УДК 004.93`1 + 004.896

АЛГОРИТМ КЛАССИФИКАЦИИ ДЛЯ ЗАДАЧ С НЕТОЧНЫМИ, ПРОТИВОРЕЧИВЫМИ ДАННЫМИ

Максимова А.Ю., Козловский В.А.
Институт прикладной математики и механики НАН Украины, г. Донецк, Украина

Рассматриваются задачи классификации, для которых классы образов обладают априорной неразделимостью. Результатомклассификациявляетсянечеткоемножество, описывающее степень принадлежности объекта каждому из классов. Для решения задачи предложен алгоритм нечеткого вывода. Рассмотрены методы построения базы правил. Для оценки качества алгоритма используются функционалы скользящего контроля.

Введение

При решении прикладных задач, связанных с обработкой баз данных исследуемых объектов часто возникает задача классификации. Множество исходных данных называют выборкой прецедентов или обучающей выборкой. Выбор типа классификатора и метода его построения происходит по результатам анализа этих данных. Разбиение на классы выполняется либо экспертом, либо методами кластер-анализа. Возможность получения хорошего классификатора зависит от свойств множества классов образов в рамках рассматриваемого признакового пространства. Хорошо изучен случай попарно линейно разделимых классов. Существует большой набор методов построения линейных классификаторов, позволяющих строить классификаторы со стопроцентным результатом распознавания. В ситуациях, когда классы образов разделимы, но имеют нелинейные, сложные границы разделения, тоже удается получать классификаторы хорошего качества.

Сложной является ситуация с пересекающимися классами,

т.к. в этих случаях всегда присутствует ошибка классификации. В работе рассматривается построение нечеткого классификатора, способного давать адекватные ответы для таких множеств разбиений классов. При отсутствии межклассовой разделимости предлагается вычислять степени соответствия образа каждому из классов. Для формализации такого подхода удобно использовать аппарат нечеткой математики. Предлагается использовать алгоритм нечеткого вывода, база правил которого формируется в результате анализа прецедентов.

Постановка задачи

Рассматривается классическая постановка задачи распознавания образов. Образы представляются векторами $x^{(i)} \in X \subset R^n$ в признаковом пространстве. Задано множество классов образов $V = \{v_i\}, i = 1,...,K$ и обучающая выборка Y как множество пар $Y = \{(x^{(i)}, v^{(i)}), x^{(i)} \in X, v^{(i)} \in V, i = 1,...,N\}$.

Необходимо построить алгоритм распознавания, который позволяет по предъявленному образу отнести его к заданному классу с определенной степенью уверенности. Результат работы алгоритма представляется в виде нечеткого множества $\tilde{y}(\bar{x}) = \sum_{i=1}^{k} \mu_i(\bar{x}) / v_i$, где $\mu_i(\bar{x})$ – степень принадлежности образа $\bar{x}$ классу v_i [1].

Известный набор признаков в общем случае не всегда обеспечивает полную разделимость всех классов: $\exists(\bar{x}_1, v_1) \in Y, (\bar{x}_2, v_2) \in Y : \bar{x}_1 = \bar{x}_2, v_1 \neq v_2$.

Основная идея построения нечеткого классификатора

Для решения данной задачи предложен упрощенный алгоритм нечеткого вывода. Исходная версия алгоритма изложена в [2]. База правил строиться по так называемым нечетким портретам рассматриваемых классов образов. Выполняется сопоставление исследуемого образа $\bar{x}$ с нечеткими портретами, в результате чего

вычисляется степень соответствия образа каждому из классов.

Качество полученного классификатора зависит от свойств нечетких портретов. Они должны обладать обобщающей способностью и обеспечивать хорошую разделимость классов образов.

В основу формирования нечетких портретов положен частотный анализ множества значений по каждому из признаков. Нечеткие портреты формально представляются лингвистическими переменными [1]. С помощью лингвистических переменных удобно записывать нечеткие предикаты базы правил для системы нечеткого вывода.

В системах нечеткого вывода для построения функций принадлежности лингвистических переменных используются несколько подходов [3]. Существуют методы построения нечетких множеств по экспертным оценкам, которые делятся на прямые и косвенные. Второй подход позволяет строить терм-множества без участия экспертов на основе анализа исходных данных. В работе предложен подход, позволяющий строить функции принадлежности без участия эксперта. В основе данного подхода лежит анализ частотных характеристик встречаемости признаков.

Рассмотрим построение функции принадлежности для множества D значений одного признака для одного класса. На первом этапе используется концепция скользящего окна для построения частотной характеристики D (рис. 1). В предельном случае такой характеристикой будет полигон частот. Вид функции зависит от выбора параметров α,β,γ, определяющих ширину скользящего окна $wind = \gamma h_{cp}$ и шаг скольжения $step = \beta h_{cp}$, где $h_{cp} = \frac{max(D) - min(D)}{|D| - 1}$ – среднее расстояние между точками множества D, $|D|$ – мощность множества, $\alpha = \frac{step}{wind}$.

На рисунке 2 представлен пример функции $\mu(\bar{x})$ для задачи классификации вин из хранилища данных UCI [4]. Полученная функция задана набором значений.

На следующем этапе функции принадлежности

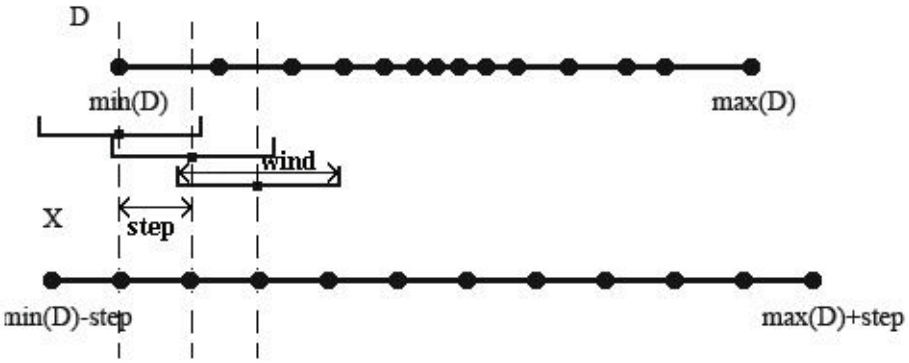

Рисунок 1 – Формирование функций принадлежности методом скользящего окна

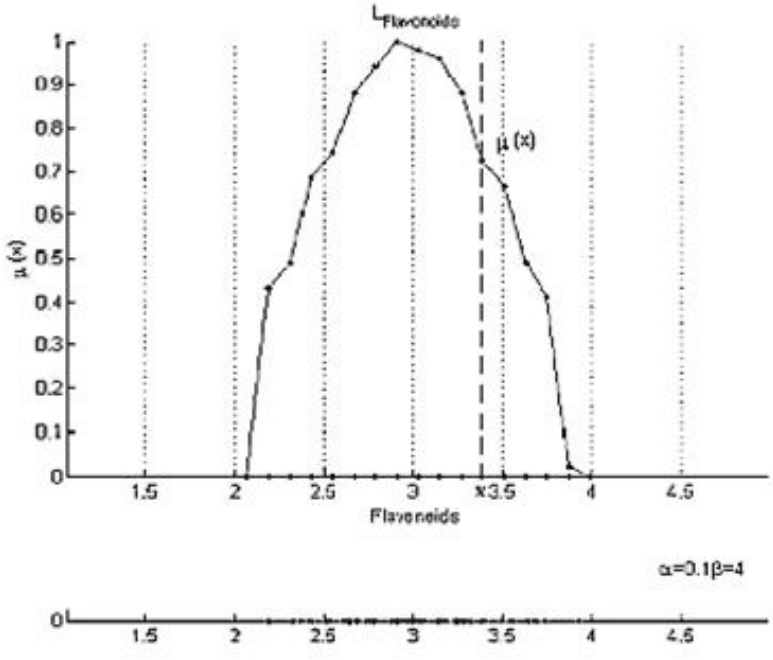

Рисунок 2 – Функция принадлежности для признака «флавоноиды» в задаче классификации вин

аппроксимируются полиномиальными функциями или комбинацией экспонент, что позволяет уменьшить объем хранимой информации.

Контроль качества и оценка обобщающей способности алгоритма выполняется методом скользящего контроля [5]. Данный подход стандартизирует функционалы качества и дает возможность сравнивать качество работы предложенного алгоритма с качеством работы других алгоритмов на конкретной задаче. Также возможно оценить работу предложенного алгоритма на нескольких задачах, для которых выборки прецедентов обладают разными свойствами и ограничениями.

4

Заключение

Предложенный в работе подход формирования классификатора позволяет решать задачу распознавания образов для задач с неточными, противоречивыми данными. В областях признакового пространства, где присутствуют противоречия, алгоритм дает дополнительную информацию по спорным образам в виде степеней принадлежности точки признакового пространства каждому из пересекающихся классов. Результат работы алгоритма представлен в виде нечеткого множества и может быть использован в таком виде на следующем этапе решения общей прикладной задачи.

Предложенный подход протестирован на задаче классификации вин из хранилища данных UCI, а также на практической задаче распознавания видов топлива от различных производителей.

Литература

[1] Заде Л. Понятие лингвистической переменной и его применение к принятию приближенных решений / Л. Заде; [под ред. Н.Н. Моисеева и С.А. Орловского]. – М.: Мир, 1976. – 168 с.

[2] Козловский В.А. Решение задачи распознавания образов по нечетким портретам классов / В.А. Козловский,

А.Ю. Максимова // Искусственный интеллект. – 2010. – № 4. – С. 221-228.

[3] Аверкин А.Н. Нечеткие множества в моделях управления и искусственного интеллекта / А.Н. Аверкин, И.З. Батыршин, А.Ф. Блишун и др.; [под ред. Д.А. Поспелова]. – М.: Наука. Гл. ред. физ. мат. лит., 1986. – 312 с.

[4] Asuncion A., Newman D.J. UCI machine learning repository. Irvine (USA): University of California, School of Information and Computer Science, 2007. – Режим доступа: http://www.ics.uci.edu/~mlearn/MLRepository.html

[5] Воронцов К.В. Комбинаторный подход к оценке качества обучаемых алгоритмов / К.В. Воронцов // Математические вопросы кибернетики. Вып. 13 : сборник статей / [под ред. О.Б. Лупанова]. – М. : Физматлит, 2004. – С. 5-36.