e-ISSN 1607-3274 Радіоелектроніка, інформатика, управління. 2020. № 3
p-ISSN 2313-688X Radio Electronics, Computer Science, Control. 2020. № 3

УДК 004.93

ОЦЕНКА ИНФОРМАТИВНОСТИ И ОТБОР ЭКЗЕМПЛЯРОВ НА ОСНОВЕ ХЭШИРОВАНИЯ

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

АННОТАЦИЯ

Актуальность. Для сокращения размерности данных при построении диагностических и распознающих моделей возникает необходимость отбора наиболее информативных экземпляров, а также отбора наиболее информативных признаков. Затраты времени на отдельную реализацию данных процедур являются высокими вследствие итеративности и взаимосвязанности данных процедур.

Цель. Цель работы – сокращение временных затрат на сокращение размерности данных путем создания метода отбора наиболее информативных экземпляров на основе хэширования.

Метод. Предложен метод расчета весов для определения хэшей экземпляров, который детерминированным способом определяет веса признаков на основе их рангов, которые, в свою очередь, определяет с учетом числа равных разбиений диапазонов признаков, минимально достаточного для выделения кластеров на оси признака с приемлемой точностью. Это позволяет исключить необходимость итеративного перебора различных комбинаций признаков, определения случайных проекций признаков, а также решение итеративных оптимизационных задач поиска наилучшей проекции признаков, что существенно сокращает затраты времени на расчет весов, при этом обеспечивая локальную чувствительность хэша. Полученные хэши возможно использовать как для отбора экземпляров, так и для отбора признаков.

Предложен метод определения индивидуальной и групповой значимости экземпляров выборки, в котором использует как меру сходства расстояния между хэшами экземпляров и по аналогии с методом потенциалов находит потенциалы, наводимые классами на каждый экземпляр, а на их основе определяет показатели значимости экземпляров, исходя из того, что экземпляр в пространстве признаков тем информативнее, чем меньше минимальная разность потенциалов классов, наводимых на экземпляр.

Предложен метод определения оценок информативности признаков, который на основе нормирования весов, полученных при формировании хэшей, определяет показатели информативности признаков, отдавая предпочтение признакам с меньшим числом разбиений.

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: экземпляр, признак, информативность, хэширование, хэш, сокращение размерности выборки.

НОМЕНКЛАТУРА

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

ε – максимально допустимое значение ошибки;

$d_*^{s,p}$ – расстояние хэшей s-го и p-го экземпляров;

$E_{j,q}$ – ошибка для каждого q-го интервала значений j-го признака;

E_j – суммарная ошибка для всех интервалов j-го признака;

F' – критерий качества полученной редуцированной выборки;

F – критерий качества исходной выборки;

G – группа экземпляров;

I_*^s – показатель индивидуальной значимости s-го экземпляра;

$I_*(G)$ – оценки групповой информативности экземпляров в группе G;

I_j – показатели индивидуальной информативности признаков;

j – номер экземпляра выборки;

k – номер класса;

K – число классов;

M – объем памяти ЭВМ, использованной при построении модели на основе исходной выборки;

M' – объем памяти ЭВМ, использованной при построении модели на основе редуцированной выборки;

N' – число признаков, характеризующих экземпляры редуцированной выборки;

N – число признаков, характеризующих экземпляры выборки;

n – размерность выборки;

n' – размерность редуцированной выборки;

$P^k(x_*^s)$ – потенциал, наводимый экземплярами разных классов на хэш s-го экземпляра;

Q – число равных по длине интервалов, на которые разбиваются диапазоны значений признаков;

Q_j – число интервалов, на которые разбивается диапазон значений j-го признака;

q – номер интервала значений признака;

r_j – ранг j-го признака;

$S_k^{j,q}$ – число экземпляров k-го класса, попавших в него;

DOI 10.15588/1607-3274-2020-3-12

S' – число экземпляров в редуцированной выборке;

s – номер экземпляра;

S – число экземпляров в выборке;

t – время построения модели на основе исходной выборки;

t' – время построения модели на основе редуцированной выборки;

w_j – вес j-го признака;

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.

ВВЕДЕНИЕ

В задачах построения диагностических и распознающих моделей по прецедентам зачастую приходится сталкиваться с проблемой большой размерности данных [1], когда число экземпляров является довольно большим, а число характеризующих их признаков также велико. Построение моделей в таких задачах, как правило, сопряжено с большими затратами машинного времени и памяти ЭВМ. Поэтому возникает необходимость сокращения размерности данных [2].

Объектом исследования являлся процесс сокращения размерности данных.

Сокращение размерности данных возможно осуществить путем отбора наиболее информативных признаков [1, 3], путем отбора наиболее значимых экземпляров [3–9], а также путем формирования искусственных признаков [10, 11].

Отбор информативных признаков (feature selection) [1, 3] является наиболее хорошо разработанным и широко используемым инструментом для сокращения размерности данных. Однако он эффективен только для ситуаций, когда среди исходных признаков имеется достаточно информативных признаков. Часто же на практике признаки могут быть индивидуально мало информативными. Кроме того, отбор признаков требует, как правило, значительных затрат времени на перебор комбинаций признаков. Временные затраты на отбор признаков существенно зависят от числа используемых экземпляров.

Выделение искусственных признаков (feature extraction) [10, 11] является узко применимым инструментом, пригодным в основном для распознавания изображений. Создание универсального для всех областей и вместе с тем эффективного метода формирования искусственных признаков представляется нерешенной и практически нереализуемой в настоящее время задачей.

Отбор наиболее значимых экземпляров (instance selection) [3–10] является мене широко используемым инструментом на практике по сравнению с отбором информативных признаков, однако он позволяет существенно сократить затраты времени на построение распознающей или диагностической модели, поскольку позволяет исключить из обучающей выборки незначимые для построения модели наблюдения, что еще до начала обучения модели обеспечивает существенное повышение ее обобщающих свойств по сравнению с использованием всей исходной выборки наблюдений.

Предметом исследования являлись методы отбора информативных экземпляров для построения диагностических и распознающих моделей.

При отборе экземпляров для формировании обучающих выборок приходится оперировать исходным набором признаков, что существенно влияет на затраты времени на отбор экземпляров. Поэтому возникает необходимость сжатия описания экземпляров, но без решения задачи отбора информативных признаков. Одним из возможных путей решения этой задачи является использование хэширования [12–14].

Целью работы являлось сокращение временных затрат на сокращение размерности данных путем создания метода отбора наиболее информативных экземпляров на основе хэширования.

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 : $|F - F'| \le \delta$.

Данная задача может быть разбита на задачу отбора наиболее информативных признаков, задачу отбора наиболее информативных экземпляров и задачу формирования искусственных признаков.

Задача отбора наиболее значимых экземпляров состоит в том, чтобы для выборки $<x, y>$ получить $<x', y'>$: $x' \subseteq x$, $y' \subseteq y$, $S' \le S$, $N'=N$.

Задача отбора наиболее значимых признаков состоит в том, чтобы для выборки $<x, y>$ получить $<x', y'>$: $x' \subseteq x$, $y'=y$, $S'=S$, $N' \le N$.

Задача формирования искусственных признаков для выборки $<x, y>$ состоит в том, чтобы получить $<x', y'>$: $x'=f(x)$, $y'=y$, $S'=S$, $N' \le N$.

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

DOI 10.15588/1607-3274-2020-3-12

2 ОБЗОР ЛИТЕРАТУРЫ

Методы отбора экземпляров можно разделить на детерминированные [3], которые основаны на жестко заданных стратегиях поиска решений и стохастические [5], которые представляют собой стратегии перебора случайно формируемых наборов решений. Обе группы методов отбора экземпляров требуют задания критерия отбора и оценки решений, критерия останова, а также способов формирования новых решений на основе ранее рассмотренных.

Хэширующие преобразования [12–14] отображают экземпляры из *N*-мерного исходного пространства признаков на одномерную ось хэша. По сути хэширование можно рассматривать как разновидность методов формирования искусственных признаков.

Ключевым свойством хэшей для задач распознавания образов является сохранение пространственной топологии исходного пространства признаков на оси хэша – признака. Поэтому среди всех известных методов расчета хэшей целесообразно ограничиться локально чувствительным хэшированием [15–17], которое стремится построить хэширующие преобразования, которые позволят отобразить расстояния между экземплярами в исходном пространстве признаков в расстояния в пространстве.

Многие методы локально чувствительного хэширования [15–17] определяют хэш для *s*-го экземпляра выборки по формуле:

$$x_*^s = \sum_{j=1}^{N} w_j x_j^s .$$

Именно способ расчета весов признаков определяет отличие методов данной группы. В большинстве известных методов веса определяются в результате итеративного перебора случайных отображений из исходного набора признаков на ось хэша, что является весьма затратным по времени. Поэтому для ускорения процесса определения весов необходимо разработать детерминированный метод.

Определив хэши экземпляров, необходимо на их основе создать показатели, позволяющие оценивать значимость экземпляров в пространстве хэша, что позволит исключить необходимость загрузки в память ЭВМ многомерной выборки, а также оперирования многомерными описаниями экземпляров. Это позволит также значительно сократить объем вычислений по сравнению с обработкой выборки во всем исходном пространстве признаков.

3 МАТЕРИАЛЫ И МЕТОДЫ

Для исключения перебора случайных проекций выборки из исходного пространства будем рассматривать 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. Тогда веса признаков определим с учетом числа интервалов, сформированных для каждого признака. Чем меньше нужно интервалов для обеспечения приемлемой точности, тем более ценным является соответствующий признак.

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

Этап Инициализации. Задать исходную выборку <*x*, *y*>, а также максимально допустимое значение ошибки $0 \le \varepsilon << S$. Нормировать значения признаков, отобразив их на интервал [0, 1]:

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

Этап задания предела разбиения признаков. Определить предельное число равных по длине интервалов *Q*, на которые разбиваются диапазоны значений признаков: не более чем *S*, но не менее чем *K*. Эвристически можно рекомендовать принять:

$$Q = S$$
или
$$Q = \max\{\lceil \log_2 S \rceil \ , \min\{S, 2K\}\}.$$

Для всех j =1, 2, ..., N принять число интервалов, на которые разбивается диапазон значений j-го признака, $Q_j = Q$.

Этап разбиения признаков. Для каждого j-го признака, j =1, 2, ..., N выполнить последовательно пункты 1–4:

1. Разбить диапазон значений j-го признака на Q_j равных по длине интервалов.

2. Для каждого q-го интервала значений j-го признака, q=1,2,..., Q_j, определить:

– число экземпляров k-го класса, попавших в него $S_k^{j,q}$, k = 1, 2, ..., K;

– ошибку для каждого q-го интервала значений j-го признака:

$$E_{j,q} = \sum_{s=1}^{S}\sum_{p=s+1}^{S}\{1 \mid -1 \le Qx_j^s - q \le 0, -1 \le Qx_j^p - q \le 0, y^s \ne y^p\} .$$

3. Определить суммарную ошибку для всех интервалов j-го признака:

$$E_j = \sum_{q=1}^{Q_j} E_{j,q} .$$

DOI 10.15588/1607-3274-2020-3-12

4. Если ошибка для всех интервалов *j*-го признака E_j является приемлемой ($E_j \le \varepsilon$), то сократить число интервалов *j*-го признака в два раза, установив $Q_j = \lceil Q_j / 2 \rceil$, и перейти к п. 1, в противном случае – вернуть предыдущее разбиение диапазона *j*-го признака.

Этап ранжирования признаков. Для j =1, 2, ..., N определить ранги признаков r_j в порядке увеличения Q_j: чем больше Q_j, тем меньше ранг *j*-го признака. Для признаков с одинаковыми значениями Q_j эвристически считать более важным (с большим рангом) тот, который индивидуально более значим для выходной переменной, или признак с меньшим номером.

Этап расчета весов. Установить веса признаков:

$$w_j = \left(\max_{j=1,2,...,N} \{Q_j\} \right)^{N-r_j},$$

либо

$$w_j = \left(2^{\left\lceil \log_2 \max_{j=1,2,...,N} \{Q_j\} \right\rceil} \right)^{N-r_j}.$$

В результате выполнения предложенного метода будет получен набор весов, позволяющий определять локально чувствительные хэши экземпляров.

Хэши экземпляров, полученных на основе весов признаков, определенных предложенным выше методом, можно использовать для оценки информативности и отбора экземпляров. Побочным результатом предложенного метода являются веса признаков, которые возможно использовать не только для определения хэшей, но также и для оценивания информативности признаков.

Поскольку хэши, рассчитанные на основе предложенного метода, являются аналогом расстояния между экземплярами, то для них подобно методу потенциалов [18] возможно определить потенциалы, наводимые классами. В свою очередь, показатель значимости экземпляра возможно определить на основе сопоставления потенциалов классов, наводимых на экземпляр. Комбинируя определенным образом показатели индивидуальной значимости экземпляров целесообразно определить показатели групповой значимости экземпляров.

Метод оценивания значимости экземпляров, реализующий описанные выше идеи, возможно представить следующим образом.

Этап инициализации. Задать исходную выборку <*x*, *y*> и нормировать её.

Этап определения весов признаков. Используя предложенный выше метод, найти веса признаков для расчета хэшей.

Этап определения хэшей. Для экземпляров выборки найти хэши $\{x_*^s\}$, s = 1, 2, ..., S, используя полученные веса.

Этап расчета потенциалов. Для каждого *s*-го хэша экземпляра, s = 1, 2, ..., S, определить потенциал, наводимый экземплярами разных классов на данный экземпляр:

$$P^k(x_*^s) = \sum_{p=1}^{S} \left\{ \frac{1}{1+d_*^{s,p}} \middle| p \ne s, y^p = k \right\},$$

$$k = 1, 2, ..., K,\ s = 1, 2, ..., S.$$

С неявно заданным учетом весов признаков определим:

$$d_*^{s,p} = \delta_*^{s,p},$$

$$\delta_*^{s,p} = \left| x_*^s - x_*^p \right|.$$

Без учета весов признаков определим расстояние хэшей следующим образом:

$$d_*^{s,p} = \sum_{j=1}^{N} \left((\delta_*^{s,p} \bmod w_{\arg\min_{m=1,2,...,N} \{r_m | r_m - r_j = 1\}}) - (\delta_*^{s,p} \bmod w_j) \right).$$

Этап определения оценок индивидуальной значимости экземпляров.

Экземпляр в пространстве признаков тем легче отделить от других экземпляров, чем больше минимальная разность потенциалов классов, наводимых на него. То есть, чем меньше минимальная разность потенциалов классов, наводимых на экземпляр, тем сложнее его отделить от экземпляров в пространстве признаков – такой экземпляр будет ценнее по сравнению с другими экземплярами для построения модели, т.к. он вероятно ближе к межклассовой границе.

Соответственно, определим показатель индивидуальной значимости *s*-го экземпляра (стратегия минимума разности потенциалов):

$$I_*^s = \frac{1}{1 + \min\limits_{k=1,...,K} \{ \min\limits_{q=k+1,...,K} \{ | P^k(x_*^s) - P^q(x_*^s) | \} \}}.$$

Данный показатель будет принимать значения от нуля до единицы. Он будет равен единице в случае, когда потенциал одного из классов, наводимый на экземпляр, не будет отличаться от потенциала другого класса, наводимого на этот же экземпляр. Чем сильнее будет удаленность экземпляра от межклассовой границы, тем меньше будет значение данного показателя.

Альтернативно показатель индивидуальной значимости *s*-го экземпляра возможно определить на основе среднего значения разностей потенциалов:

$$I_*^s = \frac{1}{1 + \dfrac{1}{0{,}5K(K-1)} \sum\limits_{k=1}^{K} \sum\limits_{q=k+1}^{K} | P^k(x_*^s) - P^q(x_*^s) |}.$$

DOI 10.15588/1607-3274-2020-3-12

Данный показатель будет принимать значения от нуля до единицы. Он будет равен единице в случае, когда в среднем потенциалы классов, наводимые на экземпляр, не будут отличаться от потенциалов других классов, наводимых на этот же экземпляр. Чем сильнее будет удаленность экземпляра от межклассовой границы, тем меньше будет значение данного показателя.

Аналогичным образом, комбинируя минимальную и среднюю стратегии объединения разностей потенциалов, наводимых разными классами на данный экземпляр, определим показатели:

– на основе стратегии минимума средней разности потенциалов:

$$I_*^S = \frac{1}{1+\min\limits_{k=1,\dots,K}\left\{\frac{1}{(K-k)}\sum\limits_{q=k+1}^{K}|P^k(x_*^S)-P^q(x_*^S)|\right\}};$$

– на основе стратегии среднего минимума разности потенциалов:

$$I_*^S = \frac{1}{1+\frac{1}{K}\sum\limits_{k=1}^{K}\min\limits_{q=k+1,\dots,K}\{|P^k(x_*^S)-P^q(x_*^S)|\}}.$$

Этап определения групповых оценок информативности экземпляров. На основе рассчитанных индивидуальных оценок информативности экземпляров определить оценки групповой информативности экземпляров на основе одной из стратегий:

– на основе стратегии минимума индивидуальных оценок информативности экземпляров в группе *G:*

$$I_*(G) = \min_{s=1,2,\dots,S}\{I_*^s \mid x^s \in G\};$$

– на основе стратегии средней индивидуальной информативности экземпляров в группе *G:*

$$I_*(G) = \frac{1}{S_G}\sum_{s=1}^{S}\{I_*^s \mid x^s \in G\}.$$

Индивидуальные и групповые оценки значимости экземпляров выборки, полученные на основе предложенного метода, могут бы использованы как в детерминированных та и в стохастических методах формирования выборок. Важной особенностью полученных показателей значимости экземпляров является то, что они не требуют построения моделей для определения значимости экземпляров.

Для оценивания информативности признаков на основе весов, полученных при формировании хэшей, пронормируем веса, получив таким образом показатели индивидуальной информативности признаков:

$$I_j = \frac{w_j}{\sum\limits_{i=1}^{N} w_i}$$

либо

$$I_j = \frac{w_j}{\max\limits_{i=1,2,\dots,N}\{w_i\}},$$

либо

$$I_j = \frac{w_j - \min\limits_{i=1,2,\dots,N}\{w_i\}}{\max\limits_{i=1,2,\dots,N}\{w_i\} - \min\limits_{i=1,2,\dots,N}\{w_i\}}.$$

Предложенный показатель индивидуальной информативности признаков будет принимать значения от нуля до единицы: чем меньше будет его значение, тем менее информативным является признак, чем больше будет его значение, тем более индивидуально информативным является признак.

4 ЭКСПЕРИМЕНТЫ

Для изучения свойств предложенных методов они были программно реализованы и исследованы путем решения ряда практических задач распознавания и диагностирования [19–21], характеристики которых приведены в табл. 1. Здесь размерность выборки $n = NS$.

Для каждой практической задачи проводились эксперименты по сокращению размерности выборки до достижения заданной приемлемой точности – отдельно на основе отбора экземпляров, отдельно на основе отбора признаков, а также совместно на основе отбора экземпляров и признаков. Модели строились на основе многослойной нейронной сети прямого распространения сигнала. Число слоев последовательно менялось от одного до трех. Число узлов входного слоя задавалось равным числу используемых признаков, на последнем слое располагался один выходной нейрон (задачи рассматривались как бинарные). Число нейронов скрытого слоя подбиралось в автоматическом режиме там, чтобы при минимальном числе нейронов обеспечить приемлемую точность.

5 РЕЗУЛЬТАТЫ

Результаты проведенных экспериментов приведены в табл. 2.

Таблица 1 – Характеристики практических задач

Задача	Описание	*N*	*S*	*n*	*K*
Iris	Классификация ирисов Фишера [19]	4	150	600	2
Plant	Классификация сельскохозяйственных растений по данным дистанционного зондирования [20]	55	248	13640	2
Bronch	Дифференциальная диагностика хронического обструктивного бронхита [21]	28	205	5740	2
Vehicle	Распознавание типа автотранспортного средства по изображению [22]	26	3992	103792	2

DOI 10.15588/1607-3274-2020-3-12

Таблица 2 – Результаты экспериментов

Задача	Тип эксперимента	S/S'	N/N'	n/n'	t/t'	M/M'
Iris	Отбор экземпляров	1,88	1,00	1,88	1,76	1,44
Plant	Отбор экземпляров	3,40	1,00	3,44	2,97	2,36
Plant	Отбор признаков	1,00	6,11	6,11	4,58	5,33
Plant	Отбор экземпляров и признаков	3,31	6,11	18,19	7,82	11,71
Bronch	Отбор экземпляров	2,01	1,00	2,03	1,57	1,51
Bronch	Отбор признаков	1,00	1,56	1,65	1,15	1,56
Bronch	Отбор экземпляров и признаков	1,92	1,47	2,82	2,78	2,04
Vehicle	Отбор экземпляров	2,18	1,00	2,18	1,66	1,59
Vehicle	Отбор признаков	1,00	1,73	1,73	1,35	1,40
Vehicle	Отбор экземпляров и признаков	2,16	1,53	3,12	2,90	2,38

6 ОБСУЖДЕНИЕ

Как видно из табл. 2, выборки, редуцированные с помощью предложенных методов, позволяют достигать приемлемой точности при существенном сокращении числа экземпляров и числа признаков, а сами хэши могут использоваться как замена или дополнение исходного или сокращенного набора признаков при построении диагностических и распознающих моделей. Использование редуцированных выборок позволяет сократить затраты времени и памяти на построение моделей, а также повышает их обобщающие свойства по сравнению с моделями, обучаемыми на основе исходных выборок.

По сравнению с методами локально чувствительного хэширования [15–17] предложенный метод определения весов не требует итеративного перебора преобразований и является детерминированным.

По сравнению со стохастическими методами формирования выборок [5, 8] предложенный метод расчета показателей информативности экземпляров не требует построения моделей, а также является детерминированным. По сравнению с детерминированными методами [4] предложенный метод позволяет в автоматическом режиме оценить значимость экземпляров без участия человека, а также с учетом топологии классов.

ВЫВОДЫ

Решена актуальная задача создания метода отбора наиболее информативных экземпляров на основе хэширования для уменьшения временных затрат на сокращение размерности данных.

Научная новизна полученных результатов состоит в том, что:

– предложен метод расчета весов для определения хэшей экземпляров, который детерминированным способом определяет веса признаков на основе их рангов, которые, в свою очередь, определяет с учетом числа равных разбиений диапазонов признаков, минимально достаточного для выделения кластеров на оси признака с приемлемой точностью. Это позволяет исключить необходимость итеративного перебора различных комбинаций признаков, определения случайных проекций признаков, а также решение итеративных оптимизационных задач поиска наилучшей проекции признаков, что существенно сокращает затраты времени на расчет весов, при этом обеспечивая локальную чувствительность хэша. Полученные хэши возможно использовать как для отбора экземпляров, так и для отбора признаков;

– предложен метод определения индивидуальной и групповой значимости экземпляров выборки, в котором использует как меру сходства расстояния между хэшами экземпляров и по аналогии с методом потенциалов находит потенциалы, наводимые классами на каждый экземпляр, а на их основе определяет показатели значимости экземпляров, исходя из того, что экземпляр в пространстве признаков тем информативнее, чем меньше минимальная разность потенциалов классов, наводимых на экземпляр;

– предложен метод определения оценок информативности признаков, который на основе нормирования весов, полученных при формировании хэшей, определяет показатели информативности признаков, отдавая предпочтение признакам с меньшим числом разбиений.

Практическая ценность полученных результатов состоит в том, что проведено экспериментальное исследование, подтвердившее работоспособность предложенных методов 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 предложенных методов на более широком классе задач, рассмотреть применимость предложенного метода для задач с вещественным выходом.

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

DOI 10.15588/1607-3274-2020-3-12

и «Virtual Master Cooperation Data Science» (VIMACS) Немецкой службы академических обменов DAAD.

ЛИТЕРАТУРА / ЛІТЕРАТУРА

1. Jensen R. Computational intelligence and feature selection: rough and fuzzy approaches / R. Jensen, Q. Shen. – Hoboken : John Wiley & Sons, 2008. – 300 p.
2. 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
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. Chaudhuri A. Survey sampling theory and methods / A. Chaudhuri, H. Stenger. – New York : Chapman & Hall, 2005. – 416 p. DOI: 10.1201/9781420028638
5. Subbotin S.A. Methods of sampling based on exhaustive and evolutionary search / S. A. Subbotin // Automatic Control and Computer Sciences. – 2013. – Vol. 47, No. 3. – P. 113–121. DOI: 10.3103/s0146411613030073
6. Lavrakas P.J. Encyclopedia of survey research methods / P. J. Lavrakas. – Thousand Oaks : Sage Publications, 2008. – Vol. 1–2. – 968 p. DOI: 10.4135/9781412963947.n159
7. Subbotin S.A. The sample properties evaluation for pattern recognition and intelligent diagnosis / S. A. Subbotin // Digital Technologies : 10th International Conference, Zilina, 9–11 July 2014 : proceedings. – Los Alamitos: IEEE, 2014. – P. 332–343. DOI: 10.1109/dt.2014.6868734
8. 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
9. Subbotin S. The Sample and Instance Selection for Data Dimensionality Reduction / S. Subbotin, A. Oliinyk // Recent Advances in Systems, Control and Information Technology. / Eds.: R. Szewczyk, M. Kaliczyńska. – Cham : Springer, 2017. – P. 97–103. DOI: 10.1007/978-3-319-48923-0_13
10. Elavarasan N. A Survey on Feature Extraction Techniques / N. Elavarasan, K. Mani // International Journal of Innovative Research in Computer and Communication Engineering. – 2015. – Vol. 3, Issue 1. – P. 52–55. DOI: 10.15680/ijircce.2015.0301009 52
11. Alpaydin E. Introduction to Machine Learning / E. Alpaydin. – London : MIT Press, 2014. –640 p.
12. 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
13. Wolfson H. J. Geometric Hashing: An Overview / H. J. Wolfson, I. Rigoutsos // IEEE Computational Science and Engineering. – 1997. – Vol. 4, № 4. – P. 10–21.
14. Fast supervised discrete hashing / [J. Gui, T. Liu, Z. Sunet al.] // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2017. – Vol. 40, № 2. – P. 490–496. DOI: 10.1109/TPAMI.2017.2678475
15. 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
16. 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.
17. 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.
18. Faure A. Perception et reconnaissance des formes / A. Faure. – Paris : Editests, 1985. – 286 p.
19. Fisher Iris dataset [Electronic resource].– Access mode: https://archive.ics.uci.edu/ml/datasets/Iris
20. The plant recognition on remote sensing results by the feed-forward neural networks / [V. Dubrovin, S. Subbotin, S. Morshchavka, D. Piza] // International Journal of Smart Engineering System Design. – 2001. – Vol. 3, No. 4. – P. 251–256.
21. Субботин С. А. Автоматическая система обнаружения и распознавания автотранспортных средств на изображении / С. А. Субботин // Программные продукты и системы. – 2010. – № 1. – С. 114–116.

Received 17.07.2020.
Accepted 22.09.2020.

УДК 004.93

ОЦІНКА ІНФОРМАТИВНОСТІ І ВІДБІР ЕКЗЕМПЛЯРІВ НА ОСНОВІ ХЕШУВАННЯ

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

АНОТАЦІЯ

Актуальність. Для скорочення розмірності даних при побудові діагностичних і розпізнавальних моделей виникає необхідність відбору найбільш інформативних екземплярів, а також відбору найбільш інформативних ознак. Витрати часу на окрему реалізацію даних процедур є високими внаслідок ітеративності і взаємопов'язаності цих процедур.

DOI 10.15588/1607-3274-2020-3-12

Мета. Мета роботи – скорочення витрат часу на скорочення розмірності даних шляхом створення методу відбору найбільш інформативних екземплярів на основі хешування.

Метод. Запропоновано метод розрахунку ваг для визначення хешів екземплярів, який детермінованим способом визначає ваги ознак на основі їх рангів, які, у свою чергу, визначає з урахуванням кількості рівних розбиттів діапазонів ознак, мінімально достатньої для виділення кластерів на вісі ознаки з прийнятною точністю. Це дозволяє виключити необхідність ітеративного перебору різних комбінацій ознак, визначення випадкових проекцій ознак, а також вирішення ітеративних оптимізаційних задач пошуку найкращої проекції ознак, що істотно скорочує витрати часу на розрахунок ваг, при цьому забезпечуючи локальну чутливість хеша. Отримані хеші можливо використовувати як для відбору екземплярів, так і для відбору ознак.

Запропоновано метод визначення індивідуальної та групової значимості екземплярів вибірки, що використовує як міру подібності відстань між хешами зразків і за аналогією з методом потенціалів знаходить потенціали, що наводяться класами на кожен екземпляр, а на їх основі визначає показники значущості екземплярів, виходячи з того, що екземпляр в просторі ознак тим інформативніше, чим менше мінімальна різниця потенціалів класів, що наводяться на екземпляр.

Запропоновано метод визначення оцінок інформативності ознак, який на основі нормування ваг, отриманих при формуванні хешів, визначає показники інформативності ознак, віддаючи перевагу ознаками з меншою кількістю розбиттів.

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

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

КЛЮЧОВІ СЛОВА: екземпляр, ознака, інформативність, хешування, хеш, скорочення розмірності вибірки.

UDC 004.93

EVALUATION OF INFORMATIVITY AND SELECTION OF INSTANCES BASED ON HASHING

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 in the diagnostic and recognition model construction, it becomes necessary to select the most informative instances, as well as to select the most informative features. The time spent on the separate implementation of these procedures is high due to the iterativity and interconnectedness of these procedures.

Objective. The purpose of this work is to reduce the time spent on reducing the data dimensionality by creating a method for selecting the most informative instances based on hashing.

Method. A method for calculating weights for determining the hashes of instances is proposed, which determines the weights of features based on their ranks in a deterministic way, which, in turn, determines, taking into account the number of equal partitions of the ranges of features, the minimum sufficient to distinguish clusters on the axis of the feature with acceptable accuracy. This eliminates the need for iterative enumeration of various combinations of features, determining random projections of features, as well as solving iterative optimization problems of finding the best projection of features, which significantly reduces the time spent on calculating weights, while ensuring the local sensitivity of the hash. The hashes obtained can be used both for the selection of instances and for the selection of features.

A method for determining the individual and group significance of sample instances is proposed, in which it uses the distance between the hashes of the instances as a measure of similarity and, by analogy with the potential method, finds the potentials induced by the classes for each instance, and on their basis determines the indicators of the significance of the instances, based on the fact that the instance in the feature space, the more informative the less the minimum potential difference of the classes induced on the specimen.

A method for determining the estimates of the informativeness of features is proposed, which, on the basis of normalizing the weights obtained during the formation of hashes, determines the indicators of the informativeness of features, giving preference to features with a smaller number of partitions.

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

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

KEYWORDS: instance, attribute, informativeness, hashing, hash, reduction of the sample size.

REFERENCES

1. Jensen R., Shen Q. Computational intelligence and feature selection: rough and fuzzy approaches. Hoboken, John Wiley & Sons, 2008, 300 p.
2. Subbotin S., Oliinyk A. Eds.: Szewczyk R., Kaliczyńska M. 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
3. Subbotin S. The instance and feature selection for neural network based diagnosis of chronic obstructive bronchi-

DOI 10.15588/1607-3274-2020-3-12

tis, *Applications of Computational Intelligence in Biomedical Technology.* Cham, Springer, 2016, pp. 215–228. DOI: 10.1007/978-3-319-19147-8_13
4. Chaudhuri A., Stenger H. Survey sampling theory and methods. New York, Chapman & Hall, 2005, 416 p. DOI: 10.1201/9781420028638
5. Subbotin S.A. Methods of sampling based on exhaustive and evolutionary search, *Automatic Control and Computer Sciences*, 2013, Vol. 47, No. 3, pp. 113–121. DOI: 10.3103/s0146411613030073
6. Lavrakas P.J. Encyclopedia of survey research methods. Thousand Oaks, Sage Publications, 2008, Vol. 1–2, 968 p. DOI: 10.4135/9781412963947.n159
7. Subbotin S.A. The sample properties evaluation for pattern recognition and intelligent diagnosis, *Digital Technologies : 10th International Conference, Zilina, 9–11 July 2014 : proceedings*. Los Alamitos, IEEE, 2014, pp. 332–343. DOI: 10.1109/dt.2014.6868734
8. 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
9. Subbotin S., Oliinyk A. Eds.: R. Szewczyk, M. Kaliczyńska The Sample and Instance Selection for Data Dimensionality Reduction, *Recent Advances in Systems, Control and Information Technology.* Cham, Springer, 2017, pp. 97–103. DOI: 10.1007/978-3-319-48923-0_13
10. Elavarasan N., Mani K. A Survey on Feature Extraction Techniques, *International Journal of Innovative Research in Computer and Communication Engineering*, 2015, Vol. 3, Issue 1, pp. 52–55. DOI: 10.15680/ijircce.2015.0301009 52
11. Alpaydin E. Introduction to Machine Learning. London, MIT Press, 2014, 640 p.
12. 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
13. Wolfson H. J., Rigoutsos I. Geometric Hashing: An Overview, *IEEE Computational Science and Engineering*, 1997, Vol. 4, № 4, pp. 10–21.
14. 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
15. 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
16. 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.
17. 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.
18. Faure A. Perception et reconnaissance des formes. Paris, Editests, 1985, 286 p.
19. Fisher Iris dataset [Electronic resource]. Access mode: https://archive.ics.uci.edu/ml/datasets/Iris
20. Dubrovin V., Subbotin S., Morshchavka S., Piza D. The plant recognition on remote sensing results by the feed-forward neural networks, *International Journal of Smart Engineering System Design*, 2001, Vol. 3, No. 4, pp. 251–256.
21. Subbotin S. A. Avtomaticheskaja sistema obnaruzhenija i raspoznavanija avtotransportnyh sredstv na izobrazhenii, *Programmnye produkty i sistemy*, 2010, No. 1, pp. 114–116.

DOI 10.15588/1607-3274-2020-3-12