УДК 004.93

Субботин С. А.

Д-р техн. наук, профессор, заведующий кафедрой программных средств Запорожского национального технического университета, Запорожье, Украина

КОМПЛЕКСНОЕ СОКРАЩЕНИЕ РАЗМЕРНОСТИ ДАННЫХ ДЛЯ ПОСТРОЕНИЯ ДИАГНОСТИЧЕСКИХ И РАСПОЗНАЮЩИХ МОДЕЛЕЙ ПО ПРЕЦЕДЕНТАМ

Решена задача сокращения размерности данных при построении диагностических и распознающих моделей. Объектом исследования являлся процесс диагностирования, управляемый данными. Предметом исследования являлись методы редукции данных для построения диагностических моделей по прецедентам. Целью работы являлось создание комплекса показателей, позволяющих количественно характеризовать ценность экземпляров и признаков, а также метода сокращения размерности выборок данных для решения задач диагностирования и распознавания. Разработано математическое обеспечение, позволяющее осуществлять формирование выборок и отбор признаков в рамках единого подхода к оценке их значимости. Предложен комплекс показателей, позволяющих количественно характеризовать индивидуальную ценность экземпляров и признаков в локальной окрестности в пространстве признаков. Получили дальнейшее развитие методы переборного поиска для сокращения размерности выборок данных при решении задач диагностирования и распознавания, которые модифицированы путем учета в поисковых операторах предложенных индивидуальных оценок информативности экземпляров и признаков. Предложенные методы и комплекс показателей программно реализованы и исследованы при решении задач сокращения размерности данных. Проведенные эксперименты подтвердили работоспособность разработанного математического обеспечения и позволяют рекомендовать его для использования на практике при решении задач неразрушающего диагностирования и распознавания образов по признакам.

Ключевые слова: выборка, экземпляр, признак, сокращение размерности данных, формирование выборки, отбор признаков, диагностирование.

НОМЕНКЛАТУРА

ЭВМ – электронная вычислительная машина;

δ – радиус окрестности;

Ω – группа признаков, рассматриваемых совместно;

E – ошибка модели;

f – критерий качества;

$F()$ – структура модели;

$I(<x', y'>)$ – показатель качества $<x', y'>$;

$I_*(x^s)$ – показатель информативности s-го экземпляра;

I^*_j – показатель информативности j-го признака,

I^*_Ω – показатель групповой информативности признаков;

j – номер текущего признака;

$I_*(x^s \mid x)$ – показатель индивидуальной информативности экземпляра x^s относительно исходного набора признаков;

$I_*(x^s \mid x \setminus x_j)$ – показатель индивидуальной информативности экземпляра относительно сокращенного набора признаков путем удаления признака x_j из исходного набора признаков;

K – число классов;

m – объем памяти ЭВМ, затраченный на формирование выборки;

n – размерности входа;

N – число входных признаков в исходной выборке;

N'– число входных признаков в редуцированной выборке;

N^* – число удаляемых признаков;

opt – условное обозначение оптимума;

$R(a, b)$ – расстояние между a и b;

s – номер текущего экземпляра;

S – число прецедентов в выборке;

S' – объем редуцированной выборки;

S^* – число удаляемых экземпляров;

S_δ – число экземпляров того же класса, что и класс экземпляра x^s, находящихся в его окрестности, не включая сам экземпляр x^s;

t – время, затраченное на формирование выборки;

w – набор значений параметров модели;

X – исходная выборка;

x – набор входных признаков в исходной выборке;

X' – редуцированная выборка;

x'– набор входных признаков в редуцированной выборке;

x_j – j-й входной признак в исходной выборке;

x^s – s-й экземпляр выборки;

x^s_j – значение j-го входного признака для s-го прецедента;

$x^{\max}_j$ – максимальное значение j-го признака;

$x^{\min}_j$ – минимальное значение j-го признака;

y – выходной признак в исходной выборке;

y' – выходной признак в редуцированной выборке;

y^s – значение выходного признака для s-го прецедента (экземпляра) выборки.

ВВЕДЕНИЕ

Для обеспечения устойчивого функционирования сложного технического оборудования, изделий наукоемкого машиностроения и электронной техники необходимо своевременно осуществлять их диагностирование [1].

Из-за новизны объектов диагностирования, присущей им динамики, нелинейностей и отсутствия или недоста-

DOI 10.15588/1607-3274-2016-4-9

точности экспертных знаний широкое применение на практике для построения автоматизированных систем диагностирования получило диагностирование, управляемое данными [2].

Объектом исследования являлся процесс диагностирования, управляемый данными.

Диагностирование, управляемое данными, предполагает построение диагностических моделей с помощью методов вычислительного интеллекта [3] на основе набора прецедентов.

Построение диагностических и распознающих моделей по прецедентам, как правило, является итеративным процессом, требующим значительных затрат времени для выборок большой размерности. Поэтому для повышения скорости построения диагностических и распознающих моделей необходимо предварительно сокращать размерность данных.

Предметом исследования являлись методы редукции данных для построения диагностических моделей по прецедентам.

Известные методы редукции данных [4–13] исходят из различных точек зрения на важность экземпляров и признаков, что может приводить противоречию между отбором экземпляров и признаков. Поэтому необходимо разработать метод редукции данных, осуществляющий отбор экземпляров и признаков исходя из одного общего представления об их информативности.

Целью данной работы являлось создание комплекса показателей, позволяющих количественно характеризовать ценность экземпляров и признаков, а также метода сокращения размерности выборок данных для решения задач диагностирования и распознавания.

1 ПОСТАНОВКА ЗАДАЧИ

Пусть мы имеем исходную выборку $X = \langle x, y\rangle$ – набор S прецедентов о зависимости $y(x)$, $x = \{x^s\}$, $y=\{y^s\}$, $s = 1, 2, ..., S$, характеризующихся набором N входных признаков $\{x_j\}$, $j = 1, 2, ..., N$, и выходным признаком y. Каждый s-й прецедент представим как $\langle x^s, y^s\rangle$, $x^s=\{x^s_j\}$, где $y^s \in \{1, 2, ..., K\}$, где $K>1$.

Тогда задача синтеза модели зависимости $y(x)$ будет заключаться в определении таких структуры $F()$ и значений параметров w модели, при которых будет удовлетворен критерий качества модели $f(F(), w, \langle x, y\rangle) \to$ opt, где *opt* – условное обозначение оптимума.

В случае, когда исходная выборка имеет большую размерность, перед построением модели необходимо решить задачу выделения обучающей выборки меньшего объема (дано: $X=\langle x, y\rangle$, надо: $X'=\langle x', y'\rangle$, $x' \in \{x^s\}$, $y'=\{y^s | x^s \in x'\}$, $S'=|y'|$, $S'<S$, $f(\langle x', y'\rangle, \langle x, y\rangle) \to$ opt).

2 ЛИТЕРАТУРНЫЙ ОБЗОР

Сокращение размерности выборки данных, как правило, обеспечивается посредством отбора информативных признаков и отбора наиболее значимых экземпляров из исходной выборки.

Известные методы отбора признаков [2, 14], как правило, основаны на переборной стратегии и оперируют некоторым показателем качества, характеризующим индивидуальную или совместную полезность признаков для решения соответствующей задачи.

Для оценки информативности признаков используют широкий спектр показателей [14, 15], которые характеризуют полезность признаков с некоторой точки зрения. В общем случае, не только количественные, но и качественные оценки данных показателей могут не совпадать.

Методы выделения выборок (отбора экземпляров) [4–13], в свою очередь, также основаны на переборной стратегии и оперируют некоторым показателем качества, характеризующим индивидуальную или совместную полезность экземпляров для решения соответствующей задачи.

Для оценки качества сформированной выборки возможно использовать широкий набор предложенных показателей [5, 6], которые на практике в общем случае качественно дают разные оценки ценности экземпляров.

Также в целом следует отметить, что подходы к оценке важности признаков не совпадают в общем случае с подходами к оценке важности экземпляров. Это затрудняет формирование единой стратегии сокращения размерности данных.

Поэтому представляется необходимым разработать показатели и методы, позволяющие давать оценки информативности и производить отбор как экземпляров, так и признаков в рамках единого подхода.

3 МАТЕРИАЛЫ И МЕТОДЫ

Поскольку масштаб значений признаков, которыми характеризуется выборка, может быть существенно различным, это может привести к подавлению одних признаков другими при сравнении экземпляров в процессе формирования выборки.

Для исключения данного негативного эффекта предлагается использовать нормированные расстояния как меру близости (меру подобия) экземпляров:

$$R(x^s, x^p) = R(x^p, x^s) = \sum_{j=1}^{N}\left(\frac{x_j^s - x_j^p}{x_j^{\max} - x_j^{\min}}\right)^2,$$

$$s = 1, 2, ..., S; p = s+1, s+2, ..., S.$$

Под локально влияющими на экземпляр x^s в окрестности радиуса δ будем понимать множество тех экземпляров обучающей выборки, которые удалены от экземпляра x^s не более чем на δ, $0<\delta \le 1$.

Показатель информативности экземпляра относительно внешних границ класса в локальной окрестности радиуса δ определим по формуле:

$$I_O(x^s) = \frac{1}{S_\delta(x^s)}\sum_{p=1}^{S}\left\{R(x^s, x^p)\,\middle|\, R(x^s, x^p) \le \delta, y^s = y^p, p \ne s\right\},$$

где S_δ – число экземпляров того же класса, что и класс экземпляра x^s, находящихся в его окрестности, не включая сам экземпляр x^s:

$$S_\delta(x^s) = \sum_{p=1}^{S}\left\{1 \,|\, R(x^s, x^p) \le \delta, y^s = y^p, p \ne s\right\}.$$

Предложенный показатель I_O будет принимать значения в интервале [0, 1]. Чем больше будет значение показателя I_O, тем ближе экземпляр x^s к внешней границе соответствующего класса в локальной окрестности радиуса δ.

Показатель информативности экземпляра x^s относительно межклассовых границ в локальной окрестности радиуса δ определим по формуле:

$$I_C(x^s) = \frac{1}{1 + \sum_{p=1}^{S} \left\{ R(x^s, x^p) \middle| R(x^s, x^p) \le \delta, y^s = y^p, p \ne s \right\}}.$$

Предложенный показатель I_C будет принимать значения в интервале $[(1+NS)^{-1}, 1]$. Чем больше будет значение показателя I_C, тем ближе экземпляр x^s к межклассовой границе в локальной окрестности радиуса δ.

Показатель информативности экземпляра относительно внутриклассового центра в локальной окрестности радиуса δ определим по формуле:

$$I_B(x^s) = \frac{1}{1 + \min\limits_{p=1,2,\ldots,S} \left\{ R(x^s, x^p) \middle| R(x^s, x^p) \le \delta, y^s \ne y^p, p \ne s \right\}}.$$

Предложенный показатель I_B будет принимать значения в интервале $[(1+N)^{-1}, 1]$. Чем больше будет значение показателя I_B, тем ближе экземпляр x^s к центру своего класса в локальной окрестности радиуса δ.

Комбинированный показатель информативности экземпляра относительно внешних границ класса и межклассовых границ в локальной окрестности радиуса δ определим по формуле:

$$I_{OC}(x^s) = \max\{I_O(x^s), I_C(x^s)\}.$$

Предложенный показатель будет принимать значения в интервале [0, 1]. Чем больше будет значение показателя I_{OC}, тем ближе экземпляр x^s к внешним границам класса и межклассовой границе в локальной окрестности радиуса δ.

Комбинированный показатель информативности экземпляра относительно внешних границ класса и внутриклассового центра в локальной окрестности радиуса δ определим по формуле:

$$I_{OB}(x^s) = \max\{I_O(x^s), I_B(x^s)\}.$$

Предложенный показатель I_{OB} будет принимать значения в интервале [0, 1]. Чем больше будет значение показателя I_{OB}, тем ближе экземпляр x^s к внешним границам класса, к центру своего класса в локальной окрестности радиуса δ.

Комбинированный показатель информативности экземпляра относительно межклассовых границ и внутриклассового центра экземпляра в локальной окрестности радиуса δ определим по формуле:

$$I_{BC}(x^s) = \max\{I_B(x^s), I_C(x^s)\}.$$

Предложенный показатель I_{BC} будет принимать значения в интервале $[(N2S+NS+N+1)^{-1}, 1]$. Чем больше будет значение показателя I_{BC}, тем ближе экземпляр x^s к межклассовой границе и ближе к центру своего класса в локальной окрестности радиуса δ.

Комбинированный показатель информативности экземпляра относительно внешних границ класса, межклассовых границ и внутриклассового центра экземпляра в локальной окрестности радиуса δ:

$$I_{OBC}(x^s) = \max\{I_O(x^s), I_B(x^s), I_C(x^s)\}.$$

Предложенный показатель I_{OBC} будет принимать значения в интервале [0, 1]. Чем больше будет значение показателя I_{OBC}, ем ближе экземпляр x^s к к внешним границам класса, межклассовой границе, ближе к центру своего класса в локальной окрестности радиуса δ.

Предложенный выше комплекс показателей может быть использован не только для отбора экземпляров, но также и для оценки информативности и отбора признаков.

Показатели индивидуальной информативности признаков можно определить по обобщенной формуле:

$$I_j^* = \frac{1}{S} \sum_{s=1}^{S} \left(I_*(x^s \mid x) - I_*(x^s \mid x \setminus x_j) \right)^2,$$

где маркер «*» заменяется обозначением типа соответствующего показателя информативности экземпляров.

Данный показатель будет принимать значения от нуля до единицы. Чем больше будет значение данного показателя, тем сильнее влияние соответствующего признака на качество выборки с точки зрения выбранного типа показателей информативности экземпляров.

Показатели групповой информативности признаков можно определить по обобщенной формуле:

$$I_\Omega^* = \frac{1}{S} \sum_{s=1}^{S} \left(I_*(x^s \mid x) - I_*(x^s \mid x \setminus \Omega) \right)^2,$$

где маркер «*» заменяется обозначением типа соответствующего показателя информативности экземпляров.

Данный показатель будет принимать значения от нуля до единицы. Чем больше будет значение данного показателя, тем сильнее влияние соответствующей группы признаков на качество выборки с точки зрения выбранного типа показателей информативности экземпляров.

Предложенный комплекс показателей может быть использован в методах редукции данных.

Наиболее точным является метод редукции на основе стратегии полного перебора [14]. Данный метод сначала выполняет перебор всех возможных комбинаций экземпляров из исходной выборки. После чего оценивается их качество и выбирается одна комбинация, содержащая наименьшее число экземпляров, обеспечивающее приемлемый уровень качества. Затем выполняется перебор всех возможных комбинаций признаков. После чего оценивается их качество и выбирается одна комбинация признаков, содержащая наименьшее число признаков, обеспечивающее приемлемый уровень качества. Формально данный метод может быть представлен следующим образом.

0. Задать исходную выборку $<x, y>$.

1. Редукция экземпляров.

1.1. Сгенерировать все возможные комбинации экземпляров $\{<x', y'>\}$ как подвыборки $<x, y>$.

1.2. Для каждой комбинации экземпляров $<x', y'>$ оценить выбранный показатель качества $I(<x', y'>)$.

1.3. В качестве итоговой сокращенной выборки принять комбинацию $<x', y'>$, содержащую наименьшее число экземпляров при приемлемом значении показателя качества $I(<x', y'>)$.

2. Редукция признаков.

2.1. Сгенерировать все возможные комбинации признаков для сокращенной выборки $<x', y'>$.

2.2. Для каждой комбинации признаков по сокращенной выборке $<x', y'>$ оценить выбранный показатель качества $I(<x', y'>)$.

2.3. Оставить в сокращенной выборке $<x', y'>$ только те признаки, которые входят в комбинацию, содержащую наименьшее число признаков при приемлемом значении показателя качества.

Данный метод потребует перебора 2^S–1 комбинаций экземпляров на этапе редукции экземпляров и 2^N–1 комбинаций признаков на этапе редукции признаков. Очевидно, что такой метод является самым медленным и вычислительно затратным. Его практическая применимость весьма ограничена.

Для устранения недостатков полного перебора возможно, оценив индивидуальную информативность признаков и экземпляров, последовательно удалять из исходной выборки некоторое подмножество наименее индивидуально информативных экземпляров и признаков, строя каждый раз по редуцированной выборки модель и оценивая показатель качества, до тех пор, пока признаков больше двух, экземпляров не меньше, чем классов, а точность модели является приемлемой.

Быстрый метод редукции данных, реализующий данные идеи, представим следующим образом.

1. Принять в качестве текущей выборки $<x', y'>$ исходную выборку $<x, y>$. Задать число удаляемых экземпляров S^* и число удаляемых признаков N^*.

2. Оценить индивидуальную информативность экземпляров и индивидуальную информативность признаков в выборке.

3. Если $S'>S^*$, то удалить S^* наименее индивидуально информативных экземпляров из текущей выборки. Если $N'>N^*$, то удалить N^* наименее информативных признаков из текущей выборки.

4. Построить распознающую модель по редуцированной выборке $<x', y'>$.

5. Оценить ошибку построенной модели E по исходной выборке $<x, y>$. Например, в качестве критерия ошибки можно использовать среднюю ошибку:

$$E = \frac{1}{S}\sum_{s=1}^{S}\{1 \mid y^s{}_{\text{р.}} \neq y^s\}.$$

6. Если ошибка E приемлемая, то принять в качестве текущей выборки редуцированную выборку и перейти к этапу 3; в противном случае – вернуть в качестве результата текущую выборку $<x', y'>$.

Такой метод при выборе достаточно больших значений S^* и N^* будет обеспечивать очень быстрое сокращение размерности выборки, однако будет достигать этого за счет потери информации, что повлечет уменьшение точности. Поскольку данный метод требует построения модели, то его эффективность также будет зависеть от эффективности используемого метода построения модели.

Поскольку одновременная редукция экземпляров и признаков может в ряде практических приложений слишком быстро приводить к потере информации и, как следствие, точности синтезируемой модели, представляется целесообразным для таких случаев последовательно редуцировать данные, синтезируя модель для контроля потери информации, и тем самым обеспечивая более тщательный контроль редукции данных.

Последовательный метод редукции данных, реализующий данные идеи, представим следующим образом.

1.Принять в качестве текущей выборки $<x', y'>$ исходную выборку $<x, y>$. Задать число удаляемых экземпляров S^* и число удаляемых признаков N^*.

2. Оценить индивидуальную информативность экземпляров в исходной выборке.

3. Если $S'>S^*$, то удалить S^* наименее индивидуально информативных экземпляров из текущей выборки $<x', y'>$.

4. Построить модель на основе текущей выборки и оценить ошибку модели по исходной выборке.

5. Если ошибка модели приемлемая и $S'>S^*$, то перейти к этапу 3; в противном случае – вернуть предыдущий набор экземпляров $<x', y'>$.

6. Оценить индивидуальную информативность признаков по редуцированной текущей выборке $<x', y'>$.

7. Если $N'>N^*$, то удалить N^* наименее информативных признаков из текущей выборки $<x', y'>$.

8. Построить распознающую модель на основе текущей выборки $<x', y'>$ и оценить ошибку модели E по исходной выборке $<x, y>$.

9. Если ошибка модели E приемлемая и $N'>N^*$, то перейти к этапу 7; в противном случае – вернуть предыдущий набор признаков $<x', y'>$.

Такой метод при выборе достаточно больших значений S^* и N^* будет обеспечивать быстрое сокращение размерности выборки, однако оно будет медленнее, чем у предыдущего метода. При этом данный метод за счет большего контроля ошибки сможет потенциально терять меньше информации, обеспечивая более тщательный отбор признаков. Тем не менее, поскольку данный метод требует построения модели, то его эффективность также будет зависеть от эффективности используемого метода построения модели.

Для комплекса предложенных показателей и методов сокращения размерности выборок данных существенным параметром является выбор размера окрестности δ.

Очевидно, что при большом значении δ в локальную окрестность экземпляра будет попадать большое число экземпляров, что сделает трудоемким расчет показателей информативности, однако позволит сопоставить соответствующий экземпляр с большим числом других экземпляров, обеспечивая более точную оценку важности экземпляра.

При малом значении δ в локальную окрестность экземпляра может не попасть ни одного экземпляра, либо попасть очень небольшое число экземпляров. Это не позволит обеспечить приемлемую точность оценивания важности экземпляров.

Предположим, что экземпляры равномерно распределены в пространстве признаков. Тогда в окрестности радиуса δ каждого экземпляра окажется порядка SV экземпляров, где $V = \pi^{0,5N}\delta^N / \Gamma(0,5N+1)$, где Γ – гамма-функция.

Очевидно, что $S>>SV$. Следовательно, $V<<1$. Зафиксировав N, получим $0<\delta<<\pi^{-0,5}\sqrt[N]{\Gamma(0,5N+1)}\le 1$.

4 ЭКСПЕРИМЕНТЫ

Для проверки работоспособности предложенного комплекса показателей информативности экземпляров и признаков, а также методов редукции данных они были программно реализованы и исследованы на наборе синтетических выборок данных.

Каждая выборка содержала экземпляры двух классов, характеризовавшиеся наборами признаков. Значения одной части признаков генерировались случайным образом. Значения другой части признаков определялись как комбинации значений некоторых признаков первой части. Характеристики синтетических выборок данных приведены в табл. 1.

Таблица 1 – Характеристики синтетических выборок данных

№ выборки	N	S	n
1	10	20	200
2	20	20	400
3	20	20	800
4	40	40	1600
5	100	100	10000
6	500	500	250000
7	1000	1000	1000000

В первой серии экспериментов осуществлялось сравнение методов по затратам ресурсов и достигнутой точности (ошибке) моделей, синтезированных по редуцированным выборкам.

Во второй серии экспериментов исследовался вопрос выбора значения δ.

5 РЕЗУЛЬТАТЫ

В табл. 2 представлены результаты сравнения затрат ресурсов предложенных методов редукции данных при решении синтетических задач редукции данных.

Как видно из табл. 2, метод полного перебора является наиболее затратным как по используемым вычислительным ресурсам, так и по ресурсам памяти. Метод быстрой редукции является наиболее эффективным с точки зрения затрат ресурсов, а метод последовательной редукции требует несколько больше вычислительных ресурсов и ресурсов памяти, по сравнению с быстрым методом редукции.

В табл. 3 представлены результаты сравнения полученной ошибки моделей, построенных на основе редуцированных данных, полученных с помощью предложенных методов.

Таблица 3 – Сравнительная характеристика методов редукции данных по ошибке модели E

№ выборки	Метод полного перебора	Метод быстрой редукции	Последовательный метод редукции
1	0	0	0
2	0	0,05	0
3		0,05	0,05
4		0,08	0,05
5		0,07	0,05
6		0,06	0,04
7		0,06	0,06

Как видно из табл. 3, метод полного перебора обеспечивает наибольшую точность, однако из-за ограничений по ресурсам не имеет широкой практической применимости. Метод быстрой редукции в среднем обеспечивает несколько большую ошибку по сравнению с последовательным методом редукции.

На рис. 1 представлены результаты проведенных экспериментов по подбору значения δ для разных значений числа используемых признаков N.

Как видно из рис. 1, с увеличением числа признаков N также возрастает практический порог для задания значения δ. При этом даже для небольших N вполне приемлемым является значение δ=0,5.

На рис. 2 представлен график усредненной зависимости ошибки полученных моделей E от величины δ.

Как видно из рис. 2, при относительно малых значениях δ наблюдается наибольшая средняя ошибка E, которая существенно сокращается с увеличением значения δ. Наиболее сильное падение ошибки наблюдается в диапазоне значений δ<0,4.

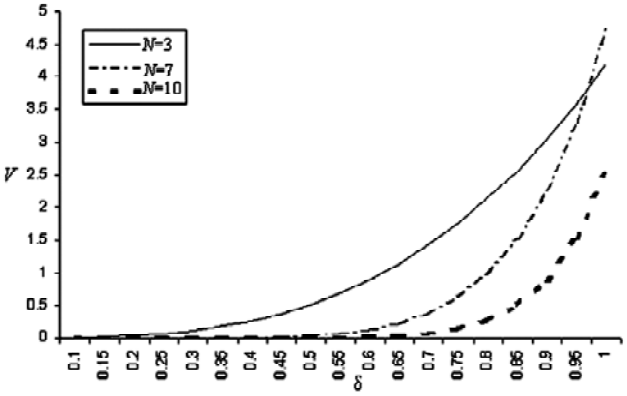

Рисунок 1 – Графики зависимостей V от δ для разных значений N

Таблица 2 – Сравнительная характеристика методов редукции данных по затратам ресурсов

№ выборки	Метод полного перебора		Метод быстрой редукции		Последовательный метод редукции	
	t, с	m, Мб	t, с	m, Мб	t, с	m, Мб
1	0,5249	40,102	0,0111	0,004	0,0212	0,009
2	1,0486	160,039	0,0161	0,008	0,0336	0,018
3			0,0433	0,015	0,0845	0,036
4			0,0642	0,031	0,1344	0,072
5			0,4006	0,191	0,8402	0,448
6			10,0032	4,802	21,0021	11,208
7			40,0109	19,109	84,1093	44,801

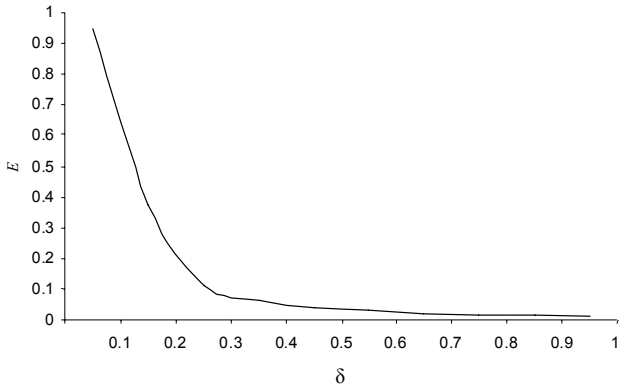

Рисунок 2 – График усредненной зависимости ошибки полученных моделей E от δ

Разработанные методы осуществляют редукцию выборки данных в рамках единого подхода к оценке индивидуальной информативности признаков и экземпляров, что позволяет сократить объем вычислений по сравнению с традиционным подходом, когда задачи отбора экземпляров и признаков решаются раздельно в рамках разных парадигм. Это позволяет существенно снизить затраты как вычислительных ресурсов, так и ресурсов памяти.

6 ОБСУЖДЕНИЕ

Метод полного перебора позволяет обеспечить наиболее точный результат решения задачи отбора информативных признаков и экземпляров минимального объема. Однако с практической точки зрения для большинства приложений данный метод оказывается не применимым.

Быстрый и последовательный методы редукции позволяют решать задачу сокращения размерности данных за приемлемое с практической точки зрения время. При этом методы обеспечивают требуемую точность при правильном подборе величины δ.

Результаты проведенных экспериментов позволяют рекомендовать задавать на практике значение δ порядка 0,5. При этом для больших выборок и малом числе признаков значение δ можно сокращать до 0,3–0,4. Для малых выборок, описываемых большим числом признаков значение δ можно задавать порядка 0,7–0,8.

Отметим также, что предложенные методы автоматически определяют размер формируемой выборки, не требуя участия человека.

ВЫВОДЫ

С целью решения задачи сокращения размерности данных при построении диагностических и распознающих моделей разработано математическое обеспечение, позволяющее осуществлять формирование выборок и отбор признаков в рамках единого подхода к оценке их значимости.

Научная новизна полученных результатов состоит в том, что:

– впервые предложен комплекс показателей, позволяющих количественно характеризовать индивидуальную ценность экземпляров и признаков в локальной окрестности в пространстве признаков;

– получили дальнейшее развитие методы переборного поиска, которые модифицированы путем учета в поисковых операторах предложенных индивидуальных оценок информативности экземпляров и признаков а также метода сокращения размерности выборок данных для решения задач диагностирования и распознавания.

Практическая значимость полученных результатов заключается в том, что предложенные методы и комлекс показателей программно реализованы и исследованы при задач сокращения размерности данных. Проведенные эксперименты подтвердили работоспособность разработанного математического обеспечения и позволяют рекомендовать его для использования на практике.

Перспективы дальнейших исследований состоят в том, чтобы определить более быстрые способы расчета предложенных показателей информативности экземпляров, изучить их взаимосвязь с качеством синтезируемых моделей, исследовать предложенное математическое обеспечение на более широком классе практических задач диагностирования и распознавания образов.

БЛАГОДАРНОСТИ

Работа выполнена в рамках госбюджетной научно-исследовательской темы Запорожского национального технического университета «Методы и средства вычислительного интеллекта и параллельного компьютинга для обработки больших объемов данных в системах диагностирования» (номер гос. регистрации 0116U007419) при частичной поддержке международнго проекта «Центры передового опыта для молодых ученых» Европейского Союза (№ 544137-TEMPUS-1-2013-1-SK-TEMPUS-JPHES) при частичной поддержке международнго проекта «Центры передового опыта для молодых ученых» Европейского Союза (№ 544137-TEMPUS-1-2013-1-SK-TEMPUS-JPHES).

СПИСОК ЛИТЕРАТУРЫ

1. Интеллектуальные информационные технологии проектирования автоматизированных систем диагностирования и распознавания образов : монография / С. А. Субботин, Ан. А. Олейник, Е. А. Гофман, С. А. Зайцев, Ал. А. Олейник ; под ред. С. А. Субботина. – Харьков : Компания СМИТ, 2012. – 318 с.
2. Russell E. L. Data-driven diagnosis Data-driven Methods for Fault Detection and Diagnosis in Chemical Processes / E. L. Russell, L. H. Chiang,R. D. Braatz. – London : Springer-Verlag, 2000. – 192 p. DOI: 10.1007/978-1-4471-0409-4
3. Computational intelligence: a methodological introduction / [R. Kruse, C. Borgelt, F.Klawonn et. al.]. – London : Springer-Verlag, 2013. – 488 p. DOI: 10.1007/978-1-4471-5013-8_1
4. Олешко Д. Н. Построение качественной обучающей выборки для прогнозирующих нейросетевых моделей / Д. Н. Олешко, В. А. Крисилов, А. А. Блажко // Штучний інтелект. – 2004. – № 3. – С. 567–573.
5. Subbotin S. A. The training set quality measures for neural network learning / S. A. Subbotin // Optical memory and neural networks (information optics). – 2010. – Vol. 19, № 2. – P. 126–139. DOI: 10.3103/s1060992x10020037
6. Субботин С. А. Критерии индивидуальной информативности и методы отбора экземпляров для построения диагностических и распознающих моделей / С. А. Субботин // Біоніка інтелекту. – 2010. – № 1. – С. 38–42.
7. Encyclopedia of survey research methods / ed. P. J. Lavrakas. – Thousand Oaks: Sage Publications, 2008. – Vol. 1–2. – 968 p. DOI: 10.1108/09504121011011879
8. Hansen M. H. Sample survey methods and theory / M. H. Hansen, W. N. Hurtz, W. G. Madow. – Vol. 1 : Methods and applications. – New York : John Wiley & Sons, 1953. – 638 p.
9. Кокрен У. Методы выборочного исследования / У. Кокрен ; пер. с англ. И. М. Сонина ; под ред. А. Г. Волкова, Н. К. Дружинина. – М. : Статистика, 1976. – 440 с.

10. Multivariate analysis, design of experiments, and survey sampling / ed. S. Ghosh. – New York : Marcel Dekker Inc., 1999. – 698 p.
11. Smith G. A deterministic approach to partitioning neural network training data for the classification problem : dissertation ... doctor of philosophy in business / Smith Gregory. – Blacksburg : Virginia Polytechnic Institute & State University, 2006. – 110 p.
12. Bernard H. R. Social research methods: qualitative and quantative approaches / H. R. Bernard. – Thousand Oaks: Sage Publications, 2006. – 784 p.
13. Chaudhuri A. Survey sampling theory and methods / A. Chaudhuri, H. Stenger. – New York : Chapman & Hall, 2005. – 416 p.
14. Субботін С. О. Інтелектуальні системи : навч. посіб. / С. О. Субботін, А. О. Олійник; під заг. ред. проф. С. О. Субботіна. – Запоріжжя : ЗНТУ, 2014. – 218 с.
15. Биргер И. А. Техническая диагностика / И. А. Биргер. – М. : Машиностроение, 1978. – 240 с.

Статья поступила в редакцию 29.06.2016.

Субботін С. О.

Д-р техн. наук, професор, завідувач кафедри програмних засобів Запорізького національного технічного університету, Запоріжжя, Україна

КОМПЛЕКСНЕ СКОРОЧЕННЯ РОЗМІРНОСТІ ДАНИХ ДЛЯ ПОБУДОВИ ДІАГНОСТИЧНИХ І РОЗПІЗНАВАЛЬНИХ МОДЕЛЕЙ ЗА ПРЕЦЕДЕНТАМИ

Вирішено завдання скорочення розмірності даних при побудові діагностичних і розпізнавальних моделей. Об'єктом дослідження є процес діагностування, керований даними. Предметом дослідження є методи редукції даних для побудови діагностичних моделей за прецедентами. Метою роботи є створення комплексу показників, що дозволяють кількісно характеризувати цінність екземплярів і ознак, а також методу скорочення розмірності вибірок даних для вирішення завдань діагностування та розпізнавання. Розроблено математичне забезпечення, що дозволяє здійснювати формування вибірок та відбір ознак в рамках єдиного підходу щодо оцінки їх значимості. Запропоновано комплекс показників, що дозволяють кількісно характеризувати індивідуальну цінність екземплярів і ознак у локальній околиці в просторі ознак. Отримали подальший розвиток методи переборного пошуку для скорочення розмірності вибірок даних при вирішенні завдань діагностування та розпізнавання, які модифіковані шляхом урахування у пошукових операторах запропонованих індивідуальних оцінок інформативності екземплярів і ознак. Запропоновані методи і комплекс показників програмно реалізовані і досліджені шляхом вирішення завдань скорочення розмірності даних. Проведені експерименти підтвердили працездатність розробленого математичного забезпечення і дозволяють рекомендувати його для використання на практиці при вирішенні завдань неруйнівного діагностування та розпізнавання образів за ознаками.

Ключові слова: вибірка, екземпляр, ознака, скорочення розмірності даних, формування вибірки, відбір ознак, діагностування.

Subbotin S. A.

Dr.Sc., Professor, Head of the Department of Software Tools, Zaporizhzhya National Technical University, Zaporizhzhya, Ukraine

THE COMPLEX DATA DIMENSIONALITY REDUCTION FOR DIAGNOSTIC AND RECOGNITION MODEL BUILDING ON PRECEDENTS

The problem of data dimensionality reduction for diagnostic and recognizing model construction is solved. The object of study is the process of data-driven diagnosis. The subject of study is the data reduction methods for diagnostic model construction on precedents. The purpose of work is to create a set of indicators to quantify the importance of instances and features, as well as a method of data sample dimensionality reduction in the diagnosis and pattern recognition and problem solving. The mathematical support for the sample formation and feature selection is developed on the base of common approach to the assessment of their significance. The set of indicators is proposed to quantify the individual informativity of instances and features in the local neighborhood in the feature space. The exhaustive search methods for data sample dimensionality reduction in the solution of recognition and diagnosis problems have been further developed. They are modified by taking into account of the offered individual estimations of informativity of instances and features in the search operators. The proposed methods and indicator complex are implemented as software and studied in the solution of data dimensionality reduction problems. The conducted experiments confirmed the efficiency of the developed mathematical tools and allow to recommend them for use in practice for solving the problems of non-destructive diagnosis and pattern recognition on features.

Keywords: sample, instance, feature, data dimensionality reduction, sampling, feature selection, diagnosis.

REFERENCES

1. Subbotin S. A., Olejnik An. A., Gofman E. A., Zajcev S. A., Olejnik Al. A.; pod red. Subbotina S. A. Intellektual'nye informacionnye tehnologii proektirovanija avtomatizirovannyh sistem diagnostirovanija i raspoznavanija obrazov : monografija. Har'kov, Kompanija SMIT, 2012, 318 p.
2. Russell E. L., Chiang L. H., Braatz R. D. Data-driven diagnosis Data-driven Methods for Fault Detection and Diagnosis in Chemical Processes. London, Springer-Verlag, 2000, 192 p. DOI: 10.1007/978-1-4471-0409-4
3. Kruse R., Borgelt C., Klawonn F. et. al. Computational intelligence: a methodological introduction. London, Springer-Verlag, 2013, 488 p. DOI: 10.1007/978-1-4471-5013-8_1
4. Oleshko D. N., Krisilov V. A., Blazhko A. A. Postroenie kachestvennoj obuchayushhej vyborki dlya prognoziruyushhix nejrosetevyx modelej, *Shtuchnyj intelekt*, 2004, No. 3, pp. 567–573.
5. Subbotin S. A. The training set quality measures for neural network learning, Optical memory and neural networks (information optics), 2010, Vol. 19, No. 2, pp. 126–139. DOI: 10.3103/s1060992x10020037
6. Subbotin S. A. Kriterii individual'noj informativnosti i metody otbora e'kzemplyarov dlya postroeniya diagnosticheskix i raspoznayushhix modelej, *Bionika intelektu*, 2010, No. 1, pp. 38–42.
7. Encyclopedia of survey research methods / ed. P. J. Lavrakas. Thousand Oaks, Sage Publications, 2008, Vol. 1–2, 968 p. DOI: 10.1108/09504121011011879
8. Hansen M. H., Hurtz W. N., Madow W. G. Sample survey methods and theory, Vol. 1, Methods and applications. New York, John Wiley & Sons, 1953, 638 p.
9. Kokren U., per. s angl. Sonina I. M.; pod red. Volkova A. G., Druzhinina N. K. Metody vyborochnogo issledovaniya. Moscow, Statistika, 1976, 440 p.
10. Ghosh S. ed. Multivariate analysis, design of experiments, and survey sampling. New York, Marcel Dekker Inc., 1999, 698 p.
11. Smith G. A deterministic approach to partitioning neural network training data for the classification problem : dissertation ... doctor of philosophy in business. Blacksburg, Virginia Polytechnic Institute & State University, 2006, 110 p.
12. Bernard H. R. Social research methods: qualitative and quantative approaches. Thousand Oaks, Sage Publications, 2006, 784 p.
13. Chaudhuri A., Stenger H. Survey sampling theory and methods. New York, Chapman & Hall, 2005, 416 p.
14. Subbotin S. O., Oliynyk A. O.; pid zag. red. prof. S. O. Subbotina Intelektual'ni systemy : navch. posib. Zaporizhzhya, ZNTU, 2014, 218 p.
15. Birger I. A. Tekhnicheskaya diagnostika. Moscow, Mashinostroenie, 1978, 240 p.