УДК 681.518.5:004.93

С.А. СУББОТИН, к.т.н., доц. ЗНТУ, Запорожье

МЕТОДЫ ФОРМИРОВАНИЯ ВЫБОРОК ДЛЯ ПОСТРОЕНИЯ ДИАГНОСТИЧЕСКИХ МОДЕЛЕЙ ПО ПРЕЦЕДЕНТАМ

Решена актуальная задача разработки математического обеспечения для формирования обучающих выборок. Получили дальнейшее развитие переборные и эволюционные методы комбинаторного поиска, которые модифицированы для формирования выборок путем введения разработанных критериев для отбора, цензурирования и псевдокластеризации экземпляров, что позволяет ускорить процесс формирования выборок и обеспечить их соответствие заданным критериям при ограниченном объеме. Определены оценки сложности разработанных методов. Библиогр.: 9 назв.

Ключевые слова: выборка, эволюционные метолды комбинаторного поиска, диагностическая модель.

Постановка проблемы и анализ литературы. Для обеспечения конкурентоспособности и высокого качества выпускаемой продукции, её безотказности в процессе эксплуатации возникает необходимость в своевременном выполнении диагностических процедур, что, в свою очередь, требует наличия диагностической модели [1]. Поскольку в подавляющем большинстве практических задач экспертный опыт является весьма ограниченным, он оказывается недостаточным для построения диагностической модели и возникает необходимость построения модели по набору прецедентов – обучающей выборке, извлекаемой из доступной исследователю исходной выборки.

Традиционным подходом для выделения обучающей выборки из исходной совокупности прецедентов является использование методов формирования случайных выборок [2 – 6], которые обладают такими недостатками, как необходимость задания объема формируемой выборки человеком (неопределенность объема выборки), возможность невключения важных и включения малозначимых экземпляров в формируемую выборку при малом объеме формируемой выборки.

Другим подходом к формированию выборки является использование процедур кластер-анализа [1, 5], позволяющих выделить все основные типы наблюдений. Однако недостатком данных методов является то, что количество кластеров (типов прецедентов) априори неизвестно, а статистические свойства (частоты экземпляров разных кластеров) в сформированной выборке могут не соответствовать исходной. Кроме того, методы [1, 5] могут выделить чрезмерно большое число кластеров, что приведет к избыточности сформированной выборки.

Наиболее точным методом формирования обучающих выборок, способным гарантированно обеспечить наилучшее решение для заданного

критерия качества, является метод полного перебора [1] всех возможных подвыборок исходной выборки. Однако метод является чрезвычайно трудоемким и для выборок большого объема не применим.

Поэтому для формирования выборок необходимо разработать методы, способные в автоматическом режиме выделять из исходной выборки подмножество экземпляров минимального объема, содержащее наиболее важные экземпляры для построения диагностической модели.

Целью данной работы являлось создание методов, позволяющих автоматизировать процесс формирования обучающих выборок для синтеза диагностических моделей по прецедентам.

Постановка задачи. Пусть задана исходная выборка $<X, Y>$ объемом S экземпляров, характеризуемых набором значений N входных (описательных) признаков X и одного выходного (целевого) признака Y. Тогда задачу формирования обучающей выборки $<x, y>$ из исходной выборки $<X, Y>$ можно представить как поиск такого минимального подмножества $<X, Y>$, для которого значение заданного функционала качества $\bar{I}(<x, y>)$ будет иметь максимальное значение. При этом функционал качества $\bar{I}(<x, y>)$ должен отражать требования относительно топологической и статистической репрезентативности формируемой подвыборки $<x, y>$ относительно $<X, Y>$.

Метод формирования выборки на основе сокращенного перебора с цензурированием и псевдокластеризацией. Для устранения недостатков метода полного перебора предлагается исходную выборку разделить на подмножества, расположенные в компактных областях пространства признаков – кластерах, и из каждого такого подмножества извлекать только те экземпляры, которые наиболее перспективны для формирования множества решений. По сути, заменить исходную выборку выборкой меньшего размера, содержащей потенциально наиболее ценные экземпляры. Далее из полученной выборки сформировать множество решений, среди которых отобрать наилучшие путем перебора с цензурированием. Для ускорения поиска предлагается кластерный анализ исходной выборки в многомерном пространстве признаков заменить на псевдокластеризацию, представляемую как объединение результатов частных кластеризаций выборки в одномерных проекциях на оси признаков. Разработанный метод включает следующие этапы.

1. Инициализация: задать исходную выборку $<X, Y>$ объемом S экземпляров, а также максимально допустимый объем $S_{ф.}$ формируемой выборки $<x, y>$. Рассчитать значение критерия качества исходной выборки $\bar{I}$.

2. Псевдокластеризация. Для каждого i-го признака (i = 1, 2, ..., N) выполнить пп. 2.1 – 2.2.

2.1. Упорядочить экземпляры исходной выборки в порядке неубывания значений i-го признака.

2.2. Просматривая упорядоченное множество экземпляров по оси i-го признака слева направо (от меньших значений к большим) попарно для каждых двух соседних экземпляров, включить оба экземпляра в выборку $<X',Y'>$, если они принадлежат к разным классам. Также включить в выборку $<X',Y'>$ крайние левый и правый экземпляры по оси значений i-го признака. При включении экземпляров в выборку $<X',Y'>$ исключить дубляж, для чего перед добавлением нового экземпляра найти расстояния от него до каждого из уже имеющихся в выборке экземпляров и включать экземпляр только тогда, когда минимальное из расстояний больше нуля.

3. Для сформированной выборки $<X', Y'>$ объемом S' сгенерировать все возможные подвыборки $<x(k), y(k)>$, содержащие комбинации не более, чем S_ϕ экземпляров, $S_\phi \le S'$. Здесь k – номер подвыборки, $x(k)$, $y(k)$ – соответственно, экземпляры k-ой выборки и сопоставленные им значения выходного признака.

4. Цензурирование и отбор решений.

4.1. Для $\forall k$ исключить из рассмотрения подвыборки, удовлетворяющие условию: $\exists q$, q = 1, 2, ..., K: $S^q(k) = 0$, где K – количество классов, $S^q(k)$ – количество экземпляров q-го класса в k-ой подвыборке.

4.2. Для оставшихся в рассмотрении подвыборок, исключить из рассмотрения те, которые удовлетворяют условию: $\exists q$, q = 1, 2, ..., K: $\left|S^q(k) - S^q\right| / S > \delta_K$, где δ_K – некоторая заранее заданная константа, $0 < \delta_K < 1$.

4.3. Для всех оставшихся подвыборок рассчитать значения критерия качества $\bar{I}(k)$ и исключить из рассмотрения те из оставшихся подвыборок, для которых: $\bar{I}(k) < \bar{I}^*$, где $\bar{I}^*$ – среднее значение критерия качества для оставшихся подвыборок. Критерий качества выборки можно определить на основе показателей качества, предложенных в [3 – 5].

5. Среди оставшихся подвыборок $\{<x(k), y(k)>\}$ в качестве решения $<x, y>$ выбрать ту подвыборку $<x(p), y(p)>$, которая наилучшим образом соответствует заранее заданному критерию выбора решения. Предлагается использовать один из следующих критериев:

– максимум качества формируемой выборки: $p = \arg\max_k \left\{\bar{I}(k)\right\}$;

– максимум соответствия качества формируемой выборки качеству исходной выборки: $p = \arg\min_k \left\{ \left| \bar{I}(k) - \bar{I} \right| \right\}$;

– минимум объема выборки: $p = \arg\min_k \{S(k)\}$;

– максимум ограниченного объема выборки: $p = \arg\max_k \{S(k)\}$;

– комбинированные критерии: $p = \arg\min_k \left\{ \left| \bar{I}(k) - \bar{I} \right| / S(k) \right\}$;

$$p = \arg\min_k \left\{ \bar{I}(k) / S(k) \right\};\ p = \arg\min_k \left\{ S(k) \left| \bar{I}(k) - \bar{I} \right| \right\};\ p = \arg\min_k \left\{ S(k) \bar{I}(k) \right\}.$$

Достоинством данного метода является то, что он перед формированием подвыборок существенно сокращает размерность исходной выборки за счет исключения малоинформативных экземпляров, сохраняя при этом экземпляры, расположенные на границе разделения классов. Таким образом, с одной стороны, существенно сокращает время поиска решений, а, с другой стороны, сохраняет наиболее важные для построения моделей прецеденты. Недостатком метода является то, что он из-за потери информации вследствие сокращения исходной выборки, может существенно изменить частоты классов в извлекаемых выборках. Другим недостатком метода является то, что информация о качестве уже сгенерированных выборок не учитывается при формировании новых выборок.

Эволюционный метод формирования выборок. Для сокращения числа перебираемых комбинаций рационально обеспечить использование информации об уже проанализированных решениях для перехода к рассмотрению новых решений, похожих на рассмотренные ранее. Также необходимо обеспечить шансы для каждого из возможных решений быть рассмотренным. Для этого предлагается использовать эволюционный подход, представляющий собой разновидность случайного поиска [7].

Метод формирования выборки на основе эволюционного поиска с псевдокластеризацией будет включать следующие этапы.

1. Инициализация: задать исходную выборку <*X*, *Y*> объемом *S* экземпляров, а также максимально допустимый объем $S_{ф.}$ формируемой выборки <*x*, *y*>. Рассчитать значение критерия качества исходной выборки $\bar{I}$ (критерий качества выборки можно определить на основе показателей качества, предложенных в [7 – 9]). Задать размер популяции решений *H*, максимальное число итераций *T*, вероятность мутации $P_{м.}$, а также приемлемое значение критерия качества результата $\bar{I}^* \le \bar{I}$.

2. Псевдокластеризация. Для каждого *i*-го признака (*i* = 1, 2, ..., *N*) выполнить пп. 2.1 – 2.3.

2.1. Упорядочить экземпляры исходной выборки в порядке неубывания значений *i*-го признака.

2.2. Просматривая упорядоченное множество экземпляров по оси *i*-го признаков слева направо (от меньших значений к большим) попарно для каждых двух соседних экземпляров, включить оба экземпляра в выборку <*X*', *Y*'>, если они принадлежат к разным классам. Также включить в выборку <*X*', *Y*'> крайние левый и правый экземпляры по оси значений *i*-го признака. При включении экземпляров в выборку <*X*', *Y*'> исключить дубляж, для чего перед добавлением нового экземпляра найти расстояния от него до каждого из уже имеющихся в выборке экземпляров и включать экземпляр только тогда, когда минимальное из расстояний больше нуля.

3. Формирование начальной популяции решений. Представим *k*-ое решение h^k как бинарную комбинацию из *S* разрядов, *s*-й разряд которой h^k_s определяет включение в решение *s*-го экземпляра исходной выборки (если $h^k_s = 0$, то *s*-й экземпляр не входит в *k*-ое решение, в противном случае, когда $h^k_s = 1$, *s*-й экземпляр входит в *k*-ое решение). Сформируем случайным образом *H* бинарных комбинаций путем выполнения п. 3.1 – 3.2 для $k = 1, 2, ..., H$; $s = 1, 2, ..., S$.

3.1. Задать вероятности включения экземпляров в *k*-ое решение:

$$P(h^k_s) = \begin{cases} 0{,}5(\lambda + \text{rand}), X^s \in < X', Y' >; \\ 0{,}5\text{rand}, X^s \notin < X', Y' >, \end{cases} \quad \lambda = \begin{cases} 1, S' \le S_\phi; \\ \min\{0{,}5; S_\phi / S'\}, S' > S_\phi, \end{cases}$$

где rand – функция, возвращающая случайное число в диапазоне [0, 1].

3.2. Двигаясь от разрядов с большими вероятностями включения экземпляров в *k*-е решение к разрядам с меньшими вероятностями, установить равными единице не более S_ϕ разрядов с наибольшими вероятностями, но не меньшими 0,5, остальные разряды установить равными нулю.

4. Проверка на окончание поиска. Для каждого *k*-го решения популяции сформировать соответствующую выборку, для которой оценить $\bar{I}(k)$. Если выполнено более чем *T* итераций или среди множества решений имеется такое решение с номером *k*, для которого $\bar{I}(k) \ge \bar{I}^*$, то прекратить поиск и вернуть в качестве результата выборку с наибольшим значением критерия качества.

5. Отбор решений для скрещивания. Рассматривая $\bar{I}(k)$ в качестве максимизируемой фитнесс-функции, сформировать родительские пары для производства решений-потомков на основе правила "колеса рулетки"

[7], обеспечивая тем самым учет $\bar{I}(k)$ для оценивания вероятности решения быть допущенным к скрещиванию.

6. Скрещивание. Реализовать скрещивание отобранных решений для производства новых решений на основе одноточечного кроссинговера.

7. Мутация. Для каждого из имеющихся решений инвертировать случайным образом не более round($P_м S$) разрядов, где round – функция округления. Для тех решений, в которых число битов, равных единице, превышает $S_ф$, инвертировать случайным образом лишние единичные биты. Исключить из текущей популяции решения, встречавшиеся ранее на предыдущих циклах работы метода. Перейти к этапу 4.

Данный метод совмещает идеи случайного формирования выборки и детерминированного поиска лучших решений. Он начинает работу с выделения наиболее перспективных для включения в решения экземпляров, однако сохраняет шансы остальных экземпляров войти в формируемые выборки, и в процессе своей работы целенаправленно улучшает рассматриваемые решения. При этом метод гарантирует, что каждая из рассматриваемых выборок будет иметь объем не более $S_ф$.

Анализ сложности методов формирования выборок. Для разработанных методов формирования выборок представляется целесообразным оценить условия их практической применимости. Очевидно, что основными показателями, определяющими сложность методов формирования выборок, являются число генерируемых подвыборок-комбинаций экземпляров исходной выборки, а также количество вычислений интегрального критерия качества. Для сравнения методов формирования выборок оценим их временную сложность.

Метод формирования выборки на основе полного перебора будет иметь сложность порядка $O(G_F G_g)$, где G_F – сложность расчета интегрального показателя качества для выборки (для выборки из S экземпляров завышенная оценка сложности составит $O(S^2)$ при эффективной реализации вычислений), G_g – количество генерируемых выборок. Для данного метода $G_g = 2^S - 1$. Поскольку генерируемые выборки будут содержать разное число экземпляров, для простоты в среднем примем: $G_F = (0{,}5S)^2$, таким образом, получим оценку сложности $O((0{,}5S)^2(2^S - 1))$. Эта оценка свидетельствует о практической пригодности полного перебора только для исходных выборок небольшого объема.

Метод формирования выборки на основе перебора с цензурированием и псевдокластеризацией имеет сложность порядка $O(G_F G_c + G_g)$, где $G_c << G_g$. В худшем случае $G_c = G_g$, а, поскольку, $G_g = 2^{S'} - 1$, то сложность метода можно оценить как $O((2^{S'} - 1)(G_F + 1))$.

Примем приближенно $G_F = S_ф^2 \approx (0{,}5S)^2$, тогда сложность метода может быть оценена как $O((2^{S'} - 1)((0{,}5S)^2+1))$. В наихудшем случае ($S_ф = S$) он будет в $(2^S - 1)(0{,}5S)^2 / ((2^{S'} - 1)((0{,}5S)^2 + 1)) \approx 2^{(S-S')}$ раз быстрее работать по сравнению с методом полного перебора. Данный метод может применяться для больших выборок, поскольку содержит процедуры устранения избыточных прецедентов перед выполнением поиска, который дополнительно еще и цензурируется.

Для эволюционного метода формирования выборки сложность будет зависеть от количества итераций, которое в худшем случае составит T, размера популяции решений H и сложности расчета показателя качества выборки G_F, который примем приближенно $G_f = (0{,}5S)^2$. В результате сложность метода приближенно может быть оценена как $O(THG_F) = O(0{,}25THS^2)$. Таким образом, для данного метода можно определить требования к значениям параметров, обеспечивающим его эффективность относительно:

– полного перебора: $TH << 2^S - 1$;

– перебора с цензурированием и псевдокластеризацией $TH << 2^{S'} - 1$. Только при соблюдении данных условий, предложенный эволюционный метод будет эффективнее этих переборных методов.

Выводы. С целью автоматизации построения диагностических моделей решена актуальная задача разработки математического обеспечения для формирования обучающих выборок.

Научная новизна работы заключается в том, что получили дальнейшее развитие переборные и эволюционные методы комбинаторного поиска, которые модифицированы для формирования выборок путем введения разработанных критериев для отбора, цензурирования и псевдокластеризации экземпляров, что позволяет ускорить процесс формирования выборок и обеспечить их соответствие заданным критериям при ограниченном объеме.

Практическая ценность результатов работы состоит в определении оценок сложности разработанных методов формирования выборок, позволяющих определить условия их применимости на практике. Использование предложенных оценок делает возможным задание критериев качества, учитывающих предпочтения пользователя при формировании выборок и синтезе моделей с учетом имеющихся ресурсов.

Дальнейшие исследования могут быть направлены на разработку процедур цензурирования решений, обеспечивающих большее сокращение пространства поиска в методах формирования выборок.

Список литературы: **1.** Интеллектуальные средства диагностики и прогнозирования надежности авиадвигателей: монография / *В.И. Дубровин, С.А. Субботин, А.В. Богуслаев, В.К. Яценко*. – Запорожье: ОАО "Мотор-Сич", 2003. – 279 с. 2. *Кокрен У.* Методы выборочного исследования / *У. Кокрен*. – М.: Статистика, 1976. – 440 с. **4.** *Bernard H.R.* Social research methods: qualitative and quantative approaches / *H.R. Bernard*. – Thousand Oaks: Sage Publications, 2006. – 784 p. **5.** *Chaudhuri A.* Survey sampling theory and methods / *A. Chaudhuri, H. Stenger*. – New York: Chapman & Hall, 2005. – 416 p. **6.** Multivariate analysis, design of experiments, and survey sampling / [ed. S. Ghosh]. – New York: Marcel Dekker Inc., 1999. – 698 p. **7.** 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. **8.** *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. **9.** *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.

Статья представлена д.т.н., проф. НТУ "ХПИ" Дмитриенко В.Д.

УДК 681.518.5:004.93

Методи формування вибірок для побудови діагностичних моделей за прецедентами / Субботін С.О. // Вісник НТУ "ХПІ". Тематичний випуск: Інформатика і моделювання. – Харків: НТУ "ХПІ". – 2011. – № 17. – С. 149 – 156.

Вирішено актуальне завдання розробки математичного забезпечення для формування навчальних вибірок. Дістали подальшого розвитку переборні й еволюційні методи комбінаторного пошуку, які модифіковані для формування вибірок шляхом уведення розроблених критеріїв для відбору, цензурування та псевдокластеризації екземплярів, що дозволяє прискорити процес формування вибірок і забезпечити їхню відповідність заданим критеріям при обмеженому обсязі. Визначено оцінки складності розроблених методів. Бібліогр.: 9 назв.

Ключові слова: вибірка, еволюційні методи комбінаторного пошуку, діагностична модель.

UDC 681.518.5:004.93

The sampling methods fordiagnostic model construction on precedents / Subbotin S.A. // Herald of the National Technical University "KhPI". Subject issue: Information Science and Modelling. – Kharkov: NTU "KhPI". – 2011. – №. 17. – P. 149 – 156.

The actual problem of mathematical support development for training sample forming is solved. The exhaustive search and evolutionary methods of combinatorial search were further developed. They are modified for sampling by the introduction of the developed criteria for exemplar selection, censoring and pseudo-clustering. This allows to speed up the process of sampling and to ensure its compliance with specific criteria in limited volume. The complexity of the developed methods is estimated. Refs.: 9 titles.

Key words: sample, evolutionary methods of combinatorial search, diagnostic model.

Поступила в редакцию 08.09.2010