Метод кластеризации на основе анализа плотности точек

Клышинский Э.С.
МИЭМ НИУ ВШЭ
eklyshinsky@hse.ru

Аннотация. В статье предлагается новый метод кластеризации, основанный на использовании информации о плотности расположения объектов в пространстве. В отличие от существующих методов, данный метод не нуждается в выборе размера окрестности, для которой будет проводиться объединение объектов в один кластер. В статье предлагается метод объединения кластеров между собой для снижения эффектов, связанных с шумом в функции плотности расположения объектов.

Ключевые слова: плотностная кластеризация, объединение кластеров

1 Введение

Кластеризация является одним из методов, используемы при анализе данных или машинном обучении. Высокая степень важности решаемых задач привела к тому, что в данной области появилась целая плеяда различных методов. Методы отличаются между собой по простоте реализации, пригодности для обработки различных данных, базовым принципам, заложенным в их основу. Так, при наличии плотных выпуклых далеко отстоящих друг от друга кластеров хорошо работают центроидные методы, например, метод k-средних [Hartigan, 1979]. В этих методах берется несколько точек, считающихся центрами кластеров. При помощи определенной процедуры точки перемещаются в центры имеющихся кластеров и в дальнейшем считается, что точка принадлежит кластеру, для которого расстояние до центра которого является минимальным.

Точность подобных методов падает при сближении кластеров и пересечении их границ. Также методы плохо работают на кластерах невыпуклой формы. В подобной ситуации применяются методы, основанные на плотности распределения объектов, например, DBSCAN [Ester, 1996]. В них кластер определяется как совокупность плотно расположенных точек, тогда как границы кластеров определяются по областям с низкой плотностью.

Описанные выше методы предполагают, что для кластеризуемых объектов существует метрика, определяющая расстояние между объектами. Однако в ряде случаев мы можем лишь определить степень сходства объектов, не обладающую транзитивностью или суперпозицией. Так, например, при кластеризации графов задать расстояние между вершинами чаще всего невозможно.

Часть недостатков описанных методов может быть решена при переходе от четких методов кластеризации, когда каждому объекту приписывается принадлежность только одному кластеру, к нечетким, когда объекту приписывается вектор коэффициентов, показывающих вероятность принадлежности объекта каждому из кластеров (c-средних, FLAME и другие [Berkhin, 2002;Bezdek, 1981; Fu, 2007; Кулажский, 2014]).

Подавляющее большинство методов предполагают априорное знание определенной информации о кластерах, которые необходимо найти. Так, например, в методе k-средних необходимо знать число кластеров. Метод DBSCAN [Ester, 1996] требует для начала работы задать максимальное возможное расстояние между объектами, принадлежащими одному кластеру. Сам выбор метода кластеризации определяется формой кластеров, их плотностью, наличию объектов на границах кластеров, расстоянием между объектами. В большинстве случаев кластеризация проводится несколько раз с вычислением критериев, характеризующих получаемые кластеры. По завершении прогонов выбираются такие характеристики кластеризации, на которых достигаются наилучшие значения этих критериев.

На практике зачастую задача состоит именно в том, чтобы определить сами характеристики кластеров. Получается замкнутый круг, выход из которого находится перебором параметров кластеризации или надеждой на то, что начальные предположения были корректны. Ситуация усложняется тем, что в ряде ситуаций может отличаться даже мнение экспертов относительно разбиения множества объектов на классы.

В данной работе сделана попытка разработать новый метод четкой кластеризации, работа которого не будет завязана на определение начальных параметров. В качестве основы было решено использовать методы, основанные на плотности распределения объектов.

2 Метод кластеризации

Пусть объект $\mathbf{o}$ задан как точка в n-мерном пространстве признаков $\mathbf{o} = \langle x_1, x_2, \ldots, x_n \rangle$. Также пусть задано множество классов $C = \{c_1, \ldots, c_k\}$. Задачей кластеризации в этом случае будет соотнести каждый объект из некоторого множества $O = \{\mathbf{o}_i\}$, $i \in [1, m]$, где m – количество объектов, с одним классом из C.

Будем считать, что имеется возможность рассчитать расстояние между двумя объектами $d_{i,j} = d(\mathbf{o}_i, \mathbf{o}_j)$. Тогда можно получить матрицу расстояний между всеми объектами $\mathbf{D} = |d_{i,j}|$, $i,j \in [1, m]$. Исход из матрицы расстояний можно посчитать плотность объектов ρ в окрестности δ заданного объекта [Маннинг и др., 2011].

В основу предлагаемого метода положено предположение, что центры кластеров находятся в областях с большими плотностями, тогда как границы кластеров пролегают по «анти-водоразделу», то есть линии, в

любой точке которой в перпендикулярном направлении функция может только возрастать. Эта линия будет проходить по имеющимся седловым точкам и будет стремиться попасть в точки минимума. Пример подобного разделения приведен на рис. 1.

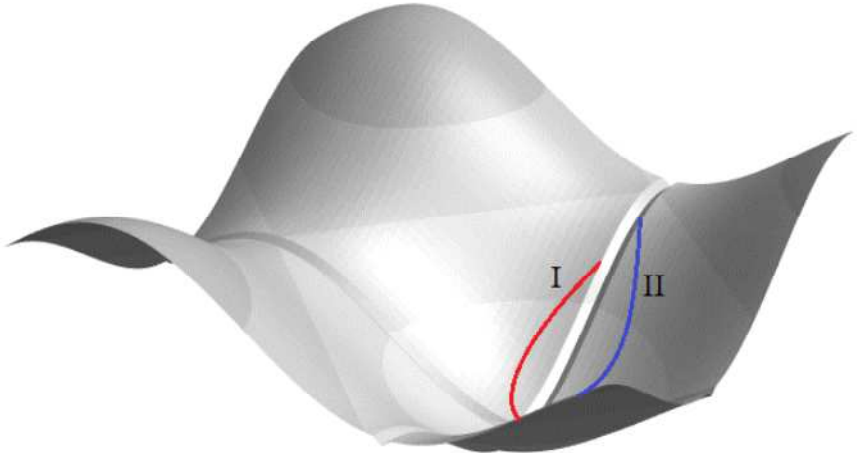


Рис. 1. Пример разделения поверхности на кластеры

Кластеризация в этом случае будет проводиться по следующему алгоритму.

Шаг 1. Помечаем все точки как не принадлежащие никакому кластеру (например, кластеру с номером *-1*). Номер текущего кластера r присваиваем равным *1*.

Шаг 2. Находим точку $\mathbf{o}_e$ с максимальным значением плотности среди точек, не отнесенных ни к одному кластеру. Если такая точка не существует, алгоритм прекращает свою работу. Найденная точка помещается в множество точек нового кластера: $\mathbf{o}_e \in Q_r$.

Шаг 3. Для всех точек в множестве $Q_r = \{\mathbf{q}_i\}$ последовательно выполняем Шаг 4. Далее переходим к Шагу 6.

Шаг 4. Для текущей точки $\mathbf{q}_i$ рассмотрим все точки $\mathbf{o}_j$, входящие в ее окрестность размером δ. Если между текущей точкой $\mathbf{q}_i$ и точкой $\mathbf{o}_j$ из ее окрестности нет других точек (см. Шаг 5) и при этом плотность в точке $\mathbf{o}_j$ меньше, чем в точке $\mathbf{q}_i$, то точка $\mathbf{o}_j$ может быть присоединена к тому же кластеру. Если точка $\mathbf{o}_j$ не принадлежит ни одному из кластеров, то она включается в множество Q_r: $\mathbf{o}_j \in Q_r$. При этом в множество связей P_r добавляется пара $< \mathbf{q}_i, \mathbf{o}_j >$: $< \mathbf{q}_i, \mathbf{o}_j > \in P_r$. Если точка $\mathbf{o}_j$ уже была помещена в некоторый кластер s, то из множества P_s извлекается пара $< \mathbf{p}_t, \mathbf{o}_j >$. Если $d(\mathbf{q}_i, \mathbf{o}_j) < d(\mathbf{p}_t, \mathbf{o}_j)$, то точка $\mathbf{o}_j$ перемещается из кластера s в кластер r, а в противном случае остается в своем кластере.

Шаг 5. Считаем, что между точками $\mathbf{q}_i$ и $\mathbf{o}_j$ нет других точек, если внутри окружности диаметром $d(\mathbf{q}_i, \mathbf{o}_j)$ нет других точек.

Шаг 6. Увеличиваем номер текущего кластера r на единицу и переходим к Шагу 2.

На практике при использовании предложенного алгоритма граница между кластерами может проходить по некоторой линии, условно представленной на рис. 1 (см. линии I и II). Эти линии отвечают требованиям, предъявленным на Шаге 4 алгоритма. При этом они не совпадают с некоторой идеальной линией (изображенной светлой линией на рис. 1).

3 Подбор параметров

Рассуждения, использованные на шаге 5 являются эмпирическими. При объединении объектов в один кластер с некоторой долей уверенности необходимо гарантировать, что между точками $\mathbf{q}_i$ и $\mathbf{o}_j$ плотность продолжает убывать. Для этого мы рассуждали следующим образом. Окружность с выбранным диаметром представляет собой геометрическое место точек, для которых сумма квадратов расстояний до точек $\mathbf{q}_i$ и $\mathbf{o}_j$ равно квадрату диаметра окружности. Если точка находится внутри окружности, нам «выгоднее» добраться сперва до нее, а лишь затем до точки $\mathbf{o}_j$, тогда как если точка находится вне выбранной окружности, то ее надо рассматривать саму по себе (см. рис. 2).

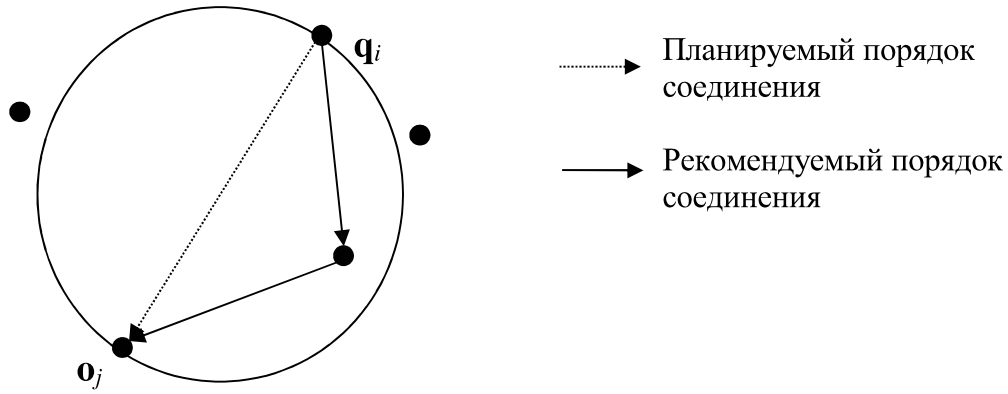


Рис. 2. Проверка на наличие других объектов при объединении

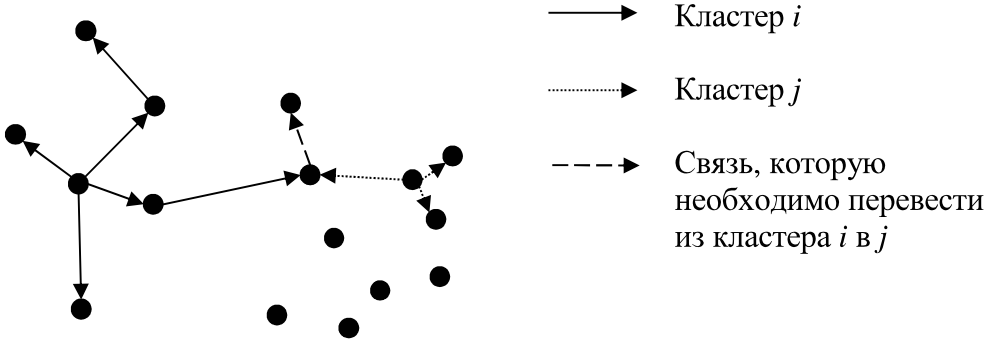


Рис. 3. Перенос точки из одного кластера в другой

Таким образом, если внутри окружности находится еще одна точка, то в первую очередь нужно попытаться добавить в кластер ее, а лишь затем переходить к точке $\mathbf{o}_j$.

Корректнее было бы использовать аппроксимацию плотности между точками, но точность подобной аппроксимации и возможность ее использования нуждается в отдельном исследовании.

На Шаге 4 мы позволяем переносить точки из кластера в кластер если расстояние до нового кластера меньше. В этом случае может возникнуть ситуация, когда мы переносим точку, которая была использована для добавления точек в другой кластер (см. рис. 3). В такой ситуации можно добавить все точки, связанные с данной, в новый кластер. Хотя, как будет показано далее, подобный перенос необходим не всегда.

Теперь вернемся к вопросу о расчете плотности расположения объектов в окрестности заданного. К данному вопросу можно подойти так, как это было сделано в методе DBSCAN [Ester, 1996], то есть взять для расчета окрестность заданного размера и считать плотность как количество объектов в ней. В этом случае перед нами снова встает вопрос выбора размеров окрестности, что требует знания структуры имеющихся данных. Заметим, что размер окрестности обязательно должен фиксироваться для всех объектов, так как при линейном росте расстояния до очередной точки площадь окрестности растет квадратично. В связи с этим плотность в окрестности большего размера будет ощутимо меньше.

Размер окрестности может быть рассчитан для ансамбля объектов на начальных этапах работы алгоритма. Так, например, можно считать, что объекты, расположенные ближе, чем среднее расстояние между всеми объектами, принадлежат одному кластеру. В связи с этим можно рассчитать среднее расстояние между объектами, расположенными ближе, чем среднее расстояние между всеми объектами. Однако подобная мера предполагает наличие компактных круглых кластеров, расположенных далеко друг от друга. Если кластеры будут находиться близко или иметь форму подковы или кольца, подобное расстояние будет слишком велико. Как следствие, плотность будет рассчитываться по слишком большой окрестности и будет излишне усреднена.

Еще одним вариантом расчета плотности является вычисление среднего расстояния до k объектов, расположенных ближе всего к данному. Можно считать, что чем меньше это расстояние, тем выше плотность расположения объектов. Подобный подход используется в методе FLAME [Fu, 2007] и неплохо себя зарекомендовал на практике. Основным его достоинством является тот факт, что нам нет необходимости заботиться о предварительном выборе размеров окрестности. Значение k обычно выбирается не очень большим – от 5 до 10 объектов.

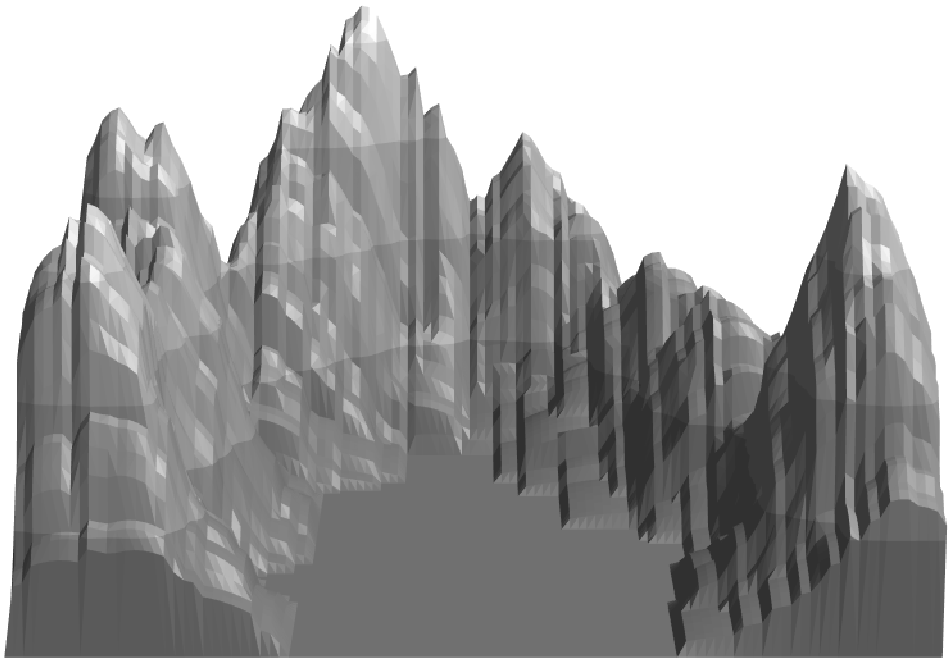


Рис. 4. Плотность объектов, расположенных случайным образом в форме подковы

В описанном выше алгоритме на Шаге 4 проводится объединение объектов в кластеры. Очевидно, что объекты, находящиеся слишком далеко друг от друга, объединяться не должны. На данном шаге фиксированная окрестность служит плохую службу, так как при высокой плотности объектов расстояние должно уменьшаться, тогда как при низкой плотности – увеличиваться до некоторого предела. В этом смысле удобным оказывается оценивать размер окрестности, внутри которой проводится объединение, индивидуально для каждого объекта. Здесь также могут использоваться ближайшие соседи объекта, при этом рекомендуется выбирать объектов больше, чем при оценке плотности (порядка 10-20). Заметим, что при малом количестве точек есть вероятность того, что у нас не получится соединиться с точками, которые расположены достаточно близко, но дальше рассчитанного порога. Как следствие кластеры будут состоять из малого количества близко расположенных точек.

4 Объединение кластеров

На практике распределение плотностей по объектам не будет гладким, а также будет содержать определенное число локальных выбросов. На рис. 4 показана аппроксимация плотности объектов, расположенных случайным образом в виде подковы. На рисунке видно, что рядом с максимумами расположены локальные выбросы. В связи с этим метод будет выделять ложные кластеры. Усреднение плотности по соседним точкам не всегда приводит к устранению ложных кластеров, хотя может несколько улучшить картину. В связи с этим нами был предложен метод

объединения соседних кластеров. Заметим, что объединение кластеров может потребоваться далеко не всегда. Так, если кластеры расположены близко и плавно перетекают один в другой, предложенный алгоритм объединит их в один. Таким образом, при применении метода перед нами стоит дилемма: получить кластеры небольшого размера, подверженные шуму, или большие кластеры, однако с вероятностью объединить два кластера в один.

Заметим также, что подобному недостатку также подвержены некоторые методы, основанные на плотности расположения точек. Так, например, в методе FLAME выбираются все точки, являющиеся локальными максимумами. В связи с этим число кластеров однозначно будет равно числу локальных максимумов, тогда как реальное количество кластеров может быть меньше.

Метод объединения кластеров заключается в следующем. Будем считать, что два кластера могут быть объединены в один, если у них имеется более b пар точек таких, что точки принадлежат разным кластерам, а расстояние между ними меньше, чем среднее расстояние до k соседних точек одной из них (рис. 5). Данный метод должен использоваться как Шаг 7 предложенного выше алгоритма.

При подобном объединении кластеров нет необходимости отслеживать связи между точками при переносе точек из одного кластера в другой (см. рис. 3). Это позволяет увеличивать число точек на границе кластеров, что в итоге приводит к тому, что они чаще объединяются. Однако если объединение кластеров не используется, алгоритм необходимо модифицировать. После переноса точки из одного кластера в другой необходимо добавить все точки, связанные с данной.

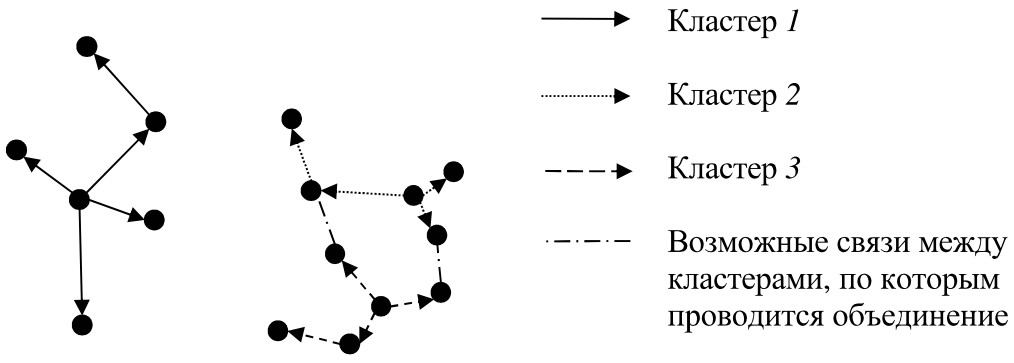
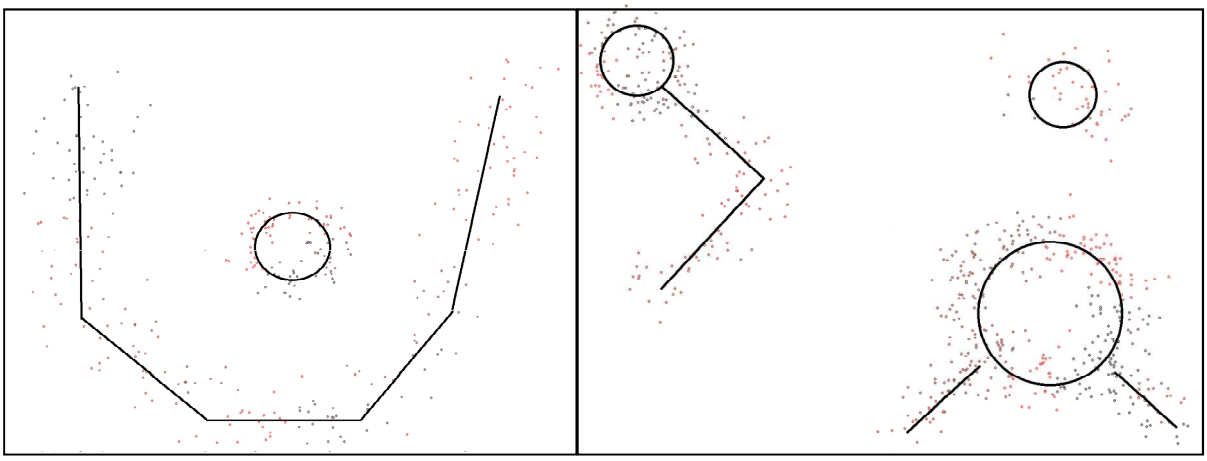
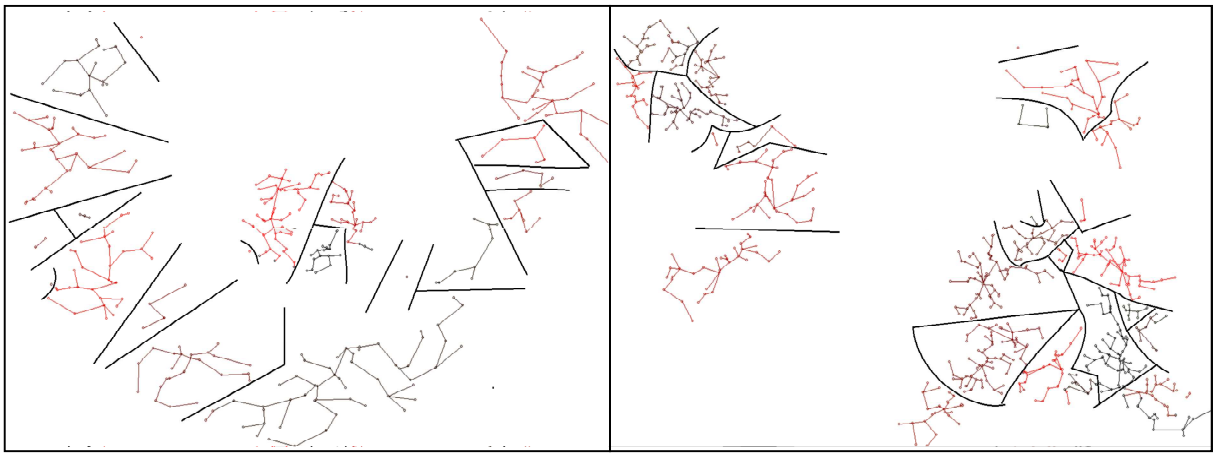


Рис. 5. Перенос точки из одного кластера в другой



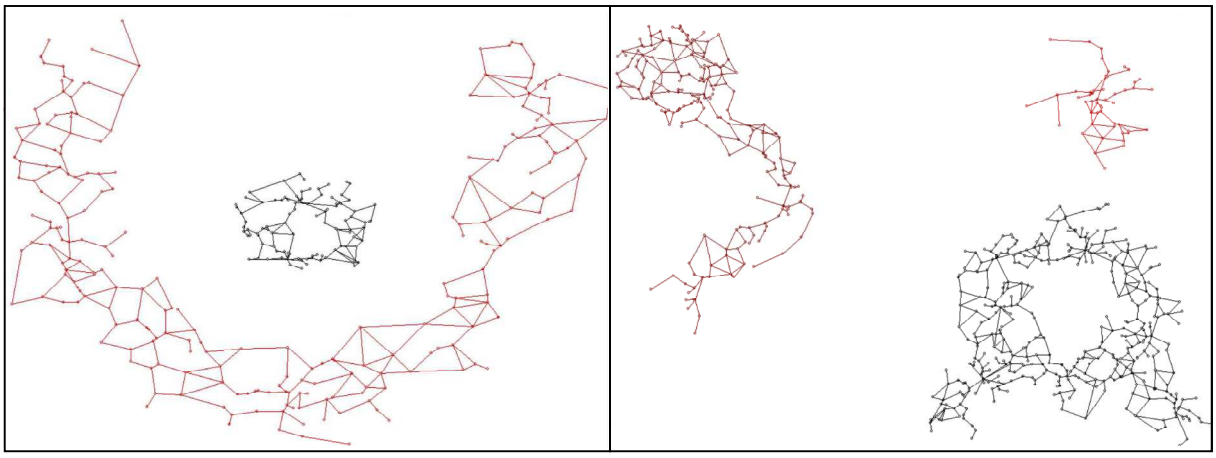
а б

Рис. 6. Примеры тестовых данных



а б

Рис. 7. Примеры кластеризации без объединения кластеров



а б

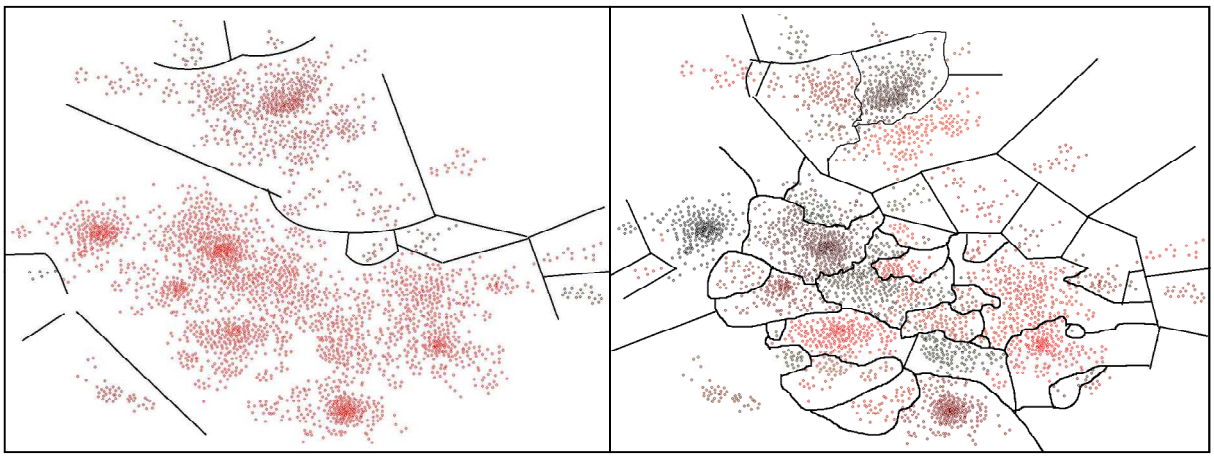
Рис. 8. Примеры кластеризации с объединением кластеров

5 Результаты экспериментов

Для проверки работоспособности метода проводилась генерация случайных последовательностей с заданным распределением точек на плоскости. Примеры последовательностей показаны на рис. 6. На нем же показаны линии, вдоль которых генерировались точки. Результаты кластеризации с объединением кластеров показаны на рис. 7. Результаты кластеризации с объединением кластеров показаны на рис. 8. Здесь линиями показаны точки, для которых проводилось объединение в кластер. Толстыми линиями показаны примерные границы кластеров.

Заметим, что F_1-мера для кластеров, показанных на рис. 7 не превышает *0,5*. При это F_1-мера для кластеров, показанных на рис. 8 равна *1*. Справедливости ради заметим, что среднее значение F_1-меры для 20 прогонов составляет порядка 0,95 с минимальным значением около 0,66 (случай объединения двух кластеров в один).

Картина разительно меняется при обработке данных с большим количеством плотных близкорасположенных кластеров. На рис. 9а показана кластеризация с объединением, тогда как на рис. 9б – без объединения (линии соединения точек убраны, чтобы была возможность оценить истинную плотность расположения точек). Заметим, что природа кластеризуемых данных предполагала наличие большого числа небольших кластеров вместо малого числа больших. В связи с этим разбиение на рис. 9б представляется более корректным.



а б

Рис. 9. Примеры кластеризации с объединением кластеров и без него

6 Выводы

Работа предложенного метода кластеризации, в отличие от метода DBSCAN, не зависит от начального выбора окрестности, в которой точки могут считаться принадлежащими одному кластеру. Размер этой окрестности выбирается динамически исходя из текущей плотности расположения точек. Кроме того, метод не требует априори задавать число

кластеров, так как оно может быть не известно. Однако шум, возникающий при вычислении плотности расположения объектов, может приводить к образованию ложных кластеров. В связи с этим в алгоритм введен шаг объединения близкорасположенных кластеров, имеющих значительную общую границу. Метод показал хорошее качество при работе со случайно сгенерированными данными, однако нуждается в проверке на стандартных наборах данных. Кроме того, в дальнейшем метод планируется применить к кластеризации слов в текстах на естественном языке [Клышинский и др., 2013].

7 Благодарности

Работа выполнена при поддержке гранта РГНФ № 12-04-00060.

8 Список литературы

[Berkhin, 2002] Berkhin P. Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, 2002. 56 p.

[Bezdek, 1981] Bezdek J.C. Pattern Recognition with Fuzzy Objective Function Algorithms, Kluwer Academic Publishers, 1981, 256 p.

[Ester, 1996] Ester M., Kriegel H.-P., Sander J., Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise// In Proc. of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231.

[Fu, 2007] Fu L., Medico E. FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data // BMC Bioinformatics. 2007. №8

[Hartigan, 1979] Hartigan J. A., Wong, M. A. Algorithm AS 136: A K-Means Clustering Algorithm // In Journal of the Royal Statistical Society, s. 28. pp. 100–108.

[Калужский, 2014] Калужский А.А. Система онлайн кластеризации на базе методов Fuzzy C-Means и MST // Сб. трудов научно-практического семинара «Новые информационные технологии в автоматизированных системах-17», М.: ИПМ им. М.В. Келдыша РАН, 2014, сс. 145-149.

[Клышинский и др., 2014] Клышинский Э.С., Кочеткова Н.А. Кластеризация семантически связанных слов из неразмеченного текста на русском языке // Сб. трудов научно-практического семинара «Новые информационные технологии в автоматизированных системах-16», М.: МИЭМ НИУ ВШЭ, 2013, сс. 311-315.

[Маннинг и др., 2011] Маннинг К.Д., Рагхаван П., Шютце Х. Введение в информационный поиск, М.: Вильямс. 528 стр.