Д.О. ВАСЯЕВА,
А.С. БЕЛЯКОВА

Обзор алгоритмов кластеризации с целью обработки медицинских данных

УДК 004.93'14

Муромский институт (филиал) ФГБОУ ВО «Владимирский государственный университет имени А.Г. и Н.Г. Столетовых», г. Муром

За последнее время медицинские исследования стали содержать в себе все больше и больше данных, которые требуют постоянного анализа и обработки с целью структурирования и подведения результатов исследования. Одним из способов обработки является кластеризация.

Кластеризация (или кластерный анализ) — это задача разбиения множества объектов на группы, называемые кластерами. Внутри каждой группы должны оказаться «похожие» объекты, а объекты разных группы должны быть как можно более отличны. Главное отличие кластеризации от классификации состоит в том, что перечень групп четко не задан и определяется в процессе работы алгоритма [1].

1. Применение кластерного анализа в общем виде сводится к следующим этапам:

2. Отбор выборки объектов для кластеризации.

3. Определение множества переменных, по которым будут оцениваться объекты в выборке. При необходимости – нормализация значений переменных.

4. Вычисление значений меры сходства между объектами.

5. Применение метода кластерного анализа для создания групп сходных объектов (кластеров).

6. Представление результатов анализа [1].

После получения и анализа результатов возможна корректировка выбранной метрики и метода кластеризации до получения оптимального результата.

Классификация алгоритмов

1. Иерархические и плоские.

Иерархические алгоритмы (также называемые алгоритмами таксономии) строят не одно разбиение выборки на непересекающиеся кластеры, а систему вложенных разбиений.

Таким образом, на выходе мы получается дерево кластеров, корнем которого является вся выборка, а листьями — наиболее мелкие кластера. Плоские алгоритмы строят одно разбиение объектов на кластеры [2].

2. Четкие и нечеткие.

Четкие (или непересекающиеся) алгоритмы каждому объекту выборки ставят в соответствие номер кластера, т.е. каждый объект принадлежит только одному кластеру. [2] Нечеткие (или пересекающиеся) алгоритмы каждому объекту ставят в соответствие набор вещественных значений, показывающих степень отношения объекта к кластерам. Т.е. каждый объект относится к каждому кластеру с некоторой вероятностью.

Алгоритмы иерархической кластеризации

Среди алгоритмов иерархической кластеризации выделяются два основных типа: восходящие и нисходящие алгоритмы. Нисходящие алгоритмы работают по принципу «сверху-вниз»: в начале все объекты помещаются в один кластер, который затем разбивается на все более мелкие кластеры [3]. Более распространены восходящие алгоритмы, которые в начале работы помещают каждый объект в отдельный кластер, а затем объединяют кластеры во все более крупные, пока все объекты выборки не будут содержаться в одном кластере. Таким образом строится система вложенных разбиений. Результаты таких алгоритмов обычно представляют в виде дерева [3]. Классический пример такого дерева – классификация животных и растений.

Для вычисления расстояний между кластерами чаще все пользуются двумя расстояниями: одиночной связью или полной связью (см. обзор мер расстояний между кластерами) [3].

К недостатку иерархических алгоритмов можно отнести систему полных разбиений, которая может являться излишней в контексте решаемой задачи.

Алгоритмы квадратичной ошибки

Задачу кластеризации можно рассматривать как построение оптимального разбиения объектов на группы. При этом оптимальность может быть определена как требование минимизации среднеквадратической ошибки разбиения:

$$e^2(X,L) = \sum_{j=1}^{K} \sum_{i=1}^{n_j} \| x_i^{(j)} - c_j \|^2$$

где c_j — «центр масс» кластера j (точка со средними значениями характеристик для данного кластера).

Алгоритмы квадратичной ошибки относятся к типу плоских алгоритмов. Самым распространенным алгоритмом этой категории является метод k-средних [4]. Этот алгоритм строит заданное число кластеров, расположенных как можно дальше друг от друга. Работа алгоритма делится на несколько этапов:

1. Случайно выбрать k точек, являющихся начальными «центрами масс» кластеров.

2. Отнести каждый объект к кластеру с ближайшим «центром масс».

3. Пересчитать «центры масс» кластеров согласно их текущему составу.

4. Если критерий остановки алгоритма не удовлетворен, вернуться к п. 2 [4].

В качестве критерия остановки работы алгоритма обычно выбирают минимальное изменение среднеквадратической ошибки. Так же возможно останавливать работу алгоритма, если на шаге 2 не было объектов, переместившихся из кластера в кластер [5].

К недостаткам данного алгоритма можно отнести необходимость задавать количество кластеров для разбиения.

Нечеткие алгоритмы

Наиболее популярным алгоритмом нечеткой кластеризации является алгоритм c-средних (c-means). Он представляет собой модификацию метода k-средних. Шаги работы алгоритма:

1. Выбрать начальное нечеткое разбиение n объектов на k кластеров путем выбора матрицы принадлежности U размера $n \times k$ [2].

2. Используя матрицу U, найти значение критерия нечеткой ошибки:

$$E^2(X,U) = \sum_{i=1}^{N} \sum_{k=1}^{K} U_{ik} \| x_i^{(k)} - c_k \|^2 ,$$

где c_k — «центр масс» нечеткого кластера k:

$$c_k = \sum_{i=1}^{N} U_k x_i .$$

3. Перегруппировать объекты с целью уменьшения этого значения критерия нечеткой ошибки.

4. Возвращаться в п. 2 до тех пор, пока изменения матрицы *U* не станут незначительными [2].

Этот алгоритм может не подойти, если заранее неизвестно число кластеров, либо необходимо однозначно отнести каждый объект к одному кластеру.

Алгоритмы, основанные на теории графов

Суть таких алгоритмов заключается в том, что выборка объектов представляется в виде графа *G=(V, E)*, вершинам которого соответствуют объекты, а ребра имеют вес, равный «расстоянию» между объектами. Достоинством графовых алгоритмов кластеризации являются наглядность, относительная простота реализации и возможность внесения различных усовершенствований, основанные на геометрических соображениях. Основными алгоритмам являются алгоритм выделения связных компонент, алгоритм построения минимального покрывающего (островного) дерева и алгоритм послойной кластеризации [4].

Алгоритм выделения связных компонент

В алгоритме выделения связных компонент задается входной параметр *R* и в графе удаляются все ребра, для которых «расстояния» больше *R*. Соединенными остаются только наиболее близкие пары объектов. Смысл алгоритма заключается в том, чтобы подобрать такое значение *R*, лежащее в диапазон всех «расстояний», при котором граф «развалится» на несколько связных компонент. Полученные компоненты и есть кластеры [6].

Для подбора параметра *R* обычно строится гистограмма распределений попарных расстояний. В задачах с хорошо

выраженной кластерной структурой данных на гистограмме будет два пика – один соответствует внутрикластерным расстояниям, второй – межкластерным расстояния [5]. Параметр *R* подбирается из зоны минимума между этими пиками. При этом управлять количеством кластеров при помощи порога расстояния довольно затруднительно.

Таблица 1

Сравнение алгоритмов кластеризации данных

Алгоритм кластеризации	Форма кластеров	Входные данные	Результаты
Иерархический	Произвольная	Число кластеров или порог расстояния для усечения иерархии	Бинарное дерево кластеров
k-средних	Гиперсфера	Число кластеров	Центры кластеров
c-средних	Гиперсфера	Число кластеров, степень нечеткости	Центры кластеров, матрица принадлежности
Выделение связных компонент	Произвольная	Порог расстояния R	Древовидная структура кластеров
Послойная кластеризация	Произвольная	Последовательность порогов расстояния	Древовидная структура кластеров с разными уровнями иерархии

В результате медицинских обследований анализируется два типа показателей здоровья:

Субъективные - самооценка человеком своего текущего состояния здоровья [2].

Объективные - выражаются в таких критериях, которые проявляются независимо от воли человека, могут быть определены и сравнимы с предыдущим состоянием и с нормативными характеристиками [3].

В результате различных обследований для анализа формируется набор в несколько сотен различных по своей природек параметров.

Использование алгоритмов кластерного анализа при интерпретации данных медицинских обследований позволит группировать пациентов по степени риска заболеваний, выделить значения характеристик состояния здоровья пациентов, наиболее характерных для рассматриваемых заболеваний, помочь врачу при постановке диагноза.

Литература

1. Артюнина Г.П., Гончар Н.Т., Игнатькова С. А.. Основы медицинских знаний: Здоровье, болезнь и образ жизни - Псков:2003, 304 с.. 2003

2. Пигалов А.П., Соловьева Н.А., Кулакова Г.А., Курмаева Е.А., Волгина С.Я. Оценка здоровья детей и подростков: Учебное пособие для студентов, врачей-интернов, ординаторов, аспирантов педиатрического факультета / Под общей ред. профессора, заведующего кафедрой поликлинической педиатрии КГМУ А.П. Пигалова. — Казань: Центр инновационных технологий, 2006. — 244 с. ISBN 5-93962 167 8

3. Сухарев А. Г. Здоровье и физическое воспитание детей и подростков. — М.: Медицина, 1991. — 272 с., ISBN 5-225-00348-6

4. Воронцов К.В. Алгоритмы кластеризации и многомерного шкалирования. Курс лекций. МГУ, 2007.

5. Jain A., Murty M., Flynn P. Data Clustering: A Review. // ACM Computing Surveys. 1999. Vol. 31, no. Котов А., Красильников Н. Кластеризация данных. 2006.

6. Информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных — www.machinelearning.ru