

# МАТЕМАТИЧЕСКИЕ ОСНОВЫ ТЕОРИИ ПРОГНОЗИРОВАНИЯ

Лектор  
Сенько Олег Валентинович  
Лекция 9

# Коллективные методы

Использование различных методов прогнозирования (распознавания), а также различных обучающих выборок или подмножеств признаков позволяет получить набор прогнозирующих (распознающих) алгоритмов:  $A_1, \dots, A_r$

Можно попытаться увеличить обобщающую способность за счёт выбора алгоритма с минимальной оценкой ошибки прогнозирования. Однако нередко более эффективной процедурой является вычисление прогноза с использованием всех алгоритмов из  $A_1, \dots, A_r$ .

# Коллективные методы

Использование коллектива (ансамбля) алгоритмов, которые строятся с помощью различных методов позволяет использовать при прогнозировании различные принципы экстраполяции, лежащих в основе этих методов.

Статистическое обоснование использованию ансамбля алгоритмов даёт анализ ошибки выпуклой комбинации прогнозов, вычисляемых членами ансамбля.

Предположим, что алгоритмы ансамбля  $A_1, \dots, A_r$  вычисляют прогноз переменной  $Y$ .

# Коллективные методы

Пусть  $f_i$  - прогноз, вычисляемый алгоритмом  $A_i$

$\Delta_i = E_{\Omega}(Y - f_i)^2$  - ошибка прогноза, вычисляемого  $A_i$ ,

$i = 1, \dots, r$ . Введём обозначение  $\rho_{i' i''} = E_{\Omega}(f_{i'} - f_{i''})^2$  - математическое ожидание квадрата отклонения друг от друга прогнозов, вычисляемых алгоритмами  $A_{i'}$  и  $A_{i''}$ .

Пусть  $c_1, \dots, c_r$  - положительные коэффициенты такие, что

$$\sum_{i=1}^r c_i = 1$$

# Коллективные методы

Обозначим через  $\hat{f}$  выпуклую комбинацию прогнозов, вычисляемых алгоритмами ансамбля  $A_1, \dots, A_r$  :  $\hat{f} = \sum_{i=1}^r c_i f_i$

Для ошибки выпуклой комбинации справедливо выражение

$$\hat{\Delta} = E_{\Omega} (Y - \hat{f})^2 = \sum_{i=1}^r c_i \Delta_i - \frac{1}{2} \sum_{i'=1}^r \sum_{i''=1}^r c_{i'} c_{i''} \rho_{i'i''}$$

Принимая во внимание, что все отклонения  $\rho_{i'i''}$  всегда неотрицательны, а коэффициенты  $c_1, \dots, c_r$  положительны

получаем неравенство  $\hat{\Delta} \leq \sum_{i=1}^r c_i \Delta_i$ .

# Коллективные методы

Рассмотрим, случай, когда все алгоритмы участвуют в построении коллективного решения равноправно. В этом случае

$$c_i = \frac{1}{r}, i = 1, \dots, m \quad \hat{f} = \frac{1}{m} \sum_{i=1}^r f_i$$

$$\hat{\Delta} = E_{\Omega} (Y - \hat{f})^2 = \frac{1}{m} \sum_{i=1}^r \Delta_i - \frac{1}{2} \frac{1}{m^2} \sum_{i=1}^r \sum_{i''=1}^r \rho^{i i''}$$

Таким образом, ошибка коллективного метода, вычисляющего средний прогноз по ансамблю равна средней ошибке по сем членам ансамбля минус средний квадрат отклонений прогнозов между участниками ансамбля.

# Комитетные методы в распознавании

Рассмотрим сначала несколько простейших эвристических методов принятия коллективных решений.

Предположим, что у нас есть ансамбль алгоритмов распознавания  $A_1, \dots, A_r$ , которые были использованы для классификации некоторого объекта  $s^*$ . Простейшим комитетным методом является метод голосования по большинству, относящий объект к тому классу, к которому он был присвоен относительным большинством алгоритмов.

# Комитетные методы в распознавании

Напомним, что произвольный распознающий алгоритм является комбинацией распознающего оператора, вычисляющего оценки за классы и решающего правила, производящего классификацию по оценкам, вычисленным распознающим оператором. Предположим, что  $\Gamma_l^i(s^*)$  - оценка за класс  $K_l$ , вычисляемая алгоритмом  $A_l$ . Коллективное решение может строиться путём вычисления коллективных оценок за классы через оценки  $\Gamma_l^i(s^*)$ , соответствующие отдельным алгоритмам.



# Комитетные методы в распознавании

1) Коллективная оценка за класс вычисляется как среднеарифметическое оценок

$$\Gamma_l^{av}(s^*) = \frac{1}{r} \sum_{i=1}^r \Gamma_l^i(s^*)$$

2) Коллективная оценка вычисляется как минимум всех оценок за данный класс полученных разными алгоритмами

$$\Gamma_l^{\min}(s^*) = \min_{i \in \{1, \dots, r\}} [\Gamma_l^i(s^*)]$$

# Комитетные методы в распознавании

3) Коллективная оценка вычисляется как вычисляется как максимум всех оценок за данный класс полученных разными алгоритмами

$$\Gamma_l^{\max}(s^*) = \max_{i \in 1, \dots, r} [\Gamma_l^i(s^*)]$$

4) Еще одним употребительным способом построения комитетного решения является произведение оценок

$$\Gamma_l^{pr}(s^*) = \prod_{i=1}^r [\Gamma_l^i(s^*)]$$

# Комитетные методы в распознавании

К достоинствам комитетных методов относится их простота, высокая быстродействие. Для применения этих методов не требуется никакой дополнительной процедуры обучения, что позволяет сразу переходить к распознаванию объектов комитетом обученных алгоритмов.

Подобными же достоинствами обладает другой известный метод построения коллективных решений – «Наивный байесовский классификатор».

# Наивный байесовский классификатор

«Наивный байесовский классификатор» является статистическим методом, основанном на оценках вероятностей принадлежности объекта классам в зависимости от результатов классификации отдельными алгоритмами.

Пусть для каждого из распознающих алгоритмов  $A_1, \dots, A_r$  известна матрица оценок условных вероятностей

$$\| \hat{\mathbf{P}}(s^* \in K_{l'} | A_i(s^*) = "s^* \in K_{l'}") \|_{L \times L}$$

# Наивный байесовский классификатор

Предположим, что алгоритмы  $A_1, \dots, A_r$  отнесли объект  $s^*$  в классы  $K_{t_1}, \dots, K_{t_r}$  соответственно.

Для вычисления коллективной оценки  $\Gamma_l^{NB}(s^*)$  объекта  $s^*$  за класс  $K_l$  формально принимается гипотеза о независимости классификаторов  $A_1, \dots, A_r$ . В результате коллективная оценка вычисляется как произведение оценок, соответствующих отдельным классификаторам

$$\Gamma_l^{NB}(s^*) = \prod_{i=1}^r \hat{\mathbf{P}}(s^* \in K_l \mid A_i(s^*) = "s^* \in K_{t_i}")$$

# Логическая коррекция

Комитетные методы и наивный байесовский классификатор являются простейшими методами коллективной коррекции, не учитывающих взаимодействие алгоритмов ансамбля или их относительную эффективность.

Требование повышения обобщающей способности ансамбля за счёт более полного учёта его структуры и использования возможностей лежащих в его основе эвристик привело к созданию средств алгебраической и логической коррекции.

Методы логической коррекции учитывают только окончательные результаты классификации.

# Логическая коррекция

Пусть у нас имеется некоторая выборка  $\tilde{S}_q = \{s_1, \dots, s_q\}$  объектов, принадлежащих классам  $K_1, \dots, K_L$ , по которой мы собираемся произвести коррекцию. Данной выборке может быть сопоставлена информационная матрица  $\|\alpha_{lj}\|_{L \times q}$ , где  $\alpha_{lj} = 1$ , если  $s_j \in K_l$  и  $\alpha_{lj} = 0$  в противном случае. Иными словами  $\alpha_{lj}$  является значением предиката  $P_l = "s \in K_l"$  на объекте  $s_j$ .

И набор матриц  $\|\beta_{lj}^i\|_{L \times q}$ , где  $\beta_{lj}^i$  значение предиката  $P_l$  на объекте  $s_j$ , вычисленное алгоритмом  $A_i$ .

,

# Логическая коррекция

Поиск оптимального логического корректора сводится к поиску такой логической функции от  $r$  переменных  $F(z_1, \dots, z_r)$ , чтобы равенство  $F(\beta_{lj}^{g(1)}, \dots, \beta_{lj}^{g(r)}) = \alpha_{lj}$  выполнялось для возможно большего числа объектов обучающей выборки, где перестановочная функция  $g(i)$  устанавливает связь между переменными  $z_1, \dots, z_r$  и алгоритмами  $A_1, \dots, A_r$ . В том случае, когда  $2^r > q$  и отсутствуют противоречия типа равенства функции  $F$  при одних и тех же значениях аргументов разным значениям элементов информационной матрицы, задача построения логического корректора сводится к задаче доопределения логической функции естественным путём заданной на выборке  $\tilde{S}_q$  на весь единичный куб  $E^r$ .



# Логическая коррекция

Приведем в качестве примера «монотонный логический корректор», в основу которого положена следующая идея. В исходном наборе  $A_1, \dots, A_r$  для каждого класса  $K_i$  выбирается поднабор  $A_{t_1}, \dots, A_{t_k}$ . Объект  $s$  относится монотонным логическим корректором в класс  $K_i$  в том и только в том случае, если он отнесён в  $K_i$  всеми алгоритмами из  $A_{t_1}, \dots, A_{t_k}$  и ещё одним алгоритмом из набора  $A_1, \dots, A_r$ , который не принадлежит  $A_{t_1}, \dots, A_{t_k}$ .

# Логическая коррекция

Построение монотонного логического корректора, правильно классифицирующего объекты выборки  $\tilde{S}_q$  сводится к построению монотонной булевой функции, для которой

$$F(\beta_{lj}^{g(1)}, \dots, \beta_{lj}^{g(r)}) = \alpha_{lj} \quad \text{для всех объектов } \tilde{S}_q .$$

# Алгебраическая коррекция

Универсальным способом построения оптимального распознающего алгоритма по набору исходных алгоритмов  $A_1, \dots, A_r$  является использование алгебраических методов коррекции. В отличие от логических методов коррекции алгебраические методы используют не только окончательные результаты классификации, содержащиеся в матрицах  $\|\beta_{lj}^i\|_{L \times q}$ , но также матрицы оценок  $\|\gamma_{lj}^i\|_{L \times q}$ , вычисляемые операторами  $R_1, \dots, R_r$ , где  $\gamma_{lj}^i = \Gamma_l^i(s_j)$  - оценка объекта  $s_j \in \tilde{S}_q$  за класс, вычисляемая оператором  $R_i$ ,  $i = 1, \dots, r$ ,  $j = 1, \dots, q$ ,  $l = 1, \dots, L$ .

# Алгебраическая коррекция

Основы теории алгебраической коррекции были разработаны Ю.И.Журавлёвым в 1976-1978 годах.

Задача распознавания в алгебраической теории рассматривается как задача построения по начальной информации  $I$  о классах  $K_1, \dots, K_L$  для предъявленной для распознавания выборки  $\tilde{S}_q = \{s_1, \dots, s_q\}$  информационной матрицы  $\|\alpha_{lj}\|_{L \times q}$ . Обозначим данную задачу как задачу  $Z(I, \tilde{S}_q, P_1, \dots, P_L)$  или просто задачу  $Z$ .

Примером начальной информации о классах является таблица признаков описаний эталонных объектов классов и их информационная матрица.

# Алгебраическая коррекция

Предположим, что у нас имеется множество алгоритмов  $\{A\}$ , переводящих пару  $\{I, \tilde{S}_q\}$  в матрицы  $\|\beta_{ij}^i\|_{L \times q}$ , составленные из элементов  $\{0, 1, \Delta\}$ , где значения  $0$  и  $1$  как и раньше являются значениями предикатов, вычисленными алгоритмами из множества  $\{A\}$ , значение  $\Delta$  соответствует отказу от вычисления значения предиката.

**Определение.** Алгоритм  $A$  называется корректным для задачи  $Z$ , если выполнено равенство  $A(I, \tilde{S}_q, P_1, \dots, P_L) = \|\alpha_{ij}\|_{L \times q}$ .

# Алгебраическая коррекция

Алгоритм, не являющийся корректным для задачи  $Z$ , называется некорректным. Совокупность  $\{A\}$  состоит из вообще говоря некорректных алгоритмов.

Алгебраический подход к решению задач распознавания включает в себя введение алгебраических операций над алгоритмами из  $\{A\}$ , позволяющих строить корректные алгоритмы по наборам алгоритмов из  $\{A\}$ . Поскольку каждый распознающий алгоритм может быть представлен как последовательное выполнение распознающего оператора и решающего правила, множеству  $\{A\}$  соответствуют множества операторов  $\{R\}$  и множество решающих правил  $\{C\}$ .

# Алгебраическая коррекция

Каждый из операторов из множества  $\{R\}$  вычисляет для задачи  $Z$  матрицу оценок за классы  $R^*(I, \tilde{S}_q) = \|\gamma_{ij}^*\|_{L \times q}$

На множестве операторов  $\{R\}$  вводятся операции сложения, умножения и умножения на скаляр.

Пусть  $R', R'' \in \{R\}$   $R'(I, \tilde{S}_q) = \|\gamma'_{ij}\|_{L \times q}$   $R''(I, \tilde{S}_q) = \|\gamma''_{ij}\|_{L \times q}$

$b$  скалярная величина.

Определим операторы  $b \bullet R'$  (умножение на скаляр),  $R' + R''$  (сложение),  $R' \bullet R''$  (умножение) следующим образом.

# Алгебраическая коррекция

$$(b \bullet R')(I, \tilde{S}_q) = \| b * \gamma'_{ij} \|_{L \times q} \quad (1)$$

$$(R' + R'')(I, \tilde{S}_q) = \| \gamma'_{ij} + \gamma''_{ij} \|_{L \times q} \quad (2)$$

$$(R' \bullet R'')(I, \tilde{S}_q) = \| \gamma'_{ij} * \gamma''_{ij} \|_{L \times q} \quad (3)$$

Использование операций (1)-(3) позволяет строить новые распознающие операторы, являющиеся полиномами от операторов из исходного множества вида  $\sum_{i=1}^{N_p} a_i R_{t(1,i)} \bullet \dots \bullet R_{t(k_i,i)}$

Функция  $t(j,i)$  - указывает на оператор, являющийся  $j$ -ым сомножителем в  $i$ -ом слагаемом полинома.



# Алгебраическая коррекция

Очевидно, что замыкание  $\mathbf{L}\{R\}$  множества операторов  $\{R\}$  относительно операций (1) и (2) является линейным векторным пространством. Обозначим через  $\mathbf{U}\{R\}$  замыкание множества  $\{R\}$

относительно операций (1)-(3) - алгебраическое замыкание.

$$Z(I, \tilde{S}_q, P_1, \dots, P_L)$$

Рассмотрим условия, существования корректного алгоритма для некоторой задачи .

# Алгебраическая коррекция

**Определение 2.** Если множество матриц  $\{R(I, \tilde{S}_q)\}$  (операторы  $R$  пробегают множество  $\tilde{\mathbf{R}}$ ) содержит базис в пространстве числовых матриц размерности  $L \times q$ , то задача  $Z(I, \tilde{S}_q, P_1, \dots, P_L)$  называется полной относительно  $\tilde{\mathbf{R}}$ .

**Определение 3.** Решающее правило  $C$  называется корректным, если для всякой выборки длины  $q$  существует хотя бы одна числовая матрица  $\|\gamma_{ij}\|_{L \times q}$  такая, что  $C(\|\gamma_{ij}\|_{L \times q}) = \|\alpha_{ij}\|_{L \times q}$

# Алгебраическая коррекция

Пусть  $\{A\}$  - множество алгоритмов вида  $A = R \otimes C^*$ , где  $R \in \{R\}$   
 $C^*$  - некоторое корректное решающее правило.

**Определение** Множества алгоритмов вида  $A = R \otimes C^*$  будут обозначаться  $\mathbf{L}\{A\}$  и  $\mathbf{U}\{A\}$ , если  $R \in \mathbf{L}\{R\}$  и  $R \in \mathbf{U}\{R\}$  соответственно.

**Теорема 1** Если множество  $\{Z\}$  состоит лишь из задач, полных относительно  $\tilde{\mathbf{R}}$ , то линейное замыкание  $\mathbf{L}\{\tilde{\mathbf{R}} \otimes C^*\}$ , где  $C^*$  - некоторое корректное решающее правило, является корректным относительно  $\{Z\}$

# Алгебраическая коррекция

**Доказательство.** При фиксированном  $q$  базис в пространстве числовых матриц размерности  $L \times q$  состоит из  $L * q$  матриц  $M_1, \dots, M_{L*q}$ . Тогда существуют числа  $c_1, \dots, c_{L*q}$  такие,

что  $M = \sum_{i=1}^{L*q} c_i * M_i$  где  $M$  является матрицей, которая может быть переведена решающим правилом  $C^*$  в информационную матрицу  $\| \alpha_{lj} \|_{L \times q}$ . Существование матрицы  $M$  следует из Корректности решающего правила  $C^*$ .

# Алгебраическая коррекция

Представление (4) возможно в силу того, что матрицы  $M_1, \dots, M_{L^*q}$  образуют базис в пространстве числовых матриц размерности  $L \times q$ . В том случае, если матрицы  $M_1, \dots, M_{L^*q}$  построены из  $\{I, \tilde{S}_q\}$  с помощью операторов  $R_1, \dots, R_{L^*q}$  из  $\tilde{\mathbf{R}}$ , корректный алгоритм может

быть представлен в виде  $\tilde{A} = \left( \sum_{i=1}^{L^*q} c_i R_i \right) \otimes C^*$ . Теорема доказана.

# Алгебраическая коррекция

**Следствие 1.** Пусть  $\{A\}$  - совокупность некорректных алгоритмов,

$\tilde{\mathbf{R}}$  - соответствующее множество операторов,  $C^*$  - некоторое фиксированное корректное решающее правило. Тогда

$\mathbf{L}\{A\} = \mathbf{L}\{\tilde{\mathbf{R}} \otimes C^*\}$  является корректным относительно  $\{Z\}$ ,  
если  $\{Z\}$

состоит из задач, полных относительно  $\tilde{\mathbf{R}}$ .

# Алгебраическая коррекция

Следствие 2. Пусть выполнены все условия следствия 1 и, кроме того,  $[\tilde{\mathbf{R}}]$  есть замыкание  $\tilde{\mathbf{R}}$  относительно операций (1) – (3). Тогда  $\mathbf{U}\{A\} = \{\mathbf{U}\{\tilde{\mathbf{R}}\} \otimes C^*\}$  является корректным относительно  $\{Z\}$ , если  $\{Z\}$  состоит из задач, полных относительно  $[\tilde{\mathbf{R}}]$

# Алгебраическая коррекция

Линейные и алгебраические замыкания могут строиться не только над конечными наборами заранее обученных алгоритмов, но также и над множествами алгоритмов, принадлежащих некоторой модели и имеющих в общем случае мощность континуума. Рассмотрим в качестве примера рассмотрим один из вариантов модели алгоритмов вычисления оценок, в котором оценки за классы вычисляются по формуле

$$\Gamma_i(s^*) = \sum_{s_\mu \in K_i} \sum_{\omega \in \Omega} \left( \sum_{j=1}^n p_j \omega_j \right) B_\omega(s^*, s_\mu, \boldsymbol{\varepsilon}) + \sum_{s_\mu \notin K_i} \sum_{\omega \in \Omega} \left( \sum_{j=1}^n p_j \omega_j \right) \bar{B}_\omega(s^*, s_\mu, \boldsymbol{\varepsilon}) \quad (5)$$



# Алгебраическая коррекция

Здесь  $\bar{B}_\omega(s^*, s_\mu, \boldsymbol{\varepsilon}) = 1 - B_\omega(s^*, s_\mu, \boldsymbol{\varepsilon})$  является функцией антиблизости объекта  $s^*$  к эталону  $s_\mu$  по опорному множеству,

описываемому бинарным характеристическим вектором

$\boldsymbol{\omega} = (\omega_1, \dots, \omega_n)$  ,  $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)$  - вектор положительных пороговых коэффициентов, задающих близость объектов по каждому из признаков,  $(p_1, \dots, p_n)$  - вектор положительных параметров, характеризующих важность признаков,  $(\gamma_1, \dots, \gamma_n)$  - вектор положительных параметров, характеризующих важность признаков

# Алгебраическая коррекция

Для того, чтобы описать условия существования корректного алгоритма в алгебраическом замыкании подмножества алгоритмов вычисления оценок введём дополнительные определения. Пусть  $\tilde{M}$  - некоторое множество допустимых объектов.

**Определение** Объекты  $s_v$  и  $s_u$  с описаниями  $(x_1^u, \dots, x_n^u)$  и  $(x_1^v, \dots, x_n^v)$  называются изоморфными относительно множества  $\tilde{M}$ , если  $\forall s \in \tilde{M}$  с описанием  $(a_1, \dots, a_n)$  выполняются равенство

$$(|x_1^u - a_1|, \dots, |x_n^u - a_n|) = (|x_1^v - a_1|, \dots, |x_n^v - a_n|)$$

# Алгебраическая коррекция

Нетрудно видеть что корректный алгоритм в рамках модели

Вычисления оценок не может существовать для задачи

$Z(I, \tilde{S}_q, P_1, \dots, P_L)$  случаях, когда

а) в выборке  $\tilde{S}_q$  существует два объекта  $s'$  и  $s''$ , изоморфных относительно выборки эталонов  $\tilde{S}_m$ , которая вместе со своей информационной матрицей образует начальную информацию  $I$ ;

б) объекты  $s'$  и  $s''$  принадлежат двум непересекающимся классам.

# Алгебраическая коррекция

Действительно в этом случае векторы оценок  $(\Gamma_1(s'), \dots, \Gamma_l(s'))$  и  $(\Gamma_1(s''), \dots, \Gamma_l(s''))$ , вычисляемые произвольным оператором из модели АВО будут одинаковы. Следовательно никакое множество операторов модели АВО не может вычислять базис в пространстве вещественных матриц размера  $L \times q$

Будем называть задачу  $Z(I, \tilde{S}_q, P_1, \dots, P_L)$  регулярной, если

а) никакие два класса полностью не совпадают, т.е.  $K_{l'} \neq K_{l''}$  при  $l' \neq l''$  ;

# Алгебраическая коррекция

б) никакие два объекта из  $\tilde{\mathcal{S}}_q$  не являются изоморфными относительно выборки эталонов  $\tilde{\mathcal{S}}_m$ , где  $I = \{\tilde{\mathcal{S}}_m, \|\alpha_{lj}\|_{L \times q}\}$  ;  $\tilde{\mathcal{S}}_q \cap \tilde{\mathcal{S}}_m = \emptyset$

в) .

Справедлива теорема.

**Теорема 2.** Алгебраическое замыкание подкласса алгоритмов модели АВО, в которой оценки за классы вычисляются по формуле (5) корректно над множеством регулярных задач.