Выявление дискретного носителя знаний в процессе глубокого обучения с помощью технологии OLAP

В. И. Пименов [1], email: v_pim@mail.ru
И. В. Пименов [2]

[1] Санкт-Петербургский государственный университет промышленных технологий и дизайна
[2] Государственный университет морского и речного флота имени адмирала С. О. Макарова

Аннотация. *Рассматривается методика построения и интерпретации OLAP-куба с бинарными мерами из данных, полученных на обученной нейросети. Такая дискретная форма имплицирует выделенные классы объектов с помощью обобщающих правил и обеспечивает когнитивную визуализацию классов в многомерном пространстве.*

Ключевые слова: *многомерный анализ данных, глубокое машинное обучение, нейронная сеть, OLAP-куб, когнитивные технологии, решающее правило*.

Стратегия развития искусственного интеллекта

В рамках стратегии развития искусственного интеллекта на период до 2030 года, национальной программы «Цифровая экономика Российской Федерации» и в других проектах уделяется внимание перспективным технологиям искусственного интеллекта, к которым относятся системы поддержки принятия решений, машинное зрение, обработка естественного языка, распознавание и синтез речи, распознавание изображений, информационный поиск, машинный перевод, биометрическая аутентификация, символьное моделирование рассуждений и другие методы. Их использование способствует повышению экономической эффективности предприятий, улучшению бизнес-процессов и формированию принципиально новых направлений деятельности в сферах управления логистикой, оптимизации планирования поставок, финансовых операций, производственных процессов, прогнозирования рисков, повышения удовлетворенности потребителей, оптимизации процессов подбора кадров, составления графика работы сотрудников, диагностики заболеваний, подбора дозировок лекарственных препаратов, оптимизации выбора профессии с помощью анализа показателей эффективности обучения [1].

В рамках слабого искусственного интеллекта используются алгоритмы глубокого машинного обучения. Они трудно интерпретируются и решают только узкие классы задач. Создание сильного искусственного интеллекта, универсального в применении к различным задачам, способного функционировать, адаптироваться и взаимодействовать с внешней средой подобно человеку, является сложной научной проблемой.

При переходе производства на цифровой формат, включения в контур управления различных датчиков, привлечения информации из внешних источников и интернета происходит взрывообразный рост объемов данных, что заставляет использовать технологии агрегирования и обработки данных, выполнять построение модели “белого ящика”, осуществлять семантическую интерпретацию решений с их последующим предоставлением клиентам. Чтобы решить все актуальные проблемы имеющихся задач искусственного интеллекта на период до 2030 года, будет достаточно существующего объема разнообразных данных из открытых источников.

В проблеме извлечения знаний из данных инструментальной базой являются методы многомерного анализа и машинного обучения [2]. Разнотипность признаков объектов, их взаимозависимость, необходимость организации данных при обработке, делают сложным комплексное, системное использование методов. При существующей многомерности в описаниях объектов предметной области, автоматическое создание баз знаний и интеллектуальных систем может основываться на универсальном алгоритме сильного искусственного интеллекта, комплексно использующем методы многомерного анализа и обеспечивающем семантическую интерпретацию полученных решений. В процессе обучения ранжируются значимые атрибуты, классифицируются данные, структурируются и отбираются понятия, представляющие объект, и устанавливаются причинно-следственные связи между свойствами объекта и его показателями эффективности, а также выполняется их когнитивная визуализация.

При обучении систем классификации и распознавания преобразование непрерывного признакового пространства в конечное множество классов может быть организовано с помощью модели, представленной многомерным OLAP-кубом с бинарными мерами. Такая дискретная форма знаний обеспечивает возможность их интерпретации на основе различных методов, например, продукционных правил, и позволяет выполнять когнитивную визуализацию многомерных классов.

1. Классифицирующая нейронная сеть как модель “черного ящика”

Разработка правил и алгоритмов, определяющих принадлежность многомерного объекта к определенному классу и осуществляющих визуализацию найденного решения, базируется на предварительной кластеризации данных на значимые части.

Индуктивное обучение нейросетевых моделей использует описания объектов ω_i, $i = \overline{1, n}$,значениями признаков $(x_{i1}, ..., x_{ij}, ..., x_{iN})$, которые характеризуют свойства $\{X_j \mid j = \overline{1, N}\}$ объектов-прецедентов. Для построения модели классификации также указывается принадлежность объектов ω_i к одному из классов Ω_m, $\Omega_m \subset \Omega$, $\Omega = \Omega_1 \cup \Omega_2 \cup ... \cup \Omega_M$. Имея достаточный априорный словарь признаков, можно сформировать сепарабельное пространство, в котором объекты обучающей выборки разделяются непересекающимися выпуклыми оболочками классов (рис. 1).

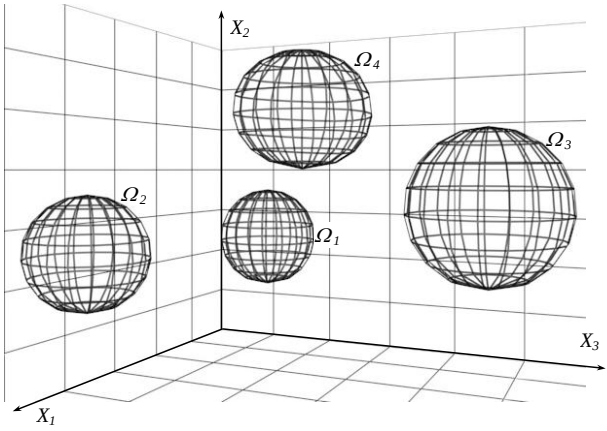

Рис. 1. Оболочки классов в многомерном пространстве

В процессе обучения нейронной сети знания формируются как набор весовых коэффициентов связей между нейронами (рис. 2). Поскольку обучение нейросетевой модели основано на подходе “черного ящика”, то, несмотря на широкий спектр решаемых задач, объяснение и вербализация полученных нейросетью результатов чрезвычайно затруднительны [3].

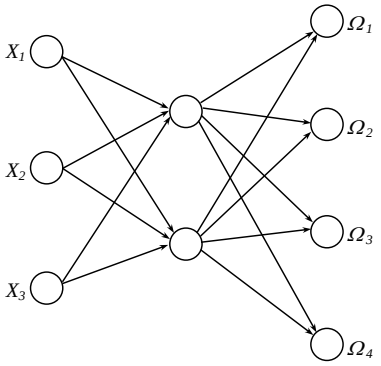

Рис. 2. Структура классифицирующей нейросети

Однозначное отображение характерных и общих признаков класса объекта в многомерных непрерывных пространствах является проблемой при визуализации решений.

2. Обученная нейросеть – источник извлечения дискретного носителя знаний

В процессе классификации нейросетевая модель преобразует непрерывное признаковое пространство в конечное, дискретное множество классов.

При интерпретации построенной нейросети примем во внимание, что закономерности обработки информации выражаются связями между нейронами, а сущности предметной области представляются вектором выходных сигналов. Для выражения в явной форме связи между комбинациями значений N признаков и классами может быть применена дискретная модель многомерного OLAP-куба с бинарными мерами (значениями ячеек) [4].

Метками измерений являются градации значений признаков.

Число градаций признаков и места расположения порогов определяются в процессе адаптивного квантования признакового пространства с помощью алгоритма минимального числа порогов, обеспечивающего разделение всех классов, непересекающихся по данному признаку [5].

Построение интервалов изменения исходных признаков $\{X_j \mid j = \overline{1, N}\}$ внутри заданных классов Ω_m выполняется посредством независимого варьирования значения X_j на входе многослойной

нейронной сети при выставленных средних значениях остальных признаков, когда срабатывает m -й выходной нейрон.

Если значения количественного признака X_j принадлежат интервалу с номером i, или у объектов m -го класса установлен бинарный признак X_j, то градации значений признаков x_{ij} для класса Ω_m принимают единичные значения в ячейках OLAP-куба

$$x_{ij}(m) = \begin{cases} 1, \exists \omega \in \Omega_m, x_j \in (d_{(i-1)j}, d_{ij}), m = \overline{1, M}, i = \overline{1, t_j}, \\ 0, \text{ в противопол ожном случае,} \end{cases}$$

где t_j – количество градаций (номинальных значений) признака X_j.

Подкуб дискретизированного многомерного пространства для класса Ω_4 представлен на рис. 3.

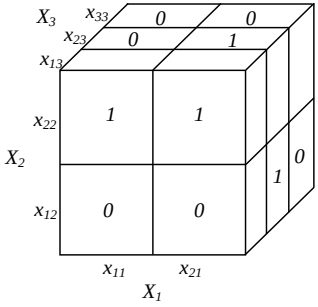

Рис. 3. Подкуб Ω_4 дискретизированного многомерного пространства

Значения признаков в таком дискретном классифицирующем пространстве выставляются в виде единичных элементов OLAP-куба и пороговых уровней. Таким способом обеспечивается легкая семантическая интерпретация решающего правила, построенного на основе обученной нейросети.

3. Интерпретация дискретного носителя знаний

Для интерпретации OLAP-куба с бинарными мерами используется система продукционных решающих правил, выбирающих класс

объектов с учетом комбинации значений дискретизированных признаков

$(X_1, X_2, ..., X_N) \xrightarrow[\text{РП}\,(X_1, X_2, ..., X_N)]{} \Omega_m$, N – глубина пространства поиска.

Правила-продукции имеют вид

$$\text{"Если}\quad (x_j \in (d_{(i-1)1}, d_{i1})_m \text{ и} \ldots x_j \in (d_{(i-1)j}, d_{ij})_m \text{ и} \ldots$$

$$x_N \in (d_{(i-1)N}, d_{iN})_m), \text{ то } \omega \in \Omega_m\text{"},$$

где x_j – значение j-го признака, $\overline{j = 1, N}$; $(d_{(i-1)j}, d_{ij})_m$ – i-й интервал квантования для класса объектов Ω_m.

Значения признаков объекта указывают на ячейки OLAP-куба. При распознавании происходит поэлементная конъюнкция ячеек, выделяющая единичную ячейку, соответствующую коду класса. Найденному объекту соответствует пространство “своих” градаций.

4. Когнитивная визуализация многомерных образов

Однозначное и точное отображение характерных и общих признаков класса объектов является проблемой при визуализации решений в многомерных непрерывных пространствах.

Если данные спроецировать на плоскость, заданную пользователем, или отобразить признаковое пространство в пространство главных компонент, то используя когнитивное облако точек, можно визуализировать решение. Однако нахождение оптимальной ориентации плоскости проекции требует значительных усилий, учитывая многомерность данных. Для однозначной идентификации класса на основе OLAP-куба требуется проанализировать $N(N-1)/2$ срезов.

Когнитивная визуализация построенного, в процессе глубокого обучения, бинарного куба выполняется в пространстве “признак – градация признака” (рис. 4). Образ класса выглядит как прямоугольная карта с ячеистой структурой размерности $N \times T$. Максимальное количество градаций T задается по наиболее дискретизированному признаку. При безошибочном разделении нейросетью объектов обучающей выборки, образы классов в дискретизированном многомерном пространстве являются уникальными.

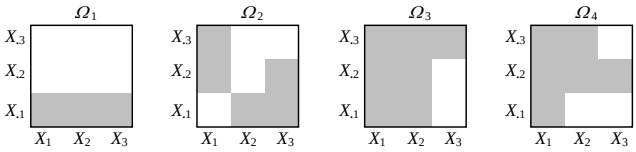

Рис. 4. Образы классов после кодирования в дискретизированном многомерном пространстве признаков

Нелинейная сигмоидная функция активации с достаточной крутизной выполняет бинаризацию суммарного сигнала и применяется для заполнения бинарной карты. Отличная от нулевой мощность характеристического множества классов на t-м интервале квантования j-го признака означает заполнение ячейки черным цветом для соответствующей t-й градации j-го признака.

Для каждого m-го образа можно определить его интегральные показатели – сумму градаций по всем дискретизированным признакам и разброс значений признаков, которые определяют местоположение и размер класса в многомерном пространстве.

Заключение

Иерархическая организация признаков, использующая множество нелинейных преобразований низкоуровневых атрибутов в высокоуровневые показатели, позволяет получить хорошие результаты, применяя методы глубокого машинного обучения. Такие методы моделируют абстрактные категории, которые сложно интерпретировать.

Происходящее в процессе обучения адаптивное квантование многомерного признакового пространства, накопление комбинаций значений дискретизированных признаков, имплицирующих выделенные классы объектов с помощью обобщающих правил и обеспечивающих семантическую интерпретацию дискретного носителя знаний, позволяют извлечь OLAP-куб с бинарными мерами из данных, полученных на обученной нейросети. Квантование и сжатие признакового пространства дают возможность интегрировать когнитивный образ в простую геометрическую фигуру.

Список литературы

1. Национальная стратегия развития искусственного интеллекта на период до 2030 года [Электронный ресурс]. – Режим доступа: http://ivo.garant.ru/#/document/72838946/paragraph/12:0

2. Пименов, В. И. Методика анализа больших данных в системах поддержки принятия решений / В. И. Пименов, И. В. Пименов // IT-

технологии: развитие и приложения: материалы XV-й междунар. конф. (Владикавказ, 12-14 декабря 2018 г.). – Владикавказ, 2018. – С. 321-332.

3. Xuan, L. Improving the interpretability of deep neural networks with knowledge distillation / L. Xuan, W. Xiaoguang, M. Stan // IEEE International Conference on Data Mining Workshops (ICDMW). – 2018. – P. 905-912.

4. Davardoost, F. Extracting OLAP cubes from document-oriented NoSQL database based on parallel similarity algorithms / F. Davardoost, Sangar A. Babazadeh, K. Majidzadeh // Canadian Journal of Electrical and Computer Engineering. – Spring, 2020. – Vol. 43. – No. 2. – P. 111-118.

5. Пименов, В. И. Когнитивная обработка многомерных данных на основе построения дискретных моделей знаний / В. И. Пименов, И. В. Пименов, Е. С. Кокорин // Сборник трудов международной научной конференции “Актуальные проблемы прикладной математики, информатики и механики” (Воронеж, 11-13 ноября 2019 г.). – Воронеж, 2020. – С. 338-344.