*Вісник Національного технічного університету "ХПІ", 2018, № 42 (1318)*
**ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)**

**УДК 004.93** **DOI: 10.20998/2411-0558.2018.42.06**

***А. А. ДАШКЕВИЧ,*** канд. техн. наук, доц., докторант, НТУ "ХПИ"

# СНИЖЕНИЕ РАЗМЕРНОСТИ ДАННЫХ НА ОСНОВЕ РАЗБИЕНИЯ ПРОСТРАНСТВА НА РЕГУЛЯРНУЮ СЕТКУ

Предлагается подход к решению задачи классификации точечных множеств на основе снижения размерности данных и разбиения пространства на регулярную сетку. Вводится понятие гиперкуба как способ представления точечных множеств. Предложен подход к снижению размерности на основе сигнатуры точечного множества. Разработанный метод даёт возможность исключить из дальнейшей классификации множество координатных осей при повышении точности классификации и уменьшении количества необходимых вычислений. Проведённые эксперименты показали работоспособность подхода на данных больших размерностей. Преимуществом подхода является быстрое определение избыточных координатных осей для произвольного набора исходных классов. Ил.: 1. Библиогр.: 17 назв.

**Ключевые слова:** гиперкуб; регулярная сетка; сигнатура; точечное множество; данные больших размерностей; классификация.

**Постановка проблемы.** В настоящее время сбор и автоматизированная обработка информации являются одними из наиболее значимых практических задач. При этом обрабатываемая информация представляет собой большие массивы многомерных данных разных типов. Существует большое количество подходов к классификации многомерных данных, среди которых можно выделить метод $k$ ближайших соседей [1], машины опорных векторов [2], деревья решающих правил [3], нейронные сети [4] и др. Одними из самых простых в реализации являются алгоритмы, основанные на поиске ближайших соседей. Однако рост размерности данных приводит к росту вычислительной сложности алгоритмов такого типа, так как увеличивается время на вычисления расстояний между объектами. Также в пространствах больших размерностей ($d > 20$) существует проблема, заключающаяся в том, что значения расстояний между точками множества оказываются сосредоточены в пределах узкого диапазона значений [5], что приводит к неэффективности классификаторов такого типа.

**Анализ последних исследований.** В работах по классификации методом поиска ближайших соседей можно выделить подходы, основанные на разбиении пространства параметров на регулярные и нерегулярные сетки [6]. Так, в работе [7] предлагается подход к классификации данных на основе адаптивных разреженных сеток и его преимущества перед регулярными сетками, а в работе [8] показана

взаимосвязь решающих правил классификатора и точек в многомерном пространстве. В работе [9] предлагаются методы классификации на основе разбиения на сетки для задач управления транспортными средствами. В работе [10] используются пространственные характеристики данных для прогнозирования преступлений. В работе [11] сеточное представление графов применяется для классификации изображений. В работе [12] предложен алгоритм пространственного хеширования на основе разбиения многомерных пространств на сетки для поиска ближайших соседей. Существующие работы, посвящённые классификации многомерных данных, в качестве решения проблемы компактного распределения точек предлагают снижение размерности исходных данных на основе сингулярного разложения матриц, метода главных компонент или факторного анализа [13-17]. К недостатку таких алгоритмов следует отнести дополнительные затраты времени вычислений на геометрические преобразования данных без учёта геометрических характеристик исходных множеств и нацеленность на дальнейшую визуализацию данных, а не точность классификации.

**Цели работы.** Разработка метода снижения размерности и алгоритма классификации точечных множеств на основе разбиения пространства на регулярную сетку.

**Основная часть.** Для решения проблемы близкого расположения точек в многомерных множествах предлагается вычисление меры заполненности пространства точками путём разбиения его на регулярную сетку на основе алгоритма пространственного хеширования [12]. При этом пространство вдоль каждой из координатных осей разбивается на $T$ (разрешение сетки) участков.

Таким образом, точечное множество может быть представлено в виде гиперкуба $C_d^T$ – $d$-мерный гиперкуб с размерностью $T_1 \times T_2 \times ... \times T_d$ на основе пространственных хешей $H_d^T$, в заполненных ячейках которого находится количество точек множества в этой ячейке, а в пустых – 0.

Гиперкубы на основе пространственных хешей позволяют приводить различные точечные множества одной размерности к регулярному представлению константного размера, что дает возможность сформулировать следующую характеристику пространственного распределения точечных множеств:

*Сигнатура точечного множества* $S$ – пространственный хеш, содержащий для каждой координатной оси $i$ заданного точечного множества с $n$ точками номер наиболее заполненной ячейки гиперкуба $C_d^T$ вдоль этой оси:

$$S = \{max(H_i^j) \mid i = 1, \ldots, d, j = 1, \ldots, n\}.$$

Так как сигнатура позволяет выявить наиболее характерные ячейки для точечных множеств, то это даёт возможность сформулировать следующую гипотезу:

*гипотеза 1* для заданных точечных множеств $P_1$ и $P_2$ разрешения сетки $T$ и множества координатных осей $A$, т.е. координатные оси гиперкуба $a_i \in A$, у которых совпадают значения в сигнатуре, могут быть исключены из дальнейшего рассмотрения при одновременном повышении точности классификации. Множество исключённых осей $A`$:

$$A` = \{a_i \mid S_i(P_1) = S_i(P_2),\ i = 1, ..., d\}.$$

На основе данного условия может быть построен алгоритм снижения размерности:

1) обучающее множество точек разбивается на гиперкуб соответствующей размерности $C_d^T$;

2) для каждого заданного $j$-го класса вычисляется сигнатура $S_j$;

3) в соответствии с условием 1 определяется множество избыточных координатных осей $A`$ и дальнейшая классификация исходных точечных множеств проводится только с учётом редуцированного множества координатных осей:

$$A^R = A \setminus A`.$$

Алгоритм классификации:

1) для каждой новой точки $p_j$, не входящей в обучающую выборку, вычисляется пространственный хеш $H_j$ и расстояние между полученным хешем и сигнатурой каждого из $i$-го 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.;

2) минимальное расстояние определяет принадлежность точки к одному из классов.

Работа предложенного алгоритма тестировалась путём снижения размерности и дальнейшей бинарной классификации многомерных точечных множеств двух классов на основе набора данных MNIST – 784-мерный, изображения рукописных цифр размерности (28×28) в оттенках серого цвета 10 классов. Разрешение сетки $T$=10. 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.

**ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)**

Примеры полученных сигнатур множеств для MNIST представлены на рис. 1, для наглядности отображения сигнатуры приведены к размерностям исходных данных, также на рис. 1 показано редуцированное множество координатных осей.

В результате проведённых экспериментов были определены редуцированные множества координатных осей для указанных наборов, при этом размерность данных для набора MNIST была снижена с 784 до 90. При этом ошибка классификации по приведённому алгоритму снизилась с ~ 0.45 до ~ 0.01.

**Выводы и перспективы дальнейших исследований.** В результате работы разработан метод определения избыточных координатных осей для снижения размерности данных и повышения точности классификации.

Предложенный метод позволяет быстро определять координатные оси, которые имеют ключевое значение для характеристики входных данных. К недостаткам подхода можно отнести небольшой коэффициент снижения размерности, что не позволяет применять его для визуализации данных больших размерностей.

Дальнейшие исследования будут направлены на повышение точности классификации и коэффициента снижения размерности.

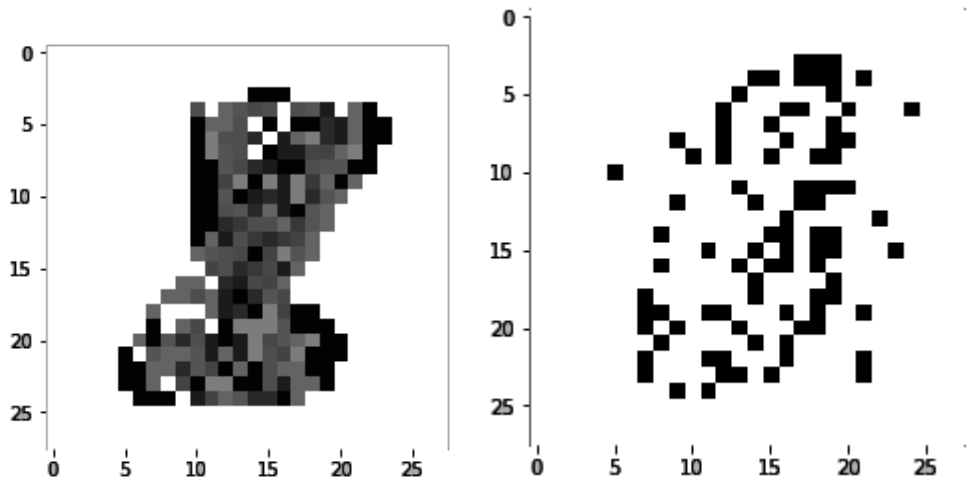


Рис. 1. Слева – пример сигнатуры для класса "Цифра 1", справа – редуцированное множество осей

**Список литературы:**

**1.** *Wang L.* An Effective Evidence Theory Based K-Nearest Neighbor (KNN) Classification / *L. Wang, L. Khan, B. Thuraisingham* // 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. – P. 797-801.

**2.** *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.

**3.** *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. – Vol. 39 (11). – P. 2142–2153.

**4.** *Najibi M.* G-CNN: An Iterative Grid Based Object Detector / *M. Najibi, M. Rastegari, L.S. Davis* // In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, NV, USA, 2016. – P. 2369-2377.

**5.** *Beyer K.* When Is "Nearest Neighbor" Meaningful? / *K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft* // International Conferenceon Database Theory: Springer, 1999. – P. 217-235.

**6.** *Garcke J.* Classification with sparse grids using simplicial basis functions / *J. Garcke, M. Griebel* // Intelligent Data Analysis, 2002. – Vol. 6. – № 6. – P. 483-502.

**7.** *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, 2007. – P. 708-715.

**8.** *Gupta P.* Algorithms for packet classification / *P. Gupta, N. McKeown* // IEEE Network, 2001. – Vol. 15. – P. 24-32.

**9.** *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.

**10.** *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 (8). – P. 298.

**11.** *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.

**12.** *Дашкевич А.А.* Алгоритм пространственного хеширования для решения задач приблизительного поиска ближайших соседей / *А.А. Дашкевич* // Науковий вісник ТДАТУ. – Вип. 8. – Т. 1. – Мелітополь, 2018. – С. 79-86.

**13.** *Cunningham J.P.* Linear Dimensionality Reduction: Survey, Insights, and Generalizations / *J.P. Cunningham, Z. Ghahramani* // Journal of Machine Learning Research, 2015. – Vol. 16. – P. 2859-2900.

**14.** *Herr D.* Visual Clutter Reduction Through Hierarchy-based Projection of High-dimensional Labeled Data / *D. Herr, Q. Han, S. Lohmann, T. Ertl* // Proceedings of the 42Nd Graphics Interface Conference, 2016. – P. 109-116.

**15.** *Feldman D.* Dimensionality Reduction of Massive Sparse Datasets Using Coresets / *D. Feldman, M. Volkov, D. Rus* // Advances in Neural Information Processing Systems, 2016. – Vol. 29. – P. 2766-2774.

**16.** *Carreira-Perpiñan M.Á.* The elastic embedding algorithm for dimensionality reduction / *M.Á. Carreira-Perpiñan* // In Proceedings of the 27th International Conference on International Conference on Machine Learning, Johannes Fürnkranz and Thorsten Joachims (Eds.). Omnipress, USA, 2010. – P. 167-174.

**17.** *Niu D.* Dimensionality Reduction for Spectral Clustering / *D. Niu, J.G. Dy, M.I. Jordan* // Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, PMLR, 2011. – Vol. 15. – P. 552-560.

**References:**

**1.** *Wang, L.*, Khan, L. and Thuraisingham, B. (2008), "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, pp. 797-801.

**2.** Wenzel, F., Galy-Fajou, T., Deutsch, M. and Kloft, M. (2017), "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, Vol 10534. Springer, Cham. – pp. 307-322.

**3.** Painsky, A., and Rosset, S. (2017), "Cross-Validated Variable Selection in Tree-Based Methods Improves Predictive Performance", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39 (11). – pp. 2142–2153.

**4.** Najibi, M., Rastegari, M., and Davis, L.S. (2016), *"G-CNN: An Iterative Grid Based Object Detector"*. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* IEEE, Las Vegas, NV, USA. – pp. 2369-2377.

**5.** Beyer K., Goldstein, J., Ramakrishnan, R. and Shaft, U. (1999), *"When Is "Nearest Neighbor" Meaningful?", In International Conferenceon Database Theory*: Springer. – pp. 217-235.

**6.** Garcke, J. and Griebel, M. (2002), "Classification with sparse grids using simplicial basis functions". *Intelligent Data Analysis*, Vol. 6, No. 6, pp. 483-502.

**7.** Pflüger, D., Muntean, I.L. and Bungartz, H.-J. (2007*), "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. – pp. 708-715.

**8.** Gupta, P. and McKeown, N. (2001), "Algorithms for packet classification". *IEEE Network*, Vol. 15, pp. 24-32.

**9.** Rieken, J., Matthaei, R., and Maurer, M. (2015), "Benefits of Using Explicit Ground-Plane Information for Grid-based Urban Environment Modeling". *18th International Conference on Information Fusion (Fusion)*, Washington, DC, pp. 2049-2056.

**10.** Lin, Y.-L., Yen, M.-F. and Yu, L.-C. (2018), "Grid-Based Crime Prediction Using Geographical Features". *ISPRS International Journal of Geo-Information*, No. 7 (8), p. 298.

**11.** Deville, R., Fromont, E., Jeudy, B., and Solnon, C. (2016), *"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, pp. 132-142.

**12.** Dashkevich, A. (2018), "Spatial hashing algorithm for approximate nearest neighbors search". *Scientific bulletin of the Tavria agrotechnological state university*, Is. 8, Vol. 1, pp. 79-86.

**13.** Cunningham, J. P. and Ghahramani, Z. (2015), "Linear Dimensionality Reduction: Survey, Insights, and Generalizations". *Journal of Machine Learning Research*, Vol. 16, pp. 2859-2900.

**14.** Herr, D., Han, Q., Lohmann, S., and Ertl, T. (2016), *"Visual Clutter Reduction Through Hierarchy-based Projection of High-dimensional Labeled Data". Proceedings of the 42Nd Graphics Interface Conference*, pp. 109-116.

**15.** Feldman, D., Volkov, M., and Rus, D. (2016), "Dimensionality Reduction of Massive Sparse Datasets Using Coresets". *Advances in Neural Information Processing Systems*, Vol. 29, pp. 2766-2774.

**16.** Carreira-Perpiñan, M.Á. (2010), "The elastic embedding algorithm for dimensionality reduction". In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, Johannes Fürnkranz and Thorsten Joachims (Eds.). Omnipress, USA, pp. 167-174.

**17.** Niu, D., Dy, J.G. and Jordan, M.I. (2011), "Dimensionality Reduction for Spectral Clustering". *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, PMLR, Vol. 15, pp. 552-560.

**ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)**

*Статью представил д-р. техн. наук, проф. Национального технического университета "Харьковский политехнический институт" А.Ю. Ницын,*

*Поступила (received) 07.12.2018*

Dashkevich Andrey, Cand. Tech, Sci.
National Technical University "KhPI",
Str. Kyrpychova, 2, Kharkiv, Ukraine, 61002
Tel.: +38 (095) 388-04-56, e-mail: dashkewich.a@gmail.com
ORCIDID:0000-0002-9963-0998

УДК 004.93

**Зниження розмірності даних на основі розбиття простору на регулярну сітку / Дашкевич А.О. //** Вісник НТУ "ХПІ". Серія: Інформатика та моделювання. – Харків: НТУ "ХПІ". – 2018. – № 42 (1318). – С. 12 – 19.

Запропоновано підхід до розв'язання задачі класифікації точкових множин на основі зниження розмірності даних і розбиття простору на регулярну сітку. Вводиться поняття гіперкубу як способу представлення точкових множин. Запропоновано підхід до зниження розмірності на основі сигнатурі точкової множини. Метод, що розроблено дозволяє виключити з подальшої класифікації множину координатних осей при підвищенні точності класифікації і зниженні кількості необхідних обчислень. Проведені експерименти показали придатність підходу на даних великих розмірностей. Перевагою підходу є швидке визначення надлишкових координатних осей для довільного набору вихідних класів. Іл.: 1. Бібліогр.: 17 назв.

**Ключові слова:** гіперкуб; регулярна сітка; сигнатура; точкова множина; дані великих розмірностей; класифікація.

УДК 004.93

**Снижение размерности данных на основе разбиения пространства на регулярную сетку / Дашкевич А.А. //** Вестник НТУ "ХПИ". Серия: Информатика и моделирование. – Харьков: НТУ "ХПИ". – 2018. – № 42 (1318). – С. 12 – 19.

Предлагается подход к решению задачи классификации точечных множеств на основе снижения размерности данных и разбиения пространства на регулярную сетку. Вводится понятие гиперкуба как способ представления точечных множеств. Предложен подход к снижению размерности на основе сигнатуры точечного множества. Разработанный метод даёт возможность исключить из дальнейшей классификации множество координатных осей при повышении точности классификации и уменьшении количества необходимых вычислений. Проведённые эксперименты показали работоспособность подхода на данных больших размерностей. Преимуществом подхода является быстрое определение избыточных координатных осей для произвольного набора исходных классов. Ил.: 1. Библиогр.: 17 назв.

**Ключевые слова:** гиперкуб; регулярная сетка; сигнатура; точечное множество; данные больших размерностей; классификация.

UDC 004.93

**Dimensionality reduction of data based on splitting space into regular grid / Dashkevich A. //** Herald of the National Technical University "KhPI". Series of "Informatics and Modeling". – Kharkov: NTU "KhPI". – 2018. – № 42 (1318). – P. 12 – 19.

The approach to classification of point sets based on dimensionality reduction and splitting space into regular grid is proposed. In paper we introduce concept of hypercube as the representation of point sets. The approach to reduce dimensionality based on point set signature as characteristic of spatial distribution of the set is proposed. The method described provides to exclude set of coordinate axis from classification with the increasing of classification precision and decreasing of computational cost. The experiments done show efficiency of our approach for multi-dimensional data. The advantage of proposed approach is fast determination of redundant coordinate axis for arbitrary set of input classes. Figs.: 1. Refs.: 17 titles.

**Keywords:** hypercube; regular grid; signature; point set; multi-dimensional data; classification.