

Автор:

Дмитриев Егор Андреевич

студент

ФГАОУ ВО «Самарский национальный исследовательский

университет им. академика С.П. Королева»

г. Самара, Самарская область

МЕТРИЧЕСКИЕ КЛАССИФИКАТОРЫ

Аннотация: в данной работе рассматриваются один из методов машинного обучения – использование метрических классификаторов. Автором также представлены типы метрических алгоритмов.

Ключевые слова: метрическое пространство, метрика, вариационный ряд, алгоритм классификации, объект, машинное обучение.

Основные определения

Определение 1. Метрическое пространство – множество, в котором любой паре векторов ставится в соответствии число, определяемое метрикой данного пространства.

Определение 2. Метрика – числовая функция, $f : L \rightarrow R$, где L – метрическое пространство, а R – множество действительных чисел, удовлетворяющая следующим аксиомам:

$$1) f(\bar{x}) \geq 0, f(\bar{x}) = 0 \Rightarrow \bar{x} = \bar{0};$$

$$2) f(\bar{x}, \bar{y}) = f(\bar{y}, \bar{x});$$

$$3) f(\bar{x}, \bar{z}) \leq f(\bar{x}, \bar{y}) + f(\bar{y}, \bar{z}).$$

Определение 3. Вариационный ряд – последовательность элементов, которые расставлены по возрастанию по какому-либо числовому признаку.

Определение 4. Алгоритм классификации – отображение $a : X \rightarrow Y$, где X – метрическое пространство признаков классифицируемого объекта, а Y – множество классов. Классифицировать объект – однозначно определить, к какому классу относится объект.

Определение 5. Машинное обучение – нахождение отображения, в частности алгоритма классификации, который строится по множеству, называемому обучающей выборкой, а качество обучения проверяется по множеству, называемому тестовой выборкой.

Введение

В последнее время к задачам машинного обучения, в частности к классификации объектов, привлекают все больше и больше внимания. В большинстве компаний, приходится иметь дело с многочисленными данными, такие как: персональные данные сотрудников и клиентов, информация на крупных веб ресурсах, научные источники, наименования болезней, растений, животных и т. д. Для того, чтобы автоматизировать различные процессы, требуется использовать обучение системы для выполнения определенных задач.

Постановка проблемы

Необходимо ознакомиться с принципами метрического классифицирования объектов, разобраться с разновидностями метрических классификаторов.

Описание метрических классификаторов

Суть задачи классификации заключается в построении алгоритма классификации по обучающей выборке. Идея метрического классификатора состоит в том, что классифицируемый объект анализирует свое ближайшее окружение, то есть те объекты обучения, для которых метрика меньше чем у других объектов обучающей выборки.

Для произвольного u из обучающей выборки X^l составляем вариационный ряд таким образом:

$$\rho(u, x_u^{(1)}) \leq \rho(u, x_u^{(2)}) \leq \dots \rho(u, x_u^{(l)})$$

$x_u^{(i)}$ – i -й сосед объекта u среди объектов обучающей выборки;

$y_u^{(i)}$ – класс, к которому относится i -й объект обучающей выборки.

Метрический алгоритм классификации выглядит таким образом:

$$a(u, X^l) = \underset{y \in Y}{\operatorname{argmax}} \sum_{i=1}^l [y_u^i = y] w(i, u)$$

$w(i, u)$ – вес, степень важности i -го соседа u , неотрицателен.

Типы метрических классификаторов

Самым простейшим и легко реализуемым классификатором является метод ближайшего соседа. Он заключается в том, что класс классифицируемого объекта определяется по первому объекту в вариационном ряду. Тем не менее, алгоритм обладает значительными недостатками: неустойчивость к шумам, плохое качество классификации, отсутствие параметров, которые можно настроить. Частично эти недостатки решаются методов k ближайших соседей.

Метод k ближайших соседей

Данный алгоритм уже менее чувствителен к шумам, и обладает параметром, который можно менять для повышения качества классификации. Используем функционал leave-one-out:

$$LOO(k, X^l) = \sum_{i=1}^l [a(x_i: X^l / \{x_i\}, k) \neq y_i] \rightarrow \min_k$$

Варьируя параметр k , добиваемся минимизации функционала.

Метод парзеновского окна переменной ширины

Данный метод заключается во взвешенном голосовании i -тых соседей в принятии решения относительно класса объекта.

$$a(u, X^l, k, K) = \operatorname{argmax}_{y \in Y} \sum_{i=1}^l [y_u^i = y] K\left(\frac{\rho(u, x_u^{(i)})}{\rho(u, x_u^{(k+1)})}\right)$$

$K(r)$ – невозрастающая функция, которая убывает переходя к следующему соседу в вариационном ряду.

Заключение

В данной работе был рассмотрен один из методов классификации – метрический алгоритм и изучены типы метрических алгоритмов.

Список литературы

1. Журавлев Ю.И. Распознавание. Математические методы. Программная система. Практические применения / Ю.И. Журавлев, В.В. Рязанов, О.В. Сенько. – М.: Фазис, 2006.

2. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. – Новосибирск: ИМ СО РАН, 1999.