Моделирование и анализ данных
2019. Том 09. № 4. С. 57–66
doi: 10.17759/mda.2019090404
ISSN: 2219-3758 (печатный)
ISSN: 2311-9454 (online)
© 2019 ФГБОУ ВО МГППУ

Modelling and Data Analysis
2019. Vol. 09, no. 4, pp. 57–66
doi: 10.17759/mda.2019090404
ISSN: 2219-3758 (print)
ISSN: 2311-9454 (online)
© 2019 Moscow State University of Psychology & Education

АНАЛИЗ ДАННЫХ

УДК 004.942, 331.101.5

Задачи анализа и прогнозирования деятельности IT компании с использованием методов машинного обучения

Алексейчук А.С.*
МАИ, Москва, Россия
alexejchuk@gmail.com

Виноградов В.И.**
МАИ, Москва, Россия
vvinogradov@inbox.ru

Данякин К.Д.****
МАИ, Москва, Россия,
kirill.danyakin@yandex.ru

В статье рассмотрены задачи применения методов машинного обучения, используемые для повышения эффективности бизнес-процессов работы с клиентами в IT-компании. Рассмотрено применение двух моделей машинного обучения: кластеризация клиентской базы и прогнозирование выручки.

Ключевые слова: машинное обучение, IT компания, сегментирование клиентской базы, прогнозирование.

Для цитаты:
Алексейчук А.С., Виноградов В.И., Данякин К.Д. Задачи анализа и прогнозирования деятельности IT компании с использованием методов машинного обучения // Моделирование и анализ данных. 2019. Том 09. № 4. С. 57–66. doi: 10.17759/mda.2019090404

****Алексейчук Андрей Сергеевич***, к.ф.-м.н., старший преподаватель каф. «Математическая кибернетика» Московского авиационного института (национального исследовательского университета), E-mail: alexejchuk@gmail.com

*****Виноградов Владимир Иванович,*** к.ф.-м.н., доцент каф. «Математическая кибернетика» Московского авиационного института (национального исследовательского университета), E-mail: vvinogradov@inbox.ru

******Данякин Кирилл Дмитриевич,*** студент, Московский авиационный институт (национальный исследовательский университет), E-mail: kirill.danyakin@yandex.ru

Введение

В данной статье описаны некоторые задачи анализа деятельности IT-компании методами машинного обучения для решения прикладных задач анализа данных по клиентам и продуктам, необходимых для повышения эффективности бизнес-процессов в IT-компании при работе с клиентами при продвижении программных продуктов и сервисов. Компания занимается разработкой и распространением электронных сервисов. В процессе ее работы был накоплен массив данных о работе компании со своими клиентами. Поэтому возникла возможность проведения анализа имеющихся данных методами машинного обучения для повышения эффективности работы компании на рынке.

В статье рассмотрены два примера использования методов машинного обучения – кластеризация клиентской базы и прогнозирование выручки от одной из групп распространяемых продуктов.

Методы и среда реализации

Для решения задач анализа данных в IT-компании используются следующие исходные данные и программные средства:

- база данных CRM на базе СУБД PostgreSQL и фреймворка Ruby on Rails, содержащая данные о 30 тыс. покупок различных программных продуктов и web-сервисов клиентами компании;
- интерпретатор Python 3.7;
- среда разработки и тестирования Jupyter Notebook;
- набор библиотек для анализа данных на Python: библиотека алгоритмов машинного обучения Scikit-Learn, библиотека численных вычислений Numpy, библиотека алгоритмов манипулирования многомерными данными Pandas, библиотека визуализации данных Matplotlib,
- RapidMiner Studio вер.7.3 – программная среда для изучения данных и машинного обучения.

Сегментация клиентской базы

Целью решения задачи сегментации клиентской базы является оптимизация работы с клиентами за счет выявления групп клиентов со схожими признаками и адресной работой с целевой аудиторией каждого сегмента (кластера). Клиенты, относящиеся к одному и тому же кластеру, имеют достаточно много общих признаков, что может позволить использовать для них общие методы клиентского обслуживания. Для каждого кластера могут быть назначены ответственные менеджеры, обученные специфике работы с клиентами, входящими в данный кластер. Экономический эффект достигается за счёт снижения затрат на работу с клиентами и за счёт улучшения показателя допродаж, поскольку структурирование клиентской базы позволяет вести работу с целевой аудиторией каждого сегмента.

Кластер – это объединение нескольких однородных элементов, которое может рассматриваться как самостоятельная единица, обладающая определёнными свойствами. Если данные понимать как точки в пространстве атрибутов, то задача кластерного анализа формулируется как выделение «сгущений» точек и разбиение совокупности имеющихся элементов на заданное число кластеров. Кластеризация – это пример обучения

без учителя, поскольку у входных данных отсутствуют метки, позволяющие построить модель кластеризации и экстраполировать ее на новые данные [1].

В качестве исходных данных для задачи сегментации клиентской базы используются записи базы данных системы CRM, используемой IT-компанией для организации работы с клиентами. В качестве атрибутов пространства признаков используется около 50 параметров, важнейшими среди которых являются:

- Сумма выручки за последний год по продуктам, разрабатываемым самой IT-компанией;
- Сумма выручки за последний год по продуктам, разрабатываемым компаниями-партнерами;
- Данные из бухгалтерского баланса клиента, такие как годовая выручка, прибыль или убытки компании-клиента, и данные из других открытых источников, например, место регистрации, количество сотрудников;
- Сведения о компании от его клиентского менеджера, такие как наличие и количество подразделений, сведения о характере деятельности клиента – например, организация на ОСНО, кредитная, бюджетная, страховая, иностранное представительство и др.

Для решения задачи кластеризации используются библиотеки scikit-learn для обработки набора данных и matplotlib для визуализации. Решению задачи кластеризации предшествует этап понижения размерности исходных данных при помощи метода главных компонент [2]. Это преобразование позволяет избавиться от сильно коррелирующих атрибутов, расположив оси координат пространства признаков вдоль направлений разброса точек исходной выборки с максимальной дисперсией.

Для осуществления кластеризации опробован ряд методов кластерного анализа: метод k-средних, DBSCAN, иерархическая кластеризация, BIRCH, affinity propagation [3]. Наилучший и относительно легко интерпретируемый результат показал метод k-средних. Этот метод состоит в разделении массива данных на заранее заданное количество кластеров k при помощи минимизации суммы квадратов расстояний от каждого элемента x_j до центроида μ_i кластера, в котором он находится:

$$\sum_{i=1}^{k} \sum_{x_j \in S} \left\| x_j - \mu_i \right\|^2 \to \min.$$

В качестве расстояния между элементами выборки используется метрика Евклида в пространстве признаков.

Фрагмент кода программы, осуществляющего кластеризацию, приведен на рис.1.

```
1 from sklearn.cluster import KMeans
2 kmeans = KMeans(n_clusters = 7, max_iter = 600)
3 kmeans.fit(data)
4 data['Cluster'] = kmeans.predict(data)
```

Рис. 1. Применение метода KMeans для кластеризации.

Переменная *data* содержит массив исходных данных в объекте класса *DataFrame* из библиотеки Pandas. *DataFrame* представляет собой двумерную таблицу, в которую помещаются данные различных типов после их выборки из базы данных. Метод *KMeans* осуществляет поиск заданного количества кластеров в массиве данных, после чего к таблице *data* добавляется столбец *Cluster*, содержащий метку кластера для каждого элемента выборки.

На рисунке 2 изображены условные границы полученных кластеров в системе координат, где по оси *x* отложена сумма покупок по продуктам, разрабатываемым самой IT-компанией, а по оси *y* – сумма покупок по продуктам партнеров. Кластеры 1 и 2 соответствуют клиентам с небольшой суммой покупок, кластер 6 – крупным клиентам, 5 – клиентам, которые покупают только продукты партнеров, 7 – клиентам, покупающим в основном продукты, разрабатываемые самой IT-компанией на крупные суммы, 3 и 4 – клиентам, занимающим промежуточное положение.

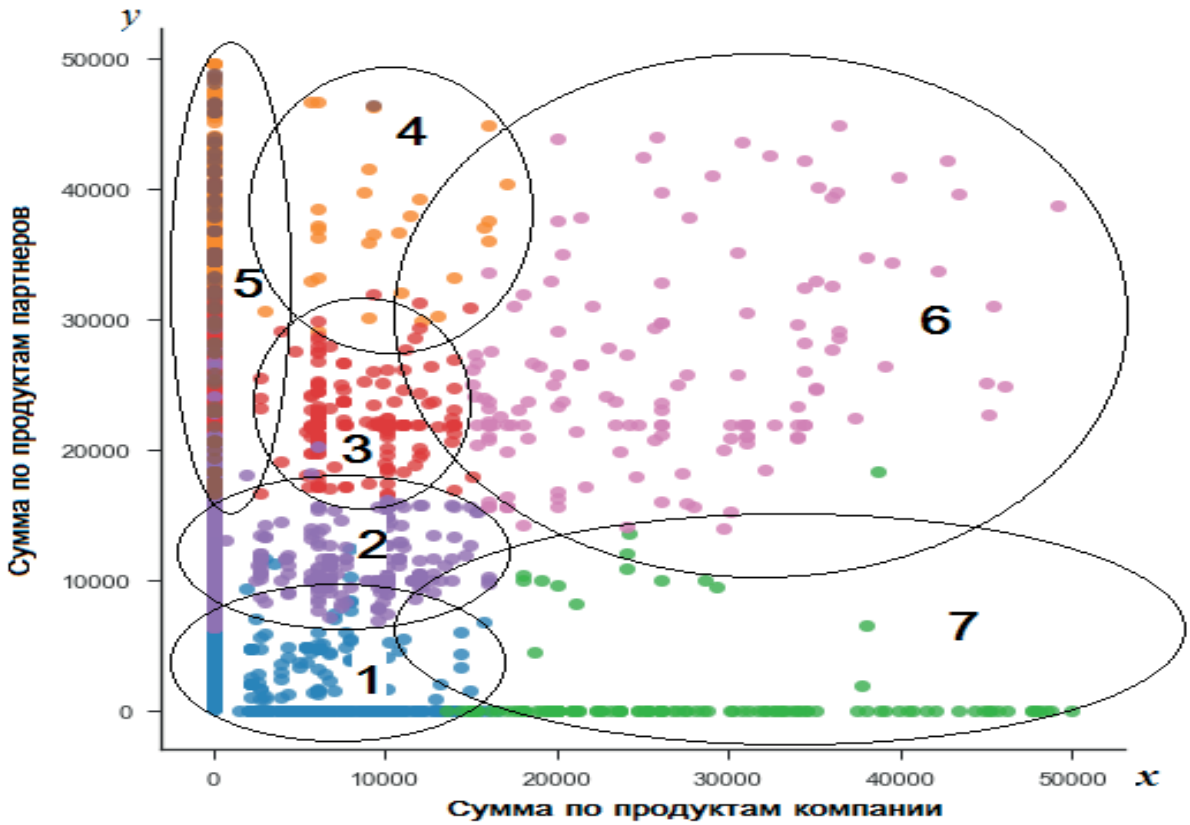

Рис. 2. Сегменты клиентской базы.

Выбор оптимального количества кластеров и оценка адекватности полученной модели могут быть произведены при помощи специальных критериев качества кластеризации (метрик), учитывающих плотность размещения элементов внутри кластеров в пространстве признаков и расстояние между соседними кластерами [4]. Метрики в той или иной мере отражают неформальные требования к результату процесса кластеризации, состоящие в том, что внутри кластеров объекты должны быть тесно связаны между собой, а объекты разных кластеров должны быть далеки друг от друга.

В данной работе используется метрика «silhouette» («силуэт»), определяемая следующим образом [5]. Пусть $a_{pj} = \frac{1}{n_{c_p} - 1} \sum_{x_k \in c_p} \left\| x_j - x_k \right\|$ – среднее расстояние от $x_j \in c_p$ до других объектов из кластера c_p; $d_{qj} = \frac{1}{n_{c_q}} \sum_{x_k \in c_q} \left\| x_j - x_k \right\|$ – среднее расстояние от x_j до объектов из другого кластера c_q, $q \neq p$. Положим $b_{pj} = \min_{q \neq p} d_{qj}$. Тогда «силуэт» эле-

мента с определяется как $S_{x_j} = \frac{b_{pj} - a_{pj}}{\max\left(a_{pj}, b_{pj}\right)}$, и оценка общего качества кластеризации вычисляется как среднее значение «силуэта» для всех N элементов выборки:

$$S = \frac{1}{N}\sum_{j=1}^{N} S_{x_j} .$$

Оптимальное количество кластеров соответствует достижению максимума значения метрики. Поскольку в методе k-means используется случайная инициализация кластеров, то при поиске оптимального количества кластеров используется усредненное значение метрики «силуэт» для 10 запусков алгоритма k-means для каждого варианта.

Поиск оптимального количества кластеров в массиве данных происходит следующим образом (рис. 3).

```
from sklearn import metrics
for i in range(3, 11):
    silhouette_score = 0
    for j in range(0, 10):
        kmeans = KMeans(n_clusters = i, max_iter = 600)
        kmeans.fit(data)
        data['Cluster'] = kmeans.predict(data)
        silhouette_score += metrics.silhouette_score(data, data['Cluster'], metric = 'euclidean')
    silhouette_score /= 10
```

```
For 3 clusters, silhouette metric is 0.5343945542162297
For 4 clusters, silhouette metric is 0.5280479127353657
For 5 clusters, silhouette metric is 0.5344349213547583
For 6 clusters, silhouette metric is 0.5374041696365525
For 7 clusters, silhouette metric is 0.5440603420263154
For 8 clusters, silhouette metric is 0.5184025109925108
For 9 clusters, silhouette metric is 0.5131951863447509
```

Рис. 3. Определение количества кластеров.

На примере видно, что для данной выборки максимум значения метрики достигается при 7 кластерах. Однако значения метрики для другого количества кластеров отличаются не сильно, что говорит о размытости границ кластеров и подверженности их изменениям.

Прогнозирование выручки от распространения электронных сервисов

Рассматриваются данные о выручке IT-компании по определенной группе продуктов по месяцам за несколько лет. Эти данные представлены таблицей, содержащей два столбца: первый столбец содержит информацию о месяце и годе, а второй – сумму выручки за соответствующий месяц.

Сформулируем задачу прогнозирования: на основе имеющихся данных построить и применить модель долгосрочного прогнозирования (больше одного года) для получения информации о потенциальной выручке в будущих периодах.

Для решения поставленной задачи необходимо, как и в любой задаче анализа данных, сначала подготовить данные, например, заменив средним или исключив элементы с пустым значением выручки.

Следующий этап состоит в том, чтобы определиться с моделью обучения. В рассматриваемой задаче используется модель Хольта-Винтерса, которая применяет метод тройного экспоненциального сглаживания в задаче прогнозирования временных рядов. В данной модели имеются три параметра для сглаживания уровня, тренда и сезонности.

Далее необходимо оптимизировать параметры модели. Сделать это можно перебирая возможные комбинации параметров с целью минимизации функции потерь. За функцию потерь была взята среднеквадратическая ошибка между предсказанным значением и истинным. Оптимизация происходит при помощи оценивания функции потерь при рассматриваемых параметрах на кросс-валидации. Стандартный процесс кросс-валидации для временных рядов не применим, так как при случайном перемешивании значений теряется временная структура ряда. Поэтому была использована кросс-валидация на скользящем окне (sliding window validation). Ее принцип состоит в следующем: происходит обучение модели на небольшом отрезке ряда от его начала до некоторого значения t, затем строится прогноз на ***t+n*** шагов вперед и рассчитывается ошибка, после чего выборка расширяется до значения ***t+n*** и производится прогноз от ***t+n*** до ***t+2•n***. Таким образом осуществляется движение тестового отрезка до последнего доступного наблюдения.

После оптимизации параметры, при которых наблюдался минимум функции потерь (среднеквадратической ошибки), можно подставить в модель Хольта-Винтерса, получив таким образом модель прогнозирования. Решение задачи проводилось в среде RapidMiner, в которой для решения задачи прогнозирования была построена следующая диаграмма процесса прогнозирования. На диаграмме отображены этапы процесса, включающие, предобработку данных, оптимизации параметров модели прогнозирования, построение модели прогнозирования, ее применение и отображение результатов.

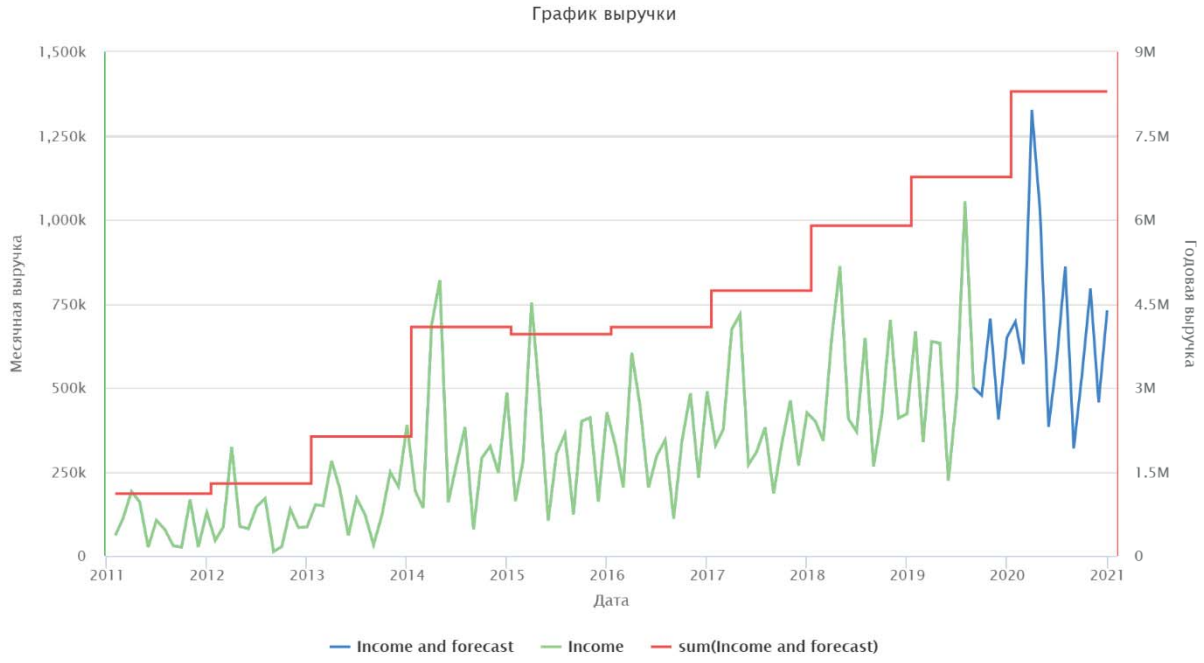

Рис. 4. Диаграмма процесса прогнозирования в среде Rapidminer

Построенная модель позволяет получить предсказание ежемесячной и годовой выручки на промежуток времени, включающий следующий полный календарный год. На рис. 5 отображены график и прогноз ежемесячной (income и income and forecast) и годовой (sum) выручки.

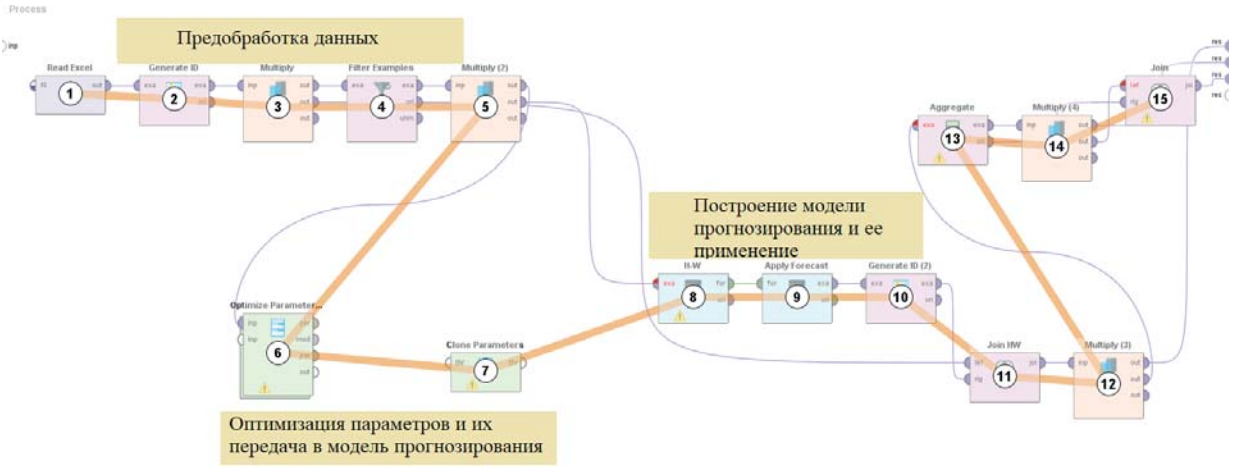

Рис. 5. График и прогноз выручки

Построенный прогноз говорит о квартальной периодичности величины выручки и устойчивом ее росте за последние 3 года и ожидаемом прогнозе сохранения такой динамики в следующем году. Построенная модель может ежемесячно пополняться сведениями о полученной выручке и осуществлять прогнозирование на следующие периоды, что позволит контролировать успешность процесса продвижения электронных сервисов в IT-компании.

Заключение

В статье показан пример использования методов машинного обучения для повышения эффективности деятельности IT-компании, приведены примеры программного кода на языке Python и диаграммы процесса анализа в среде RapidMiner, а также результаты расчетов. В результате применения алгоритмов машинного обучения в IT-компании выявлены закономерности в структуре базы клиентов и в структуре их покупательской корзины, которые могут быть понятны для интерпретации их маркетологами компании на практике. На основе алгоритмов машинного обучения разработаны программные модули, позволяющие в дальнейшем использовать их для повышения эффективности деятельности.

Реализация алгоритмов Data Mining встроена в CRM, что позволяет использовать их маркетологами и клиентскими менеджерами в их повседневной работе. В процессе деятельности компании информация о клиентах и их покупках регулярно обновляется, поэтому периодически производится перерасчет обновленных данных по приведенным алгоритмам. Это позволяет поддерживать актуальность структуры сегментов клиентов и наборов ассоциативных правил.

Благодарности

Авторы благодарят за предоставленные данные генерального директора компании «ОВИОНТ ИНФОРМ» Артамонова Г.Ф.

Литература

1. *Dubes R.C., Jain A.K.* Algorithms for Clustering Data. Englewood Cliffs: Prentice Hall, 1988.
2. *Shlens J.* A tutorial on principal component analysis. Institute for Nonlinear Science, UCSD, 2005.

3. *Rui Xu, Wunsch D.* Survey of clustering algorithms. IEEE Transactions on Neural Networks, vol. 16, no. 3, 2005. pp. 645–678.
4. *Wang L., Leckie C., Ramamohanarao K., Bezdek J.* Automatically Determining the Number of Clusters in Unlabeled Data Sets. IEEE Transactions on Knowledge and Data Engineering, vol. 21, 2009. p. 335–350.
5. *Rousseeuw Peter J.* Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. Computational and Applied Mathematics, vol. 20, 1987. p. 53–65.
6. *Лукашин Ю.П. А*даптивные методы краткосрочного прогнозирования временных рядов. – М.: Финансы и статистика, 2003.

The tasks of analysis and forecasting the activities of IT companies using Machine Learning methods

Aleksaychuk A.S.*
MAI, Moscow, Russia
alexejchuk@gmail.com

Vinogradov V.I.**
MAI, Moscow, Russia
vvinogradov@inbox.ru

Danyakin K.D.****
MAI, Moscow, Russia
kirill.danyakin@yandex.ru

The article describes applying machine learning methods for improving the efficiency of business processes when working with clients in an IT company. Two models of machine learning are considered: clustering the customer base and revenue forecasting.

Keywords: Machine Learning, IT Company, customer base segmentation, forecasting.

Acknowledgements
The authors are grateful to G.F. Artamonov, General Director of OVIONT INFORM, for the data provided.

References

1. Dubes R.C., Jain A.K. Algorithms for Clustering Data. Englewood Cliffs: Prentice Hall, 1988.
2. Shlens J. A tutorial on principal component analysis. Institute for Nonlinear Science, UCSD, 2005.
3. Rui Xu, Wunsch D. Survey of clustering algorithms. IEEE Transactions on Neural Networks, vol. 16, no. 3, 2005. pp. 645–678.

For citation:
Aleksaychuk A.S., Vinogradov V.I., Danyakin K.D. The tasks of analysis and forecasting the activities of IT companies using Machine Learning methods. *Modelirovanie i analiz dannykh = Modelling and Data Analysis*, 2019. Vol. 09, no. 4, pp. 57–66. doi: 10.17759/mda.2019090404 (In Russ., abstr. in Engl.)

***Aleksaychuk Andrej Sergeevich,** Ph.D., Senior Lecturer, Moscow Aviation Institute (National Research University), Moscow, Russia. E-mail: alexejchuk@gmail.com

****Vinogradov Vladimir Ivanovich,** Ph.D., Associate Professor, Moscow Aviation Institute (National Research University), Moscow, Russia. E-mail: vvinogradov@inbox.ru

*****Danyakin Kirill Dmitrievich,** Student, Moscow Aviation Institute (National Research University), Moscow, Russia. E-mail: kirill.danyakin@yandex.ru

4. Wang L., Leckie C., Ramamohanarao K., Bezdek J. Automatically Determining the Number of Clusters in Unlabeled Data Sets. IEEE Transactions on Knowledge and Data Engineering, vol. 21, 2009. p. 335–350.
5. Rousseeuw Peter J. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. Computational and Applied Mathematics, vol. 20, 1987. p. 53–65.
6. Lukashin Yu.P. Adaptivnye metody kratkosrochnogo prognozirovaniya vremennyh ryadov. [Adaptive methods of short-term forecasting of time series.] – M.: Finansy i statistika, 2003. (In Russ., Abstr. in Engl.)