Б.А .Есипов, О.В. Москвичёв, Н.С. Складнев, А.О. Алёшинцев

РАЗРАБОТКА И ИССЛЕДОВАНИЕ АЛГОРИТМА КЛАСТЕРИЗАЦИИ С ПРОЕКЦИЕЙ ДЛЯ РЕШЕНИЯ ЗАДАЧ ОПТИМИЗАЦИИ ТРАНСПОРТНОЙ ИНФРАСТРУКТУРЫ

(Самарский национальный исследовательский университет имени С.П.Королева, Самарский государственный университет путей сообщения)

Основной идеей для повышения эффективности транспортных сетей является создание многоуровневой инфраструктуры с центрами обслуживания на каждом уровне. Так, например, для решения общей задачи выбора оптимальной двухуровневой сети транспортных объектов, реализующих технологию контейнерных поездов, предлагается на первом уровне все производства с контейнеропригодной продукцией привязать к ж/д контейнерным пунктам (КП), а на втором уровне создать контейнерные накопительно-распределительные центры (КНРЦ), к которым будут привязаны подмножества КП. В нашей работе предлагается для целей оптимального выбора мест расположения КП и КНРЦ применить универсальную методологию разбиения множества объектов с заданными свойствами на подмножества при заданных критериях разбиения и получения «центров» этих подмножеств, обладающих оптимальными свойствами. В качестве такой универсальной процедуры предлагается использовать математические методы кластеризации объектов, известные как *кластерный анализ* [1,2].

Действительно, геометрическая близость объектов от центра гарантирует минимизацию расстояний при перевозке, а учет «веса» каждого объекта, выражающего объем перерабатываемой объектом продукции оптимизирует общие затраты перевозок в тонно-километрах. Большое достоинство кластерного анализа и в том, что он позволяет производить разбиение объектов не по одному параметру, а по целому набору признаков. Анализ литературы по кластерному анализу и опыт использования стандартных программных средств кластерного анализа позволяет утверждать, что принципиально возможно решать поставленные задачи для практических задач большой размерности (для федеральных округов и всей страны в целом).

При подходе к решению практических задач оптимизации местоположения объектов КТС возникают новые научные задачи, развивающие сами методы кластерного анализа. Так, например, при применении алгоритмов кластеризации по известному методу k-means считается, что оптимальный «центр» может находиться в любой точке пространства параметров, определяющих объекты. Если параметры – это геометрические координаты центров производства, то «центр» лежит в любой точке плоскости. На практике следует рассмотреть случай, когда «центр» обязательно должен находиться в одной из заданных точек (например ж.д. линии или станции). Т.о. при решении задачи определения мест КП и КНРЦ приходится решать задачу кластеризации « с проекцией на функ-

цию», когда «центр» обязательно должен находиться на ж.д. магистрали или «с проекцией на точки».

В работе предлагается новый метод кластеризации с проекцией на множество точек «k-means pro» и исследуется возможность его применения в практических задачах проектирования транспортной инфраструктуры.

Входными данными является множество объектов кластеризации $X=\{\mathbf{x_1, \dots, x_n}\}$, их веса $V = \{v_1, \dots, v_n\}$ и допустимое множество проекций $Y=\{\mathbf{y_1, \dots, y_p}\}$. Каждый j –й объект и каждая допустимая точка-проекция задан в G-мерном пространстве R^G, т.е. $x_j = (x_{j1},\dots x_{jG})$ и $y_r = (y_{r1},\dots,y_{rG})$.

Единственным управляющим параметром является число кластеров k, на которые производится разбиение $S = \{S_1,...,S_k\}$ множества X. В результате, получается несмещенное разбиение $S^* = \{S_1^*,...,S_k^*\}$, центры которого являются оптимальным множеством проекций $C^* \subseteq Y$.

Введем обозначения : n – количество объектов кластеризации, p – количество точек допустимого множества проекций, i, i' – номер кластера, j – номер объекта, r – номер точки множества проекций, l – номер координаты точки, m – текущая итерация, G – размерность пространства, в котором выполняется кластеризация.

Расстояние между точками в G-мерном пространстве определяется по евклидовой метрике, где t_1 и t_2 – две любые точки пространства R^G

$$d(t_1, t_2) = \sqrt{\sum_{l=1}^{G} (t_{1l} - t_{2l})^2}.$$

1. Выберем начальное разбиение $S^0 = \{S_1^0,...,S_k^0\}$

$$S_i^0 = \left\{x_{i1}^0,...,x_{in}^0\right\}, \bigcup_{i=1}^{k} S_i^0 = X, S_i^0 \cap S_{i'}^0 = \varnothing,\ i \neq i'.$$

2. Пусть построено m-е разбиение $S^m = \{S_1^m,...,S_k^m\}$.

Вычислим набор векторов средних $E^m = \{e_1^m,...,e_k^m\}$ т.е. $e_i^m = (e_{i1}^m,...,e_{iG}^m)$, где

$$e_{il}^m = \frac{\sum_{j=1}^{n_i} v_j x_{jl}}{\sum_{j=1}^{n} v_j}.$$

n_i – количество точек i-го кластера

3. Определим множество проекций средних для текущего разбиения

$$C^m = \{y \in Y : \forall i, \quad d^*(y, e_i^m) = \min_{1 \le r \le p} d(y, e_i^m)\}$$

4. Построим минимальное дистанционное разбиение, порождаемое множеством C^m и возьмем его в качестве $S^{m+1} = (S_1^{m+1},...,S_k^{m+1})$, т.е. для 1-ого

$$S_1^{m+1} = \left\{x \in X : d(x, c_1^m) = \min_{1 \le i' \le k} d(x, c_{i'}^m)\right\}$$

и далее $$S_i^{m+1} = \left\{ x \in X \setminus \bigcup_{i'=1}^{i-1} S_{i'}^{m+1} : d(x, c_i^m) = \min_{1 \le i' \le k} d(x, c_{i'}^m) \right\}, 2 \le i \le k,$$

5. Если $S^{m+1} \neq S^m$, то переходим к п.2, заменив m на $m+1$, если $S^{m+1} = S^m$, то полагаем $S^m = S^*$, $C^m = C^*$ и заканчиваем работу алгоритма.

Разработана программа, реализующая вышеприведенный алгоритм с возможностью визуализации получаемых кластеров и вычисления разнообразных параметров. Для сравнения взят классический алгоритм k-means Мак-Куина из программного пакета WEKA [3,4].Сравнение производилось на тестовых выборках точек, распределенных равномерно на плоскости. На рис.1 показаны кластеры, полученные классическим k-means, а на рис.2 алгоритмом k-means pro. В данном примере ж/д магистраль представлена в виде «синусоиды».(+ отмечены центры кластеров, о – станции ж/д).

Разработанный алгоритм применен для решения задачи оптимального выбора мест расположения КП для заданных 900 производств и 137 ж/д станций Приволжского федерального округа (ПФО). Производства определялись географическими координатами и объемом контейнеропригодной продукции. Множество ж/д станций задано на сети 7 железных дорог, расположенных на территории ПФО. Результат при k = 20 представлен на рис.5

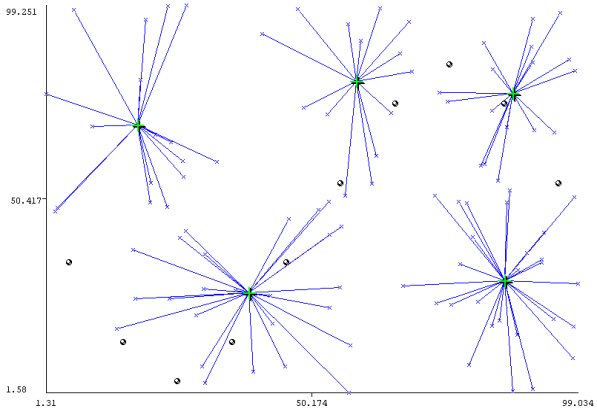

Рис. 1

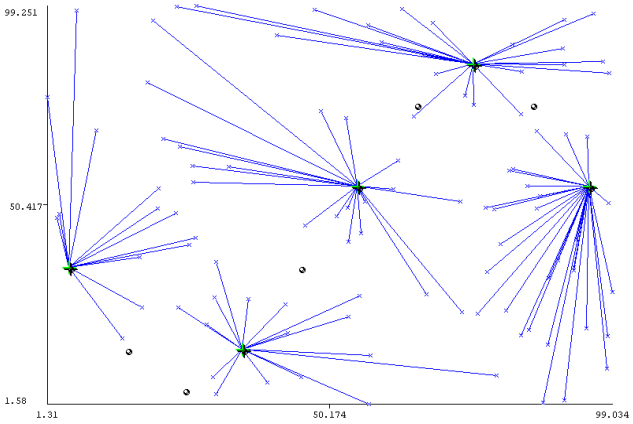

Рис. 2

Представляет интерес соотношение показателей качества кластеризации для обычного алгоритма и алгоритма «кластеризации с проекцией». В первом случае центры кластеров определяются исключительно из свойств расположения объектов (точек) и критерия оптимальности кластеризации *D*. Такую кластеризацию можно назвать свободной .

В нашем случае центры кластеров обязательно должны находиться на ж-д линии и это является ограничением для самого процесса кластеризации. Алгоритм каждый раз проектирует центры кластеров на ж-д линию. В результате получаем вариант кластеризации с проекцией и, очевидно, с другим значением критерия D_{np}.

Для классического алгоритма k – means

$$D=\sum_{i=1}^{k}\sum_{j=1}^{n_i} d(x_{ij},e_i),\ d(x_{ij},e_i)=\sqrt{(x_{ij1}-e_{i1})^2+(x_{ij2}-e_{i2})^2}$$

Для k-means pro

$$D_{np}=\sum_{i=1}^{k}\sum_{j=1}^{n_i} d(x_{ij},c_i^*),\ d(x_{ij},c_i^*)=\sqrt{(x_{ij1}-c_{i1}^*)^2+(x_{ij2}-c_{i2}^*)^2}$$

На рис. 3 изображена зависимости D и D_{np} от k для примера выше, а на рис.4 для ПФО.

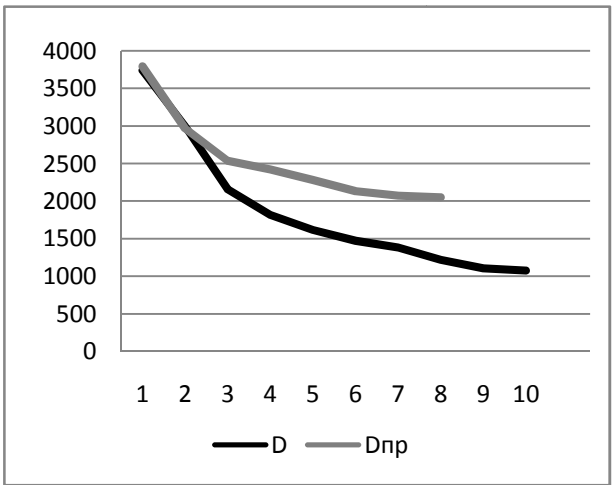

Рис. 3

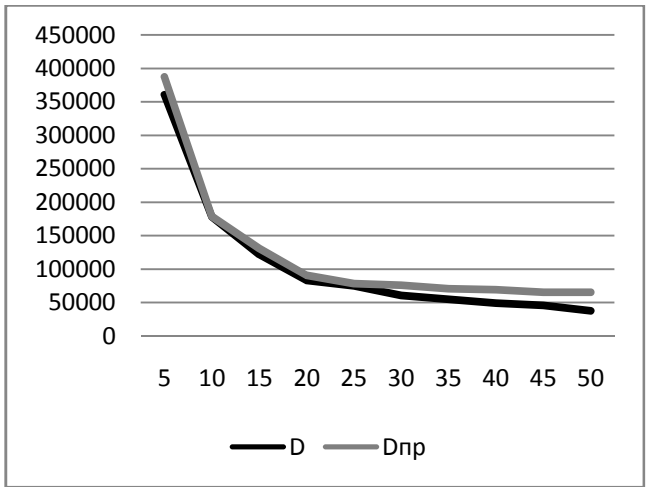

Рис. 4

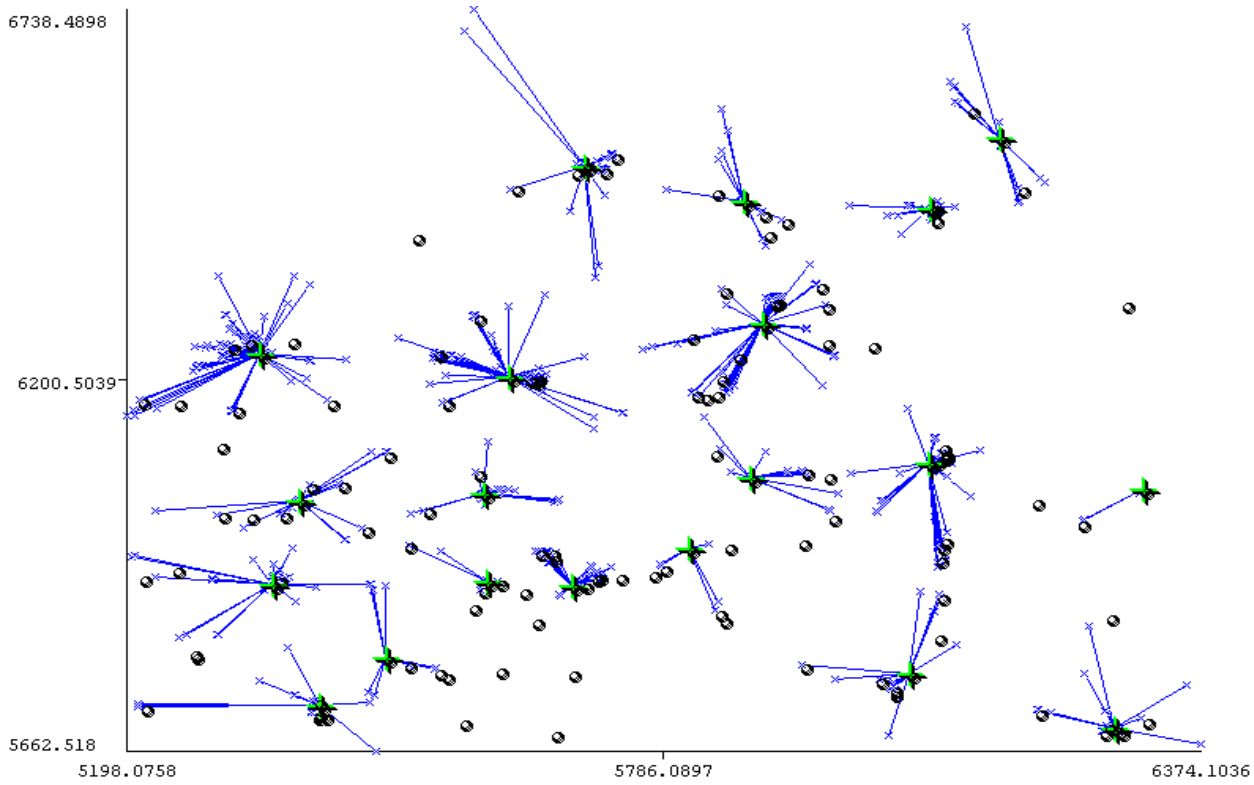

Рис. 5

Назовем дефектом проекции разницу критериальных величин качества свободной кластеризации и кластеризации «с проекцией» $\Delta = D_{пр} - D$. Зависимость Δ / D от k для производств ПФО представлена на рис.6 .

На рис.7 изображены кривые затрат на проект $E = sD_{np} + ck$,где s – тариф перевозки, c – удельная стоимость одного КП на единицу объема перевозки. E1(c=5000), E2(c=10000), E3(c=20000). Из графика видно, что, например, для c = 20000 оптимальным решением будет создание 10 КП.

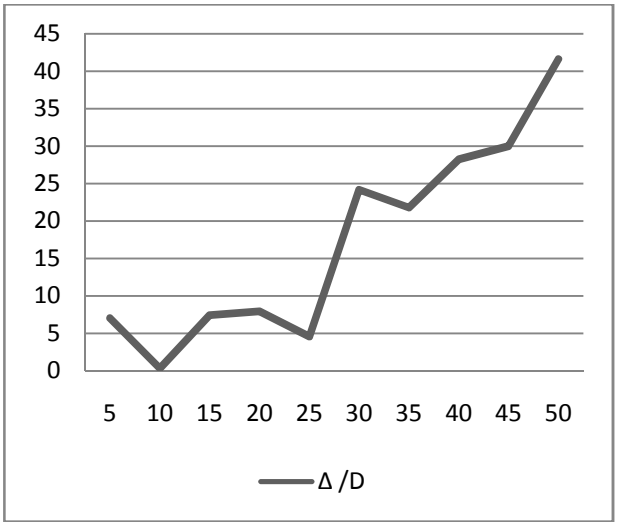

Рис. 6

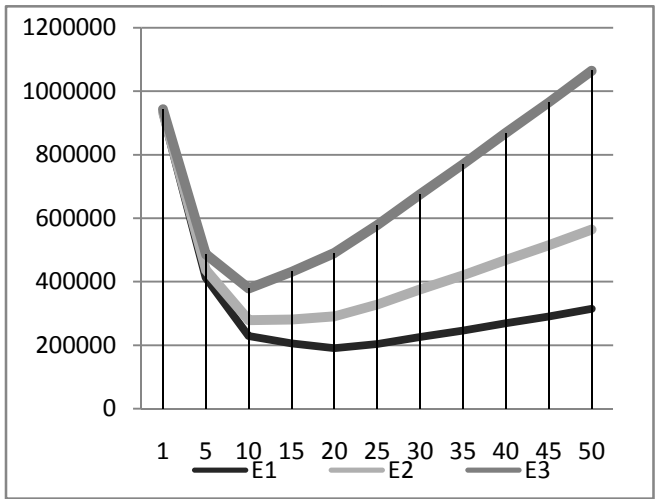

Рис.7

Литература

1.Прикладная статистика. Классификация и снижение размерности [Текст] под ред С.А.Айвазяна. –М: Финансы и статистика,1989. – 607с.

2. Кластерный анализ [Текст] Мандель И.Д.. – М:Финансы и статистика, 1988.- 177с

3. Witten, I.H. and Frank, E. (2005) Data Mining: Practical machine learning tools and techniques. 2nd edition Morgan Kaufmann, San Francisco.

4. WekaWiki – http://weka.wikispaces.com/

А.М. Зиятдинов, Р.М. Зиятдинова, А.В. Клепиков

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ НА ТРАНСПОРТЕ: АНАЛИЗАТОРЫ ЭКСПЛУАТАЦИОННОЙ РАБОТЫ В ЖЕЛЕЗНОДОРОЖНОЙ ОТРАСЛИ

(Филиал ФГБОУ ВО «Уфимский государственный нефтяной технический университет» в г. Октябрьском, Российская Федерация)

Федеральный закон от 27.12.2002 г. №184-ФЗ (ред. от 05.04.2016) "О техническом регулировании" включает ряд положений касательно транспортных условий и устанавливает требования в части условий перевозочного процесса. На сегодняшний день отечественный транспортный рынок представлен такими участниками и владельцами процесса, как трубопроводный транспорт, железнодорожный, автомобильный, воздушный и водный. В рамках нашего повест-