Алгоритм обобщения, работающий с зашумлёнными данными

Куликов А. В., Фомина М. В.
m_fomina2000@mail.ru
Москва, МЭИ (ТУ)

Обнаружение знаний в базах данных является стремительно увеличивающейся областью, развитие которой вызвано большим интересом к настоятельным практическим, социальным и экономическим нуждам. Современные базы данных содержат так много данных, что практически невозможно вручную проанализировать их для извлечения ценной информации, помогающей принимать важные решения. Отсюда следует, что люди нуждаются в помощи интеллектуальных систем для повышения своих аналитических возможностей.

Шум в обучающей выборке

Индуктивные экспертные системы, обрабатывающие реальные массивы данных, обычно работают в условиях наличия шума во входных данных. Шум возникает из-за таких причин, как, например, некорректное измерение входного параметра, неверное описание значения параметра экспертом, использование испорченных измерительных приборов, потеря данных при пересылке и хранении информации.

Шум вызывает две проблемы: сначала при построении обобщённых правил, а затем при классификации объектов с использованием этих правил.

Мы исследуем две модели шума:

1. Шум связан с исчезновением значений атрибутов.
2. Шум связан с искажением некоторых значений атрибутов в обучающей выборке. При этом истинное значение заменяется на одно из допустимых, но ошибочных значений (значения перемешаны).

Зашумлённые обучающие выборки должны обрабатываться алгоритмом обобщения в соответствии с процедурой «обучения с учителем» Бонгарда [1].

Предсказание неизвестных значений методом ближайшего соседа

Пусть дана выборка с шумом, K', причём искажениям подвергаются атрибуты, принимающие как дискретные, так и непрерывные значения. Рассмотрим проблему использования объектов обучающей выборки K' при построении решающего дерева T и при проведении экзамена с использованием решающего дерева T.

Пусть $X \in K'$ — очередной объект выборки; $X = (x_1, \ldots, x_n)$. Среди всех значений его атрибутов имеются атрибуты со значением N

(Not known). Это могут быть как дискретные, так и непрерывные атрибуты [2].

Наличие неизвестных значений в примерах обучающей выборки затрудняет как обучение, так и экзамен, поскольку часть примеров может быть отвергнута, либо при классификации получен неоднозначный результат. Предлагается восстановить эти неизвестные значения, используя аналог метода «ближайшего соседа» [3, 4].

Основная идея алгоритма в следующем. Если пример X обучающей выборки K' содержит неизвестные значения, определяем на основе введенной в [3, 4] метрики p ближайших к нему примеров, не имеющих неизвестных значений. На основе анализа этих примеров, имеющих максимальное сходство с X, восстанавливаем значения признаков этого объекта.

Рассмотрим стратегию определения неизвестного значения.

1. Признак — количественный. Определяем неизвестное значение как среднее арифметическое значений его ближайших p примеров, для которых определены значения признака.
2. Признак — качественный. Определить неизвестное значение признака как наиболее часто встречающееся среди ближайших p примеров.

Использование процедуры восстановления при построении дерева решений

Рассмотрим возможность использования процедуры восстановления для решения задач индуктивного формирования понятий. Предлагается алгоритм IDTUV (Induction of Decision Tree with restoring Unknown Values), который включает процедуру восстановления неизвестных значений при наличии в обучающей выборке примеров, содержащих шум. Когда неизвестные значения атрибутов восстановлены, используется один из алгоритмов построения деревьев решений. Примеры, для которых не удалось восстановить неизвестные значения, удаляются из обучающей выборки.

Ниже приводится псевдокод алгоритма IDTUV.

Результаты классификации примеров с шумом

Был проведен ряд экспериментов на следующих четырех группах данных из известной коллекции тестовых наборов данных Machine Learning Repository кафедры информатики и вычислительной техники Калифорнийского университета UCI.

Поскольку главной задачей эксперимента было оценить влияние шума на результаты построения классификационных правил и распознавания тестовых примеров, было принято решение отказаться от хаотичного

Алгоритм 1. IDTUV

Вход: $K = K^+ \cup K^-$;
Выход: дерево решений T;
получение $K = K^+ \cup K^-$;
для всех информативных атрибутов K
  пока имеется неизвестное значение атрибута
    применить алгоритм ВОССТАНОВЛЕНИЕ;
если информативные атрибуты имеют непрерывные значения **то**
  применить алгоритм C4.5;
иначе
  применить алгоритм ID3;
вернуть T — дерево решений.

внесения шума в поля любых признаков в тестовых таблицах. Для внесения искажений был использован наиболее информативный признак таблицы, который размещается в корне дерева решений. Были рассмотрены и проанализированы ситуации полного отсутствия шума и наличия шума в 5%, 10% и 20% по выбранному признаку. На каждом тестовом множестве проводился ряд экспериментов. Затем результаты усреднялись.

Влияние шума на построение дерева решений. Первая группа опытов предназначалась для проверки того, как шум в обучающей выборке влияет на построение дерева решений. Опыты проводились по следующей схеме. В обучающую выборку вносится некоторое количество неизвестных значений. Затем производится восстановление таких значений по методу ближайшего соседа. На полученной обучающей выборке вновь строится дерево решений, которое необходимо сравнить с деревом, построенным на основе выборки без шума.

Полученные результаты показали, что в некоторых случаях, даже при отсутствии до 20% значений наиболее информативного признака, не происходило изменения дерева решений, и, следовательно, не изменялись правила классификации (либо эти изменения были крайне незначительны). Это свидетельствует о высокой эффективности метода восстановления.

Влияние шума на классификацию примеров. Второй этап проверки заключался в установлении того, как шум влияет на успешность классификации примеров. Были использованы две модели шума: шум, как отсутствие значений, и шум, заключающийся в перепутывании определенного количества значений в тестовой выборке. В обоих случаях для классификации использовались правила, полученные на обучающей выборке, не содержащей шума. Полученные результаты показали, что алго-

ритм IDTUV в сочетании с алгоритмами восстановления позволяет повысить точность классификации примеров с отсутствующими значениями признаков в 3–4 раза по сравнению с классическими алгоритмами ID3 и C4.5 [5, 6]. При использовании шума типа «перемешивание значений» алгоритм IDTUV повышает точность классификации примеров в 2 раза. Из проведенных опытов можно сделать вывод, что алгоритм IDTUV способен успешно работать с зашумлённой информацией.

Работа выполнена при поддержке РФФИ, проект № 05-01-00818.

Литература

[1] *Бонгард М. М.* Проблема узнавания. — М.: Наука, 1967. — 320 с.

[2] *Вагин В. Н., Куликов А. В., Фомина М. В.* Методы теории приближенных множеств в решении задачи обобщения понятий // Известия РАН. Теория и системы управления. — 2004. — № 6. — с. 52–66.

[3] *Бериша А. М., Вагин В. Н.* Использование алгоритма построения деревьев решений для зашумлённых данных // Международный форум информатизации — 2004: Труды международной конференции «Информационные средства и технологии». Т. 1. — М.: Янус-К, 2004. — с. 171–174.

[4] *Бериша А. М., Вагин В. Н., Куликов А. В., Фомина М. В.* Методы обнаружения знаний в «зашумленных» базах данных // Известия РАН. Теория и системы управления. — 2005. — № 6. — с. 143–158.

[5] *Quinlan J. R.* C4.5: Programs for Machine Learning. — San Mateo, CA: Morgan Kaufmann Publishers, 1993.

[6] *Quinlan J. R.* Induction of Decision Trees // Machine Learning, 1986. — № 1. — Pp. 1–81.