УДК 519.23

ИСПОЛЬЗОВАНИЕ МЕТОДОВ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА В ЗАДАЧАХ БИНАРНОЙ КЛАССИФИКАЦИИ

© 2014 В.А. Алексеева

Ульяновский государственный технический университет

Поступила в редакцию 22.06.2014

Рассматривается задача бинарной классификации объектов. Для решения предлагается использование методов интеллектуального анализа, таких, как деревья решений, нейронные сети, нечеткая логика, метод ближайших соседей. Проводится сравнительный анализ эффективности рассматриваемых методов при бинарной классификации объектов.

Ключевые слова: *бинарная классификация, интеллектуальный анализ, нейронные сети, деревья решения, нечеткие методы, метод ближайших соседей*

Рассмотрим задачу бинарной классификации объектов, в которой каждый объект $K_i(i=1,\ldots,N)$ характеризуется m-мерным вектором признаков $(X_1 \ldots X_m)$. Эти факторы могут принимать как числовые, так и нечисловые значения и образуют обучающую выборку для дальнейших исследований. Необходимо на основании значений признаков предсказать выходную характеристику объектов, принимающую два значения (0 и 1). В качестве примера может служить задача обнаружения сигналов. В бинарных задачах на интервале наблюдения может передаваться один из двух сигналов. Частным случаем бинарной задачи является обнаружение факта передачи или отсутствия единственного сигнала. При решении задач классификации широкое распространение получили методы интеллектуального анализа. К ним можно отнести:

- деревья принятия решений;
- нейронные сети;
- генетические алгоритмы;
- методы нечеткой логики;
- метод ближайших соседей;
- логико-вероятностные подходы.

Деревья принятия решений. Цель метода состоит в том, чтобы создать модель, которая предсказывает значения целевой бинарной переменной на основе нескольких переменных на входе. На основе данных за прошлые периоды строится дерево [1]. Структура дерева представляет собой следующее: «листья» и «ветки». На ребрах («ветках») дерева решения записаны атрибуты, от которых зависит целевая функция, в «листьях» записаны значения целевой функции, а в остальных узлах — атрибуты, по которым различаются случаи. При этом класс каждой из ситуаций, на основе которых строится дерево, заранее известен. При построении дерева все известные ситуации обучающей выборки сначала попадают в верхний узел, а потом распределяются по узлам, которые в свою очередь также могут быть разбиты на дочерние узлы. В качестве критерия разбиения можно использовать различные значения одного из входных факторов. Выбирается тот узел, по которому устраняется больше неопределенности.

Метод дерева принятия решений имеет несколько достоинств: прост в понимании и интерпретации, не требует подготовки данных, способен работать с любым типом переменных, позволяет оценить модель (и тем самым ее надежность) при помощи статистических тестов и т.д. Имеется также ряд недостатков: алгоритмы построения деревьев не могут обеспечить оптимальность всего дерева в целом, существует проблема «переобучения» и т.д.

Нейронные сети. Нейронные сети широко используются в задачах классификации. Построение нейросетевой модели [6] включает в себя следующие этапы.

1) Постановка задачи. На этом этапе алгоритма определяются цели моделирования, устанавливаются входные и выходные параметры модели, устанавливается структура (состав и длина) входного вектора X и выходного вектора Y. Входным вектором будем считать вектор факторов. Выходным параметром модели является бинарная номинальная переменная Y, соответствующая двум классам.

2) Формирование данных. Формируется содержимое входных и выходных векторов. Создается множество обучающих пар из имеющейся информации(Xq;Yq), где q = 1, ..., N, N – количество исследуемых объектов. Все множество данных разбивают на две подвыборки: обучающую и тестирующую. На обучающей выборке производится обучение сети, на тестирующей – проверка построенной нейросетевой модели.

Алексеева Венера Арифзяновна, кандидат технических наук, доцент кафедры «Прикладная математика и информатика». E-mail: v.fashutdinova@mail.ru

3) Первоначальное проектирование сети. Нейронные сети могут быть однослойными и многослойными. Многослойными персептронами называют нейронные сети прямого распространения: входной сигнал распространяется в прямом направлении, от слоя к слою. Рассмотрим трехслойный персептрон, состоящий из входного, выходного и одного скрытого слоя. Алгебраически трехслойный персептрон можно задать следующим образом:

$$z_k = F_1 \sum_{q=0}^{p} w_{kq} x_q,$$

где F_1 - активационная функция для первого скрытого слоя, w_{kq} - синаптический вес, связывающий k-ый нейрон и q-ый входной сигнал, z_k, где k=1,…,r - это выходы из первого скрытого слоя. Поскольку выход одного слоя – это вход в следующий слой, мы можем записать конечный результат в виде:

$$y_v = F_2\left(\sum_{k=1}^{r} K_{vk} z_k\right) = F_2\left(\sum_{k=1}^{r} K_{vk}\left(F_1\left(\sum_{q=0}^{p} w_{kq} x_q\right)\right)\right),$$

где y_v - выход нейрона v из выходного слоя, и K_{vk} - это синаптический вес, который объединяет нейрон k в скрытом слое и нейрон v в выходном слое, F_2 - активационная функция для выходного слоя.

4) Обучение сети. Цель обучения – подобрать синаптические веса так, чтобы на каждый входной вектор Xq (q=1,…,N) множества обучающих примеров сеть выдавала вектор, минимально отличающийся от заданного выходного вектора Yq. Эта цель достигается путем использования алгоритмов обучения нейронной сети. Одним из самых распространенных алгоритмов обучения нейросетей является алгоритм обратного распространения ошибки [6].

5) Проверка и оптимизация сети. Проверка обобщающих свойств сети производится на контрольной выборке. Если нейросеть показывает улучшение аппроксимации и на обучающей, и на тестирующей выборках, то обучение сети происходит в правильном направлении. Иначе может снижаться ошибка на обучающей выборке, но происходить ее увеличение на тестирующей, что означает «переобучение» сети. В этом случае сеть не может быть использована для прогнозирования или классификации. В этом случае немного изменяются веса нейронов, чтобы вывести сеть из окрестности локального минимума ошибки. Если погрешность сети окажется неприемлемо большой, то надо попытаться оптимизировать сеть. Оптимизация сети состоит в подборе наиболее подходящей для данной задачи структуры сети – количества скрытых слоев, количества скрытых нейронов, количества синаптических связей, вида и параметров активационных функций нейронов. Результатом оптимизации и проверки сети является готовая к использованию нейросетевая математическая модель бинарной классификации.

6) Исследование предметной области. Основное достоинство этого метода заключается в способности алгоритма подстраивать структуру сети под новые наблюдения и объяснять довольно сложные связи между значениями факторов и выходной характеристикой.

Нечеткая логика. В настоящее время нечеткая логика является распространенным инструментом решения задач классификации, управления и принятия решений [5]. Главные достоинства этого метода:

- возможность отказа от сложных систем управления, где это позволяет требуемая точность вычислений;
- описание процесса принятия решений на естественном языке, с использованием привычных для человека качественных оценок, и привязка этих оценок к строгому математическому аппарату.

Исходная задача классификации задается с помощью нечеткой базы знаний, в которой характеристики процесса представлены в виде лингвистических переменных с заданными функциями принадлежности. Нечеткая база знаний формируется экспертом, имеет структуру: *если … и(или) … , то …*. Задача решается с использованием нечеткого логического вывода Мамдани [7].

Метод ближайших соседей. Основным принципом метода ближайших соседей является то, что объект присваивается тому классу, который является наиболее распространённым среди соседей данного элемента. Это пример подхода «ленивого обучения», когда обучение сводится к добавлению новых случаев в базу данных. Пусть имеется n наблюдений, каждому из которых соответствует запись в таблице, включающая в себя значения факторов и класса объекта (0 или 1). Прогнозирование в данном методе заключается в определении класса для новой записи по известным значениям факторов. Сначала задается новый объект с рядом признаков по каждой характеристике. Из всей исходной выборки находятся k записей с минимальным расстоянием до вектора признаков нового объекта (поиск соседей) [2]. Объект классифицируется в зависимости от того, к какому классу относится большинство его соседей.

Логико-вероятностный подход. Метод основан на методах математической логики и теории вероятностей [3]. Рассмотрим этапы построения логико-вероятностной модели:

1) структурно-логическая постановка задачи: разделение всей рассматриваемой системы на конечное число элементов, каждый из которых представляется в модели событием с двумя

возможными состояниям и заданными вероятностными параметрами; определение условий реализации и/или не реализации выходных функций для каждого элемента в системе; описание и задание с помощью отдельных или групповых выходных функций логических критериев функционирования системы;

2) логическое моделирование: с помощью специальных методов преобразования осуществляется построение логической функции работоспособности системы;

3) вероятностное моделирование: с помощью преобразования функции работоспособности системы осуществляется построение многочлена расчетной вероятностной функции.

Для исследования применимости описанных выше интеллектуальных методов анализ данных рассмотрим задачу бинарной классификации, представляющую собой выборку, состоящую из 1000 объектов, каждый из которых характеризуется 11 признаками. Исходная выборка была разделена случайным образом на обучающую выборку (700 объектов) и на тестовую выборку (300 объектов). Результаты применения методов представим в виде таблицы. Сравним методы по двум показателям:

- процент совпадений (доля верных предсказанных значений выходной бинарной характеристики);

- показатель AUC – площадь, ограниченная ROC-кривой и осью доли ложных положительных классификаций. ROC-кривая – график, позволяющий оценить качество бинарной классификации, отображает соотношение между долей верных положительных классификаций от общего числа положительных классификаций с долей ошибочных положительных классификаций от общего числа отрицательных классификаций. Чем выше показатель AUC, тем качественнее классификатор.

Результаты показывают, что наилучшим образом данную задачу бинарной классификации объектов решает метод нейронных сетей (показатель AUC не ниже 80%).

Таблица 1. Результаты применения методов интеллектуального анализа в задаче бинарной классификации

Метод бинарной классификации	**Процент совпадений**	**AUC**
деревья принятия решений;	75,2 %	0,74
нейронные сети;	83,2%	0,82
методы нечеткой логики;	62,9%	0,67
метод ближайших соседей;	46,7%	0,56
логико-вероятностный подход.	85,6%	0,78

Работа выполнена в рамках задания Минобрнауки России №2014/232.

СПИСОК ЛИТЕРАТУРЫ:

1. *Quinlan, J.R.* C4.5: Programs for Machine learning. Morgan Kaufmann Publishers 1993. 324 p.
2. Алгоритм ближайшего соседа [Электронный ресурс]. – Режим доступа: http://www.basegroup.ru/library/ analysis/regression/knn/, свободный.
3. *Алексеев, В.В.* Логико-вероятностное моделирование риска портфеля ценных бумаг / *В.В. Алексеев, Е.Д. Соложенцев* // Информационно-управляющие системы. 2007. № 6. С. 49-56.
4. *Клячкин, В.Н.* Сравнительный анализ точности нелинейных моделей при прогнозировании состояния системы на основе марковской цепи / *В.Н. Клячкин, Ю.С. Донцова* // Известия Самарского научного центра Российской академии наук. 2013. Т. 15, № 4(4). С. 924-927.
5. *Круглов, В.В.* Интеллектуальные информационные системы: компьютерная поддержка систем нечеткой логики и нечеткого вывода / В.В. Круглов, М.И. Дли. – М.: Физматлит, 2002. 256 с.
6. *Ясницкий, Л.Н.* Введение в искусственный интеллект. – М.: Издательский центр «Академия», 2005. 176 с.
7. *Штовба, С.Д.* Проектирование нечетких систем средствами MATLAB. – М: Горячая линия-Телеком, 2007. 288 с.

USING OF MINING TECHNIQUES IN PROBLEMS OF BINARY CLASSIFICATION

© 2014 V.A. Alekseeva

Ulyanovsk State Technical University

The problem of objects binary classification is considered. To solve this problem we suggest the use of mining techniques, such as decision trees, neural networks, fuzzy logic, the method of nearest neighbors. It is performed a comparative analysis of the effectiveness of these methods in the binary classification of objects.

Key words: *binary classification, mining, neural networks, decision trees, fuzzy methods, the method of the nearest neighbors*

Venera Alekseeva, Candidate of Technical Sciences, Associate Professor at the Department «Applied Mathematics and Computer Science». E-mail: v.fashutdinova@mail.ru