Р.Б. Сергиенко

Метод формирования нечеткого классификатора самонастраивающимися коэволюционными алгоритмами

Аннотация. Предлагается новый метод формирования нечеткого классификатора, объединяющий два основных подхода к проектированию нечетких систем генетическими алгоритмами – Питтсбургский и Мичиганский. Для автоматической настройки параметров генетического алгоритма предлагается коэволюционный подход. Проведены исследования метода генерирования нечеткого классификатора на ряде практических задач классификации.

Ключевые слова: нечеткий классификатор, коэволюционный алгоритм, классификация, извлечение знаний.

Введение

Одна из наиболее распространенных задач анализа данных – классификация. Потребность в классификации возникает в самых разнообразных технических, экономических и организационных системах. К задачам классификации сводится диагностирование заболеваний в медицине, принятие решения о выдаче кредита клиенту банка, распознавание изображений и звука и многие другие задачи. На сегодняшний день создано большое количество разнообразных алгоритмов классификации [1], разработаны формализованные теории конструирования высокоэффективных композиций из этих алгоритмов [2]. Однако недостатком большинства из существующих алгоритмов классификации является работа по принципу «черного ящика» - невозможность явной интерпретации закономерностей, приводящих к отнесению объекта классификации к тому или иному классу.

Такого недостатка лишен нечеткий классификатор [3], представляющий собой базу нечетких правил. Каждое нечеткое правило – выражение причинно-следственной закономерности отнесения объекта к какому-либо классу в лингвистической форме. Т.е. нечеткое правило – это знание, доступное для прямого восприятия экспертом в соответствующей проблемной области. Таким образом, нечеткий классификатор представляет собой инструмент интеллектуального анализа данных, позволяющий одновременно проводить классификацию и извлекать экспертные знания, связанные с процессом классификации.

Формирование нечеткого классификатора сводится к задаче оптимизации – выбору наилучшей базы нечетких правил из множества существующих. Данная задача оптимизации отличается существенной вычислительной и алгоритмической сложностью, обусловленной крайне высокой размерностью решаемой задачи оптимизации, процедурно заданной целевой функцией (вычисление которой, как правило, трудоемко), наличием дискретных переменных и др. В такой ситуации использование классических методов оптимизации неприемлемо. Для задачи формирования нечеткого классификатора целесообразно использовать эволюционные алгоритмы [4], которые хорошо зарекомендовали себя при решении именно такого рода задач оптимизации.

Однако применение эволюционных алгоритмов оптимизации сопряжено с другой серьезной проблемой – высокой сложностью и трудоемкостью настройки параметров алгоритма оптимизации на решаемую задачу в связи с большим числом возможных комбинаций таких параметров (селекции, мутации, скрещивания и некоторых других). Эффективность одной и

той же настройки на разных задачах и различных настроек на одной и той же задаче может изменяться в очень широком диапазоне. Поэтому выбор настроек наугад является неприемлемым, так как многие комбинации параметров алгоритма оказываются неработоспособными, а тщательная настройка под новую задачу является чрезмерно трудоемкой. Ввиду вычислительной затратности формирования нечеткого классификатора данная проблема становится особенно актуальной. Для решения указанной проблемы предлагается использование так называемого коэволюционного алгоритма для адаптации стратегии оптимизации [5], позволяющего отказаться от выбора настроек эволюционного алгоритма за счет взаимодействия (конкуренции и кооперации) множества эволюционных алгоритмов с различными настройками.

Таким образом, разработка и исследование схемы формирования нечетких классификаторов самонастраивающимися коэволюционными алгоритмами для одновременного решения задач классификации и извлечения знаний, является актуальной научно-технической задачей.

1. Схема работы нечеткого классификатора

Нечеткий классификатор [3] – алгоритм классификации, основанный на извлечении нечетких правил из массивов данных. Каждое правило в нечетком классификаторе содержит значения лингвистических термов для всех информативных признаков (включая терм «игнорирования»), номер класса, которому соответствует данное правило, а также уровень значимости в интервале [0;1], показывающий степень достоверности правила. Рассмотрим алгоритм работы классификатора.

Пусть $X = (x_1, x_2, .. , x_k)$ – вектор информативных признаков задачи классификации. Для каждого i-го информативного признака в нечетком правиле R соответствует определенное нечеткое число, по которому можно вычислить функцию принадлежности $\mu_i(x_i)$. Далее определяем значение функции принадлежности по всему правилу как минимальное значение среди всех $\mu_i(x_i)$: $\mu(X) = \min \{\mu_1(x_1), \mu_2(x_2), .. , \mu_k(x_k)\}$. Для определения класса, соответствующему правилу R, для каждого класса по обучающей выборке вычисляется среднее значение функции принадлежности по правилу R [3]:

$$\beta_{Classh}(R) = \frac{\sum_{X \in class\ h} \mu(X)}{n_{Class\ h}},$$

где h=1..L – номер класса, L – количество классов, X – элемент обучающей выборки, $n_{Class\ h}$ - количество элементов класса h в обучающей выборке. Тогда класс C, соответствующий нечеткому правилу R, определяется как класс, обеспечивающий максимальное значение $\beta_{Class\ h}$:

$$\beta_{Class\ C}(R) = \max\left\{\beta_{Class\ 1}(R), \beta_{Class\ 2}(R), .., \beta_{Class\ L}(R)\right\}.$$

Уровень значимости нечеткого правила CF определяется по формуле:

$$CF = \frac{\beta_{Class\ C}(R) - \frac{\sum_{\substack{h=1,\\ h \neq c}}^{L} \beta_{Class\ h}(R)}{L-1}}{\sum_{h=1}^{L} \beta_{Class\ h}(R)}.$$

Данная формула позволяет определить, насколько среднее значение функции принадлежности по правилу для класса-«победителя» отличается от аналогичных значений для других классов. Значение CF=1 в случае $\beta_{Class\ C}$ = 1 и равенстве нулю остальных $\beta_{Class\ h}$. CF=0 в случае равенства значений средней функции принадлежности для всех классов – эта ситуация означает, что правило полностью неинформативно и не позволяет идентифицировать никакие характерные особенности класса.

Для осуществления нечеткого вывода по базе нечетких правил классификатора при поступлении объекта X для классификации определяется правило-победитель W, удовлетворяющее следующему условию [3]:

$$\mu_W(X) \cdot CF_W = \max_{i=1..S} \mu_i(X) \cdot CF_i$$

где S – число правил в базе.

Таким образом, победителем назначается правило, обладающее максимальным значением произведения функции принадлежности по правилу для данного объекта классификации на уровень значимости правила.

Основное преимущество нечеткого классификатора перед другими алгоритмами классификации (например, нейронными сетями, работающими по принципу «черного ящика»)

заключается в том, что база нечетких правил представляет собой лингвистические знания, доступные для понимания и интерпретации экспертами в проблемной области решаемой задачи. Т.е. нечеткий классификатор представляет собой один из инструментов Data Mining для извлечения знаний. Кроме того, для нечеткого классификатора характерна высокая обобщающая способность. Следовательно, для него не характерна проблема переобучения, имеющая место при использовании тех же нейронных сетей.

Выступая одновременно алгоритмом классификации и алгоритмом извлечения знаний, нечеткий классификатор должен удовлетворять трем требованиям:

1) эффективность классификации должна быть максимальной;

2) количество правил в базе должно быть минимальным, так как только незначительное число правил будет восприниматься человеком и представлять лингвистические знания;

3) правила должны быть максимально общими, т.е. обладать как можно большим числом термов «игнорирования» признаков (иными словами, наименьшей длиной), так как знания по определению должны отражать наиболее общие закономерности.

Основная сложность в использовании нечетких систем, в т.ч. нечеткого классификатора состоит в генерации эффективной базы правил. Ввиду алгоритмической и вычислительной сложности данной задачи целесообразно использование эволюционных алгоритмов оптимизации.

Существует два основных подхода в использовании генетических алгоритмов (ГА) в задачах конструирования нечетких систем: Питтсбургский и Мичиганский [6]. В Мичиганском методе индивиды генетического алгоритма представляют собой отдельные правила, в Питтсбургском – базу нечетких правил в целом. Недостаток Мичиганского метода – противоречие между целевой функцией для индивидов и эффективностью базы правил в целом. Питтсбургский метод лишен этого недостатка, однако требует значительных вычислительных ресурсов – размерность решаемой задачи оптимизации возрастает многократно. Поэтому гибридизация Питтсбургского и Мичиганского методов в задачах генерации нечетких систем является актуальной задачей.

2. Новая схема формирования нечеткого классификатора

В [7] задача формирования нечеткого классификатора эволюционными алгоритмами сформулирована в виде трехкритериальной задачи оптимизации (критерии – надежность классификации, число правил, средняя длина правил). В работе предлагается использование Питтсбургского подхода с использованием Мичиганского метода в качестве оператора мутации [8]. Недостатками предлагаемой схемы являются:

- алгоритмическая и вычислительная сложность реализации многокритериальной оптимизации, необходимость использования методов сужения множества Парето после работы собственно нечеткого классификатора;

- отсутствие строго формализованного метода отбора правил для стартовой базы (в [7] предлагается, или использование всех возможных правил для задач низкой размерности, или отбор только правил с незначительной длиной предикатной части – их число существенно меньше общего числа всех возможных правил - однако, данный подход не является целесообразным для всех задач с точки зрения эффективности классификации, кроме того, число даже только таких правил может быть неприемлемо большим для задач высокой размерности);

- отсутствие этапа строго направленного улучшения стартовых правил (за исключением оператора мутации);

- использование при реализации Питтсбургского подхода целочисленных хромосом переменной длины, что существенно повышает вычислительную сложность метода генерирования классификатора;

- отсутствие автоматического выбора настроек параметров эволюционного алгоритма, используемого для генерирования нечеткого классификатора.

Для устранения указанных недостатков предлагается схема генерирования нечеткого классификатора с новым методом гибридизации Мичиганского и Питтсбургского методов и использованием конкурирующего коэволюционного алгоритма для адаптации стратегии оптимизации в качестве инструмента автоматизации выбора настроек эволюционного алгоритма [5]. Задача формирования нечеткого классификатора сформулирована в виде комбинации задач безусловной и условной однокритериаль-

ной оптимизации без использования многокритериального подхода.

Новая схема формирования нечеткого классификатора включает три основных этапа [9] (помимо этапа фаззификации информативных признаков, осуществляемой тривиальным способом – равномерное заполнение нечеткими числами интервалов варьирования признаков): формализованная процедура отбора стартовых правил с использованием априорной информации из обучающей выборки, этап улучшения стартовых правил Мичиганским методом (задача однокритериальной безусловной оптимизации), этап сокращения найденного множества правил Питтсбургским методом (задача однокритериальной условной оптимизации). Рассмотрим этапы более подробно.

Формирование начальной популяции для Мичиганского этапа (отбор стартовых правил). Данная операция очень важна, так как случайное генерирование правил для начального заполнения популяции неприемлемо – при значительном числе информативных признаков в задаче классификации вероятность случайной генерации правила, которому соответствовал хотя бы один элемент из обучающей выборки, становится крайне малой. Эта проблема становится существенной уже при размерности четыре и выше. Полный перебор всех возможных правил практически невозможен, так как общее их число при данном числе лингвистических термов для каждого признака исчисляется 6^n, где n – число информативных признаков задачи классификации. Поэтому необходимо использовать априорную информацию из обучающей выборки. Процедура формирования начальной популяции в таком случае выглядит следующим образом.

Пусть n – число нечетких правил, необходимое для заполнения популяции для Мичиганского этапа генерирования, k – число классов, l – число информативных признаков в задаче классификации. Последовательность шагов:

1) вычислить $m = \text{int}(n/k)$;

2) отсортировать элементы обучающей выборки по номеру класса в порядке возрастания и определить границы для каждого класса в отсортированном массиве;

3) от i:=1 до k от j:=1 до m;

выполнить случайный выбор элемента класса i из обучающей выборки;

от t:=1 до l;

определить ближайший центр нечеткого числа по признаку t;

назначить соответствующее нечеткое число в нечеткое правило;

изменить нечеткое число на терм «игнорирование» с вероятностью 0,33;

4) заполнить популяцию сгенерированными нечеткими правилами.

Мичиганский этап формирования нечеткого классификатора. Поиск множества нечетких правил, обладающих высоким доверительным уровнем.

Индивиды представляют собой отдельные нечеткие правила. Длина хромосомы равна числу информативных признаков, каждый ген – число от 1 до 6, соответствующее нечеткому числу (для каждого признака используется пять значащих нечетких чисел, а также терм «игнорирования» признака). Функция пригодности индивидов – доверительный уровень правила, вычисляемая по обучающей выборке. Представляется целесообразным включение в пригодность штрафного слагаемого, пропорционального длине правила, с невысоким коэффициентом значимости. Реализуется безусловная однокритериальная оптимизации.

Модифицирован метод формирования нового поколения. Среди родителей и потомков для каждого класса отбирается необходимое число неповторяющихся правил с наибольшим значением пригодности (отбор с вытеснением для каждого отдельного класса). Такой метод обеспечивает поддержание разнообразия правил для каждого отдельного класса и разнообразие классов в популяции. На каждом поколении вычисляется надежность классификации по обучающей выборке множества правил в целом. Популяция с наибольшей надежностью классификации используется на следующей стадии генерирования нечеткого классификатора.

Питтсбургский этап формирования нечеткого классификатора. После того как мы сгенерировали множество правил с заданным уровнем доверительной вероятности необходимо сгенерировать базу правил, состоящую из подмножества сгенерированного множества правил, и обладающую по возможности максимальной эффективностью (наименьшей ошибкой классификации по имеющейся выборке). Вовсе не обязательно, что совокупность всех найденных правил обладает максимальной эффективностью. Кроме того, множество правил может оказаться достаточно большим и избы-

точным. Необходимо сгенерировать базу правил, обладающую по возможности минимальным числом правил.

Индивиды представляют собой базу нечетких правил целиком. Длина хромосомы равна числу правил, найденных на Мичиганском этапе. Хромосомы бинарные, бит «1» означает использование соответствующего нечеткого правила, найденного на предыдущем этапе, бит «0» - исключение правила из базы. Пригодность – надежность классификации базы правил. Вводится ограничение на максимально допустимое число правил, используемых в базе. Таким образом, Питтсбургский этап позволяет сформировать из полученных на предыдущем этапе правил базу, состоящую из существенно меньшего числа правил. Компактные базы правил гораздо предпочтительнее с точки зрения задачи извлечения знаний.

Использование на Питтсбургском этапе бинарных строк фиксированной длины (при фиксированных нечетких правилах и незначительном их числе) существенно уменьшает вычислительную и алгоритмическую сложность генерирования нечеткого классификатора в сравнении со схемой по Ишибучи [7].

В ходе реализации схемы генерирования нечеткого классификатора был рассмотрен вопрос о ситуации, в которой для объекта, поступающего для классификации, не соответствует ни одного правила в базе (нулевой нечеткий вывод по всем правилам). Здесь возможны два варианта:

1) вывод сообщения о невозможности классифицировать объект;

2) автоматическое отнесение к первому классу – такой подход эквивалентен добавлению в базу правила «*Иначе – класс 1*».

В данной работе использовался второй подход, так как он позволяет существенно сократить число используемых правил (особенно данный метод актуален при наличии двух классов в постановке задачи классификации). Недостаток – невозможность идентифицировать «инородный» объект, не относящийся ни к одному из классов. Однако для преодоления данного недостатка достаточно определить для каждого класса границы варьирования информативных признаков по обучающей выборке. При поступлении объекта для классификации осуществляется его проверка на «корректность» по всем классам. В случае существенного выхода за допустимые границы во всех ситуациях, можно заключать об «инородности» объекта.

Следует отметить, что нечеткий классификатор позволяет работать не только с вещественными или целочисленными информативными признаками, но и бинарными, ранговыми и качественными. Для этого достаточно использовать дискретные значения указанных переменных в качестве лингвистических переменных. Однако в случае значительного числа значений ранговых переменных целесообразно проводить фаззификацию.

3. Коэволюционный алгоритм при формировании нечеткого классификатора

Одна из важнейших проблем в использовании эволюционных алгоритмов (в т.ч. для генерирования нечетких систем) – проблема выбора настроек параметров алгоритма. Существует большое число возможных комбинаций параметров генетического алгоритма (селекции, мутации, скрещивания и некоторых других). Эффективность одной и той же настройки на разных задачах и различных настроек на одной и той же задаче может изменяться в очень широком диапазоне. Поэтому выбор настроек наугад является неприемлемым, так как многие комбинации параметров алгоритма оказываются неработоспособными, а тщательная настройка под новую задачу является чрезмерно трудоемкой.

Для решения указанной проблемы предлагались различные подходы, одним из которых являются конкурирующие подпопуляции [10]. Этот подход получил дальнейшее развитие в коэволюционном алгоритме [5], в котором параллельно работают, при этом взаимодействуя между собой, индивидуальные генетические алгоритмы с различными настройками (подпопуляции). «Конкуренция» и «кооперация» (в отличие от [10]) индивидуальных алгоритмов обеспечивает самонастройку эволюционного поиска на решаемую задачу в ходе ее однократного решения и снимает проблему «ручного» выбора наилучшего алгоритма. Стандартный коэволюционный алгоритм состоит из следующих этапов:

- выбор индивидуальных алгоритмов;
- задание параметров коэволюционного алгоритма (размер общего ресурса, величина ин-

тервала адаптации, размер штрафа «проигравшего» алгоритма, размер «социальной карты»);

- независимая работа выбранных алгоритмов в течение интервала адаптации (обычно около пяти поколений);
- оценка алгоритмов;
- перераспределение ресурсов;
- миграция лучших индивидов во все подпопуляции.

Ключевыми этапами работы коэволюционного алгоритма являются перераспределение ресурсов и миграция, которые обеспечивают «конкуренцию» и «кооперацию» между индивидуальными генетическими алгоритмами соответственно.

Ранее были разработаны и исследованы коэволюционные алгоритмы для решения задач безусловной оптимизации. На основе этих исследований [11] был сделан общий вывод, что коэволюционный алгоритм обеспечивает более высокую эффективность в сравнении со средним генетическим алгоритмом, однако в большинстве случаев уступает по эффективности наилучшему индивидуальному алгоритму. Наилучший и средний по эффективности индивидуальные алгоритмы определяются исчерпывающим перебором всех возможных комбинаций настроек с многократным прогоном на каждой задаче и статистическим оцениванием показателей эффективности.

Автором разработан коэволюционный алгоритм условной оптимизации. Адаптация коэволюционного алгоритма на класс задач условной оптимизация связана с добавлением к числу настраиваемых параметров алгоритмов метода учета ограничений, что влечет за собой модификацию одного из ключевых этапов коэволюции – миграции индивидов между индивидуальными алгоритмами. Дело в том, что миграция основана на группировании индивидов из всех подпопуляций, сортировке их по пригодности и «раздаче» всем алгоритмам наилучших индивидов. Использование же различных методов учета ограничений перестает делать пригодность единым критерием оценивания индивидов из различных подпопуляций. По результатам проведенных исследований [12] наилучшей схемой миграции для коэволюционного алгоритма условной оптимизации признана так называемая «пропорционально-групповая». Суть заключается в объединении подпопуляций с одинаковым методом учета ограничений в группы, сортировке индивидов внутри групп и миграции лучших индивидов из каждой группы во все алгоритмы пропорционально доле группы в общем размере популяции.

По результатам исследований [13] выяснено, что коэволюционный алгоритм условной оптимизации не менее эффективен, чем генетический алгоритм с наилучшими настройками. На некоторых задачах эффективность коэволюционного алгоритма значительно превышает эффективность наилучшего генетического алгоритма. Кроме того, на всех тестовых задачах коэволюционный алгоритм превосходит наилучший ГА по скорости сходимости. Таким образом, на задачах условной оптимизации наблюдается значительный положительный эффект от взаимодействия индивидуальных генетических алгоритмов с различными настройками.

В схеме генерирования нечеткого классфикатора коэволюционный алгоритм безусловной оптимизации используется на Мичиганском этапе, коэволюционный алгоритм условной оптимизации – на Питтсбургском этапе.

4. Результаты исследований на практических задачах классификации

Для апробации гибридной схемы генерирования нечеткого классификатора с использованием коэволюционного алгоритма были взяты следующие практические задачи классификации из репозитория UCI [14]:

1. Credit (Australia-1) - задача о выдаче банковского кредита, австралийский вариант, 14 признаков, 2 класса;
2. Credit (Germany) - задача о выдаче банковского кредита, немецкий вариант, 24 признака, 2 класса;
3. Liver Disorder - диагностирование заболевания печени, 6 признаков, 2 класса;
4. Iris - классификация видов ириса, 4 признака, 3 класса;
5. Yeast - классификация типов дрожжей, 8 признаков, 10 классов;
6. Glass Identification - классификация сортов стекла по содержанию химических элементов, 9 признаков, 7 классов;
7. Landsat Images - распознавание типов земель по спутниковым изображениям, 36 признаков, 6 классов.

Детальное описание информативных признаков и классов задач классификации доступны в [14].

Для задачи Landsat Images методом главных компонент размерность решаемой задачи была снижена до четырех признаков [15].

Для всех задач приводятся результаты применения алгоритма генерирования нечеткого классификатора (параметры, надежность классификации после каждого этапа) при варьировании уровня ограничения на число используемых правил. Для первых трех задач классификации приведено сравнение надежности классификации, полученной при использовании сгенерированных баз нечетких правил с эффективностью других алгоритмов классификации. Сравнение приведено с Байесовским подходом, многослойным персептроном [16], бустингом [17], бэггингом [18], методом случайных подпространств (Random Subspace Method, RSM) [19], коэволюционным методом обучения алгоритмических композиций [20].

Из Табл. 1 видно, что нечеткий классификатор по надежности решения задач превосходит многие современные алгоритмы классификации, что доказывает целесообразность использования данного алгоритма.

По данным Табл. 2 можно сделать выводы об особенностях работы каждого этапа гибридной схемы формирования нечеткого классификатора. Видно, что уже после формирования стартовых правил из обучающей выборки мы получаем вполне работоспособную базу. При этом число сгенерированных правил ничтожно по сравнению с общим количеством всех возможных правил. Далее, Мичиганский этап позволяет несколько увеличить надежность классификации по базе нечетких правил. Таким образом, Мичиганский этап необходим, в первую очередь, для сглаживания эффектов случайности при формировании начальной базы нечетких правил.

Питтсбургский этап позволяет сформировать из полученных на предыдущем этапе правил базу, состоящую из существенно меньшего числа правил. Компактные базы правил гораздо предпочтительнее с точки зрения задачи извлечения знаний. Кроме того, во многих случаях

Табл. 1. Сравнение различных алгоритмов классификации на задачах из репозитория UCI по надежности классификации

Алгоритм	Credit (Australia-1)	Credit (Germany)	Liver Disorder
Нечеткий классификатор	**0,891**	**0,794**	**0,725**
Байесовский подход	0,847	0,679	0,629
Многослойный персептрон	0,833	0,716	0,693
Бустинг	0,760	0,700	0,656
Бэггинг	0,847	0,684	0,630
Метод случайных подпространств	0,852	0,677	0,632
Коэволюционный метод обучения алгоритмических композиций	0,866	0,746	0,644

Табл. 2. Результаты исследования гибридной схемы генерирования нечеткого классификатора на задачах из репозитория UCI

Параметр	Задача						
	1	2	3	4	5	6	7
Начальное число правил	200	200	100	50	200	200	100
Средняя надежность классификации начальной базы правил	0,854	0,782	0,595	0,945	0,417	0,640	0,793
Средняя надежность классификации после Мичиганского этапа	0,870	0,791	0,653	0,979	0,495	0,687	0,812
Средняя надежность классификации после Питтсбургского этапа (в скобках – ограничение на число используемых правил)	0,827(10) 0,861(20) 0,873(30)	0,762(50) 0,790(80)	0,666(10) 0,682(15) 0,692(30)	0,908(3) 0,951(4) 0,971(5) 0,975(6)	0,573(20) 0,586(30) 0,593(60)	0,737(20) 0,781(30)	0,838(10) 0,847(15) 0,849(20)
Макисмальная надежность классификации после Питтсбургского этапа (в скобках – ограничение на число используемых правил)	0,870(10) 0,890(20) 0,891(30)	0,767(50) 0,794(80)	0,687(10) 0,710(15) 0,725(20)	0,947(3) 0,973(4) 0,987(5) 0,987(6)	0,598(20) 0,606(30) 0,626(60)	0,757(20) 0,827(30)	0,849(10) 0,857(15) 0,857(20)

надежность классификации после Питтсбургского этапа превосходит надежность классификации, полученной на предыдущем этапе, несмотря на существенное сокращение числа правил. Это означает, что «плохие» правила оказывают негативное влияние на эффективность базы в целом, что подтверждает целесообразность сокращения числа используемых правил даже с точки зрения собственно задачи классификации (не говоря уже о том, что классификация по базе правил меньшего объема требует меньшего числа вычислений).

В Табл. 2 приведены значения надежности классификации при поэтапном увеличении количества используемых правил в случаях, когда такое увеличение приводит к соответствующему увеличению надежности классификации. Дальнейшее увеличение числа используемых правил будет приводить к падению надежности классификации.

Разработанный метод генерирования нечеткого классификатора обладает статистической устойчивостью – при проведении многократных запусков алгоритма наблюдается незначительный разброс в значениях надежности классификации полученных баз нечетких правил, а также генерация одних и тех же или семантически близких нечетких правил.

Заключение

Таким образом, разработан и исследован на практических задачах классификации новый метод комбинирования Мичиганского и Питтсбургского подходов в задаче формирования нечеткого классификатора с использованием коэволюционного алгоритма для адаптации стратегии оптимизации в качестве метода автоматизации выбора настроек эволюционного алгоритма, используемого для генерирования классификатора. Преимущества разработанного метода заключаются в следующем.

1) Используется формализованная процедура отбора стартовых правил с использованием априорной информации из обучающей выборки, существенно повышающая эффективность метода. Данная процедура позволяет сразу же генерировать работоспособный классификатор из незначительного числа правил.

2) Постановка задачи в виде комбинирования безусловной и условной оптимизации без использования многокритериального подхода повышает быстродействие процедуры генерирования нечеткого классификатора и не требует привлечения дополнительных методов для выбора окончательной базы правил ЛПР.

3) Используемый в схеме генерирования классификатора Мичиганский этап позволяет улучшить стартовую базу правил и сглаживать случайные эффекты на этапе отбора стартовых правил.

4) Питтсбургский этап в методе позволяет существенно сократить число используемых нечетких правил. При этом использование бинарных строк фиксированной длины и фиксированных правил (при этом незначительного их числа) существенно уменьшает вычислительные затраты в сравнении с классическим методом использования Питтсбургского подхода

5) Использование на двух основных этапах формирования нечеткого классификатора коэволюционного алгоритма для адаптации поисковой стратегии генетического алгоритма позволяет решить проблему выбора оптимальных настроек ГА.

В ходе проведения апробации разработанной схемы формирования нечеткого классификатора были выявлены следующие особенности метода:

- существует определенный уровень ограничения на число используемых нечетких правил, при котором достигается максимальная надежность классификации; дальнейшее увеличение числа используемых в базе правил будет приводить к уменьшению надежности классификации (эффект влияния «плохих» правил);

- при выборе такого уровня ограничения будет наблюдаться увеличение эффективности классификации на каждом этапе генерирования классификатора;

- наблюдается статистическая устойчивость по надежности классификации и повторяемости генерируемых нечетких правил при многократном запуске алгоритма;

Сравнение нечеткого классификатора с другими методами показывает, что нечеткий классификатор превосходит по эффективности современные алгоритмы классификации. Кроме того, нечеткий классификатор – это также инструмент извлечения знаний. В ходе апробации алгоритма на практических задачах классификации получены компактные базы лингвистических правил, представляющие интерес для специалистов в соответствующих проблемных областях.

Литература

1. Айвазян, С.А. Прикладная статистика. Классификация и снижение размерности / С.А. Айвазян и др. - М.: Финансы и статистика, 1989 .– 607 с.
2. Журавлев, Ю.И. Об алгебраическом подходе к решению задач распознавания или классификации / Ю.И. Журавлев // Проблемы кибернетики. – 1978. - Т.33. – С.5-68.
3. H. Ishibuchi, T. Nakashima, and T. Murata. Performance evaluation of fuzzy classifier systems for multidimensional pattern classification problems, IEEE Trans. on Systems, Man, and Cybernetics, vol.29, pp. 601-618, May 1999.
4. Michalewicz, Z. Genetic Algorithms + Data Structures = Evolution Programs (3rd Edn.), New York: Springer-Verlag, 1996.
5. Жукова, М. Н. Коэволюционный алгоритм решения сложных задач оптимизации: дисс. … канд. техн. наук / М. Н. Жукова – Красноярск: СибГАУ, 2004. – 126 с.
6. O. Cordón, F. Herrera, F. Gomide, F. Hoffmann and L. Magdalena. Ten years of genetic-fuzzy systems: a current framework and new trends, Proc. of Joint 9th IFSA World Congress and 20th NAFIPS International Conference, 2001, Vancouver - Canada, pp. 1241-1246.
7. H. Ishibuchi, T. Nakashima, and T. Murata. Multiobjective optimization in linguistic rule extraction from numerical data. Evolutionary Multi-Criterion Optimization: Springler-Verlag Berlin, 2001, pp. 588-618.
8. H. Ishibuchi, T. Nakashima, and T. Kuroda. A hybrid fuzzy GBML algorithm for designing compact fuzzy rule-based classification systems. Proc. of 9th IEEE International Conference on Fuzzy Systems, 2000, pp. 706-711.
9. Сергиенко, Р.Б. Гибридная схема генерирования нечеткого классификатора с использованием коэволюционного алгоритма / Р. Б. Сергиенко // Теория и практика системного анализа. Труды I Всероссийской научной конференции молодых ученых. – Т.1. – Рыбинск: РГАТА им. П.А. Соловьева, 2010. – С. 11-19.
10. D. Schlierkamp-Voosen and H. Mühlenbein. Strategy adaptation by competing subpopulations. Parallel Problem Solving from Nature III, Springer-Verlag, 1994.
11. Емельянова М.Н. Исследование эффективности коэволюционного алгоритма / М.Н. Емельянова, Е.С. Семенкин // Вестник Сибирского государственного аэрокосмического университета им. академика М.Ф. Решетнева. - № 6. – 2004. - С. 28-34.
12. Сергиенко, Р. Б. Коэволюционный генетический алгоритм решения сложных задач условной оптимизации / Р. Б. Сергиенко, Е. С. Семенкин // Вестник Сибирского государственного аэрокосмического университета им. академика М.Ф. Решетнева. - №2 (23). – 2009. – С. 17-21.
13. Сергиенко, Р. Б. Исследование эффективности коэволюционного генетического алгоритма условной оптимизации / Р. Б. Сергиенко // Вестник Сибирского государственного аэрокосмического университета имени академика М.Ф. Решетнева. - №3 (24). – 2009. – С. 31-36.
14. UCI Machine Learning Repository http://kdd.ics.uci.edu/.
15. Сергиенко, Р. Б. Нечеткий генетический классификатор в задаче распознавания спутниковых изображений / Р. Б. Сергиенко // Информационные технологии и математическое моделирование. Материалы VIII Всероссийской научно-практической конференции в 2-х частях. – Томск: Издательство Томского университета, 2009, Часть 2. – С. 272-276.
16. Тестирование многослойного персептрона http://poligon.machinelearning.ru/Testing/.
17. R. Schapire. The boosting approach to machine learning: An overview. MSRI Workshop on Nonlinear Estimation and Classification, Berkeley, CA, 2001.
18. L. Breiman. Arcing classifiers. The Annals of Statistics, vol. 26, pp. 801-849, Mar. 1998.
19. T. K. Ho. The random subspace method for constructing decision forests. IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 20, pp.832-844, Aug. 1998.
20. Воронцов, К. В. Коэволюционный метод обучения алгоритмических композиций / К. В. Воронцов, Д. Ю. Каневский // Таврический вестник информатики и математики. - №2. – 2005. – С.51-66.

Сергиенко Роман Борисович. Аспирант Сибирского государственного аэрокосмического университета им. ак. М.Ф. Решетнева (СибГАУ, г. Красноярск). Окончил магистратуру СибГАУ в 2010 г. Автор 19 печатных работ. Область научных интересов: интеллектуальные информационные технологии, стохастические методы оптимизации, экспертные системы, интеллектуальный анализ данных. E-mail: romaserg@list.ru.