Метод построения дерева решений, основанного на похожести объектов в задаче распознавания томографических изображений

Р.М. Козинец[1], В.Б. Бериков[1,3], И.А. Пестунов[2], С.А. Рылов[2]

[1]Новосибирский государственный университет, Пирогова 2, Новосибирск, Россия, 630090
[2]Институт вычислительных технологий СО РАН, Лаврентьева 6, Новосибирск, Россия, 630090
[3]Институт математики имени С.Л. Соболева СО РАН, Академика Коптюга 4, Новосибирск, Россия, 630090

Аннотация. В работе предложен метод распознавания образов с применением модификации класса логических решающих функций, представленных в виде дерева решений. Вместо стандартных высказываний, соответствующих вершинам дерева, в которых проверяется принадлежность некоторой переменной тем или иным множествам ее значений, используется более общий тип высказываний относительно близости рассматриваемой точки к различным подмножествам наблюдений. При этом для определения степени похожести могут выбираться различные метрики и подпространства признаков. Этот тип дерева решений позволяет получить более сложные границы принятия решений, которые в то же время имеют понятную пользователю логическую интерпретацию. Рассмотрено несколько стратегий построения дерева: на основе преобразования данных с использованием опорных точек, выделенных с помощью процедур Relief, SVM и k-средних. Метод был применен для анализа томографических изображений. Эксперименты показали, что предложенный метод дает более точные прогнозы, чем алгоритмы CART, SVM, kNN и глубокая сверточная нейронная сеть (AlexNet).

1. Введение

В данной работе рассматривается задача распознавания образов (классификации "с учителем"), постановка которой выглядит следующим образом.

Пусть X – множество (генеральная совокупность) объектов *x*, $\boldsymbol{x} = (x^{(1)}, \dots, x^{(m)})$ – признаковое описание объекта, $\boldsymbol{x} \in \mathbf{X}$, где m – размерность пространства признаков $\mathbf{X}$, Y – множество классов. Для простоты в работе рассматривается задача двуклассового распознавания: $Y = \{+1, -1\}$, хотя полученные результаты легко распространяются на многоклассовый случай. Пусть $X^d \subset X$ – выборка объёма d, $y^*: X \to Y$ целевая функция, значения которой известны на конечном множестве (обучающей выборке) X^l. Требуется построить решающую функцию $a: \mathbf{X} \to Y$ из некоторого заданного класса, приближающую y^* на X^d и минимизирующую оценку вероятности ошибки – $Q(a, X^d) = \frac{1}{d}\sum_i^d L(a, x_i)$, где $L(a, x_i)$ – ошибка алгоритма a на объекте x_i.

Одними из наиболее часто используемых методов классификации являются методы построения логических решающих функций, представленных в виде дерева решений. По сравнению с другими методами распознавания образов, эти методы обладают рядом положительных свойств:

– позволяют проводить анализ разнотипной информации (т.е. при наличии количественных и качественных признаков, описывающих объекты);

– находят вероятностные логические закономерности, отражающие причинно-следственные связи изучаемого явления;

– автоматически определяют наиболее информативные признаки, по которым принимается решение;

– в сочетании с коллективным подходом, способны находить достаточно устойчивые решения, обладающие высокой обобщающей способностью (метод решающего леса, бустинг деревьев решений [14, 15]).

Обзор существующих методов построения деревьев решений изложен, например, в работах [10, 11]. Несмотря на достаточно большое число известных подходов, существует потребность в разработке методов построения деревьев решений, которые позволяли бы улучшать качество прогнозирования. Можно рассмотреть несколько способов повышения качества. Первый способ заключается в нахождении критериев, позволяющих усилить прогнозирующую способность решений за счет оптимального сочетания точности и сложности дерева [12] Второй способ связан с разработкой более сложных способов представления дерева (например, с использованием в узлах линейных разделяющих гиперплоскостей [13]) и применением более «глубоких» алгоритмов поиска оптимального решения [11].

«Классическое» дерево решений – это граф-дерево, в узлах которого проверяется условие: больше или меньше значение конкретного признака объекта некоторого порогового значения, а листьям присвоены значения целевой переменной или метки классов. Такой подход имеет существенный недостаток: разбиение выборки происходит строго параллельно признаковым осям, даже если реальная граница между классами линейной формы (рисунок 1). Предлагаемый в данной работе метод направлен на устранение этого недостатка.

Для экспериментального исследования метода использована задача распознавания злокачественных новообразований в легких по изображениям компьютерной томографии.

Статья имеет следующую структуру. Во втором разделе описывается предлагаемая модификация дерева решений, основанного на сходстве объектов (Similarity Based Decision Tree, SBDT), а также несколько стратегий для методов его построения. В третьем разделе проводится экспериментальное исследование метода на реальной задаче анализа данных и его сравнение с рядом других методов. В заключении формулируются основные выводы работы.

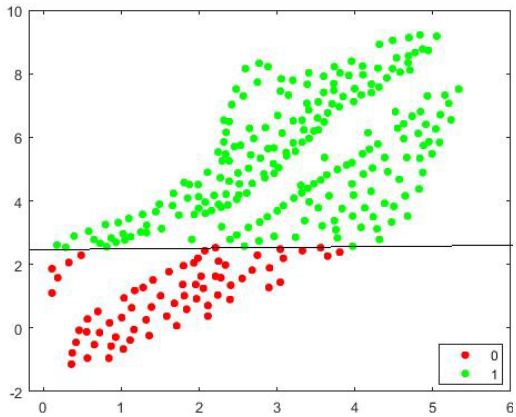


Рисунок 1. Разбиение сгенерированных данных классическим деревом решений.

2. Дерево решений на основе похожести объектов (SBDT)

Пусть $X^d = X^l \cup X_{test}$ – выборка объема d из X. Множества объектов A и B называются наборами опорных точек классов $+1, -1$ соответственно если:

$$A \subset X^l \ \forall x_i \in A;\ y^*(x_i) = +1,\ i = 1, \dots, t,\ t < l,$$
$$B \subset X^l \ \forall x_i \in B;\ y^*(x_i) = -1,\ i = 1, \dots, t,\ t < l.$$

Разработанный метод основан на идее о построении дерева решений, сравнивающего в узлах не значения признака $f_i(x)$ объекта $x \in X^l$, а степень «близости» x между заранее сформированными опорными точками из наборов A и B или, другими словами, в каждой вершине дерева для объекта $x \in X$ сравниваются расстояния между x и опорными точками.

Пусть T – бинарное дерево, $Y = \{+1, -1\}$ – метки классов. При этом $\{A_i\}, \{B_j\}$ – множество опорных точек классов +1, -1 соответственно. Для каждой вершины v_i определим предикат следующим образом: "$x \in M_1^{v_i}$", где $M_1^{v_i}$ – множество точек из признакового пространства находящихся ближе к A^{v_i} чем к B^{v_i} (рисунок 2). Таким образом, происходит линейное разделение выборки (рисунок 3).

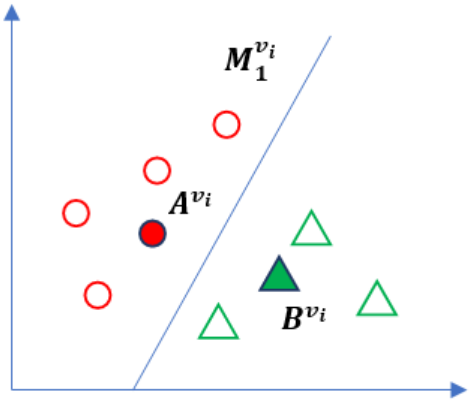


Рисунок 2. Пример однократного разбиения.

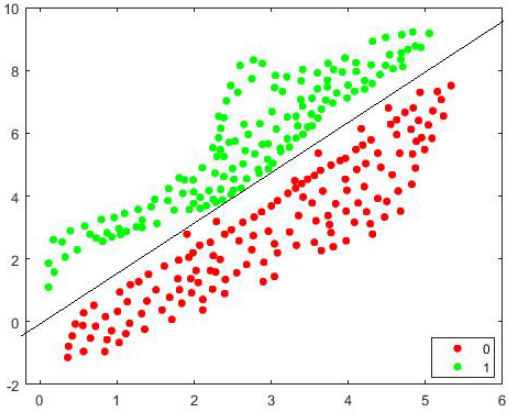


Рисунок 3. Разбиение сгенерированных данных с помощью SBDT на 2 класса.

Для $\forall x \in X^d$ определим матрицу M_x:

$$m(i,j) = \begin{cases} 1, & \rho(x, A_i) - \rho(x, B_j) < 0 \\ 0, & \rho(x, A_i) - \rho(x, B_j) > 0 \end{cases},$$

где ρ – метрика в пространстве признаков **X**.

Выполним преобразование матрицы M_x в вектор $\overrightarrow{M_x}$, "растянув" её в одну строку длиной $p * n$, где p и n – число опорных точек $\{A_i\}$ и $\{B_j\}$ соответсвенно. Тогда каждый объект x описывается вектором $\overrightarrow{M_x}$ размера $O(l^2)$, полученный преобразованием матрицы M_x в признаковый вектор. Таким образом $\boldsymbol{X'} = \{\bigcup_{x \in X^d} \overrightarrow{M_x}\}$ – новое признаковое описание выборки X^d.

Рассмотрим построение новых признаков для ситуации из рисунка 2. Пусть A, B – опорные точки класса +1 и –1. Так как число всевозможных пар опорных точек равно единице, матрица M_x является единичным элементом. Для объектов-кругов $M_x = 1$, для объектов-треугольников $M_x = 0$.

2.1. *Отбор опорных точек*

Предложенный метод классификации разделяет объекты выборки на основе их взаимного расположения. Выбор так называемых «опорных точек» должен основываться на их информативности для данной выборки X^d. В работе реализовано три способа отбора опорных точек.

Первый способ основан на алгоритме Relief, предложенном в работе [3]. Пусть X_+ и X_- подвыборки из X^l классов +1 и –1. Тогда положим $\{A_i\} = X_+$, $\{B_j\} = X_-$. После получения

векторов $\overrightarrow{M_x}$ для $\forall\, x \in X^d$, применение алгоритма Relief позволяет выбрать наиболее информативные пары $P_i = \{\langle A_i, B_i \rangle\}$, уменьшая признаковую размерность.

Второй подход использует метод опорных векторов (SVM), предложенный в работе [7]. SVM строит гиперплоскость с максимальным расстоянием между объектами разных классов. Те объекты, которые находятся на границе, принято называть *опорными векторами.* При обучении SVM на X^l, происходит разделение X^l на опорные и не опорные вектора. В качестве $\{A_i\}$ и $\{B_j\}$ возьмем опорные вектора соответствующих классов.

Третий способ основан на алгоритме кластеризации k-средних. Сгенерируем S подвыборок из X^d длины L так, чтобы в каждой подвыборке X_i содержалось $\gamma * L$ объектов из X^l и $(1-\gamma) * L$ из X_{test}, где $\gamma \in (0{,}1)$. Применим к каждой подвыборке алгоритм k-средних с разделением на 2 кластера и извлечем центры тяжести из каждого класса, которые обозначим через $A_i, B_j,\ i,j = 1, \dots, S$. Наборы $\{A_i\}$ и $\{B_j\}$ примем за искомые опорные точки.

В данном способе отбора также применялся алгоритм k-средних с использованием ядра (kernel k-means). При помощи ядра совершается переход в другое пространство большей размерности, в котором исходная конфигурация точек претерпевает изменения, зачастую приобретая более простую форму.

2.2. *Обучение и построение дерева решений*

В качестве алгоритма, определяющего структуру дерева и способ его построения, использовался CART алгоритм, описанный в работе [6], с применением критерия Gini как функционала качества разбиения подвыборок в узлах дерева. Этот алгоритм можно записать в виде последовательности трех шагов.

Шаг 1. Нахождение опорных точек $\{A_i\}$ и $\{B_j\}$ классов +1, –1.

Шаг 2. Вычисление векторов $\overrightarrow{M}_x$ для всех $x \in X^d$.

Шаг 3. Построение дерева решений в признаковом пространстве $\{\bigcup_{x \in X^d} \overrightarrow{M_x}\}$ с помощью алгоритма CART.

При использовании методов отбора SVM и k-средних без использования ядра получаем линейную трудоёмкость, а при использовании алгоритмов Relief и kernel k-means имеем квадратичную трудоемкость от объема исходной выборки.

3. Исследование на экспериментальных данных

Предложенный метод исследовался на прикладной задаче классификации раковых опухолей легких по изображениям компьютерной томографии. Исходные данные взяты с открытого ресурса www.cancerimagingarchive.net.

3.1. *Исходные данные*

Данные представлены набором 370 медицинских изображений в формате DICOM компьютерной томографии поперечных срезов легких с крупноклеточной раковой опухолью на стадиях I – III двух распространенных типов крупноклеточного рака легких (NSLC): 1) Adenocarcinoma (170 изображений), 2) Squamous Cell Carcinoma (200 изображений).

Каждому пациенту соответствует комплект 2D изображений. При формировании набора данных отбираются изображения с явно различимыми опухолями замкнутой формы, находящиеся строго внутри области легкого. Таким образом, каждое изображение (рисунок 4) представляет из себя поперечный срез легкого с замкнутой опухолью внутри легкого.

3.2. *Сегментация области интереса*

Для выделения области интереса (опухоли легкого) применялась двухэтапная процедура сегментации с помощью алгоритма «наращивания областей». На первом этапе выделялась область легких, а на втором этапе производилось выделение области опухоли.

Алгоритм сегментации легких записывается в виде последовательности трех шагов.

Шаг 1. Анализ гистограммы яркости пикселей и поиск «точек кристаллизации».

Шаг 2. Применение метода «наращивания областей».
Шаг 3. Применение морфологических операций к выделенной области.

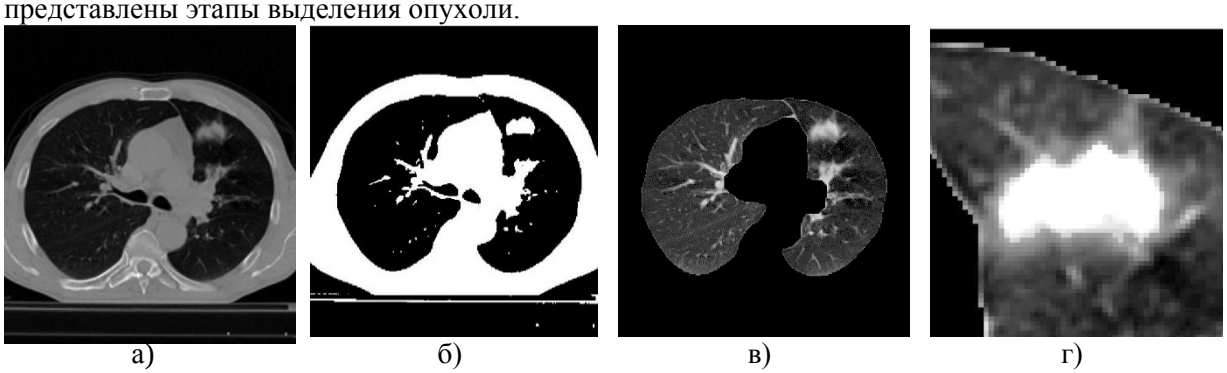
Рисунок 4. Пример исходного изображения.

Анализ гистограммы изображения показал, что интересующие области легочной ткани имеют яркости пикселей в диапазоне [0.01, 0.5], а ткани органов и внутренней полости тела – в диапазоне [0.6, 0.9]. Для получения начальных точек производилась бинаризация методом Отсу [5]. После бинарное изображение разбивалось на компоненты связности, где наиболее крупными являлись искомые области легких (рисунок 5). Далее из этих компонент отбирались точки, играющие роль начальных центров для алгоритма «наращивания областей».

а) б) в)

Рисунок 5. (а) – исходное изображение; (б) – бинаризация; (в) – нахождение наибольших компонент.

При выделении области опухоли начальная точка выбирается вручную. На рисунке 6 представлены этапы выделения опухоли.



а) б) в) г)

Рисунок 6. Этапы выделения опухоли: (а) – исходное изображение; (б) – бинаризация; (в) – выделенная область легких; (г) – выделенная область опухоли.

Весь набор извлекаемых признаков, которые извлекаются из выделенной области опухоли и далее используются как входные данные для классификатора, можно разделить на четыре

группы: геометрические, морфологические, текстурные и гистограмные. Все эти признаки описаны в работе [1]. Алгоритмы извлечения признаков были реализованы в среде MATLAB.

3.3. *Результаты экспериментов*

Исходная выборка разделена на обучающую и тестовую с сохранением пропорций классов и отсутствием изображений одного пациента в разных выборках: X^l – обучающая выборка (260 изображений), X_{test} – тестовая выборка (110 изображений).

3.3.1. *Эксперимент 1*

Метод SBDT сравнивался с методами опорных векторов (SVM), деревом решений (CART) и методом k-ближайших соседей (kNN). В качестве ядра для kernel k-means и SVM использовалось полиномиальное ядро. Результаты классификации представлены метками классов для объектов из тестовой выборки X_{test}. В таблице 1 приведено сравнение оценок точностей.

Таблица 1. Сравнение точности методов.

Метод	**Точность, %**
SBDT + SVM	83.4
SBDT + Relief	88
SBDT + k-средних	88.3
SBDT + kernel k-means	90
SVM	79.8
kNN	72.5
Дерево решений	84

3.3.2. Применение сверточной нейронной сети

Было проведено исследование ~~о~~ возможности применения методов глубокого обучения для рассматриваемой задачи. Популярным на данный момент методом является сверточная нейронная сеть, предложенная в [9]. В качестве исходной сети была использована предобученная сеть AlexNet. Была выполнена тонкая настройка параметров и дообучение на обучающем наборе с аугментацией (поворотами) изображений опухолей. Наибольшая точность составила 81%.

Таблица 2. Архитектура сети AlexNet.

[227x227x3] INPUT [55x55x96]
CONV1: 96 11x11 filters at stride 4, pad 0 [27x27x96]
MAX POOL1: 3x3 filters at stride 2 [27x27x96]
NORM1: Normalization layer [27x27x256]
CONV2: 256 5x5 filters at stride 1, pad 2 [13x13x256]
MAX POOL2: 3x3 filters at stride 2 [13x13x256]
NORM2: Normalization layer [13x13x384]
CONV3: 384 3x3 filters at stride 1, pad 1 [13x13x384]
CONV4: 384 3x3 filters at stride 1, pad 1 [13x13x256]
CONV5: 256 3x3 filters at stride 1, pad 1 [6x6x256]
MAX POOL3: 3x3 filters at stride 2 [4096]
FC6: 4096 neurons [4096]
FC7: 4096 neurons [1000]
FC8: 1000 neurons (class scores)

Исходя из результатов эксперимента можно сделать вывод, что в условиях небольшого объема обучающих данных применение сверточной сети нецелесообразно.

4. Заключение

В работе был предложен метод построения дерева решений, основанного на похожести объектов. Особенностью данного метода является использование опорных точек, полученных с помощью алгоритма Relief, метода опорных векторов (SVM) и алгоритма k-средних. Метод исследовался на задаче классификации раковых опухолей двух типов. Предложенный метод показал более высокое качество распознавания по сравнению с методом опорных векторов (SVM), деревом решений (CART) и методом k-ближайших соседей (kNN). В дальнейшем будет проведено исследование метода и на других типах задач анализа данных.

5. Литература

[1] Satrajit, B. Developing Predictive Model for Lung Tumor Analysis. – Graduate Theses and Dissertations, 2012.
[2] Gonzalez, R.C. Digital image processing using MATLAB / R. Gonzales, R.E. Woods, S.L. Eddins // Upper Saddle River. – New Jersey: Pearson-Prentice-Hall, 2004. – 624 p.
[3] Kira, K. A Practical Approach to Feature Selection / K. Kira, L. Rendell // Machine Learning Proceedings, 1992. – P. 249-256.
[4] Duda, R.O. Pattern classification and scene analysis / R.O. Duda, P. E. Hart. – New York: Wiley, 1973.
[5] Otsu, N. A threshold selection method from gray-level histograms // IEEE transactions on systems, man, and cybernetics. – 1979. – Vol. 9(1). – P. 62-66.
[6] Breiman, L. Classification and Regression Trees / L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone. – New York: Routledge, 1984. – 368 p.
[7] Vapnik, V. Support-vector networks / C. Cortes, V. Vapnik // Machine learning. – 1995. – Vol. 20(3). – P. 273-297.
[8] Breiman, L. Random forests // Machine learning. – 2001. – Vol. 45(1). – P. 5-32. DOI: 10.1023/A:1010933404324
[9] LeCun, Y. Backpropagation applied to handwritten zip code recognition / Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard // Neural computation. – 1989. – Vol. 1(4). – P. 541-551.
[10] Kotsiantis, S.B. Decision trees: a recent overview // Artificial Intelligence Review. – 2013. – Vol. 39(4). – P. 261-283. DOI: 10.1007/s10462-011-9272-4.
[11] Лбов, Г.С. Устойчивость решающих функций в задачах распознавания образов и анализа разнотипной информации / Г.С. Лбов, В.Б. Бериков. – Новосибирск: Ин-т математики СО РАН, 2005.
[12] Бериков, В.Б. Выбор оптимальной сложности класса логических решающих функций в задачах распознавания образов / Г.С. Лбов, В.Б. Бериков // Докл. РАН. – 2007. – Т. 417, № 1. – С. 26-29.
[13] Murthy, S.K. A system for induction of oblique decision trees / S.K. Murthy, S. Kasif, S. Salzberg // Journal of artificial intelligence research. – 1994. – Vol. 2. – P. 1-32.
[14] Breiman, L. Bagging predictors // Machine learning. – 1996. – Vol. 24(2). – P. 123-140.
[15] Schapire, R.E. The boosting approach to machine learning: An overview // Nonlinear estimation and classification. – Springer, New York, 2003. – P. 149-171.

Благодарности

Работа проведена при частичной поддержке РФФИ (гранты 18-07-00600а, 18-29-0904мк).

A method for similarity-based decision tree induction in the problem of recognition of tomographic images

R.M. Kozinets[1], V.B. Berikov[1,3], I.A. Pestunov[2], S.A. Rylov[2]

[1]Novosibirsk State University, Pirogova 2, Novosibirsk, Russia, 630090
[2]Institute of Computational Technologies SB RAS, Lavrentiev avenue 6, Novosibirsk, Russia, 630090
[3]Sobolev Institute of Mathematics SB RAS, Acad. Koptyug avenue 4, Novosibirsk, Russia, 630090

Abstract. The paper proposes pattern recognition method using a modification of the class of logical decision functions presented in the form of a decision tree. Instead of standard statements corresponding to the tree nodes, in which a variable is tested for certain sets of its values, a more general type of statements is used regarding the proximity of the point in question to different subsets of observations. At the same time, to determine the degree of similarity, various metrics and subspaces of features can be used. This type of decision tree allows one to get more complex decision-making boundaries, which at the same time have clear logical interpretation for the user. Several tree induction strategies are considered: based on data transformation using reference points selected with Relief, SVM, and K-means procedures. The method is experimentally investigated in the problem of tomographic images analysis. Experiments have shown that the proposed method gives more accurate predictions than CART, SVM, kNN algorithms and deep convolutional neural network (AlexNet).