

Генетический выбор частичных порядков на множестве значений признаков в задаче классификации

Сорокин Олег, 317

ММП ВМК МГУ

Спецсеминар
28 февраля 2023 г.

1 Постановки задач

Задача классификации по прецедентам

Определения и обозначения

Постановка задачи для произведения частичных порядков

2 Упорядочение признаков и построение алгоритмов

Определения и обозначения

Общая схема работы алгоритма

Теоремы

Процедуры упорядочения признаков

3 Генетический подход к поиску покрытий

Схема ГА, применение ГА к задаче о покрытии

Формирование начальной популяции

Скрещивание и мутация

Селекция (функции приспособленности)

Программная реализация

Задача классификации по прецедентам

Пусть задано некоторое множество объектов M , представимое в виде объединения / непересекающихся множеств-классов K_1, \dots, K_l . Элементы множества M есть признаковые описания вида x_1, \dots, x_n , где каждый из признаков принимает конечное число значений. Имеется $\{S_1, \dots, S_m\} \subset M$ — множество объектов, принадлежность которых к определённым классам известна. Такие объекты называются прецедентами.

Требуется по предъявленному набору значений признаков $(b_1, \dots, b_n) \in M$ определить класс объекта.

Частичные порядки в признаковых пространствах

Особый интерес представляют задачи со сложными отношениями на множествах значений признаков. Существуют эффективные подходы (в смысле качества классификации), основанные на задании линейных порядков.

Определения и обозначения

Определение

Элементы x, y из частично упорядоченного множества P называются сравнимыми, если x предшествует y (запись $x \leq y$).

Определение

Пусть $P = P_1 \times \dots \times P_n$, где P_1, \dots, P_n — конечные частично упорядоченные множества.

Элемент $x = (x_1, \dots, x_n) \in P$ следует за элементом $y = (y_1, \dots, y_n) \in P$, если x_i следует за y_i ($i = 1, 2, \dots, n$)

Определения и обозначения

Определение

Пусть $R \subset P$, R^+ — множество элементов, следующих за элементами из R . Элемент $x \in P \setminus R^+$ называется независимым от R элементом множества P .

Если же кроме того $\forall u \in P \setminus R^+$ не выполнено отношение $x < u$, то x — максимальный независимый от R элемент множества P .

Постановка задачи для произведения частичных порядков

Аналогично предыдущей постановке, пусть

$M = \bigcup_{n=1}^l K_n$, где $K_i \cap K_j = \emptyset$ при $i \neq j$.

Пусть теперь M представимо в виде $N_1 \times \dots \times N_n$, где N_i ($i \in \{1, 2, \dots, n\}$) — конечное множество допустимых значений признака x_i . Не ограничивая общности, можно считать, что N_i имеет наибольший элемент k_i .

Пусть также задан набор прецедентов

$S_1 = (a_{11}, \dots, a_{1n})$, $S_2 = (a_{21}, \dots, a_{2n})$, ..., $S_m = (a_{m1}, \dots, a_{mn})$.

Требуется по предъявленному набору значений признаков (a_1, \dots, a_n) объекта $S \in M$ (класс которого, вообще говоря, неизвестен) определить этот класс.

Определения и обозначения

Определение

Пусть $R(K)$ и $R(\bar{K})$ — множества прецедентов из класса K и не из класса K соответственно.

Будем говорить, что алгоритм A классифицирует объект из $R(K)$ правильно, если A относит его к классу K .

Определение

Алгоритм A называется корректным на M алгоритмом, если A правильно классифицирует каждый прецедент S_1, \dots, S_m .

Определения и обозначения

Определение

Пусть $H = \{x_{j_1}, \dots, x_{j_r}\}$, $\sigma = (\sigma_1, \dots, \sigma_r)$, $\sigma_i \in N_{j_i}$ ($i = 1, 2, \dots, r$). Пара (σ, H) называется элементарным классификатором (эл. кл.) ранга r .

Замечание

Эл. кл. порождает набор $S_{(\sigma, H)} = (\gamma_1, \dots, \gamma_n)$, где $\gamma_{j_i} = \sigma_i$ ($i = 1, 2, \dots, r$) и $\gamma_t = k_t$ при $t \notin \{j_1, \dots, j_r\}$.

Определение

$$\hat{B}(\sigma, S, H) = \begin{cases} 1, & a_{j_i} \leq \sigma_i (i = 1, 2, \dots, r) \\ 0, & otherwise \end{cases}$$

Определения и обозначения

Определение

Эл. кл. (σ, H) называется корректным для класса K , если нельзя указать пару объектов $S' \in R(K)$ и $S'' \in R(\bar{K})$:
 $\hat{B}(\sigma, S', H) = \hat{B}(\sigma, S'', H) = 1$.

Определение

Корректный для класса K эл. кл. (σ, H) называется тупиковым, если $\forall(\sigma', H')$: $S_{(\sigma, H)} < S_{(\sigma', H')}$ не является корректным для класса K .

Определения и обозначения

Определение

(Тупиковый) корректный эл. кл. называется (тупиковым) представительным для класса K , если хотя бы один прецедент из класса K содержит данный эл. кл.

Общая схема работы алгоритма

- 1 Обучение: для каждого класса K строится некоторое множество представительных эл. кл. $C^A(K)$.
- 2 Процедура голосования: вычисление оценок вида

$$\Gamma(S, K) = \frac{1}{|C^A(K)|} \sum_{(\sigma, H) \in C^A(K)} P_{(\sigma, H)} * \hat{B}(\sigma, S, H)$$

Здесь $P_{(\sigma, H)}$ — веса, обычно это число объектов из $R(K)$, содержащих (σ, H) .

Теоремы

Теорема 1.

Пусть $C^A(K)$ содержит все тупиковые представительные эл. кл. класса K . Алгоритм A правильно классифицирует объект $S' \in R(K)$ тогда и только тогда, когда S' — независимый от $R(\bar{K})$ элемент множества M .

Теорема 2.

Пусть $\phi : M \rightarrow M \times \tilde{M}$ дублирует описание объекта с обратным отношением порядка. Если классы множества M не пересекаются, то любой прецедент из класса $\phi(K)$ содержит представительный эл. кл. класса $\phi(K)$.

Быстрая процедура независимого линейного упорядочения значений признаков

- Пусть $\mu_{ij}^{(1)}(a)$ ($i \in \{1, 2, \dots, l\}, j \in \{1, 2, \dots, \}, a \in N_j$) — доля прецедентов класса K_i , у которых признак x_j принимает значение a . Аналогично определим $\mu_{ij}^{(2)}(a)$ для прецедентов не из класса K .
- Введём $\mu_{ij}(a) = \mu_{ij}^{(1)}(a) - \mu_{ij}^{(2)}(a)$ — вес значения a .
- $\forall u, z \in N_j$ считаем $u \leq z$ тогда и только тогда, когда $\mu_{ij}(u) \geq \mu_{ij}(z)$.

Замечание

Порядок на множестве значений каждого признака выбирается независимо от выбора порядков для других признаков.



Процедура корректного упорядочения значений признаков

- Пусть для любого класса K множество $C^A(K)$ содержит все тупиковые представительные эл. кл. класса K .
- Построим булеву матрицу B_K :
 - 1 Каждой строке соответствует пара объектов $S' \in R(K)$, $S'' \in R(\overline{K})$, а каждому столбцу соответствует тройка (j, a, b) , где $j \in \{1, 2, \dots, n\}$, $a, b \in N_j$, $a \neq b$.
 - 2 Элемент на пересечении строки (S', S'') и столбца (j, a, b) равен 1, если признак x_j равен a и b у объектов S' и S'' соответственно.

Процедура корректного упорядочения значений признаков

Определение

Частичный порядок на M называется (A, K) -корректным, если алгоритм A правильно классифицирует каждый объект из $R(K)$.

Теорема 3.

Частичный порядок, заданный на множестве M , является (A, K) -корректным тогда и только тогда, когда существует неприводимое покрытие H матрицы B_K такое, что $\forall j \in \{1, 2, \dots, n\}$ и $\forall a, b \in N_j$ ($a < b$) столбец (j, b, a) не входит в H .

Одна из схем генетического алгоритма

- 1 Создаётся начальная популяция заданного объёма N_p . Для каждого индивида вычисляется приспособленность.
- 2 Скрещивание. Из популяции выбираются два родителя. К ним применяется оператор скрещивания, получается потомок.
- 3 Мутация. Потомок с заданной вероятностью подвергается мутации.
- 4 Отбор. Вычисляется приспособленность потомка. Одна из менее приспособленных особей заменяется.
- 5 Если не выполнено условие останова, то переход к п.2.

Одна из схем генетического алгоритма

Возможные способы представления особей:

- Бинарное. Код особи есть бинарный вектор $g = (g_1, \dots, g_n)$, где $g_i = 1$ тогда и только тогда, когда i -й столбец входит в набор столбцов H .
- Целочисленное. Код особи есть целочисленный вектор $g = (g_1, \dots, g_m)$, где i -я компонента равна номеру столбца, который покрывает строку с номером i .

Формирование начальной популяции

- 1
 - В бинарном случае все компоненты выбираются случайно. Если полученный набор столбцов не является покрытием, то он дополняется новыми столбцами.
 - Для целочисленного случая каждый раз случайно выбирается столбец из числа покрывающих нужную строку.
- 2 По вектору g восстанавливается набор столбцов.
- 3 Для каждого набора в порядке убывания весов столбцов проверяется, является ли $H \setminus \{j\}$ покрытием. Если да, то столбец исключается.
- 4 Если в P нет особи (H, g) , то она добавляется в P .
- 5 Если сгенерировано достаточно особей, то процесс завершается. Иначе переход к п.1.

Выбор родителей

- Панмиксия — все особи имеют одинаковую вероятность.
- Инбридинг — первый случайно, второй наиболее похожий в каком-то смысле.
- Аутбридинг — первый случайно, второй наиболее отличный в каком-то смысле.
- Селективное скрещивание — устанавливается порог приспособленности для возможности скрещивания.

Оператор скрещивания (кроссовер)

- Одноточечный — в наборе хромосом происходит разрыв по случайной точке, а затем обмен получившимися частями.
- Многоточечный — выбираются несколько таких точек.
- Однородный — каждая хромосома копируется от одного из родителей случайным образом. Для бинарного случая

$$g_i = \begin{cases} g_i^1, p_1 = \frac{f_2}{f_1 + f_2} \\ g_i^2, p_2 = \frac{f_1}{f_1 + f_2} \end{cases}$$

Мутация

В зависимости от модификации алгоритма, может изменяться одна или несколько случайно выбранных хромосом. При этом возможен выход из локального минимума.

Также можно изменять количество мутируемых хромосом со временем. Например,

$$k(t) = k_0 * \left(1 - \frac{1}{C * t + 1}\right)$$

Восстановление допустимости решения

При применении операторов скрещивания и мутации может возникнуть набор, не являющийся (неприводимым) покрытием. Пусть не покрыты строки M_H . В этом случае необходимо произвести процедуру восстановления допустимости решения:

- 1 H дополняется до покрытия матрицы последовательным добавлением столбцов, покрывающих M_H и минимизирующих $\frac{w_j}{|M_H \cap M_j|}$.
- 2 Из H конструируется неприводимое покрытие: убираем столбец j (в порядке убывания весов), если $H \setminus \{j\}$ является покрытием.

Функции приспособленности из статей (покрытие минимального веса)

$$1 \quad f = \sum_{i=1}^n [g_i = 1] * w_i$$

(в целочисленном случае аналогично, вычисляется вес покрытия).

$$2 \quad f_i = w_i - \min_{j \in \{1, \dots, N\}} w_j + 1$$

(исправляет проблему предыдущей функции).

Функции приспособленности, которые можно попробовать для поиска любых покрытий

- 1 Пусть B' — матрица, составленная из столбцов H .

$$f_1 = \sum_{i=1}^m [\sum_{j=1}^n B'_{ij} = 0] + 1$$

- 2 Некоторые строки покрываются малым числом столбцов. Если ГА их не включает, то застревает в локальном минимуме. Идея: покрытие таких строк должно сильнее минимизировать функцию. Можно перейти ко взвешенной задаче.

Классический ГА для поиска минимальных покрытий

