Скороход Сергей Васильевич
Технологический институт федерального государственного автономного образовательного учреждения высшего профессионального образования «Южный федеральный университет» в г. Таганроге.
E-mail: sss64@mail.ru.
347923, г. Таганрог, ул. Инструментальная, 19/2, кв. 48.
Тел.: 88634648891.

Skorokhod Sergey Vasilievitch
Taganrog Institute of Technology – Federal State-Owned Educational Establishment of Higher Vocational Education "Southern Federal University".
E-mail: sss64@mail.ru.
19/2, ap. 48. Instrumentalynaya street, Taganrog, 347923, Russia.
Phone: +78634648891.

УДК 519.712.2

Л.А. Гладков, Н.В. Гладкова

НОВЫЕ ПОДХОДЫ К ПОСТРОЕНИЮ СИСТЕМ АНАЛИЗА И ИЗВЛЕЧЕНИЯ ЗНАНИЙ НА ОСНОВЕ ГИБРИДНЫХ МЕТОДОВ*

Приводятся основные отличия методов Data Mining от традиционных методов анализа, обсуждаются их преимущества и недостатки и приводятся предложения по их решению на основе использования гибридных интеллектуальных технологий и методов вычислительного интеллекта. Рассмотрены основные аспекты применения нечетких генетических алгоритмов для решения задач извлечения знаний. Описаны основные компоненты организации и процесса взаимодействия генетического алгоритма и нечеткого логического контроллера. Приводятся основные определения и принципы использования методов эволюционного проектирования и моделирования, многоагентных систем и нечетких математических моделей при создании гибридных компонентов интеллектуальных систем. Обсуждаются преимущества и недостатки традиционных методов и приводятся предложения по их решению на основе использования гибридных интеллектуальных технологий и методов вычислительного интеллекта. Приводится обоснование актуальности разработки новых гибридных методов анализа и извлечения данных.

Анализ и извлечение знаний; нечеткий генетический алгоритм; нечеткий логический контроллер; фаззификация; дефаззификация; мультиагентная система; эволюционное проектирование; модели эволюции.

L.A. Gladkov, N.V. Gladkova

NEW APPROACHES TO CONSTRUCTION OF DATA MINING SYSTEMS ON THE BASIS OF HYBRID METHODS

In work the basic differences of methods Data Mining from traditional methods of the analysis are resulted. Also advantages and lacks of methods Data Mining are discussed and offers under their decision on the basis of use of hybrid intellectual technologies and methods of computing intelligence are resulted.Also the basic aspects of application of fuzzy genetic algorithms for the decision of problems of extraction of knowledge are considered. The basic components of the organization and process of interaction of genetic algorithm and the fuzzy logic controller are described.The basic definitions and principles of use of methods of evolutionary designing and modeling, multiagent systems and fuzzy mathematical models at creation of hybrid components of intellectual systems are resulted. Also advantages and lacks of traditional methods are discussed

* Работа выполнена при поддержке: РФФИ (гранты № 08-01-00473), г/б № 2.1.2.1652.

and offers under their decision on the basis of use of hybrid intellectual technologies and methods of computing intelligence are resulted. In the conclusion the substantiation of an urgency of working out of new hybrid methods of the analysis and extraction of data is resulted.

Data mining; fuzzy genetic algorithm; fuzzy logic controller; fuzzification; defuzzification; multiagent systems; evolutionary designing; evolution models.

Введение. Проблема создания эффективных систем интеллектуального анализа и извлечения знаний (Data Mining) из имеющихся массивов данных сегодня чрезвычайно актуальна. Задача состоит в разработке новых эффективных технологий выявления в больших массивах данных неявной и неструктурированной информации, неочевидных, но полезных закономерностей. Понятие «знание» определяется как совокупность фактов, закономерностей и эвристических правил, с помощью которых решается поставленная задача [1]. Отличительными особенностями знаний, как особой понятийной категории являются: структурированность, компактность и внутренняя непротиворечивость. При этом к понятию компактности знаний можно отнести и такое свойство как лаконичность, отсутствие посторонних, не относящихся к изучаемому предмету данных, и удобство доступа и усвоения новых знаний и т.д.

Основными проблемами при построении современных систем анализа и извлечения знаний являются [1]:

- сложность разработки и эксплуатации;
- сложность подготовки данных;
- большой процент недостоверных или бессмысленных решений;
- высокая стоимость.

По мнению различных экспертов из-за существенных различий между инструментами разработчиков программного обеспечения технологии анализа и извлечения знаний перед применением необходимо тщательно изучить на предмет их совместимости и корректности будущих результатов. При этом считается, что результаты применения технологий Data Mining на восемьдесят процентов зависят от уровня подготовки исходных данных, который выполняется до начала работы собственно алгоритма.

Для решения задачи анализа и извлечения знаний, эффективным представляется использование гибридных технологий, основанных на использовании методов вычислительного интеллекта: нейросетевые алгоритмы, нечеткие модели и методы, биоинспирированные алгоритмы, экспертные системы [2]. Эти технологии уже давно и эффективно используются при решении различных задач анализа и принятия решений в условиях нечеткой, плохо формализованной, а зачастую и противоречивой входной информации.

Нечеткие генетические алгоритмы. С концептуальной точки зрения, создаваемые системы интеллектуального анализа и извлечения знаний можно классифицировать как смешанные искусственные системы, т.е. системы созданные человеком и объединяющие искусственные и естественные подсистемы [3]. Они также являются целеориентированными системами, т.е. системами основой функционирования которых является факторы целесообразности [4].

Модель, соответствующая уровню бионических систем может быть представлена следующим образом [5]:

$$SYS = (GN, KD, MB, EV, FC, RP),$$

где GN – генетическое начало (создание стартового множества решений);
KD – условия существования;
MB – обменные явления (эволюционные и генетические операторы);
EV –развитие (стратегия эволюционирования);

FC – функционирование;

RP – репродукция.

В данный момент основными направлениями разработки гибридных методов нечетких генетических алгоритмов (НГА) считаются следующие [6]:

1) Применение механизмов генетических и эволюционных алгоритмов для решения проблем оптимизации и поиска в условиях нечеткой, неопределенной или недостаточной информации об объекте, параметрах и критериях решаемой задачи, совместно с использованием систем, основанных на нечетких правилах (genetic fuzzy rule-based system – GFRBS). При этом полученные гибридные системы используются для обучения и настройки различных компонент системы нечетких правил: автоматической генерации базы знаний GFRBS, ее проверки и настройки выходной функции [7];

2) Использование нечетких инструментов и методов, основанных на нечеткой логике для моделирования различных компонентов и операторов генетических алгоритмов, а также для адаптации и управления основными параметрами генетического алгоритма для динамической настройки и улучшения работы ГА.

Как правило, под нечетким генетическим алгоритмом (НГА) понимают гибридные структуры, относящиеся ко второй области. Таким образом, нечеткий генетический алгоритм можно определить как алгоритм, сочетающий поисковые возможности генетических алгоритмов и возможности математического аппарата нечеткой логики. Нечеткий генетический алгоритм должен обладать следующими свойствами: нечеткое кодирование; нечеткие генетические операторы; нечеткие правила. Рассмотрим подробнее основные направления создания и модификации НГА.

Для контроля и динамического изменения соответствующих параметров генетического алгоритма в систему вводится Нечеткий Логический Контроллер (НЛК), который используя опыт и знание экспертов в рассматриваемой области, соответствующим образом динамически изменяет параметры генетического поиска в ходе выполнения алгоритма для того, чтобы избежать преждевременной сходимости.

НЛК преобразует заданные параметры к нечеткому виду, затем на основе имеющихся в системе знаний и правил определяет управляющее воздействие и возвращает скорректированные значения контрольных параметров [8].

Система выработки правил на основе знаний экспертов и используя рассуждения, делает определенный вывод, который после дефаззификации, превращается из нечеткого правила в воздействие на параметры алгоритма. Изменение параметров алгоритма влечет за собой изменение процесса поиска и текущих результатов, которые затем в блоке фаззификации из переменных состояния преобразуются в нечеткие множества [9].

На выходе блока выработки решения формируется одно или несколько нечетких множеств с соответствующими функциями принадлежности. Соответственно необходимо решить задачу преобразования этих результирующих нечетких множеств (нечеткого множества) в управляющее воздействие на объекты управления НЛК. Такое преобразование называется дефаззификацией (defuzzification).

Важным фактором является зависит выбора управляющих параметров генетического алгоритма. Использование нечетких логических контроллеров, для изменения параметров генетического алгоритма позволяет улучшить работу генетического алгоритма за счет более осторожного, взвешенного и целенаправленного контроля.

Вероятности кроссинговера и мутации могут определяться НЛК исходя из оценки не всей популяции, а по определенной выборке решений учитывающей значения функции пригодности и разнообразие популяции. Также могут одновременно использоваться несколько НЛК [10]. В общем случае, схему взаимодействия блоков нечеткого генетического алгоритма можно представить следующим образом (рис. 1):

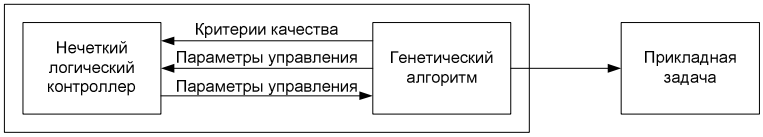

Рис. 1. Схема взаимодействия блоков нечеткого генетического алгоритма

Классическое бинарное представление решений, когда гены принимают значения ноль или единица может быть преобразовано в нечеткое представление, где гены принимают значения в интервале между нолем и единицей. Это позволит выражать более сложные особенности, как генотипа, так и фенотипа различных решений популяции наподобие тех, которые встречаются в природе.

Использование в нечетких генетических алгоритмах нечеткой информации требует использования новых, отличных от классических, ориентированных на битовую строку, генетических операторов [11]. Нечеткие ГО построены с использованием элементов нечеткой логики. Оператор мутации совместно с кроссинговером обеспечивает разнообразие в популяции. Классическая битовая мутация не подходит для НГА. В этом случае может использоваться, например, следующая модификация оператора мутации: $\boldsymbol{x}^* = \boldsymbol{x} \pm \Delta$, где x $\in$ [0; 1], а $\Delta \in$ [0; 0,1]. Нечеткий оператор кроссинговера выполняется на основе логических операций, например, таких как конъюнкция и дизъюнкция. В современной теории нечетких множеств логико-лингвистические связки «И» и «ИЛИ» определяются в виде треугольных норм и конорм. Обобщенные нечеткие операции конъюнкции и дизъюнкции называются треугольными нормами и конормами. Треугольные нормы T и треугольные конормы S – это бинарные операции в теории нечетких множеств, удовлетворяющие условиям ограниченности, монотонности, коммутативности и ассоциативности.

Нечеткие связки и треугольные функции распределения вероятности можно использовать для создания эффективных операторов кроссинговера, которые устанавливают адекватные уровни разнообразия популяции и таким образом позволяют решать проблему преждевременной сходимости.

В литературе описывается большое количество различных операторов кроссинговера для генетических алгоритмов с вещественным кодированием [12].

Применяется также ряд гибридных методик и технологий организации процесса поиска, сочетающих достоинства локальных методов поиска и методов генетического поиска. К ним относятся модели минимального разрыва поколений (Minimal generation gap – MGG) и обобщения поколений (G3-Generalized generation gap). В таких моделях из нескольких родительских решений создаются несколько сотен решений-потомков, из которых отбираются два лучших решения.

Мультиагентные системы. Еще одним перспективным подходом к организации структуры методов Data Mining является использование мультиагентных архитектур.

Под «агентом» может пониматься все, что способно воспринимать свою среду обитания с помощью датчиков (сенсоров) и воздействовать на нее с помо-

щью исполнительных механизмов [13]. Например, программное обеспечение, выступающее в роли агента, в качестве входных данных получает коды нажатия клавиш, содержимое файлов и сетевые пакеты, а его отклик выражается в выводе данных на экран, записи и передаче файлов.

Понятие агента применительно к различным системам может трактоваться по-разному. Многоагентная система может рассматриваться как популяция простых и независимых агентов, каждый агент которой самостоятельно реализуется в локальной среде и взаимодействует с другими агентами. Связи между различными агентами являются горизонтальными, а глобальное поведение агентов определяется на основе расплывчатых правил.

В настоящее время известны различные подходы и методы построения искусственных агентов и многоагентных систем (МАС), в частности, методологии восходящего проектирования на основе ролей агентов и взаимодействий между ними (Gaia, MASE, PASSI, TROPOS и др.), методологии нисходящего проектирования в многомерном пространстве критериев и т.п.

Классическая методология требует построения множества моделей, которые определяют спецификацию многоагентной системы. Каждая модель состоит из компонентов и взаимоотношений между ними. Разрабатываемые модели можно разделить на внешние и внутренние. Внешние модели относятся к системному уровню описания: основными компонентами в них являются сами агенты, взаимодействия между которыми описываются с использованием отношений наследования, агрегации и т.п. Внутренние модели предлагаются для каждого отдельного класса агентов и описывают внутренние структуры агентов: их мнения, цели, планы и т.д. [14].

Обычно выделяются два основных вида внешних моделей: модель агентов и модель взаимодействий, определяющая способы связи (коммуникации) между агентами. Модель агентов разделяется на модель классов агентов и модель экземпляров агентов. Эти две модели определяют классы агентов и их возможные реализации, связанные между собой отношениями наследования, агрегации и др. Классы агентов определяют различные атрибуты агентов, включая атрибуты, задающие мнения, цели и планы агента.

Назначение модели агентов состоит в описании различных типов агентов, существующих в системе. Типы агентов определяются множеством ролей. Поэтому разработчик может предложить объединить несколько сходных ролей в один тип агентов. Главным критерием на этой стадии является эффективность реализации: разработчик прежде всего стремится к оптимизации решений, и объединение нескольких ролей в один тип есть один из способов достижения этой эффективности [14].

Модель взаимодействия агентов включает в себя описание услуг (сервисов), взаимосвязей и обязательств, существующих между агентами. Она состоит из множества протоколов, определяемых для каждого межролевого взаимодействия. Здесь под протоколом мы понимаем схему взаимодействия. Общее понятие протокола включает следующий набор атрибутов:

- назначение: краткое описание смысла взаимодействия (например, «запрос информации», «выдача задания»);
- инициатор: роль, ответственная за инициирование взаимодействия;
- респондент: роль(и), с которой(ыми) осуществляется взаимодействие;
- входы: информация, используемая инициатором для начала взаимодействия;
- выходы: информация, предоставляемая респондентом в ходе взаимодействия.

Предполагается, что реализация протокола будет вызывать серию взаимодействий.

Данная схема определяется формально, абстрагируясь от конкретной схемы реализации (непосредственной последовательности шагов). Подобное рассмотрение взаимодействий означает, что основное внимание уделяется природе и назначению взаимодействия, а не точной схеме обмена сообщениями.

Внутренняя модель, представляющая мнения, цели и планы конкретного класса агентов, является непосредственным расширением объектно-ориентированных моделей (мнения и цели) и динамических моделей (планы) [14].

Кроме указанных видов моделей, могут строиться также:

- модели, описывающие задачи, которые могут выполняться агентами (исходные цели, варианты их декомпозиции, методы решения задач, и пр.);
- модели организации (например, описание сообщества агентов или характеристики организации, куда должна внедряться данная МАС);
- модели коммуникации, которые уточняют характеристики партнерского интерфейса человека с компьютером.

Однако практически все известные методологии требуют предварительного определения функций и типов агентов, а также опираются на достаточно жесткие, заранее заданные протоколы коммуникации. По сути, они не учитывают различные механизмы самоорганизации, эволюции, кооперации агентов в МАС. Поэтому, большую актуальность имеет создание нового класса методологий и методов проектирования агентов и МАС, основанных на использовании бионических принципов, методов и моделей, в частности, идей и технологий эволюционного проектирования. Под эволюционным проектированием (ЭП) искусственной (технической) системы понимается целенаправленное использование компьютерных моделей эволюции на всех стадиях разработки системы. Эволюционное проектирование является подходом, лежащим на границе теории проектирования и теории самоорганизации. Любая самоорганизация предполагает кооперацию агентов в многоагентной системе, она также связана с адаптацией агента к среде и некоторой схемой эволюции. Возможны разные подходы к эволюционному проектированию агентов и МАС, которые могут опираться на различные модели эволюции. Естественным основанием для классификации концепции и стратегий ЭП может служить анализ причин развития агента или МАС: внешних или внутренних [15].

В первом случае эволюционное проектирование МАС рассматривается как процесс ее эволюционной адаптации к внешней среде. Здесь внешняя среда есть причина эволюции разрабатываемой системы и ее важнейшая движущая сила. Тогда главным направлением развития создаваемой МАС полагается ее соответствие текущим условиям среды, которое может достигаться путем прямого приспособления системы к среде.

В частности, отправной момент эволюции МАС может быть связан с наступлением кризисных условий среды. Такие условия нарушают естественное функционирование МАС и ее агентов. В этой ситуации мутация (например, приобретение нового гена) позволяет агенту выжить и адаптироваться к изменившимся условиям. Эта категория мутаций наиболее перспективна и направлена на исправление функциональной недостаточности.

Во втором случае причины изменения МАС усматриваются в ней самой; они могут быть связаны с целеустремленностью агентов, их приспособлением для достижения общей цели и т.п.

Здесь и далее под эволюционным проектированием агента будем понимать процессы формирования его наследственной изменчивости и эволюционной адаптации к внешней среде. Иными словами, ЭП определяется как процесс формиро-

вания и развертывания, как генотипа, так и фенотипа агента. Генотип агента соответствует всей наследственной (генетически обусловленной) информации, которую агент получает от родителей, а фенотип содержит набор структур агента (определяемых ситуативными правилами), которые возникают в результате развития генотипа в определенной среде. При этом часто требуется обрабатывать качественную нечеткую информацию и рассматривать различные стратегии и компьютерные модели эволюции.

Формально проблему эволюционного проектирования (ЭП) искусственных систем можно представить в виде [15]:

$$ED = \langle E, K, O, Q \rangle,$$

где E – множество моделей эволюции; K – множество критериев ЭП; O – множество объектов ЭП; Q – множество процедур ЭП.

Эволюционная теория и эволюционное моделирование вместе с нечеткой логикой позволяют создать алгоритм, который определяет взаимодействия агентов. Эти агенты имеют параметры, определяемые в интервале [0, 1]; поэтому с помощью нечеткой логики мы модифицируем генетические операторы и оператор мутации в алгоритме. В результате алгоритма скрещивания агентов-родителей образуются агенты-потомки, которые в совокупности с агентами-родителями образуют семью (агентство). Для наглядного отображения агентств и общей структуры многоагентной системы используются графы.

Очевидно, что процесс развития любых, в том числе многоагентных систем, складывается как из постепенных изменений, которые могут длиться на протяжении жизни многих поколений, так и резких, быстрых скачков.

Создание общей теории эволюции агентов и МАС предполагает рассмотрение ряда принципиальных проблем, включая:

1) анализ общих причин и движущих сил эволюции агентов в МАС;
2) исследование механизмов развития приспособлений (адаптации) агентов к среде и ее изменениям;
3) определение причин и механизмов возникновения разнообразия типов агентов и агентств;
4) изучение основных методов и средств имитационного моделирования эволюционных процессов.

Существует достаточно много различных схем и моделей эволюционного процесса. Анализ рассмотренных теорий эволюции показывает, что ни одна из них не свободна от недостатков. Все они описывают отдельные аспекты эволюции. Для окончательного выбора общей схемы и модели эволюции применимой к задачам теории агентов необходимо изучить эти проблемы, а также рассмотреть другие современные эволюционные учения.

Заключение. Опыт последних лет показал, что применение в информатике однородных методов, т.е. методов, соответствующих одной научной парадигме, для решения сложных проблем, далеко не всегда приводит к успеху. В гибридной архитектуре, объединяющей несколько парадигм, эффективность одного подхода может компенсировать слабость другого. Комбинируя различные подходы, можно обойти недостатки, присущие каждому из них в отдельности. Поэтому одной из ведущих тенденций, определяющей развитие современной информатики и автоматизированного проектирования стало распространение интегрированных и гибридных систем. Подобные системы состоят из различных элементов (компонентов), объединенных в интересах достижения поставленных целей.

Интеграция различных направлений и методов вычислительного интеллекта и создание на этой основе новых гибридных технологий решения слабоформали-

зованных задач одно из перспективных направлений исследований в области Data Mining. Основой для подобной интеграции является их терпимость к нечеткости и противоречивости используемых данных, гибкость и относительно низкая себестоимость. Активная разработка новых форм и направлений подобной интеграции сейчас активно ведется как России, так и за рубежом.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Чубукова И.А.* Data Mining. Учебное пособие. – М.: Интернет-Университет Информационных Технологий; БИНОМ. Лаборатория знаний, 2006.
2. *Курейчик В.М., Курейчик В.В., Родзин С.И.* Концепция эволюционных вычислений, инспирированных природными системами // Известия ЮФУ. Технические науки. – 2009. – № 4 (93). – С. 16-25.
3. *Перегудов Ф.И., Тарасенко Ф.П.* Введение в системный анализ. – М.: Высшая школа, 1989.
4. *Прангишвили И.В.* Системный подход и общесистемные закономерности. – М.: СИНТЕГ, 2000.
5. *Борисов В.В., Круглов В.В., Федулов А.С.* Нечеткие модели и сети. – М.: Горячая линия – Телеком, 2007.
6. *Гладков Л.А., Гладкова Н.В.* Особенности использования нечетких генетических алгоритмов для решения задач оптимизации и управления // Известия ЮФУ. Технические науки. – 2009. – № 4 (93). – С. 130-136.
7. *Ярушкина Н.Г.* Основы теории нечетких и гибридных систем. – М.: Финансы и статистика, 2004.
8. *Herrera F., Lozano M.* Fuzzy Adaptive Genetic Algorithms: design, taxonomy, and future directions. // Soft Computing 7(2003), Springer-Verlag, 2003. – P. 545-562.
9. *Гладков Л.А., Курейчик В.В., Курейчик В.М.* Биоинспирированные методы оптимизации. – М.: Физматлит, 2009.
10. *Herrera F., Lozano M.* Adaptation of genetic algorithm parameters based on fuzzy logic controllers. In: F. Herrera, J. L. Verdegay (eds.) Genetic Algorithms and Soft Computing, Physica-Verlag, Heidelberg, 1996. – P. 95-124.
11. *Курейчик В.М.* Модифицированные генетические операторы // Известия ЮФУ. Технические науки. – 2009. – № 12 (101). – С. 7-15.
12. *Deb K., Joshi D., Anand A.* Real-Coded Evolutionary Algorithms with Parent-Centric Recombination. Kanpur Genetic Algorithms Laboratory (KanGAL), Kanpur, PIN 208 016, India. KanGAL Report No. 2001003.
13. *Рассел С., Норвиг П.* Искусственный интеллект: современный подход. – М.: Издательский дом «Вильямс», 2006.
14. *Тарасов В.Б.* От многоагентных систем к интеллектуальным организациям. – М.: Эдиториал УРСС, 2002.
15. *Тарасов В.Б., Голубин А.В.* Эволюционное проектирование: на границе между проектированием и самоорганизацией // Известия ТРТУ. – 2006. – № 8 (63). – С. 77-82.

Гладков Леонид Анатольевич
Гладкова Надежда Викторовна
Технологический институт федерального государственного автономного образовательного учреждения высшего профессионального образования «Южный федеральный университет» в г. Таганроге.
E-mail: leo@tsure.ru.
347928, г. Таганрог, пер. Некрасовский, 44.
Тел.: 88634371625.

Gladkov Leonid Anatolievich
Gladkova Nadegda Viktorovna
Taganrog Institute of Technology – Federal State-Owned Educational Establishment of Higher Vocational Education “Southern Federal University”.
E-mail: leo@tsure.ru.
44, Nekrasovskiy, Taganrog, 347928, Russia.
Phone: +78634371625.