УДК 004.93 DOI: 10.31388/2220-8674-2018-2-48

АЛГОРИТМ ПОСТРОЕНИЯ РАЗДЕЛЯЮЩЕЙ ПОВЕРХНОСТИ ДВУХ ТОЧЕЧНЫХ МНОЖЕСТВ МЕТОДОМ РАЗБИЕНИЯ ПРОСТРАНСТВА НА РЕГУЛЯРНУЮ СЕТКУ

Дашкевич А. А., к.т.н.
Национальный технический университет «Харьковский политехнический институт»
Тел. (057) 707–64–31

Аннотация – в работе рассмотрен подход к решению задачи классификации данных двух точечных множеств на основе построения их разделяющей поверхности. Предлагается понятие гиперкуба, как расширение метода пространственного хеширования. Обобщенный подход к построению разделяющей поверхности двух точечных множеств заключается в разбиении пространства, занимаемого множествами на регулярную сетку с помощью метода пространственного хеширования, построения гиперкуба для полученной сетки и нахождения значений в ячейках гиперкуба методом проведения дискретизированных гиперпрямых для нахождения средней ячейки гиперкуба между двумя ячейками, принадлежащими разным классам. Наиболее вероятный класс для новых точек определяется знаком и модулем значения в той ячейке гиперкуба, в которой находится эта точка. Преимуществом предложенного подхода является простота вычислений и возможность расширения для данных произвольной размерности.

Ключевые слова **– гиперкуб, пространственное хеширование, разделяющая поверхность, регулярная сетка, точечное множество, классификация.**

Постановка проблемы. В настоящее время существует большое количество подходов к классификации многомерных данных, среди которых можно выделить такие направления: методы основанные на поиске ближайших соседей [1], методы поиска разделяющих гиперплоскостей на основе машин опорных векторов [2], построение деревьев решающих правил [3], нейронные сети [4] и др.

При этом, алгоритмы, основанные на поиске ближайших соседей являются одними из самых простых в реализации и наглядности результата с точки зрения геометрической структуры данных. Недостатком таких методов является недостаточная гибкость в настройке алгоритма для различных типов данных и размерностей.

---

* Науковий консультант: Шоман О. В. д.т.н., проф.

Анализ последних исследований. В большинстве работ, посвящённых методам классификации, не уделяется достаточно внимания геометрической структуре исходных данных. Можно выделить подходы, основанные на разбиении пространства параметров на регулярные и нерегулярные сетки [5]. Так, в работе [6] предлагается подход к классификации данных на основе адаптивных разреженных сеток и его преимущества перед регулярными сетками, а в работе [7] показана взаимосвязь решающих правил классификатора и точек в многомерном пространстве. В работе [8] предлагаются методы классификации на основе разбиения на сетки для задач управления транспортными средствами. В работе [9] используются пространственные характеристики данных для прогнозирования преступлений. В работе [10] сеточное представление графов применяется для классификации изображений. В работе [11] предложен алгоритм пространственного хеширования на основе разбиения многомерных пространств на сетки для поиска ближайших соседей.

Формулирование целей статьи. Разработка метода построения оболочки точечного множества на дискретной сетке и алгоритма классификации на её основе.

Основная часть. В работе предлагается расширение алгоритма [11] - концепция линейного хеша H_d^T — пространственный хеш, в котором на одно пространственное измерение d_i приходится один разряд хеша, T — разрешение сетки, максимальное целочисленное значение в её ячейках. Примеры линейных хешей:

- 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 $\{0, 1\}$ и т.д.

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. Другим преимуществом является быстрый поиск соседних ячеек - у линейного хеша 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 $H_2^{10} = [4, 2]$ ближайшие 8 хеш-ячеек: $[3, 1]$, $[3, 2]$, $[3, 3]$, $[4, 1]$, $[4, 3]$, $[5, 1]$, $[5, 2]$, $[5, 3]$.

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 даёт возможность проводить, например, прямое сравнение точечных множеств без учёта различного количества точек в них.

В работе предлагается метод построения разделяющей поверхности для двух точечных множеств на основе дискретизации пространства и алгоритма пространственного хеширования [11]. Общая схема алгоритма следующая:

1) Инициализируется гиперкуб, необходимой размерности, все ячейки заполняются нулями;

2) Вычисление хеш-функции для всех попарных сочетаний точек из двух множеств;

3) Для каждой пары хешей вычисляется ячейка гиперкуба, находящаяся посередине между ними;

4) Нахождение всех ячеек гиперкуба вдоль дискретной гиперпрямой, проходящей через крайние ячейки пары через центральную ячейку;

5) Все ячейки линии с шага (4), находящиеся ближе к одной крайней ячейке хеша увеличиваются на *+1*, а ячейки, находящиеся по другую сторону — на *-1*, в центральную ячейку устанавливается *0*;

6) В результате ячейки гиперкуба, относящиеся к одному множеству будут иметь положительные значения, относящиеся к другому — отрицательные. Разделяющие ячейки — значения близкие к *0*. Максимальные по модулю значения, будут находиться в ячейках гиперкуба, через которые прошло максимальное количество линий — центроиды множеств.

Схематическое изображение построения разделяющей поверхности показано на рис. 1. На рис. 2 показаны исходные точечные множества для классификации. На рис. 3 показано разбиение пространства на области при значении разрешения сетки хешей *T=64*.

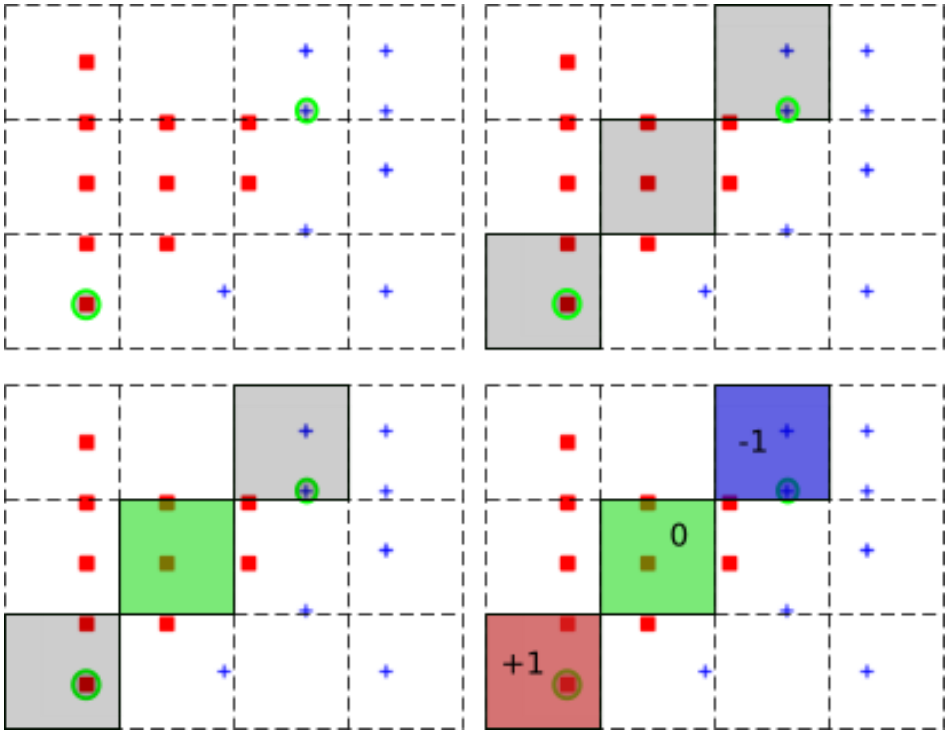

Рис. 1. Иллюстрация работы алгоритма

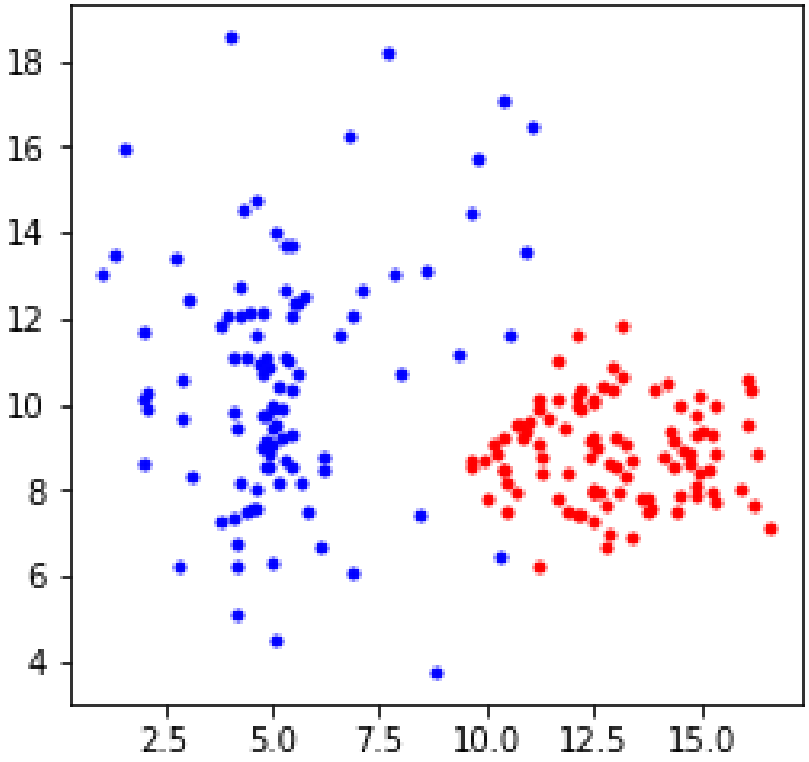

Рис. 2. Исходные точечные множества

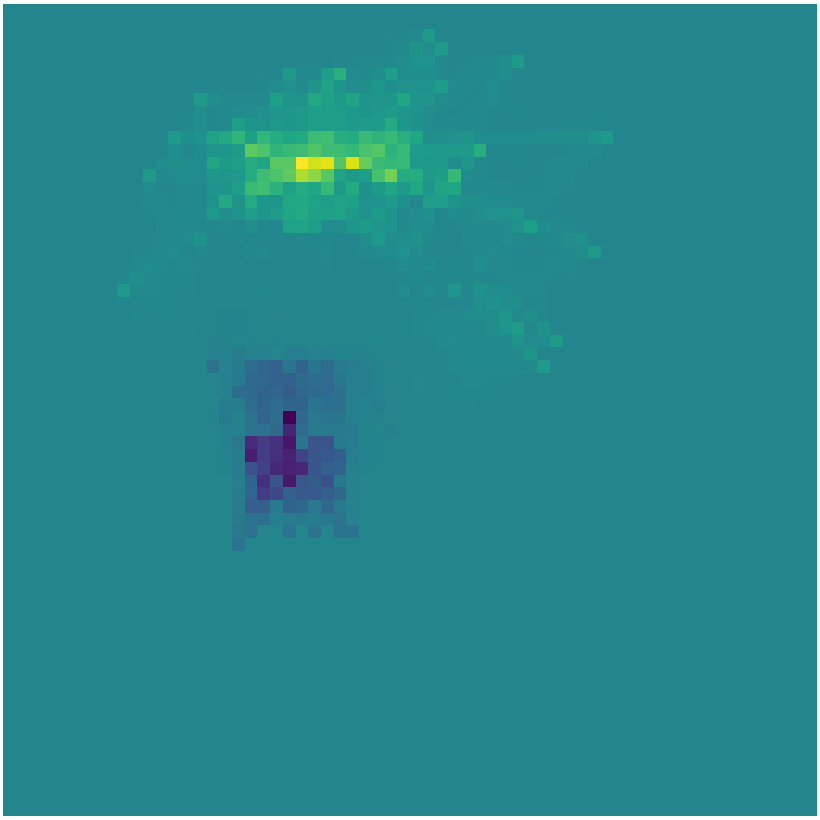

Рис. 3. Результат работы алгоритма при T = 64 (изображение увеличено)

Представленный алгоритм реализован в виде программного приложения на языке Python с использованием библиотеки NumPy.

Выводы и перспективы дальнейших исследований. В результате работы разработан метод построения разделяющей поверхности двух точечных множеств на основе дискретизации пространства на регулярную сетку.

Предложенный алгоритм обладает простотой вычислений и программной реализации. Метод может быть применён для классификации данных произвольной размерности.

К недостаткам метода следует отнести невозможность построения разделяющей поверхностей и классификации при числе классов >2.

Дальнейшие исследования будут направлены на повышение точности классификации и расширение алгоритма на случай мультиклассовой классификации.

Література

1. Ougiaroglou, S., Nanopoulos, A., Papadopoulos, A. N., Manolopoulos, Y., & Welzer-Druzovec, T. (2007). Adaptive k-Nearest-Neighbor Classification Using a Dynamic Number of Nearest Neighbors.

ADBIS'07 Proceedings of the 11th East European conference on Advances in databases and information systems, 4690, 66-82.

2. Wenzel, F., Galy-Fajou, T., Deutsch, M., & Kloft, M. (2017). Bayesian Nonlinear Support Vector Machines for Big Data. *Machine Learning and Knowledge Discovery in Databases*. ECML PKDD, 307-322. doi: 10.1007/978-3-319-71249-9_19.
3. Painsky, A., & Rosset, S. (2017). Cross-Validated Variable Selection in Tree-Based Methods Improves Predictive Performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(11), 2142-2153. doi: 10.1109/TPAMI.2016.2636831.
4. Najibi, M., Rastegari, M., & Davis, L.S. (2016). G-CNN: An Iterative Grid Based Object Detector. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2369-2377. doi: 10.1109/CVPR.2016.260.
5. Garcke, J., & Griebel, M. (2002). Classification with sparse grids using simplicial basis functions. *Intelligent Data Analysis*, *6*(6), 483-502.
6. Pflüger, D., Muntean, I. L., & Bungartz, H.-J. (2007). Adaptive Sparse Grid Classification Using Grid Environments. *Computational Science – ICCS 2007*. Springer Berlin Heidelberg, Berlin, Heidelberg, 708–715. doi: 10.1007/978-3-540-72584-8_94.
7. Gupta, P., & McKeown, N. (2001). Algorithms for packet classification. *IEEE Network*, *15*(2), 24-32. doi: 10.1109/65.912717.
8. Rieken, J., Matthaei, R., & 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, 2049-2056.
9. Lin, Y.-L., Yen, M.-F., & Yu, L.-C. (2018). Grid-Based Crime Prediction Using Geographical Features. *ISPRS International Journal of Geo-Information*, *7*, 298. doi: 10.3390/ijgi7080298.
10. Deville, R., Fromont, E., Jeudy, B., & Solnon, C. (2016). GriMa: A Grid Mining Algorithm for Bag-of-Grid-Based Classification. *Structural, Syntactic, and Statistical Pattern Recognition*: Joint IAPR International Workshop, S+SSPR 2016, Mérida, Mexico, November 29 - December 2, 2016, Proceedings, 132-142. doi: 10.1007/978-3-319-49055-7_12.
11. 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.

АЛГОРИТМ ПОБУДОВИ РОЗДІЛЬНОЇ ПОВЕРХНІ ДЛЯ ДВОХ ТОЧКОВИХ МНОЖИН МЕТОДОМ РОЗБИТТЯ ПРОСТОРУ НА РЕГУЛЯРНУ СІТКУ

А. О. Дашкевич

Анотація

В роботі розглянуто підхід до розв'язання задачі класифікації даних двох точкових множин на основі побудови роздільної поверхні. Пропонується поняття гіперкуба, як розширення метода просторового хешування. Узагальнений підхід до побудови роздільної поверхні полягає в розбитті простору, що займають точкові множини, на регулярну сітку з використанням методу просторового хешування, побудови гіперкуба для отриманої сітки і знаходження значень в клітинах гіперкубу методом проведення дискретизованих гіперпрямих для знаходження середньої клітини гіперкуба між двома клітинами, що належать різним класам. Найбільш імовірний клас для нових точок визначається знаком та абсолютним значенням в тій клітині гіперкуба, в якій знаходиться ця точка. Перевагою запропонованого підходу є простота обчислень і можливість розширення для даних довільної розмірності.

Ключові слова: гіперкуб, просторове хешування, поверхня, що розділяє, регулярна сітка, точкова множина, класифікація.

ALGORITHM OF CONSTRUCTING OF SEPARATING SURFACE FOR TWO POINT SETS BY SPLITTING OF THE SPACE INTO REGULAR GRID

A. Dashkevich

Summary

In our work the approach to solve classification problem by construction of separating surface for two point sets data is considered. The concept of hypercube as the extension of spatial hashing technique is proposed. Generalized approach to construct separating surface is splitting of the space, occupied by two point sets, into regular grid by the spatial hashing approach. Then we compute hypercube based on grid and find values in hypercube cells by drawing discrete hyperlines to find medial hypercube cell between two different classes cells. The most probable class for new points is defined by the sign and absolute value of the hypercube cell the point belongs to. The advantage of the approach proposed is computational and implementation simplicity and possibility of extending the algorithm to the data of arbitrary dimensionality.

Keywords: hypercube, spatial hashing, separating surface, regular grid, point set, classification.