Вычислительные технологии, 2024, том 29, № 1, с. 45–58. © ФИЦ ИВТ, 2024
Computational Technologies, 2024, vol. 29, no. 1, pp. 45–58. © FRC ICT, 2024
ISSN 1560-7534
eISSN 2313-691X

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

DOI:10.25743/ICT.2024.29.1.005

Слабо-контролируемая групповая классификация

В. Б. Бериков[1,*], О. А. Кутненко[1], И.А. Пестунов[2]
[1]Институт математики им. С.Л. Соболева СО РАН, 630090, Новосибирск, Россия
[2]Федеральный исследовательский центр информационных и вычислительных технологий, 630090, Новосибирск, Россия
*Контактный автор: Бериков Владимир Борисович, e-mail: berikov@math.nsc.ru
Поступила 19 июля 2023 г., принята в печать 06 сентября 2023 г.

В работе решается задача слабо-контролируемого обучения в постановке групповой бинарной классификации. Предполагается, что каждый объект выборки может включать набор подобъектов, относящихся к разным классам. Предлагаемый метод решения основан на выборе информативного признакового пространства и фильтрации обучающей выборки. Описывается применение разработанного метода для прогнозирования степени поражения участков головного мозга у пациентов с ишемическим инсультом по снимкам компьютерной томографии.

Ключевые слова: слабо-контролируемое обучение, групповая классификация, информативные признаки, фильтрация объектов выборки, компьютерная томография.

Цитирование: Бериков В.Б., Кутненко О.А., Пестунов И.А. Слабо-контролируемая групповая классификация. Вычислительные технологии. 2024; 29(1):45–58. DOI:10.25743/ICT.2024.29.1.005.

Введение

В теории и методах машинного обучения существует несколько основных направлений. В рамках направления “обучение с учителем” предполагается, что в выборке наблюдений для каждого объекта указана “метка” (распознаваемый образ, класс или числовое значение целевой переменной), полученная от некоторого “учителя”. Набор наблюдений за объектами анализа представляется в виде таблицы, содержащей значения определенного набора признаков. Требуется построить решающее правило для прогнозирования целевой переменной для новых объектов по их признаковым описаниям, оптимальное по заданному критерию (например, минимизирующее оценку вероятности ошибки классификации).

При решении задач “обучения без учителя” анализируется неразмеченная выборка. Требуется выявить структурные свойства данных, такие как группы похожих между собой объектов (кластеров), оценить плотность распределения выборки или найти значимые в определенном смысле комбинации признаков.

В постановке задачи полуконтролируемого обучения (semi-supervised learning) [1] предполагается, что часть объектов размечена, а другая часть (как правило, гораздо большего объема) не размечена. Частичное отсутствие разметки объясняется большими затратами на ее проведение, связанными, например, с необходимостью привлечения высококвалифицированных специалистов или выполнения дорогостоящих исследований.

В то же время неразмеченная часть выборки может дать дополнительную информацию о структуре данных, повысив тем самым качество прогноза. В парадигме "активного обучения" (active learning) [2] для наиболее "важных" объектов может быть запрошена дополнительная информация об их разметке.

Слабо-контролируемое обучение (weakly-supervised learning) подразумевает возможную неопределенность или нечеткость разметки. Эта неопределенность может пониматься по-разному (см. обзор [3]).

Одна из постановок задачи предполагает наличие неопределенности в указании точной метки класса, возникающей из-за ошибок разметки или в силу ограниченности самого метода наблюдения. Для решения такой задачи предложен ряд подходов. Один из них основан на поиске потенциально ошибочных разметок и их корректировке ("цензурировании" выборки) [4, 5]. Другой способ решения применяется в случае, когда разметка проводится множеством независимых пользователей (метод краудсорсинга), среди которых могут встречаться как опытные, так и неопытные (и даже преднамеренно ошибающиеся). Для решения задачи используются вероятностные или ансамблевые методы [6, 7]. Следующий прием основан на минимизации теоретической оценки риска с учетом случайной ошибки разметки. Так, в [8] предложен алгоритм, использующий тот факт, что функционал эмпирического риска может быть разделен на две части, где первая часть не зависит от зашумления, а вторая часть подвержена влиянию зашумленных меток. Известны также подходы [9], основанные на предположениях о кластерной структуре данных или о том, что данные принадлежат некоторому многообразию в признаковом пространстве. В [10] с применением предварительного разбиения данных на кластеры найдены верхние границы показателей зашумленности меток, а также предложен алгоритм оценивания степени зашумленности. Разработаны алгоритмы слабо-контролируемой классификации и регрессии [11], базирующиеся на методе регуляризации многообразия и кластерном ансамбле.

В другой постановке задачи слабо-контролируемого обучения метки определены для множеств объектов [12]. Каждый объект из сответствующего множества может иметь свое признаковое описание. Например, объект-молекула относится к классу обладающих лекарственным эффектом, если среди множества форм молекулы (конформаций) содержится хотя бы один вариант, который взаимодействует с целевым сайтом связывания некоторого белка. Требуется предсказать наличие или отсутствие лекарственного эффекта для новых молекул, представленных как набор конформаций. Известны несколько вариантов постановки данной задачи, называемой обучением на множествах примеров (multi-instance learning) [13], групповым обучением [14] или обучением на мультимножествах [15]. Для решения рассматриваемого круга задач разработан ряд методов. Так, в [16] представлен алгоритм, основанный на модификации метода опорных векторов.

В методах решения задачи слабо-контролируемой классификации используются предположения о модели неточности разметки, проверка которых не всегда возможна в силу ограниченного объема выборки.

В предлагаемой работе рассматривается задача слабо-контролируемого обучения в постановке групповой классификации, когда каждый объект может включать набор подобъектов, относящихся к различным классам. Способ решения задачи основан на выборе информативного признакового пространства и удалении из обучающей выборки объектов-выбросов (фильтрации выборки). Метод не требует задания модели неточности разметки. Рассматривается случай задачи бинарной классификации. Разработан ал-

горитм, который был применен для решения задачи прогнозирования степени поражения участков головного мозга при ишемическом инсульте с использованием снимков компьютерной томографии. В указанной задаче неточность разметки характерна для областей изображений, находящихся на границах контура очага поражения.

Настоящая работа имеет следующую структуру. В разд. 1 дается формальная постановка задачи, вводятся необходимые обозначения. Раздел 2 посвящен определению метрик качества решающей функции для рассматриваемой задачи. В разд. 3 описываются используемые способы выбора информативного признакового пространства и фильтрации объектов обучающей выборки, формулируются основные шаги предлагаемого метода. В разд. 4 представлена задача распознавания области поражения при ишемическом инсульте на основе анализа КТ-изображений мозга. Характеризуется неточность, возникающая при разметке. Приводятся результаты применения предложенного алгоритма и его сравнения с рядом других известных алгоритмов. В заключении делаются основные выводы работы, намечаются возможные направления дальнейших исследований.

1. Описание проблемы и обозначения

Пусть имеется генеральная совокупность Γ объектов $a \in \Gamma$, описываемых набором признаков $X = \{X_1, \ldots, X_d\}$, где d — размерность признакового пространства. Обозначим через $\mathbf{x} = X(a)$ ($\mathbf{x} \in \mathbb{R}^d$) набор наблюдений признаков для объекта a, где $\mathbf{x} = (x_1, \ldots, x_d)$, $x_j = X_j(a)$, $j = 1, \ldots, d$. Дана метрика r, позволяющая вычислять расстояние $r(a, b)$ между любой парой объектов a и b как в пространстве, соответствующем набору признаков X, так и в любом его подпространстве.

Следуя специфике задачи слабо-контролируемого обучения, предположим, что для каждого объекта может быть указана нечеткая принадлежность к образам A, B. Пусть для объекта a задана величина $y_A(a)$ — степень принадлежности к образу A. Будем полагать, что $y_A(a) \in [0, 1]$. Соответственно определена степень принадлежности к образу B как величина $y_B(a) = 1 - y_A(a)$.

Требуется для произвольного наблюдения $\mathbf{x} = X(a)$ предсказать его принадлежность $f(\mathbf{x})$ к образу A. При этом определена функция потерь $L(\widehat{y}, y)$, возникающих при использовании прогноза (решающей функции) $\widehat{y} = f(\mathbf{x})$ в случае истинного значения прогнозируемой величины равного y.

Пусть прогнозируемый объект выбирается из Γ случайным образом, тогда можно полагать, что существует вероятностное распределение $p_X(\mathbf{x})$, которому подчиняются наблюдения. Кроме того, предполагается, что процедура определения степени принадлежности к образам также имеет случайный характер. Поэтому можно считать, что величина $y_A(a)$ при фиксированном значении $\mathbf{x}$ имеет некоторое распределение $p_Y(y|\mathbf{x})$. Естественно потребовать, чтобы оптимальная решающая функция давала минимально возможное значение ожидаемых потерь (риска) $R(f) = \mathbb{E}_{\mathbf{x},y} L(f(\mathbf{x}), y)$.

Как правило, вероятностные распределения рассматриваемых величин неизвестны, поэтому для построения решающей функции используется случайная обучающая выборка объектов $s = \{a_1, \ldots, a_N\}$, где N — объем выборки. Предположим, объекты извлекаются из Γ независимо друг от друга в соответствии с одним и тем же распределением. Пусть этим объектам соответствует набор наблюдений признаков $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. Элементы выборки $\mathbf{X}$ являются слабо размеченными, т. е. для них указан набор $Y = \{y_1, \ldots, y_N\}$ степеней принадлежности к образу A, где $y_i = y_A(a_i)$, $i = 1, \ldots, N$.

Пусть для построения решающей функции f используется некоторый метод обучения, в котором оптимизируется определенный функционал качества метода $Q(f)$:

$$f = \arg\min_{f \in \Phi} Q(f),$$

где Φ — заданный класс решающих функций. Например, в методе обучения путем минимизации эмпирического риска используется критерий

$$\widetilde{R}(f) = \frac{1}{N} \sum_{i=1}^{N} L(f(\mathbf{x}_i), y_i). \tag{1}$$

Оценки качества, полученные по обучающей выборке, как правило, являются смещенными. Несмещенная оценка качества построенной решающей функции $f(\mathbf{x})$ может быть определена как оценка риска по тестовой выборке s_t, сформированной независимо от обучающей выборки s:

$$\widehat{R}(f) = \frac{1}{N_t} \sum_{\substack{\mathbf{x}, y: \\ \mathbf{x} \in \mathbf{X}_t,\, y \in Y_t}} L(f(\mathbf{x}), y),$$

где N_t — объем тестовой выборки, $\mathbf{X}_t$ — набор наблюдений объектов тестовой выборки, Y_t — набор слабых разметок (степеней принадлежности к образу A) для объектов из s_t. В случае абсолютной функции потерь соответствующий критерий MAE (mean absolute error) выглядит как

$$\mathrm{MAE}(f) = \frac{1}{N_t} \sum_{\substack{\mathbf{x}, y: \\ \mathbf{x} \in \mathbf{X}_t,\, y \in Y_t}} |f(\mathbf{x}) - y|. \tag{2}$$

Таким образом, цель состоит в том, чтобы предложить (построить) модель, позволяющую прогнозировать целевой признак для новых примеров, описанных в том же признаковом пространстве. Рассматриваемая проблема актуальна для многих прикладных задач, поскольку аннотация имеющихся данных может быть неточной из-за слабой изученности проблемы, нехватки ресурсов для тщательной маркировки объектов, наличия случайных искажений, возникающих в процессе идентификации метки, а также в силу других причин, определяющих специфику решаемой задачи.

2. Метрики качества решающей функции для задачи группового обучения

Рассмотрим вариант постановки задачи слабо-контролируемого обучения (задачу группового обучения), когда каждый объект может включать набор подобъектов как образа A (“положительного”), так и образа B (“отрицательного”). Определим метрики качества бинарной классификации. Пусть величины $n_A(a)$, $n_B(a)$ означают соответственно количество подобъектов образов A и B для объекта $a \in \Gamma$. Общее число подобъектов для a обозначим как $n(a) = n_A(a) + n_B(a)$. Если степень принадлежности объекта a образу A равна $y_A(a)$, то должны выполняться $n_A(a) = y_A(a)n(a)$, $n_B(a) = (1 - y_A(a))n(a)$.

Предположим, что для произвольного объекта a с использованием решающей функции $f(\mathbf{x})$, где $\mathbf{x} = X(a)$, получен прогноз степени принадлежности классу A. Обозначим через $\hat{n}_A(a)$ прогнозируемую частоту подобъектов образа A : $\hat{n}_A(a) = f(\mathbf{x})n(a)$. Тогда можно определить следующие характеристики:

– количество истинно положительных результатов

$$\mathrm{TP} = \sum_{a:X(a)\in\mathbf{X}_t} \min(n_A(a), \hat{n}_A(a));$$

– число истинно отрицательных меток

$$\mathrm{TN} = \sum_{a:X(a)\in\mathbf{X}_t} \min(n_B(a), \hat{n}_B(a));$$

– количество ложноположительных прогнозов

$$\mathrm{FP} = \sum_{a:X(a)\in\mathbf{X}_t} \max(\hat{n}_A(a) - n_A(a), 0);$$

– число ложноотрицательных случаев

$$\mathrm{FN} = \sum_{a:X(a)\in\mathbf{X}_t} \max(\hat{n}_B(a) - n_B(a), 0).$$

Для оценки качества решающей функции на тестовой выборке вычисляются следующие характеристики:

– точность

$$\mathrm{accuracy} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \cdot 100\ \%;$$

– чувствительность — способность решающей функции давать правильный результат (доля истинно положительных прогнозов)

$$\mathrm{sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \cdot 100\,\%;$$

– специфичность — способность классификатора не давать ложноположительных результатов (доля истинно отрицательных меток)

$$\mathrm{specifisity} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \cdot 100\,\%;$$

– прецизионная точность — доля объектов, названных положительными и при этом действительно являющихся положительными,

$$\mathrm{precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \cdot 100\,\%;$$

– сбалансированная точность

$$\mathrm{BA} = \frac{\mathrm{sensitivity}+\mathrm{specifisity}}{2}.$$

Указанные выше показатели часто используются в случае несбалансированных данных, когда частоты встречаемости классов значительно отличаются друг от друга.

3. Предлагаемый метод

При решении задач классификации на практике, как правило, приходится предварительно решать две задачи: выбор информативного признакового пространства и формирование обучающей выборки. Существует две стратегии повышения информативности обучающей выборки, различающиеся своими целями и методами. Первая стратегия — это выбор наименьшего возможного числа обучающих объектов, достаточного для построения надежного классификатора. При решении рассматриваемой задачи используется вторая стратегия формирования обучающей выборки — фильтрация (удаление) объектов, плохо описываемых моделью, используемой для классификации.

Как выбор информативных признаков, так и удаление "шумовых" объектов ("выбросов") осуществляются на основе анализа локального окружения объектов. Данный подход опирается на гипотезу локальной компактности [17]. При решении используется метод k-ближайших соседей (kNN). Критерий качества решения указанных задач — экстремум некоторого заданного функционала, связанного с ошибкой прогнозирования целевого признака. Прогнозное значение зададим следующим образом.

Пусть $\mathfrak{N}(a)$ — множество, состоящее из некоторого заданного числа k ближайших соседей по метрике r для объекта $a \in \Gamma$, которому соответствует набор признаков $\mathbf{x} = X(a)$. Определим степень сходства объекта a и образа A как величину

$$P(A|a) = \frac{1}{C} \sum_{z \in \mathfrak{N}(a)} n_A(z) e^{-\gamma\, r(z,a)}, \tag{3}$$

где C — нормализующая константа:

$$C = \sum_{z \in \mathfrak{N}(a)} n(a) e^{-\gamma\, r(z,a)},$$

$\gamma > 0$ — заданный параметр. Аналогично вводится и степень сходства с образом B. Заметим, что приведенное выражение для $P(A|a)$ основано на непараметрической оценке плотности распределения с экспоненциальным ядром. Возможны и другие способы задания сходства.

Определим решающую функцию для прогнозирования целевого признака по наблюдению $\mathbf{x} = X(a)$ как $f(\mathbf{x}) = P(A|a)$.

3.1. Выбор информативного признакового пространства

При анализе мало изученного материала часто используется большое количество характеристик, описывающих объекты. Не все признаки несут информацию, полезную для решаемой задачи: среди них могут содержаться дублирующие друг друга данные или просто шумовые. Из исходного множества признаков необходимо выбрать подмножество наиболее информативных характеристик. Для решения этой задачи разработано большое количество алгоритмов, обзор которых можно найти в [18, 19].

Предлагается использовать алгоритм AdDel [20], основанный на комбинации двух известных алгоритмов: Ad [21] — метода последовательного добавления признаков и Del [22] — метода последовательного сокращения признаков. Эти алгоритмы дают оптимальное решение на каждом шаге, но не обеспечивают глобального оптимума.

С целью ослабления влияния ошибок на первых шагах алгоритма AdDel применяется релаксационный метод: сначала алгоритмом Ad набирается некоторое количество

d_1 информативных с точки зрения используемого критерия признаков, затем d_2 из них ($d_2 < d_1$) исключаются методом Del. После этого алгоритмом Ad размерность набора информативных признаков наращивается на величину d_1. В этот момент снова включается алгоритм Del, который удаляет из системы d_2 "наименее ценных" признаков. Такое чередование алгоритмов Ad и Del, которое получило название алгоритма AdDel, продолжается до момента достижения экстремального значения критерия (либо до достижения заданного количества признаков).

Таким образом, алгоритм AdDel сочетает в себе идеи методов "последовательного добавления наиболее ценных" (Addition) и "последовательного удаления наименее ценных" (Deletion) признаков. Алгоритм позволяет указать как состав, так и наилучшее количество характеристик. Отметим, что значительное увеличение числа выбранных признаков для достижения экстремума критерия при обучении приводит к переобучению модели и, соответственно, к снижению качества распознавания на контроле.

Добавлять и исключать признаки можно как по одному, так и группами (гранулами), состоящими из нескольких признаков. В используемом при построении информативного признакового пространства алгоритме GRAD [23] алгоритм AdDel работает на множестве наиболее информативных гранул, состоящих из v признаков, $v = 1, \ldots, V$. Гранулы мощности 1 — это исходные признаки, из них методом полного перебора формируются гранулы мощности $2, \ldots, V$. Затем весь список из заданного количества самых информативных гранул подается на вход алгоритма AdDel. Выбор величины V делается исходя из двух соображений: первое основано на учете возможностей компьютера, второе — на гипотезе о преобладании простых закономерностей над сложными.

3.2. Фильтрация объектов обучающей выборки

Наличие разного рода ошибок в наблюдениях приводит к ухудшению качества получаемых закономерностей. Проблема редактирования и очистки данных (data editing, data cleaning) актуальна для многих прикладных задач. Погрешности описания данных могут возникать из-за сложности решаемых проблем, недостаточной квалификации экспертов, проводящих разметку, из-за ограниченной точности измерительных приборов. На качество распознавания также влияет и неполное признаковое описание данных, связанное со слабой изученностью проблемы. Стратегия удаления объектов обучающей выборки, плохо описываемых с помощью используемой для классификации модели, себя оправдывает, если данных много и даже если значительное уменьшение объема выборки после процедуры фильтрации не сказывается на ее представительности.

В случае рассмотрения целевого признака $y_A(a)$ как характеризующего степень принадлежности объекта a образу A "выброс" — это такой объект, что $L(\hat{y}, y) \geq l^*$, где $\hat{y}$ — прогнозируемая степень принадлежности объекта a образу A, l^* — пороговое значение ошибки. Фильтрация подобных объектов позволяет повысить информативность обучающей выборки и уменьшить ошибку прогнозирования контрольных объектов. Для получения количественной оценки информативности обучающей выборки предлагается использовать описанный выше критерий (1).

Так как с ростом числа исключенных объектов неизбежно возникает переобучение модели и результаты прогноза становятся все менее достоверными, для учета этого эффекта используется нормирующий (штрафной) коэффициент $(N/N^*)^q$, где N — исходное число объектов в выборке, а N^* — число объектов, оставшихся после удаления объектов-выбросов; $q > 0$ — параметр. Итоговая оценка информативности обучающей выборки задается как

$$I = \left(\frac{N}{N^*}\right)^q \widetilde{R}(f). \tag{4}$$

Таким образом, процедура фильтрации состоит в следующем. Поочередно перебираются наиболее нетипичные объекты обучающей выборки. Нетипичными будем считать объекты a_i, $i=1,\ldots,N$, для которых ошибка прогноза целевого признака $l_i = l(\hat{y}_i, y_i) \geq l^*$. Варьирование порогового значения влияет на соотношение числа ошибок первого рода (пропуск цели) и второго рода (ложная тревога). Для выбранного нетипичного объекта вычисляются значения критерия качества обучающей выборки при удалении этого объекта. Процесс останавливается при достижении точки перегиба в значении критерия или при отсутствии нетипичных объектов в оставшейся части выборки.

Конкретная стратегия процедуры фильтрации и критериев ее остановки зависит от решаемой прикладной задачи. Ниже приведен алгоритм фильтрации, используемый при проведении описанных в работе экспериментов.

Введем обозначения. Пусть l_0 — текущий уровень ошибки для фильтрации объектов, $\bar{l}$ — шаг изменения уровня ошибки, I_0 — исходная информативность выборки, I — текущая информативность. Критерии остановки алгоритма: доля удаляемых объектов не больше δ от исходного размера выборки, т. е. $N - N^* \leq \delta N$, $I > I_0$, отсутствие в выборке нетипичных объектов.

Процедура фильтрации проводилась следующим образом.

Шаг 0. Задание начальных значений l_0, l^*, $\bar{l}$, q, а также $I_0 := \widetilde{R}(f)$.

Шаг 1. $l_0 := l_0 - \bar{l}$. Если $l_0 < l^*$, то перейти на шаг 3. Если число удаленных объектов и объектов с $l \geq l_0$ больше величины δN, то переход на шаг 3. Если нет объектов с $l \geq l_0$, то перейти на шаг 1.

Шаг 2. Удаляются объекты с ошибкой $l \geq l_0$. Вычисляется значение I по (4). Если $I > I_0$, то переход на шаг 3, иначе $I_0 := I$ и перейти на шаг 1.

Шаг 3. Конец алгоритма.

Отметим, что при уменьшении значения l^* число ошибок первого рода уменьшается и, соответственно, увеличивается число ошибок второго рода.

3.3. Основные шаги предлагаемого метода

На вход поступают данные в виде таблицы объект – признак, где целевой признак — степень принадлежности объектов образу A.

Шаг 0. Выбор метрики, задание начальных значений и диапазонов возможных значений параметров алгоритмов. Предварительная обработка данных, которая включает их нормализацию по признакам. Деление данных на три части — обучающую, валидационную и контрольную выборки в пропорции, определяемой пользователем.

Шаг 1. Формирование по обучающей выборке и текущим значениям параметров алгоритмом GRAD (или AdDel) информативного признакового пространства.

Шаг 2. Фильтрация обучающей выборки с текущими значениями параметров.

Шаг 3. Вычисление текущего значения критерия (4).

Шаг 4. По валидационной части выборки с помощью перебора различных комбинаций параметров из заданных диапазонов проводится подбор параметров алгоритмов (выбора признаков, фильтрации и kNN), для которых достигается минимальное значение критерия (4).

Шаг 5. Распознавание методом kNN контрольных объектов и оценка качества решения по заданным метрикам качества.

4. Экспериментальное исследование

Оценивание эффективности предлагаемого метода решения задачи слабо-контролируемого обучения (в задаче групповой классификации) проводилось на реальных медицинских данных, полученных в Международном томографическом центре СО РАН. Автоматизированная диагностика ишемического поражения мозга в перспективе даст возможность обрабатывать данные компьютерной томографии (КТ) для потока пациентов, позволяя обращать внимание врача в первую очередь на наиболее сложные и опасные для пациентов случаи.

4.1. Исходные данные

Данные представляют собой анонимизированные цифровые снимки бесконтрастной компьютерной томографии головного мозга у пациентов с ишемическим инсультом. Зоны поражения на снимках были аннотированы специалистами-рентгенологами. На рис. 1 представлен пример КТ-изображения среза головного мозга и соответствующей сегментационной маски. Для каждого пациента получено порядка 300 изображений-срезов в формате DICOM. Размер исходных снимков 512×512 пикселей. Предварительно была проведена предобработка изображений — выделение зоны головного мозга, контрастирование и нормализация по яркости.

Для сегментации зоны поражения в [24] использованы глубокие сверточные нейронные сети с архитектурой U-Net. Однако решение, принимаемое этими сетями, является “черным ящиком”, поэтому для получения более интерпретируемых выводов (пусть обладающих несколько меньшей точностью) применяются и другие методы анализа изображений, в частности основанные на текстурных признаках. Эти признаки вычисляются с помощью анализа распределения интенсивности пикселей внутри некоторых участков изображения. Для формирования признаков использовалась процедура их извлечения из анализируемых изображений, разбиваемых на участки — квадраты одинакового размера. Подробное описание процедуры и исходного 30-мерного признакового пространства приведено в работе [11].

Всего в настоящей работе использованы снимки 24 пациентов. Выборка представлена 8043 объектами — участками КТ-снимков головного мозга со степенью поражения от нуля до ста процентов. Целевой признак — степень поражения области, вычисленная по уровню пересечения с сегментационной маской. Требуется определить степень поражения участков головного мозга для новых пациентов по их КТ-снимкам. Поскольку каждый участок содержит подобъекты — пиксели, относящиеся к двум образам, решаемая задача относится к классу задач группового обучения.

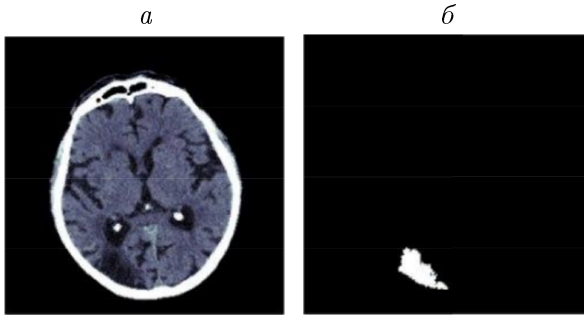

Рис. 1. Пример КТ-изображения (*а*) и сегментационной маски (*б*)
Fig. 1. Example of CT image (*а*) and segmentation mask (*б*)

В [11] с применением текстурного подхода был разработан алгоритм прогнозирования степени поражения участков, использующий метод регуляризации многообразия и лапласиан графа сходства. Ниже проводится сравнение результатов предлагаемого алгоритма и алгоритма [11], а также ряда других алгоритмов на указанном наборе данных.

4.2. Результаты экспериментов

Для оценки качества предложенного метода определения степени поражения вычислялись указанные в разд. 2 метрики качества информативности диагностических методов и ошибка определения степени поражения (2). При решении использован метод kNN; лучшие результаты получены при $k = 7$. В (3) выбрано $\gamma = 45$. Критерий качества решения описанных выше задач выбора информативного признакового пространства и формирования обучающей выборки — минимум ошибки прогнозирования целевого признака.

Для каждого пациента задавался набор участков головного мозга с разной степенью поражения. Предварительно данные были нормализованы к диапазону $[0, 1]$; для вычисления расстояний между объектами использовалась метрика Евклида. По обучающей выборке алгоритмом GRAD выбрано информативное признаковое пространство. Параметры алгоритма GRAD: $d_1 = 2$, $d_2 = 1$, $V = 3$; параметры алгоритма фильтрации: $l_0 = 1$, $l^* = 0.5$, $\bar{l} = 0.1$, $q = 5$, $\delta = 0.1$.

Поскольку число пациентов было сравнительно невелико, оценка качества метода проводилась с помощью процедуры скользящего экзамена по пациентам. Отметим, что сформированные в ходе данной процедуры наборы признаков были устойчивыми для всех обучающих выборок.

Гистограмма, представленная на рис. 2, показывает частоту появления указанного уровня ошибки MAE (2) на контрольных выборках, формируемых на скользящем экзамене. Усредненное значение MAE было равно 0.182, стандартное отклонение среднего 0.005.

В таблице для предложенного алгоритма и ряда других алгоритмов приведены усредненные метрики качества решений. Рассматривались алгоритмы, основанные на регуляризации многообразия [11], случайном лесе (random forest) с деревьями регрессии или классификации, байесовский классификатор в предположении нормального рас-

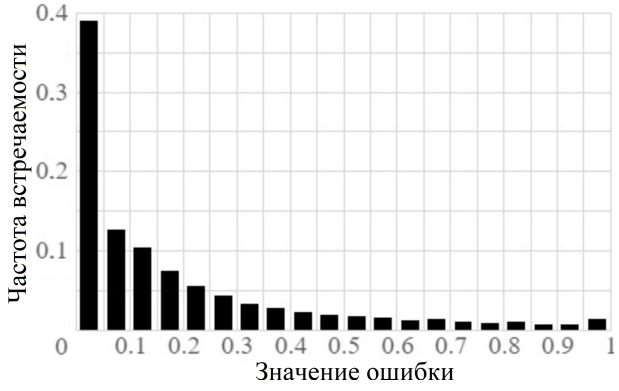

Рис. 2. Гистограмма уровня ошибки
Fig. 2. Histogram of error level

Усредненные метрики качества алгоритмов
Averaged quality metrics for algorithms

Название алгоритма	Значения метрик, %		
	sensitivity	specificity	BA
Предложенный алгоритм	86.8	73.5	80.2
Алгоритм [11]	82.7	77.3	80.0
Random Forest (regression trees)	77.0	77.4	77.2
Random Forest (classification trees)	29.4	82.5	55.9
Normal Bayes classifier	36.2	60.9	48.5
SVM (histogram intersection kernel)	68.8	34.3	51.5
SVM (linear kernel)	48.9	52.4	50.6
SVM (polynomial kernel)	31.9	67.3	49.6
SVM (sigmoid kernel)	47.7	53.3	50.5
kNN ($k = 2$)	50.3	56.7	53.5
kNN ($k = 4$)	42.8	61.1	51.9
kNN ($k = 13$)	28.7	74.8	51.7

пределения признаков, метод опорных векторов SVM с различными вариантами ядра и стандартный метод kNN для ряда значений числа ближайших соседей k.

По результатам сравнения можно сделать вывод, что для предложенного метода качество решений, определяемое по чувствительности, оказалось наилучшим. По сбалансированной точности результаты сравнимы с ближайшими по качеству алгоритмами. Более низкая чувствительность метода означает, что у части пациентов очаг поражения не будет идентифицирован, что может привести к нежелательным последствиям в виде осложнений течения заболевания. Можно предположить, что для повышения точности прогноза целесообразно привлекать дополнительную информацию о местонахождении очага поражения в структуре головного мозга, клинических характеристиках пациентов и т. д.

Заключение

Предложен метод решения задачи слабо-контролируемой групповой классификации с использованием выбора информативного признакового пространства и фильтрации объектов обучающей выборки. Проведены численные эксперименты в задаче анализа томографических снимков головного мозга для прогнозирования степени поражения его участков при инсульте. Результаты экспериментального исследования и сравнения с рядом известных алгоритмов машинного обучения подтвердили достаточно высокую эффективность разработанного метода. В отличие от других аналогичных алгоритмов, метод позволяет сформировать набор наиболее информативных признаков с целью улучшения интерпретируемости решений и снижения эффекта переобучения. Планируются дальнейшие исследования для повышения надежности и устойчивости слабо-контролируемого распознавания, в частности с применением комбинации предложенного метода, ансамблевых и нейросетевых алгоритмов.

Благодарности. Работа выполнена при финансовой поддержке РНФ (грант № 22-21-00261).

Список литературы

[1] **Van Engelen J.E., Hoos H.H.** A survey on semi-supervised learning. Machine Learning. 2020; 109(2):373–440.

[2] **Cohn D.A., Ghahramani Z., Jordan M.I.** Active learning with statistical models. Journal of Artificial Intelligence Research. 1996; (4):129–145.

[3] **Zhou Z.-H.** A brief introduction to weakly supervised learning. National Science Review. 2018; 5(1):44–53.

[4] **Muhlenbach F., Lallich ., Zighed D.** Identifying and handling mislabelled instances. Journal Intelligent Information Systems. 2004; (22):89–109.

[5] **Борисова И.А., Кутненко О.А.** Исправление диагностических ошибок в целевом признаке с помощью функции конкурентного сходства. Математическая биология и биоинформатика. 2018; 13(1):38–49. DOI:10.17537/2018.13.38.

[6] **Raykar V.C., Yu S., Zhao L.H., Florin C., Bogoni L., Moy L.** Learning from crowds. Journal of Machine Learning Research. 2010; (11):1297–1322.

[7] **Zhou Z.-H.** Ensemble methods: foundations and algorithms. Boca Raton: CRC Press; 2012: 218.

[8] **Gao W., Wang L., Li Y.F., Zhou Z.-H.** Risk minimization in the presence of label noise. Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix, AZ; 2016: 30(1).

[9] **Huang K., Shi Y., Zhao F., Zhang Z., Tu S.** Multiple instance deep learning for weakly supervised visual object tracking. Signal Processing: Image Communication. 2020; (84):115807.

[10] **Gao W., Zhang T., Yang B.-B., Zhou Z.-H.** On the noise estimation statistics. Artificial Intelligence. 2021; (293):103451.

[11] **Berikov V., Litvinenko A., Pestunov I., Sinyavskiy Yu.** On a weakly supervised classification problem. Lecture Notes in Computer Science. Springer; 2022; (13217):315–329. DOI:10.1007/978-3-031-16500-9_26.

[12] **Foulds J., Frank E.** A review of multi-instance learning assumptions. The Knowledge Engineering Review. 2010; 25(1):1–25.

[13] **Dietterich T.G., Lathrop R.H., Lozano-Pèrez T.** Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence. 1997; 89(1–2):31–71.

[14] **Abusev R.A.** On group choice procedures for problems of classification and reliability in the case of lognormal variance. Journal Mathematical Sciences. 2013; 189(6):911–918. DOI:10.1007/s10958-013-1231-y.

[15] **Petrovsky A.B.** Methods for the group classification of multi-attribute objects (part 1). Scientific and Technical Information Processing. 2010; 37(5):346–356.

[16] **Xiao Y., Zijian Y., Bo L.** A similarity-based two-view multiple instance learning method for classification. Knowledge-based Systems. 2020; (201):105661.

[17] **Аркадьев А.Г., Браверман Э.М.** Обучение машины распознаванию образов. М.: Наука; 1964: 112.

[18] **Загоруйко Н.Г.** Прикладные методы анализа данных и знаний. Новосибирск: Издательство Института математики СО РАН; 1999: 270.

[19] **Li Y., Li T., Liu H.** Recent advances in feature selection and its applications. Knowledge and Information Systems. 2017; (53):551–577. DOI:10.1007/s10115-017-1059-8.

[20] **Загоруйко Н.Г., Кутненко О.А.** Методы распознавания, основанные на алгоритме AdDel. Сибирский журнал индустриальной математики. 2004; 7(1(17)):39–47.

[21] **Барабаш Ю.Л., Варский Б.В., Зиновьев В.Т.** Автоматическое распознавание образов. Киев: Издательство КВАИУ; 1963: 168.

[22] **Merill T., Green O.M.** On the effectiveness of receptors in recognition systems. IEEE Transactions on Information Theory. 1963; (IT–9):11–17.

[23] **Загоруйко Н.Г.** Когнитивный анализ данных. Новосибирск: Академическое издательство ГЕО; 2013: 186.

[24] **Kalmutskiy K., Tulupov A., Berikov V.** Recognition of tomographic images in the diagnosis of stroke. Del Bimbo A. et al. (Eds.) Pattern Recognition. ICPR International Workshops and Challenges. ICPR 2021. Lecture Notes in Computer Science. Springer, Cham; 2021; (12665). DOI:10.1007/978-3-030-68821-9_16.

Вычислительные технологии, 2024, том 29, № 1, с. 45–58. © ФИЦ ИВТ, 2024 ISSN 1560-7534
Computational Technologies, 2024, vol. 29, no. 1, pp. 45–58. © FRC ICT, 2024 eISSN 2313-691X

INFORMATION TECHNOLOGIES

DOI:10.25743/ICT.2024.29.1.005

Weakly supervised group classification

V. B. Berikov[1,*], O. A. Kutnenko[1], I. A. Pestunov[2]

[1]Sobolev Institute of Mathematics SB RAS, 630090, Novosibirsk, Russia
[2]Federal Research Center for Information and Computational Technologies, 630090, Novosibirsk, Russia
[*]Corresponding author: Vladimir B. Berikov, e-mail: berikov@math.nsc.ru
Received July 19, 2023, accepted September 06, 2023.

Abstract

Weakly supervised learning implies possible uncertainty or fuzziness of the labelling. Current study addresses this problem using the formulation of group binary classification. It is assumed that each sample object may include a set of sub-objects belonging to one of two classes. Objects are described by a set of features; the predicted feature determines the degree to which an object belongs to the "positive" class. It is required to construct a decision function from the training sample to predict the target feature for new objects.

The proposed method is based on the selection of informative feature space and filtering the training sample. Both the selection of informative features and the removal of noise observations are carried out on the basis of analysis of the local environment of objects. The degree of similarity between the object and the class is determined by the k nearest neighbours of the object, taking into account their degree of belonging to the target class. For an experimental study of the developed method, the real problem of analyzing tomography images of the brain to predict the degree of damage to its areas in stroke is solved. The results are compared with a number of known methods.

A method for constructing a decision function for predicting the degree of belonging of an object to the target class has been developed. The results of an experimental study and comparison with a number of well-known machine learning algorithms (random forest, support vector machine, kNN) confirmed the efficiency of the method for solving the problem of predicting the degree of damage to brain areas in stroke patients. Unlike other similar algorithms, the proposed method allows establishing a set of the most informative features in order to improve the interpretability of the solution and reduce the effect of overfitting.

Keywords: weakly supervised learning, group classification, informative features, filtering of sample objects, computed tomography.

Citation: Berikov V.B., Kutnenko O.A., Pestunov I.A. Weakly supervised group classification. Computational Technologies. 2024; 29(1):45–58. DOI:10.25743/ICT.2024.29.1.005. (In Russ.)

Acknowledgements. The work is supported by Russian Science Foundation, grant No. 22-21-00261.

References

1. **Van Engelen J.E., Hoos H.H.** A survey on semi-supervised learning. Machine Learning. 2020; 109(2):373–440.
2. **Cohn D.A., Ghahramani Z., Jordan M.I.** Active learning with statistical models. Journal of Artificial Intelligence Research. 1996; (4):129–145.
3. **Zhou Z.-H.** A brief introduction to weakly supervised learning. National Science Review. 2018; 5(1):44–53.
4. **Muhlenbach F., Lallich ., Zighed D.** Identifying and handling mislabelled instances. Journal Intelligent Information Systems. 2004; (22):89–109.
5. **Borisova I.A., Kutnenko O.A.** The problem of correction diagnostic errors in the target attribute with the function of rival similarity. Mathematical Biology and Bioinformatics. 2018; 13(1):38–49. (In Russ.)
6. **Raykar V.C., Yu S., Zhao L.H., Florin C., Bogoni L., Moy L.** Learning from crowds. Journal of Machine Learning Research. 2010; (11):1297–1322.
7. **Zhou Z.-H.** Ensemble methods: foundations and algorithms. Boca Raton: CRC Press; 2012: 218.
8. **Gao W., Wang L., Li Y.F., Zhou Z.-H.** Risk minimization in the presence of label noise. Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix, AZ; 2016: 30(1).
9. **Huang K., Shi Y., Zhao F., Zhang Z., Tu S.** Multiple instance deep learning for weakly supervised visual object tracking. Signal Processing: Image Communication. 2020; (84):115807.
10. **Gao W., Zhang T., Yang B.-B., Zhou Z.-H.** On the noise estimation statistics. Artificial Intelligence. 2021; (293):103451.
11. **Berikov V., Litvinenko A., Pestunov I., Sinyavskiy Yu.** On a weakly supervised classification problem. Lecture Notes in Computer Science. Springer; 2022; (13217):315–329. DOI:10.1007/978-3-031-16500-9_26.
12. **Foulds J., Frank E.** A review of multi-instance learning assumptions. The Knowledge Engineering Review. 2010; 25(1):1–25.
13. **Dietterich T.G., Lathrop R.H., Lozano-Pèrez T.** Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence. 1997; 89(1–2):31–71.
14. **Abusev R.A.** On group choice procedures for problems of classification and reliability in the case of lognormal variance. Journal Mathematical Sciences. 2013; 189(6):911–918. DOI:10.1007/s10958-013-1231-y.
15. **Petrovsky A.B.** Methods for the group classification of multi-attribute objects (part 1). Scientific and Technical Information Processing. 2010; 37(5):346–356.
16. **Xiao Y., Zijian Y., Bo L.** A similarity-based two-view multiple instance learning method for classification. Knowledge-based Systems. 2020; (201):105661.
17. **Arkad'ev A.G., Braverman E.M.** Obuchenie mashiny raspoznavaniyu obrazov [Machine learning for pattern recognition]. Moscow: Nauka; 1964: 112. (In Russ.)
18. **Zagoruiko N.G.** Prikladnye metody analiza dannykh i znaniy [Applied methods of data and knowledge analysis]. Novosibirsk: Izdatel'stvo Instituta Matematiki SO RAN; 1999: 270. (In Russ.)
19. **Li Y., Li T., Liu H.** Recent advances in feature selection and its applications. Knowledge and Information Systems. 2017; (53):551–577. DOI:10.1007/s10115-017-1059-8.
20. **Zagoruiko N.G., Kutnenko O.A.** Recognition methods based on the AdDel algorithm. Siberian Journal of Industrial Mathematics. 2004; 7(1(17)):39–47. (In Russ.)
21. **Barabash Yu.L., Varskiy B.V., Zinoviev V.T.** Avtomaticheskoe raspoznavanie obrazov [Automatic pattern recognition]. Kiev: Izdatel'stvo KVAIU; 1963: 168. (In Russ.)
22. **Merill T., Green O.M.** On the effectiveness of receptors in recognition systems. IEEE Transactions on Information Theory. 1963; (IT–9):11–17.
23. **Zagoruiko N.G.** Kognitivnyy analiz dannykh [Cognitive data analysis]. Novosibirsk: Akademicheskoe Izdatel'stvo GEO; 2013: 186. (In Russ.)
24. **Kalmutskiy K., Tulupov A., Berikov V.** Recognition of tomographic images in the diagnosis of stroke. Del Bimbo A. et al. (Eds.) Pattern Recognition. ICPR International Workshops and Challenges. ICPR 2021. Lecture Notes in Computer Science. Springer, Cham; 2021; (12665). DOI:10.1007/978-3-030-68821-9_16.