ИЗВЛЕЧЕНИЕ ЯВНЫХ ЗНАНИЙ ИЗ РАЗНОТИПНЫХ ДАННЫХ С ПОМОЩЬЮ НЕЙРОННЫХ СЕТЕЙ

Н. А. Игнатьев
Национальный университет Узбекистана, Ташкент
e-mail tin000@tashsu.silk.org

The methods of extraction of implicit knowledge from various data types at synthesis of neural networks in the form of the solution of minimum cover by objects — measurement standards of learning sample problem are considered

Введение

Автоматизация процесса извлечения явных знаний из больших объемов данных — одно из приоритетных направлений в области искусственного интеллекта. Извлечение знаний с помощью обучаемых нейронных сетей (НС) может выражаться в форме объяснения процесса принятия решений в задачах распознавания образов, прогнозирования. Пример такого объяснения — определение количества нейронов и структуры НС, необходимых для решения конкретной задачи.

В работах [1, 2] показано, что введение определенных ограничений над множеством значений признаков позволяет использовать точные методы для вычисления синаптических весов нейронов и определять их количество. Поиск минимальной конфигурации однослойной НС осуществлялся в форме решения задачи минимального покрытия обучающей выборки объектами-эталонами.

Согласно методу, описанному в [2], состав множества объектов покрытия определяется выбранной схемой (порядком) подачи объектов — кандидатов на удаление из обучающей выборки процедурой “последовательного исключения”. Изменение порядка подачи, как правило, приводит к разным множествам объектов, образующих покрытие.

В настоящей работе рассматривается метод построения НС, ориентированный на случай, когда число объектов в выборке достаточно велико, чтобы использовать последовательный алгоритм, описанный в [2]. Генетический алгоритм, реализующий этот метод, позволяет получить локально-оптимальное покрытие выборки объектами-эталонами, используя различные рекомбинации состава множества объектов покрытия. Так же, как и в [3], результаты генетического алгоритма могут использоваться при построении минимально допустимой обучающей выборки для нейросетевой системы принятия решений.

Теоретический и практический интерес представляет проверка гипотезы, что множество локально-оптимальных покрытий эталонами обучающей выборки замкнуто относи-

тельно операций объединения и пересечения. В качестве нового источника знаний предлагается использовать:

— частотные характеристики номинальных признаков, определяющие внутриклассовое сходство и межклассовое различие;

— значения синаптических весов НС для отбора информативных наборов признаков.

1. Генетический алгоритм построения минимального покрытия обучающей выборки объектами-эталонами

Постановка задачи и изложение метода имеют некоторую аналогию с описанными в [2]. Считается, что множество допустимых объектов $E_0 = \{S_1, ..., S_m\}$ — обучающая выборка, в которой заданы представители l непересекающихся классов $K_1, ..., K_l$. Каждый допустимый объект описывается с помощью n признаков, r из которых количественные $(0 \le r \le n)$, $n - r$ — номинальные.

Пусть I, J — множества номеров соответственно количественных и номинальных признаков, используемых для описания допустимых объектов и объект $S_j \in E_0\,(S_j = (x_{j1}, ..., x_{jn}))$ является эталоном выборки. Веса количественных признаков эталона вычисляются как $w_{jt} = x_{jt}\,\forall t \in I$ и $w_{j0} = -\dfrac{1}{2}\sum\limits_{t\in I} w_{jt}^2$.

Значения весов номинальных признаков определяются так же, как и в [2], на основе предположения об одинаковом характере различий между количественными и номинальными признаками, т.е. считается, что максимальное различие объектов по количественным признакам соответствует максимальному возможному различию по каждому номинальному признаку. Вычисляются предельные значения, — общие для всех номинальных признаков

$$w_{\max} = \max_{S_j \in E_0}(-2w_{j0}/r),$$

$$\lambda_{\max} = \sum_{t=1}^{l} \mid K_t \mid (\mid K_t \mid -1),$$

$$\beta_{\max} = \sum_{t=1}^{l} \mid K_t \mid (m - \mid K_t \mid).$$

Обозначим через p число градаций признака $c \in J$, через g_{dc}^t — количество значений t-й $(1 \le t \le p)$ градации c-го признака в описании объектов класса K_d. Тогда

$$\lambda_c = \sum_{i=1}^{l}\sum_{t=1}^{p} g_{ic}^t(g_{ic}^t - 1), \tag{1}$$

$$\beta_c = \sum_{i=1}^{l}\sum_{t=1}^{p} \begin{cases} g_{ic}^t(\mid CK_i \mid - b_{ic}^t), g_{ic}^t \ne 0, \\ b_{ic}^t \mid K_i \mid, g_{ic}^t = 0, \end{cases} \tag{2}$$

где b_{ic}^t — количество значений t-й градации c-го признака в CK_i-дополнении класса K_i. Вес каждого номинального признака $c \in J$ определяется по формуле

$$w_{jc} = \left(\frac{\lambda_c}{\lambda_{\max}}\right)\left(\frac{\beta_c}{\beta_{\max}}\right) w_{\max}. \tag{3}$$

Значение взвешенной суммы по объекту-эталону $S_j \in E_0$ для произвольного допустимого объекта $S = (a_1, ..., a_n)$ вычисляется как

$$\varphi(S, S_j) = \sum_{i \in I} w_{ji} a_i + \sum_{i \in J, x_{ji} = a_i} w_{ji} + w_{j0}. \tag{4}$$

Согласно решающему правилу, реализующему принцип "победитель забирает все", объект S принадлежит к тому классу, значение взвешенной суммы (4) объекта-эталона которого максимально.

Алгоритм построения НС, изложенный в [2], рассчитан на обработку обучающей выборки с ограниченными, заранее заданными размерами. При значительном числе объектов обучающей выборки (1000 и более) из-за больших затрат вычислительных ресурсов машинная реализация такого алгоритма практически неприемлема.

В предлагаемом здесь генетическом алгоритме для нахождения локально-оптимального покрытия обучающей выборки E_0 эталонами последовательность выполнения действий такова. Из выборки E_0 случайным образом формируется подвыборка данных E_0^0 (популяция индивидуумов). Состав объектов E_0^0 предопределяет начальное приближение покрытия F^0 всей обучающей выборки, получаемого в результате выполнения процедуры "последовательного исключения". Краткая суть работы этой процедуры на i-м ($i = 0, 1, 2, ...$) шаге итерации заключается в следующем.

Изначально все объекты E_0^i считаются эталонами покрытия F^i, т. е. $F^i = E_0^i$. Если при использовании $F^i \backslash \{S_t\}$ в качестве эталонов алгоритм распознавания — корректный (не делает ошибок) на E_0^i, то производится удаление объекта S_t из F^i. Условие, что исключение любого объекта-эталона из F^i и последующее использование F^i алгоритмом распознавания приводят к ошибкам на E_0^i, служит критерием останова выполнения процедуры на i-м шаге.

Пусть $P^i (P^i \in E_0)$ — множество некорректно распознанных на E_0 объектов при использовании F^i в качестве эталонов в (4). Подвыборка E_0^{i+1} формируется путем включения покрытия F^i и случайным образом отобранных объектов из P^i. Описанный процесс повторяется до тех пор, пока на некотором r-м ($r \geq 0$) шаге итерации не выполнится условие $P^r = \emptyset$.

2. Системы, основанные на знаниях

Одной из разновидностей систем, основанных на знаниях (СОЗ), является автоматическое гипотезирование (порождение гипотез), реализуемое в форме обучения НС. Считается, что система с элементами искусственного интеллекта обладает априорными знаниями и реализует кумулятивный (накопительный) вид обучения, когда накопление знаний, вообще говоря, улучшает способность к обучению с учителем или без него [4].

В качестве одного из источников знаний при реализации описанного выше метода синтеза НС можно рассматривать значения величин, характеризующих внутриклассовое

сходство ($\lambda_c/\lambda_{\max}$) и межклассовое различие ($\beta_c/\beta_{\max}$), вычисляемых соответственно с помощью (1), (2). Множество значений этих величин лежит в интервале $[0, 1]$, что облегчает их интерпретацию в терминах нечетких логик. Упорядочение значений (3) позволяет эксперту-исследователю делать заключения о том, какие из номинальных признаков с малыми значениями весов следует считать неинформативными.

Локально-оптимальные покрытия обучающей выборки E_0 эталонами получаются как результат различных схем подачи объектов — кандидатов на вход процедуры "последовательного исключения", число этих схем не превышает $m!$. С позиций извлечения знаний представляют интерес такие вопросы:

1. Совпадает ли объединение всех покрытий с исходной выборкой?

2. Как можно сравнивать два и более покрытия?

3. Какие отношения можно ввести на множестве различных покрытий, какие свойства имеют эти отношения и какие классы отношений они образуют?

Пусть испытано k схем подачи кандидатов на исключение и $M_1, ..., M_k$ — множества объектов покрытия, получаемые по каждой из k схем. В первую очередь интерес представляют результаты применения на $M_1, ..., M_k$ классических операций пересечения и объединения множеств. Объединение $\bigcup_{i=1}^{k} M_i \neq E_0$ может быть проинтерпретировано как наличие:

а) объектов, не определяющих структуру классов;

б) ограниченного множества объектов $Z_0(| Z_0 |<| E_0 |)$, локально-оптимальные покрытия которого такие же, как и на E_0.

Очевидно, что объекты линейных оболочек [5] также образуют одно из покрытий (не обязательно локально-оптимальное), которое можно считать базовым, неизменным и фиксированным для конкретной выборки. Значения расстояний от объектов покрытия до ближайших объектов линейной оболочки из противоположных классов дают характеристику конкретного локально-оптимального покрытия, которая может применяться для сравнительного анализа. Число различных локально-оптимальных покрытий фиксированной выборки можно рассматривать в качестве меры ее дихотомизационной мощности.

Результаты экспериментов по использованию различных схем подачи объектов — кандидатов на удаление — позволяют по-другому рассматривать вопрос разделения выборки на обучающую и контрольную. Действительно, выбрав (осознанно или чисто случайно) объекты локально-оптимального покрытия в качестве обучающей выборки, а остальные — в качестве контрольной, в результате получим абсолютную точность распознавания.

Поэтому имеет смысл точность распознавания на данных, которые не предъявлялись на обучении, определять в форме теоретической оценки, например, как математическое ожидание вероятности ошибки на контроле. При использовании традиционной технологии разбиения на обучающую и контрольную выборку следует особо подчеркивать, что обучающая выборка не является локально-оптимальным покрытием.

Вопросы сравнения двух и более покрытий, введения отношений на множестве различных покрытий и изучения их свойств — предмет дальнейших исследований.

В заключение приведем результаты вычисления синаптических весов номинальных признаков, получаемых при диагностике аномальной сетевой активности [6] по следующим показателям:

1) протокол, связанный с событием (TCP = 0, UPD = 1, ICMP = 2 и Unknown = 3);

2) IP — адрес источника;

3) номер порта источника;

4) IP — адрес получателя;

5) номер порта получателя;
6) тип ICMP пакета;
7) кодовое поле (или поле кода) из ICMP пакета;
8) длина данных в пакете.

Обучение сети проводилось на 116 746 событиях, каждое из которых описывалось восемью приведенными выше показателями. 74 070 события определяли нормальную сетевую активность (класс 1), 42 676 — аномальную (класс 2). Результаты вычисления значений синаптических весов номинальных признаков приведены ниже.

Номер признака	Вес	Внутриклассовое сходство	Межклассовое различие
1	13454.01	0.3793	0.0473
3	514.51	0.0049	0.1392
5	480.63	0.0048	0.1341
6	2778.59	0.4072	0.0091
7	2637.81	0.4079	0.0086

Большое разнообразие номеров портов источника и получателя, используемых в процессе передачи сообщений в сети, служит объяснением малых значений величин внутриклассового сходства и синаптических весов для номинальных признаков с номерами 3 и 5.

Список литературы

[1] Игнатьев Н.А. Выбор минимальной конфигурации нейронных сетей // Вычисл. технологии. 2001. Т. 6, №1. С. 23–28.

[2] Игнатьев Н.А. К вопросу построения эффективных нейронных сетей для данных, описываемых разнотипными признаками // Вычисл. технологии. 2001. Т. 6, №5. С. 34–38.

[3] Беркульцев М.Б., Дьячук А.К., Оркин С.Д. Применение генетического алгоритма к построению минимально допустимой обучающей выборки для нейросетевой системы принятия решений // Изв. РАН. Теория и системы управления. 1999. №5. С. 172–176.

[4] Васильев С.Н. От классических задач регулирования к интеллектуальному управлению // Изв. РАН. Теория и системы управления. 2001. Т. 1, №1. С. 5–22.

[5] Игнатьев Н.А. Распознающие системы на базе метода линейных оболочек // Автоматика и телемеханика. 2000. №3. С. 168–172.

[6] Cannady J. Artifical Neural Network for Misuse Detection. School of Computer and Information Sciences. Nova Southeastern Univ., 1998.

Поступила в редакцию 25 июля 2002 г.,
в переработанном виде — 21 ноября 2002 г.