АНАЛИЗ МЕТОДОВ КЛАССИФИКАЦИИ И КЛАСТЕРИЗАЦИИ ДАННЫХ

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Нахратьянц Д. А., Сапунов А. К., Глебов Д. А.

Ролич О. Ч. – канд.техн.наук, доцент

В современной разработке программных продуктов все более важное значение имеет решение задач классификации и кластеризации данных. В ходе данного исследования были изучены основные методы решения данных задач.

Методы классификации и кластеризации данных в машинном обучении можно разделить на 3 основные категории: контролируемое, неконтролируемое и подкрепляемое обучение.

Наивная байесовская классификация представляет собой семейство простых вероятностных классификаторов, которые основаны на применении Теоремы Байеса со строгими (наивными) предположениями о независимости функций.

Линейная регрессия — используемая в статистике регрессионная модель зависимости одной (объясняемой, зависимой) переменной y от другой или нескольких других переменных (факторов, регрессоров, независимых переменных) x с линейной функцией зависимости.

Логистическая регрессия представляет собой мощный статистический способ прогнозирования вероятности возникновения некоторого события с одной или несколькими независимыми переменными. Логистическая регрессия определяет степень зависимости между категориальной зависимой и одной или несколькими независимыми переменными путем использования логистической функции, являющейся аккумулятивным логистическим распределением.

Метод опорных векторов — это набор алгоритмов, использующихся для задач классификации и регрессионного анализа. Учитывая, что в N-мерном пространстве каждый объект принадлежит одному из двух классов, метод генерирует (N-1)-мерную гиперплоскость с целью разделения этих точек на 2 группы. Среди наиболее масштабных проблем, которые были решены с помощью метода опорных объектов (и его модифицированных реализаций) выделяют отображение рекламных баннеров на сайтах, распознавание пола на основании фотографии.

Задача кластеризации состоит в группировании множества объектов таким образом, чтобы поместить максимально похожие между собой элементы в одну группу (кластер).

Алгоритмы кластеризации используются в биологии, социологии и информационных технологиях. Например, в биоинформатике с помощью кластеризации анализируются сложные сети взаимодействующих генов, состоящие порой из сотен или даже тысяч элементов.

Метод главных компонент — это статистическая процедура, которая использует ортогональное преобразование с целью конвертации набора наблюдений за возможно коррелированными переменными в набор значений линейно некоррелированных переменных, называемых главными компонентами.

Сингуля́рное разложе́ние — определённого типа разложение прямоугольной матрицы. Имеющее широкое применение, в силу своей наглядной геометрической интерпретации, при решении многих прикладных задач. Переформулировка сингулярного разложения, так называемое разложение Шмидта имеет приложения в квантовой теории информации, например в запутанности.

Метод главных компонент является простым применением сингулярного разложения. Первые алгоритмы компьютерного виденья использовали PCA и SVD, чтобы представить лица в виде суммы базисных компонент, выполнить уменьшение размерности, а затем сопоставить их с изображениями из обучающей выборки. И хотя современные методы характеризуются более сложной реализацией, многие из них по-прежнему работают на базе подобных алгоритмов.

Анализ независимых компонент представляет собой статистический метод выявления скрытых факторов, которые лежат в основе множества случайных величин, сигналов и прочих измерений. ICA определяет порождающую модель для исследуемых многофакторных данных, которые обычно подаются в виде большой базы данных образцов. В модели переменные подаются как линейная смесь некоторых скрытых переменных, а любая информация о законах смешивания отсутствует.

Список использованных источников:

1. 10 главных алгоритмов машинного обучения [Электронный ресурс] – http://ru.datasides.com/code/algorithms-machine-learning/ – Дата доступа: 09.04.2019.

2. Обзор методов машинного обучения [Электронный ресурс] – http://elib.ict.nsc.ru/jspui/bitstream/ICT/1455/1/2005_diss_ageev.pdf – Дата доступа: 10.04.2019.

3. Обзор самых популярных методов машинного обучения [Электронный ресурс] – https://tproger.ru/translations/top-machine-learning-algorithms// – Дата доступа: 10.04.2019.