Классификация данных на основе горной кластеризации

АРИМПИЛОВ С.Н., ВОРОЖЦОВ С.А.

Проект ГПО-1604 Нечеткие аппроксиматоры, руководитель Сарин К.С.

Аннотация. В докладе предложен алгоритм классификации на основе горной кластеризации, приведены сравнения точности и времени выполнения классификации с общеизвестным алгоритмом *k*-ближайших соседей.

Введение

Проблема классификации объектов возникает в различных сферах человеческой деятельности. Задачи классификации являются предметом исследования многих дисциплин. Решением проблем классификации на основании прецедентов занимаются машинное обучение и интеллектуальный анализ данных. Существует множество инструментов, позволяющих строить классификаторы на основе экспериментальных данных, например алгоритм *k*-ближайших соседей, деревья решений [1], нечеткие классификаторы [2]. Важным свойством классификации является точность, которая обычно выражается процентом правильной классификации:

$$E = \frac{\sum_i^m f_i}{m} * 100\%,$$

где $f_i = \begin{cases} 1, \text{ если для объекта } \mathbf{x_i} \text{ был правильно определен класс} \\ 0, \text{ если для объекта } \mathbf{x_i} \text{ был НЕ правильно определен класс} \end{cases}$, m – количество всех экспериментальных данных.

В данной работе предлагается алгоритм классификации по прецедентам, основанный на алгоритме горной кластеризации.

Алгоритм горной кластеризации

Идея алгоритма основана на визуальном представлении человека процесса формирования кластера [3]. Особенностью данного алгоритма является то, что количество кластеров определяется самим алгоритмом, центры кластеров здесь выбираются из множества экспериментальных данных. Для описания алгоритма будем использовать следующее обозначение: d_{min} – минимальное расстояние между кандидатом и существующими центрами кластеров, c_1 – номер экземпляра данных, который оказался первым кластером, $d_{i,j}$ – расстояние между *i*-м и *j*-м экземпляром данных.

Ниже представлен алгоритм горной кластеризации в виде последовательности шагов.

Начало

Шаг 1. Рассчитать потенциал каждой точки $\mathbf{x}_i$ экспериментальных данных $P_i = \sum_{j=1}^{m} e^{-\frac{4 \cdot d_{i,j}^{\ 2}}{r_a^{\ 2}}}$.

Шаг 2. Найти точку с максимальным потенциалом $new = \arg\max_{i=1}^{m} P_i$.

ЕСЛИ $P_{new} > \bar{\varepsilon} \cdot P_{c_1}$, ТО *new* новый центр кластера,

ИНАЧЕ ЕСЛИ $P_{new} < \underline{\varepsilon} \cdot P_{c_1}$, ТО завершить работу алгоритма,

ИНАЧЕ ЕСЛИ $\frac{d_{\min}}{r_a} + \frac{P_{new}}{P_{c_1}} \geq 1$ ТО *new* новый центр кластера,

ИНАЧЕ $P_{new} = 0$ и перейти на начало шага 2.

Шаг 3. Выполнить перерасчет потенциалов всех точек данных $P_i = P_i - P_{new} \cdot e^{-\frac{4 \cdot d_{i,new}^{\ 2}}{r_b^{\ 2}}}$, перейти на шаг 2.

Конец

В алгоритме используются следующие параметры: $\underline{\varepsilon}$ и $\bar{\varepsilon}$ определяют границы «серой зоны», в которой проходит проверка кандидата в кластеры на выполнения компромисса

между достаточным значением потенциала и достаточной удаленность от существующих центров кластеров. Предлагается устанавливать их значения 0,15 и 0,5 соответственно. Параметры r_a и r_b величины, определяющие размер кластера. Большее их значение ведет к появлению меньшего количества кластеров, поскольку с их увеличением увеличивается и степень близости к центру кластера. Рекомендуется устанавливать $r_b = 1.5 \cdot r_a$. Параметр же r_a, радиус кластера, варьируется в диапазоне [0.15, 0.9].

Классификация на основе алгоритма горной кластеризации

Практическое применение алгоритма горной кластеризации опробовано на примере построения классификатора. Идея построения классификатора очень проста. На пространстве обучающих данных находятся центры кластеров с помощью алгоритма горной кластеризации. Каждому центру присваивается та метка класса, которая соответствует большинству экземпляров данных, принадлежащих данному кластеру. Принадлежность данных кластеру определяется минимальным расстоянием до центров кластеров.

Когда необходимо определить класс поступивших данных, то он определяется классом центра кластера, наиболее близко расположенного к данному экземпляру.

Классификация алгоритмом *k* – ближайших соседей

В основе алгоритма *k* ближайших соседей (*k*-nearest neighbor algorithm, KNN) лежит сходство объектов. Алгоритм способен выделить среди всех наблюдений k известных объектов (*k*-ближайших соседей), похожих на новый неизвестный ранее объект. Выделение *k* объектов происходит по степени близости к искомому объекту. На основе классов ближайших соседей выносится решение касательно нового объекта. Важной задачей данного алгоритма является подбор коэффициента *k* – количество записей, которые будут считаться похожими. Данный алгоритм работает по достаточно простой схеме:

Шаг 1. Вычислить расстояние от классифицируемого объекта до каждого из элементов обучающей выборки, в данной работе использовалось Евклидово расстояние, между точками i и j оно равно $\mathrm{D}_{ij} = \sqrt{\sum_{p=1}^{n}(x_{ip} - x_{jp})^2}$, где n – размерность пространства данных, x_{ip} – значение i-го признака p-го экземпляра данных.

Шаг 2. Выбрать *k* элементов, расстояние до которых минимально, где *k* – количество соседей, выбираемых пользователем.

Шаг 3. Класс классифицируемого объекта – это класс, наиболее встречающийся среди *k* ближайших соседей, $Class = \arg\max_j Neighbor_j$, где $Neighbor_j$ – количество соседей j-го класса. Если количество соседей одинаково для двух или более классов, то класс классифицируемого объекта выбирается случайным образом, либо по минимальному суммарному расстоянию до соседей одного класса.

На Рисунке 1 показан пример классификации для $k = 7$, количество классов равно 3. Классифицируемый объект будет относиться к классу треугольников. D1, …, D7 обозначены расстояния от классифицируемого объекта до ближайших объектов обучающей выборки.

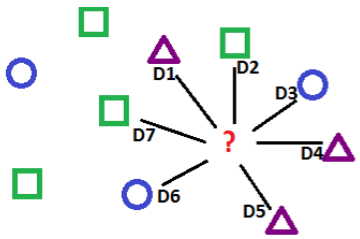

Рисунок 1 – Алгоритм классификации.

Эксперименты с реальными данными

На основание экспериментальных данных взятых из репозитория KEEL (http://keel.es), были построены классификатор *k*-ближайших соседей и классификатор на основе горной кластеризации. Полученные значения точности классификации представлены в Таблице 1.

Таблица 1 Точность классификации на реальных наборах данных

Наборы данных	Объем выборки	Количество признаков	*E*,«k-ближайших соседей»	*E*,«горная кластеризация»
Iris	150	4	100	97
Balance	625	4	90	83
Bupa	345	4	80	67
Glass	214	9	76	73

Важной характеристикой классификации является её время выполнения. Следует отметить, что для алгоритма *k*-ближайших соседей время выполнения классификации в среднем больше в 500 раз, чем для алгоритма на основе горной кластеризации. На Рисунке 2 приведены зависимости времени выполнения классификации обоих алгоритмов от количества входных признаков на примере набора данных Letters.

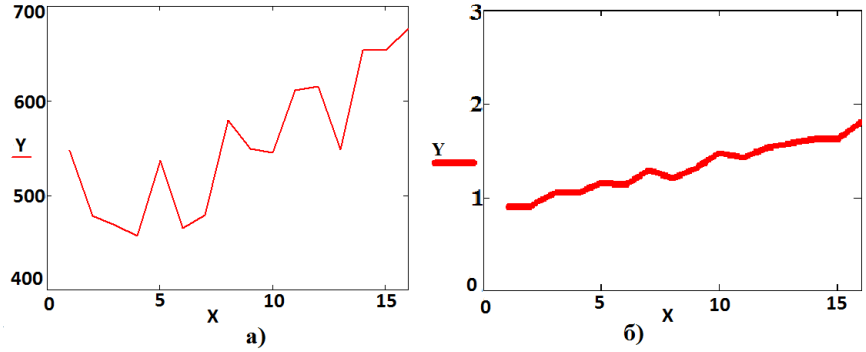

Рисунок 2 – График зависимости времени выполнения классификации от количества входных параметров: а) алгоритм *k*-ближайших соседей, б) алгоритм горной кластеризации.

Выводы

Рассмотрен алгоритм горной кластеризации, особенностью которого является то, что количество кластеров определяется самим алгоритмом, а не задается фиксировано. Предложена классификация данных с использованием горной кластеризации, принцип работы которой основан на выборе класса ближайшего центра кластера.

Рассмотрен общеизвестный алгоритм классификации *k*-ближайших соседей, в котором класс точки определяется классом большинства ближайших данных.

Проведены тесты обоих классификаторов на реальных данных. Работа классификатора на основе горной кластеризации оказалась значительно быстрее по сравнению с алгоритмом *k*-ближайших соседей, а точность классификации уменьшилась в среднем на 6.5%. Поскольку данный алгоритм классификации планируется использовать для настройки нечетких классификаторов, такой процент снижения точности является несущественным. На этапах оптимизации нечетких классификаторов процент правильной классификации будет повышаться алгоритмами оптимизации [2].

Список литературы

1. Witten I. H. Data Mining: Practical Machine Learning Tools and Techniques / I. H. Witten, E. Frank, M. A. Hall — 3rd Edition. — Morgan Kaufmann, 2011. — 664p.

2. Ходашинский И.А. Метаэвристические методы оптимизации параметров нечетких классификаторов / И.А. Ходашинский, А.Е. Анфилофьев, М.Б. Бардамова и др. // Информационные и математические технологии в науке и управлении. – 2016. – №1, Т.27. – С.73-81.

3. Chiu S., Fuzzy model identification based on cluster estimation // Journal of Intelligent & Fuzzy Systems. – 1994. – Vol. 2, № 3 – P. 267-278.