Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.8_____

Виланский
Арсений Юрьевич

МОДЕЛИ СРЕДСТВА И МЕТОДЫ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА СПОРТИВНЫХ РЕЗУЛЬТАТОВ

АВТОРЕФЕРАТ

на соискание степени магистра технических наук
по специальности 1-31 80 10 теоретические основы информатики

_____

Научный руководитель
Степанова М.Д.
Кандидат технических
наук, доцент

_____

г. Минск 2017

Нормоконтроль

_____

(фамилия имя, отчество)

_____

(дата, подпись)

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Данная работа посвящена прогнозированию результатов спортивных соревнований. Прогнозирования исхода с матча является вариантом анализа развития социального процесса. Это позволяет надеяться, что опыт, полученный при построении моделей прогнозирования результатов футбольных соревнований, может быть полезен при построении компьютерных систем для анализа разнообразных социальных процессов. Кроме того, прогнозирование результатов спортивных соревнований представляет непосредственный интерес для таких бизнес организаций, как букмекерские компании, а также для всех любителей спорта.

Целью исследования является разработка модели прогнозирования результатов спортивных соревнований, превосходящей существующие модели.

Предметом исследования является модели прогнозирования результатов спортивных соревнований.

Объектом исследования является статистика результатов игр.

В исследовании были поставлены и решены следующие задачи –

- Изучение существующих моделей прогнозирования результатов спортивных соревнований.
- Реализация некоторых из моделей прогнозирования.
- Разработка новой модели прогнозирования результатов и сравнение ее с имеющимися моделями.

Общий объем магистерской диссертации составляет 65 страниц, включая 16 иллюстраций, 3 таблицы, библиографический список из 28 наименований. Работа состоит из четырех глав, введения, заключения и двух приложений.

В первой главе приводится информация о трех применяемых в спорте моделях, а также некоторые необходимые сведения из теории вероятности и математической статистики.

Во второй главе определяется формальная модель предсказания результатов спортивных соревнований и два способа сравнения моделей; подробно рассматривается модель Диксона-Кола и модель рейтинга Эло для футбольных турниров; предлагаются три варианта модификации этой модели; на ос-

новании модифицированных и исходных моделей строится комбинированная модель.

В третьей главе приводятся результаты сравнения рассмотренных моделей. В четвертой главе предложена система прогнозирования спортивных соревнований и рассмотрены средства для реализации такой системы. В заключении приводятся рекомендации по использованию рассмотренных моделей.

В приложениях приводятся примеры статистических данных и фрагменты исходного кода для реализации анализа моделей.

Материалы по теме диссертации были апробированы на XIX Международной научно-практической конференции "Наука и образование в условиях социально-экономической трансформации общества" в г. Минске 01.12.2016 и опубликованы в сборнике материалов данной конференции [1-А.].

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Некоторые математические модели, используемые при прогнозировании результатов спортивных матчей

Модель Диксона-Кола

Одной из популярных моделей для прогнозирования результатов спортивных соревнований является модель (Dixon and Coles model) [2]. Эта модель учитывает атакующие и оборонительные свойства игрока (команды), а также, "гостевой" фактор. Модель основана на пуассоновском распределении.

Пусть имеется N команд (игроков). Пусть команда i принимает на своем поле команду j. Через X_{ij} и Y_{ij} обозначим, соответственно, количество голов забитых принимающей и гостевой командой, α_i – параметр, учитывающий "атакующие" свойства i-ой команды, β_i – параметр, учитывающий "оборонительные" свойства i-ой команды и γ – параметр, учитывающий "домашний фактор".

Тогда вероятность, что i-ая команда забьет X_{ij} голов и j-ая команда забьет Y_{ij} голов в соответствии с моделью Диксона-Колеса равна –

$$Pr(X_{ij} = x, Y_{ij} = y) = \tau_{\lambda_{ij},\mu_{ij}}(x, y)\frac{\lambda_{ij}^x \exp(-\lambda_{ij})}{x!}\frac{\mu_{ij}^y \exp(-\mu_{ij})}{y!}, \qquad (1)$$

где $\lambda_{ij} = \alpha_i\beta_i\gamma$ и $\mu_{ij} = \alpha_i\beta_i$, а

$$\tau_{\lambda,\mu} = \begin{cases} 1 - \lambda\mu\rho, & x = y = 0; \\ 1 + \lambda\rho, & x = 0, y = 1; \\ 1 + \mu\rho, & x = 1, y = 1; \\ 1 - \rho, & x = y = 1; \\ 1, & \text{в остальных случаях.} \end{cases}$$

Параметр ρ, используемый в определении $\tau_{\lambda,\mu}$ принимает значения $\min\{-\frac{1}{\lambda}, -\frac{1}{\mu}\} \leq \rho \leq \max\{\frac{1}{\lambda\mu}, 1\}$ и позволяет вводить оценку для учета зависимости событий: $\rho = 0$ – соответствует независимым событиям.

Параметры α_k, β_k находятся с помощью метода наибольшего правдоподобия.

Система рейтинга Эла

Рейтинг Эла впервые был использован Международной шахматной федерацией в 1970 году для ранжирования шахматистов. Рейтинговая система была предложена Arpad E. Elo. Она частично основана на более ранней работе, выполненной Антоном Хоеслингером (Anton Hoesslinger). Официальное описание системы, как она используется в шахматы можно найти по адресу [11]. Рейтинг используется для прогнозировании вероятности выигрыша.

В данной системе для каждого игрока (команды) вычисляется некоторая величина, характеризующая его (ее) уровень – рейтинг R_p. Основная идея в системе заключается в том, чтобы постоянно менять рейтинг игроков R_p, в зависимости от результатов соревнований.

Важным усовершенствованием Эло к этой системе была добавка, так называемой, "функции вероятности победы" (Win Expectancy Function) W_e, которая определяется как

$$W_e(\Delta R) = \left(1 + 10^{\frac{\Delta R}{\alpha}}\right)^{-1}, \quad \Delta R = R_p - R_e. \tag{2}$$

Здесь R_p – рейтинг игрока, а R_e – рейтинг его противника в мачте, для которого вычисляется функция.

Если игроком с рейтингом R_p играется турнир из M матчей, то вычисляется новый рейтинг $R_{p\ new}$, по правилу

$$R_{p\ new} = R_p + K\left(S - \sum_i W_{e,i}(\Delta R_i)\right), \quad S = N_w + \frac{1}{2}N_t, \tag{3}$$

где R_t – количество связанных игр, K – в принципе постоянная, но в действительности может отличаться в зависимости от турнира.

Рейтинг Месси

Другой системой экспертных оценок является рейтинг Месси [6, 10, 24, 25].

Метод Мэсси основан на решении системы уравнений по методу наименьших квадратов. Основная идея этого метода заключается в следующем:

Составим уравнение $r_i - r_j = y$, где r_i и r_j являются рейтингами команд i и j и y – разностная точка в игре сыгранной между командами i и j.

Если для каждой игры, сыгранной между двумя командами i и j вычислить разностное значение (differentials pont), то получим уравнение

$$\mathbf{I} - \mathbf{J} = \mathbf{S}, \tag{4}$$

где **I** и **J** – очки, заработанные командами i и j, и **S** – счет.

Таким образом, если в общей сложности m игр играют между n командам, то можно построить аналогичное уравнение для каждой игры, в результате получим систему уравнений из m уравнений с n неизвестными. Исходя из этого, можно построить $m \times n$ матрицу (m строк, потому что есть m уравнений, по одному для каждой игры, и n столбцов, по одному для каждой команды). Обозначим эту матрицу **X**.

В каждой строке матрицы **X**, для команд i и j, которые играли друг с другом, проставляется 1 в ячейке i и -1 в ячейках j, указывающая, что эти команды играли друг с другом, и команда i обыграла команду j, и нули на всех остальных.

Для того, чтобы определить рейтинги команд, рассмотрим линейное уравнение $\mathbf{Xr} = \mathbf{y}$, где **X** – наша матрица, **r** – $m \times 1$ вектор рейтингов, которые мы хотим определить и **s** – $m \times 1$ вектор разностных значений для каждой игры.

Хотя система не имеет решения, мы можем умножить обе части матричного уравнения на $\mathbf{X}^T$ (транспонированную матрицу) слева с обеих сторон, и вместо того, чтобы решать исходное уравнения $\mathbf{Xr} = \mathbf{s}$, рассмотрим уравнение

$$\mathbf{X^T Xr} = \mathbf{X^T s}.$$

Для построения системы дающей единственное решение попытаемся привести исходную систему к системе из n уравнений с n неизвестными.

Обозначим $\mathbf{M} = \mathbf{X^T X}$ и вычислим $\mathbf{p} = \mathbf{X^T y}$.

Система $\mathbf{Mr} = \mathbf{p}$ имеет бесконечное число решений. Однако, для того, чтобы определить рейтинг каждой команды, система должна иметь единственное решение.

Для решения этой проблемы, Мэсси предложил заменить последнюю строку матрицы набором из всех единиц и соответствующее разностное значение в векторе **p** на ноль.

Таким, образом вычисление рейтингов сводится к решению уравнения

$$\hat{\mathbf{M}}\mathbf{r} = \hat{\mathbf{p}},$$

где матрица $\hat{\mathbf{M}}$ получена из матрицы $\mathbf{X^T X}$, заменой последней строки на единичную, и вектор $\hat{\mathbf{p}}$ получен из вектора $\mathbf{X^T s}$ заменой n-ого элемента на 0.

$$\mathbf{r} = \hat{\mathrm{M}}^{-1}\hat{\mathrm{p}}.$$

МОДЕЛИ ДЛЯ ПРЕДСКАЗАНИЯ РЕЗУЛЬТАТОВ СОРЕВНОВАНИЙ

Формальная модель для прогнозирования результата спортивной игры и оценка качества прогноза

Пусть имеется $N \in \mathbb{N}$ команд. Результатом некоторого, конкретного спортивного матча между i и j командой представляет собой пару $(g_i, g_j) \in \mathbb{N} \times \mathbb{N}$, где g_i – количество голов забитых i-ой командой в ворота j-ой команды и g_j – количество голов забитых j-ой командой в ворота i-ой команды.

Таким образом спортивный матч, может быть описан набором

$$R_k = \{t_k, i_k, j_k, g_{i_k}, g_{j_k}\},$$

$k = 1, \ldots M, i_k \neq j_k$, где t_k – штамп времени k-ой игры в турнире, i_k – номер (идентификатор) команды, играющей "на своем поле", j_k – номер (идентификатор) команды, играющей "в гостях", M – число всех сыгранных матчей.

Статистикой турнира будем называть множество $\mathbf{T} = \bigcup_{k=1}^{M} R_k$.

По результатам матча, можно получать различные производные показатели. Например, в футболе, команде начисляются очки по правилу

$$S(R_k) = \frac{1 + \mathrm{sign}(g_{i_k} - g_{j_k})}{2}. \tag{5}$$

Модель Диксона-Кола

Модель Диксона-Кола, описанная в разделе 1.3.1, основана на расчете вероятности забитых и пропущенных голов (1.14). Очевидно, что эта модель позволяет предсказывать не только выигрыш (проигрыш) команды но и ничью. Вероятность победы команды (игрока) i над командой (игроком) j будет равна

$$Pr(i,j) = e^{-(\lambda_{ij}+\mu_{ij})} \sum_{k=0}^{\infty} \frac{\mu_{ij}^k}{k!} \sum_{n=1}^{\infty} \frac{\lambda_{ij}^{k+n}}{(k+n)!}, \tag{6}$$

где $\lambda_{ij} = \alpha_i \beta_j \gamma$, $\mu_i j = \lambda_j \beta_i$, α_i, α_j – "атакующие" параметры для i-ой, j-ой команды, β_i, β_j – "оборонительные" параметры для i-ой, j-ой команды, γ – "гостевой" фактор.

На практике, для вычисления вероятности победы по формуле (2.9) нет необходимости рассматривать бесконечные суммы, а достаточно рассмотреть $k = 1, \ldots, 10$, $n = 1, \ldots, 10$, т. е. первые 10 членов каждого ряда, что эквивалентно 100 элементам.

Определим функции предсказания, ошибки и точности предсказания.

$$\mathrm{PREDICT}(i,j) = \mathbf{I}\Big(Pr(i,j) > 1/2\Big). \tag{7}$$

$$Err(i,j,S) = 1 - \mathbf{I}\Big(\mathrm{PREDICT}(i,j) = S\Big), \quad S \in \{0,1\}. \tag{8}$$

$$P_E(T_N) = \frac{1}{N}\sum_{k=1}^{N}\big(1 - Err(i_k, j_k, S_k)\big). \tag{9}$$

Применение рейтинга Эла для прогнозирования результатов игр

Рейтинг Эло, описанный в разделе 1.3.2, позволяет прогнозировать вероятность победы команды, исключая возможность предсказания ничейного результата.

Рассмотрим использование системы рейтингов Эло для предсказания результатов футбольных матчей. Для футбольных соревнований обычно используется несколько измененная процедура определения рейтинга [26]. Рейтинг команды пересчитывается после каждого матча по формуле:

$$R_{p\ new} = R_p + KG\Big(S - W_e(\Delta R + H)\Big), \tag{10}$$

где K некоторая константа, зависящая от турнира, обычно $K = 30$, параметр "функции побед" $\alpha = 400$,

$$S = \begin{cases} 0, & \text{если команда проиграла} \\ 0.5, & \text{если команды сыграли вничью} \\ 1, & \text{если команда выиграла} \end{cases},$$

$$G_i = \begin{cases} 1, & \text{при преимуществе не более чем в 1 гол} \\ 1.5, & \text{при преимуществе в 2 гола} \\ \dfrac{11+m}{8}, & \text{при преимуществе в } m > 2 \text{ голов} \end{cases},$$

$$H = \begin{cases} 100, & \text{если команда играла дома} \\ -100, & \text{если команда играла в гостях} \end{cases}.$$

Для исследования системы Эло, определим функцию PREDICT – предсказание победы команды, играющей дома, рейтинг которой отличается на ΔR от команды противника:

$$\mathrm{PREDICT}(\Delta R) = \mathbf{I}\Big(W_e(\Delta R + H) > 1/2\Big), \tag{11}$$

где $H = 100$.

Определим функцию ошибки, как

$$Err(\Delta R, S) = 1 - \mathbf{I}\Big(\mathrm{PREDICT}(\Delta R) = S\Big), \;\; S \in \{0, 1\}. \tag{12}$$

Тогда, для турнира T, в котором сыграно N матчей введем функцию $P_E(N)$ – отношение верно угаданных результатов к общему количеству сыгранных матчей

$$P_E(T_N) = \frac{1}{N}\sum_{i=1}^{N}\big(1 - Err(\Delta R_i, S_i)\big), \tag{13}$$

Модификация системы Эло за счет одноразового подбора параметров

При определении стандартного рейтинга Эло используются фиксированные значения параметров K, и $\alpha = 400$, а также $H = 100$. Рассмотрим вариант при котором параметры K, α и H подбирается таким образом, чтобы результат прогнозирования некоторого набора матчей из предыдущих сезонов был оптимален. Полученные значения используются для прогнозирования результатов в течение всего текущего сезона.

Для каждого турнира введем два значения – $N_{OptStart}, N_{OptEnd}$ – соответственно номер первого и последнего матча, использующихся для выбора оптимальных параметров. На начало сезона, для всех команд установим рейтинг $R_i = 500$. Оптимальное решение будем искать на множестве Ω :

$$\Omega = \left\{ \begin{array}{rll} & K = 2k, & k = 5, \ldots, 20; \\ (K, \alpha, H), & \alpha = 25a, & a = 2, \ldots, 20; \\ & H = 20h, & h = 0, \ldots, 10. \end{array} \right\} \tag{14}$$

В качестве $N_{OptStart}$ и N_{OptEnd} возьмем соответственно начало и середину текущего сезона.

Подбор параметров после каждого тура

Анализ результатов предыдущего раздела позволяет развить предложенный в предыдущем разделе способ. В предыдущем разделе оптимальный элемент $\omega^* = \{K^*, \alpha^*, H^*\} \in \Omega$ (2.18), рассчитывался один раз перед началом турнира. Можно предположить, что если подбирать ω^* в процессе турнира, то можно улучшить результаты прогнозирования.

Обычно, турнир состоит из нескольких туров, в течение которого, каждая команда сыграет 2 раза, естественно попробовать выполнить обновление ω^*, после каждого тура.

Назовем такой способ – модификацией системы Эло посредством подбора параметров после каждого тура (ППТ).

Композиционный способ модификации рейтинга Эло

Пусть дана некоторая последовательность $\{b_n\}_N$, $b_n \in \mathbb{R}$, состоящая из N элементов и последовательность $\{w_k\}_K$, $k \in \mathbb{R}$ состоящая из K элементов, где $K < N$. Определим оператор сдвигающего среднего $\mathbf{M_a}$ следующим образом:

$$\mathbf{M_a}\big(\{b_n\}_N, \{w_k\}_K\big) = \left\{ \frac{\sum\limits_{i=1}^{K} w_i b_{n-K+i}}{\sum\limits_{i=1}^{K} w_i} \right\}_{N-K+1}. \qquad (15)$$

Последовательность $\{w_k\}_K$ будем называть параметрами оператора сдвигающего среднего или просто параметрами сдвигающего среднего.

Таким образом, результатом применения оператора сдвигающего среднего к некоторой последовательности $\{b_n\}_N$ с параметрами $\{w_k\}_K$ будет последовательность $\{c_m\}_{N-K+1} = \mathbf{M_a}\big(\{b_n\}_N, \{w_k\}_K\big)$, каждый элемент которой получается из K последовательных элементов по правилу:

$$c_m = \frac{\sum\limits_{i=1}^{K} w_i b_{m-K+i}}{\sum\limits_{i=1}^{K} w_i}.$$

В предыдущих разделах были рассмотрены три модели, предсказывающие вероятность исхода матча на основе рейтинга Эло.

Пусть имеется M моделей $\mu_k(i, j, \mathbf{T}_{|t}) = pWin_{i,j} \in \mathbb{R}$, $0 \leq pWin_{i,j} \leq 1$. Для статистики некоторого турнира $\mathbf{T}$, каждая k-ая модель порождает последовательность $\{p_i^{(k)}\}_N$ – предсказанных результатов i-го матча.

Построим для каждого i-го матча $i \leq 5$ упорядоченный набор $\{w_1(i), \ldots, w_M(i)\}$ по следующему правилу

$$w_k(i) = \frac{1}{5} \sum_{m=i-5}^{i-1} \frac{p_m^{(k)}}{\sum_{j=1}^{M} p_i^j}. \tag{16}$$

Композицией моделей будет называть модель, которая на основании предсказанных другими моделями значений будет формировать новое значение как взвешенную сумму предсказаний моделей

$$p(i) = \frac{\sum_{k=1}^{M} w_k(i) p_i^{(k)}}{\sum_{k=1}^{M} w_k(i)}. \tag{17}$$

СРАВНЕНИЕ МОДЕЛЕЙ ПРОГНОЗИРОВАНИЯ

В данной главе приведены результаты сравнения качества различных моделей прогнозирования, полученные на статистике разных сезонов и турниров.

Сравнение выполнялось как на основе (2.7), так и (2.8).

Анализ поведения моделей на основе статистических данных различных турниров

В разделе 2.7 был определен композиционный способ предсказания результата спортивного соревнования. Данным способом, комбинировались результаты моделирования по трем рассмотренным ранее системам. Очевидно, что этот способ легко дополнить четвертой моделью, созданной на основе системы Диксона-Кола.

Определим весовые коэффициентов w_4 в формуле 2.20, рассчитывая p_m по формуле 2.9 и будем вычислять $p(i)$ – вероятность победы в i-ом матче принимающей команды по четырем моделям: стандартной Эло, ППТ,ОПП и Диксона-Кола. Как и ранее будем называть этот способ композиционным.

В работе представлена таблица с значениями для количества верно предсказанных результатов, полученная посредством применения различных моделей для первой 1000 матчей всех сезонов турнира Английская Премьер-лига. Данные таблицы свидетельствуют о преимуществе композиционной модели над другими рассмотренными моделями.

Заметим, что все рассмотренные модели предсказывают результат лучше, чем простое угадывание.

СРЕДСТВА РЕАЛИЗАЦИИ МОДЕЛЕЙ ПРЕДСКАЗАНИЯ

Система для предсказания результатов игр

В предыдущих частях были рассмотрены несколько моделей предсказания результатов спортивных соревнований. Эти модели могут быть созданы различными способами, но любая система, реализующая их, будет иметь компоненты, обеспечивающие получение результатов спортивных соревнований, подготовку данных, их обработку для выдачи предсказания, модуль проверки истинности предсказания, а также, возможно, модуль корректировки системы

Такая система может быть интересна любителям спорта, букмекерам, исследователям социальных процессов.

Эта система должна обрабатывать большие объемы данных, и реализовать вычислительно сложные алгоритмы. При этом не обойтись без определенного инструментария.

В процессе подготовки данной работы был реализован прототип такой системы. Данный прототип построен по модульному принципу. Функциональная схема системы представлена на рис. 0.1.

ЗАКЛЮЧЕНИЕ

В диссертации были изучены современные модели прогнозирования результатов спортивных соревнований.

Предложена формальная модель для прогнозирования результатов спортивных соревнований и два критерии сравнения различных моделей.

На базе стандартной модели Эло и модели Диксона-Кола были разработаны три новых модели прогнозирования результатов футбольных матчей : модификация стандартной модели Эло посредством одноразового подбора параметров (ОПП), модификация системы Эло посредством подбора параметров

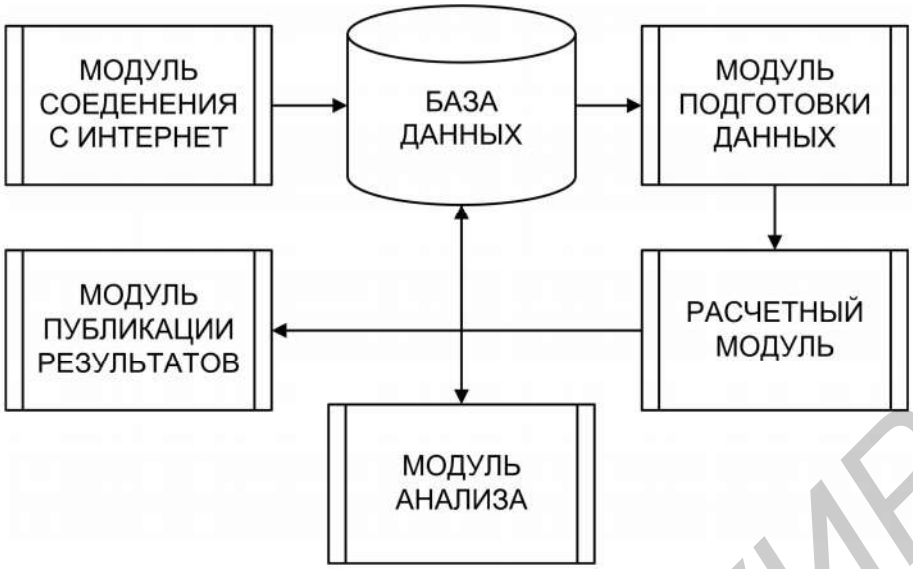

Рисунок 0.1 – Функциональная схема системы, реализующей модель предсказания результатов соревнования

после каждого тура (ППТ) и композиционная модель, позволяющая использовать для предсказания результатов совокупность нескольких моделей.

Данные модели были реализованы и проверены на статистике нескольких футбольных турниров. Результаты проверки показывают, что предложенные модели, в среднем, превосходят стандартную модель Эло и модель Диксона-Кола. При этом композиционная модель, в среднем, лучше ППТ, ППТ, в свою очередь, в среднем, лучше ОПТ. Кроме того, синтезированная композиционная модель почти всегда однозначно лучше других моделей.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

Виланский А.Ю., Прогнозирование результатов футбольных соревнований с помощью модифицированного рейтинга Эло. / А.Ю. Виланский // Наука и образование в условиях социально-экономической трансформации общества: материалы XIX межд. научн.-практ. конф.: Частное учреждение образования "Институт современных знаний имени А.М. Широкова"; – Минск, 2016. С. 299-302.