ПРИБЛИЖЕННЫЙ МЕТОД РЕШЕНИЯ ЗАДАЧИ ОБОБЩЕНИЯ МНОЖЕСТВА НЕЧЕТКИХ ПРОДУКЦИЙ

Шатохина Н.К., Шатохин П.А.
Донецкий национальный технический университет,
факультеты ВТ, КИТА
E-mail: palsan@kita.dgtu.donetsk.ua

Abstract. ***Shatokhina N.K., Shatokhin P.A. A problem of generalization of given initial set of experts' reasonings rules is considered.*** *The rules are presented as diffuse productions. The generalization purposed is a such unredundant productions set, which is less by the capacity but allows to obtain all reasonings rules from the initial set of them. In this paper a search of several generalizations by approximate method with the use of genetic algorithms ideas is considered.*

Введение. *С* точки зрения практических применений наполнение базы знаний (БЗ) корректной информацией является сложным и важным этапом проектирования экспертной системы (ЭС). По содержанию БЗ состоит из правил рассуждений экспертов (фактов), общее число которых может быть достаточно велико. Кроме того, некоторые факты могут представлять собой частные факты. Поэтому возникают задачи обобщения знаний, т.е. задачи получения более общих фактов по нескольким частным с условием сохранения знаний исходного множества фактов. В силу природы данной задачи получение точного решения является трудоемкой процедурой. В данной работе рассмотрена задача получения некоторой совокупности приближенных решений, каждое из которых обобщает исходное множество фактов.

Основные понятия. Рассматривается множество элементарных фактов некоторой предметной области и числовое множество некоторых качественных оценок, описывающих коэффициенты уверенности выполнения фактов. Обозначим через $U = \{u_1, u_2, \ldots, u_n\}$ множество фактов, а через K — множество качественных оценок. Продукцией будем называть выражение $(W_i \to y, k_i)$, которое описывает высказывание: "если справедливы факты из множества W_i, то с уверенностью k_i можно считать, что выполняется целевой факт y", где $W_i \subseteq U$, $y \in U$ и $k_i \in K$.

Конечное множество нечетких продукций такого вида далее будем считать БЗ некоторой предметной области.

В качестве правил вывода новых продукций по имеющимся будем использовать правила, предложенные в [1], которые в сокращенном виде таковы:

$(W \to y, k) \to (Q \to y, h)$ тогда и только тогда, когда $h \le k$ *и* $Q \supseteq W$.

Данное правило позволяет работать с каждым целевым фактом независимо от других. Все построения и выводы, сделанные относительно множества продукций одного целевого факта автоматически распространяются на множества продукций остальных целевых фактов. Поэтому без потери общности, будем рассматривать множества продукций одного целевого факта.

Исходя из вышесказанного, будем рассматривать пространство $P= 2^U \times K$ всех продукций *(W,k)* с одним и тем же целевым фактом y , где 2^U обозначает булеан *U*.

В [2,3] были исследованы предложенные правила вывода. Рассмотрено расширение "$\to$" правил вывода на множества продукций. В [2,3] исследовался вопрос: какая алгебраическая структура [4] описывается этими правилами вывода, если их применять не к отдельным продукциям, а к множествам продукций.

Пусть $H,G \in 2^P$, тогда $H \to G \Leftrightarrow \forall_{e \in H} \exists_{c \in G} [c \to e]$. Множество *H* будем называть индуктивным обобщением (ИО) *G*. Показано, что расширение правил вывода на множество продукций *P* задает предпорядок $\angle$ на 2^P. Следует заметить, что если на некотором множестве определен предпорядок, то для него не возможно определить единственный наименьший элемент. Но предпорядок позволяет получить некоторое множество минимальных элементов, которые для данной задачи и основой для построения ИО. Показано, что для рассмотренного предпорядка $\angle$ минимальным элементом является множество продукций, являющихся ИО *P,* вида: $\{(u_1,h), (u_2,h), \dots,(u_n,h)\}$, где *h* — наибольшее значение коэффициентов уверенности из *K,* а $u_1, u_2, \dots, u_n \in U$. Значит, вместо любого исходного множества фактов Е в качестве БЗ следует использовать множества фактов *H,* состоящие из минимальных элементов каждой продукции.

Следует заметить, что для исходного множества продукций Е таких множеств может быть достаточно много. Интерес представляют только те *H*, которые согласно [1] являются неизбыточными и характеристическими ИО (ХИО).

Неизбыточность множества H означает, что при удалении любой продукции из него, H перестает быть ИО E: если $S \to H$ влечет $H \subseteq S$.

Множество H является характеристическим, если из него невозможно вывести противоречивые относительно E продукции: если $(W,k) \in E$, то для всех $h>k$ в H не существует (Q,h) таких, что $(Q,h) \to (W,k)$.

В [2,3] была приведена формальная постановка задачи и алгоритм ее решения, а также сделано предположение о принципиальной трудности решаемых задач. Последнее означает, что для реальных ЭС перебор всех вариантов ИО из-за большого количества продукций практически невозможен, поэтому представляет интерес создание подхода, позволяющего эффективно находить если не точные ХИО, то хотя бы некоторые "лучшие". В данной работе предлагается подход, основанный на применении генетических алгоритмов для решения данной задачи. Рассматривается задача построения совокупности индуктивных обобщений для заданного исходного множества продукций, состоящей из r решений. Поиск нескольких решений целесообразен для создания некоторого резервного количества "лучших" решений, каждое из которых может быть выбрано с учетом дополнительных требований. Значение r является одним из параметров в данном алгоритме. В частном случае r может быть равным 1.

Метод решения. Сформулируем рассматриваемую задачу. Пусть имеется множество продукций $E=\{(W,k)\}$, $E \subseteq 2^U \times K$, $U=\{u_1, u_2, \ldots, u_n\}$ — множество элементарных фактов, называемых далее буквами. Необходимо построить множество из r "лучших" ХИО E.

Согласно генетическим алгоритмам [5] поиск решения ведется по аналогии с принципами естественного отбора и генетики. Основной идеей их является принцип выживания наиболее перспективных особей — вариантов решения, который реализуется удалением особей, имеющих худшие показатели относительно сформулированной фитнес-функции. Для порождения новых особей моделируются процессы наследования и мутации, непременно включающие в себя элемент случайности подобно тому, как это имеет место в живой природе.

Для представления некоторого варианта решения (далее особи) в генетических алгоритмах используются коды, которые могут представлять собой строку элементов, например, строку бит, называемую в дальнейшем геном. При работе алгоритма интерпретация гена не учитывается. Она рассматривается только в начале и в конце работы алгоритма.

В алгоритме рассматривается несколько особей одновременно (так называемая популяция) из пространства поиска решений. Размер q популяции является одним из параметров алгоритма и задается заранее. Исходная популяция может быть создана либо некоторым оптимизационным алгоритмом, либо на основе использования случайного закона.

Построим пространство поиска решений, которое будет использоваться для порождения первых r популяций (по числу решений).

Будем называть букву u элитной, если она включена лишь в одну продукцию (W,k) исходного множества E. Заметим, что она является единственным минимальным элементом (u,k) продукции (W,k), а значит, будет включена в любое ХИО E. Следовательно в дальнейшем все элитные буквы будут включаться в любое решение.

Разобьем множество E на m подмножеств по числу различных значений $k\in K$. Через $E(k)$ будем обозначать множество всех таких $W\subseteq U$, что $(W,k)\subseteq E$, и будем называть $E(k)$ k-ым слоем E. Обозначим через n_k мощность слоя $E(k)$. Поставим в соответствие слою $E(k)$ кортеж $T(k)=(t_{k1}, t_{k2},...,t_{kn})$, в котором $t_{kj}=1$, если буква u_j встречается в продукциях слоя k, остальные значения в кортеже равны 0.

Далее для каждого слоя $E(k)$ создается множество генов $T_j(k)$, составляющих k-ую часть пространства поиска решений. Каждый ген $T_j(k)$ особи может иметь меньшее количество единиц, чем $T(k)$, которые будут находиться лишь в тех позициях, где находятся единичные значения в $T(k)$. Алгоритмы построения $T_j(k)$ первого и последующих слоев отличаются.

Для слоя $E(k_1)$, соответствующего минимальному значению из K, применяется следующий оптимизационный алгоритм.

Пусть $W_1, W_2, ... , W_r$ — набор фактов, составляющих продукции из $E(k_1)$. В множество $T_j (k_1)$ прежде всего включаются элитные буквы, которые присутствуют в продукциях в единственном экземпляре. Причем, если буква u_i является элитной, то создается ген с единственной единицей в i-ой позиции. Если же некоторый факт W содержит более одной элитной буквы, то для каждой из них создается ген с “1” в соответствующей позиции. После чего из $E(k_1)$ удаляются все продукции с элитными буквами.

На основе оставшихся после удаления фактов строится кортеж $T(k_1)$, состоящий из "1" в местах букв из фактов без элитных букв. Далее в построенное множество $T_j\,(k_1)$ добавляются гены, содержащие по одной "1" , для оставшихся фактов. Для построения этих добавляемых генов создадим таблицу, строки которой отмечены фактами $W_1, W_2, \ldots, W_r$ из E_1, а столбцы буквами из U. На пересечении строки W_i и столбца u_j стоит 1, если буква u_j входит в W_i., в противном случае стоит 0. Для каждого столбца подсчитывается количество стоящих в нем единиц. Из всех столбцов выбирается столбец p, в котором находится наибольшее число единиц, и удаляются строки с единицами в этом столбце и p-ый столбец не удаляется. Во множество генов $T_j\,(k_1)$ добавляется ген, в котором имеется единственная единица в p-ой позиции. Алгоритм представляет собой многократное повторение этих шагов. Процесс построения продолжается до тех пор, пока в таблице не остается ни одной строки.

Видно, что для первого слоя создадутся гены, содержащие минимально возможное количество "1", и количество генов не превосходит суммы количества элитных букв и количества продукций без элитных букв.

Все остальные слои вначале также просматриваются с целью определения элитных букв. Если слой $E(p)$ содержит хотя бы одну элитную букву $u_i \in W$, то для него создается ген с единственной единицей в i-ой позиции. Заметим, если некоторый факт W содержит более одной элитной буквы, то соответствующие им гены помещаются также в множество $T_j(k_p)$, для каждой элитной буквы отдельный ген..

Из слоя $E(p)$ удаляются на данном шаге факты W с элитными буквами. Далее, если слой $E(p)$ не пуст, то для него строится кортеж $T(p)$, с единицами в позициях, соответствующих буквам из оставшихся фактов слоя. По $T(p)$ строятся $q{\cdot}r{-}t$ генов, имеющих по s единиц, распределенных в позициях, соответствующих единичным значениям исходного кортежа $T(p)$. Здесь t — число элитных букв, т.е. уже построенных кортежей, а s – минимальное значение такое, что

$$\sum_{i=2}^{s} C^{i}_{|U|-t} > r\cdot(|E|-t),$$

где $|E|$ — мощность множества E. Нетрудно видеть, что эта сумма соответствует тем генам, которые с одной стороны имеют наименьшее количество единиц, а с другой стороны их достаточно, чтобы по r раз сопоставить с каждой продукцией исходного множества E. Построенные гены имеют “1” в тех же позициях, где имеются единицы в $T(k_1)$. Процесс создания генов следует организовывать в порядке возрастания количества единиц. Причем переход к генам с тремя единицами происходит только после того, как будут сгенерированы все гены с двумя единицами и т.д.

Заметим, что все построенные слои состоят из трех частей: 1-я часть содержит гены для элитных букв, содержащихся в продукциях в единственном экземпляре, 2-я часть состоит из генов элитных букв, входящих по несколько в продукцию и 3-я часть состоит из генов остальных продукций слоя.

В результате получим поисковое пространство, изображенное на рис. 1.

1-ая зона	…	r-тая зона	
Буквы из U		Буквы из U	
0 1 0 … 0		1 0 0 … 0 0	Слой $E(k_1)$
…………		…………	
0 0 0 … 1		0 0 1 … 0 1	
…………		…………	
1 1 0 … 1		0 1 0 … 1 0	Слой $E(k_m)$
…………		…………	
0 1 1 … 0		0 1 0 … 0 1	

Рисунок 1 — Представление пространства решений

На основе полученного пространства строим первоначальные популяции T_r для r множеств решений. Опишем процесс построения первоначальной популяции для первого решения, которое строится из генов первой зоны.

В слое E_1 из первой части выбираются все гены, из второй части — случайным образом один из генов для каждой продукции, и из третьей части — все гены.

В остальных слоях $E(k_i)$ из первой части выбираются также все гены, из второй части — случайным образом один из генов для каждой продукции, и из третьей части случайным образом выбираются по $a < n_i$ генов. В качестве особи первого решения будем рассматривать объединение выбранных наборов генов по всем слоям 1-ой зоны. Таким же образом строим особи остальных $r–1$ решений.

Заметим, что количество генов, составляющий одну особь, не превосходит $m=|E|$. Количество особей, составляющих размер популяции для любого из r решений T_r, будет равным $q=a(m-1)+n_1$. В результате выполнения этого этапа каждая популяция может быть представлена набором из q особей.

Аналогично строятся первоначальные популяции для последующих решений из генов соответствующих зон.

После того как создано первые популяции, включается механизм отбора, который реализован за счет применения операторов: скрещивания, мутации и уничтожения неперспективных особей.

Новые популяции в процессе работы алгоритма получаются в результате скрещивания родительских пар. Выбор особей для скрещивания осуществляется согласно целевой функции особи, которая оценивает близость особи к оптимальному решению. Поэтому, если заранее определены элитные особи, то они переходят в следующую популяцию автоматически.

Каждому гену $T_{ri}(p)$ слоя $E(p)$ ставится в соответствие оценочная функция:

$$g(T_{rj}(p)) = 1 - \frac{g_1(T_{rj})}{g_2(p)},$$

где $g_1(p)$, $g_2(p)$ — эвристические функции, определяющие:

$g_1(T_{ri})$ — количество генов из слоев с меньшими номерами, имеющие единицы в тех же позициях, что и $T_{ri}(p)$;

$g_2(p)$ — общее количество фактов из слоев с меньшими номерами, чем p.

Каждой особи H_{rj} ставится в соответствие функция

$$g(H_{ri}) = \sum_{j=1}^{|H_i|} g(T_{rj}),$$

большие значения которой соответствуют более качественным особям в смысле сохранения знаний.

На основе предложенной функции случайным образом выбираются из каждого слоя $E(p)$ по n_p продукций для каждого родителя. Элитные кортежи также участвуют в построении родительских пар.

Скрещивание родительских пар производится для каждого решения в отдельности, хотя возможен вариант использования родителей из различных зон. В процессе скрещивания особей i-го решения преобразуется каждый слой в отдельности. В каждом слое выбираются произвольно пары родителей, для которых случайным образом определяется пара точек x_1 и x_2, с помощью которых производится сечение родительских генов и выделяются фрагменты для обмена. Нетрудно заметить, что этот оператор позволяет получать в каждом слое большее и меньшее количество фактов.

Пары особей скрещиваются путем обмена между собой фрагментов генов. Точки разрыва (одна или несколько) выбираются случайным образом. Цель применения оператора — реализовать принцип выживаемости сильнейших особей.

Далее применяется оператор мутации, который с небольшой вероятностью изменяет один из битов, в котором изменяется случайно выбранный ген. При этом могут быть осуществлены либо перестановка элементов строки, либо замена значения на другое допустимое значение, например, инверсия бит, либо удаление и/или добавление гена в популяцию. Целью применения оператора является расширение области поискового пространства, не допускающее попадания в локальный экстремум целевой функции. Внесение случайных изменений может привести также к сокращению времени работы алгоритма.

Оператор мутации обеспечивает появление нового качества в особях и имеет две разновидности, выбор каждой из них происходит случайно. Возможны преобразования вида: инвертирование отдельного бита в допустимых позициях слоев любого из родителей, или удаление гена, выбранного случайно из числа, не имеющих элитных букв у любого из родителей.

После этого работает оператор уничтожения, который убирает особи с наименьшим значением целевой функции. Понятно, что в результате применения этого оператора исключаются самые слабые особи. Из числа полученных потомков удаляются те особи H, которым соответствуют минимальное значение функции $g(H)$. Если среди отобранных имеется несколько особей с минимальным значением $g(H)$, то удаляются те особи, в которые включено наибольшее число генов.

Процесс работы алгоритма представляет собой многократное применение этих операторов, которые осуществляют постепенное изменение исходной популяции в сторону ее улучшения. Завершается алгоритм при получении приемлемого качества популяции.

После завершения алгоритма из полученной популяции выбирается особь, которой соответствует экстремальное значение целевой функции.

Аналогично алгоритм повторяется многократно для каждого решения, постепенно улучшая популяцию каждого решения.

Заключение. В данной работе предложен подход к решению задачи обобщения заданного множества нечетких продукций. Подход позволяет получить не точное решение (множество всех решений), а набор “лучших” решений. Использование генетических алгоритмов позволяет уйти от перебора при решении рассмотренной задачи. Возможным дальнейшим развитием этого подхода, по мнению авторов, является использование набора решений в задаче “обучения” ЭС.

Литература

1. Горчинская О. Ю., Рубашкин В.А. Метод индуктивного построения базы знаний для экспертных систем, моделирующих нечеткие рассуждения. Построения БЗ с нечетким механизмом вывода // Автоматика и телемеханика. — 1991. — №3. — С.113–120.
2. Шатохина Н.К., Шатохин П.А. Об индуктивном построении базы знаний экспертных систем // Наукові праці Донецького державного технічного університету. Серія: Обчислювальна техніка та автоматика, випуск 12: Донецьк: ДонДТУ, ТОВ "Лебідь", 1999. — С.158–164.
3. Грунский И.С., Шатохина Н.К. Об индуктивном обобщении нечетких заключений. Серія: Обчислювальна техніка та автоматика, випуск 25: Донецьк: ДонДТУ, ТОВ "Лебідь", 2001. — С.154–160.
4. Общая алгебра т.1: Справочник/ по ред. Л.А. Скорылкова — М.:Наука,1990. — 592с.
5. Корнеев В.В., Гареев А.Ф., Васютин С.В., Райх В.В. Базы данных. Интелектуальная обработка информации. — М.:” Нолидж ”, 2000. — 352с.

Сдано в редакцию: .03.2003г.

Рекомендовано к печати: д.т.н., проф. Скобцов Ю.А.