УДК 004.93

Субботин С.А.

Канд. техн. наук, доцент, Запорожский национальный технический университет, Украина, E-mail: subbotin.csit@gmail.com.

ФОРМИРОВАНИЕ И РЕДУКЦИЯ ВЫБОРОК ДЛЯ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

Решена задача формирования и редукции выборок для интеллектуального анализа данных. Предложен метод формирования и редукции выборок, который обеспечивает сохранение в сформированной подвыборке важнейших топологических свойств исходной выборки, не требуя при этом загрузки в память ЭВМ исходной выборки, а также многочисленных проходов исходной выборки, что позволяет сократить объем выборки и уменьшить требования к ресурсам ЭВМ.

Ключевые слова: выборка, отбор экземпляров, редукция данных, интеллектуальный анализ данных, сокращение размерности данных.

ВВЕДЕНИЕ

При решении задач интеллектуального анализа данных [1], в частности, задач построения моделей принятия решений на основе нейронных и нейро-нечетких сетей, а также деревьев решений [2–4] в различных прикладных областях нередко приходится оперировать выборками данных большого объема. Это влечет за собой существенные затраты времени на обработку данных, а также требует наличия значительных объемов оперативной и дисковой памяти ЭВМ. Поэтому актуальной задачей является сокращение размерности выборок данных [1–5].

Традиционным и наиболее широко применяемым подходом при решении данной задачи является использование методов отбора информативных признаков [1–5] (удаляют из исходного набора наименее информативные признаки) и методов конструирования признаков [5, 6] (заменяют исходный набор признаков рассчитанным на его основе набором искусственных признаков меньшего размера). Однако, если изначально заданный набор признаков не является избыточным, либо объем выборки (число экземпляров в ней) чрезвычайно велик для представления и обработки в памяти ЭВМ, применение этих методов оказывается чрезвычайно затруднительным, а результаты их работы либо приводят к потере существенной для дальнейшего анализа информации, либо не позволяют сохранить исходную интерпретабельность данных.

Другим, существенно реже используемым на практике, подходом при решении данной задачи является сокращение объема выборки. Как правило, это реализуется посредством извлечения случайных подвыборок из исходной выборки [7–9], что может приводить к формированию нерепрезентативных в топологическом смысле выборок вследствие невключения в них редко встречающихся экземпляров на границах классов, представленных в исходной выборке.

В [10–13] автором предложены переборные и эволюционные методы формирования выборок, а также модель (комплекс критериев) качества выборки, которые позволяют обеспечить формирование из исходной выборки подвыборок меньшего объема, обладающих в системе используемых критериев наилучшими свойствами. Однако для выборок очень большого объема применение данных методов и модели оказывается весьма затратным как с вычислительной точки зрения, так и с точки зрения ресурсов оперативной и дисковой памяти.

Целью данной работы является создание метода формирования и редукции выборок, позволяющего обрабатывать исходные выборки большого объема.

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 $\{x_j\}$, $j = 1, 2, ..., N$, где j – номер признака, и выходным признаком y. Каждый s-й прецедент представим как $\langle x^s, y^s \rangle$, $x^s = \{x^s_j\}$, где x^s_j – значение j-го входного, а y^s – значение выходного признака для s-го прецедента (экземпляра) выборки, $y^s \in \{1, 2, ..., K\}$, где K – число классов, $K>1$.

Тогда задача сокращения объема выборки может быть представлена как задача формирования (выделения) из исходной выборки $X = \langle x, y \rangle$ подвыборки X^*, $X^* \subset X$, меньшего объема $S^* < S$, обладающей наиболее важными свойствами исходной выборки.

Поскольку для задач интеллектуального анализа данных, связанных с автоматизацией поддержки принятия решений, наиболее важным является сохранение топологии классов, то формируемая подвыборка должна обеспечивать сохранение экземпляров исходной выборки, находящихся на границах классов.

2. МЕТОД ФОРМИРОВАНИЯ И РЕДУКЦИИ ВЫБОРОК БОЛЬШОГО ОБЪЕМА

Для обнаружения экземпляров, находящихся на границах классов, в общем случае необходимо решить задачу кластер-анализа, что требует определения расстояний между всеми экземплярами выборки. Это, в свою очередь, требует либо загрузки всей выборки в память ЭВМ (что не всегда возможно из-за ограниченного объема оперативной памяти), либо многократных проходов по исходной выборке (что вызывает значительные затраты машинного времени), а также приводит к необходимости хранить и обрабатывать матрицу расстояний между экземплярами большой размерности.

Для устранения отмеченных недостатков предлагается заменить обработку экземпляров на обработку их описаний в виде числовых скаляров, которые характеризуют положение экземпляров в пространстве признаков. При этом, заменив экземпляры, характеризующиеся N признаками, на представления в виде скаляров, мы отобразим N-мерное пространство признаков в одномерное пространство.

Исходная выборка, будучи отображенной в одномерное пространство, позволит выделить на одномерной оси интервалы ее значений, соответствующие кластерам разных классов в исходном N-мерном пространстве. Определив границы интервалов на одномерной оси, можно найти ближайшие к ним экземпляры, которые и составят формируемую подвыборку.

Приведенные выше идеи лежат в основе предлагаемого метода.

Этап инициализации. Задать исходную выборку данных $X=<x, y>$.

Этап анализа характеристик классов. Разбить выборку X на K подвыборок $X(k)$, отдельных для экземпляров каждого класса:

$$X(k)=\{X(k)\cup < x^s, y^s > | \; y^s = k\}, s = 1,2,..., S; k = 1,2,..., K.$$

Для каждой подвыборки $X(k)$ определить по каждому признаку его минимальное $\min\{x_j^s \mid x^s \in X(k)\}$, максимальное $\max\{x_j^s \mid x^s \in X(k)\}$ и среднее значения для экземпляров соответствующего класса:

$$C_j^k = \frac{1}{S^k}\sum_{s=1}^{S}\{x_j^s \mid y^s = k\}.$$

Для каждой подвыборки $X(k)$, $k = 1, 2, ..., K$, определить:

– частные поосевые нормированные расстояния от экземпляров до центров классов:

$$R_{(k)}(s)_j = \frac{C_j^k - x_j^s}{\max\limits_s\{x_j^s \mid x^s \in X(k)\} - \min\limits_s\{x_j^s \mid x^s \in X(k)\}},$$
$$s = 1,2,..., S, j = 1,2,..., N;$$

– нормированные расстояния экземпляров до центров классов:

$$R_{(k)}(s) = \frac{1}{N}\sum_{j=1}^{N}\left|R_{(k)}(s)_j\right|, s = 1,2,..., S;$$

– нормированные расстояния между экземплярами:

$$R_{(k)}(s,s) = 0, s = 1,2,..., S;$$

$$R_{(k)}(s,p) = R_{(k)}(p,s) = \frac{1}{2N}\sum_{j=1}^{N}\left|R_{(k)}(s)_j - R_{(k)}(p)_j\right|,$$
$$s = 1,2,..., S, \; p = s+1, s+2,..., S.$$

Этап устранения дублирующихся экземпляров. Целью этапа является выделение подмножеств эквивалентных и существенно похожих экземпляров и замена каждого такого подмножества на один его экземпляр-представитель.

Четкий дубляж: из каждой группы одинаковых экземпляров каждой подвыборки следует оставить только один экземпляр:

$$X(k) = X(k) \setminus \{x^p \mid x^s \in X(k), x^p \in X(k),$$
$$R_{(k)}(s) = R_{(k)}(p), R_{(k)}(s,p) = 0\},$$
$$S^k > 1, k = 1,2,..., K, s = 1,2,...S, p = s+1,..., S.$$

Нечеткий дубляж: из каждой группы неодинаковых подобных экземпляров каждой подвыборки следует оставить только один экземпляр:

$$X(k) = X(k) \setminus \{x^p \mid p \neq s, x^s \in X(k), x^p \in X(k),$$
$$\left|R_{(k)}(s) - R_{(k)}(p)\right| \le \varepsilon_1(k), R_{(k)}(s,p) \le \varepsilon_2(k,s,p)\},$$
$$S^k > 1, k = 1,2,..., K, s = 1,2,...S, p = s+1,..., S,$$

где

$$\varepsilon_1(k) = \frac{1}{\ln S^k}, \; \varepsilon_2(k,s,p) = \exp(-\left|R_{(k)}(s) - R_{(k)}(p)\right| \ln S^k).$$

Этап выделения граничных экземпляров. Целью данного этапа является выявление экземпляров, находящихся вблизи границ классов, что позволит устранить остальные экземпляры, находящиеся внутри области класса.

Вначале необходимо определить индексы для всех экземпляров выборки относительно центров всех подвыборок:

$$I^s(k) = \text{round}(R_{(k)}(s)\ln S^k) + \frac{1}{\pi}\arccos\left(\frac{\sum\limits_{j=1}^{N} C_j^k x_j^p}{\sqrt{\sum\limits_{j=1}^{N}\left(C_j^k\right)^2}\sqrt{\sum\limits_{j=1}^{N}\left(x_j^P\right)^2}}\right),$$

где round – функция округления до ближайшего целого числа.

Это позволит отобразить исходную выборку на одномерные оси $I(k)$ (заметим, что при этом произойдет потеря части информации вследствие неявного квантования пространства признаков при преобразовании).

Просматривая каждую одномерную ось $I(k)$ можно выделить скопления (области пространства) близко расположенных экземпляров одного класса, выделив интервалы для каждого из них $I(k)=\{i_l(k)\}$, где $i_l(k)$ – l-й интервал k-й оси, либо для простоты разбить эту ось на несколько равных интервалов и определить доминирующий класс в каждом из них.

До тех пор, пока $\exists k, k=1,2,...,K: X(k)\neq\varnothing$, выполнять в цикле:

– если для области пространства, где расположены экземпляры k-го класса не существует попавших в нее экземпляров других классов ($\neg\exists s, s=1,2,...,S: y^s=k, I^s(k)\le I^p(k), \forall p=1,2,...,S, s\neq p, y^p=k$), то данный класс расположен компактно и отделен от других классов. Следовательно, из экземпляров k-го класса в новую выборку следует включить лишь те экземпляры, которые находятся вблизи его внешней границы:

$$X^*=X^*\cup\{x^p \mid p=1,2,...,S: x^p\in X(k),$$
$$I^p(k)\ge\frac{\alpha}{S^k}\sum_{s=1}^{S}\{I^s(k)\mid y^s=k\},$$

где α – задаваемый пользователем коэффициент, регулирующий долю помещаемых в новую выборку экземпляров k-го класса (например, можно рекомендовать задавать $\alpha=1$).

После чего необходимо исключить экземпляры k-го класса из дальнейшего рассмотрения:

$$X(k)=X(k)\setminus\{x^p \mid y^p=k, p=1,2,...,S\};$$

– если для области пространства, где расположены экземпляры k-го класса, существуют попавшие в нее экземпляры других классов ($\exists s, p: s=1,2,...,S, p=1,2,...,S, s\neq p, x^s\notin X^*, x^p\notin X^*, y^s\neq k, y^p=k, I^s(k)\le I^p(k)$), но число экземпляров других классов $S^*(k)$, попавших в область k-го класса невелико: $S^*(k)\le\beta S(k)$, где $S(k)$ – число экземпляров k-го класса, β – заданный коэффициент ($0<\beta<1$), то из экземпляров k-го класса в новую выборку следует включить лишь те экземпляры, которые находятся вблизи его внешней границы, а также экземпляры, ближайшие к экземплярам других классов:

$$X^*=X^*\cup\{x^p \mid p=1,2,...,S: x^p\in X(k), I^p(k)\ge\frac{\alpha}{S^k}\sum_{s=1}^{S}\{I^s(k)\mid y^s=k\},$$

$$X^*=X^*\cup\{x^q \mid q=1,2,...,S, x^q\notin X^*, y^q=k, \exists p, p=1,2,...,S: y^p\neq k,$$
$$|I^q(k)-I^p(k)|\le|I^s(k)-I^p(k)|, \forall s=1,2,...,S, y^s=k, s\neq q\}.$$

Все экземпляры других классов, попавшие в область k-го класса, также следует включить в новую выборку:

$$X^*=X^*\cup\{x^s \mid s=1,2,...,S, p=1,2,...,S, s\neq p,$$
$$x^s\notin X^*, y^s\neq k, y^p=k, I^s(k)\le I^p(k)\};$$

– если для области пространства, где расположены экземпляры k-го класса, существуют попавшие в нее экземпляры других классов ($\exists s, p: s=1,2,...,S, p=1,2,...,S, s\neq p, x^s\notin X^*, x^p\notin X^*, y^s\neq k, y^p=k, I^s(k)\le I^p(k)$), но число экземпляров других классов, попавших в область k-го класса велико ($S^*(k)>\beta S(k)$), то на оси $I(k)$ следует выделить отдельные скопления экземпляров каждого класса и включить в новую выборку лишь те экземпляры, которые находятся вблизи его внешней границы, а также граничные экземпляры каждого интервала и экземпляры, ближайшие к ним:

$$X^*=X^*\cup\{x^p \mid p=1,2,...,S: x^p\in X(k), I^p(k)\ge$$
$$\ge\frac{\alpha}{S^k}\sum_{s=1}^{S}\{I^s(k)\mid y^s=k\},$$

$$X^*=X^*\cup\{x^q\cup x^p \Big| q,p=1,2,...,S,$$
$$x^q\notin X^*, x^p\notin X^*, y^q=k, y^p\neq k,$$
$$\left|\tau(x^q,k)-\tau(x^p,k)\right|=1, \left|I^q(k)-I^p(k)\right|\le$$
$$\le\gamma(k,\tau(x^q,k),\tau(x^p,k))\left|I^S(k)-I^g(k)\right|,$$
$$y^S\in i_{\tau(x^q,k)}(k), y^g\in i_{\tau(x^p,k)}(k), s,g=1,2,...,S\},$$

где $\tau(x^s,k)=\begin{cases}l, x^s\in i_l(k), l=1,2,...,L(k);\\ 0, \text{в противном случае};\end{cases}$

$L(k)$ – число интервалов значений, на которые разбита ось $I(k)$, γ – заданный коэффициент, регулирующий размер области вблизи межклассовых границ, экземпляры из которой включаются в формируемую выборку, $0<\gamma(k,l,m)<\omega(k,l,m)$, где

$$\omega(k,l,m)=\begin{cases}\max\limits_{\substack{s=1,2,...,S;\\ g=1,2,...,S}}\left\{I^s(k)-I^g(k)\,\middle|\, y^s\in i_l(k), y^g\in i_m(k)\right\}\\ \min\limits_{\substack{s=1,2,...,S;\\ g=1,2,...,S}}\left\{I^s(k)-I^g(k)\,\middle|\, y^s\in i_l(k), y^g\in i_m(k)\right\}\end{cases}, \min\limits_{\substack{s=1,2,...,S;\\ g=1,2,...,S}}\left\{I^s(k)-I^g(k)\,\middle|\, y^s\in i_l(k), y^g\in i_m(k)\right\}>0;$$
$$0, \min\limits_{\substack{s=1,2,...,S;\\ g=1,2,...,S}}\left\{I^s(k)-I^g(k)\,\middle|\, y^s\in i_l(k), y^g\in i_m(k)\right\}=0.$$

После чего из дальнейшего рассмотрения следует исключить экземпляры k-го класса, а также те экземпляры остальных классов, которые были включены в новую выборку:

$$X(q) = X(q) \setminus \{x^s \mid s = 1,2,...,S, p = 1,2,...,S, s \neq p, x^s \notin X^*,$$
$$y^s = q, y^s \neq k,$$
$$y^p = k, I^s(k) \leq I^p(k)\}, q = 1,2,...,K,$$

$$X(k) = X(k) \setminus \{x^p \mid x^p \in X^*, p = 1,2,...,S\}.$$

В результате выполнения предложенного метода будет сформирована выборка $X^* \subseteq X$.

3. АНАЛИЗ ВЫЧИСЛИТЕЛЬНОЙ И ПРОСТРАНСТВЕННОЙ СЛОЖНОСТИ МЕТОДА

Предложенный метод не требует хранения в памяти ЭВМ всей исходной выборки. На этапе анализа характеристик классов метод делает один проход по исходной выборке для определения значений ее характеристик. При этом для каждого класса определяются минимальное, максимальное и среднее значения каждого признака. Таким образом, при максимальной экономии оперативной памяти (исходная выборка размещается во внешней памяти) пространственная сложность этих действий составит $O(3KN)$, а вычислительная – $O(3KNS)$ без учета системных затрат на доступ ко внешней памяти.

Определение расстояний до центров классов характеризуется пространственной сложностью $O(KNS + SK + S^2)$ и вычислительной сложностью $O(2KNS + 2NS^2)$.

Используя символ Ландау «O» в так называемом «мягком виде», оценим общую сложность данного этапа: пространственную – $O(3KN+KNS+SK+S^2)$, вычислительную – $O(5KNS + 2NS^2)$.

Этап устранения дубляжа не требует существенных затрат оперативной памяти, а его вычислительная сложность в предельном случае составит $O(KS^2)$.

На этапе выделения граничных экземпляров метод делает один проход по исходной выборке для расчета значений одномерных индексов. Его вычислительная сложность составит $O(6NSK)$, а пространственная – $O(SK)$.

Далее метод оперирует только множеством индексов. Его пространственной сложностью можно пренебречь, а вычислительную сложность грубо оценим как $O(2S^2)$.

В итоге оценим общую сложность метода: вычислительную – $O(11KNS + S^2(2N+K+2))$, а пространственную – $O(KN(S+3)+2SK+S^2)$. Полагая из практических соображений для простоты K=2, N<<S (например, $N \approx 0{,}01S$) и, обозначив размерность исходной выборки $n=NS$, получим оценки сложности метода: вычислительную – $O(0{,}02S^3+4S^2+0{,}22S) \approx O(20n\sqrt{n}+400n+2{,}2\sqrt{n})$, пространственную – $O(1{,}02S^2+4{,}06S)O(102n+40{,}6)$.

4. ЭКСПЕРИМЕНТЫ И РЕЗУЛЬТАТЫ

Для экспериментальной проверки работоспособности предложенного метода была разработана его программная реализация на языке пакета MATLAB, с помощью которой проводились эксперименты по сокращению объема выборок данных для различных практических задач [14–16], характеристики которых приведены в табл. 1.

Результаты проведенных экспериментов подтвердили работоспособность и практическую применимость предложенного метода, а также программного обеспечения, реализующего его. Как видно из таблицы, использование предложенного метода позволяет существенно сократить объем выборки (в 7,7–12,5 раз), не требуя при этом загрузки в память ЭВМ исходной выборки, а также многочисленных проходов по исходной выборке, что существенно снижает требования к ресурсам ЭВМ, обеспечивая при этом сохранение в сформированной подвыборке важнейших для последующего анализа топологических свойств исходной выборки.

ЗАКЛЮЧЕНИЕ

В работе решена актуальная задача формирования и редукции выборок для интеллектуального анализа данных.

Научная новизна результатов работы заключается в том, что впервые предложен метод формирования и редукции выборок, который обеспечивает сохранение в сформированной подвыборке важнейших для последующего анализа топологических свойств исходной выборки, не требуя при этом загрузки в память ЭВМ исходной выборки, а также многочисленных проходов по исходной выборке, что позволяет существенно сократить объем выборки, существенно уменьшает требования к ресурсам ЭВМ.

Практическая значимость результатов работы состоит в том, что разработано программное обеспечение, реализующее предложенный метод формирования и редукции выборок, а также проведены эксперименты по их исследованию при решении практических задач, результаты которых позволяют рекомендовать разработанный метод для использования на практике при решении задач интеллектуального анализа данных.

Таблица 1. Характеристики исходных и сформированных выборок

Задача	K	N	S	n	S^*	S^*/S
Классификация автотранспортных средств по изображению [14]	2	26	1062	27612	139	0,13
Диагностирование патологий плода по кардиотокограмме [15]	3	23	2126	48898	182	0,09
Предсказание типа лесного покрова [16]	7	54	581012	31374648	49386	0,08

СПИСОК ЛИТЕРАТУРЫ

1. *Олійник, А. О.* Інтелектуальний аналіз даних : навчальний посібник / А. О. Олійник, С. О. Субботін, О. О. Олійник. – Запоріжжя : ЗНТУ, 2012. – 271 с.
2. *Рутковская, Д.* Нейронные сети, генетические алгоритмы и нечеткие системы / Д. Рутковская, М. Пилиньский, Л. Рутковский ; пер. с польск. И. Д. Рудинского. – М. : Горячая линия – Телеком, 2004. – 452 с.
3. 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 : ООО «Компания Смит», 2012. – 317 с.
4. Прогрессивные технологии моделирования, оптимизации и интеллектуальной автоматизации этапов жизненного цикла авиационных двигателей : монография / [А. В. Богуслаев, Ал. А. Олейник, Ан. А. Олейник, Д. В. Павленко, С. А. Субботин ; под ред. Д. В. Павленко, С. А. Субботина]. – Запорожье : ОАО «Мотор Сич», 2009. – 468 с.
5. *Субботин, С. А.* Формирование выборок и анализ качества моделей на основе нейронных и нейро-нечетких сетей в задачах диагностики и распознавания образов / С. А. Субботин : монография. – Saarbrücken: LAP Lambert academic publishing, 2012. – 232 с.
6. *Jensen, R.* Computational intelligence and feature selection: rough and fuzzy approaches / R. Jensen, Q. Shen. – Hoboken: John Wiley & Sons, 2008. – 339 p.
7. *Chaudhuri, A.* Survey sampling theory and methods / A. Chaudhuri, H. Stenger. – New York: Chapman & Hall, 2005. – 416 p.
8. Encyclopedia of survey research methods / ed. P. J. Lavrakas. – Thousand Oaks: Sage Publications, 2008. – Vol. 1–2. – 968 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. *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.
11. *Субботин, С. А.* Комплекс характеристик и критериев сравнения обучающих выборок для решения задач диагностики и распознавания образов / С. А. Субботин // Математичні машини і системи. – 2010. – № 1. – С. 25–39.
12. *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.
13. *Субботин, С. А.* Методы формирования выборок для построения диагностических моделей по прецедентам / С. А Субботин // Вісник Національного технічного університету «Харківський політехнічний інститут» : зб. наук. праць. – Харків: НТУ «ХПІ», 2011. – № 17. – С. 149–156.
14. *Субботин, С. А.* Синтез нейро-нечетких моделей для выделения и распознавания объектов на сложном фоне по двумерному изображению / С. А. Субботин // Комп'ютерне моделювання та інтелектуальні системи : збірник наукових праць за ред. Д. М. Пізи, С. О. Субботіна. – Запоріжжя : ЗНТУ, 2007. – С. 68–91.
15. Cardiotocography Data Set [Electronic resource]. – Access mode: http://archive.ics.uci.edu/ml/datasets/Cardiotocography.
16. Covertype Data Set [Electronic resource]. – Access mode: http://archive.ics.uci.edu/ml/datasets/Covertype.

Стаття надійшла до редакції 03.09.2012.

Субботін С. О.
Канд. техн. наук, доцент, Запорізький національний технічний университет, Україна

ФОРМУВАННЯ І РЕДУКЦІЯ ВИБІРОК ДЛЯ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

Вирішено задачу формування і редукції вибірок для інтелектуального аналізу даних. Запропоновано метод формування і редукції вибірок, що забезпечує збереження у сформованій підвибірці найважливіших топологічних властивостей вихідної вибірки, не вимагаючи при цьому завантаження у пам'ять ЕОМ вихідної вибірки, а також численних проходів вихідної вибірки, що дозволяє скоротити обсяг вибірки і зменшити вимоги до ресурсів ЕОМ.

Ключові слова: вибірка, відбір екземплярів, редукція даних, інтелектуальний аналіз даних, скорочення розмірності даних.

Subbotin S. A.
Doctor of philosophy (Cand. Tech. Sc.), associate professor (docent), Zaporizhian National Technical University, Ukraine

SAMPLE FORMATION AND REDUCTION FOR DATA MINING

In data mining problem solving it has to operate with a large amount of data samples. This entails a significant amount of time to process the data. Therefore, an urgent task is to reduce the dimensionality of the data samples. The aim of paper is to provide a method for the formation and reduction of samples, allowing to handle a large amount of the original sample.

The problem of sample formation and reduction for data mining was solved.

The scientific novelty of the work lies in the fact that the method of sample formation and reduction is firstly proposed. It provides a saving of the most important topological properties of original sample in the formed sub-sample without the need for downloading the original sample to the computer memory, and without numerous passages of the original sample. It allows to reduce the size of the sample and to reduce the resource requirements of a computer.

The practical significance of the work lies in the development of software, which implements the proposed method of sample formation and reduction, also as conducting of experiments on research of proposed method to solve practical problems, the results of which allows to recommend the developed method for use in practice in solving problems of data mining.

Using the proposed method one can significantly reduce the amount of a sample (in 7,7–12,5 times), without the need to download the original sample into computer memory, providing preservation in the generated sub-sample the most important for analysis of the topological properties of the original sample.

Keywords: sample, example selection, data reduction, data mining, data dimensionality reduction.

REFERENCES

1. Olijnik A. O., Subbotin S. O., Olijnik O. O. Intelektual'nij analiz danih : navchal'nij posibnik. Zaporizhzhja, ZNTU, 2012, 271 p.
2. Rutkovskaja D., Pilin'skij M., Rutkovskij L.; per. s pol'sk. I. D. Rudinskogo. Nejronnye seti, geneticheskie algoritmy i nechjotkie sistemy. Moscow, Gorjachaja linija, Telekom, 2004, 452 p.
3. Subbotin S. A., Olejnik An. A., Gofman E. A., Zajcev S. A., Olejnik Al. A.; pod red. S. A. Subbotina Intellektual'nye informacionnye tehnologii proektirovanija avtomatizirovannyh sistem diagnostirovanija i raspoznavanija obrazov : monografija. Har'kov, OOO «Kompanija Smit», 2012, 317 p.
4. Boguslaev A. V., Olejnik Al. A., Olejnik An. A., Pavlenko D. V., Subbotin S. A.; pod red. D. V. Pavlenko, S. A. Subbotina. Progressivnye tehnologii modelirovanija, optimizacii i intellektual'noj avtomatizacii jetapov zhiznennogo cikla aviacionnyh dvigatelej : monografija. Zaporozh'e, OAO «Motor Sich», 2009, 468 p.
5. Subbotin S. A. Formirovanie vyborok i analiz kachestva modelej na osnove nejronnyh i nejro-nechjotkih setej v zadachah diagnostiki i raspoznavanija obrazov : monografija. Saarbrücken, LAP Lambert academic publishing, 2012, 232 p.
6. Jensen R., Shen Q. Computational intelligence and feature selection: rough and fuzzy approaches. Hoboken, John Wiley & Sons, 2008, 339 p.
7. Chaudhuri A., Stenger H. Survey sampling theory and methods, New York, Chapman & Hall, 2005, 416 p.
8. Encyclopedia of survey research methods. ed. P. J. Lavrakas, Thousand Oaks, Sage Publications, 2008, Vol. 1–2, 968 p.
9. Kokren U.; per. s angl. I. M. Sonina ; pod red. A. G. Volkova, N. K. Druzhinina. Metody vyborochnogo issledovanija. Moscow, Statistika, 1976, 440 p.
10. 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.
11. Subbotin S. A. Kompleks harakteristik i kriteriev sravnenija obuchajuwih vyborok dlja reshenija zadach diagnostiki i raspoznavanija obrazov. *Matematichni mashini i sistemi*, 2010, No. 1, pp. 25–39.
12. Subbotin S. A. Kriterii individual'noj informativnosti i metody otbora jekzempljarov dlja postroenija diagnosticheskih i raspoznajuwih modelej. *Bionika intelektu,* 2010, No.1, pp. 38–42.
13. Subbotin S. A. Metody formirovanija vyborok dlja postroenija diagnosticheskih modelej po precedentam. *Visnik Nacional'nogo tehnichnogo universitetu «Harkivs'kij politehnichnij institut» : zb. nauk. prac'*, Harkiv, NTU «HPI», 2011, No. 17, pp. 149–156.
14. Subbotin S. A. Sintez nejro-nechetkih modelej dlja vydelenija i raspoznavanija objektov na slozhnom fone po dvumernomu izobrazheniju. *Komp'juterne modeljuvannja ta intelektual'ni sistemi : zbirnik naukovih prac',* za red. D. M. Pizi, S. O. Subbotina, Zaporizhzhja, ZNTU, 2007, pp. 68–91.
15. Cardiotocography Data Set [Electronic resource]. – Access mode: http://archive.ics.uci.edu/ml/datasets/Cardiotocography.
16. Covertype Data Set [Electronic resource]. – Access mode: http://archive.ics.uci.edu/ml/datasets/Covertype.