УДК 004.93

С. А. Субботин
Запорожский национальный технический университет
ул. Жуковского, 64, 69063 Запорожье, Украина

Разбиение исходной выборки большого объема для решения задач диагностики и распознавания образов на основе методов вычислительного интеллекта

Предложен новый метод разбиения исходной выборки на обучающую и тестовую, сохраняющий в сгенерированной подвыборке наиболее важные топологические свойства исходной выборки и не требующий ее загрузки в память. Он обеспечивает последовательную обработку экземпляров, а также выполняет преобразование многомерных координат в одномерные и дискретизацию для улучшения обобщающих свойств. Метод позволяет значительно уменьшить размер выборки и снизить требования к ресурсам компьютера.

Ключевые слова: *выборка, сокращение размерности данных, отбор экземпляров, распознавание, диагностика.*

Введение

Для построения диагностических и распознающих моделей широкое распространение получили методы вычислительного интеллекта (искусственные нейронные и нейро-нечеткие сети, метрические методы обучения распознаванию образов, деревья решений), которые для построения модели требуют наличия обучающей выборки наблюдений-прецедентов [1].

При этом на практике в ряде задач приходится сталкиваться с необходимостью обработки больших объемов имеющихся данных, не позволяющих загружать их полностью в память ЭВМ, а также тем, что время построения модели существенным образом зависит от объема используемой обучающей выборки.

Поэтому актуальной является задача сокращения объема обрабатываемой выборки, решаемая посредством выделения из имеющейся исходной выборки большого объема обучающей и тестовой выборок меньшего размера.

Известные методы формирования выборок на основе переборного [2, 3] и случайного поиска [3, 4] предполагают перебор большого числа возможных комбинаций экземпляров, что для исходных выборок большого объема приводит к т.н. «комбинаторному взрыву». Данные методы также требуют задания критериев, позволяющих оценивать качество разбиений. Несмотря на наличие подоб-

ных критериев [5–7], на практике для больших исходных выборок их применение приводит к значительным затратам времени ЭВМ на их расчет.

Целью данной работы являлось создание метода, позволяющего в автоматическом режиме разбивать исходную выборку большого объема на обучающую и тестовую выборки с учетом ограничений памяти ЭВМ.

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 $\langle x^s, y^s \rangle$, $x^s = \{x_j^s\}$, 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 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 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.

Метод формирования выборок

Наиболее очевидным методом выделения наиболее значимых экземпляров из исходной выборки данных является реализация ее кластер-анализа [8] с последующим определением экземпляров, находящихся на границах классов. Однако данный метод характеризуется рядом существенных недостатков.

Первым недостатком данного метода является его практическая применимость в основном для небольших по объему выборок вследствие необходимости расчета и хранения в памяти ЭВМ матриц расстояний между экземплярами для выбора. Поэтому для выборок наблюдений большого объема предлагается реализовать последовательную обработку экземпляров так, чтобы не нужно было хранить расстояния между всем экземплярами и тем самым обеспечить экономию памяти ЭВМ.

Вторым недостатком рассмотренного метода является сложность определения экземпляров, находящихся на границах классов в многомерном пространстве признаков. Поэтому для устранения данного недостатка предлагается заменить многомерный набор координат на одномерный, который также дискретизировать для повышения обобщающих свойств метода.

Третьим недостатком метода является неопределенность числа кластеров при осуществлении кластер-анализа. Как правило, в большинстве задач пользователь не может знать заранее число кластеров, а его автоматическое определение тре-

бует перебора большого числа вариантов разбиений, а также расчета и хранения в памяти матрицы расстояний между всем экземплярами. Для устранения данного недостатка предлагается вначале эвристически определить ограничения на число кластеров и задать координаты их центров так, чтобы покрыть все участки пространства признаков, после чего выполнить распознавание экземпляров исходной выборки относительно набора центров кластеров, присваивая центрам кластеров номера классов ближайших к ним экземпляров, а при возникновении коллизий — ситуаций, когда центр ближайшего к распознаваемому экземпляру кластера относится к отличному от него классу — формировать новые кластеры, занося в их центры координаты распознаваемых экземпляров, вступающих в коллизию с центром ближайшего кластера. После чего для сформированного набора кластеров предлагается выполнить объединение всех ближайших друг к другу кластеров, принадлежащих к одному и тому же классу. Это позволит, с одной стороны, обеспечить изначально более высокий уровень обобщения данных по сравнению с четким кластер-анализом с сокращением числа кластеров, а, с другой стороны — ускорить вычисления по сравнению с четким кластер-анализом с сокращением числа кластеров.

Рассмотренные идеи лежат в основе предлагаемого метода формирования выборок.

Этап инициализации. Задать исходную выборку X. Определить минимальное $x_j^{\min}$ и максимальное $x_j^{\max}$ значения каждого j-го признака, $j = 1, 2, ..., N$. Создать Q центров кластеров $C^q = \{C_j^q\}$, $q = 1,2,...,Q$, $j = 1, 2, ..., N$, где C_j^q — значение j-го признака для центра q-го кластера: $x_j^{\min} \le C_j^q \le x_j^{\max}$. При этом следует задать $K \le Q << S$.

В простейшем случае C_j^q можно сгенерировать как псевдослучайные числа, однако более предпочтительным является использование формулы:

$$C_j^q = \begin{cases} \dfrac{1}{S^q}\sum\limits_{s=1}^{S}\{x_j^s \mid y^s = q\}, q \le K; \\ C_j^{q-K} + (rand_1 - rand_2)(x_j^{\max} - x_j^{\min}), x_j^{\min} < C_j^{q-K} + \\ +(rand_1 - rand_2)(x_j^{\max} - x_j^{\min}) < x_j^{\max}, K < q; \\ x_j^{\min} + rand_1(x_j^{\max} - x_j^{\min}), \text{ в противном случае,} \end{cases}$$

где $rand_1$, $rand_2$ — два случайных числа: $rand_1$, $rand_2 \in [0, 1]$; Q задается пользователем, либо автоматически выбирается по формуле

$$Q = \begin{cases} K, K \ge \text{round}(\ln(S)); \\ \text{round}(\ln(S)), \text{round}(\ln(S)) > K, \end{cases}$$

где round — функция округления к ближайшему целому числу.

После чего установить для каждого кластера C^q, q = 1,2,...,Q, номер класса: $Y^q = 0$, а также число экземпляров исходной выборки, попавших в кластер: $S_q = 0$.

Этап формирования разбиения выборки на кластеры. Просматривая исходную выборку для каждого экземпляра x^s, s = 1,2,...,S:

— определить расстояние от экземпляра x^s до каждого центра кластера C^q:

$$R(x^s, C^q) = R(C^q, x^s) = \sqrt{\sum_{j=1}^{N} (x_j^s - C_j^q)^2}\ , q = 1, 2, ..., Q;$$

— найти номер ближайшего к экземпляру x^s центра кластера:

$$q = \arg \min_{g=1,2,...,Q} \{R(x^s, C^g)\};$$

— если центр не имеет метки класса ($Y^q = 0$), то присвоить ему класс экземпляра x^s: $Y^q = y^s$, $S_q = 1$;

— если класс экземпляра x^s и центра кластера C^q совпадают ($Y^q = y^s$), то принять: $S_q = S_q + 1$;

— если класс экземпляра x^s и центра кластера C^q не совпадают ($Y^q \neq y^s$), то добавить новый кластер: $Q = Q + 1$, $C^q = x^s$, $Y^q = y^s$, $S_q = 1$.

Этап редукции множества кластеров. Просматривая множество сформированных центров кластеров для q-го кластера, q = 1,2,...,Q:

— если число попавших в него экземпляров равно нулю ($S_q = 0$), то удалить q-й кластер, скорректировав соответствующим образом число кластеров: $Q = Q - 1$, а также перенумеровав элементы $\{Y^q\}$ и $\{S_q\}$ и перейти к следующему кластеру;

— найти расстояния от q-го кластера до всех остальных кластеров:

$$R(C^q, C^g) = R(C^g, C^q) = \sqrt{\sum_{j=1}^{N} (C_j^g - C_j^q)^2}\ , g = 1, 2, ..., Q, g \neq q;$$

— найти ближайший к q-му кластер C^p:

$$p = \arg \min_{g=1,2,...,Q} \{R(C^q, C^g)\};$$

— если $Y^q = Y^p$, то объединить q-й и p-й кластеры по формуле:

$$C_j^q = \frac{C_j^q + C_j^p}{2}, j = 1, 2, ..., N,$$

либо с учетом числа их экземпляров по формуле:

$$C_j^q = \frac{S_q C_j^q + S_p C_j^p}{S_q + S_p}, j = 1, 2, ..., N,$$

после чего скорректировать число кластеров: $Q = Q - 1$, а также перенумеровать элементы $\{Y^q\}$ и $\{S_q\}$ и перейти к следующему кластеру.

Этап разрешения коллизий — добавления кластеров. Установить $S_q = 0$. Просматривая исходную выборку для каждого экземпляра x^s, $s = 1, 2, ..., S$:

— определить расстояния от него до центра каждого кластера:

$$R(x^s, C^q) = R(C^q, x^s) = \sqrt{\sum_{j=1}^{N} (x_j^s - C_j^q)^2}, q = 1, 2, ..., Q;$$

— найти номер ближайшего к экземпляру x^s центра кластера:

$$q = \arg \min_{g=1,2,...,Q} \{R(x^s, C^g)\};$$

— если он принадлежит к тому же классу, что и экземпляр x^s ($Y^q = y^s$), то установить в качестве координаты экземпляра $x_*^s = q$, $y_*^s = Y^q$, $S_q = S_q + 1$;

— если класс экземпляра x^s и ближайшего к нему кластера C^q не совпадают ($Y^q \neq y^s$), то добавить новый кластер: $Q = Q + 1$, $C^q = x^s$, $Y^q = y^s$, $S_q = 1$, установить в качестве координаты экземпляра на обобщенной оси: $x_*^s = q$, $y_*^s = Y^q$.

Этап оценки индивидуальной информативности признаков. Вначале упорядочить (перенумеровав) экземпляры $\{< x_*^s, y_*^s\}$ по обобщенной оси в порядке возрастания значения x_*. Затем последовательно просматривая экземпляры по обобщенной оси x_*^s, $s = 1,2,...,S$:

— найти минимальное и максимальное значения каждого признака в исходном пространстве признаков для экземпляров с одинаковой координатой по обобщенной оси отдельно для каждого класса:

$$x_j^{\min,q} = \min_{p=1,2,..,S} \{x_j^p \mid y^p = q, x_*^s = x_*^p\}, \; x_j^{\max,q} = \max_{p=1,2,..,S} \{x_j^p \mid y^p = q, x_*^s = x_*^p\};$$

— по каждому j-му признаку, $j = 1,2,...,N$, определить число экземпляров каждого q-го класса S_j^{q+} и число экземпляров других классов S_j^{q-}, попадающих по j-му признаку в интервал его значений для q-го класса в исходном пространстве признаков для конкретной координаты по обобщенной оси:

$$S_j^{q+} = \sum_{s=1}^{S}\sum_{p=s}^{S}\{1 \mid x_j^{\min,q} \le x_j^s \le x_j^{\max,q}, x_j^{\min,q} \le x_j^p \le x_j^{\max,q}, y_*^s = q, y_*^p = q, x_*^s = x_*^p\},$$

$$S_j^{q-} = \sum_{s=1}^{S}\sum_{p=s+1}^{S}\{1 \mid x_j^{\min,q} \le x_j^s \le x_j^{\max,q}, x_j^{\min,q} \le x_j^p \le x_j^{\max,q}, y_*^s = q, y_*^p \ne q, x_*^s = x_*^p\}.$$

После этого для каждого j-го признака, $j = 1, 2, ..., N$:

— определить вес (индивидуальную оценку значимости) j-го признака для q-го кластера, $q = 1, 2, ...,Q$:

$$w_j^q = \begin{cases} 1, S_j^{q-} = 0, \\ \dfrac{S_j^{q+}}{S_j^{q-}}, S_j^{q-} > 0; \end{cases}$$

— определить вес (индивидуальную оценку значимости) j-го признака на всем множестве кластеров:

$$w_j = \max_{q=1,2,...,K}\{w_j^q\} \text{ либо } w_j = \frac{\max\limits_{q=1,2,...,K}\{w_j^q\}}{\max\limits_{\substack{q=1,2,...,K \\ j=1,2,...,N}}\{w_j^q\}} \text{ либо } w_j = \frac{\frac{1}{K}\sum\limits_{q=1}^{K} w_j^q}{\max\limits_{\substack{q=1,2,...,K \\ j=1,2,...,N}}\{w_j^q\}}.$$

Этап разбиения исходной выборки на обучающую и тестовую выборки:

— создать пустые обучающую X^* и тестовую X' выборки: $X^* = \varnothing$, $X' = \varnothing$;

— найти все экземпляры в исходном пространстве признаков с одинаковой координатой по обобщенной оси, но разными номерами класса и занести их в формируемую обучающую выборку X^*:

$$X^* = \bigcup_{p=1}^{S}\{< x^p, y^p > \mid < x^p, y^p > \notin X^*, x_*^p = x_*^s, y^p \ne y^s\}, s = 1,2,...,S;$$

— найти все экземпляры в исходном пространстве признаков с одинаковой координатой по обобщенной оси и одинаковыми номерами классов и занести из них в формируемую обучающую выборку тот, который ближе всего по расстоянию к центру соответствующего кластера:

$$X^* = X^* \cup < x^q, y^q >,$$

$$q = \arg\min_{g=1,2,...,Q}\{R(x^p, C^g) \mid < x^p, y^p > \in \Omega^s\},$$

$$R_w(x^s, C^q) = R_w(C^q, x^s) = \sqrt{\sum_{j=1}^{N} w_j (x_j^s - C_j^q)^2},$$

$$\Omega^s = \bigcup_{p=1}^{S} \{< x^p, y^p > | < x^p, y^p > \notin X^*, x_*^p = x_*^s, y^p = y^s\}, s = 1,2,...,S;$$

— оставшиеся экземпляры занести в тестовую выборку: $X' = X \setminus X^*$.

Разработанный метод формирования выборок позволяет выделить из исходной выборки большого объема обучающую и тестовую выборки, обеспечивая при этом незначительное число проходов по исходной выборке и не требуя ее загрузки в память ЭВМ, а также хранения в памяти матрицы расстояний между экземплярами исходной выборки.

Дополнительным результатом работы метода являются координаты сформированных центров кластеров, которые можно использовать для задания последующего построения диагностических и распознающих моделей. Также полученные индивидуальные оценки значимости признаков позволяют рассматривать предложенный метод не только как метод отбора экземпляров, но и как метод оценки информативности признаков. Использование оценок информативности признаков возможно в методах отбора признаков, а также в некоторых методах построения диагностических и распознающих моделей.

Анализ сложности метода формирования выборок

Для оценки временной и пространственной сложности предложенного метода будем исходить из его реализации на основе ЭВМ с последовательными вычислениями, а размерность памяти будем исчислять в ячейках, содержащих вещественные числа. При этом оценки сложности будем давать в т.н. «мягком» виде, когда отсутствует подавление слагаемых меньшего порядка слагаемыми больших порядков.

Для этапа инициализации временная сложность составит $O(6NS + 2Q)$, а пространственная сложность — $O(NS + NQ + 2N + 2Q)$. Для этапа формирования разбиения выборки на кластеры временная сложность составит $O(2SQN+4SQ)$, а пространственная сложность — $O(NQ)$. Для этапа редукции множества кластеров временная сложность составит $O(Q + 2Q^2N + Q^2 + QN)$, а пространственная сложность — $O(Q^2)$. Для этапа разрешения коллизий — добавления кластеров временная сложность составит $O(SNQ + SQ + 5S)$, а пространственная сложность — $O(NQ + 3S)$. Для этапа оценки индивидуальной информативности признаков временная сложность составит $O(2SNQ + 20NS^2 + 2N + KN + KN^2)$, а пространственная сложность — $O(2NQ + 2NQ + NQ + N)$. Для этапа разбиения исходной выборки на обучающую и тестовую выборки временная сложность составит $O(4S^2 + 4S^2 + 4QN + Q)$, а пространственная сложность — $O(2S + 2S + Q)$. При этом не учтен объем памяти для сформированных обучающей и тестовой выборок, которые можно хранить как в оперативной, так и на внешней памяти ЭВМ. Очевидно, что суммарный объем памяти для хранения сформированных выборок не превысит размерности исходной выборки NS.

Таким образом, суммарная временная сложность метода составит $O(6NS + 4Q + 5SQN + 5SQ + 5S + 2Q^2N + Q^2 + 5QN + 20NS^2 + 2N + KN + KN^2 + 8S^2)$, а пространственная сложность — $O(NS + 8NQ + 3N + 3Q + Q^2 + 7S)$.

Для упрощения аналитических оценок примем оправданные с практической точки зрения обозначения и соотношения параметров: $n = NS$, $K = 2$, $N \approx 0{,}25S \approx \approx 2\sqrt{n}$, $Q \approx 0{,}25S \approx 0{,}0625N \approx 0{,}125\sqrt{n}$.

В результате с учетом принятых допущений и округляя, получим относительно размерности входа задачи n оценки сложности предложенного метода: временной — $O(162{,}0625\ n\sqrt{n} + 148{,}25n + 56\sqrt{n})$ и пространственной — $O(3{,}015625n + 68\sqrt{n})$.

Эксперименты и результаты

Для проверки практической применимости предложенного метода была разработана его программная реализация, которая использовалась при решении практических задач диагностирования и распознавания образов [9–11]. Характеристики исходных выборок для решавшихся задач, а также результаты проведенных экспериментов по исследованию предложенного метода приведены в таблице. Здесь $n_{об.}$ — размерность сформированной обучающей выборки.

Характеристики исходных выборок и результаты экспериментов по формированию выборок

Название задачи	N	S	K	n	$n_{об.}/n$
Прогнозирование повышения поверхностной прочности лопаток газотурбинных авиадвигателей (с дискретизированным выходом) [9]	12	59	2	708	0,39
Автоматическая классификация транспортных средств по изображению (в системе исходных и конструируемых признаков) [10]	4122	1062	3	4377564	0,23
Распознавание сельскохозяйственных растений на культурные и сорные [11, 12]	256	3226	2	825856	0,12

Проведенные эксперименты подтвердили работоспособность предложенного метода и реализующего его программного обеспечения.

Как видно из таблицы, предложенный метод позволяет существенно сокращать размерность обучающих данных, экономя тем самым время на последующее построение моделей на основе методов вычислительного интеллекта, а также обеспечивая повышение обобщающих свойств синтезируемых моделей относительно размерности исходных выборок данных.

Разработанный метод, кроме формирования выборок, также позволяет определять индивидуальные оценки информативности признаков, которые могут быть использованы для сокращения размерности обучающих выборок и синтезируемых на их основе моделей.

Выводы

В работе решена актуальная задача автоматизации разбиения исходной выборки на обучающую и тестовую выборки для решения построения диагностических моделей по прецедентам.

Научная новизна результатов работы заключается в том, что впервые предложен метод формирования обучающих и тестовых 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 выборок, а также проведены эксперименты по его 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 направлены на разработку эффективных реализаций предложенного метода для многопроцессорных (многоядерных) ЭВМ, работающих в параллельном режиме.

Работа выполнена в Запорожском национальном техническом университете в рамках научно-исследовательской темы «Интеллектуальные информационные технологии автоматизации проектирования, моделирования, управления и диагностирования производственных процессов и систем» (№ гос. регистрации 0112U005350).

1. *Интеллектуальные* информационные технологии проектирования автоматизированных систем диагностирования и распознавания образов: монография / [С.А. Субботин, Ан.А. Олейник, Е.А. Гофман и др.]; под ред. С.А. Субботина. — Харьков: ООО «Компания Смит», 2012. — 317 с.

2. *Chaudhuri A.* Survey Sampling Theory and Methods / A. Chaudhuri, H. Stenger. — New York: Chapman & Hall, 2005. — 416 p.

3. *Subbotin S.A.* Methods of Sampling Based on Exhaustive and Evolutionary Search / S.A. Subbotin // Automatic Control and Computer Sciences. — 2013. — Vol. 47, N 3. — P. 113–121.

4. *Encyclopedia* of Survey Research Methods / ed. P. J. Lavrakas. — Thousand Oaks: Sage Publications, 2008. — Vol. 1–2. — 968 p.

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. — N 2. — P. 126–139.

6. *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.

7. *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.

8. *Прикладная* статистика: Классификации и снижение размерности: справ. изд. / [С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин]; под ред. С.А. Айвазяна. — М.: Финансы и статистика, 1989. — 607 с.

9. *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.

10. *Субботин С.А.* Автоматическая система обнаружения и распознавания автотранспортных средств на изображении / С.А. Субботин, К.Ю. Бойченко // Программные продукты и системы. — 2010. — № 1. — С. 114–116.

11. *The Plant* Recognition on Remote Sensing Results by the Feed-Forward Neural Networks // [V.I. Dubrovin, S.A. Subbotin, S.V. Morshchavka, D.M. Piza] // Intelligent engineering systems through artificial neural networks. — New York : ASME Press, 2000. — Vol. 10: Smart engeneering systems design: neural networks, fuzzy logic, evolutionary programming, data mining, and complex systems: Artificial neural networks in engineering conference ANNIE–2000, St. Louis, 5–8 November 2000; eds.: C.H. Dagli [et al.]. — P. 697–702.

12. *Шама Є.О.* Побудова класифікатора рослинних об’єктів за допомогою нейронних мереж / Є.О. Шама, С.О. Субботін, С.В. Морщавка // Радіоелектроніка. Інформатика. Управління. — 2013. — № 1. — С. 55–61.

Поступила в редакцию 07.10.2013