e-ISSN 1607-3274 Радіоелектроніка, інформатика, управління. 2020. № 4
p-ISSN 2313-688X Radio Electronics, Computer Science, Control. 2020. № 4

УДК 004.93

ХЭШИРОВАНИЕ НА ОСНОВЕ ПОЛЯРНЫХ КООРДИНАТ ДЛЯ СОКРАЩЕНИЯ РАЗМЕРНОСТИ ДАННЫХ

Субботин С. А. – д-р техн. наук, профессор, заведующий кафедрой программных средств Национального университета «Запорожская политехника», Запорожье, Украина.

АННОТАЦИЯ

Актуальность. Для сокращения размерности данных в задачах распознавания и диагностирования на основе хэширования возникает необходимость сокращение временных затрат на формирование хэширующего преобразования.

Цель. Цель работы – уменьшение временных затрат на сокращение размерности данных путем создания метода хэширования, не требующего решения оптимизационной задачи поиска наилучшего случайного преобразования, а также уменьшение потерь локальных свойств признакового пространства.

Метод. Предложен метод формирования хэша, который переводит координаты экземпляров из исходной системы признаков в многомерную полярную систему координат, на основе которых, дискретизируя полярные координаты, с помощью эвристик различным способами кодирует и комбинирует значения дискретизированных полярных координат, формируя хэши экземпляров, из которых в качестве результирующего преобразования выбирает наилучшее в системе заданных критериев на основе минимизации числа коллизий, при которых экземпляры разных классов и разными значениями исходных признаков, получают одинаковые хэши. Это позволяет автоматизировать формирование хэширующих преобразований, исключить необходимость решения оптимизационных задач перебора случайных проекций, обеспечив сокращение затрат времени, а также делает хэширующее преобразование более свободным от навязывания данным разбиения признакового пространства, неприсущей им природы, что позволяет повысить обобщающие свойства и точность преобразований. Предложены критерии оценивания качества хэширующих преобразований, включающие определение числа позитивных и негативных коллизий, а также оценивания на их основе вероятностей соответствующих коллизий. Это позволяет автоматизировать анализ и выбор хэширующих преобразований для сокращения размерности данных в задачах распознавания и диагностирования.

Результаты. Проведено экспериментальное исследование, подтвердившее работоспособность предложенных методов при решении практических задач.

Выводы. Разработанное математическое обеспечение может быть рекомендовано для решения задач сокращения размерности данных.

КЛЮЧЕВЫЕ СЛОВА: хэширование, хэш, сокращение размерности выборки, полярные координаты.

НОМЕНКЛАТУРА

δ – заданная пользователем константа, регулирующая допустимое расхождение значений критериев качества редуцированной и исходной выборок;

ρ^s – расстояние от s-го экземпляра до центра нормированных координат (радиальная координата);

ρ_b^s – значение b-го бита целочисленного расстояния (или номера интервала расстояния) от центра полярной координатной системы до s-го экземпляра;

P – число бит для представления максимального целочисленного расстояния;

φ_j^s – j-я угловая координата s-го экземпляра;

$\varphi^s = \{\varphi_j^s\}$ – набор угловых координат s-го экземпляра;

φ_{jb}^s – значение b-го бита j-го целочисленного угла в полярной координатной системе для s-го экземпляра;

Ф – число бит для представления максимального целочисленного угла;

a – кодируемое число;

$a_{\min}$ – минимальное возможное значение переменной a;

$a_{\max}$ – максимальное возможное значение переменной a;

a_* – значение кода иерархического бинарного разбиения;

$\breve{a}$ – нижняя граница области;

$\hat{a}$ – верхняя граница области;

F – критерий качества исходной выборки;

F' – критерий качества полученной редуцированной выборки;

$f(x)$ – преобразование;

G – число групп битов;

j – номер экземпляра выборки;

k – номер класса;

K – число классов;

L – длина машинного слова;

N' – число признаков, характеризующих экземпляры редуцированной выборки;

N – число признаков, характеризующих экземпляры выборки;

N_{col-} – число нежелательных коллизий;

N_{col+} – число положительных коллизий;

P_{col-} – вероятность нежелательной коллизии;

P_{col+} – вероятность положительной коллизии;

s – номер экземпляра;

S – число экземпляров в выборке;

S' – число экземпляров в редуцированной выборке;

w_j – вес j-го признака;

DOI 10.15588/1607-3274-2020-4-12

w^{ρ} – вес полярной координаты;

w_j^{φ} – вес j-й угловой координаты;

x – набор экземпляров исходной выборки;

x' – набор экземпляров редуцированной выборки;

x^s – s-й экземпляр выборки;

x_j^s – значение j-го входного признака, сопоставленное s-му экземпляру выборки;

x_*^s – хэш s-го экземпляра выборки;

y – набор значений исходной признака;

y' – набор значений выходного признака, сопоставленных экземплярам редуцированной выборки;

y^s – значение выходного признака, сопоставленное s-му экземпляру выборки.

Z – число первых углов, включаемых хэш.

ВВЕДЕНИЕ

Сокращение размерности данных [1–4] представляет собой процесс замены исходного описания данных на сокращенное, полученное на основе исходного описания. Данный процесс имеет чрезвычайно важное значение для построения моделей зависимостей по прецедентам в условиях данных большой размерности, поскольку сокращение описания данных позволяет уменьшить сложность моделей, а также сократить временные затраты на их построение.

Объектом исследования являлся процесс сокращения размерности данных.

Одним из подходов к сокращению размерности данных является хэширование [5–10], которое представляет собой преобразование данных из исходного многомерного пространства описательных признаков в одномерное пространство хэша.

Предметом исследования являлись методы хэширующих преобразований для сокращения размерности данных.

Известные методы хэширования [5–10] основаны на итеративной процедуре перебора случайных преобразований, обеспечивающих приемлемое отображение выборки данных. Однако они являются чрезвычайно затратными по времени и могут приводить к потере локальных свойств признакового пространства.

Целью работы являлось уменьшение временных затрат на сокращение размерности данных путем создания метода хэширования, свободного от отмеченных выше недостатков.

1 ПОСТАНОВКА ЗАДАЧИ

Пусть задана исходная выборка наблюдений $<x, y>$, $x = \{x^s\}$, $x^s = \{x_j^s\}$, $y = \{y^s\}$, $s = 1, 2, ..., S$, $j = 1, 2, ..., N$.

Тогда задача сокращения размерности выборки $<x, y>$ состоит в том, чтобы получить $<x', y'>$: $x' \subseteq x$, $y' \subseteq y$, $S' \le S$, $N' \le N$. При этом критерий качества полученной редуцированной выборки F' должен принимать приемлемое значение относительно значения критерия качества для исходной выборки F: $|F–F'| \le \delta$.

Для заданной выборки наблюдений $<x, y>$ задача формирования хэширующего преобразования [11] состоит в том, чтобы получить $<x', y'>$: $x'=f(x)$, $y'=y$, $S'=S$, $N'=1$.

2 ОБЗОР ЛИТЕРАТУРЫ

Известные методы хэширования [5–10] определяют хэш экземпляра как взвешенную сумму значений его признаков, причем значения весов рассчитываются в итеративном режиме, что является весьма затратным по времени.

В [11] предложен эвристический метод хэширования, который для исключения перебора случайных проекций выборки из исходного пространства рассматривает иерархию разбиений пространства признаков на области, заменяя в общем случае вещественные значения признаков на дискретные номера интервалов по оси признака, стремясь для каждого признака найти такое разбиение на интервалы, при котором число интервалов будет наименьшим, но обеспечивающим требуемую точность. Здесь веса признаков определяются с учетом числа интервалов, сформированных для каждого признака. Чем меньше нужно интервалов для обеспечения приемлемой точности, тем более ценным является соответствующий признак. Недостатком данного подхода является навязывание данным неприсущей им природы прямоугольного разбиения, что ограничивает точность работы метода, либо может существенно снижать обобщающие свойства разбиения при необходимости обеспечения высокой точности.

Для обеспечения учета в преобразовании близости расположения экземпляров возможно использование многомерной полярной системы координат вместо исходной координатной системы.

Многомерная полярная система координат – система координат, в которой каждая точка определяется полярным радиусом и полярными углами. Полярная система координат задаётся лучом, который называют нулевым лучом, или полярной осью. Точка, из которой выходит этот луч, называется началом координат, или полюсом [12].

Радиальная координата соответствует расстоянию от точки до начала координат. Радиальная координата может принимать значения от нуля до бесконечности.

Угловая координата также называется полярным углом или азимутом и равна углу, на который нужно повернуть против часовой стрелки полярную ось для того, чтобы попасть в эту точку. Угловая координата изменяется в пределах от 0° до 360°.

Поскольку признаки в исходной системе координат могут иметь существенно различный масштаб шкалы, перед отображением в полярную систему их целесообразно нормировать.

Пронормируем значения признаков, отобразив их на интервал [0, 1]:

DOI 10.15588/1607-3274-2020-4-12

$$x_j^s = \frac{x_j^s - \min\limits_{i=1,2,..,N}\{x_i^s\}}{\max\limits_{i=1,2,..,N}\{x_i^s\} - \min\limits_{i=1,2,..,N}\{x_i^s\}}.$$

Для *s*-го экземпляра определим расстояние от него до центра нормированных координат (радиальную координату):

$$\rho^s = \sqrt{\sum_{j=1}^{N}\left(x_j^s\right)^2}.$$

Далее определим углы экземпляра относительно координатных осей в исходной системе координат-признаков в радианах [12]:

$$\varphi_j^s = \arccos\frac{x_j^s}{\sqrt{\sum\limits_{i=j}^{N}\left(x_i^s\right)^2}},\ j = 1, 2, ..., N–1.$$

Полученное отображение позволит оперировать экземплярами выборки в полярных координатах. Однако само по себе оно не сократит размерность данных. Поэтому актуальной задачей является разработка методов формирования хэшей на основе полярных координат экземпляров для сокращения размерности данных.

3 МАТЕРИАЛЫ И МЕТОДЫ

Для упрощения процесса создания хэширующего преобразования на основе полярных координат экземпляров предлагается использовать эвристические методы, которые будут по-разному комбинировать полярные координаты, получая на их основе хэши. Из набора таких преобразований возможно будет отобрать наилучшее, задав соответствующие критерии.

Формально предложенный метод хэширования, реализующий описанные выше идеи, можно представить следующим образом.

Этап инициализации. Задать исходную выборку данных <*x*, *y*>. Пронормировать значения признаков.

Этап преобразования координат в полярную систему. Определить на основе нормированных значений признаков полярные координаты экземпляров $<\{<\rho^s,\varphi^s>\},y>$, где $\varphi^s = \{\varphi_j^s\}$. После чего сократить описание данных путем перехода к целочисленным значениям угловых координат:

$$\varphi_j^s = \left\lfloor \frac{\varphi_j^s 180}{\pi} \right\rfloor.$$

Этап формирования хэширующего преобразования. Хэш *s*-го экземпляра, отображенного в полярную систему координат, возможно получить одним из следующих способов.

Способ 1. Составим хэш следующим образом: первая часть хэша – целочисленное значение или номер интервала значений квантованного расстояния, вторая часть хэша – последовательно по признакам целочисленные значения или номера интервалов значений углов экземпляра.

Число бит для представления максимального целочисленного расстояния составит:

$$\mathrm{P} = \left\lceil \log_2 \max_{s=1,2,..,S}\{\rho^s\} \right\rceil.$$

Для *N* признаков в полярной *N*-мерной системе координат получим:

$$\mathrm{P} \le \left\lfloor \log_2 \sqrt{N} \right\rfloor.$$

Число бит для представления максимального целочисленного угла составит:

$$\Phi = \left\lceil \log_2 \max_{p=1,2,..,S} \left\{ \max_{j=1,N-1} \{\varphi_j^p\} \right\} \right\rceil.$$

Для *N* признаков в полярной *N*-мерной системе координат получим: $\Phi \le \log_2 90$.

При составлении хэша необходимо обеспечить, чтобы $L \ge \mathrm{P} + (N-1)\Phi$ (для современных ЭВМ, как правило, *L* = 64 бит):

$$L \ge \left\lfloor \log_2 \sqrt{N} \right\rfloor + (N-1)\left\lceil \log_2 90 \right\rceil.$$

Поскольку при большом числе признаков *N* (уже при $N \ge 10$ для *L*= 64) данное условие не будет выполняться из-за малого значения *L*, то целесообразно квантовать расстояние и углы, например, путем разбиения их диапазонов значений на интервалы и замены значения номером интервала. Это, с одной стороны, приведет к потере информации, но, с другой стороны, повысит уровень обобщения данных.

Применительно к обучающей выборке эвристически определим, что расстояние следует квантовать не менее чем на *K* и не более чем на $\lceil \log_2 S \rceil$ уровней. Таким образом, получим:

$$\mathrm{P} = \lceil \log_2 \max(K, S) \rceil,$$

тогда

$$\Phi = \left\lfloor \frac{L - \mathrm{P}}{N-1} \right\rfloor,\ N > 1.$$

Если $\Phi < 1$, то хэш не сможет представить все углы и необходим специальный механизм их объединения.

В простейшем случае можно ограничиться включением в хэш только первых *Z* углов (*Z*>0):

$$\Phi = \left\lfloor \frac{L - \mathrm{P}}{Z} \right\rfloor > 1.$$

Решая данное неравенство, получим:

$$Z = \left\lfloor \frac{L - \mathrm{P}}{2} \right\rfloor.$$

С другой стороны, углы можно разбить на группы и заменить значения углов каждой группы на среднее или максимальное или минимальное значение угла в группе, после чего заменить это значение на номер интервала значений среднего угла в группе. Конструктивно возможно предложить широкий набор средств для объединения значений углов.

Если $\Phi \ge 1$, то каждый угол будет представлен не более чем 2^Φ интервалами значений.

Для перехода от реального значения расстояния к номеру интервала можно использовать формулу:

DOI 10.15588/1607-3274-2020-4-12

$$\rho^s = \left\lfloor \frac{\rho^s}{\frac{\sqrt{N}}{2^{\mathrm{P}}}} \right\rfloor = \left\lfloor \frac{\sqrt{\sum_{j=1}^{N}\left(x_j^s\right)^2}}{\frac{\sqrt{N}}{2^{\mathrm{P}}}} \right\rfloor = \left\lfloor \frac{2^{\mathrm{P}}}{\sqrt{N}} \sqrt{\sum_{j=1}^{N}\left(x_j^s\right)^2} \right\rfloor .$$

Для перехода от реального значения угла в градусах к номеру интервала угла можно использовать формулу.

$$\varphi_j^s = \left\lfloor \frac{\varphi_j^s}{\frac{90^\circ}{2^\Phi}} \right\rfloor = \left\lfloor \frac{2^\Phi \varphi_j^s}{90^\circ} \right\rfloor .$$

В итоге хэш *s*-го экземпляра определим по формуле:

$$x_*^s = w^\rho \rho^s + \sum_{j=1}^{N-1} w_j^\varphi \varphi_j^s ,$$

$$w^\rho = 2^{L-P+1},$$

$$w_j^\varphi = 2^{L-\mathrm{P}-\Phi j+1} .$$

Способ 2. Представим хэш *s*-го экземпляра последовательностью групп битов, где в каждой группе первый бит – соответствующий группе бит целого значения или номера интервала квантованного расстояния, а последующие биты – соответствующие группе биты целочисленных значений или номеров интервалов углов экземпляра:

$$x_*^s = \sum_{g=1}^{G} 2^{(G-g)(Z+1)} \left(\rho_g^s + \sum_{j=1}^{Z} 2^{Z-j} \varphi_{jg}^s \right),$$

$$\rho_b^s = (\rho^s \bmod 2^{b+1} - \rho^s \bmod 2^b),$$

$$\varphi_{jb}^s = (\varphi_j^s \bmod 2^{b+1} - \varphi_j^s \bmod 2^b),$$

где число учитываемых углов Z: $N{-}1 \ge Z \ge 1$, число групп битов G: $1 \le \lceil G(Z+1) \rceil \le L$ (следовательно, стоит задавать: $G = \lfloor L/(Z+1) \rfloor$), ρ_b^s – значение *b*-го бита целочисленного расстояния (или номера интервала расстояния) от центра полярной координатной системы до *s*-го экземпляра, φ_{jb}^s – значение *b*-го бита *j*-го целочисленного угла в полярной координатной системе для *s*-го экземпляра.

Способ 3. Переведем номер интервала или значение расстояния экземпляра в формат кода иерархического бинарного разбиения: старший разряд указывает номер одной из двух равных по длине областей, на которые разбит диапазон значений расстояния или диапазон значений номера интервала расстояния, в которую попал экземпляр по расстоянию, затем каждый последующий разряд аналогичным образом указывает в какую из подобластей области старшего разряда попал экземпляр. Подобным же образом переведем в формат кода иерархического бинарного разбиения значения углов или номеров интервалов углов экземпляра.

Обобщенно перевод числа *a* будет осуществляться следующим образом. Задать кодируемое число *a*, его минимальное $a_{\min}$ и максимальное $a_{\max}$ возможные значения, длину разрядной сетки ЭВМ *L*. Установить начальные значения кода иерархического бинарного разбиения: $a_* = 0$, а также переменных границ областей $\breve{a} = a_{\min}$, $\hat{a} = a_{\max}$. Для *i*=1, 2, ..., *L* в цикле повторять: установить: $\bar{a} = (\breve{a} + \hat{a})/2$; если $a > \bar{a}$, то принять $\breve{a} = \bar{a}$, $a_* = 2a_* + 1$, в противном случае – принять: $\hat{a} = \bar{a}$, $a_* = 2a_*$.

После перевода расстояния и углов представим по аналогии со вторым вариантом хэш экземпляра последовательностью групп битов, где в каждой группе первый бит – соответствующий группе бит иерархического кода целого значения или номера интервала квантованного расстояния, а последующие биты – соответствующие группе биты кодов целочисленных значений или номеров интервалов углов экземпляра. Формулы в данном случае будут аналогичными способу 2, но в качестве ρ_b^s будет использоваться значение *b*-го бита иерархического бинарного кода целочисленного расстояния (или номера интервала расстояния) от центра полярной координатной системы до *s*-го экземпляра, а в качестве φ_{jb}^s – значение *b*-го бита иерархического бинарного кода *j*-го целочисленного угла в полярной координатной системе для *s*-го экземпляра.

Способ 4. Данный способ будет аналогичен способу 1, но вместо значения расстояния или номера его интервала и вместо значений или номеров интервалов углов будем использовать их иерархические коды, получаемые подобно способу 3. В итоге составим хэш следующим образом: первая часть хэша – иерархический код значения или номера интервала значений квантованного расстояния, вторая часть хэша – последовательно по признакам коды целочисленных значений или номеров интервалов значений углов экземпляра.

Этап оценивания качества хэширующего преобразования. Для сформированных хэшей выборки оценим значение критериев качества хэширующих преобразований. Далее как результирующее преобразование выберем то, которое обеспечит наилучшее значение заданного критерия.

Критерии оценки качества хэша определим следующим образом.

Качество хэша возможно оценить числом коллизий для одной и той же выборки [13].

Коллизией называют ситуацию, когда несколько экземпляров с разными значениями признаков получают одинаковое значение хэша.

Очевидно, что не все коллизии являются плохими, поскольку, если несколько экземпляров, принадлежащих к одному и тому же классу, но имеющих разные значения признаков, получат одинаковое значение хэша, это не только не ухудшит точность синте-

DOI 10.15588/1607-3274-2020-4-12

зируемых на основе такого хэша моделей, но, наоборот, повысить их обобщающие свойства. Поэтому для оценки качества хэша будем учитывать отдельно число нежелательных коллизий, т.е. таких, когда экземпляры, принадлежащие к разным классам, получат одинаковые хэши:

$$N_{col-} = \sum_{s=1}^{S} \sum_{p=s+1}^{S} \{1 \mid x_*^s = x_*^p, y^s \neq y_*^p\},$$

а также число положительных коллизий, т.е. таких, когда когда экземпляры, принадлежащие к одному классу, получат одинаковые хэши:

$$N_{col+} = \sum_{s=1}^{S} \sum_{p=s+1}^{S} \{1 \mid x_*^s = x_*^p, y^s = y_*^p\}$$

Вероятности таких коллизий оценим, соответственно, как

$$P_{col-} = \frac{N_{col-}}{S(S-1)} \text{ и } P_{col+} = \frac{N_{col+}}{S(S-1)}.$$

Хэш будет тем лучше для одной и той же выборки, чем меньше будет вероятность негативных коллизий и выше вероятность позитивных коллизий. При этом заметим, что минимизация негативных коллизий предпочтительнее максимизации позитивных коллизий.

Также при сравнении хэшей возможно использовать такие меры, как показатели индивидуальной информативности хэшей экземпляров по отношению к выходному признаку и между собой [11].

Если один хэш будет теснее связан с выходным признаком по сравнению со связью другого хэша с выходным признаком, то первый хэш, очевидно, индивидуально более информативен.

Если несколько хэшей тесно связаны между собой, то, очевидно, что из них стоит выбрать то, который теснее всего связан с выходным признаком.

Заметим, что для хэшей, как и для первичных признаков индивидуальная и групповая информативности отличаются. Поэтому даже индивидуально малоинформативные хэши могут совместно оказаться высокоинформативными.

Аналогичным образом можно сравнивать хэши с исходными признаками. Использование хэша имеет смысл, если он лучше (информативнее) любого из оригинальных признаков и (или) информативнее любого из учитываемых в нем оригинальных признаков.

4 ЭКСПЕРИМЕНТЫ

Для изучения свойств предложенных преобразований они были программно реализованы и исследованы путем решения практических задач [14–16], характеристики которых приведены в табл. 1.

Для каждой практической задачи проводились эксперименты по расчету хэшей экземпляров и оцениванию вероятностей коллизий.

В случае, если исходный размер признаков не помещался в хэш, выбирался набор первых признаков, который позволял рассчитать хэш без переполнения.

Таблица 1 – Характеристики практических задач

Задача	Описание	N	S	K
Iris	Классификация ирисов Фишера [14]	4	150	2
Aritmia	Диагностика сердечной аритмии [15]	279	452	2
Acutediag	Урологическая диагностика [16]	6	120	4

5 РЕЗУЛЬТАТЫ

Результаты проведенных экспериментов приведены в табл. 2.

Таблица 2 – Результаты экспериментов

Задача	Способ расчета хэша	N_{col-}	N_{col+}	P_{col-}	P_{col+}
Iris	1	0	1	0	$4{,}4743\times10^{-5}$
	2	130	54	0,0058	0,0024
	3	0	1	0	$4{,}4743\times10^{-5}$
	4	0	1	0	$4{,}4743\times10^{-5}$
Aritmia	1	0	0	0	0
	2	19	48	$9{,}4036\times10^{-5}$	$2{,}3756\times10^{-4}$
	3	8	20	$3{,}9594\times10^{-5}$	$9{,}8985\times10^{-5}$
	4	0	0	0	0
Acute-diag	1	19	24	0,0013	0,0017
	2	175	128	0,0123	0,0090
	3	19	24	0,0013	0,0017
	4	21	26	0,0015	0,0018

На рис. 1–3 приведены результаты расчета хэшей.

6 ОБСУЖДЕНИЕ

Как видно из табл. 2 и рис 1–3, для всех рассмотренных задач удалось на основе предложенных преобразований получить хэши, обеспечивающие малые вероятности негативных коллизий хотя бы на основе одного из предложенных способов.

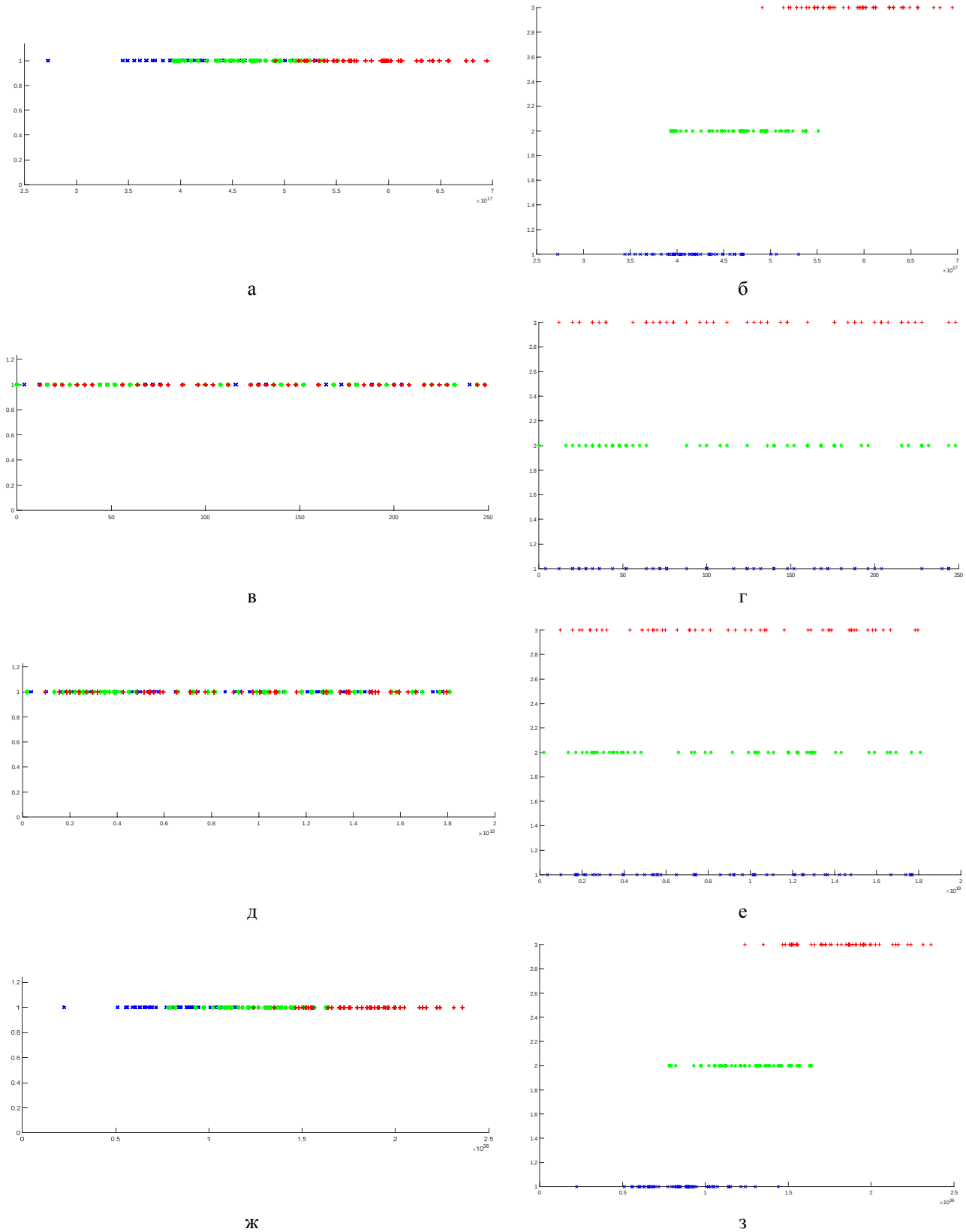
Вместе с тем, следует отметить, что ни одно из предложенных преобразований само по себе не гарантирует наилучшего результата для всех задач.

Поэтому на практике рекомендуется определять хэши всеми доступными способами, рассчитывать показатели качества для каждого вида хэшей и отбирать как результирующие искусственные признаки те хэши, которые будут обеспечивать наилучшее качество преобразования данных.

Важной особенностью предложенных хэширующих преобразований является их интерпретабельность относительно системы полярных координат, в то время, как преобразования [4-9] не обеспечивают сохранение связи с исходными признаками.

По сравнению с методами [4–9] предложенные преобразования не требуют итеративного перебора случайных значений весовых коэффициентов для выбора наилучшего преобразования, что позволяет за один проход определить параметры хэша.

DOI 10.15588/1607-3274-2020-4-12



а б в г д е ж з

Рисунок 1 – Результаты расчета хэшей для задачи Iris:
а, б – способ 1, в, г – способ 2, д, е – способ 3, ж, з – способ 4
(а, в, д, ж – экземпляры всех классов отображены на одной оси,
б, г, е, з – экземпляры разных классов разнесены по оси ординат)

DOI 10.15588/1607-3274-2020-4-12

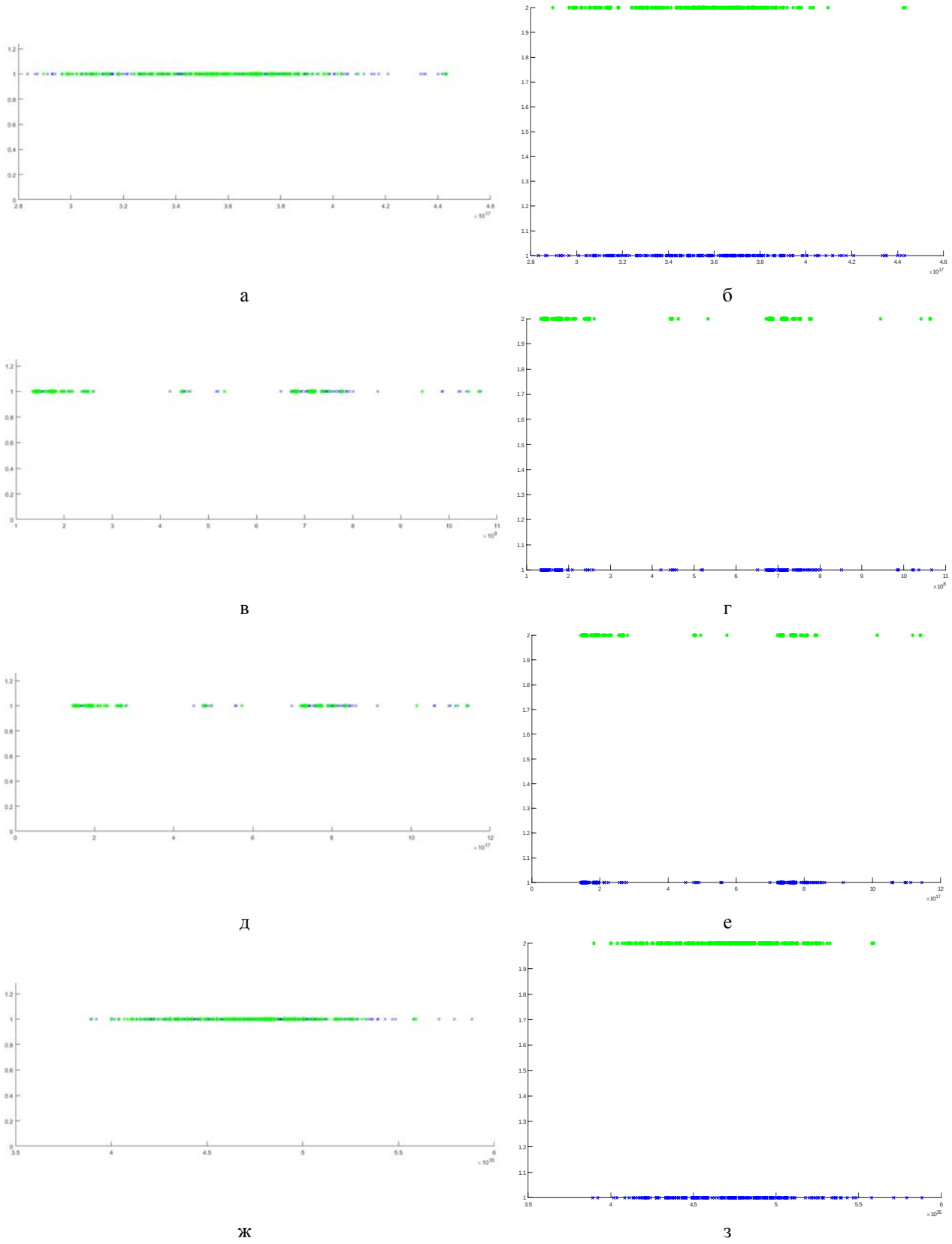


Рисунок 2 – Результаты расчета хэшей для задачи Aritmia:
а, б – способ 1, в, г – способ 2, д, е – способ 3, ж, з – способ 4
(а, в, д, ж – экземпляры всех классов отображены на одной оси,
б, г, е, з – экземпляры разных классов разнесены по оси ординат)

DOI 10.15588/1607-3274-2020-4-12

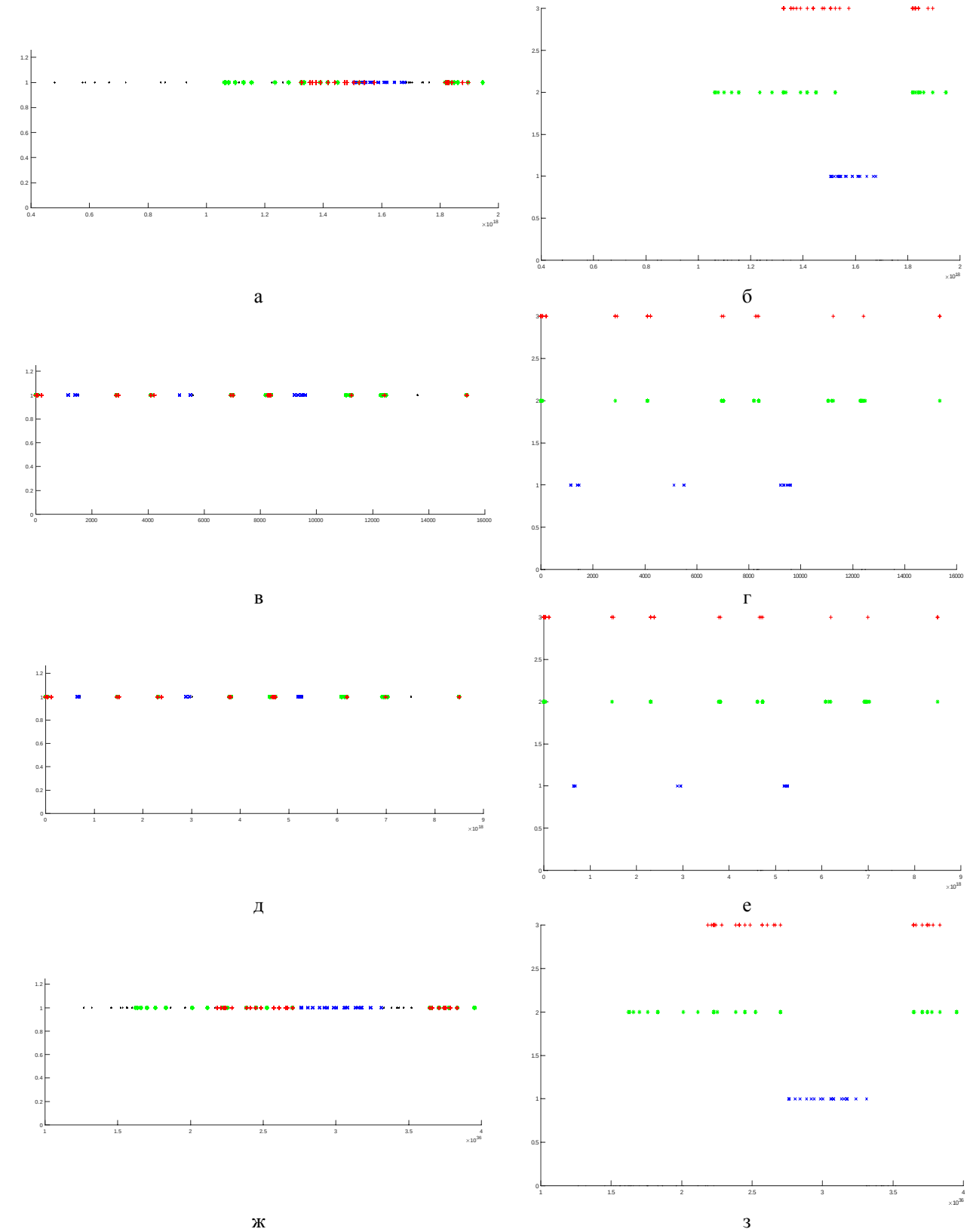


Рисунок 3 – Результаты расчета хэшей для задачи Acutediag:
а, б – способ 1, в, г – способ 2, д, е – способ 3, ж, з – способ 4
(а, в, д, ж – экземпляры всех классов отображены на одной оси,
б, г, е, з – экземпляры разных классов разнесены по оси ординат)

DOI 10.15588/1607-3274-2020-4-12

ВЫВОДЫ

Решена актуальная задача создания метода формирования хэша для сокращения размерности данных.

Научная новизна полученных результатов состоит в том, что:

– предложен метод формирования хэша, который переводит координаты экземпляров из исходной системы признаков в многомерную полярную систему координат, на основе которых, дискретизируя полярные координаты, с помощью эвристик различным способами кодирует и комбинирует значения дискретизированных полярных координат, формируя хэши экземляров, из которых в качестве результирующего преобразования выбирает наилучшее в системе заданных критериев на основе минимизации числа коллизий, при которых экземпляры разных классов и разными значениями исходных признаков, получают одинаковые хэши. Это позволяет автоматизировать формирование хэширующих преобразований, исключить необходимость решения оптимизационных задач перебора случайных проекций, обеспечив сокращение затрат времени, а также делает хэширующее преобразование более свободным от навязывания данным разбиения признакового пространства, неприсущей им природы, что позволяет повысить обобщающие свойства и точность преобразований;

– предложены критерии оценивания качества хэширующих преобразований, включающие определение числа позитивных и негативных коллизий, а также оценивания на их основе вероятностей соответствующих коллизий. Это позволяет автоматизировать анализ и выбор хэширующих преобразований для сокращения размерности данных в задачах распознавания и диагностирования.

Практическая ценность полученных результатов состоит в том, что проведено экспериментальное исследование, подтвердившее работоспособность предложенного метода при решении практических задач распознавания и диагностирования. Разработанное математическое обеспечение может быть рекомендовано для решения задач сокращения размерности данных.

Перспективы дальнейших исследований состоят в том, чтобы изучить работоспособность предложенного метода на более широком классе задач.

БЛАГОДАРНОСТИ

Работа выполнена в рамках госбюджетной научно-исследовательской темы «Интеллектуальные методы и программные средства диагностирования и неразрушающего контроля качества техники военного и гражданского назначения» (гос. рег. № 0119U100360) Национального университета «Запорожская политехника» при частичной поддержке международных проектов «Innovative Multidisciplinary Curriculum in Artificial Implants for Bio-Engineering BSc/MSc degrees» программы «Эразмус+» Европейского Союза и «Virtual Master Cooperation Data Science» (VIMACS) Немецкой службы академических обменов DAAD.

ЛИТЕРАТУРА / ЛІТЕРАТУРА

1. Subbotin S. The Dimensionality Reduction Methods Based on Computational Intelligence in Problems of Object Classification and Diagnosis / S. Subbotin, A. Oliinyk // Recent Advances in Systems, Control and Information Technology / Eds.: R. Szewczyk, M. Kaliczyńska . – Cham: Springer, 2017. – P. 11–19. DOI: 10.1007/978-3-319-48923-0_2
2. Jensen R. Computational intelligence and feature selection: rough and fuzzy approaches / R. Jensen, Q. Shen. – Hoboken: John Wiley & Sons, 2008. – 300 p.
3. Subbotin S. The instance and feature selection for neural network based diagnosis of chronic obstructive bronchitis / / S. Subbotin // Applications of Computational Intelligence in Biomedical Technology. – Cham: Springer, 2016. – P. 215–228. DOI: 10.1007/978-3-319-19147-8_13
4. Łukasik S. An algorithm for sample and data dimensionality reduction using fast simulated annealing / S. Łukasik, P. Kulczycki // Advanced Data Mining and Applications, Lecture Notes in Computer Science. – Berlin : Springer, 2011. – Vol. 7120. – P. 152–161. DOI: 10.1007/978-3-642-25853-4_12
5. Feature Hashing for Large Scale Multitask Learning / [K. Weinberger, A. Dasgupta, J. Langford et al.] // 26th Annual International Conference on Machine Learning (ICML '09) Montreal, June 2009 : proceedings. – New York : ACM, 2009. – P. 1113–1120. DOI: 10.1145/1553374.1553516
6. Wolfson H. J. Geometric Hashing: An Overview / H. J. Wolfson, I. Rigoutsos // IEEE Computational Science and Engineering. – 1997. – Vol. 4. – № 4. – P. 10–21.
7. Indyk P. Approximate nearest neighbors: towards removing the curse of dimensionality / P. Indyk; R. Motwani // The 30th annual ACM symposium on Theory of computing (STOC'98), Dallas, 23-26 of May 1998 : proceedings. – 1998. — P. 604-613. DOI:10.1145/276698.276876
8. Fast supervised discrete hashing / [J. Gui, T. Liu, Z. Sun et al.] // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2017. – Vol. 40. – № 2. – P 490–496. DOI: 10.1109/TPAMI.2017.2678475
9. Zhao K. Locality Preserving Hashing / K. Zhao, H. Lu, J. Mei // Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI'14), Québec, 27–31 of July 2014 : proceedings. – Palo Alto: AAAI Press, 2014. – P. 2874–2880.
10. Tsai Y.-H. Locality preserving hashing / Y.-H. Tsai, M.-H. Yang // 2014 IEEE International Conference on Image Processing (ICIP), Paris, 27–30 of October 2014: proceedings. – Los Alamitos: IEEE, 2014. – P. 2988–2992. DOI: 10.1109/ICIP.2014.7025604.
11. Субботин С. А. Оценка информативности и отбор экземпляров на основе хэширования / С. А. Субботин // Radio Electronics, Computer Science, Control. – 2020. – № 3. – P. 129–137. DOI: 10.15588/1607-3274-2020-3-12
12. Blumenson L. E. A Derivation of n-Dimensional Spherical Coordinates / L. E. Blumenson // The American Mathematical Monthly. – 1960. – Vol. 67, № 1. – P. 63–66. DOI:10.2307/2308932. JSTOR 2308932.
13. Subbotin S. A. Methods and characteristics of locality-preserving transformations in the problems of computational intelligence / S. A. Subbotin // Radio Electronics, Computer Science, Control. – 2014. – № 1. – P. 120–128. DOI: 10.15588/1607-3274-2014-1-17

DOI 10.15588/1607-3274-2020-4-12

14. Fisher Iris dataset [Electronic resource]. – Access mode: https://archive.ics.uci.edu/ml/datasets/Iris
15. Arrhythmia dataset [Electronic resource]. – Access mode: http://archive.ics.uci.edu/ml/datasets/Arrhythmia
16. Acute inflammations data set [Electronic resource]. – Access mode: https://archive.ics.uci.edu/ml/datasets/Acute+ Inflammations

Received 25.09.2020.
Accepted 19.10.2020.

УДК 004.93

ГЕШУВАННЯ НА ОСНОВІ ПОЛЯРНИХ КООРДИНАТ ДЛЯ СКОРОЧЕННЯ РОЗМІРНОСТІ ДАНИХ

Субботін С. О. – д-р техн. наук, професор, завідувач кафедри програмних засобів Національного університету «Запорізька політехніка», Запоріжжя, Україна.

АНОТАЦІЯ

Актуальність. Для скорочення розмірності даних в задачах розпізнавання та діагностування на основі гешування виникає необхідність скорочення витрат часу на формування гешувального перетворення.

Мета. Мета роботи – зменшення витрат часу на скорочення розмірності даних шляхом створення методу гешування, що не потребує вирішення оптимізаційної задачі пошуку найкращого випадкового перетворення, а також зменшення втрат локальних властивостей простору ознак.

Метод. Запропоновано метод формування гешу, який переводить координати екземплярів з вихідної системи ознак у багатовимірну полярну систему координат, на основі яких, дискретизуючи полярні координати, за допомогою евристик різним способами кодує і комбінує значення дискретизованих полярних координат, формуючи геші екземплярів, з яких в якості результуючого перетворення вибирає найкраще у системі заданих критеріїв на основі мінімізації кількості колізій, при яких екземпляри різних класів і різних значеннях вихідних ознак, отримують однакові геши. Це дозволяє автоматизувати формування гешувальних перетворень, виключити необхідність вирішення оптимізаційних задач перебору випадкових проекцій, забезпечивши скорочення витрат часу, а також робить гешувальне перетворення більш вільним від нав'язування даним розбиття простору ознак, непритаманної їм природи, що дозволяє підвищити узагальнюючі властивості і точність перетворень. Запропоновано критерії оцінювання якості гешувальних перетворень, що містять визначення кількостей позитивних і негативних колізій, а також оцінювання на їхній основі ймовірностей відповідних колізій. Це дозволяє автоматизувати аналіз і вибір гешувальних перетворень для скорочення розмірності даних в задачах розпізнавання та діагностування.

Результати. Проведено експериментальне дослідження, яке підтвердило працездатність запропонованих методів при вирішенні практичних завдань.

Висновки. Розроблене математичне забезпечення може бути рекомендовано для вирішення завдань скорочення размірності даних.

КЛЮЧОВІ СЛОВА: хэшування, хэш, скорочення розмірності вибірки, полярні координати.

UDC 004.93

THE POLAR COORDINATES BASED HASHING FOR DATA DIMENSIONALITY REDUCTION

Subbotin S. A. – Dr. Sc., Professor, Head of the Department of Software Tools at the National University “Zaporizhzhia Polytechnic”, Zaporizhzhia, Ukraine.

ABSTRACT

Context. To reduce the data dimensionality of in recognition and diagnostics problems based on hashing, it becomes necessary to reduce the time spent on generating a hashing transformation.

Objective. The purpose of the work is to reduce the time spent on reducing the dimension of data by creating a hashing method that does not require solving the optimization problem of finding the best random transformation, as well as reducing the loss of local properties of the feature space.

Method. A hash generation method is proposed. It converts the instance coordinates from the original feature system into a multidimensional polar coordinate system, on which basis discretize polar coordinates using heuristics, in various ways encodes and combines the values of the discretized polar coordinates, forming hashes of instances, from which as the resulting transformation selects the best one in the system of given criteria based on minimizing the number of collisions in which instances of different classes and different values of the original features receive the same hashes. This makes possible to automate the formation of hashing transformations, eliminate the need to solve optimization problems of enumerating random projections, ensuring a reduction in time consumption, and also makes the hashing transformation freer from imposing the data on the partitioning of the feature space, of a non-inherent nature, which allows increase the generalizing properties and accuracy of transformations. Criteria for evaluating the quality of hashing transformations are proposed, including determining the number of positive and negative collisions, as well as evaluating the probabilities of the corresponding collisions on their basis. This makes it possible to automate the analysis and selection of hashing transformations to reduce the dimension of the data in the problems of recognition and diagnosis.

Results. An experimental study has been carried out, which has confirmed the efficiency of the proposed methods in solving practical problems.

Conclusions. The developed mathematical support can be recommended for solving problems of data dimension reduction.

KEYWORDS: hashing, hash, sample size reduction, polar coordinates.

DOI 10.15588/1607-3274-2020-4-12

REFERENCES

1. Subbotin S., Oliinyk A. Eds.: R. Szewczyk, M. Kaliczyńska The Dimensionality Reduction Methods Based on Computational Intelligence in Problems of Object Classification and Diagnosis, *Recent Advances in Systems, Control and Information Technology*. Cham, Springer, 2017, pp. 11–19. DOI: 10.1007/978-3-319-48923-0_2
2. Jensen R., Shen Q. Computational intelligence and feature selection: rough and fuzzy approaches. Hoboken, John Wiley & Sons, 2008, 300 p.
3. Subbotin S. The instance and feature selection for neural network based diagnosis of chronic obstructive bronchitis, *Applications of Computational Intelligence in Biomedical Technology*. Cham, Springer, 2016, pp. 215–228. DOI: 10.1007/978-3-319-19147-8_13
4. Łukasik S., Kulczycki P. An algorithm for sample and data dimensionality reduction using fast simulated annealing, *Advanced Data Mining and Applications, Lecture Notes in Computer Science.* Berlin, Springer, 2011, Vol. 7120, pp. 152–161. DOI: 10.1007/978-3-642-25853-4_12
5. Weinberger K., Dasgupta A., Langford J., Smola A., Attenberg J. Feature Hashing for Large Scale Multitask Learning, *26th Annual International Conference on Machine Learning (ICML '09) Montreal, June 2009 : proceedings*. New York, ACM, 2009, pp. 1113–1120. DOI: 10.1145/1553374.1553516
6. Wolfson H. J., Rigoutsos I. Geometric Hashing: An Overview, *IEEE Computational Science and Engineering*, 1997, Vol. 4, № 4, pp. 10–21.
7. Indyk P., Motwani R. Approximate nearest neighbors: towards removing the curse of dimensionality, *The 30th annual ACM symposium on Theory of computing (STOC'98), Dallas, 23–26 of May 1998 : proceedings*, 1998, pp. 604–613. DOI:10.1145/276698.276876
8. Gui J., Liu T., Sun Z., Tao D., Tan T. Fast supervised discrete hashing, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, Vol. 40, No. 2, pp. 490–496. DOI: 10.1109/TPAMI.2017.2678475
9. Zhao K., Lu H., Mei J. Locality Preserving Hashing, *Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI'14), Québec, 27–31 of July 2014 : proceedings*, Palo Alto, AAAI Press, 2014, pp. 2874–2880.
10. Tsai Y.-H., Yang M.-H. Locality preserving hashing, *2014 IEEE International Conference on Image Processing (ICIP), Paris, 27–30 of October 2014: proceedings*. Los Alamitos: IEEE, 2014, pp. 2988–2992. DOI: 10.1109/ICIP.2014.7025604.
11. Subbotin S. A. Otsenka informativnosti i otbor ekzemplyarov na osnove kheshirovaniya, *Radio Electronics, Computer Science, Control*, 2020, No. 3. pp. 129–137. DOI: 10.15588/1607-3274-2020-3-12
12. Blumenson L. E. A Derivation of n-Dimensional Spherical Coordinates, *The American Mathematical Monthly*, 1960, Vol. 67, No. 1, pp. 63–66. DOI:10.2307/2308932. JSTOR 2308932.
13. Methods and characteristics of locality-preserving transformations in the problems of computational intelligence, *Radio Electronics, Computer Science, Control*, 2014, No. 1, pp. 120–128. DOI: 10.15588/1607-3274-2014-1-17
14. Fisher Iris dataset [Electronic resource]. Access mode: https://archive.ics.uci.edu/ml/datasets/Iris
15. Arrhythmia dataset [Electronic resource]. Access mode: http://archive.ics.uci.edu/ml/datasets/Arrhythmia
16. Acute inflammations data set [Electronic resource]. Access mode: https://archive.ics.uci.edu/ml/datasets/Acute+Inflammations

DOI 10.15588/1607-3274-2020-4-12