Секция «Теория вероятностей и математическая статистика»

О возможности кластеризации после применения методов снижения размерности.

Свинин Глеб Евгеньевич
Студент (специалист)
Московский государственный университет имени М.В.Ломоносова,
Механико-математический факультет, Кафедра теории вероятностей, Москва, Россия
E-mail: gleb.svinin@math.msu.ru

Одними из способов поиска структуры в данных являются так называемые «методы обучения без учителя». Их преимущество состоит в том, что использование этих методов не предполагает предварительных знаний о данных [1]. Работа этих методов зависит только от задания гиперпараметров - настраиваемых параметров, определяющих работу алгоритма. Таким образом, получаемые результаты определяются гиперпараметрами и исходной структурой данных. Целью работы является изучение структуры данных эпидемиологического исследования с использованием «методов обучения без учителя» для выделения областей высокой плотности (кластеров) в исходном высокоразмерном пространстве биомаркеров. Вначале снижается размерность исходных данных с помощью алгоритмов t-SNE [2] и PCA [3]. Затем полученное двумерное представление визуализируется на плоскости и производится его кластеризация при помощи алгоритмов DBSCAN [4] и kNN [1]. Такая последовательность применения алгоритмов объясняется тем, что гиперпараметры алгоритмов кластеризации проще выбирать для двумерных представлений. Для выявления при кластеризации структуры подмножества, объем которого на несколько порядков меньше полного объема данных, нами предложен новый подход. Его смысл заключается в искусственном увеличении объема такого подмножества с последующим использованием t-SNE. При таком подходе повышается эффективность применения алгоритмов кластеризации, что подтверждается модельными примерами. Применение описанных выше методов в эпидемиологическом исследовании позволило по 12 биомаркерам (систолическое давление, диастолическое давление, общий холестерин, липопротеиды высокой плотности, триглицериды, глюкоза, мочевая кислота, ц-реактивный белок, частота сердечных сокращении, индекс массы тела, охват талии и отношение APOA к APOB) разделить участников на две группы. Используя известные данные, которые не учитывались алгоритмом, выявлена удвоенная частота инфаркта миокарда в одной из групп. Данные предоставлены отделом эпидемиологии хронических неинфекционных заболеваний ФГБУ «НМИЦ ТПМ» МЗ России.

Источники и литература

1) Hastie T. et al. The elements of statistical learning: data mining, inference, and prediction. – New York : springer, 2009. – Т. 2. – С. 1-758.
2) Van der Maaten L., Hinton G. Visualizing data using t-SNE //Journal of machine learning research. – 2008. – Т. 9. – №. 11. MLA
3) Pearson K. LIII. On lines and planes of closest fit to systems of points in space //The London, Edinburgh, and Dublin philosophical magazine and journal of science. – 1901. – Т. 2. – №. 11. – С. 559-572.
4) Ester M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise //kdd. – 1996. – Т. 96. – №. 34. – С. 226-231.