Для цитирования: *Глазкова А.В. Оценка результативности применения расстояний Евклида и Махаланобиса для решения одной из задач классификации текстов. Вестник Дагестанского государственного технического университета. Технические науки. 2017; 44 (1):86-93. DOI:10.21822/2073-6185-2017-44-1-86-93*
For citation: *Glazkova A.V. Efficiency assessment of Euclidean and Makhalanobis distances for solving a major text classification problem. Herald of Dagestan State Technical University. Technical Sciences. 2017; 44 (4):86-93. (In Russ.) DOI:10.21822/2073-6185-2017-44-1-86-93*

ТЕХНИЧЕСКИЕ НАУКИ
ИНФОРМАТИКА, ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА И УПРАВЛЕНИЕ

УДК 004.912
DOI:10.21822/2073-6185-2017-44-1-86-93

ОЦЕНКА РЕЗУЛЬТАТИВНОСТИ ПРИМЕНЕНИЯ РАССТОЯНИЙ ЕВКЛИДА И МАХАЛАНОБИСА ДЛЯ РЕШЕНИЯ ОДНОЙ ИЗ ЗАДАЧ КЛАССИФИКАЦИИ ТЕКСТОВ

Глазкова А.В.
Тюменский государственный университет,
625003, г. Тюмень, ул. Перекопская, д. 15а,
e-mail: anna_glazkova@yahoo.com

Резюме: Цель. *Целью работы является проведение сравнения эффективности применения метрик Евклида и Махаланобиса для решения задачи определения категории потенциальных адресатов текста. Актуальность поставленной задачи определена необходимостью развития средств идентификации адресата электронного документа, возросшей в связи с введением возрастных ограничений на контент интернет-страниц и содержимое текстовых ресурсов, а также малой освещенностью данной проблемы в работах российских исследователей.* ***Метод.*** *Сравнение эффективности использования расстояний Евклида и Махаланобиса проведено в рамках реализации интеллектуальной системы автоматической классификации текстов на основании возрастной категории их адресатов.* ***Результат.*** *Рассмотрены основные подходы к установлению меры близости объектов, представленных в виде наборов классификационных признаков, а также обоснован выбор метрик Евклида и Махаланобиса для проведения численного сравнения результатов классификации. Приведено описание выборок текстов, предоставленных для вычислительного эксперимента, и классификационных признаков, характеризующих категории. Проведен вычислительный эксперимент с использованием текстов, входящих в состав Национального корпуса русского языка.* ***Вывод.*** *Вычислительный эксперимент позволяет выбрать наиболее эффективный метод решения задачи определения возрастной категории потенциальных адресатов текста. Результаты эксперимента показали возможность использования метрик Евклида и Махаланобиса для решения задач классификации текстов, а также подтвердили предпочтительность использования метрики Махаланобиса для оценивания расстояний объектами, представленными коррелированными признаками. Представленное сравнение проведено в рамках реализации интеллектуальной системы автоматической классификации текстов на основании возрастной категории их адресатов.*

Ключевые слова: *расстояние Евклида; расстояние Махаланобиса; классификация документов; обработка естественного языка; характеристики текста; текст; классификационный признак*

TECHICAL SCIENCE
COMPUTER SCIENCE, COMPUTER ENGINEERING AND MANAGEMENT

EFFICIENCY ASSESSMENT OF EUCLIDEAN AND MAKHALANOBIS DISTANCES FOR SOLVING A MAJOR TEXT CLASSIFICATION PROBLEM

Glazkova A.V.
Tyumen State University
15a Perekopskaya Str., Tyumen 625003, Russia,
e-mail: anna_glazkova@yahoo.com

Abstract. Objectives *The aim is to compare the efficiency of using the Euclidean and Mahalanobis metrics to solve the problem of determining the category of potential text recipients. The relevance of the task is determined by the need to develop a means of identifying the recipients of electronic documents. This has been complicated with the introduction of age restrictions on the content of Internet webpages and text resources. Moreover, there has been little coverage of this issue in the works of Russian researchers.* ***Method*** *A comparison of the relative efficiencies of using Euclid and Mahalanobis distances was carried out within the framework of the implementation of an intelligent system for text automatic classification based on the age category of their recipients.* ***Results*** *The main approaches to establishing proximity measures of objects represented as sets of classification characteristics are discussed and the choice of Euclidean and Mahalanobis metrics for numerical comparison of classification results is justified. A description of the sample texts and characteristics of category designations are given for a computational experiment. The computational experiment was carried out using texts included in the National Corpus of the Russian language.* ***Conclusion*** *The computational experiment allows the most effective method for solving the problem of determining the age category of potential text recipients to be selected. The results of the experiment showed the possibility of using Euclidean and Mahalanobis metrics for solving text classification problems; the preference for using Mahalanobis metrics for estimating distances by objects represented by correlated features was also confirmed. The presented comparison of the relative efficiencies of Euclid and Mahalanobis distances was carried out within the framework of the implementation of an intelligent system for automatic text classification based on the age category of their recipients.*

Keywords: *Euclidean distance, Mahalanobis distance, document classification, natural language processing, text characteristics, text, classification feature*

Введение. Решение вопросов обработки текстов на естественном языке является важным направлением развития информационного поиска [1-2]. Актуальными проблемами классификации естественно-языковых текстов являются идентификация автора и адресата текста. Механизмы решения данных задач применяются при создании диалоговых и поисковых систем, систем электронного обучения и фильтрации спама.

Проблемам атрибуции (установлению авторства) текста посвящены работы многих российских и зарубежных учёных (в частности, [3-5]). Вопрос определения характеристик адресата текста в настоящее время является менее освещенным и затрагивается преимущественно зарубежными исследователями (работы [6-7]). В то же время задача идентификации адресата текста приобретает высокую актуальность в связи с введением возрастных ограничений на контент интернет-страниц и содержимое текстовых ресурсов.

Важным этапом построения классификатора является создание набора информативных признаков [8-9]. На основании полученного набора признаков проводится разбиение объектов обучающей выборки и обучение классификатора, использующее детерминированные линейные методы [3; 10] или нелинейные методы, построенные на использовании деревьев решений и нейронных сетей [11-12]. Преимущество детерминированных методов состоит в большей прозрачности процесса классификации, что создает возможность пользователю системы

классификации проанализировать степень зависимости результатов от значений различных классификационных признаков.

Постановка задачи. В данной работе рассматривается задача определения *возрастной* категории потенциальных адресатов текста.

Набор классификационных признаков, характеризующих тексты, предназначенные различным возрастным категориям читателей, с большой долей вероятности имеет следующие особенности: значения признаков могут быть представлены в интервальной шкале;

признаки являются коррелированными (в частности, значения признаков «Средняя длина слов текста» и «Количество многосложных слов в тексте» связаны между собой).

На основании предположения о существовании перечисленных особенностей были рассмотрены существующие пути вычисления меры близости объектов. В рамках решения задачи классификации часто используются понятия меры близости, характеризующие взаимное расположение классов и расстояния между объектами, подлежащими классификации. Количественная оценка сходства объектов связана с понятием метрики. При этом объекты представляются в виде точек координатного пространства, а размерность пространства определяется количеством признаков, использованных для описания объектов.

В табл. 1 в обобщенном виде приводится перечень наиболее часто применяемых метрик и коэффициентов ассоциативности, используемых для установления меры близости объектов, описанных бинарными переменными [13-14].

Таблица 1. Методы установления меры близости объектов
Table 1. Methods for establishing the proximity of objects

Мера близости	Шкала измерения признаков	Примечание
Евклидово расстояние	Количественные шкалы	Представляет собой геометрическое расстояние в многомерном пространстве признаков.
Квадратичное евклидово расстояние	Количественные шкалы	Придает большие веса расстояниям между более отдаленными объектами.
Расстояние Чебышева	Количественные шкалы	Позволяет определить различность двух объектов, отличающихся по одному признаку.
Манхэттенское расстояние	Количественные шкалы	В сравнении с евклидовым расстоянием влияние отдельных больших разностей уменьшается.
Расстояние Махаланобиса	Количественные шкалы	Применяется в случае ненулевой корреляции переменных.
Процент несогласия	Качественные шкалы	Применяется в случае, когда признаки, характеризующие объект, являются категориальными.
Простой коэффициент совстречаемости	Номинальная (бинарная) шкала	Учитывает одновременное отсутствие признака у рассматриваемых объектов.
Коэффициент Жаккара	Номинальная (бинарная) шкала	Не учитывает одновременного отсутствия признака.
Коэффициент Гауэра	Качественные и количественные шкалы	Допускает одновременное использование переменных, измеренных по различным шкалам.

Методы исследования. Выбор метода оценивания меры близости является важным моментом исследования, влияющим на результат классификации объектов и зависящим от конкретной решаемой задачи. В данной работе, исходя из особенностей поставленной задачи, а также предположений о составе набора классификационных признаков, для вычисления меры

близости текстов были выбраны расстояние Евклида и расстояние Махаланобиса. Обе меры близости неоднократно применялись для решения задач классификации [15-17] и в зависимости от условий постановки задачи демонстрировали ту или иную степень предпочтительности своего использования.

Расчет расстояния Евклида проводился по классической формуле вычисления меры близости объектов, представленных точками в многомерном пространстве:

$$\rho_E(x,y)=\sqrt{\sum_{i=1}^{k}(x_i-y_i)^2}, \quad (1)$$

где x_i, y_i — значения i-го признака объектов x и y;

k – общее количество признаков.

Для расчета расстояния Махаланобиса использовалась формула:

$$\rho_{Mx}(x,y)=\sqrt{(x-y)^T S^{-1}(x-y)}, \quad (2)$$

где x_i, y_i — значения i-го признака объектов x и y;

S – матрица ковариаций.

Корпус текстов. В ходе вычислительного эксперимента использовались база данных «Морфологический стандарт Национального корпуса русского языка» и «База данных метатекстовой разметки Национального корпуса русского языка» (коллекция детской литературы)» [18].

Тексты, составляющие Национальный корпус русского языка [19], размечены по различным лингвистическим параметрам.

Базы содержат заведомо качественные и максимально разнообразные тексты на русском языке, возрастная категория потенциальных читателей которых – взрослая или детская – определена на основании мнений экспертов. Объем выборки – 532 текста художественной литературы и 510 текстов детской литературы. В базах данных представлены тексты 372 авторов. Средняя длина в корпусе составляет 471 слово.

В исследовании, в соответствии с выборкой, предоставленной для эксперимента, используется деление текстов на детские и взрослые.

Набор классификационных признаков. Анализ данных показал возможность использования следующего набора классификационных признаков [20]:

- средняя длина слов текста (кроме стоп-слов);
- среднее количество слов в предложении;
- количество многосложных слов в тексте (более трех слогов, %);
- количество особых глагольных форм в тексте (%);
- среднее количество грамматических основ в предложении;
- количество числительных в тексте (%);
- доля простых предложений с двумя главными членами (относительно простых предложений, %);
- доля служебных слов (%);
- количество глаголов в тексте (%);
- количество прилагательных в тексте (%).

Для приведения значений к единому диапазону было дополнительно проведено нормирование.

Обсуждение результатов. Вычислительный эксперимент. Целью вычислительного эксперимента явилось сравнение эффективности использования метрик Евклида и Махаланобиса для определения категории текстов. Каждому тексту тестовой выборки было сопоставлено признаковое описание – набор значений признаков и их весовых коэффициентов. Под расстоянием от текста до категории подразумевалось расстояние от набора (вектора) признаков, характеризующих текст, до центра масс категории.

Обучение классификаторов проводилось на 75% текстов имеющихся выборок, тестирование проводилось на оставшихся 25%. После проведения *n* разбиений исходной выборки на обучающую и контрольную (*n*=5) были вычислены средние значения по всем разбиениям.

Результатом классификации является процент правильно классифицированных записей на контрольной выборке. Число распознаваний с использованием метрики Махаланобиса составило 74,16% (среднеквадратическое отклонение – 5,88%), с использованием метрики Евклида – 68,42% (среднеквадратическое отклонение – 5,38%).

Использование других мер близости, подходящих для решения задач классификации объектов, представленных в виде векторов количественных признаков, – манхэттенского расстояния и квадратичного евклидова расстояния – показывает результаты, сравнимые с результатом применения метрики Евклида. Это обосновано тем, что формулы расчета данных метрик являются модифицированными формулами вычисления расстояния Евклида.

Сравнение с расстоянием Чебышева и коэффициентом Гауэра не проводилось. Это обусловлено тем, что расстояние Чебышева высчитывается как абсолютное значение максимальной разности последовательных пар значений признаков, характеризующих тексты. То есть оно применимо в случае, когда необходимо определить два объекта как различные, исходя из значений одного признака [21].

Коэффициенты ассоциативности, в отличие от мер сходства, предназначены для сравнения объекта не с эталоном, а для определения некой взаимной упорядоченности объектов [13]. Для проведения же классификации по известным классам необходимо вычисление именно меры близости объекта с эталоном, то есть с центроидом класса.

Вывод. Описанный в работе вычислительный эксперимент призван определить наиболее эффективный метод решения задачи определения возрастной категории потенциальных адресатов текста. Результаты эксперимента показали возможность использования метрик Евклида и Махаланобиса для решения поставленной задачи, а также подтвердили предпочтительность использования метрики Махаланобиса для оценивания расстояний объектами, представленными коррелированными признаками.

Представленное сравнение проведено в рамках реализации интеллектуальной системы автоматической классификации текстов на основании возрастной категории их адресатов. Говоря о сложности сравнимых алгоритмов оценки близости текстов, можно отметить, что время выполнения классификации с использованием метрики Евклида составляет O(n), в то время сложность алгоритма классификации с применением расстояния Махаланобиса равна $O(n^2)$, где n – количество классификационных признаков.

Указанное различие обусловлено необходимостью расчета матрицы ковариации классификационных признаков при вычислении метрики Махаланобиса. Однако, поскольку в реальных задачах количество признаков, как правило, не превышает десяти-двадцати, на практике различие во времени выполнения классификации с использованием расстояний Евклида и Махаланобиса не проявляется.

Библиографический список:

1. Кадиев, П.А. Пакет программ для скремблирования информационного потока / П.А. Кадиев, И.П. Кадиев, Т.М. Мирзабеков // Вестник Дагестанского государственного технического университета. Технические науки. – 2016. – № 2. – С. 83-92.
2. Шихиев, Ф.Ш. Графовая модель синтаксиса / Ф.Ш. Шихиев // Вестник Дагестанского государственного технического университета. Техническиенауки. – 2012. – № 25. – С. 32-37.
3. Nguyen, D. Author Age Prediction from Text using Linear Regression / D. Nguyen, N. Smith, C. Rose // Proc. of ICASSP. – New-York, 2011. – P. 267-276.
4. Кубарев, А.И. Сравнительный анализ эффективности распознавания авторского стиля текстов различными классификаторами / А.И. Кубарев, К.А. Михалева, В.В. Поддубный // Известия высших учебных заведений. Физика. – 2015. – Т. 58. № 11-2. – С. 252-258.

5. Муха, А.В. Автоматизированный подход к определению авторства текста / А.В. Муха, В.Л. Розалиев, Ю.А. Орлова, А.В. Заболеева-Зотова // Известия Волгоградского государственного технического университета. – 2013. – Т. 17. № 14 (117). – С. 51-54.
6. Akker, R. A comparison of addressee detection methods for multiparty conversations / R. Akker, D. Traum // Proc. of methods for multiparty conversations. – Amsterdam, 2009. – P. 99-106.
7. Choi, D. Text Analysis for Detecting Terrorism-Related Articles on the Web / D. Choi, B. Ko, H. Kim, P. Kim // Journal of Network and Computer Applications. – 2013. – Vol. 8, №5. – P. 37-46.
8. Колесникова, С.И. Методы анализа информативности разнотипных признаков / С.И. Колесникова // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. – 2009. – №1(6). – С. 69-80.
9. Поляков, И.В. Проблема классификации текстов и дифференцирующие признаки/ И.В. Поляков, Т.В. Соколова, А.А. Чеповский, А.М. Чеповский // Вестник Новосибирского государственного университета. Серия: Информационные технологии. – 2015. – Т. 13. № 2. – С. 55-63.
10. Толчеев, В.О. Модифицированный и обобщенный метод ближайшего соседа для классификации библиографических текстовых документов / В.О. Толчеев // Заводская лаборатория. Диагностика материалов. – 2009. – №7. – С. 63-70.
11. Мешкова, Е.В. Методика построения классификатора текста на основе гибридной нейросетевой модели / Е.В. Мешкова // Известия ЮФУ. Технические науки. – 2008. – № 4 (81). – С. 212-215.
12. Козоброд, А.В. Анализ архитектур гибридных нейросетевых моделей в задачах автоматической классификации текстовой информации / А.В. Козоброд, В.Е. Мешков, Е.В. Мешкова // Известия ЮФУ. Технические науки. – 2010. – № 12 (113). – С. 185-190.
13. Ким, Дж.-О. Факторный, дискриминантный и кластерный анализ: Пер. с англ. / Дж.-О. Ким, Ч.У. Мьюллер, У.Р. Клекка, М.С. Олдендерфер, Р.К. Блэшфилд. – М.: Финансы и статистика, 1989. – 215 с.
14. Хачумов, М.В. Расстояния, метрики и кластерный анализ / М.В. Хачумов // Искусственный интеллект и принятие решений. – 2012. – №1. – С. 81-89.
15. Толмачев, И.Л. Бинарная классификация на основе варьирования размерности пространства признаков и выбора эффективной метрики / И.Л. Толмачев, М.В. Хачумов // Искусственный интеллект и принятие решений. – 2010. – №2. – С. 3-10.
16. Хачумов, М.В. Применение нейрона и расстояния Евклида-Махаланобиса в задаче бинарной классификации / М.В. Хачумов // Наука и современность. – 2010. – №2-3. – С. 82-86.
17. Шумская, А.О. Оценка эффективности метрик расстояния Евклида и расстояния Махаланобиса в задачах идентификации происхождения текста / А.О. Шумская // Доклады Томского государственного университета систем управления и радиоэлектроники. – 2013. – №3 (29). – С. 141-145.
18. «База данных метатекстовой разметки Национального корпуса русского языка» (коллекция детской литературы)». 2014.
19. Национальный корпус русского языка [Электронный ресурс]. 2015. URL: http:// ruscorpora.ru/ (дата обращения: 26.07.2016).
20. Глазкова, А.В. Проверка информативности классификационных признаков в задаче автоматической классификации текстов на естественном языке / А.В. Глазкова // Открытые семантические технологии проектирования интеллектуальных систем (OSTIS-2015): материалы конференции. – 2015. – С. 541-544.
21. Буреева, Н.Н. Многомерный статистический анализ с использованием ППП “STATISTICA” / Н.Н. Буреева. – Нижний Новгород: Нижегородский государственный университет им. Н.И. Лобачевского, 2007. – 112 с.

References:

1. Kadiev P.A., Kadiev I.P., Mirzabekov T.M. Paket programm dlya skremblirovaniya informatsionnogo potoka. Vestnik Dagestanskogo gosudarstvennogo tekhnicheskogo universiteta. Tekhnicheskie nauki. 2016; 2:83-92. [Kadiev P.A., Kadiev I.P., Mirzabekov T.M. Software package for scrambling the information flow. Herald of Daghestan State Technical University. Technical Sciences. 2016; 2:83-92. (in Russ.)]
2. Shikhiev F.Sh. Grafovaya model' sintaksisa. Vestnik Dagestanskogo gosudarstvennogo tekhnicheskogo universiteta. Tekhnicheskie nauki. 2012; 25:32-37. [Shikhiev F.Sh. Graph model of syntax. Herald of Daghestan State Technical University. Technical Sciences. 2012; 25:32-37. (in Russ.)]
3. Nguyen D., Smith N., Rose C. Author Age Prediction from Text using Linear Regression. Proc. of ICASSP. New-York; 2011. P. 267-276.
4. Kubarev A.I., Mikhaleva K.A., Poddubnyy V.V. Sravnitel'nyy analiz effektivnosti raspoznavaniya avtorskogo stilya tekstov razlichnymi klassifikatorami. Izvestiya vysshikh uchebnykh zavedeniy. Fizika. 2015; 58(11-2):252-258. [Kubarev A.I., Mikhaleva K.A., Poddubnyy V.V. Comparative analysis of efficiency of author's style recognition of texts by various classifiers. Russian Physics Journal. 2015; 58(11-2):252-258. (in Russ.)]
5. Mukha A.V., Rozaliev V.L., Orlova Yu.A., Zaboleeva-Zotova A.V. Avtomatizirovannyy podkhod k opredeleniyu avtorstva teksta. Izvestiya Volgogradskogo gosudarstvennogo tekhnicheskogo universiteta. 2013; 17(14-117):51-54. [Mukha A.V., Rozaliev V.L., Orlova Yu.A., Zaboleeva-Zotova A.V. Automated approach to determining the authorship of the text. Izvestia VSTU. 2013; 17(14-117):51-54. (in Russ.)]
6. Akker R.A., Traum D. Comparison of addressee detection methods for multiparty conversations. Proc. of methods for multiparty conversations. Amsterdam; 2009. P. 99-106.
7. Choi D., Ko B., Kim H., Kim P. Text Analysis for Detecting Terrorism-Related Articles on the Web. Journal of Network and Computer Applications. 2013; 8(5):37-46.
8. Kolesnikova S.I. Metody analiza informativnosti raznotipnykh priznakov. Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitel'naya tekhnika i informatika. 2009; 1(6):69-80. [Kolesnikova S.I. Methods for analysing the informativeness of different types of signs. Tomsk State University Journal of Control and Computer Science. 2009; 1(6):69-80. (in Russ.)]
9. Polyakov I.V., Sokolova T.V., Chepovskiy A.A., Chepovskiy A.M. Problema klassifikatsii tekstov i differentsiruyushchie priznaki. Vestnik Novosibirskogo gosudarstvennogo universiteta. Seriya: Informatsionnye tekhnologii. 2015; 13(2):55-63. [Polyakov I.V., Sokolova T.V., Chepovskiy A.A., Chepovskiy A.M. The problem of text classification and differentiating features. Novosibirsk State University Journal of Information Technologies. 2015; 13(2):55-63. (in Russ.)]
10. Tolcheev V.O. Modifitsirovannyy i obobshchennyy metod blizhayshego soseda dlya klassifikatsii bibliograficheskikh tekstovykh dokumentov. Zavodskaya laboratoriya. Diagnostika materialov. 2009; 7:63-70. [Tolcheev V.O. B.O. Modified and generalised method of the nearest neighbor for the classification of bibliographic text documents. Industrial Laboratory. Materials Diagnostics. 2009; 7:63-70. (in Russ.)]
11. Meshkova E.V. Metodika postroeniya klassifikatora teksta na osnove gibridnoy neyrosetevoy modeli. Izvestiya YuFU. Tekhnicheskie nauki. 2008; 4(81):212-215. [Meshkova E.V. Method for constructing a text classifier based on a hybrid neural network model. Izvestiya SFedU. Engineering sciences. 2008; 4(81):212-215. (in Russ.)]
12. Kozoborod A.V., Meshkov V.E., Meshkova E.V. Analiz arkhitektur gibridnykh neyrosetevykh modeley v zadachakh avtomaticheskoy klassifikatsii tekstovoy informatsii. Izvestiya YuFU. Tekhnicheskie nauki. 2010; 12 (113):185-190. [Kozoborod A.V., Meshkov V.E., Meshkova E.V. Architecture analysis of hybrid neural network models in problems of automatic classification of textual information. Izvestiya SFedU. Engineering sciences. 2010; 12(113):185-190. (in Russ.)]

13. Kim Dhz.-O., Myuller Ch.U., Klekka U.R., Oldenderfer M.S., Bleshfild R.K. Faktornyy, diskriminantnyy i klasternyy analiz: Per. s angl.Moscow: Finansy i statistika; 1989. 215 p. [Kim Dhz.-O., Myuller Ch.U., Klekka U.R., Oldenderfer M.S., Bleshfild R.K. Factor, discriminant and cluster analysis: translated from English. Moscow: Finansy i statistika; 1989. 215 p. (in Russ.)]
14. Khachumov M.V. Rasstoyaniya, metriki i klasternyy analiz. Iskusstvennyy intellekt i prinyatie resheniy. 2012; 1:81-89. [Khachumov M.V. Distances, metrics and cluster analysis. Iskusstvennyy intellekt i prinyatie resheniy. 2012; 1:81-89. (in Russ.)]
15. Tolmachev I.L., Khachumov M.V. Binarnaya klassifikatsiya na osnove var'irovaniya razmernosti prostranstva priznakov i vybora effektivnoy metriki. Iskusstvennyy intellekt i prinyatie resheniy. 2010; 2:3-10. [Tolmachev I.L., Khachumov M.V. Binary classification based on variation of the feature space dimension and the choice of an effective metric. Iskusstvennyy intellekt i prinyatie resheniy. 2010; 2:3-10. (in Russ.)]
16. Khachumov M.V. Primenenie neyrona i rasstoyaniya Evklida-Makhalanobisa v zadache binarnoy klassifikatsii. Nauka i sovremennost'. 2010; 2-3:82-86. [Khachumov M.V. The application of the neuron and the Euclidean-Mahalanobis distance in the binary classification problem. Science and Modernity. 2010; 2-3:82-86. (in Russ.)]
17. Shumskaya A.O. Otsenka effektivnosti metrik rasstoyaniya Evklida i rasstoyaniya Makhalanobisa v zadachakh identifikatsii proiskhozhdeniya teksta. Doklady Tomskogo gosudarstvennogo universiteta sistem upravleniya i radioelektroniki. 2013; 3(29):141-145. [Shumskaya A.O. Estimation of the effectiveness of Euclidean distance metrics and the Mahalanobis distance in the problems of text origin identification. Proceedings of TUSUR University. 2013; 3(29):141-145. (in Russ.)]
18. "Baza dannykh metatekstovoy razmetki Natsional'nogo korpusa russkogo yazyka» (kollektsiya detskoy literatury)". 2014. ["Database of metatext marking of the National Corpus of the Russian language "(collection of children's literature))". 2014. (in Russ.)]
19. Natsional'nyy korpus russkogo yazyka [Elektronnyy resurs]. 2015. URL: http:// ruscorpora.ru/ (data obrasjcheniya: 26.07.2016). [The National Corpus of the Russian language [Electronic resource]. 2015. URL: http:// ruscorpora.ru/ (access date: 26.07.2016).]
20. Glazkova A.V. Proverka informativnosti klassifikatsionnykh priznakov v zadache avtomaticheskoy klassifikatsii tekstov na estestvennom yazyke. Materialy konferentsii "Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh sistem (OSTIS-2015)". Minsk; 2015. S. 541-544. [Glazkova A.V. Checking the informativeness of classification characteristics in the task of text automatic classification in natural language. Proceedings of conference "Open Semantic Technology for Intelligent Systems (OSTIS-2015)". Minsk; 2015. P. 541-544. (in Russ.)]
21. Bureeva N.N. Mnogomernyy statisticheskiy analiz s ispol'zovaniem PPP "STATISTICA". Nizhny Novgorod: Nizhegorodskiy gosudarstvennyy universitet im. N.I. Lobachevskogo; 2007. 112 s. [Bureeva N.N. Multidimensional statistical analysis using "STATISTICA". Nizhniy Novgorod: Lobachevsky State University of Nizhni Novgorod; 2007. 112 p. (in Russ.)]

Сведения об авторе.
Глазкова Анна Валерьевна – ассистент кафедры программного обеспечения.
Information about the author.
Anna V. Glazkova – Assistant, department of software.

Конфликт интересов **Conflict of interest**

Автор заявляет об отсутствии конфликта интересов. The author declare no conflict of interest.

Поступила в редакцию 23.01.2017. **Received** 23.01.2017.
Принята в печать 20.02.2017. **Accepted for publication** 20.02.2017.