УДК 681.3: 519.68

Д.Н. Олешко, В.А. Крисилов, А.А. Блажко
Одесский национальный политехнический университет, Украина, d_vader@farlep.net

Построение качественной обучающей выборки для прогнозирующих нейросетевых моделей

В данной работе рассмотрены основные требования, которым должна отвечать обучающая выборка, а также предложены параметры для оценки качества построенной выборки для решения задачи прогнозирования. На базе требований разработан подход к решению задачи построения качественной обучающей выборки.

Введение

Эффективность функционирования любой обучающейся системы существенно зависит от качества тех данных, на которых проходит обучение. Некачественность обучающего материала не только отразится на функционировании системы в будущем, но может привести к невозможности обучиться в принципе.

Широкое распространение в различных предметных областях получили обучающиеся модели, построенные с использованием аппарата искусственных нейронных сетей (НС). Одним из частых применений НС является их использование для получения прогнозов поведения различных процессов в АСУ, задачах, связанных с электронным бизнесом и экономическим мониторингом. Как и любая другая обучающаяся система, НС должна быть обеспечена исходными данными для обучения, формирующими ее обучающую выборку (ОВ). Однако зачастую процессу формирования ОВ не уделяют должного внимания, кроме того, отсутствует единая методика оценки качества построенной ОВ.

При построении ОВ для нейросетевых моделей, прогнозирующих поведение временного ряда, используется так называемый «Метод скользящих окон», основанный на теореме Такенса [1]. Если результаты обучения не удовлетворяют разработчика, то ответные действия обычно ограничиваются поиском такого размера окна, при котором ошибка обучения нейронной сети (НС) достигнет приемлемого значения. Либо проектировщик НС дополнительно использует различные предварительные преобразования данных (вейвлеты, промежутки стабильности и т.п.) для достижения необходимого результата [2]. Но в распоряжении разработчика нет критериев, по которым он мог бы оценить качество построенной ОВ до обучения. Единственным косвенным способом оценки качества построенной ОВ является обучение на ней нейронной сети. И только после многократных попыток уменьшить ошибку обучения за счет увеличения числа эпох обучения и изменения параметров обучающего алгоритма разработчик пытается решить проблему переформированием ОВ.

Отсутствие методики оценки ОВ приводит к увеличению затрат на синтез прогнозирующей модели в целом. Причинами этого являются и увеличение времени сходимости процесса обучения на некачественной ОВ, и затраты на поиск альтернативных методов построения ОВ в случае неуспешного обучения.

Увеличение же времени, затрачиваемого на построение или на актуализацию прогнозирующей модели, делает данный метод прогнозирования неэффективным, поскольку большинство экономических и технических процессов имеют достаточно динамичный характер и требуют частого обновления НС-моделей.

Данная статья посвящена проблеме построения *качественной* ОВ для прогнозирующих НС. В ней предложены характеристики, позволяющие оценить качество построенной ОВ, а также алгоритм построения ОВ в соответствии с этими характеристиками.

Требования к качеству ОВ

Наиболее распространенной моделью НС, используемой для прогнозирования, является многослойный персептрон с обучающим алгоритмом обратного распространения ошибки [3]. Такая НС предполагает возможность решать задачу прогнозирования временного ряда двумя способами:

1) решение задачи прогнозирования как задачи регрессии – и входы и выход сети представлены непрерывными величинами;
2) решение задачи прогнозирования как задачи классификации – входы являются непрерывными, а на выходе – дискретное множество распознаваемых классов изменения прогнозируемой величины, которое задается множеством эталонных значений ОВ.

В любом случае НС фактически исполняет роль памяти и распознавателя, и если в процессе обучения удается выявить некоторые закономерности изменения временного ряда, то решение задачи прогнозирования становится возможным.

Известно, что для того, чтобы задача распознавания образов на некотором множестве объектов была решена достоверно, необходимо, чтобы распознаваемые классы были обеспечены достаточным количеством примеров, описывающих все многообразие объектов данного класса. То есть для каждого класса должно быть соблюдено требование его представительности.

То же справедливо и для ОВ прогнозирующей НС. Однако в данном случае под *представительностью* ОВ понимается то, что данные об изменении прогнозируемой величины должны быть взяты за период времени, достаточный для построения достоверного прогноза.

Пусть ОВ сформирована, а исходные данные содержат достаточно информации о характере прогнозируемого процесса. Рассмотрим ситуацию, когда количество обучающих наборов (ОН) в одном классе (*A*) существенно превышает количество ОН в другом классе (*B*). Очевидно, что поскольку НС будет чаще учиться на ОН класса *A*, то и в процессе функционирования она будет лучше и увереннее распознавать объекты этого класса. Но в такой ситуации получить достоверный прогноз достаточно проблематично, поскольку НС в большинстве случаев будет «прогнозировать» ситуацию, описанную классом *A*. Этот пример демонстрирует тот факт, что требование *представительности* ОВ является необходимым, но не достаточным для получения от прогнозирующей модели достоверных результатов.

Таким образом, для того чтобы модель смогла строить достоверный прогноз, необходимо стремиться к тому, чтобы количество ОН в классах было соизмеримо, т.е. вторым требованием к ОВ можно обозначить *равномерность*. Однако следует понимать, что злоупотребление этим требованием может привести к недопустимому искажению картины естественного распределения объектов и значительному огрублению точности решения поставленной задачи.

Для оценки неравномерности была рассмотрена случайная величина N_{OH} – количество ОН в классе. В качестве величины, характеризующей неравномерность построенной ОВ, предлагается среднеквадратичное отклонение N_{OH}:

$$R_{OB} = \sigma(N_{OH}).$$

Однако представительная и равномерная ОВ не всегда служит залогом качественного обучения НС. Такая выборка может в себе содержать одинаковые объекты, но принадлежащие разным классам и делающие ОВ *противоречивой*. Противоречивость обучающих данных является серьезным недостатком для ОВ. Для любой обучающейся системы, в том числе и человека, обучение на таком материале приводит к тупиковым ситуациям и не является успешным. НС в силу своей адаптивности способны к обучению на ОВ с противоречивыми объектами, однако качество обучения от этого существенно снижается. И чем больше таких объектов в ОВ, тем меньше вероятность того, что процесс обучения будет успешным.

Основываясь на этом, можно выдвинуть третье требование к ОВ – *непротиворечивость*.

На сегодня существует методика оценки противоречивости ОВ [4]. Однако в ней понятие противоречивости рассматривается в дискретном пространстве, и непрерывные величины, являющиеся характеристиками объектов, предварительно дискретизируются. Такие действия вносят определенную погрешность в вычисления.

В связи с этим предлагается рассмотреть противоречивость как непрерывную величину.

Пусть два обучающих набора заданы парами вида

$$\left(\left\{a_1^i, \text{K}, a_k^i\right\}, A_m\right) \text{ и } \left(\left\{a_1^j, \text{K}, a_k^j\right\}, A_n\right),$$

где a_k^i – свойства i-го объекта в k-мерном пространстве, образующие вектор входных значений для НС, а A_m – соответственно центроид m-го класса – эталонное значение в обучающем наборе.

Тогда $\Delta A_{mn} = \left|A_m - A_n\right|$ – расстояние между центроидами соответственно m-го и n-го классов. А расстояние между объектами этих классов будет вычисляться по следующей формуле:

$$\Delta a_{ij} = \sqrt{(1/Da_1)\left(\Delta a_1^{ij}\right)^2 + \text{K} + (1/Da_k)\left(\Delta a_k^{ij}\right)^2}, \qquad (1)$$

где $\Delta a_k^{ij} = a_k^i - a_k^j$, Da_k – дисперсия свойства k-го измерения по всей ОВ.

Теперь, введя два расстояния – расстояние между объектами и между центроидами классов, к которым они принадлежат, можно определить понятие противоречивости.

Пусть C_{ij} – парная противоречивость – противоречивость двух обучающих наборов i-го и j-го, принадлежащих соответственно классам A_m и A_n. Тогда очевидно, что C_{ij} возрастает, если возрастает ΔA_{mn} или убывает Δa_{ij}.

На основании данных рассуждений предложена следующая формула для вычисления C_{ij}:

$$C_{ij} = \frac{\Delta A_{mn}}{\Delta a_{ij} + \Delta A_{mn}}.$$

Согласно этой формуле противоречивость двух объектов лежит в диапазоне [0; 1], достигает максимума при совпадении характеристик объектов, принадлежа-

щих разным классам, и становится равной 0 в случае, если рассматриваются объекты одного класса. Противоречивостью всей ОВ будет среднее всех C_{ij}:

$$C_{OB} = \left(\sum_{1}^{n} C_{ij} \right) / n ,$$

где n – количество всех парных противоречивостей в ОВ.

Таким образом, для обеспечения качественного обучения, а также достоверного прогноза ОВ должна отвечать следующим требованиям.

1. ОВ должна быть представительной.
2. ОВ должна быть равномерной.
3. ОВ должна быть непротиворечивой.

Как показывает практика, наиболее трудным для соблюдения является второе требование, поскольку чаще всего встречается нормальное распределение объектов по классам. Некоторые подходы по разрешению проблем неравномерности и противоречивости изложены в [5].

Предложенные подходы, связанные с реорганизацией классов, в большинстве случаев дают хорошие результаты. Однако в реальных задачах часто, кроме перечисленных трех требований, на ОВ могут накладываться дополнительные требования, связанные с постановкой задачи. Одно из них – требование к *точности представления данных*. Это требование может существенно ограничить свободу в переформировании множества распознаваемых классов. Здесь следует отметить, что злоупотребление *точностью* приводит к следующим последствиям:

– во-первых, если стремиться максимально сохранить картину естественного распределения объектов, которое чаще всего является нормальным, то при этом существенно *повышается неравномерность* ОВ;
– во-вторых, если под точностью понимать величину ошибки преобразования ОВ в процессе построения множества классов, то тогда неоправданно возрастает количество распознаваемых классов (N_{CLASS}) и как результат *увеличивается размерность* НС.

В обоих случаях возрастают затраты времени на обучение НС. Поэтому разработчику всегда необходимо определять оптимальное соотношение для данной задачи, руководствуясь принципом достаточности [5], [6].

Критерий оценки качества ОВ

Рассмотренные в предыдущем разделе характеристики позволяют получить информацию о некоторых свойствах ОВ, но не позволяют сделать обобщенный вывод о ее качественности. Для этого предлагается ввести следующий критерий.

Разработанный критерий оптимальности базируется на рассмотренных ранее характеристиках ОВ и включает в себя несколько дополнительных. Рассмотрим их подробнее.

Как и для любого преобразования данных, в данном случае важно уметь оценить его погрешность и стремиться выбирать решение с минимальной величиной этой погрешности. Пусть множество пар вида $\left(\left\{ a_1^i, \text{K} , a_k^i \right\}, Y^i \right)$ представляют собой множество ОН исходной ОВ, а множество пар вида $\left(\left\{ a_1^i, \text{K} , a_k^i \right\}, A_m^i \right)$ – множество ОН преобразованной ОВ после разбиения множества значений Y^i на классы. Тогда ошибкой преобразования на одном наборе будет следующая величина:

$$D_i = A_m^i - Y^i ,$$

а среднеквадратичная ошибка преобразования будет вычислена следующим образом:

$$D_{OB} = \sqrt{\sum_{1}^{n} D_i^2 \Big/ n},$$

где n – количество ОН.

Немаловажным фактом для определения оптимальности полученного разбиения является то, что величина D_{OB} позволяет оценить компактность разбиения объектов по классам.

Кроме того, как было сказано ранее, ограничивающим фактором для понижения R_{OB} служит величина N_{CLASS}. Таким образом, она также должна быть включена в искомый критерий.

На основе данных рассуждений предложен следующий критерий оценки качества ОВ. Пусть Q_{OB} – некоторый функционал, позволяющий оценить качество ОВ на каждой отдельной итерации, вычисляется по формуле

$$Q_{OB} = C_{OB}^{w_C} * R_{OB}^{w_R} * D_{OB}^{w_D} * N_{CLASS}^{w_N}, \quad (2)$$

где w_C, w_R, w_D, w_N – показатели, характеризующие важность той или оной характеристики для разработчика в рамках текущей задачи.

Тогда оптимальная ОВ будет выбираться согласно следующему критерию:

$$Q_{OB}^{*} = \min\left(Q_{OB}\right). \quad (3)$$

Данный критерий позволяет формализовать процедуру построения качественной ОВ. В формуле (2) эвристичность процесса построения ОВ становится проявленной и выражена показателями степеней каждой из характеристик. Выбор же конкретных значений данных показателей должен осуществляться исходя из требований и условий поставленной задачи и опираться на понятие достаточности.

Алгоритм построения качественной ОВ

Разработанный алгоритм построения качественной ОВ стал результатом объединения подходов [5], подчиненных ограничениям, накладываемым требованиями к точности представления данных и качеству построенной ОВ.

Входными данными для рассматриваемого алгоритма служит ОВ, построенная «Методом скользящих окон». Задача алгоритма – построить оптимальное разбиение на классы множества эталонных значений ОВ.

Согласно алгоритму, эталонные значения A_i предварительно нормируются в диапазон от – 100 % до 100 % и упорядочиваются по возрастанию для удобства проведения таксономии. Максимальное количество объектов в классах ограничивается некоторым значением $N_{\max}$, которое достаточно мало на начальной итерации и увеличивается до некоторого предела на последующих итерациях алгоритма. Кроме того, пользователем задается ограничение на точность представления данных. Оно представляет собой допустимое максимальное расстояние между объектами класса для заданной шкалы.

Для сглаживания неравномерностей алгоритм должен менее жестко следовать ограничению по точности в областях с низкой плотностью распределения объектов

(ПРО), а в областях с высокой плотностью ограничивать мощность формируемых классов. Поэтому решение о том, добавлять текущий объект в уже созданный класс или формировать новый, принимается с учетом следующих условий:

а) заданного ограничения по точности;
б) ПРО вокруг анализируемого объекта;
в) заполненности уже сформированного класса.

Окрестность, на которой вокруг анализируемого объекта оценивается ПРО, является параметром алгоритма и называется *горизонтом разведки*.

Сам алгоритм можно описать следующими шагами.

1) Множество исходных эталонных значений ОВ нормируется и упорядочивается по возрастанию.
2) Задаются значения горизонта разведки, допустимого разброса объектов внутри класса, $N^0_{\max}$ и предел его изменения N_{TOP}.
3) Осуществляется перебор множества объектов от минимального к максимальному. Первый объект образует первый класс с центроидом, равным значению объекта.
4) Далее для каждого нового объекта анализируются условия а – в и в соответствии с ними принимается решение об отнесении текущего объекта к уже имеющемуся классу или об образовании нового класса. При добавлении объекта в класс центроид такого класса принимает значение, равное среднему значению величин объектов, его составляющих.
5) По окончании разбиения объектов на классы для преобразованной ОВ вычисляется значение Q_{OB}.
6) Увеличиваем на единицу значение $N_{\max}$.
7) Повторяем шаги 1 – 6, пока $N_{\max}$ не достигнет N_{TOP}.
8) Из полученного множества решений выбирается то, которое соответствует критерию (3).

Заключение

Итак, в данной работе были предложены основные требования для оценки качества ОВ, строящихся для прогнозирующих НС.

1. ОВ должна быть представительной, т.е период, который охватывает ОВ, должен содержать данные, которые позволили бы выявить закономерность изменения прогнозируемой величины и получить достоверный прогноз.
2. Распределение ОН между классами ОВ должно быть близким к равномерному, но не в ущерб заданным ограничениям точности представления данных.
3. ОВ должна быть максимально непротиворечивой.

Для вычисления противоречивости выборки предложены формулы, оперирующие с непрерывными величинами и не требующие их предварительной дискретизации, как это делалось ранее.

Предложенный критерий оценки качества построенной ОВ позволяет осуществить объективную оценку и дает разработчику наглядную возможность оперировать степенями важности характеристик ОВ в процессе ее построения.

Разработанный алгоритм автоматического построения качественной ОВ в значительной степени облегчает труд проектировщика прогнозирующей нейросетевой модели. К его достоинствам можно отнести следующее:

1. На порядок сокращаются затраты времени на построение обучающей выборки, отвечающей всем заданным требованиям.
2. Построенная в результате выборка максимально отвечает установленным требованиям качества, что обуславливает снижение в разы затрат времени на обучение НС и повышение достоверности прогноза на 10 – 20 %.
3. Компактное разбиение множества эталонных значений ОВ на классы позволяет выиграть в размерности сети, что также приводит к снижению затрат времени на синтез прогнозирующей НС.

Литература

1. Научная сессия МИФИ-2003. V Всероссийская науч.-техн. конф. «Нейроинформатика – 2003»: Лекции по нейроинформатике. Ч. 1. – М.: МИФИ, 2003. – 188 с.
2. Крисилов В.А., Чумичкин К.В., Кондратюк А.В. Представление исходных данных в задачах нейросетевого прогнозирования // V Всероссийская науч.-техн. конф. «Нейроинформатика-2003»: Сб. науч. трудов. – Том 1. – Москва: МИФИ.– 2003. – С. 184-191.
3. Нейронные сети. Statistica Neural Networks: Пер. с англ. – М.: Горячая линия – Телеком, 2000. – 182 с.
4. Тарасенко Р.А. Метод анализа и повышения качества обучающих выборок нейронных сетей для прогнозирования временных рядов: Дис… канд. техн. наук. – ОНПУ, 2002.
5. Олешко Д.Н., Крисилов В.А. Повышение качества и скорости обучения нейронных сетей в задаче прогнозирования поведения временных рядов // Праці «МКІМ – 2002». – Львів. – 2002. – Секції 4, 5. – С. 76-81.
6. Krissilov V.A., Krissilov A.D., Oleshko D.N. Application of the sufficiency principle in acceleration of neuron networks training // X-th Internation Conf. “Knowledge-Dialogue-Solution” (KDS-2003), June 16-26. – Varna (Bulgaria). – P. 164-168.

Д.М.Олешко, В.А. Крісілов, О.О. Блажко

Побудова якісної навчальної вибірки для прогнозуючих нейромережних моделей.
У даній роботі розглянуті основні вимоги, яким повинна відповідати навчальна вибірка, а також запропоновані параметри для оцінки якості побудованої вибірки для розв‘язання задачі прогнозування. На базі вимог розроблений підхід до рішення задачі побудови якісної навчальної вибірки.

This paper is devoted to consideration of requirements to which training sample for neural networks in forecast tasks should meet. Also there are suggested some parameters for estimation of quality of training sample and a criterion based on these parameters. The result of this work consists in developed automated algorithm based on this criterion for creation of qualitative training sample.

Статья поступила в редакцию 19.07.2004.