

УДК 004.691.54

Выбор и настройка параметров модели классификации спортсменов для формирования оптимальных по составу велокоманд

Землянская С.Ю., Крайний Д.С.
Донецкий национальный технический университет
zsaa@ya.ru, krdima92@gmail.com

Землянская С.Ю., Крайний Д. С. Выбор и настройка параметров модели классификации спортсменов для формирования оптимальных по составу велокоманд. Описан принцип классификации спортсменов, проведен анализ результатов работы различных методов классификации на различных входных наборах данных, разработана модель классификации спортсменов с учетом значимых параметров на основе нейронной сети.

Постановка задачи

Задача формирования состава велокоманды заключается в корректном подборе спортсменов на конкретную гонку. Для этого необходимо определиться со специализацией спортсменов и специализацией гонки. Выбор специализации велоспортсмена можно сделать, основываясь на результатах его предыдущих тренировок, а специализацию гонки определить по её характеристикам. Таким образом, необходимой частью процесса формирования состава команды велоспортсменов на гонку является решение задачи классификации.

На данный момент процесс учета физического состояния профессионального спортсмена является обязательным фактором для достижения максимального результата. Это связано с тем, что профессиональный велоспорт характеризуется высокой конкурентностью между спортсменами [3], что определяет высокие требования к уровню физической подготовки.

Актуальность поставленной проблемы заключается в том, что благодаря развитию компьютерных технологий все данные, характеризующие тренировку и физическое состояние спортсмена, находятся в электронном виде. Это в свою очередь позволяет легко манипулировать ими, сводить данные в таблицы и графики, а также производить расчеты для составления рекомендаций спортсмену.

Учитывая вышеуказанное, разработка модели, позволяющей оценить степень соответствия физического состояния спортсмена той или иной гонке для достижения максимальных результатов, является актуальной.

Применяя модель классификации спортсменов, можно оценить физическую подготовку спортсмена на данный момент, на

основании временной статистики его тренировок. Эта информация может использоваться как для

составления спортивного календаря на год, так и для распределения ролей гонщиков в команде.

Цель классификации заключается в выявлении специализации спортсмена, что может помочь спортивному директору выбрать максимально эффективный состав спортсменов на определённую гонку, чтобы команда смогла показать лучший результат, что, вероятно, приведет к новым вложениям спонсоров [4].

Цель статьи

Целью данной статьи является выбор метода для решения задачи классификации велоспортсменов по возможным специализациям, а также настройка работы выбранного метода, посредством изменения различных параметров для достижения максимально точного результата.

Начальные исследования

Поскольку возможность получить для анализа реальные данные отсутствует, входные параметры генерируются в диапазонах значений, согласованных с экспертом в данной области.

Выделены три основные классификации спортсменов – спринтер, горняк, раздельщик. Эти классификации характеризуются определенным набором правил, представляющих собой верхние и нижние границы для анализируемых характеристик спортсмена, полученных при моделировании. Например, значение текущей скорости у горняков при наборе более 300 метров будет варьироваться от 25 до 35 км/ч, в то время как спринтеры смогут показать там скорость 10-20 км/ч, а раздельщики 15-25 км/ч. Подобным образом генерируются значения всех входных данных нейронной сети, основываясь на наборе либо потере высоты и теоретической специализации гонщика.

В качестве входного набора для метода классификации будут использоваться следующие параметры: ID (уникальный номер спортсмена), возраст, вес, рост, набор высоты, потеря высоты, средняя скорость, максимальная скорость, средняя

мощность, максимальная мощность, средний пульс, максимальный пульс, средний каденс, максимальный каденс. Последние 10 показателей относятся только к одной тренировке.

Имея входной набор параметров, можно протестировать различные методы классификации чтобы выбрать наиболее подходящий для решения поставленной задачи классификации [6]. Входной файл будет представлять собой таблицу, где все столбцы, кроме последнего – характеристика спортсмена, а последний столбец – ожидаемый результат, если он необходим для обучения.

Формализация классификации спортсменов

Классификация, в данном случае представляет собой выбор подходящей специализации спортсмена, данный выбор делается, основываясь на диапазонах значений всех 14 входных параметров, описанных ниже. Принадлежность спортсмена к той или иной специализации определяется на основании попадания параметров тренировки в определенные диапазоны:

$$\begin{aligned} N_i = (&lbA_i \leq A \leq ubA_i) \& \\ (&lbH_i \leq H \leq ubH_i) \& (&lbW_i \leq W \leq ubW_i) \\ (&lbD_i \leq D \leq ubD_i) \& (&lbEG_i \leq EG \leq ubEG_i) \\ (&lbEL_i \leq EL \leq ubEL_i) \& (&lbMS_i \leq MS \leq ubMS_i) \\ (&lbAS_i \leq AS \leq ubAS_i) \& (&lbMPo_i \leq MPo \leq ubMPo_i) \\ (&lbAPo_i \leq APo \leq ubAPo_i) \& (&lbMPu_i \leq MPu \leq ubMPu_i) \\ (&lbAPu_i \leq APu \leq ubAPu_i) \& (&lbMC_i \leq MC \leq ubMC_i), \end{aligned} \quad (1)$$

где N_i – принадлежность спортсмена к i -ому типу специализации, префиксы lb и ub представляют собой значения нижней и верхней границы рассматриваемого параметра, соответственно, D – значение длительности тренировки, A – переменная, обозначающая возраст спортсмена, H – значение роста спортсмена, W – значение веса спортсмена, EG – набор высоты за тренировку, EL – потеря высоты за тренировку, MS – максимальная скорость за тренировку, AS – средняя скорость за тренировку, MPo – максимальная мощность за тренировку, APo – средняя мощность за тренировку, MPu – максимальный пульс за тренировку, APu – средний пульс за тренировку, MC – максимальный каденс за тренировку, AC – средний каденс за тренировку.

Стоит отметить, что классификация выполняется с учетом всех тренировок, затем данные группируются по спортсменам, и для каждого спортсмена считается среднее значение его специализации по формуле:

$$Spec = \frac{\sum_{i=1}^N s_i}{N}, \quad (2)$$

где $Spec$ – окончательное значение специализации для спортсмена, которое в дальнейшем используется в процессе формирования состава велокоманд, s_i – специализация спортсмена, полученная по результатам i -ой тренировки, N – количество проведенных тренировок.

Выбор метода решения проблемы

Для решения задач классификации существует множество методов [5], рассмотрим несколько наиболее популярных среди них и выберем наиболее подходящий для решения поставленной задачи.

Метод k взвешенных ближайших

Метод k взвешенных ближайших соседей – это метрический алгоритм классификации, основанный на оценивании сходства объектов. Классифицируемый объект относится к тому классу, которому принадлежат ближайшие к нему объекты обучающей выборки [9].

Пусть $X \in R^n$ – множество объектов; Y – множество допустимых ответов. Задана обучающая выборка:

$$\{(x_i, y_i)\}_{i=1}^l \quad (3)$$

Задано множество объектов:

$$X^m = \{x_i\}_{i=1}^m \quad (4)$$

Требуется найти множество ответов y_i для объектов x_i ;

На множестве объектов задается евклидова функция расстояния:

$$p(x, x') = \sum_{i=1}^n (x_i - x'_i)^2 \quad (5)$$

Для произвольного объекта x расположим объекты обучающей выборки x_i в порядке возрастания расстояний до x (6).

$$p(x, x_{1;x}) \leq p(x, x_{2;x}) \leq \dots \leq p(x, x_{m;x}), \quad (6)$$

где через $x_{i;x}$ обозначается тот объект обучающей выборки, который является i -м соседом объекта x . Аналогичное обозначение введём и для ответа на i -м соседе: $y_{i;x}$.

Таким образом, произвольный объект x порождает свою перенумерацию выборки. В наиболее общем виде алгоритм ближайших соседей есть (7).

$$a(x) = \arg \max \left(\sum_{i=1}^m [x_{i;x} = y] \omega(i, x) \right), \quad (7)$$

где $\omega(i, x)$ – заданная весовая функция, которая оценивает степень важности i -го соседа для классификации объекта. Так, при $\omega(i, x) = 1$ при $i \leq k$ алгоритм соответствует методу k ближайших соседей. Но в задаче с

несколькими возможными ответами максимальная сумма голосов может достигаться на нескольких классах одновременно. Неоднозначность можно устранить, если в качестве весовой функции взять нелинейную последовательность, например, геометрическую прогрессию: в рассматриваемом примере:

$$\omega(i, x) = [i \leq k] * q^2, \quad (8)$$

что соответствует методу k экспоненциально взвешенных ближайших соседей, причем предполагается:

$$0.5 \leq q \leq 1 \quad (9)$$

Но, несмотря на простоту реализации, данный алгоритм приводит к неэффективному расходу памяти и чрезмерному усложнению решающего правила в следствии необходимости хранения обучающей выборки целиком и сравнения классифицируемого объекта со всеми объектами выборки, а в данном случае она представляет собой значительный объем данных.

Иерархические методы кластеризации различаются правилами построения кластеров.

В качестве правил выступают критерии, которые используются при решении вопроса о "схожести" объектов при их разделении на группы (дивизимные методы) [7].

Суть дивизимной иерархической кластеризации состоит в последовательном разделении больших кластеров на меньшие.

Критерием для определения схожести и различия кластеров является расстояние между точками на диаграмме рассеивания. Это сходство можно "измерить", оно равно расстоянию между точками на графике. Способов определения меры расстояния между кластерами, называемой еще мерой близости, существует несколько. Наиболее распространенный способ - вычисление евклидова расстояния между двумя точками i и j на плоскости, когда известны их координаты X и Y :

$$D_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}, \quad (10)$$

чтобы узнать расстояние между двумя точками, надо взять разницу их координат по каждой оси, возвести ее в квадрат, сложить полученные значения для всех осей и извлечь квадратный корень из суммы.

Когда осей больше, чем две, расстояние рассчитывается таким образом: сумма квадратов разницы координат состоит из столько слагаемых, сколько осей (измерений) присутствует в нашем пространстве. Например, если нам нужно найти расстояние между двумя точками в пространстве трех измерений, формула (9) приобретает вид:

$$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}, \quad (11)$$

В начале работы алгоритма все объекты принадлежат одному кластеру, который на последующих шагах делится на меньшие кластеры, в результате образуется последовательность расщепляющих групп.

Иерархические алгоритмы связаны с построением дендрограмм (от греческого dendron - "дерево"), которые являются результатом иерархического кластерного анализа [10]. Дендрограмма описывает близость отдельных точек и кластеров друг к другу, представляет в графическом виде последовательность объединения (разделения) кластеров.

Принцип работы описанного выше метода в виде дендрограммы показан на рис. 1.

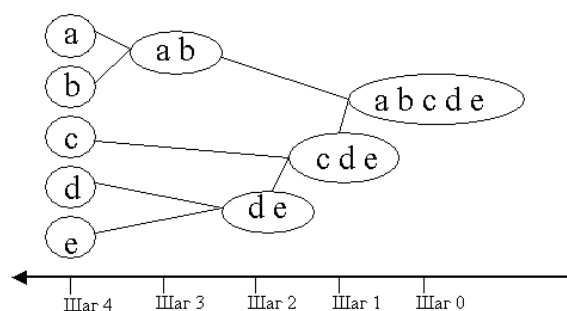


Рисунок 1 – Принцип работы метода деления иерархического дерева кластеров

Преимуществом иерархических методов кластеризации является их наглядность, однако иерархические методы кластерного анализа используются при небольших объемах наборов данных.

Искусственная нейронная сеть

Искусственная нейронная сеть или просто нейронная сеть — это математическая модель, а также ее программные или аппаратные реализации, построенная в некотором смысле по образу и подобию сетей нервных клеток живого организма [1].

Как и линейные методы классификации и регрессии, по сути нейронные сети выдают ответ вида:

$$y(x, \omega) = f \left(\sum_{j=1}^N \omega_j * \varphi_j(x) \right), \quad (12)$$

где f — нелинейная функция активации, ω — вектор весов, φ — нелинейные базисные функции. Обучение нейронных сетей состоит в настройке весов, а также базисных функций.

Несмотря на недостатки нейронных сетей вроде возможной тупиковой ситуации [2] после обучения, они имеют ряд достоинств, важных при

решении данной задачи: решений задач при неизвестных закономерностях – система получает на вход множество параметров, взаимосвязь между которыми сложно отследить; устойчивость к шумам во входных данных – плохие погодные условия на тренировках, плохое самочувствие спортсмена и множество подобных факторов может внести неясность при принятии решения; адаптированные к изменениям – со временем спортсмен может переклассифицироваться в гонщика другой специализации, что будет видно по данным его тренировок; высокое быстродействие – система должна работать с множеством показателей каждой тренировки для каждого спортсмена команды, следовательно быстродействие будет являться значительным плюсом.

Существует ряд нейронных сетей, подходящих для выполнения задачи классификации.

Выбор метода классификации

Для выбора метода классификации сравним результаты работы выбранных методов классификации с ожидаемым значением специализации велоспортсмена. В качестве входных данных используем тестовую выборку из 50 тренировок нескольких спортсменов, сгенерированную с учетом заданных границ значений параметров спортсмена и тренировки для разных специализаций. Для наглядности рассчитаем среднюю ошибку результатов, как сумму разниц в значениях специализации деленную на количество спортсменов. Результаты сравнения методов классификации приведены в таблице 1.

Основываясь на средней ошибке результатов работы методов классификации, которые приведены в таблице 1, можно сделать вывод, что для решения поставленной задачи лучше всего подходит нейронная сеть.

Таблица 1. Сравнение методов классификации

Значение	Средняя ошибка
Метод k взвешенных ближайших	0,667
Иерархический метод кластеризации	1,085
Нейронная сеть	0,033

Выбор структуры и параметров нейронной сети

Анализ результатов заключается в том, чтобы сравнить показатели производительности и

погрешность нейронных сетей, и на основании этих данных определиться с типом нейронной сети и сгенерировать наиболее подходящую структуру для решения поставленной задачи.

Определим, как влияет количество входных данных нейросети на полученные результаты. Запустим модуль STATISTICA Neural Network и в качестве входного набора данных укажем файл с 25 тренировками различных спортсменов. В таблице 2 приведены результаты работы лучших нейронных сетей различных типов.

Таблица 2. Описание нейронных сетей с входным набором из 25 записей

Type	Error	Inputs	Hidden	Performance
MLP	0.38962	6	4	0.02576
RBF	0.91534	4	4	0.10245
PNN	0.40613	13	13	0.41798

В таблице 2 столбец Type обозначает тип нейронной сети, в данном случае приведены 3 типа сетей с лучшими результатами, MLP – многослойный персептрон, RBF – радиал-базисные функции, PNN – вероятностная нейронная сеть [8]. Столбец Error обозначает значение средней ошибки на выходе нейронной сети, столбец Inputs – количество входов нейронной сети, столбец Hidden – количество скрытых слоёв, столбец Performance – производительность нейронной сети, что является показателем скорости её работы.

На основании данных таблицы 2 можно сказать что для поставленной задачи лучше всего подходит нейронная сеть с топологией многослойный персептрон.

Теперь повторим операцию, только в качестве входного набора данных используем файл с 50 тренировками и сравним результаты.

Таблица 3. Описание нейронной сети с входным набором из 50 записей

Type	Error	Inputs	Hidden	Performance
MLP	0.19131	13	8	0.1597355

Сравнивая лучшую нейронную сеть из таблицы 2 и нейронную сеть таблицы 3 можно отметить тот факт, что средняя ошибка выходных данных значительно уменьшилась, более чем в 2 раза, а также возросла производительность. Также количество входов нейронной сети увеличилось до 13, что говорит о том, что в первом случае выбор делался, используя меньше половины входных параметров, что могло приводить к ошибочным результатам. Но, даже с увеличением входного набора данных нейронная сеть не использует все 14 входных параметров, что говорит о незначимости одного из них, либо об отсутствии пересечения границ значений некоторых параметров, что

искусственно делает неучитываемый параметр не значимым. В данном случае незначимым оказался рост спортсменов, что противоречит поставленным данным классификации гонщиков. Пересмотрим границы параметров классификаций в программе генерации тренировок и запустим модуль STATISTICA Neural Network с полученным набором данных. Результаты работы новой нейронной сети на отредактированных данных приведены в таблице 4.

Таблица 4. Описание окончательной нейронной сети с входным набором из 50 записей

Type	Error	Inputs	Hidden	Performance
MLP	0.1162049	14	8	0.1636307

Данная нейронная сеть использует все входные параметры для своей работы, также имеет небольшую погрешность, учитывая вариативность выходных данных. Точность работы нейросети может быть увеличена при тренировке нейронной сети п квазиныютоновскому методу.

Выходным значением нейронной сети является специализация спортсмена в числовом виде: 1 – спринтер, 2 – горняк, 3 – раздельщик.

Структура полученной нейронной сети изображена на рисунке 2.

Литература

1. Беркинблит М. Б. Нейронные сети. М.: МИРОС и ВЗМШ РАО, 1993. – 254 с.
2. Уоссермен. Ф. Нейрокомпьютерная техника: Теория и практика, 1992. – 152 с.
3. Шеннон Совндаль, Анатолия велосипедиста – Попурри, 2011. – с. 89-93.
4. Полищук Д.А. «Велосипедный спорт», 1997. – 203 с.
5. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: классификация и снижение размерности. – М.: Финансы и статистика, 1989. – с. 55-74.
6. Вапник В.Н. Восстановление зависимостей по эмпирическим данным. – М.: Наука, 1979. – 119 с.

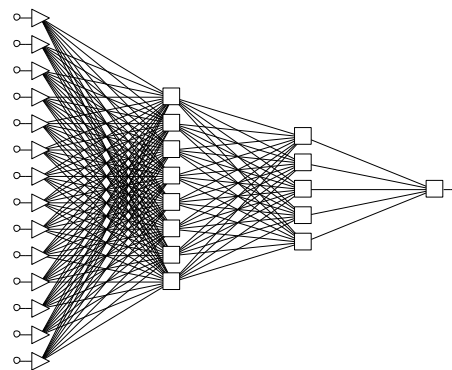


Рисунок 2 – Структура нейронной сети

Выводы

В статье описаны существующие методы классификации, осуществлен выбор метода и параметров нейронной сети, описаны входные и выходные данные нейронной сети, приведены таблицы с результатами точности и производительности нейронных сетей, проведен анализ полученных значений, приведено описание исследуемого объекта.

В дальнейшем результаты исследования планируется использовать при реализации системы поддержки принятия решения по формированию состава велокоманды на предстоящие гонки.

7. Жамбю М. Иерархический кластер-анализ и соответствия: Пер. с фр. М.: Финансы и статистика, 1988. – 342 с.

8. Круглов В.В., Борисов В.В. Искусственные нейронные сети. Теория и практика. – М.: Горячая линия - Телеком, 2001. – 382 с.

9. Бериков В.С., Лбов Г.С. Современные тенденции в кластерном анализе // Всероссийский конкурсный отбор обзорно-аналитических статей по приоритетному направлению «Информационно-телекоммуникационные системы», 2008. — 26 с.

10. Олдендерфер М.С., Блэшфилд Р. К. Кластерный анализ // Факторный, дискриминантный и кластерный анализ: пер. с англ.; Под. ред. И. С. Енюкова. – М.: Финансы и статистика, 1989 – 215 с.

Землянська С.Ю., Крайній Д. С. Вибір і налаштування параметрів моделі класифікації спортсменів для формування оптимальних за складом велокоманд. Описано принцип класифікації спортсменів, проведено аналіз результатів роботи різних методів класифікації на різних вхідних наборах даних, розроблено модель класифікації спортсменів з урахуванням значущих параметрів на основі нейронної мережі.

Ключові слова: велоспортсмен, спеціалізація, модель, параметри, класифікація, нейронна мережа

Zemlyansky S. Y., Krayniy D. S. Sportsmen classification model selection and parameters tuning for the Cycling teams optimal formation. The principle of sportsmen classification is described, the analysis of different classification methods results on different input data sets is made, the sportsmen classification model with relevant parameters based on neural network is developed.

Keywords: cyclist, racer specialization, model, parameters, classification, neural network

Стаття поступила в редакцію 20.09.2015

Рекомендована к публікації д-ром техн. наук В.Н. Павлышом