3. Рябинин И. А. Структурно-сложные системы и их формализация с помощью функций алгебры логики // Биосфера. 2011. Т. 3, № 4. С. 455–461.

4. Гарсиа-Молина Г., Ульман Дж. Д., Уидом Дж. Системы баз данных. Полный курс. М.: Издательский дом «Вильямс», 2003. С. 839–841.

5. Документация PostgeSql. URL: https://www.postgresql.org/docs/12/index.html (дата обращения 20.12.2019).

6. База данных «ключ–значение». URL: https://en.wikipedia.org/wiki/Key-value_database (дата обращения 20.12.2019).

7. Садаладж П. Дж., Фаулер М. NoSQL Новая методология разработки нереляционных баз данных. М.: Издательский дом «Вильямс», 2013. С. 101–102.

8. Кораблев Ю. А. Проектирование гибридных интеллектуальных систем отказоустойчивого управления. СПб., 2019. С. 3.

9. Попович В. В. Интеллектуальные географические информационные системы. СПб.: Наука, 2013. С. 37–50.

10. Кондратьев С. А., Сычев И. О. Применение семантических технологий для хранения и обработки данных в корабельных информационно-управляющих системах // Морская радиоэлектроника. 2019. № 2 (68). С. 38–41.

11. Сычев И. О., Кораблев Ю. А., Звягин Л. С. Применение семантических технологий для обработки данных в геоинформационных системах // Всерос. науч. конф. по пробл. управления в технических системах. СПб.: Изд-во СПбГЭТУ «ЛЭТИ», 2019. Т. 1. С. 322–325.

I. O. Sychev, S. A. Kondratev, Ju. A. Korablev
Saint Petersburg Electrotechnical University

LOGGING MANAGEMENT METHODS IN INFORMATION MANAGEMENT SYSTEMS BASED ON EMBEDDED DATABASES

The issues of information logging management in information management systems are considered. Relevance of logging issues is related to the need to store and process large amounts of data in systems of various classes, particularly in automated monitoring and diagnostics systems. The general requirements for loggings software module are defined. The analysis of methods for log storage and tools for their collecting and processing are carried out. The method for storing logs using «key-value» database is proposed. The implementation variants using DBMS limbmdbx and SQLite are considered. Storage structure for logs is developed. Logging software module is developed; module architecture is carried out.

Documenting, logging, data base, diagnostics, NoSQL

УДК 519.67

Д. В. Козлов, Я. А. Бекенёва
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В. И. Ульянова (Ленина)

Интеллектуальный анализ данных в области определения пригодности собак к выполнению специфической службы

Проведение различного рода тестирований широко используется в качестве инструмента оценки в различных областях деятельности. Результаты тестирований, как правило, хранятся в специальных базах и содержат информацию о тестируемых объектах, результатах для каждого отдельного теста и итоговый результат. Объем данных о результатах тестирований может быть достаточно большим и с трудом поддаваться обработке вручную. Объектом разработки и исследования служат математические модели классификации, применяемые к исходному набору данных с характеристиками собак, выполняющих определенный род деятельности. Целью статьи является исследование возможности применения методов классификации для оценки пригодности собак к выполнению службы собаки-поводыря. Были исследованы несколько методов классификации: k ближайших соседей, дерево классификации, логистическая регрессия. Разработана программа для подготовки исходного набора данных к дальнейшему интеллектуальному анализу с помощью выбранных методов. Проведены эксперименты, связанные с оценкой точности каждого метода, а также выявлены критерии, имеющие наибольшее влияние на результаты тестирования собак.

Классификация, метод *k* ближайших соседей, дерево классификации, логистическая регрессия

Одним из основных подходов к машинному обучению является обучение с учителем (supervised learning) [1]. Данный подход характеризуется наличием некоторой выборочной совокупно-

сти, каждый элемент которой представляет собой набор данных – например, набор характеристик, временных сигналов или изображений. Каждому набору данных соответствует определенный отклик. Различают несколько видов откликов, и для каждого из них существует определенный класс задач [1].

В случае, когда множество откликов безгранично (например, ответом может быть любое вещественное число), для построения математической модели применяются методы регрессионного анализа.

Когда по отклику можно характеризовать дальнейшее поведение нужного процесса, модель, построенная по данной выборочной совокупности, решает задачу прогнозирования.

В случае, когда имеется ограниченное число вариантов отклика, принято говорить о задачах классификации.

В машинном обучении проблема классификации представляет собой проблему идентификации категории, к которой принадлежит элемент, по его входным характеристикам. В качестве базиса для идентификации используется тренировочное множество (выборочная совокупность), для каждого элемента которого уже определена категория. Математическая модель, решающая задачу классификации, называется классификатором [2].

Существует еще несколько подходов к машинному обучению – обучение без учителя (unsupervised learning), обучение с подкреплением (reinforcement learning) и т. п., но они в данном исследовании не рассматриваются ввиду того, что не подходят для решения требуемой задачи.

Метод *k* ближайших соседей. kNN (k-Nearest Neighbors) – это метрический алгоритм построения классификатора или модели [3]. Основная идея алгоритма состоит в применении пространства признаков, где каждое измерение представляет собой входную характеристику – независимую переменную. Классификатор хранит в пространстве признаков точки из тренировочного множества, для каждой из которых определен класс. При получении входного набора характеристик классификатор строит по ним точку в пространстве признаков и для определения класса берет тот, который чаще всего встречается среди k ближайших соседей этой точки из тренировочной выборки.

Для проверки точности метода kNN происходит изменение двух параметров. Первый представляет собой отношение размера тренировочной выборки к размеру тестовой, а второй – количество соседей k. Меняя эти параметры для дальнейшего сравнения с другими методами, возьмем параметр с наилучшими результатами точности. Также для сравнения с логистической регрессией будет проведен эксперимент с двумя состояниями независимой переменной.

Для теста возьмем размеры тренировочной выборки от 25 до 75 % с шагом в 5 % и для каждого размера – классификаторы с различным значением k. Для верификации модели сравним предсказанные значения из тестовой выборки с реальными – в качестве оценки точности модели будет выступать процент попаданий; аналогичную проверку проведем и для тренировочной выборки.

В качестве точности модели с конкретным размером выборки возьмем наилучшую точность среди всех проверяемых значений k. На рис. 1 представлен график зависимости точности полученной модели от размера выборки для тренировочного (кривая *1*) и тестового (кривая *2*) множеств.

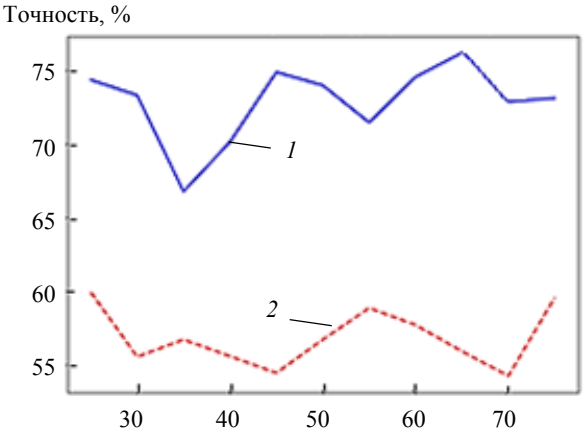

Рис. 1

Падение точности говорит о том, что в выборке имеются объекты, которые по каким-то причинам находятся в разных классах несмотря на одинаковые или похожие характеристики. Аналогичное изменение точности, характеризуемое кривой *1* на рис. 1, можно увидеть и при проверке самой тренировочной группы на данной модели, что только подтверждает наличие шумовых объектов в исходных данных.

Поиск оптимального количества соседей. Для поиска оптимального количества соседей проведем проверку точности модели на различных значениях k. Для этого возьмем значения в пределах от 3 до 9. На рис. 2 представлен график

зависимости точности получившейся модели от значения k при оптимальном размере тестовой выборки, где кривая *1* иллюстрирует результаты проверки тренировочной, а кривая *2* – тестовой выборок.

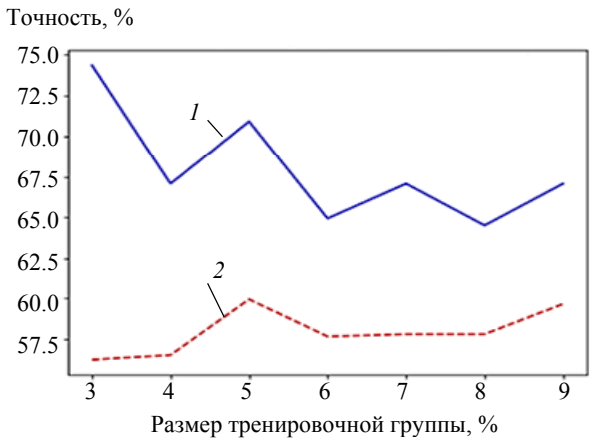

Рис. 2

Как можно видеть по графикам на рис. 2, при проверке тестовой выборки наибольшая точность достигается при $k = 5$. Также можно заметить, что точность тренировочной выборки падает при увеличении k.

Дерево классификации. Дерево классификации представляет собой набор узлов и листьев. Каждый узел содержит в себе условие, связанное с определенным признаком, от выполнения или невыполнения которого для входного элемента зависит, на какой дочерний узел перейдет этот элемент [4]. Каждый лист содержит в себе информацию о том, к какому классу требуемый элемент принадлежит. Таким образом, переходя от узла к узлу, проверяя соответствующие параметры входного элемента, можно определить, к какому классу он относится, дойдя до одного из листьев дерева. К основным проблемам деревьев классификации относится слишком большая ветвистость при большем количестве признаков, из-за чего могут возникать правила, которые распространяются только на несколько объектов и, соответственно, имеют низкую ценность. Для решения данной задачи применяется отсечение ветвей. Преимущество деревьев классификации состоит в удобстве интерпретации, а также в возможности выделить с их помощью наиболее информативные признаки.

Проверим работу алгоритма дерева решений для всех классов. Аналогично тестированию алгоритма kNN найдем для него оптимальный размер тренировочной выборки – график зависимости точности от размера выборки, показан на рис. 3.

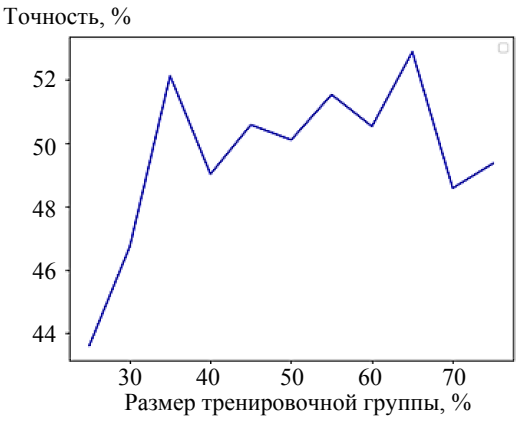

Рис. 3

Большой разброс точности говорит о плохой устойчивости алгоритма к шумовым объектам. Также из рисунка видно, что максимальная точность значительно ниже, чем для модели, построенной алгоритмом k ближайших соседей, и достигает лишь 54 %. Это обусловлено большой ветвистостью смоделированного дерева, что сказывается на его точности. В дереве много поддеревьев, до которых из тренировочной выборки доходят лишь несколько объектов. Пример такого поддерева представлен на рис. 4.

В узлах поддерева сверху обозначено условие, по которому происходит деление. Правый потомок обозначает невыполнение условия, а ле-

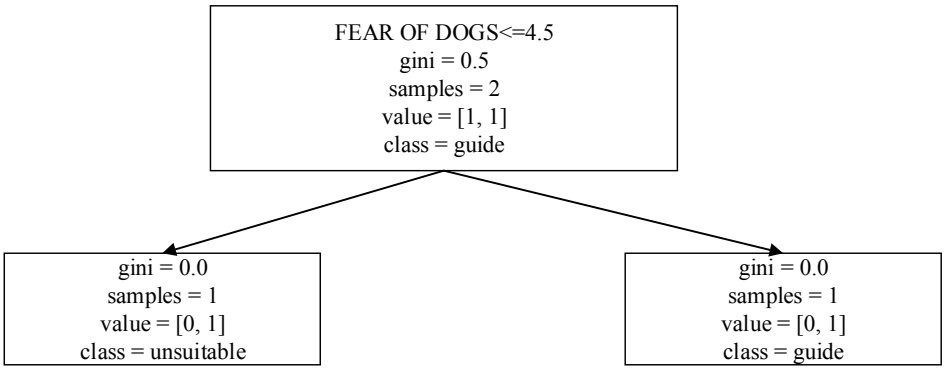

Рис. 4

вый – его выполнение. Ниже обозначены индекс gini, количество объектов, которые дошли до данного узла, их распределение по классам и название класса, которому принадлежит наибольшее количество объектов.

Из-за того что при моделировании возникает большое количество поддеревьев, аналогичных представленному на рис. 4, многие объекты из тестовой выборки имеют больше шансов попасть в лист неправильного класса.

Теперь проверим данный алгоритм для двух классов: ставшая поводырем, и не прошедшей тестирование. На рис. 5 представлен график зависимости точности от размера тренировочной выборки.

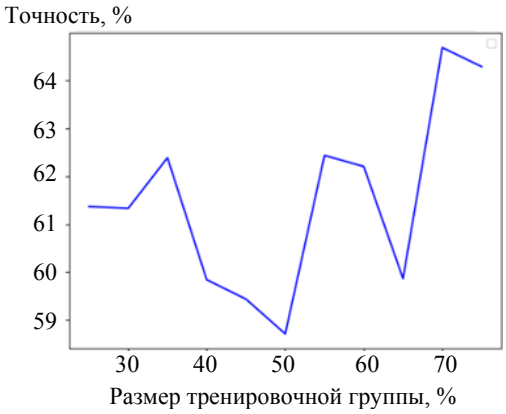

Рис. 5

Алгоритм дерева решений достигает наибольшей точности при размере тестовой выборки 70 %, если он применяется для двух классов. Максимальная точность применения алгоритма дерева классификации равна 65 %.

Для наиболее точной тестовой выборки было построено изображение дерева. По тому, какие характеристики находятся на его верхушке (рис. 6), можно определить ключевые характеристики для того или иного класса.

Из полученного результата можно вывести, что следующие характеристики наиболее важны для определения того, сможет ли собака стать поводырем:

– Not willing – отсутствие у собаки желания выполнять команды;

– Bad behavior around the home – плохое поведение в помещении;

– Kennels poorly – неспособность собаки прижиться к определенному месту.

Логистическая регрессия. Эта статистическая модель широко применяется во многих областях. Несмотря на название, это – линейная модель классификации, а не регрессии. Логистическая регрессия применяется для решения задач классификации в тех случаях, когда в качестве отклика имеется только два состояния (задачи с бинарным откликом) [5]. Данная модель рассчитывает вероятность принадлежности входного элемента к определенному классу. Построение статистической модели логистической регрессии представляет собой решение задачи построения модели линейной регрессии, т. е. поиска коэффициентов регрессии по тренировочному набору данных. С помощью коэффициентов регрессии можно определить, какая из независимых переменных имеет наибольшее влияние на конечный результат в выборе класса [5].

Также существует полиномиальная логистическая регрессия (multinomial logistic regression), которая обобщает логистическую регрессию для задач с количеством классов, большим, чем 2. Этот метод основан на предположении о независимости нерелевантных альтернатив [6]. Данное предположение гласит, что вероятность выбора одного класса из нескольких не должна зависеть

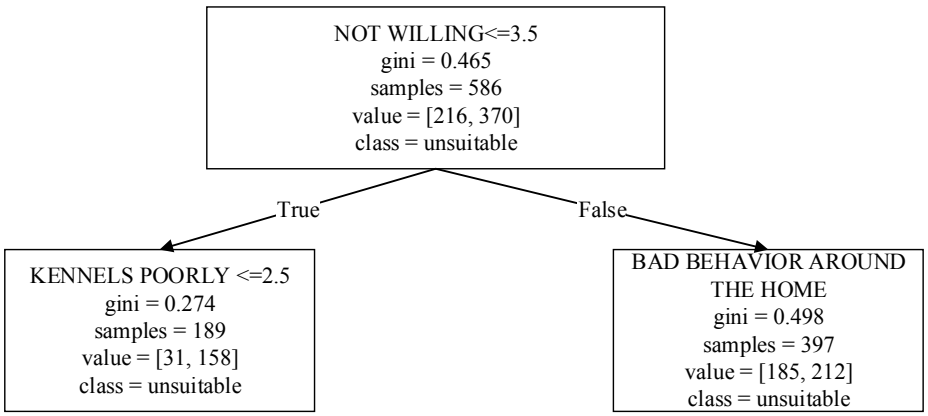

Рис. 6

от наличия или отсутствия других альтернатив. Одним из самых простых способов создания модели полиномиальной логистической регрессии, который возможен благодаря предположению о независимости альтернатив для множества откликов с мощностью k, является создание $k-1$ независимой модели логистической регрессии. Один из откликов берется в качестве опорного, и после этого для остальных вариантов отклика строится модель логистической регрессии относительно него.

На рис. 7 представлен график зависимости точности модели, построенной с помощью алгоритма полиномиальной логистической регрессии, от размера тренировочной выборки. Как видно из графика, модель, построенная по данному алгоритму, повышает точность при увеличении тренировочной выборки, что говорит о большей шумоустойчивости алгоритма. Наибольшая точность достигается на размере тренировочной выборки 70 % и равна 66 %.

В таблице сравниваются полученные результаты точности алгоритмов классификации. Из нее видно, что наибольшую точность в проведенных экспериментах показал алгоритм логистической регрессии, а наименьшую – метод дерева классификации.

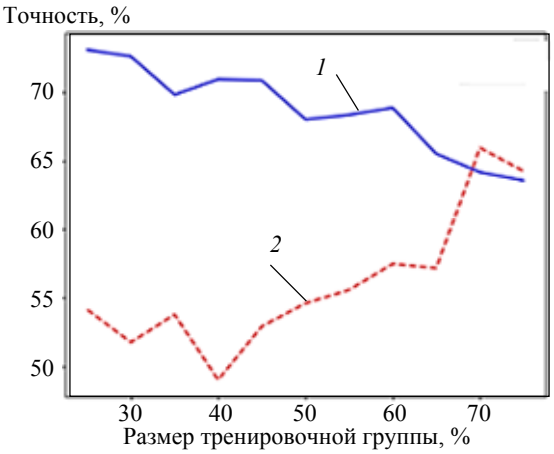

Рис. 7

Имя метода	Точность, %	Размер тренировочного множества, %
kNN	60	25
Дерево классификации	54	70
Лог. регрессия	66	70

В дальнейшем планируется исследование нелинейных методов анализа данных применительно к решению представленной задачи.

СПИСОК ЛИТЕРАТУРЫ

1. Машинное обучение. URL: https://en.wikipedia.org/wiki/Machine_learning (дата обращения 20.03.2019).

2. Классификация. URL: https://en.wikipedia.org/wiki/Statistical_classification (дата обращения 22.03.2019).

3. Метод *k* ближайших соседей. URL: https://scikit-learn.org/stable/modules/neighbors.html (дата обращения 23.03.2019).

4. Алгоритмы деревьев решений. URL: https://scikit-learn.org/stable/modules/tree.html (дата обращения 23.03.2019).

5. Способы реализации логистической регрессии. URL: https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression (дата обращения 23.03.2019).

6. Независимость нерелевантных альтернатив. URL: https://en.wikipedia.org/wiki/Independence_of_irrelevant_alternatives (дата обращения 28.03.2019).

D. V. Kozlov, Ya. A. Bekeneva
Saint Petersburg Electrotechnical University

DATA ANALISYS IN DETERMINATION OF APPLICABILITY OF DOGS FOR SPECIFIC SERVICE

The object of development and research is mathematical classification models that are applied to the initial data set with the characteristics of dogs performing a certain type of activity. The purpose of the work is to investigate the possibility of using classification methods to assess the suitability of dogs for service dogs. As a result of the work, several classification methods were investigated: k-nearest neighbors, classification tree, logistic regression. A program has been developed to prepare the initial data set for further mining using selected methods. Experiments have been carried out related to assessing the accuracy of each method, and the criteria that have the greatest influence on the results of testing dogs have been identified.

Classification, k-nearest neighbors, classification tree, logistic regression