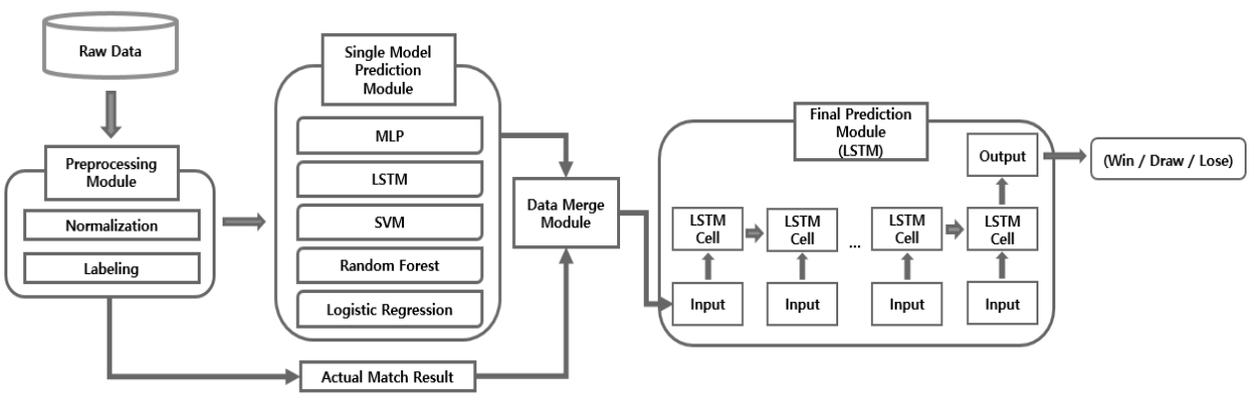

Fig. 5. Structure of the Proposed Ensemble Model in Our Experiment

Table 9. Prediction Accuracy of Single Model Prediction Module and Final Prediction Module

Model	Prediction Accuracy	
	Single Model Prediction Module	Final Prediction Module
Logistic Regression	0.5645	0.5753
SVM	0.5753	0.578
Random Forest	0.5753	0.578
MLP	0.5726	0.594
LSTM	0.5887	0.6102

본 연구에서는 제안모델의 성능을 비교하기 위해 2개의 Baseline 모델을 선정하였다. Baseline_1은 전체 수집데이터의 특성상 홈팀이 승리한 경우의 경기 수가 무승부의 경우와 패배했을 경우보다 많았기 때문에 모든 경기에 대해서 홈팀이 승리한다고 예측했을 때의 정확도이다. Baseline_2는 경기가 일어나는 시점의 두 팀의 리그 내 순위를 기반으로 순위가 높은 팀이 이길 것이라고 예측했을 때의 정확도이다. Baseline_1 모델과 Baseline_2 모델의 성능은 각각 47.8%, 54.1%로 나타났다.

본 연구에서 제안한 앙상블 모델과 각 단일모델 및 Baseline 모델과의 예측 정확도를 비교한 결과는 Table 10과 같다.

Table 10. Prediction Accuracy of the Proposed Model and Comparative Models

Model	Accuracy(%)	Remarks
Baseline_1	47.8	Expect the Home Team Will Win All Games
Baseline_2	54.1	Ranking Based Predictions
LSTM	56.1	
MLP	55.7	
SVM	56.2	
Logistic Regression	56.5	
Random forest	51.5	
Ensemble model	56.9	Proposed model

Table 10에 따르면 제안 앙상블 모델이 다른 비교모델들보다 높은 예측정확도를 보이는 것을 확인할 수 있다. 제안 앙상블 모델은 예측하고자하는 새로운 경기에 대하여 각 단일모델의 예측결과 및 최종 예측 모델의 예측을 통하여 경기예측을 수행한다. 즉 각각의 단일모델이 같은 경기에 대한 예측이 다를수록 더 유의미한 분석 및 예측을 수행할 수 있다. 본 실험에서는 앙상블 모델에 사용된 각 단일 모델들이 예측하고자 하는 test data 740개의 경기 중 189개의 경기에 대해 서로 다른 예측 값을 보였고, 이로 인해 각 예측모델간의 상호보완이 적용되어 앙상블 모델이 가장 좋은 성능

을 보일 수 있었던 것으로 추측할 수 있다.

5. 결론 및 향후 계획

본 연구에서는 잉글리시 프리미어리그의 경기결과를 예측하기 위해 새로운 앙상블 모델을 제안하였다. 제안하는 앙상블 모델에는 LSTM 모델, MLP 모델, SVM 모델, Random Forest 모델, Logistic Regression 총 5개의 단일모델이 사용되었다. 제안하는 앙상블 모델은 각 단일모델들의 경기 예측결과와 실제 경기결과를 병합한 데이터로부터 최종예측모델을 학습하여 경기 승부예측을 수행한다.

프리미어리그의 2개의 시즌에 해당하는 740개의 경기에 대한 경기결과 예측을 진행한 결과, 제안 앙상블 모델이 다른 단일모델과 Baseline보다 높은 예측정확도를 보였다. 본 연구에서 사용되지 않은 통계학적 모델인 베이지안 모델, 푸아송 모델과 딥러닝 모델인 CNN 등과 같은 단일모델을 추가적으로 활용했을 때의 예측 정확도의 향상이 있을지에 대한 연구가 필요하다. 또한 축구 경기에 대한 데이터를 잉글리시 프리미어리그 공식홈페이지에서의 수집 이외의 각 잉글리시 프리미어리그 구단의 데이터 수집 시스템을 활용하여 수집된 데이터로 경기결과에 대한 예측을 진행할 시 모델의 정확도의 향상을 기대해 볼 수 있을 것이다.

References

[1] Swung Hwan Gu, Hyun Soo Kim, and Seong Yong Jang, "A Comparison Study On the Prediction Models For the Professional Basketball Game," *Korean Journal of Sport Science*, Vol.20, No.4, pp.704-711, 2009.

[2] Andreas Groll, Thomas Kneib, Andreas Mayr, and Gunther Schauberger, "On the Dependency of Soccer Scores - A Sparse Bivariate Poisson Model for the UEFA European Football Championship 2016," *Journal of Quantitative Analysis In Sports*, Vol.14, No.2, pp.65-79, 2018.

[3] Owramipur Farzin, Eskandarian Parinaz, and Sadat Mozneb Faezeh, "Football Result Prediction With Bayesian Network In Spanish League-Barcelona Team," International Journal of Computer Theory And Engineering, pp.812-815, 2013.

[4] Daniel Petterson and Robert Nyquist, "Football Match Prediction Using Deep Learning," CHALMERS UNIVERSITY OF TECHNOLOGY Gothenburg, Sweden 2017 EX031/2017, 2017. B. Lenat, "Programming articitical interlligence," in Understanding Artificial Intelligence, Scientific American, Ed., New York: Warner Books Inc., pp.23-29, 2002.

[5] Hyung Joon Choi, "Prediction of Game Results Using ANN(Artificial Neural Networks) Within The Wimbledon Tennis Championship 2005," *The Korean Journal of Physical Education*, Vol.45, No.3, pp.459-468, 2006.

[6] Jae Hyun Yi and Soo Won Lee, "Prediction of English Premier League Game Results By Using Deep Learning Techniques," *ISSAT International Conference Data Science in Business, Finance and Industry*, pp.96-98, 2019.

[7] Chong Sun Hong, Min Sub Jung, and Jae Hyoung Lee, "Prediction Model Analysis of 2010 South Africa World Cup," *The Korean Data & Information Scienece Society*, Vol.21, No.6, pp.1137-1146, 2010.

[8] Darwin Prasetio, "Predicting Football Match Results With Logistic Regression," *2016 International Conference on Advanced Informatics: Concepts, Theory and Application (ICAICTA)*, George Town, pp.1-5, 2016.

[9] Albina Yezus, "Predicting Outcomes of Soccer Matches Using Machine Learning," Saint-Petersburg University, Saint-Petersburg State University Mathematics And Mechanics Faculty, 2014.

[10] Ben Ulmer and Matthew Fernandez, "Predicting Soccer Match Results In The English Premier League," Doctoral Dissertation, Ph. D. Dissertation, Stanford, 2013.

[11] Joo Hak Kim, Gap Taik Ro, Jong Sung Park, and Won Hi Lee, "The Development of Soccer Game Win Lost Prediction Model Using Neural Network Analysis -FIFA World Cup 2006 Germany-," *Korean Journal of Sport Science*, Vol.18, No.4, pp.54-63, 2007.

[12] Youn Hak Oh, Han Kim, Jae Sub Yun, and Jong Seok Lee, "Using Data Mining Techniques To Predict Win-Loss In Korean Professional Baseball Games," *Journal of the Korean Institute of Industrial Engineers*, Vol.40, No.1, pp.8-17, 2014.

[13] Hyong Jun Choi and Yun Soo Lee, "The Prediction of Game Outcomes Based On Match Data Within Soccer World Cup," *Korean Journal of Sports Science*, Vol.28, No.1, pp.1317-1325, 2019.

[14] Soo Hyun Cho and Soo Won Lee, "Winner Prediction of A Curling Game based On A Hybrid Machine Learning Model," Master Thesis, Soongsil University Graduate School of Software Specialization: Software 2017. 2, 2017.

[15] Tianxiang Cui, Jingpeng Li, and John Woodward, "An Ensemble Based Genetic Programming System To Predict English Football Premier League games," *IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, Singapore, 2013, pp.138-143, 2013.

[16] Ahmet Emin Saricaoğlu, Abidin Aksoy, and Tolga Kaya, "Prediction of Turkish Super League Match Results Using Supervised Machine Learning Techniques," Intelligent And Fuzzy Techniques In Big Data Analytics And Decision Making. INFUS 2019. *Advances In Intelligent Systems And Computing*, Vol.1029, 2019.

[17] Vincent Hoekstra, Pieter Bison, and Guszti Eiben, "Predicting Football Results With An Evolutionary Ensemble

Classifier," Master Thesis, Business Analytics In VU University, Amsterdam, 2012.
[18] Diederik P. Kingma and Jimmy Ba, "Adam: A Method For Stochastic Optimization," ArXiv preprint ArXiv:1412.6980, 2014.

이 재 현

https://orcid.org/0000-0002-7371-590X
e-mail : dlwogus2525@gmail.com
2017년 상명대학교 미디어소프트웨어학과 (학사)
2018년 ~ 현 재 숭실대학교 융합소프트웨어학과 석사과정
관심분야 : Data Science & Artificial intelligence

이 수 원

https://orcid.org/0000-0001-5863-1188
e-mail : swlee@ssu.ac.kr
1982년 서울대학교 계산통계학과(학사)
1984년 한국과학기술원 전산학과(석사)
1994년 University of Southern California 전산학과(박사)
2003년 ~ 2004년 한국정보과학회 인공지능연구회 분과위원장
2008년 ~ 2009년 한국정보과학회 논문지(SA) 편집위원장
2008년 ~ 현 재 한국BI데이터마이닝학회 부회장
1995년 ~ 현 재 숭실대학교 소프트웨어학부 교수
관심분야 : Data Science & Artificial intelligence