ISSN 2658 – 7505
Выпуск №4, 2019
Электронный научный журнал «Вестник молодёжной науки России»

УДК 004.415.2

ФОРМАЛИЗАЦИЯ ЗАДАЧИ КЛАСТЕРИЗАЦИИ ФАКТОРОВ ВОДОСНАБЖЕНИЯ И ВОДООТВЕДННИЯ

Халимов В.А., *grrenvich@mail.ru,* ***Ковалевский В.Н.,*** *petrov@mail.ru*
Южно-Российский государственный политехнический университет (НПИ)
имени М.И. Платова, г. Новочеркасск

В данной работе рассмотрен механизм кластеризации факторов водоснабжения и водоотведения, реализованного на основе алгоритма *k-means* (*k*-средних). В качестве функциональной платформы этого алгоритма был выбран аналитический программный продукт «*Deductor*», включающий функцию кластерного анализа. Областью применения этого метода является информационно-аналитическая система для предприятия по водоснабжению и водоотведению. Задача состоит в исследовании районов и параметров, влияющих на уровень потребления данных ресурсов. Выбранный метод позволяет отнести многообразие факторов к различным кластерам по определенным признакам.

FORMALIZATION OF THE PROBLEM OF THE CLUSTERING OF FACTORS OF WATER SUPPLY AND WATER DISPOSAL

V.A.Halimov, V.N.Kovalevsky
Platov South-Russian State Polytechnic University (NPI), Novocherkassk

B In this work the mechanism of a clustering of factors of the water supply and water disposal implemented on the basis of an algorithm k-means (k-averages) is considered. As the functional platform of this algorithm the analytical Deductor software product including function of cluster analysis was selected. A scope of this method is the information-analytical system for the enterprise for water supply and water disposal. The task consists in a research of the areas and parameters affecting level of consumption of resource data. The selected method allows to refer variety of factors to different clusters on certain signs.

Данная работа является продолжением исследования по анализу информационной системы (ИС) абонентского отдела предприятия водоснабжения [1] и посвящен аналитической обработке данных, поступающих в эту систему. Дальнейшее проектирование ИС связано с внедрением в нее ниже рассмотренного блока анализа данных, построенного на основе *кластерного* алгоритма. Он подразумевает под собой упорядочение объектов по схожести, что позволяет разбивать общий поток данных на сегменты (группы) со похожими свойствами и параметрами. Такой подход предусматривает причисление, например, районов города, обеспечиваемых водными ресурсами, к определенным группам, по определенным факторам. В данном случае этими факторами являются такие элементы, как средний расход воды, стоимость воды, уровень жизни в районе, наличие промышленного объекта и т.д.

Задачей кластерного анализа является организация наблюдаемых данных в наглядные структуры [2]. В процессе кластеризации осуществляется группировка объектов по различным факторам и параметрам. Состояние исследуемого объекта может быть описано с помощью вектора дескрипторов или многомерного набора зафиксированных на нём признаков, Например, $X = \{x^1, x^2, \ldots, x^p\}$, где x^1 – средний уровень жизни (от 0 до 10), x^2- наличие промышленного объекта, x^3- средняя температура окружающей среды (за 3 месяца, зима), x^4- наличие сельхоз объ-

ISSN 2658 – 7505
Выпуск №4, 2019
Электронный научный журнал «Вестник молодёжной науки России»

екта), x^5- средний расход воды и т.д. Часть признаков может носить количественный характер и принимать любые действительные значения. Другая часть носит качественный характер и позволяет упорядочивать объекты по степени проявления какого-либо качества (например, бинарный признак, отображающий присутствие или отсутствие данного свойства).

Алгоритмов кластеризации существует множество. Один из наиболее широко применяемых — алгоритм (или метод) *k*-средних (*k-means clustering*), также называемый быстрым кластерным анализом. В отличие от иерархических методов, которые не требуют предварительных предположений относительно числа кластеров, для возможности использования этого метода необходимо иметь гипотезу о наиболее вероятном количестве кластеров.

Метод *k*-средних это метод кластерного анализа, цель которого является разделение *m* наблюдений на *k* кластеров, при этом каждое наблюдение относится к тому кластеру, к центру которого оно ближе всего. Центр кластера представляет точку пересечения значений всех его факторов. т.е *центроид*. Этот алгоритм основан на оптимизации суммы квадратов взвешенных отклонений координат объектов от центров искомых кластеров. В качестве меры близости используется Евклидово расстояние [2]:

$$\rho(x, y) = \|x - y\| = \sqrt{\sum_{p=1}^{n}(x_p - y_p)^2},$$

где x^1 – средний уровень жизни (от 0 до 10), x^2- наличие промышленного объекта, x^3- средняя температура окружающей среды (за 3 месяца, зима) и т.д,
$x, y \in R^n$ - рассматриваемый ряд наблюдений $(x^{(1)}, x^{(2)}, \ldots, x^{(m)})$, $x^{(j)} \in R^n$.

Метод *k*-средних разделяет *m* наблюдений на *k* групп (или кластеров) ($k \leq m$) $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$ так, чтобы минимизировать суммарное квадратичное отклонение точек кластеров от центроидов этих кластеров:

$$\min \left[\sum_{i=1}^{k} \sum_{x^{(j)} \in S_i} \left\| x^{(j)} - \mu_i \right\|^2\right],$$

где $x^{(j)} \in R^n, \mu_i \in R^n$, μ_i - центроид для кластера S_i.

Последовательность алгоритма кластеризации состоит в следующем.

1. Первоначальное распределение объектов по кластерам.

Задается число *k* наблюдений (объектов) и на первом шаге они считаются центрами кластеров. Каждому кластеру соответствует один центр. Выбор этих центров может осуществляться одним из следующих способов:

- выбор первых *k*-наблюдений;
- случайный выбор *k*-наблюдений;
- выбор *k*-наблюдений для максимизации начального расстояния.

В результате каждый объект будет назначен определенному кластеру. Рассмотрим первоначальный набор *k* - центроидов $\mu_1, \ldots, \mu_k$ в кластерах $S_1, S_2, \ldots, S_k$. Отнесем наблюдения к тем кластерам, чье среднее к ним ближе всего. Каждое наблюдение принадлежит только к одному кластеру, даже если его можно отнести к двум и более кластерам.

2. Итеративный процесс.

Вычисляются центроиды каждого *i*-го кластера по следующей формуле:

ISSN 2658 – 7505
Выпуск №4, 2019
Электронный научный журнал «Вестник молодёжной науки России»

$$\mu_j = \frac{1}{s_j} \sum_{x^{(j)} \in s_j} x^{(j)}$$

Процесс их вычисления и перераспределения объектов (наблюдений) продолжается до тех пор, пока не выполнено одно из условий:

- кластерные центры стабилизировались, т.е. все наблюдения принадлежат кластеру, которому принадлежали до текущей итерации - когда значения μ_i не меняются: $\mu_i^{\text{mar } t} = \mu_i^{\text{mar } t+1}$
- число итераций равно максимальному заданному числу итераций.

Таким образом, алгоритм *k*-средних заключается в перевычислении на каждом шаге центроида для каждого кластера, полученного на предыдущем шаге. Неправильный выбор первоначального числа кластеров *k* может привести к некорректным результатам. Именно поэтому при использовании метода *k*-средних важно сначала провести проверку подходящего числа кластеров для данного набора данных.

Особенности метода *k*-средних:

1) в качестве метрики используется Евклидово расстояние;
2) число кластеров заранее не известно и выбирается исследователем заранее;
3) качество кластеризации зависит от первоначального разбиения.

Рассмотрим практическую реализацию механизма кластеризации на основе алгоритма *k-means*, основываясь на данных наблюдений, собранных в информационной системе абонентского отдела предприятия водоснабжения. Задача состоит в исследовании районов города и факторов, влияющих на водопотребление и водоотведение (см. таблицу 1).

Таблица 1

Факторы, влияющие на водопотребление и водоотведение

код	Поле
1	Название Района
2	Средний Уровень Жизни (от 0 до 10)
3	Наличие Промышленного Объекта
4	Средняя Температура Окружающей Среды (за 3 месяца, Зима)
5	Средняя Температура Окружающей Среды (за 3 месяца, Тёплое)
6	Наличие Сельхоз Объекта
7	Среднее Водопотребление Холодной Воды (в месяц на человека, кубы)
8	Водоотведение (в месяц на человека, кубы)

Эти данные наблюдений были собраны в файле «*Факторы02.txt»*. Далее необходимо осуществить импорт этого файла с помощью мастера импорта в аналитический блок *«Deductor»* [3]. В результате в основном окне появится таблица, заполненная из указанного файла и представленная на рис. 1.

Далее необходимо запустить Мастер обработки данных, в окне раздела *DataMining* выбрать метод обработки «Кластеризация» Кластеризация алгоритмом *k-means* (см.рис. 2).

ISSN 2658 – 7505
Выпуск №4, 2019
Электронный научный журнал «Вестник молодёжной науки России»

Код	Название Района	Средний Уровень Жизни(от 0 до 10)	Наличие Промышленного Объекта	Средняя Температура Окружающей Среды(за 3 месяца,Зима)	Средняя Температура Окружающей Среды(за 3 месяца,Теплое)	Наличие Сельхоз Объекта	Среднее Водопотребление Холодной Воды(в месяц на человека, кубы)	Водоотведение (в месяц на человека, кубы)
1	Октябрьский	6	0	-1	14,5	0	7,2	4
2	Соцгород	6	1	-1	14,7	0	8,1	5
3	Центр	9	0	-3	13	0	6,5	7
4	Черемушки	7	0	-3,2	13,1	0	6	6,9
5	Молодёжка	5	1	-4	11	1	9	4
6	Хотунок	5	0	-2	14	0	5,6	4,1
7	Персиановский	4	0	-2	16	1	13,2	4
8	Восточный	6	1	-2,1	14	0	7	4
9	Донской	4	0	-3	17	1	8	5

Рис.1 – Исходные данные

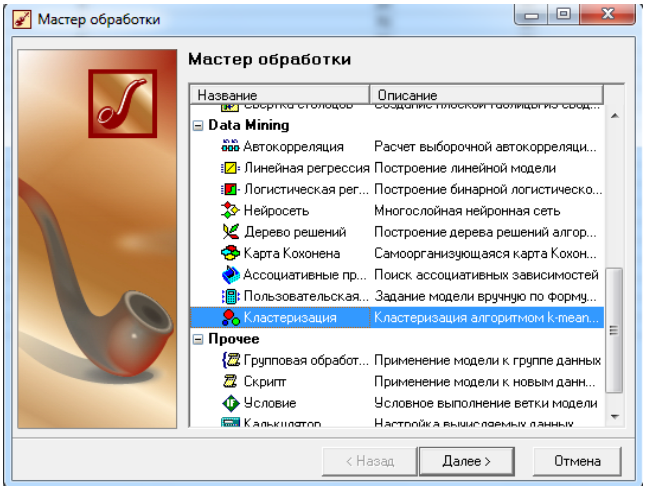

Рис.2 – Мастер обработки

На вкладке «Настройка значения столбцов» необходимо задать назначения столбцов данных, т.е. выбрать свойства, по которым будет происходить группировка объектов (см.рис.3). Также надо указать столбцам "Номер кластера" и "Расстояние до центра кластера» назначение "Выходное", а остальным столбцам – "Входное".

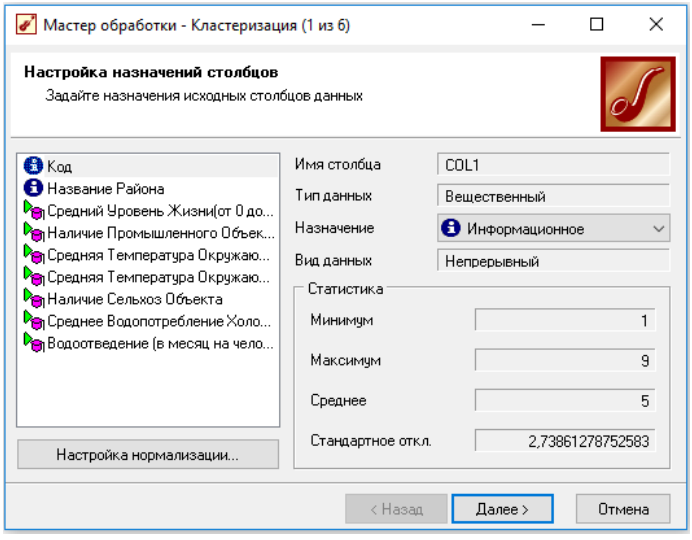

Рис. 3 – Настройка назначения полей

ISSN 2658 – 7505
Выпуск №4, 2019
Электронный научный журнал «Вестник молодёжной науки России»

На следующем шаге необходимо настроить способ разделения исходного множества данных на тестовое и обучающее, а также количество примеров в том и другом множестве. Для данных обоих множеств этот способ определяется случайный образом, а количество примеров для обучающего множества -100%. Следующий шаг предназначен для настройки параметров кластеризации, т.е. для определения количества кластеров (см.рис. 4).

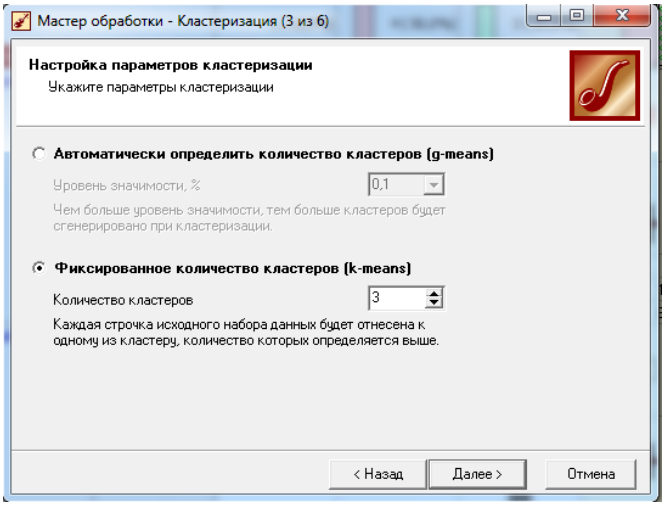

Рис.4 – Настройка параметров кластеризации

Для отображения полученных групп кластеров необходимо выбрать в обработчике "Кластеризация» из списка визуализаторов способы отображения данных: "Что-если" для решения задачи классификации, отнесения объекта к одному из кластеров; «Таблица" для наглядного просмотра объекта, вошедших в кластеры; "Статистика" для просмотра статистических данных по объектам; «Профили кластеров» для определения структуры формирования группы кластеров и «Куб» для наглядного просмотра полученных результатов. Для настройки визуализатора «Куб» необходимо выбрать рассматриваемые свойства как факты, а также номер кластера и код как измерение. В дальнейших настройках надо задать отображение фактов как среднее по рассматриваемой группе (см.рис. 5).

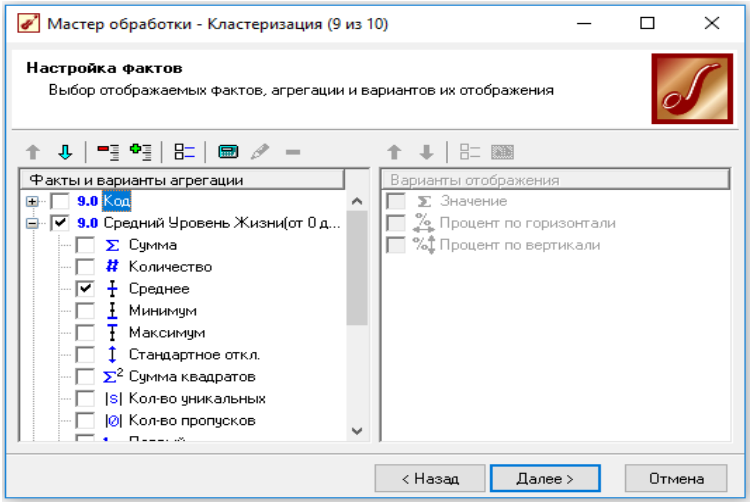

Рис.5 – Настройка фактов

ISSN 2658 – 7505
Выпуск №4, 2019
Электронный научный журнал «Вестник молодёжной науки России»

Общую структуру сформированных алгоритмом кластеров можно просмотреть в визуализаторе "Профили кластеров". В нем представлены все рассматриваемые свойства вместе с характером влияния их на состав кластера. Основным определяющим состав кластера фактором является значимость свойств, выраженная в процентах. Общая значимость рассматриваемого поля определяется вариабельностью ее рассматриваемых параметров. Значимость для непрерывных и дискретных полей определяется по-разному. Значимость для непрерывных полей устанавливается взависимости от отклонения среднего значения рассматриваемой группы кластеров от общего среднего значения всей выборки. Чем больше выражено данное отклонение, тем больше его значимость. Значимость для дискретных полей определяется наличием индивидуальных различий, между рассматриваемыми группами. Чем больше выражены различия, тем больше значимость. Для каждого рассматриваемого свойства в кластере вычисляются: доверительный интервал, среднее, стандартное отклонение и стандартная ошибка (см.рис. 6).

Таким образом алгоритм автоматически разбил районы на три кластера с разной поддержкой и разными процентами значимости свойств. Первый кластер содержит высшие показатели по всем параметрам. Наиболее ярко выраженными кластерами по заданным свойствам является первый и второй. Они максимально отличаются от нулевого значениями свойств и минимальной поддержкой.

		Кластеры			
		0	1	2	Итого
		6 (66,7%)	2 (22,2%)	1 (11,1%)	
Поля	Показатели				
9.0 Водоотведение (в месяц на человека, кубы)	Значимость	79,4%	94,9%	6,6%	100,0%
	Доверительный интервал				
	Среднее	4,183333333	6,95	5	4,888888889
	Стандартн. откл.	0,402079361	0,07071067812	0	1,240407639
	Стандартн. ошиб.	0,16414763	0,05	0	0,413469213
9.0 Средний Уровень Жизни(от 0 до 10)	Значимость	70,7%	90,1%	10,4%	97,1%
	Доверительный интервал				
	Среднее	5	8	6	5,777777778
	Стандартн. откл.	0,894427191	1,414213562	0	1,56347192
	Стандартн. ошиб.	0,3651483717	1	0	0,5211573066
9.0 Средняя Температура Окружающей	Значимость	2,4%	64,7%	76,3%	73,5%
9.0 Наличие Промышленного Объекта	Значимость	0,0%	61,1%	75,8%	70,4%

Рис.6 – Профили кластеров

ISSN 2658 – 7505
Выпуск №4, 2019
Электронный научный журнал «Вестник молодёжной науки России»

Результаты по сформированным кластерам наиболее удобно рассматриваются с помощью визуализатора "Куб", в котором встроена кросс-диаграмма, изображающая полученные кластеры в графическом виде, что существенно упрощает анализ (см.рис. 7). Исходя из данных кросс-диаграммы можно сделать выводы о влиянии на включение в кластеры такого показателя, как уровень жизни, являющимся основополагающим для Центра города и «Черемушек».

После окончания кластерного анализа, разбиения первичных данных на три кластера для полученная точечной информации был проведен «Корреляционный анализ». Данный вид анализа позволяет понять какие факторы влияют на друг друга в большей степени. Выбранный фактор изучения ставится в параметрах данных как «выходные», а остальные указываются как «входные». Далее следует выбрать один из двух методов корреляционного анализа. В данном случае будет использован «Метод Пирсона», на основе которого получена таблица, показанная на рисунке 8.

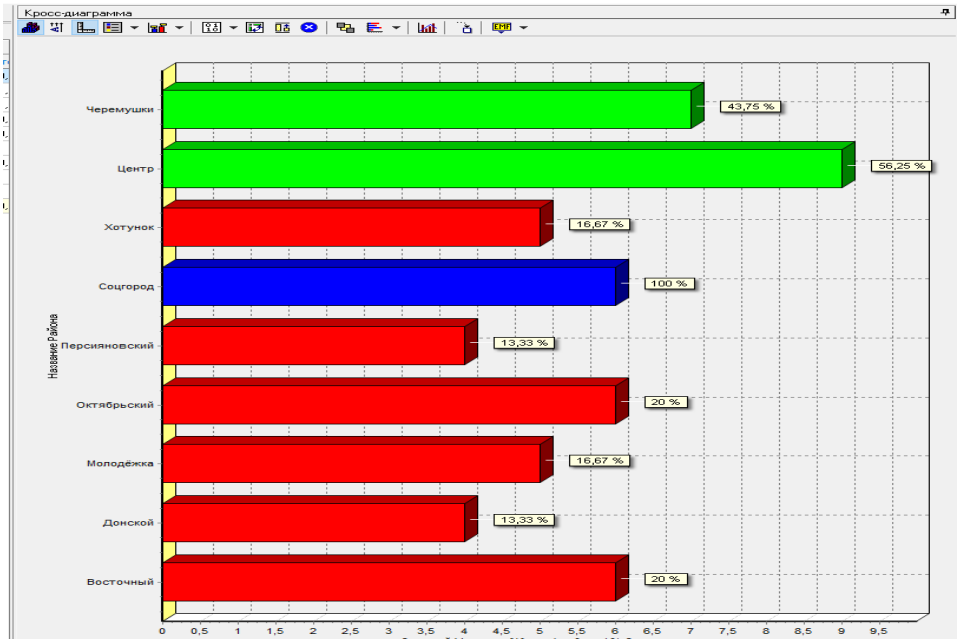

Рис.7 – Кросс-диаграмма

	Входные поля	Корреляция с выходными полями
№	Поле	Среднее Водопотребление Холодн...
1	Средний Уровень Жизни(от 0 до 10)	-0,555
2	Наличие Промышленного Объекта	0,062
3	Средняя Температура Окружающе...	0,043
4	Средняя Температура Окружающе...	0,333
5	Наличие Сельхоз Объекта	0,732
6	Водоотведение (в месяц на челове...	-0,407

Рис.8 – Корреляционный анализ «Среднее Водопотребление Холодной Воды»

Исходя из представленных в таблице данных, можно сделать вывод, что наличие сельхоз объекта больше всего влияет на количество потребляемой холодной воды. Выводить данные нужно отдельно для каждого вида факторов, чтобы установить все зависимости и их влияние.

Таким образом, в ходе выполнения данной работы было рассмотрено применение кластеризации для группового анализа данных. С помощью этой задачи

жилые районы были разделены на кластеры и представлены в удобной для анализа визуализации Куб. С помощью Куба была построена Кросс-диаграмма данных, на основании которой был сделан вывод об основных показателях взаимосвязей районов и их интенсивности использования водных ресурсов Также был проведен корреляционный анализ, с помощью которого были выявлены наиболее влияющие факторы.

Список цитируемой литературы

1. Халимов В.А., Ковалевский В.Н. Моделирование информационной системы абонентского отдела типового предприятия водоснабжения// Фундаментальные основы, теория, методы и средства измерений, контроля и диагностики: материалы 19-ой Междунар. молод. науч.-практ. конф., г. Новочеркасск, 27-28 февр. 2018г. / Южно-Российский государственный политехнический университет (НПИ) имени М.И. Платова. – Новочеркасск: Лик, 2018. - С.341-346.

2. Чубукова И. *DataMining.* Интернет-университет информационных технологий, Бином. Лаборатория знаний, 2008. - 384с.

3.Карпузова В.И., Скрипченко Э.Н., Стратонович Ю.Р., Чернышева К.В. Методические указания по *Deductor*. Москва: Изд-во ФГОУ ВПО РГАУ-МСХА, 2010.- 80с.