АНАЛИЗ СОВРЕМЕННЫХ КЛАССИФИКАТОРОВ ДЛЯ МНОГОМЕРНЫХ ДАННЫХ

А. А. Цыкунова

Белорусский государственный университет, Минск;
tsykunova.ane4ka@gmail.com;
науч. рук. – А. М. Недзьведь, д-р техн. наук

В последнее время исследователи, занимающиеся машинным обучением и интеллектуальным анализом данных, уделяют повышенное внимание классификации многомерных данных. В многомерном наборе данных (МНД) каждый экземпляр связан с несколькими значениями классов. Из-за своей сложной природы выбор характеристик и классификатора, построенного на основе МНД, обычно более трудоемкий процесс. Поэтому нам нужна надежная методика выбора признаков для выбора оптимального единственного подмножества признаков МНД для дальнейшего анализа и разработки классификатора. В данной статье проведен анализ методов классификации и определены атрибуты данных для качественной классификации.

Ключевые слова: классификация; извлечение признаков; размерность; правило независимости.

ПРЕДСТАВЛЕНИЕ ДАННЫХ

Необработанные данные, полученные из разных источников в разных форматах, очень чувствительны к шуму, нерелевантным атрибутам, пропущенным значениям и противоречивым данным [1]. Поэтому предварительная обработка данных является важной фазой, которая помогает подготовить высококачественные данные для эффективного анализа в больших наборах данных. Предварительная обработка улучшает результаты и упрощает процесс интеллектуального анализа данных. Отсутствующие значения существуют во многих ситуациях, когда для некоторых переменных нет доступных значений. Поэтому важно обрабатывать пропущенные значения для повышения точности классификатора в задачах интеллектуального анализа. Для обработки используется набор данных, соответствующий такому условию [2].

В работе используется набор данных, состоящий из трех типов объектов: (а) спецификация автомобиля с точки зрения различных характеристик, (б) присвоенный ему рейтинг страхового риска, (в) нормированные потери в использовании по сравнению с другими автомобилями.

Второй показатель соответствует степени риска автомобиля относительно цены. Первоначально автомобилям присваивается символ фактора риска, связанный с его ценой. Затем, если риск увеличивается или уменьшается, этот символ корректируется путем перемещения его вверх

(или вниз) по шкале. Диапазон значений для фактора выбран от -3 до +3. Значение + 3 указывает, что автомобиль находится в зоне риска, -3 что этот автомобиль наиболее безопасен.

Третьим фактором является относительная средняя выплата убытков за застрахованное транспортное средство в год. Это значение нормируется для всех автомобилей в рамках определенной классификации размера (двухдверные малые, универсалы, спортивные/специальные и т.д.), и представляет собой средние потери за машину в год [3].

СРАВНЕНИЕ КЛАССИФИКАТОРОВ ДЛЯ МНОГОМЕРНЫХ ДАННЫХ

Рассматриваемая в данной работе модель, построенная на основе выбранного набора данных, является избыточной. Если обратить внимание на значения каждого коэффициента, эта модель предполагает, что для моделирования могут понадобиться только 13 из 23 признаков. Для уменьшения количества объектов в модели используется информационный критерий Акаике (AIC). AIC – критерий для выбора лучшей из нескольких статистических моделей, построенных на одном и том же наборе данных и использующих логарифмическую функцию правдоподобия. Цель состоит в том, чтобы минимизировать AIC.

$$-2\ln\left(p\left(x\middle|\hat{\theta}\right)\right). \quad (1)$$

Величина (1) иногда называется отклонением модели. Отклонение является мерой относительной вероятности модели и обобщением дисперсии. Фактически, отклонение следует измерять относительно полноценной модели (количество параметров = количество наблюдений), но этот шаг часто пропускается.

После преобразования модель имеет 17 признаков или параметров. Значение p статистики F значительно меньше 0.05, что говорит о том, что предлагаемая модель достаточно хорошо интерпретирует данные. Кроме того, R^2_{adj} этой модели немного лучше, чем у линейной модели ранее.

После классификации методом *k*-ближайших соседей видно, что в модели классификатора каждый раз есть улучшения. Улучшение может быть сделано с точки зрения использования различных значений параметра *k* и выбора одного с максимальной точностью.

Для тестирования классификации выбрано $k = 4$, основываясь на экспериментальных рассуждениях.

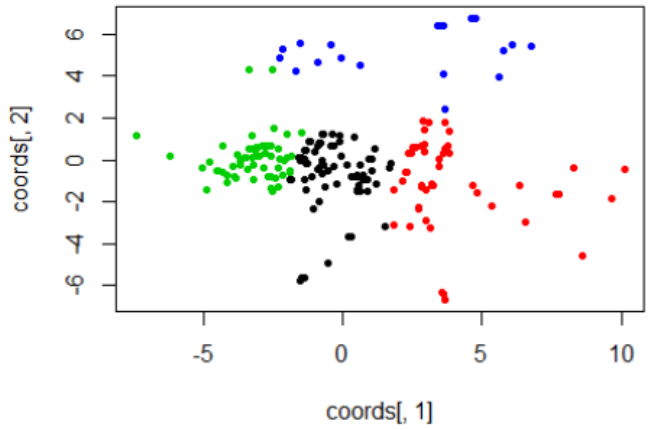

Рис. 2. Результат выполнения метода к-ближайших соседей

Согласно изображению, у нас есть 4 класса, которые основаны на главном параметре «цена», но результат зависит и от других параметров.

Ценовой параметр со средним значением был использован для классификации с деревьями решений. До начала работы с деревом решений нужно перевести значение параметра в двоичную переменную.

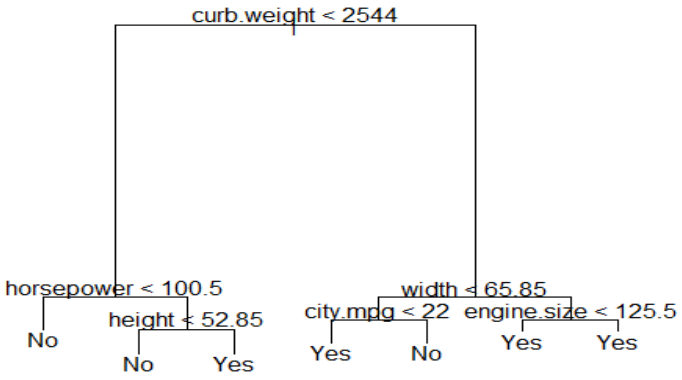

Рис. 3. Дерево решений множества данных об автомобилях

В результате получается выбор между двумя группами, которые зависят от цены: больше, чем средняя цена, и меньше, чем средняя цена. Шаг за шагом следует уменьшать размер дерева. На последнем шаге остаются только значимые параметры. Можно считать, что желаемое дерево получено и можно интерпретировать результаты.

Для наивного байесовского классификатора была поставлена цель классифицировать автомобили по цене в зависимости от их характеристик.

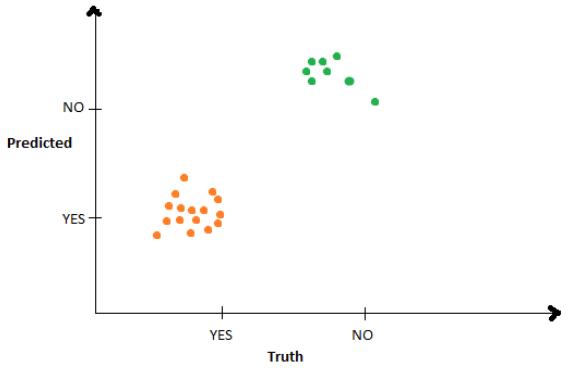

Рис. 4. Предсказанные и наблюдаемые значения из набора данных с использованием наивного байесовского классификатора

ВЫВОДЫ

Преимуществом деревьев решений является их удобочитаемость по сравнению с другими моделями. Однако самая большая проблема, которая есть с этим классификатором – это попытка обучить деревья решений, используя наборы данных с большим количеством категориальных данных. Для многомерных данных нужно использовать модификацию: на каждом шаге проводить разбиение на группы признаков и далее задача такая, что нужно выявить какие предикторы и в какой степени влияют на совокупную изменчивость количественных соотношений между отдельными компонентами.

При подборе алгоритма *k*-ближайших соседей нужно указать количество соседей (*k*), которые следует учитывать в алгоритме. Выбор значения *k* значительно влияет на результат работы данного метода.

Наивный байесовский алгоритм предполагает, что данные имеют атрибуты, независимые друг от друга. Если это предположение о независимости имеет место, Наивный Байесовский метод работает лучше, чем другие модели. Если все атрибуты данных являются категориальными, наивный байесовский метод работает очень хорошо. С помощью Наивного Байесовского метода была получена лучшую модель по цене.

Библиографический ссылки

1. *Read J.* A pruned problem transformation method for multi-label classification // Proceedings of the 6th New Zealand Computer Science Research Student Conference, April 2008. P. 143–150.
2. *Zhang D., Chen S., Zhou Z.* Constraint score: a new filter method for feature selection with pairwise constraints // Pattern Recognition. 2008. № 41 (5). P.1440–1451.
3. *Tsykunova A., Kopanja L., Kulinkovich V.* Analysis of modern classifiers for multidimensional data [Electronic resource] // Proceedings of the 14th International Conference (PRIP'2019), Minsk, May 2019. URL: https://prip.bsuir.by (date of access: 30.05.2019).