УДК 004.62; 004.04 DOI: 10.18698/2541-8009-2018-01-230

ОБЗОР ВОЗМОЖНОСТЕЙ КЛАСТЕРНОГО АНАЛИЗА ДАННЫХ В ПРОГРАММНОМ ПАКЕТЕ STATISTICA ADVANCED

И.А. Тихонов quiteman@mail.ru
SPIN-код: 5941-2719

МГТУ им. Н.Э. Баумана, Москва, Российская Федерация

Аннотация

Работа посвящена обзору возможностей кластеризации данных в программном пакете STATISTICA. Представлено описание методов кластеризации и особенностей работы с ними в данном продукте с практической точки зрения. Рассмотрено такое свойство, как мера расстояния между элементами исходного множества и некоторые методы разбиения исходного множества наблюдений на кластеры, а также результаты кластерного анализа после применения алгоритмов пакета STATISTICA Advanced. Практическая значимость и актуальность применения кластерного анализа к данным не вызывает сомнения, поскольку в современном информационном обществе данные и результаты их анализа играют все большую роль, а кластеризация позволяет лучше понять эти данные.

Ключевые слова

Анализ данных, кластерный анализ, кластеризация, классификация без учителя, STATISTICA, евклидово пространство, иерархические и неиерархические методы кластеризации, joining/tree clustering, k-mean clustering, two-way joining

Поступила в редакцию 30.10.2017

Первые работы по кластерному анализу появились в конце 1930-х годов. Позднее в связи с развитием ЭВМ и ПК, то есть с ростом объема обрабатываемых данных и сложностью их обработки, кластеризация стала основой анализа данных.

Американский профессор Йельского университета John A. Hartigan в своей работе «Clustering Algorithms» выделил следующие значимые сферы, в которых применяется кластерный анализ: природа, медицина, психиатрия, археология, антропология, экономика и исследования рынка, астрология, образование [1].

Применение принципов кластеризации в данных сферах является очень важным. Например, в медицине, кластеризация имеющихся данных (наблюдений) может помочь спасти человеку жизнь. На основании имеющихся данных врач сможет соотнести состояние с той или иной категорией больных и быстрее назначить лечение.

Также в последнее время в связи с ростом объемов и ценности информации многие компании начали использовать методы анализа данных с целью извлечения прибыли.

Постановка задачи. Кластерный анализ, кластеризация, или естественная классификация, — это процедура или комплекс процедур, которые осуществляют разбиение исходного множества данных на подмножества, обладающие максимально схожими характеристиками и в то же время максимально отличными от других подмножеств [2]. Задача кластеризации решается на начальных

этапах исследования, ее решение помогает лучше понять данные и их природу. Формальная постановка задачи кластеризации имеет следующий вид:

1) $X\{x_1,\dots,x_i,\dots,x_j,\dots,x_n\}$ — исходное множество объектов, характеризующихся некоторым набором атрибутов $A\{a_1,\dots,a_i,\dots,a_j,\dots,a_n\}$;

2) $\rho(x_i,x_j)$ — функция расстояния, характеризующая меру близости между объектами исходного множества;

3) $C\{c_1,\dots,c_i,\dots,c_j,\dots,c_n\}$ — искомое множество, являющееся совокупностью непересекающихся подмножеств (кластеров), состоящих из объектов множества X и являющихся близкими в соответствии с метрикой ρ: $c_i=\{x_i,x_j \mid x_i,x_j \in X \text{ и } \rho(x_i,x_j)<\sigma\}$, где σ — величина, определяющая меру близости.

Таким образом, выбор меры расстояния является важным этапом при кластерном анализе. Выделяют меры следующих типов:

– евклидовы основаны на местоположении точек в евклидовом пространстве;

– неевклидовы основаны на свойствах точек, но не на их положении в пространстве.

Рассмотрим поподробнее первый тип [3, 4]. К нему относят такие меры как (формулы представлены в общем виде)

• евклидово расстояние:

$$\rho(x_i,x_j)=\sqrt{\sum_{m=1}^{n}(x_{im}-x_{jm})^2},$$

где n — размерность пространства;

• квадрат евклидова расстояния. Позволяет задать большее расстояние отдаления между исследуемыми объектами:

$$\rho(x_i,x_j)=\sum_{m=1}^{n}(x_{im}-x_{jm})^2;$$

• расстояние городских кварталов (Manhattan distances). В отличие от евклидова расстояния, для этой меры влияние отдельных больших разностей (выбросов) уменьшается:

$$\rho(x_i,x_j)=\sum_{m=1}^{n}|x_{im}-x_{jm}|;$$

• расстояние Чебышёва. Позволяет задать такое расстояние, в соответствии с которым из исходной выборки будут определены максимально отличные друг от друга объекты:

$$\rho(x_i,x_j)=\max(|x_{im}-x_{jm}|).$$

Для иллюстрации изложенного выше предположим, что необходимо найти расстояния между точками A и C в двумерном пространстве. Точки имеет координаты x, y (рис. 1).

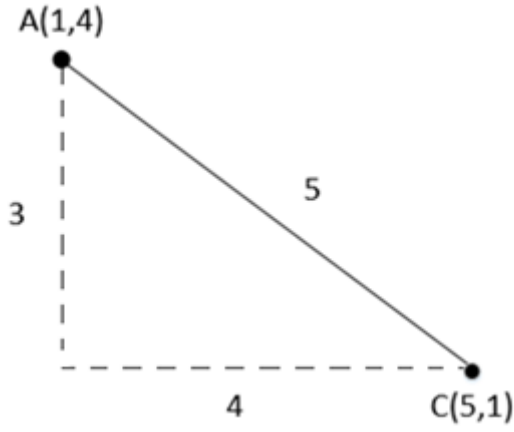

Рис. 1. Расстояние между точками в двумерном пространстве

С учетом представленных выше формул и того, что пространство является двумерным, расстояния между этими точками будут иметь следующие значения:

$$\rho_{\text{евклид}} = \sqrt{(5-1)^2+(4-1)^2} = 5;$$
$$\rho_{\text{кв.евклид}} = (5-1)^2+(4-1)^2 = 25;$$
$$\rho_{\text{гор.кв}} = |5-1| + |4-1| = 7;$$
$$\rho_{\text{чеб}} = \max(|5-1| + |4-1|) = 4.$$

Следующим этапом является определение мощности множества кластеров, то есть числа кластеров. В зависимости от целей и методов, применяемых при решении поставленной задачи, количество кластеров может быть либо не определено заранее, либо определено. Принято выделять три метода разбиения исходного множества на кластеры [5].

1. Иерархический метод: число кластеров заранее неизвестно. Предположения об их количестве делаются субъективно, на основе дендрограмм и/или динамики порога расщепления/слияния кластеров. Дендрограмма — это граф, отображающий связи между объектами исходного множества. В этой группе методов выделяют подвиды [6]:

– агломеративные методы: элементы исходного множества, которые сначала представляют собой отдельные кластеры, в дальнейшем объединяются в группы, тем самым уменьшая число кластеров до тех пор, пока не будет получен один- единственный кластер;

– дивизимые методы: изначально все элементы принадлежат одному- единственному кластеру, а с увеличением числа шагов разбиения количество кластеров увеличивается, то есть такие методы противоположны агломеративным.

2. Неиерархический: для достижения цели необходимо заранее определиться с результирующим количеством кластеров, а также с методами кластеризации.

3. Нечеткий метод: объекту исходной выборки не ставят в соответствие какой-либо конкретный кластер, а определяют степень принадлежности объекта к тому или иному кластеру.

Решение задачи кластеризации. Как было отмечено выше, кластерный анализ применяют в сфере экономики. В связи с этим принято решение провести кластеризацию на основе экономических показателей для 40 стран мира. Целью анализа является определение классов стран со схожими характеристиками и выявление свойств, оказывающих наибольшее влияние на каждый конкретный кластер. К важным экономическим показателям относят:

– уровень доходов (SAL);
– размер валового национального продукта (GNP);
– уровень инфляции (INFL);
– размер государственного долга (DEBT).

Часть исходных данных с нормализованными значениями приведена на рис. 2.

	Анализ экономики 40 стран мира			
	1 SAL	2 GNP	3 INFL	4 DEBT
Algeria	1,04	-1,5	0,9	1,43
Australia	-0,99	0,7	-0,9	-0,70
Austria	-1,03	0,8	-1,0	-0,85
Belgium	-0,83	0,8	-0,8	0,12
Britain	-0,70	1,2	-0,6	-1,03
Brazil	0,91	-0,7	0,8	1,03
China	0,21	-1,1	0,2	0,46
Czech Republic	0,21	-0,1	-0,2	-0,32
Denmark	-0,76	1,1	-0,9	-0,97
Egypt	1,78	-1,4	1,8	1,77
Ethiopia	1,82	-1,7	2,3	1,48
Finland	-0,76	0,3	-0,7	-0,81
France	-1,14	0,9	-1,1	-0,89
Germany	-1,10	1,4	-1,1	-1,00
Greece	0,44	-0,1	0,6	0,06
Haiti	2,40	-1,7	2,5	1,90
Hungary	0,00	-0,2	0,1	-0,61
India	0,95	-1,1	1,0	1,32
Indonesia	0,13	0,2	-0,2	0,24

Рис. 2. Нормализованные значения исходных данных анализа

Для анализа был выбран программный пакет STATISTICA, который предоставляет широкий спектр возможностей по кластерному анализу.

Программный математический пакет STATISTICA. Пакет STATISTICA американской компании StatSoft Inc. предоставляет мощные инструменты для решения статистических задач различного характера, в том числе и графического анализа данных. Его активно используют в рамках учебных программ во многих высших учебных заведениях России, например МГУ им. М.В. Ломоносова, НИУ ВШЭ, МГТУ им. Н.Э. Баумана, ННГУ им. Н.И. Лобачевского.

К основным преимуществам пакета STATISTICA относят [7]:

- удобный интерфейс;
- наличие одного из лучших графических модулей;
- высокую точность вычислений;
- простоту интеграции, совместимости и доступа к базам данных.

В STATISTICA кластеризация данных поддерживается в редакциях Data Miner и Advanced. Режим Data Miner предназначен для работы с большим объ-

емом данных и для автоматического определения числа кластеров, используя алгоритмы k-mean и EM. Режим Advanced включает в себя такие алгоритмы кластеризации, как joining/tree clustering, k-mean clustering и two-way joining.

Joining/tree clustering (объединение/древовидная кластеризация) — иерархический агломеративный метод, результатом применения которого является дендрограмма, определяющая возможное количество кластеров. Помимо выбора типа меры расстояния между кластерами, метод предполагает и более тонкую настройку путем выбора правил объединения данных в кластеры. Ниже представлены некоторые из них.

1. Правило одиночной связи: расстояние между двумя кластерами определяется как расстояние между двумя наиболее близкими объектами в различных кластерах. Результирующие кластеры имеют тенденцию объединяться в цепочки.

2. Правило полных связей: расстояния между кластерами определяются наибольшим расстоянием между любыми двумя объектами.

3. Правило невзвешенного попарного среднего: расстояние между двумя кластерами определяется как среднее расстояние между всеми парами объектов в них.

4. Правило взвешенного попарного среднего. Метод идентичен предыдущему, за исключением того, что при вычислении размер соответствующих кластеров используется в качестве весового коэффициента.

При анализе исходных данных выбранным методом были указаны следующие параметры: мера — квадрат евклидового расстояния, правило связи — полная связь. Эти значения позволяют получить более крупные объединения наблюдений в кластеры за счет достаточно большого значения расстояния объединения (Linkage Distance). Меняя параметры можно достигнуть уменьшения этого значения. Дендрограмма результатов кластеризации наблюдений представлена на рис. 3.

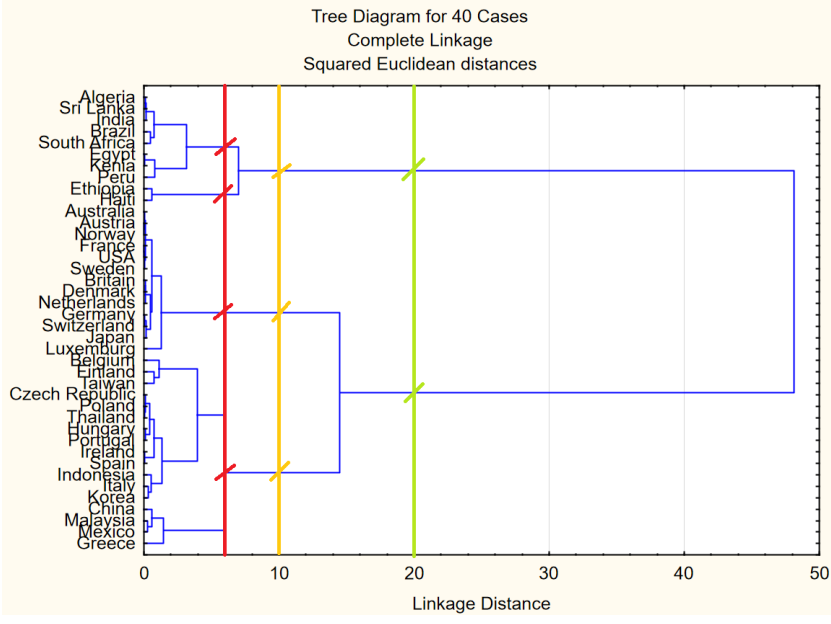

Рис. 3. Дендрограмма результатов кластеризации наблюдений

Как было отмечено выше, количество кластеров при использовании данного метода оценивается субъективно, однако для большей конкретики можно выбрать некоторое значение расстояния связи на оси Linkage Distance и провести перпендикуляр (количество точек пересечения соответствует числу кластеров). На рисунке 3 отмечены линии, с помощью которых можно четко определить число кластеров. На расстоянии 6 единиц число кластеров равно 4, 10 единиц — 3 кластерам, а 20 — всего 2.

Данный метод рекомендуется применять при первоначальной оценке данных, когда нет четкого представления о классах исходного множества. Однако важно понимать, что результат зависит от выбранной меры и правила объединения данных. При большом объеме данных, когда определение числа кластеров по дендрограмме может быть затруднено, следует воспользоваться описательными статистиками или графиком зависимости расстояния связи от числа шагов объединения.

K-mean clustering (алгоритм k-средних) — это неиерархический метод, предполагающий знание конечного числа групп. В результате исходное множество разбивается на k максимально удаленных друг от друга кластеров [8, 9]. Данный алгоритм основан на минимизации среднеквадратичного отклонения точек кластеров от центров этих кластеров. В качестве настроек здесь выступают число кластеров, итераций, а также способ выбора центра кластеров. Основными результатами являются график кластеров со средними значениями анализируемых параметров, данные по принадлежности элементов исходного множества к одному из кластеров, а также описательная информация по каждому кластеру (среднее, среднеквадратическое отклонение, дисперсия).

На основании результатов предыдущего метода, то есть представлении о возможном числе кластеров в системе (принято равным трем), с помощью метода k-средних были определены параметры, по которым кластеры существенно отличаются друг от друга (рис. 4).

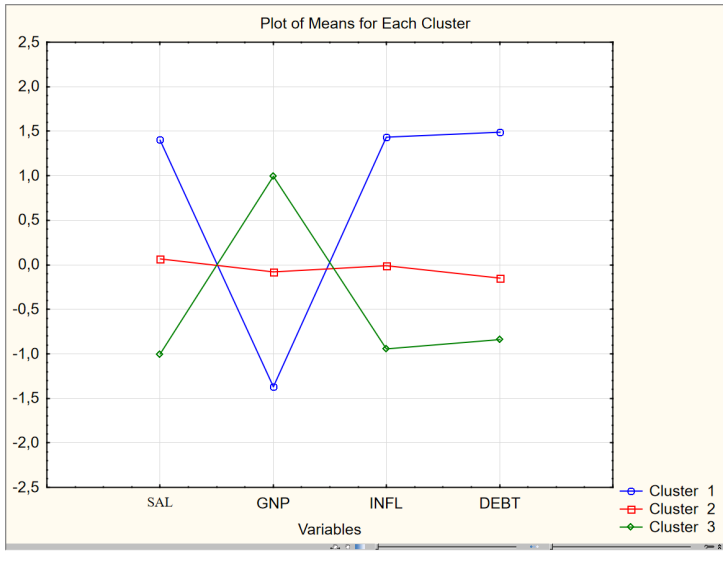

Рис. 4. Кластеризация наблюдений методом k-средних

Из рисунка видно, что выделенные кластеры существенно отличаются друг от друга. Для элементов кластера *1* характерны высокий размер заработной платы, инфляция, большой размер государственного долга, но при этом самый маленький объем валового национального продукта (Гаити, ЮАР, Бразилия). Кластер *3* обладает практически диаметральными характеристиками (Бельгия, Финляндия, Люксембург). Кластер *2* занимает промежуточное положение между *1* и *3* (Китай, Тайвань, Италия и Мексика).

Достоинствами алгоритма *k*-средних являются простота реализации и интуитивная понятность, недостатками — необходимость знать заранее число кластеров и зависимость результата от инициализации центров кластеров.

Tow-way joining (двухвходовое объединение) — это метод, который одновременно классифицирует элементы исходного множества и параметры, которыми они характеризуются. Имеются ограничения количества объектов наблюдения (cases), а также возможность более тонкой настройки путем указания порогового значения, влияющего на принадлежность элемента к кластеру. По умолчанию предлагается порог равный половине среднеквадратичного отклонения. Визуализация результатов анализа представлена на рис. 5, где по оси *X* расположены анализируемые параметры, по оси *Y* — анализируемые объекты, цвет на пересечении — принадлежность к определенному кластеру.

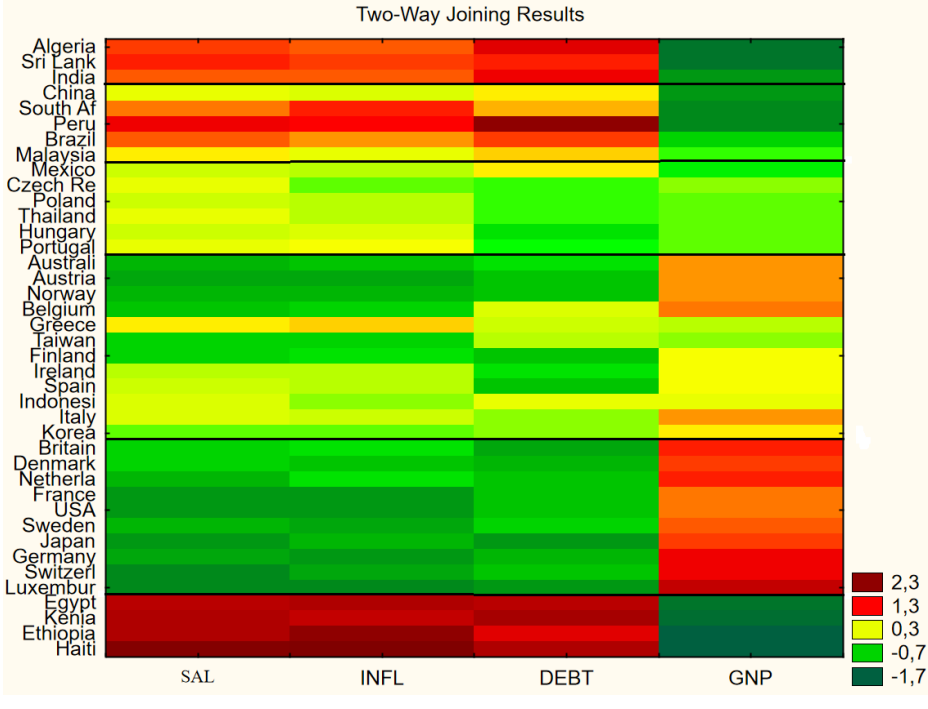

Рис. 5. Кластеризации наблюдений методом двухвходового объединения

Несмотря на некоторую сложность интерпретации данных вследствие того, что полученные кластеры зачастую неоднородны [10] и специфичны, можно эмпирически выделить кластеры наблюдений.

Как уже было отмечено, данный метод позволяет выделить группы данных и на основе параметров, и на основе наблюдений. Например, если кластеризовать исходное множество по параметру размера заработной платы (SAL), то можно выделить кластер с уровнем — 0,7 (Германия, Швеция, Бельгия, Австралия), если по наблюдениям (строкам) — группы кластеров (отделены горизонтальными линиями, см. рис. 5).

Выводы. Проведена кластеризация данных на конкретном примере тремя методами, представленными в редакции программного пакета STATISTICA Advanced. Данные методы принадлежат к разным классам и предназначены для решения разных типов задач: первый позволяет определить все возможные объединения наблюдений в кластеры, второй — сгруппировать наблюдения в соответствии с заранее известным число кластеров, третий — осуществить кластеризацию одновременно по наблюдениям и свойствам (параметрам). Рассмотренные методы можно использовать как по отдельности, так и вместе.

Литература

[1] Hartigan J.A. Clustering algorithms. John Wiley & Sons, Inc., 1975. 369 p.

[2] Дюран Б., Оделл П. *Кластерный анализ.* М.: Статистика, 1977. 128 с.

[3] *Анализ данных и процессов* / А.А. Барсегян, М.С. Куприянов, И.И. Холод, М.Д. Тесс, С.И. Елизаров. СПб.: БХВ-Петербург, 2009. 512 с.

[4] Калинина В.Н., Соловьев В.И. *Введение в многомерный статистический анализ.* М.: ГУУ, 2003. 66 с.

[5] *Прикладная статистика: Классификации и снижение размерности* / А.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д Мешалкин. М.: Финансы и статистика, 1989. 607 с.

[6] Методы кластерного анализа. Иерархические методы. URL: http://www.intuit.ru/studies/courses/ 6/6/lecture/182?page=2 (дата обращения 10.10.2017).

[7] Обзор STATISTICA. URL: http://statsoft.ru/products/overview/#advantages (дата обращения 25.09.2017).

[8] Data clustering: A Review A.K. Jain Michigan State University M.N. Murty // Indian Institute of Science AND P.J. FLYNN The Ohio State University. ACM Computing Surveys. Vol. 31. No. 3. September 1999. URL: http://users.eecs.northwestern.edu/~yingliu/datamining_papers/survey.pdf (дата обращения 26.09.2017).

[9] Smola Alex, Vishwanathan S.V.N. Introduction to machine learning. Cambridge University Press, 2008. 234 p.

[10] Буреева Н.Н. Многомерный статистический анализ с использованием ППП “STATISTICA”. Учебно-методический материал по программе повышения квалификации «Применение программных средств в научных исследованиях и преподавании математики и механики». Нижний Новгород, 2007. 112 с.

Тихонов Иван Андреевич — магистрант кафедры «Системы обработки информации и управления», МГТУ им. Н.Э. Баумана, Москва, Российская Федерация.

Научный руководитель — Виноградова Мария Валерьевна, кандидат технических наук, доцент кафедры «Системы обработки информации и управления», МГТУ им. Н.Э. Баумана, Москва, Российская Федерация.

AN OVERVIEW OF CAPABILITIES FOR CLUSTER ANALYSIS OF DATA FOUND IN THE STATISTICA ADVANCED SOFTWARE PACKAGE

I.A. Tikhonov

quiteman@mail.ru
SPIN-код: 5941-2719

Bauman Moscow State Technical University, Moscow, Russian Federation

Abstract

The article reviews data clustering capabilities of the STATISTICA software package. We describe the clustering methods found in this product and the specifics of working with them from a practical standpoint. We consider the concept of distance measure between elements of the initial set and certain methods of clustering the initial set of observations, as well as cluster analysis results produced by the algorithms implemented in the STATISTICA Advanced package. There is no doubt that cluster analysis of data is highly relevant and pertinent at present, since data and data analysis results play an increasingly significant role in the information society of today, and clustering provides a better understanding of these data.

Keywords

Data analysis, cluster analysis, clustering, unsupervised classification, STATISTICA, Euclidean space, hierarchical and non-hierarchical clustering methods, joining/tree clustering, k-mean clustering, two-way joining

References

[1] Hartigan J.A. Clustering algorithms. John Wiley & Sons, Inc., 1975. 369 p.

[2] Duran B., Odell P. Cluster analysis. A survey. Springer-Verlag. Berlin – Heidelberg – N.Y. 1974.

[3] Barsegyan, A.A., M.S. Kupriyanov, I.I. Kholod, M.D. Tess, S.I. Elizarov. Analiz dannykh i protsessov [Analysis of data and processes]. St. Petersburg, Peterburg Publ., 2009. 512 p.

[4] Kalinina V.N., Solov'ev V.I. Vvedenie v mnogomernyy statisticheskiy analiz [Introduction to multivariate statistical analysis]. Moscow, GUU Publ., 2003. 66 p.

[5] Ayvazyan A.A., Bukhshtaber V.M., Enyukov I.S., Meshalkin L.D. Prikladnaya statistika: Klassifikatsii i snizhenie razmernosti [Applied statistics: Classification and reduction of dimensionality]. Moscow, Finansy i statistika Publ., 1989. 607 p.

[6] Metody klasternogo analiza. Ierarkhicheskie metody. Available at: http://www.intuit.ru/studies/courses/6/6/lecture/182?page=2 (accessed 10.10.2017).

[7] Obzor STATISTICA. Available at: http://statsoft.ru/products/overview/#advantages (accessed 25.09.2017).

[8] Data clustering: A Review A.K. Jain Michigan State University M.N. Murty. *Indian Institute of Science AND P.J. FLYNN The Ohio State University. ACM Computing Surveys.* Vol. 31, no. 3, September 1999. Available at: http://users.eecs.northwestern.edu/~yingliu/datamining_papers/survey.pdf (accessed 26.09.2017).

[9] Alex Smola and S.V.N. Vishwanathan. Introduction to machine learning. Cambridge University Press, 2008. 234 p.

[10] Bureeva N.N. Mnogomernyy statisticheskiy analiz s ispol'zovaniem PPP “STATISTICA”. Uchebno-metodicheskiy material po programme povysheniya kvalifikatsii “Primenenie programmnykh sredstv v nauchnykh issledovaniyakh i prepodavanii matematiki i mekhaniki”. Nizhniy Novgorod, 2007. 112 p (in Russ.).

Tikhonov I.A. — Master's Degree student, Department of Information Processing and Control Systems, Bauman Moscow State Technical University, Moscow, Russian Federation.

Scientific advisor — M.V. Vinogradova, Cand. Sc. (Eng.), Assoc. Professor, Department of Information Processing and Control Systems, Bauman Moscow State Technical University, Moscow, Russian Federation.