# СКС ЭВОЛЮЦИОННОГО ПРОГНОЗИРОВАНИЯ РЕВМАТОИДНОГО АРТРИТА

Балабанов В.Н., группа КСД-01а
Руководитель проф. каф. АСУ Скобцов Ю.А.

Диагностика ревматоидного артрита приобретает все большую актуализацию по мере освоения человеком окружающей среды. Ревматоидный артрит по праву считается центральной проблемой современной ревматологии, поскольку это заболевание является наиболее распространенным во всем мире и наиболее тяжелым из воспалительных заболеваний суставов. Ревматоидный артрит характеризуется существенным снижением показателей качества жизни и представляет собой чрезвычайно серьезную социально-экономическую проблему даже в таких экономически развитых странах мира, как США.

Разработка инструмента, проводящего диагностику ревматоидного артрита, который позволит по наличию различных симптомов с большой долей достоверности, своевременно и достаточно быстро определить необходимые стратегии лечения пациентов, все еще является насущной необходимостью.

Прогнозирование — одна из самых востребованных, но при этом и самых сложных, задач анализа. Существуют различные алгоритмы поиска закономерностей в существующих данных. Наряду со стандартными методами, использующими параметрические модели, в последнее время для этих целей стали применяться другие подходы, в частности, нейросетевые и эволюционные методы. Генетические алгоритмы (ГА) есть поисковые алгоритмы, основанные на механизмах натуральной селекции и натуральной генетики. Они реализуют «выживание сильнейших» среди рассмотренных структур, формируя и изменяя поисковый алгоритм на основе моделирования эволюции. Простой генетический алгоритм был впервые описан Гольдбергом на основе работ Холланда. Механизм простого ГА (ПГА) несложен. Он копирует последовательности и переставляет их части. Предварительно ГА случайно

генерирует популяцию последовательностей — стрингов (хромосом). Затем ГА применяет множество простых операций к начальной популяции и генерирует новые популяции. ПГА состоит из 3 операторов: репродукция, кроссинговер, мутация. Репродукция — процесс, в котором хромосомы копируются согласно их целевой функции (ЦФ). Копирование хромосом с «лучшим» значением ЦФ имеет большую вероятность для их попадания в следующую генерацию. Оператор репродукции (ОР), является искусственной версией натуральной селекции, "выживания сильнейших" по Дарвину. После выполнения ОР оператор кроссинговера (ОК) может выполниться в 3 шага. На первом шаге члены нового репродуцированного множества хромосом выбираются сначала. Далее каждая пара хромосом (стрингов) пересекается по следующему правилу: целая позиция $k$ вдоль стринга выбирается случайно между $l$ и длиной хромосомы меньше единицы, т.е. в интервале $(1,L-1)$. Длина $L$ хромосомы это число значащих цифр в его двоичном коде. Число $k$, выбранное случайно между первым и последним членами, называется точкой ОК или разделяющим знаком. Механизм ОР и ОК включает случайную генерацию чисел, копирование хромосом и частичный обмен информацией между хромосомами. Далее, согласно схеме классического ПГА, выполняется оператор мутации. Считают, что мутация — вторичный механизм в ГА.

Во многих проблемах имеются специальные знания, позволяющие построить апроксимационную модель. При использовании ГА это может уменьшить объем и время вычислений и упростить моделирование функций, сократить число ошибок моделирования. ГА — это мощная стратегия выхода из локальных оптимумов. Она заключается в параллельной обработке множества альтернативных решений с концентрацией поиска на наиболее перспективных из них. Причем периодически в каждой итерации можно проводить стохастические изменения в менее перспективных решениях. Временная сложность алгоритмов зависит от параметров генетического поиска и числа генераций. Системы, использующие в своем алгоритме модели эволюционных алгоритмов, в отличие от различного вида экспертных систем, не требуют вмешательства человека в процесс обучения, а

руководствуются только данными из обучающей выборки. Это позволяет исключить погрешность, вносимую экспертом. Таким образом, одной из наиболее адекватных моделей для автоматизации задачи диагностирования, является модель, основанная на представлении об эволюционных алгоритмах. Данная модель была использована при разработке автоматизированной системы диагностирования ревматоидного артрита на основании существующей статистической информации.

Использование ГА в прогнозировании — бурно развивающаяся отрасль эволюционных вычислений. Анализ данных и предикция часто могут формулироваться как проблемы поиска, например, поиска модели, представляющей данные, поиска прогностических правил, или поиска специфической структуры или сценария, хорошо предсказывающего по имеющимся данным. Основные парадигмы ГА были рассмотрены выше, однако, применительно к прогнозированию, проблема формализации задачи и предварительного проектирования стоит особенно остро, так как именно эти факторы зачастую становятся решающими и определяют степень успешности реализации всего проекта. Использование ГА в задачах предикции сопряжено с некоторыми специфическими проблемами. Например, в [1] описано два проекта, в которых генетический алгоритм используется для решения таких проблем поиска, как прогнозирование динамических систем и предсказание структуры белков.

Так, в 1990 Norman Packard разработал форму ГА, направленную на решение этой проблемы и применил свой метод к нескольким задачам анализа данных и прогнозирования. Общая задача может быть заявлена следующим образом: ряд наблюдений некоторого процесса (например, физической системы или формальной динамической системы) принимают форму набора пар,

$$\{(\bar{x}^1, y^1), \ldots, (\bar{x}^N, y^N)\}, \qquad (1)$$

где $\bar{x}^i = (x_1^i, \ldots x_N^i)$ является независимой переменной и $y^i$ является зависимой переменной. В задачах прогнозирования на фондовой бирже независимыми переменными могут быть $\bar{x} = \big(x(t_1), x(t_2), \ldots x(t_n)\big)$, представляющие значения

стоимости отдельных акций (т.н. «переменная состояния») в последовательные шаги времени, а зависимой переменной может быть величина $y = x(t_n + k)$, представляющая стоимость акции через некоторое время в будущем. (В этих примерах для каждого вектора независимых переменных $\bar{x}$ имеется только одна зависимая переменная, более общий случай задачи предполагает вектор зависимых переменных для каждого вектора независимых переменных.)

Packard использовал ГА для поиска в пространстве наборов условий независимых переменных для таких наборов условий, которые хорошо прогнозируют зависимую переменную. К примеру, в задаче прогнозирования на фондовой бирже, особь в популяции ГА может быть задана таким набором условий как:

$$\begin{aligned} C = \{&(\$20 \le \mathrm{Price\ of\ Xerox\ stock\ on\ day}\ 1) \\ &\wedge\ (\$25 \le \mathrm{Price\ of\ Xerox\ stock\ on\ day}\ 2 \le \$27) \\ &\wedge\ (\$22 \le \mathrm{Price\ of\ Xerox\ stock\ on\ day}\ 2 \le \$25)\}, \end{aligned} \qquad (2)$$

где "$\wedge$" — логический оператор "И". Эта особь представляет все наборы трех дней, в которых заданные условия встречаются (возможен пустой набор, если условия никогда не пересекутся). Набор условий *C*, таким образом, определяет специфическое подмножество точек данных (здесь, набор всех 3-дневных периодов). Цель Packard состояла в том, чтобы использовать ГА для поиска наборов условий, которые хорошо предсказывают что-то, другими словами, искать наборы условий, которые определяют подмножества точек данных, чьи значения зависимых переменных близки к тому, чтобы быть равномерными. В примере для фондовой биржи, если ГА нашел такой набор условий, для которого все дни, удовлетворяющие этому набору, следуют за днями, в которые цена акций Xerox росла приблизительно до 30$, тогда мы можем с уверенностью предсказать, что если эти условия удовлетворяют сегодняшнему дню, то цена акций Xerox повысится.

Фитнесс каждой особи *C* вычисляется пропуском всех данных $(\bar{x}y)$ в обучающем наборе через *C* и для каждого $\bar{x}$, который удовлетворяет *C*, запоминается соответствующий $y$. После того как это сделано, измерение состоит из

равномерных результирующих значений $y$. Если все значения $y$ близки к соответствующему значению $\dot{A}$, тогда $C$ является кандидатом, хорошо прогнозирующем $y$ (можно надеяться, что новый $\bar{x}$, удовлетворяющий $C$, также будет соответствовать значению $y$, близкому к $\dot{A}$). С другой стороны, если значения $y$ очень отличаются одно от другого, тогда $x$, удовлетворяющий $C$, не является правилом, прогнозирующим что-либо соответствующее значению $y$.

Одна из самых многообещающих и быстро развивающихся областей применения ГА — анализ данных и прогнозирование в молекулярной биологии. ГА использовался, между прочим, для интерпретации данных ядерно-магнитного резонанса и определении структуры ДНК (Lucasius и Kateman в 1989), находя правильный порядок для неупорядоченной группы фрагментов ДНК (Parsons, Forrest и Burks).

Schulze-Kremer взял последовательность аминокислот белка Крамбина и использовал ГА, для поиска в пространстве возможных структур той, которая будет хорошо соответствовать последовательности аминокислот Крамбина. Он описал структуры белка, используя "углы скручивания" — грубо говоря, углы, образованные пептидными цепями, соединяющими аминокислоты и углы, образованные цепями в аминокислотах ответвлений. Schulze-Kremer использовал 10 углов скручивания, чтобы описать каждую из $N$ (46 в случае Крамбина) аминокислот в последовательности данного белка. Эта коллекция из $N$ наборов 10 углов скручивания полностью определяет трехмерную структуру белка. Хромосома, представляя структуру кандидата с $N$ аминокислотами, таким образом, содержит $N$ наборов десяти вещественных чисел. Это представление проиллюстрировано на рис. 1. Следующим шагом явилось определение фитнесс-функции в пространстве хромосом. Цель состоит в том, чтобы найти структуру, которая имеет низкую потенциальную энергию для данной последовательности аминокислот. В начальных экспериментах Schulze-Kremer использовал чрезвычайно упрощенную модель, в которой потенциальная энергия структуры предполагалась как функция только углов

скручивания, электростатических парных взаимодействий между атомами, и парных взаимодействий Ван-дер-Ваальса между атомами.

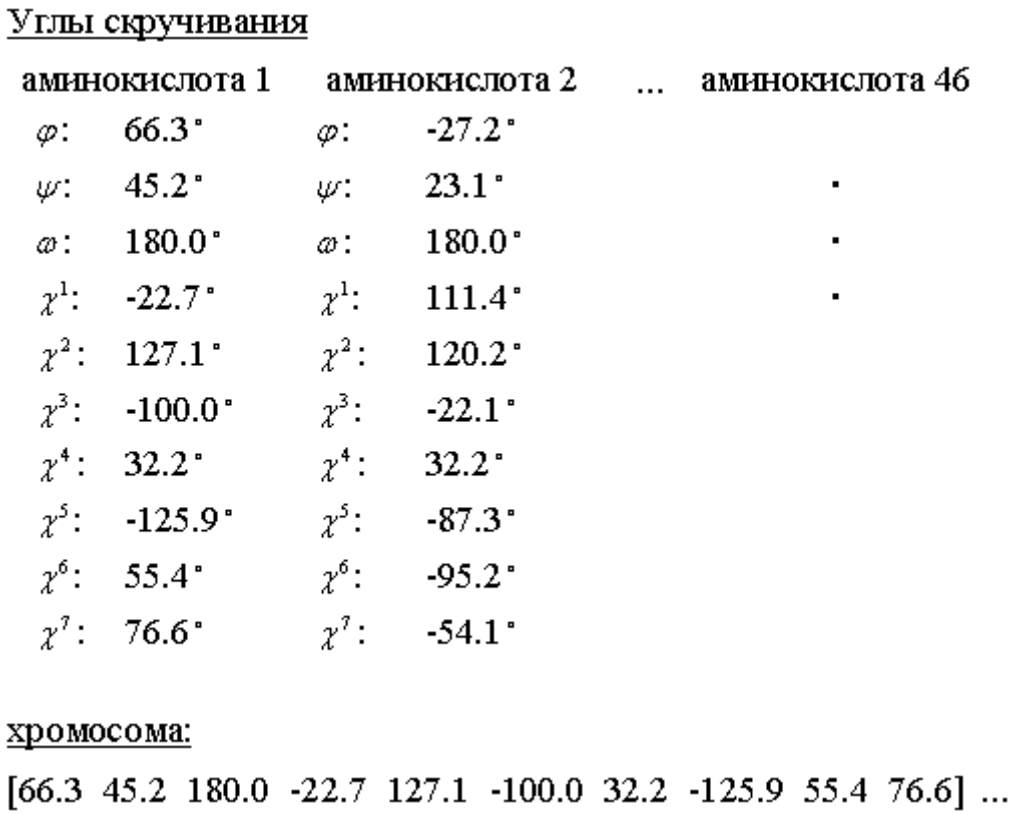


Рисунок 1 — Представление структуры белка, использованное Schulze-Kremer

Поиск ГА произвел множество структур с весьма низкой потенциальной энергией. К сожалению, ни одна из произведенных особей не была структурно подобна Крамбину. Препятствием явилось то, что для ГА было слишком легко найти структуры с низкой энергией при использовании функции упрощенной потенциальной энергии; то есть, фитнесс-функция не была в достаточной мере эффективной для того, чтобы вынудить ГА находить реальную целевую структуру. Приведенный пример свидетельствует о том, что подчас весьма сложно, а зачастую и вовсе невозможно построить явную целевую функцию.

Для решаемых в рамках диагностической системы прогностических задач наиболее употребим подход, описанный в [1], при котором целью является выбор с помощью ГА таких наблюдений из множества данных, которые имеют

сходные тенденции изменений независимых переменных, и близкие значения зависимых переменных. Таким образом, генетические алгоритмы выбирают из тех фактов, которые уже состоялись в прошлом, такие, которые имеют достаточно много общего с фактом, который имеет место в настоящем. Этот подход был реализован в разрабатываемой прогностической системе, после того, как прошел некоторую адаптацию. Уточнение применяемых модификаций ГА невозможно без четкой формализации поставленной задачи, выявления факторов, которые являются исходными для вынесения адекватного прогноза.

Обучающая выборка структурирована и содержит данные для некоторого количества пациентов. Для каждого пациента имеется совокупность параметров, обычно 7–10 наборов, прослеживающих течение болезни в динамике. Каждому набору соответствует комплекс препаратов, применяемых в промежутках между посещениями. Количество параметров (анализов), входящих в набор, значительно колеблется в пределах от 20-ти до 100. Так как вычислительная сложность при использовании большого параметров весьма значительна, то согласно выдвинутой гипотезе о важности системы нейропептидов в лечении ревматоидного артрита [5], вначале отбираются параметры, связанные с этой гипотезой и наиболее общие медицинские показатели функционального состояния пациента. Таким образом, получена некая совокупность переменных, с которой будет работать ГА. На начальных стадиях проектирования используется набор параметров, представленный на рис. 2. Под параметром понимается значение анализа (вещественное число). Обычно $N<5$, а $M$ на начальном этапе принимается равным в 5–10.

После определения исходного набора параметров, с которым будет работать ГА, решается еще одна важная проблему, от которой зачастую зависит успешность применения ГА — кодирование. В приведенном на рис. 2 наборе входных параметров определено несколько различных типов данных — целочисленный, вещественный и логический. Первый и последний можно

обобщить в один тип данных — целочисленный. Таким образом, возраст и пол пациента — целые числа. Препарат (его наличие или отсутствие), продолжительность приема, доза (нормированием приведенная к целочисленному значению) также представляют собой целые числа. В то же время для параметров характерно вещественное представление. Имеет смысл реализовать генетические операторы над вещественными числами, что сулит значительный выигрыш в вычислительной сложности.

$$\left\{\begin{array}{l}(Возраст;пол),\\ (Препарат^1;продолжительность\ применения;доза),\\ (Препарат^2;продолжительность\ применения;доза),\\ \vdots\\ (Препарат^N;продолжительность\ применения;доза),\\ (параметр^1;параметр^2;\ldots;параметр^M)\end{array}\right\}$$

Рисунок 2 — Представление набора входных параметров в виде хромосомы

Используется следующая стратегия применения ГА. На практике задача предикции состоит в назначении адекватного медикаментозного лечения и получении качественного прогноза состояния больного как следствие совокупности факторов текущего состояния и назначенного лечения. При поступлении больного производится некоторый набор анализов. Кроме того, имеется информация из анамнеза пациента, которую также весьма важно учесть при назначении лечения. На этом этапе необходимо сделать заключение о стратегиях лечения пациента, т.е. назначить ему некий комплекс препаратов. Именно нахождение этого комплекса препаратов и представляется точкой приложения всей мощи ГА в этой задаче. С помощью традиционных генетических операторов или их адаптированных модификаций продуцируется набор хромосом, который пропускается через обучающую выборку и производится поиск похожих наборов (анамнез; параметры), когда

примененная стратегия лечения приводила к ремиссии или улучшению состояния больного. Степень ремиссии задается программно. Таким образом, подбирается комплекс препаратов и назначается лечение. Затем делается качественный прогноз, исходя из назначенного лечения. Здесь в качестве хромосом уже выступает совокупность параметров, так как сочетание (комплекс препаратов; анамнез) уже известно. Важно помнить, что экспериментальные данные прослеживают течение заболевания в динамике, поэтому необходимо учитывать состояние пациента в предыдущие циклы лечения.

Дальнейшее развитие прогностической системы заключается в уточнении модели и дополнении ее другими параметрами, исключении малозначащих факторов и создание программной реализации, пригодной для использования в медицинских учреждениях. После продолжительного тестирования и доказанной адекватности предлагаемых системой стратегий лечения предполагается ее использовать в больницах и научно-исследовательских институтах, как систему поддержки принятия решений при диагностировании ревматоидного артрита.

Перечень ссылок

1. Mitchell, M. (1996). An Introduction to Genetic Algorithms. Cambridge, MA: MIT Press. — P. 42–49.

2. Курейчик В.М. Генетические алгоритмы. Обзор и состояние// Новости ИИ. 1998. — №3. — С. 14–63.

3. Goldberd David E. Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley Publishing Company, Inc. 1989, 412 p.

4. Chambers L.D., Practical Handbook of Genetic Algorithms. CRS Press, Boca Ration FL, 1995, v. 1, 560 p., v. 2, 448 p.

5. Гнилорыбов А.М. Нейропептиды и нейрогенные механизмы артритов. Государственный медицинский университет. / Электронный ресурс. Способ доступа: URL: http://rheumatology.org.ua/?area_id=13&stuff_id=66.