УДК 004.93

Т. Ланге, К. Мослер, П. Можаровский

Эффективная глубинная классификация с помощью проективного инварианта классовой принадлежности

Предложен новый непараметрический метод *DD*α-процедура для автоматической классификации на $\theta \geq 2$ классов по многомерным данным. Классификатор *DD*α применен на имитационных и реальных данных. Приведено сравнение частоты ошибок при применении *DD*α и других дискриминантных методов.

The *DD*α-procedure is a nonparametric method for the supervised classification of multidimensional objects originating from $\theta \geq 2$ classes. The behavior of the *DD*α-classifier is investigated on the simulated as well as real data. The new procedure outperforms many existing discrimination methods, including *SVM*, regarding training speed, while its error rate is comparable with those techniques.

Запропоновано новий непараметричний метод *DD*α-процедура для автоматичної класифікації на $\theta \geq 2$ класів багатовимірних даних. Класифікатор *DD*α застосовано до імітаційних та реальних даних. Наведено порівняння частоти помилок при застосуванні *DD*α та інших дискримінантних методів.

Введение. Статья посвящена всемирно известному ученому в области самоорганизации и самообучения в распознавании образов [1–5] академику Алексею Григорьевичу Ивахненко. Одна из авторов, Т. Ланге, с большой благодарностью и радостью вспоминает педагогический талант своего первого научного руководителя А.Г. Ивахненко, который поддержал и развил научный интерес у сотен своих аспирантов.

В статье описана идея предлагаемого метода с исторической справкой; вводится понятие преобразования глубины, которое переводит данные из *d*-размерного объектного пространства в θ-размерное пространство глубины. Также предложено первое обсуждение проблемы «аутсайдеров» (не «*outlier*») – точек, имеющих нулевой вектор глубин; рассмотрена модификация α-процедуры в некоторых деталях, а также несколько теоретических результатов поведения *DD*α-процедуры на эллиптических и зеркальных симметрических распределениях; приведены результаты имитации и сравнения; вычисления на примерах реальных данных.

Описание идеи метода

В настоящее время наблюдается новая волна интереса к теории самообучения, связанная с появлением непараметрического инструментария в статистике. При этом особенно выделяют построение функций глубины, таких как предложенная в 1974 году полупространственная глубина [6] или выдвинутая в 1990 году симплексная глубина [7]. Несмотря на то, что только в последнее десятилетие функции глубины начали применяться в классификации с использованием обучающих последовательностей (т.е. с учителем), они дали новый и мощный импульс теории распознавания образов.

В случае обучения с учителем классифицирующая функция строится с помощью обучающего набора данных в *d*-мерном пространстве, где каждая точка описывает один объект, принадлежащий определенному классу. В обучающем наборе данных «учитель» маркирует каждый объект меткой, указывающей на то, к какому классу он принадлежит. Таким образом, имеется два или больше маркированных разными «стиккерами» облаков обучающих данных в *d*-мерном пространстве. Глубина данных измеряет центрированность определенной точки относительно каждого облака. Она определяет степень близости к каждому «стиккеру» любой точки в *d*-мерном пространстве.

Для задач классификации это может быть решено по-разному. Многие авторы использовали идеи глубины данных в области классификации с учителем. В 1999 году впервые были отмечены полезность и многосторонность глубинных трансформаций в многомерном анализе [8]. Была введена нотация «*DD*-диаграммы» для двухмерного представления многомерных объектов с помощью их глубин данных относительно двух заданных распределений.

На плоскости ($\theta = 2$ для двух классов) абсциссами (без ординат) служат значения глубин.

Для каждого объекта (точки) на *DD*-диаграмме рассчитывается его глубина относительно каждого из двух классов. Принадлежность объекта к какому-либо классу можно определять по тому, в каком классе он имеет максимальную глубину.

Многие авторы [9–11] применяли этот или подобные методы. В [12, 13] описана разделяющая функция, линейная в плоскости, базирующейся на оценках ядра проекции глубины, соответственно «L_p-глубины». В настоящее время авторы [14] используют полиномиальные разделяющие функции на *DD*-диаграмме для классификации объектов с помощью представления их глубины.

Все эти методы отличаются разным использованием глубины, и все они открыты для изменений и расширений. Общим во всех упомянутых методах есть то, что многомерное пространство объектов (размерность которых иногда может быть очень большой) преобразовывается в пространство редуцированной размерности значений глубин этих объектов, и задача классификации решается в новом пространстве глубин.

Однако существует много нерешенных вопросов, связанных с глубиной данных. Какая из известных мер глубины – наилучшая? Какой классифицирующий алгоритм лучше всего применить на данных, описывающих глубину? И, наконец, можно ли расширить процедуру для случаев, где имеется больше чем два класса ($\theta > 2$)? Упомянутые источники по-разному отвечают на эти вопросы.

Полупространственные и симплексные глубины, использованные в работах [8, 9, 14], зависят только от комбинаторной структуры данных. При этом они устойчивы в охваченной области данных, т.е. они робастны к выбросам, но вычисления могут быть громоздкими или даже невозможными. С другой стороны, глубина Махаланобиса [15], тоже применяемая в упомянутых трудах, требует мало вычислений, но в большей степени неробастна. Более того, она зависит только от первых двух моментов и не отражает несимметричности данных. Более робастные формы глубины Махаланобиса также остаются нечувствительными к несимметричности данных.

Использованная в [9] L_1-глубина имеет аналогичные недостатки. В [12] применяется форма L_p-глубины, которая легко вычисляется при известном p. Можно определить p с помощью адаптивной процедуры, но для этого необходим большой объем вычислений. В [11] используется максимальная глубина зоноида в комбинации с глубиной Махаланобиса. Тот и другой метод могут эффективно использоваться даже в случае большой размерности, но оба страдают отсутствием робастности.

В [14] описан другой подход, в котором проблема классификации на *DD*-диаграмме решается конструированием полиномиальной линии, разделяющей единичный квадрант с минимальной частотой ложных классификаций. Степень полинома (до трех) выбирается путем перекрестной проверки. Подобным образом (методом валидации) определяются разделяющие функции в [12, 13]. При наличии $\theta > 2$ классов обычно используется двухэтапная процедура классификации: в начале применяются $\binom{\theta}{2}$ классификаторов, выбранных попарно для всех классов; на втором этапе точку относят к тому классу, к которому ее чаще всего относили на этапе 1.

В данной статье применяется глубина зоноида [16, 17], так как она эффективно вычисляется для многомерных данных (включая $d = 20$ и более), а также имеет отличные теоретические свойства – непрерывность и статистическое заключение. Однако глубина зоноида имеет и слабые стороны. Например, если важна робастность, то необходимо предварительно обработать данные процедурой нахождения выбросов.

Для финальной классификации в пространстве глубин используется разновидность α-процедуры [18–20], которая просто и очень эффективно работает на пространствах с небольшим числом измерений, а пространство глубин таким и есть. Для классификации двух классов используем *DD*-диаграммы, а для $\theta > 2$ классов

применяем θ-мерные глубинные диаграммы. Принадлежность точки к классу определяется по максимальному значению одного из $\binom{\theta}{2}$ бинарных классификаторов в θ-мерном пространстве глубин.

Следует обратить внимание на то, что при каждой бинарной классификации используется вся информация о глубинах, которая относится ко всем θ классам. Такой метод назовем *DD*α-методом и применим его как на искусственных, так и на реальных данных. Полученные результаты отличаются от представленных в [12–14].

Особенности процедуры классификации:

- эффективность процедуры вычисления для объектов большой размерности;
- высокая скорость процедуры классификации *D*-трансформированных данных;
- использование полной многомерной информации для классификации θ > 2 классов.

Преобразование глубины

Функция глубины – это функция, показывающая, насколько близко к «центру» конечного множества $X \in R^d$ находится заданная точка *x*, т.е. насколько «глубоко» она находится в этом множестве. Функция глубины – функция

$$(x, X) \mapsto D_X(x) \in [0,1], \quad x \in R^d, \quad X \subset R^d,$$

удовлетворяющая следующим ограничениям: аффинно инвариантна; полунепрерывна сверху; квазивогнута в точке x (т.е. имеет выпуклые верхние Лебеговы множества); стремится к нулю при $\| x \| \to \infty$.

Иногда налагаются два более слабых ограничения: ортогональная инвариантность; убывание вдоль лучей, которые начинаются в точке с максимальной глубиной (т.е. имеет верхние Лебеговы множества в виде звезды). Исследование этих ограничений и многих специальных понятий глубины данных описано в [17, 21–24].

Пусть данные в R^d необходимо классифицировать на θ ≥ 2 классов, и $X_1, \ldots, X_q \subset R^d$ – обучающие множества для этих классов, имеющие конечные размерности $n_j = | X_j |$; *D* – глубина данных. Функцию $R^d \to [0,1]^q$, отображающую:

$$x \mapsto d := (D_{X_1}(x), \ldots, D_{X_q}(x)), \qquad (1)$$

назовем представлением глубины. Каждый объект характеризируется вектором, в котором θ компонент указывают его глубину (близость) относительно θ классов. В частности, обучающие наборы $X_j \subset R^d$ переходят в множества в пространстве $[0,1]^\theta$, характеризующие исходные классы в глубинном пространстве. Заметим, что «близость» точек в исходном пространстве переходит в «близость» в их представлении. Теперь задачей классификации становится разбиение глубинного пространства $[0,1]^\theta$ на θ частей.

Покажем сказанное на примере (рис. 1), где исходное *d*-мерное пространство признаков трансформируется в θ-мерное глубинное пространство θ-классов.

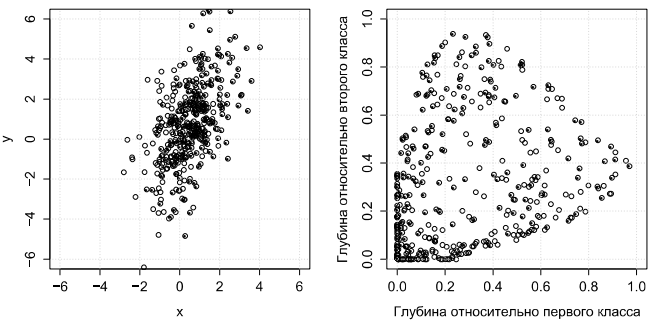

Рис. 1. Исходное пространство признаков; *DD*-диаграмма: глубинное пространство двух классов

В качестве простого правила разбиения в новом пространстве можно, например, предложить следующее: точка классифицируется как принадлежащая тому классу, в котором она имеет наибольшее значение глубины [9, 10]). Это значит, что пространство глубин разбито на θ частей, разделенных между собой с помощью θ секущих гиперплоскостей. Такая классификация с помощью максимальной глубины есть линейным правилом. Нелинейное правило классификации было применено в [14], где рассмотрено построение разделяющей полиномиальной линии до третьего порядка для пространства $[0, 1]^2$, т.е. в случае, если θ = 2 [12, 13].

Учитывая несколько важных свойств глубины данных, $D_X(x)$ обращается в ноль вне выпуклой оболочки *X*. Это происходит в случае с

полупространственной и симплексной глубиной, а также глубиной зоноида, но не с глубиной Махаланобиса и L_p-глубиной. Точка, которая не находится внутри выпуклой оболочки по крайней мере одного из обучающих множеств, отображается на начало координат пространства глубин. Такую точку будем называть «аутсайдер». Конечно, она не может быть ни правильно классифицирована, ни проигнорирована. Рассмотрим три возможных подхода для классификации этой точки. Каждый из подходов содержит в себе несколько различных вариантов реализации:

- классифицировать случайным образом с вероятностями, равными ожидаемой доле входных точек быть классифицированными;
- использовать алгоритм *k*-ближайших соседей с правильно подобранным расстоянием – евклидовым, L_p-расстоянием, Махаланобиса с оценками моментов, расстоянием Махаланобиса с робастными оценками *MCD*, например [25];
- классифицировать с использованием максимальной глубины Махаланобиса (используя оценки моментов или *MCD*) или с использованием максимума другой глубины, которая должным образом расширена вне выпуклой оболочки, как, например, в [11].

В [26] доказываются следующие утверждения относительно глубины.

Утверждение 1. Пусть D – аффинно инвариантная функция глубины, а P – эллиптическое распределение. Тогда для любого $\alpha \in (0, 1]$ верхнее Лебегово множество

$$D_\alpha(P) = \{x \in R^d \mid D_P(x) \geq \alpha\} \qquad (2)$$

есть эллипсоид.

Утверждение 2. 1) Пусть D – глубина зоноида, а $P = Ell(\mu, BB', r)$ – унимодальное эллиптическое распределение. Тогда для любого непустого верхнего Лебегова множества функции плотности $\{x \in R^d \mid f(x) \geq \beta\}$ существует $\alpha = \varphi(\beta)$ такое, что:

$$\{x \in R^d \mid f(x) \geq \beta\} = D_\alpha(P). \qquad (3)$$

2) Если к тому же носитель функции r представляет собой замкнутый интервал, то φ – непрерывная, строго возрастающая функция. Тогда справедливо $D_P(x) = \varphi(f(x))$ и, следовательно,

$$f(x) \geq f(y) \Leftrightarrow D_P(x) \geq D_P(y). \qquad (4)$$

Альфа-процедура

Вначале проводится предварительный отбор и оставляется для дальнейшей обработки только те m свойств p_q, для которых можно найти два пороговых значения x_{1q}^0 и x_{2q}^0, разделяющие ось значения свойства p_q в три сегмента (рис. 2).

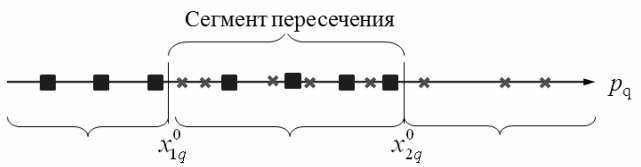

Рис. 2. Пересечение объектов, относящихся к двум классам

После этого вводим понятие *разделяющей силы*, определяемой для *отдельного свойства* как $F(p_q) = \frac{\omega_q}{l}$, где ω_q – число объектов, расположенных вне сегмента пересечения свойства p_q, l – длина обучающей выборки, т.е. число объектов.

Для синтеза пространства шаг за шагом выбираются признаки, имеющие наилучшие разделяющие мощности.

Каждый новый признак редуцирует сегмент пересечения и увеличивает число правильно классифицированных объектов. Применяется следующее определение *разделяющей силы* признака на *k*-м шагу селекции:

$$F(x_k) = \frac{\omega_k - \omega_{k-1}}{l} = \frac{\Delta\omega_k}{l}, \qquad \omega_0 = 0,$$

где ω_{k-1} – аккумулированное число правильно классифицированных объектов *перед* выбором *k*-го признака, а ω_k – такое же число *после* выбора *k*-го признака.

На первом шагу выбирается свойство с наилучшей разделяющей мощностью как базовый признак f_0 (рис. 3) и представляется вместе с его численными значениями для всех объектов как ось.

На следующем шаге к системе координат добавляется второе свойство p_k и определяются позиции всех объектов на плоскости, определяемой осями f_0 и p_k. Далее в этой плоскости формируется новая ось, которая вращается во-

круг начала системы координат *на угол* α до тех пор, пока *проекции* объектов на нее не дадут наилучшего разделения объектов.

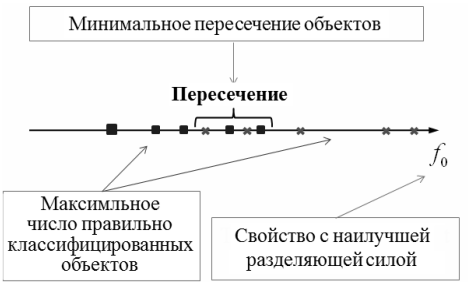

Рис. 3. Альфа-процедура, первый шаг

Эта процедура повторяется для всех оставшихся свойств, и из них выбирается *свойство*, дающее наилучшее разделение объектов на соответствующей оси $\widetilde{f}_1$. Это новое свойство берется как следующий *признак***,** и таким образом строится первый репер (рис. 4).

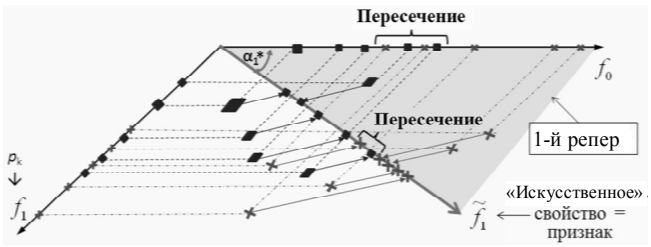

Рис. 4. Альфа-процедура, второй шаг

На третьем шаге добавляется следующее свойство p_j как третья ось и определяются позиции всех объектов на *новой плоскости*, которая образуется осями $\widetilde{f}_1$ и p_j. Затем строим *новую ось* в этой *новой плоскости* и вращаем эту ось вокруг начала системы координат на угол α до тех пор, пока *проекции* объектов на нее не дадут лучшее разделение объектов. Эта процедура повторяется для всех оставшихся свойств, и из них выбирается наилучшее, которое формирует второй репер (рис. 5). В нашем простом примере уже третий шаг ведет к безошибочному разделению объектов.

В случае, если после перебора всех признаков полное разделение объектов не достигнуто, используется специальный критерий остановки [27]. Вектор Дарбу разделяющей решающей плоскости описывается последовательностью значений:

$$\left(\prod_{k=2}^{n}\cos\alpha_k^0, +\sin\alpha_2^0\prod_{k=3}^{n}\cos\alpha_k^0, ..., + \right.$$

$$\left. +\sin\alpha_q^0\prod_{k=q+1}^{n}\cos\alpha_k^0, ..., +\sin\alpha_n^0\right);\ n \le n_0,$$

где позиция номера в последовательности соответствует номеру шага процедуры.

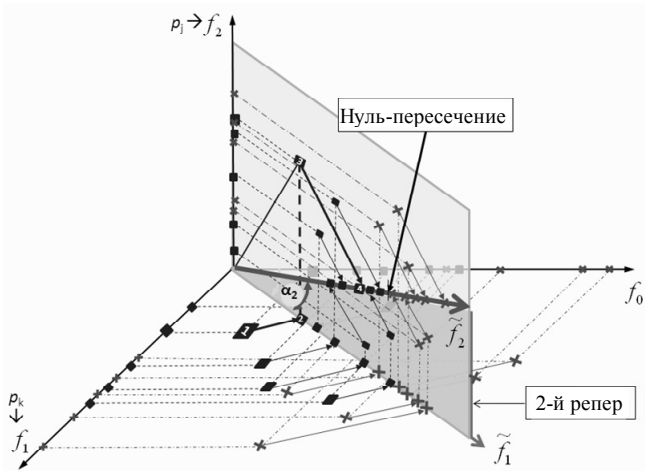

Рис. 5. Альфа-процедура, третий шаг

Учитывая, что порядок элементов вектора Дарбу определяется индукцией процедуры, они должны приписываться к своим свойствам в *обратном порядке* в практических задачах классификации. Пример разделяющей решающей плоскости и декомпозиции соответствующего направляющего вектора показан на рис. 6 и 7. Свойства, не выбранные процедурой, не учитываются для классификации.

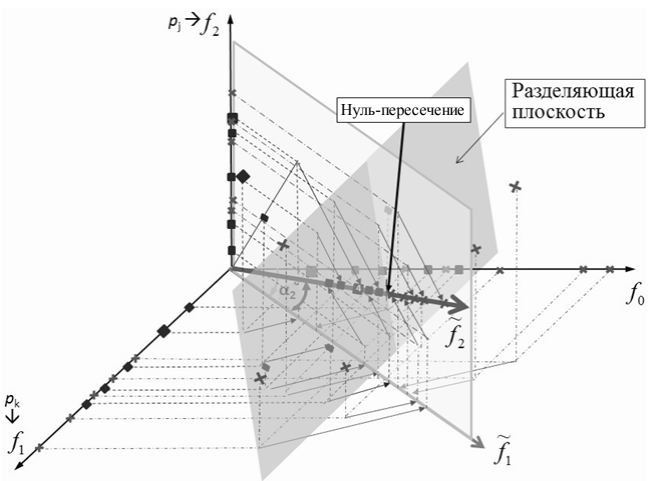

Рис. 6. Разделяющая решающая плоскость

Примечание. Если разделение объектов невозможно в оригинальном пространстве свойств, то следует расширить это пространство конструированием расширенных свойств типа $x_{iq} \cdot x_{ir}$; $i = 1, \ldots, l$; $q, r = 1, \ldots, m$.

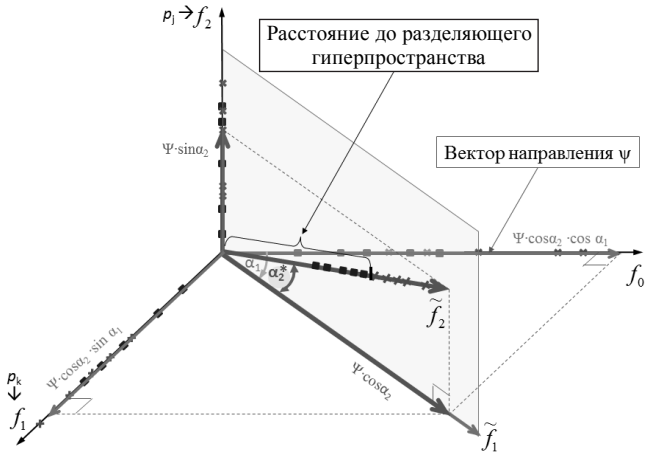

Рис. 7. Декомпозиция направляющего вектора разделяющей решающей плоскости

В [26] доказаны следующие теоремы относительно алгоритма *DD*α.

Теорема 1 (Байесово правило). Пусть F и G – распределение вероятностей в R^d с плотностями f и g соответственно; H – такая гиперплоскость, что G – зеркальное отражение F по отношению к H, и $f \geq g$ в одном из отделяемых этой гиперплоскостью полупространств. Тогда, приняв за основу $50:50$ независимых примеров из F и G, *DD*α-процедура асимптотически придет к линейному разделяющему правилу, соответствующему биссектрисе *DD*-диаграммы.

Теорема 2. Пусть F, G – унимодальные эллиптические распределения вероятностей, $F = Ell(\mu_F, BB', r)$, $G = Ell(\mu_G, BB', r)$. Тогда, приняв за основу $50:50$ независимых примеров из F и G, *DD*α-процедура асимптотически придет к линейному разделяющему правилу, соответствующему биссектрисе *DD*-диаграммы.

Имитация

Исследуем обобщающую способность *DD*α-классификатора путем моделирования, и сравнение его как с известными традиционными, так и с глубинными классификаторами, предлагаемыми в более поздней литературе. При этом внимание уделим трем случаям:

- «идеализированные» условия нормального распределения;
- условия «тяжелых хвостов» на примере распределения Коши;
- внутри- и межклассовая асимметрия (под последним термином будем понимать принадлежность классов к разным семействам распределений).

Для этой цели *DD*α-классификатор запрограммирован в среде статистического программирования «*R*» с реализацией критических по скорости модулей на *C*++; программная часть может быть получена по запросу у авторов.

Сравнение для каждой отдельной задачи обучения распознаванию образов (для каждой пары распределений, представляющих классы) происходит следующим образом. Вначале генерируется обучающая выборка, содержащая по 200 наблюдений в каждом классе, на которых и осуществляется обучение всех рассматриваемых классификаторов. После этого распознается проверочная выборка, содержащая по 500 наблюдений в каждом классе, и для каждого классификатора подсчитывается допущенная ошибка (в долях единицы). Повторение изложенной последовательности действий 100 раз формирует для каждого классификатора массив из 100 значений – ошибок распознавания, отображаемый бокс-диаграммой из «усатых ящиков». Для каждой задачи бокс-диаграммы рассматриваемых классификаторов представляют комбинированную диаграмму ошибок распознавания – компактный рисунок, удобный для визуального сравнения.

Обобщающая способность *DD*α-классификатора сравнивается в первую очередь с тремя традиционными классификаторами: линейным (*LDA*) и квадратичным (*QDA*) дискриминантным анализом и классификатором k ближайших соседей (*KNN*), где k определяется методом скользящего контроля. Рассматриваются также классификаторы максимальной глубины [10] на основе глубины Махаланобиса (*MM*), симплексной (*MS*) и полупространственной (*MH*) глубин, а также предлагаемые в [14] *DD*-классификаторы с тем же набором глубин (*DM*, *DS* и

DH соответственно). На каждой комбинированной бокс-диаграмме ошибки распознавания перечисленных классификаторов выстроены вертикально в порядке упоминания и дополнены *DD*α-классификатором (*DD*α).

Поскольку в данном разделе рассматриваются двумерные распределения, симплексная и полупространственная глубины вычисляются точно при помощи *R*-пакета «*depth*», так как на плоскости существует естественный порядок углов, а для глубины зоноида в *DD*α-классификаторе используется симплексный алгоритм [28], работающий эффективно в любой размерности. При этом исключается проблема отличающихся априорных вероятностей классов, составляющая существенные трудности для классификаторов максимальной глубины дискриминантного анализа. Для *DD*-классификаторов применяются оригинальные настройки из [14]: разделение на *DD*-диаграмме выполняется полиномиальной кривой, проходящей через начало координат, степень {1,2,3} которой выбирается методом перекрестной проверки с 10 частями, а константа логистической функции, сглаживающей эмпирический риск (что позволяет его минимизировать), принимается t=100. В качестве аргумента полиномиальной функции принимается глубина относительно «первого» класса (оси *DD*-диаграммы не меняются местами). Аутсайдеры, возникающие при использовании симплексной (в *MS* и *DS*) и полупространственной (в *MH* и *DH*) глубин и глубины зоноида (в *DD*α), относятся случайным образом к одному из классов, что можно рассматривать как наихудший, но все же «честный», вариант их классификации; это применимо для *LDA*, *QDA*, *KNN*, *MM* и *DM*.

В первом случае рассматривается нормальное распределение в двух постановках задачи классификации: с разницей *а*) только в положении и *б*) в положении и масштабе. В *а*) классы генерируются из нормальных распределений $N(\mu_1, \Sigma_1)$ и $N(\mu_2, \Sigma_2)$ с одинаковой ковариационной матрицей $\Sigma_1 = \Sigma_2 = \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}$, отличающихся лишь математическим ожиданием: $\mu_1 = (0,0)^T$ и $\mu_2 = (1,1)^T$ (комбинированные диаграммы ошибок распознавания приведены на рис. 8, вверху). Наиболее успешно справляется *LDA*, так как именно на такую модель он и рассчитан, слегка превосходя *QDA*, оперирующее двумя ковариационными матрицами, и *MM* и *DM*, допускающие эллиптичность распределения. Хотя остальные классификаторы (включая *DD*α) несколько уступают упомянутым, их применение повышает ошибку распознавания лишь незначительно – на единицы процентов.

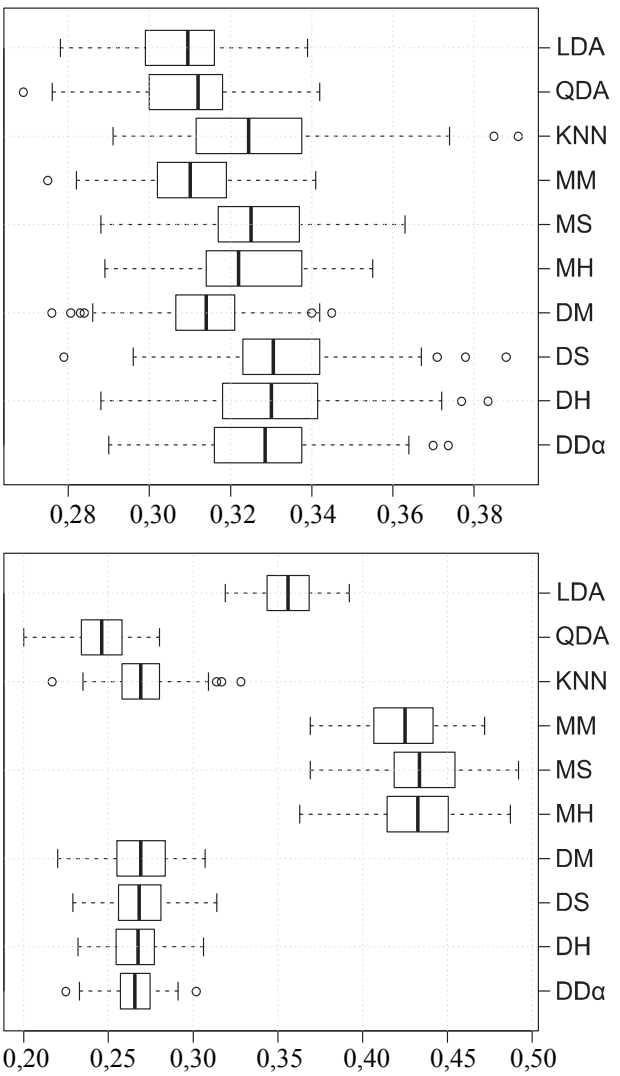

Рис. 8. Комбинированные диаграммы ошибок распознавания для нормальных распределений, отличающихся положением (вверху), положением и масштабом (внизу)

Для *б*) разными принимаются также и ковариационные матрицы: Σ_1 та же, $\Sigma_2 = 4\Sigma_1$ (см.

рис. 8, внизу). Это увеличивает ошибку классификаторов максимальной глубины, а также *LDA*, поскольку функциональная зависимость между глубиной и плотностью распределения различна для двух классов. *QDA*, как и ожидалось, несколько превосходит остальные классификаторы, также показывающие вполне приемлемые результаты.

Второй случай также рассматривает аналогичную разницу *а*) в положении и *б*) в положении и масштабе, но для распределений с «тяжелыми хвостами», допускающих выбросы. Для этого используется двумерное распределение Коши: *а*) $Cauchy\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1&1\\1&4\end{bmatrix}\right)$ для одного класса и $Cauchy\left(\begin{bmatrix}1\\1\end{bmatrix},\begin{bmatrix}1&1\\1&4\end{bmatrix}\right)$ для другого (см. рис. 9, вверху), и *б*) $Cauchy\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1&1\\1&4\end{bmatrix}\right)$ и $Cauchy\left(\begin{bmatrix}1\\1\end{bmatrix},\begin{bmatrix}4&4\\4&16\end{bmatrix}\right)$ соответственно (см. рис. 9, внизу). Классификаторы, использующие моментные оценки (*LDA*, *QDA*, *MM* и *DM*), показывают высокий уровень ошибки классификации на *а*), дополняясь, как и в первом случае, классификаторами максимальной глубины (*MS* и *MH*) на *б*). *DD*α-классификатор на *а*) допускает больше ошибок, чем классификаторы максимальной глубины и *DD*-классификаторы (за исключением использующих глубину Махаланобиса), из-за невысокой робастности используемой им глубины зоноида, но «исправляется» и показывает вполне приемлемую ошибку на *б*).

В третьем случае для моделирования внутриклассовой асимметрии используется *а*) двумерное экспоненциальное распределение Маршалла–Олкина: $(\min\{Z_1, Z_{12}\}, \min\{Z_2, Z_{12}\})$ для одного класса и $(\min\{Z_2, Z_{12}\}+0{,}5,\ \min\{Z_1, Z_{12}\}+0{,}5)$ для другого, где $Z_1 \sim \mathrm{Exp}(1)$, $Z_2 \sim \mathrm{Exp}(0{,}5)$ и $Z_{12} \sim \mathrm{Exp}(0{,}75)$ (рис. 10, вверху). Межклассовая асимметрия *б*) представлена нормальным $N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1&0\\0&1\end{bmatrix}\right)$ и экспоненциальным (Exp(1), Exp(1)) распределениями с независимыми маргиналами (см. рис. 10, внизу). В обоих случаях диаграммы ошибок схожи, приемлемые ошибки показывают *KNN* и *DD*-классификаторы; *DD*α-классификатор также справляется с задачей.

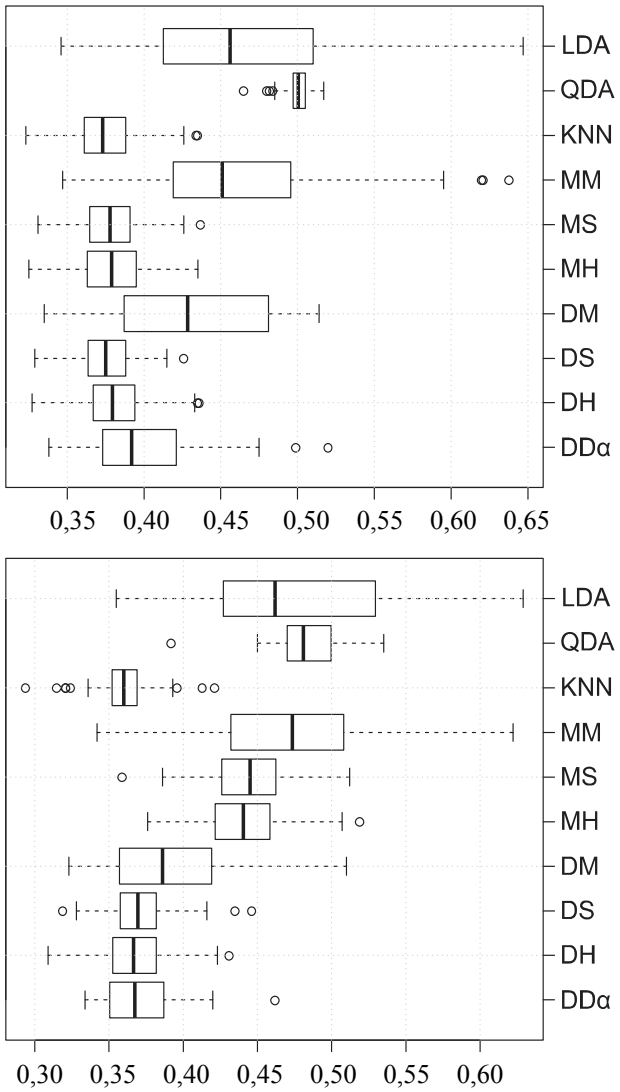

Рис. 9. Комбинированные диаграммы ошибок распознавания для распределений Коши, отличающихся положением (вверху) и положением и масштабом (внизу)

По этим результатам моделирования можно сделать вывод, что *DD*α-классификатор обладает хорошей обобщающей способностью на эллиптически распределенных данных, устойчив против выбросов и демонстрирует хороший потенциал обучения распознаванию асимметричных распределений и классов, порожденных распределениями разных семейств.

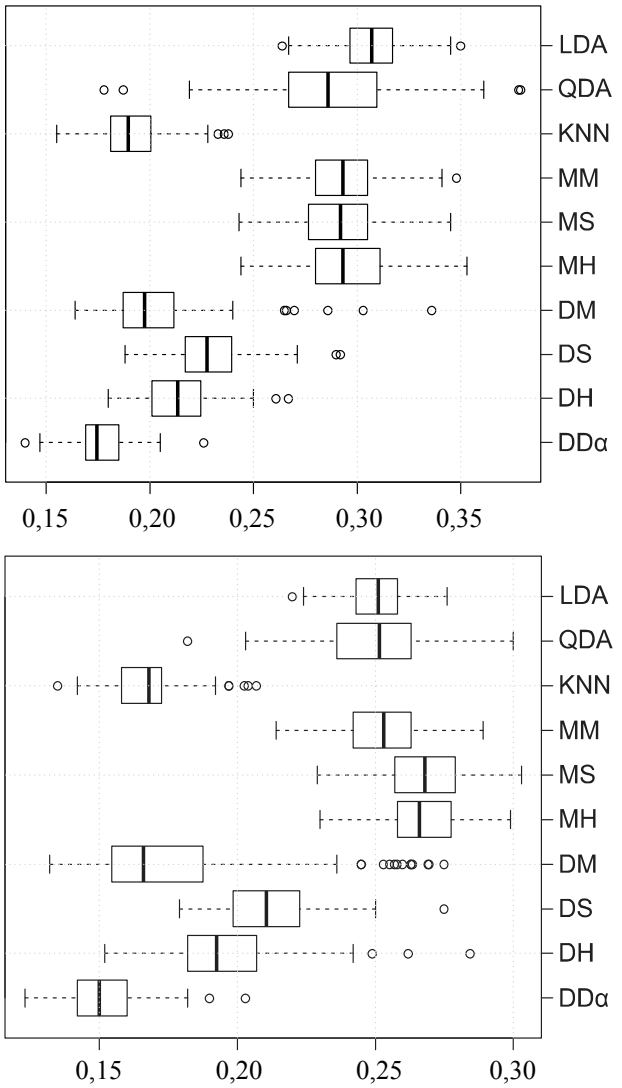

Рис. 10. Комбинированные диаграммы ошибок распознавания для внутри- (вверху) и меж- (внизу) классовой асимметрии

Эксперименты с реальными данными

Проведем сравнение обобщающей способности предложенного классификатора с использованием реальных данных.

Набор данных *Synthetic* представляет собой два класса, каждый из которых содержит двумерные наблюдения из двух нормальных распределений, отличающихся положением и масштабом; обучающая выборка состоит из 1000 точек, проверочная – из 250.

Glass насчитывает 214 образцов стекла, описанных содержанием восьми оксидов и значением показателя преломления; для сравнения классификаторов используется лишь пять наиболее значимых атрибутов и два класса, представленных большинством наблюдений (в целом 146 образцов, из них 70 – в одном классе и 76 – во втором).

Из 209 рассматриваемых в *Biomedical* образцов крови, представленных четырьмя атрибутами, внимание уделяется 194 образцам, не содержащим отсутствующих значений. Еще один набор двумерных данных *Hemophilia* содержит 75 пар логарифмированных измерений у женщин, 45 из которых – носительницы гемофилии типа *A*. *Diabetes* содержит три класса (33 + + 36 + 76 = 145) индивидуумов, страдающих разными типами диабета, охарактеризованных пятью показателями. Набор данных *BloodTransfusion* представлен тремя атрибутами (число месяцев со времени первого и последнего донорства и общее число донорств) 748 доноров крови, разделенных на два класса по признаку донорства в марте 2007 года. Более подробное описание данных можно найти в сети «Интернет» на общественном репозитории *http://archive.ics.uci.edu/ml* [29] и в пристатейном списке.

*DD*α-классификатор сравнивается со стандартными классификаторами: линейным (*LDA*) и квадратичным (*QDA*) дискриминантным анализом и классификатором ближайших соседей (*k-NN*), а также с глубинными классификаторами, основанными на проекционной (*PD*) [13] и L_p-глубине (L_pD) [12], и различными классификаторами, использующими глубину Махаланобиса (*MD*), на основе данных из [12, 13]. Применение *DD*α-классификатора, как и любого другого, использующего глубинное представление и функцию глубины, нулевую вне выпуклой оболочки, должно быть дополнено соответствующим методом классификации аутсайдеров. В качестве таких методов используются: классификатор ближайших соседей, всегда учитывающий только одно обучающее наблюдение, расположенное наиболее близко исходя из евклидова расстояния; классификатор максимальной глубины [10], основанный на быстро рассчитываемой глубине Махаланобиса, с использованием как моментных, так и робастных (*Minimum Covariance Determinant – MCD*) оценок положения и масштаба. Для обоих классификаторов максимальной глубины априорные

вероятности аппроксимируются пропорциями классов в обучающей выборке.

В [13] предложены методы робастной классификации с использованием проекционной глубины (*PD*) с одно- и многомасштабным способом юстировки классификатора (*Single-* и *Multi-scale*), проводится их сравнение с традиционными классификаторами (*LDA*, *QDA* и *k-NN*) и классификатором максимальной глубины Махаланобиса, использующим моментные (*MD*) и робастные (MD_{MCD}, коэффициент робастности $\alpha = 0{,}75$) оценки (табл. 1).

Для *Synthetic*, где обучающая и проверочная выборки заданы, приводится ошибка (в процентах) на проверочной выборке. Поскольку для *Glass* и *Biomedical* разделение на обучающую и проверочную выборки отсутствует, обучающая формируется случайным выбором 100 наблюдений (по 50 из каждого класса) из *Glass* и 150 наблюдений (100 и 50) из *Biomedical* (значения для набора данных *Biomedical* не специфицируется в [13] и выбраны авторами). Проверочная выборка формируется из оставшихся наблюдений, и такое случайное разделение проводится 500 раз. В табл. 1 приведены усредненные значения ошибок (в процентах), полученных на проверочной выборке, дополненные стандартным отклонением (в скобках внизу), также представляющим собой в данном случае важный показатель. Последние три колонки табл. 1 отображают результаты для *DD*α-классификатора с применением описанных методов классификации аутсайдеров, обозначенных 1 *NN*, *Mah.depth* (*Mom.*) и *Mah.Depth* (*MCD*) соответственно.

Из табл. 1 следует, что хотя на *Synthetic* *DD*α-классификатор несколько уступает *PD*-классификатору, для *Biomedical* он показывает вполне сравнимые результаты. Для *Glass*-результатов, сравнимых с *PD*, но с меньшими стандартными отклонениями, удается достичь при использовании расстояния Махаланобиса в 1-*NN* (ошибки и стандартные отклонения равны 25,79 (0,17) и 24,73 (0,18) для моментных и робастных оценок), в то время как *Mah.depth* в этом случае показывает значительно худшие результаты.

В табл. 2 приведено сравнение *DD*α-классификатора с классификатором, основанным на L_p-глубине [12], с использованием тех же методов классификации аутсайдеров, что и в табл. 1, и трех дополнительных наборов данных (*Hemophilia*, *Diabetes* и *BloodTransfusion*). При этом обучающая выборка формируется случайным образом и насчитывает 100 наблюдений для *Glass*, *Biomedical* и *Diabetes*, 50 для *Hemophilia*, а для *BloodTransfusion* – половину наличных данных. Оставшиеся наблюдения используются для проверки обобщающей способности, и разделение, как и выше, повторяется 500 раз. На всех шести наборах данных *DD*α-классификатор в общем уступает L_pD-классификатору, хотя его обобщающая способность во многом зависит от метода классификации аутсайдеров. При его правильном выборе ошибка, допускаемая *DD*α-классификатором, близка к таковой

Таблица 1. Сравнение на реальных данных с классификатором, основанным на проекционной глубине

Набор данных	*LDA*	*QDA*	*k-NN*	*MD*		MD_{MCD}		*PD*		1-*NN*	*DD*α	
				Single-scale	*Multi-scale*	*Single-scale*	*Multi-scale*	*Single-scale*	*Multi-scale*		*Mah. depth*	
											Mom.	*MCD*
Synthetic	10,80	10,20	11,70	13,00	11,60	10,30	10,40	10,00	10,50	12,10	11,90	12,00
Glass	30,59 (0,25)	36,13 (0,26)	22,88 (0,24)	26,59 (0,25)	26,14 (0,25)	24,92 (0,25)	24,43 (0,25)	25,70 (0,34)	25,24 (0,33)	29,45 (0,20)	30,09 (0,18)	35,06 (0,22)
Biomedical	15,66 (0,14)	12,57 (0,12)	17,88 (0,15)	12,44 (0,13)	12,04 (0,12)	14,25 (0,13)	14,03 (0,14)	12,37 (0,14)	12,18 (0,13)	13,51 (0,14)	12,91 (0,14)	15,23 (0,16)

Таблица 2. Сравнение на реальных данных с классификатором на основе L_p-глубины

Набор данных	*LDA*	*QDA*	*k-NN*	*MD*		L_pD		*DD*α		
				Mom.	*MCD*	*Mom.*	*MCD*	1-*NN*	*Mah. depth*	
									Mom.	*MCD*
Synthetic	10,80	10,20	11,70	10,20	10,60	9,60	10,70	12,10	11,90	12,00
Hemophilia	15,22 (0,27)	15,47 (0,26)	15,79 (0,30)	15,84 (0,30)	17,13 (0,32)	15,39 (0,32)	16,43 (0,32)	16,63 (0,20)	18,65 (0,22)	19,39 (0,22)
Glass	30,69 (0,25)	36,13 (0,25)	22,78 (0,24)	26,80 (0,26)	24,80 (0,29)	27,64 (0,29)	24,75 (0,26)	30,13 (0,19)	32,88 (0,22)	36,82 (0,23)
Biomedical	15,64 (0,12)	12,57 (0,12)	17,81 (0,14)	12,35 (0,14)	14,48 (0,15)	12,68 (0,15)	15,11 (0,15)	13,74 (0,09)	14,34 (0,12)	17,28 (0,14)
Diabetes	10,46 (0,18)	9,39 (0,18)	10,04 (0,18)	8,22 (0,18)	11,49 (0,22)	9,39 (0,21)	11,92 (0,27)	10,77 (0,12)	12,70 (0,18)	15,90 (0,19)
Blood Transfusion	23,08 (0,03)	22,61 (0,05)	27,69 (0,09)	22,75 (0,07)	22,17 (0,08)	22,30 (0,07)	22,06 (0,07)	23,11 (0,06)	22,59 (0,06)	22,17 (0,06)

для L_pD-классификатора (для набора данных *Glass* ее можно значительно уменьшить, используя расстояние Махаланобиса в 1-*NN*). Так, для *Hemophilia*, например, рекомендуется 1-*NN*, когда для *BloodTransfusion* все предложенные методы показывают неплохие результаты.

Заключение. Предложенный новый метод обучения распознаванию образов отличается полной непараметричностью. *DD*α-классификатор, используя глубину зоноида, отображает данные в θ-мерное глубинное пространство, в полиномиальном расширении которого классы обучающей выборки разделяются линейным решающим правилом при помощи α-процедуры. Благодаря этому сочетанию *DD*α-классификатор обладает высокой скоростью как обучения, так и классификации, превосходя существующие непараметрические классификаторы. Он также показывает вполне конкурентоспособную ошибку в задачах обучения распознаванию образов с эллиптически распределенными классами, проявляя при этом устойчивость к выбросам в данных. Эксперименты показывают применимость предлагаемого метода к асимметричным данным, а также при наличии классов, порожденных разными семействами распределений.

В отличие от большинства предлагаемых в литературе процедур, *DD*α-классификатор рассматривает полное пространство глубин (относительно всех рассматриваемых в задаче классов), выполняя бинарное разделение. Метод не нуждается в юстировке отдельных определенных параметров модели. *DD*α-классификатор легко может быть применим в комбинации с другими методами классификации, а также для конструирования новых нотаций глубины, включая функциональную.

Теоретические результаты показывают, что метод сходится к Байесовскому классификатору в случаях, когда два класса являются зеркальным отражением друг друга, и для эллиптически распределенных классов, отличающихся лишь положением. Применяемая в методе глубина зоноида обладает рядом преимуществ как в теоретическом, так и в вычислительном смысле: в первую очередь она может быть эффективно рассчитана для многомерных данных, даже в размерностях порядка 15–20 и более. Хотя тот факт, что ее максимум достигается в математическом ожидании, и влечет потерю робастности для глубины, *DD*α-классификатор справляется с загрязненными выбросами данными. Это можно объяснить тем, что обучающая выборка сперва отображается в компактное пространство – единичный гиперкуб, и лишь потом разделяется α-процедурой, которая, оперируя координатными осями, отличается высокой степенью робастности.

Перспективным для будущих исследований есть более детальное теоретическое изучение свойств *DD*α-классификатора, его обобщающей способности на асимметричных распределениях и распределениях с «тяжелыми хвостами», а также выбор, возможно адаптивный, более эффективных методов классификации аутсайдеров и нотаций глубины.

1. *Ивахненко А.Г.* Самообучающиеся системы распознавания и автоматического управления. – Киев: Техніка, 1969. – 392 с.
2. *Ивахненко А.Г.* Системы эвристической самоорганизации в технической кибернетике. – Там же, 1971. – 372 с.
3. *Персептрон* – система распознавания образов / Под ред. А.Г. Ивахненко. – Киев: Наук. думка, 1975. – 432 с.
4. *Self-organizing* methods in modeling: GMDH type algorithms / Ed. S.J. Farlow. – New York, Basel: Marcel Decker Inc., 1984. – 350 p.
5. *Ивахненко А.Г.*, *Степашко В.С.* Помехоустойчивость моделирования. – Киев: Наук. думка, 1985. – 216 с.
6. *Tukey J.W.* Mathematics and the picturing of data // Proc. of the Int. Congr. of Mathematicians / Ed. R.D. James. – **2**. – Montreal: Canadian Mathematical Congress, 1975. – P. 523–531.
7. *Liu R.Y.* On a notion of data depth based on random simplices // Annals of Statistics. – 1990. – **18**. – P. 405–414.
8. *Liu R.Y.*, *Parelius J.*, *Singh K.* Multivariate analysis of the data depth: descriptive statistics and inference // Annals of Statistics – 1999. – **27**. – P. 783–858.
9. *Jornsten R.* Clustering and classification based on the L1 data depth // J. of Multivariate Analysis. – 2004. – **90**. – P. 67–89.
10. *Ghosh A.K.*, *Chaudhuri P.* On maximum depth and related classifiers // Scandinavian J. of Statistics. – 2005. – **32**. – P. 327–350.

11. *Mosler K.*, *Hoberg R.* Data analysis and classification with the zonoid depth // Data depth: robust multivariate analysis, computational geometry and applications / Eds. R. Liu, R. Serfling, D. Souvaine. – New York: American Mathematical Society, 2006. – P. 49–59.
12. *Dutta S.*, *Ghosh A.K.* On classification based on L_p depth with an adaptive choice of p / Technical Report Number R5. – Indian Statistical Institute. – Kolkata: Statistics and Mathematics Unit, 2011. – 16 p.
13. *Dutta S.*, *Ghosh A.K.* On robust classification using projection depth // Annals of the Institute of Statistical Mathematics. – 2012. – **64**. – P. 657–676.
14. *Li J.*, *Cuesta-Albertos J.A.*, *Liu R.Y.* DD-classifier: nonparametric classification procedure based on DD-plot // J. of the American Statistical Association. – 2012. – **107**. – P. 737–753.
15. *Mahalanobis P.* On the generalized distance in statistics // Proc. of the National Institute of Sciences of India. – 1936. – **2**(1). – P. 49–55.
16. *Koshevoy G.*, *Mosler K.* Zonoid trimming for multivariate distributions // Annals of Statistics. – 1997. – **25**. – P. 1998–2017.
17. *Mosler K.* Multivariate dispersion, central regions and depth: the lift zonoid approach. – New York: Springer, 2002. – 291 p.
18. *Vasil'ev V.I.* The reduction principle in pattern recognition learning (PRL) problem // Pat. Recog. and Image Analysis. – 1991. – N 1. – P. 23–52.
19. *Васильев В.И.* Принцип редукции в задачах обнаружения закономерностей. I // Кибернетика и системный анализ. – 2003. – № 5. – С. 69–81.
20. *Васильев В.И.*, *Ланге Т.И.* Принцип дуальности в проблемах обучения распознавания образов // Кибернетика и вычислительная техника. – 1998. – **121**. – С. 7–17.
21. *Zuo Y.J.*, *Serfling R.* General notions of statistical depth function // Annals of Statistics. – 2000. – **28**. – P. 461–482.
22. *Dyckerhoff R.* Data depths satisfying the projection property // Advances in Statistical Analysis. – 2004. – **88**. – P. 163–190.
23. *Serfling R.* Depth functions in nonparametric multivariate inference // Data depth: robust multivariate analysis, computational geometry and applications / Eds. R. Liu, R. Serfling, D. Souvaine. – New York: American Mathematical Society, 2006. – P. 1–16.
24. *Cascos I.* Data depth: Multivariate statistics and geometry // New perspectives in stochastic geometry / Eds. W. Kendall, I. Molchanov. – Oxford: Oxford University Press, 2010. – P. 398–423.
25. *Hubert M.*, *Van Driessen K.* Fast and robust discriminant analysis // Computational Statistics and Data Analysis. – 2004. – **45**. – P. 301–320.
26. *Lange T.*, *Mosler K.*, *Mozharovskyi P.* Fast nonparametric classification based on data depth. – www.wisostat.ini–Koeln.de/Forschung/Papers/DPO112.pdf
27. *Васильев В.И.* Теория редукции в проблемах экстраполяции // Проблемы управления и информатики. – 1996. – № 1–2. – С. 239–251.
28. *Dyckerhoff R.*, *Koshevoy G.*, *Mosler K.* Zonoid data depth: theory and computation // Proc. in computationnal statistics / Ed. A. Prat. – Heidelberg: Physica-Verlag, 1996. – P. 235–240.
29. *Asuncion A.*, *Newman D.* Machine Learning Repository / Univ. of Calif., Center for Machine Learning and Intell. Syst., Irvine, 2007. – Access mode. – http://archive.ics.uci.edu/ml/ (18.12.2012). – Machine Learning Repository.

E-mail: *tatjana.lange@hs-merseburg.de*, *mosler@statistik.uni-koeln.de*, *mozharovskyi@statistik.uni-koeln.de*