САМАРСКИЙ УНИВЕРСИТЕТ
SAMARA UNIVERSITY

Отбор признаков в задачах анализа данных

Доцент кафедры технической кибернетики
Самарского университета
Гайдель Андрей Викторович

Самара 2019

Набор данных в большинстве задач представляет собой таблицу, в которой строки являются **объектами наблюдения**, а столбцы – **признаками** (атрибутами) этих объектов.

Виды признаков:

- бинарные (принимают одно из двух значений);
- категориальные (принимают значение из конечного неупорядоченного множества);
- порядковые признаки (принимают значение из упорядоченного, но не метрического множества);
- числовые (вещественные, непрерывные) (значением являются числа);
- неструктурированные признаки (всё остальное).

Обучающая выборка $U \in \mathbf{R}^{N \times M}$
N объектов, M признаков

M

Имя	Раса	Рост	Масса	Цвет глаз	Цвет меча	Злой
Luke Skywalker	Human	172	77	blue	green	No
Obi-Wan Kenobi	Human	182	81	gray-blue	blue	No
Darth Vader	Human	203	120	blue	red	Yes
Darth Maul	Zabrak	175	80	red	red	Yes

N

ЗАДАЧА КЛАССИФИКАЦИИ

Всё множество объектов наблюдения $\boldsymbol{\Omega}$ разделено на L классов $\{\boldsymbol{\Omega}_l\}$, так что классы не пересекаются и каждый объект принадлежит одному из классов:

$$\forall i, j \in [1; L] \cap \mathbb{N}: \quad \boldsymbol{\Omega}_i \cap \boldsymbol{\Omega}_j = \varnothing;$$

$$\bigsqcup_{l=1}^{L} \boldsymbol{\Omega}_l = \boldsymbol{\Omega}.$$

Идеальный классификатор Φ переводит объект в его класс:

$$\forall l \in [1; L] \; \forall \omega \in \boldsymbol{\Omega}_l : \Phi(\omega) = \boldsymbol{\Omega}_l.$$

Решить задачу классификации – значит построить классификатор $\hat{\Phi}$, который тоже пытается перевести объект в его класс, но не владеет информацией о том, из какого класса этот объект, а использует лишь обучающую выборку U.

Имя	Раса	Рост	Масса	Цвет глаз	Цвет меча	Злой
Luke Skywalker	Human	172	77	blue	green	No
Obi-Wan Kenobi	Human	182	81	gray-blue	blue	No
Darth Vader	Human	203	120	blue	red	Yes
Darth Maul	Zabrak	175	80	red	red	Yes

Векторы $U(i, :)$ (строки таблицы) представляют собой точки в m-мерном **признаковом пространстве**.

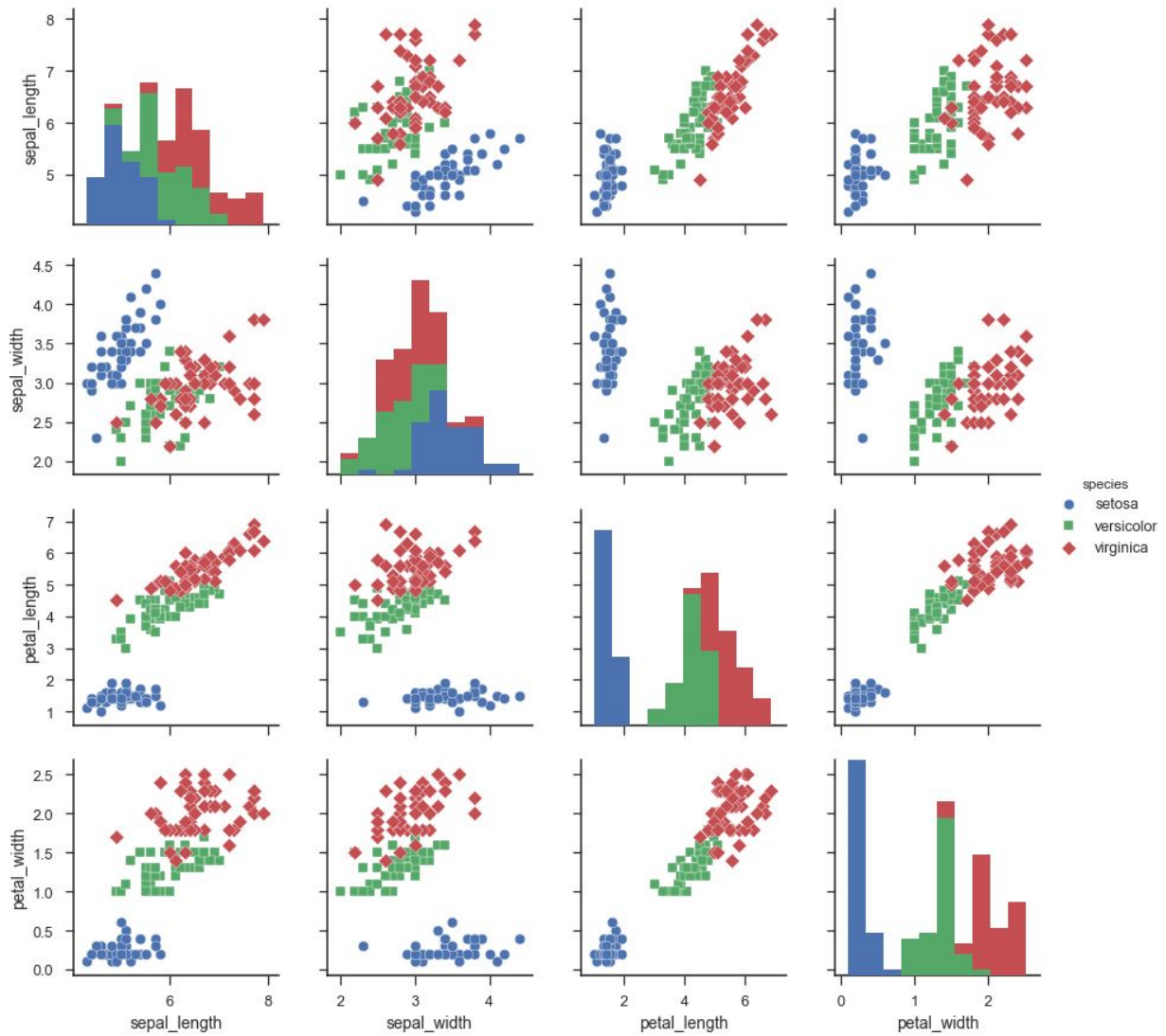
Отбор признаков – это процесс выбора некоторого конечного подмножества признаков с целью улучшения качества соответствующего признакового пространства и повышения эффективности дальнейшего решения задачи анализа данных (например, задачи классификации). При этом определённые столбцы исключаются из исходной обучающей выборки и далее не рассматриваются.

Назначение отбора признаков:

- упрощение модели для визуализации и понимания человеком,
- сокращение времени работы алгоритмов анализа данных,
- принципиальное снижение сложности исходной задачи анализа данных, устранение избыточности,
- повышение эффективности решения задач анализа данных за счёт компенсации недостатков алгоритмов решения этих задач хорошим качеством признакового пространства.

Имя	Раса	Рост	Масса	Цвет глаз	Цвет меча	Злой
Luke Skywalker	Human	172	77	blue	green	No
Obi-Wan Kenobi	Human	182	81	gray-blue	blue	No
Darth Vader	Human	203	120	blue	red	Yes
Darth Maul	Zabrak	175	80	red	red	Yes

ИРИСЫ ФИШЕРА



ПОСТАНОВКА ЗАДАЧИ ОТБОРА ПРИЗНАКОВ

Решить задачу отбора признаков – значит выбрать множество признаков $A \subseteq [1; M] \cap \mathbf{Z}$, так чтобы некоторый показатель J качества признакового пространства достигал своего максимального значения. Этот показатель качества зависит от выбранного множества признаков A и от обучающей выборки U.

$$\arg\max_{A \in 2^{[1;L] \cap \mathbf{Z}}} J(A, U)$$

Задача отбора признаков в большинстве постановок является **NP-сложной**, что впервые было показано ещё в [Amaldi, E. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems / E. Amaldi, V. Kann // Theoretical Computer Science. – 1998. – Vol. 209(1). – P. 237-260].

Постановки задачи отбора признаков могут отличаться характером признаков и видом показателя качества J.

Оптимальный по вычислительной сложности алгоритм для большинства постановок задачи отбора признаков: $[1; M] \cap \mathbf{Z}$.
полный перебор всех подмножеств A множества признаков

Вычислительная сложность большинства таких алгоритмов отбора признаков: $O(N M 2^M)$.

КЛАССИФИКАЦИЯ АЛГОРИТМОВ ОТБОРА ПРИЗНАКОВ

По характеру показателя качества признакового пространства J:

- **Обёртки** (*wrappers*) – методы отбора признаков, использующие в качестве показателей качества значения эффективности предсказательных моделей, например, классификаторов.
  Достоинство: подбирается оптимальный набор признаков для конкретного классификатора.
  Недостаток: высокая вычислительная сложность, поскольку для каждого набора признаков требуется заново обучать классификатор.
  Примеры: достоверность классификации, коэффициент детерминации, F-мера Ван Ризбергена
- **Фильтры** (*filters*) – методы отбора признаков, использующие в качестве показателей качества свойства самого признакового пространства, в котором лежат объекты из обучающей выборки.
  Достоинство: выбирается универсальный набор признаков, не зависящий от классификатора.
  Недостаток: предсказательная сила конкретного классификатора может быть хуже, чем при использовании обёртки.
  Примеры: взаимная информация, корреляция признаков, критерии дискриминантного анализа.
- **Встроенные методы** (*embedded*) – это методы отбора признаков, естественным образом встроенные в сами предсказательные модели.
  Достоинство: каждый такой метод имеет хорошую основу и может получить уникальные результаты для конкретной задачи.
  Недостаток: если требуется только отобрать признаки, то вся остальная часть предсказательной модели не используется.
  Примеры: решающий лес, регуляризованные регрессионные модели.

$$\mathrm{TP} = \left|\left\{ i \in [1;N] \cap \mathbf{Z} \,|\, \tilde{\Phi}\left(\tilde{U}(i,:)\right) = 1 \wedge \Phi\left(\tilde{U}(i,:)\right) = 1 \right\}\right|$$

$$\mathrm{FP} = \left|\left\{ i \in [1;N] \cap \mathbf{Z} \,|\, \tilde{\Phi}\left(\tilde{U}(i,:)\right) = 1 \wedge \Phi\left(\tilde{U}(i,:)\right) = 0 \right\}\right|$$

$$\mathrm{FN} = \left|\left\{ i \in [1;N] \cap \mathbf{Z} \,|\, \tilde{\Phi}\left(\tilde{U}(i,:)\right) = 0 \wedge \Phi\left(\tilde{U}(i,:)\right) = 1 \right\}\right|$$

$$\mathrm{TN} = \left|\left\{ i \in [1;N] \cap \mathbf{Z} \,|\, \tilde{\Phi}\left(\tilde{U}(i,:)\right) = 0 \wedge \Phi\left(\tilde{U}(i,:)\right) = 0 \right\}\right|$$

$\tilde{U}$ – контрольная выборка

TP	FP
FN	TN

Матрица ошибок (*Confusion Matrix*)

Достоверность классификации (*accuracy*):

$$J = \frac{1}{N}\left|\left\{ i \in [1;N] \cap \mathbf{Z} \,|\, \tilde{\Phi}\left(\tilde{U}(i,:)\right) = \Phi\left(\tilde{U}(i,:)\right) \right\}\right| = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+\mathrm{TN}}$$

Точность (*precision*):

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$$

Чувствительность (*sensitivity, recall*):

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

Специфичность (*specificity, selectivity*):

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}$$

F-мера Ван Ризбергена (F_1-*score, F-score, F-measure*):

$$J_F = 2\frac{\mathrm{PPV}\cdot\mathrm{TPR}}{\mathrm{PPV}+\mathrm{TPR}}$$

Матрица рассеяния внутри класса l:

$$R_l(i,j)=\frac{1}{N}\sum_k\left(U_l(k,i)-\overline{x}_l(i)\right)\left(U_l(k,j)-\overline{x}_l(j)\right),$$

где

$U_l=\left(U(i,j)\right)_{\substack{i\in\{i\in[1;N]\cap\mathbf{Z}\mid\Phi(U(i,:))=l\}\\ j\in[1;M]\cap\mathbf{Z}}}$ – часть обучающей выборки из класса $\mathbf{\Omega}_l$,

$\overline{x}_l=\frac{1}{N}\sum_i U_l(i,:)$ – среднее значение признаков в классе $\mathbf{\Omega}_l$.

Внутриклассовая матрица рассеяния:

$$R_\Sigma=\frac{1}{L}\sum_{l=1}^{L}R_l.$$

Матрица рассеяния между классами:

$$R(i,j)=\frac{1}{N}\sum_{k=1}^{N}\left(U(k,i)-\overline{x}(i)\right)\left(U(k,j)-\overline{x}(j)\right),$$

$\overline{x}=\frac{1}{N}\sum_{i=1}^{N}U(i,:)$ – среднее значение признаков во всей выборке.

Критерии дискриминантного анализа:

$$J_1 = \mathrm{tr}\left(R_\Sigma^{-1} R\right),$$

$$J_2 = \ln\left(\left|R\right| / \left|R_\Sigma\right|\right),$$

$$J_3 = \mathrm{tr}\left(R\right) - \mu\left(\mathrm{tr}\left(R_\Sigma\right) - c\right),$$

$$\boxed{J_4 = \frac{\mathrm{tr}\left(R\right)}{\mathrm{tr}\left(R_\Sigma\right)}.}$$

Другие способы выбора внутриклассовой матрицы рассеяния:

$$R_\Sigma = \sum_{l=1}^{L} \frac{\left|\left\{i \in [1;N] \cap \mathbf{Z} \mid \Phi\left(U\left(i,:\right)\right) = l\right\}\right|}{N} R_l.$$

Другие способы выбора матрицы рассеяния между классами:

$$R = \sum_{l=1}^{L} \frac{\left|\left\{i \in [1;N] \cap \mathbf{Z} \mid \Phi\left(U\left(i,:\right)\right) = l\right\}\right|}{N} \left(\overline{x}_l - \overline{x}\right)\left(\overline{x}_l - \overline{x}\right)^T.$$

[Fukunaga, K. Introduction to Statistical Pattern Recognition / K. Fukunaga. – Academic Press, 1990. – 592 p.]

Взаимная информация (*mutual information*):

$$I(X;Y) = H(Y) - H(Y \mid X)$$

X – признаки случайного объекта, Y – класс этого объекта

Энтропия (*entropy*):

$$H(Y) = -\sum_{l=1}^{L} \mathrm{P}(Y = l) \log_2 \mathrm{P}(Y = l)$$

Условная энтропия (*conditional entropy*):

$$H(Y \mid X) = -\sum_{k} \mathrm{P}(X = x_k) \sum_{l=1}^{L} \mathrm{P}(Y = l \mid X = x_k) \log_2 \mathrm{P}(Y = l \mid X = x_k)$$

$$H(Y \mid X) = H(X \mid Y) - H(X) + H(Y)$$

Для непрерывных признаков

$$H(X \mid Y) = -\sum_{l=1}^{L} P(Y = l) \int_{\Omega} f(x \mid Y = l) \log_2 f(x \mid Y = l) \, \mathrm{d}x$$

$f(x \mid Y = l)$ – условная плотность вероятности вектора признаков в точке x для объектов из класса l

По способу перебора подмножеств признаков:

- Жадные алгоритмы отбора признаков
  - Упорядочение и отбор признаков (*best first*)
  - Жадный прямой отбор признаков (*greedy forward selection*)
  - Жадное обратное исключение признаков (*greedy backward elimination*)
- Кластеризация признаков
- Эволюционные алгоритмы
  - Генетический алгоритм
  - Роевой алгоритм
- Алгоритмы оптимизации на графах
- Полный перебор

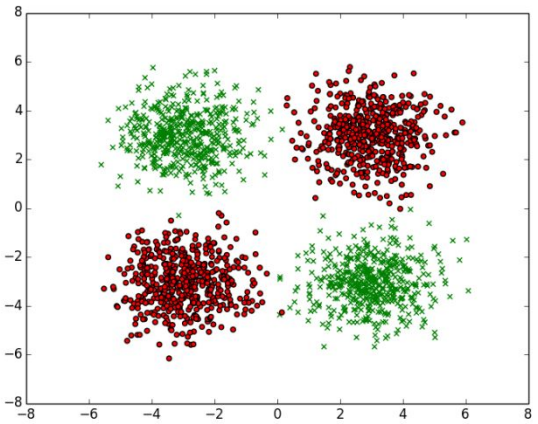
Входные данные: количество признаков k, обучающая выборка U, показатель качества J
Выходные данные: подмножество признаков A
Алгоритм:

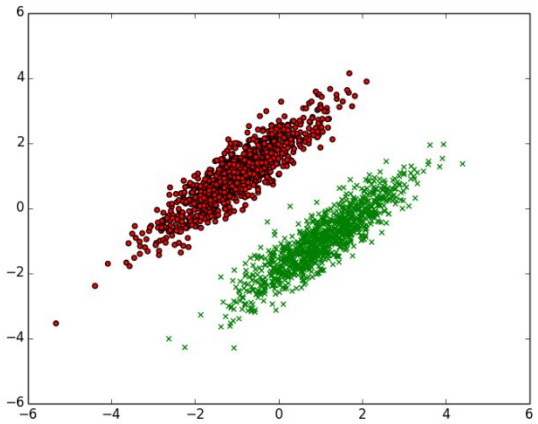
1. Для каждого из M признаков вычислить индивидуальный показатель качества $J(j)$ этого признака.
2. Упорядочить признаки по убыванию показателя качества $J(j)$.
3. Выбрать k признаков с самым высоким показателем качества $J(j)$.

Вычислительная сложность: $O(N\,M \log M)$

Достоинство: высокая скорость работы по сравнению с другими алгоритмами.

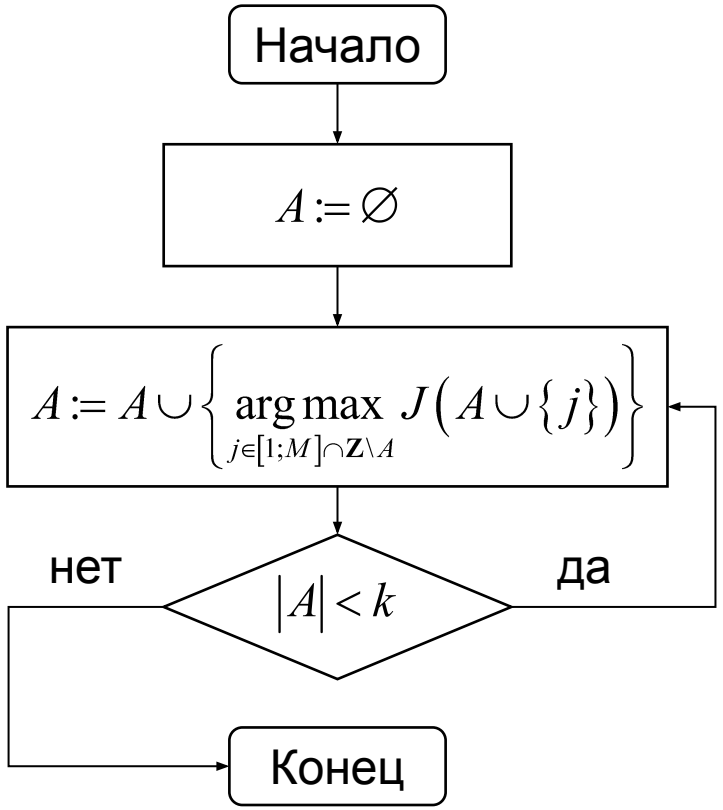
Недостаток: никак не учитывает качество подмножеств из нескольких признаков.



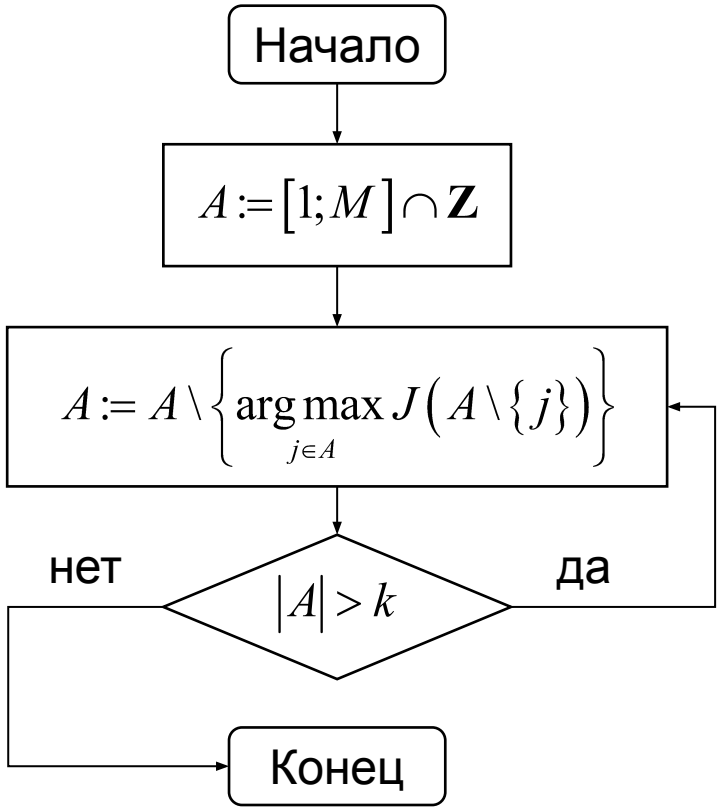


Входные данные: количество признаков k, обучающая выборка U, показатель качества J
Выходные данные: подмножество признаков A

Жадный прямой отбор признаков



Жадное обратное исключение признаков

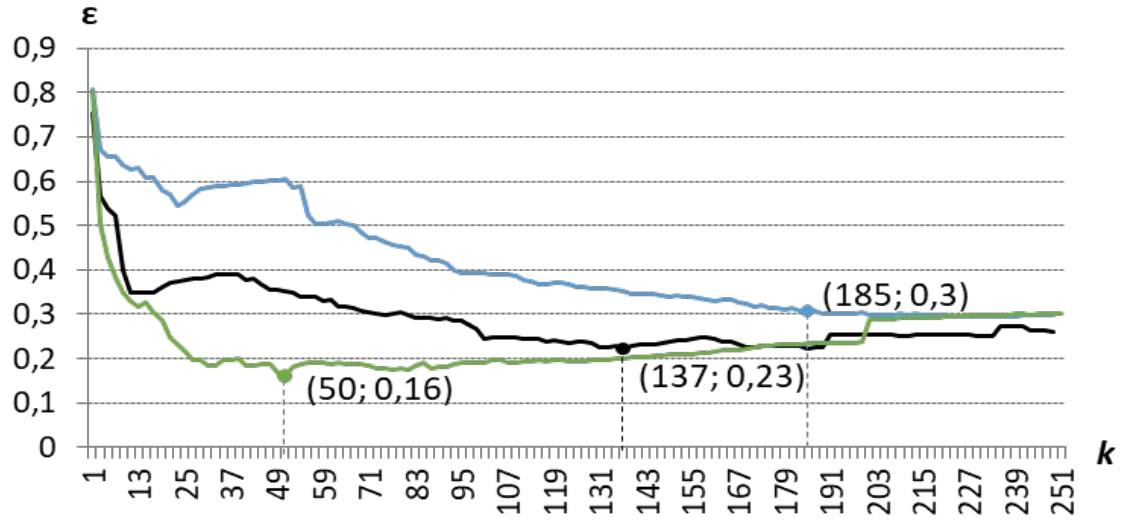


Модификации жадных алгоритмов могут быть основаны на их комбинировании.

$$A := A \cup \left\{ \arg\max_{j \in [1;M] \cap \mathbf{Z} \setminus A} J\left(A \cup \{j\}\right) \right\}$$ – шаг вперёд

$$A := A \setminus \left\{ \arg\max_{j \in A} J\left(A \setminus \{j\}\right) \right\}$$ – шаг назад

Результаты работы алгоритма по схеме «два шага вперёд – один назад» для распознавания аэрофотоснимков:

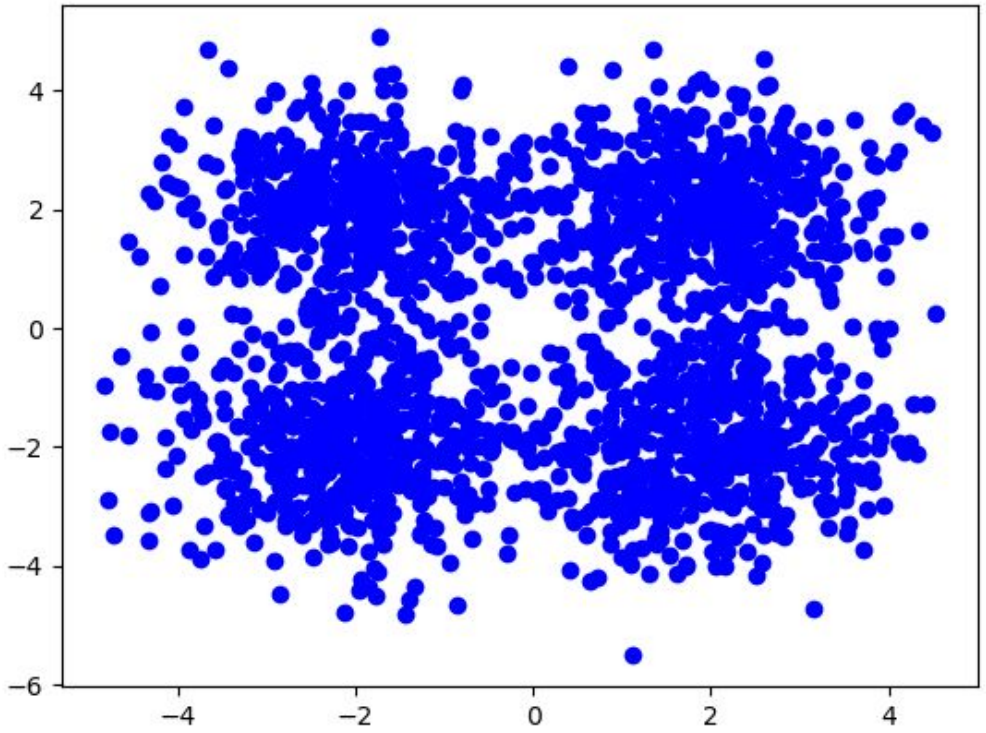


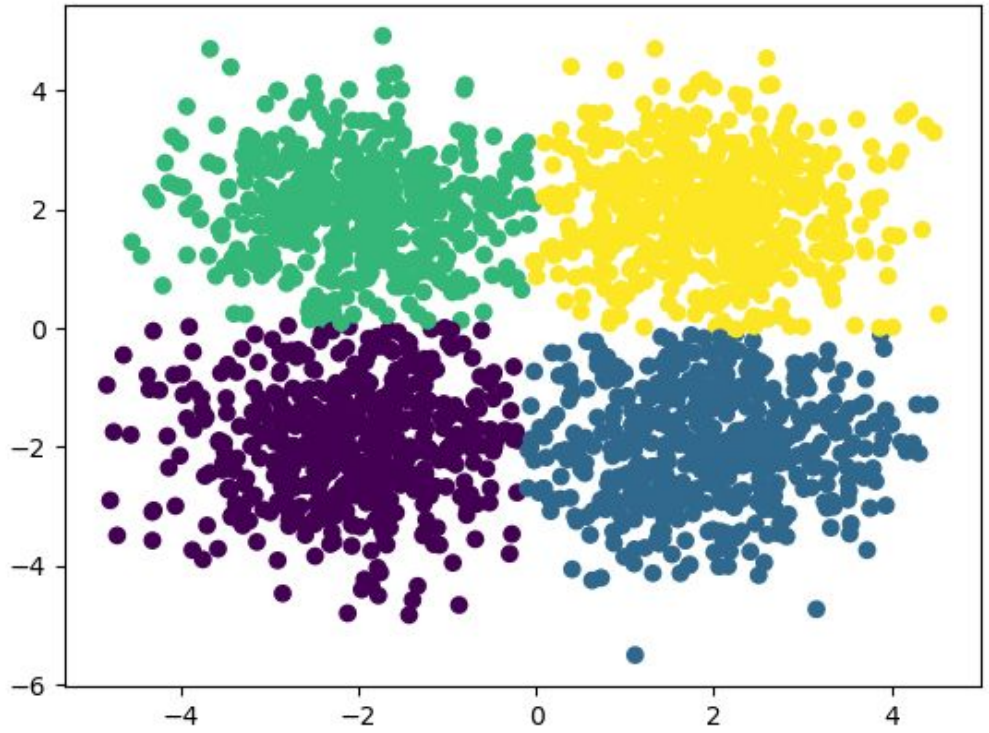
[Goncharova, E.F. Greedy algorithms of feature selection for multiclass image classification / E.F. Goncharova, A.V. Gaidel // CEUR Workshop Proceedings. – 2018. – Vol. 2210. – P. 38-46]

k – количество признаков в наборе.

ε – оценка вероятности ошибочной классификации

Кластеризация (*clustering*) – это процедура разбиения множества векторов признаков в метрическом признаковом пространстве на непересекающиеся подмножества (кластеры), так чтобы расстояния между элементами одного и того же кластера были как можно меньше, а расстояния между элементами разных кластеров были как можно больше.





Входные данные: количество признаков k, обучающая выборка U, расстояние между признаками ρ
Выходные данные: подмножество признаков A
Идея алгоритма:

- Разбить множество признаков на k кластеров и выбрать по одному признаку из каждого кластера. Это обеспечит уникальность, важность, непохожесть на другие признаки для каждого признака из результирующего набора.
- В качестве алгоритма кластеризации использовать, например, алгоритм k внутригрупповых средних (*k-means*)
- В качестве расстояния между признаками использовать некоторую меру зависимости, например, модуль коэффициента корреляции Пирсона, вычтенный из единицы:

$$\rho(i,j) = 1 - \left| \frac{R(i,j)}{\sqrt{R(i,i)R(j,j)}} \right|$$

- Выбирать в итоговый набор признаков те признаки, которые ближе всего к центрам своих кластеров.

Вычислительная сложность формально:

$$O\left(NMk^M\right)$$

[Chormunge, S. Correlation based feature selection with clustering for high dimensional data / S. Chormunge, S. Jena // Journal of Electrical Systems and Information Technology. – 2018. – Vol. 5(3). – P. 542-549]

Генетический алгоритм (*genetic algorithm*) – это эвристический алгоритм оптимизации, основанный на использовании таких механизмов естественной эволюции, как естественный отбор, скрещивание и мутация.

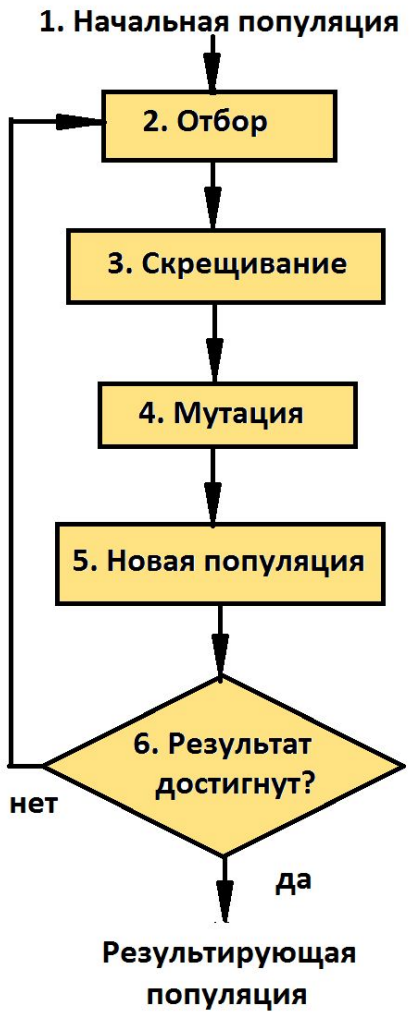
Особь – это одно из возможных решений задачи.

Популяция – это текущее множество особей.

Отбор – это процедура исключения из популяции особей с наименьшими значениями целевой функции.

Скрещивание – это процедура формирования новых особей на основании пары уже существующих в популяции особей.

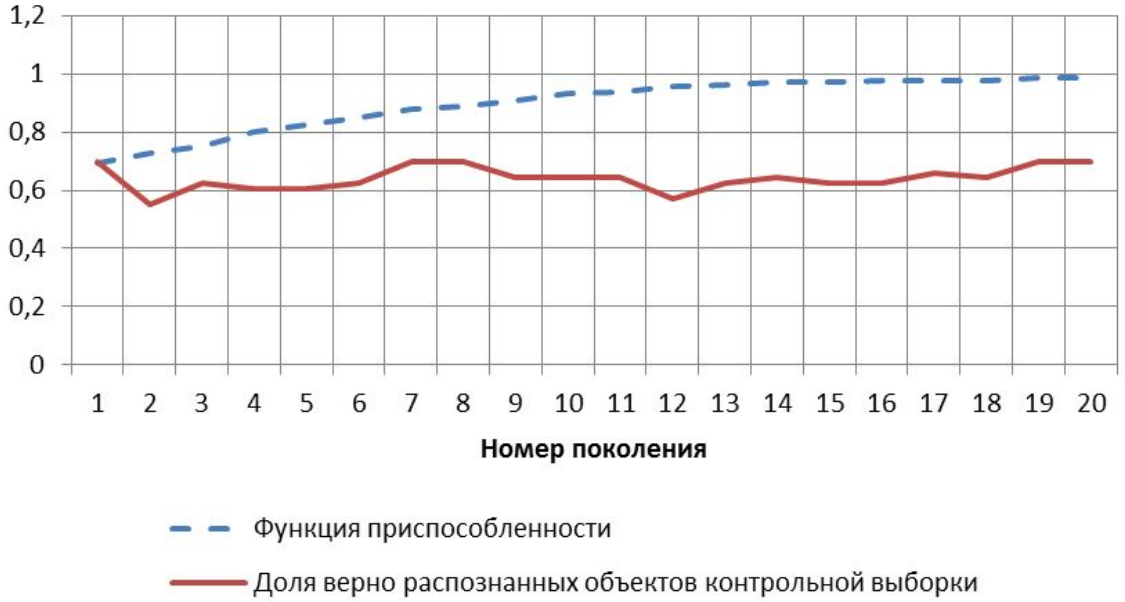
Мутация – это процедура формирования новых особей путём внесения случайных изменений в уже существующие в популяции особи.



ГЕНЕТИЧЕСКИЙ АЛГОРИТМ ОТБОРА ПРИЗНАКОВ

Входные данные: количество признаков k, обучающая выборка U, показатель качества J
Выходные данные: подмножество признаков A
Идея генетического алгоритма:

- Особь – это одно из подмножеств k признаков.
- ***Скрещивание*** двух особей реализовано таким образом, чтобы в новый набор признаков вошли все признаки из объединения родительских наборов признаков и случайные признаки из оставшихся в родительских наборах, так чтобы общее количество признаков у потомка было ровно k.
- ***Мутация*** особи представляет собой замену случайного признака в наборе на случайный признак, не входящий в набор.



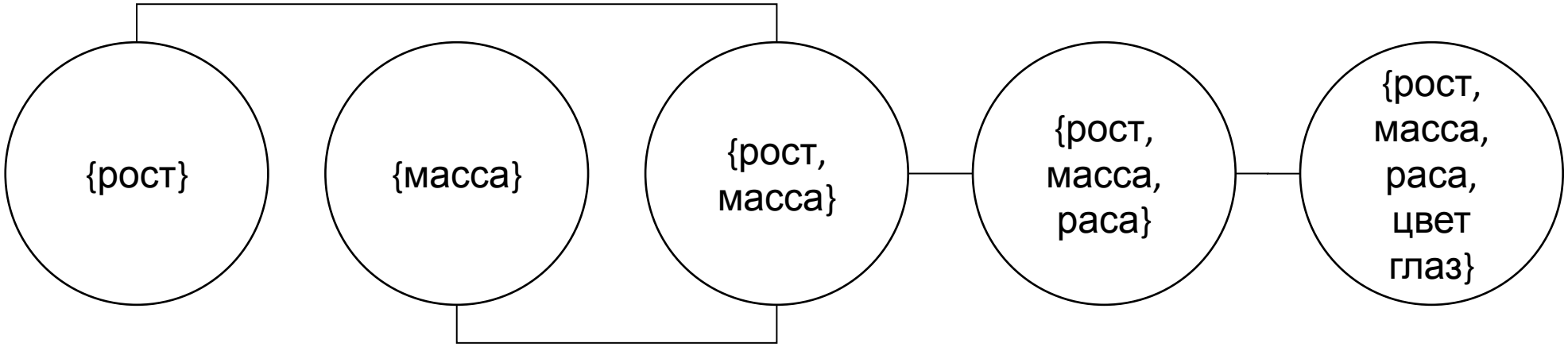
Распознавание рака лёгких по генетическим данным.
Доля верно распознанных объектов: 0,69

Рассмотрим граф:

- вершины – подмножества признаков,
- рёбра – близкие между собой подмножества признаков (например, отличающиеся добавлением или удалением одного признака),
- для каждой вершины задан показатель качества J.

Гипотеза: значения показателя качества J близкие у соседних вершин.

Задача отбора признаков заключается в поиске вершины с максимальным значением показателя качества путём обхода графа по рёбрам от вершины к вершине. Решения основаны на алгоритмах **восхождения к вершине** (*hill climbing*).



СПИСОК ЛИТЕРАТУРЫ

- Guyon, I. An Introduction to Variable and Feature Selection / I. Guyon, A. Elisseeff // The Journal of Machine Learning Research. – 2003. – Vol. 3. – P. 1157-1182.
- Fukunaga, K. Introduction to Statistical Pattern Recognition / K. Fukunaga. – Academic Press, 1990. – 592 p.
- Theodoridis, S. Pattern Recognition / S. Theodoridis, K. Koutroumbas. – Academic Press, 2008. – 984 p.
- Tou, J.T. Pattern Recognition Principles / J.T. Tou, R.C. Gonzalez. – Addison-Wesley Publishing Company, 1974. – 378 p.
- Bennasar, M. Feature selection using Joint Mutual Information Maximisation / M. Bennasar, Yu. Hicks, R. Setchi // Expert Systems with Applications. – 2015. – Vol. 42 (22). – P. 8520-8532.
- Bolón-Canedo, V. Recent advances and emerging challenges of feature selection in the context of big data / V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos // Knowledge-Based Systems. – 2015. – Vol. 86. – P. 33-45.
- Глумов, Н.И. Метод отбора информативных признаков на цифровых изображениях / Н.И. Глумов, Е.В. Мясников // Компьютерная оптика. – 2007. – Т. 31, № 3. – С. 73-76.
- Кутикова, В.В. Исследование методов отбора информативных признаков для задачи распознавания текстурных изображений с помощью масок Лавса / В.В. Кутикова, А.В. Гайдель // Компьютерная оптика. – 2015. – Т. 39, № 5. – С. 744-750.
- Гайдель, А.В. Отбор признаков для задачи диагностики остеопороза по рентгеновским изображениям шейки бедра / А.В. Гайдель, В.Р. Крашенинников // Компьютерная оптика. – 2016. – Т. 40, № 6. – С. 939-946.

ЗАКЛЮЧЕНИЕ

- Когда подбор параметров классификатора уже не помогает повысить эффективность классификации, отбор признаков в некоторых случаях всё ещё может помочь.
- Отбор признаков устраняет избыточность данных, повышает качество признакового пространства, и за счёт этого положительно влияет на эффективность решения конкретной задачи анализа данных.
- Отбор признаков до сих пор представляет собой сложную задачу, которая может быть полностью решена только полным перебором, что предоставляет большой простор для исследования различных эвристических алгоритмов отбора признаков.

САМАРСКИЙ УНИВЕРСИТЕТ
SAMARA UNIVERSITY

БЛАГОДАРЮ ЗА ВНИМАНИЕ

Доцент кафедры технической кибернетики
Самарского университета
Гайдель Андрей Викторович
andrey.gaidel@gmail.com