Сходство по ансамблю, деревья и леса решений, основанные на сходстве, в задачах анализа данных, мультиспектральных и КТ изображений

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

1.Институт математики им. С.Л. Соболева СО РАН,
2.Новосибирский государственный университет
3.Институт вычислительных технологий СО РАН

ММРО-19

Краткое содержание

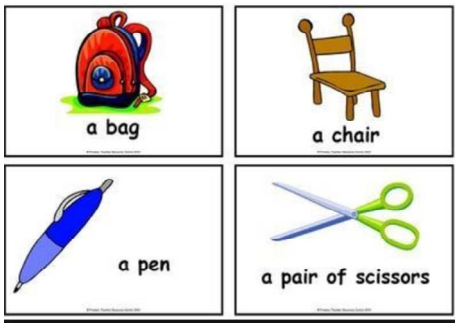
- постановка задачи;
- кластерные ансамбли и сходство по ансамблю;
- спектральный кластерный анализ и частично контролируемое обучение;
- деревья решений, основанные на сходстве;
- примеры.

Обучение с учителем в задачах распознавания образов, классификации и регрессии

$A = \{a_1, ..., a_n\}$ – множество объектов;

$X = (X_1, ..., X_d)$ – набор признаков, $x_{i,j} = X_j(a_i)$; таблица данных $\mathbf{X} = (x_{i,j})$.

Заданы метки классов для всех объектов:
$\{y_1, ..., y_N\}$, $y_i \in \{c_1, ..., c_K\}$ либо $y_i \in R$

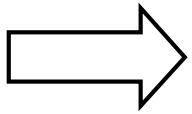


Требуется построить решающее правило:

$$\forall x \to y(x),$$

оптимальное по некоторому критерию качества

(например, $P[error] \to \min$).

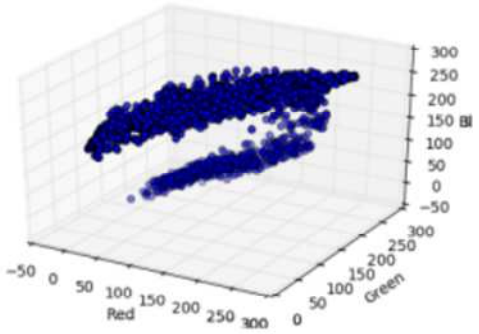


a pen

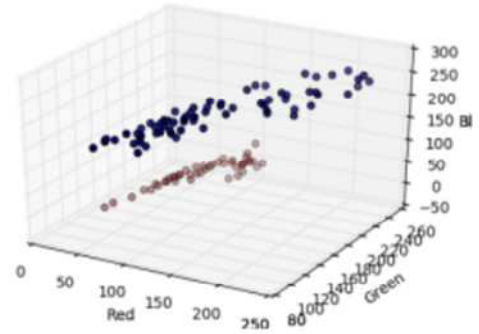
Обучение с частичным привлечением учителя
Semi-supervised learning

Размечена относительно небольшая часть данных:

X



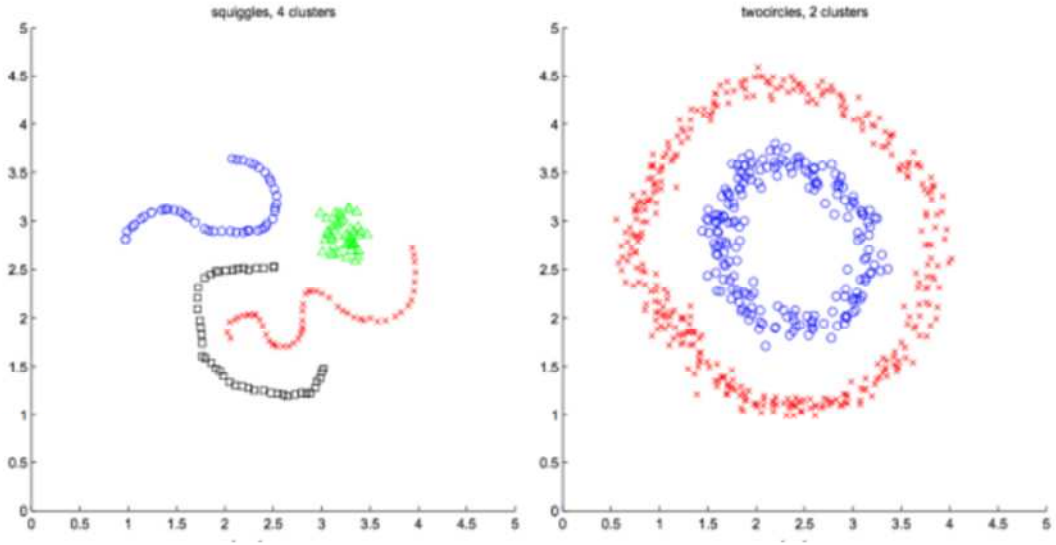
X_1



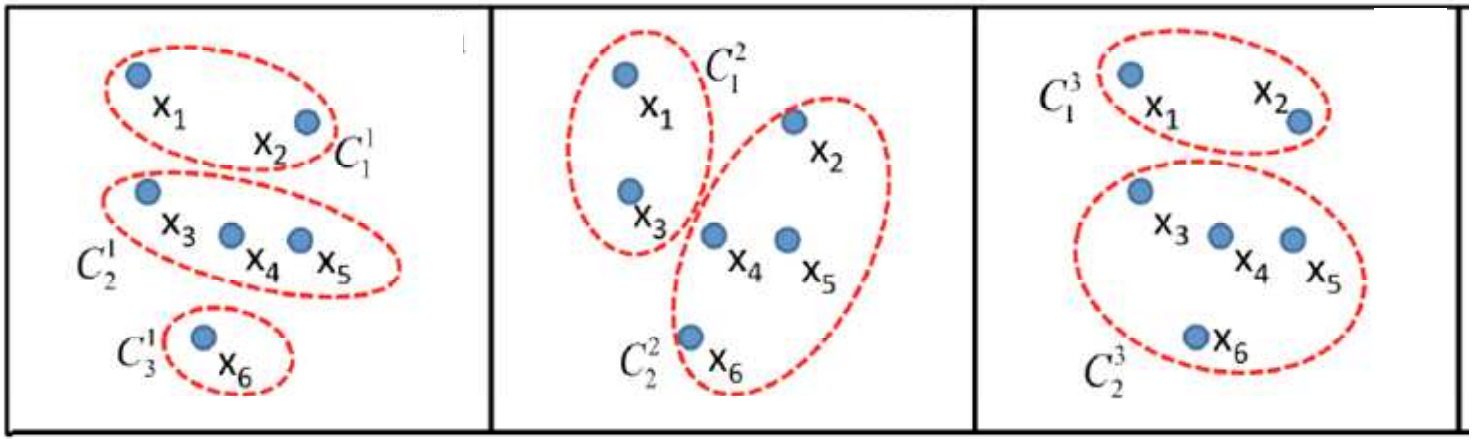
Требуется построить решающее правило для точек из неразмеченной части X_0 (трансдуктивное обучение) либо для новых объектов (индуктивное обучение).

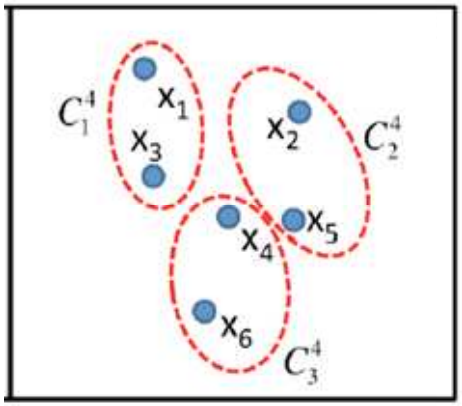
Классификация без учителя (кластерный анализ, автоматическая группировка)

Требуется найти разбиение $P = \{C_1, ..., C_K\}$ множества A на $K << n$ кластеров, наилучшее по некоторому критерию качества; K - либо задано, либо требует автоматического определения.



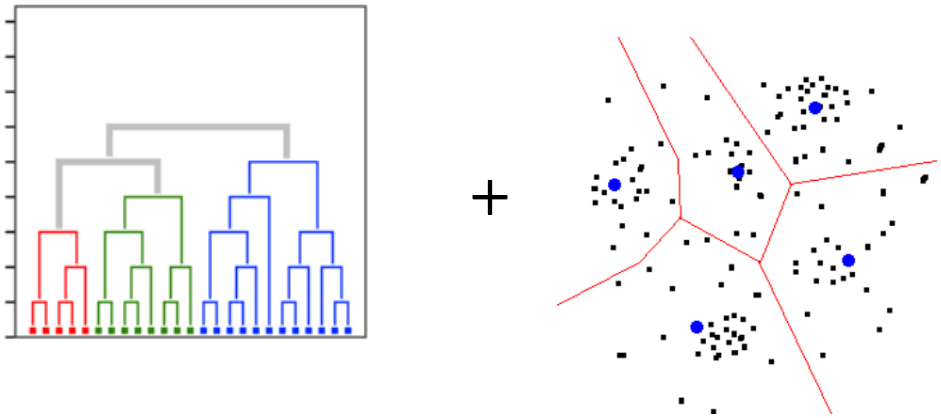
Кластерный ансамбль





консенсусное разбиение

- Однородный ансамбль – включает один алгоритм, формирующий решения для разных параметров (число кластеров, инициализации, подмножества переменных и т.п.);
- Разнородный ансамбль – включает несколько алгоритмов различных типов.



Преимущества ансамблевой кластеризации:

- Известно, что комбинация решений полезна при обучении с учителем (boosting, bagging, decision forest);
- Улучшение качества группировки

  много «слабых» решений -> «сильное» решение;
- Снижение зависимости от неудачного выбора параметров алгоритма;
- Возможность проведения распределенных вычислений (при различном местоположении подмножеств объектов или переменных; Big data).

Усредненная коассоциативная матрица

l-й вариант разбиения на кластеры $\rightarrow$ бинарная матрица

$$H_l = \{ h_l(i, j) \},$$

где $h_l(i, j) = 1$, если a_i и a_j принадлежат одному кластеру;

$h_l(i, j) = 0$, иначе, $i, j = 1, 2, \ldots, n$.

Усредненная с весами коассоциативная матрица $H^* = \{ h^*(i, j) \}$,

$$h^*(i, j) = \sum_{l=1}^{L} w_l h_l(i, j),\ w_l \geq 0,\ \sum_l w_l = 1.$$

Нахождение по H^* итогового варианта:

используется некоторый произвольный алгоритм, который в качестве входной информации использует меры сходства между парами объектов.

Свойства усредненной коассоциативной матрицы

1. $h^*(i, j)$ определяет псевдометрику.

$\Rightarrow$ элементы H^* могут рассматриваться как аналоги мер расстояния или сходства между наблюдениями.

2. H^* является неотрицательно определенной (удовлетворяет условиям теоремы Мерсера).

$\Rightarrow$ может быть использована в «ядерных» методах обучения, в частности, в SVM, kernel NN, kernel Fisher Discriminant.

3. При некоторых условиях регулярности, вероятность ошибки классификации по кластерам с использованием H^*, для произвольной пары точек, стремится к нулю при росте числа элементов ансамбля[1].

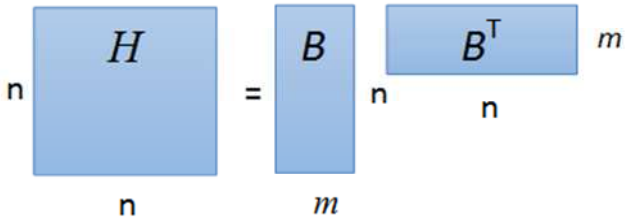
1.Berikov V., Pestunov I. Ensemble clustering based on weighted co-association matrices: Error bound and convergence properties // Pattern Recognition. 2017. Vol. 63. P. 427-436.

4. Усредненная матрица коассоциации может быть представлена в малоранговой форме[2]:

$$H^* = BB^T,\ B = [B_1 B_2 \dots B_L]$$

где B - блочная матрица, $B_l = \sqrt{w_l}\, A_l$, A_l - $n \times K_l$ матрица ассоциации для l-го зазбиения: $A_l(i,k) = \mathbb{I}[c(x_i) = k]$, $i = 1,\dots,n$, $k = 1,\dots,K_l$.

Как правило, $m = \sum_l K_l \ll n$



$\Rightarrow$ экономия памяти при хранении $n \times m$ разряженной матрицы вместо полной $n \times n$ матрицы.

Сложность умножения $H^* \cdot x$ уменьшается с $O(n^2)$ до $O(nm)$.

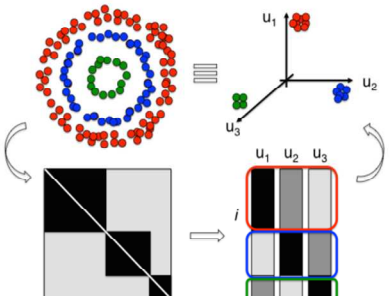
---

2. Berikov V. Semi-supervised Classification Using Multiple Clustering and Low-Rank Matrix Operations // LNCS. 2019, Vol 11548, pp. 529-540.

Предлагаемый подход

Усредненая матрица коассоциации рассматривается как матрица сходства объектов

- Если точки часто объединяются в один кластер, значит между ними имеется сходство, даже если кластер имеет сложную (вытянутую и т.п.) форму (предполагается, что существует преобразование, в результате которого кластеры сложной формы окажутся компакными);



- Существуют параллели из когнитивной психологии[3]
- Веса элементов ансамбля могут определяться оптимальным образом, с учетом оценок качества разбиений[4];
- Возможность выявить структурные мета-признаки для «похожих» данных $\Rightarrow$ трансферное обучение.

3. Bruner J. Beyond the information given: Studies in the Psychology of Knowing. W. W. Norton & Company, 1973.
4. Berikov V. Construction of an optimal collective decision in cluster analysis on the basis of an averaged co-association matrix and cluster validity indices // Pattern Recognition and Image Analysis. 2017. Vol. 27(2), P. 153-165.

Эксперименты: cluster ensemble kernel + SVM; UCI datasets; без зашумления и с зашумлением

Noise level is $\sigma = 0$.

Data	KCCE	SVM	KerFish
iris	0.943	**0.973**	0.96
canc	0.958	**0.967**	0.957
glas	0.619	**0.705**	0.615
park	0.846	0.891	**0.938**
gest	0.66	**0.719**	0.592
kidn	0.98	**1**	0.903
CTG	0.934	**0.986**	0.942
lung	**0.766**	0.675	0.708
P-b	0.938	**0.964**	0.94
thr	**0.85**	**0.85**	0.486
tran	**0.765**	0.761	0.65
sale	0.899	**0.906**	0.824

Noise level is $\sigma = 1.25$.

Data	KCCE	SVM	KerFish
iris	**0.701**	0.692	0.57
canc	0.885	**0.909**	0.607
glas	**0.44**	0.423	0.182
park	**0.768**	0.752	0.712
gest	0.51	**0.548**	0.392
kidn	**0.910**	0.736	0.303
CTG	**0.823**	0.778	0.236
lung	**0.738**	0.675	0.708
P-b	0.906	**0.913**	0.454
thr	**0.85**	**0.85**	0.343
tran	**0.763**	0.761	0.583
sale	**0.751**	0.734	0.589

Алгоритм спектрального ансамблевого кластерного анализа

Вход: $H^* = BB^T$ - матрица схожести объектов множества A; K– число кластеров в искомом разбиении.

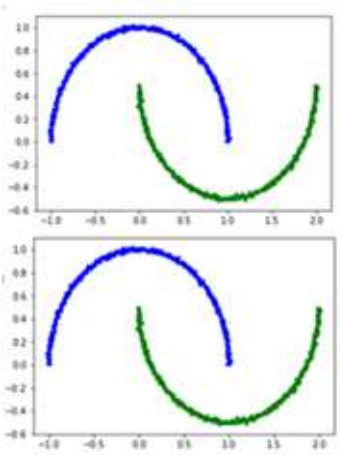
Выход: разбиение A на кластеры $C_1,...,C_K$.

1. вычислить Лапласиан графа сходства в малоранговом представлении: $\mathbf{L} = D - BB^T$, где $D = diag(d_1,...,d_n)$, $d_i = \sum_{j=1}^{n} h^*(i,j)$.
2. с помощью метода степенной итерации вычислить первые K собственных векторов $u_1,\ldots,u_K$ Лапласиана, соответствующих наименьшим собственным числам;
3. определить матрицу V размерности $n \times K$, столбцами которой являются $u_1,\ldots,u_K$;
4. найти разбиение $C_1,...,C_K$ по V (например, с помощью алгоритма K-средних);

end.

Трудоемкость алгоритма: близка к $O(n)$**, память:** $O(n)$**.**

Сравнение кластеризаций, полученных стандартным SC (sklearn) и малоранговым алгоритмами (L=10)



Набор точек	n	Время работы алгоритма, с	
		Полн. ранг	Мал. ранг
2 полумесяца	10^3	0.69	0.08
	$3\cdot10^3$	5.46	0.38
	$5\cdot10^3$	13.67	0.65
	10^4	57.63	1.74
	$3\cdot10^4$	-	5.2
	10^5	-	26.1

Обучение с частичным привлечением учителя: Graph Laplacian regularization (transductive regression)

$X = \{x_1, \ldots, x_n\} = X_1 \cup X_0$, где $X_1 = \{x_1, \ldots, x_{n_1}\}$ - размеченная часть с метками $Y_1 = \{y_1, \ldots, y_{n_1}\}$, X_0 - неразмеченная часть.

Задача: найти $f^* = (f_1^*, ..., f_N^*)$:

$$f^* = \arg\min_{f \in \mathbf{R}^n} Q(f) = \frac{1}{2}\Big(\underbrace{\sum_{x_i \in X_1} (f_i - y_i)^2}_{fitting\ error} + \alpha \underbrace{\sum_{x_i, x_j \in \mathbf{X}} h^*(i,j)(f_i - f_j)^2}_{smoothing\ term} + \beta \| f \|^2 \Big)$$

где f - вектор прогнозов, $\alpha, \beta > 0$ параметры регуляризации.

Обозначим $Y_{1,0} = (y_1, \ldots, y_{n_1}, \underbrace{0, \ldots, 0}_{n-n_1})^T$, $S = G + \alpha D$, $G = diag(G_{11} \ldots, G_{nn})$,

$$G_{ii} = \begin{cases} \beta + 1, i = 1, ..., n_1 \\ \beta, i = n_1 + 1, ..., n \end{cases}, \quad \frac{\partial Q}{\partial f} = 0 \Rightarrow \text{решение:}$$

$$f^* = (G + \alpha L)^{-1} Y_{1,0} = (S - \alpha B B^T)^{-1} Y_{1,0}.$$

V. Berikov and A. Litvinenko, Semi-supervised regression using cluster ensemble and low-rank co-association matrix decomposition under uncertainties, UNCECOMP 2019. Pp. 229-242.

Тождество Вудбери (матричная алгебра):

$$(S+UV)^{-1} = S^{-1} - S^{-1}U(I+VS^{-1}U)^{-1}VS^{-1}$$

где $S \in \mathbf{R}^{n\times n}$ - обратимая матрица, $U \in \mathbf{R}^{n\times m}$ и $V \in \mathbf{R}^{m\times n}$. Тогда

$S^{-1} = \mathrm{diag}(1/(G_{11}+\alpha D_{11}),\ldots,1/(G_{nn}+\alpha D_{nn})) \Rightarrow$

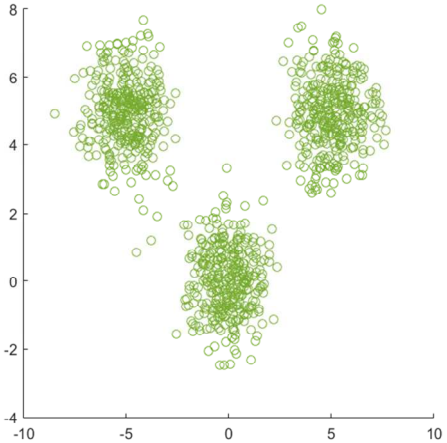
$$\boxed{f^* = (S^{-1} + \alpha S^{-1}B(I-\alpha B^T S^{-1}B)^{-1}BS^{-1})\, Y_{1,0}}$$

$\Rightarrow$ Требуется обратить $m\times m$ матрицу вместо $n\times n$, где $m \ll n$

Трудоемкость $O(nm+m^3)$.

Возможно также получение прогнозного вектора путем численного решения СЛАУ.

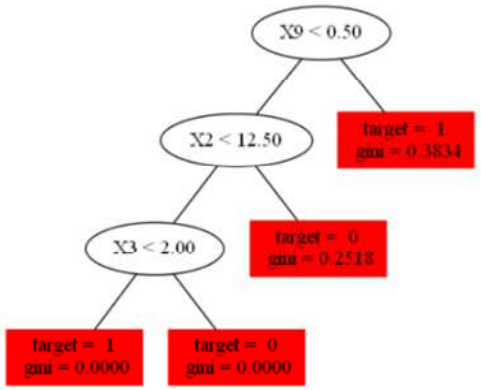
Результаты МК экспериментов

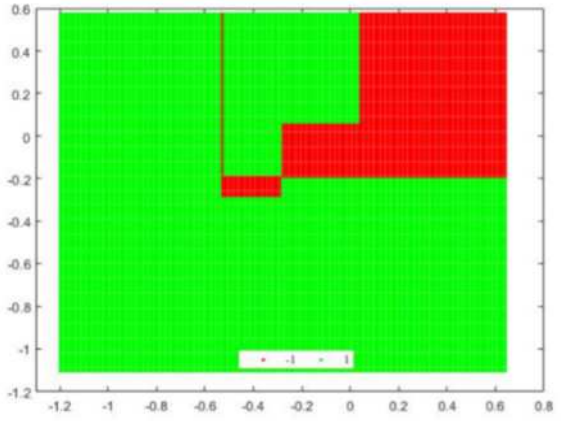


n	σ_ε	SSR-LRCM			SSR-RBF	
		RMSE	t_{ens} (sec)	t_{matr} (sec)	RMSE	time (sec)
1000	0.01	**0.052**	0.06	0.02	**0.085**	0.10
	0.1	**0.054**	0.04	0.04	**0.085**	0.07
	0.25	**0.060**	0.04	0.04	**0.102**	0.07
3000	0.01	**0.049**	0.06	0.02	**0.145**	0.74
	0.1	**0.051**	0.06	0.02	**0.143**	0.75
	0.25	**0.053**	0.07	0.02	**0.150**	0.79
7000	0.01	**0.050**	0.16	0.08	**0.228**	5.70
	0.1	**0.050**	0.16	0.08	**0.229**	5.63
	0.25	**0.051**	0.14	0.07	**0.227**	5.66
10^5	0.01	0.051	1.51	0.50	-	-
10^6	0.01	0.051	17.7	6.68	-	-

Деревья решений на основе сходства

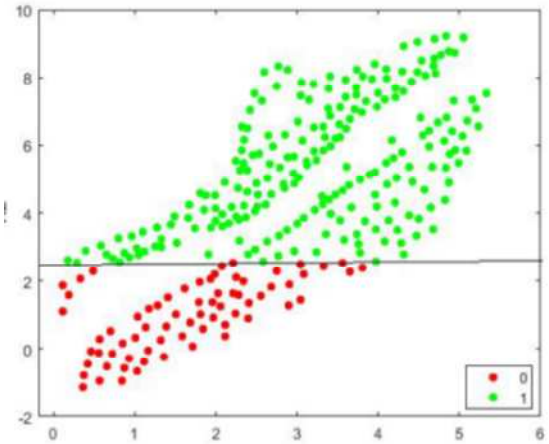
Стандартное дерево решений:





Деревья с линейными условиями (oblique DT):

$$\sum \beta_j X_j(a) + \beta_0 < 0$$

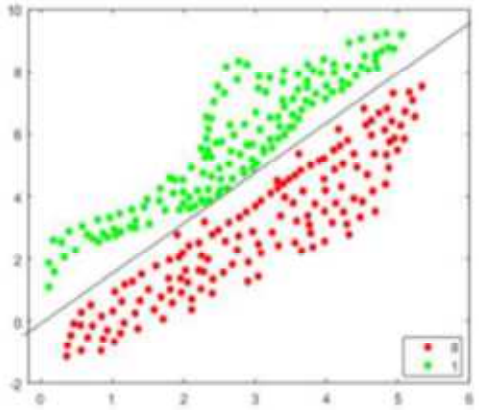


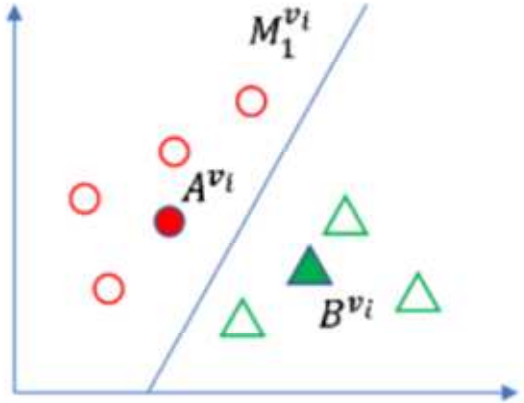
- сложность интерпретации.

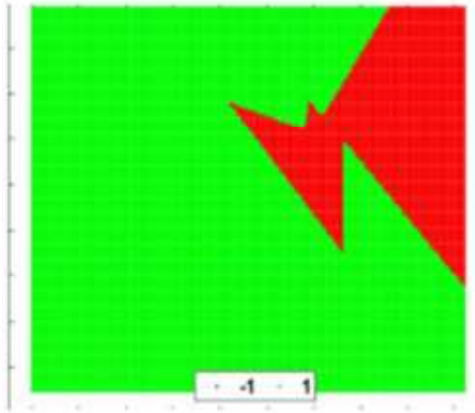
Деревья решений на основе сходства

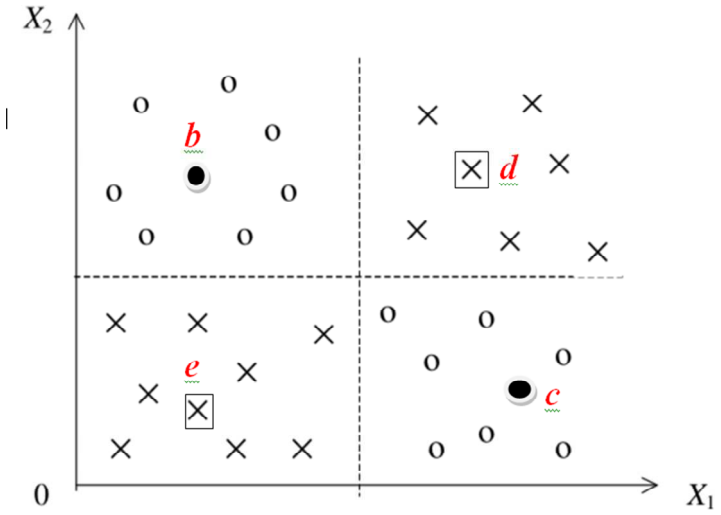
В узлах дерева проверяются высказывания вида:

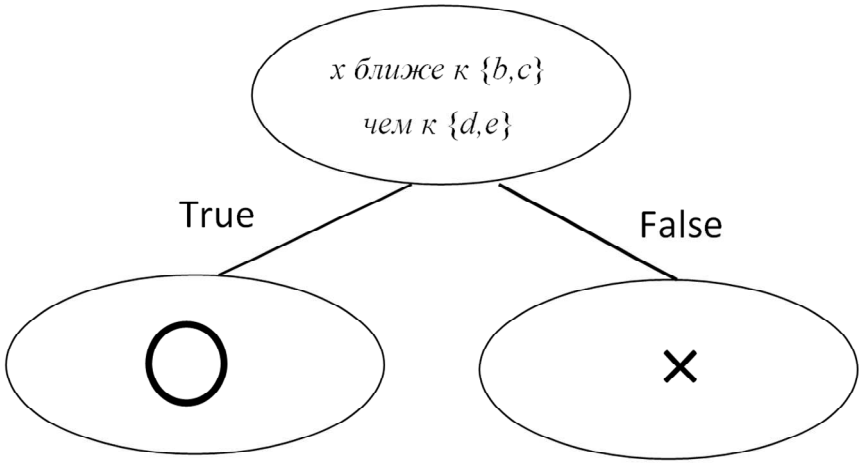
«объект x более схож с $\{B\}$ чем с $\{C\}$ в признаковом подпространстве X' по метрике μ»











Построение SBDT

Схема алгоритма

1. Найти «опорные» точки $s_1,...,s_M$ и сформировать множество допустимых для сравнения пар $G=\{(s_j,s_k)\}$, $r=|G|$;

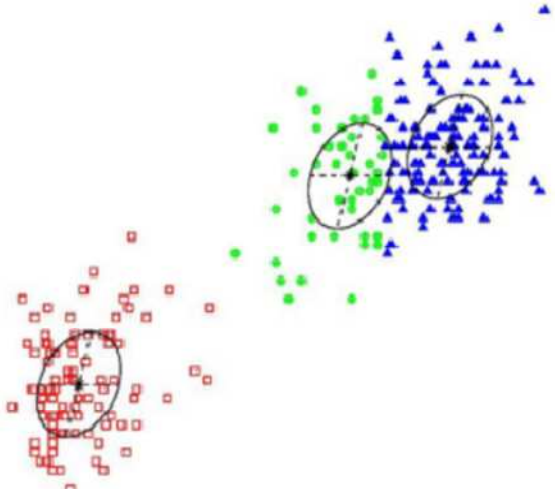
2. Преобразовать исходные данные: $\mathbf{X}_{n\times d} \Rightarrow \tilde{\mathbf{X}}_{n\times r}$:

$$x_i \Rightarrow \tilde{x}_{ijk} = \begin{cases} 1, \ x_i \ \textit{ближе к} \ s_j \ \textit{чем к} \ s_k \\ 0 \ \textit{иначе} \end{cases}$$

3. Построить дерево решений некоторым «стандартным» методом по преобразованным данным $\tilde{\mathbf{X}} \Rightarrow$ сформировать соответствующее SBDT.

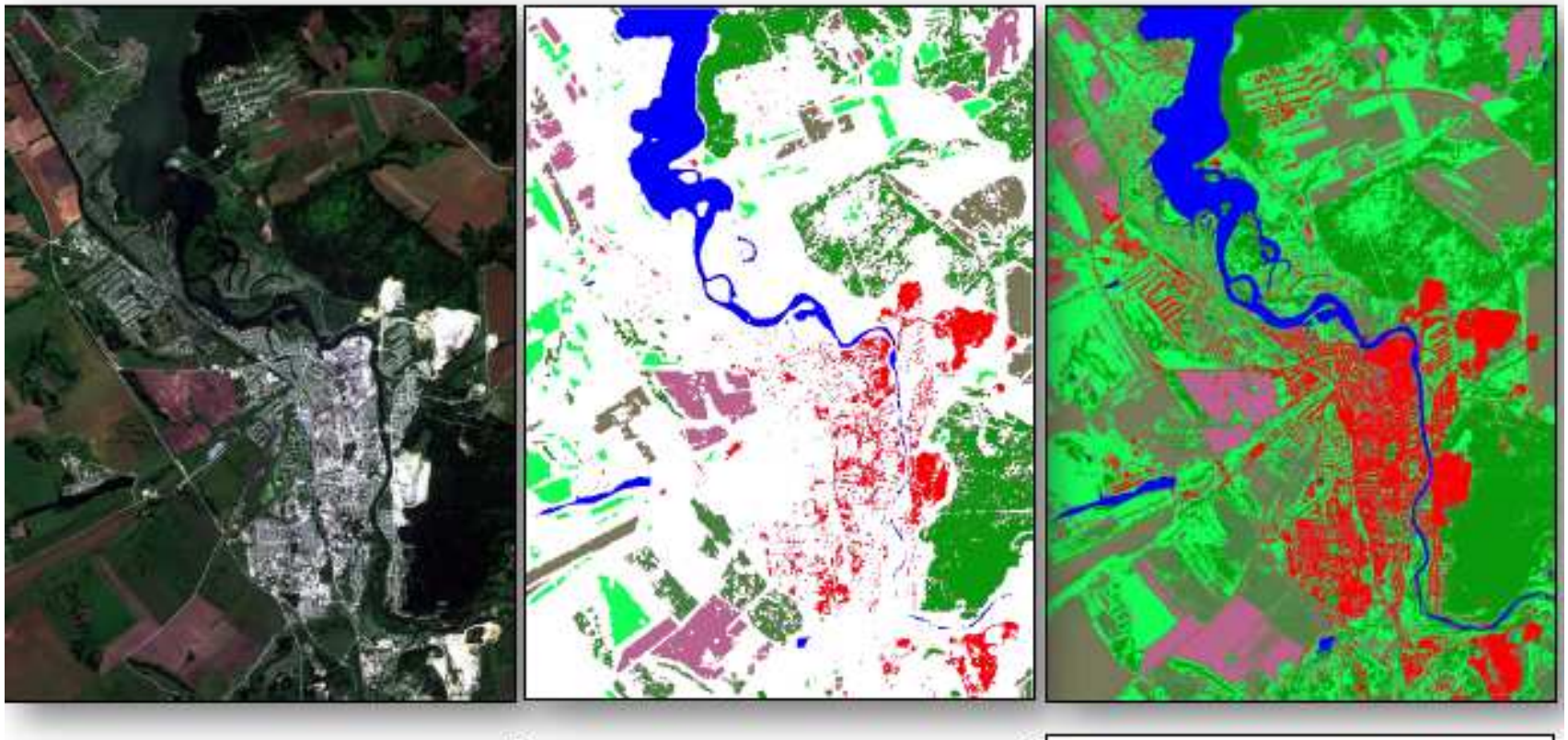
Отбор опорных точек

1. Селекция с помощью алгоритма поиска информативных признаков Relief[4].
2. Использование опорных векторов SVM.
3. Центроиды кластеров K-means.



4. Kira, K. A Practical Approach to Feature Selection / K. Kira, L. Rendell // Machine Learning Proceedings, 1992. – P. 249-256.

Эксперименты: мультиспектральное изображение; сравнение SBDT и CART



Спутниковое изображение Landsat-8	Тестовая выборка (10% – для обучения)	**Результат классификации** Точность = 98% (SBDT) Точность = 88% (CART)

Эксперименты: распознавание КТ изображений с использованием а) «классических» признаков, б) CNN

- Распознавание типа опухоли легких
- Набор данных

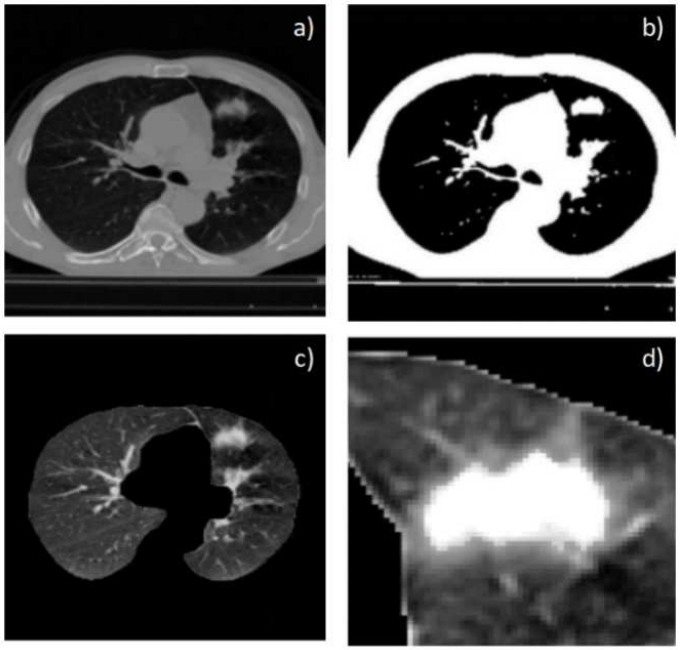
  371 DICOM-изображений 512x512. Два класса:

  - Adenocarcioma 171 снимок
  - Squamous Cell Carcinoma 200 снимков

- Схема решения:

Сегментация легкого и опухоли

a) исходный снимок; b) бинаризация; c) сегментация легкого;
d) – сегментация опухоли



Формирование признаков

Четыре группы признаков[7] (всего 60):

1.Геометрические
2.Морфологические
3.Текстурные
4.Гистограммные

7.Satrajit, B. Developing Predictive Model for Lung Tumor Analysis / B. Satrajit – Graduate Theses and Dissertations. 2012.

Результаты экспериментов

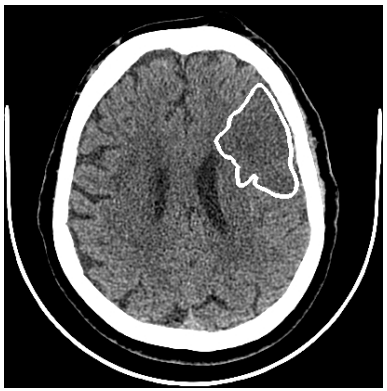
- Объем обучающей выборки: 260
- Объем контрольной выборки: 111

Algorithm	Accuracy, %
SBDT + k-means	90.
SBDT + Relief	88
CART decision tree	84
SBDT + SVM(linear)	83
AlexNet	81.5
SVM(linear kernel)	79
kNN (k=5)	72.5

Другие применения

- лес SBDT, SBDT + XGBoost;

- Распознавание ишемического и геморрагического инсульта по КТ изображениям с использованием разрабатываемого подхода; объяснение выводов сети.



Выводы

- С использованием усредненной коассоциативной матрицы кластерного ансамбля в качестве матрицы сходства объектов:

  а) разработан алгоритм ансамблевого спектрального кластерного анализа с применением малорангового представления матрицы сходства по ансамблю;

  б) предложен алгоритм частично контролируемого обучения на основе малорангового представления;

- Разработан алгоритм построения дерева решений по сходству;

- Эффективность разработанных алгоритмов экспериментальна подтверждена на тестовых задачах и реальных задачах анализа изображений.

Спасибо за внимание!