УДК 004.032.26

МЕТОДЫ АНАЛИЗА И СЕМАНТИЧЕСКОЙ ИНТЕРПРЕТАЦИИ ПРОЦЕССОВ ПРИНЯТИЯ РЕШЕНИЯ В КЛАССИФИКАЦИОННЫХ НЕЙРОСЕТЕВЫХ МОДЕЛЯХ МАШИННОГО ОБУЧЕНИЯ С УЧИТЕЛЕМ

А.В. КУРОЧКИН, канд. техн. наук, доц. В.С. САДОВ
(Белорусский государственный университет, Минск)

В работе рассматривается проблема интерпретации характера поведения обученной классификационной нейросетевой модели прямого распространения, а также предлагаются решения по анализу и объяснению получаемых результатов на основе локальных линейных аппроксимаций.

Ключевые слова: *нейронные сети, системы поддержки принятия решений, анализ работы нейросетевых моделей.*

Искусственные нейронные сети являются на сегодняшний день одним из самых распространенных инструментов для решения задач машинного обучения с учителем. Нейронные сети прямого распространения с нелинейной функцией активации являются универсальным параметрическим аппроксиматором, т.е. на основе достаточно сложной конфигурации слоев и нейронов при помощи регулирования параметров сети – весов связей между нейронами соседнего слоя – нейронные сети могут смоделировать сколь угодно сложную функцию с заданной размерностью вектора входных и выходных данных; при этом выходной результат может быть дополнительно ограничен, например, при помощи сигмоидальной функции [1].

Ключевой этап построения нейросетевой модели – обучение. При наличии достаточно репрезентативной обучающей выборки – набора из достоверных входных и ожидаемых выходных значений модели – целью алгоритма обучения нейронной сети является подбор параметров сети (весов) таким образом, чтобы сформировать разнообразные совокупности линейных и нелинейных комбинаций входных признаков, которые в последующем используются для получения итогового результата, причем подбор параметров должен осуществляться таким образом, чтобы минимизировать ошибку выходного значения, формируемого моделью, на предоставленной обучающей выборке. Другими словами, внешний вид универсальной параметрической аппроксимирующей функции, определяемой нейросетевой моделью, из-за подстройки весов в процессе обучения с каждым шагом будет все больше приближаться к такому виду, который максимально корректно (в смысле ошибки) описывает предоставленную обучающую выборку. При этом сами алгоритмы обучения являются агностическими по отношению как к характеру решаемой задачи, так и к виду предоставленной обучающей выборки. Фактически, одну и ту же нейросетевую архитектуру прямого распространения (конфигурацию слоев) можно обучить решению задачи из принципиально другой предметной области, если она описывается таким же количеством входных и выходных признаков и для нее существует обучающая выборка.

Таким образом, на основании этих факторов можно сказать, что, хотя нейросетевые модели и являются универсальными аппроксиматорами, сам процесс принятия решения в них опирается исключительно на статистические распределения, которые смогли или не смогли быть установлены в процессе обучения сети. В этой связи в процессе анализа обученных нейросетевых моделей для их последующего практического применения возникает две ключевые задачи:

1) проверка статистической корректности обученной нейросетевой модели для решаемой задачи;
2) проверка семантической корректности обученной нейросетевой модели для решаемой задачи.

Под проверкой статистической корректности понимается непосредственно анализ генерируемого обученной моделью выходного значения в контексте известных результатов для решаемой задачи. Статистическая корректность должна отвечать на вопрос: «Насколько корректно полученная модель ведет себя при решении известных задач предметной области?». Под семантической корректностью понимается анализ генерируемого обученной моделью выходного значения в контексте семантики решаемой задачи. Семантическая корректность должна отвечать на вопросы: «Почему полученная модель ведет себя таким образом?» и «Насколько корректным является поведение полученной модели?» [2, 3].

Задача проверки статистической корректности является хорошо изученной, поскольку на основании статистической корректности, помимо всего прочего, происходит непосредственно обучение сети – ошибку на обучающей выборке можно также считать одной из метрик статистической корректности.

Одной из наиболее распространенных проблем статистической корректности является проблема переобучения. Поскольку основная задача нейросетевой модели состоит в определении выходных значений, лежащих вне точек, которые использовались в обучающей выборке, в целом ожидается, что полученная модель может предоставлять достаточно качественные обобщения. С другой стороны, поскольку моделируемая функция является универсальной относительно изменения весов нейронной сети, если единственной целью

обучения поставить необходимость минимизации ошибки на известных данных, на выходе может быть получена переобученная модель – излишне сложная функция, которая ведет себя предсказуемо только в очень близких окрестностях тех точек, которые присутствовали в обучающей выборке, и при этом демонстрирует сложное и непредсказуемое поведение вне этих точек. Типовым приемом для выявления этой проблемы является перекрестная валидация – использование некоторого случайного подмножества (порядка 70%) обучающей выборки непосредственно для обучения, а затем оставшихся элементов – для оценки статистической корректности в тех точках, которые не присутствовали в обучающей выборке. Если модель показывает низкую ошибку на обучающей выборке, но высокую ошибку перекрестной валидации, то полученная модель является переобученной. После установления факта переобучения для его предотвращения на этапе обучения могут использоваться различные методы регуляризации параметров и ансамблирования моделей.

Другой важной проблемой статистической корректности является проблема скошенных классов при использовании бинарной классификации, т.е. неравномерная репрезентация экземпляров двух классов в обучающей выборке. Из-за этого ошибка на обучающей выборке может являться ненадежным критерием статистической корректности модели, и вместо этого для анализа полученной модели может применяться ROC-анализ, а также использоваться такие метрики, как F1-мера, точность, полнота и т.д. [1]

Статистическая корректность является агностической относительно предоставляемых данных, используемой модели и решаемой задачи. В этой связи оценка статистической корректности полученной модели во многих случаях может осуществляться автоматизировано, на основании некоторых критериев. При выявлении недостаточной статистической корректности повторное обучение может быть подстроено соответствующим образом, и для этого, как правило, также не требуется знание семантического наполнения решаемой задачи.

Проверка семантической корректности нейросетевых моделей подразумевает попытку понять, на основании чего исследуемая модель принимает решения и является ли ее процесс принятия решения корректным с точки зрения рассматриваемой задачи и известных экспертных знаний в предметной области. В отличие от проверки статистической корректности, семантическая корректность не может трактовать полученную модель как «черный ящик». Кроме того, сам процесс обучения нейросетевой модели никак не принимает во внимание семантическую корректность в контексте решаемой задачи. В частности, при обучении традиционные нейросетевой архитектуры никак не позволяют учитывать экспертные знания в формализованном виде: поиск итоговой аппроксимирующей функции производится только исходя из статистических характеристик обучающей выборки. В этой связи актуальной является, с одной стороны, проблема поддержки формализованных экспертных знаний в процессе принятия решения нейросетевой моделью, а с другой стороны – проблема формального описания процесса принятия решения в полученной модели.

Для оценки семантической корректности требуется, чтобы процесс принятия решения в полученной модели мог быть описан в понятном человеку виде. Во многих случаях такое описание построить невозможно. Интуитивно ожидается, что в процессе обучения нейронная сеть сможет выделить такие совокупности значений признаков, которые совместно предоставляют более сложное абстрактное описание входных данных. Например, в задаче распознавания символов печатного шрифта латинского алфавита, при использовании бинаризированных яркостей пикселей как входных параметров, один из нейронов второго слоя может активизироваться по совокупности значений пикселей в центральном ряду по горизонтали, что соответствует абстрактному описанию «горизонтальная черта в центре в начертании символа»; в дальнейшем, такой нейрон может иметь более значительный вес при связи с выходными нейронами, которые соответствуют буквам «A», «B», «E», «F», «H», «P», «R». В то же время, для принятия решения могут использоваться более сложные нелинейные совокупности признаков, которые не будут иметь такой простой интерпретации [3].

Еще одной проблемой является тот факт, что, даже при наличии интерпретации процесса принятия решения в понятном человеку виде, для оценки семантической корректности модели по определению не существует объективных критериев. Любой анализ корректности может производиться только субъективно, на основании экспертных знаний. С другой стороны, если статистической корректности достаточно для решения задачи в описываемой предметной области, анализ семантической корректности предоставляет широкие возможности по решению обратной задачи – описанию и формализации новых экспертных данных, которые выявлены моделью в процессе обучения. Иначе говоря, анализ семантической корректности может рассматриваться не с точки зрения применимости обученной модели для решения поставленной задачи, а с точки зрения интерпретации полученной модели с целью улучшить понимание зависимостей, присутствующих в данных, и сформировать принципиально новые знания в доменной области, что может быть полезно экспертам в этой сфере [4, 5].

Для формирования интерпретаций по обученной модели в первую очередь требуется определить, какой вид описания обученной модели может легко восприниматься человеком. При статистическом анализе по нескольким входным признакам одним из самых распространенных подходов является метод би-

нарной классификации на основании линейного порогового разделения. Сам подход может быть сформулирован следующим образом. Пусть имеется обучающая выборка *T* из *k* элементов в задаче бинарной классификации по *n* признакам:

$$T = \{\boldsymbol{X}, \vec{y}\}, \tag{1}$$

где $\boldsymbol{X} \in \mathrm{R}^{k \times n}$ – матрица входных значений,

$\vec{y} \in \mathrm{B}^k$ – вектор ожидаемых выходных значений,

$\mathrm{R}^{k \times n}$ – множество матриц рациональных чисел R размером $k \times n$,

B^k – множество k-мерных векторов значений булева множества $\mathrm{B} = \{0;1\}$,

k – количество элементов обучающей выборки,

n – количество признаков

Для произвольного классификатора $C(\vec{x}) : \mathrm{R}^n \to \mathrm{B}$ введем некоторую меру корректности на обучающей выборке (1), например, F_1-меру:

$$F_1[C,T] = \frac{2P^+}{2P^+ + N^+ + P^- + N^-}, \tag{2}$$

где $P^+ = \sum_{i=1}^{k} \left[C\left(\overrightarrow{x_i}\right) = 1 \middle| y_i = 1 \right]$ – количество корректно определенных положительных результатов,

$N^+ = \sum_{i=1}^{k} \left[C\left(\overrightarrow{x_i}\right) = 0 \middle| y_i = 0 \right]$ – количество корректно определенных отрицательных результатов,

$P^- = \sum_{i=1}^{k} \left[C\left(\overrightarrow{x_i}\right) = 1 \middle| y_i = 0 \right]$ – количество ошибок первого рода,

$N^- = \sum_{i=1}^{k} \left[C\left(\overrightarrow{x_i}\right) = 0 \middle| y_i = 1 \right]$ – количество ошибок второго рода,

$\overrightarrow{x_i}$ – i-я строка матрицы $\boldsymbol{X}$,

y_i – i-й элемент вектора $\vec{y}$.

Для реализации линейного порогового разделения вводится n однопараметрических классификаторов, в которых выходное значение определяется как

$$C_j(\vec{x}, p) = [x_j > p], \tag{3}$$

где x_j – значение j-го признака входного вектора $\vec{x}$,

p – пороговое значение.

Другими словами, линейное пороговое разделение подразумевает рассмотрение каждого из n признаков как линейного разделителя между двумя классами. Значение критерия (2) может использоваться, чтобы найти оптимальное пороговое значение p для каждого классификатора. Кроме того, регулируя пороговое значение для таких классификаторов может осуществляться ROC-анализ [1].

Такая реализация классификатора является крайне примитивной, поскольку каждый из классификаторов учитывает только один признак и не рассматривает их в совокупности, а также реализует простое линейное отсечение по значению этого признака. С другой стороны, такая классификация является интуитивно понятной: решение принимается на основании конкретного признака и конкретного порогового значения.

Обученные нейросетевые модели для классификации имеют намного более сложный вид. Тем не менее, поскольку именно такое представление является наиболее понятным, в работе предлагается использовать аналогичный подход для анализа семантической корректности и объяснения процесса принятия решения в некоторой локальной окрестности с помощью метода локальных линейных аппроксимаций.

Локальная линейная аппроксимация служит для построения порогового описания, аналогично определению порогового классификатора (3), по окрестности некоторой точки, которая подается на вход обученному нейросетевому классификатору. Другими словами, такой подход позволяет в схожем виде объяснить, какие из признаков в окрестности точки наибольшим образом повлияли на полученный результат. Для бинарного классификатора на основе нейронной сети прямого распространения $C_f(\vec{x}, \boldsymbol{\Theta}) : \mathrm{R}^n \to \mathrm{B}$ с набором параметров $\boldsymbol{\Theta}$, которые подобраны в результате обучения на обучающей выборке (1), локальная линейная

аппроксимация для интерпретации поведения модели в окрестности произвольной точки $\vec{x}^*$ может быть построена следующим образом. Из входных векторов обучающей выборки (1), задаваемых матрицей $\boldsymbol{X}$, выбирается вектор, ближайший к $\vec{x}^*$. В качестве расстояния для поиска ближайшего вектора может использоваться любая функция векторного расстояния, например, функция Евклидова расстояния:

$$d\left(\vec{x}_1, \vec{x}_2\right) = \sqrt{\sum_i^n \left(x_{1i} - x_{2i}\right)^2}. \quad (4)$$

После нахождения ближайшей точки и минимального расстояния d_{min} осуществляется сэмплирование окрестности $\vec{x}^*$ – случайная генерация набора точек $\left\{\vec{x}^{(s)}\right\}$ фиксированной длины m, которые удалены от $\vec{x}^*$ не больше, чем на d_{min}, в соответствии с расстоянием (4). Для каждой из m полученных точек $\left\{\vec{x}^{(s)}\right\}$ определяется соответствующее значение, вычисленное обученным классификатором:

$$y_i^{(s)} = C_f\left(\vec{x}_i^{(s)}, \boldsymbol{\Theta}\right). \quad (5)$$

Набор точек $\left\{\vec{x}^{(s)}\right\}$ и соответствующих им выходных значений классификатора $\left\{y_i^{(s)}\right\}$ характеризуют непосредственную окрестность точки $\vec{x}^*$ и поведение классификатора C_f в окрестностях этой точки, соответственно. Для построения локальной линейной аппроксимации для этих точек осуществляется построение простого линейного классификатора $C^*\left(\vec{x}, \vec{\theta}\right): \mathrm{R}^n \to \mathrm{B}$ с использованием этих точек в качестве обучающей выборки. Для самого классификатора может использоваться любой подходящий метод линейной классификации, в том числе линейной регрессией при помощи метода наименьших квадратов, метод Фишера, логистическая регрессия, а также искусственная нейронная сеть прямого распространения с линейной функцией активации. Основная задача построения линейного классификатора состоит в нахождении такой гиперплоскости $p^*\left(\vec{x}\right) = \theta_0 + \vec{\theta} \cdot \vec{x}$, которая разделяет исходное n-мерное пространство признаков на два подпространства в окрестности точки $\vec{x}^*$, одно из которых в пределах расстояния d_{min} с наибольшей вероятностью содержит экземпляры одного класса классификатора C_f, а другое – другого. Эта гиперплоскость задается при помощи $(n+1)$-мерного вектора параметров классификатора $\vec{\theta}$ и может быть описана аналитически.

По локальному линейному классификатору C^* и соответствующей ему разделяющей гиперплоскости $p^*\left(\vec{x}\right)$ может быть рассмотрена точка пересечения этой гиперплоскости с осью каждого из n исходных признаков. Пусть угол между единичным вектором оси x_n и вектором нормали к плоскости $p^*\left(\vec{x}\right)$ составляет φ_{n0}, а точка пересечения гиперплоскости $p^*\left(\vec{x}\right)$ с осью x_n, если она существует, находится в $x_n = x_{n0}$. Тогда по углу φ_{n0} можно определить влияние n-го признака на результат классификации в окрестности точки $\vec{x}^*$.

Если разделяющая гиперплоскость $p^*\left(\vec{x}\right)$ проходит перпендикулярно оси x_n, т.е. $\varphi_{n0} = 0$, то можно утверждать, что в окрестности точки $\vec{x}^*$ решение о выборе конкретного класса обученным классификатором C_f принимается исключительно по значению признака x_n, то есть в такой окрестности этот признак является наиболее значимым. Аналогично, если разделяющая гиперплоскость не пересекает ось x_n, т.е. $\varphi_{n0} = \frac{\pi}{2}$, то можно утверждать, что обученный классификатор C_f принимает решение о выборе конкретного класса вне зависимости от того, какое значение примет признак x_n, то есть в такой окрестности этот признак никак не влияет на результат. Таким образом, в качестве значения «веса» влияния n-го признака на результат классификации точки $\vec{x}^*$ (и ее небольшой окрестности) может использоваться значение $w_n^* = \cos\varphi_{n0}$. Для исключения влияния размерности пространства полученные значения для каждой из осей могут нормироваться на максимальное значение.

Семантически полученные значения весов w_i^* и точек пересечения x_{n_0} могут быть интерпретированы в виде логических высказываний следующего вида: «Для точки $\vec{x}^*$ обученный классификатор C_f принял решение $y_f^* = C_f\left(\vec{x}^*, \boldsymbol{\Theta}\right)$, причем с весом w_i^* на это решение повлиял тот факт, что i-й признак принял значение x_i^*, большее (меньшее), чем x_{i0}». Полученные значения весов w_i^* могут быть отсортированы по убыванию чтобы определить, какие из признаков имели наибольшее влияние на выходной результат.

Полученные локальные аппроксимации предоставляют собой, по сути, описание характера поведения функции классификатора C_f в окрестности конкретной точки в виде строгого логического высказывания. Следует отметить, что предоставленные описания справедливы исключительно для окрестности точки $\vec{x}^*$. Поскольку сам вид полученной функции классификатора C_f может быть сколь угодно сложный, нельзя утверждать, что поведение обученной модели будет корректно описано линейным классификатором для всего набора данных. Таким образом, для более строгой формализации процесса принятия решений в обученном бинарном классификаторе на основании нейронной сети прямого распространения схожие интерпретации следует получить в других характерных точках, где влияния исходных признаков на конечный результат могут принимать другие значения, а затем определить границы корректности для каждого описания. Такая формализация, по сути, будет представлять собой кусочно-линейное описание функции, которую аппроксимирует классификатор C_f после обучения, поэтому ее можно использовать для преобразования обученной нейросетевой модели в пару деревьев принятия решений: вначале по набору значений признаков определяется характерная точка $\vec{x}^*$, к окрестности которой относится входной вектор, а затем по полученной локальной окрестности этой точки принимается конечное решение на основании аппроксимирующего линейного классификатора C^*, который статистически справедлив для этой окрестности. Проанализировав полученные деревья принятия решений, во многих случаях можно убедиться в семантической корректности принимаемого решения и получить новые знания в предметной области на основании описанного в дереве процесса вывода.

Таким образом, метод локальных линейных аппроксимаций может использоваться для описания поведения классификационной нейросетевой модели в окрестности некоторой точки в линейном виде, что упрощает понимание того процесса, на который опирается модель при принятии конкретного решения. Хотя такой подход не предоставляет полное описание процесса принятия решения, проводимый анализ позволяет для некоторой совокупности входных признаков получить представление о том, какие из этих признаков являлись определяющими в вычислении итогового значения по этой модели. Представленный алгоритм может использоваться для построения интуитивно понятной интерпретации процесса формирования выходного результата по конкретной совокупности признаков. Полученные интерпретации могут использоваться экспертами в предметной области как с целью проверки семантической корректности процесса принятия решения, так и с целью синтеза новых знаний и формирования более глубокого понимания тех зависимостей, которые присущи исследуемым данным и которые смогли быть выявлены в процессе обучения.

ЛИТЕРАТУРА

1. Goodfellow, I. Deep learning / I. Goodfellow, B. Yoshua, C. Aaron. – Cambridge : The MIT Press, 2016. – 775 p.
2. Курочкин, А.В. Оптимизация процесса принятия решений в медицинских экспертных системах на базе нечеткой логики с использованием исторических данных / А.В. Курочкин, В.С. Садов, О.М. Демиденко // Проблемы физики, математики и техники. – 2019. – № 1 (38). – с. 78–84.
3. Ioannou, Y. Decision Forests, Convolutional Networks and the Models in-Between / Y. Ioannou, D. Robertson, D. Zikic, [et al.] // arXiv:1603.01250[cs] [Электронный ресурс]. – 2016. – Режим доступа: https://arxiv.org/abs/1603.01250. – Дата доступа: 08.10.2019.
4. Lundberg, S.M. A Unified Approach to Interpreting Model Predictions / S.M. Lundberg, S.-I. Lee // Advances in Neural Information Processing Systems. – 2017. – № 30. – P. 4765–4774.
5. Knowledge Acquisition for Expert Systems : A Practical Handbook / ed. by A.L. Kidd. – Springer Science & Business Media, 2012. – 208 p.

Поступила 12.09.2019

ANALYSIS AND SEMANTIC INTERPRETATION METHODS OF DECISION-MAKING PROCESS IN NEURAL NETWORK SUPERVISED LEARNING CLASSIFICATION MODELS

A. KURACHKIN, V. SADAU

The paper focuses on the problem of interpreting the behavior of a feedforward neural network classifier model after learning, and proposes solutions to analyze and describe model output based on local linear approximations.

Keywords: *neural networks, decision support systems, neural network model analysis.*