УДК 004.93

Субботин С. А.

Д-р техн. наук, профессор, профессор кафедры программных средств Запорожского национального технического университета, Украина

БЫСТРЫЙ МЕТОД ВЫДЕЛЕНИЯ ОБУЧАЮЩИХ ВЫБОРОК ДЛЯ ПОСТРОЕНИЯ НЕЙРОСЕТЕВЫХ МОДЕЛЕЙ ПРИНЯТИЯ РЕШЕНИЙ ПО ПРЕЦЕДЕНТАМ

Решена задача формирования обучающих выборок для автоматизации построения нейросетевых моделей по прецедентам. Предложен метод формирования выборок, который автоматически выделяет из исходной выборки обучающую и тестовую выборки, не требуя загрузки всей исходной выборки в память ЭВМ, осуществляя поэкземплярную обработку исходной выборки с хэширующим преобразованием на одномерную ось, формирует эталоны кластеров на обобщенной оси, минимизируя их число, что позволяет повысить скорость формирования выборок, снизить требования к вычислительным ресурсам и памяти ЭВМ и обеспечить приемлемый уровень точности синтезируемых моделей. Разработанный метод не требует многократных проходов по выборке, ограничиваясь всего тремя просмотрами. При этом метод хранит в оперативной памяти только один текущий экземпляр и набор сформированных одномерных эталонов, который минимизирован по объему. В отличие от методов на основе случайного отбора и кластер-анализа предложенный метод автоматически определяет размер формируемых обучающей и тестовой выборок, не требуя участия человека. Разработано программное обеспечение, реализующее предложенный метод, на основе которого решена практическая задача построения модели принятия решений для индивидуального прогнозирования состояния пациента, больного гипертонией.

Ключевые слова: выборка, формирование выборки, экземпляр, нейронная сеть, индивидуальное прогнозирование, обучение по прецедентам.

НОМЕНКЛАТУРА

ЭВМ – электронная вычислительная машина;

$c_*^q(k)$ – k-й хэш-эталон q-го класса;

C_*^k – набор хэш-эталонов k-го класса;

d – расстояние;

E – ошибка модели;

$E_{об.}$ – ошибка обученной нейромодели при распознавании обучающей выборки;

$E_{расп.}$ – ошибка обученной нейромодели при распознавании тестовой выборки;

f – критерий качества;

$F()$ – структура нейросетевой модели;

i^* – номер ближайшего хэш-эталона к хэшу рассматриваемого экземпляра;

Ind – набор номеров экземпляров класса;

j – номер текущего признака;

K – число классов;

k^q – указатель числа эталонов q-го класса;

M – объем использованной оперативной памяти;

n – размерности входа;

N – число входных признаков;

O – символ Ландау;

opt – условное обозначение оптимума;

Q – число кластеров;

q – номер класса;

$Realmax$ – максимальное вещественное число, представимое в разрядной сетке ЭВМ;

s – номер текущего экземпляра

S – число прецедентов в выборке;

S' – объем сформированной обучающей выборки;

t – время, затраченное на формирование выборки;

$t_{об.}$ – время, затраченное на обучение;

w – набор значений параметров нейронной сети;

X – исходная выборка;

x' – набор входных признаков в обучающей выборке;

x – набор входных признаков в исходной выборке;

x_j – j-й входной признак в исходной выборке;

$x_j^{\max}$ – максимальное значение j-го признака;

$x_j^{\min}$ – минимальное значение j-го признака;

x^s – s-й экземпляр выборки;

x_*^s – хэш s-го экземпляра;

x_j^s – значение j-го входного признака для s-го прецедента;

$X_{об.}$ – обучающая выборка;

$X_{тест.}$ – тестовая выборка;

y – выходной признак в исходной выборке;

y^s – значение выходного признака для s-го прецедента;

y' – выходной признак в обучающей выборке;

χ – вычислительная сложность хэш-преобразования одного экземпляра.

ВВЕДЕНИЕ

Для автоматизации поддержки принятия решений в диагностике возникает необходимость построения диагностических моделей. Зачастую на практике из-за отсутствия или недостатка экспертных знаний построение диагностических моделей предполагает использование экспериментальных наблюдений за состоянием диагностируемого объекта, в процессе которого с помощью методов искусственного интеллекта осуществляется извлечение знаний из данных. Искусственные нейронные и нейро-нечеткие сети [1] являются наиболее широко применяемым классом методов искусственного интеллекта при построении моделей по прецедентам.

DOI 10.15588/1607-3274-2015-1-6

Объектом исследования является процесс построения диагностических моделей на основе нейронных сетей.

Методы обучения нейросетей [1, 2], как правило, характеризуются высокой итеративностью, а также значительными затратами времени на обучение нейросетей при большой размерности обучающих данных. В таких случаях применение нейросетевых технологий оказывается затруднительным. Это вызывает необходимость сокращения объема данных, используемых при обучении нейромоделей.

Предмет исследования составляют методы формирования обучающих выборок из исходных выборок большого объема для обучения нейросетевых моделей.

Целью данной работы являлась разработка метода, позволяющего сократить время обучения нейросетей при построении модели по большому объему прецедентов за счет разбиения исходной выборки большого объема на обучающую и тестовую выборки, обеспечивая минимизацию объема обучающей выборки и ее топологическую репрезентативность относительно исходной выборки.

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 синтеза нейросетевой модели зависимости $y(x)$ будет заключаться в определении таких структуры $F()$ и значений параметров w нейронной сети, при которых будет удовлетворен критерий качества модели $f(F(), w, \langle x, y\rangle) \rightarrow opt$, где opt – условное обозначение оптимума [1]. Обычно критерий качества обучения нейросетей определяют как функцию ошибки модели:

$$\overline{E} = \frac{1}{2}\sum_{s=1}^{S}(y^s - F(w, x^s))^2 \rightarrow \min.$$

Для задач с дискретным выходом ошибку обученной сети можно характеризовать также формулой:

$$E = \frac{100\%}{S}\sum_{s=1}^{S}\left|y^s - F(w, x^s)\right| \rightarrow \min.$$

В случае, когда исходная выборка имеет большую размерность, перед построением нейромодели необходимо решить задачу выделения обучающей выборки меньшего объема (дано: $\langle x, y\rangle$, надо: $\langle x', y'\rangle$, $x' \in \{x^s\}$, $y'=\{y^s | x^s \in x'\}$, $S'=|y'|$, $S'<S$, $f(\langle x', y'\rangle, \langle x, y\rangle) \rightarrow opt$).

Для оценки качества сформированной выборки возможно использовать широкий набор предложенных показателей [3–5]. Однако их расчет требует существенных затрат вычислительных ресурсов, поэтому для упрощения расчетов данную задачу можно рассматривать в конструктивистской постановке (дано: $\langle x, y\rangle$, надо: $\langle x', y'\rangle$, $x' \in \{x^s\}$, $y'=\{y^s | x^s \in x'\}$, $S'=|y'|$, $S'<S$).

2 ЛИТЕРАТУРНЫЙ ОБЗОР

Методы извлечения выборок [6–12] выделяют: вероятностные и детерминированные.

Вероятностные методы [6–8, 12] предполагают случайное извлечение набора экземпляров из исходной выборки, причем каждый экземпляр исходной выборки имеет ненулевую вероятность, которая может быть точно определена, быть включенным в формируемую выборку. К вероятностным методам извлечения выборок относят:

– простой случайный отбор (simple random sampling): из исходной выборки случайным образом отбирается заданное число экземпляров;

– систематический отбор (systematic sampling): исходная выборка упорядочивается определенным образом и разбивается на последовательные группы экземпляров, в каждой из которых выбирается для включения в формируемую выборку объект с заданным порядковым номером в группе;

– стратифицированный отбор (stratification sampling): исходная выборка разделяется на непересекающиеся однородные подмножества – страты, представляющие все виды экземпляров, в каждом из которых применяется случайный или систематический отбор;

– вероятностный пропорциональный к объему отбор (probability proportional to size sampling): используется, когда имеется «вспомогательная переменная» или «метрика объема», которая предполагается связанной с интересующей переменной для каждого экземпляра, вероятность выбора для каждого элемента исходной выборки будет пропорциональна его метрике объема;

– отбор на основе кластер-анализа (cluster sampling): исходная выборка разделяется на кластеры, из группы экземпляров каждого кластера случайно выбирается некоторое подмножество экземпляров для формируемой выборки.

Достоинствами данных методов являются их относительная простота и возможность оценки ошибки выборки, а недостатками – то, что они не гарантируют, что сформированная выборка малого объема будет хорошо отображать свойства исходной выборки, а также не будет избыточной и не будет искусственно упрощать задачу.

Детерминированные методы формирования выборок [6, 9–11] предполагают извлечение экземпляров на основе предположений об их полезности (информативности), при этом некоторые экземпляры могут не быть выбраны или вероятность их выбора не может быть точно определена; они, как правило, основаны на кластерном анализе и стремятся обеспечить топологическое подобие исходной выборке. К детерминированным методам формирования выборок относят методы:

– удобного отбора (convenience sampling): формирует нерепрезентативную выборку из наиболее легко доступных для исследования объектов;

– квотного отбора (quota sampling): исходная выборка разделяется на отличающиеся свойствами подгруппы, после чего из каждой подгруппы выбираются объекты на основе заданной пропорции);

– целевого отбора (judgmental (purposive) sampling): объекты извлекаются из исходной выборки исследователем в соответствии с его мнением относительно их пригодности для исследования.

Недостатком данных методов является невозможность оценивания ошибки сформированных выборок. Достоинством детерминированных методов является то, что они могут выявить наиболее значимые для решения задачи построения диагностической модели прецеденты, которые также могут быть использованы для инициализации распознающих моделей и ускорения процесса обучения. Поэтому для достижения цели, поставленной в работе, в качестве базиса для формирования выборок предлагается выбрать детерминированные методы.

Однако следует отметить, что детерминированные методы, основанные на кластерном анализе, являются сложно применимыми для выборок большого объема, поскольку предполагают расчет расстояний между всеми экземплярами и манипуляции с матрицей расстояний. Следовательно, для повышения скорости обработки данных необходимо сократить объем вычислений за счет исключения необходимости расчета расстояний, что также позволит сократить требования к памяти ЭВМ.

3 МАТЕРИАЛЫ И МЕТОДЫ

Для того, чтобы при ограниченном объеме оперативной памяти ЭВМ обеспечить обработку исходной выборки большой размерности, предлагается осуществлять поэкземплярную обработку исходной выборки, загружая в память только один текущий экземпляр на каждой итерации. При этом заменять N-мерное представление экземпляра на одномерное посредством отображения его координат на обобщенную ось с использованием хэширующего преобразования, сохраняющего топологию исходного пространства признаков в синтезируемом одномерном пространстве.

Для экономии как ресурсов памяти, так и вычислительных ресурсов предлагается вместо расчета расстояний между всеми экземплярами целесообразно оперировать только расстояниями между текущим рассматриваемым экземпляров и сформированными центрами кластеров, причем в одномерном пространстве хэш-кодов.

Для сокращения влияния человеческого фактора на результаты формирования выборки число эталонов предлагается не задавать, а определять автоматически, начиная с одного и наращивая по мере необходимости.

С учетом изложенных выше идей, предложенный метод может быть представлен следующим образом.

Этап инициализации. Установить: s=1, $C_*^{y^s}=\varnothing$, k^q=0, $q=1,...,K$.

Этап хэширующего преобразования. Если $s<S$, тогда считать с внешнего носителя памяти очередной экземпляр выборки x^s в оперативную память. Определить на основе значений признаков экземпляра x^s его хэш x_*^s, используя одно из преобразований, предложенных в [13], в противном случае – перейти к этапу разбиения выборки.

Этап формирования набора хэш-эталонов (хэшей центров кластеров). Если в наборе хэш-эталонов для класса y^s нет ни одного эталона, т. е. $C_*^{y^s}=\varnothing$ и $k^{y^s}=0$, тогда записать хэш текущего экземпляра как эталон в набор хэш-эталонов для класса y^s по формуле (1):

$$k^{y^s}=k^{y^s}+1,\ c_*^{y^s}(k^{y^s})=x_*^s, C_*^{y^s}=C_*^{y^s}\cup c_*^{y^s}(k^{y^s}), \quad (1)$$

записать номер текущего экземпляра s в набор номеров экземпляров класса y^s: $Ind(y^s,k^{y^s})=s$, после чего найти расстояние от нового хэш-эталона до существующих хэш-эталонов всех кластеров всех классов по формуле (2):

$$d(c_*^{y^s}(k^{y^s}),c_*^{y^s}(k))=d(c_*^{y^s}(k),c_*^{y^s}(k^{y^s}))=$$
$$=\left|c_*^{y^s}(k^{y^s})-c_*^{y^s}(k)\right|, k=1,...,|C_*^q|, q=1,...,K, \quad (2)$$

установить: $d(c_*^{y^s}(k^{y^s}),c_*^{y^s}(k^{y^s}))=Realmax$ и перейти к этапу обработки нового экземпляра.

В случае если набор хэш-эталонов непустой ($C_*^{y^s}\neq\varnothing$), тогда определить расстояния от хэша x_*^s рассматриваемого экземпляра x^s до хэш-эталонов всех кластеров данного класса:

$$d(x_*^s,c_*^{y^s}(k))=\left|x_*^s-c_*^{y^s}(k)\right|, k=1,...,|C_*^{y^s}|,$$

затем найти среди имеющихся в наборе хэш-эталонов кластеров класса y^s, номер ближайшего хэш-эталона к хэшу x_*^s рассматриваемого экземпляра x^s:

$$i^*=\arg\min_k\{d(x_*^s,c_*^{y^s}(k))\}, k=1,...,|C_*^{y^s}|,$$

после чего если $d(x_*^s,c_*^{y^s}(i^*))<d(c_*^{y^s}(i^*),c_*^q(p))$, $q=1,...,K$, $p=1,...,|C_*^q|$, тогда пропустить s-й экземпляр и перейти к этапу обработки нового экземпляра, в противном случае – добавить новый хэш-эталон на основе s-го экземпляра по формуле (1), записать номер текущего экземпляра s в набор номеров экземпляров класса y^s: $Ind(y^s,k^{y^s})=s$, после чего найти расстояние от нового хэш-эталона до существующих хэш-эталонов всех кластеров всех классов по формуле (2), установить: $d(c_*^{y^s}(k^{y^s}),c_*^{y^s}(k^{y^s}))=Realmax$ и перейти к этапу обработки нового экземпляра.

Этап обработки нового экземпляра. Установить s=s+1, перейти к этапу хэширующего преобразования.

Этап разбиения выборки. Все экземпляры, исходной выборки, на основе которых сформированы хэш-эталоны кластеров, занести в обучающую выборку $X_{\text{об.}}$, а остальные – в тестовую выборку $X_{\text{тест.}}$:

$$X_{\text{об.}}=\{<x^s,y^s>|\ Ind(y^s,k^{y^s})=s, s=1,...,S, k=1,...,|C_*^{y^s}|\},$$
$$X_{\text{тест.}}=X\setminus X_{\text{об.}}$$

Предложенный метод позволяет загружать в оперативную память ЭВМ полное признаковое описание только одного текущего экземпляра и не требует расчета матрицы расстояний между экземплярами, заменяя их расстояниями от рассматриваемого экземпляра до одномерных хэш-эталонов, что позволяет существенным образом снизить требования к ресурсам оперативной памяти и осуществлять обработку выборок очень большого размера.

Для предложенного метода вычислительная сложность в так называемом «мягком смысле» может быть оценена как $O(2NS+4SQ+S\chi)$, где χ – вычислительная сложность хэш-преобразования одного экземпляра, которая, очевидно, является функцией числа признаков N. Исходя из практического опыта, положим, что $\chi=N$, $Q = \ln S$, $N=0{,}01S$. Тогда вычислительная сложность метода может быть оценена как $O(0{,}01S^2+4S \ln S)$.

Пространственная сложность метода может быть оценена как $O(NS+SK+3N+Q^2+Q)$. Приняв $K=2$ и $Q = \ln S$, $N=0{,}01S$, получим оценку пространственной сложности $O(0{,}01S^2+2{,}03S+(\ln S)^2+\ln S)$.

В терминах размерности входа $n=NS\approx 0{,}01S^2$, приняв для упрощения $\ln S\approx 0{,}5N\approx 0{,}05S$, мы получим грубую оценку вычислительной сложности предложенного метода порядка $O(21n)$ и грубую оценку пространственной сложности – $O(n+2{,}08S+0{,}0025S^2)$.

4 ЭКСПЕРИМЕНТЫ

Для проверки работоспособности предложенного метода он был программно реализован как дополнение к компьютерной программе «Автоматизированная система синтеза нейросетевых и нейро-нечетких моделей для неразрушающей диагностики и классификации образов по признакам» (Свидетельство о регистрации авторского права на произведение № 35431 от 21.10.2010).

Разработанное математическое обеспечение использовалось для проведения экспериментов по решению задачи индивидуального прогнозирования состояния здоровья больного гипертонической болезнью по результатам экспериментально полученных наблюдений за состоянием здоровья пациента и погодными условиями.

Исходная выборка данных была получена в г. Запорожье (Украина) и содержала наблюдения за период с 2002 г. по 2009 г., где каждый экземпляр представлял собой набор данных, характеризующих состояние пациента в определенную часть дня. В качестве временных характеристик использовались дата (год, месяц, день), код дня недели, время (час) наблюдения, код части дня (0 – утро, 1 – вечер). В качестве объективных клинико-лабораторных показателей использовались наблюдаемое артериальное давление (систолическое и диастолическое, мм. рт. ст.), пульс (ударов в минуту), сведения о приеме лекарств (Амло (0 – нет, 1 – да), Эгилок (0 – нет, 1 – да), Берлиприл (0 – нет, 1 – да)), В качестве субъективных показателей использовались характеристики самочувствия (наличие экстросистолы (0 – нет, 1 – есть), наличие боли в голове (0 – нет, 1 – есть), наличие боли в затылке (0 – нет, 1 – есть), наличие пульсации (0 – нет, 1 – есть), наличие боли в левом боку (0 – нет, 1 – есть), наличие боли в области сердца (0 – нет, 1 – есть), нехватка воздуха (0 – нет, 1 – есть), наличие боли в животе (0 – нет, 1 – есть), общая слабость (0 – нет, 1 – есть)). В качестве метеорологических характеристик [14] использовались (температура воздуха (°С), атмосферное давление (мм. рт. ст.), тип облачности (0 – нет, 1 – малая, 2 – облачно, 3 – пасмурно), наличие грозы (0 – нет, 1 – есть), направление ветра (0 – штиль, 1– северный, 2 – северо–восточный, 3 – восточный, 4 – юго–восточный, 5 – южный, 6 – юго–западный, 7 – западный, 8 – северо-западный), скорость ветра (м/с), данные солнечной активности (Mg II индекс [15]). Фрагмент исходных данных в графическом виде представлен на рис. 1.

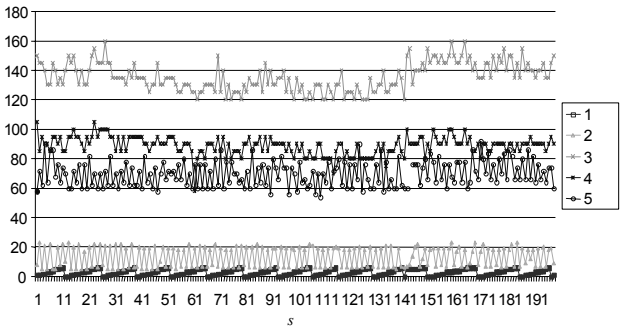

Рисунок 1 – Визуализация фрагмента выборки данных: 1 – день недели, 2 – время, 3 – систоличесокое давление, 4 – диастолическое давление, 5 – пульс

Полученные наблюдения методом «окон» были использованы для получения выборки для решения задачи качественного прогнозирования состояния пациента на ближайшую вторую половину суток по данным предыдущих наблюдений: в качестве входных признаков использовались данные за предыдущие (утро и вечер) и текущие сутки (утро), а в качестве выхода – состояние пациента вечером в текущие сутки (0 – нормальное, 1 – ухудшение состояния, сопровождающееся повышением артериального давления).

Сформированная выборка использовалась для формирования обучающей и контрольной выборок на основе предложенного метода, а также ряда известных методов формирования выборок.

После чего для каждой из сформированных выборок строились прогнозирующие модели на основе трехслойных нейронных сетей прямого распространения сигнала. Каждая нейронная сеть содержала на входе $N = 3\times 26 = 78$ признаков, число нейронов входного слоя – 10, число нейронов среднего слоя – 5, число нейронов выходного слоя – 1. Все нейроны использовали весовую (дискриминантную) функцию взвешенная сумма, а функцию активации – тангенциальный сигмоид.

Перед подачей на входы сети сигналы нормировались по формуле: $x_j^s = \dfrac{x_j^s - x_j^{\min}}{x_j^{\max} - x_j^{\min}}$.

Обучение нейросетей осуществлялось на основе метода Левенберга-Марквардта [1, 2].

5 РЕЗУЛЬТАТЫ

Результаты проведенных экспериментов представлены в табл. 1.

Как видно из табл. 1, предложенный метод позволяет существенно сократить объем обучающей выборки, обеспечивая при этом высокую точность как обучения, так и распознавания (качественного прогнозирования).

Разработанный метод не требует многократных проходов по выборке, ограничиваясь всего тремя просмотрами. Причем, если заранее известны граничные значения признаков, то число просмотров выборки можно сократить до двух. При этом метод хранит в оперативной памяти только один текущий экземпляр и набор сформированных одномерных эталонов, который минимизирован по объему. Это позволяет существенно снизить затраты как вычислительных ресурсов, так и ресурсов памяти.

Таблица 1 – Результаты экспериментов

Метод формирования выборки	S'/S	t, сек.	M, Мегабайт	$E_{об.}$, %	$t_{об.}$, сек.	$E_{расп.}$, %
Случайный отбор (без загрузки всей исходной выборки в память)	0,5	1,35	1,62	0	1768,6	2,01
	0,25	0,68	0,81	0	936,6	4,14
	0,1	0,27	0,38	0	373,2	7,42
	0,05	0,14	0,16	0	191,3	12,97
Кластер-анализ	0,26	690,37	101,34	0	920,4	0,41
Предложенный метод	0,21	93,62	0,79	0	912,5	0,49

6 ОБСУЖДЕНИЕ

Предложенный метод формирования выборок по сравнению с методом на основе случайного отбора [6] обеспечивает существенно большую точность прогнозирования, сокращая при этом время обучения. Однако предложенный метод требует больших затрат времени на формирование выборки, чем метод случайного отбора [6].

По сравнению с методом формирования выборок на основе кластер-анализа [9] предложенный метод обеспечивает сопоставимую точность прогнозирования, сокращая при этом время формирования выборки. Однако предложенный метод требует существенно меньших затрат времени и ресурсов памяти при формировании выборки, чем метод на основе кластер-анализа [9].

Также, в отличие от методов на основе случайного отбора и кластер-анализа, предложенный метод автоматически определяет размер формируемых обучающей и тестовой выборок, не требуя участия человека.

Недостатком предложенного метода является то, что он требует задания преобразования на обобщенную ось и зависит от его требований к вычислительным ресурсам.

Эффективность применения разработанного метода будет тем выше, чем больше признаков будет характеризовать исходный набор данных и чем больше будет экземпляров в исходной выборке данных. При небольшом объеме исходной выборки эффект от применения разработанного метода будет незначительным.

ВЫВОДЫ

В работе решена задача формирования обучающих выборок для автоматизации построения нейросетевых моделей по прецедентам.

Научная новизна результатов, полученных в статье, состоит в том, что впервые предложен метод формирования выборок, который, автоматически выделяет из исходной выборки обучающую и тестовую выборки, не требуюя загрузки всей исходной выборки в память ЭВМ, осуществляя поэкземплярную обработку исходной выборки с хэширующим преобразованием на одномерную ось, формирует эталоны кластеров на обобщенной оси, минимизируя их число, что позволяет повысить скорость формирования выборок, снизить требования к вычислительным ресурсам и памяти ЭВМ и обеспечить приемлемый уровень точности синтезируемых моделей.

Практическая значимость полученных результатов заключается в том, что разработано программное обеспечение, реализующее предложенный метод, на основе которого решена практическая задача построения модели принятия решенй для индивидуального прогнозирования состояния пациента, больного гипертонией.

Перспективы дальнейших исследований состоят в том, чтобы определить эффективные по времени и затратам памяти хэширующие преобразования экземпляров выборки, сохраняющие топологию классов в пространстве признаков, исследовать предложенный метод на более широком классе задач количественного и качественного прогнозирования и распознавания образов.

БЛАГОДАРНОСТИ

Работа выполнена в рамках госбюджетной научно-исследовательской темы Запорожского национального технического университета «Интеллектуальные информационные технологии автоматизации проектирования, моделирования, управления и диагностирования производственных процессов и систем» (номер гос. регистрации 0112U005350) при частичной поддержке международного проекта «Центры передового опыта для молодых ученых» программы Tempus Европейской Комиссии (№ 544137-TEMPUS-1-2013-1-SK-TEMPUS-JPHES).

СПИСОК ЛИТЕРАТУРЫ

1. Субботін С. О. Нейронні мережі : навчальний посібник / С. О. Субботін, А. О. Олійник ; під заг. ред. проф. С. О. Субботіна. – Запоріжжя : ЗНТУ, 2014. – 132 с.
2. Computational intelligence: a methodological introduction / [R. Kruse, C. Borgelt, F.Klawonn et. al.]. – London : Springer-Verlag, 2013. – 488 p. DOI: 10.1007/978-1-4471-5013-8_1
3. Олешко Д. Н. Построение качественной обучающей выборки для прогнозирующих нейросетевых моделей / Д. Н. Олешко, В. А. Крисилов, А. А. Блажко // Штучний інтелект. – 2004. – № 3. – С. 567–573.
4. 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. DOI: 10.3103/s1060992x10020037
5. 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.
6. Encyclopedia of survey research methods / ed. P. J. Lavrakas. – Thousand Oaks: Sage Publications, 2008. – Vol. 1–2. – 968 p. DOI: 10.1108/09504121011011879
7. Hansen M.H. Sample survey methods and theory / M. H. Hansen, W. N. Hurtz, W. G. Madow. – Vol. 1 : Methods and applications. – New York: John Wiley & Sons, 1953. – 638 p.
8. 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.
9. Multivariate analysis, design of experiments, and survey sampling / ed. S. Ghosh. – New York: Marcel Dekker Inc., 1999. – 698 p.
10. Smith G. A deterministic approach to partitioning neural network training data for the classification problem : dissertation ... doctor of philosophy in business / Smith Gregory. – Blacksburg: Virginia Polytechnic Institute & State University, 2006. – 110 p.
11. Bernard H. R. Social research methods: qualitative and quantative approaches / H. R. Bernard. – Thousand Oaks: Sage Publications, 2006. – 784 p.
12. Chaudhuri A. Survey sampling theory and methods / A. Chaudhuri, H. Stenger. – New York : Chapman & Hall, 2005. – 416 p.

13. Subbotin S. A. Methods and characteristics of locality-preserving transformations in the problems of computational intelligence / S. A. Subbotin // Радіоелектроніка, інформатика, управління. – 2014. – № 1. – С. 120–128.
14. Дневник погоды [Электронный ресурс]. – Москва : ООО «НПЦ «Мэп Мейкер», 2014. – Режим доступа: http://www.gismeteo.ru/diary/5093
15. Weber M. Observations of Solar Activity (Mg II Index) by GOME, SCIAMACHY, and GOME-2 [Electronic resource]. – Bremen: University of Bremen, 2014. – Access mode: http://www.iup.uni-bremen.de/gome/gomemgii.html

Статья поступила в редакцию 15.12.2014.
После доработки: 20.01.2015.

Субботін С. О.
Д-р техн. наук, професор, професор кафедри програмних засобів Запорізького національного технічного університету, Запоріжжя, Україна

ШВИДКИЙ МЕТОД ВИДІЛЕННЯ НАВЧАЛЬНИХ ВИБІРОК ДЛЯ ПОБУДОВИ НЕЙРОМЕРЕЖЕВИХ МОДЕЛЕЙ ПРИЙНЯТТЯ РІШЕНЬ ЗА ПРЕЦЕДЕНТАМИ

Вирішено завдання формування навчальних вибірок для автоматизації побудови нейромережевих моделей за прецедентами. Запропоновано метод формування вибірок, який автоматично виділяє з вихідної вибірки навчальну та тестову вибірки, не вимагаючи завантаження всієї вихідної вибірки у пам'ять ЕОМ, здійснюючи поекземплярну обробку вихідної вибірки з гешуючим перетворенням на одновимірну вісь, формує еталони кластерів на узагальненій осі, мінімізуючи їх число, що дозволяє підвищити швидкість формування вибірок, знизити вимоги до обчислювальних ресурсів і пам'яті ЕОМ і забезпечити прийнятний рівень точності синтезованих моделей. Розроблений метод не вимагає багаторазових проходів по вибірці, обмежуючись лише трьома переглядами. При цьому метод зберігає в оперативній пам'яті тільки один поточний екземпляр і набір сформованих одновимірних еталонів, який мінімізовано за обсягом. На відміну від методів на основі випадкового відбору та кластер-аналізу запропонований метод автоматично визначає розмір формованих навчальної та тестової вибірок, не вимагаючи участі людини. Розроблено програмне забезпечення, що реалізує запропонований метод, на основі якого вирішена практична задача побудови моделі прийняття рішень для індивідуального прогнозування стану пацієнта, хворого на гіпертонію.

Ключові слова: вибірка, формування вибірки, екземпляр, нейронна мережа, індивідуальне прогнозування, навчання за прецедентами.

Subbotin S. A.
Dr.Sc., Professor, Professor of Department of Software Tools, Zaporizhzhya National Technical University, Zaporizhzhya, Ukraine

THE QUICK METHOD OF TRAINING SAMPLE SELECTION FOR NEURAL NETWORK DECISION MAKING MODEL BUILDING ON PRECEDENTS

The problem of training sample forming is solved to automate the construction of neural network models on precedents. The sampling method is proposed. It automatically selects the training and test samples from the original sample without the need for downloading the entire original sample to the computer memory. It processes an initial sample for each one instance with hashing transformation to a one-dimensional axis, forming cluster templates on the generalized axis, minimizing their number. This allows to increase the speed of sampling, to reduce the requirements to computing resources and to computer memory and to provide an acceptable level of accuracy of the synthesized models. The developed method does not require multiple passes through the sample, being limited by only three viewing. At the same time the method keeps in a random access memory only the current instance and the generated set of one-dimensional templates, which is minimized by volume. Unlike the methods based on random sampling and cluster analysis the proposed method automatically determines the size of the formed training and test samples without the need for human intervention. Software realizing proposed method is developed. On its basis the practical task of decision-making model building to predict the individual state of the patient with hypertension is resolved.

Keywords: sample, sampling, instance, neural network, individual prediction, training on precedents.

REFERENCES

1. Subbotin S. O., Olijnik A. O. Nejronni merezhi : navchal'nyj posibnik ; pid zag. red. prof. S. O. Subbotina. Zaporizhzhya, ZNTU, 2014, 132 p.
2. Kruse R., Borgelt C., Klawonn F. et. al. Computational intelligence: a methodological introduction. London, Springer-Verlag, 2013, 488 p. DOI: 10.1007/978-1-4471-5013-8_1
3. Oleshko D. N., Krisilov V. A., Blazhko A. A. Postroenie kachestvennoj obuchayushhej vyborki dlya prognoziruyushhix nejrosetevyx modelej. Shtuchnyj intelekt, 2004, No. 3, pp. 567–573.
4. 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. DOI: 10.3103/s1060992x10020037
5. Subbotin S. A. Kriterii individual'noj informativnosti i metody otbora e'kzemplyarov dlya postroeniya diagnosticheskix i raspoznayushhix modelej, *Bionika intelektu,* 2010, No. 1, pp. 38–42.
6. Encyclopedia of survey research methods. Ed. P. J. Lavrakas. Thousand Oaks, Sage Publications, 2008, Vol. 1–2, 968 p. DOI: 10.1108/09504121011011879
7. Hansen M. H., Hurtz W. N., Madow W. G. Sample survey methods and theory, Vol. 1, Methods and applications. New York, John Wiley & Sons, 1953, 638 p.
8. Kokren U. Metody vyborochnogo issledovaniya; per. s angl. I. M. Sonina ; pod red. A. G. Volkova, N. K. Druzhinina. Moscow, Statistika, 1976, 440 p.
9. Multivariate analysis, design of experiments, and survey sampling. Ed. S. Ghosh. New York, Marcel Dekker Inc., 1999, 698 p.
10. Smith G. A deterministic approach to partitioning neural network training data for the classification problem : dissertation ... doctor of philosophy in business. Blacksburg, Virginia Polytechnic Institute & State University, 2006, 110 p.
11. Bernard H. R. Social research methods: qualitative and quantative approaches. Thousand Oaks, Sage Publications, 2006, 784 p.
12. Chaudhuri A., Stenger H. Survey sampling theory and methods. New York, Chapman & Hall, 2005, 416 p.
13. Subbotin S. A. Methods and characteristics of locality-preserving transformations in the problems of computational intelligence, *Radioelektronika, informatika, upravlinnya,* 2014, No. 1, pp. 120–128.
14. Dnevnik pogody [E'lektronnyj resurs]. Moscow, OOO «NPC «Me'p Mejker», 2014, Rezhim dostupa: http://www.gismeteo.ru/diary/5093
15. Weber M. Observations of Solar Activity (Mg II Index) by GOME, SCIAMACHY, and GOME-2 [Electronic resource]. Bremen, University of Bremen, 2014, Access mode: http://www.iup.uni-bremen.de/gome/gomemgii.html