DOI: 10.12737/article_58e6133853c808.40302316

Янчуковский В.Н., ст. преп.
Иркутский национальный исследовательский технический университет

СИСТЕМА РАСПРЕДЕЛЕННОГО КЛАСТЕРНОГО АНАЛИЗА ОБЪЕКТОВ С ОДНОРОДНЫМИ ПРИЗНАКАМИ – МОДЕЛЬ И РЕАЛИЗАЦИЯ

V.Yanchukovsky@gmail.com

В статье описывается система распределённого кластерного анализа объектов с однородными признаками: основные требования, блоки и элементы, ее графоаналитическая и математическая модели. Математическая модель представлена в системе Пи-исчисления. Приводится ее программная реализация на языке Matlab. Для больших объемов данных предусмотрены параллельные вычисления. Представлены методы визуализации полученных результатов.

Ключевые слова: *кластерный анализ, параллельные вычисления, вычислительный эксперимент, Пи-исчисление.*

Введение. В последнее десятилетие, благодаря развитию сетевых технологий, наблюдается экспоненциальный рост количества доступной и обрабатываемой информации. В связи с этим относительно недавно появился термин Big Data [1], сочетающий в себе такие подходы, как кластерный анализ – многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы [2], и параллельные вычисления - способ организации компьютерных вычислений, при котором программы разрабатываются как набор взаимодействующих вычислительных процессов, работающих параллельно (одновременно) [3]. Результатом совместного применения описанных выше методов является система распределенного кластерного анализа и её программная реализация.

Общее описание системы. Исходя из специфики методов кластерного анализа и параллельных вычислений сформулируем основные требования к системе:

1. Поскольку среди алгоритмов кластерного анализа нет какого-либо универсального, то, для повышения точности и эффективности, следует использовать несколько наиболее распространенных алгоритмов;

2. В случае если объем исходных данных будет значительным, для анализа результатов и восприятия их графического представления следует использовать метод понижения размерности для визуализации результатов получаемого разбиения и методы параллельных вычислений;

3. Некоторые распространенные алгоритмы требуют задавать количество кластеров как входной параметр, и в случае, когда, при предварительном анализе, пользователь не может сам оценить это количество, следует использовать двухуровневую схему, когда на первом шаге менее точный алгоритм вычисляет количество кластеров, а более точный алгоритм на втором шаге производит окончательное разбиение;

4. Для более высокой эффективности анализа полученных результатов следует использовать несколько методов визуализации разбиений на кластеры, поскольку нет какого-либо одного универсального метода.

Исходя из всего вышеперечисленного, следует, что система будет включать следующие блоки:

- Блок алгоритмов кластерного анализа (K-means, FCM, иерархический, субтрактивный, ФОРЭЛ);
- Блок различных методов визуализации результатов в виде дендрограмм, силуэтов и графиков в координатах главных компонент;
- Блок двухуровневого кластерного анализа;
- Блок параллельных расчетов.

Для удобства формализации структуры системы и перехода к математической модели в системе пи-исчисления представим структуру в упрощенном виде как набор из основных конструкций – графоаналитически.

Пи-исчисление – математическая модель процессов, взаимосвязи которых изменяются. Основной вычислительный шаг – передача канала связи между двумя процессами; после этого получатель может использовать канал для дальнейшего взаимодействия с другими участвующими сторонами. Именно эта особенность исчисления делает его крайне удобным для моделирования систем, в которых доступные ресурсы изменяются с течением времени [4].

Примитивными сущностями пи-исчисления являются имена. Их бесконечно много, они лишены внутренней структуры. Имена записываются как символьные строки,

начинающиеся со строчной буквы: x, y,...$\in X$. Процесс P (выражение пи-исчисления) представляет собой одно из следующего списка[5]:

1) $c(x).P$ – входной префикс, получение данных **x** из канала **c**;

2)$\bar{c}\langle y\rangle.P$– выходной префикс, передача данных **y** по каналу **c**;

3) $P\ |Q$ – параллельный запуск двух процессов;

4) $!P$ – репликация процесса;

5) $(\nu x)P$ – объявление канала и последующее выполнение процесса;

6) τ_P – внутреннее действие процесса;

7) 0 —— пустой процесс.

На рис.1 представлена модель описанной системы в графоаналитическом виде. В данном случае все действия (задачи) представлены в виде процессов, переходы между действиями заменены на именованные потоки процесса, подсистемы заменены на блоки параллельного разделения, синхронизации и выбора, все блоки принадлежат одному из трех типов [5]:

• блок параллельного разделения, если из него выходит несколько потоков процесса;

• блок синхронизации, если в него входит несколько потоков процесса;

• блок выбора, если из него выходит несколько процессов.

Любой процесс можно представить, как набор из основных конструкций [5]. Процессы, показанные на рис. 1, могут быть представлены в терминах пи-исчисления следующим образом:

$$
\begin{aligned}
&A = !a(x).\tau_A.(\bar{b}\langle x\rangle.0 + \bar{i}\langle x\rangle.0);\\
&B = !b(x).\tau_B.\bar{c}\langle x\rangle.0;\\
&C = !c(x).\tau_C.\bar{d}\langle x\rangle.0;\\
&D = !d(x).\tau_D.\overline{e1}\langle x\rangle.0;\\
&E = e1(x).!2(x).\tau_E.\overline{f1}\langle x\rangle.0;\\
&F = (!f1(x)\,|\,f2(x)).\tau_F.\bar{g}\langle x\rangle.0;\\
&G = !g(x).\tau_G.\bar{h}\langle x\rangle.0;\\
&H = !h(x).\tau_H.(\bar{a}\langle x\rangle.0 + 0);\\
&I = !i(x).\tau_I(\overline{d2}\langle x\rangle.0 + \overline{e2}\langle x\rangle.0 + \overline{f2}\langle x\rangle.0\ );
\end{aligned}
\tag{1}
$$

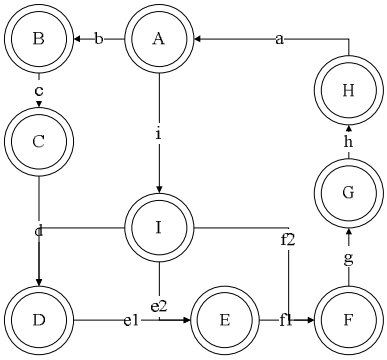

Рис. 1. Упрощенный вид структуры системы, где A – пользователь, B – смежная система формирования исходных данных, C – входные данные, D – блок параллельных вычислений, E – блок расчета, F – блок визуализации, G – выходные данные, H –процесс анализа данных, I – смежная система формирования управляющих параметров

Таким образом, вся модель системы может быть описана в виде следующих ниже выражений:

$$P = A\,|\,(B\,|\,C + I)\,|\,D\,|\,E\,|\,F\,|\,G\,|\,H\,|\,(A + 0) \tag{2}$$

Или, подставляя соответствующие выражения:

$$
\begin{aligned}
&P = !a(x).\tau_A.(\bar{b}\langle x\rangle.0 + \bar{i}\langle x\rangle.0)\,|\,(!b(x).\tau_B.\bar{c}\langle x\rangle.0\,|\\
&!c(x).\tau_C.\bar{d}\langle x\rangle.0 + !i(x).\tau_I(\overline{d2}\langle x\rangle.0 + \overline{e2}\langle x\rangle.0 + \overline{f2}\langle x\rangle.0\ ))\,|\\
&!d(x).\tau_D.\overline{e1}\langle x\rangle.0\,|\,e1(x).!2(x).\tau_E.\overline{f1}\langle x\rangle.0\,|\,(!f1(x)\,|\,f2(x)).\tau_F.\bar{g}\langle x\rangle.0\,|\\
&!g(x).\tau_G.\bar{h}\langle x\rangle.0\,|\,(!a(x).\tau_A.(\bar{b}\langle x\rangle.0 + \bar{i}\langle x\rangle.0) + 0)
\end{aligned}
\tag{3}
$$

В качестве элемента модели будем рассматривать процесс – совокупность взаимосвязанных действий, преобразующих входящие данные в исходящие.

Все элементы системы имеют общую концептуальную модель (рис. 2), за исключением двухуровневого элемента, который представляет собой синтез двух алгоритмов (рис. 3).

Представим данную модель в терминах пи-исчисления:

или:

$$
\begin{aligned}
&A = !a(x).\tau_A.\bar{b}\langle x\rangle.0;\\
&B = !b(x).\tau_B.\bar{c}\langle x\rangle.0;\\
&C = !c(x).\tau_C.\bar{d}\langle x\rangle.0;\\
&D = !d(x).\tau_D.\bar{a}\langle x\rangle.0;
\end{aligned}
\tag{4}
$$

Таким образом, данный элемент может быть записан в виде:

$$P = A\,|\,B\,|\,C\,|\,D\,|\,(A + 0) \tag{5}$$

$$P = !a(x).\tau_A.\overline{b}\langle x\rangle.0 \mid !b(x).\tau_B.\overline{c}\langle x\rangle.0 \mid !c(x).\tau_C.\overline{d}\langle x\rangle.0 \mid !d(x).\tau_D.\overline{a}\langle x\rangle.0 \mid (!a(x).\tau_A.\overline{b}\langle x\rangle.0 + 0) \quad (6)$$

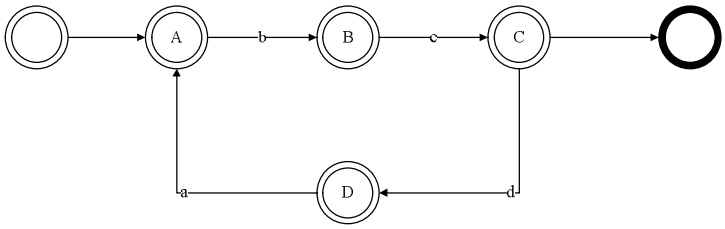

Рис. 2. Концептуальная модель элемента Системы распределенного кластерного анализа, где A – процесс изменения, B – процесс кластерного анализа, C - визуализация, D – блок принятия решений

Поскольку двухуровневый элемент состоит из двух процессов, упрощенная графоаналитическая модель примет следующий вид (рис. 4):

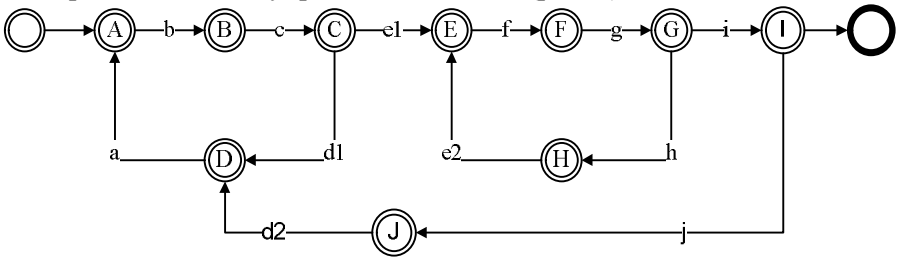

Рис. 3. Упрощенная графоаналитическая модель двухуровневого элемента, где A – процесс изменения, B – субтрактивный процесс, C - входные данные, D – блок обратной связи, E – процесс изменения, F – процесс k-means, G- выходные данные, H – блок обратной связи, I – блок визуализации, J – пользователь

Представим модель в терминах пи-исчисления:

$$\begin{aligned}
&A = !a(x).\tau_A.\overline{b}\langle x\rangle.0;\\
&B = !b(x).\tau_B.\overline{c}\langle x\rangle.0;\\
&C = !c(x).\tau_C.(\overline{d1}\langle x\rangle.0 + \overline{e1}\langle x\rangle.0);\\
&D = d1(x).!d2(x)\tau_D.\overline{a}\langle x\rangle.0;\\
&E = e1(x).!e2(x)\tau_E.\overline{f}\langle x\rangle.0;\\
&F = !f(x).\tau_F.\overline{g}\langle x\rangle.0;\\
&G = !g(x).\tau_G.(\overline{i}\langle x\rangle.0 + \overline{h}\langle x\rangle.0);\\
&H = !h(x).\tau_H.\overline{e2}\langle x\rangle.0;\\
&I = !i(x).\tau_I.\overline{j}\langle x\rangle.0;\\
&J = !j(x).\tau_I.\overline{d2}\langle x\rangle.0;
\end{aligned} \quad (7)$$

Таким образом, данный элемент может быть записан в виде:

$$P = A \mid B \mid C \mid ((D+0)+E) \mid F \mid G \mid (H \mid (E+0)+I) \mid J \mid (D+0) \quad (8)$$

или:

$$\begin{aligned}
&P = !a(x).\tau_A.\overline{b}\langle x\rangle.0 \mid !b(x).\tau_B.\overline{c}\langle x\rangle.0 \mid !c(x).\tau_C.(\overline{d1}\langle x\rangle.0 + \overline{e1}\langle x\rangle.0) \mid\\
&((d1(x).!d2(x)\tau_D.\overline{a}\langle x\rangle.0 + 0) + e1(x).!e2(x)\tau_E.\overline{f}\langle x\rangle.0) \mid\\
&!f(x).\tau_F.\overline{g}\langle x\rangle.0 \mid !g(x).\tau_G.(\overline{i}\langle x\rangle.0 + \overline{h}\langle x\rangle.0) \mid (!h(x).\tau_H.\overline{e2}\langle x\rangle.0 \mid\\
&(e1(x).!e2(x)\tau_E.\overline{f}\langle x\rangle.0 + 0) + !i(x).\tau_I.\overline{j}\langle x\rangle.0) \mid\\
&!j(x).\tau_I.\overline{d2}\langle x\rangle.0 \mid (d1(x).!d2(x)\tau_D.\overline{a}\langle x\rangle.0 + 0)
\end{aligned} \quad (9)$$

Графоаналитическая модель блока параллельных вычислений может быть представлена в следующем виде (рис.4):

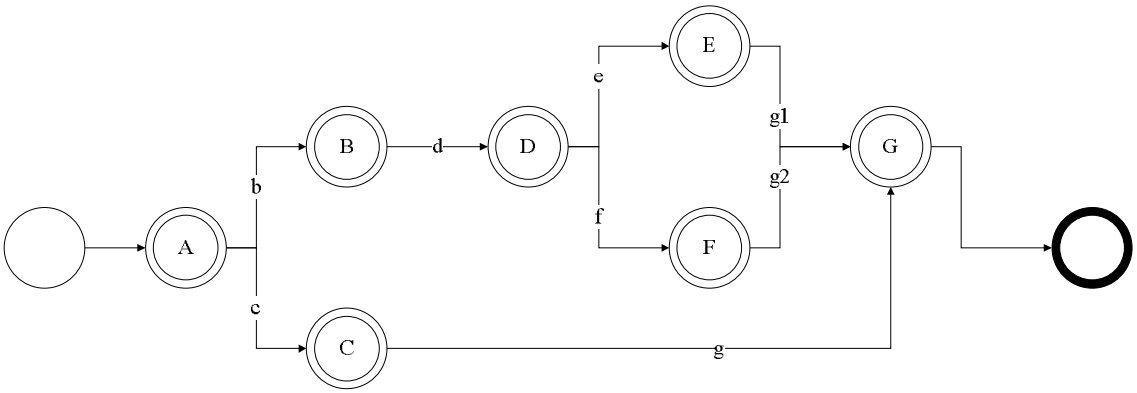

Рис.4. Графоаналитическая модель блока параллельных вычислений, где A – смежная система формирования исходных данных, B – процесс запуска параллельного режима, C – исходные данные, D – процесс разбиения данных на параллельные процессы, E,F – параллельные процессы, G – блок визуализации

Представим эту модель в терминах пи-исчисления:

$$\begin{aligned} A &= a(x).\tau_A.(\overline{b}\langle x\rangle.0 + \overline{c}\langle x\rangle.0);\\ B &= b(x).\tau_B.\overline{d}\langle x\rangle.0;\\ C &= c(x).\tau_C.\overline{g}\langle x\rangle.0;\\ D &= d(x).\tau_D.(\overline{e}\langle x\rangle.0 \mid \overline{f}\langle x\rangle.0);\\ G &= g(x).(!g1(x) \mid g2(x)).\tau_G.0;\\ E &= !e(x).\tau_E.\overline{g}1\langle x\rangle.0;\\ F &= !f(x).\tau_F.\overline{g}2\langle x\rangle.0; \end{aligned} \tag{10}$$

Таким образом, данный элемент может быть записан в виде:

$$P = A \mid (B + C) \mid D \mid (E \mid F) \mid (G + 0) \tag{11}$$

Или:

$$\begin{aligned} P = {} & a(x).\tau_A.(\overline{b}\langle x\rangle.0 + \overline{c}\langle x\rangle.0) \mid d(x).\tau_D.(\overline{e}\langle x\rangle.0 \mid \overline{f}\langle x\rangle.0) \mid \\ & (!e(x).\tau_E.\overline{g}1\langle x\rangle.0 \mid !f(x).\tau_F.\overline{g}2\langle x\rangle.0) \mid (g(x).(!g1(x) \mid g2(x)).\tau_G.0 + 0) \end{aligned} \tag{12}$$

На основе разработанных моделей системы распределенного кластерного анализа представим следующую общую структуру (рис.5).

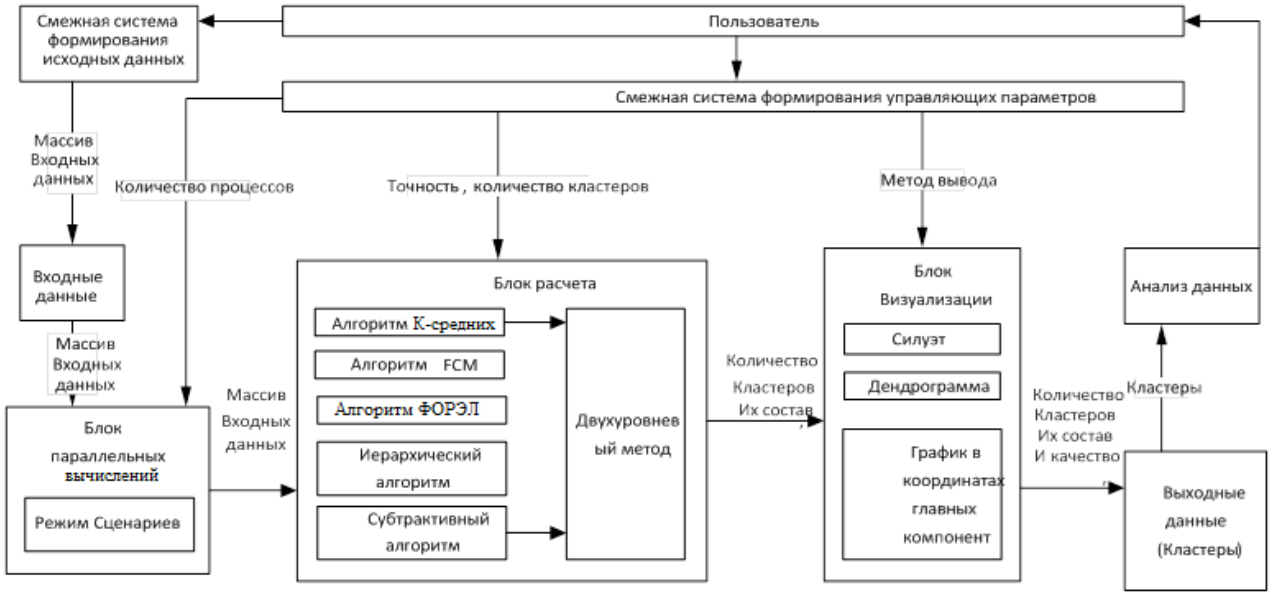

Рис. 5. Структура системы распределенного кластерного анализа

Все блоки служат для достижения общей цели системы. Цель блока распараллеливания - ускорение получения выходных данных системы. Цель блока расчёта – формирование выходных кластеров. В зависимости от выбранного метода достигается необходимая точность в определенной предметной области. Цель блока визуализации – представление выходных данных в удобной для пользователя форме. Для этого используется несколько различных методов, в совокупности позволяющих предоставить наиболее полную информацию о полученных результатах.

В блоке визуализации использованы следующие методы визуализации:

- График на основе главных компонент.
- График силуэтов кластеров.
- График дендрограмм.

Их возможности по визуализации отображены в табл. 1.

Таблица 1

Блок визуализации

График	Качество	Состав	Количество
Дендрограмма	+	+	-
График Силуэта	+	-	+
График главных компонент	-	+	+

Основные методы визуализации, используемые в системе распределенного кластерного анализа, такие как дендрограмма и силуэты, дают неполное представление о получаемых классах. Силуэты позволяют оценить качество кластеров, но не их состав[6]. Дендрограмма в свою очередь представляет собой дерево, то есть граф без циклов, построенный по матрице мер близости, и позволяет изобразить взаимные связи между объектами из заданного множества[7]. Помимо этого, в случае, когда объекты кластеризации имеют более двух признаков, для удобства восприятия и визуализации необходимо понижать размерность данных. С этой целью реализован метод главных компонент[8] и визуализация кластеров, полученных в результате его применения[9].

Программная реализация. На основе разработанной формальной модели системы распределенного кластерного анализа, в среде MATLAB разработан программный комплекс[10]. В системе MATLAB созданы основные формы графического приложения, представленные на рис.6:

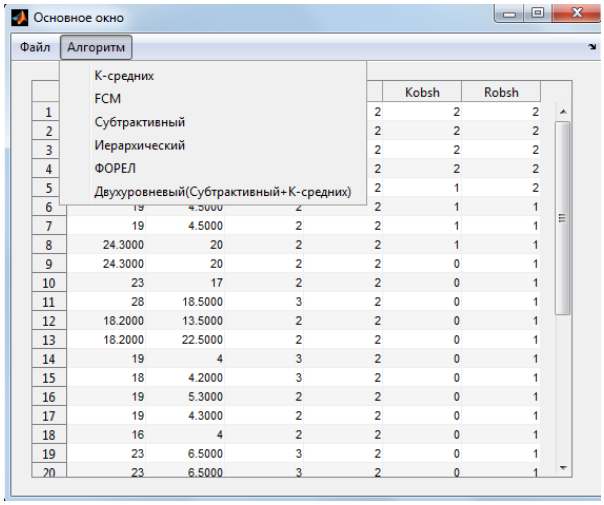

Рис. 6. Главная форма программного комплекса кластерного анализа

Параллельные вычисления в программном комплексе для встроенных алгоритмов реализованы на основе сценариев – последовательного набора команд встроенного языка программирования MATLAB, записанных в специализированном исполняемом файле сценариев – m-файле.

Для параллельной версии алгоритма ФОРЭЛ[11], ввиду того, что он генерирует разное число классов в разных задачах, выбран режим spmd («одна программа — много данных») [12]. Параллельная версия двухуровневого метода кластерного анализа реализована в режиме parfor (параллельный цикл for). Такой выбор обусловлен тем, что позволяет минимизировать изменения, вносимые в последовательный вариант программы.

Для иерархического кластерного анализа удобным средством визуализации результатов является функция dendrogram, которая выводит дерево дендрограммы (рис.5). Для всех использованных автором методов кластерного анализа можно посчитать величину силуэта и вывести его график (рис.7).

Для удобства визуализации результатов кластеризации в координатах главных компонент используется встроенная в MATLAB функция gscatter. Она реализует график рассеяния двух переменных – двух главных компонент, образующих двумерную систему координат, сгруппированных по значениям третьей переменной – массиву номеров классов, которым принадлежат объекты исходной выборки в соответствии с тем или иным алгоритмом (рис. 8).

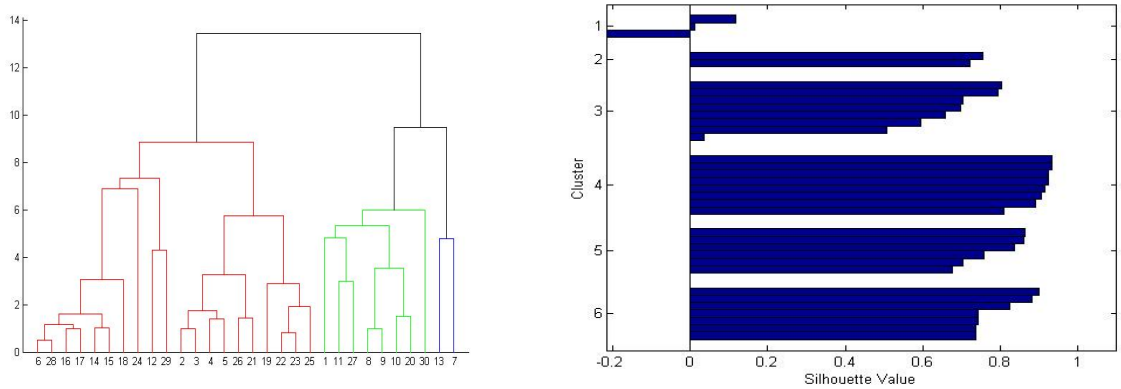

Рис. 7. График дендрограммы и график силуэтов кластеров

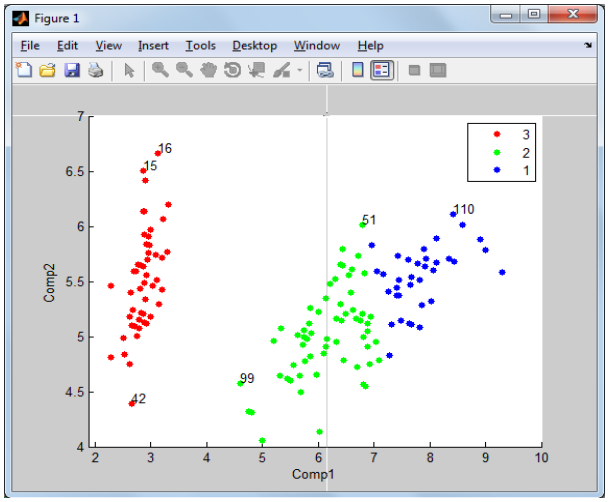

Рис. 8. Визуализация результатов с использованием главных компонент

Заключение. Представленная формальная модель и программная реализация системы распределенного кластерного анализа была апробирована на группе изделий машиностроительного производства – комплексных деталях[13, 14] и наблюдениям по концентрациям некоторых газов в атмосфере за несколько лет – результатам трехлетнего мониторинга содержания газов SO_2 и CO в воздушной атмосфере центрального района г. Санкт-Петербурга [15].

Результаты работы характеризуют разработанную систему как универсальную систему кластерного анализа, которая может быть использована во многих отраслях.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Черняк Леонид. Большие Данные – новая теория и практика (рус.) // Открытые системы. СУБД. М.: Открытые системы, 2011. № 10. С.18–26.

2. Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: Классификация и снижение размерности. М.: Финансы и статистика, 1989. 607 с.

3. Воеводин В. В., Воеводин Вл. В. Параллельные вычисления. СПб: БХВ-Петербург, 2002. 608 с.

4. Milner R. Communicating and Mobile Systems: the PI-Calculus. Cambridge University Press. 1999, 159 p.

5. Parrow J. An Introduction to the π-Calculus, chapter 8, pages 479–543. Handbook of Process Algebra. Elsevier, 2001.

6. Rouseeuw P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. // Journal of Computational and Applied Mathematics. Vol. 20. №. 1. 1987. Pp. 53–65.

7. Жамбю М. Иерархический кластер-анализ и соответствия. М.: Финансы и статистика, 1988. 345 с.

8. Jolliffe I.T. Principal Component Analysis, Series: Springer Series in Statistics, 2nd ed., Springer, NY, 2002. №XXIX, 487 p. 28 .

9. Зиновьев А. Ю., Визуализация многомерных данных, Красноярск // Изд. КГТУ, 2000. 180 с.

10.Янчуковский В.Н., Сосинская С.С. Программный комплекс «Система распределенного кластерного анализа», свидетельство о регистрации в реестре программ для ЭВМ №2016615937 от 02 июня 2016 г.

11.Загоруйко Н.Г. Прикладные методы анализа данных и знаний. // Новосибирск: Издательство Института математики, 1999. 270 с.

12.Parallel Computing Toolbox™ 5 User's Guide // The MathWorks, Inc. Natik, 2007, 730 p.

13.Соколовский А. П. «Автоматизация технологических процессов механической обработки металлов», Автомат. и телемех., 1938, № 3. С.117–139.

14.Митрофанов С.П. Научная организация машиностроительного производства. 2-е изд., доп. и перераб. Л. : Машиностроение, 1976. 712 с.

15.Янчуковский В.Н., Сосинская С.С., Козловский А.С., Челибанов В.П. Двухуровневый кластерный анализ в среде MATLAB с применением параллельных вычислений // Вестник БГТУ им. В.Г. Шухова. Шухова. №5. 2014. С. 201–205.

Yanchukovskiy V.N.

MODEL AND IMPLEMENTATION OF HOMOGENEOUS ATTRIBUTES OBJECTS DISTRIBUTED CLUSTER ANALYSIS SYSTEM

Homogeneous attributes objects distributed cluster analysis system described: basic requirements, blocks and elements, its graphoanalytical and mathematical models. The mathematical model is represented in the Pi-calculus. Its software implementation in Matlab is given. In case of big data, parallel computing is provided . Methods for visualizing the results are presented.

Key words: *cluster analysis, parallel computing, computational experiment, PI-calculus.*

Янчуковский Владислав Николаевич, старший преподаватель кафедры вычислительной техники.
Иркутский национальный исследовательский технический университет
Адрес: Россия, 664074, г. Иркутск, ул. Лермонтова 83.
E-mail: V.Yanchukovsky@gmail.com