СПЕЦКУРС
Логический анализ данных в распознавании
(Logical data analysis in recognition)

лектор д.ф.-м.н. Елена Всеволодовна Дюкова

Спецкурс посвящён вопросам применения аппарата дискретной математики в задачах интеллектуального анализа данных. Излагаются общие принципы, лежащие в основе логического подхода к задачам машинного обучения. Описываются методы конструирования процедур классификации по прецедентам с использованием понятий теории булевых функций и теории покрытий булевых матриц. Рассматриваются основные модели логических процедур классификации, вопросы сложности их реализации и качества решения прикладных задач.

Спецкурс для бакалавров 2-4 курсов ВМК МГУ им. М.В. Ломоносова.

По спецкурсу издано учебное пособие:
http://www.ccas.ru/frc/papers/djukova03mp.pdf

Лекция 11
Методы повышения эффективности дискретных процедур распознавания

- Одним из подходов к повышению качества распознавания и снижению вычислительных затрат является проведение предварительного анализа обучающей информации. Целью такого анализа является оценка основных информативных характеристик обучающей выборки, в частности, оценка информативности признаков и выделение «шумящих» признаков, а также выделение наиболее представительных (типичных) для своего класса обучающих объектов.

1. Методы оценки информативности признаков

• Традиционно в качестве меры важности признака $\boldsymbol{x_j}$, $\boldsymbol{j} \in \{\boldsymbol{1}, \boldsymbol{2}, \dots, \boldsymbol{n}\}$, рассматривалась величина

$$\boldsymbol{I_j} = \left|\boldsymbol{C_j^A}\right| / \left|\boldsymbol{C^A}\right|,$$

где $\boldsymbol{C_j^A}$ - подмножество таких элементарных классификаторов из $\boldsymbol{C^A}$, в которых содержится признак $\boldsymbol{x_j}$. При этом в качестве $\boldsymbol{A}$ использовался тестовый алгоритм или алгоритм голосования по представительным наборам.

• Во многих случаях такой способ оценки информативности признаков даёт не очень хорошие результаты. Например, если признак является тупиковым тестом, тогда он входит лишь в один тупиковый тест и согласно формуле для $\boldsymbol{I_j}$ имеет очень маленький вес. Однако, очевидно, что указанный признак является важным и обладает большой информативностью.

• С другой стороны, многозначные признаки, являющиеся малоинформативными «шумящими» признаками порождают большое число тестов (представительных наборов) и, следовательно, имеют большой вес $\boldsymbol{I_j}$.

• Пусть **A** - алгоритм голосования по представительным наборам. Частично проблему шумящих признаков можно решить, введя порог $\boldsymbol{p_{min}}$ минимальной встречаемости представительного набора в обучающей выборке. То есть рассматривать только те представительные наборы, которые в обучающей выборке встречаются не менее $\boldsymbol{p_{min}}$ раз. Используя это дополнительное условие, шумящие многозначные признаки не получают самой большой оценки по информативности. Однако не всегда можно взять достаточно большой порог $\boldsymbol{p_{min}}$ из-за того, что структура класса такова, что каждый представительный набор встречается не очень большое число раз. В данном случае предлагается следующее.

- Положим $\overline{\boldsymbol{K}} = \{\boldsymbol{K_1}, \dots, \boldsymbol{K_l}\} \setminus \boldsymbol{K}$. Зададим целые числа $\boldsymbol{p}$ и $\boldsymbol{q}$ такие, что
- $\boldsymbol{1 \le p \le \min\limits_{1 \le i \le l} |K_i|} \quad \boldsymbol{0 \le q \le \min\limits_{1 \le i \le l} |\overline{K}_i|}, \quad \boldsymbol{p > q}$.
- Элементарный классификатор $(\boldsymbol{\sigma}, \boldsymbol{H})$ назовём $(\boldsymbol{p}, \boldsymbol{q})$ - представительным набором для класса $\boldsymbol{K_i}$, если не менее чем для $\boldsymbol{p}$ объектов $\boldsymbol{S'}$ из класса $|\boldsymbol{K_i}|$ справедливо $\boldsymbol{B(\sigma, S', H) = 1}$ и не более чем для $\boldsymbol{q}$ объектов $\boldsymbol{S''}$ из $\overline{\boldsymbol{K}}_{\boldsymbol{i}}$ справедливо $\boldsymbol{B(\sigma, S'', H) = 1}$. В частности, представительный набор является **(1,0)**-представительным набором.

- Пусть $(\boldsymbol{\sigma}, \boldsymbol{H})$ является $(\boldsymbol{p}, \boldsymbol{q})$-представительным набором класса $\boldsymbol{K}$, $\boldsymbol{p_\sigma}$- его встречаемость в обучающей выборке в классе $\boldsymbol{K}$, $\boldsymbol{q_\sigma}$ - встречаемость в остальных классах $(\boldsymbol{p_\sigma \ge p}, \boldsymbol{q_\sigma \le q})$. Тогда информативный вес представительного набора $(\boldsymbol{\sigma}, \boldsymbol{H})$ будем вычислять по следующей формуле

$$\boldsymbol{\nu(\sigma, H) = (p_\sigma - p) + (q - q_\sigma)}.$$

- За счет варьирования $\boldsymbol{p}$ и $\boldsymbol{q}$ можно существенно снизить влияние шумящих признаков. Дело в том, что разрешив представительным наборам встречаться в других классах (увеличив $\boldsymbol{q}$), можно увеличить минимальную встречаемость в своём классе (увеличить $\boldsymbol{p}$). В практической реализации для снижения вычислительных затрат используется следующий приём. Множество $(\boldsymbol{p}, \boldsymbol{q})$ -представительных наборов строится по случайной подвыборке, которая составляет **60 - 70%** обучающей выборки. После этого осуществляется проверка, являются ли построенные представительные наборы $(\boldsymbol{p}, \boldsymbol{q})$ -представительными наборами для всей обучающей выборки, и в случае положительного результата вычисляются их информативные веса. Далее вычисляется информативный вес каждого признака как отношение суммы весов $(\boldsymbol{p}, \boldsymbol{q})$ - представительных наборов, которые он порождает, к сумме весов всех $(\boldsymbol{p}, \boldsymbol{q})$ - представительных наборов для данного класса.

• Для решения проблемы шумящих признаков не обязательно добиваться того, чтобы оценка их информативности была низкой на любой подвыборке. Достаточно, чтобы оценка информативности шумящих признаков при различных разбиениях была неустойчива по разбиению информации на две подвыборки, при том, что по остальным признакам наблюдается некоторая устойчивость.

• Достаточно эффективным и не требующим больших вычислительных затрат является предлагаемый ниже метод оценки информативности признаков и отдельных значений признаков, основанный на вычислении близости между парами объектов по отдельным признакам.

• Пусть $\boldsymbol{S' \in K_i}$, $\boldsymbol{i \in \{1, 2, \dots, l\}}$, $\boldsymbol{j \in \{1, 2, \dots, n\}}$. Положим

$$\boldsymbol{\mu_{ij}^{(1)}(S') = \frac{1}{|K_i|} \sum_{S'' \in K_i} B\left(S', S'', \{x_j\}\right)},$$

$$\boldsymbol{\mu_{ij}^{(2)}(S') = \frac{1}{|\overline{K_i}|} \sum_{S'' \in \overline{K}_i} B\left(S', S'', \{x_j\}\right)},$$

• Величины $\boldsymbol{\mu_{ij}^{(1)}(S')}$ и $\boldsymbol{\mu_{ij}^{(2)}(S')}$ характеризуют близость объекта $\boldsymbol{S'}$ соответственно к своему классу и к другим классам по признаку $\boldsymbol{x_j}$. Величину

$$\boldsymbol{\mu_{ij}(S') = \mu_{ij}^{(1)}(S') - \mu_{ij}^{(2)}(S')}$$

назовём весом значения признака $\boldsymbol{x_j}$, для объекта $\boldsymbol{S'}$. Будем говорить, что значение признака $\boldsymbol{x_j}$ является типичным для объекта $\boldsymbol{S'}$, если $\boldsymbol{\mu_{ij}(S') > \mu}$, где $\boldsymbol{\mu}$ - порог информативности значения признака, $\boldsymbol{-1 < \mu < 1}$.

• Например, можно положить $\boldsymbol{\mu = 0}$. Тогда значение признака будет являться типичным для класса, если в этом классе оно встречается чаще, чем в остальных.

• Множество типичных значений признаков в таблице обучения образует так называемую информативную зону. Далее при построении множества элементарных классификаторов имеет смысл анализировать только те значения признаков, которые попадают в информативную зону, тем самым уменьшается перебор при построении распознающего алгоритма. Кроме того, в информативную зону не попадают значения признаков, которые очень редко встречаются во всех классах (шумящие). Поэтому использование информативной зоны позволяет также снизить влияние шумящих признаков.

• Исследование типичности значений признаков позволяет также оценить сложность (в смысле возможности построения качественного алгоритма распознавания) решаемой задачи.

2. Выделение типичных объектов в классе

- При решении прикладной задачи распознавания интересно попытаться оценить эффективность построенного алгоритма при распознавании объектов, не входящих в обучающую выборку. Например, воспользоваться хорошо известным методом скользящего контроля. К сожалению, в ряде прикладных задач алгоритмы не всегда показывают достаточно высокую эффективность. Такая ситуация возникает, когда классы плохо отделяются друг от друга (в каждом классе есть много объектов, описания которых похожи на объекты, не принадлежащие данному классу). В этом случае построенный алгоритм зачастую, хотя и хорошо распознает "известные" ему объекты (объекты, которые участвовали в построении алгоритма), но плохо распознает "новые" объекты. В данном разделе предлагается подход, позволяющий повысить качество распознающих алгоритмов за счёт выделения "типичных" для своих классов объектов. Этот подход продемонстрирован ниже на примере модели голосования по представительным наборам.

- Объекты, лежащие на границе между классами, плохо распознаются и, по-видимому, они не позволяют строить короткие представительные наборы. Пусть описание обучающего объекта, не принадлежащего классу ***K***, похоже на описания некоторых объектов из ***K***. Тогда данный объект "лишает" класс ***K*** некоторого множества коротких представительных наборов, и это существенно снижает эффективность алгоритма. Для решения указанной проблемы предлагается разбить обучающую выборку на две подвыборки, по первой (базовой) построить множество представительных наборов, по второй (контрольной) вычислить их веса. Причем разбить нужно таким образом, чтобы объекты, находящиеся на границе между классами, попали в контрольную подвыборку, а все остальные (типичные) объекты - в базовую подвыборку. Практические эксперименты на прикладных задачах подтверждают гипотезу о том, что такое разбиение увеличивает число коротких представительных наборов и тем самым позволяет повысить качество алгоритма распознавания.

• Описанный метод позволяет очень быстро оценить типичность обучающих объектов и отдельных их фрагментов по отношению к своим классам. Его недостатком является то, что информативность (типичность) значения признака вычисляется независимо от других признаков. Не учитывается тот факт, что, вообще говоря, фрагмент описания некоторого объекта может быть типичен для одного из классов (в этом классе данный фрагмент встречается значительно чаще, чем в остальных классах), но при этом значения признаков, которые составляют этот фрагмент, встречаются в разных классах примерно одинаковое число раз и не являются типичными ни для одного из классов. Кроме того, необходимо настраивать параметры $\boldsymbol{\mu}$ и $\boldsymbol{p}$, что не очень удобно.

• Ниже предлагается метод выделения типичных объектов на основе проведения процедуры скользящего контроля, который заключается в следующем.

• Из обучающей выборки исключается один из объектов, например объект $\boldsymbol{S_i}$, $\boldsymbol{i} \in \{\mathbf{1}, \mathbf{2}, \dots, \boldsymbol{m}\}$. По подвыборке $\{\boldsymbol{S_1}, \dots, \boldsymbol{S_m}\} \backslash \boldsymbol{S_i}$ строится распознающий алгоритм (например, используется модель алгоритма голосования по представительным наборам в классическом варианте). Далее этот алгоритм применяется для распознавания объекта $\boldsymbol{S_i}$. Объект $\boldsymbol{S_i}$ считается типичным для своего класса, если построенный алгоритм распознал его правильно, и нетипичным для своего класса, если алгоритм отнёс его к другому классу или отказался от распознавания. Описанная процедура повторяется для всех объектов обучающей выборки.

• Очевидно, что для больших задач процедура скользящего контроля в алгоритме голосования по представительным наборам требует существенных вычислительных затрат. Существует быстрый способ вычисления требуемых оценок, позволяющий сократить время счета примерно в $\boldsymbol{m}$ раз (см. *Дюкова Е.В., Песков Н.В. Построение распознающих процедур на базе элементарных классификаторов // Математические вопросы кибернетики. 2005. № 14. С. 57-92*, http://library.keldysh.ru/mvk.asp?id=2005-57).

3. Вычисление информативности эл.кл.

• Способ вычисления информативности эл.кл. проиллюстрируем на примере представительного набора. Разделим обучающую выборку на две подвыборки: базовую и контрольную. По базовой подвыборке построим множество представительных наборов. Сопоставим каждому построенному представительному набору некий вес, который вычислим по контрольной подвыборке.

• Пусть $(\boldsymbol{\sigma}, \boldsymbol{H})$ - представительный набор класса $\boldsymbol{K}$, $\boldsymbol{K} \in \{\boldsymbol{K_1}, \dots, \boldsymbol{K_l}\}$, порождаемый объектом из базовой выборки, и пусть $\boldsymbol{N}(\boldsymbol{K}, \boldsymbol{\sigma}, \boldsymbol{H})$ - число объектов в контрольной выборке, за которых данный представительный набор голосует "правильно", $\bar{\boldsymbol{N}}(\boldsymbol{K}, \boldsymbol{\sigma}, \boldsymbol{H})$ - число объектов в контрольной выборке, за которых он голосует "неправильно". Тогда в качестве веса представительного набора $(\boldsymbol{\sigma}, \boldsymbol{H})$ можно рассматривать, например, следующие величины:

$$\boldsymbol{\nu_1}(\boldsymbol{\sigma}, \boldsymbol{H}) = \boldsymbol{N}(\boldsymbol{K}, \boldsymbol{\sigma}, \boldsymbol{H}) ;$$

$$\boldsymbol{\nu_2}(\boldsymbol{\delta}, \boldsymbol{H}) = \frac{\boldsymbol{1} + \boldsymbol{N}(\boldsymbol{K}, \boldsymbol{\sigma}, \boldsymbol{H})}{\boldsymbol{1} + \bar{\boldsymbol{N}}(\boldsymbol{K}, \boldsymbol{\sigma}, \boldsymbol{H})} .$$

- Принадлежность объекта $\boldsymbol{S}$ классу $\boldsymbol{K}$ будем оценивать величиной

$$\Gamma_{\boldsymbol{i}}(\boldsymbol{S},\boldsymbol{K}) = \frac{\mathbf{1}}{|\boldsymbol{C^A}(\boldsymbol{K})|} \sum_{(\boldsymbol{\sigma},\boldsymbol{H})\in \boldsymbol{C^A}(\boldsymbol{K})} \boldsymbol{\nu_i}(\boldsymbol{\sigma},\boldsymbol{H})\boldsymbol{B}(\boldsymbol{\sigma},\boldsymbol{S},\boldsymbol{H}),$$

- В качестве информативного веса признака $\boldsymbol{x_j}$ будем рассматривать величину

$$\boldsymbol{I_j} = \frac{\sum_{(\boldsymbol{\sigma},\boldsymbol{H})\in \boldsymbol{C^A}(\boldsymbol{K}),\, \boldsymbol{x_j}\in \boldsymbol{H}} \boldsymbol{\nu}(\boldsymbol{\sigma},\boldsymbol{H})}{\sum_{(\boldsymbol{\sigma},\boldsymbol{H})\in \boldsymbol{C^A}(\boldsymbol{K})} \boldsymbol{\nu}(\boldsymbol{\sigma},\boldsymbol{H})}$$

- В общем случае функция, по которой вычисляется вес представительного набора, должна обладать следующим свойством. Эта функция должна монотонно возрастать по числу объектов из контрольной выборки, за которые представительный набор голосует правильно, и монотонно убывать по числу объектов, за которые он голосует не правильно.

- УПРАЖНЕНИЯ

- 1. Пусть один класс один представлен прецедентами **{(1, 0, 0, 1), (1, 0, 1, 1), (1, 1, 0, 1)},** а другой **{(1, 1, 0, 0), (1, 0, 0, 0), (1, 0, 1, 0)}**. Найдите все **(2, 1)** представительные наборы.

- 2. Будет ли для $(\boldsymbol{p}, \boldsymbol{q})$ элементарного классификатора применима для оценки информативности $\boldsymbol{\nu_1}(\boldsymbol{\delta}, \boldsymbol{H})$? $\boldsymbol{\nu_2}(\boldsymbol{\delta}, \boldsymbol{H})$?