электронный журнал
МОЛОДЕЖНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
Издатель ФГБОУ ВПО "МГТУ им. Н.Э. Баумана". Эл No. ФС77-51038.

УДК 00

Модуль обработки чертежей сапр линейного транспорта: 2. задача кластеризации

 # 09, сентябрь 2012

Горин Я.А.

Научный руководитель к.т.н. Волосатова Т.М.
Кафедра РК6, МГТУ им. Н.Э. Баумана, Москва, Россия

МГТУ им. Н.Э. Баумана
bauman@bmstu.ru

Введение

При обработке чертежей плана возникает задача фильтрации исходных данных с целью избавления от выбросов. Выбросы – точки (или тексты), лежащие вне коридора проектирования трассы, но имеющие такое же графическое обозначение, как и точки (или тексты), лежащие внутри коридора проектирования трассы [5].

Одним из подходов к решению данной задачи является кластеризация точек рельефа. Помимо избавления от выбросов кластеризация позволяет разбить задачу на подзадачи в случае размещения нескольких коридоров проектирования трассы на одном чертеже.

Основные обозначения

Пусть множество $I = \{I_1, I_2, \ldots, I_n\}$ обозначает n точек на плоскости. Каждая точка из I обладает некоторым множеством *наблюдаемых* показателей или характеристик $C = (C_1, C_2, \ldots, C_p)^T$. В общем случае наблюдаемые характеристики могут быть как *количественными,* так и *качественными*. Количественные данные иногда называют *измерениями*. Результат измерения i-ой характеристики I_j точки будем обозначать символом x_{ij}, а вектор $X_j = [x_{ij}]$ размерности $p \times 1$ будет отвечать каждому ряду измерений (для j-ой точки). Таким образом, для множества точек I исследователь располагает множеством векторов измерений $X = \{X_1, X_2, \ldots, X_n\}$, которые описывают множество I.

Выбор наблюдаемых характеристик (определение признаков оценки объектов). В качестве наблюдаемых характеристик выбираем координаты x и y геодезических точек. Координата z, во-первых, отсутствует в явном виде на изыскательском чертеже, во-вторых, должна быть исключена из рассмотрения по определению *коридора проектирования трассы* (см. [5]). Единицы измерение x и y совпадают. Следовательно, результат не будет зависеть от масштабов, от единиц измерения признаков. В нормировании наблюдаемых характеристик нет необходимости.

Задача кластерного анализа

Пусть m – целое число, меньшее, чем n. Задача кластерного анализа заключается в том, чтобы на основании данных, содержащихся в множестве X, разбить множество точек I на m кластеров (подмножеств) $\pi_1, \pi_2, \ldots, \pi_m$ так, чтобы каждый объект I_i принадлежал одному и только одному подмножеству разбиения и чтобы объекты, принадлежащие

одному и тому же кластеру, были *сходными*, в то время как объекты, принадлежащие разным кластерам, были разнородными (*несходными*) [1].

Решением задачи кластерного анализа является разбиение, удовлетворяющее некоторому критерию оптимальности. Этот критерий может представлять собой некоторый функционал, выражающий уровни желательности различных разбиений и группировок. Этот функционал называют *целевой функцией* [1].

Целями кластеризации приминительно к решаемой задаче являются:

1. Понять структуру множества объектов X^l, разбив его на группы схожих объектов. Упростить дальнейшую обработку данных и принятия решений, работая с каждым кластером по отдельности (стратегия “разделяй и властвуй”). В этом случае число кластеров стараются сделать как можно меньшим.

2. Выделить нетипичные объекты, которые не подходят ни к одному из кластеров. Эту задачу называют одноклассовой классификацией, обнаружением нетипичности или новизны (novelty detection).

Расстояния между объектами (метрика).

Для определения сходства между объектами используется понятие расстояния $d(X_i, X_j)$. Чем меньше расстояние, тем более похожими считаются объекты. Наиболее употребительные функции расстояний приведены в [1].

Учитывая специфику решаемой задачи, остановимся на использовании *евклидова расстояния* в качестве меры близости между точками на плоскости:

$$d_2(X_i, X_j) = \left[\sum_{k=1}^{p} (x_{ki} - x_{kj})^2\right]^{\frac{1}{2}}$$

В качестве **расстояний между кластерами** используются длины ребер, соединяющих кластеры.

Реализованный метод кластеризации

Реализованный метод кластеризации состоит из двух этапов:

1. Кластеризация на основе теории графов,
2. Иерархическая агломеративная кластеризация на основе кластеров полученных на первом этапе.

Таким образом, мы получаем более тонкую, двухуровневую, структуру кластеров: каждый кластер верхнего уровня распадается на более мелкие подкластеры нижнего уровня.

Определение числа кластеров. В разработанном алгоритме число кластеров зависит от параметров кластеризации на первом и втором этапе - α_1 и α_2.

Кластеризация на основе теории графов

Обширный класс алгоритмов кластеризации основан на представлении выборки в виде графа. Вершинам графа соответствуют объекты выборки, а рёбрам - попарные расстояния между объектами $d_{ij} = d(X_j, X_i)$.

На плоскости заданы n точек. Необходимо построить дерево, вершинами которого являются все заданные точки и суммарная длина всех ребер минимальна. Такое дерево называется *Евклидовым Минимальным Остовным Деревом* (*ЕМОД, EMST - Euclidean Minimum Spanning Tree*).

ЕМОД множества S из n точек на плоскости может быть построено на основании триангуляции Делоне множества S, за оптимальное время $O(n)$. В свою очередь триангуляция Делоне множества S может быть построена за время $O(n \cdot log(n))$. Следовательно, ЕМОД множества S из n точек на плоскости может быть построено за оптимальное время $O(n \cdot log(n))$.

Детальное описание построение ЕМОД можно найти в [3].

Удаление любого из ребер ЕМОД разбивает дерево на лес. Существует множество способов выбрать ребро для удаления.

В реализованном алгоритме используется модификация идеи изложенной в [4]. При выборе ребра – кандидата на удаление мы сравниваем его длину с длинами других ребер, прилегающих к его вершинам. Назовем ребро r *несовместимым*, если его длина l значительно больше $\hat{l}$, где $\hat{l}$ - длина минимального ребра, смежного ребру r.

В реализованном алгоритме удаляются все ребра, чья длина $l > \alpha_1 \cdot \hat{l}$. На момент написания статьи оптимизация параметра α_1 не проводилась, $\alpha_1 = 5$.

В качестве дополнительного критерия для удаления ребра может выступать – длина отклонения от *диаметрального пути*. Диаметральный путь – самый длинный путь по дереву. Когда точки данных располагаются в длинные цепочки, минимальное покрывающее дерево образует естественный скелет для цепочки.

Иерархическая агломеративная кластеризация

Среди алгоритмов иерархической кластеризации различаются два основных типа. Дивизимные или нисходящие алгоритмы разбивают выборку на всё более и более мелкие кластеры. Более распространены агломеративные или восходящие алгоритмы, в которых объекты объединяются во всё более и более крупные кластеры.

На данном этапе работы алгоритма уже имеется определенное число кластеров, полученных с помощью кластеризации на основе теории графов. В качестве связей между кластерами будем рассматривать удаленные на первом этапе ребра. Будем возвращать ребро r между кластерами U и V, удаленное на предыдущем этапе, если его длина l значительно меньше как $\overline{l_U}$, так и $\overline{l_V}$, где $\overline{l_U}$ и $\overline{l_V}$ – средние длины ребер в соседних ребру r кластерах.

В реализованном алгоритме возвращаются все ребра, удаленные на предыдущем этапе, чья длина $l < \alpha_2 \cdot \overline{l_U}$ и $l < \alpha_2 \cdot \overline{l_V}$. На момент написания статьи оптимизация параметра α_2 не проводилась, $\alpha_2 = \alpha_1$.

Функционалы качества кластеризации

Для сравнения качества разбиения на классы используется ряд функционалов качества. Задачу кластеризации можно ставить как задачу дискретной оптимизации: необходимо так приписать номера кластеров Y_i объектам X_i, чтобы значение выбранного функционала качества приняло наилучшее значение. По сути дела, каждый метод кластеризации можно рассматривать как точный или приближённый алгоритм поиска оптимума некоторого функционала. Существует много разновидностей функционалов качества кластеризации, наиболее употребляемые из них могут быть найдены в [1,2].

Так как основой данного метода кластеризации является кластеризация на основе теории графов, можно воспользоваться какой-либо статистикой, получаемой из МОД. Одной из полезных статистик, получаемой из МОД, является *распределения длин ребер*. В качестве функционала предлагается использовать функцию от *распределения длин ребер* МОД, построенного на точках относящихся к одному кластеру.

Достоинства и недостатки.

Достоинствами графовых алгоритмов кластеризации являются:

1. наглядность,
2. относительная простота реализации,
3. возможность вносить различные усовершенствования, опираясь на простые геометрические соображения.

Результаты работы алгоритма

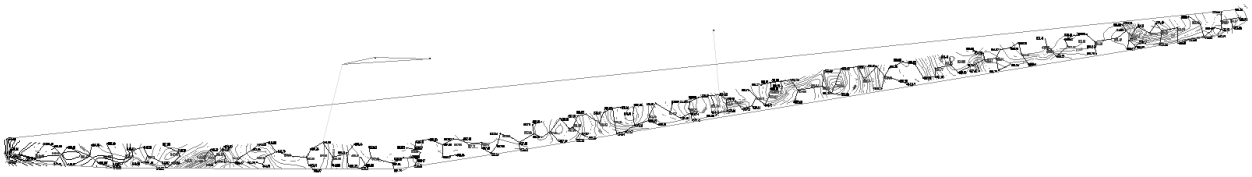

Рис. 1 Единственный кластер и наличие “выбросов”

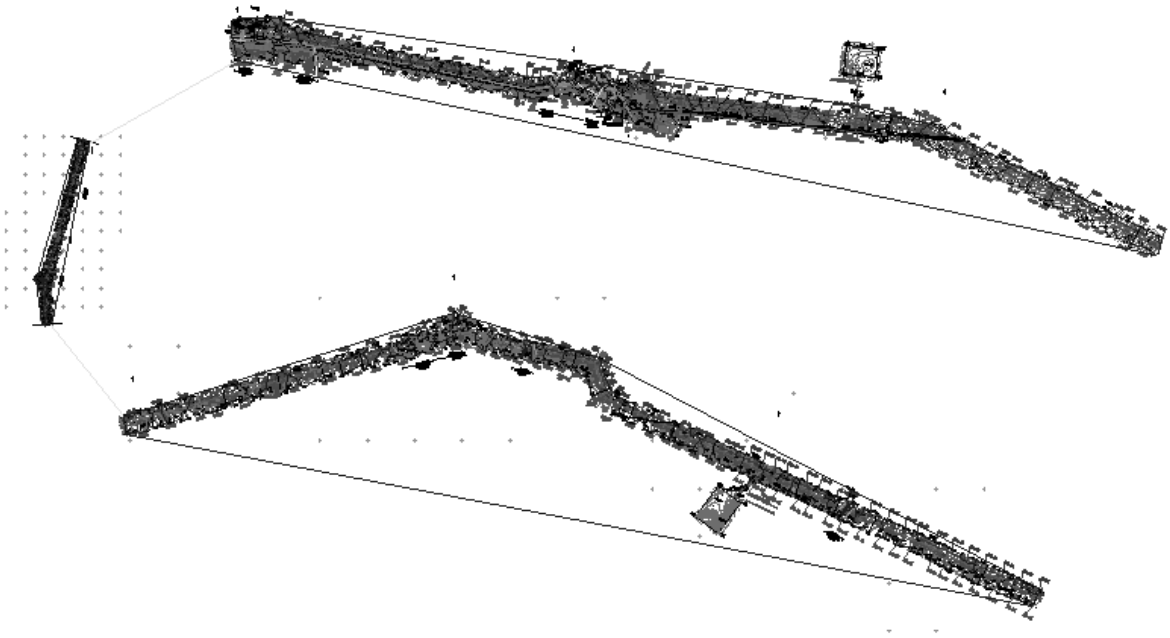

Рис. 2 Три кластера и отсутствие “выбросов”

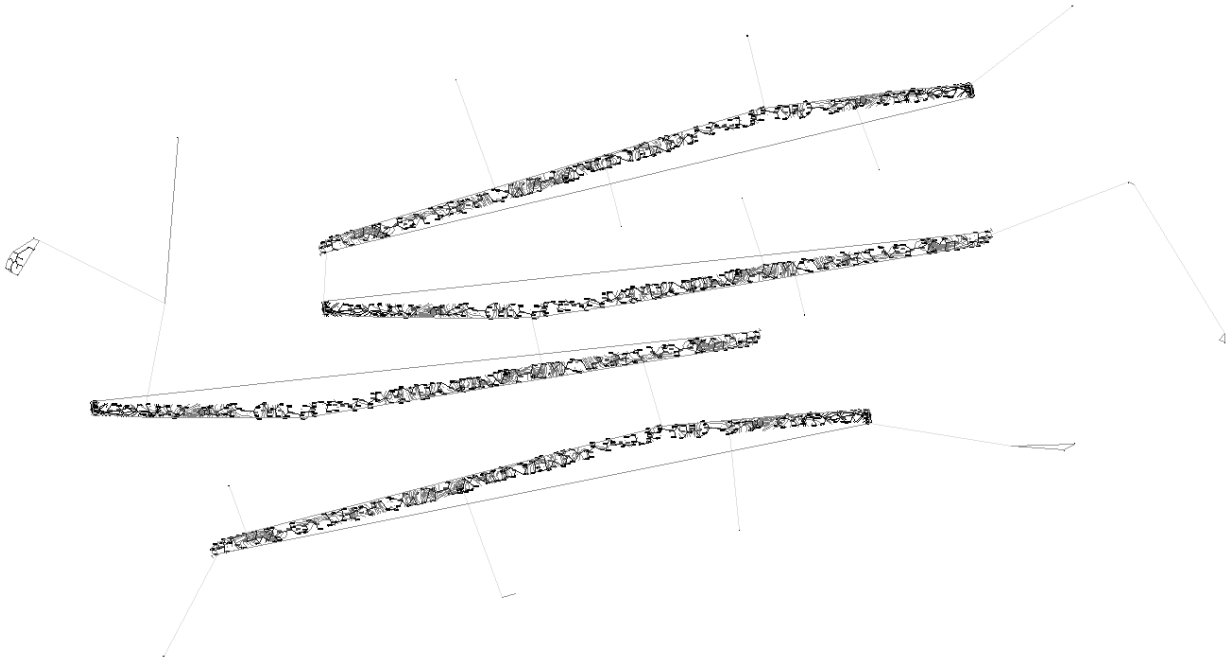

Рис. 3 Четыре кластера и наличие “выбросов”

Достоинством реализованного графового алгоритма также является *устойчивость при масштабировании*. Под устойчивостью к масштабированию здесь понимается адекватность кластеризации в случае различных масштабов коридоров проектирования трассы, расположенных в одном пространстве.

Алгоритм обладает цепочечным эффектом, когда независимо от формы кластера к нему присоединяются ближайшие к границе объекты.

К н**едостаткам** алгоритма можно отнести неадекватность кластеризации при наличии разреженного фона или “узких перемычек” между кластерами.

Литература

1. Дюран Б., Оделл П. Кластерный анализ
2. Воронцов К. В. Лекции по алгоритмам кластеризации и многомерного шкалирования. 21 декабря 2007 г.
3. Препарата Ф., Шеймос М. Вычислительная геометрия: Введение. – М.: Мир, 1989.
4. Дуда Р., Харт П. Распознавание образов и анализ сцен. - М.: Мир, 1976.
5. Горин Я.А. Модуль обработки чертежей САПР линейного транспорта: чертежи плана. – Сборник статей 14-ой международной научно-технической конференции "Наукоемкие технологии и интеллектуальные системы", Москва, 2012
6. Загоруйко Н. Г. Прикладные методы анализа данных и знаний. – Новосибирск: ИМ СО РАН, 1999.
7. Загоруйко Н. Г., Ёлкина В. Н., Лбов Г. С. Алгоритмы обнаружения эмпирических закономерностей. - Новосибирск: Наука, 1985.
8. Кулаичев А. П. Методы и средства комплексного анализа данных. – М: ИНФРА- М, 2006.
9. Лагутин М. Б. Наглядная математическая статистика. - М.: П-центр, 2003.
10. Мандель И. Д. Кластерный анализ. - М.: Финансы и Статистика, 1988.
11. Уиллиамс У. Т., Ланс Д. Н. Методы иерархической классификации // Статистические методы для ЭВМ / Под ред. М. Б. Малютов. - М.: Наука, 1986. - с. 269–301.
12. Jain A., Murty M., Flynn P. Data clustering: A review // ACM Computing Surveys. - 1999.- Vol. 31, no. 3.- Pp. 264–323. http://citeseer.ifi.unizh.ch/jain99data.html.
13. Lance G. N., Willams W. T.A general theory of classification sorting strategies. 1. hierarchical systems //Comp. J. 1967. - no. 9. - Pp. 373–380.