Янчуковский В. Н., электроник,
Сосинская С. С., канд. техн. наук, доц.
Национальный исследовательский Иркутский государственный технический университет
Козловский А. С., аспирант
Российский государственный педагогический университет им. А.И. Герцена
Челибанов В. П., канд. хим. наук, ген. директор.
Приборостроительное предприятие «ОПТЭК»

ДВУХУРОВНЕВЫЙ КЛАСТЕРНЫЙ АНАЛИЗ В СРЕДЕ MATLAB С ПРИМЕНЕНИЕМ ПАРАЛЛЕЛЬНЫХ ВЫЧИСЛЕНИЙ

V.Yanchukovsky@gmail.com

Предлагается технология организации двухуровневого кластерного анализа, основанная на применении методов субтрактивной кластеризации и алгоритма К-средних. В качестве исходных данных использованы результаты трехлетних измерений концентрации газов (SO_2, CO) в воздушной атмосфере Санкт-Петербурга. Рассматриваемый набор данных характеризуется большим объемом – порядка сотни тысяч значений (измерения проводились в среднем раз в 20 минут) и крайне нечеткой границей между объектами. По этой причине предлагается использовать параллельные вычисления для сокращения времени обработки.

Ключевые слова: *Кластерный анализ, параллельные вычисления, вычислительный эксперимент, нечеткий алгоритм.*

Введение. Кластерный анализ - метод разбиения заданной выборки объектов (ситуаций) на подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались. Общий вопрос, задаваемый исследователями во многих областях, состоит в том, как организовать наблюдаемые данные в наглядные структуры, то есть визуализировать данные. Всякий раз, когда необходимо классифицировать «горы» информации в пригодные для дальнейшей обработки группы, кластерный анализ оказывается весьма полезным и эффективным, особенно совместно с распараллеливанием вычислений. Однако существуют такие наборы данных, классифицировать которые достаточно сложно, ввиду их специфики. Для повышения точности кластерного анализа, а также, для упрощения интерпретации результатов предлагается применить двухуровневый кластерный анализ.

Постановка задачи. Для проведения кластерного анализа были взяты результаты трехлетнего мониторинга содержания газов SO_2 и CO в воздушной атмосфере центрального района Санкт-Петербурга. Концентрации компонентов воздушной среды были измерены аттестованным на Федеральный Знак качества программно-аппаратным аналитическим комплексом "Скат" производства ЗАО «ОПТЭК», установленным в техническом помещении Некрополя 18 века Музея городской скульптуры. Измерения проводили с интервалом в 20 минут. В результате были получены порядка сотни тысяч значений, на основании которых рассчитывали среднемесячные концентрации газов, которые и являлись исходной матрицей наблюдений для алгоритмов двухуровневой кластеризации. Анализируемый набор данных характеризовался крайне нечеткой границей между объектами.

Методы кластеризации. Существует более ста методов кластеризации. В большинстве алгоритмов количество кластеров является одним из входных параметров, однако в данном случае этот параметр сложно назначить, исходя из внешних соображений. После анализа большого числа алгоритмов было предложено проведение двухуровневой кластеризации. Суть технологии заключается в том, что методом субтрактивной кластеризации определяются центры кластеров и их число, Основу алгоритма составляют идеи горного метода кластерного анализа, который был предложен Рональдом Ягером (RonaldYager) и Димитаром Филевым (DimitarFilev). Особенностью метода является отсутствие необходимости задания количества кластеров до начала работы алгоритма.

Задача нахождения центров кластеров ставится следующим образом.

Дано множество $X = (X_1, X_2, ..., X_n)$ объектов, подлежащих кластеризации, где n – количество объектов. Каждый объект $X_K = (x_{K1}, x_{K1}, ..., x_{KP})$ представляет собой точку в р-мерном пространстве признаков $(K = \overline{1,n})$. Необходимо найти центры кластеров, то есть координаты центров скопления объектов, заданных множеством X.

Идея метода заключается в следующем. Объекты рассматриваются как потенциальные центры кластеров. Для каждого объекта рассчитывается значение так называемого потенциала, характеризующего плотность расположения других объектов в его окрестности. Чем гуще соседние объекты расположены к данному объекту, тем больше значение его потенциала. Значение потенциала для объекта $X_K = (x_{K1}, x_{K1}, ..., x_{KP})$ рассчитывается по формуле

$$P(X_K) = \sum_{i=1,n} \exp\left(-4 \cdot \sum_{j=1,P} \omega_j \cdot \left(X_{K_j} - X_{i_j} \right)^2 \right), \quad (1)$$

где ω_j - вес j-й координаты. В случае, когда объект задан двумя признаками, графическое изображение распределения потенциалов будет представлять собой поверхность, напоминающую горный рельеф. Отсюда и название - горный метод. В качестве центров кластеров выбирают координаты “горных” вершин. Для этого, центром первого кластера назначают объект с наибольшим потенциалом. Затем центр кластера, а также близко расположенные к нему объекты исключают из дальнейшего рассмотрения, то есть из “горного массива” вычленяют наивысшую “гору”. Значения потенциалов оставшихся объектов пересчитывают, и вновь в качестве центра кластера выбирают объект с максимальным значением потенциала. Итерационная процедура выбора центров кластеров продолжается до тех пор, пока не будут исключены все объекты[3].

К плюсам данного метода следует отнести то, что количество кластеров определяется во время выполнения алгоритма, а также достаточную простоту алгоритма и в некотором роде универсальность.

Недостатками являются достаточно низкая точность алгоритма и не очень наглядное представление полученных результатов[2].

Полученное в результате работы алгоритма количество кластеров используется как входной параметр для метода *K*-средних[1], который является одним из наиболее популярных методов кластеризации. Алгоритм представляет собой модификацию *EM*-алгоритма для разделения смеси гауссиан. Он разбивает множество элементов векторного пространства на заранее известное число кластеров k. Действие алгоритма таково, что он стремится минимизировать дисперсию на точках каждого кластера:

$$V = \sum_{i=1}^{K} \sum_{x_j \in S_i} \left(x_j - \mu_i \right)^2 \quad (2)$$

где k - число кластеров, S_i - полученные кластеры, $i = 1, 2, ..., K$ и μ_i - центры масс векторов $x_j \in S_i$.

Основная идея заключается в том, что на каждой итерации перевычисляется центр масс для каждого кластера, полученного на предыдущем шаге, затем векторы разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе по выбранной метрике. Алгоритм завершается, когда на какой-то итерации не происходит изменения кластеров.

Плюсами метода являются наглядность при представлении кластеров и высокая точность.

К минусам следует отнести то, что необходимо заранее знать количество кластеров. Помимо этого, алгоритм очень чувствителен к выбору начальных центров кластеров и не справляется с задачей, когда объект принадлежит к разным кластерам в равной степени или не принадлежит ни одному[2].

В качестве среды для проведения исследования был выбран пакет Matlab.

Эксперимент проводился на одном ПК следующей конфигурации:

1. Процессор - четырехядерный, с частотой 2,3 Ггц.

2. Оперативная память объема 8 Гб.

При выборе способа распараллеливания в среде Matlab был выбран метод запуска параллельной программы с явным заданием пула – matlabpool, с объявлением количества ядер процессора, то есть выделение необходимого числа процессов на локальной машине или на кластере. В качестве режима распараллеливания был выбран режим –parfor. В основе режима parfor лежит тот же принцип, что и в цикле for: Matlab выполняет последовательность команд в теле цикла. В этом режиме программный код распределяется между клиентским процессом (client) и рабочими процессами (worker). Основная часть вычислений производится на workers, затем результаты вычислений отправляются на client и объединяются воедино. Режим parfor больше подходит для случаев, когда необходимо большое количество итераций для решения простой задачи[4].Оператор parfor очень удобен, когда необходимо использовать все ядра локального компьютера. Разовое выполнение кода в теле цикла parfor представляет собой независимую итерацию.

Такой выбор обусловлен тем, что необходимо было сократить объем данных. В параллельном режиме все данные каждого месяца использовались в отдельном процессе для нахождения среднеарифметических значений по концентрации каждого газа. Скорость вычисления среднего значения измерялась с помощью ко-

манд tic…toc, как в обычном, так и в параллельном режиме.

Результаты эксперимента. Эксперимент, проведенный с применением двухуровневого кластерного анализа, показал, что скорость анализа существенно увеличивается (табл. 1).

Данные, представленные в таблице 1, свидетельствуют о том, что в результате применения параллельного режима скорость решения задачи возросла примерно в 3 раза по отдельным годам, и в 3.5 раз при расчете по массиву данных за все три года.

Таблица 1

Скорость выполнения в обычном и параллельном режимах

Год	Обычный режим, секунды	Параллельный режим, секунды
Первый	14.114188	4.7
Второй	19.962332	6.65
Третий	15.380173	5.1
Все	66.5953	19

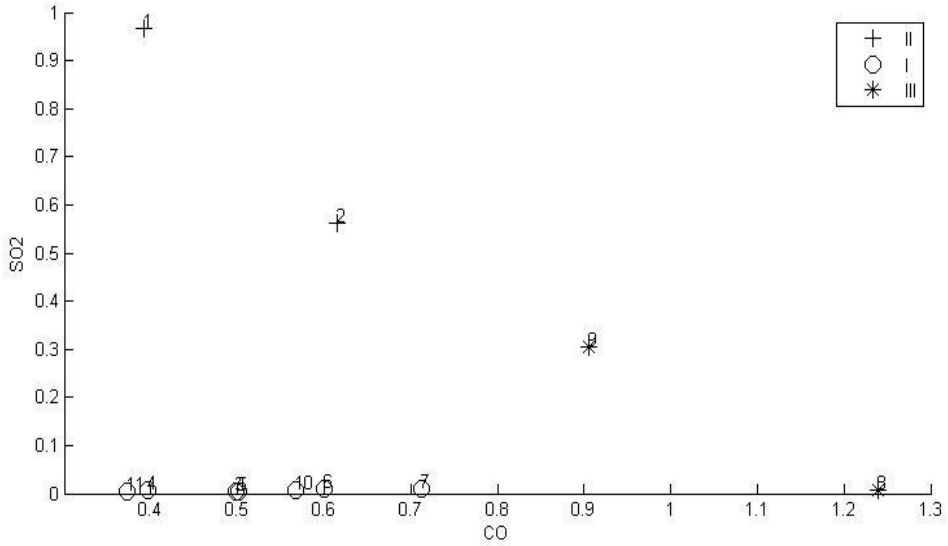

Рис.1. Распределение среднемесячных концентраций газов за первый год измерений

Алгоритм субтрактивной кластеризации разделил данные по каждому году наблюдения на три кластера (I-III, рис.1-3, табл. 2). На рис.1 изображено распределение по кластерам среднемесячных концентраций за первый год наблюдений как результат алгоритма k-среднего в виде попарных зависимостей в двумерном пространстве. На графике отсутствует 12-ый месяц, поскольку за первый год по этому месяцу нет измерений.

Разбиения концентраций газов в остальные годы проводились аналогичным способом (рис. 2, 3).

Таблица 2

Характеристика кластеров по среднемесячным концентрациям газов за каждый год измерений

№	Содержание (мг/м^3)		Месяцы
	CO	**SO_2**	
1 год измерений (06.2006 - 05.2007)			
1	0,373-0,713	0,004-0,010	3-7,10,11
2	0,392-0,615	0,563-0,968	1,2
3	0,905-1,240	0,008-0,305	8,9
2 год измерений (06.2007 - 05.2008)			
1	0,354-0,412	0,003-0,009	3,5,7,9-11
2	0,460-0,546	0,006-0,010	4,8
3	0,296-0,313	0,006-0,007	1,2,12
3 год измерений (06.2008 - 05.2009)			
1	0,182-0,240	0,002-0,004	4,5,10-12
2	0,278-0,298	0,004-0,006	1,3,6
3	0,342-0,396	0,002-0,005	2,7,8,9

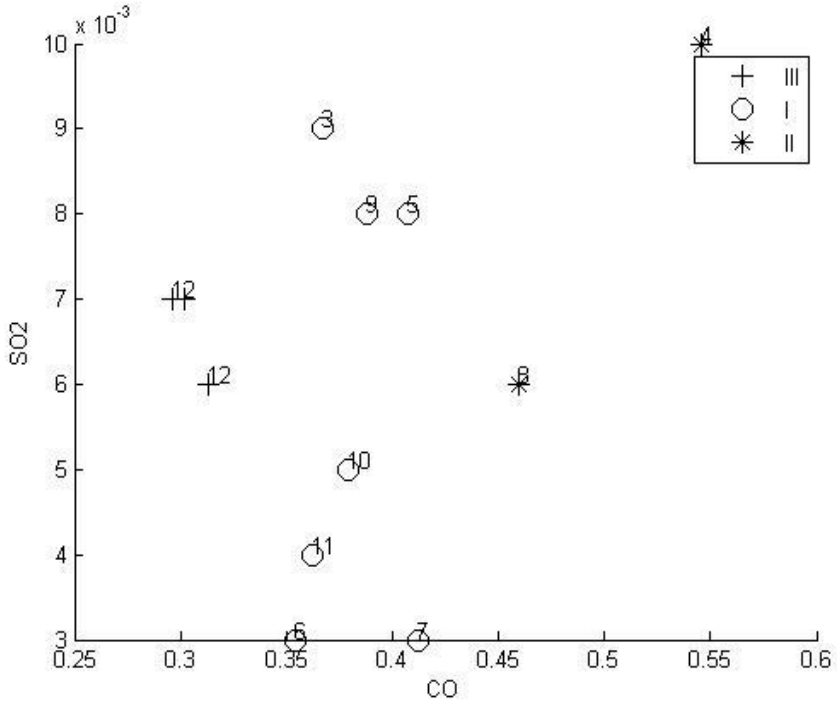

Рис. 2. Распределение среднемесячных концентраций газов за второй год измерений

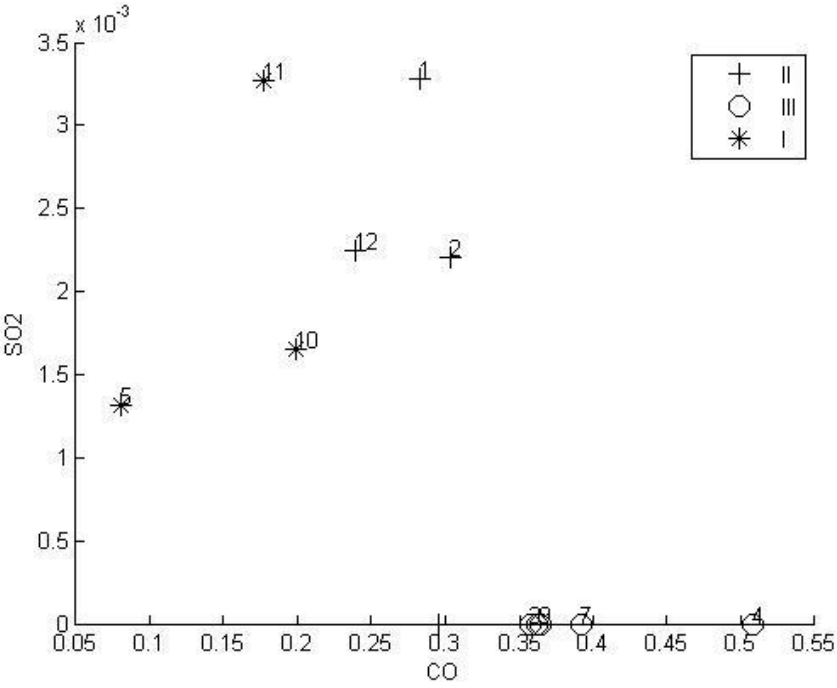

Рис. 3. Распределение среднемесячных концентраций газов за третий год измерений

Также был проведен анализ среднемесячных концентраций за все три года измерений. Для обеспечения большей точности, была использована выборка из 35 месячных значений, переданных на вход алгоритма кластеризации по порядку, начиная с первого года. Результаты представлены на рис.4 и продублированы в таблице 3.

Таблица 3

Характеристика кластеров по среднемесячным концентрациям газов за три года измерений

№ Класса	Содержание		№ Месяца
	CO	SO_2	
I	0,713-1,240	0,008-0,305	7,8,9
II	0,182-0,600	0,002-0,010	3,4,5,6,10,11,12,13,14,15,16,17,18,19,20,21, 22,23,24,25,26,27,28,29,30,31,32,33,34,35
III	0,392-0,615	0,563-0,968	1,2

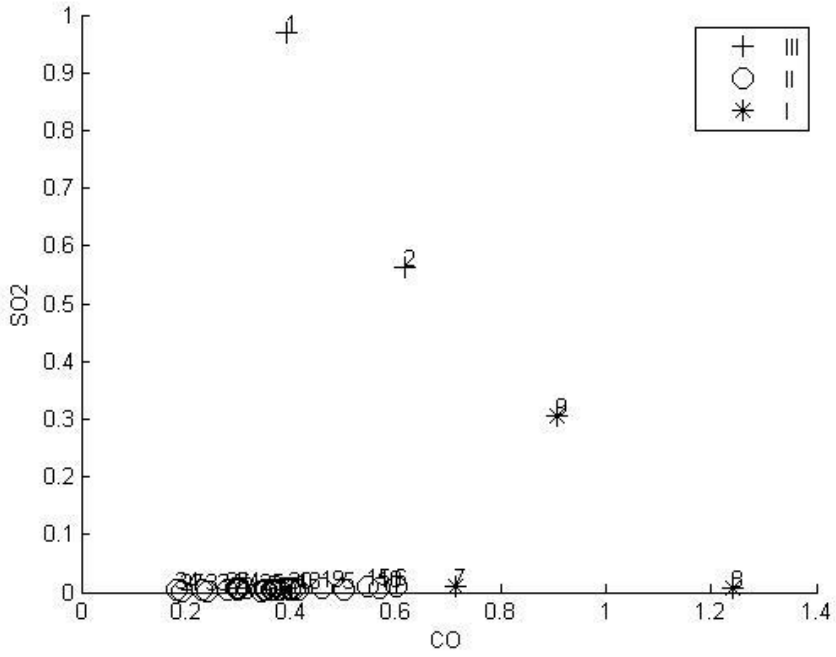

Рис. 4. Распределение среднемесячных концентраций за три года измерений

Результаты кластерного анализа четко выявили особенности вариаций содержаний исследуемых газов (табл. 2,3). Первый год наблюдений характеризовался существенными вариациями как CO так и SO2 (табл. 2, рис1). Соответственно получены три кластера: с аномально высоким содержанием SO_2 (≥ 0,6 мг/м3, кластер II), с максимально высоким содержанием CO (≥ 0,9 мг/м3, кластер III) и с промежуточными значениями концентраций газов (кластер I). Во второй и третий года наблюдений содержание серы снизилось и существенно меньше варьировало. Поэтому выделенные кластеры отличаются незначительно только по содержанию CO (рис 2,3). В первый год наблюдений - ~0,3; 0,4; 0,5 (в кластерах III,I и II, соответственно). Во второй год наблюдений - ~0,2; 0,3; >0.3-0.4 (в кластерах I,II и III, соответственно).

Результаты кластерного анализа за три года наблюдений показали, что в большинстве случаев (за исключением 1,2,7,8,9 месяцев первого года) среднемесячные концентрации газов изменялись незначительно. Объединяющий эти данные кластер II характеризуется минимальными значениями содержания SO2 (≤0,01 мг/м3), содержание CO варьирует от 0,2 до 0,6 мг/м3. Остальные кластеры соответствуют аномальным концентрациям анализируемых газов. Кластер I характеризуется максимальным содержанием CO (≥0,7 мг/м3) и включает с 7 по 9 месяцы первого года наблюдений. Кластер III - максимальным содержанием SO2 (≥0,6 мг/м3) и включает 1 и 2 месяцы первого года наблюдений.

Заключение. Применение параллельных вычислений позволило сократить время выполнения поставленной задачи.

В целом, можно сделать вывод о том, что двухуровневый кластерный анализ – достаточно эффективный метод повышения точности кластерного анализа, позволяющий в некоторой степени нивелировать недостатки одного метода кластеризации, за счет применения другого метода на следующем уровне.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Миркин Б.Г. Методы кластер-анализа для поддержки принятия решений: обзор.препринт WP7/2011/03 // Национальный исследовательский университет Высшая школа экономики. М. : Изд. дом Национального исследовательского университета Высшая школа экономики., 2011. 88 с. 150 экз.

2. Янчуковский В.Н. Использование параллельных вычислений в кластерном анализе для формирования комплексных деталей. // Вестник Иркутского государственного технического университета №6(65)., Иркутск: ИрГТУ, 2012. С. 25-30.

3. Центр компетенций MathWorks Нахождение центров кластеров данных с использованием субтрактивного алгоритма [Электронный ресурс] // MATLAB.Exponenta: [сайт]. [2001-2014]. URL: http://matlab.exponenta.ru/fuzzylogic/book2/1/subclust.php (Дата обращения: 20.01.2014).

4. The MathWorks, Inc. MATLAB Parallel Computing Toolbox 5 User's guide. – Natick, 2010. 713p.