УДК 004.93

С.А. СУББОТИН

МЕТОД ИЗВЛЕЧЕНИЯ ОБУЧАЮЩИХ ВЫБОРОК ИЗ ИСХОДНЫХ ВЫБОРОК БОЛЬШОГО ОБЪЕМА ДЛЯ ДИАГНОСТИРОВАНИЯ И РАСПОЗНАВАНИЯ ОБРАЗОВ

Анотація. *Вирішено задачу автоматизації формування виборок для побудови діагностичних і розпізнавальних моделей за прецедентами. Запропоновано метод витягу навчальних виборок, що забезпечує збереження у сформованій підвиборці 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.*

Аннотация. *Решена задача автоматизации формирования выборок для построения диагностических и распознающих моделей по прецедентам. Предложен метод извлечения обучающих 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 объём выборки и уменьшить требования к ресурсам ЭВМ.*
Ключевые слова: *выборка, отбор экземпляров, редукция данных, интеллектуальный анализ данных, сокращение размерности данных.*

Abstract. *The task of sample formation automaticity for diagnostic and recognizing model building on precedents is solved. Extraction method of training samples is offered. It maintains saving the important topological properties of the original sample in a formed sub-sample, and does not require download of the original sample to the computer memory, and the numerous passages of the original sample. This reduces the size of the sample and reduces the resource requirements of a computer.*
Keywords: *sample, example selection, data reduction, data mining, data dimensionality reduction.*

1. Введение

При решении задач построения диагностических и распознающих моделей на основе нейронных и нейро-нечётких сетей [1–4], а также деревьев решений [3], зачастую необходимо использовать выборки данных большого объема. Это приводит к необходимости использования ЭВМ с большим объемом оперативной памяти, а также существенно увеличивает затраты машинного времени на обработку данных. Поэтому актуальной задачей является сокращение размерности выборок данных.

Традиционным и наиболее широко применяемым подходом при решении данной задачи является использование методов отбора информативных признаков [1–5], которые удаляют из исходного набора наименее информативные признаки, и методов конструирования признаков [5, 6], которые 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 чрезвычайно велик для представления и обработки в памяти ЭВМ, применение этих методов оказывается на практике затруднительным, а результаты их работы 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 объёма 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] 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 автоматического извлечения обучающих выборок из исходных выборок большого объёма.

2. Постановка задачи

Пусть мы имеем исходную выборку $X =< x, y >$ – 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 $< x^s, y^s >$, $x^s = \{x_j^s\}$, где x_j^s – 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 объёма выборки может быть представлена как задача формирования (выделения) из исходной выборки $X =< x, y >$ подвыборки $X^*, X^* \subset X$, меньшего объёма $S^* < S$, обладающей наиболее важными свойствами исходной выборки.

Поскольку для задач автоматизации поддержки принятия диагностических решений, а также задач автоматической классификации наиболее важным является сохранение топологии классов, то формируемая подвыборка должна обеспечивать сохранение экземпляров исходной выборки, находящихся на границах классов.

3. 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 объёма 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 отображённой в одномерное пространство, позволит выделить на одномерной оси интервалы её 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 выборки. Вначале, путём просмотра всей исходной выборки, определить для каждого j-го признака, $j = 1,2,...,N$, его максимальное и минимальное значения:

$$x_j^{\max} = \max_{s=1,2,...,S} \{x_j^s\},\ x_j^{\min} = \min_{s=1,2,...,S} \{x_j^s\},$$

а также координаты центров классов:

$$C_j^q = \frac{1}{S^q} \sum_{s=1}^{S} \{x_j^s \mid y^s = q\},\ q = 1,2,...,K,$$

где S^q – число экземпляров исходной выборки, принадлежащих к q-му классу.

После чего определить:

– размах диапазона значений j-го признака:

$$\delta_j = x_j^{\max} - x_j^{\min},\ j = 1,2,...,N;$$

– расстояния между центрами классов по j-му признаку:

$$d_j(q,p) = d_j(p,q) = \left| C_j^q - C_j^p \right|,\ j = 1,2,...,N,\ q,\ p = 1,2,...,K;$$

– число интервалов значений j-го признака, $S \geq K$:

$$n_j = \begin{cases} \text{round}\left(\ln\left(\dfrac{\delta_j S}{\min\limits_{\substack{q=1,2,...,K;\\ p=q+1,q+2,...,K}} \{d_j(q,p)\}} \right) \right), & \min\limits_{\substack{q=1,2,...,K;\\ p=q+1,q+2,...,K}} \{d_j(q,p)\} > 0; \\ 0{,}5S, & \min\limits_{\substack{q=1,2,...,K;\\ p=q+1,q+2,...,K}} \{d_j(q,p)\} = 0; \end{cases}$$

– длину интервала значений j-го признака: $\theta_j = \dfrac{\delta_j}{n_j}$.

Этап преобразования выборки. Для каждого s-го экземпляра x^s, $s = 1,2,...,S$ определить:

– номер интервала, в который попадает экземпляр x^s по оси значений j-го признака:

$$r_j(x^s) = \begin{cases} \text{round}\left(1 + \dfrac{x_j^s - x_j^{\min}}{\theta_j} \right), \theta_j > 0; & j = 1,2,...,N; \\ 1, \theta_j = 0, \end{cases}$$

– нормированное интервальное расстояние от экземпляра x^s до начала отсчета системы координат:

$$R(x^S)=\sqrt{\sum_{j=1}^{N} r_j(x^S)^2} \text{ либо } R(x^S)=\min_{j=1,2,...,N}\{r_j(x^S)\};$$

– угол, определяющий положение экземпляра x^s в пространстве интервалов значений признаков:

$$\alpha(x^s)=\frac{1}{\pi}\arccos\left(\frac{\sum_{j=1}^{N} r_j(x^s)}{\sqrt{\sum_{j=1}^{N}\left(r_j(x^s)\right)^2}}\right);$$

– индекс экземпляра x^s: $I^S=R(x^S)+\alpha(x^S)$.

Это позволит отобразить исходную выборку на одномерную ось I. Заметим, что при этом произойдет потеря части информации вследствие неявного квантования пространства признаков при преобразовании.

После чего следует сформировать набор $X'=\{x'^s\}$, элементы которого $x'^s=<I^s,y^s,s>$ необходимо отсортировать в порядке возрастания значений I^s.

Этап выделения граничных экземпляров выборки. По сформированной одномерной оси I можно выделить скопления (области пространства) близко расположенных экземпляров одного класса, выделив интервалы для каждого из них.

Для этого следует:

– определить границы интервалов её значений, внутрь которых попадают экземпляры, принадлежащие только к одному классу. Вначале установить число интервалов: $k=0$ и номер текущего экземпляра $s=1$. Затем до тех пор, пока $s<S$, выполнять в цикле: принять: $k=k+1$, установить левую границу k-го интервала: $l_k=I^s$, установить номер класса k-го интервала: $K_k=y^s$, далее до тех пор, пока $s<S$ и $y^{s+1}=K_k$, наращивать $s:s=s+1$, после чего установить правую границу текущего интервала: $r_k=I^s$. Занести число интервалов в k_I;

– из элементов X' оставить только экземпляры, ближайшие к границам интервалов:

$$X'=X'\setminus\{<I^S,y^S,s>|\neg k=1,2,...,k_I: l_k=I^S\vee r_k=I^S\}.$$

Этап формирования новой выборки. При просмотре исходной выборки X занести в формируемую обучающую выборку X^* те экземпляры из X, номера которых содержатся в X':

$$\forall s=1,2,...,S: X^*=X^*\cup\{<x^S,y^S>|s\in X'\}.$$

Из экземпляров X, не вошедших в X^*, при необходимости можно сформировать тестовую выборку.

4. Анализ вычислительной и пространственной сложности метода

Предложенный метод не требует хранения в оперативной памяти ЭВМ всей исходной выборки: необходимо хранить только текущий обрабатываемый экземпляр и набор индексов

$X^{'}$, причём для ЭВМ с малым объемом оперативной памяти возможно хранение набора $X^{'}$ во внешней памяти (это, однако, замедлит скорость работы метода).

Предложенный метод при эффективной программной реализации делает всего три прохода по исходной выборке (один – на этапе анализа выборки, один – на этапе преобразования выборки и один – на этапе формирования новой выборки) и порядка $2+S\ln S$ проходов по оси индексов (один – на этапе выделения граничных экземпляров выборки, один – на этапе формирования новой выборки, а остальные – при сортировке на этапе преобразования выборки).

Его пространственную сложность можно оценить как $O\left(SN+5S+6N+K^2-K\right)$ – при полной загрузке исходной выборки в оперативную память и $O\left(5S+7N+K^2-K\right)$ – при поэкземплярном доступе к исходной выборке, хранящейся во внешней памяти, при условии, что формируемая выборка хранится во внешней памяти.

Вычислительная сложность метода может быть оценена как $O\left(14NS+6S+S\ln S+N+\left(N+1\right)\left(K^2-K\right)\right)$ без учёта затрат на доступ во внешнюю память, которые определяются особенностями конкретной ЭВМ и программной реализации метода.

Полагая из практических соображений для простоты $K=2, N<<S$ (например, $N\approx 0,001S$) и обозначив размерность исходной выборки $n=NS\approx 0,001S^2$, а также, полагая $S\ln S\approx 8S$, получим оценки сложности метода: вычислительной – $O\left(0,014S^2+14,003S\right)\approx O\left(14n+442,81\sqrt{n}\right)$, пространственной при поэкземплярном доступе – $O\left(5,007S\right)\approx O\left(158,34\sqrt{n}\right)$.

5. 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.

Таблица 1. Характеристики исходных и сформированных выборок

Задача	K	N	S	n	S^*	S^*/S
Диагностирование патологий плода по кардиотокограмме [14]	3	23	2126	48898	236	0,11
Предсказание типа лесного покрова [15]	7	54	581012	31374648	51926	0,09

Результаты проведенных экспериментов подтвердили работоспособность и практическую применимость предложенного метода, а также программного обеспечения, реализующего его. Как видно из табл. 1, использование предложенного метода позволяет существенно сократить объём выборки (в 9–11 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.

6. Заключение

В работе решена актуальная задача автоматизации формирования выборок для построения диагностических и распознающих моделей по прецедентам.

Научная новизна результатов работы заключается в том, что впервые предложен метод извлечения обучающих 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 объём выборки, существенно уменьшить 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.

Дальнейшие исследования могут быть сосредоточены на разработке новых способов формирования описаний экземпляров в виде обобщённых показателей, разработке реализаций предложенного метода для параллельных вычислительных систем и распределенной обработки данных.

Работа выполнена в рамках госбюджетных научно-исследовательских тем Запорожского национального технического университета "Методы, модели и устройства принятия решений в системах распознавания образов" (№ гос. регистрации 0111U000059) и "Интеллектуальные информационные технологии автоматизации проектирования, моделирования, управления и диагностирования производственных процессов и систем".

СПИСОК ЛИТЕРАТУРЫ

1. Руденко О.Г. Штучні нейронні мережі / О.Г. Руденко, Є.В. Бодянський. – Харків: Компанія СМІТ, 2006. – 404 с.
2. Рутковская Д. Нейронные сети, генетические алгоритмы и нечёткие системы / Д. Рутковская, М. Пилкий, 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 / [С.А. Субботин, Ан.А. Олейник, Е.А. Гофман и др.; 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 / [А.В. Богуслаев, Ал.А. Олейник, Ан.А. Олейник и др.; 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 нейронечётких сетей в задачах диагностики и распознавания образов / С.А. Субботин. – 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, N 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. Cardiotocography Data Set [Електронний ресурс]. – Режим доступу: http://archive.ics.uci.edu/ml/datasets/Cardiotocography.
15. Covertype Data Set [Електронний ресурс]. – Режим доступу: http://archive.ics.uci.edu/ml/datasets/Covertype.

Стаття надійшла до редакції 03.10.2012