Л.О. Шашкин

Применение методов эволюционного моделирования для оптимизации множества ДСМ-гипотез

Аннотация. Анализируется применимость генетических алгоритмов для работы с множеством гипотез, порожденных ДСМ-методом. Решается задача поиска небольшого количества причин исследуемого свойства методами эволюционного моделирования.

Ключевые слова: ДСМ -метод, генетические алгоритмы.

Введение

Необходимость выявления причинно-следственных связей между структурой объекта исследования и его свойствами возникает в самых разных областях, в том числе фармакологии, медицине. В качестве объектов здесь могут рассматриваться химические формулы соединений, множество сведений о пациенте, а в качестве свойств - биологическая активность вещества либо, соответственно, результат лечения.

Один из способов установления связей такого рода предоставляет ДСМ-метод автоматического порождения гипотез [1]. Гипотезы о наличии или отсутствии определенного свойства строятся на основе анализа групп примеров. В качестве возможной причины рассматриваются общие элементы, присутствующие в структуре исследуемых объектов. Таким образом, и объектам, и гипотезам соответствуют некоторые совокупности элементов структуры.

ДСМ-система, действуя формально, порождает все возможные гипотезы, которые могут объяснять наличие либо отсутствие у объектов определенного свойства. Количество полученных гипотез зависит от конкретной ситуации (количества и вида примеров, а также настроек алгоритма ДСМ-метода) и в некоторых случаях оказывается чрезмерно большим. В дальнейшем может потребоваться выбор некоторого подмножества наиболее полезных гипотез, который происходит при участии эксперта.

Возможно ли с помощью некоторой дополнительной процедуры облегчить работу эксперта, уменьшив множество рассматриваемых им гипотез за счет отсеивания посторонних? Роль такой процедуры должна заключаться в повышении доли «интересных» гипотез по сравнению с исходным соотношением.

Далее мы рассмотрим возможность решения этой задачи с использованием эволюционных методов поиска.

1. Кодирование элементов структуры

Изучая различные методы преобразования множества гипотез, построенного ДСМ-системой, полезно выбрать конкретную задачу, чтобы иметь возможность оценить результаты работы рассматриваемых алгоритмов. В качестве модельного примера рассмотрим гипотезы, построенные на данных из области фармакологии.

Структура химического соединения может быть описана на языке ФКСП (фрагментарный код суперпозиции подструктур) [2]. В этом случае соединение кодируется набором дескрипторов, характеризующих тип групп атомов и их взаимное расположение. Часть используемых в качестве примеров соединений оказывает одинаковое специфическое действие, в то вре-

мя как у остальных это свойство отсутствует, что позволяет разделить примеры на положительные и отрицательные. На активность соединения может влиять присутствие в молекуле определенных фрагментов, поскольку именно они участвуют в химическом взаимодействии. Так как общие для некоторой группы объектов дескрипторы рассматриваются ДСМ-системой в качестве возможной причины наличия либо отсутствия у соответствующей группы соединений заданного свойства, можно надеяться, что некоторые из найденных гипотез соответствуют реальным причинам биологической активности. При этом можно ожидать, что количество таких причин будет небольшим.

2. Кодирование объектов и гипотез

При описании структуры химического соединения один дескриптор может упоминаться несколько раз, поэтому набор дескрипторов, кодирующих объект, нельзя считать множеством в обычном смысле. Это не позволяет использовать, в частности, теоретико-множественную операцию пересечения для поиска общих элементов.

Зафиксируем и упорядочим полный список дескрипторов, применявшихся для кодирования всех рассмотренных в задаче примеров. После этого объекты (химические соединения) и гипотезы возможно записывать при помощи кортежей одинаковой длины, равной числу элементов списка всех дескрипторов. *k*-й элемент кортежа равен количеству вхождений *k*-го дескриптора из упорядоченного списка в ФКСП-код примера или гипотезы (если дескриптор не используется, значение элемента равно нулю). Например, ФКСП-код одной из гипотез содержит дескрипторы {6,06; 3101310; 3101320; 3101320; 3201320; 5,06M1M2; 3100331}. Учитывая их расположение в полном списке, содержащем 195 дескрипторов, получаем кортеж {0, 1, 1, 2, 1, 1, 0, 1, 0, 0,…, 0}, где многоточие заменяет нулевые элементы.

Такой способ кодирования обеспечивает удобство работы с данными при программной реализации алгоритма. В частности, один набор дескрипторов содержится в другом тогда и только тогда, когда каждый элемент первого кортежа не превосходит соответствующий элемент второго. Для получения пересечения необходимо из двух кортежей составить новый, сравнивая пары элементов, соответствующих одному дескриптору, и выбирая наименьший элемент.

3. Простой генетический алгоритм

Уточним постановку задачи. Имеется большое множество гипотез, порожденных ДСМ-системой при анализе обучающих примеров. Желательно выделить несколько гипотез, которые объясняли бы те же примеры, что и исходное множество гипотез. Считаем, что гипотеза объясняет пример, если все соответствующие ей дескрипторы содержатся (с учетом количества вхождений) среди дескрипторов, кодирующих пример. Каждой гипотезе при этом сопоставляется некоторое множество примеров, которые она объясняет, а множеству гипотез – объединение множеств примеров. В каждом случае будем рассматривать одно конкретное свойство и гипотезы (и примеры) определенного типа (положительные либо отрицательные).

Цели поиска могут различаться. Можно либо искать наименьшее множество, либо ограничиться требованием некоторого уменьшения мощности множества гипотез, с учетом дальнейшей работы с ним специалиста, хорошо знающего предметную область. В обоих случаях возможно и допустимо предъявление нескольких различных решений, чтобы не ограничивать эксперта единственным вариантом.

Для поиска оптимального решения может применяться направленный, случайный или комбинированный перебор. Рассмотрим особенности задачи, которые позволяют выбрать в качестве метода решения простой генетический алгоритм [2]. Такие алгоритмы при подходящей формализации задачи позволяют сочетать преимущества случайного и целенаправленного поиска.

В начальный момент должно быть задано некоторое множество потенциальных решений – популяция хромосом. Начальная популяция обычно выбирается случайно. Механизм работы с двоичными строками является универсальным. Решения, принадлежащие новому поколению, могут быть получены одним из следующих способов:

- мутация – изменяется случайный двоичный символ в некоторой хромосоме;
- кроссинговер – в двух хромосомах начальные фрагменты одинаковой длины меняются местами;

- репродукция – хромосома копируется без изменений;

- процедура селекции, учитывающая приспособленность хромосом (значение целевой функции), позволяет исключить из популяции лишние решения.

Перечисленные процедуры называются генетическими операторами. Их циклическое применение позволяет создавать популяции хромосом, относящиеся к следующим поколениям, в которых должны накапливаться положительные изменения.

В рассматриваемом примере поиск решения происходит на конечном множестве всех подмножеств множества ДСМ-гипотез. Подмножества могут быть закодированы обычным способом в виде двоичных строк, которые и будем называть хромосомами.

Допустимым решением будем считать любое подмножество гипотез, объясняющее все примеры, сформулированные исходным множеством гипотез.

Поскольку ставится задача уменьшить количество гипотез, в качестве целевой функции (функции приспособленности) выберем мощность множества гипотез, которая вычисляется как число единиц в хромосоме (двоичном коде множества).

Любое движение в правильном направлении облегчит дальнейшую работу эксперта, поэтому допустима остановка алгоритма по числу итераций (по времени). Это избавляет от необходимости проверять факт получения наилучшего решения, для чего пришлось бы оценивать наименьшую мощность множества гипотез, объясняющих необходимое количество примеров.

Применение генетических операторов к хромосомам, являющихся кодами множеств, сводится, по сути, к удалению и добавлению элементов этих множеств. Такие изменения множества гипотез могут проводиться поэтапно, с использованием достигнутого результата, что и способен обеспечить генетический алгоритм.

4. Анализ работы стандартных операторов

Существует, однако, проблема появления недопустимых решений. Например, применение оператора мутации, изменяющего одно из двоичных значений в хромосоме, должно привести к добавлению или удалению некоторой гипотезы. В последнем случае примеры, прежде объяснявшиеся удаленной гипотезой, должны быть сформулированы оставшимися. Такая ситуация возможна на начальном этапе работы алгоритма, но для множеств, содержащих небольшое число гипотез, она является маловероятной. Те же соображения относятся к кроссинговеру, который позволяет хромосомам обмениваться своими фрагментами. Эта операция должна обеспечить соединение частей наиболее удачных решений для создания новых с наилучшими значениями целевой функции. Однако получение решений, близких к оптимальным, означает максимальное устранение дублирования со стороны гипотез при объяснении примеров. Множества примеров, соответствующие различным гипотезам, не совпадают, замена некоторой части гипотез другими почти наверняка приведет к увеличению количества необъясненных примеров.

Решить проблему допустимости можно в духе метода штрафных функций. Объявив допустимыми любые подмножества гипотез, соответственно изменим целевую функцию. При ее вычислении будем учитывать не только мощность рассматриваемого множества гипотез, но и мощность множества объясненных ими примеров. Приспособленность хромосом при этом окажется тем выше, чем больше примеров будет объяснено и чем меньше гипотез для этого потребуется.

Тем не менее, такой подход не устраняет проблему целиком, она превращается в проблему эффективности генетических операторов. Целью работы алгоритма является получение подмножеств небольшой мощности. Задача может иметь несколько решений, соединение их частей приведет либо к излишнему увеличению числа гипотез, либо к уменьшению числа объясненных примеров. По этой причине потомки хромосом с высокой приспособленностью окажутся хуже своих родителей. Другая причина снижения эффективности заключается в том, что близкие к оптимальным хромосомы содержат мало ненулевых элементов. Мутация в этом случае будет чаще увеличивать размер множества гипотез, заменяя ноль единицей в его коде. Кроссинговер же часто будет переставлять нулевые фрагменты, не порождая новых решений.

Такая отрицательная обратная связь затрудняет целенаправленный поиск решения. Неэффективность кроссинговера на хромосомах с высокими значениями целевой функции ставит под сомнение способность генетического алгоритма найти оптимальное решение задачи.

5. Модификация алгоритма поиска

Попробуем модифицировать алгоритм, сохранив идеи эволюционной модели поиска решений. Выявленные проблемы показывают, что следует отказаться от использования стандартных генетических операторов. Методы преобразования множества гипотез должны учитывать структуру и свойства самих гипотез, а не рассматривать их как однородные элементы.

Стремление минимизировать количество гипотез означает, что алгоритм должен выбирать те из них, которые дают объяснение как можно большему числу примеров. Поскольку для объяснения некоторого множества примеров все элементы структуры гипотезы должны содержаться в каждом из объектов, преимущества имеют «короткие» гипотезы, содержащие небольшое число элементов.

Рассмотрим ситуации, в которых удаление гипотезы из множества может происходить с наименьшими потерями из-за имеющегося «дублирования информации».

Наиболее простым является случай, когда дескрипторы (элементы структуры) одной гипотезы целиком содержатся среди дескрипторов другой. Более длинной гипотезе соответствуют только те примеры, которые объясняет также и короткая. В этом смысле длинная гипотеза оказывается бесполезной и может быть удалена.

Если одна гипотеза «почти содержится» в другой, после удаления более длинной могут появиться необъясненные примеры. Однако следует учитывать присутствие в множестве, являющимся потенциальным решением, и других гипотез, которыми эти примеры могут объясняться. Наконец, наличие значительного пересечения у двух гипотез также позволяет рассмотреть возможность удаления одной из них.

Построим эволюционный алгоритма поиска, который использует перечисленные соображения.

Изучаемые объекты и гипотезы о причинах их свойств кодируются кортежами одинаковой длины, как было предложено выше.

Для действий с решениями предполагается использовать только стандартные операции над множествами, поэтому нет необходимости кодировать множества гипотез каким либо специальным образом. Эти множества в дальнейшем будем также называть хромосомами.

Работа алгоритма начинается с создания начальной популяции. Следует задать ее размер, а также начальный размер хромосом. В качестве хромосомы можно взять как все исходное множество гипотез, порожденных ДСМ-методом, так и его подмножество, элементы которого выбираются случайно. В процессе работы алгоритма размер хромосом будет изменяться как в меньшую, так и в большую сторону.

Модификацию хромосомы реализуем следующим образом. Выбрав две случайные гипотезы, найдем их пересечение. После этого более длинная гипотеза может быть удалена из множества с вероятностью, которая зависит от отношения количества дескрипторов, содержащихся в пересечении и в более короткой гипотезе. Если одна из гипотез целиком содержится в другой, отношение равно единице, если пересечение пусто – нулю. Такой подход позволяет единым образом учитывать содержательные причины удаления длинных гипотез, рассмотренные выше. Интенсивность удаления гипотез (количество рассмотренных пар) является параметром процедуры, которая применяется к каждому из потенциальных решений. Вероятностный характер этого оператора означает возможность получения различного результата даже в случае одинаковых исходных множеств. Заметим, что данный оператор, в отличие от обычного оператора мутации, предназначен для целенаправленного улучшения популяции решений.

Чтобы избежать резкого уменьшения количества объясненных примеров и обеспечить возможность возвращения «в работу» удаленных ранее гипотез, к популяции добавляется некоторое количество хромосом, полученных как результат объединения случайно выбранных решений. Это также позволяет «полезным» гипотезам перемещаться в другие решения. Такой случайный «кроссинговер» необходим для компенсации «целенаправленной мутации».

Поскольку «мутация» обрабатывает каждое множество гипотез, найденные ранее удачные решения желательно заранее скопировать, что-

бы гарантировано использовать в следующем поколении.

Для определения приспособленности хромосомы сортируются по убыванию числа объясненных примеров, а при одинаковых значениях этой характеристики – по возрастанию мощности множества гипотез. Селекция отбирает начало этого списка, возвращая популяцию хромосом к исходному размеру. Такой подход к оцениванию решений позволяет работать с любыми подмножествами множества гипотез как с допустимыми.

Остановка алгоритма выполняется после смены заданного числа поколений хромосом, что не гарантирует отыскания наилучшего решения. Выбирая другой критерий остановки, следует иметь в виду, что для числа объясненных примеров, одного из двух параметров, характеризующих оптимальность решения, максимальное значение известно точно. Оно равно количеству примеров, объясненных исходным множеством ДСМ-гипотез. Достижения максимума по количеству объясненных примеров можно добиться выбором подходящих настроек алгоритма.

Заметим, что в любом случае окончательный выбор наиболее интересных гипотез должен осуществлять эксперт, хорошо знакомый с предметной областью.

Таким образом, алгоритм поиска оптимального множества ДСМ-гипотез состоит из следующих шагов.

Предварительный этап

1. Составляется упорядоченный список дескрипторов, использованных при кодировании объектов: $D = \{d_1,\ldots, d_n\}$.

2. Гипотезы, порожденные ДСМ-системой, кодируются кортежами вида $h_i = \{ h_{i1},\ldots, h_{in}\}$, где h_{ik} – количество вхождений дескриптора d_k в гипотезу h. Аналогичная процедура применяется к примерам.

Выбор параметров алгоритма поиска

mx – мощность каждого из начальных решений ($0 < mx \le mh$, где mh – количество всех гипотез, порожденных ДСМ-меодом);

nx – размер популяции хромосом (количество потенциальных решений);

pm и pc – неотрицательные параметры, задающие интенсивность процессов модификации хромосом и их объединения;

N – число шагов алгоритма (количество поколений).

Алгоритм поиска оптимального множества гипотез

Создание начальной популяции:

3. Из исходного множества выбираются mx случайных гипотез, образующих потенциальное решение. Процедура повторяется nx раз.

4. Решения сортируются по убыванию количества объясненных ими примеров.

Создание новой популяции:

5. Копируется лучшее решение (первое в списке).

6. Удаление гипотез. Выполняется для каждого решения из текущей популяции.

Выбирается случайная пара гипотез h_i и h_j, строится пересечение int (вычисляются элементы кортежа $int_k = \min(h_{ik}, h_{jk})$, $k = 1,\ldots, n$). Вычисляются суммы s_i, s_j и s_{int} элементов кортежей h_i, h_j и int соответственно. Кортеж с большей суммой удаляется с вероятностью, равной отношению s_{int} и суммы элементов другого кортежа. Процедура повторяется $pm*m$ раз, где m – количество гипотез в рассматриваемом решении.

7. Объединение решений. Выбирается случайная пара модифицированных решений M_i и M_j. Новое решение $M_i \cup M_j$ добавляется к имеющимся. Процедура повторяется $pc*nx$ раз.

8. Селекция. Рассматривается список решений, полученных при выполнении шагов 5 – 7. Решения сортируются по убыванию количества объясненных примеров, в случае равенства – по возрастанию количества гипотез.

Первые nx решений образуют новую популяцию, остальные решения удаляются.

9. Условие остановки. Если число поколений достигло N, алгоритм заканчивает работу, первое множество гипотез из списка выдается в качестве решения задачи. В противном случае переход к шагу 5.

Вероятностный характер применяемых процедур означает, что повторный запуск алгоритма может привести к появлению решения, отличного от найденного ранее. Относительно выбора параметров алгоритма, позволяющих настроить его на решаемую задачу, необходимо сделать некоторые замечания. Если исходное множество гипотез не слишком велико, начальную популяцию желательно составить из нескольких экземпляров этого множества. Число примеров, объясненных наилучшим в популяции решением, не убывает при смене поколений, поэтому найденное алгоритмом множество гипотез будет объяснять максимально возмож-

Поколение	Количество гипотез	Количество необъясненных примеров	Лучшее решение
	{26, 26, 26, 26, 26, 26}	{0, 0, 0, 0, 0, 0}	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26}
1	{10, 10, 11, 13, 13, 13}	{0, 0, 0, 0, 0, 0}	{2, 4, 6, 7, 8, 9, 11, 14, 18, 22}
2	{6, 7, 7, 9, 10, 10}	{0, 0, 0, 0, 0, 0}	{4, 7, 11, 14, 15, 19}
3	{4, 5, 6, 6, 8, 11}	{0, 0, 0, 0, 0, 0}	{4, 7, 8, 14}
4	{4, 4, 5, 5, 5, 7}	{0, 0, 0, 0, 0, 0}	{4, 7, 8, 14}
5	{3, 3, 3, 3, 3, 4}	{0, 0, 0, 0, 0, 0}	{4, 7, 9}
6	{3, 3, 3, 2, 2, 2}	{0, 0, 0, 4, 4, 4}	{4, 7, 9}

ное количество примеров. Если оптимальное решение задачи имеет небольшую мощность, выбор больших значений параметра *pm*, означающий многократное сравнение гипотез для возможного удаления одной из них, позволяет уже в первых поколениях решений переходить к рассмотрению небольших множеств гипотез.

Выбор в качестве начальных решений подмножеств небольшой мощности ускоряет работу алгоритма. При этом следует учесть, что одновременно уменьшается вероятность достижения алгоритмом наилучшего решения, поскольку какая-либо из входящих в его состав гипотез может не попасть при случайном отборе ни в одно из множеств начальной популяции.

Вернемся к задаче, связанной с поиском гипотез о причинах биологической активности соединений. ДСМ-метод порождает некоторое количество гипотез. Рассмотрим для определенности положительные гипотезы, количество которых равняется 26. Множество положительных примеров содержит 14 объектов, все они объясняются имеющимися гипотезами.

Для кодирования используется язык ФКСП, полный список дескрипторов содержит 195 элементов. Таким образом, гипотезы и примеры записываются в виде кортежей длины 195, для удобства идентификации они перенумерованы.

Применим описанный выше алгоритм для поиска оптимального множества гипотез. В качестве начальных решений возьмем все множество положительных ДСМ-гипотез. Алгоритм не допускает уменьшения приспособленности наилучшего в популяции решения, следовательно, количество содержащихся в нем гипотез может только уменьшиться при сохранении количества объясненных примеров. Выберем $nx = 6$ (размер популяции).

Проследим за динамикой процесса поиска. В каждом поколении хромосом будем фиксировать количество гипотез, образующих решение, количество примеров, оставшихся необъясненными и лучшее решение. Для удобства наилучшее решение будем показывать в виде списка номеров гипотез. Типичный результат применения алгоритма представлен в таблице.

Рассмотрев подробнее последнее поколение хромосом ({4, 7, 9}, {4, 7, 14}, {4, 7, 14}, {4, 7}, {4, 7}, {4, 7}), можно обнаружить не только предъявленное решение {4, 7, 9}, но и решение {4, 7, 14}, также объясняющее все множество примеров. При повторных запусках алгоритма получаются решения {4, 7, 10}, {7, 11, 14} и другие. Цель, заключающуюся в поиске нескольких небольших множеств гипотез, объясняющих все положительные примеры, следует считать достигнутой.

Заключение

Алгоритм, разработанный с применением идей эволюционного моделирования, оказался способен решить поставленную задачу оптимизации множества ДСМ-гипотез. Преимуществом эволюционных методов является поэтапное преобразование решений с отбором наиболее удачных и «обменом информацией» между ними. Однако их успешное применение потребовало разработки специальных процедур, приспособленных для решения конкретной задачи. Рассмотрение универсальных генетических операторов показало, что они не могут обеспечить «самонастройку» алгоритма при работе с множествами гипотез. Предложенные методы преобразования решений работают со структурой самих гипотез, учитывают особенности их использования в рамках ДСМ-метода.

Литература

1. Аншаков О.М. Об одной интерпретации ДСМ-метода автоматического порождения гипотез // ДСМ-метод автоматического порождения гипотез: Логические и эпистемологические основания. – М.: Книжный дом «ЛИБРОКОМ», 2009. С. 81-95.
2. Блинова В.Г., Добрынин Д.А. Языки представления химических структур в интеллектуальных системах

для конструирования лекарств // Автоматическое порождение гипотез в интеллектуальных системах. – М.: Книжный дом «ЛИБРОКОМ», 2009. С. 294-309.

3. Гладков Л.А., Курейчик В.В., Курейчик В.М. Генетические алгоритмы / Под ред. В.М. Курейчика. М.: ФИЗМАТЛИТ, 2006. – 320 с.

Шашкин Леонид Олегович. Старший преподаватель Российского государственного гуманитарного университета. Окончил Московский государственный университет им. М.В. Ломоносова в 1987 году. Имеет 17 печатных работ. Область научных интересов: искусственный интеллект, генетические алгоритмы. E-mail: shashkin@lenta.ru.