рицательный результат, в то время как выражения на самом деле являются ограниченно синтаксически эквивалентными (адаптивно унифицируемыми). Вместе с тем достоверность результата может быть повышена за счет расширения базы знаний необходимыми утверждениями в форме равенства или равносильности.

ЛИТЕРАТУРА

1. *Гаврилова Т.Л., Клещев А.С.* Внутренняя модель математической практики для систем автоматизированного конструирования доказательств теорем // Проблемы управления. – 2006. – № 4. – Ч. 1. – С.32-35; № 5. – Ч. 2. – С.68-73; № 6. – Ч. 3. – С.68-71.
2. *Клещев А.С., Тимченко В.А.* Алгоритм унификации для расширяемой модели математического диалекта // Информатика и системы управления. − 2012. − № 1(31). − С.155-165.
3. Wolfram Mathematica 9. [Электронный ресурс]. URL: http://www.wolfram.com/mathematica (дата обращения 14.12.2012).
4. Maple 16. [Электронный ресурс]. URL: http://www.maplesoft.com/products/Maple (дата обращения 14.12.2012).
5. Matlab. The language for Technical Computing. [Электронный ресурс]. URL: http://www.mathworks.com/products/matlab (дата обращения 14.12.2012).
6. *Клещев А.С., Тимченко В.А.* Задача применения подстановки для расширяемой модели математического диалекта // Информатика и системы управления. − 2011. − № 3(29). − С. 80-88.

E-mail:
Клещев Александр Сергеевич – kleschev@iacp.dvo.ru;
Тимченко Вадим Андреевич – rakot2k@mail.ru.

УДК 004.93

© 2013 г. **С.А. Субботин,** канд. техн. наук
(Запорожский национальный технический университет, Украина)

ФОРМИРОВАНИЕ ВЫБОРОК ДЛЯ РЕШЕНИЯ ЗАДАЧ КЛАССИФИКАЦИИ ПО ПРИЗНАКАМ*

Решена задача формирования выборок для автоматизации решения задач классификации объектов по признакам. Предложен новый метод формирования обучающих выборок, который обеспечивает сохранение в сформированной подвыборке топологических свойств исходной выборки и не требует при этом загрузки всей выборки в память ЭВМ, что позволяет сократить объем выборки и уменьшить требования к ресурсам ЭВМ.
Ключевые слова: выборка, отбор экземпляров, редукция данных, классификация, распознавание образов, сокращение размерности данных.

* Работа выполнена в рамках госбюджетных научно-исследовательских тем Запорожского национального технического университета "Методы, модели и устройства принятия решений в системах распознавания образов" и "Интеллектуальные информационные технологии автоматизации проектирования, моделирования, управления и диагностирования производственных процессов и систем".

Введение

Построение моделей принятия решений на основе нейронных и нейро-нечетких сетей, а также деревьев решений [1 – 4] в различных прикладных областях нередко предполагает необходимость оперировать выборками данных большого объема. Это влечет за собой существенные затраты времени на обработку данных, а также требует значительных объемов оперативной и дисковой памяти ЭВМ. Поэтому актуальна задача сокращения размерности выборок данных [1 – 5].

Традиционным и наиболее широко применяемым подходом при решении данной задачи является использование методов отбора информативных признаков (удаляют из исходного набора наименее информативные признаки) [1 – 5] и методов конструирования признаков (заменяют исходный набор признаков рассчитанным на его основе набором искусственных признаков меньшего размера) [5, 6]. 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 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.

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; 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 формирования обучающей выборки может быть представлена

как задача 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 автоматизации классификации данных наиболее важным является сохранение топологии классов, то формируемая подвыборка должна обеспечивать сохранение экземпляров исходной выборки, находящихся на границах классов.

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 $X = \langle x, y \rangle$.

Этап анализа характеристик выборки. Определить $x_j^{\min}$ и $x_j^{\max}$ – соответственно, минимальное и максимальное значения j-го признака, $j = 1, 2, ..., N$. Определить число интервалов для каждого признака: $k = K \ln S$, а также длины интервалов: $\delta_j = (x_j^{\max} - x_j^{\min}) / k$.

Этап расчета обобщенных признаков. Для каждого s-го экземпляра, $s = 1, 2, ..., S$:

определить k_j – номер интервала значений по каждому j-му признаку, j=1, 2, ..., N, в который попадает s-й экземпляр

$$k_j = 1 + \left\lfloor \frac{x_j^s - x_j^{\min}}{\delta_j} \right\rfloor;$$

рассчитать координату s-го экземпляра по обобщенной оси

$$I^s = \sum_{j=1}^{N}(k_j - 1)^2 + \frac{1}{\pi}\arccos\left(\frac{\sum_{j=1}^{N} k_j}{\sqrt{N\sum_{j=1}^{N}(k_j)^2}}\right).$$

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

Этап анализа обобщенной оси. Сформировать набор кортежей $I = \{<I^s, y^s, s>\}$.Упорядочить набор I в порядке неубывания значений I^s. Просматривая обобщенную ось в порядке увеличения ее значений, определить граничные значения ее интервалов $<l_q, r_q>$, в которых номер класса y^s остается неизменным, где l_q, r_q – соответственно левое и правое граничные значения q-го интервала обобщенной оси. Обозначим: K_q – номер класса, соответствующий q-му интервалу обобщенной оси; k_I – число интервалов обобщенной оси.

Этап анализа характеристик интервалов. Для каждого q-го интервала обобщенной оси, $k_q = 1, 2,..., k_I$, определить S_q – число попавших в него экземпляров, а также номера этих экземпляров.

Этап формирования обучающей выборки. Среди экземпляров q-го интервала включить в обучающую выборку X^* все экземпляры:

его класса, находящиеся на одной из границ интервала

$$X^* = X^* \cup \{< x^s, y^s > | \, y^s = K_q, x^s = l_q \vee x^s = r_q\}, s = 1, 2,..., S, q = 1,2,..., k_I;$$

его класса, ближайшие к одной из границ интервала:

$$X^* = X^* \cup \{< x^s, y^s > | \, y^s = K_q, ((x^s - l_q) < \alpha) \vee ((r_q - x^s) < \alpha)\}, 0 < \alpha < 1,$$

$$s = 1, 2,..., S, q = 1, 2,..., k_I,$$

где α – пороговый коэффициент, регулирующий близость экземпляров к границам интервала (например, можно задать: $\alpha = 0{,}1(r_q - l_q)$);

интервалов с малым числом экземпляров:

$$X^* = X^* \cup \{< x^s, y^s > | \, y^s = K_q, l_q \le x^s \le r_q\}, S_q < \beta\overline{S}, s = 1, 2,..., S, q = 1,2,..., k_I,$$

где β – некоторый пороговый коэффициент, $0 < \beta < 1$ (например, можно задать: $\beta = 0{,}1$); $\overline{S}$ – среднее число экземпляров в интервале обобщенной оси.

Этап устранения избыточности обучающей выборки. Определить расстояния между всеми экземплярами, вошедшими в сформированную обучающую выборку, сформировав матрицу расстояний R (для упрощения и ускорения вычислений будем оперировать квадратами расстояний):

$$R(s, p) = \sum_{j=1}^{N}(x_j^s - x_p^s)^2, \; s = 1, 2, ..., S^*, p = 1, 2, ..., S^*.$$

Заметим, что $R(s, p) = R(p, s)$, а $R(s, s) = 0$.

До тех пор, пока $\exists R(g, p) > 0, g \ne p,$ выполнять в цикле действия:

найти в матрице расстояний два экземпляра с наименьшим расстоянием между собой:

$$g, p = \arg \min_{\substack{g=1,2,...,S^* \\ p=g+1,...,S^*}} \{R(g,p) \mid R(g,p) \geq 0\};$$

если два ближайших экземпляра принадлежат к одному и тому же классу, то оставить в обучающей выборке только тот из них, который находится ближе к экземплярам других классов, а другой исключить из нее

$$X^* = X^* / \{< x^q, y^q >, q = \arg\max\{R(g,q_g), R(p,q_p)\} \mid y^g = y^p\},$$

$$q_g = \arg \min_{\substack{s=1,2,...,S^*; \\ s \neq g, s \neq p}} \{R(g,s) \mid y^g \neq y^s, R(g,s) \geq 0\},$$

$$q_p = \arg \min_{\substack{s=1,2,...,S^*; \\ s \neq g, s \neq p}} \{R(p,s) \mid y^p \neq y^s, R(p,s) \geq 0\}.$$

Скорректировать соответствующим образом элементы матрицы R, установив:

$R(s, q) = R(q, s) = -1$;

если два ближайших экземпляра принадлежат к разным классам, то перейти к выполнению этапа дополнения (уточнения) обучающей выборки.

Этап дополнения (уточнения) обучающей выборки. Определить разность исходной и сформированной выборок

$X' = X / X^*$.

Последовательно для каждого s'-го экземпляра $<x^{s'}, y^{s'}>$ выборки X', $s' = 1, 2, ..., S'$ относительно экземпляров сформированной выборки X^*:

найти расстояние (квадрат расстояния) от него до каждого экземпляра выборки X^*:

$$R^*(s', s*) = \sum_{j=1}^{N} (x_j^{s'} - x_j^{s*})^2;$$

если ближайший к s'-му экземпляру экземпляр сформированной выборки принадлежит к другому классу, то включить его в выборку X^*:

$$X^* = X^* \cup \left\{ < x^{s'}, y^{s'} > \mid y^{s'} \neq y^{q'}, q = \arg \min_{s^*=1,2,...,S^*} \{R^*(s', s*)\} \right\}.$$

В результате выполнения данного метода для исходной выборки X получим сформированную обучающую выборку X', которая будет обладать основными топологическими свойствами исходной выборки. При этом также из исходной выборки можно получить также тестовую выборку как разность исходной и сформированной обучающей выборок.

Анализ сложности метода

Для определения целесообразности применения предложенного метода для конкретной задачи на практике используем нотацию Ландау в так называемом "мягком виде" и оценим сложность этапов предложенного метода.

Для этапа инициализации вычислительной сложностью можно пренебречь, а пространственная сложность может быть оценена как $O(NS)$. Для этапа анализа

характеристик выборки вычислительная сложность составит $O(2NS)$, а пространственная – $O(4N)$. Для этапа расчета обобщенных признаков вычислительная сложность может быть оценена как $O(9NS)$, а пространственная – $O(N+S)$.

Для этапа анализа обобщенной оси вычислительная сложность может быть оценена как $O(2S^2+2S)$, а пространственная – $O(3S+3K\ln S)$. Для этапа анализа характеристик интервалов вычислительная сложность может быть оценена как $O(3KS\ln S)$, а пространственная – $O(K\ln S)$.

Для этапа формирования обучающей выборки вычислительная сложность может быть оценена как $O(20KS\ln S)$, а пространственная в виде $O(0{,}2K\ln S + 0{,}2S)$.

Для этапа устранения избыточности обучающей выборки вычислительная и пространственная сложность могут быть оценены соответственно как $O(0{,}0016S^4+(0{,}44+0{,}08N)S^2)$ и $O(0{,}04S^2)$. Для этапа дополнения (уточнения) обучающей выборки вычислительная и пространственная сложность оцениваются соответственно как $O((0{,}48N+0{,}32)S^2)$ и $O(0{,}8S)$.

Таким образом, общая сложность метода может быть оценена как:

вычислительная – $O(0{,}0016S^4 + 0{,}56NS^2 + 2{,}76S^2 + 11NS + 2S + 23KS\ln S)$;

пространственная – $O(0{,}04S^2+NS+5N+5S+4{,}2K\ln S)$.

Для упрощения оценок сложности метода введем следующие допущения. Поскольку $N<<S$, примем, например, $N=0{,}01S$. В простейшем случае $K = 2$. С учетом принятых допущений получим оценки сложности метода:

вычислительной – $O(0{,}0016S^4 + 0{,}0056S^3 + 2{,}87S^2 + 46S\ln S+2S)$;

пространственной – $O(0{,}05S^2+5{,}05S+8{,}4\ln S)$.

Обозначим размерность обучающей выборки $n = NS \approx 0{,}01S^2$, тогда, с учетом принятых допущений, округляя, получим оценки сложности метода:

вычислительной – $O(16n^2+5{,}6n\sqrt{n}+2870n+2{,}3\sqrt{n}\ln n+0{,}2\sqrt{n})$;

пространственной – $O(5n+0{,}505\sqrt{n}+4{,}2\ln n)$.

Эксперименты и результаты

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

Задача	K	N	S^*/S
Классификация автотранспортных средств по изображению [14]	2	26	0,13
Диагностирование патологий плода по кардиотокограмме [15]	3	23	0,09
Предсказание типа лесного покрова [16]	7	54	0,08

Результаты проведенных экспериментов подтвердили работоспособность и практическую применимость предложенного метода, а также реализующего его программного обеспечения.

Как следует из таблицы, использование предложенного метода позволяет в среднем в 10 раз сократить объем выборки, 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 предложено новое решение актуальной научно-практической задачи формирования выборок для автоматизации классификации данных.

Научная новизна результатов работы заключается в том, что впервые предложен метод формирования 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 по исходной выборке, что позволяет существенно сократить объем выборки и 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.

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

ЛИТЕРАТУРА

1. *Олійник А.О., Субботін С.О., Олійник О.О.* Інтелектуальний аналіз даних : навчальний посібник. – Запоріжжя : ЗНТУ, 2012.
2. *Рутковская Д., Пилиньский М., Рутковский Л.* Нейронные сети, генетические алгоритмы и нечеткие системы / пер. с польск. И. Д. Рудинского. – М.: Горячая линия – Телеком, 2004.
3. *Субботин С.А., Олейник Ан.А., Гофман Е.А. и др.* Интеллектуальные информационные технологии проектирования автоматизированных систем диагностирования и распознавания образов. – Харьков: ООО «Компания Смит», 2012.
4. *Богуслаев А.В., Олейник Ал.А., Олейник Ан.А. и др.* Прогрессивные технологии моделирования, оптимизации и интеллектуальной автоматизации этапов жизненного цикла авиационных двигателей. – Запорожье: ОАО "Мотор Сич", 2009.
5. *Субботин С. А.* Формирование выборок и анализ качества моделей на основе нейронных и нейро-нечетких сетей в задачах диагностики и распознавания образов. – Saarbrücken: LAP Lambert academic publishing, 2012.
6. *Jensen R., Shen Q.* Computational intelligence and feature selection: rough and fuzzy approaches. – Hoboken: John Wiley & Sons, 2008.
7. *Chaudhuri A., Stenger H.* Survey sampling theory and methods. – New York: Chapman & Hall, 2005.
8. *Encyclopedia* of survey research methods / ed. P. J. Lavrakas. – Thousand Oaks: Sage Publications, 2008. – Vol. 1-2.
9. *Кокрен У.* Методы выборочного исследования / пер. с англ. И. М. Сонина, под ред. А. Г. Волкова, Н. К. Дружинина. – М.: Статистика, 1976.
10. *Subbotin S.A.* The training set quality measures for neural network learning // 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 // 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 // Біоніка інтелекту. – 2010. – № 1. – С.38-42.
13. *Субботин С.А.* Методы формирования выборок для построения диагностических моделей по прецедентам // 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 // Комп'ютерне моделювання та інтелектуальні системи : зб. наук. праць. – 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.

Статья представлена к публикации членом редколлегии Е.Л. Ереминым.

E-mail:
Субботин Сергей Александрович – subbotin@zntu.edu.ua.

УДК 62-501

© 2013 г. **К.А. Числов**, канд. техн. наук
(Институт автоматики и процессов управления ДВО РАН, Владивосток)

ОПТИМАЛЬНОЕ ОЦЕНИВАНИЕ ПАРАМЕТРОВ ВРАЩЕНИЯ ТЕХНОЛОГИЧЕСКОЙ ПЛАТФОРМЫ*

Предложен и исследован нейроморфный алгоритм оценки параметров вращения подвижной технологической платформы, основанный на интерпретации фильтра Калмана. Представлены результаты численного исследования.
***Ключевые слова*:** астроинерциальная система, технологическая платформа, нейронные сети, синаптические коэффициенты, фильтр Калмана.

Введение

Функционирование многоцелевых подвижных технологических платформ (ТП авиационного, космического или морского базирования) в значительной степени обеспечивается выполнением требуемых условий на движение, что, как известно, достигается управлением по наблюдениям – с помощью обратной связи.

Настоящая работа посвящена модели астроинерциальной системы (АИС)

* Исследование выполнено при поддержке РФФИ-ДВО (грант № 11-01-98501-р_восток_а) и ДВО РАН (грант № 12-1-П17-01).