УДК 004.93 doi:10.20998/2413-4295.2018.45.12

СИГНАТУРА ТОЧЕЧНОГО МНОЖЕСТВА И АЛГОРИТМ КЛАССИФИКАЦИИ НА ЕЁ ОСНОВЕ

А. А. ДАШКЕВИЧ

кафедра геометрического моделирования и компьютерной графики, НТУ «ХПИ», Харьков, УКРАИНА
e-mail: dashkewich.a@gmail.com

АННОТАЦИЯ *На данный момент существует большое количество задач по автоматизированной обработке многомерных данных, например, классификация, кластеризация, прогнозирование, задачи управления сложными объектами. Соответственно, возникает необходимость в развитии математического и алгоритмического обеспечения для решения возникающих задач. Целью исследования является развитие алгоритмов классификации точечных множеств на основе их пространственного распределения. В работе предлагается рассматривать данные как точки в многомерном метрическом пространстве. В работе рассмотрены подходы к описанию характеристик точечных множеств в пространствах высокой размерности и предлагается подход к описанию точечного множества на основе сигнатур, которые представляют собой характеристику заполненности точечного множества на основе расширения понятия пространственного хеширования. Обобщенный подход к вычислению сигнатур точечных множеств заключается в разбиении пространства, занимаемого множеством на регулярную сетку с помощью метода пространственного хеширования, вычисления геометрических характеристик множества в полученных ячейках и определения наиболее заполненных ячеек по каждому из пространственных измерений. Предлагается новый подход к классификации на основе сигнатур множества, который заключается в нахождении сигнатур для точек с известным значением принадлежности к некоторым классам, а для новых точек вычисляется расстояние от хеша точки до сигнатуры каждого из известных множеств, на основе чего определяется наиболее вероятный класс точки. В качестве используемых метрик предлагаются Евклидово расстояние и метрика городских кварталов. В работе проведён сравнительный анализ используемых метрик с точки зрения точности классификации. Преимуществами предложенного подхода являются простота вычислений и высокая степень точности классификации для равномерно распределенных точек. Представленный алгоритм реализован в виде программного приложения на языке Python с использованием библиотеки NumPy. Также рассмотрены варианты использования предложенного подхода для задач с нечисловыми данными, такими как строковые и булевы значения. Для таких данных предложено использовать метрику Хэмминга, проведённые эксперименты показали работоспособность алгоритма для таких типов данных*

Ключевые слова: *пространственное хеширование; классификация; точечное множество; метрическое пространство; сигнатура точечного множества; Евклидово расстояние; расстояние городских кварталов; метрика Хэмминга*

POINT SET SIGNATURE AND ALGORITHM OF CLASSIFICATIONS ON ITS BASIS

A. DASHKEVICH

Geometrical modeling and computer graphics department, National technical university «Kharkiv polytechnic institute», Kharkiv, UKRAINE

ABSTRACT *There are many unsolved problems in the field of automatic multi-dimensional data processing, for example, classification, clustering, regression, and control of complex objects. This leads to the need of development of mathematical and algorithmical background for such problems. In our research we aim to development of classification algorithms of point sets based on their spatial distribution. We propose to consider data as points in multi-dimensional metric space. The approaches to describe point set features in high dimensional spaces are viewed. The algorithm of describing of point set based on their signatures, that are spatial distribution of point set is considered. In our approach we extend spatial hashing technique. The generalized method of computation of point set signatures is to split space, occupied by point set into regular grid by the spatial hashing algorithm, then we evaluate geometrical characteristics of the set in cells of the grid and define cells, that contain most of the points for the all of coordinate axis. The new approach to classification by means of point set signatures is developed that is to find signatures of known points with the classes defined and then we compute spatial hashes for unknown points and their distance to the signatures of classes. The probable class of the tested point is defined by the minimal distance among all distances to each signature. To define distance in our approach we use Manhattan and Euclidean metric. The comparative study of impact of metrics used to the classification error is provided. The main advantage of our method is computation simplicity and low classification error for evenly distributed points. Prototype implementation of our algorithm was written in order to test this algorithm for practical classification applications. The implementation was coded in Python with use NumPy library. The use of our algorithm to the classification of non-numerical data such as texts and booleans is viewed. For such data types we propose use of Hamming distance and experiments done show practical viability for such data types.*

Keywords*: spatial hashing; classification; point set; metric space; point set signature; Euclidean distance; Manhattan distance; Hamming distance*

Введение

Одной из наиболее часто встречающихся задач в области обработки данных является задача классификации – задача соотнесения входных данных одним из заранее заданных классов. Существует большое количество подходов к решению задачи классификации, среди которых можно выделить такие

ISSN 2079-5459 (print)
ISSN 2413-4295 (online)

направления: методы основанные на мере близости объектов и поиска ближайших соседей [1-4], методы поиска разделяющих гиперплоскостей на основе машин опорных векторов [5], построение деревьев решающих правил [6], нейронные сети [7] и др. Алгоритмы, основанные на поиске ближайших соседей являются одними из самых простых в реализации, однако не обладают достаточной точностью на данных большой размерности. С другой стороны, нейронные сети могут давать приемлемую точность вычислений для многомерных входных данных, но при этом процесс выбора архитектуры сети и настройки её весовых коэффициентов может быть довольно затратным с точки зрения вычислительной сложности и не может быть точно формализован.

Выбор метода классификации чаще всего осуществляется эмпирическим путём, так как не существует единой методики по определению наиболее подходящих алгоритмов в зависимости от задачи. В работе [8] предложена методика выбора и оценки алгоритма классификации. Также, в большинстве работ не уделяется достаточно внимания геометрической структуре исходных данных. Можно выделить подходы, основанные на разбиении пространства параметров на регулярные и нерегулярные сетки [9]. Так, в работе [10] предлагается подход к классификации данных на основе адаптивных разреженных сеток и его преимущества перед регулярными сетками, а в работе [11] показана взаимосвязь решающих правил классификатора и точек в многомерном пространстве. В работе [12] предлагаются методы классификации на основе разбиения на сетки для задач управления транспортными средствами. В исследовании [13] используются пространственные характеристики данных для прогнозирования преступлений. В работе [14] сеточное представление графов применяется для классификации изображений. В публикации [15] предложен алгоритм пространственного хеширования на основе разбиения многомерных пространств на сетки для поиска ближайших соседей.

Цель работы

Разработка метода представления и алгоритма классификации точечных множеств на основе их пространственного распределения.

Изложение основного материала

Одной из проблем при работе с точечными множествами является различная мощность множеств, для решения практических задач часто требуется приведение различных точечных множеств к единой размерности. Для решения этой проблемы предлагается расширение алгоритма [15] – концепция линейного хеша H_d^T – пространственный хеш, в котором на одно пространственное измерение d_i приходится один разряд хеша, T – максимальное целочисленное значение в ячейках сетки, например, H_3^{10} – трёхмерный десятичный линейный хеш, в котором для каждого из трёх пространственных измерений допустимыми значениями хеша являются $\{0, .., 9\}$, H_4^2 – четырехмерный двоичный линейный хеш, в котором для каждого из четырёх пространственных измерений допустимыми значениями хеша являются $\{0, 1\}$.

Преимуществом пространственных хешей и, в частности, линейных хешей, является то, что они могут быть представлены в виде единственного целого числа, что позволяет их использовать в качестве ключей хеш-таблицы с константным временем поиска элементов. Также хеши могут быть представлены в виде одномерных массивов или строковых переменных.

Для линейных хешей H_d^T ближайшие ячейки имеют значения по каждому из пространственных измерений, отличающиеся не более, чем на 1 от соответствующих значений заданного хеша.

Например, для двумерного десятичного хеша $H_2^{10} = [5, 3]$ ближайшие 8 хеш-ячеек: *[4, 2], [4, 3], [4, 4], [5, 2], [5, 4], [6, 2], [6, 3], [6, 4]*.

Таким образом точечное множество может быть представлено в виде гиперкуба на основе линейных хешей C_d^T - d-мерный гиперкуб на основе линейных хешей H_d^T с размерностью $T_1 \times T_2 \times ... \times T_d$, в заполненных ячейках которого находится значение 1, а в пустых – 0. Другим вариантом заполнения ячеек предлагается количество точек в ячейке.

Гиперкубы на основе линейных хешей позволяют приводить различные точечные множества одной размерности к регулярному представлению константного размера, что дает возможность проводить, например, прямое сравнение точечных множеств без учета различного количества точек в них.

Сигнатура точечного множества S – линейный хеш, в ячейках которого содержится номер наиболее заполненной ячейки заданного точечного множества.

Алгоритм классификации на основе сигнатур:

1) обучающее множество точек для каждого i-го класса разбивается на гиперкуб соответствующей размерности C_i;

2) для каждого гиперкуба вычисляется сигнатура S_i;

3) для каждой новой точки P_j, не входящей в обучающую выборку, вычисляется линейный хеш и расстояние между полученным хешем и сигнатурой каждого из классов:

$dist_i = m(H_j, S_i)$,

где m – некоторая метрика, в качестве которой может выступать Евклидово расстояние, расстояние городских кварталов (Манхеттенская метрика) и др.;

4) минимальное расстояние определяет принадлежности точки к одному из классов.

Обсуждение результатов

По предложенному алгоритм проведена бинарная классификация многомерных точечных множеств на основе наборов данных «Ирисы Фишера» (четырёхмерный) и «MNIST» (784-мерный). Значение *T* для каждого из наборов последовательно выбиралось из ряда *{4, 5, 8, 10, 16, 32, 64, 100}*.

Результаты классификации и соответствующие сигнатуры для «Ирисов Фишера» и «MNIST» представлены на рис. 1 и рис. 2, соответственно. В качестве метрики была использована метрика городских кварталов. Также проводилось исследование точности классификации с использованием Евклидового расстояния, однако полученная точность при этом сравнима с манхеттенской метрикой при более высокой вычислительной сложности. На рис. 3 показан пример полученных сигнатур для классов из набора «MNIST», для наглядности отображения сигнатуры приведены к размерности исходных данных *(28×28)*.

Можно увидеть, что на наборе «Ирисы Фишера» ошибка классификации при некоторых значениях *T* падает до *0*.

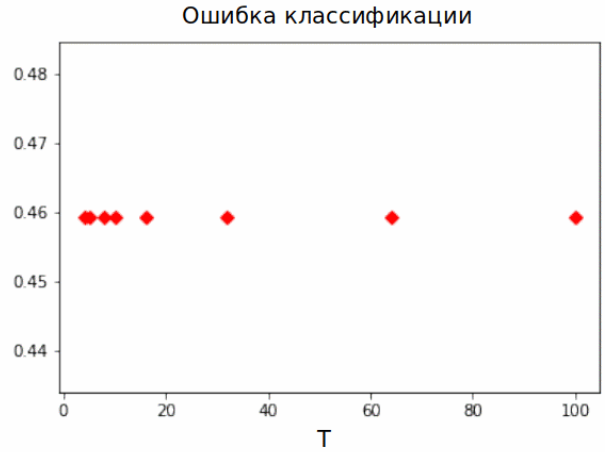

Рис. 1 – Точность классификации для набора «Ирисы Фишера»

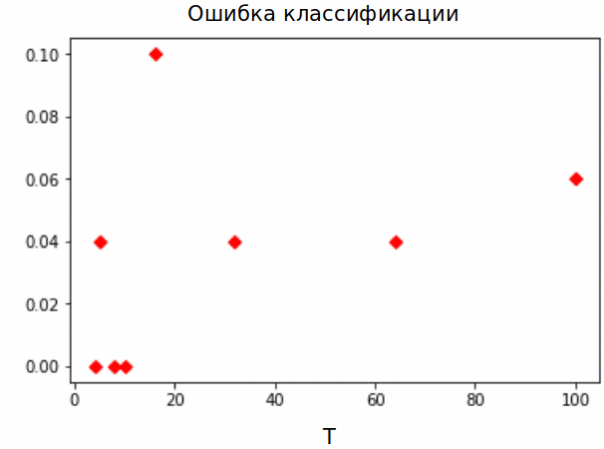

Рис. 2 – Точность классификации для набора «MNIST»

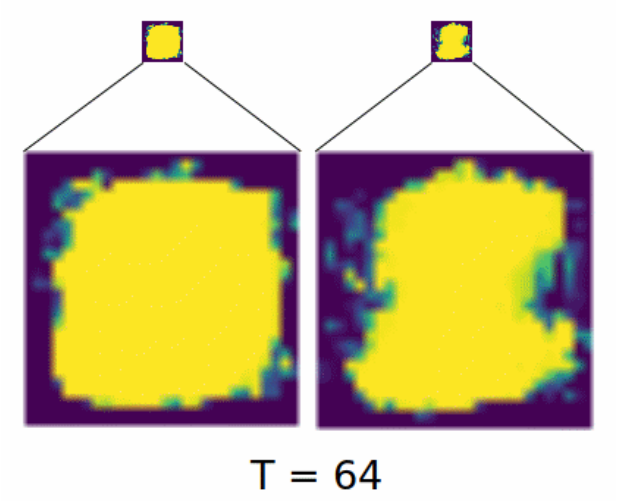

Рис. 3 – Сигнатуры классов набора «MNIST»

Для набора «MNIST» независимо от *T* точность принимает значения *~0,46*, что является неприемлемым уровнем. Низкий уровень точности классификации может быть объяснён тем, что исходные точки разных классов в указанном наборе имеют небольшой разброс значений вдоль координатных осей, поэтому значения их хешей попадают в одни и те же ячейки. Для повышения точности в таких случаях может быть применено снижение размерности данных.

Также проводилась классификация по указанному алгоритму текстовых массивов с использованием метрики городских кварталов и булевых данных с использованием расстояния Хэмминга.

Выводы

В результате работы разработан метод определения характеристики точечного множества на основе его сигнатур и алгоритм классификации методом сравнения расстояния от хеша точки до сигнатуры множества.

Предложенный алгоритм обладает высокой степенью точности классификации для точечных множеств, точки которых равномерно рассеяны в пространстве, а для множеств, имеющих малый разброс точек в пределах разных классов степень точности классификации падает.

Дальнейшие исследования будут направлены на повышение точности классификации и алгоритмы снижения размерности точечных множеств на основе разбиения этих множеств на сетки.

Список литературы

1. **Ougiaroglou, S.** Adaptive k-Nearest-Neighbor Classification Using a Dynamic Number of Nearest Neighbors / **S. Ougiaroglou, A. Nanopoulos, A. N. Papadopoulos, Y. Manolopoulos, T. Welzer-Druzovec** // *In: Ioannidis Y., Novikov B., Rachev B. (eds) Advances in*

ISSN 2079-5459 (print)
ISSN 2413-4295 (online)

Databases and Information Systems. ADBIS, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg. – 2007. – Vol. 4690. – P. 66-82.

2. **Law, Y.-N.** An Adaptive Nearest Neighbor Classification Algorithm for Data Streams / **Y.-N. Law, C. Zaniolo** // *in: Jorge, A.M., Torgo, L., Brazdil, P., Camacho, R., Gama, J. (Eds.),* Knowledge Discovery in Databases: PKDD, Springer Berlin Heidelberg, Berlin, Heidelberg. – 2005. – P. 108-120. – doi: 10.1007/11564126_15.
3. **Wang, L.** An Effective Evidence Theory Based K-Nearest Neighbor (KNN) Classification / **L. Wang, L. Khan, B. Thuraisingham** // *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.* Presented at the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, IEEE, Sydney, Australia. – 2008. – P. 797-801. – doi: 10.1109/WIIAT.2008.411.
4. **Song, Y.** IKNN: Informative K-Nearest Neighbor Pattern Classification / **Y. Song, J. Huang, D. Zhou, H. Zha, C.L. Giles** // in: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (Eds*.), Knowledge Discovery in Databases:* PKDD 2007. Springer Berlin Heidelberg, Berlin, Heidelberg. – 2007. – P. 248-264. – doi: 10.1007/978-3-540-74976-9_25.
5. **Wenzel, F.** Bayesian Nonlinear Support Vector Machines for Big Data / **F. Wenzel, T. Galy-Fajou, M. Deutsch, M. Kloft** // In: Ceci M., Hollmén J., Todorovski L., Vens C., Džeroski S. (eds) *Machine Learning and Knowledge Discovery in Databases*. ECML PKDD 2017. Lecture Notes in Computer Science, vol 10534. Springer, Cham, 2017. – P. 307-322. – doi: 10.1007/978-3-319-71249-9_19.
6. **Painsky, A.** Cross-Validated Variable Selection in Tree-Based Methods Improves Predictive Performance / **A. Painsky, S. Rosset** // *IEEE Transactions on Pattern Analysis and Machine Intelligence.* – 2017. – 39(11). – P. 2142-2153. – doi: 10.1109/TPAMI.2016.2636831.
7. **Najibi, M.** G-CNN: An Iterative Grid Based Object Detector / **M. Najibi, M. Rastegari, L.S. Davis** // 2016 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* Las Vegas, NV, USA. – 2016. – P. 2369-2377. – doi: 10.1109/CVPR.2016.260.
8. **Ali, S.** On learning algorithm selection for classification / **S. Ali, K.A. Smith** // *Applied Soft Computing.* – 2006. – 6. – P. 119-138. – doi: 10.1016/j.asoc.2004.12.002.
9. **Garcke, J.** Classification with sparse grids using simplicial basis functions / **J. Garcke, M. Griebel** // *Intelligent Data Analysis.* – 2002. – Vol. 6. – № 6. – P. 483-502.
10. **Pflüger, D.** Adaptive Sparse Grid Classification Using Grid Environments / **D. Pflüger, I.L. Muntean, H.-J. Bungartz** // in: Shi, Y., van Albada, G.D., Dongarra, J., Sloot, P.M.A. (Eds.), *Computational Science – ICCS 2007*. Springer Berlin Heidelberg, Berlin, Heidelberg. – P. 708-715. – doi: 10.1007/978-3-540-72584-8_94.
11. **Gupta, P.** Algorithms for packet classification / **P. Gupta, N. McKeown** // *IEEE Network.* – 2001. – 15. – P. 24-32. – doi: 10.1109/65.912717.
12. **Rieken, J.** Benefits of Using Explicit Ground-Plane Information for Grid-based Urban Environment Modeling / **J. Rieken, R. Matthaei, M. Maurer** // *18th International Conference on Information Fusion* (Fusion), Washington, DC. – 2015. – P. 2049-2056.
13. **Lin, Y.-L.** Grid-Based Crime Prediction Using Geographical Features / **Y.-L. Lin, M.-F. Yen, L.-C. Yu** // *ISPRS International Journal of Geo-Information.* – 2018. – 7. – 298. – doi: 10.3390/ijgi7080298.
14. **Deville, R.** GriMa: A Grid Mining Algorithm for Bag-of-Grid-Based Classification / **R. Deville, E. Fromont, B. Jeudy, C. Solnon** // in: Robles-Kelly, A., Loog, M., Biggio, B., Escolano, F., Wilson, R. (Eds.), *Structural, Syntactic, and Statistical Pattern Recognition.* Springer International Publishing, Cham. – 2016. – P. 132-142. – doi: 10.1007/978-3-319-49055-7_12.
15. **Дашкевич, А. А.** Алгоритм пространственного хеширования для решения задач приблизительного поиска ближайших соседей / **А. А. Дашкевич** // *Науковий вісник ТДАТУ*. – 2018. – Вип. 8. – Т. 1. – С. 79-86.

References (transliterated)

1. **Ougiaroglou, S., Nanopoulos, A., Papadopoulos, A.N., Manolopoulos, Y., Welzer-Druzovec, T.** Adaptive k-Nearest-Neighbor Classification Using a Dynamic Number of Nearest Neighbors. In: Ioannidis Y., Novikov B., Rachev B. (eds) *Advances in Databases and Information Systems*. ADBIS, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2007, **4690**, 66-82.
2. **Law, Y.-N. Zaniolo, C.** An Adaptive Nearest Neighbor Classification Algorithm for Data Streams. In: Jorge, A.M., Torgo, L., Brazdil, P., Camacho, R., Gama, J. (Eds.), *Knowledge Discovery in Databases*: PKDD, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, 108-120, doi: 10.1007/11564126_15.
3. **Wang, L., Khan, L., Thuraisingham, B.** An Effective Evidence Theory Based K-Nearest Neighbor (KNN) Classification. In: 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. Presented at the 2008 IEEE/WIC/ACM *International Conference on Web Intelligence and Intelligent Agent Technology*, IEEE, Sydney, Australia, 2008, 797-801, doi: 10.1109/WIIAT.2008.411.
4. **Song, Y., Huang, J., Zhou, D., Zha, H., Giles, C. L.** IKNN: Informative K-Nearest Neighbor Pattern Classification. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (Eds.), *Knowledge Discovery in Databases*: PKDD 2007. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007, 248-264, doi: 10.1007/978-3-540-74976-9_25.
5. **Wenzel, F., Galy-Fajou, T., Deutsch, M., Kloft, M.** Bayesian Nonlinear Support Vector Machines for Big Data. In: Ceci M., Hollmén J., Todorovski L., Vens C., Džeroski S. (eds) *Machine Learning and Knowledge Discovery in Databases*. ECML PKDD 2017. Lecture Notes in Computer Science, **10534**. Springer, Cham, 2017, 307-322, doi: 10.1007/978-3-319-71249-9_19.
6. **Painsky, A., Rosset, S.** Cross-Validated Variable Selection in Tree-Based Methods Improves Predictive Performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39** (11), 2142-2153, doi: 10.1109/TPAMI.2016.2636831.
7. **Najibi, M., Rastegari, M., Davis, L. S.** G-CNN: An Iterative Grid Based Object Detector. 2016 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA, 2016, 2369-2377, doi: 10.1109/CVPR.2016.260.
8. **Ali, S., Smith, K. A.** On learning algorithm selection for classification. *Applied Soft Computing* **6**, 2006, 119-138, doi: 10.1016/j.asoc.2004.12.002.

9. **Garcke, J., Griebel, M.** Classification with sparse grids using simplicial basis functions. *Intelligent Data Analysis*, 2002, **6**, 6, 483-502.
10. **Pflüger, D., Muntean, I.L., Bungartz, H.-J.** Adaptive Sparse Grid Classification Using Grid Environments. In: Shi, Y., van Albada, G.D., Dongarra, J., Sloot, P.M.A. (Eds.), *Computational Science – ICCS 2007*. Springer Berlin Heidelberg, Berlin, Heidelberg, 708-715, doi: 10.1007/978-3-540-72584-8_94.
11. **Gupta, P., McKeown, N.** Algorithms for packet classification. *IEEE Network*, 2001, **15**, 24-32, doi: 10.1109/65.912717.
12. **Rieken, J., Matthaei, R., Maurer, M.** Benefits of Using Explicit Ground-Plane Information for Grid-based Urban Environment Modeling. *18th International Conference on Information Fusion* (Fusion), Washington, DC, 2015, 2049-2056.
13. **Lin, Y.-L., Yen, M.-F., Yu, L.-C.** Grid-Based Crime Prediction Using Geographical Features. *ISPRS International Journal of Geo-Information*, 2018, **7**, 298, doi: 10.3390/ijgi7080298.
14. **Deville, R., Fromont, E., Jeudy, B., Solnon, C.** GriMa: A Grid Mining Algorithm for Bag-of-Grid-Based Classification. In: Robles-Kelly, A., Loog, M., Biggio, B., Escolano, F., Wilson, R. (Eds.), *Structural, Syntactic, and Statistical Pattern Recognition*. Springer International Publishing, Cham, 2016, 132-142, doi: 10.1007/978-3-319-49055-7_12.
15. **Dashkevich, A.** Algoritm prostranstvennogo heshirovaniya dlya resheniya zadach priblizitelnogo poiska blizhayshih sosedey [Spatial hashing algorithm for solving approximate search problems for nearest neighbors]. *Scientific bulletin of the Tavria agrotechnological state university*, TASU, 2018, **8**, 1, 79-86.

Сведения об авторах (About authors)

Дашкевич Андрей Александрович – кандидат технических наук, доцент, Национальный технический университет «Харьковский политехнический институт», докторант кафедры геометрического моделирования и компьютерной графики; г. Харьков, Украина; ORCID: 0000-0002-9963-0998; e-mail: dashkewich.a@gmail.com.

Andrey Dashkevich – Ph. D., docent, doctoral student of geometrical modeling and computer graphics department, National technical university «Kharkiv polytechnic institute», Kharkiv, Ukraine; ORCID: 0000-0002-9963-0998; e-mail: dashkewich.a@gmail.com.

Пожалуйста, ссылайтесь на эту статью следующим образом:

Дашкевич, А. А. Сигнатура точечного множества и алгоритм классификации на её основе / **А. А. Дашкевич** // *Вестник НТУ «ХПИ», Серия: Новые решения в современных технологиях*. – Харьков: НТУ «ХПИ». – 2018. – № 45 (1321). – С. 93-97. – doi:10.20998/2413-4295.2018.45.12.

Please cite this article as:

Dashkevich, A. Point set signature and algorithm of classifications on its basis. *Bulletin of NTU "KhPI". Series: New solutions in modern technologies*. – Kharkiv: NTU "KhPI", 2018, **45** (1321), 93–97, doi:10.20998/2413-4295.2018.45.12.

Будь ласка, посилайтесь на цю статтю наступним чином:

Дашкевич, А. О. Сигнатура точкової множини та алгоритм класифікації на її основі / **А. О. Дашкевич** // *Вісник НТУ «ХПІ», Серія: Нові рішення в сучасних технологіях*. – Харків: НТУ «ХПІ». – 2018. – № 45 (1321). – С. 93-97. – doi:10.20998/2413-4295.2018.45.12.

АНОТАЦІЯ *На даний момент існує велика кількість задач з автоматизованої обробки багатовимірних даних, наприклад, класифікація, кластеризація, прогнозування, задачі з керування складними об'єктами. Відповідно, виникає необхідність в розвитку математичного та алгоритмічного забезпечення для розв'язання таких задач. Метою дослідження є розвиток алгоритмів класифікації точкових множин на основі їх просторового розподілу. В дослідженні пропонується розглядати дані як точки в багатовимірному метричному просторі. В роботі розглянуто підходи до опису характеристик точкових множин в просторах високої розмірності та пропонується підхід до опису точкової множини на основі сигнатур, які представляють характеристику заповненості точкової множини на основі розширення поняття просторового хешування. Узагальнений підхід до обчислення сигнатур точкових множин полягає в розбитті простору, що займає множина на регулярну сітку з використанням методу просторового хешування, обчислення геометричних характеристик множини в отриманих клітинах сітки та визначення найбільш заповнених клітин за кожним з просторових вимірів. Пропонується новий підхід до розв'язання задачі класифікації на основі сигнатур множин, який полягає в визначенні сигнатур для точок з відомою належністю до заданих класів, а для невідомих точок обчислюється відстань від хешу цієї точки до сигнатур усіх заданих класів, на основі відстані визначається найбільш вірогідний клас точки. В якості метрик пропонується використання Евклідової відстані та метрики міських кварталів. У роботі проведений порівняльний аналіз використаних метрик з точки зору точності класифікації. До переваг розробленого підходу можна віднести простоту обчислень та високий ступінь точності класифікації для рівномірно розподілених точок. Представлений алгоритм реалізовано у вигляді програмного додатку на мові програмування Python з використанням бібліотеки NumPy. Також розглянуто варіанти використання запропонованого підходу для задач з нечисловими даними, такими як текстові та булеві значення. Для таких типів даних запропоновано використання метрики Геммінга, проведені експерименті показали доцільність використання алгоритму для таких типів даних.*

Ключові слова: *просторове хешування; класифікація; точкова множина; метричний простір; сигнатура точкової множини; Евклідова відстань; відстань міських кварталів; метрика Геммінга*

Поступила (received) 07.11.2018