УДК 519.7

А.А. ГАЛКИН

ГЛУБИННЫЙ МЕТОД КЛАССИФИКАЦИИ НА ОСНОВЕ УДАЛЕННОЙ МЕРЫ КОНЦЕНТРАЦИИ ДЛЯ ОБРАБОТКИ АСИММЕТРИЧНЫХ ДАННЫХ

Аннотация. Разработан и исследован глубинный метод классификации на основе удаленной меры концентрации для обработки асимметричных данных. Мотивацией построения метода стала неэффективность использования большинства аффинно-инвариантных классификаторов при их сочетании с функциями глубины, которые обращаются в нуль за пределами выпуклой оболочки данных. Идея предложенного метода заключается в отображении дистанционного пространства с использованием удаленной меры концентрации, меры удаленности Штахеля–Донохью и меры скорректированной удаленности.

Ключевые слова: функция глубины, удаленная мера концентрации, многомерная классификация.

ВВЕДЕНИЕ

Проблема потенциальных последствий выбросов и экстремальных значений при решении современных задач распознавания требует поиска новых устойчивых к выбросам непараметрических методов. В большинстве случаев выбросы являются допустимыми элементами, которые поступают из различных множеств данных. В задачах классификации с учителем метки классов некоторых элементов данных в учебном множестве могут присваиваться ошибочно. Большинство методов классификации являются эффективными только при применении к данным с эллиптической симметрией или с многомерным нормальным распределением. Большинство существующих методов, которые дают возможность классифицировать асимметричные многомерные данные, реализованы на основе функций глубины. Однако такие классификаторы часто имеют достаточно низкую производительность, поскольку функции глубины обращаются в нуль за пределами выпуклой оболочки данных.

Учитывая актуальность исследуемой проблематики, данная статья посвящается разработке и исследованию нового непараметрического метода классификации, который обеспечивает возможность обработки асимметричных многомерных данных. Предложенный метод относится к классу обучения с учителем и основывается на концепции дистанционного пространства.

ОПРЕДЕЛЕНИЕ ГЛУБИННЫХ ОБЛАСТЕЙ НА КОНЕЧНЫХ ВЫБОРКАХ

Исходя из требований статистической функции глубины, функция полупространственной глубины является монотонно убывающей вдоль линий, исходящих из центра, а также является аффинно-инвариантной. Кроме того, функция полупространственной глубины равна нулю на бесконечности и достигает своего максимального значения в центре симметрии [1].

Функция полупространственной глубины $\forall\, z \in \mathrm{R}^p$ относительно H_X определяется как минимальная вероятностная группа, содержащаяся в закрытом полупространстве с пределом по z, а именно

$$F_d\,(z, H_X) = \inf_{\|b\|=1} H_X\,\{b'X \geq b'z\},$$

где X — случайная величина на R^p с распределением H_X.

Областью φ-глубины E_φ является множество точек, глубина которых составляет не менее φ, т.е.

$$E_\varphi = \{z \in \mathrm{R}^p\}$$

для $\forall \varphi \in [0,1]$ и $E(z, H_X) \geq \varphi$. Отметим, что профилем φ-глубины является предел E_φ.

Полупространственное средневзвешенное значение определяется как центр веса наименьшей области, содержащей точки с максимальной полупространственной глубиной (непустая область глубины). Заметим, что для возрастающего φ области полупространственной глубины являются выпуклыми, вложенными и замкнутыми. Кроме того, определение глубинных областей на конечных выборках, а также полупространственного средневзвешенного значения можно получить путем замены H_X эмпирическим вероятностным распределением H_m.

В настоящей статье используем точечную диаграмму, которая обобщает одномерную диаграмму разброса для двумерных данных [2]. Концентрация данных первого типа является наименьшей глубинной областью, имеющей не менее 50% вероятностной группы, т.е. $V = E_{\overline{\varphi}}$ такое, что $H_X(V) \geq 0.5$ и $H_X(E_\varphi) < 0.5$ для всех $\varphi > \overline{\varphi}$. Отметим, что внутри концентрации находится полупространственное средневзвешенное значение. Зигзагообразное частично упорядоченное множество, которое не является самообращающимся, можно получить расширением диаграммы на коэффициент 3 относительно средневзвешенного значения, при этом точки данных за ее пределами обозначаются как выбросы. Петля, образованная из данных второго типа, является выпуклой оболочкой точек внутри зигзагообразного частично упорядоченного множества.

Использование диаграммы концентрации данных мотивировано тем, что она не зависит от предполагаемой симметрии. Поэтому она одинаково эффективна для симметричных и асимметричных данных. Заметим, что средневзвешенное значение не обязательно должно быть расположенным внутри концентрации данных, а сама концентрация данных не обязательно должна быть эллиптической формы.

УДАЛЕННАЯ МЕРА КОНЦЕНТРАЦИИ ДЛЯ ОБРАБОТКИ АСИММЕТРИЧНЫХ ДАННЫХ

Идея предложенного подхода состоит в вычислении статистического расстояния многомерной точки $z \in \mathrm{R}^p$ в направлении H_X на основе функции полупространственной глубины. Для вычисления данного расстояния используются центр и дисперсия H_X; для оценки дисперсии применяется концентрация V.

Следует отметить, что $g(z) = g_z$ определяется как пересечение границы V и линии от полупространственного средневзвешенного значения ω через z. В результате удаленная мера концентрации z в X определяется соотношением эвклидовой метрики z в направлении полупространственного средневзвешенного значения и эвклидовой метрики g_z к полупространственному средневзвешенному значению, а именно $\Delta(z, H_X) = 0$, если $z = \omega$, и $\Delta(z, H_X) = ||z - \omega|| / ||g_z - \omega||$ в противном случае. Заметим, что знаменатель в уравнении $\Delta(z, H_X) = ||z - \omega|| / ||g_z - \omega||$ определяет дисперсию H_X в направлении z. В данном случае удаленная мера концентрации является аффинно-инвариантной, но не предполагает наличия симметрии [3]. Отметим также, что удаленная мера концентрации неявно используется в диаграмме концентрации данных, а зигзагообразное частично упорядоченное множество состоит из точек, удаленная мера концентрации которых не более трех.

Далее определим обобщенную норму как функцию $c: \mathrm{R}^p \to [0, \infty[$ такую, что $c(0) = 0$ и $c(z) \neq 0$ для $z \neq 0$; она удовлетворяет $c(\beta z) = \beta c(z)$ для $\forall z$ и $\forall \beta > 0$.

В частности, имеет место обобщенная норма

$$c(z) = \sqrt{z' \Xi^{-1} z}$$

для гауссовского распределения $N(0, \Xi)$ с положительно-определенной величиной Ξ.

Предположим, что компактное множество V имеет форму звезды в окрестности нуля, т.е. для $\forall z \in V$ и $0 \le \beta \le 1$ имеем $\beta z \in V$. Для $\forall z \ne 0$ строим точку g_z, что является сечением между линией, которая выходит из нуля в направлении z, и границей к V.

Также предположим, что нуль находится внутри V, т.е. существует такое $\mu > 0$, для которого выполняется включение $V(0, \mu) \subset V$. Тогда $||g_z|| > 0$ при $z \ne 0$. Далее имеем $c(z) = 0$, если $z = 0$, и $c(z) = ||z|| / ||g_z||$ в противном случае. Поскольку $\beta > 0$ такое, что $\beta^{-1} z$ расположено на границе V, можем определить $c(z)$. Таким образом, нет необходимости в получении евклидовой нормы. Кроме того, можно проверить, является ли $c(\cdot)$ обобщенной нормой, которая может не быть непрерывной функцией [4].

Лемма 1. Функция c является выпуклой и непрерывной, если $0 \in \text{int}(V)$, а множество V является компактным и выпуклым.

Доказательство. Необходимо показать, что для $\forall z, x \in \mathrm{R}^p$ и $0 \le \theta \le 1$ имеет место неравенство $c(\theta z + (1-\theta)x) \le \theta c(z) + (1-\theta)c(x)$. Отметим, что функция c, которая ограничена этой прямой, является выпуклой, поскольку линейно возрастает в обоих направлениях под разными углами и равна нулю в начале координат. Заметим, что выпуклость функции c имеет место в случае, когда векторы $\{0, z, x\}$ являются коллинеарными. В противном случае $\{0, z, x\}$ образуют треугольник.

Введем следующие обозначения: $y := \theta z + (1-\theta)x$, $x = c(x) g_x$, $z = c(z) g_z$. Можно проверить, что выпуклой комбинацией g_z и g_x является выражение

$$\bar{y} := (\theta c(z) + (1-\theta) c(x))^{-1} y.$$

Поскольку $g_z, g_x \in V$, что следует из компактности V, тогда $\bar{y} \in V$, что следует из выпуклости V. Принимая во внимание, что $||g_y|| = ||g_{\bar{y}}|| \ge ||\bar{y}||$, имеет место равенство

$$c(y) = \frac{||y||}{||g_y||} \le \frac{||y||}{||\bar{y}||} = \theta c(z) + (1-\theta) c(x).$$

Лемма доказана.

Используя лемму 1, а также неравенство

$$c(z+x) = 2c\left(\frac{1}{2} z + \frac{1}{2} x\right) \le 2\frac{1}{2} c(z) + 2\frac{1}{2} c(x) = c(z) + c(x),$$

можно утверждать, что функция c удовлетворяет неравенству треугольника. Отсюда следует, что функция c и удаленная мера концентрации удовлетворяют следующим условиям:

а) функция $c(z) \ge 0 \ \forall z \in \mathrm{R}^p$;

б) функция $c(z) = 0 \Rightarrow z = 0$;

в) функция $c(\beta z) = \beta c(z)$, $\beta \ge 0$ и $\forall z \in \mathrm{R}^p$;

г) функция $c(z+x) \le c(z) + c(x) \ \forall z, x \in \mathrm{R}^p$.

Заметим, что получение нормы возможно при добавлении $c(-z) = c(z)$ для $\forall z \in \mathrm{R}^p$. Кроме того, обобщая полученный результат, можно проводить асимметричную дисперсию для удаленной меры концентрации. Также вычислив $f(z) = (c(z) + c(-z))/2$, можно получить норму этой функции.

В результате имеет место равенство

$$c(z) = ||z|| / ((z' \Xi^{-1} z)^{-1/2} ||z||) = \sqrt{z' \Xi^{-1} z},$$

где $V = \{z; z' \Xi^{-1} z \le 1\}$, а $g_z = (z' \Xi^{-1} z)^{-1/2} z$ для $\forall z \ne 0$. Таким образом, можно утверждать, что функция c обобщает расстояние Махаланобиса в обобщенной норме: $c(z) = \sqrt{z' \Xi^{-1} z}$.

Заметим, что лемма 1 имеет место, когда V является выпуклым множеством. В данном случае использование функции экстраполяционной глубины является альтернативой глубинным областям Тьюки [5]. В одномерном случае компактное выпуклое множество V в лемме 1 становится замкнутым интервалом, который можно определить как $V = \left[-\frac{1}{v}, \frac{1}{w}\right]$ при $v, w > 0$; $c(z) = wz^+ + vz^-$. В линейной регрессии при минимизации $\sum_i^m c(d_i)$ имеем регрессионный квантиль $w/(w+v)$.

При решении практических задач классификации, когда удаленную меру концентрации нужно вычислить для множества точек, необходимо сначала вычислять концентрацию данных, а затем точку пересечения g_z. Заметим, что вычисление удаленной меры концентрации точки z относительно случайной выборки возможно при использовании данных малой размерности.

В некоторых случаях, в частности при использовании данных большой размерности, вычисление удаленной меры концентрации требует значительно больших ресурсов [6]. Учитывая тот факт, что функция полупространственной глубины является монотонно убывающей на прямой, применение алгоритма бисекции является эффективным инструментом поиска многомерной точки g^* на прямой от ω через z, где $F_d(g^*, H_m) = \underset{i}{\Omega}\{F_d(x_i, H_m)\}$. В данном случае x_i являются точками данных, а Ω — средневзвешенным значением.

ПРОЕКТИРОВАНИЕ ДАННЫХ НА ОСНОВЕ ОДНОМЕРНОЙ МЕРЫ ОТДАЛЕННОСТИ

Аффинно-инвариантная функция экстраполяционной глубины является обратной по отношению к функции отдаленности Штахеля–Донохью. Согласно геометрическому толкованию этой функции многомерный выброс должен быть отдаленным как минимум в одном направлении. Предложенный подход заключается в проектировании данных на множество прямых с использованием одномерной меры отдаленности на проекциях.

Множество данных функции отдаленности Штахеля–Донохью произвольной точки z относительно случайной величины X с распределением H_X определяется как

$$D_{SH}(z, H_X) = \sup_{\|b\|=1} \frac{|b'z - \Omega(b'X)|}{\Sigma(b'X)},$$

где Σ — среднее абсолютное отклонение. Отсюда имеем функцию экстраполя-

ционной глубины

$$F_e(z, H_X) = \frac{1}{1 + D_{SH}(z, H_X)}.$$

Заметим, что функция отдаленности Штахеля–Донохью является более подходящей для симметричных распределений, поскольку имеет среднее абсолютное отклонение в знаменателе и абсолютное отклонение в числителе.

В случае асимметричных распределений эффективным инструментом является функция скорректированной отдаленности на основе метода независимых компонент [7]. В качестве надежной меры асимметрии функция скорректированной отдаленности использует M-статистику одномерного множества данных $Y = \{y_1, \dots, y_m\}$, определяемую как

$$\mathrm{M}(y_1, \dots, y_m) = \underset{i,l}{\Omega} \frac{(y_l - \Omega_t y_t) - (\Omega_t y_t - y_i)}{y_l - y_i},$$

где $-1 \le \mathrm{M} \le 1$, а i и l удовлетворяют таким условиям: $y_i \le \Omega_t(y_t) \le y_l$ и $y_i \ne y_l$. Заметим, что $\mathrm{M} < 0$ и $\mathrm{M} > 0$ означают левую и правую асимметрию соответственно, а при $\mathrm{M} = 0$ имеем случай симметричных распределений.

Далее введем понятие меры скорректированной отдаленности O:

$$\mathrm{O}(z, H_X) = \sup_{\|b\|=1} \mathrm{O}_1(b'z, H_{b'X}),$$

где O_1 — мера одномерной скорректированной отдаленности,

$$\mathrm{O}_1(y, Y) = \frac{y - \Omega(Y)}{a_2(Y) - \Omega(Y)}, \text{ если } y > \Omega(Y)$$

и

$$\mathrm{O}_1(y, Y) = \frac{\Omega(Y) - y}{\Omega(Y) - a_1(y)}, \text{ если } y \le \Omega(Y).$$

Заметим, что знаменатель в выражениях $\mathrm{O}_1(y, Y)$ соответствует зигзагообразному частично упорядоченному множеству одномерной скорректированной диаграммы концентрации данных. Кроме того, имеют место выражения $a_1(Y) = W_1(Y) - 1.5e^{-4\mathrm{M}(Y)}Q(Y)$ и $a_2(Y) = W_3(Y) + 1.5e^{+3\mathrm{M}(Y)}Q(Y)$, где Q является межквартильным диапазоном. В случае, если $\mathrm{M}(Y) < 0$, заменяем (y, Y) на $(-y, -Y)$. Если $\mathrm{M}(Y) \ge 0$, имеем $Q(Y) = W_3(Y) - W_1(Y)$, где $W_1(Y)$ и $W_3(Y)$ определяют первую и третью квартили Y.

Определим функцию асимметрично скорректированной экстраполяционной глубины

$$\widetilde{F}_e(z, H_X) = \frac{1}{1 + \mathrm{O}(z, H_X)}.$$

Ввиду невозможности использования всех направлений b применение приближенных алгоритмов является эффективным инструментом для вычисления функции конечно-выборочной асимметрично скорректированной экстраполяционной глубины [8]. В результате при рассмотрении направлений b, которые являются ортогональными к аффинной гиперплоскости через $p+1$ случайную точку данных, был получен комплексный аффинно-инвариантный подход.

Метод k-ближайших соседей является одним из наиболее эффективных непараметрических классификаторов, который для каждого нового элемента находит k точек данных, ближайших к нему, и присваивает его к преобладаю-

щей группе среди этих соседей. Наиболее часто для минимизации коэффициента ошибочной классификации используется метод перекрестной проверки для выбора значения k.

Подход на основе функции максимальной глубины может быть применен к двум и более группам и позволяет присваивать новый элемент к группе, в которой он имеет наибольшую глубину. Однако когда функция глубины тождественно равна нулю на больших интервалах, имеет место наличие множества узлов, что является недостатком такого подхода. Заметим, что использование функции экстраполяционной глубины позволило решить данную проблему.

Для развития метода классификации на основе функции максимальной глубины автором предложен и исследован новый Σ-классификатор. Итак, пусть H_1 и H_2 — эмпирические распределения двух групп данных. Используя статистическую функцию глубины d_s, выполняем отображение произвольной точки данных до двумерной точки $(d_s(z, H_1), d_s(z, H_2))$, где $z \in \mathrm{R}^{\mathrm{p}}$. Полученные двумерные точки образуют Σ-схему, в которой две группы элементов данных имеют различные метки, на основе этой схемы проводится классификация данных.

Метод классификации на основе функции максимальной глубины базируется на концепции разделения данных относительно прямой, проходящей через начало координат. Если элемент данных располагается выше многочлена, то он относится к первой группе, в противном случае — ко второй группе. Отметим, что недостатками Σ-классификатора является необходимость применения метода мажоритарного голосования при наличии более двух групп данных, а также вычислительная сложность нахождения наиболее эффективного разделительного многочлена.

Ввиду эффективности аффинно-инвариантности при решении многоклассовых задач классификации суть данного подхода заключается в синтезе функции скорректированной отдаленности и удаленной меры концентрации. Эти функции, которые являются стойкими к выбросам и экстремальным значениям, можно использовать для асимметричных данных.

Предположим, что H_c является эмпирическим распределением данных из группы $c = 1, \ldots, C$, где C может быть больше двух. Если величина $\widetilde{d}(z, H_c)$ является обобщенным расстоянием или мерой отдаленности точки z в направлении c-й выборки данных, отображаем точку $z \in \mathrm{R}^p$ в направлении C-мерной точки $(\widetilde{d}(z, H_1), \ldots, \widetilde{d}(z, H_C))$ вместо трансформации глубины $(d_s(z, H_1), d_s(z, H_2))$. В данном случае размерность C может быть меньше, больше или равной исходной размерности p.

В результате для отображения расстояния $(\widetilde{d}(z, H_1), \ldots, \widetilde{d}(z, H_C))$ можно применять произвольный многомерный классификатор, т.е. линейный или квадратичный дискриминантный анализ, метод классификации на основе минимального расстояния и т.д. [10]. Заметим, что последний метод присваивает только элемент z к группе с наименьшими координатами в $(\widetilde{d}(z, H_1), \ldots, \widetilde{d}(z, H_C))$.

С учетом неэффективности применения метода мажоритарного голосования для всех отображенных точек используем метод k-ближайших соседей в сочетании с аффинно-инвариантностью, которая получена благодаря отображению данных. В результате экспериментальных исследований было установлено, что рассмотренный метод на основе удаленной пространственной меры имеет достаточно низкую частоту ошибок. Соответствующие результаты были получены в процессе отображения расстояния $(\widetilde{d}(z, H_1), \ldots, \widetilde{d}(z, H_C))$ с использованием метода k-ближайших соседей.

РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТАЛЬНОГО ИССЛЕДОВАНИЯ

Процедура отображения дистанционного пространства была исследована на реальном примере. Использованы данные банковских клиентов в специализированных компьютерных системах. Процедура волнового преобразования применялась к данным по отношению к 1114 платежеспособным и 954 неплатежеспособным клиентам.

На рис. 1 отображен график дистанционного пространства этих данных на основе удаленной меры концентрации. Можно увидеть, что данные о платежеспособных клиентах Σ_d образуют плотный сектор по сравнению с данными о неплатежеспособных клиентах $\hat{\Sigma}_d$. Заметим, что данные о платежеспособных и неплатежеспособных клиентах являются эффективно разделенными.

На рис. 2 показана полупространственная Σ-схема исследуемых данных, где полупространственная глубина данных из одной группы относительно данных другой группы равна нулю. Заметим, что такая процедура не является эффективной для классификации данных, расположенных за пределами обеих оболочек, несмотря на то, что выпуклые оболочки обеих групп не пересекаются.

Согласно проведенным экспериментальным исследованиям были использованы учебная и тестовая выборки данных, которые содержали соответственно 100 и 1000 элементов для каждого запуска алгоритма. Поскольку M является общим размером учебной выборки, m_c — числом элементов данных группы c в учебном множестве, а e_c — процент ошибочно классифицированных элементов данных группы c в тестовом множестве, коэффициент ошибочной классификации (в процентах) вычисляется как $\sum_{c=1}^{C} e_c m_c / M$. Это позволило оценить эффективность каждого исследуемого классификатора. Повторялась данная операция 1500 раз для каждого случая и было проведено взвешивание коэффициентов (в процентах) ошибочной классификации относительно априорных вероятностей в тестовом множестве.

Рассмотрим два случая. Первый охватывает двумерную нормальную асимметрию, когда $C = 2$. Отметим, что при рассмотрении двух двумерных распределений первая группа G_1 была сгенерирована из стандартного нормального рас-

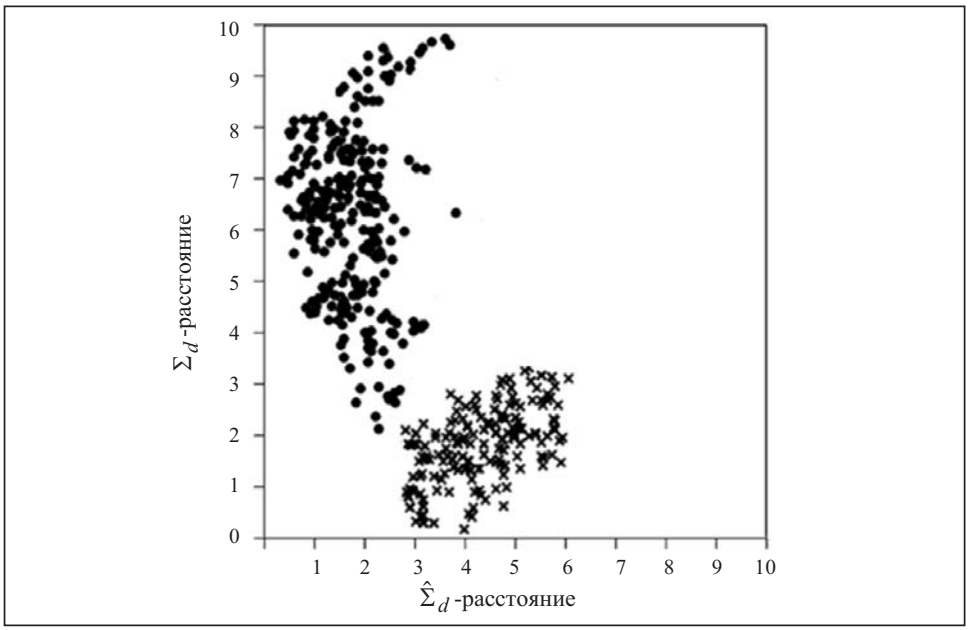

Рис. 1. Σ-схема банковских данных (на основе удаленной меры концентрации)

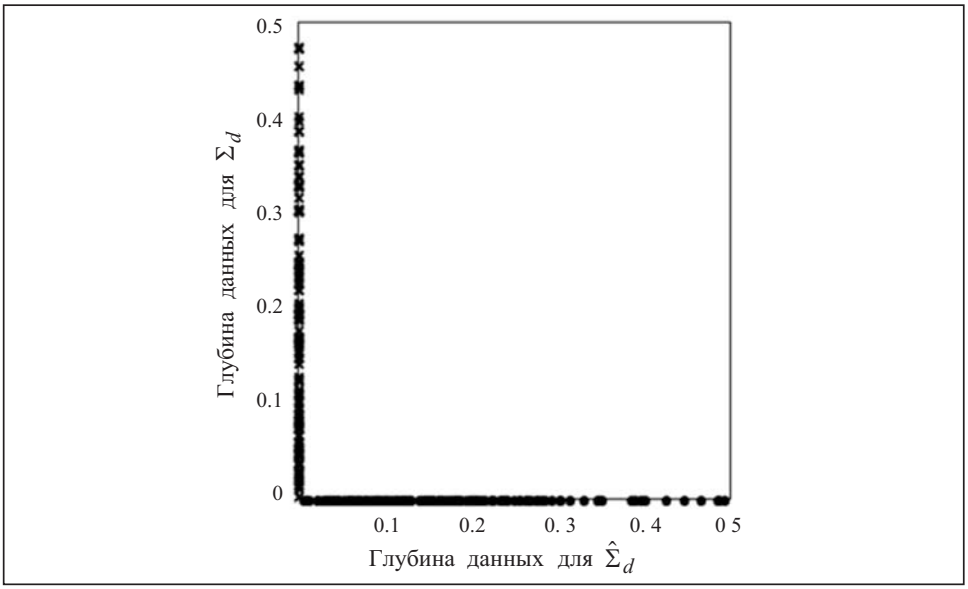

Рис. 2. Σ-схема банковских данных (на основе полупространственной глубины)

пределения, в то время как координаты второй группы являлись независимыми с экспоненциальным распределением и коэффициентом единица, т.е.

$$G_1 \approx N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right),\ G_2 \approx \begin{bmatrix}\mathrm{Exp}(1)\\ \mathrm{Exp}(1)\end{bmatrix}.$$

Во втором случае была проведена нормализация данных, при которой использовалась форма списка и средневзвешенное абсолютное отклонение. Заметим, что в данном случае $C = 2$.

В процессе исследования глубинных классификаторов был проведен сравнительный анализ классификатора на основе функции полупространственной глубины, классификатора на основе функции экстраполяционной глубины и классификатора на основе функции асимметрично скорректированной экстраполяционной глубины. Кроме того, был проведен сравнительный анализ классификатора, основанного на удаленной мере концентрации, классификатора на основе функции отдаленности Штахеля–Донохью и классификатора на основе функции скорректированной отдаленности. Отметим, что учебная выборка являлась случайным образом сгенерированным подмножеством, состоящим из 868 элементов данных, а тестовая выборка состояла из оставшихся 1200 элементов данных.

В случае двумерной нормальной асимметрии результаты экспериментальных исследований свидетельствуют о том, что Σ-классификатор явно превосходит классификатор максимальной глубины, а классификатор на основе дистанционного пространства аналогично превосходит классификатор на основе функции минимального расстояния. Учитывая, что двумерная нормальная асимметрия содержит асимметричную группу, классификаторы на основе удаленной меры концентрации и скорректированной отдаленности имеют преимущество перед классификатором на основе функции отдаленности Штахеля–Донохью, которая предполагает симметрию данных.

Следует отметить, что в процессе применения Σ-классификатора для функции асимметрично скорректированной экстраполяционной глубины, а также классификатора на основе дистанционного пространства были получены наиболее низкие коэффициенты ошибочной классификации.

Анализируя результаты для второго случая, отметим, что низкий процент ошибочной классификации получен Σ-классификатором с использованием функции экстраполяционной глубины и функции асимметрично скорректированной экстраполяционной глубины, а также классификатором на основе дистанционного пространства.

ЗАКЛЮЧЕНИЕ

В случае, когда дисперсия данных обусловлена направлением, в котором она измеряется, использование большинства непараметрических методов распознавания может иметь низкую эффективность при работе с многомерными данными. Эта проблема может быть решена с помощью классификатора максимальной глубины, а также Σ-классификатора благодаря их аффинно-инвариантности. Однако эти классификаторы демонстрируют низкую производительность при сочетании с функциями глубины, которые обращаются в нуль за пределами выпуклой оболочки данных.

Учитывая актуальность указанной проблематики, был предложен глубинный метод классификации с использованием удаленной меры концентрации данных. Были исследованы свойства этого метода, позволяющие отображать асимметрию данных. Суть предложенного метода заключается в отображении дистанционного пространства с использованием удаленной меры концентрации данных, отдаленности Штахеля–Донохью и скорректированной отдаленности.

В результате проведенного исследования было установлено, что для классификации модифицированных данных после применения Σ-классификатора разделительный полиномиальный метод имеет низкую производительность вследствие необходимости применения метода мажоритарного голосования при наличии более двух групп, а также занимает много времени ввиду выбора многочлена. Установлено, что наиболее высокую производительность продемонстрировали Σ-классификатор и классификатор на основе дистанционного пространства с использованием метода k-ближайших соседей, который применяется к модифицированным данным. Предложенный аффинно-инвариантный метод классификации может быть эффективно применен к многомерным данным и является надежным инструментом для решения многих практических задач распознавания.

СПИСОК ЛИТЕРАТУРЫ

1. Kong L., Zuo Y. Smooth depth contours characterize the underlying distribution // Journal of Multivariate Analysis. — 2010. — **101**, N 9. — P. 2223–2225.
2. Liu R. On a notion of data depth based on random simplices // The Annals of Statistics. — 1990. — **18**, N 1. — P. 406–412.
3. Pigoli D., Sangalli L. Wavelets in functional data analysis: estimation of multidimensional curves and their derivatives // Computational Statistics and Data Analysis. — 2012. — **56**, N 6. — P. 1483–1497.
4. Zuo Y., Serfling R. Structural properties and convergence results for contours of sample statistical depth functions // The Annals of Statistics. — 2000. — **28**, N 2. — P. 484–497.
5. Lange T., Mosler K., Mozharovskyi P. Fast nonparametric classification based on data depth // Statist. Papers. — 2014. — **55.** — P. 53–67.
6. Oja H., Paindaveine D. Optimal signed-rank tests based on hyperplanes // Journal of Statistical Planning and Inference. — 2005. — **135.** — P. 307–321.
7. Romanazzi M. Influence function of halfspace depth // Journal of Multivariate Analysis. — 2001. — **77**. — P. 140–159.

8. Rousseeuw P.J, Struyf A. Characterizing angular symmetry and regression symmetry // Journal of Statistical Planning and Inference. — 2004. — **122.** — P. 163–171.
9. Struyf A., Rousseeuw P.J High-dimensional computation of the deepest location // Computational Statistics and Data Analysis. — 2000. — **34**, N 4. — P. 419–425.
10. Mizera I., Volauf M. Continuity of halfspace depth contours and maximum depth estimators: diagnostics of depth-related methods // Journal of Multivariate Analysis. — 2002. — **83**, N 2. — P. 367–386.

Надійшла до редакції 30.11.2015

О.А. Галкін

ГЛИБИННИЙ МЕТОД КЛАСИФІКАЦІЇ НА ОСНОВІ ВІДДАЛЕНОЇ МІРИ КОНЦЕНТРАЦІЇ ДЛЯ ОБРОБКИ АСИМЕТРИЧНИХ ДАНИХ

Анотація. Розроблено та досліджено глибинний метод класифікації на основі віддаленої міри концентрації для обробки асиметричних даних. Мотивацією побудови методу стала неефективність використання більшості афінно-інваріантних класифікаторів при їх поєднанні з функціями глибини, які перетворюються в нуль за межами опуклої оболонки даних. Ідея запропонованого методу полягає у відображенні дистанційного простору з використанням віддаленої міри концентрації, міри віддаленості Штахеля–Донохью та міри скоректованої віддаленості.

Ключові слова: функція глибини, віддалена міра концентрації, багатовимірна класифікація.

O.A. Galkin

THE DEPTH-BASED CLASSIFICATION METHOD BASED ON REMOTE CONCENTRATION MEASURE FOR ASYMMETRIC DATA PROCESSING

Abstract. The author develops and investigates the depth-based classification method based on remote concentration measure for asymmetric data processing. The motivation for the construction of the method was inefficient use of affine invariant classifiers in combination with depth functions, which vanish outside the convex hull. The idea of the proposed method is to map a remote space using a remote concentration measure, Stahel–Donoho remoteness measure, and adjusted remoteness measure.

Keywords: depth function, remote concentration measure, multi-dimensional classification.

Галкин Александр Анатольевич,
кандидат физ.-мат. наук, ассистент кафедры Киевского национального университета имени Тараса Шевченко, e-mail: galkin.o.a@gmail.com.