


**ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»**

Факультет компьютерных наук
Департамент программной инженерии

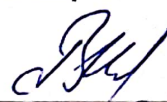
СОГЛАСОВАНО

Научный руководитель,
старший преподаватель департамента
программной инженерии факультета
компьютерных наук


_____ А.В. Меликян
«22» _____ 2019 г.

УТВЕРЖДАЮ

Академический руководитель
образовательной программы
«Программная инженерия»
профессор департамента программной
инженерии, канд. техн. наук


_____ В.В. Шилов
«22» _____ 2019 г.

**ПРОГРАММА ДЛЯ КЛАСТЕРИЗАЦИИ РОССИЙСКИХ ВУЗОВ ПО
ПОКАЗАТЕЛЯМ ИХ НАУЧНО-ОБРАЗОВАТЕЛЬНОЙ ДЕЯТЕЛЬНОСТИ
НА ОСНОВЕ ИЕРАРХИЧЕСКОГО АГЛОМЕРАТИВНОГО МЕТОДА**

Пояснительная записка

ЛИСТ УТВЕРЖДЕНИЯ

RU.17701729.04.13-01 81 01-1-ЛУ

Подп. и дата	
Инв. № дубл.	
Взам. инв. №	
Подп. и дата	
Инв. № подл.	

Исполнитель
студент группы БПИ181
/А.А. Матевосян /
«22» _____ 2019 г.

Москва 2019

УТВЕРЖДЕН
RU.17701729.04.13-01 81 01-1-ЛУ

**ПРОГРАММА ДЛЯ КЛАСТЕРИЗАЦИИ РОССИЙСКИХ ВУЗОВ ПО
ПОКАЗАТЕЛЯМ ИХ НАУЧНО-ОБРАЗОВАТЕЛЬНОЙ ДЕЯТЕЛЬНОСТИ
НА ОСНОВЕ ИЕРАРХИЧЕСКОГО АГЛОМЕРАТИВНОГО МЕТОДА**

Пояснительная записка

RU.17701729.04.13-01 81 01-1

Листов 41

<i>Подп. и дата</i>	
<i>Инв. № дубл.</i>	
<i>Взам. инв. №</i>	
<i>Подп. и дата</i>	
<i>Инв. № подл</i>	

Москва 2019

АННОТАЦИЯ

В данном программном документе приведена пояснительная записка к программе «Cluzterizer» («Программа для кластеризации российских вузов по показателям их научно-образовательной деятельности на основе иерархического агломеративного метода»), предназначенной для выполнения кластерного анализа данных по агломеративному методу.

В разделе «Введение» указано наименование программы, краткое наименование программы и документы, на основании которых ведется разработка.

В разделе «Назначение и область применения» указано функциональное назначение программы, эксплуатационное назначение программы и краткая характеристика области применения программы.

В разделе «Технические характеристики» содержатся следующие подразделы:

- постановка задачи на разработку программы;
- описание алгоритма и функционирования программы с обоснованием выбора схемы алгоритма решения задачи и возможные взаимодействия программы с другими программами;
- описание и обоснование выбора метода организации входных и выходных данных;
- описание и обоснование выбора состава технических и программных средств.

В разделе «Ожидаемые технико-экономические показатели» указана предполагаемая потребность и экономические преимущества разработки по сравнению с отечественными и зарубежными образцами или аналогами

Настоящий документ разработан в соответствии с требованиями:

- 1) ГОСТ 19.101-77 Виды программ и программных документов [1];
 - 2) ГОСТ 19.102-77 Стадии разработки [2];
 - 3) ГОСТ 19.103-77 Обозначения программ и программных документов [3];
 - 4) ГОСТ 19.104-78 Основные надписи [4];
 - 5) ГОСТ 19.105-78 Общие требования к программным документам [5];
 - 6) ГОСТ 19.106-78 Требования к программным документам, выполненным печатным способом [6];
 - 7) ГОСТ 19.404-79 Пояснительная записка. Требования к содержанию и оформлению [7].
- Изменения к Пояснительной записке оформляются согласно ГОСТ 19.603-78 [8], ГОСТ 19.604-78 [9].

Перед прочтением данного документа рекомендуется ознакомиться с терминологией, приведенной в Приложении 1 настоящего технического задания.

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

СОДЕРЖАНИЕ

1.	ВВЕДЕНИЕ	5
1.1.	Наименование программы	5
1.2.	Документы, на основании которых ведется разработка	5
2.	НАЗНАЧЕНИЕ И ОБЛАСТЬ ПРИМЕНЕНИЯ	6
2.1.	Назначение программы	6
2.1.1.	Функциональное назначение	6
2.1.2.	Эксплуатационное назначение	6
2.2.	Краткая характеристика области применения	6
3.	ТЕХНИЧЕСКИЕ ХАРАКТЕРИСТИКИ	8
3.1.	Постановка задачи на разработку программы	8
3.2.	Описание алгоритма и функционирования программы	8
3.2.1.	Описание функционирования программы	8
3.2.2.	Описание алгоритма кластеризации данных	8
3.2.3.	Описание алгоритма расчета рекомендуемого числа кластеров	9
3.2.4.	Описание алгоритма расчета рекомендуемого числа кластеров	10
3.2.5.	Описание алгоритма расчета дистанции между кластерами	10
3.2.6.	Описание алгоритма расчета дистанции между объектами	12
3.2.7.	Описание алгоритма Min-Max нормализации данных	13
3.2.8.	Описание алгоритма Z-Score нормализации данных	13
3.2.9.	Обоснование выбора алгоритма решения задачи	13
3.2.10.	Возможные взаимодействия программы с другими программами	13
3.3.	Описание и обоснование выбора метода организации входных и выходных данных	13
3.2.11.	Описание метода организации входных и выходных данных	13
3.2.12.	Обоснования выбора метода организации входных и выходных данных	14
3.4.	Описание и обоснование выбора состава технических и программных средств	15
3.2.13.	Состав технических и программных средств	15
3.2.14.	Обоснование выбора технических и программных средств	15
4.	ТЕХНИКО-ЭКОНОМИЧЕСКИЕ ПОКАЗАТЕЛИ	16
4.1.	Предполагаемая потребность	16
4.2.	Ориентировочная экономическая эффективность	16
4.3.	Экономические преимущества разработки по сравнению с отечественными и зарубежными аналогами	16
5.	СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ	17
	ПРИЛОЖЕНИЕ 1 - ТЕРМИНОЛОГИЯ	18
	ПРИЛОЖЕНИЕ 2 - ДИАГРАММА ПРЕЦЕДЕНТОВ	19
	ПРИЛОЖЕНИЕ 3 - ПРИМЕР КОНФИГУРАЦИОННОГО ФАЙЛА DATACONFIG.XML	20
	ПРИЛОЖЕНИЕ 4 - ПРИМЕР ВХОДНОГО ФАЙЛА	21
	ПРИЛОЖЕНИЕ 5 - ПРИМЕР ВЫХОДНОГО ФАЙЛА ТАБЛИЦЫ КЛАСТЕРОВ	22

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

ПРИЛОЖЕНИЕ 6 - ПРИМЕР ВЫХОДНОГО ФАЙЛА ОПИСАТЕЛЬНОЙ СТАТИСТИКИ КЛАСТЕРОВ.....	23
ПРИЛОЖЕНИЕ 7 - ПРИМЕР ВЫХОДНОГО ФАЙЛА КАРТИНКИ ДЕНДОГРАММЫ ...	24
ПРИЛОЖЕНИЕ 8 - ДИАГРАММА КЛАССОВ	25
ПРИЛОЖЕНИЕ 9 - ОПИСАНИЕ И ФУНКЦИОНАЛЬНОЕ НАЗНАЧЕНИЕ КЛАССОВ ..	27
ПРИЛОЖЕНИЕ 10 - ОПИСАНИЕ И ФУНКЦИОНАЛЬНОЕ НАЗНАЧЕНИЕ ПОЛЕЙ МЕТОДОВ И СВОЙСТВ	28
ЛИСТ РЕГИСТРАЦИИ ИЗМЕНЕНИЙ.....	40

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

1. ВВЕДЕНИЕ

1.1. Наименование программы

Наименование программы: «Программа для кластеризации российских вузов по показателям их научно-образовательной деятельности на основе иерархического агломеративного метода» («Program for Clustering Russian Universities on their Educational and Research Indicators on the Basis of Agglomerative Hierarchical Method»).

Краткое наименование программы: («Clusterizer»)

1.2. Документы, на основании которых ведется разработка

Разработка «Программа для кластеризации российских вузов по показателям их научно-образовательной деятельности на основе иерархического агломеративного метода» ведется на основании Приказа № 2.3-02/0812-01 от 08.12.16 «Об утверждении тем, руководителей курсовых работ студентов образовательной программы Программная инженерия факультета компьютерных наук».

Разработка выполняется в рамках темы курсовой работы в соответствии с учебным планом подготовки бакалавров Национального исследовательского университета «Высшая школа экономики» по направлению 09.03.04 «Программная инженерия».

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

2. НАЗНАЧЕНИЕ И ОБЛАСТЬ ПРИМЕНЕНИЯ

2.1. Назначение программы

2.1.1. Функциональное назначение

Данная программа является инструментом для кластерного анализа данных используя агломеративный метод иерархической кластеризации. Она обладает следующим функционалом:

- Кластеризация данных по выбранным показателям
- Построения дерева кластеров из исходных данных
- Построение таблицы объектов с описанием кластеров, которым данные объекты принадлежат
- Нормализация данных
- Вывод описательной статистики по полученным кластерам

2.1.2. Эксплуатационное назначение

Кластерный анализ является одним из ведущих направлений в сфере описательной статистики и машинного обучения. И реализации решении задач кластеризации часто используются в быту.

Например, кластерный анализ может использоваться в следующих сферах:

- В качестве метода для классификации различных объектов
- В качестве метода для группирования некоторых объёмных запросов в интернете в более компактные кластеры для дальнейшей обработки
- В качестве метода для нахождения закономерностей в данных
- В качестве метода для генерации подобных данных
- В качестве метода для обработки больших данных и разделение их на категории для дальнейшей обработки

2.2. Краткая характеристика области применения

«Программа для кластеризации российских вузов по показателям их научно-образовательной деятельности на основе иерархического агломеративного метода» - программа, которая является инструментом для кластерного анализа данных.

Классы или концептуально смысловые группы являются объектами, которые разделяют общие характеристики, что играет важную роль в том, как люди понимают и описывают окружающий мир. С самого рождения мы обучены тому, чтобы разделять объекты на некоторые группы. В контексте понимания данных кластеры потенциальные классы, а кластерный анализ является техникой для автоматического нахождения этих классов.

Кластерный анализ и ее различные реализации могут быть использованы в разных областях для решения задач, таких как:

1. Классификация различных биологических видов, элементов

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

2. В разбиении результатов по запросу в поисковых системах (Google, Yahoo, Yandex...)
3. В различных исследованиях для выявления закономерностей
4. В нахождении вариации различных болезней в медицине
5. В бизнес среде, для обработки большого объёма данных

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

3. ТЕХНИЧЕСКИЕ ХАРАКТЕРИСТИКИ

3.1. Постановка задачи на разработку программы

Программа должна решать следующие задачи:

- Рассчитать число рекомендуемых кластеров по индексу Цалиньски-Харабаша
- Выполнить кластеризацию по выбранным параметрам
- Построить дендограмму
- Построить таблицу кластеров и таблицу описательной статистики

3.2. Описание алгоритма и функционирования программы

3.2.1. Описание функционирования программы

Программа выполняет все функциональные блоки диаграммы вариации (см. Приложении 2)

3.2.2. Описание алгоритма кластеризации данных

В качестве основы был использован алгоритм AGNES (Agglomerative Nesting Hierarchical Clustering см. [17] Глава 5, [15]).

Из данных таблицы создаются одиночные(синглтон) кластеры, т. е. кластеры, в которых нет вложенных кластеров и добавляются в список кластеров.

```
1. private void BuildDissimilarityMatrix()
2. {
3.     _dissimilarityMatrix = new DissimilarityMatrix();
4.
5.     for (int i = 0; i < _clusters.Count - 1; i++)
6.     {
7.         for (int j = i + 1; j < _clusters.Count; j++)
8.         {
9.             var clusterPair = new ClusterPair(_clusters.GetCluster(i),
10. _clusters.GetCluster(j));
11.
12.             var distanceBetweenTwoClusters = ClusterDistance.ComputeDistance(
13. clusterPair.Cluster1, clusterPair.Cluster2, _distanceMetric);
14.             _dissimilarityMatrix.AddClusterPairAndDistance(clusterPair,
15. distanceBetweenTwoClusters); // adds distance to matrix
16.         }
17.     }
18. }
```

Создается матрица различия, которая представляет собой структуру данных для хранения расстояний между пар кластеров. После этого выполняется иерархическая кластеризация пока число кластеров не будет совпадать с заданным числом кластеров [16].

```
1. public ClusterSet ExecuteClustering(int k, bool isWithIndex = false)
2. {
3.     BuildHierarchicalClustering(_clusters.Count, k, isWithIndex);
4.     return _clusters;
5. }
```

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

В каждом шагу итерации алгоритма методом подбора всех всевозможных пар кластеров выбирается пара кластеров с наименьшей дистанцией. Создаётся новый кластер путем объединения пары кластеров. Старые кластеры удаляются из списка кластеров, обновляется матрица различия по ново созданному кластеру.

```

1. private void BuildHierarchicalClustering(int indexNewCluster, int k, bool isWithIndex =
   false)
2. {
3.     ClusterPair closestClusterPair = _dissimilarityMatrix.GetClosestClusterPair();
4.     // gets the clusterpair with minimal distance
5.
6.     // creates new cluster by merging clusters from closestClusterPair
7.     Cluster newCluster = new Cluster();
8.     newCluster.AddSubCluster(closestClusterPair.Cluster1);
9.     newCluster.AddSubCluster(closestClusterPair.Cluster2);
10.    newCluster.Id = indexNewCluster;
11.    newCluster.SetCentroid();
12.
13.    // removes cluster pair from _clusters
14.    _clusters.RemoveClusterPair(closestClusterPair);
15.    _updateDissimilarityMatrix(newCluster);
16.    // add new cluster to _clusters
17.    _clusters.AddCluster(newCluster);
18.
19.    if (isWithIndex) // checks is executed for calculating CH index
20.    {
21.        _chValue.Add(GetCHIndex()); // adds index to array of CH values
22.        _chIndex.Add(_clusters.ClustersList.Count);
23.        // adds number of clusters for current CH value
24.    }
25.
26.    // exit point of algorithm (Where _clusters count is equal to k)
27.    if (_clusters.Count > k)
28.        BuildHierarchicalClustering(indexNewCluster + 1, k, isWithIndex);
29. }

```

3.2.3. Описание алгоритма расчета рекомендуемого числа кластеров

Для расчета рекомендуемого числа кластеров выполняется кластеризация данных описанный выше, но при этом при каждом итерации, программа считает индекс Цалиньски-Харабаша(далее СН) и останавливается, когда число кластеров равна двум. При этом значения индекса и текущего числа кластеров добавляются в списки.

После этого ищется локальный максимум в множестве точек из значений СН.

```

1. public int GetRecommendedCountOfClusters()
2. {
3.     int maxIndex = _initialNumberOfClusters - 1; // index of Local Max CH
4.     double maxCoeff = 0; // local Max CH
5.
6.     // finds local Max of CH Values
7.     for (int i = 1; i < _chValue.Count - 1; i++)
8.     {
9.         if (_chValue[i] > _chValue[i -
10.        1] && _chValue[i] > _chValue[i + 1] && _chValue[i] > maxCoeff)
11.        {
12.            maxCoeff = _chValue[i];
13.            maxIndex = _chIndex[i];
14.        }
15.    }
16. }

```

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

```
16.     return maxIndex;
17. }
```

3.2.4. Описание алгоритма расчета рекомендуемого числа кластеров

CH индекс вычисляется по следующей математической формуле:

$$\frac{SS_B}{SS_w} \times \frac{N - k}{k - 1}$$

k – текущее число кластеров

N – общее число объектов

SS_w – общая сумма квадратичного отклонения в нутри кластера

SS_B – общая сумма квадратичного отклонения кластера

```
1. private double GetCHIndex()
2. {
3.     // merging all clusters into one
4.     Cluster overallCluster = new Cluster();
5.     _clusters.ClustersList.ForEach(c => overallCluster.AddSubCluster(c));
6.     overallCluster.SetCentroid();
7.
8.     int currentNumberOfClusters = _clusters.Count;
9.     if (_clusters.ClustersList.Count < 2) // CH can't be computed for one cluster
10.        return double.NaN;
11.
12.     double withinSumOfSquares = 0,
13.        betweenSumOfSquares = 0;
14.
15.     foreach (var cluster in _clusters.ClustersList)
16.     {
17.         // computes sum of squares within cluster
18.         withinSumOfSquares += cluster.GetSumOfSquaredError(_distanceMetric);
19.         // computes som of squares with overallcluster (outside of cluster)
20.         betweenSumOfSquares += Math.Pow(Distance.GetDistance(overallCluster.Centroid, c
21. luster.Centroid, _distanceMetric), 2);
22.     }
23.     // checks if withinSumOfSquares is less then epsilon (CH is NaN)
24.     // else returns CH using formula
25.     return Math.Abs(withinSumOfSquares) < double.Epsilon
26.         ? double.NaN
27.         : (betweenSumOfSquares / withinSumOfSquares / (currentNumberOfClusters -
28. 1)) *
29.         (_initialNumberOfClusters - currentNumberOfClusters);
30. }
```

3.2.5. Описание алгоритма расчета дистанции между кластерами

Для объединения кластеров используются данные стратегии объединения:

- Одиночная связь – расстояние между двумя кластерами является наименьшее расстояние двух объектов, принадлежащих двум разным кластерам.
- Полная связь - расстояние между двумя кластерами является наибольшее расстояние двух объектов, принадлежащих двум разным кластерам.

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

- Невзвешенное попарное среднее – расстояние между двумя кластерами вычисляется, как среднее расстояние из всевозможных пар объектов принадлежавшим двум разным кластерам.
- Взвешенное попарное среднее – вычисляется так же, как и Невзвешенное попарное среднее с добавлением весового коэффициента из числа объектов кластеров.
- Метод центроидов – расстояние между двумя кластерами вычисляется как расстояние между их центроидами.
- Метод Варда – расстояние между двумя кластерами вычисляется как сочетание суммы квадратических отклонений кластеров из суммы квадратической ошибки объединённого кластера.

```

1. public static double ComputeDistance(Cluster cluster1, Cluster cluster2, DissimilarityM
   atriX dissimilarityMatrix, MergeStrategy strategy)
2. {
3.     double distance = 0;
4.     var distance1 = dissimilarityMatrix.ReturnClusterPairDistance(new ClusterPair(clust
   er1, cluster2.GetSubCluster(0)));
5.     var distance2 = dissimilarityMatrix.ReturnClusterPairDistance(new ClusterPair(clust
   er1, cluster2.GetSubCluster(1)));
6.
7.     // computes distance by using merge strategy
8.     switch (strategy)
9.     {
10.         case MergeStrategy.SingleLinkage:
11.             distance = _MinValue(distance1, distance2); // Min(x, y)
12.             break;
13.         case MergeStrategy.CompleteLinkage:
14.             distance = _MaxValue(distance1, distance2); // Max(x, y)
15.             break;
16.         case MergeStrategy.AveragelinkageWpma:
17.             distance = (distance1 + distance2) / 2; // Avg(x, y)
18.             break;
19.         case MergeStrategy.AveragelinkageUpma:
20.             distance = ((cluster2.GetSubCluster(0).QuantityOfDataPoints * distance1) /
   cluster2.QuantityOfDataPoints)
21.                 + ((cluster2.GetSubCluster(1).QuantityOfDataPoints * distance2)
   / cluster2.QuantityOfDataPoints); // WeightedAvg(x, y)
22.             break;
23.         case MergeStrategy.CentroidMethod:
24.             cluster1.SetCentroid();
25.             cluster2.SetCentroid();
26.             distance = Distance.GetDistance(cluster1.Centroid, cluster2.Centroid,
   DistanceMetric.SquareEuclidianDistance); // Distance of centroids
27.             break;
28.         case MergeStrategy.WardsMethod:
29.
30.             Cluster newCluster = new Cluster();
31.             newCluster.AddSubCluster(cluster1);
32.             newCluster.AddSubCluster(cluster2);
33.             newCluster.SetCentroid();
34.
35.             distance = newCluster.GetSumOfSquaredError(DistanceMetric.EuclidianDistance
   )
36.                 -
37.                 cluster1.GetSumOfSquaredError(DistanceMetric.EuclidianDistance)
38.                 -
39.                 cluster2.GetSumOfSquaredError(DistanceMetric.EuclidianDistance);
40.             // SEO(xy) - SEO(x) - SEO(y)
41.             break;
42.     }

```

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

```
43.     return distance;
44.
45. }
```

3.2.6. Описание алгоритма расчета дистанции между объектами

Для расчета расстояния между двумя объектами используются следующие меры расстояния:

- Евклидово расстояние – геометрическое расстояние в многомерном пространстве
- Квадрат евклидова расстояния – квадрат геометрического расстояния в многомерном пространстве
- Манхэттенское расстояние – средняя разность по координатам
- Расстояние Чебышева – максимальный модуль разности по координатам

```
1. public static double GetDistance(DataPoint x, DataPoint y, DistanceMetric distanceMetric)
2. {
3.     double distance = 0;
4.     double diff;
5.
6.     // checks for dimensions match
7.     if (x.Count != y.Count)
8.         throw new ArgumentException("Неравное количество точек.");
9.
10.    switch (distanceMetric)
11.    {
12.        case DistanceMetric.EuclidianDistance: //calculates by using Euclidian Distance
13.            for (var i = 0; i < x.Count; i++)
14.            {
15.                diff = x[i] - y[i];
16.                distance += diff * diff;
17.            }
18.
19.            distance = Math.Sqrt(distance);
20.            break;
21.        case DistanceMetric.SquareEuclidianDistance: // calculates by using Square of Euclidian Distance
22.            for (var i = 0; i < x.Count; i++)
23.            {
24.                diff = x[i] - y[i];
25.                distance += diff * diff;
26.            }
27.
28.            break;
29.        case DistanceMetric.ManhattanDistance: // calculates by using Manhattan Distance
30.            for (var i = 0; i < x.Count; i++)
31.            {
32.                diff = x[i] - y[i];
33.                distance += Math.Abs(diff);
34.            }
35.
36.            break;
37.        case DistanceMetric.ChebyshevDistance: // calculates by using Chebyshev Distance
38.            for (var i = 0; i < x.Count; i++)
39.            {
40.                diff = Math.Abs(x[i] - y[i]);
41.                distance = distance > diff ? distance : diff;
42.            }
43.
44.            break;
```

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

```
45.     }
46.
47.     return distance;
48. }
```

3.2.7. Описание алгоритма Min-Max нормализации данных

Вычисляется максимальное и минимальное значение в множестве точек. Потом от каждой точки вычитается минимальный элемент и делится на разность максимального и минимального значения [15].

```
1. public static void MinMaxNormalize(ref double[] arr)
2. {
3.     var max = arr.Max();
4.     var min = arr.Min();
5.
6.     for (var i = 0; i < arr.Length; i++)
7.         arr[i] = (arr[i] - min) / (max - min);
8. }
```

3.2.8. Описание алгоритма Z-Score нормализации данных

Вычисляется среднее значения множества точек. Из каждой точки множества вычитается среднее значение, и точка делится на среднеквадратическое отклонение.

Среднеквадратическое отклонение считается как сумма квадратов разности точки и среднего значения [15].

```
1. public static void ZScoreNormalize(ref double[] arr)
2. {
3.     var mean = arr.Sum() / arr.Length;
4.     double bigSum = 0;
5.     foreach (var d in arr) bigSum += Math.Pow(d - mean, 2);
6.
7.     var standartDeviation = Math.Sqrt(bigSum / (arr.Length - 1));
8.
9.     for (var i = 0; i < arr.Length; i++)
10.         arr[i] = (arr[i] - mean) / standartDeviation;
11. }
```

3.2.9. Обоснование выбора алгоритма решения задачи

Сама по себе программа использует агломеративный метод иерархической кластеризации данных. Выбор был обусловлен функциональными требованиями программы.

3.2.10. Возможные взаимодействия программы с другими программами

В целом программа работает самостоятельно. Возможно употребление текстовых или других редакторов для изменения входных и выходных данных. Также понадобится программа для просмотра изображений, для просмотра дендограмм.

3.3. Описание и обоснование выбора метода организации входных и выходных данных

3.2.11. Описание метода организации входных и выходных данных

В качестве входных данных принимаются файлы формата CSV ([14]) в специальном формате(см. пример в Приложение 4), которая задается через конфигурационный XML файл dataconfig.xml(см. пример в Приложение 3).

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

В dataconfig.xml содержатся следующие элементы:

- StringHeadings – строковой массив, элементы которого задают описание объекта кластеризации (далее строковые данные). Т. е. элементы, по которым не будет проведена кластеризация. Например, такие характеристики как название объекта, местоположение объекта, тип объекта.
- NumericHeadings – строковой массив, элементы которого задают имена характеристик объекта по которым может быть проведена кластеризация (далее показатели).
- GroupNames – строковой массив, элементы которого задают имена групп по которым разделяются показатели.
- GroupItemsCount – массив натуральных чисел, элементы которого задают количество показателей в каждой группе.

При этом каждый из элементов не может быть пустым и сумма элементов GroupItemsCount должна равна сумме количества элементов StringHeadings и NumericHeadings. В качестве первого элемента StringHeadings указывается обусловленное название объекта (далее название кластера).

В качестве выходных данных программы выступают:

- Открытый программой входной файл в формате заданном выше
- Таблица кластеров программы, сохраняемая в формате CSV (см. Приложение 5), которая состоит из трех столбцов: название кластера, изначальный кластер и выходной кластер.
- Таблица описательной статистик сохраняемая в формате CSV (см. Приложение 6), показывающая среднее значение показателей в выходном кластере, а также количество элементов в кластере. Столбцами таблицы являются показатели и количество кластеров.
- Картинка дендограммы сохраняемая в формате PNG (см. Приложение 7)

3.2.12. Обоснования выбора метода организации входных и выходных данных

Была реализовано возможность создания конфигурационного файла, так как нет конкретного формата данных, есть бесконечное множество вариации структуры данных.

При этом программа работает не с неизменяемым форматом данных, а наоборот.

Формат хранения CSV является распространённым форматом для хранения таблицы данных и может быть использован во многих программах.

Выходные файлы программы могут использоваться пользователем для дальнейшей обработки данных.

Также пользователь может посмотреть диаграмму любой программой для просмотра фотографии в формате PNG.

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

3.4. Описание и обоснование выбора состава технических и программных средств

3.2.13. Состав технических и программных средств

Для надёжной и бесперебойной работы программы требуется следующий состав технических средств [19]:

- 1) персональный компьютер, оснащенный 32-разрядным (x86) или 64-разрядным (x64) процессором с тактовой частотой 1 ГГц и выше
- 2) 1 ГБ для x86 и 2 ГБ для x64 оперативной памяти или больше
- 3) не менее 16 ГБ для x86 и 20 ГБ для x64 свободного места на жестком диске
- 4) видеокарта и монитор с разрешением не менее чем 1366x768 точек
- 5) мышь или совместимое указывающее устройство
- 6) клавиатура

Для работы программы необходим следующий состав программных средств:

- 1) операционная система Microsoft Windows 7 SP1 или более поздняя версия;
- 2) установленный Microsoft .NET Framework 4.7.1, требующий Windows Installer 5.0 или более поздняя версия

3.2.14. Обоснование выбора технических и программных средств

В момент написания курсовой работы последней версией .NET Framework был .NET Framework 4.7.1.

Windows Installer 5.0 нужно для установки программы.

Для работы программы была выбрана операционная система Windows 7 SP1 с минимальными техническими характеристиками, что являются минимальными требованиями для работы .NET Framework 4.7.1 [18].

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

4. ТЕХНИКО-ЭКОНОМИЧЕСКИЕ ПОКАЗАТЕЛИ

4.1. Предполагаемая потребность

Кластерный анализ является одним из востребованных на сегодняшний день направлений для статической обработки данных и машинного обучения. Его могут использовать каждый, кому нужно будет кластеризовать большой объём данных.

4.2. Ориентировочная экономическая эффективность

В рамках данной работы расчёт экономической эффективности не предусмотрен.

4.3. Экономические преимущества разработки по сравнению с отечественными и зарубежными аналогами

Существует много алгоритмов кластеризации, и каждый из них по-своему уникален, нет хорошего или плохого. Чаше всего используются готовые библиотеки кластеризации, которые настраиваются под конкретные данные. Так же для кластеризации используют платную программу IBM SPSS [16].

Данная программа может кластеризовать любые данные, соответствующие заданной конфигурации. Она легкая в использовании, не потребляет много ресурсов и распространяется бесплатно.

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

5. СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

- 1) ГОСТ 19.101-77 Виды программ и программных документов. //Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
- 2) ГОСТ 19.102-77 Стадии разработки. //Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
- 3) ГОСТ 19.103-77 Обозначения программ и программных документов. //Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
- 4) ГОСТ 19.104-78 Основные надписи. //Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
- 5) ГОСТ 19.105-78 Общие требования к программным документам. //Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
- 6) ГОСТ 19.106-78 Требования к программным документам, выполненным печатным способом. //Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
- 7) ГОСТ 19.404-79 Пояснительная записка. Требования к содержанию и оформлению. //Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
- 8) ГОСТ 19.603-78 Общие правила внесения изменений. //Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
- 9) ГОСТ 19.604-78 Правила внесения изменений в программные документы, выполненные печатным способом. //Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
- 10) Жамбю М. Иерархический кластер-анализ и соответствия. — М.: Финансы и статистика, 1988. — 345 с.
- 11) Мандель И. Д. Кластерный анализ. — М.: Финансы и статистика, 1988. — 176 с.
- 12) Орлов А. И. Прикладная статистика. Учебник. — М.: Экзамен, 2006. — 671 с.
- 13) Шрейдер Ю. А. Что такое расстояние? — М.: Физматлит, 1963. — 76 с.
- 14) Common Format and MIME Type for Comma-Separated Values (CSV) Files. [Электронный ресурс] / SolidMatrix Technologies, Inc.: <https://tools.ietf.org/html/rfc4180>, свободный(дата обращения: 19.04.2019).
- 15) Finding groups in data with Agglomerative Clustering [Электронный ресурс] / Codeproject. Режим доступа: <https://www.codeproject.com/Articles/1120804/Finding-groups-in-data-with-Csharp-Agglomerative-C>, свободный(дата обращения: 19.04.2019).
- 16) IBM SPSS Software - РФ. [Электронный ресурс] / IBM. Режим доступа: <https://www.ibm.com/ru-ru/analytics/spss-statistics-software>, свободный(дата обращения: 19.04.2019).
- 17) Kaufman L., Rousseeuw P.J. Finding Groups in Data: an introduction to cluster analysis / L. Kaufman, P.J. Rousseeuw. — Wiley, 1990 – 368 с.
- 18) The .NET Framework 4.7.1 offline installer for Windows– Windows Help. [Электронный ресурс] / Microsoft. Режим доступа: <https://support.microsoft.com/en-us/help/4033342/the-net-framework-4-7-1-offline-installer-for-windows>, свободный(дата обращения: 19.04.2019).
- 19) Windows 7 system requirements – Windows Help. [Электронный ресурс] / Microsoft. Режим доступа: <https://support.microsoft.com/en-us/help/10737/windows-7-system-requirements>, свободный(дата обращения: 19.04.2019).

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

ТЕРМИНОЛОГИЯ

Ниже приведен список необходимых терминов для ознакомления [10], [11], [12], [13].

Агломеративный метод – один из методов иерархической кластеризации, в котором создание новых кластеров выполняется путем объединения малочисленных кластеров в более крупные кластеры, таким образом дерево созданным методом имеет направление от листьев к стволу, которая называется деревом кластеров.

Дендрограмма – совокупность древовидных диаграмм дерева кластеров.

Иерархические алгоритм – группа алгоритмов кластеризации, которая упорядочивает данные путем создания иерархии(дерева) вложенных кластеров.

Индекс Цалиньски Харабаша – критерия для оценки обусловленного качества выполненной кластеризации.

Кластер — группа однородных объектов.

Кластеризация (или кластерный анализ) — задача группирования множества объектов так, чтобы объекты, которые принадлежат одной группе были более похожими(однородными), а объекты разных групп должны максимально быть различны. Сама кластеризация не является алгоритмом, а общей задачей для решения.

Матрица различия — матрица в котором хранятся значения расстояния между двумя кластерам.

Мера расстояния — метрика, которая описывает расстояние между двумя объектами.

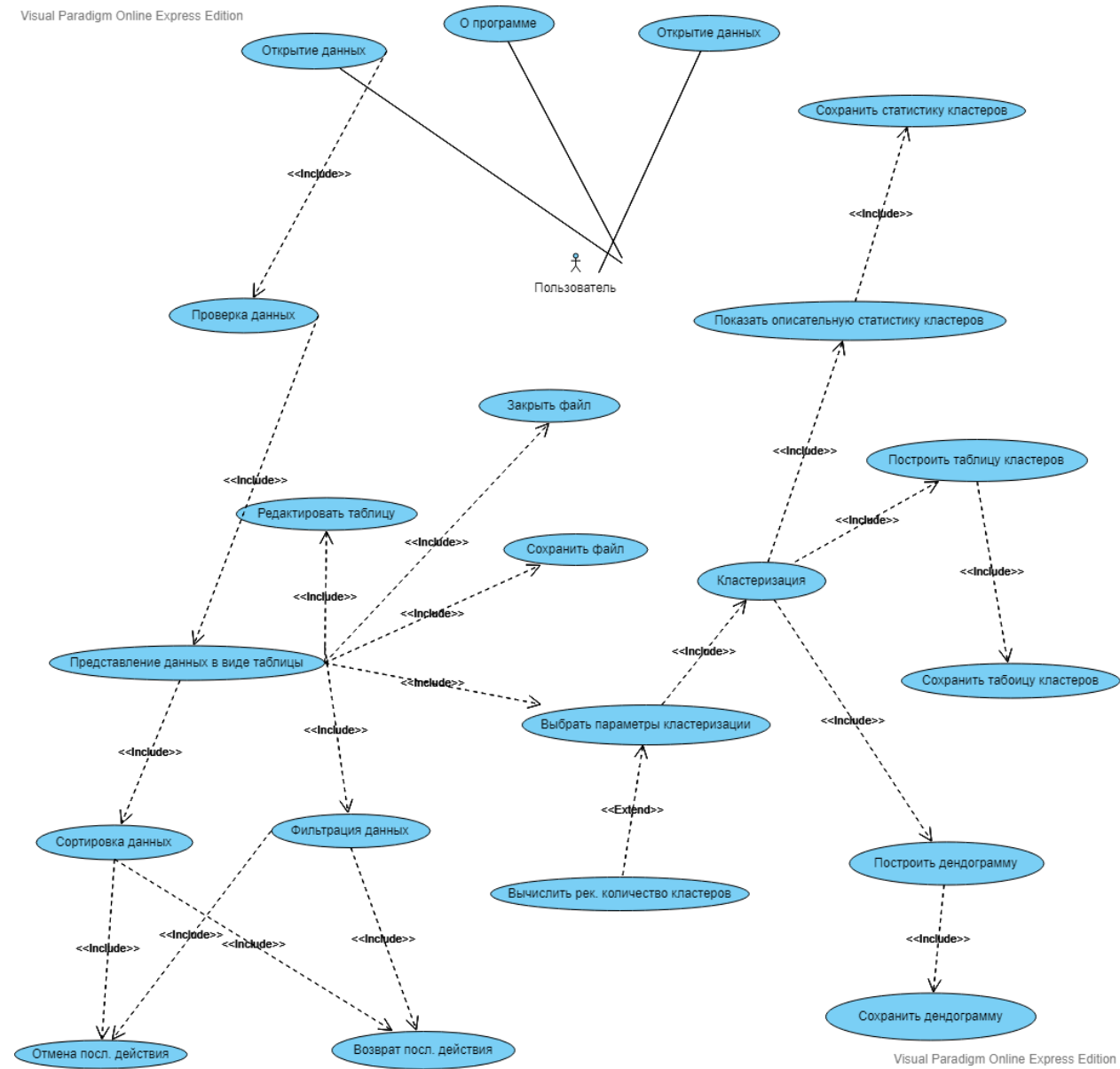
Стратегия объединения — алгоритм объединения двух кластеров.

Сумма квадратического отклонения кластера – сумма квадратов расстояния объектов кластера от его центроида.

Центроид – центр тяжести кластера. Представляет собой кластер, значения точек(характеристик) которой равны среднему значению по каждой точке.

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

ДИАГРАММА ПРЕЦЕДЕНТОВ



Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

ПРИМЕР КОНФИГУРАЦИОННОГО ФАЙЛА dataconfig.xml

```
<?xml version="1.0"?>
<Configuration xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:xsd="http://www.w3.org/2001/XMLSchema">
  <StringHeadings>
    <string>Название вуза</string>
    <string>Округ</string>
    <string>Субъект РФ</string>
    <string>Город</string>
    <string>Ведомственная принадлежность</string>
    <string>Профиль организации</string>
  </StringHeadings>
  <NumericHeadings>
    <string>Удельный вес выпускников, трудоустроившихся в течение календарного года,
    следующего за годом выпуска, в общей численности выпускников образовательной
    организации обучавшихся по основным образовательным программам высшего
    образования</string>
    <string>Удельный вес НПП, имеющих ученую степень кандидата наук, в общей
    численности НПП</string>
    <string>Удельный вес НПП имеющих ученую степень доктора наук, в общей
    численности НПП</string>
    <string>Удельный вес НПП, имеющих ученую степень кандидата и доктора наук, в
    общей численности НПП образовательной организации (без совместителей и работающих
    по договорам гражданско-правового характера)</string>
    <string>Число НПП, имеющих ученую степень кандидата и доктора наук, в расчете на
    100 студентов</string>
    <string>Доля штатных работников ППС в общей численности ППС</string>
  </NumericHeadings>
  <GroupNames>
    <string>Трудоустройство</string>
    <string>Кадровый состав</string>
  </GroupNames>
  <GroupItemsCount>
    <int>1</int>
    <int>5</int>
  </GroupItemsCount>
</Configuration>
```

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

ПРИЛОЖЕНИЕ 4**ПРИМЕР ВХОДНОГО ФАЙЛА**

Российский новый университет;1;1;Москва;1;1;0;50.15;21.49;67.78;1.25;79.55
 Адыгейский государственный университет;6;60;Майкоп;5;1;0;74.19;15.9;90.93;6.77;87.56
 Алтайский государственный медицинский университет Министерства здравоохранения
 Российской Федерации;3;9;Барнаул;9;4;0;57.59;22.1;78.7;9.59;80.08
 Алтайский государственный технический университет им. И.И.
 Ползунова;3;9;Барнаул;5;1;75;59.64;10.95;70.39;4.55;93.35
 Алтайский государственный университет;3;9;Барнаул;5;1;0;64.24;16.72;82.11;4.81;87.11
 Астраханский государственный медицинский университет Министерства
 здравоохранения Российской Федерации;6;57;Астрахань;9;4;65;51.65;19.65;70.47;8.77;75.5
 Астраханский государственный технический
 университет;6;57;Астрахань;20;1;70;59.07;14.14;74.07;4.05;76.52
 Астраханский государственный
 университет;6;57;Астрахань;5;1;70;60.14;15.3;75.32;4.64;77.32
 Балтийский федеральный университет имени Иммануила
 Канта;5;49;Калининград;5;1;70;48.67;14.96;66.87;5.77;82.86
 Башкирский государственный медицинский университет Министерства здравоохранения
 Российской Федерации;9;77;Уфа;9;4;85;64.9;26.98;91.37;10.35;76.09
 Башкирский государственный университет;9;77;Уфа;5;1;75;63.63;23.03;84.11;5.07;85.7
 Белгородский государственный национальный исследовательский
 университет;1;30;Белгород;5;1;0;60.21;16.07;78.52;5.44;83.07
 Белгородский государственный технологический университет им. В.Г.
 Шухова;1;30;Белгород;5;1;75;58.89;17.07;70.37;3.93;87.32
 Благовещенский государственный педагогический
 университет;4;21;Благовещенск;5;1;75;70.73;9.24;80.43;4.07;92.93
 Брянский государственный аграрный университет;1;31;село
 Кокино;6;3;65;63.4;19.79;82.83;3.35;96.94
 Владивостокский государственный университет экономики и
 сервиса;4;25;Владивосток;5;1;65;62.57;9.34;67.72;3.15;83.53
 Владимирский государственный университет имени Александра Григорьевича и Николая
 Григорьевича Столетовых;1;32;Владимир;5;1;80;58.27;13.28;72.48;3.84;84.04
 Волгоградский государственный медицинский университет Министерства
 здравоохранения Российской
 Федерации;6;58;Волгоград;9;4;80;53.47;15.42;71.44;11.69;72.4

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

ПРИЛОЖЕНИЕ 5**ПРИМЕР ВЫХОДНОГО ФАЙЛА ТАБЛИЦЫ КЛАСТЕРОВ**

Адыгейский государственный университет;Cluster1;Cluster26
 Благовещенский государственный педагогический университет;Cluster13;Cluster26
 Российский новый университет;Cluster0;Cluster29
 Алтайский государственный медицинский университет Министерства здравоохранения
 Российской Федерации;Cluster2;Cluster29
 Алтайский государственный университет;Cluster4;Cluster29
 Белгородский государственный национальный исследовательский
 университет;Cluster11;Cluster29
 Астраханский государственный медицинский университет Министерства
 здравоохранения Российской Федерации;Cluster5;Cluster32
 Балтийский федеральный университет имени Иммануила Канта;Cluster8;Cluster32
 Астраханский государственный технический университет;Cluster6;Cluster32
 Астраханский государственный университет;Cluster7;Cluster32
 Алтайский государственный технический университет им. И.И.
 Ползунова;Cluster3;Cluster32
 Владивостокский государственный университет экономики и сервиса;Cluster15;Cluster32
 Белгородский государственный технологический университет им. В.Г.
 Шухова;Cluster12;Cluster32
 Владимирский государственный университет имени Александра Григорьевича и Николая
 Григорьевича Столетовых;Cluster16;Cluster32
 Башкирский государственный университет;Cluster10;Cluster32
 Брянский государственный аграрный университет;Cluster14;Cluster32
 Башкирский государственный медицинский университет Министерства здравоохранения
 Российской Федерации;Cluster9;Cluster32
 Волгоградский государственный медицинский университет Министерства
 здравоохранения Российской Федерации;Cluster17;Cluster32

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

ПРИЛОЖЕНИЕ 6

ПРИМЕР ВЫХОДНОГО ФАЙЛА ОПИСАТЕЛЬНОЙ СТАТИСТИКИ КЛАСТЕРОВ

Cluster26;37.5;72.46;12.57;85.68;5.42;90.245;2

Cluster29;0;58.0475;19.095;76.7775;5.2725;82.4525;4

Cluster32;72.9166666666667;58.6916666666667;16.6591666666667;74.7866666666667;5.763
3333333333;82.6308333333333;12

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

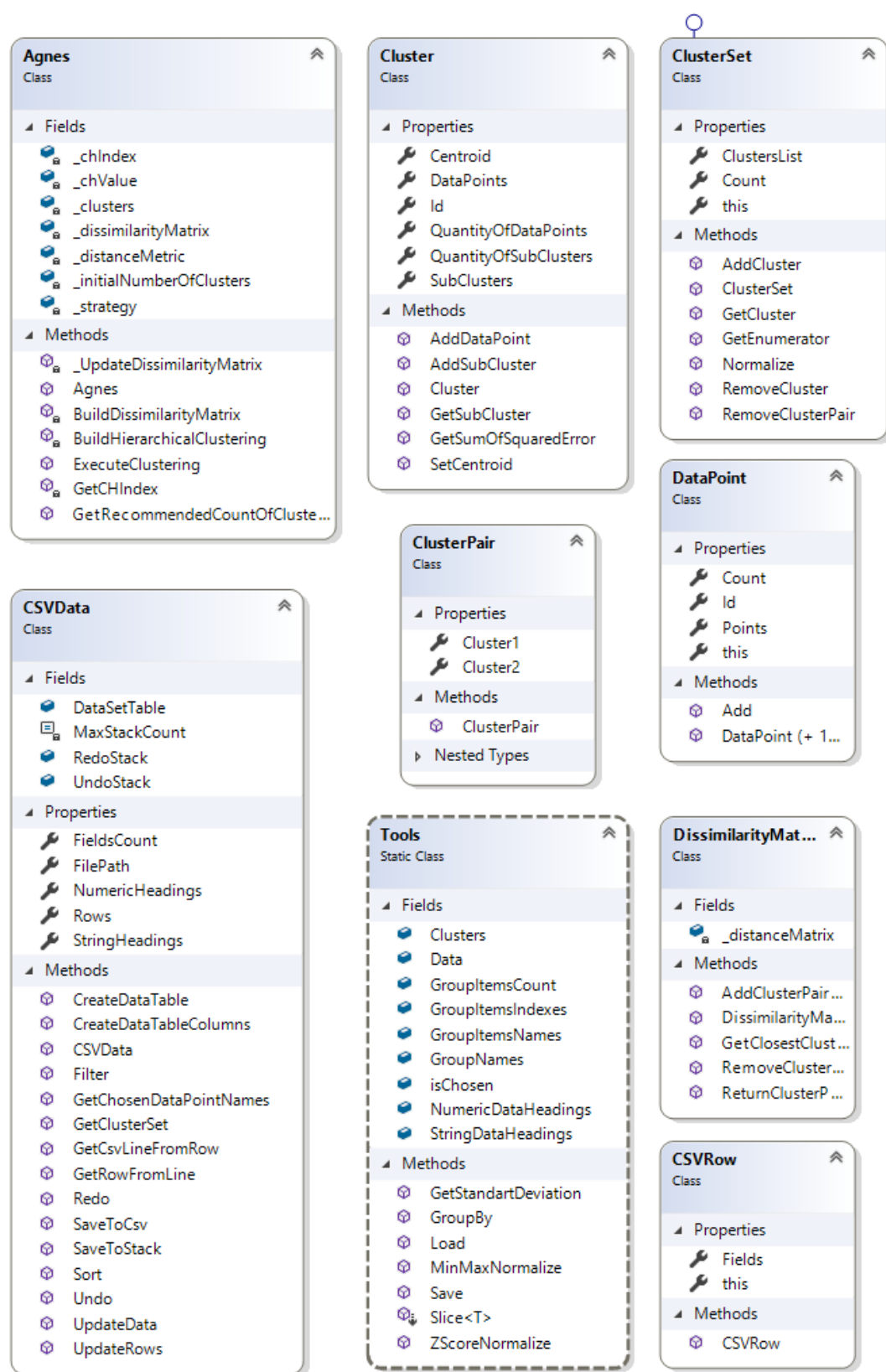
ПРИЛОЖЕНИЕ 7

ПРИМЕР ВЫХОДНОГО ФАЙЛА КАРТИНКИ ДЕНДОГРАММЫ

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

ПРИЛОЖЕНИЕ 8

ДИАГРАММА КЛАССОВ



Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

ОПИСАНИЕ И ФУНКЦИОНАЛЬНОЕ НАЗНАЧЕНИЕ КЛАССОВ

Таблица 9.1

Описание и функциональное назначение классов

Класс	Назначение
AboutBox	Окно показа информации о программе
Agnes	Класс, реализующий кластеризацию данных по алгоритму AGNES
Cluster	Класс кластер
ClusterDistance	Статистический класс для расчета расстояний между кластерами
MergeStrategy	Стратегия объединения кластеров
CSVRow	Класс для представления строки в CSV файле
CSVData	Класс для работы с входными данными
ClusterizeForm	Окно для выбора параметров кластеризации
ClusterPair	Класс пары кластеров
ClusterSet	Класс множества кластеров
NormalizeMethod	Метод нормализации
Configuration	Класс конфигурации входного файла
DendrogramForm	Окно для построения дендограммы
DissimilarityMatrix	Матрица различий
Distance	Статистический класс для расчета расстояний между двумя объектами
DistanceMetric	Мера расстояний
MainForm	Главное окно программы
Node	Элемент дерева кластеров
DataPoint	Класс множества точек
StatisticsForm	Окно для представления описательной таблицы кластеров
EqualityComparer	Класс для сравнения значений пар кластеров

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

ПРИЛОЖЕНИЕ 10

ОПИСАНИЕ И ФУНКЦИОНАЛЬНОЕ НАЗНАЧЕНИЕ ПОЛЕЙ МЕТОДОВ И СВОЙСТВ

Таблица 10.1

Описание полей методов и свойств класса AboutBox

Поля				
Имя	Модификатор доступа	Тип	Назначение	
components	private	IContainer	Генерируемые компоненты программы	
tableLayoutPanel	private	TableLayoutPanel	Сверточная таблица формы	
logoPictureBox	private	PictureBox	Лого программы	
labelProductName	private	Label	Продуктное название	
labelVersion	private	Label	Версия программы	
labelCopyright	private	Label	Копирайт	
labelCompanyName	private	Label	Имя компании	
textBoxDescription	private	TextBox	Описание программы	
okButton	private	Button	Кнопка ОК	
Свойства				
Имя	Модификатор доступа	Тип	Назначение	
AssemblyTitle	public	String	Название Assembly	
AssemblyVersion	public	String	Версия Assembly	
AssemblyDescription	public	String	Описание Assembly	
AssemblyProduct	public	String	Продукт Assembly	
AssemblyCopyright	public	String	Копирайт Assembly	
AssemblyCompany	public	String	Компания Assembly	
Методы				
Имя	Модификатор доступа	Тип	Аргументы	Назначение
InitializeComponent	private	Void		Инициализация компонент

Таблица 10.2

Описание полей методов и свойств класса CSVRow

Свойства			
Имя	Модификатор доступа	Тип	Назначение
Fields	public	List<string>	Строковые поля

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

Таблица 10.3

Описание полей методов и свойств класса Agnes

Поля				
Имя	Модификатор доступа	Тип	Назначение	
_clusters	private	ClusterSet	Множество кластеров	
_dissimilarityMatrix	private	DissimilarityMatrix	Матрица различий	
_distanceMetric	private	DistanceMetric	Мера расстояния	
_strategy	private	MergeStrategy	Стратегия объединения	
_initialNumberOfClusters	private	Int32	Начальное число кластеров	
_chIndex	private	List<int>	Количество кластеров при текущем значении СН	
_chValue	private	List<double>	Значение СН	
Методы				
Имя	Модификатор доступа	Тип	Аргументы	Назначение
BuildDissimilarityMatrix	private	Void		Строит таблицу различий
_UpdateDissimilarityMatrix	private	Void	Cluster newCluster	Обновляет таблицу различий
BuildHierarchicalClustering	private	Void	Int32 indexNewCluster, Int32 k, Boolean isWithIndex	Выполняет шаг кластеризации
ExecuteClustering	public	ClusterSet	Int32 k, Boolean isWithIndex	Выполняет кластеризацию
GetCHIndex	private	Double		Возвращает СН индекс
GetRecommendedCountOfClusters	public	Int32		Возвращает рекомендуемое количество кластеров

Таблица 10.4

Описание полей методов и свойств класса CustomException

Свойства			
Имя	Модификатор доступа	Тип	Назначение
Text	public	String	Текст
Caption	public	String	Дополнение

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

Таблица 10.5

Описание полей методов и свойств класса Cluster

Свойства				
Имя	Модификатор досутпа	Тип	Назначение	
SubClusters	public	List`1	Подкластеры	
DataPoints	public	List`1	Множество данных	
Centroid	public	DataPoint	Центроид	
Id	public	Int32	ID	
QuantityOfDataPoints	public	Int32	Количество данных	
QuantityOfSubClusters	public	Int32	Количество подкластеров	
Методы				
Имя	Модификатор досутпа	Тип	Аргументы	Назначение
AddDataPoint	public	Void	DataPoint dataPoint	Добавляет множество точек
AddSubCluster	public	Void	Cluster subCluster	Добавляет подкластер
GetSubCluster	public	Cluster	Int32 index	Возвращает подкластер
SetCentroid	public	Void		Определяет центроид
GetSumOfSquaredError	public	Double	DistanceMetric distanceMetric	Возвращает сумму квадратичного отклонения

Таблица 10.6

Описание полей методов и свойств класса ClusterDistance

Методы				
Имя	Модификатор доступа	Тип	Аргументы	Назначение
ComputeDistance	public static	Double	Cluster cluster1, Cluster cluster2, DistanceMetric distanceMetric	Расчитывает расстояние между кластерами
ComputeDistance	public static	Double	Cluster cluster1, Cluster cluster2, DissimilarityMatrix dissimilarityMatrix, MergeStrategy strategy	Расчитывает расстояние между кластерами
_MinValue	private static	Double	Double value1, Double value2	Возвращает минимальное значение
_MaxValue	private static	Double	Double value1, Double value2	Возвращает максимальное значение

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

Таблица 10.7

Описание полей методов и свойств класса MergeStrategy

Поля			
Имя	Модификатор доступа	Тип	Назначение
SingleLinkage	static public	MergeStrategy	Одиночная связь
CompleteLinkage	static public	MergeStrategy	Полная связь
AverageLinkageWpgma	static public	MergeStrategy	Невзвешанное среднее
AverageLinkageUpgma	static public	MergeStrategy	Взвешанное среднее
CentroidMethod	static public	MergeStrategy	Метод Центроидов
WardsMethod	static public	MergeStrategy	Метод Варда

Таблица 10.8

Описание полей методов и свойств класса ClusterizeForm

Поля				
Имя	Модификатор доступа	Тип	Назначение	
strategy		MergeStrategy	Стратегия объединения	
distanceMetric		DistanceMetric	Мера расстояний	
normalizeMethod		NormalizeMethod	Метод нормализации	
countOfClusters		Int32	Количество кластеров	
isParametersSelected		Boolean	Состояние выбора параметров	
_isTreeViewBusy	private	Boolean	Состояние недоступности TreeView	
distanceSelectComboBox	private	ComboBox	Combobox выбора мер расстояний	
strategySelectComboBox	private	ComboBox	Combobox выбора стратегии объединения	
doClusteringButton	private	Button	Кнопка кластеризации	
pointsSelectTreeView	private	TreeView	Дерево выбора показателей	
clusterCountTextBox	private	TextBox	Текстовое поле ввода количества кластеров	
calculateClusterCountButton	private	Button	Кнопка расчёта рек количества кластеров	
normalizeMethodSelectComboBox	private	ComboBox	Combobox выбора метода нормализации	
Методы				
Имя	Модификатор доступа	Тип	Аргументы	Назначение
CheckNodes	private	Void	TreeNode node, Boolean check	Клик на TreeView
InitializeComponent	private	Void		Инициализация

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

Таблица 10.9

Описание полей методов и свойств класса ClusterSet

Свойства				
Имя	Модификатор досутпа	Тип	Назначение	
Count	public	Int32	Число кластеров	
ClustersList	public	List<Cluster>	Список калстеров	
Методы				
Имя	Модификатор досутпа	Тип	Аргументы	Назначение
AddCluster	public	Void	Cluster cluster	Добавляет кластер
RemoveCluster	public	Void	Cluster cluster	Удаляет кластер
GetCluster	public	Cluster	Int32 index	Возвращает кластер под индексом
RemoveClusterPair	public	Void	ClusterPair clusterPair	Удаляет пару кластеров
GetEnumerator	public	IEnumerator		Возвращает IEnumerator
Normalize	public	Void	NormalizeMethod normalizeMethod	Нормализует кластеры

Таблица 10.10

Описание полей методов и свойств класса NormalizeMethod

Поля			
Имя	Модификатор доступа	Тип	Назначение
None	static public	NormalizeMethod	Никакой нормализации
MinMax	static public	NormalizeMethod	MinMax нормализация
ZScore	static public	NormalizeMethod	Z-Score нормализация

Таблица 10.11

Описание полей методов и свойств класса Configuration

Поля			
Имя	Модификатор доступа	Тип	Назначение
StringHeadings	public	String[]	Строковые данные
NumericHeadings	public	String[]	Показатели
GroupNames	public	String[]	Имена групп
GroupItemsCount	public	Int32[]	Количество элементов в группе

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

Таблица 10.11

Описание полей методов и свойств класса EqualityComparer

Методы				
Имя	Модификатор доступа	Тип	Аргументы	Назначение
Equals	public	Boolean	ClusterPair x, ClusterPair y	Сравнивает две пары кластеров
GetHashCode	public	Int32	ClusterPair x	Возвращает Хэш код пары кластеров

Таблица 10.12

Описание полей методов и свойств класса DissimilarityMatrix

Поля				
Имя	Модификатор доступа	Тип	Назначение	
_distanceMatrix	private	<ClusterPair, double>	Матрица различий	
Методы				
Имя	Модификатор доступа	Тип	Аргументы	Назначение
AddClusterPairAndDistance	public	Void	ClusterPair clusterPair, Double distance	Добавляет пару кластеров
RemoveClusterPair	public	Void	ClusterPair clusterPair	Удаляет пару кластеров
GetClosestClusterPair	public	ClusterPair		Возвращает пару кластеров с минимальным расстоянием
ReturnClusterPairDistance	public	Double	ClusterPair clusterPair	Возвращает расстояние пары кластеров

Таблица 10.13

Описание полей методов и свойств класса DistanceMetric

Поля			
Имя	Модификатор доступа	Тип	Назначение
EuclidianDistance	static public	DistanceMetric	Евклидовое расстояние
SquareEuclidianDistance	static public	DistanceMetric	Квадрат Евклидогого расстояния
ManhattanDistance	static public	DistanceMetric	Расстояние городских кварталов
ChebyshevDistance	static public	DistanceMetric	Расстояние Чебышева

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

Таблица 10.14

Описание полей методов и свойств класса Tools

Поля				
Имя	Модификатор доступа	Тип	Назначение	
StringDataHeadings	static public	String[]	Строковые значения	
NumericDataHeadings	static public	String[]	Числовые значения	
GroupNames	static public	String[]	Имена групп	
GroupItemsCount	static public	Int32[]	Количество элементов в группе	
GroupItemsNames	static public	String[][]	Заголовки в группах	
GroupItemsIndexes	static public	Int32[][]	Индексы в группах	
isChosen	static public	Boolean[]	Массив состояний выбора показателя	
Data	static public	CSVData	Данные	
Clusters	static public	ClusterSet	Множество кластеров	
Методы				
Имя	Модификатор доступа	Тип	Аргументы	Назначение
MinMaxNormalize	public static	Void	Double[]& arr	Min-Max нормализация
ZScoreNormalize	public static	Void	Double[]& arr	Z-Score нормализация
GetStandartDeviation	public static	Double	Double[] points	Возвращает среднеквадратическое отклонение
GroupBy	public static	Int32[]	Double[] points	Группирует элементы
Slice	public static	T[]	T[] arr, Int32 indexFrom, Int32 indexTo	Разделяет массив
Save	public static	Void		Сохраняет данные
Load	public static	Void		Загружает данные

Таблица 10.15

Описание полей методов и свойств класса DataPoint

Свойства				
Имя	Модификатор досутпа	Тип	Назначение	
Points	public	List<Double>	Точки	
Id	public	Int32	ID	
Count	public	Int32	Количество точек	
Item	public	Double	Элемент	
Методы				
Имя	Модификатор досутпа	Тип	Аргументы	Назначение
Add	public	Void	Double point	Добавляет точку

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

Таблица 10.16

Описание полей методов и свойств класса Node

Свойства			
Имя	Модификатор доступа	Тип	Назначение
Contents	public	T	Контент
ChildrenNodes	public	List<Node<T>>	Дети

Таблица 10.17

Описание полей методов и свойств класса StatisticsForm

Поля				
Имя	Модификатор доступа	Тип	Назначение	
Clusters	internal	ClusterSet	Множество кластеров	
contentsHeadings	internal	String[]	Заголовки	
_dataTable	private	DataTable	Таблица данных	
_contents	private	StatisticPoint[][]	Контент данных	
_colors	private	Color[]	Цвета	
clustersOverviewGridView	private	DataGridView	DataGridView данных	
Методы				
Имя	Модификатор доступа	Тип	Аргументы	Назначение
Setup	public	Void		
StatisticsForm_FormClosing	private	Void	Object sender, FormClosingEventArgs	Событие перед закрытием окна
StatisticsForm_Load	private	Void	Object sender, EventArgs	Событие при загрузке окна
InitializeComponent	private	Void		Инициализация компонентов

Таблица 10.18

Описание полей методов и свойств класса Node

Свойства			
Имя	Модификатор доступа	Тип	Назначение
Contents	public	T	Контент
ChildrenNodes	public	List<Node<T>>	Дети

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

Описание полей методов и свойств класса CSVData

Поля				
Имя	Модификатор доступа	Тип	Назначение	
DataSetTable	public	DataTable	Таблицы данных	
UndoStack	public	Stack<List<>>	Стек Undo	
RedoStack	public	Stack<List<>>	Стек Redo	
MaxStackCount	static private	Int32	Максимальное количество стека	
Свойства				
Имя	Модификатор доступа	Тип	Назначение	
FieldsCount	public	Int32	Количество столбцов	
Rows	public	List<CSVRow>	Столбцы	
StringHeadings	public	String[]	Строковые заголовки	
NumericHeadings	public	String[]	Числовые заголовки	
FilePath	public	String	Путь файла	
Методы				
Имя	Модификатор доступа	Тип	Аргументы	Назначение
GetClusterSet	public	ClusterSet	Boolean[] isChosen	Возвращает множество кластеров
GetRowFromLine	public	CSVRow	String line	Возвращает строки из Row
GetCsvLineFromRow	public	String	CSVRow csvRow	Возвращает Row из строки
UpdateRows	public	Void		Обновляет Rows
CreateDataTableColumns	public	Void		Создает заголовки в таблице
CreateDataTable	public	Void		Создает таблицу
UpdateData	public	Void		Обновляет таблицу
GetChosenDataPointNames	public	String[]	Boolean[] isChosen	Возвращает выбранные таблицы
SaveToCsv	public static	Void	CSVData data, String filePath	Сохраняет в CSV файл
Sort	public	Void	Int32 index, Boolean isAscending	Сортирует данные

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

Продолжение таблицы 10.19

Filter	public	Void	String expression, Int32 index, Int32 selectedOperati on	Фильтрует данные
Redo	public	Void		Отмена последнего действия
Undo	public	Void		Повтор последнего действия
SaveToStack	public	Void		Сохранить состояние в стек

Таблица 10.20

Описание полей методов и свойств класса DendrogramForm

Поля				
Имя	Модификатор доступа	Тип	Назначение	
_drawarea	private	Graphics	Поле отрисовки	
_root	private	Node<string>	Корневой Node	
_leaves	private	Int32	Число уровней	
_levels	private	Int32	Число листьев	
_color	private	Color	Цвет отрисовки	
_widthPerLevel	private	Int32	Ширина уровня	
_maxLevel	private	Int32	Максимальный уровень	
_currentY	private	Int32	Текущий Y	
_height	private	Int32	Высота	
_width	private	Int32	Ширина	
_rootList	private	List`1	Список корней	
_leavesList	private	List`1	Список листьев	
_levelsList	private	List`1	Список уровней	
_colors	private	List`1	Список цветов	
dendogramControl	private	PictureBox	PictureBox отрисовки	
random	static private	Random	Генератор случайных чисел	
HeightPerLeaf	static private	Int32	Высота листа	
DrawingAreaMargin	static private	Int32	Отступ от отрисовки	
ContestOffset	static private	Int32	Отступ от контента	
DrawingAreaOffset	static private	Int32	Отступ от окна	
Методы				
Имя	Модификатор доступа	Тип	Аргументы	Назначение
Create	private	Node<string>	String contents	Создает лист
Create	private	Node<string>	Node<string> child0, Node<string> child1	Создает лист

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

Продолжение таблицы 10.20

CountLeaves	private	Int32	Node<string> node	Считывает листья
CountLevels	private	Int32	Node<string> node	Считывает уровни
BuildDendrogram	private	Node<string>	Cluster[] clusters	Строит дендограммы
GetNodeFromCluster	private	Node<string>	Cluster cluster	Возвращает Node
Setup	public	Void		Настройка окна
Draw	private	Point	Graphics g, Node`1 node, Int32 y	Отрисовка
DendrogramForm_SizeChanged	private	Void	Object sender, EventArgs e	Событие смены размера окна
DendrogramForm_FormClosing	private	Void	Object sender, FormClosingEventArgs e	Событие перед закрытием окна
InitializeComponent	private	Void		Инициализация окна

Таблица 10.21

Описание полей методов и свойств класса MainForm

Поля			
Имя	Модификатор доступа	Тип	Назначение
_dataPointsBeforeNormalization	private	List<DataPoint>	Множество точек перед нормализацией
statisticsForm	private	StatisticsForm	Окно StaticsForm
menuStrip	private	MenuStrip	Меню
tabControl	private	TabControl	Табличный контроль
preprocessPage	private	TabPage	Основная страница
toolStrip	private	ToolStrip	Панель инструментов
dataTableGridView	private	DataGridView	DataGridView таблицы данных
clusterizeTabPage	private	TabPage	Страница кластеризации
splitContainer	private	SplitContainer	Разделяющий контейнер
clusterTableGridView	private	DataGridView	DataGridView таблицы кластеров
clusterTreeView	private	TreeView	TreeView дерева кластеров

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

Продолжение таблицы 10.21

Методы				
Имя	Модификатор доступа	Тип	Аргументы	Назначение
ResetComponents	private	Void		Сброс компонент
ResetData	private	Void		Сброс данных
RestoreClusterSet	private	Void	Cluster cluster	Восстанавливает множество кластеров
DeleteTempFiles	public	Void		Удаляет временные файлы
LoadData	public	Void		Загружает данные
BuildResultTable	private	Void		Строит таблицу кластеров
BuildTreeView	private	Void		Строит дерево кластеров
AddNodes	private	Void	Cluster[] clusters, TreeNode node	Добавляет Nodes
DisableClustering	private	Void		Отключает кластеризацию
EnableClustering	private	Void		Включает кластеризацию
MainForm_FormClosing	private	Void	Object sender, FormClosingEventArgs e	Событие перед закрытием окна
Dispose		Void	Boolean disposing	Dispose
InitializeComponent	private	Void		Инициализация
<clusterizeToolStripMenuItem_Click>b__21_0	private	Void		

Таблица 10.22

Описание полей методов и свойств класса Distance

Методы				
Имя	Модификатор доступа	Тип	Аргументы	Назначение
GetDistance	public static	Double	DataPoint x, DataPoint y, DistanceMetric distanceMetric	Возвращает расстояние между двумя множествами точек

Таблица 10.23

Описание полей методов и свойств класса CSVRow

Свойства			
Имя	Модификатор доступа	Тип	Назначение
Fields	public	List<string>	Строковые поля

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

[illegible]

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.04.13-01 81				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата