Методы анализа данных в задаче прогнозировани спортивных результатов

А. Д. Журавлев[1]

В данной статье рассматривается задача возможности предсказывать спортивные результаты с помощью методов анализа данных и машинного обучения и определения качества такого прогноза. Также приводитя сравнение построенной модели с моделью используемой букмекерскими конторами.

Ключевые слова: Ключевые слова: машинное обучение, прогнозирование спортивных результатов, анализ данных, классификация.

1. Введение

Данная статья продолжает работу по построению модели, описывающей некоторые параметры хокккейного матча, начатую в [1]. Будет рассмотрен один из возможных подходов к моделированию, использующий в качестве основного инструмента методы машинного обучения. Весь процесс исследования будет происходить на реальных исторических данных.

2. Подготовка данных

2.1. Описание данных

Различные исторические данные по хоккейным матчам предоставляются на официальных сайтах турниров. Например, на сайте www.khl.ru Континентальной Хоккейной Лиги можно найти информацию об игроках, результатах матчей. В частности, можно получить такие данные, как количество забитых шайб командами как в матче целиком, так и за определенный период.

В качестве набора данных была получена статистическая информация о матчах Континентальной Хоккейной Лиги сезонов 2014-2020. Доступными оказались следующие признаки: домашняя и гостевая коман-

[1] *Журавлев Артем Дмитриевич* — аспирант каф. математической теории интеллектуальных систем мех.-мат. ф-та МГУ, e-mail: artemzhuravlev.msu@gmail.com.

Zhuravlev Artem Dmitrievich — graduate student, Lomonosov Moscow State University, Faculty of Mechanics and Mathematics, Chair of Mathematical Theory of Intellectual Systems.

ды, дата матча, итоговый счет, количество бросков, результат каждого из 3-х периодов, заброшенные шайбы в большинстве и меньшинстве, число численных преимуществ, число выигранных вбрасываний, штрафное время и различный коэффициенты БК.

2.2. Отбор переменных и обработка данных

Чтобы улучшить качество переменных, можно использовать генерирование накопленной статистики. Например, подходят следующие признаки: средняя реализация бросков, среднее количество нанесенных бросков, среднее количество заброшенных шайб, среднее количество пропущенных шайб, среднее количество пропущенных бросков. Такие значения накапливаются для каждой команды, исходя из ее предыдущих результатов, и обновляются после каждого сыгранного матча.

Распределение признаков с заброшенными шайбами будет хорошо моделироваться распределением Пуассона - распределение дискретного типа случайной величины, представляющей собой число событий, произошедших за фиксированное время, при условии, что данные события происходят с некоторой фиксированной средней интенсивностью и независимо друг от друга. Введем определение распределения Пуассона:

Выберем фиксированное число $\lambda > 0$ и определим дискретное распределение, задаваемое следующей функцией вероятности:

$$\mathrm{p(k)} \equiv \mathbb{P}(Y = k) = \frac{\lambda^k}{k!}\,\mathrm{e}^{-\lambda},$$

где k! обозначает факториал числа k, e = 2.718281... — основание натурального логарифма. Тот факт, что случайная величина Y имеет распределение Пуассона с математическим ожиданием λ, записывается: $Y \sim \mathrm{P}(\lambda)$.

Таким образом, для каждого матча можем получить значения различных вероятностей, которые его описывают.

Также в признаках присутствуют коэффициенты БК на события, при которых в матче будет забито больше или меньше некоторого значения суммарного количества забитых шайб обеими командами, и само количество - тотал.

В качестве целевой переменной будем рассматривать суммарное количество голов, забитых обеими командами в матче. Целевая переменная будет принимать значение 1, если реальное суммарное количество голов больше, чем тотал, предложенный БК, и значение 0, если меньше, соответственно.

Временно́й ряд (или ряд динамики) — собранный в разные моменты времени статистический материал о значении каких-либо параметров (в простейшем случае одного) исследуемого процесса. Каждая единица

статистического материала называется измерением или отсчётом, также допустимо называть его уровнем на указанный с ним момент времени. Во временном ряде для каждого отсчёта должно быть указано время измерения или номер измерения по порядку.

Так как у нас имеются статистические показатели, которые формируются на основе предыдущих результатов команды, то такие данные необходимо анализировать с помощью временных рядов - для построения алгоритма необходимо использовать только уже прошедшие события до некоторого время t, а для оценки события, которые происходят после.

3. Постановка задачи классификации

Пусть X — множество описаний объектов, Y — множество номеров (или наименований) классов. Существует неизвестная целевая зависимость — отображение $y^*\colon X \to Y$, значения которой известны только на объектах конечной обучающей выборки
$X^m = \{(x_1, y_1), \ldots, (x_m, y_m)\}$. Требуется построить алгоритм $a\colon X \to Y$, способный классифицировать произвольный объект $x \in X$, сопоставив объект $y \in Y$.

Нам необходимо отличать хороший алгоритм классификации от плохого, для этого нам нужно ввести метрики для оценки качества модели. Модель также будем оценивать по тому, сколько денег мы получим или потеряем, если будем ставить согласно ее предсказаниям. Для этого введем следующую метрику: *ROI (return on investment)* - $ROI = \frac{P_n * 100}{s * n}$, где P_n - прибыль на дистанции в n матчей, s - сумма одной ставки, n - количество ставок. То есть ROI является отношением заработанных денег к потраченным. ROI - основной показатель эффективности прогностической модели. Таким образом, будем "играть в плюс"в том случае, если $ROI > 0$.

3.1. Перекресная проверка

Кросс-валидация (cross-validation) - метод оценки аналитической модели и её поведения на независимых данных. При оценке модели имеющиеся в наличии данные разбиваются на k частей. Затем на $k-1$ частях данных производится обучение модели, а оставшаяся часть данных используется для тестирования. Процедура повторяется k раз; в итоге каждая из k частей данных используется для тестирования. В результате получается оценка эффективности выбранной модели с наиболее равномерным использованием имеющихся данных.

Улучшим качество с помощью кросс-валидации. Так как мы будем строить предсказательную модель с помощью временных рядов, то необ-

ходимо правильно использовать кросс-валидацию. Будем сдвигать обучающую выборку на n шагов, а не увеличивать ее, добавляя прошедшие матчи, чтобы не учитываять слишком старые результаты, которые уже потеряли свою информативность и могут только помешать при обучении.

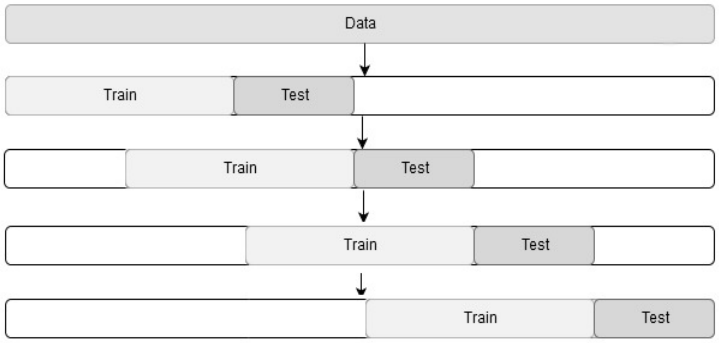

Рис. 1. Кросс-валидация

После проведенного исследования для выбора оптимальных t и n , основываясь на размере доступных исторических данных - 2770 матчей, оказались значения 600 и 100 соответственно - Рис.2. Их и будем использовать для построения модели.

3.2. Предсказание модели

Так как в исходных признаках у нас есть коэффициенты на целевые события от БК - коэффциенты k_0 и k_1 для исхода "Тотал меньше" и "Тотал больше" какого-то количества шайб соответственно, то будем строить вероятностные модели для улучшения предсказательной способности алгоритма, то есть модель возвращает вероятность принадлежности к классу 0 или 1 - p_0 и p_1 соответственно. Введем величины $b_i = p_i * k_i$ для $i = 0, 1$ - отношение вероятности наступления исхода i, предсказанной нашей моделью, к вероятности, которую дает БК. Данные величины - математическое ожидание наших ставок.

Теперь можно рассмотреть различные пороговые значения b для отбора значений b_i, при которых хотим ставить на матч. Такой порог b должен быть больше 1, так как иначе мы будем играть с отрицательным математическим ожиданием. В предположении, что p_i - вероятности событий, то изменяя порог b принятия решений, мы будем влиять на выигрыш.

	train_size	test_size	train_roi
0	500	50	-0.0161334
1	500	100	-0.0295533
2	500	150	0.0188796
3	500	200	-0.033247
4	600	50	-0.0224978
5	600	100	0.0209019
6	600	150	0.0154973
7	600	200	0.0127306
8	700	50	-0.0425073
9	700	100	-0.0155961
10	700	150	-0.0233171
11	700	200	-0.0658967
12	800	50	-0.00506778
13	800	100	-0.00514799
14	800	150	-0.0229847
15	800	200	-0.0341113

Рис. 2. Поиск оптимальных t и n

Поэтому и хотим искать именно вероятности, а не просто построить бинарный классификатор. При верной гипотезе, можно искать подходящие коэффициенты от БК для того чтобы увеличить свой выигрыш.

Перейдем к доказательству гипотезы. Для потенциального улучшения предсказательной способности алгоритма рассмотрим введение следующих пороговых значений: 1, 1.01, 1.02, 1.03, 1.04, 1.05.

4. Техническая реализация

4.1. Поиск наилучшего алгоритма

Разделим выборку на 2 множества, предварительно отсортировав по дате: $X_1 = \{\text{первые 2000 матчей}\}, X_2 = \{\text{оставшиеся 770 матчей}\}$. Модель будет обучаться с помощью ранее введенной кросс-валидации на множестве X_1, а после с соответствующими окнами оценена на множестве X_2, где на первом шаге выборка для обучения будет составлять последние 600 матчей множества X_1. В качестве моделей выбраны сле-

дующие алгоритмы: LightGBM, XGBoost - одни из сильных алгоритмов машинного обучения. Также был использован алгоритм случайного поиска лучших параметров классификатора по сетке - Random Search вместо перебора всевозможных вариантов - Grid Search по причине его трудоемкости.

4.2. Анализ результатов

Для каждого алгоритма проводим перебор гиперпараметров, перебор различных порогов и выбираем лучшую модель, согласно выбранной метрике. Получили следующие результаты - Рис.3

D	E	F	G	H	I	J	K	L
model_typ	feats	bord	test_money	test_pa	test_lo	test_w	test_be	test_roi
lgbm	default	1	-4.37000000000001	189	294	288	582	-0.007508591065292
lgbm	default	1.01	1.07999999999999	235	269	267	536	0.002014925373134
lgbm	default	1.02	8.07	295	236	240	476	0.016953781512605
lgbm	default	1.03	4.55	339	216	216	432	0.010532407407407
lgbm	default	1.04	10.78	386	190	195	385	0.028
lgbm	default	1.05	13.05	426	169	176	345	0.037826086956522

Рис. 3. Поиск лучшего порога - LGBM

Таким образом, можно сделать вывод, что:

1) LightGBM с порогом 1.05 - наилучший алгоритм с точки зрения метрики ROI

2) итоговый ROI на X_2 составил 3.8%- это означает, что если бы мы ставили по предсказниям нашей модели, то переиграли бы БК.

3) При увеличении порога отсечения происходит увеличение метрики ROI

5. Заключение

Итоги проведенного исследования показывают, что методы машинного обучения можно применять для анализа спортивных исторических данных. Также они подтверждают сформулированную гипотезу о том,

что можно увеличить ROI с помощью использования пороговых значений для отбора вероятностных предсказаний.

Полученную модель можно использовать для игры на рынке спортивных ставок против БК. Более того, стоит заметить, что в качестве коэффициентов БК в данных использовались коэффициенты на момент закрытия приема ставок, которые значительно ниже коэффициентов на live-ставки, то есть прибыль может быть увеличена.

Список литературы

[1] Журавлев А.Д., "Возможный подход к задаче прогнозирования спортивных результатов методами анализа данных", *Интеллектуальные системы: Теория и приложения том 25, № 1,*, 2021, 63-69.

[2] https://www.sports.ru/tribuna/blogs/teilnahme/1098481.html, "Хоккейная аналитика".

Data analysis methods in the problem predicting sports results
Zhuravlev A. D.

This article discusses the problem of predicting sports results using data analysis and machine learning methods and determining the quality of such a forecast. Also giving a comparison of the constructed model with the model used by bookmakers.

Keywords: machine learning, sports performance prediction, data analysis, classification.

References

[1] Zhuravlev A. D., "Possible approach to the problem of predicting sports results using data analysis methods", 2021, № Intelligent Systems: Theory and Applications Volume 25, No. 1,, 63-69.

[2] https://www.sports.ru/tribuna/blogs/teilnahme/1098481.html, "Hockey analytics".