УДК 621. 37/39. 061. 2/4

МЕТОДЫ АНАЛИТИЧЕСКОЙ ОБРАБОТКИ ИНФОРМАЦИИ

ГВОЗДИНСКИЙ А.Н., КЛИМКО Е.Г., СОРОКОВОЙ А.И.

Проводится аналитический обзор методов интеллектуального анализа данных (также называют: ИАД, data mining, обнаружение знаний в базах данных) с учетом использования определенного метода для условий Украины. Обзор методов аналитической обработки информации в сложных информационных системах рассматривается с точки зрения скорости извлечения данных, сбора обобщенной информации и повышения достоверности процесса.

Процесс интеллектуального анализа данных — это аналитическое исследование больших объемов информации в целях определения закономерностей и взаимосвязей между переменными, которые можно в дальнейшем применить к новым данным. Полученные сведения преобразуются до уровня информации, которая характеризуется как знание. Этот процесс состоит из трех основных этапов [1]:

– исследование (выявление закономерностей);

– использование выявленных закономерностей для построения модели ;

– анализ исключений для обнаружения и объяснения отклонений в найденных закономерностях.

Нахождение нового знания средствами ИАД – новое и быстро развивающееся направление, использующее методы искусственного интеллекта, математики, статистики. Этот процесс включает в себя следующие шаги [2]:

– определение проблемы (постановка задачи);

– подготовка данных;

– сбор данных: оценка их, объединение и очистка, отбор и преобразование;

– построение модели: оценка и интерпретация, внешняя проверка;

– использование модели;

– наблюдение за моделью.

Построить модель и улучшить ее качество помогает формальная проверка данных с помощью последовательности запросов или предварительного интеллектуального анализа данных. Средства такого анализа включают следующие основные методы: нейронные сети, деревья решений, генетические алгоритмы, а также их комбинации [2-5].

Нейронные сети относят к классу нелинейных адаптивных систем, строением они условно напоминают нервную ткань из нейронов.

Это набор связанных друг с другом узлов, получающих входные данные, осуществляющих их обработку и вырабатывающих на выходе некоторый результат. На узлы нижнего слоя подаются значения входных параметров, на их основе производятся вычисления, необходимые для принятия решений, прогнозирования развития ситуации и т.д. Эти значения рассматривают как сигналы, которые передаются в вышележащий слой, усиливаясь или ослабляясь в зависимости от числовых значений (весов), приписываемых межнейронным связям. На выходе нейрона самого верхнего слоя вырабатывается значение, которое рассматривается как ответ, реакция всей сети на введенные начальные значения. Так как каждый элемент нейронной сети частично изолирован от своих соседей, у таких алгоритмов имеется возможность для распараллеливания вычислений. На рис. 1 показано условное строение нейронной сети.

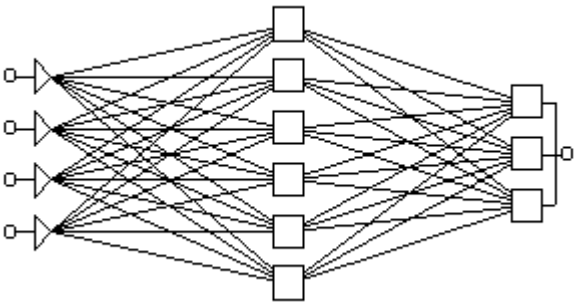

Рис.1. Нейронная сеть

Размер и строение сети должны соответствовать существу исследуемого явления. Построенная сеть подвергается процессу так называемого “обучения”. Нейроны сети обрабатывают входные данные, для которых известны и значения входных параметров, и правильные ответы на них. Обучение состоит в подборе весов межнейронных связей, которые обеспечивают наибольшую близость ответов сети к известным правильным ответам. После обучения на имеющихся данных сеть готова к работе и может быть использована для построения прогнозов поведения объекта в будущем, опираясь на данные его развития в прошлом, производить анализ, выявлять отклонения и сходства. Достоверные прогнозы могут формироваться, не уточняя вид зависимостей, на базе которых он основан.

Нейронные сети используются для решения задач прогнозирования, классификации или управления.

Достоинство – сети могут аппроксимировать любую непрерывную функцию, нет необходимости заранее принимать какие-либо предположения относительно модели. Исследуемые данные могут быть неполными или зашумленными.

Недостаток – необходимость иметь большой объем обучающей выборки. Окончательное решение зависит от начальных установок сети. Данные должны быть обязательно преобразованы к числовому виду. Полученная модель не объясняет обнаруженные знания (так называемый “черный ящик”).

Деревья решений используют разбиение данных на группы на основе значений переменных. В результате получается иерархическая структура операторов “Если... То...”, которая имеет вид дерева. Для классификации объекта или ситуации нужно ответить на вопросы, стоящие в узлах этого дерева, начиная от его корня. Если ответ положительный, переходят к правому узлу следующего уровня, если отрицательный – к левому узлу и т.д. Заканчивая ответы, доходят до одного из конечных узлов, где

указывается, к какому классу надо отнести рассматриваемый объект.

Деревья решений предназначены для решения задач классификации и поэтому весьма ограничено применяются в области финансов и бизнеса.

Достоинство метода – простое и понятное представление признаков для пользователей. В качестве целевой переменной используются как измеряемые, так и не измеряемые признаки – это расширяет область применения метода.

Недостаток – проблема значимости. Данные могут разбиваться на множество частных случаев, возникает "кустистость" дерева, которое не может давать статистически обоснованных ответов. Полезные результаты получают только в случае независимых признаков.

Генетические алгоритмы имитируют процесс естественного отбора в природе. Для решения задачи, более оптимального с точки зрения некоторого критерия, все решения описываются набором чисел или величин нечисловой природы. Поиск оптимального решения похож на эволюцию популяции индивидов, которые представлены их наборами хромосом. В этой эволюции действуют три механизма, представленных на рис. 2.

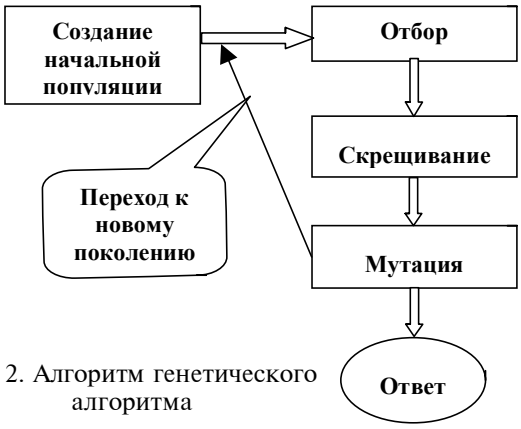

Рис. 2. Алгоритм генетического алгоритма

Можно выделить следующие механизмы:

– отбор сильнейших наборов хромосом, которым соответствуют наиболее оптимальные решения;

– скрещивание – получение новых индивидов при помощи смешивания хромосомных наборов отобранных индивидов;

– мутации – случайные изменения генов у некоторых индивидов популяции.

В результате смены поколений вырабатывается такое решение поставленной задачи, которое уже нельзя дальше улучшить.

Достоинство – метод удобен для решения различных задач комбинаторики и оптимизации, предпочтителен больше как инструмент научного исследования.

Недостаток – возможность эффективно сформулировать задачу, определить критерий отбора хромосом и сама процедура отбора являются эвристическими и под силу только специалисту. Постановка задачи в терминах не дает возможности проанализировать статистическую значимость получаемого с их помощью решения.

Компьютерные технологии интеллектуальной аналитической обработки данных позволяют использовать методы искусственного интеллекта, статистики, теории баз данных и дают возможность создавать современные интеллектуальные системы.

В настоящее время остро стоит вопрос о создании информационных хранилищ (хранилище данных, data warehouse) – оптимально организованных баз данных, которые обеспечивают наиболее быстрый и удобный доступ к информации, необходимой для принятия решений. Хранилище накапливает достоверную информацию из различных источников за большой промежуток времени, которая остается неизменной. Данные объединены и хранятся в соответствии с теми областями, которые они описывают (предметно-ориентированы) и удовлетворяют требованиям всего предприятия (интегрированы).

Учитывая сравнительно небольшой срок существования большинства отечественных предприятий, немногочисленность анализируемых данных, нестабильность предприятий, которые подвержены переменам в связи с изменением законодательной базы, возникает трудность в выработке эффективной стратегии принятия решений с помощью систем интеллектуального анализа данных. Поэтому наиболее приемлемым методом исследования данных в области финансов и бизнеса прогнозируются генетические алгоритмы, а для задач классификации образов и фактов лучше использовать методы деревьев решений или нейронные сети.

Литература: 1. *Щавелев Л.В.* Интеллектуальный анализ данных. http://www.citforum.ru/seminars/cis99/sch_04.shtml, 2. *Буров К.* Обнаружение знаний в хранилищах данных // Открытые системы. 1999. №5-6., http://www.osp.ru/os/1999/05-06/14.htm. 3. *Киселев М., Соломатин Е.* Средства добычи знаний в бизнесе и финансах // Открытые системы. 1997. №4. С. 41-44. 4. *Кречетов Н., Иванов П.* Продукты для интеллектуального анализа данных // Computer Week – Москва. 1997. №14-15. С. 32-39. 5. *Edelstein H.* Интеллектуальные средства анализа и представления данных в информационных хранилищах // Computer Week – Москва. 1996. №16. С. 32-35.

Поступила в редколлегию 22.06.2000

Рецензент: д-р техн. наук, проф. Путятин В.П.

Гвоздинский Анатолий Николаевич, канд. техн. наук, профессор кафедры искусственного интеллекта ХТУРЭ. Научные интересы: оценка эффективности сложных информационных систем управления. Увлечения и хобби: классическая музыка, туризм. Адрес: Украина, 61166, Харьков, ул. акад. Ляпунова, 7, кв. 9, тел. 32-69-08.

Климко Елена Генриховна, ассистент кафедры компьютерных технологий и информационных систем Полтавского государственного технического университета имени Юрия Кондратюка. Аспирантка (без отрыва от производства) кафедры искусственного интеллекта ХТУРЭ. Научные интересы: аналитический анализ данных. Увлечения и хобби: чтение, вязание на спицах. Адрес: Украина, 36021, Полтава, ул. Алмазная, 1-А, кв. 34, тел. (053-22) 3-43-12.

Сороковой Александр Иванович, канд. техн. наук, доцент кафедры компьютерных технологий и информационных систем Полтавского государственного технического университета имени Юрия Кондратюка. Научные интересы: KDD (обнаружение знаний). Увлечения и хобби: собаки. Адрес: Украина, 36022, Полтава, пер. Ломаный, 37А , тел.(053-2) 18-60-87, e-mail: kdd_s@mail.ru