УДК 519.254

О критериях оценки качества кластеризации

В.В. Журавлева, К.Е. Аюпов
АлтГУ, г. Барнаул

Методы кластерного анализа являются мощнейшими средствами в области анализа данных. Уже несколько десятилетий множество различных алгоритмов кластеризации с успехом используются в самых различных научных областях, начиная от экономики и заканчивая биологией и психологией. Кластеризацию используют и как самостоятельный инструмент анализа данных, и как предварительный этап для других методов анализа [1, 2].

Общая схема процесса кластеризации данных включает пять основных этапов:

– выделение существенных характеристик исследуемых объектов;

– определение меры сходства;

– разбиение множества объектов на кластеры;

– оценка качества кластеризации;

– представление и интерпретация результатов.

На каждом из первых трех этапов мы можем допустить существенные ошибки, которые могут исказить результат. По этой причине следующий этап – оценка качества кластеризации – является не менее значимым.

Очевидно, что «абсолютно объективной» кластеризации не существует. Все реальные объекты имеют бесконечное число свойств, и выделение некоторого конечного подмножества этих свойств субъективно. Меры близости также выбираются субъективно. Если известна цель, для достижения которой строится разбиение, то качество проверяется тем, хорошо ли кластеризация способствует достижению этой цели. Эта проверка носит объективный характер, но выбор суперцели опять-таки субъективен [1].

Основная задача кластеризации формулируется так: разделить объекты на группы таким образом, чтобы объекты одной группы имели большое сходство, а сходство между объектами разных групп было малым. Согласно приведенного определения формулируются основные критерии качества кластеризации: компактность, отделимость и, редко используемый, концентрация.

Компактность означает, что элементы одного кластера должны быть как можно ближе друг к другу (обладать высокой степенью сходства). Это свойство можно выразить через расстояния между элементами в кластере, плотностью внутри кластера или же объемом, занимаемым кластером в пространстве [2].

Свойство ***отделимости*** значит, что элементы разных кластеров должны быть как можно дальше друг от друга (обладать низкой степенью сходства) [2]. Расстояние между кластерами обычно измеряется одним из трех способов:

– расстояние между ближайшими элементами кластеров;

– расстояние между наиболее удаленными элементами кластеров;

– расстояние между кластерными центрами.

Концентрация означает, что элементы кластера должны быть сконцентрированы вокруг центра кластера. Этот пункт используется гораздо реже, потому что далеко не во всех алгоритмах кластеризации используется понятие центра кластера [2].

Для самих показателей качества кластеризации обычно вводят следующую классификацию: внешние, внутренние и относительные. К внутренним показателям относятся те, которые учитывают априорную информацию о структуре кластеров в рассматриваемом множестве данных. К внешним относят показатели, которые не имеют априори знаний о структуре классов и при оценке опираются только на ту информацию, которую можно получить из самого разбиения. Относительные показатели оценивают качество, сравнивая несколько кластерных структур между собой, не имея априорной информации. Обзор основных подходов к оценке качества кластеризации проведен в работе Сивоголовко Е.В. [2].

В среде электронных таблиц Microsoft Excel (на VBA) реализован алгоритм кластеризации, позволяющий строить классы «необычной» формы (путем объединения малых «сфер»-кластеров) [3]. Данный алгоритм применялся авторами для построения кластерной структуры данных по количеству вызовов скорой помощи и комплексу геофизических факторов [4, 5]. Для этого алгоритма возникает

проблема оценки качества кластеризации (а также при выборе оптимального количества кластеров). В общем случае описанные выше критерии качества не применимы.

Используем в качестве показателя компактности кластеров среднее расстояние между соседними вершинами минимального остовного дерева (МОД), построенного на всех объектах-точках кластера [5].

Для построения МОД взвешенного связного неориентированного графа можно использовать алгоритм Прима. Построение МОД начинается с произвольной вершины. Рост дерева происходит до тех пор, пока не будут исчерпаны все вершины графа. Данная стратегия является «жадной», то есть на каждом шаге к дереву добавляется ребро, которое вносит минимально возможный вклад в общий вес. Результатом алгоритма является остовное дерево с минимальным суммарным весом [5].

Вернемся к проблеме оценки качества кластеризации. Для кластеров произвольной формы (например, ленточных) удобно сравнивать степень компактности кластеров через среднее значение длины ребер в МОД кластера. Наиболее компактными будем считать кластеры с наименьшим значением данного показателя.

Критерий качества кластеризации будем вычислять как среднее значение показателей компактности для построенной кластеризации. Наиболее качественной будем считать кластерную структуру с наименьшим значением выбранного критерия.

Итак, оценка качества кластеризации заслуженно считается сложной областью анализа данных. Проблему качества сложно выразить семантически и также сложно подогнать ее под математическую модель. Как правило, для любого показателя качества кластеризации существует такое множество, на котором его оценка является верной. Но даже лучший показатель ошибается на определенных тестовых множествах. Универсального решения в этом вопросе не существует, однако отказ от оценки качества кластеризации неприемлем. В целом, для повышения эффективности в оценке качества кластеризации и получения объективного результата лучше пользоваться не одним показателем, а их совокупностью [2].

Библиографический список

1. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. – Новосибирск: ИМ СО РАН, 1999. – 270 с.

2. Сивоголовко Е.В. Методы оценки качества четкой кластеризации. // Компьютерные инструменты в образовании. – Тверь, 2011 – Вып. 4 (96) – С. 14–31.

3. Журавлева В.В., Бондарева А.А. Описание одного алгоритма кластеризации типа Forel // МАК-2015 : сборник трудов восемнадцатой всероссийской конференции по математике. Барнаул, 1-5 июля, 2015. – Барнаул: Изд-во Алт. ун-та, 2015. – С. 142–144.

4. Журавлева В.В. Исследование связи между состоянием геомагнитного поля и обострением сердечно-сосудистых заболеваний // Известия АГУ. – Барнаул, 2011. – №1-1(69). – С. 98–100.

5. Журавлева В.В., Аюпов К.Е. Применение метода кластерного анализа для обнаружения зависимости обострений сердечно-сосудистых заболеваний от геофизических факторов // Ломоносовские чтения на Алтае: фундаментальные проблемы науки и образования : сборник научных статей международной конференции. Барнаул, 20-24 октября, 2015. – Барнаул : Изд-во Алт. ун-та, 2015. – С. 831–834.

УДК 532.135

Исследование течения расплава полимера в канале с внезапным сужением

А.Е. Кузнецов
АлтГТУ, г. Барнаул

Изучению течений полимерных расплавов и растворов в различных сходящихся каналах посвящено большое число работ. В этих работах часто отмечают возникновение вторичных течений (или вихрей) во входной области щелевого канала. Размеры вторичных течений могут зависеть от таких факторов как температура расплава, скорость течения и некоторых других [1–3]. Кроме того, такие течения могут показывать трехмерный характер течения, когда размеры вихря зависят от положения секущей плоскости [1].

В настоящее время для описания течений расплавов линейных и разветвленных полимеров часто используются уравнения, учитывающие в той или иной мере существенные особенности строения полимерных жидкостей [4, 5]. Учет этих особенностей может вызывать затруднения, поэтому наиболее востребованными являются модели, в основе которых лежит мезоскопический подход. В этом случае поведение полимерной макромолекулы заменяется поведением одного или нескольких релаксаторов, а переход к макроскопическому описанию осуществляется методами статистической меха-