На правах рукописи

Рукавицын Андрей Николаевич

СРЕДСТВА КЛАСТЕРИЗАЦИИ РАСПРЕДЕЛЕННЫХ ДАННЫХ НА ОСНОВЕ НЕЙРОННЫХ СЕТЕЙ КОХОНЕНА

05.13.11. – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

АВТОРЕФЕРАТ

диссертации на соискание учёной степени

кандидата технических наук

Санкт-Петербург – 2020

Работа выполнена на кафедре вычислительной техники федерального государственного автономного образовательного учреждения высшего образования «Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)» (СПбГЭТУ «ЛЭТИ»).

Научный руководитель:	**Шоров Андрей Владимирович**, кандидат технических наук, федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)» (СПбГЭТУ «ЛЭТИ»), кафедра вычислительной техники, ведущий научный сотрудник.
Официальные оппоненты:	**Прокопчина Светлана Васильевна,** доктор технических наук, профессор, федеральное государственное образовательное бюджетное учреждение высшего образования «Финансовый университет при Правительстве Российской Федерации», профессор кафедры «Системного анализа в экономике» (г. Москва); **Чечулин Андрей Алексеевич,** кандидат технических наук, доцент, Федеральное государственное бюджетное учреждение науки Санкт-Петербургский институт информатики и автоматизации Российской академии наук, лаборатория проблем компьютерной безопасности, ведущий научный сотрудник (г. Санкт-Петербург)
Ведущая организация:	Государственный научный центр РФ «Центральный научно-исследовательский и опытно-конструкторский институт робототехники и технической кибернетики» (ГНЦ РФ ЦНИИ РТК) (г. Санкт-Петербург)

Защита состоится «04» марта 2020 года в «15:30» на заседании диссертационного совета Д 212.238.01, созданном при Санкт-Петербургском государственном электротехническом университете «ЛЭТИ» им. В.И. Ульянова (Ленина) по адресу: 197376, Санкт-Петербург, ул. Профессора Попова, 5.

С диссертацией можно ознакомиться в библиотеке ФГАОУ ВО «Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)» и на сайте университета www.etu.ru в разделе «Подготовки кадров высшей квалификации» - «Объявление о защитах»

Автореферат разослан «31» декабря 2019 года.

Ученый секретарь
диссертационного совета Д 212.238.01 / А.А. Пазников /
к.т.н., доцент

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы. Развитие информационных и мобильных технологий, а также популяризация Интернета привели к высокому росту количества источников данных. В результате данные могут храниться на разных и независимо работающих устройствах, которые могут быть связаны друг с другом через локальные или глобальные сети. При этом датчики могут быть расположены географически в разных местах: охранные системы, мобильные или корпоративные сети. Подобные системы распределенных источников данных используются в различных сферах жизнедеятельности: защита окружающей среды, медицина, безопасность и др. Обработка полученных данных выполняется с использованием интеллектуального анализа данных. В результате роста количества источников информации появились методы распределенного интеллектуального анализа данных. Одним из наиболее распространенных подходов является централизация данных в едином локальном хранилище, к которым применяются традиционные методы интеллектуального анализа данных. Традиционные способы обработки имеют недостатки, связанные с конфиденциальностью, высокой стоимостью централизации данных, ограниченной пропускной способностью и высокой нагрузкой. Поэтому для работы с распределенными источниками данных необходимо адаптировать известные методы. Наиболее распространенными задачами в распределенных системах является сегментация и детектирование выбросов, которые обычно решаются методами кластеризации. Существующие алгоритмы кластеризации имеют ряд особенностей, которые могут сказываться на работе с распределенными источниками данных в распределенных системах с множеством устройств. В связи с этим, актуальной задачей является исследование в области анализа существующих методов кластеризации и подходов адаптации удовлетворяющего метода для выполнения в распределенной среде с распределенными источниками данных.

Целью диссертационной работы является разработка средств кластеризации распределенных данных с учетом типа их распределения алгоритмами, использующие нейронные сети Кохонена. Для достижения заявленной цели в работе решаются следующие **задачи:**

1. Анализ существующих средств кластеризации, в том числе распределенных данных;
2. Разработка формальной модели декомпозиции алгоритмов кластеризации, использующих нейронные сети Кохонена для горизонтально и вертикально распределенных данных;
3. Разработка метода объединения промежуточных результатов, полученных при анализе распределенных данных, с учетом типа их распределения для алгоритмов кластеризации, использующих нейронные сети Кохонена;

4. Разработка методики выполнения кластеризации на основе нейронных сетей Кохонена на распределенных источниках данных с учетом типа распределения данных.
5. Программная реализация алгоритмов кластеризации, использующих нейронные сети Кохонена для выполнения на распределенных источниках данных с учетом типа распределения данных;
6. Экспериментальная проверка полученных результатов.

Объектом исследования является процесс выполнения кластеризации в распределенных системах мониторинга.

Предметом исследования являются средства выполнения алгоритмов интеллектуального анализа данных в системах с распределенными источниками данных.

Методы исследования: методы проектирования программного обеспечения, модель системы с распределенными источниками данных, модель кластеризации.

Положения, выносимые на защиту:

1. Формальная модель декомпозиции алгоритмов кластеризации, использующих нейронные сети Кохонена для горизонтально и вертикально распределенных данных;
2. Метод объединения промежуточных результатов алгоритмов кластеризации, использующих нейронные сети Кохонена для горизонтального и вертикального распределения данных;
3. Методика выполнения алгоритмов кластеризации, использующих нейронные сети Кохонена на распределенных источниках данных.

Научная новизна работы заключается в следующем:

1. Предложена формальная модель декомпозиции алгоритмов кластеризации, использующих нейронные сети Кохонена в виде композиции функций, позволяющая выполнять распределенную кластеризацию данных в зависимости от типа распределения данных;
2. Предложен метод объединения промежуточных результатов алгоритмов кластеризации, использующих нейронные сети Кохонена для горизонтального и вертикального распределения данных.

Практическая значимость:

1. Методика выполнения алгоритмов кластеризации, использующих нейронные сети Кохонена на распределенных источниках данных, на основе предложенной модели и метода, учитывающая условия кластеризации;
2. Программная реализация модели алгоритмов кластеризации, использующих нейронные сети Кохонена и метода кластеризации для

горизонтально и вертикально распределенных данных в системах с множеством источников.

Апробация работы. Основные положения и результаты диссертационной работы докладывались и обсуждались на международных конференциях ElConRus, St. Petersburg, Russia, 2019 г, SCM, St. Petersburg, Russia, 2018 г., ElConRus, St. Petersburg, Russia, 2018 г, NEW2AN , St. Petersburg, Russia, 2017 г, ElConRus, St. Petersburg, Russia, 2017 г., SCM, St. Petersburg, Russia, 2017 г.

Обоснованность и достоверность представленных в диссертационной работе научных положений обеспечивается проведением анализа состояния исследований в данной области, подтверждается согласованностью теоретических результатов с практическими, полученными при компьютерной реализации, а также апробацией основных теоретических положений в печатных трудах и докладах на научных конференциях. Достоверность результатов диссертационной работы подтверждается программной разработкой метода и модели кластерного анализа и системы с распределенными источниками данных, протестированной в лаборатории облачных вычислений кафедры Вычислительной техники.

Публикации

Основные теоретические и практические результаты диссертации опубликованы в 22 трудах, среди них: 4 научные статьи, опубликованные в журнале, входящие в рекомендуемый перечень ВАК, 7 научных публикаций в журналах и сборниках трудов, входящих в базы цитирования Web of Science и Scopus, 4 публикации в сборниках конференций, и 7 свидетельств о регистрации программы ЭВМ.

Личный вклад соискателя состоит в непосредственном участии в получении исходных данных и научных экспериментах, разработке формальной модели и метода объединения, методики построения алгоритма, подготовке ключевой части публикаций по выполненной работе и представлению результатов работы на конференциях различного уровня, в том числе международных.

Структура и объем диссертации. Диссертационная работа состоит из введения, четырех глав, заключения, одного приложения, списка литературы (103 наименований). Общий объем работы составляет 118 страниц машинописного текста, который включает 29 рисунков, 8 таблиц, 1 приложение.

Соответствие паспорту специальности. Данное диссертационное исследование выполнено в соответствии с паспортом специальности 05.13.11 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей, отрасль – технические науки. Диссертация соответствует пунктам 8, 9 паспорта специальности 05.13.11: п. 8 – Модели и методы создания программ и программных систем для параллельной и распределенной обработки данных, языки и инструментальные средства параллельного программирования; п. 9 – Модели, методы, алгоритмы и

программная инфраструктура для организации глобально распределенной обработки данных.

ОСНОВНОЕ СОДЕРЖАНИЕ ДИССЕРТАЦИИ

Во введении дано обоснование актуальности темы диссертационного исследования, сформулированы цель и задачи работы, ее научная новизна и практическая значимость, представлены положения, выносимые на защиту.

В первой главе дается общий обзор методов кластеризации данных в распределенных системах мониторинга. Рассматриваются особенности интеллектуального анализа распределенных источников данных. Вводится формальная модель для распределенных источников данных с целью определения модели интеллектуального анализа за каждый промежуток времени. Данные, собранные на разных источниках, могут быть распределены либо однородно (то есть каждый узел наблюдает общее подмножество данных), либо гетерогенно (то есть каждый узел наблюдает соответствующее подмножество данных). Гетерогенная система рассматривает два вида распределения данных: горизонтальное и вертикальное. Установлено, что целевой является гетерогенная система, где устройства независимы и являются или имеют собственные источники данных. Выделяются требования к методу кластеризации для распределенных источников данных:

1. Однопроходность по данным – алгоритм должен выполнять кластеризацию за один проход по всем данным.
2. Поддержка разных типов входных данных – данные могут быть дискретными, непрерывными, категориальными и другими.
3. Визуализация кластеров, связей и возможность интерпретации полученной визуализации – одно из основных требований, позволяющее на полученных результатах производить дальнейший анализ и диагностику специалистам не только для детектирования выбросов или сегментации, но и выявления причин.
4. Поддержка online режима и адаптация к данным при изменении среды – в системах мониторинга должна учитывать условия изменения данных от среды.
5. Проецирования многомерного пространства в пространство с более низкой размерностью (чаще всего двухмерное) – увеличивает скорость обработки многомерных данных, получаемых от множества датчиков системы.
6. Выделение наиболее значимых атрибутов.
7. Масштабирование больших объемов данных.
8. Входные параметры не ограничены априорными знаниями, например, количеством кластеров.
9. Анализ выполняется без предположений о распределении входных данных.
10. Обнаружение аномалий.

11. Анализ данных на источниках информации без их передачи третьей стороне.
12. Поддержка анализа распределенных данных.

Проводится сравнительная характеристика методов кластеризации алгоритмов в распределенных системах мониторинга (Таблица 1.1).

Таблица 1.1 – Сравнительная характеристика методов кластеризации для систем мониторинга

Требование\ Метод	**Иерархические**	**Центроидные**	**Плотностные**	**Нейронные сети Кохонена**
Однопроход-ность	–	+	+	+
Работа с высоко-размерными данными	–	–	–	+
Поддержка разных типов входных данных	–	+	+	+
Поддержка online режима (адаптация к данным при их изменении)	–	+	+	+
Работа с большим количеством данных	–	+	+	+
Визуализация связей	–	+	–	+*
Визуализация высоко-размерных данных	–	–	–	+*
Графическая интерпретация данных	–	–	–	+*
Не требуются предварительные данные	+	–	+	+
Работа с распределенными данными	+*	+	+*	+*

* - требует предобработку специальными методами.

По результатам сравнительной характеристики (Таблица 1.1) выбран метод кластеризации на основе нейронных сетей Кохонена.

Во второй главе представлен подход описания формальных моделей декомпозиции алгоритмов кластеризации, использующих нейронные сети Кохонена. Предложена формальная модель декомпозиции для алгоритмов Growing neural gas (GNG) и Self-organizing map (SOM).

Методы кластеризации на основе нейронных сетей Кохонена используют алгоритмы Winner Takes All (WTA) или Winner Takes Most (WTM) и их производные. Можно выделить общую последовательность шагов:

1. Инициализация нейронов
2. Для каждого вектора
   a. Определение нейрона победителя
   b. Вычисление окрестности
   c. Корректировка весов и конфигурации сети (также вычисление ошибок и соединений нейронов, корректировка топологии путем добавления и удаления нейронов)

Следующие шаги предполагают непосредственную работу с данными:

- определение нейрона победителя, т.к. выполняется вычисление расстояния от каждого вектора до нейрона;
- корректировка весов, так как в вычислениях участвуют значения атрибутов векторов.

Можно отметить, что алгоритмы SOM и GNG имеют аналогичные блоки:

- цикл по входным векторам;
- цикл по нейронам;
- цикл по атрибутам (вычисление расстояния);
- калибровка нейронов.

Отличия:

- калибровка ошибок в GNG;
- изменение связей в GNG.

В результате обобщенный алгоритм представлен на рисунке 2.1.

```
1.  for i=1 to n
2.      for k=1 to p
3.          μ[i].ω[k] = random()
4.  for j=1 to z
5.      for i=1 to n
6.          for k=1 to p
7.              μ[i].δ[j]=μ[i].δ[j]+(d[j,k]-μ[i].ω[k])^2
8.          μ[i].δ[j] = sqrt(μ[i].δ[j])
9.      endfor
10.     i_w = 1
11.     for i=2 to n
12.         if μ[i].δ[j]> μ[i_w].δ[j] then i_w = i
13.     for i=1 to n
14.         for k=1 to p
15.             μ[i].ω[k]=μ[i].ω[k]+η·G(i,i_w,j)·(d[j,k]-μ[i].ω[k])
16. endfor
```

Рисунок 2.1 – Псевдокод обобщенного алгоритма

Обобщенный алгоритм можно представить в виде композиции функциональных блоков:

Kohonen = $fd_1 \circ f_0$ = (loopr 1 z ($fd_9 \circ f_7 \circ f_6 \circ fd_2$) d)° f_0 = (loopr 1 z (loopn 1 n fd_{10})° (loopn 2 n f_8)° f_6° (loopn 1 n ($f_5 \circ fd_3$)) d)°f_0 = (loopr 1 z (loopn 1 n (loopc 1 p fd_{11}))°(loopn 2 n f_8)° f_6° (loopn 1 n (f_5°(loopc 1 p fd_4))) d)° f_0

Функции формируются для каждой строки алгоритма следующим образом:

- f_1 - цикл для строк набора данных (строка 4 на рисунке 2.1):
  f_1 = loopr 1 z ($fd_9 \circ f_7 \circ f_6 \circ fd_2$);
- fd_2 цикл по нейронам (строка 5 на рисунке 2.1):

fd_2= loopn 1 n ($f_5 \circ fd_3$);

- fd_3 цикл по данным (строка 6 на рисунке 2.1):
  fd_3 = loopc 1 p fd_4;
- fd_4 вычисление суммы расстояний (строка 7 на рисунке 2.1);
- f_5 вычисление расстояния (строка 8 на рисунке 2.1);
- f_6 инициализация индекса победителя (строка 10 на рисунке 2.1);
- f_7 цикл по нейронам (строка 11 на рисунке 2.1):

f_7= loopn 2 n f_8;

- f_8 определение победителя (строка 12 на рисунке 2.1);
- fd_9 цикл по нейронам (строка 13 на рисунке 2.1):

fd_9= loopn 1 n fd_{10};

- fd_{10} цикл по данным (строка 14 на рисунке 2.1):

fd_{10} = loopc 1 p fd_{11};

- fd_{11} корректировка весов (строка 15 на рисунке 2.1).

На основе полученной декомпозиции обобщенного алгоритма можно представить **формальную модель декомпозиции SOM** как следующую композицию функций:

som = (loopr 1 z (loopn 1 n (loopc 1 p f'd_{11}))° (loopn 2 n f_8)° f_6° (loopn 1 n (f_5°(loopc 1 p f'd_4))) d)° f_0.

Аналогичным образом выразим **формальную модель декомпозиции GNG**:

gng = f_{28} ° (loopr 1 c (((loopc 1 p f_{26}) (loopq 1 n f_{24}) (loopn 1 n f_{22}) ° f_{20}) ° (loopn 1 n f_{18}) ° (loopn 1 n (loopq 1 n f_{16})) ° f_{13} ° (loopn 1 n (loopc 1 p f'd_{12})) ° f_9 ° (loopn 2 n f'$_8$) ° f_6 ° (loopn 1 n (f_5 ° (loopc 1 p f'd_4)))) d) ° f0

В третьей главе представлены формальные модели декомпозиции алгоритмов SOM и GNG для выполнения на распределенных источниках данных. Представлен метод объединения промежуточных результатов для алгоритмов кластеризации, использующих нейронные сети Кохонена. Описана методика построения кластеризации с использованием нейронных сетей Кохонена в распределенных системах мониторинга. Предложены следующие стратегии распределенного выполнения:

1. Промежуточная синхронизация нейрона-победителя, когда результаты отправляются для определения нейрона-победителя на общий узел, где они

объединяются и обобщенная модель рассылается на узлы источников для корректировки весов.

2. Слияние нейронных сетей после обработки всех или части векторов, когда нейронные сети строятся на каждом узле по отдельности, а после этого объединяются.

В случае первой стратегии для каждого вектора выполняется два взаимодействия узлов источников с общим узлом, что замедляет работу алгоритма и увеличивает сетевой трафик.

Работа при второй стратегии зависит от частоты слияния *n*. Так при *n=1* слияние нейронных сетей выполняется для каждого вектора. Это повышает точность модели, но увеличивает число передач данных по сети и, как следствие, время анализа и сетевой трафик. При такой стратегии необходимо учитывать время завершения обработки самого медленного источника.

На рисунке 3.1 представлена схема передачи для кластеризации распределенных данных, где:

1. Инициализация нейронов;
2. Определение победителя;
3. Вычисление окрестности;
4. Выполнение корректировки весов;
5. Объединение нейронов.

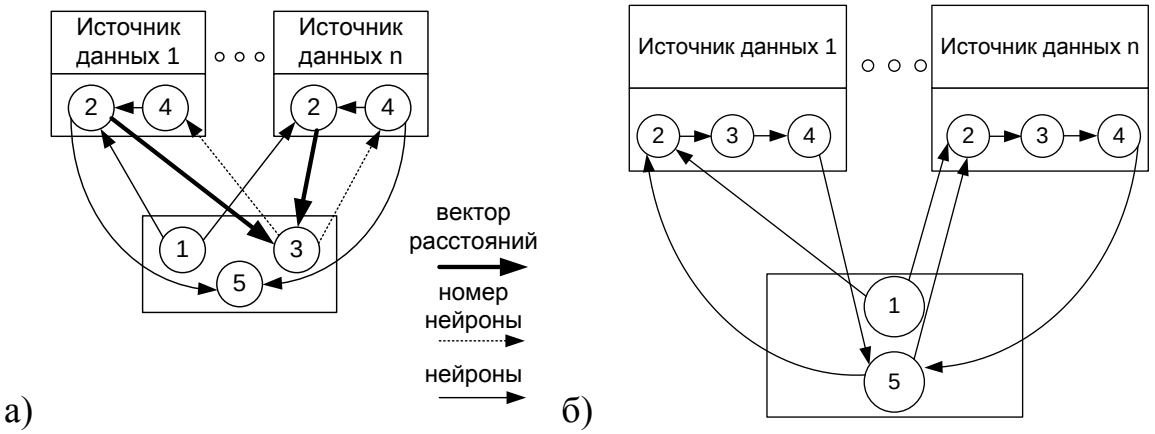


Рисунок 3.1 – Схема передачи данных в распределенном алгоритме:
а) промежуточная синхронизация; б) слияние нейронных сетей

Выбор стратегии и параметров слияния зависит от условий среды выполнения кластеризации и способа распределения данных.

При горизонтальном распределении и использовании первой стратегией первичная синхронизация результатов выполняется после определения победителя на каждом узле-источнике и заключается в выборе единого победителя. Однако, корректировка весов на следующем шаге для выбранного единого нейрона-победителя, будет верной только для своего входного вектора (вектора для которого он выбран). Корректировка его веса по векторам от других источников приведет к неверному смещению, т.к. для них победителями являются другие нейроны. Следовательно, первая стратегия при горизонтальном распределении данных не может использоваться. При вертикальном

распределении данных расстояния от вектора до нейронов с использованием Евклидового расстояния будут вычисляться по частям:

$$d_j^1 = \sqrt{\sum_{i=1}^{n-1}(w_i - x_{ji})^2}; d_j^w = \sqrt{\sum_{i=n}^{m}(w_i - x_{ji})^2}$$

При синхронизации будет вычислено общее расстояние: $d_j = d_{j1}+\ldots+d_{jw}$.

$$d'_j \neq d^1{}_j + d^2{}_j, d_j = \sqrt{\sum_{i=1}^{m}(w_i - x_{ji})^2}$$

При этом отношение неравенства сохраняется:

$$d_j < d_{j+1}, (d^1{}_j + d^2{}_j) < (d^1{}_{j+1} + d^2{}_{j+1})$$

Нейроны объединяются по атрибутам:

$$w_i = \{a_1, a_2, a_3, a_4, a_5, a_6\};$$

$$w_i = \{a_1, a_2, a_3, 0, 0, 0\} + \{0, 0, 0, a_4, a_5, a_6\}.$$

При второй стратегии синхронизация результатов осуществляется после завершения обработки n векторов и заключается в вычислении весов для каждого нейрона сети. Вычисления выполняются для нейронов с одинаковыми индексами:

$$w_i = f\left(w^1{}_i,\ w^2{}_i,\ \ldots,\ w^w{}_i\right),\ 0 < i < n$$

Функция вычисления общего веса f может быть разной. В простейшем случае это может быть функция вычисления среднего:

$$f\left(w^1{}_i, w^2{}_i, \ldots, w^w{}_i\right) = \left(w^1{}_i + w^2{}_i + \ldots + w^w{}_i\right) / w$$

В соответствии с этим **метод объединения** можно представить в виде псевдокода для промежуточной синхронизации (Рисунок 3.2).

```
function mergeSum (distance, i) {
  var delta = 0.0;
  for (var j = 1; j < μ[i].δ.size(); j++) {
    delta += μ[i].δ[j] - distance;
  }
  distance += delta;
  return distance;
};
function mergeWinner (i) {
  var index = winner[1].iNeuron;
  var distance = winner[1].distanceToNeuron;
  for (var j = 1; j < μ[i].size(); j++) {
    if (distance > μ[i].δ[j]) {
      index = i;
      distance = μ[i].δ[j];
    }
  }
  return index;
};
```

Рисунок 3.2 – Псевдокод метода объединения на основе промежуточной синхронизации

При слиянии нейронных сетей метод объединения представлен на рисунке 3.3.

```
function mergeMap (m)  {
  for (var n = 0; n < m.size(); n++) {
    for (var i = 0; i < m[i].size(); i++) {
      for (var j = 0; j < m[i].size(); j++) {
        μ[i].ω[j] += maps[i].ω[j] / m.size();
      }
    }
  }
}
```

Рисунок 3.3 – Псевдокод метода объединения на основе слияния нейронных сетей

Условия применение стратегий в зависимости от вида распределения данных представлены в таблице 3.1.

Таблица 3.1 – Условия применения стратегий для разных видов распределения данных

Распределение	**Стратегия**	**Условия использования**
Горизонтальное	1я стратегия	Не применима
	2я стратегия	При использовании небольшой окрестности
Вертикальное	1я стратегия	Применима
	2я стратегия	При наличии зависимости между атрибутами

При горизонтальном распределении **формальную модель декомпозиции SOM** можно представить:

kohonenHpar = (loopr 1 z (loopn 1 n (**paralleld** [loopc 1 p f'd_{11}]))° f_7° f_6° (loopn 1 n (f_5°(**paralleld** [loopc 1 p f'd_4]))) d)°f_0

Тогда **формальная модель декомпозиции GNG** для горизонтального распределения:

hGNG = f28 ° paralleld(loopr 1 c ((parallels[loopc 1 p f26]) ° (parallels[loopq 1 n f24]) ° (parallels[loopn 1 n f22]) ° f20) ° (parallels[loopn 1 n f18]) ° (parallels[loopn 1 n (parallels[loopq 1 n f16])] ° f13 °(parallels[loopn 1 n (parallels[loopc 1 p **fd12**])]) ° f9 ° (parallels[loopn 2 n f8]) ° f6 ° (parallels[loopn 1 n (f5 ° (parallels[loopc 1 p **fd4**]))])) °) ° f0

При вертикальном распределении **формальная модель декомпозиции SOM** будет выглядеть следующим образом:

kohonenVpar'=(**paralleld**[loopc 1 p fd_9])° f_4'°(**paralleld** [loopc 1 p fd_2 d]) °f_0

Для **GNG формальная модель декомпозиции** при вертикальном распределении представляет собой следующую композицию функций:

GNG = f28°
(loopr 1 c (
(parallels[loopc 1 p f26]) °
(parallels[loopq 1 n f24]) °
(parallels[loopn 1 n f22]) ° f20)) °
(parallels[loopn 1 n f18]) °
(parallels[loopn 1 n (parallels[loopq 1 n f16])]) °
f13 °
[paralleld(loopc 1 p (loopn 1 n (loopr 1 c (**fd12**))))] °
(loopr 1 c (f9 ° (parallels[loopn 2 n f8]) ° f6 °)) °
[paralleld(loopc 1 p (loopn 1 n (f5 (loopr 1 c (**fd4**)))))]

В соответствии с предложенными стратегиями, описанными условиями работы и декомпозицией алгоритмов для построения кластеризации на основе нейронных сетей в распределённых системах мониторинга предлагается **методика** (Рисунок 3.4).

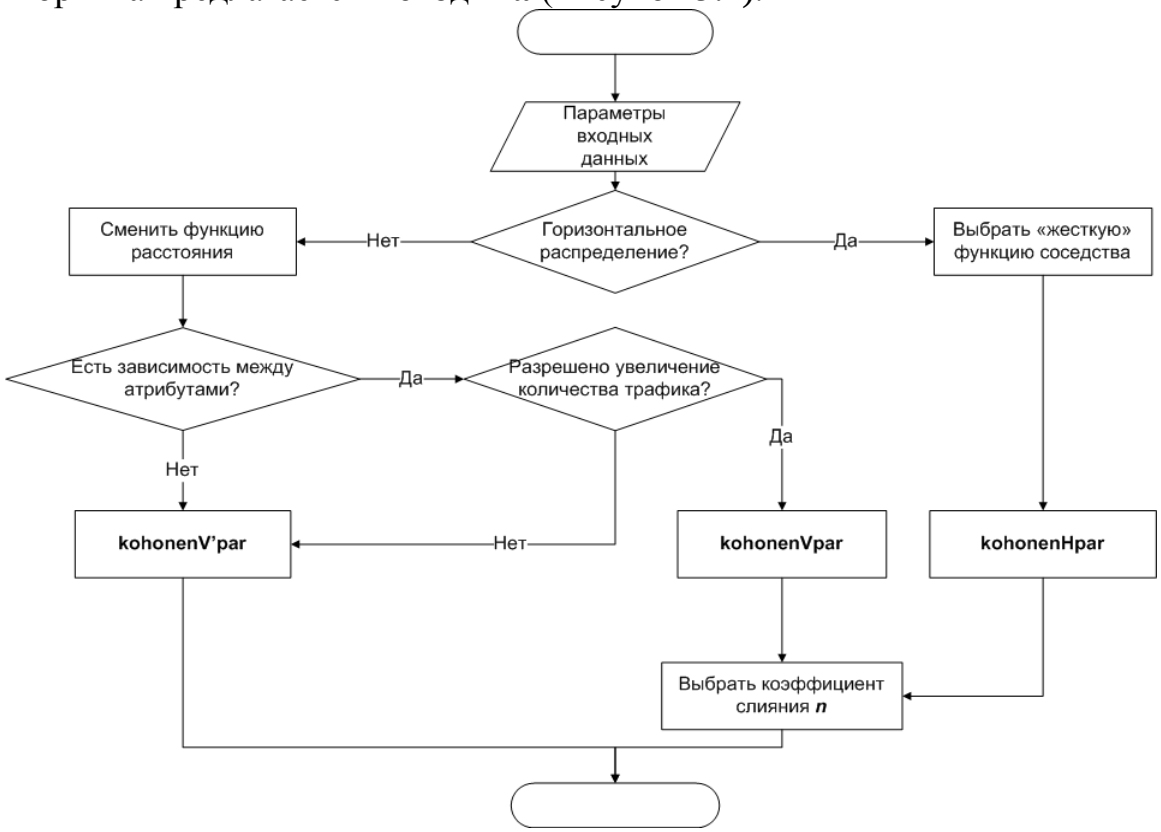


Рисунок 3.4 – Схема построения алгоритма в распределенных системах мониторинга

Методика (Рисунок 3.4) построения алгоритма основывается на параметрах входных данных и предполагает, что конфигурация работы мониторинга будет после предварительного анализа системы и на основе типа распределения данных.

В четвертой главе представлена программная реализация алгоритмов в распределенных системах мониторинга, на основе которой проводится экспериментальная оценка.

Программную реализацию алгоритмов кластеризации, использующие нейронные сети Кохонена в распределенных системах мониторинга можно представить в виде диаграммы (Рисунок 4.1).

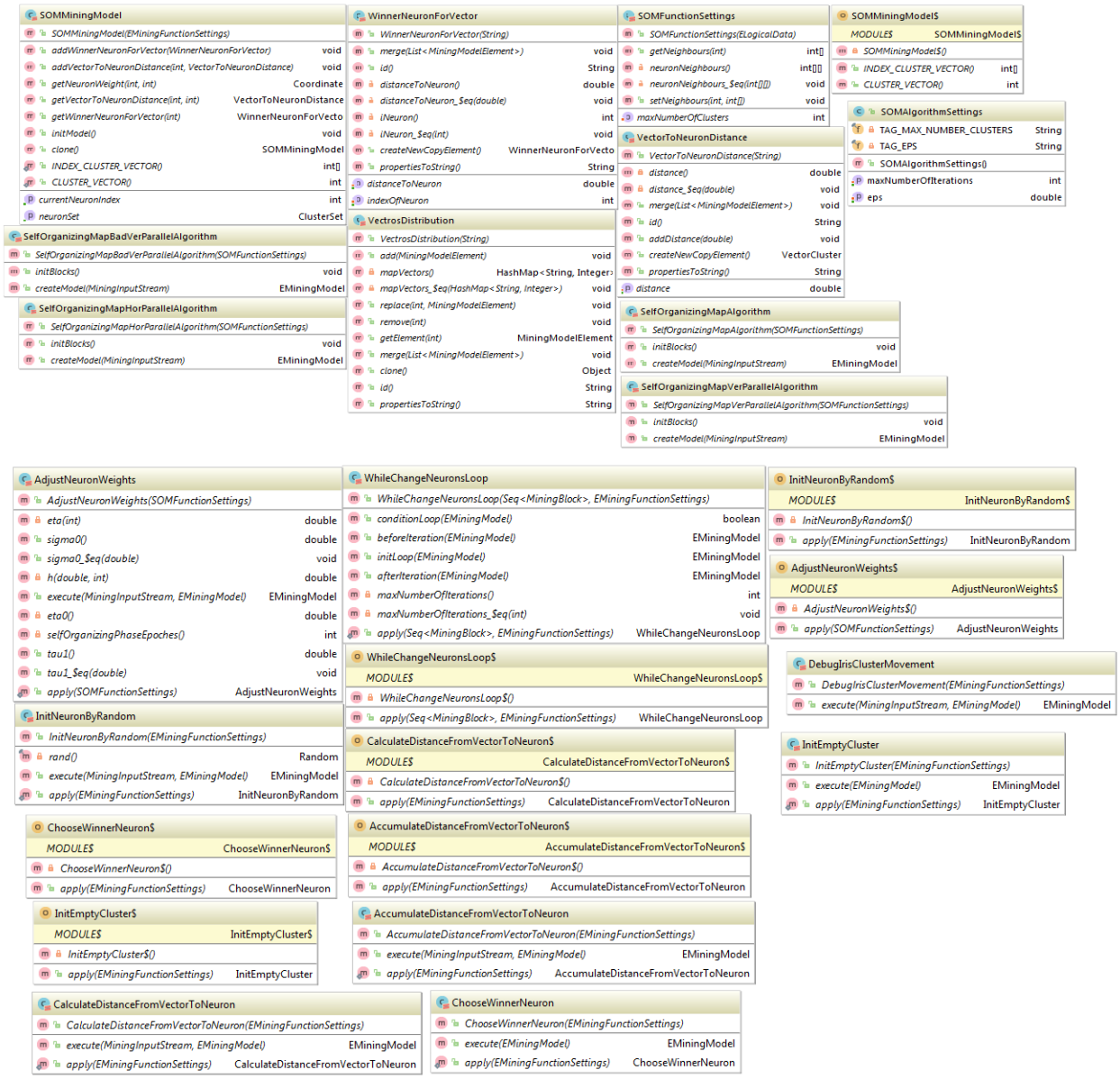
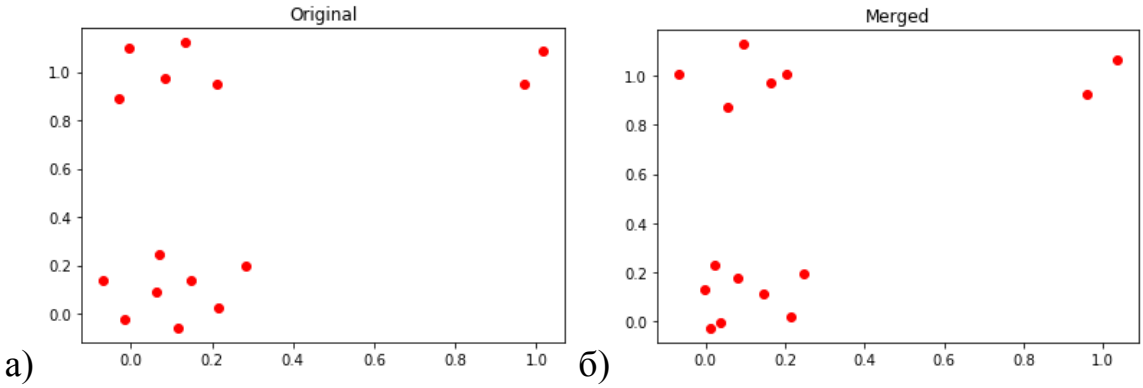


Рисунок 4.1 – Программная реализация блоков алгоритма

Атрибуты данных являются независимыми. Матрица была разбита на 2 и 4 части по строкам (для имитации горизонтального распределения) и по колонкам (для имитации вертикального распределения).

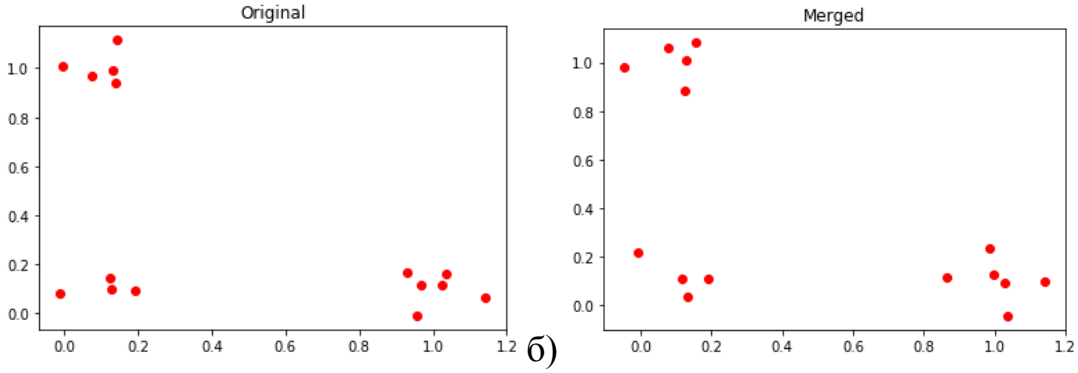
Результаты кластеризации представлены на рисунке 4.2.



а) б)

Рисунок 4.2 – Отображение нейронов при:
а) последовательном выполнении; б) горизонтально распределенном выполнении

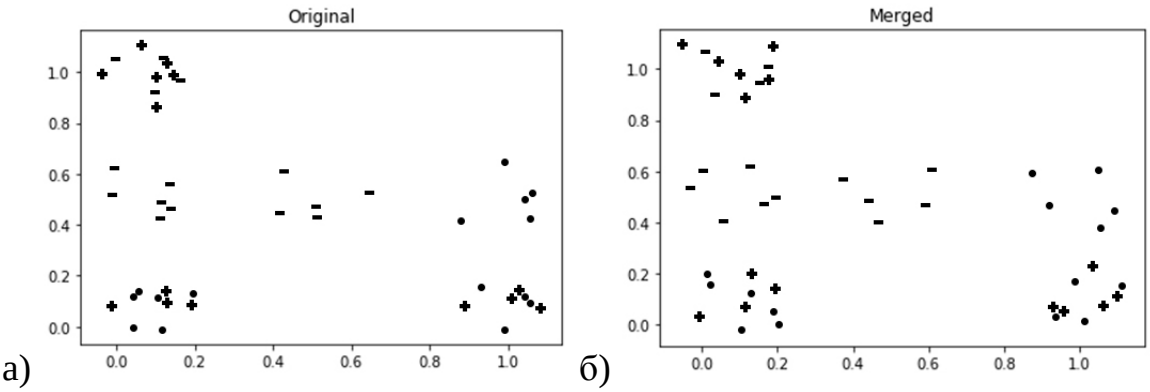
Как видно по рисунку 4.2, оригинальное разбиение кластерных областей и количества нейронов в них сохраняется. На рисунке 4.3 для наглядности представления в двухмерном пространстве представлена проекция нейронов по двум атрибутам.



а) б)

Рисунок 4.3 – Отображение нейронов при:
а) последовательном выполнении; б) вертикально распределенном выполнении

На рисунке 4.4 представлены выборки проекции нейронов попарно разных атрибутов.



а) б)

Рисунок 4.4 – Отображения нейронов разных атрибутов попарно при:
а) последовательном выполнении; б) вертикально распределенном выполнении

Полученные графики показывают, что при анализе на распределенных источниках данных кластера формируются в соответствии с заданными при генерации центрами и незначительно отличаются от результатов при анализе на одном источнике данных.

В заключении сформулированы основные результаты работы:

1. Выполнен анализ существующих средств кластеризации, в том числе распределенных данных показывающий, что основным требованиям для распределенных систем мониторинга соответствуют алгоритмы кластеризации, использующие нейронные сети Кохонена;
2. Разработана формальная модель декомпозиции алгоритмов кластеризации, использующих нейронные сети Кохонена для горизонтально и вертикально распределенных данных;
3. Разработан метод объединения промежуточных результатов полученных при анализе распределенных данных с учетом типа их распределения для алгоритмов кластеризации, использующих нейронные сети Кохонена;
4. Разработана методика выполнения кластеризации, использующая нейронные сети Кохонена на распределенных источниках данных с учетом типа распределения данных;
5. Выполнена программная реализация алгоритма кластеризации, использующий нейронные сетей Кохонена для распределенных данных с учетом метода объединения полученных результатов;
6. Выполнена экспериментальная проверка на основе программной реализации, подтверждающая верность представленного метода объединения и методики выполнения алгоритмов кластеризации, использующие нейронные сети Кохонена в распределенных системах мониторинга.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ ПО ТЕМЕ ДИССЕРТАЦИИ

Статьи, опубликованные в изданиях, включенных в перечень ВАК:

1. Рукавицын А.Н. Кластеризация данных в распределенных системах мониторинга / А. Н. Рукавицын // Information & Control Systems/Informazionno-Upravlyaushie Sistemy. – 2019. – Т. 99. – №. 3. – С. 35-43.

2. Рукавицын А.Н. Разработка модели классификации веб-страниц с использованием методов интеллектуального анализа данных / А. Н. Рукавицын // ИЗВЕСТИЯ СПбГЭТУ «ЛЭТИ». – 2016. – №4. – С. 12-20.

3. Маннанов Э.Р. Повышение качества электрической энергии при резкопеременной нагрузке / Э.Р. Маннанов, А.Н. Рукавицын // ИЗВЕСТИЯ СПбГЭТУ «ЛЭТИ». – 2016. – №2. – С. 55-59.

4. Маннанов Э.Р. Анализ и оптимальный выбор типа гибкой системы электропередачи переменного тока / Э.Р. Маннанов, А.Н. Рукавицын, С.А. Галунин, Т.П. Козулина // ИЗВЕСТИЯ СПбГЭТУ «ЛЭТИ». – 2015. – №9. – С. 50-57.

Статьи, опубликованные в зарубежных изданиях, включённых в системы цитирования Scopus и WebOfScience:

5. Kholod I., Efimova M., Rukavitsyn A., Shorov A. Time Series Distributed Analysis in IoT with ETL and Data Mining technologies, Proceedings of 17th International Conference NEW2AN 2017, 10th Conference ruSMART 2017, Third Workshop NsCC 2017, St. Petersburg, Russia, August 28–30, 2017, pp. 97-108.

6. Kholod I. I., Rukavitsyn A.N., Shorov A.V. Accident predicting method by data from multiple sensors, Proceedings of the 2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering, EIConRus 2018, Saint Petersburg, pp. 318 – 321.

7. Rukavitsyn A., Borisenko K., Shorov A. Self-learning Method for DDoS Detection Model in Cloud Computing, 2017 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering, 2017 ElConRus, Saint Petersburg, 1-3 Feb, 2017.

8. Rukavitsyn A., Kholod I., Shorov A. Neural network training with distributed data, Proceedings of 2017 XX IEEE International Conference on Soft Computing and Measurements, SCM 2018, Saint Petersburg.

9. Rukavitsyn A., Borisenko K., Holod I., Shorov A. The method of ensuring confidentiality and integrity data in cloud computing, Proceedings of 2017 XX IEEE International Conference on Soft Computing and Measurements, SCM 2017, Saint Petersburg, pp. 272-274.

10. Rukavitsyn A.N., Kupriyanov M.S., Shorov A.V., Petukhov I.V. Investigation of website classification methods based on data mining techniques, Proceedings of the 19th International Conference on Soft Computing and Measurements, SCM 2016, Saint Petersburg, 25-27 May, pp. 333 - 336.

11. Borisenko K., Rukavitsyn A., Gurtov A., Shorov A. Detecting the origin of DDoS attacks in OpenStack cloud platform using data mining techniques, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9870 LNCS, pp. 303-315.

Статьи, опубликованные в других изданиях и материалах конференций:

12. А.Н. Рукавицын. Обучение нейронных сетей по распределенным данным / А.Н. Рукавицын [и др.] // XXI международная конференция по мягким вычислениям и измерениям, г. Санкт-Петербург, с 23 по 25 мая 2018 г. – Т. 1. – С. 740-757.

13. А.Н. Рукавицын. Методика обеспечения конфиденциальности и целостности данных в облачных вычислительных средах / А.Н. Рукавицын [и др.] // XX международная конференция по мягким вычислениям и измерениям, г. Санкт-Петербург, с 24 по 26 мая 2017 г. – Т. 2. – С. 412-416.

14. Исследование методов категоризации ВЕБ-страниц на основе методов интеллектуального анализа данных / А.Н. Рукавицын [и др.] // XIX международная конференция по мягким вычислениям и измерениям, г. Санкт-Петербург, с 25 по 27 мая 2016 г. – Т. 2. – С. 333 - 336.

15. Холод И.И. Параллельное выполнения алгоритмов интеллектуального анализа данных с использованием балансировки узлов распределенной сети на основе сетей Петри / Холод И.И., Рукавицын А.Н. //

Свидетельства о государственной регистрации программы для ЭВМ

16. Программа для мультиклассовой мягкой категоризации веб-страниц: Свидетельство о государственной регистрации программы для ЭВМ № 2016615025, 13 мая 2016г.

17. Программа сбора данных для получения количественных характеристик трафика в высоконагруженных компьютерных сетях: Свидетельство о государственной регистрации программы для ЭВМ № 2016614418, 22 апреля 2016г.

18. Программа для предварительного анализа и выбора типа гибких систем электропередачи переменного тока: Свидетельство о государственной регистрации программы для ЭВМ № 2015619931, 22 июля 2015г.

19. Автоматизированная информационная система «Прием заявок типографии»: Свидетельство о государственной регистрации программы для ЭВМ № 2013610379, 19 ноября 2012г.

20. Программа для автоматического резервного копирования и восстановления настроек, образов, виртуальных машин облачной вычислительной среды № 2017663038, 2017г.

21. Программа для обучения и тестирования моделей интеллектуального анализа данных, классифицирующих сетевой трафик для ЭВМ № 2017663039, 2017г.

22. Программа для обеспечения защиты виртуализированных компьютерных сетей облачных вычислительных сред от внешних DDOS-атак: Свидетельство о государственной регистрации программы для ЭВМ № 2016661066, 28 сентября 2016г.