

ПОСТРОЕНИЕ ДЕРЕВЬЕВ РЕШЕНИЙ ДЛЯ ЗАДАЧ КЛАССИФИКАЦИИ В R

Интеллектуальный анализ данных (англ. Data Mining) – это процесс обработки и анализа больших объемов данных с целью выявления скрытых в них закономерностей и шаблонов. Результатом интеллектуального анализа данных является получение ценной информации, которая может быть использована для улучшения бизнес-планирования и принятия обоснованных управленческих решений.

Интеллектуальный анализ данных играет ключевую роль в современном маркетинге, поскольку помогает лучше понять поведение потребителей и оптимизировать маркетинговые кампании. Он представляет собой мощный инструмент, который может быть применен в данной деятельности для идентификации трендов и возможностей [5]. Преимуществом интеллектуального анализа данных в сфере маркетинга является возможность выявления скрытых тенденций и закономерностей, которые могут отражать потенциальные изменения в спросе потребителей или их предпочтениях. К примеру, на основе анализа информации о поведении посетителей сайта или аккаунтов в социальных сетях можно определить наиболее востребованные товары или услуги, а также установить факторы, влияющие на решение о покупке. Также это возможность оптимизировать маркетинговые кампании и рекламные стратегии. Анализируя данные о проведенных ранее кампаниях, их результатах и поведении потребителей, можно найти оптимальные способы взаимодействия с целевой аудиторией. Благодаря этому можно улучшить таргетинг рекламных сообщений, сделать их более релевантными и полезными для пользователей, а также повысить конверсию и отдачу от маркетинговых усилий.

Одной из задач интеллектуального анализа данных является задача классификации, где основной целью является отнесение объектов к определенным классам или категориям на основе их характеристик, представленных в наборе данных. Для этого используются различные методы машинного обучения: байесовская классификация, метод k ближайших соседей, метод опорных векторов, логистическая регрессия, деревья принятия решений и прочие.

Постановка задачи классификации приведена на рис. 1. Пусть выполняется анализ некоторого набора данных ($n_1, n_2, \dots, n_N$), объекты соответствуют строкам таблицы.

	X_1	X_2	...	X_k	Y	
n_1	x_{11}	x_{12}	...	x_{1k}	y_1	эталонная выборка
n_2	x_{21}	x_{22}	...	x_{2k}	y_2	
...	...	...	...	...	...	
n_L	x_{L1}	x_{L2}	...	x_{Lk}	y_L	
$n_{(L+1)}$	$x_{(L+1)1}$	$x_{(L+1)2}$	...	$x_{(L+1)k}$	?	тестируемая выборка
...	...	...	...	...	...	
n_N	x_{N1}	x_{N2}	...	x_{Nk}	?	

Рис. 1. Постановка задачи классификации

Допускается, что для L объектов, которых называют эталонной или обучающей выборкой, значения зависимой $y_1, y_2 \dots y_L$ переменной, расположенной в последнем столбце, известны и дискретны, а для $N-L$ объектов значения $y_{L+1}, y_{L+2} \dots y_N$ неизвестны и должны быть определены. Независимые переменные $x_1, x_2 \dots x_K$ описывают признаки объектов и могут иметь любое распределение. Группа объектов $n_{L+1}, n_{L+2} \dots n_N$ называется тестируемой выборкой. Задача классификации состоит в определении неизвестных значений зависимой переменной $y_{L+1}, y_{L+2}, \dots y_N$ и отнесении $N-L$ объектов в m , заранее определенных классов.

Деревья принятия решений представляют собой иерархическую структуру, состоящую из узлов, которые принимают решения на основе определенных условий, и ветвей, которые и соединяют эти узлы. В каждом узле дерева происходит проверка условия, при выполнении которого объект относится к соответствующему классу. Таким образом, дерево принятия решений можно представить как последовательность вопросов, на которые необходимо ответить, чтобы отнести объект к определенному классу [1].

Преимущества деревьев принятия решений заключаются в их простоте и наглядности, а также в том, что они способны обрабатывать большие объемы данных. Кроме того, деревья принятия решений обладают высокой точностью классификации и не требуют большого количества вычислительных ресурсов.

Для того, чтобы использовать эти методы, тем, кто не является специалистами в области прикладной математики и программирования, предлагаются различные библиотеки и функции, реализующие эти алгоритмы на языке программирования R. С одной стороны – это язык программирования, а с другой – рабочая среда, в которую встроены средства статистического и интеллектуального анализа данных. Для R есть удобная среда разработки Rstudio.

Рассмотрим построение дерева классификации в R на примере набора данных Spam7, содержащего информацию о 4 601 электронном письме, из которых 1 813 отнесены к спаму. Набор данных содержит 5 столбцов, в которых по каждому письму дана следующая информация: *crl.tot* – общая длина слов, написанных большими буквами; *dollar* – относительное число вхождений в тело письма символа “\$”; *bang* – относительное число вхождений в тело письма символа “!”; *money* – относительное число вхождений в тело письма слова “money”; *n000* – относительное число вхождений в тело письма строки “000”; *make* – относительное число вхождений в тело письма слова “make”; *yesno* – символ “n” для письма, не идентифицированного как спам, символ “y” – для письма являющегося спамом. Зависимой переменной или меткой класса в этом наборе данных является *yesno*. На рис. 2 приведен пример кода в R, который устанавливает пакеты, подключает библиотеки, считывает из файла набор данных Spam7, отбирает 2/3 данных для обучающей выборки и 1/3 для тестируемой и функцией *rpart* строит дерево классификации.

```
install.packages("rpart")
install.packages("maptree")
library(maptree)
library(rpart)

df = read.csv("spam7.csv")

train_indices = sample(1:nrow(df), floor(2/3 * nrow(df)))
train_data = data[train_indices, ]
test_data = data[-train_indices, ]

tree_model = rpart(yesno ~ crl.tot + dollar + bang + money + n000 + make,
                    data = train_data, method = "class")
draw.tree(tree_model)
```

Рис. 2. Пример кода

На рис. 3 приведено построенное дерево для классификации электронных писем.

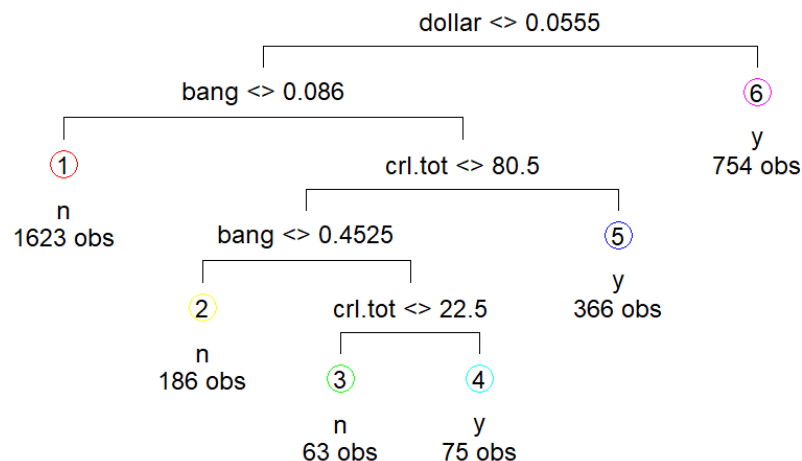


Рис. 3. Дерево для классификации электронных писем

С помощью функции *predict()* и построенного дерева была классифицирована тестируемая выборка. Точность составила практически 86 %.

Использование деревьев принятия решений позволяет маркетологам более точно определить, к какому сегменту относится клиент, основываясь на его предпочтениях, поведении и других характеристиках. Для предложения наиболее подходящего товара или услуги используются деревья принятия решений, которые помогают определить, какие продукты или услуги предпочитают потребители на основе их прошлых покупок и демографических данных.

Кроме того, деревья принятия решений будут очень эффективны в прогнозировании продаж, так как с их помощью маркетологи могут предсказать будущие продажи, используя данные о прошлых продажах, маркетинговой активности и других факторах, и в прогнозировании спроса и цен, по той причине, что методы классификации используются для предсказания спроса на товары и услуги, основываясь на исторических данных о продажах, сезонные изменения и другие факторы, а деревья принятия решений позволяют определить, как изменение цен может повлиять на продажи и прибыль компании. Это помогает планировать производство, распределение и управление запасами, а также помогает менеджерам принимать решения о ценообразовании, чтобы максимизировать прибыль и конкурентоспособность.

В маркетинге бинарная и мультиклассовая классификации позволяют разделить клиентов на группы на основе их характеристик и предпочтений. Бинарная классификация используется для разделения клиентов на две группы, например, можно разделить клиентов на тех, кто покупает продукт, и тех, кто не покупает. Для построения бинарного дерева принятия решений нужно задать вопрос, который будет разделять клиентов на два класса, и определить, какие характеристики наиболее важны для классификации. Мультиклассовая классификация позволяет разделить клиентов на несколько групп, соответствующих разным предпочтениям или сегментам рынка. Например, если нужно классифицировать клиентов по их предпочтениям в продуктах, то будет несколько классов, соответствующих различным предпочтениям. Для построения мультиклассового дерева принятия решений необходимо задать несколько вопросов, которые будут разделять клиентов на разные классы, и определить наиболее важные характеристики для каждого класса.

Подводя итоги, можно сказать, что применение интеллектуального анализа данных в маркетинге способно существенно увеличить эффективность деятельности компаний и по-

мочь им достичь конкурентных преимуществ. В результате, использование интеллектуального анализа данных становится все более важным инструментом для выявления тенденций и возможностей в области маркетинга. Этот подход позволяет компаниям принимать обоснованные решения, оптимизировать маркетинговые стратегии и улучшать взаимодействие с клиентами, что в конечном итоге положительно сказывается на эффективности бизнеса и его конкурентоспособности.

Список использованных источников

1. Microsoft learn [Электронный ресурс]. – Режим доступа: <https://learn.microsoft.com/ru-ru/>. – Дата доступа: 01.04.2024.
2. Мультиклассовые и множественные алгоритмы вывода [Электронный ресурс]. – Режим доступа: <https://scikit-learn.ru/1-12-multiclass-and-multioutput-algorithms/>. – Дата доступа: 01.04.2024.
3. Binary classification [Электронный ресурс]. – Режим доступа: <https://medium.com/learning-machine-learning-by-gc-nit/binary-classification-35590076ca69/>. – Дата доступа: 01.04.2024.
4. Package ‘rpart’ [Электронный ресурс]. – Режим доступа: <https://cran.r-project.org/web/packages/rpart/rpart.pdf>. – Дата доступа: 01.04.2024.
5. Применение интеллектуального анализа данных в торговых предприятиях для решения маркетинговых задач [Электронный ресурс]. – Режим доступа: <https://sciup.org/primenenie-intellektualnogo-analiza-dannyh-v-torgovyh-predpriyatijah-dlja-140266463>. – Дата доступа: 01.04.2024.