*А.В. Гаврилов*
Новосибирский государственный технический университет

# Проблемы обработки символьной информации в нейронных сетях

The problem of coding input symbol data and decoding of output binary vectors in applications based on neural networks is discussed in this report. The methods of coding with saving of semantic affinity between coded conceptions are proposed, in particular, based on lingustic variable.

При практическом использовании нейронных сетей при построении экспертных систем [1], для обработки текстовой информации или для анализа баз данных [2] одной из трудностей является ориентация нейронных сетей на обработку сигналов, а не символьной информации. В наиболее развитых моделях искусственных нейронных сетей (ИНС) входная информация представляется в виде двоичного вектора. При использовании таких моделей имеет место проблема кодирования входной информации ИНС и декодирования выходного вектора ИНС.

Если кодировать символьную информацию на входе ИНС "беспорядочно", т.е. не заботиться о корреляции между значениями двоичных векторов и соответствующими им символьными значениями, то близкие по семантике символьные значения могут кодироваться совершенно разными двоичными векторами, отстоящими друг от друга на очень большое расстояние в пространстве состояний нейронной сети. Это затрудняет обучение нейронной сети и может приводить к ошибкам при функционировании обученной ИНС. К такому же эффекту могут приводить и орфографические ошибки во входной информации, когда искаженное слово воспринимается как новое, а также, выход из строя нейроподобных элементов в случае аппаратной реализации ИНС. Кроме того, при использовании неполносвязных моделей ИНС, где ненулевая вероятность ошибки является особенностью архитектуры ИНС, желательно свести к минимуму эффект этой ошибки. Например, наверное, допустимо, если вместо решения "старый" на выходе ИНС появится семантически близкое значение "пожилой", но совершенно не допустимо, если ИНС сформирует решение "молодой".

Для исключения этих недостатков при использовании нейронных сетей для обработки символьной информации предлагается использовать следующие принципы :

- разбиение входного вектора на подвекторы, кодирующие разные компоненты символьной информации, поступающей на нейронную сеть (например, разные поля реляционной базы данных или разные аспекты контекста обрабатываемого нейронной сетью текста ), при этом для кодирования подвекторов необходимо использовать тезаурусы с фиксированным количеством слов в каждом из них,
- использование представления лингвистической переменной для кодирования семантически близких значений, которые могут быть связаны с метрической шкалой,
- использование классификации понятий и определение семантических шкал для них ( задание отношений частичного порядка на множестве понятий, семантически близких в определенном контексте, задаваемом классом и признаком классификации).

При кодировании символьной информации на входе ИНС необходимо использовать фиксированный тезаурус, свой для каждого подвектора входного вектора ИНС. Конечно, можно кодировать входные слова как произвольные последовательности символов. В этом случае набор используемых слов ничем не ограничен, и в процессе функционирования системы могут появляться но-

вые, ранее не использованные слова. Но в этом случае они могут восприниматься нейронной сетью только как сигналы, и ни о каком использовании семантической близости понятий при работе ИНС не может быть речи. Вся тяжесть построения разделяющей гиперповерхности для трудно разделяемых входных векторов в пространстве признаков ложится в этом случае на нейронную сеть. К тому же, в этом случае теряются основания для разбиения входного вектора на подвекторы, т.к. заранее не известны их длины.

При использовании значений лингвистической переменной (ЛП) в качестве входной информации для нейронной сети с бинарными входами целесообразно кодировать значения ЛП так, чтобы расстояние между максимальными значениями функции принадлежности на метрической шкале взаимно однозначно соответствовало расстоянию Хэмминга между соответствующими двоичными входными векторами ИНС, и отношение частичного порядка на множестве этих максимальных значений сохранялось на множестве соответствующих расстояний Хэмминга В этом случае можно предположить, что вероятность ошибок при распознавании значений ЛП будет минимальной. Естественно, что сохранение семантической близости двоичных векторов при таком кодировании приводит к избыточности разрядов. При кодировании " в лоб" достаточно int($\log_2 n$) двоичных разрядов, где n - количество значений ЛП, int - округление до большего целого. При кодировании с сохранением семантической близости при строгом подходе требуется (n-1) двоичных разрядов.

Например, пусть нейронная сеть должна обрабатывать значения ЛП "возраст", принимающую значения "дитя", "ребенок", "юный", "молодой", "зрелый", "пожилой", "старый", "очень старый". Для кодирования "в лоб" ( порядке перечисления слов ) достаточно 8 двоичных разрядов и код для "дитя" будет 000, а для "зрелый" - 100. Если нейронная сеть ошибется в одном (2-м) разряде, это приведет к тому, что вместо "дитя" мы получим "зрелый" или наоборот. При кодировании с сохранением семантической близости можно использовать следующие коды :

"дитя" - 0000000,
"ребенок" - 0000001,
"юный" - 0000011,
"молодой" - 0000111,
"зрелый" - 0001111,
"пожилой" - 0011111,
"старый" - 0111111,
"очень старый" - 1111111.

Алгоритм кодирования, используемый здесь, очевиден.

Можно уменьшить избыточность, сняв требование строгого соответствия между расстоянием Хэмминга и расстоянием на метрической шкале:

"дитя" - 00000,
"ребенок" - 00001,
"юный" - 00011,
"молодой" - 00111,
"зрелый" - 01111,
"пожилой" - 11111,
"старый" - 11110,
"очень старый" - 11100.

При этом способе кодирования сначала расстояние Хэмминга от первого значения до текущего кодируемого растет, а с некоторого значения начинает падать. Уровень избыточности можно задавать ограничением на расстояние Хэмминга между крайними на шкале значениями ЛП. Любопытно, что в этом случае расстояние между значениями "дитя" и "очень старый" меньше, чем между "дитя" и "зрелый".

В случае использования в качестве входной информации произвольных символьных значений, которые не возможно представить в виде значений лингвистической переменной, можно использовать разбиение их на классы и определение для каждого класса своей семантической шкалы в контексте признака классификации и, может быть, признака, по которому оценивается се-

мантическая близость между представителями заданного класса. На семантической шкале определяется отношение частичного порядка между значениями, принадлежащими данному классу, и семантическое расстояние между двумя значениями, равное количеству значений, находящихся между ними на шкале, увеличенному на 1. Например, класс "мебель" можно представить следующими значениями в порядке их расположения на семантической шкале: "кровать", "диван", "кресло", "стул", "журнальный столик", "письменный стол", "обеденный стол", "кухонный стол", "буфет", "шкаф". Семантическое расстояние между понятиями "кровать" и "диван" равно 1, между "кровать" и "стул" - 3, между "кровать" и "обеденный стол" - 6. Наименование класса и признака классификации кодируются отдельно. В этом случае классификация может производиться другой нейронной сетью, что может быть реализовано в известных ансамблевых моделях ИНС [3].

На выходе нейронной сети полученный в результате работы ИНС двоичный вектор необходимо декодировать, т.е. преобразовать его в одно или несколько символьных значений. При этом также можно использовать его разбиение на подвекторы, каждому из которых соответствует компонент решения со своим множеством возможных символьных значений.

При декодировании выходного вектора надо учитывать одно из возможных требований, которые могут предъявляться к решению, получаемому нейронной сетью :

- исключить или уменьшить вероятность ложного ( неверного ) решения при сохранении возможности не получить никакого,
- исключить отсутствие какого-либо решения, может быть, в ущерб качеству, при этом предполагается, что отсутствие ошибки гарантируется качеством обучения и особенностями архитектуры нейронной сети.

В первом случае необходимо вводить избыточность в кодирование значений символьных решений и двоичные вектора, не соответствующие кодам значений из тезауруса решений, не декодировать ( случай "отсутствия решения" ).

Во втором случае избыточность при кодировании не требуется, а если она используется ( например, для обеспечения симметричности методов кодирования на входе и выходе сети ), в случае несоответствия выходного вектора какому-либо из значений тезауруса выбирается и декодируется ближайший (по Хэммингу) код.

Предложенные подходы в настоящее время используются при создании "двухполушарной" [1] экспертной системы и программы анализа баз данных на основе ИНС [2].

Литература


1. Gavrilov A.V. The method of the combination of logic and associative proceeses in Expert Systems. - / Труды межд. семинара "Мягкие вычисления-96", Казань, 1996.-С. 84-86.
2. Гаврилов А.В., Канглер В.М., Катомин М.Н., Коротенко А.И. Обнаружение ассоциативных взаимосвязей между полями в базах данных с использованием нейронной сети. - / Труды межд. н.-т. конф. "Научные основы высоких технологий", Том 2, Новосибирск, 1997.- С.210-211.
3. Куссуль Э.М. Ассоциативные нейроподобные структуры. - Киев, Наукова думка, 1990.