Лекция 5

Прицип частичной прецедентности, тестовый алгоритм, модель ABO

Лектор – *Сенько Олег Валентинович*

Курс «Математические основы теории прогнозирования»
4-й курс, III поток

Содержание лекции

Принцип частичной прецедентности. Тестовый алгоритм.

Существует ряд методов распознавания, основанных на принципе частичной прецедентности. Данный принцип подразумевает поиск по обучающей выборке фрагментов описаний, позволяющих разделить распознаваемые классы $K_1, \ldots, K_L$. Распознаваемый объект оценивается по совокупности найденных фрагментов. Одной из первых реализаций принципа частичной прецедентности является тестовый алгоритм, предложенный в 1966 году. Данный алгоритм основан на понятии тупикового теста. Исходный вариант тестового алгоритма предназначен для распознавания объектов, описываемых с помощью бинарных или категориальных признаков $X_1, \ldots, X_n$. Иными словами $X_i \in \{a_1^i, \ldots, a_{k_i}^i\}$, , $i = 1, 2, \ldots, n$.

Тестовый алгоритм

Пусть обучающая выборка $\widetilde{S}_t$, содержит объекты из классов $K_1, \ldots, K_L$. При этом общее число объектов равно m. Выборке $\widetilde{S}_t$ ставится в соответствие таблица $\mathbb{T}_{nmL}$, j-ой строкой которой являются значения признаков для объекта s_j.

Определение 1. Тестом таблицы $\mathbb{T}_{nmL}$ называется совокупность столбцов $i_1, \ldots, i_r$ таких, что после удаления из $\mathbb{T}_{nmL}$ всех столбцов, за исключением $i_1, \ldots, i_r$, в полученной таблице все пары строк, соответствующие разным классам различны хотя бы по одному признаку

Определение 2. Тест $T = \{i_1, \ldots, i_r\}$ называется тупиковым, если никакое его отличное от T подмножество (собственное подмножество) тестом не является.

На этапе обучения ищется множество всевозможных тупиковых тестов $\widetilde{T}(\widetilde{S}_t)$ для таблицы $\mathbb{T}_{nmL}$. Предположим что нам требуется распознать объект s_* с векторным описанием $(x_{*1}, \ldots, x_{*n})$.

Тупиковые тесты

Выделим в векторном описании фрагмент $(x_{*i_1}, \ldots, x_{*i_r})$, соответствующий тесту

$$T = \{i_1, \ldots, i_r\}, T \in \widetilde{T}(\widetilde{S}_t).$$

Фрагмент $(x_{*i_1}, \ldots, x_{*i_r})$ сравнивается с множеством фрагментов строк таблицы $\mathbb{T}_{nmL}$, соответствующих классу K_l : $\{(x^t_{ji_1}, \ldots, x^t_{ji_r}) \mid s_j \in K_l\}$. В случаях, когда выполняются равенства

$$x_{*i_1} = x^t_{ji_1}, \ldots, x_{*i_r} = x^t_{ji_r}$$

фиксируем полное совпадение распознаваемого объекта s_* с объетом $s_j \in K_l$ тесту T. Обозначим число полных совпадений через $G_l(\mathbb{T}_{nmL}, s_*)$. Оценка объекта s_* за класс K_l вычисляется по формуле:

$$\gamma(s_*) = \frac{1}{m_l} \sum_{T \in \widetilde{T}(\widetilde{S}_t)} G_l(\mathbb{T}_{nmL}, s_*),$$

где m_l - число
объектов обучающей выборки из класса K_l .

Тупиковые тесты

Классификация объекта s_* может производится по вектору оценок $[\gamma_1(s_*), \ldots, \gamma_L(s_*)]$ с помощью стандартного решающего правила, т.е. объект относится в тот класс, оценка за который максимальна. Задача нахождения множества всех тупиковых тестов таблицы $\mathbb{T}_{nmL}$ сводится к задаче поиска всех тупиковых покрытий матрицы сравнений $\mathbb{C}_{nmL}$, которая строится по матрице $\mathbb{T}_{nmL}$. Каждой паре классов K_l и K'_l в матрице $\mathbb{C}_{nmL}$ сопоставлена подматрица $\mathbb{C}^{ll'}_{nmL}$, состоящая из $m_l m_{l'}$ строк. Пусть $(c^{ll'}_{f1}, \ldots, c^{ll'}_{fn})$ строка матрицы $\mathbb{C}^{ll'}_{nmL}$ соответствует сравнению описаний объекта $\boldsymbol{x}_g$ из класса K_l и описание объекта $\boldsymbol{x}_{g'}$ из класса $K_{l'}$. Элемент $c^{ll'}_{fi} = 0$, если $x_{gi} = x_{g'i}$, и $c^{ll'}_{fi} = 1$, если $x_{gi} \neq x_{g'i}$. Таким образом $\mathbb{C}_{nmL}$ имеет размерность $M \times n$, где$\sum_{l=1}^{L} \sum_{l'=1}^{l-1} m_l m_{l'}$.

Методы поиска тупиковых тестов. Задача о покрытии.

Мы будем говорить, что столбец с номером j матрицы $\mathbb{C}_{nmL}$ покрывает строку $(c_{f1}, \ldots, c_{fn})$, если $c_{fj} = 1$. Набор столбцов $\{j_1, \ldots, j_r\}$ образует покрытие матрицы $\mathbb{C}_{nmL}$, если при любом $f \in \{1, \ldots, M\}$ существует такое $j' \in \{j_1, \ldots, j_r\}$, что $c_{fj'} = 1$. Покрытие $\widetilde{J}$ называется тупиковым, если его произвольное собственное подмножество, покрытием не является. Очевидно, что для произвольного набора столбцов обладание свойством тупикового набора для $\mathbb{T}_{nmL}$ эквивалентно обладанию свойством тупикового покрытия для $\mathbb{C}_{nmL}$. Таким образом задача о поиске всевозможных тупиковых тестов сводится к известной задаче о поиске всевозможных тупиковых покрытий. Нахождение всех тупиковых тестов является сложной комбинаторной задачей. Однако эффективные алгоритмы поиска разработаны для некоторых типов таблиц. При решении практических задач эффективен подход, основанный на вычислении только части тупиковых тестов.

Предстаительные наборы

Другим известным классом алгоритмов распознавания, основанным на принципе частичной прецедентности, являются алгоритмы типа КОРА. В отличие от тестового алгоритма, где в качестве информативных элементов используются несжимаемые наборы признаков – тупиковые тесты, в алгоритмах типа КОРА в качестве информативных элементов используются несжимаемые фрагменты описаний эталонных объектов обучающей выборки.

Определение 3. Пусть $(x_{v1}, \ldots, x_{v_n})$ - признаковое описание объекта $s_v \in \widetilde{S}_t \bigcap K_l$. Набор $(x_{vj_1}, \ldots, x_{vj_r})$ называется представительным набором для класса K_l при выполнении следующего условия. Пусть $(x_{uj_1}, \ldots, x_{uj_r})$ произвольная строка таблицы $\mathbb{T}_{nmL}$, которая соответствует объекту s_u, не принадлежащему $\widetilde{S}_t \bigcap K_l$. Тогда существует такое $j' \in \{j_1, \ldots, j_r\}$, что $x_{uj'} \neq x_{vj'}$.

Представительные наборы

Определение 4. Представительный набор называется тупиковым, если никакое его собственное подмножество представительным набором не является.

На этапе обучения для каждого из классов $K_1, \ldots, k_L$ по таблице $\mathbb{T}_{nmL}$ ищется множество всевозможных тупиковых представительных наборов. Пусть $\widetilde{V}_l$ - множество всевозможных представительных наборов для класса K_l.

Предположим, что нам требуется распознать объект s_* с описанием $(x_{*1}, \ldots, x_{*n})$. Пусть $\boldsymbol{v} = (x_{vj_1}, \ldots, x_{vj_r})$ - представительный набором Функция $B(s_*, \boldsymbol{v})$ равна 1, при выполнении всех неравенств $(x_{*j_1} = x_{vj_1}, \ldots, x_{*j_r} = x_{vj_r})$ и $B(s_*, \boldsymbol{v})$ равна 0 в противном случае. Оценка s_* за класс K_l вычисляется по формуле

$$\Gamma(s_*, K_l) = \frac{1}{\mid \widetilde{V}_l \mid} \sum_{\boldsymbol{v} \in \widetilde{V}_l} B(s_*, \boldsymbol{v})$$

Представительные наборы

Для нахождения тупиковых представительных наборов для класса K_l , содержащихся в эталонном описании $\boldsymbol{x}_v$ объекта s_v формируются матрица сравнения $\mathbb{C}^{lv}_{nmL}$ со всеми описаниями других классов таблицы $\mathbb{T}_{nmL}$. Пусть строка $(c^{lv}_{f1}, \ldots, c^{lv}_{fn})$ матрицы $\mathbb{C}^{lv}_{nmL}$ соответствует сравнению $\boldsymbol{x}_v$ с описанием $\boldsymbol{x}_g$ объекта s_g, не принадлежащему $K_l \bigcap \widetilde{S}_t$. Элемент $c^{lv}_{f1} = 0$, если $x_{vj} = x_{gj}$, и $c^{lv}_{f1} = 1$, если $x_{vj} \neq x_{gj}$. Таким образом матрица $\mathbb{C}^{lv}_{nmL}$ имеет размер $(m - m_l)n$. Тупиковые покрытия матриц сравнения $\mathbb{C}^{lv}_{nmL}$ определяют тупиковые представительные наборы, являющиеся фрагментами описания $\boldsymbol{x}_v$. Полное множество представительных наборов для класса K_l является объединением множеств представительных наборов, найденных для описаний всех объектов обучающей выборки из K_l, Таким образом задача поиска всех представительных наборов сводится к решению m_l задач поиска тупиковых покрытий для матриц сравнения размера $(m - m_l)n$.

Представительные наборы

Первоначальные варианты тестового алгоритма и алгоритма типа КОРА были разработаны для бинарных или категориальных переменных. Они не могут быть напрямую использованы в задачах с признаками, принимающими значения из интервалов вещественной оси. Для того, чтобы обеспечить возможность работы с подобной информацией могут быть использованы два подхода.
Первый подход основан на разбиении области возможных значений каждого веществennозначного признака на k связных подмножеств (интервалов, полуинтервалов, отрезков). Значению признака, принадлежащего j-ому элементу разбиения присваивается значение j. Разбиение оптимизируется с целью достижения максимального разделения классов. Выбирается такое число элементов разбиения k, при котором достигается максимальная точность распознавания.

Веществознозначная информация

Другой подход основан на модификации понятий теста и представительного набора с использованием пороговых параметров $(\varepsilon_1, \ldots, \varepsilon_n)$, которые задаются для признаков $X_1, \ldots, X_n$.

Определение 5. Тестом таблицы $\mathbb{T}_{nmL}$ называется совокупность столбцов $i_1, \ldots, i_r$ таких, что после удаления из $\mathbb{T}_{nmL}$ всех столбцов, за исключением столбцов $i_1, \ldots, i_r$ в полученной таблице $\mathbb{T}_{rmL}$ для всех пар строк, соответствующих разным классам абсолютная величина различий хотя бы по одному признаку X_j превышает ε_j .

Аналогичным образом вводится модифицированное определение представительного набора.

Определение 6. Пусть $(x_{v1}, \ldots, x_{vn})$- признаковое описание объекта $s_v \in \widetilde{S}_t \bigcap K_l$. Набор $(x_{vj_1}, \ldots, x_{vj_r})$ называется представительным набором для класса K_l при выполнении следующего условия. Пусть $(x_{uj_1}, \ldots, x_{uj_r})$ произвольная строка таблицы $\mathbb{T}_{nmL}$ соответствующая объекту s_u, не принадлежащему $\widetilde{S}_t \bigcap K_l$. Тогда существует такое $j' \in \{j_1, \ldots, j_r\}$, что $\mid x_{uj'} - x_{vj'} \mid > \varepsilon_{j'}$.

Главным требованием при выборе ε-порогов является достижение максимальной отделимости объектов разных классов при сохранении сходства внутри классов. Поиск тупиковых тестов и тупиковых представительных наборов при модифицированных определениях аналогичен их поиску в первоначальных вариантах методов.

Алгоритмы вычисления оценок

Тестовый алгоритм и алгоритм с представительными наборами являются частью более общей конструкции, основанной на принципе частичной прецедентности и носящей название алгоритмов вычисления оценок. Существует много вариантов моделей данного типа. Причём конкретный вид модели определяется выбранными способами задания различных её элементов. Рассмотрим основные составляющие модели.

Задание системы опорных множеств. Под опорными множествами модели ABO понимается наборы признаков, по которым осуществляется сравнение распознаваемых и эталонных объектов. Примером системы опорных множеств является множество тупиковых тестов. Система опорных множеств Ω_A некоторого алгоритма A может задаваться через систему подмножеств множества $\{1, \ldots, n\}$ или через систему характеристических бинарных векторов.

Алгоритмы вычисления оценок

Каждому подмножеству $\{1, \ldots, n\}$ может быть сопоставлен бинарный вектор размерности n . Пусть $\{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\}$. . Тогда$\{i_1, \ldots, i_k\}$ сопоставляется вектор $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)$, все компоненты которого равны 0 кроме равных 1 компонент $(\omega_{i_1}, \ldots, \omega_{i_k})$. . Теоретические исследования свойств тупиковых тестов для случайных бинарных таблиц показали, что характеристические векторы для почти всех тупиковых тестов имеют асимптотически (при неограниченном возрастании размерности таблицы обучения) одну и ту же длину. Данный результат является обоснованием выбора в качестве системы опорных векторов всевозможных наборов, включающих фиксированное число признаков k или

$$\Omega_A = \{\boldsymbol{\omega} :| \omega |= k\}$$

Оптимальное значение k находится в процессе обучения или задаётся экспертом. .

Другой часто используемой системе опорных множеств соответствует множество всех подмножеств $\{1, \ldots, n\}$ за исключением пустого множества. Иными словами в систему опорных множеств входит произвольный набор признаков или $\Omega = \{\boldsymbol{\omega}\} \backslash \omega_0$, где ω_0 - вектор, все компоненты которого равны 0.

Задание функции близости. Пусть опорное множество $\{i_1, \ldots, i_k\}$ соответствует характеристическому вектору $\boldsymbol{\omega}$. Фрагмент $(x_{\mu i_1}, \ldots, x_{\mu i_k})$ описания $(x_{\mu 1}, \ldots, x_{\mu n})$ объекта s_μ называется $\boldsymbol{\omega}$-частью объекта s_μ.

Под функцией близости $B_{\boldsymbol{\omega}}(s_\mu, s_\nu)$ понимается функция от соответствующих $\boldsymbol{\omega}$-частей сравниваемых объектов, принимающая значения 1 (объекты близки) или 0 (объекты удалены). Функции близости обычно задаются с помощью порговых параметров $(\varepsilon_1, \ldots, \varepsilon_n)$, характеризующих близость объектов по отдельным признакам.

Пример функций близости.

1) $B_{\boldsymbol{\omega}}(s_\mu, s_\nu) = 1$, если при произвольном $i \in \{1, \ldots, n\}$, при котором $\omega_i = 1$, всегда выполняется неравенство

$$| x_{\mu i} - x_{\nu i} | < \varepsilon_i.$$

$B_{\boldsymbol{\omega}}(s_\mu, s_\nu) = 0$,, если существует такое $i' \in \{1, \ldots, n\}$, что одновременно $\omega'_i = 1$ и $| x_{\mu i'} - x_{\nu i} | > \varepsilon'_i$.

2) Пусть ε - скалярный пороговый параметр. Функция $B_{\boldsymbol{\omega}}(s_\mu, s_\nu) = 1$, если выполняется неравенство

$$[\sum_{i=1}^{n} \omega_i \mid x_{\mu i} - x_{\nu i} \mid] < \varepsilon.$$

В противном случае $B_{\boldsymbol{\omega}}(s_\mu, s_\nu) = 0$.

Алгоритмы вычисления оценок (АВО)

Важным элементом АВО является оценка близости распознаваемого объекта s_* к эталону s_μ по заданной $\boldsymbol{\omega}$- части. Данная оценка близости, которая будет обозначаться $\Gamma_{\boldsymbol{\omega}}(s_*, s_\mu)$, формируется на основе введённых ранее функций близости и, возможно, дополнительных параметров.

$$A)\Gamma_{\boldsymbol{\omega}}(s_*, s_\mu) = B_{\boldsymbol{\omega}}(s_*, s_\mu).$$
$$B)\Gamma_{\boldsymbol{\omega}}(s_*, s_\mu) = p_{\boldsymbol{\omega}} B_{\boldsymbol{\omega}}(s_*, s_\mu),$$

где $p_{\boldsymbol{\omega}}$- параметр, характеризующий информативность опорного множества с характеристическим вектором $\boldsymbol{\omega}$.

$$C)\Gamma_{\boldsymbol{\omega}}(s_*, s_\mu) = \gamma_\mu(\sum_{j=1}^{n} p_j \omega_j) B_{\boldsymbol{\omega}}(s_*, s_\mu),$$

где γ_μ - параметр, характеризующий информативность эталонного объекта s_μ , параметры $p_1, \ldots, p_n$ характеризуют информативность отдельных признаков.

Алгоритмы вычисления оценок (ABO)

Оценка объекта s_* за класс K_l при фиксированном $\boldsymbol{\omega}$. Оценка объекта s_* за класс K_l при фиксированном характеристическом векторе $\boldsymbol{\omega}$ может вычисляться как среднее значение близости s_* к эталонным объектам из класса K_l

$$\Gamma_{\boldsymbol{\omega}}^{l}(s_*) = \frac{1}{m_l} \sum_{s_\mu \in K_l} \Gamma_{\boldsymbol{\omega}}(s_*, s_\mu).$$

Общая оценка s_* за класс K_l вычисляется как сумма оценок $\Gamma_{\boldsymbol{\omega}}^{l}(s_*)$ по опорным множествам из системы Ω_A:

$$\Gamma^{l}(s_*) = \sum_{\omega \in \Omega_A} \Gamma_{\boldsymbol{\omega}}^{l}(s_*). \tag{1}$$

Наряду с формулой (1) используется формула

$$\Gamma^{l}(s_*) = \nu_l \sum_{\omega \in \Omega_A} \Gamma_{\boldsymbol{\omega}}^{l}(s_*). \tag{2}$$

Алгоритмы вычисления оценок (ABO)

Использование взвешивающих параметров $\nu_1, \ldots, \nu_L$ позволяет регулировать доли правильно распознанных объектов $K_1, \ldots, K_L$. Прямое вычисление оценок за классы по формулам (1) и (2) в случаях, когда в качестве систем опорных множеств используются наборы с фиксированным числом признаков или всевозможные наборы признаков, оказывается практически невозможным при сколь либо высокой размерности признакового пространства из-за необходимости вычисления огромного числа значений функций близости. Однако при равенстве весов всех признаков существуют эффективные формулы для вычисления оценок по формуле (1). Предположим, что оценки близости распознаваемого объекта s_* к эталону s_μ по заданной $\boldsymbol{\omega}$ - части вычисляются по формуле (A). Тогда оценка по формуле (1) принимает вид

$$\Gamma^l(s_*) = \sum_{\boldsymbol{\omega} \in \Omega_A} \sum_{s_\mu \in K_l} B_{\boldsymbol{\omega}(s_*, s_\mu)}$$

Рассмотрим сумму $\sum_{\boldsymbol{\omega}\in\Omega_A} B_{\boldsymbol{\omega}}(s_*, s_\mu)$. Предположим, что общее число признаков, по которым объект s_* близок к объекту s_μ равно $\boldsymbol{d}(s_*, s_\mu)$. Иными словами $\boldsymbol{d}(s_*, s_\mu) =\mid D(s_*, s_\mu) \mid$, где $D(s_*, s_\mu) = \{i :\mid x_{*i} - x_{\mu i} \mid< \varepsilon_i\}$. Очевидно функция близости $B_{\boldsymbol{\omega}}(s_*, s_\mu) = 1$ тогда и только тогда, когда опорное множество, задаваемое характеристическим вектором $\boldsymbol{\omega}$, полностью входит в множество $D(s_*, s_\mu)$. Во всех остальных случаях $B_{\boldsymbol{\omega}}(s_*, s_\mu) = 0$.

Алгоритмы вычисления оценок (ABO)

Предположим, что система опорных множеств удовлетворяет условию $\Omega_A = \{\boldsymbol{\omega} :| \boldsymbol{\omega} :|= k\}$. Очевидно, что число опорных множеств в Ω_A , удовлетворяющих условию $B_{\boldsymbol{\omega}}(s_*, s_\mu) = 1$, равно $C^k_{\boldsymbol{d}(s_*,s_\mu)}$. Откуда следует, что $\sum_{\boldsymbol{\omega}\in\Omega_A} B_{\boldsymbol{\omega}}(s_*, s_\mu) = C^k_{\boldsymbol{d}(s_*,s_\mu)}$. Следовательно оценка по формуле (1) может быть записана в виде

$$\Gamma^l(s_*) = \frac{1}{m_l} \sum_{s_\mu\in K_l} \gamma_\mu C^k_{\boldsymbol{d}(s_*,s_\mu)} \tag{3}$$

Предположим, что система Ω_A включает в себя всевозможные опорные множества. В этом случае число опорных множеств в Ω_A , удовлетворяющих условию $B_{\boldsymbol{\omega}}(s_*, s_\mu) = 1$, равно $2^{\boldsymbol{d}(s_*,s_\mu)} - 1$. Следовательно оценка по формуле (1) может быть записана в виде

$$\Gamma^l(s_*) = \frac{1}{m_l} \sum_{s_\mu\in K_l} \gamma_\mu [2^{\boldsymbol{d}(s_*,s_\mu)} - 1].$$

Обучение АВО

Для обучения алгоритмов АВО в общем случае может быть использован тот же самый подход, который используется для обучения в методе «Линейная машина». Предположим, что решается задача обучения алгоритмов для распознавания объектов , принадлежащих классам $K_1, \ldots, K_L$ При правильного распознавания объекта $s_i \in \widetilde{S}_t \bigcap K_l$ должен выполняться блок неравенств

$$\Gamma^l(s_i) > \Gamma^{l\prime}(s_i), \tag{4}$$

где $l' \in \{1, \ldots, L\} \setminus \{l\}$. При использовании для вычисления оценок формулы (3) блок (4) приводится к виду

$$\frac{1}{m_l} \sum_{s_\mu \in K_l} \gamma_\mu C^k_{\boldsymbol{d}(s_*, s_\mu)} > \frac{1}{m_l} \sum_{s_\nu \in K_{l'}} \gamma_\mu C^k_{\boldsymbol{d}(s_*, s_\nu)} \tag{5}$$

где $l' \in \{1, \ldots, L\} \setminus \{l\}$.

Обучение АВО

Обучение сводится к поиску такого набора весовых коэффициентов $\gamma_1, \ldots, \gamma_m$ при которых выполняется по возможности максимальное число блоков неравенств вида (4). Поиск максимальной оптимальных коэффициентов может производиться с использованием эвристического релаксационного метода, аналогичного тому, что был использован при обучении алгоритма «Линейная машина».