МАТЕМАТИЧЕСКИЕ ОСНОВЫ ТЕОРИИ ПРОГНОЗИРОВАНИЯ

Лектор

Сенько Олег Валентинович

Лекция 5

Принцип частичной прецедентности

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.

Тестовый алгоритм

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 к описываемых с помощью бинарных или категориальных признаков $X_1, \ldots, X_n$. Иными словами $X_i \in \{a_1^i, \ldots, a_{k_i}^i\}$, где k_i - число значений, принимаемых признаком X_i

$i = 1, 2, \ldots, n$

Тестовый алгоритм

Обучающей выборке $\tilde{S}_t$ ставится в соответствие таблица $\mathbf{T}_{nmL}$, j-ой строкой которой являются значения признаков для объекта s_j .

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

Тестовый алгоритм

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 $\tilde{T}(\tilde{S}_t)$ для таблицы $\mathbf{T}_{nmL}$.

Предположим что нам требуется распознать объект S_* с веторным описанием $(x_{*1}, \ldots, x_{*n})$. Выделим в векторном описании фрагмент $(x_{*i_1}, \ldots, x_{*i_r})$, соответствующий тесту $T = \{i_1, \ldots, i_r\}, T \in \tilde{T}(\tilde{S}_t)$

Тестовый алгоритм

Фрагмент $(x_{*i_1},\ldots,x_{*i_r})$ сравнивается с множеством фрагментов строк $(x^{\mathbf{T}}_{ji_1},\ldots,x^{\mathbf{T}}_{ji_r})$ таблицы $\mathbf{T}_{nmL}$, соответствующих классу K_i : $\{(x^{\mathbf{T}}_{ji_1},\ldots,x^{\mathbf{T}}_{ji_r}) | s_j \in K_i\}$. В случаях, когда выполняются равенства $(x_{*i_1} = x^{\mathbf{T}}_{ji_1},\ldots,x_{*i_r} = x^{\mathbf{T}}_{ji_r})$ фиксируем полное совпадение.

Обозначим число полных совпадений через $G_i(T, s_*)$

Тестовый алгоритм

Оценка объекта s_* за класс K_i вычисляется по формуле:

$$\gamma_i(s_*) = \frac{1}{m_i} \sum_{T \in \tilde{T}(\tilde{S}_t)} G_i(T, s_*)$$

, где m_i - число

объектов обучающей выборки из класса K_i .

Тестовый алгоритм

- Классификация объекта s_* может производится с помощью 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 $\mathbf{T}_{nmL}$ сводится к задаче поиска всех тупиковых покрытий матрицы сравнений $\mathbf{C}_{nmL}$, которая строится по матрице $\mathbf{T}_{nmL}$.

Тестовый алгоритм

Каждой паре классов K_i и $K_{i'}$ в матрице $\mathbf{C}_{nmL}$ сопоставлена подматрица $\mathbf{C}_{nmL}^{ii'}$, состоящая из $m_i m_{i'}$ строк. Пусть строка $(c_{f1}^{ii'},\ldots,c_{fn}^{ii'})$ матрицы $\mathbf{C}_{nmL}^{ii'}$ соответствует сравнению описаний объекта $\mathbf{x}_g$ из класса K_i и объекта $\mathbf{x}_{g'}$ из $K_{i'}$ класса.

Элемент $c_{fj}^{ii'}=0$, если $x_{gj}=x_{g'j}$, и $c_{fj}^{ii'}=1$, если $x_{gj}\neq x_{g'j}$.

Таким образом $\mathbf{C}_{nmL}$ имеет размерность $M\times n$, где

$$M=\sum_{i=1}^{L}\sum_{i'=1}^{i-1} m_i m_{i'}$$

Тестовый алгоритм

Мы будем говорить, что столбец с номером j матрицы $\mathbf{C}_{nmL}$ покрывает строку $(c_{f1},\ldots,c_{fn})$, если $c_{fj}=1$. Набор столбцов $\{j_1,\ldots,j_r\}$ образует покрытие матрицы $\mathbf{C}_{nmL}$,

если $\forall f \in\{1,\ldots\ldots,M\}\,\exists j' \in\{j_1,\ldots,j_r\}$ такое, что $c_{fj'}=1$.

Покрытие $\tilde{J}$ тупиковым, если его произвольное собственное подмножество, покрытием не является. Очевидно что для произвольного набора столбцов обладание свойством тупикового набора для $\mathbf{T}_{nmL}$, эквивалентно обладанию свойством тупикового покрытия для $\mathbf{C}_{nmL}$

Тестовый алгоритм

- 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.

Представительные 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.

Представительные наборы

Определение 3. Пусть $(x_{v1},\ldots,x_{vn})$ - признаковое описание объекта $s_v \in K_i$. Набор $(x_{vj_1},\ldots,x_{vj_r})$ называется представительным набором для класса K_i , если для произвольной строки $(x_{u1},\ldots,x_{un})$ таблицы $\mathbf{T}_{nmL}$ соответствующей объекту $s_u \notin K_i$ $\exists j'$ $j' \in \{j_1,\ldots,j_r\}$ такое, что $x_{vj'} \neq x_{uj'}$

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 $\mathbf{T}_{nmL}$ ищется множество всевозможных тупиковых представительных наборов. Пусть $\tilde{V}_i$ - множество всевозможных представительных наборов для класса K_i

Представительные наборы

Предположим, что нам требуется распознать объект s_* с описанием $(x_{*1},\ldots,x_{*n})$. Пусть $u=(x_{v1},\ldots,x_{vn})$ - представительный набор.

Функция $\mathrm{B}(s_*,u)$ равна 1, если $(x_{*i_1}=x_{vi_1},\ldots,x_{*i_r}=x_{vi_r})$ и $\mathrm{B}(s_*,u)$ равна 0 в противном случае.

Оценка s_* вычисляется по формуле $\Gamma_i(s_*)=\frac{1}{|\tilde{V}_i|}\sum_{u\in\tilde{V}_i}\mathrm{B}(s_*,u)$

Представительные наборы

Для нахождения тупиковых представительных наборов для класса K_i, содержащихся в эталонном описании $\mathbf{x}_v$ объекта $s_v \in K_i$ формируются матрица сравнения $\mathbf{C}_{nmL}^{iv}$ со всеми описаниями других классов таблицы T_{nmL}.

Пусть строка $(c_{f1}^{iv},\ldots,c_{fn}^{iv})$ матрицы $\mathbf{C}_{nmL}^{iv}$ соответствует сравнению $\mathbf{x}_v$ с описанием $\mathbf{x}_g$ объекта $s_g \notin K_i$

Элемент $c_{fj}^{iv}=0$, если $x_{vj}=x_{gj}$, и $c_{fj}^{iv}=1$, если $x_{vj}\neq x_{gj}$.

Таким образом матрица $\mathbf{C}_{nmL}^{iv}$ имеет размер $(m-m_i)\times n$.

Представительные наборы

Тупиковые покрытия матриц сравнения $\mathbf{C}_{nmL}^{iv}$ определяют тупиковые представительные наборы, являющиеся фрагментами описания $\mathbf{x}_v$. Полное множество представительных наборов для класса K_i является объединением множеств представительных наборов, найденных для описаний всех объектов обучающей выборки из K_i, Таким образом задача поиска всех представительных наборов сводится к решению m_i задач поиска тупиковых покрытий для матриц сравнения размера $(m-m_i)\times n$

Обобщение на задачи с вещественнозначной информации

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.

Обобщение на задачи с вещественнозначной информацией

- Первый подход основан на разбиении области возможных значений каждого вещественнозначного 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.

Обобщение на задачи с вещественнозначной информацией

Другой подход основан на модификации понятий теста и представительного набора с использованием пороговых параметров $(\varepsilon_1,\ldots,\varepsilon_n)$

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

Обобщение на задачи с вещественнозначной информацией

- Аналогичным образом вводится модифицированное определение представительного набора.

Определение 6. Пусть $(x_{v1},\ldots,x_{vn})$ - признаковое описание объекта $s_v \in K_i$. Набор $(x_{vj_1},\ldots,x_{vj_r})$ называется представительным набором для класса K_i , если для произвольной строки $(x_{u1},\ldots,x_{un})$ таблицы $\mathbf{T}_{nmL}$ соответствующей объекту $s_v \notin K_i \quad \exists j' \quad j' \in \{j_1,\ldots,j_r\}$ такое, что $|x_{vj'} - x_{uj'}| \geq \varepsilon_{j'}$.

Обобщение на задачи с вещественнозначной информацией

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.

Алгоритмы вычисления оценок

- **Задание системы опорных множеств.** Под Опорными множествами модели АВО 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 $\{i_1,\ldots,i_k\}\in\{1,\ldots,n\}$. Тогда $\{i_1,\ldots,i_k\}$ сопоставляется вектор $\boldsymbol{\omega}=(\omega_1,\ldots,\omega_n)$, все компоненты которого равны 0 кроме равных 1 компонент $\omega_{i_1},\ldots,\omega_{i_r}$.

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.

Алгоритмы вычисления оценок

Данный результат является обоснованием выбора в качестве системы опорных векторов всевозможные наборы, включающие фиксированное число признаков k или $\Omega_A = \{\omega : |\omega| = k\}$. 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.

Алгоритмы вычисления оценок

Иными словами в систему опорных множеств входит произвольный набор признаков или $\Omega_A = \{\omega\} \setminus \omega_0$, где ω_0 - вектор, все компоненты которого равны 0.

Задание функции близости. Пусть опорное множество $\{i_1, \ldots, i_k\}$ соответствует характеристическому вектору ω. Фрагмент $(x_{\mu i_1}, \ldots, x_{\mu i_k})$ описания $(x_{\mu 1}, \ldots, x_{\mu n})$ объекта S_μ называется ω - частью объекта и обозначается ωS_μ .

Алгоритмы вычисления оценок

Под функцией близости $\mathrm{B}_{\omega}(S_{\mu}, S_{\nu})$ понимается функция от соответствующих ω -частей сравниваемых объектов, принимающая значения 1 (объекты близки) или 0 (объекты удалены).

Примеры функций близости

$$\mathrm{B}_{\omega}(S_{\mu}, S_{\nu}) = \begin{cases} 1, |\, x_{i\mu} - x_{i\nu} \,| < \varepsilon_i \ \forall i : \omega_i = 1 \\ 0, \text{в противном случае} \end{cases}$$

где $(\varepsilon_1, \ldots, \varepsilon_n)$ - пороговые параметры для различий по соответствующим признакам.

Алгоритмы вычисления оценок

$$B_{\omega}(S_{\mu}, S_{\nu}) = \begin{cases} 1, [\sum \omega_i \mid x_{i\mu} - x_{i\nu} \mid] < \varepsilon \\ 0, \text{в противном случае} \end{cases}$$

, где ε - пороговый параметр.

Оценки близости распознаваемого объекта S_* к эталону S_μ по заданной ω - части. Данная оценка близости формируется на основе введённых ранее функций близости и, возможно, дополнительных параметров

Алгоритмы вычисления оценок

а) $\Gamma_{\omega}(S_*, S_\mu) = \mathrm{B}_{\omega}(S_*, S_\mu)$

б) $\Gamma_{\omega}(S_*, S_\mu) = \mathrm{p}_{\omega}\mathrm{B}_{\omega}(S_*, S_\mu)$, где p_{ω} - параметр, характеризующий информативность опорного множества с характеристическим вектором ω .

в) $\Gamma_{\omega}(S_*, S_\mu) = \gamma_\mu(\sum_{j=1}^{n} p_j\omega_j)\mathrm{B}_{\omega}(S_*, S_\mu)$, где γ_μ - параметр, характеризующий информативность эталонного объекта S_μ , параметры $(p_1, \ldots, p_n)$ характеризуют информативность отдельных признаков.

Алгоритмы вычисления оценок

Оценка объекта S_* за класс K_j по заданной ω - части.

Функция $\Gamma_{\omega}^{j}(S_*, K_j) = \frac{1}{m_i} \sum_{S_\mu \in K_j} \Gamma_{\omega}(S_*, S_\mu)$ является оценкой близости распознаваемого объекта к классу K_j по опорному множеству, характеризуемому бинарным вектором ω.

Оценка объекта S_* за класс K_j. Данная функция вычисляет суммарную степень близости распознаваемого объекта S_* к классу K_j. Приведём обычно используемые выражения:

а) $$\tilde{\Gamma}^{j}(S_*, K_j) = \sum_{\omega \in \Omega_A} \Gamma_{\omega}^{j}(S_*, K_j) \qquad (1)$$

Алгоритмы вычисления оценок

б) $\tilde{\Gamma}^j(S_*, K_j) = \mathrm{v}_j \sum_{\omega \in \Omega_A} \Gamma^j_\omega(S_*, K_j)$, где – “вес” класса K_j

Параметр v_j позволяют регулировать точность распознавания

Прямое вычисление оценок за классы по формуле (1) в случаях, K_j 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.

Алгоритмы вычисления оценок

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 ω - части вычисляются по формуле (а).

Тогда оценка по формуле (1) принимает вид

$$\tilde{\Gamma}^j(S_*, K_j) = \frac{1}{m_i} \sum_{S_\mu \in K_j} \sum_{\omega \in \Omega_A} \mathrm{B}_\omega(S_*, S_\mu)$$

Алгоритмы вычисления оценок

Рассмотрим сумму $\sum_{\omega \in \Omega_A} \mathrm{B}_\omega(S_*, S_\mu)$. Предположим, что общее число признаков, по которым объект S_* близок к объекту S_μ равно $\mathbf{d}(S_*, S_\mu)$. Иными словами $\mathbf{d}(S_*, S_\mu) = |\tilde{D}(S_*, S_\mu)|$, где $\tilde{D}(S_*, S_\mu) = \{t : |x_{*t} - x_{\mu t}| < \varepsilon_t\}$. Очевидно функция близости $\mathrm{B}_\omega(S_*, S_\mu) = 1$ тогда и только тогда, когда опорное множество, задаваемое характеристическим вектором ω , полностью входит в множество $\tilde{D}(S_*, S_\mu)$. Во всех остальных случаях $\mathrm{B}_\omega(S_*, S_\mu) = 0$.

Алгоритмы вычисления оценок

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

$$\tilde{\Gamma}^j(S_*, K_j) = \frac{1}{m_i} \sum_{S_\mu \in K_j} C^k_{\mathbf{d}(S_*, S_\mu)}$$

Алгоритмы вычисления оценок

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

$$\tilde{\Gamma}^j(S_*, K_j) = \frac{1}{m_i} \sum_{S_\mu \in K_j} [2^{\mathbf{d}(S_*, S_\mu)} - 1]$$

Алгоритмы вычисления оценок

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 $S_i \in K_j$ должна выполняться система неравенств

$\tilde{\Gamma}_j(S_i) > \tilde{\Gamma}_{j'}(S_i),$ где $j' \in \{1,\ldots,L\} \setminus j$

Алгоритмы вычисления оценок

В наиболее общем из приведённых выше вариантов модели АВО обучение может быть сведено к поиску максимальной совместной подсистемы системы неравенств

$$\forall S_i \in K_j \cap \tilde{S}_t$$

$$\frac{1}{m_j}\sum_{S_\mu \in K_j}\gamma_\mu\left(\sum_{t=1}^{n} p_t\omega_t\right)\mathrm{B}_{\boldsymbol{\omega}}(S_i,S_\mu) > \frac{1}{m_{j'}}\sum_{S_\nu \in K_{j'}}\gamma_\nu\left(\sum_{t=1}^{n} p_t\omega_t\right)\mathrm{B}_{\boldsymbol{\omega}}(S_i,S_\nu) \quad (2)$$

$$j' \in \{1,\ldots,L\} \setminus j$$

Алгоритмы вычисления оценок

- Поиск максимальной совместной подсистемы системы (2) может производиться с использованием эвристического релаксационного метода, аналогичного тому, что был использован при обучении алгоритма «Линейная машина».