UDC 004.8

Гамбарова Тамара
студентка II курса
Харьковский национальный университет радиоэлектроники, Украина

Биленко Егор Александрович
студент II курса
Харьковский национальный университет радиоэлектроники, Украина

КЛАСТЕРИНГ. ПРИМЕНЕНИЕ КЛАСТЕРНОГО АНАЛИЗА В БИОИНФОРМАТИКЕ

Что такое кластеризация? Отличия от классификации.

Кластеризация - задача классифицирования набора данных, их разделения, в результате чего мы получаем подмножества, так называемые кластеры, сгруппированные по определенному критерию [1].

На первый взгляд может быть неочевидно, чем задача кластеризации отличается от задачи классификации.

Кластеризация - обучение без учителя, вариант обучения системы, без какого-либо вмешательства, в отличии от классификации, обучения системы с учителем, для каждого объекта заранее задан «правильный ответ», и требуется найти зависимость между стимулами и реакциями системы.

Другим важным отличием является то, что критерии, по которым проводится классификация, так называемые метки объектов, могут быть неизвестны изначально, неизвестным может быть даже само множество.

Результат кластеризации зачастую неоднозначен (по причине неоднозначности критерия разделения), и, несмотря на наличие определенных критериев оценки качества в задаче кластеризации (Ф-индекс, индекс Гудмана-Крускала, Ф-мера и др.), для определения ее результатов необходим эксперт предметной области.

Цели кластеризации.

Преимуществом проведения кластеризации является “уменьшение размерностей” выборки, что упрощает дальнейший анализ. Также кластеризация

позволяет определить объекты, которые не подходят ни в один из кластеров. Эти объекты требуют отдельной обработки и их обнаружение может значительно упростить последующую работу с выборкой.

Эти важные возможности, которые предоставляются процессом кластеризации, делают его весьма актуальным во многих областях: экономика, биология, медицина, маркетинг. С развитием новых технологий, а особенно после появления таких геномных технологий, как олигонуклеотидные и кДНК-микрочипы, прикладное применение кластеризация нашла и в биоинформатике.

Применение в биоинформатике.

Типичной задачей биоинформатики является выявление сходств и различий исследуемых последовательностей или структур. Одним из способов зафиксировать обнаруженные закономерности является распределение объектов по группам.

В биоинформатике при анализе набора взаимодействующих генов, состоящих из огромного количества элементов, содержащих большой объем информации, кластерный анализ позволяет выделить общее и различное в свойствах каждого гена, его влияние на тот или иной фактор. Гены, относящиеся к одному и тому же кластеру, обычно отвечают за определенный физиологический процесс или относятся к одному и тому же молекулярному комплексу.

Актуальные способы проведения кластеризации в биоинформатике.

Схема кластерного анализа:

- формирование множества характеристик, по которым будет проводиться анализ;
- применение метода кластеризации для выделения групп объектов;
- проверка и применение данного решения.

Существуют некоторые алгоритмы кластеризации: вероятностные алгоритмы, графовые алгоритмы, спектральная кластеризация [2]. В биоинформатике наиболее используемыми являются иерархический и алгоритм k-средних.

Первый метод основан на построении дендрограммы, которая строится от листьев(изначально каждый элемент выборки) к корню. На каждом шаге алгоритма происходит итеративный процесс слияния двух ближайших кластеров. Алгоритм работает до нахождения необходимого количества подмножеств.

На каждом шаге необходимо уметь вычислять расстояние между кластерами и пересчитывать расстояние между новыми. Простым способом решения задачи является Евклидова мера. Однако эта формула не совсем практична при обработке, например, шкалированных выборок, а также при отрицательной корреляции. Исходные данные такого типа подвергаются Z-нормализации.

Другими вариантами являются использование коэффициента корреляции Пирсона (недостаток: чувствителен к выбросам), ранговой корреляции Спирмена.

Алгоритм k-средних напротив является не иерархическим.

В данном алгоритме дан набор наблюдений и количество кластеров k, на которые наш набор должен быть разбит. Алгоритм работает так, чтобы минимизировать сумму квадратов расстояний от каждой точки кластера до его центра (центр масс кластера). Алгоритм завершается, когда на какой-то итерации не происходит изменения кластеров. Существуют различные эвристики и улучшения алгоритма [3], однако главными его недостатками являются необходимость заранее знать количество кластеров, неоднозначность выбора начальных центров кластеров.

Список источников:

1. Миркин, Б. Г. Методы кластер-анализа для поддержки принятия решений: обзор : препринт WP7/2011/03 [Текст] / Б. Г. Миркин ; Национальный исследовательский университет «Высшая школа экономики». – М. : Изд. дом Национального исследовательского университета «Высшая школа экономики», 2011. – 88 с. – 150 экз.
2. CHAPMAN & HALL Clustering in Bioinformatics and Drug Discovery, Mathematical and Computational Biology Series, 2011.
3. Dan Pelleg, Andrew Moore. Accelerating exact k-means Algorithms with geometric reasoning.