ЗАДАЧИ КЛАССИФИКАЦИИ В ИНТЕЛЛЕКТУАЛЬНОМ АНАЛИЗЕ ДАННЫХ

Князева Л.А., Маракулина А.Р.

Вятский государственный университет, РФ, 610000, г. Киров, ул. Московская, 36,
liliknyaz@mail.ru

Интеллектуальный анализ данных (data mining) необходим для поиска в имеющихся данных скрытых нетривиальных и полезных закономерностей, которые позволяют получить новые знания об исследуемых данных. Интеллектуальный анализ данных опирается на прошлый опыт и алгоритмы, определенные с помощью существующего программного обеспечения и пакетов, причем с различными методами ассоциируются разные специализированные инструменты. Несколько основных методов, которые используются для интеллектуального анализа данных, описывают тип анализа и операцию по восстановлению данных. К таким методам относят: классификацию, ассоциацию, кластеризацию и др. В данной работе будет использоваться классификация как метод, помогающий принять решение покупателю в выборе товара. Для классификации в Data Mining используются модели, при построении которых применяется обучение с учителем, когда выходная переменная задана для каждого наблюдения. Формально классификация производится на основе разбиения пространства признаков на области, в пределах 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.

Объектом исследования данной работы является рынок подержанных автомобилей, представленный на европейском ресурсе eBay-Kleinanzeigen. Для анализа были взяты ряды распределения пробега (kilometer), цены (price) и мощности автомобилей (powerPS). Цель исследования – классифицировать данные по автомобилям с помощью байесовского классификатора, используя переменные цены, пробега и мощности автомобилей, чтобы сформировать базу для принятия решения покупателем. Для получения результата используется наивный байесовский классификатор, который аппроксимируется с помощью метода линейного дискриминантного анализа (LDA). Данные для исследования были взяты с платформы Kaggle. Размер выборки составляет 371824 строк, это позволяет сделать вывод о том, что расчеты адекватны ввиду достоверности данных. Для реализации использовались русерсы R – Studio [1].

Литература.

1. Джеймс Г. Уиттон Д., Хасти Т., Тибширани Р. Введение в статистическое обучение с примерами на языке R. Пер. с англ. С.Э. Мастицкого - М.: ДМК Пресс, 2016. - 450 с.