

ПОДХОД К РЕШЕНИЮ ЗАДАЧИ ОБОБЩЕНИЯ МНОЖЕСТВА НЕЧЕТКИХ ПРОДУКЦИЙ С ПОМОЩЬЮ ГЕНЕТИЧЕСКИХ АЛГОРИТМОВ

Шатохина Н.К., Шатохин П.А.

Донецкий национальный технический университет,
palsan@kita.dgtu.donetsk.ua

Abstract

Shatokhina N.K., Shatokhin P.A., An approach to the solving problem of generalization of diffuse productions set by means of genetic algorithms. A way of solving generalization problem for the diffuse productions set is proposed. The way is based on the ideas of evolution simulation by means of genetic algorithms.

Введение

При построении экспертной системы (ЭС) сложным и важным с точки зрения практических применений является процесс создания базы знаний (БЗ). В процессе создания базы возникает множество правил рассуждений (фактов) экспертов. Некоторые из них могут представлять собой частные случайные факты, которые следует объединить в один общий факт, т.е. индуктивно обобщить. Рассмотрена задача обобщения - построения наиболее общих фактов по нескольким примерам частных фактов, позволяющая сохранить знания заданного множества E , представленного в виде совокупности нечетких продукций.

Основные понятия

Пусть задано множество $U = \{u_1, u_2, \dots, u_n\}$ элементарных фактов некоторой предметной области (ПО) и числовое множество некоторых качественных оценок K , описывающих коэффициенты уверенности выполнения фактов. Продукцией будем называть выражение $(W_i \rightarrow y, k_i)$, которое можно интерпретировать как "если справедливы факты из множества W_i , то с уверенностью k_i можно считать, что выполняется целевой факт y ". Здесь $W_i \subseteq U$, $y \in U$ и $k_i \in K$. Под БЗ некоторой ПО далее будем понимать некоторое множество нечетких продукций такого вида.

В качестве правил вывода новых продукций используются правила, предложенные в [1], которые более кратко могут быть записаны следующим образом:

$(W \rightarrow y, k) \rightarrow (Q \rightarrow y, h)$ тогда и только тогда, когда $h \leq k$ и $Q \supseteq W$.

Заметим, поскольку приведенное правило сформулировано относительно одного и того же целевого факта, то, без потери общности, будем рассматривать множества продукций одного целевого факта. Это значит, что 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 .

Исследование предложенных правил вывода дало следующие результаты.

В [2,3] рассмотрено расширение \rightarrow правил вывода на множества продукций. Пусть $H, G \in 2^P$, тогда $H \rightarrow G \Leftrightarrow \forall c \in H \exists e \in G [c \rightarrow e]$. 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 на нем хотя и не существует единственного наименьшего элемента, но предпорядок позволяет получить некоторое множество минимальных элементов. Показано, что в данном случае минимальным элементом является множество продукций, являющихся ИО $P - \{(u_1, h), (u_2, h), \dots, (u_n, h)\}$, где h – наибольшее значение из K , а $u_1, u_2, \dots, u_n \in U$. Значит, вместо любого исходного множества фактов E в качестве БЗ следует использовать множества фактов H , состоящие из минимальных элементов каждой продукции.

Следует заметить, что для любого E таких множеств может быть достаточно много. Интерес представляют 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 $S \rightarrow H$ влечет $H \subseteq S$.

Свойство H быть характеристическими для E означает невозможность вывода противоречивых относительно E продукции: если $(W, k) \in E$, то для всех $h > k$ в H не существует (Q, h) таких, что $(Q, h) \rightarrow (W, k)$.

В [2,3] приведены формальные постановки ряда задач обобщения E , предложены алгоритмы построения ХИО для них, и сделано предположение о принципиальной трудности решаемых задач.

Ясно, что для реальных ЭС перебор всех вариантов ИО просто невозможен, поэтому представляет интерес создание подхода, позволяющего эффективно находить если не точное ХИО, то хотя бы “лучшее”. В данной работе предлагается подход, основанный на применении генетических алгоритмов.

Основные положения

Генетические алгоритмы представляют собой алгоритмы поиска, построенные по аналогии с принципами естественного отбора и генетики. Основной парадигмой их является принцип выживания наиболее перспективных особей – вариантов решения и порождение особей, моделирующее процессы наследования и мутации, в которых непременно присутствует элемент случайности.

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

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

Алгоритм использует три оператора: 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. Пары особей скрещиваются путем обмена между собой фрагментов строк. 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 генов, в котором изменяется случайно выбранный элемент. 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, чтобы не допустить попадания в локальный экстремум целевой функции. Внесение случайных изменений может привести также к сокращению времени работы алгоритма.

После этого работает оператор уничтожения, который убирает особи с наименьшим значением целевой функции. Понятно, что в результате применения этого оператора исключаются самые слабые особи.

Процесс работы алгоритма представляет собой многократное применение этих операторов, которые осуществляют постепенное изменение исходной популяции в сторону ее улучшения, что и является целью генетического алгоритма. Завершается алгоритм или по достижении заданного числа итераций, или при получении приемлемого качества популяции.

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

Метод решения

Сформулируем рассматриваемую задачу. Пусть имеется множество продукций $E = \{(W, k)\}$, $E \subseteq 2^U \times R$, $U = \{u_1, u_2, \dots, u_n\}$ - множество элементарных фактов, называемых далее буквами. Необходимо построить одно "лучшее" ХИО E .

Опишем алгоритм построения пространства поиска решений, необходимого для порождения первой популяции.

Будем называть букву u элитной, если она включена лишь в одну продукцию (W, k) исходного множества E . Действительно, она является единственным минимальным элементом (u, k) продукции (W, k) , а значит, будет включена в любое ХИО E .

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 $(W, k) \in E$, и будем $E(k)$ называть слоем k . Обозначим через n_k мощность слоя $E(k)$. Поставим в соответствие слою $E(k)$ кортеж $T(k) = (t_{k1}, t_{k2}, \dots, t_{kn})$, в котором $t_{kj} = 1$, если буква u_j встречается в продукциях слоя k , остальные значения кортежа - 0.

Для каждого слоя $E(k)$ создается множество генов $T_j(k)$, составляющих часть особи. Каждый ген $T_j(k)$ особи может иметь меньшее количество единиц, которые будут находиться лишь в тех позициях, где находятся единичные значения в кортеже $T(k)$.

Для слоя $E(k_1)$, соответствующего минимальному значению из K , применяется следующий оптимизационный алгоритм. Пусть W_1, W_2, \dots, W_r - набор фактов, составляющих продукции из $E(k_1)$. Построим таблицу, строки которой отмечены фактами W_i ,

W_2, \dots, W_r , 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 сам p -ый столбец. Во множество генов $E(k_1)$ 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 не остается ни одной строки. Добавим построенное множество $T(k_1)$ до $\max(n_p)$ двухбитовыми генами. Двухбитовые гены имеют единицы в единичных позициях $T(k_1)$, выбранных случайным образом.

Видно, что для первого слоя создадутся факты, содержащие минимально возможное количество 1.

Все остальные слои вначале 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 выбирается случайным образом одна из них. 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 $T(p)$, в с единицами в позициях, соответствующих буквам u_i из фактов слоя $E(p)$ за исключением элитных. По $T(p)$ строятся $n_p - t$ генов по s единиц, распределенных случайным образом в позициях, соответствующих единичным значениям исходного кортежа $T(p)$ за исключением мест элитных букв. Здесь t – число элитных букв, т.е. уже построенных кортежей, а s – минимальное значение такое, что

$$\sum_{i=1}^s C_{|U|}^i > |E|,$$

где $|E|$ - мощность множества E . Нетрудно видеть, что эта сумма представляет собой те продукции булеана 2^U , которые с одной стороны имеют наименьшее количество букв, а с другой стороны их достаточно, чтобы сопоставить с каждой продукцией исходного множества E .

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

		Буквы из U				
Зона слоя $E(k_1)$		0	1	0	...	0
					
		0	0	0	...	0
Зона слоя $E(k_m)$					
		1	1	0	...	1
					
		0	1	1	...	0

Рис. 1 - Представление пространства решений.

На основе полученного пространства строим первоначальную популяцию. Из зоны генов $E(k_i)$ выбирается случайном образом n_i кортежей. В качестве особи будем рассматривать объединение выбранных наборов генов по всем слоям. Таким же образом строим остальные особи. Заметим, что количество генов, составляющий одну особь, не

превосходит $m = |E|$. Количество особей, составляющих размер популяции, будет равным $r = \max(n_k) + 1$, по всем $n_k = |E_k|$. Представление популяции изображено на рис. 2.

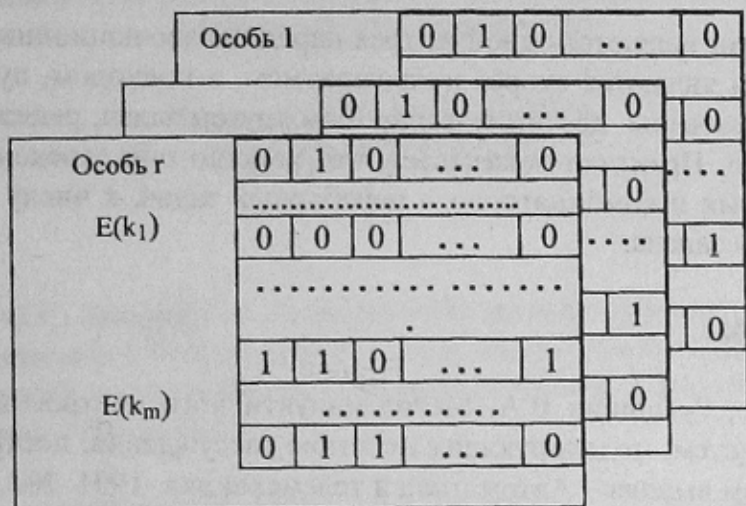


Рис. 2 - Представление популяции.

Каждому гену $T_i(p)$ слоя $E(p)$ ставится в соответствие оценочная функция: $f(T_i(p)) = 1 - f_1(T_i)/f_2(p)$, где $f_1(p)$, $f_2(p)$ - эвристические функции, определяющие:

$f_1(T_i)$ - количество генов из слоев с меньшими номерами, имеющие единицы в тех же позициях, что и $T_i(p)$;

$f_2(p)$ - общее количество фактов из слоев с меньшими номерами, чем p .

Каждой особи H_j ставится в соответствие функция

$$f(H_j) = \sum_{j=1}^{|H_j|} f(T_j),$$

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 продуктов для каждого родителя. Элитные кортежи также участвуют в построении родительских пар.

При скрещивании родительских пар преобразуется каждый слой в отдельности. Для каждого слоя первого и второго родителя в выбирается случайным образом пара точки сечения x_1 и x_2 , с помощью которых выделяются фрагменты для обмена. Нетрудно заметить, что этот оператор позволяет получать в каждом слое большее и меньшее количество фактов.

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.

Далее из числа полученных потомков удаляются те особи H , которым соответствуют минимальное значение функции $f(H)$. Если с минимальным значением $f(H)$ имеется несколько особей, то удаляются особи максимальной мощности.

Алгоритм повторяется многократно, постепенно улучшая популяцию.

Заключение

Генетические алгоритмы являются одной из трех парадигм эволюционных вычислений. В связи с тем, что они являются скорее не алгоритмом, а подходом, путем адаптации они могут быть использованы достаточно широким кругом задач, решение которых на ЭВМ было затруднено. Практика показывает, что хорошо они зарекомендовали себя при решении численных и комбинаторно – переборных задач, к числу которых относится рассматриваемая задача.

Список литературы

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.

Поступила в редакційну колегію 1.02.2002 р.