

УДК 681.5.017

ИСПОЛЬЗОВАНИЕ ПРИНЦИПОВ КЛАСТЕРИЗАЦИИ ДЛЯ ВЫБОРА РЕМОНТНЫХ УЧАСТКОВ МАГИСТРАЛЬНЫХ НЕФТЕПРОВОДОВ ПО ДАННЫМ ДИАГНОСТИКИ

Богданов Р. М.

Институт механики УНЦ РАН, Уфа

Аннотация. На основе классификации образов с помощью функций расстояния и принципа кластеризации решается задача определения ремонтных участков магистрального нефтепровода. Критерии, характеризующие кластер, определяются некоторыми задаваемыми величинами, на основе сравнения с которыми определяется принадлежность дефекта к данному кластеру, обеспечиваются процедуры перераспределения дефектов по кластерным зонам, а также производится изменение параметров самих кластерных зон. Приведенные расчеты демонстрируют диапазон изменения плотности дефектов в зависимости от участков нефтепроводов с одной стороны и какими универсальными возможностями, по компоновке линейных объектов с произвольной плотностью, обладает кластерный анализ.

В нефтепроводном транспорте существует проблема по составлению рационального плана капитального ремонта магистрального нефтепровода [1] по данным внутритрубной диагностики. Под ним подразумевается организация работ по ремонту наибольшего числа потенциально опасных дефектов (ПОД) при фиксированной сумме выделенных средств [1]. В постановке задачи существенным аспектом является определение участков трубопровода с высокой степенью концентрации дефектов. В данной работе задача определения ремонтных участков решалась с использованием алгоритма, построенного на основе клас-

сификации образов с помощью функций расстояния и принципа кластеризации. Использование альтернативной классификации с помощью решающих функций приводит лишь к частичному решению поставленной задачи, так как она основана на построении функций, разделяющих образы на классы. В рамках этой методики невозможно определение параметров, характеризующих классы или кластеры.

Выбор функций расстояния в качестве методологической базы построения алгоритма является естественным следствием того факта, что наиболее очевидный способ введения меры сходства для образов любой природы — это определение их близости в пространстве признаков. В качестве образов далее будем рассматривать дефекты, а под кластером будем подразумевать группу дефектов, образующих в пространстве — по длине трубопровода — компактную область, и удовлетворяющую набору капитального ремонта с учетом этого показателя. Основная задача оптимизации в данном контексте состоит в организации работ по ремонту критериев, к описанию которых мы приступаем.

Критерии, характеризующие кластер, определяются некоторыми задаваемыми величинами, на основе сравнения с которыми определяется принадлежность дефекта к данному кластеру, обеспечиваются процедуры перераспределения дефектов по кластерным зонам, а также производится изменение параметров самих кластерных зон.

К указанным критериям относятся:

- число дефектов в кластере — θ_N ;
- плотность дефектов (среднее число дефектов в кластере на единицу длины) — θ_P ;
- величина *max* среднеквадратичного отклонения — θ_S ;
- величина компактности — θ_C .

Основными исходными параметрами являются координаты дефектов — X_i , полученные в ходе внутритрубной диагностики, число дефектов N на рассматриваемом участке магистрального нефтепровода и его длина AL .

Основными вычисляемыми параметрами являются координаты

наты центров кластеров — Z_c и расстояния между дефектами и центрами кластеров

$$D_{i,j} = |X_i - Z_{cj}|,$$

где $i \in [1, N]$; $j \in [1, N_c]$; N_c — число кластеров на данном шаге процедуры кластеризации.

Построение алгоритма дефектов по длине трубопровода осуществим на основе известного алгоритма по выявлению кластеров ISODATA. Название алгоритма является аббревиатурой Iterative Self-Organizing Data Analysis Techniques (итеративный самоорганизующийся метод анализа данных). Он достаточно нагляден и эффективен; обладает свойством разумной достаточности, сочетая достоинства простоты идей, заложенных в него, и широту возможностей для его применения. Следует отметить, что в данной работе предложена модификация алгоритма ISODATA с учетом введения дополнительного критерия образования кластеров по плотности дефектов.

Выявление алгоритма начинается с задания исходных центров кластеров: $Z_{c1}, Z_{c2}, Z_{c3}, \dots, Z_{cN_c}$.

Алгоритм имеет блоки, описывающие следующие процедуры:

– включение очередного дефекта в тот или иной кластер в соответствии с критерием наименьшего расстояния от центра кластера;

– ликвидация кластеров, не удовлетворяющих критериям кластерообразования;

– образование кластеров в зонах группирования дефектов, в которых удовлетворяются критерии кластерообразования;

– изменение параметров, характеризующих кластеры, путем исключения или вовлечения в них «пограничных» дефектов.

Модифицированный алгоритм ISODATA — ISODAT-Mod имеет следующую структуру и последовательность действий.

1. Ввод исходной информации.

1.1. Введение массива данных по результатам внутритрубной диагностики $\{X_1, X_2, \dots, X_N\}$.

1.2. Введение характеристических признаков, определяющих процесс кластеризации $\theta_N, \theta_P, \theta_S, \theta_C$.

1.3. Введение параметров, характеризующих процесс кластеризации: K_c — необходимое число кластеров; N_c — исходное число кластеров; L_c — максимальное число пар центров кластеров, подлежащих объединению за один шаг итерации; I — допустимое число циклов итерации.

2. Разбиение участка магистрального нефтепровода на начальные кластеры:

вариант 1: равномерное разбиение всей длины участка на N_c исходных кластера.

вариант 2: разбиение участка на исходные кластеры на основе построения гистограммы плотности дефектов по длине участка нефтепровода.

Определение границ кластеров Z_{ej}, Z_{rj} . Выделение дефектов, с критическим расстоянием друг от друга.

3. Распределение дефектов по кластерам, соответствующим исходным центрам этих кластеров.

Принадлежность дефекта X_i к кластеру C_j определяется по степени близости данного дефекта к центру кластера, то есть:

$$X \in C_j \quad \forall |X_i - Z_{cj}| < |X_i - Z_{ck}|, \quad j, k = 1, 2, \dots, N_c, \quad j \neq k$$

4. Определение состоятельных кластеров.

4.1. Вычисление средней плотности дефектов P_j в каждом кластере:

$$P_j = N_j / al_j,$$

где al_j — длина j -го кластера.

4.2. Ликвидация кластеров, в состав которых входит менее θ_N дефектов или их плотность менее θ_P :

$$C_j : N_j < Q_N \cup P_j < Q_P \Rightarrow C_{j+1} \rightarrow C_j, \quad N_c \rightarrow N_c - N_e,$$

где N_e — число устраняемых кластеров за один шаг итерации.

5. Вычисление среднего отклонения в кластерах и обобщенного среднего отклонения.

5.1. Корректировка центров кластеров

$$C_j : Z_{cj} = \frac{1}{N_j} \sum_{X_i \in C_j} X_i, \quad j = 1, 2, \dots, N_c,$$

$$i \in [i_{left}; i_{right}], \quad X_i \in [Z_{lj}; Z_{rj}].$$

5.2. Вычисление среднего расстояния DA_j между дефектами j -го кластера и его центром Z_{cj}

$$DA_j = \frac{1}{N_j} \sum_{X_i \in C_j} |X_i - Z_{cj}|, \quad j = 1, 2, \dots, N_c, \quad X_i \in [Z_{lj}; Z_{rj}].$$

5.3. Вычисление обобщенного среднего расстояния DAC по всем кластерам:

$$DAC = \frac{1}{N} \sum_{j=1}^{N_c} N_j DA_j.$$

6. Вычисление среднего квадратичного отклонения дефектов от центра соответствующего кластера:

$$\sigma_j = \sqrt{\frac{1}{N_j} \sum_{X_i \in C_j} (X_i - Z_{cj})^2}.$$

7. Расщепление кластеров.

7.1. Условия расщепления кластера:

$$C_j : a) \sigma_j > \theta_S \cap DA_j > DAC \cap N_j > 2(\theta_N + 1);$$

$$\text{или б) } \sigma_j > \theta_S \cap N_c \leq k/2.$$

7.2. Результаты расщепления кластера.

При расщеплении кластера количество кластеров увеличивается на единицу

$$N_c \rightarrow N_c + 1, \quad Z_{cj}^+ = Z_{cj} + k\sigma_j, \quad Z_{cj}^- = Z_{cj} - k\sigma_j,$$

где $0 < k \leq 1$.

Пересчет центров кластеров:

$$Z_{cj}^- \rightarrow Z_{cj}, \quad Z_{cj}^+ = Z_{cj+1}, \quad Z_{cj+1} \rightarrow Z_{cj+2}.$$

7.3. Возврат к пункту 3 если данная итерация не является последней.

8. Объединение кластеров.

8.1. Вычисление расстояний D_{ij} между всеми парами центров кластеров

$$D_{jk} = |Z_{cj} - Z_{ck}|, \quad j = 1, 2, \dots, N_c - 1, \quad k = j + 1, \dots, N_c.$$

8.2. Сравнение межцентровых расстояний кластеров с признаком компактности θ_C .

$$C_j : \forall D_{jk} < \theta_C \Rightarrow D_{jk}^1 < D_{jk}^2 < \dots < D_{jk}^L$$

(ранжирование кластеров, для которых выполняется условие объединения).

8.3. Слияние кластеров. Кластеры с центрами Z_{cje} и Z_{cke} ($j = 1, 2, \dots, L$) объединяются, причем центр нового кластера определяется по формуле:

$$Z_{c*} = \frac{1}{N_{jl} + N_{kl}} (N_{jl} Z_{cjl} + N_{kl} Z_{ckl}).$$

Центры кластеров Z_{cje} и Z_{cke} ликвидируются, а общее число кластеров сокращается на единицу:

$$Z_{c*} \rightarrow Z_{cj}, \quad Z_{cj+2} \rightarrow Z_{cj+1}, \dots, \quad N_c \rightarrow N_c - 1.$$

На этом, если текущий цикл итераций — последний, выполнение алгоритма прекращается, иначе следует вернуться к следующему шагу итерации начиная с пункта 3. Алгоритм прекращает вычисления, если в очередном цикле итераций параметры процесса остались неизменными.

Представленный алгоритм не является свободным от некоторых эвристических констант — признаков кластеризации θ_N ,

θ_P , θ_S , θ_C . Их значения невозможно подобрать, исходя из рациональных соображений, теоретически. Данная ситуация весьма характерна для задач, связанных с принятием решений. Это те самые величины, которые олицетворяют такие понятия как опыт и интуиция. Главное, чтобы результат, в нашем случае — разбиение дефектов на группы, отражал реальную ситуацию и позволял бы принимать обоснованное решение о местах дислокации ремонтных бригад для проведения оптимального ремонта нефтепровода.

Таким образом, указанные константы целесообразно определять с помощью решения тестовых задач с применением алгоритма ISODAT-Mod, в которых эти константы принимают различные значения. При этом, в каждом случае степень адекватности результата может быть предметом экспертных оценок.

На рис. 1, 2 и в табл. 1, 2 приведены результаты разбиения нефтепроводов на кластеры (линейные объекты) для участков нефтепроводов Самара–Лисичанск и Рязань–Москва. Разбиение на начальные кластеры проводилось нами по второму варианту пункта 2 последовательности действий алгоритма ISODAT-Mod, т.е. на основе построения гистограммы плотности дефектов по длине. С этой целью трасса рассматриваемого участка нефтепровода разбивалась нами на участки длиной порядка 50 м. Для нефтепровода Рязань–Москва число разбиений составило 4000, а для участка Бородаевка–левый берег р. Волги нефтепровода Самара–Лисичанск — 2000 разбиений. Результаты построения гистограммы представлены на рис. 1, 2. Далее проводилась процедура кластеризации при следующих параметрах, для нефтепровода Рязань–Москва: $\theta_N = 0$, $\theta_P = 60$ шт/км, $\theta_S = 10$ м, для нефтепровода Самара–Лисичанск: $\theta_N = 19$ шт, $\theta_P = 300$ шт/км, $\theta_S = 10$ м. Кластеризация проводилась только для ПОД нефтепровода Рязань–Москва. Результаты разбиения на линейные объекты для нефтепровода Рязань–Москва представлены в табл. 1. Необходимо отметить, что нефтепровод Рязань–Москва является одним из самых благополучных с точки зрения плотности дефектов. Примером одного из самых

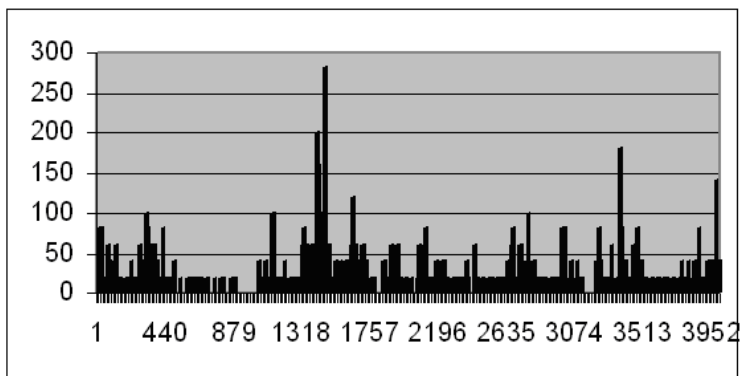


Рис. 1. Гистограмма распределения плотности дефектов (Q_p) в зависимости от числа разбиений участка нефтепровода Рязань–Москва

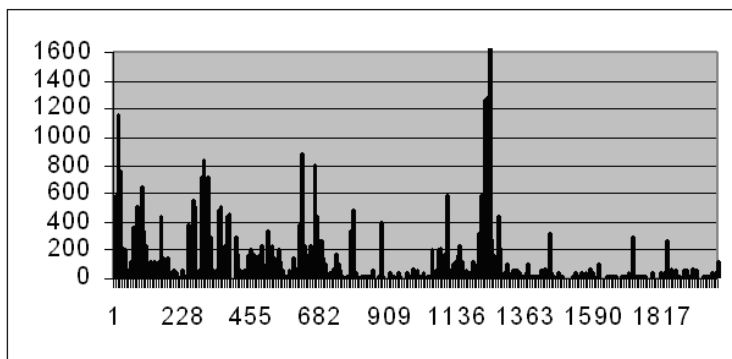


Рис. 2. Гистограмма распределения плотности дефектов (Q_p) в зависимости от числа разбиений участка нефтепровода Самара–Лисичинск (Бородаевка–лев. берег р. Волги)

неблагополучных участков нефтепроводов по АК «Транснефть» является участок Бородаевка–левый берег р. Волги нефтепровода Самара–Лисичанск. Результаты кластеризации по этому нефтепроводу представлены на табл. 2. Необходимо отметить, что для наглядности параметры для расчета этого нефтепровода были подобраны таким образом, чтобы число линейных объектов, полученных при расчете нефтепровода Самара–Лисичанск, совпадало с числом линейных объектов на нефтепроводе Рязань–Москва. Приведенные расчеты демонстрируют (см.табл. 1, 2), диапазон изменения плотности дефектов в зависимости от участков нефтепроводов с одной стороны и какими универсальными возможностями, по компоновке линейных объектов с произвольной плотностью, обладает кластерный анализ.

Автор выражает благодарность С. Ф. Урманчееву за плодотворное обсуждение результатов исследования.

Таблица 1. Результат кластеризации дефектов на магистральном нефтепроводе Рязань–Москва

№ кластера	Плотность (шт/км)	Координаты границ		Длина (м)	Кол-во (шт.)
		левая (м)	правая (м)		
1	60	49	147	98	6
2	60	344	393	49	3
3	80	688	737	49	4
4	60	1130	1179	49	3
5	120	3585	3634	49	6
6	80	15274	15372	98	8
7	70	65368	65466	98	7
8	170	69444	69543	99	17
9	65	70083	70255	172	11
10	125	70525	70770	245	31
11	90	80838	80937	99	9
12	80	147287	147385	98	8
13	65	164820	164992	172	11
14	70	194975	195147	172	12

Таблица 2. Результат кластеризации дефектов на магистральном нефтепроводе Самара–Лисичанск (Бородаевка–лев. берег р. Волги)

№ кластера	Плотность (шт./км)	Координаты границ		Длина (м)	Кол-во (шт.)
		левая (м)	правая (м)		
1	363	200	451	251	91
2	458	752	1152	400	184
3	380	1152	1202	50	19
4	760	3307	3357	50	39
5	313	3808	3908	100	31
6	320	4208	4309	101	32
7	652	4559	4885	326	212
8	337	15681	15857	176	59
9	307	18888	18988	100	31
10	570	30912	31237	325	186
11	620	33417	33517	100	62
12	440	60671	60997	326	143
13	600	61623	62625	1002	601
14	320	72244	72344	100	32

Список литературы

- [1] Лисин Ю. В., Бахтизин Р. Н., Ахатов И. Ш., Богданов Р. М., Хисматуллин Д. Б. Оптимизация планирования капитального ремонта магистральных нефтепроводов // Газовая промышленность. Москва. 1999. № 7. С. 51–53.
- [2] Гонсалес Р., Ту Дж. Принципы распознавания образов. М: «Мир». 1978. 412 с.