DOI: 10.12737/article_5968b452793d15.68658529

Дорошенко А.Ю., аспирант
Курский государственный университет

О ПОДХОДЕ К РЕШЕНИЮ ЗАДАЧИ КЛАССИФИКАЦИИ ДАННЫХ НА ОСНОВЕ ПОСТРОЕНИЯ РАЗДЕЛЯЮЩЕЙ ГИПЕРПОВЕРХНОСТИ

doroshenkoay@bk.ru

В статье рассматривается предложенный нами ранее в серии работ подход к решению задачи машинного обучения классификации данных. Его основная идея состоит в нахождении разделяющей классы гиперповерхности с помощью адаптированного метода вычисления срединной оси, основанного на многомерной триангуляции Делоне. Получаемая таким способом гиперповерхность располагается на равном удалении от множеств объектов классов, что в общем случае является более предпочтительным. Рассматривается принцип контроля обобщающей способности данного классификатора, реализуемый с помощью метода сглаживания Лапласа. Приводятся некоторые результаты экспериментального исследования программной реализации предложенных методов на реальных данных, что дает возможность составить общее представление об эффективности решений в целом. Вкратце анализируются основные преимущества и недостатки подхода.

Ключевые слова: *классификация, машинное обучение, триангуляция Делоне, гиперповерхность, переобучение, срединная ось.*

Выбор модели для машинного обучения классификации данных всегда определяется условиями конкретной задачи и предполагает поиск компромиссного по различным критериям решения. Существующее многообразие методов и алгоритмов [1–4] достаточно полно покрывает пространство встречающихся задач, включающих, например, разработку новых лекарств [5–7], биометрическую идентификацию [8], оценку кредитоспособности [9], геостатистику [10, 11]. Тем не менее, совершенствование методологии и разработка новых идей, предполагающих получение оптимального при имеющихся ограничениях решения, являются актуальной задачей. Рассматриваемый в настоящей работе подход [12, 13] в некоторых случаях позволяет получать более предпочтительные модели, обладающие меньшей средней ошибкой распознавания.

В общей постановке задачи обучению классификации требуется построить (обучить) математическую модель (классификатор), аппроксимирующую некоторую зависимость по известному набору данных (обучающему), заданному в виде пар

$$(x_1, y_1), \dots, (x_n, y_n), \qquad (1)$$

где $x_i \in X$ – d-мерный вектор пространства признаков (объектов) X, отождествляемого здесь с Евклидовым пространством E^d, а $y_i \in Y$ – значение (класс) пространства ответов $Y = \{1, \dots, q\}$. Важно, чтобы с помощью полученного классификатора можно было бы с приемлемой точностью распознать не только данные обучающего набора, но также и неизвестные. В противном случае модель является бесполезной и называется «переобученной» или не способной к обобщению. В дальнейшем изложении будет описываться предлагаемый подход на примере разделения двух классов, поскольку обучение модели для q (q больше или равно двум) классов сводится к построению $(q-1)$ двухклассовых моделей по схеме «один против всех» – i-й классификатор отделяет класс i от классов $(i+1), \dots, q$, где $i = 1, \dots, (q-1)$. Также отметим, что приведенные в статье иллюстрации поясняют текст на примерах двумерного пространства, но предлагаемые решения применимы для пространств произвольной размерности.

Формализация задачи машинного обучения классификации образов может также интерпретироваться как задача реконструкции гиперповерхности, разделяющей классы в пространстве признаков. В отсутствии какой-либо априорной информации о распределении классов, предпочтительной более обоснованно можно считать равноудаленную от границ классов гиперповерхность. Учитывая недостаточную полноту обучающего набора, дающего возможно лишь приближенное представление об истинном распределении, такой выбор гарантирует наиболее уверенную классификацию объектов обоих классов. Вычисление данной гиперповерхности, обозначаемой здесь и далее HS, можно свести к задаче вычисления срединной оси (см. [14, 15]) незамкнутой фигуры, состоящей из гиперповерхностей границ классов (см. рисунок 1).

Для оценивания граничных гиперповерхностей классов, требуется установить некоторый критерий необходимой сложности их структуры, поскольку при излишней или недостаточной выбранной детальности представления, результиру-

ющий классификатор может оказаться не способным к обобщению. В такой постановке вопроса можно пытаться найти наиболее простую возможную форму описания, обеспечивающую максимальный отступ между ближайшими точками различных классов. Приближенное к такому представление может быть получено с помощью многомерной триангуляции Делоне (см. [14, 15]).

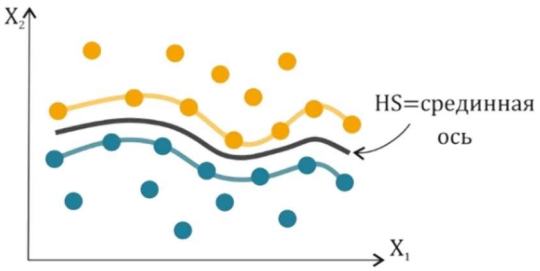

Рис. 1. Незамкнутая фигура, образованная границами гиперповерхностей (линий) классов, и её срединная ось

Рассмотрим триангуляцию Делоне $DT(x_1, \dots, x_n)$ для векторов $x_1, \dots, x_n$ обучающей выборки. Симплексы (фигура в d-мерном пространстве, имеющая $(d+1)$ вершин, не лежащих в одной гиперплоскости) этой триангуляции, имеющие среди своих вершин точки различных классов, очевидно, располагаются между классами (рис. 2а) и далее будут называться виртуальными (равно как и их ребра, имеющие вершины в точках различных классов). Предлагается совокупность их вершин считать вершинами гиперповерхностей границ классов, при этом для вычисления срединной оси, как целевой разделяющей гиперповерхности HS, с использованием такого представления можно применять соответствующие алгоритмы, основанные на триангуляции Делоне [16, 17]. Адаптированный с учетом специфики задачи алгоритм описывался в работе [12].

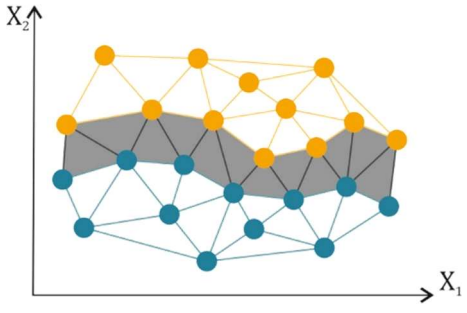

Рис. 2. Представление незамнутой фигуры, образованной виртуальными симплексами триангуляции Делоне

Разделяющая гиперповерхность задается в виде совокупности сегментов, где каждый сегмент является сечением виртуального симплекса гиперплоскостью, проходящей через середины виртуальных ребер этого симплекса. Виртуальные ребра симплекса определены множеством пар его вершин

$$\{(p_i, q_j)\}_{i=1,j=1}^{l,\ m}, \quad (2)$$

то множество вершин соответствующего симплексу сегмента задается как

$$\{(p_i + q_j)0.5\}_{i=1,j=1}^{l,m}, \quad (3)$$

где $p_1, \dots, p_l$ и $q_1, \dots, q_m$ – вершины симплекса, являющиеся элементами первого и второго классов. Отдельно заметим, что общее число вершин, определяемое произведением lm оценивается как

$$d \le lm \le \left\lfloor \frac{(d+1)^2}{4} \right\rfloor, \quad (4)$$

где $\lfloor \cdot \rfloor$ – целая часть значения, заключенного в скобки.

В рассматриваемом подходе сегмент задается пересечением $k \in \{d, d+1\}$ полупространств $\beta_1, \dots, \beta_k$, ограниченных проходящими через его стороны гиперплоскостями, с гиперплоскостью α, в которой он лежит (Рисунок 3).

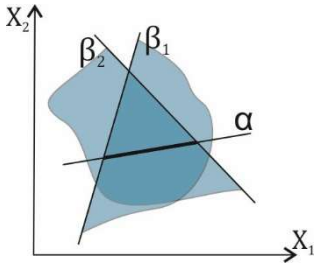

Рис. 3. Представление сегмента: β_1, β_2 – полупространства, проходящие через стороны сегмента, α – плоскость, в которой лежит сегмент

Формально сегмент можно описать в виде множества точек

$$S = \{x \in E^d | x \in (\alpha \cap \beta_1 \cap \dots \cap \beta_k)\}, \quad (5)$$

или, подставив соответствующие неравенства полупространств и уравнение гиперплоскости, и обозначив через $u \cdot v$ операцию скалярного произведения некоторых векторов u и v:

$$S\{x \in E^d | w_\alpha \cdot x + b_\alpha = 0, w_{\beta_i} \cdot x + b_{\beta_i} \ge 0, i = 1, \dots, k\}, \quad (6)$$

где $w_\alpha, w_{\beta_1}, \dots, w_{\beta_k}$ – нормальные векторы гиперплоскости α и ограничивающих полупространства $\beta_1, \dots, \beta_k$ гиперплоскостей соответственно, $b_\alpha, b_{\beta_1}, \dots, b_{\beta_k}$ – смещение от начала координат.

Для вычисления параметров уравнения гиперплоскости α в d-мерном пространстве находятся d некомпланарных точек среди вершин сегмента. Для исключения полного перебора всех возможных вариантов, предлагается выбирать комбинацию из вершин сегмента, вычисленных

из виртуальных ребер, множество концевых точек которых включает все вершины симплекса. В качестве гиперплоскостей, ограничивающих полупространства $\beta_1, \dots, \beta_m$, выбираются гиперплоскости, проходящие через грани симплекса, и для определения параметров неравенств полупространств используются вершины этих граней.

Рассмотрим применение разработанного метода для классификации объектов. Очевидно, что данная задача сводится к определению положения относительно гиперповерхности произвольной точки, для чего предлагается использовать адаптированный метод трассировки луча (*even-odd rule*) [14]. В отличие от оригинального алгоритма в измененной версии устанавливается, лежит ли распознаваемая точка p с той же стороны, относительно гиперповерхности, что и некоторая точка q. Как следствие, если известно, к какому классу принадлежит q, что можно обеспечить, выбрав её из обучающей выборки, можно указать и класс точки p. Формально идея заключается в определении числа пересечений отрезка прямой линии (p, q) с сегментами гиперповерхности. При четном количестве пересечений или их отсутствии точка p находится с той же стороны, что и q (см. рис. Рис. 4(а) и 4(б)), при нечетном – с разных (рис. Рис. 4(в)).

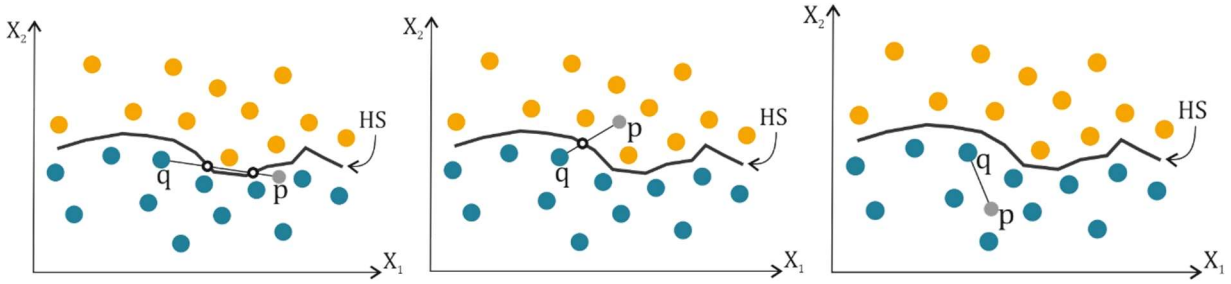

Рис. 4. Идея определения положения точки p относительно разделяющей гиперповерхности: при четном количестве пересечений (а) или их отсутствии (в) точка p лежит с той же стороны, что и q, а при нечетном (б), соответственно, по разные

.

Для проверки пересечения отрезка и сегмента вычисляется точка пересечения отрезка с плоскостью α сегмента, и, если она принадлежит полупространствам $\beta_1, \dots, \beta_k$, принимается, что пересечение есть. Формально данное условие записано в (5) и (6). Отдельно следует обсудить случай, когда отрезок проходит через грань сегмента. Следуя общей рекомендации для классического алгоритма трассировки луча, выполняется малое смещение (поворот) рассматриваемого отрезка прямой линии в произвольную сторону так, чтобы было или явное пересечение, или его отсутствие. В целом такие действия не влияют на точность результата. В данной работе также предлагается смещать отрезок не в произвольном направлении, а в направлении нормали гиперплоскости, задающей грань, через которую проходит отрезок. Тем самым доставляется пересечение с сегментом, что повышает точность классификации в случае прохождения отрезка через краевые сегменты разделяющей гиперповерхности.

Качество получаемого описанным методом классификатора сильно зависит от полноты и качества представленной для обучения выборки, поэтому существует вероятность получения модели, не способной к обобщению для распознавания неизвестных объектов. Эффективным статистическим подходом снижения такого риска является поиск модели, оптимальной по двум противоречивым критериям – количеству ошибок на обучающем наборе и её сложности [18, 19]. Второй критерий выбирается из априорных предположений о целевой функции, например, что она должна быть гладкой или, что её параметр по модулю не должен превышать некоторого значения. Поскольку в предлагаемом методе используется геометрический подход к построению классификатора, его трудно явно свести к такой постановке. Однако, применяя методы сглаживания поверхностей, использующие информацию о смежных точках, при различных конфигурациях параметров, можно получить более предпочтительное исходному решение. В настоящей работе для достижения поставленной цели рассматривается метод Лапласа (Laplacian smoothing) [20–22].

Пусть поверхность задана в виде (V, E), где V содержит её вершины $x_1, \dots, x_n$, а E определяет множества пар индексов связанных между собой точек (ребра), а также для каждой из вершин известны индексы смежных точек

$$N_i = \{j | (i, j) \in E\},\ i = 1, \dots, n. \qquad (7)$$

Обозначив исходный набор точек как $x_1(0), \dots, x_n(0)$, метод сглаживания Лапласа на итерации $k \in \{1, \dots, R\}$ можно формализовать в следующем виде:

$$x_i(k) = x_i(k-1) + \beta|N_i|^{-1}\sum_{j\in N_i}(x_j - x_i(k-1)),\ i = 1,\dots,n, \qquad (8)$$

где $|\cdot|$ – операция определения числа элементов (мощности) множества, заключенного в скобки, а R и β – параметры сглаживания, определяющие количество итераций и величину (вес) смещения соответственно. Как видно, приведенные формулы (7) и (8) не зависят от размерности поверхности сглаживания, поэтому их можно применять в пространствах произвольной мерности.

Используемое представление разделяющей гиперповерхности не позволяет корректировать позиции вершин сегментов, поскольку, согласно (4) их может быть больше чем размерность пространства. Переместив какую-нибудь из них, возникнет неоднозначность при определении сегмента, так как его вершины будут лежать в разных гиперплоскостях. Поэтому предлагается сглаживать не разделяющую гиперповерхность, а граничные гиперповерхности классов, получаемые с помощью многомерной триангуляции Делоне (см. рис. 2). При этом также решается и задача нахождения смежных точек – точки одного класса, являющиеся вершинами виртуального симплекса, будут смежными друг другу. Подбор параметров сглаживания требует выполнения ряда тестов для получения приемлемого результата [13]. В ходе экспериментов установлено, что для многих задач достаточно 2–10 итераций сглаживания, а величина β лежит диапазоне $[0.001,1]$ и обратно пропорциональна количеству итераций R.

Для определения практической значимости предложенного подхода разработана его программная реализация с использованием объектно-ориентированного языка программирования Java в свободно распространяемой среде программирования с открытым исходным кодом Eclipse IDE for Java Developers версии Luna Service Release 2 (4.4.2), предоставляющей широкие возможности и удобный интерфейс для решения поставленной задачи. Тестирование проводилось на реальных данных, доступных в базе данных UCI [23]. В силу существующих ограничений рассматриваемого метода выбирались наборы данных, имеющих до восьми признаков, описание которых отражено в таблице ниже.

Таблица 1

Используемые для тестирования наборы данных, доступные в базе UCI

Наименование	Количество признаков	Количество классов	Размер выборки	Тип данных
Balance Scale	4	3	625	категориальные
Banknote Authentification	4	2	1372	вещественные
Haberman’s Survival	3	2	306	целые
Iris	4	3	150	вещественные
Qualitative Bankruptcy	6	2	250	категориальные
Skin Segmentation	3	2	245057	вещественные
Blood Transfusion Service Center	4	2	748	вещественные

Из исходного набора данных 70 % элементов использовались для обучения, а оставшиеся 30 % для тестирования модели. В качестве аналогов использовались известные реализации различных моделей классификации, включенные в свободно распространяемые программные библиотеки WEKA [24] и LIBSVM [25]. В частности, применялись метод опорных векторов (SVM) с радиальной базисной функцией в качестве ядра, метод k-ближайших соседей (kNN), обучаемые алгоритмом C4.5 решающие деревья (J48), метод «Random forest» (RF), ансамблевый алгоритм (бустинг) AdaBoost. Оптимальные параметры, при которых достигалась наибольшая точность классификации для каждой модели подбирались экспериментально. Полученные результаты приведены в табл. 2

Как видно из таблицы на некоторых наборах данных точность модели, построенной предложенным методом, выше, чем у аналогов. Наибольшее преимущество достигается при распознавании классов, описываемых вещественными признаками.

Рассмотренный выше подход наиболее близок к методу ближайшего соседа. Действительно, разбиение пространства признаков, полученное методом ближайшего соседа, соответствует двойственной триангуляции Делоне структуре – диаграмме Вороного (см. [14, 15]), использование которой для аппроксимации срединной линии и привело бы к идентичному методу (рис. 5).

Таблица 2

Точность классификации различных алгоритмов на наборах реальных данных, доступных в базе UCI

Наименование	HS	AB	J48	RF	kNN	SVM
Balance Scale	88.3%	75.2%	80%	81.1%	91.7%	**100%**
Banknote Authentification	**100%**	**100%**	99%	99.4%	**100%**	**100%**
Haberman's Survival	**79%**	75.8%	75.8%	71%	**79%**	77.4%
Iris	97.1%	97.1%	97.1%	97.1%	**100%**	**100%**
Qualitative Bankruptcy	100%	100%	100%	100%	100%	100%
Skin Segmentation	**99.7%**	98.8%	98.11%	98.7%	99.5%	99.3
Blood Transfusion Service Center	80.7%	80.7%	80.7%	71.1%	**82.4%**	79.8%

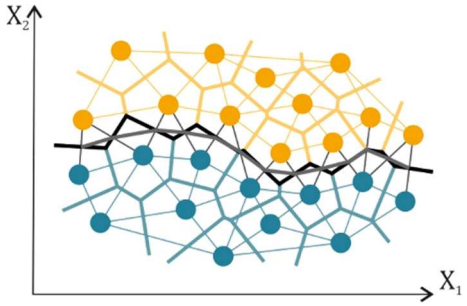

Рис. 5. Разделение точек классов с помощью Диаграммы Вороного. Серая линия показывает разделение с помощью предлагаемого метода

Однако, точность распознавания классификатора, получаемого выбранным способом, на разных выборках превышает точность метода ближайшего соседа и его обобщения – метод k ближайших соседей. Применяя метод сглаживания Лапласа можно избежать проблемы переобучения, хотя данная процедура требует проведения экспериментов для подбора параметров, что повышает время обучения.

Область применения предложенного метода ограничена пространствами 2–8 измерений. Данное утверждение отталкивается от верхней оценки количества симплексов триангуляции $O(n^{\lceil d/2 \rceil})$ [26] (где $\lceil u \rceil$ означает округление u к большему целому числу), откуда видно, что уже для 100 точек (которых может быть недостаточно для адекватной оценки существующей зависимости) в пространстве с девятью измерениями будет порядка 10^{10} симплексов. Несмотря на указанное ограничение, метод может эффективно применяться в некоторых задачах классификации, включающих классы со сложными линейно неразделимыми конфигурациями их пространственных образов.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Marshland S. Machine learning: an algorithmic perspective. CRC press, 2009. 390 pp.

2. Хайкин С. Нейронные сети: полный курс, 2-е издание: Пер. с англ. Москва: Издательский дом Вильямс, 2006. 1104 с.

3. Shalev-Shwartz S., Ben-David S. Understanding machine learning: From theory to algorithms. Cambridge University Press, 2014. 449 pp.

4. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. Springer Series in Statistics, 2009. 745 pp.

5.Murphy R.F. An active role for machine learning in drug development // Nat. Chem. Biol. 7. 2011. 327–330 pp.

6.Lavecchia A. Machine-learning approaches in drug discovery: methods and applications // Drug Discov. today. 2015. No. 20. 318–331 pp.

7.Gawehn E., Hiss J.A., Schneider G. Deep learning in drug discovery // Mol. Inform. 2016. No. 35 (1). 3–14 pp.

8.Banerjee S.P., Woodard D.L. Biometric authentication and identification using keystroke dynamics: A survey // Journal of Pattern Recognition Research. 2012. Vol. 7. №. 1. 116–139 pp.

9.Lessmann S., Seow H.-V., Baesens B., Thomas L. C. Benchmarking state-of-the-art classification algorithms for credit scoring: A ten year update // Eur. J. Oper. Res. 2015. Vol. 247. №. 1. 124–136 pp.

10.Белозеров Б.В., Бочков А.С., Ушмаев О.С., Фукс О.М. Использование метода ближайших соседей при восстановлении обстановки осадконакопления // Машинное обучаение и анализ данных. 2014. Т. 1. № 9. С. 1319–1329.

11. Kanevski M., Pozdnoukhov A., Timonin V. Machine Learning Algorithms for GeoSpatial Data. Applications and Software Tools // Integrating sciences and information technology for enviromental assessment and decision making. 2008. Vol. 1. Pp 320–327.

12. Довгаль В.М., Дорошенко А.Ю. Об алгоритме построения разделяющей

гиперповерхности для решения задачи классификации при линейной неразделимости классов образов // Auditorium. Электронный научный журнал Курского государственного университета. 2016. Вып. №3(7). 12 с. URL: http://auditorium.kursksu.ru/pdf/012-013.pdf (дата обращения: 04.06.2017)

13. Дорошенко А.Ю., Довгаль В.М. Об одном подходе к решению проблемы переобучения классификатора // Auditorium. Электронный научный журнал Курского государственного университета. 2017. Вып. №1(13). 9 с. URL: http://auditorium.kursksu.ru/pdf/013-009.pdf (дата обращения: 04.06.2017)

14. Ласло М. Вычислительная геометрия и компьютерная графика на C++: Пер. с англ. Москва: БИНОМ, 1997. 304 с.

15. Препарата Ф., Шеймос М. Вычислительная геометрия: Введение / Пер. с англ. М.: Мир, 1989. 478 с.

16. Kimmel R., Shaked D., Kiryati N., Bruckstein A.M. Skeletonization via distance maps and level sets // Computer Vision and Image Understanding. 1995, 62:3. 382-391 pp.

17. Zou J.J, Chang H.-H., Yan H. A new skeletonization algorithm based on constrained Delaunay Triangulation // in Proc. 5th ISSPA. 1999. Vol. 2. Australia. 927-930 pp.

18. Vapnik V.N. Statistical Learning Theory // John Wiley and Sons, Inc. New York, 1998. 768 p.

19.Vapnik V.N. The Nature of Statistical Learning Theory // Springer-Verlag. New York, 1995. 314 p.

20. Buell W.R., Bush B.A. Mesh generation – a survey // Trans. ASME, J. Eng. Ind. 1973. 332-338 pp.

21. Kobbelt L., Campagna S., Vorsatz J., Seidel H.-P. Interactive multiresolution modeling on arbitrary meshes // Computer Graphics (SIGGRAPH 98 Proceedings). 1998. 105–114 pp.

22. Vollmer J., Mencl R., Müller H. Improved Laplacian Smoothing of Noisy Surface Meshes // Computer graphics forum. Vol. 18. Wiley Online Library. 1999. 131–138 pp.

23. Linchman M. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml] // Irvine, CA: University of California. School of Information and Computer Science

24. Hall M., Frank E., Holmes G., Pfahringer B., Peter R., Witten I. H. The WEKA data mining software: an update // ACM SIGKDD explorations newsletter. 2009. 11(1). 10-18 pp.

25. Chang C.C., Lin C.H. LIBSVM: a library for support vector machines // ACM Trans. Intell. Syst. Technol. 2. 3. Article 27. 2011. 27 p.

26. Seidel R. The upper bound theorem for polytopes: an easy proof of its asymptotic version // Computation Geometry. 1995. 5(2). 115–116 pp.

Doroshenko A.Y.

ABOUT SEPARATING HYPERSURFACE METHOD FOR SOLVING THE PROBLEM OF DATA CLASSIFICATION

The article considers the approach to solve the problem of data classification that we proposed earlier. Its main idea is to find the separating classes hypersurface using an adapted method for calculating the medial axis, based on the multidimensional Delaunay triangulation. The hypersurface thus obtained is located at an equal distance from the sets of objects of classes, which in general is more preferable. The principle of control of the generalizing ability of this classifier based on the Laplacian smoothing method is considered. Some results of an experimental research of the implementation of the proposed methods on real data are given, which make it possible to draw up a general idea of the effectiveness of solutions as a whole. A brief discussion of the advantages and disadvantages is given.

Key words: *classification, machine learning, Delaunay triangulation, hypersurface, overfitting, medial axis.*

Дорошенко Александр Юрьевич, аспирант кафедры программного обеспечения и администрирования информационных систем.
Курский государственный университет.
Адрес: Россия, 305000, г. Курск, ул. Радищева, 33.
E-mail: doroshenkoay@bk.ru.