УДК 681.3

С.В. Хмелевой, Т.А. Васяева

ФИТНЕСС-ФУНКЦИИ ГЕНЕТИЧЕСКОГО АЛГОРИТМА РЕШЕНИЯ ЗАДАЧИ КЛАССИФИКАЦИИ

Введение.

Задача классификации является одной из базовых классических задач, актуальность которой бесспорна. Одним из естественным подходов для её решения является создание набора правил, на основе которого и осуществляется классификация. Эффективным механизмом для автоматической генерации такого набора является аппарат генетических алгоритмов (ГА). Важнейшей задачей при построении ГА является определение фитнесс-функции. Этот актуальный вопрос и является предметом данного обзора.

Постановка задачи и ее решение.

Классификация является наиболее популярной задачей для Data Mining. Задачей классификации является отнесение целевого атрибута входного сигнала к одному классу из заранее определенного множества на основании входных атрибутов. Естественным подходом для этого будет получение набора продукций – условий на входной сигнал. Условия чаще всего имеют вид [1]:

$$C = \bigcap_{i=1}^{n} \left(X_{i\min} < X_i < X_{i\max} \right), \tag{1}$$

где $\bigcap$ – логический оператор «И», X_i – некоторые влияющие факторы из n возможных.

Продукции могут быть получены автоматически с помощью эволюционного алгоритма.

Для конкретного входного воздействия принимается решение на основании такого набора продукций. Вид принятия решения и форма наложения условий на множество примеров определяются конкретной реализацией генетического алгоритма. Например, наиболее распространенным является простое или взвешенное усреднение решения, выдаваемого правилами.

Фитнесс-функции, применяемые для оценки точности прогнозирования. Критерием качества фитнесс-функции должна быть оценка точности прогноза, наиболее распространенной из которых является среднеквадратичная ошибка (MSE), характеризуемая чувствительностью к значительным отклонениям от прогнозируемого значения:

$$E = \frac{1}{p} \sum_{i=1}^{p} (x_i - \tilde{x}_i)^2 , \tag{2}$$

где x_i – действительное значение прогнозной величины; $\tilde{x}_i$ – спрогнозированная величина; i – индекс входного воздействия; p – общее количество входных воздействий.

Для устранения зависимости погрешности от размерности данных может быть взято абсолютное отклонение прогноза $\tilde{x}_i$ от истинного значения x_i , деленное на диапазон входных значений, что есть сутью нормирования:

$$E = \frac{1}{p} \sum_{i=1}^{p} \left| \tilde{x}_i - x_i \right| / (x_{i\max} - x_{i\min}) , \tag{3}$$

Функция (4) возвращает ошибку с нормализованной дисперсией:

$$E = \frac{1}{\sigma^2} * \frac{1}{p} * \sum_{i=1}^{p} (x_i - \tilde{x}_i)^2 , \tag{4}$$

Также широко распространено применение средней арифметической ошибки (MAE):

$$E = \frac{1}{p} \sum_{i=1}^{p} \left| x_i - \tilde{x}_i \right| , \tag{5}$$

Вариаций таких функций может быть достаточно много.

Составляющие фитнесс–функции, используемые при продукционном подходе. В сложных задачах фитнесс-функция мультикритериальна, состоит из аддитивных или мультипликативных составляющих. Кроме точностных характеристик прогноза важную роль могут играть составляющие, оценивающее качество самого правила. Качество правила можно оценить следующим образом. Имеем правило, представленное в форме: ЕСЛИ А ТО С, где А – условие, С – вывод. Для точки, условие которой соответствует А, вывод может соответствовать С или не соответствовать С. Классификация правильности прогнозирования может быть описана матрицей *2 x 2*, как показано в таблице 1 [2]:

Таблица 1. Классификация правильности прогнозирования

Для точки, условие которой соответствует A		Действительный класс	
		C	не C
Прогнозируемый класс	C	*TP*	*FP*
	не C	*FN*	*TN*

Здесь:

TP = true positives – число примеров, удовлетворяющих и A, и C (правильный прогноз);

FP = false positive – число примеров, удовлетворяющих A, но не совпадающих с C (ошибка 1 рода, неправильный прогноз);

FN = false negative – число примеров, не удовлетворяющих A, но удовлетворяющих C (ошибка 2 рода, возможный, но не реализованный правильный прогноз);

TN = true negative – число примеров, не удовлетворяющих ни A, ни C (правильный отрицательный прогноз).

Хорошие правила имеют высокие *TP* и *TN*, и низкие *FP* и *FN*.

Можно также определить показатель CF , отвечающей за ошибки 1 рода (т.е. точность), в терминах таблицы 1:

$$CF = \frac{TP}{TP + FP}. \tag{6}$$

Кроме точности необходимо, чтобы правило обладало полнотой. Мера полноты правила $Comp$ отвечает за ошибки 2 рода:

$$Comp = \frac{TP}{TP + FN}. \tag{7}$$

Среди других критериев можно выделить: количество ситуаций, отобранных правилом; близость правил в выборке; максимизация энтропии данных, отбираемых правилом; максимизация разброса отбираемых правилом данных; максимизация логарифма отношения дисперсии отбираемых правилом данных, к дисперсии всех данных; минимизация размера правила.

Например, в [3, 4, 5] показаны следующие составляющие:

Высокое значение покрытия примеров правилом. Частота употребления *i*-го нечеткого правила R_i , вычисляемая по множеству обучающих примеров EP, определяется:

$$\Psi_{E_p}(R_i) = \frac{\sum_{l=1}^{p} R_i(e_l)}{p}, \tag{8}$$

где e_l – примеры из обучающей выборки.

Малая степень покрытия отрицательных примеров. Определив количество правильно отобранных примеров через n_ω^+, а количество неправильно отобранных примеров (отобранное частью условием, но не отобранное частью - выводом), через:

$$n_\omega^- = \left| E_\omega^-(R_i) \right| \text{, где } E_\omega^-(R_i) = \{e_l \in E_p / R_i = 0, при \quad A_i(ex_l) > 0\}, \tag{9}$$

где $A_i(ex_l)$ - значение активации части – условия l-го правила.

Также определяется штрафная составляющая фитнесс-функции для уменьшения степени покрытия отрицательных примеров.

$$g_n(R_i^-) = \begin{cases} 1, \text{ если } n_\omega^-(R_i) \le k * n_\omega^+(R_i); \\ \dfrac{1}{n_\omega^-(R_i) - k * n_\omega^+(R_i) + \exp(1)}, \text{ иначе.} \end{cases} \tag{10}$$

В [6] предложена фитнесс-функция, основанная на *J*-мере (11).

$$\begin{cases} b = \dfrac{|C \,\&\, P|}{|C|} \\ a = \dfrac{|P|}{N} \\ J = \dfrac{|C|}{N}\left(b * \log\left(\dfrac{b}{a}\right) + (1-b) * \log\left(\dfrac{1-b}{1-a}\right)\right) \end{cases}, \tag{11}$$

где $|C|$ – количество записей данных из обучающей выборки (ОВ), отобранных частью-условием правила,

$|P|$ – количество записей данных из ОВ, отобранных частью-выводом правила,

$|C \,\&\, P|$ – количество записей данных из ОВ, отобранных полным правилом,

N - общее количество данных в ОВ.

Таким образом, можно отметить, что:

- переменная b представляет собой отношение покрытия всего правила к покрытию условия;
- переменная a представляет собой долю отобранных выводом записей данных; в выражении J она играет роль смещения, помогающего обобщению правила;
- общее выражение J взято из теории информации [7], основано на энтропии точек, отбираемых правилом, и по существу является мерой расстояния между апостериорной и априорной величиной доверия к правилу.

Такая фитнесс-функция, как и многие похожие из источников, указанных в [2], имеет составляющие, прямо зависящее от точности прогноза, получаемого правилом. Кроме этого, она учитывает качество каждой из частей правила (условия и вывода).

Модификация составляющих фитнесс-функции в случае нечеткого кодирования хромосомы. В случае, если хромосома кодирует нечеткое правило, необходимо учитывать, что и такие величины, как, например, количество примеров, отбираемых правилом, также не будет являться целым числом, поскольку каждый пример может быть отобран правилом только с некоторой степенью истинности. Величина количества примеров, правильно отобранных правилом, используемая в **Ошибка! Источник ссылки не найден.** модифицируется с учетом нецелого количества примеров: $n_\omega^+ = \left|E_\omega^+(R_i)\right|$, где $E_\omega^+(R_i)$ – множество степеней активации положительных примеров (в правилах, где степени активации этих примеров больше некоторого определяемого пользователем порога ω):

$$E_\omega^+(R_i) = \{e_l \in E_p \,/\, R_i \geq \omega\}\,. \tag{12}$$

Значение активации части – условия правила $A_i(ex_l)$ также не будет целой величиной, соотвественно, модифицируется и значение $n_\omega^- = \left|E_\omega^-(R_i)\right|$, также применяемое в **Ошибка! Источник ссылки не найден.**. На основании этих модифицированных величин можно определить еще одно точностная составляющая, специфичная только для нечетких правил: большая степень покрытия положительных примеров, может быть определена из (13):

$$G_\omega(R_i) = \sum_{e_l \in E_\omega^+(R_i)} R_i(e_l) / n_\omega^+(R_i)\,, \tag{13}$$

где $R_i(e_l)$ – степень активации правила точкой e_l из обучающей выборки.

В [8, 9, 10] рассмотрено еще одно точностное составляющее: степень нечеткости или энтропии системы нечеткого вывода:

$$R_\phi = \frac{1}{n} \bullet \sum_{i=1}^{n} R_{cur,i} \tag{14}$$

где $R_{cur,i}$ – число активных правил для i-го обучающего примера.

Если число активных правил минимально (одно), система нечеткого вывода максимально понятна. Система нечеткого вывода с высоким числом активных правил ведет себя как нейронная сеть, где большое число нейронов определяет выходное значение. Чтобы понизить энтропию системы нечеткого вывода и частичное совпадение функций принадлежности, включенных в подусловия различных правил, максимальное число R_{max} и желательное число R_{act} активных нечетких правил, может быть определено и еще несколько составляющих фитнесс-функции. Рассмотренные далее компоненты специфичны только для нечеткого кодирования хромосомы [3, 4, 5], и отвечают за форму функций принадлежности подусловий правил.

Малая степень взаимного перекрытия с другими правилами множества. Имеем центры подусловий *i*-х правил $N_i=(N_{ix}, N_{iy})$, определенных ранее в процессе работы алгоритма $(i=1..d_r)$, где d_r – количество шагов

алгоритма по генерации правил. Степень взаимного перекрытия правил можно определить как (15).

$$LNIR(R_i) = 1 - NIR(R_i), \tag{15}$$

где $NIR(R_i) = Max_i\{H_i\}$

$H_i = *(A(N_i x), B(N_i y)), i = 1..d_r$,

$A(N_i x) = *(A_1(N_i x_1), ..., A_n(N_i x_n))$ – часть условия правила,

$R_i : ЕСЛИ \bigcap_{j=1}^{n}(x_j \in A_j)\ ТО\ y \in B$.

Здесь: * – *t*-норма (чаще всего функция минимум),

n – количество входных переменных (или подусловий правила).

Малые значения *LNIR* получаются, если в популяции уже созданы правила, подобные данному. Если существуют аналогичные правила, значение *LNIR*=0.

Высокая симметричность правил:

$$RS(R_i) = \frac{1}{ds_i}, \tag{16}$$

где $ds_i = Max_{j=1..n+m}\{ds_i^j\}$,

$$ds_i^j = Max\left\{\frac{ds_{i1}^j}{ds_{i2}^j}, \frac{ds_{i2}^j}{ds_{i1}^j}\right\},$$

$ds_{i1}^j = b_{ij} - a_{ij}, ds_{i2}^j = c_{ij} - b_{ij}$.

Здесь: *m* – число выходных переменных,

a – крайняя левая точка *j*-го подусловия *i*-го правила, которую можно считать значимой (точка «начала подусловия правила»),

b – точка максимума *j*-го подусловия *i*-го правила («центра»),

c – крайняя правая точка *j*-го подусловия *i*-го правила.

Величина *RS* равняется 1 при симметричном подусловии правила.

В [8, 9, 10] можно выделить такие составляющие, отвечающие за форму правил:

$$OF^E = \frac{1}{\left(\frac{R_\phi}{R_{\max}} - 1\right)a + \left(\frac{R_\phi}{R_{act}} - 1\right)b + 1}, \tag{17}$$

где *a* и *b* – предварительно определенные весовые факторы.

Для уменьшения числа различных функций принадлежности подусловий правил системы нечеткого вывода в случае свободно изменяемой формы функций принадлежности подусловий может быть применено следующее составляющее:

$$OF^S = \left(\frac{s_{no}}{(n+m)R_{total}}\right)\gamma + 1, \tag{18}$$

где S_{no} – число подобных функций принадлежности,

Rtotal – общее число правил,

$\gamma \in R$ – предварительно определенный весовой фактор для числа подобных функций принадлежности.

Также следует выделять правила, которые не активируются ни на какой точке обучающей последовательности, или активируются всегда. В первом случае правило может быть удалено, во втором — может быть удалено одно подусловие из правила. Составляющее, ограничивающее число таких функций принадлежности, с предварительно определенными весовыми факторами $\zeta, \eta \in R$, может быть определено из выражения (19):

$$OF^{UZ} = \left(\frac{u_{no}}{(n+m)R_{total}}\right)\zeta + \left(\frac{z_{no}}{(n+m)R_{total}}\right)\eta + 1 \tag{19}$$

где u_{no} – число функций принадлежности, активирующихся на всех примерах, z_{no} – ни на одном примере.

Также возможны дополнительные масштабные, штрафные составляющие для функции, например, отвечающие за регулировку размера особи, за накладывание ограничений на пространство поиска. В последнем случае необходимо помнить, что штрафные значения влияют на корреляцию ландшафта фитнесс-функции, могут значительно уменьшить длину автокорреляционной функции, что может привести или к концентрации всех особей «в одном углу», или к неоправданному увеличению размера популяции.

Зачастую, кроме точностных составляющих, в фитнесс-функции присутствует минимум одна составляющая, отвечающая за простоту правил. Соотношение между точностью классификации и числом

используемых факторов (если нужно сокращение числа используемых факторов), осуществляется с помощью назначения веса (*Q1* и *Q2*, $Q_1, Q_2 \in (0,1)$) для каждой из составляющих, которые объединяются аддитивно [5] (20) .

$$F = \left(\frac{X_i}{X_n}\right) \cdot Q_1 + \left(\frac{E_i - E_n}{E_n}\right) \cdot Q_2, \qquad (20)$$

где X_i – количество отобранных факторов для i-й хромосомы,
X_n – общее количество факторов,
E_i – ошибка классификации для i-й хромосомы,
E_n – ошибка классификации при использовании полного количества факторов.

На основании рассмотренных составляющих может быть составлена фитнесс-функция генетического алгоритма, строящего набор правил нечеткого вывода для решения любой многокритериальной задачи, где влияние каждого из критериев определяется эмпирическим или опытным путем.

Выводы.

Показаны правила построения генетического алгоритма для создания набора продукций применительно к задаче классификации.

Проанализированы фитнесс-функции для оценки точности прогнозирования, применимые для любого прогнозирующего алгоритма. Они основаны на оценках, подобных MSE и MAE и считаются для каждой точки временного ряда.

Выделены составляющие, преимущественно точностного или энтропийного характера, которые могут учавствовать в фитнесс-функциях любого ГА, формирующего оптимальный набор продукций.

Выделены составляющие, которые могут быть задействованы в фитнесс-функциях ГА, создающего набор нечетких правил, входящих в систему нечеткого вывода.

Выделены дополнительные штрафные составляющие, отвечающие за регулировку размера особи, за простоту правил.

ЛИТЕРАТУРА:

1. M.Mitchel. An Introduction to Genetic Algorithms. MIT Press, 1998.
2. Alex A. Freitas. A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery // http://citeseer.ist.psu.edu/freitas01survey.html .- Б.и., 2001.- 27с.
3. F.Herrera, M.Lozano, J.L.Verdegay. A Learning Process for Fuzzy Control Rules using Genetic Algorithms // Fuzzy Sets and Systems. - 100 (1998). - 143-158.
4. O.Cordon, F.Herrera. A Three-Stage Evolutionary Process for Learning Descriptive and Approximative Fuzzy Logic Controller Knowledge Bases from Examples.// International Journal of Approximate Reasoning Vo.- 17-4 (1997).- 369-407.
5. O.Cordon, F.Herrera. Hybridizing Genetic Algorithms with Sharing Scheme and Evolution Strategies for Designing Approximate Fuzzy Rule-Based Systems. // Fuzzy Sets and Systems .- 118:2 (2001).- 235-255.
6. Smyth P. and Goodman R.M. – “Rule induction using information theory”. In Piatesky-Shapiro G. and Frawley J. (editors), Knowledge Discovery in Databases, pp. 159-176, Cambridge: MIT press, 1991.
7. Cover, T.M.: Thomas. J.A. Elements of Information Theory. John Wiley&Sons. 1991.
8. H.Surmann, M.Maniadakis. Learning feed-forward and recurrent fuzzy systems: A genetic approach. Journal of System Architecture .- 47(2001), pp. 649-662.
9. H.Surmann, A.Selenschtschikow. Automatic Generation of fuzzy logic rule bases: Examples I. – Proc. of the NF2002: first international ICSC conference on neuro-fuzzy technologies. Pp 75-81, CUBA 16-19 jan, 2002.
10. H.Surmann. Learning a fuzzy rule based knowledge representation. – Proc. of 2 ICSC symp. on neural computation, NC`2000. Berlin, 23-26 may. 2000, pp. 349-355.

Хмелевой С.В. – к.т.н., доцент каф. АСУ, ДонНТУ, hmelevoy_sergey@ukr.net, 0506142940
Научные интересы: нечеткие методы, генетические алгоритмы, нейронные сети, прогнозирование временных рядов.

Васяева Т.А. – к.т.н., доцент каф. АСУ, ДонНТУ, vasyaeva_tanya@tr.dn.ua, 0504728558
Научные интересы: искусственный интеллект, компьютерные технологии

УДК 681.3/ С.В. Хмелевой, А.Н. Васяева / Составляющие фитнесс-функции алгоритма решения задачи прогнозирования с использованием продукционного похода // … 5 стр. Библ.: 10., рус.
Показано, что наиболее использованным подходом для решения задачи прогнозирования с помощью генетических алгоритмов является продукционный подход. В работе для решения задачи прогнозирования создавался набор нечетких продукций. Для генетического алгоритма решения задачи прогнозирования выполнен анализ применимости разнообразных фитнесс-функций и их частей. Выделены части, которые отвечают за точность прогнозирования и части, которые отвечают за качество правил. Выделены части, которые применимые к произвольным продукциям и части, применимые лишь к нечетким продукциям.

УДК 681.3/ С.В. Хмільовий, Т.О. Васяєва / Складові фітнес-функції алгоритму рішення задачі прогнозування з використанням продукційного підходу //… … 5 стр. Библ.: 10., рус.
Показано, що найбільш застосовним для рішення задачі прогнозування за допомогою генетичних алгоритмів є продукційний підход. В даній роботі для рішення задачі прогнозування створювався набір нечітких продукцій. Для генетичного алгоритму рішення задачі прогнозування виконано аналіз застосовності різноманітних фітнес-функцій та їх частин. Виділено частини, що відповідають за точність прогнозування та частини, що відповідають за якість правил. Виділено частини, що застосовні до довільних продукцій та частини, застосовні лише до нечітких продукцій.

УДК 681.3 / Khmilovyy S.V, Vasyaeva T.A. / Making fitness functions of algorithm of the forecasting problem decision with use fuzzy rules //… … 5 стр. Библ.: 10., рус.
It is shown that the most applicable for the decision of a forecasting problem by means of genetic algorithms is the rules set. In the current work for the decision of a forecasting problem the set of fuzzy rules was created. For genetic algorithm decision of a forecasting problem the analysis of applicability of various fitness functions and their parts is made. It is allocated parts which are responsible for accuracy of forecasting and a part which are responsible for quality of rules. It is allocated parts, applicable to any rules and parts, applicable only to the fuzzy.