|  |
| --- |
| https://sun9-52.userapi.com/impg/c854424/v854424968/1b56e7/0eZxz5zmBGY.jpg?size=2189x2160&quality=96&sign=3350b1482476eb1b48e740cdd1073352&type=album **МБОУ гимназия «Пущино»**  |

 Автор: Шухтина Александра Николаевна, 11 класс

ЗАДАЧИ И МЕТОДЫ КЛАССИФИКАЦИИ. СПОСОБЫ ОЦЕНКИ КАЧЕСТВА КЛАССИФИКАЦИИ

|  |
| --- |
| **Пущино, 2021** |

**СОДЕРЖАНИЕ**

[ВВЕДЕНИЕ 3](#_Toc72510516)

[1.1. Классификация: понятие и задачи 5](#_Toc72510517)

[1.2. Существующие базовые методы классификации данных 8](#_Toc72510518)

[1.2.1. Наивный Байесовский классификатор 8](#_Toc72510519)

[1.2.2. Семейство методов ближайших соседей 9](#_Toc72510520)

[1.2.3. Метод «Деревья решений» 10](#_Toc72510521)

[1.3. Изучение способов оценки качества классификации данных 12](#_Toc72510522)

[1.4 Выводы 16](#_Toc72510523)

[ЗАКЛЮЧЕНИЕ 17](#_Toc72510524)

[БИБЛИОГРАФИЧЕСКИЙ СПИСОК 17](#_Toc72510525)

[ДНЕВНИК ИССЛЕДОВАНИЯ 20](#_Toc72510526)

# ВВЕДЕНИЕ

В настоящее время растет значимость анализа данных, который применяется повсеместно. Анализ данных подразумевает подходы и алгоритмы, решающие различные задачи, одной из которых является задача классификации. Для решения задачи классификации применяются различные подход, наиболее распространенные из которых: деревья решений, семейство методов ближайших соседей, наивный Байесовский классификатор, нейросетевые классификаторы, и другие.

Аннотация:

**Актуальность работы** обусловлена:

* повышенным вниманием отечественных исследователей к компьютеризированному анализу данных;

**Объектом** **исследования** является анализ данных.

**Предмет** **исследования** – решение задач классификации данных.

**Задачи**:

1. определить понятие «классификация» и ее задачи;
2. описать существующие базовые методы классификации данных (наивный Байесовский классификатор, семейство методов ближайших соседей, метод «Деревья решений»);
3. изучить способы оценки качества классификации данных;

Для решения поставленных заданий в работе применена комплексная методика исследования. Основными **методами** анализа являются анализ, синтез, обобщение, метод классификации.

**Практическое значение** данной работы дает возможность использовать ее основные выводы и положения в курсах по программированию и теории автоматического управления.

**Структура работы.** Индивидуальный проект состоит из введения, основной части, заключения и списка использованной литературы.

В настоящей работе определяется понятие «классификация» и ее задачи, описываются существующие базовые методы классификации данных (наивный Байесовский классификатор, семейство методов ближайших соседей, метод «Деревья решений»). Также изучаются способы оценки качества классификации данных.

В *заключении* обобщаются основные выводы работы и намечаются перспективы для проведения дальнейших исследований в исследуемой области.

*Библиография* включает 16 источников использованной справочной литературы по теме исследования.

1.1. Классификация: понятие и задачи

Актуальность задачи классификации данных невозможно переоценить.

Так как с каждым годом объем информации, распространяемой в электронном виде увеличивается, обработка больших массивов данных, выдвигает задачу разработки методов интеллектуального анализа текстовых данных, в число самых приоритетных в области информатики.

Классификация является наиболее простой и одновременно наиболее часто решаемой задачей Data Mining [1].

Ввиду распространенности задач классификации необходимо четкое понимание сути этого понятия.

Примеры задач, решаемые в том или ином виде методами классификации:

1. задачи диагностики:
* постановки диагноза в медицине и ветеринарии;
* задачи технической диагностики (определение технического состояния объектов);
* задачи диагностики производственной деятельности предприятия (ситуационный анализ);
* педагогическая диагностика (совокупность приёмов контроля и оценки, направленных на решение задач оптимизации учебного процесса) и т.д.;
1. задачи атрибуции:
* кредитный скоринг;
* контроль качества и т.д. [2; 3]

Следовательно, классификация является одной из важнейших задач интеллектуального анализа данных, которая применяется в маркетинге при оценке кредитоспособности заемщиков, определении лояльности клиентов, распознавании образов, медицинской диагностике и многих других сферах. Если аналитику известны свойства объектов каждого класса, то в случае отношения нового наблюдения к определенному классу, данные свойства автоматически распространяются и на него [4].

В математическом виде, задачу классификации можно представить следующим образом:

Задано конечное множество классов и множество объектов, для конечного подмножества имеется соотношения объект-класс, которое называется обучающей выборкой. Для остальных объектов это соотношение неизвестно. Требуется составить алгоритм, способный классифицировать все объекты из заданного множества.

Классифицировать объект – означает указать номер или имя класса, к которому относится данный объект. Формальная запись постановки задачи: задано множество объектов , конечное множество классов , конечное множество обучающей выборки . Требуется составить алгоритм, способный классифицировать произвольный объект [5]

Если число классов ограничено двумя, то имеет место бинарная классификация, к которой могут быть сведены более сложные задачи. Например, вместо определения таких степеней кредитного риска как «высокий», «средний» или «низкий», можно использовать всего два – «выдать» или «отказать».

При решении задач классификации допускается несколько способов подачи данных. Наиболее распространенным является тот, при котором образец представляется вектором. Компоненты этого вектора представляют собой различные характеристики образца, которые влияют на принятие решения о том, к какому классу можно отнести данный образец. Например, для медицинских задач компонентами этого вектора могут быть данные из медицинской карты больного. Таким образом, на основании некоторой информации об образце, необходимо определить, к какому классу его можно отнести. Классификатор таким образом относит объект к одному из классов в соответствии с определенным разбиением N-мерного пространства, которое называется пространством входов, и размерность этого пространства определяется количеством компонентов вектора [6].

Для решения данной задачи, прежде всего, необходимо определить уровень сложности системы, что довольно непросто в ситуации, когда количество образцов значительно ограничено. Можно выделить три основных уровня сложности. Первый (самый простой) связан с той ситуацией, когда классы можно разделить прямыми – линейная разделимость. Во втором случае классы невозможно разделить линиями (плоскостями), но их возможно отделить с помощью более сложного деления – нелинейная разделимость. В третьем случае классы пересекаются и можно говорить лишь о вероятностной разрешение.

После предварительной обработки необходимо получить линейно раздельную задачу для упрощения построения классификатора. Однако при решении реальных задач наличествует ограниченное количество образцов, на основе которых проводится построение классификатора. При этом нельзя провести такую предобработку данных, при которой будет достигнуто линейное разрешение образцов.

В интеллектуальном анализе данных для решения задачи классификации используются следующие подходы [7; 8; 9, 5]:

1. нейронные сети;
2. деревья решений;
3. вероятностный подход (Наивный Байесовский метод);
4. машины опорных векторов (support vector machine, SVM);
5. семейство методов ближайших соседей.

Формально классификация проводится на основе разбиения пространства признаков на области, в каждой из которых многомерные векторы рассматриваются как идентичные. Иными словами, если объект попал в область пространства, ассоциированную с определенным классом, он к нему и относится, как показано на рис. 1.1, где а – это линейноразделимый случай SVM, б – деревья решений и в – нейронные сети.



Рисунок 1.1. Решение задачи классификации тремя методами

В работе более подробно будут описаны следующие наиболее распространенные подходы:

* наивный Байесовский классификатор,
* метод ближайших соседей,
* деревья решений.

1.2. Существующие базовые методы классификации данных

1.2.1. Наивный Байесовский классификатор

Наивные Байесовские классификаторы относятся к семейству простых вероятностных классификаторов и основаны на теореме Байеса с наивными предположениями о независимости между признаками.

Этот классификатор нашел свое широкое использование в начале 1960 гг. и остается до сих пор популярным в различных областях. С соответствующей предварительной обработкой данный классификатор является конкурентоспособным с более продвинутыми методами, включая SVM [10].

Существует целое семейство алгоритмов, основанных на общем принципе: все наивные Байесовские классификаторы предполагают, что значение конкретного признака не зависит от значения любого другого признака, учитывая переменную класса. Например, плод может считаться яблоком, если он красный, круглый и около 10 см в диаметре. Наивный классификатор Байеса рассматривает каждый из этих признаков независимо от вероятности того, что этот плод является яблоком, независимо от возможных корреляций между цветом, округлости и значениями диаметра.

Для некоторых типов вероятностных моделей, наивные Байесовские классификаторы могут быть очень эффективными в управляемом обучения с учителем. Во многих практических применениях оценка параметров для наивных Байесовских моделей использует метод максимальной вероятности. Иными словами, можно работать с наивной моделью Байеса, не используя байесовскую вероятность или любые байесовские методы.

Для oпределения вероятности принaдлежнoсти объекта Х к классу *Q*, который харaктеризуется признакaми *x(i)* (*i=1,..,M*), мoжет быть использoвaна следующая фoрмулa:

 (1)

Где:

 априoрная верoятность того, что объект отнoсится к клaссу Q. – условнaя вероятность того, что признaк *x(i)* появится в oбъектах клaсса (т.е. oценкa, насколькo признaк хaрaктерен для клaссa).

Преимуществом наивного Байеса является то, что для оценки параметров, необходимых для классификации, требуется лишь небольшое количество обучающих данных [11].

1.2.2. Семейство методов ближайших соседей

В методе ближайших соседей для классификации нового наблюдения  проводится упорядочивание исходных элементов выборки по какой-либо метрике [12]. Метрика – это понятия расстояния на произвольные пространства. Самой простоя метрикой является евклидова метрика (или квадрат евклидовой метрики), вычисляемая как сумма квадратов разностей признаков объектов [5]. При этом определяется не один ближайший сосед, а группа соседей, наиболее близких к новому наблюдению. Число соседей *k* является настраевым на стадии обучения параметром метода. Решение об отнесении  к классу *Qk* (*k*=1,. . ., *K*) принимается путём голосования его ближайших *k-*соседей с помощью простого подсчёта голосов. Если более половины его *k* – ближайших соседей принадлежат к классу *Qk*, то  также относится к этому классу [13].

1.2.3. Метод «Деревья решений»

Метод дерева решений применяется в задачах классификации и прогнозирования, когда решения приходится принимать в условиях риска, неопределённости и исход событий зависит от вероятностей. На каждое решение влияют какие-то определённые факторы, и у каждого решения есть свои последствия, которым присущ вероятностный характер. В этих условиях процесс принятия решений является последовательным и метод дерева решений предполагает определять, какие действия следует предпринять в каждой вершине дерева [14].

Дерево решений – математическая модель, которая задаёт процесс принятия решений так, что будут отображены каждое возможное решение, предшествующие и последующие этим решениям события или другие решения и последствия каждого конечного решения.

Дерево решений состоит из следующих элементов: дуг, узлов решений, узлов событий и конечных узлов (исходов).

Основополагающие идеи, послужившие толчком к появлению и развитию деревьев решений, были заложены в 1950-х годах в области исследований моделирования человеческого поведения с помощью компьютерных систем. Среди них следует выделить работы К. Ховеленда «Компьютерное моделирование мышления» [15] и Е. Ханта и др. «Эксперименты по индукции» [16].

Выбор способа выявления наиболее информативного признака, по которому проводится разбиение, является ключевым элементом алгоритма. МДР относится к так называемым «жадным» алгоритмам (Greedy Algorithms), для которых невозможно вернуться на предыдущий шаг и изменить разбиение, поменяв ранее выбранный.

В настоящее время разработано значительное число различных алгоритмов обучения МДР, отличающихся способом выбора классообразующего признака. Алгоритмы обычно измеряют однородность целевой переменной на подмножествах, используя некоторые метрики (критерии), а затем определяют качество разбиения.

Одним из важных моментов является выбор признака, по которому должно проходить очередное разбиение на подмножества. Идея состоит в том, чтобы минимизировать разброс ответов в одном или обоих подмножествах, полученных при разбиении. Часто для этого используют коэффициенты Джини и прироста информации, которые стараются минимизировать.

Ниже представлены формулы коэффициента Джини (Gini impurity):

** (2)

И прироста информации:

 (3)

Где *pi* - верoятность (oтнoсительнaя частота) класса i в рaссматривaемoй пoдвыбoрке.

Таким образом, критерий Джини характеризует вероятность неправильной классификации нового объекта, если этот объект был проклассифицирован классификатором, выдающим решения случайным образом, причем вероятность выдать класс i = pi.

Критерий прироста информации (его также называют энтропийным критерием), показывает меру отличия классов от вырожденного, т.е. такого, в котором все объекты относятся к одному классу (энтропия = 0). Соответственно, бОльшее значение энтропии соответствует равномерному распределению.

1.3. Изучение способов оценки качества классификации данных

Для обучения модели и оценки ее качества в задаче классификации требуется размеченная выборка, в которой каждому объекту сопоставляется истинный класс, назначенный экспертом (экспертный подход). При этом обычно часть выборки используется для обучения модели (обучающее множество), а другая часть – для оценки ее качества (тестовое множество).

На практике используются три основных подхода для получения оценки ошибки обощения.

1) Оценка по тестовой выборке. Исследуемую выборку делят на две части. По первой части (обучающая выборка) производится оценка параметров и построение решающего правила (обучающего метода), по второй части (тестовая выборка) – производится определение оценки вероятности ошибок. К недостаткам этого способа следует отнести уменьшение объема выборки, используемой при классификации (соответсвенно уменьшается и возможное число оцениваемых параметров в алгоритме). К достоинствам этого способа относятся его несмещенность и универсальность.

2) Оценка с помощью скользящего контроля (кросс-валидация - cross-validation). Для оценки параметров вся выборка делится на k частей (k-fold cross-validation), далее все части кроме одной используются для обучения, а оставшаяся часть используют для проверки. Затем эта часть включается в общую выборку, а для контроля извлекается другая часть. Данная процедура повторяется для всех частей исходной выборки. Недостатком скользящего контроля является существенный рост количества необходимых вычислений, а достоинством – несмещённость полученных оценок.

3) Метод статистического моделирования (Bootstrap). Предложен в рамках разработки методов коллективной классификации (boosting and bagging) для формирования обучающих и тестовых выборок в условиях ограниченного количества наблюдений с целью проведения многократного обучения и тестирования. При этом обчающая выборка принимается за генеральную совокупность и из нее случайным образом произвадиться составление обучающих и тестовых подвыборок.

Рассмотрим способы определения качества классификации по тестовой выборке в простейшем случае бинарной классификации [5]. Составим матрицу ошибок (Confusion Matrix, матрица неточностей), представленную в таблице 2.

Таблица 1. Матрица ошибок

|  |  |  |
| --- | --- | --- |
|  | Фактический класс Q=1 | Фактический класс Q=0 |
| Предсказанный класс =1  | TP Истинно положительные (True Positive)  | FP Ложно положительные (False Positive) |
| Предсказанный класс =0 | FN Ложно отрицательные (False Negative) | TN Истинно отрицательные (True Negative)  |

TP – число правильно определенных меток класса Q=1;

TN – число правильно определенных меток класса Q=0;

FN – число неправильно определенных примеров, когда класс объектов Q=1 проклассифицирован как =0 (ошибка первого рода);

FP – число неправильно определенных меток, когда класс объектов Q=0 проклассифицирован как =1 (ошибка второго рода).

Таким образом, матрица ошибок показывает, сколько объектов интересующего нас («положительного») класса Q=1 были отнесены к («отрицательному») классу Q=0 (и наоборот).

В теории вероятностей и математической статистике FP (False Positive) и FN (False Negative) называются соответственно ошибками I-го рода (ложная тревога, ложно положительное срабатывание, т.е. отвергается правильная гипотеза Н0) и II-го рода (пропуск цели, ложно отрицательное срабатывание, т.е. принимается неправильная гипотеза Н0).

На основе матрицы ошибок рассчитываются следующие показатели качества:

1) Аккуратность δ (Accuracy, также можно встретить термин верность) и ошибка классификации Δ:

|  |  |
| --- | --- |
| δ =   | (4) |
| Δ=1-δ=  | (5) |

2) Полнота R (Recall) и точность P (Precision):

|  |  |
| --- | --- |
| R=  - полнота  | (6) |
| P=  - точность | (7) |

3) Ф-мера (f-score) – унифицированный показатель общей точности классификации, который рассчитывается на основании двух предыдущих параметров – точности и полноты:

|  |  |
| --- | --- |
| f-score =  | (8) |

# 1.4 Выводы

В данной работе были изучены основные понятия классификации данных и задачи, которые решаются её помощью. Приведена постановка задачи классификации.

Также были рассмотрены базовые методы классификации данных, а именно Наивный Байесовский классификатор, метод ближайших соседей и метод деревьев решений.

Были изучены способы оценки качества полученных результатов, которые являются неотъемлемой частью решения задачи классификации данных.

# ЗАКЛЮЧЕНИЕ

В рамках данного исследования была достигнута поставленная цель и выполнены все исследовательские задачи, а именно определено понятие «классификация» и ее задачи, описаны существующие базовые методы классификации данных (наивный Байесовский классификатор, семейство методов ближайших соседей, метод «Деревья решений»), изучены способы оценки качества классификации данных.

Рассмотрение теоретических аспектов классификации данных показывает, что классификация является одной из важнейших задач интеллектуального анализа данных, которая применяется в маркетинге при оценке кредитоспособности заемщиков, определении лояльности клиентов, распознавании образов, медицинской диагностике и многих других сферах.

# БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Картиев С.Б., Курейчик В.М. Алгоритм классификации, основанный на принципах случайного леса, для решения задачи прогнозирования [Электронный ресурс] // Программные продукты и системы. – 2016. Режим доступа: <https://cyberleninka.ru/article/n/algoritm-klassifikatsii-osnovannyy-na-printsipah-sluchaynogo-lesa-dlya-resheniya-zadachi-prognozirovaniya>. Загл. с экрана.
2. Замятин А.В. Интеллектуальный анализ данных: учеб. пособие. – Томск: Издательский Дом Томского государственного университета, 2016. – 120 с.
3. Семенова А., Ганеева Л. Введение в машинное обучение [Электронный ресурс] / А. Семенова, Л. Ганеева // Казанский (приволжский) федеральный университет. Режим доступа: <http://elibrary.sgu.ru/uch_lit/1141.pdf>. Загл. с экрана.
4. Курейчик В.М. Особенности построения систем поддержки принятия решений // Изв. ЮФУ: Технич. науки. – 2012. – № 7. – С. 92-98.
5. Маннинг К.Д., Рагхаван П., Шютце Х. «Введение в информационный поиск». – М.: «Вильямс», 2014
6. Мохов А.С., Толчеев В.О., Бородкин А.А. Анализ и обработка текстовых данных: практикум / А.С. Мохов, В.О. Толчеев, А.А. Бородкин. – М.: Издательство МЭИ, 2019. – 56 с.
7. Луньков А.Д., Харламов А.В. Луньков А.Д., Харламов А.В. Интеллектуальный анализ данных: учебно-методическое пособие [Электронный ресурс]. Ч. 1. – Саратов: СГУ. – 96 с. Режим доступа: <https://www.researchgate.net/publication/321686595_Vvedenie_v_masinnoe_obucenie>. Загл. с экрана.
8. Rish I. An empirical study of the naive Bayes classifier. In IJCAI Workshop on Empirical Methods in AI, 2001.
9. Rish I. An empirical study of the naive Bayes classifier // IJCAI 2001 workshop on empirical methods in artificial intelligence. – New York, 2001. – Vol. 3. – No. 22. – pp. 41-46.
10. Russell S., Norvig P. Artificial Intelligence: A Modern Approach (2nd ed.). Prentice Hall, 2003.
11. Семерякова Е.Г., Берестнева О.Г., Макарова Л.С. Математические методы в задачах медицинской диагностики [Электронный ресурс] / Е.Г. Семерякова, О.Г. Берестнева, Л.С. Макарова // Современные проблемы науки и образования. – 2012. – № 6. Режим доступа: <https://www.science-education.ru/ru/article/view?id=7472>. Загл. с экрана.
12. Замятин А.В. Интеллектуальный анализ данных: учеб. пособие. – Томск: Издательский Дом Томского государственного университета, 2016. – 120 с.
13. Нечаев В.В., Трофименко В.М. Анализ методов семантического поиска информационных ресурсов [Электронный ресурс] / В.В. Нечаев, В.М. Трофименко //Образовательные ресурсы и технологии. – 2014. Режим доступа: <https://cyberleninka.ru/article/n/analiz-metodov-semanticheskogo-poiska-informatsionnyh-resursov>. Загл. с экрана.
14. К. Ховеленд // Компьютерные исследования и моделирование, 2020, т. 12, № 3, с. 471-473
15. Е. Хант // «Эксперименты по индукции», 1962. – 36 с.
16. .Гусева А.И., Киреев В.С., Кузнецов И.А., Бочкарёв П.В. Исследование

 алгоритмов многомерной классификации научных данных [Электронный

 ресурс] / А.И. Гусева, В.С. Киреев, И.А. Кузнецов, П.В. Бочкарев //

 Фундаментальные исследования. – 2015 – № 11 (часть 5) – С. 868-874.

 Режим доступа: https://www.fundamental-research.ru/ru/article/view?id=39524.

 Загл. с экрана.

# ДНЕВНИК ИССЛЕДОВАНИЯ

|  |  |  |
| --- | --- | --- |
| **№** |  **Содержание** | **Дата** |
| 1. | Выбор темы | Ноябрь 2020 г. |
| 2. |  Работа с теоретическим материалом (литературой) | Ноябрь 2020г.- Январь 2021 г. 2021 г. |
| 3. | Сбор и анализ информации  | Февраль 2021 г. |
| 4. | Систематизация материала | Март - Апрель 2021 г. |
| 5. | Представление научного проекта | Май 2021 г.   |