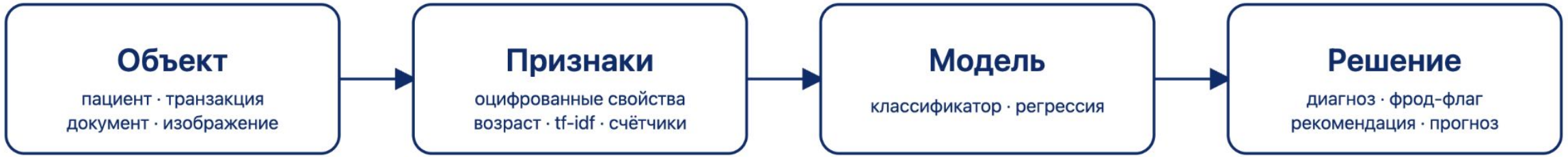
Efficient Feature Selection via Analysis of Relevance and Redundancy

Подготовили студенты:
Подолеев Вадим
Стёпкин Арсений
Брель Мария
Пастухова Эрика
Дергунов Дмитрий
Завьялов Гордей

Введение

Признаки — способ превратить объект в числовое описание для решения прикладной задачи (диагностика, фрод, рекомендации, поиск, промышленная аналитика)

Объект можно охарактеризовать набором признаков: $x = (f_1, \ldots, f_N)$



Проблема высокой размерности

«Проклятие размерности»: мало наблюдений при большом кол-ве признаков ⇒ переобучение и нестабильность.

Две главные причины вреда:

- Нерелевантность (шум)
- Избыточность (дублирование/сильная корреляция между фичами)

Цена: взрыв времени/памяти, снижение обобщающей способности, сложная интерпретация.
Примеры: BoW в NLP (десятки тысяч токенов), экспрессия генов (десятки-сотни тысяч), сетевые логи/кибербез…

Задача отбора признаков

Найти минимальное подмножество $G \subseteq F$, такое что распределение классов «не портится»:

$$P(C \mid G) \approx P(C \mid F).$$

Практические цели:

- Повышение точности и устойчивости
- Уменьшение времени обучения и памяти
- Проще интерпретация

Интуитивно: оставить минимальный, но достаточный набор информативных и неизбыточных признаков.

Классические решения 1

ReliefF – признак полезен, если для объекта он «отталкивает» ближайших соседей другого класса и «сближает» соседей своего класса

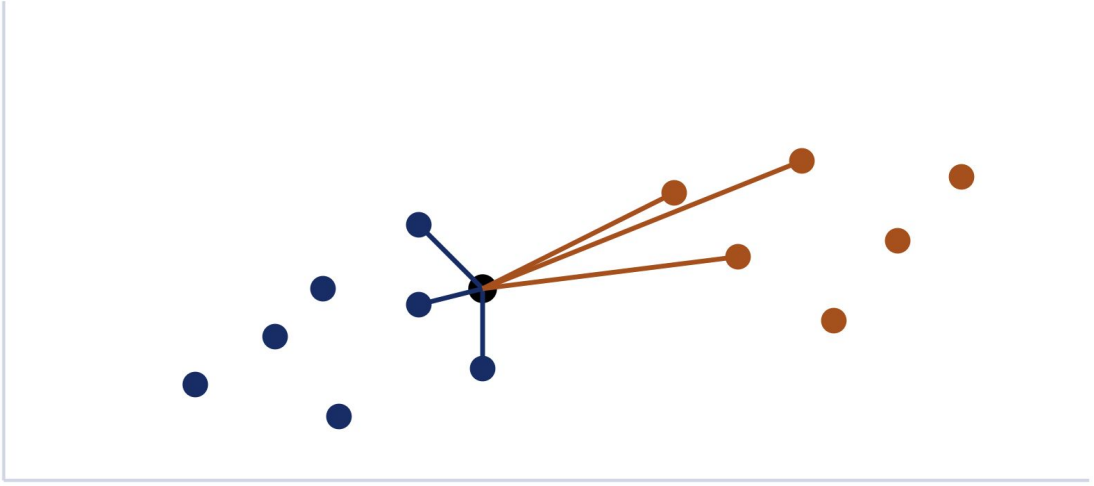
Идея: выбираем k (число соседей) и m (число опорных объектов). Для каждого признака считаем его оценку “полезности”

$$w_j \mathrel{+}= \frac{1}{m}\left(\underbrace{\tfrac{1}{k}\sum \mathrm{diff}_j(x^{(t)}, \mathrm{miss})}_{\text{хорошо, если далеко}} - \underbrace{\tfrac{1}{k}\sum \mathrm{diff}_j(x^{(t)}, \mathrm{hit})}_{\text{плохо, если далеко}}\right)$$

, где x - элемент из выборки размера m для всех объектов. Далее выкидываем самые бесполезные

Плюсы: очень быстро из-за выборочной оценки, линейная сложность

Минусы: не устраняет избыточность – близнецы по смыслу тоже попадут в топ.



Классические решения 2

CFS-SF (Correlation-based Feature Selection с жадным SFS) – выбрать набор признаков, где каждый хорошо связан с классом, а между собой признаки мало коррелируют → минимум избыточности

Идея: жадно добавляем признаки в “идеальное” подмножество для максимизации, пока это дает какой-то результат

$$\mathrm{Merit}(S) = \frac{k\,\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}},$$

где $k = |S|$, $\overline{r_{cf}}$ — средняя корреляция фич с классом, $\overline{r_{ff}}$ — средняя межфичевая корреляция.

Плюсы: явно борется с избыточностью

Минусы: риск локального оптимума (из-за жадного алгоритма), квадратичная сложность

P.S. SFS = Sequential Forward Selection

Классические решения 3

FOCUS-SF (минимально разделяющее подмножество + SFS) – перебор всех подмножеств признаков с той же консистентностью, но вместо полного перебора - жадное добавление новых признаков для максимизации совпадений классов.

Идея: жадно добавляем признаки в “идеальное” подмножество для максимизации кол-ва совпадений классов, пока это дает какой-то результат

Плюсы: нацелен на **минимальность** набора, достаточного для разделения

Минусы: может найти локальную оптимизацию из-за SFS, сложность O(N^2+)

Идея статьи

Разделяем **релевантность** и **избыточность**

Шаг 1: быстрым фильтром убрать нерелевантные признаки

Шаг 2: среди оставшихся убрать избыточные — те, которые «покрываются» другими

Ключевое отличие: делать это без дорогого поиска по подмножествам

Релевантность

1. **Сильно релевантные:** признак абсолютно незаменим для точного предсказания C. Его информация уникальна и не содержится ни в каком сочетании других признаков. Удаление такого признака обязательно ухудшит качество модели.

$$\mathbf{P}(C \mid F_i, S_i) \neq \mathbf{P}(C \mid S_i) .$$

2. **Слабо релевантные:** признак сам по себе не незаменим, если есть все другие признаки. НО его информация может быть полезной и необходимой, если каких-то других признаков не хватает. Он может частично дублировать другие.

$$\mathbf{P}(C \mid F_i, S_i) = \mathbf{P}(C \mid S_i), \text{ and}$$

$$\exists\, S_i' \subset S_i, \textit{ such that } \mathbf{P}(C \mid F_i, S_i') \neq \mathbf{P}(C \mid S_i') .$$

3. **Нерелевантные:** признак совершенно бесполезен для предсказания C. Его информация не влияет на результат ни при каких условиях. Такие признаки только добавляют шум.

$$\forall\, S_i' \subseteq S_i, \mathbf{P}(C \mid F_i, S_i') = \mathbf{P}(C \mid S_i') .$$

Марковское одеяло

Definition 3 (Markov blanket) *Given a feature F_i, let $M_i \subset F$ $(F_i \notin M_i)$, M_i is said to be a Markov blanket for F_i iff*

$$\mathbf{P}(F - M_i - \{F_i\}, C \mid F_i, M_i) = \mathbf{P}(F - M_i - \{F_i\}, C \mid M_i) .$$

Это такой набор признаков M_i (не включающий сам F_i), который полностью "заменяет" F_i.

Если у нас есть значение M_i, то знание самого F_i не дает никакой новой информации ни о целевом классе C, ни о любых других признаках в наборе данных.

Сильно релевантные признаки никогда не смогут иметь Марковского покрытия

Выбора признаков **Markov Blanket Filtering** (получаем оптимальный поднабор признаков, сохраняющий всю полезную информацию для предсказания C):

1. Это алгоритм обратного отбора: начинаем со всех признаков (G = F)
2. Шаг: Просматриваем текущий набор G. Если для какого-то признака F_i находится Марковское покрытие M_i среди других признаков, все еще находящихся в G, то F_i можно безопасно удалить из G
3. Безопасность: знаем, что если признак был удален на раннем шаге (потому что тогда для него нашлось марковское покрытие), то он останется ненужным и на всех последующих шагах, даже после удаления других признаков

Избыточный признак

Definition 4 (Redundant feature) *Let G be the current set of features, a feature is redundant and hence should be removed from G iff it is weakly relevant and has a Markov blanket M_i within G .*

Признак считается избыточным в текущем наборе признаков G и должен быть удалён тогда и только тогда, когда он:

1. Слабо релевантен: Он может быть полезен для предсказания, но не незаменим
2. Имеет Марковское покрытие внутри текущего набора G: Существует группа других признаков из текущего G, которая полностью заменяет его

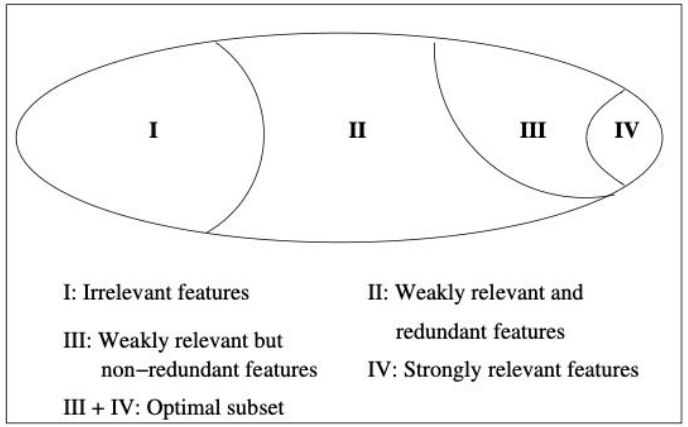


Figure 1: A view of feature relevance and redundancy.

Стандартный выбор признаков

1. **Генерация подмножества:** предлагает набор признаков S.
2. **Оценка подмножества:** насколько S хорош? (например, точность классификатора на S).
3. **Сравнение и обновление:** Если S лучше текущего лучшего подмножества, он становится новым лучшим.

Учитывает избыточность, но работает медленно – O(2^N) или O(N^2) если оптимизировать.

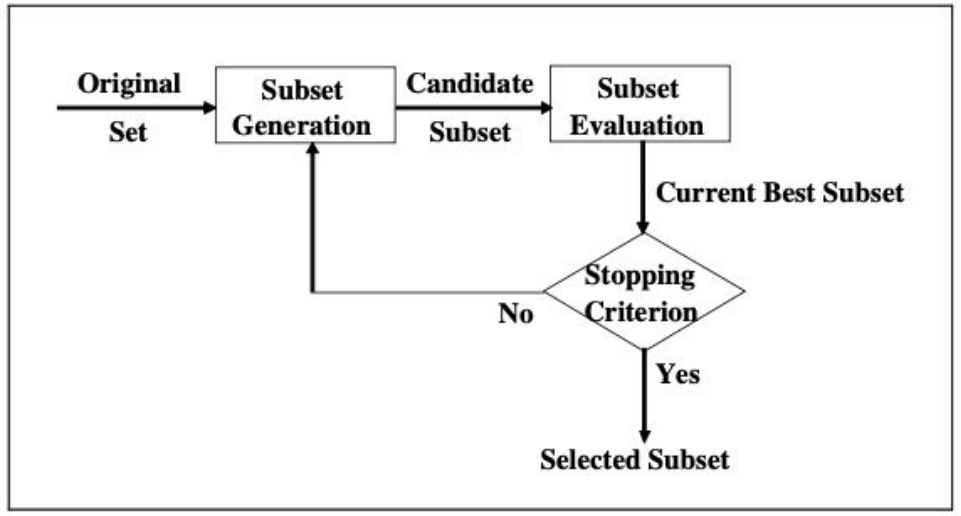


Figure 2: A traditional framework of feature selection.

Эффективный выбор признаков

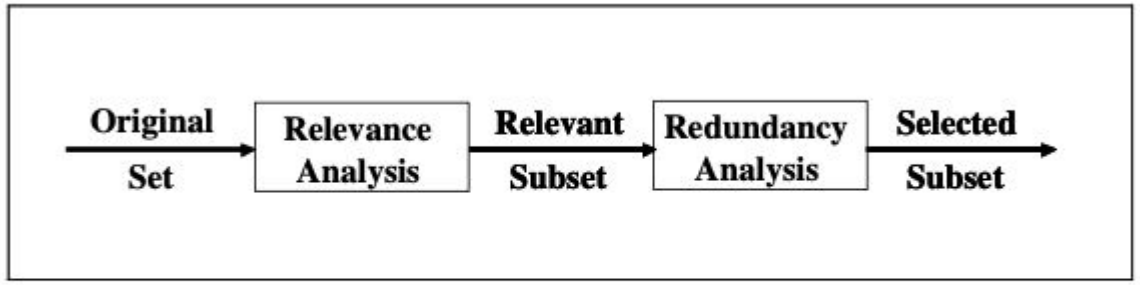


Figure 3: A new framework of feature selection.

1. Анализ релевантности => остается подмножество релевантных признако – все, что хоть как-то связано с целью.
2. Анализ Избыточности => выявлены и удалены избыточные признаки из релевантных, оставив только неизбыточные слабо релевантные и сильно релевантные.

По сравнению с традиционным:

1. Быстрее
2. Эффективнее
3. Менее склонен к переобучению

Энтропия

$$\rho = \frac{\sum_i (x_i - \overline{x_i})(y_i - \overline{y_i})}{\sqrt{\sum_i (x_i - \overline{x_i})^2} \sqrt{\sum_i (y_i - \overline{y_i})^2}} ,$$

Коэфф. линейной корреляции

The entropy of a variable X is defined as

$$H(X) = -\sum_i P(x_i) \log_2(P(x_i)) ,$$

and the entropy of X after observing values of another variable Y is defined as

$$H(X|Y) = -\sum_j P(y_j) \sum_i P(x_i \mid y_j) \log_2(P(x_i \mid y_j)) ,$$

Энтропия: Мера "неопределенности" или "непредсказуемости" случайной величины X. Чем выше энтропия, тем хаотичнее X.

Условная энтропия: Неопределенность X, после того как мы узнали значение Y. Если Y помогает предсказать X, то H(X|Y) будет меньше H(X).

Симметричная неопределенность

$$IG(X \mid Y) = H(X) - H(X \mid Y) \ .$$

Информационный выигрыш

$$SU(X, Y) = 2 \left[\frac{IG(X \mid Y)}{H(X) + H(Y)} \right] \ .$$

Симметричная неопределенность: устраняет смещение в пользу признаков с высокой энтропией (часто из-за многих уникальных значений).

C- и F- корреляция

C-Correlation (SU_i,c): Сила связи признака Fi с целевой переменной C (релевантность). Вычисляется как SU(Fi, C).

F-Correlation (SU_i,j): Сила связи признака Fi с другим признаком Fj (потенциальная избыточность). Вычисляется как SU(Fi, Fj).

Приближенное марковское одеяло

Definition 7 (Approximate Markov blanket) *For two relevant features* F_i *and* F_j $(i \neq j)$, F_j *forms an approximate Markov blanket for* F_i *iff* $SU_{j,c} \geq SU_{i,c}$ *and* $SU_{i,j} \geq SU_{i,c}$.

Для двух релевантных признаков Fi и Fj (Fi ≠ Fj), Fj образует приближенное Марковское покрытие для Fi тогда и только тогда:

1. SU_j,c ≥ SU_i,c (релевантность Fj к цели C не меньше, чем релевантность Fi к C).
2. SU_i,j ≥ SU_i,c (связь между Fi и Fj не слабее, чем связь Fi с целью C).

FCBF алгоритм


```
input:  S(F_1, F_2, ..., F_N, C)   // a training data set
        δ                          // a predefined threshold
output: S_best                     // a selected subset

1  begin
2    for i = 1 to N do begin
3      calculate SU_{i,c} for F_i;
4      if (SU_{i,c} > δ)
5        append F_i to S'_list;
6    end;
7    order S'_list in descending SU_{i,c} value;
8    F_j = getFirstElement(S'_list);
9    do begin
10     F_i = getNextElement(S'_list, F_j);
11     if (F_i <> NULL)
12       do begin
13         if (SU_{i,j} ≥ SU_{i,c})
14           remove F_i from S'_list;
15         F_i = getNextElement(S'_list, F_i);
16       end until (F_i == NULL);
17     F_j = getNextElement(S'_list, F_j);
18   end until (F_j == NULL);
19 S_best = S'_list;
20 end;
```

Figure 4: FCBF Algorithm.

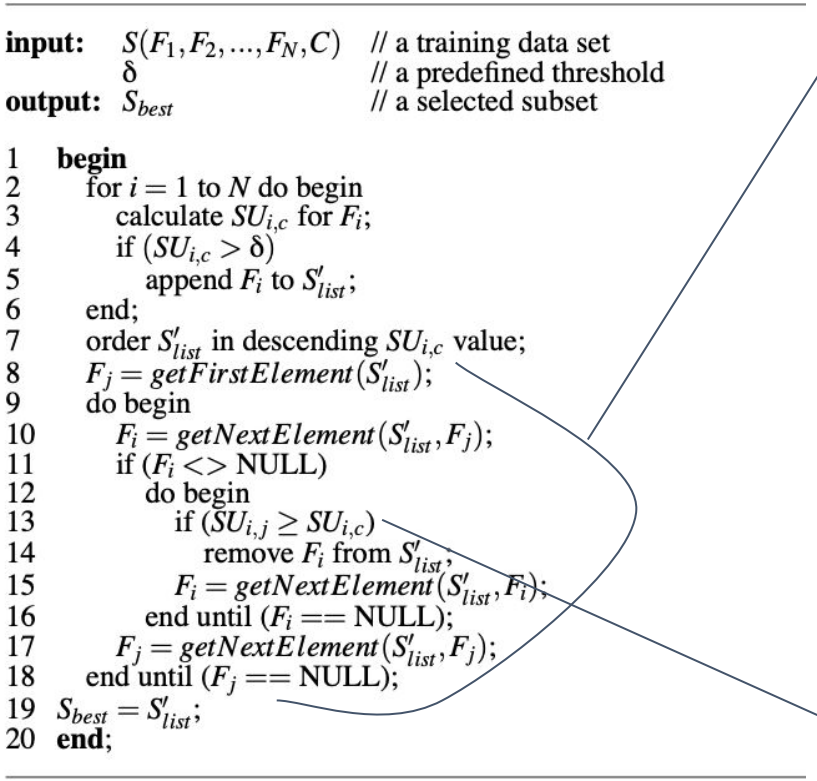
Вход: Обучающая выборка S с признаками F_1, F_2, ..., F_N и целевой переменной C. Порог релевантности δ (задаётся пользователем).

Выход: Оптимальное подмножество признаков S_best.

Отбор релевантных признаков: Для каждого F_i:

1. Вычислить симметричную неопределенность с целевой переменной C
2. Отфильтровать: Добавить F_i в список S'_list только если SU_i, c > δ
3. Это удаляет нерелевантные признаки
4. Отсортировать S'_list по убыванию SU_i,c
5. Признаки с наибольшей релевантностью — в начале списка

FCBF алгоритм



```
input:  S(F1, F2, ..., FN, C)   // a training data set
        δ                       // a predefined threshold
output: Sbest                   // a selected subset

1   begin
2     for i = 1 to N do begin
3       calculate SUi,c for Fi;
4       if (SUi,c > δ)
5         append Fi to S'list;
6     end;
7     order S'list in descending SUi,c value;
8     Fj = getFirstElement(S'list);
9     do begin
10      Fi = getNextElement(S'list, Fj);
11      if (Fi <> NULL)
12        do begin
13          if (SUi,j ≥ SUi,c)
14            remove Fi from S'list;
15          Fi = getNextElement(S'list, Fi);
16        end until (Fi == NULL);
17      Fj = getNextElement(S'list, Fj);
18    end until (Fj == NULL);
19  Sbest = S'list;
20  end;
```

Figure 4: FCBF Algorithm.

Удаление избыточных признаков

1. Начать с первого признака в S'_list – это самый релевантный признак
2. Проверить все признаки F_i, которые идут после F_j в списке (менее релевантны). Если SU_i, j >= SU_i, c (сила связи Fi с Fj >= силы связи Fi с C), удаляем Fi из списка S'_list. F_j считается приближенным марковским покрытием для F_i, делая F_i избыточным
3. Перейти к следующему признаку F_j в оставшемся списке S'_list.
4. Повторять шаги 1-2, пока не будут обработаны все признаки в списке.
5. Результат (S_best): Оставшиеся в S'_list признаки – это неизбыточные релевантные признаки

Связь между признаком Fi и "доминирующим" признаком Fj сильнее или равна связи самого Fi с целевой переменной C. Это гарантирует, что Fj содержит всю информацию, которую Fi несёт о C, делая Fi избыточным

Пример

SU_6, 3 >= SU_6, c => удаляем F_6

SU_2,1 >= SU_2, c => удаляем F_2

список признаков отсортирован по убыванию

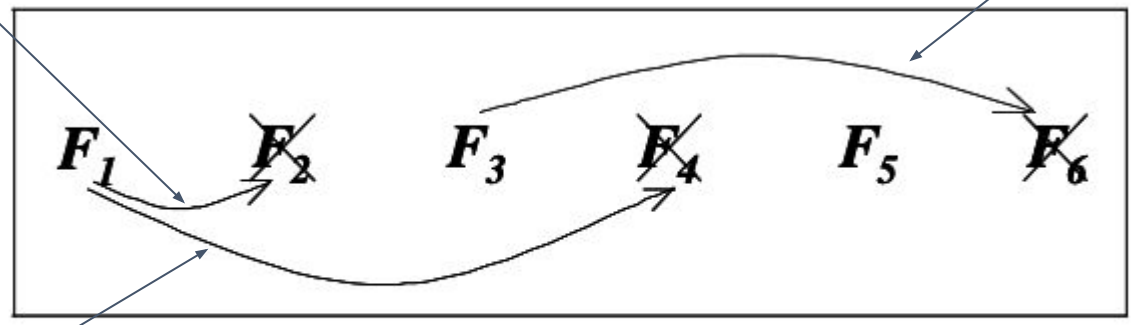


Figure 5: Selection of predominant features

SU_4,1 >= SU_4,c => удаляем F_4

Эксперименты

- Оценка эффективности и результативности метода FCBF для отбора признаков.
- Сравнение с популярными алгоритмами: ReliefF, CFS-SF, FOCUS-SF.
- Эксперименты на синтетических и реальных данных высокой размерности.

Эксперименты

- Эффективность: время работы алгоритма на разных наборах данных
- Результативность:
  - Для синтетических данных: сравнение выбранного подмножества с оптимальным
  - Для реальных данных: точность прогноза модели на выбранных признаках
- Для сравнения: анализируются только filter-алгоритмы
- Используемые алгоритмы:
  - ReliefF (индивидуальная оценка)
  - CFS-SF и FOCUS-SF (оценка подмножеств, последовательный прямой отбор)
  - Порог релевантности в FCBF подбирается эвристически и сравнивается с вариантом $\gamma = 0$

Эксперименты

Оценка точности и инструменты

- Прогностическая точность оценивается с помощью алгоритмов NBC (наивный байесовский классификатор) и C4.5 (дерево решений)
- Все алгоритмы реализованы в среде Weka (FCBF также реализован в Weka)
- Для реальных данных ключевой показатель — точность прогноза на выбранных признаках

Эксперименты

Определения синтетических наборов данных

- Corral

  6 булевых признаков: A0, A1, B0, B1, I, R

  a. Класс: Y = (A0 ∧ A1) ∨ (B0 ∧ B1)
  b. A0, A1, B0, B1 — независимые, I — случайный, R совпадает с классом в 75% случаев
  c. Оптимальный набор: A0, A1, B0, B1; I — нерелевантный, R — избыточный
- Corral-47

  a. 47 булевых признаков: 5 основных (A0, A1, B0, B1, R), 14 нерелевантных, 28 избыточных
  b. 14 нерелевантных: 2 случайных и 12 полностью коррелированных с ними
  c. Для каждого из A0, A1, B0, B1 добавлено по 7 избыточных признаков с разной степенью корреляции
- Corral-46

  a. То же, что Corral-47, но без признака R

Эксперименты

Результаты на синтетических результатах

- Corral
    a. Все алгоритмы удаляют нерелевантный признак (I), но не справляются с удалением избыточного признака (R)
    b. FCBF(log) пропускает несколько релевантных признаков при некорректном выборе порога
- Corral-47 (47 признаков, много избыточных и нерелевантных)
    a. ReliefF, CFS-SF, FOCUS-SF, FCBF(0), FCBF(log) полностью удаляют нерелевантные признаки
    b. Только FCBF(0), FCBF(log) и CFS-SF успешно удаляют все дополнительные избыточные признаки, но не признак R
    c. Порог в FCBF(log) не влияет на результат при высокой размерности и множестве дублирующих признаков
- Corral-46 (аналог Corral-47 без R)
    a. Только FCBF(0) и FCBF(log) отбирают оптимальное подмножество признаков

Эксперименты

Эталонные данные

Title	Features	Instances	Classes
Lung-cancer	56	32	3
Promoters	57	106	2
Splice	60	3190	3
USCensus90	67	9338	3
CoIL2000	85	5822	2
Chemical	150	936	3
Musk2	166	6598	2
Arrhythmia	279	452	16
Isolet	617	1560	26
Multi-features	649	2000	10

Title	$FCBF_{(log)}$	$FCBF_{(0)}$	ReliefF	CFS-SF	FOCUS-SF
Lung-cancer	4	6	5	8	4
Promoters	6	6	4	4	4
Splice	9	22	11	6	10
USCensus90	3	4	2	1	13
CoIL2000	3	5	12	10	29
Chemical	4	5	7	7	11
Musk2	2	2	2	10	11
Arrhythmia	5	12	25	25	24
Isolet	5	32	23	137	11
Multi-Features	27	130	14	87	7
Average	7	22	11	30	12

Table 4: Number of features selected by each feature selection algorithm on **UCI** data.

Title	$FCBF_{(log)}$	$FCBF_{(0)}$	ReliefF	CFS-SF	FOCUS-SF
Lung-cancer	0.001	0.02	0.09	0.05	0.08
Promoters	0.001	0.02	0.06	0.03	0.16
Splice	0.20	0.55	0.89	0.55	16.59
USCensus90	0.30	0.50	2.94	0.52	77.67
CoIL2000	0.25	0.50	4.25	1.98	143.94
Chemical	0.05	0.05	1.36	0.28	6.56
Musk2	0.53	0.88	9.55	4.84	85.78
Arrhythmia	0.06	0.08	1.19	0.78	13.70
Isolet	0.42	3.05	10.05	93.94	107.33
Multi-Features	1.19	19.42	11.42	71.00	67.56

Table 3: Running time (seconds) for each feature selection algorithm on **UCI** data.

Эксперименты

Эталонные данные

Title	$FCBF_{(log)}$	$FCBF_{(0)}$		Full Set		ReliefF		CFS-SF		FOCUS-SF	
	Acc	Acc	p-Val	Acc	p-Val	Acc	p-Val	Acc	p-Val	Acc	p-Val
Lung-cancer	83.33	86.67	0.34	78.33	0.34	84.17	0.85	86.67	0.34	87.5	0.46
Promoters	93.27	93.27	1	91.55	0.55	87.82	0.25	95.18	0.17	90.45	0.40
Splice	93.95	96.14	**0.00**$^{+}$	95.52	**0.00**$^{+}$	91.32	**0.00**$^{-}$	93.54	0.24	94.36	**0.08**$^{+}$
USCensus90	97.94	97.88	0.19	93.49	**0.00**$^{-}$	97.97	0.17	97.99	0.65	97.87	0.44
CoIL2000	93.94	93.92	0.34	78.68	**0.00**$^{-}$	93.89	0.66	92.92	**0.01**$^{-}$	83.22	**0.00**$^{-}$
Chemical	71.91	67.73	**0.02**$^{-}$	60.90	**0.00**$^{-}$	71.26	0.77	70.51	0.35	66.35	**0.00**$^{-}$
Musk2	84.59	84.59	1	84.78	0.51	84.59	1	64.87	**0.00**$^{-}$	83.53	**0.01**$^{-}$
Arrhythmia	67.48	65.73	0.45	60.88	**0.01**$^{-}$	55.79	**0.00**$^{-}$	69.05	0.45	69.06	0.56
Isolet	50.06	83.33	**0.00**$^{+}$	84.10	**0.00**$^{+}$	60.90	**0.00**$^{+}$	87.31	**0.00**$^{+}$	71.03	**0.00**$^{+}$
Multi-feat	95.9	95.65	0.50	94.1	**0.01**$^{-}$	67.65	**0.00**$^{-}$	96.15	0.64	93.7	**0.02**$^{-}$
L/W/T	-	**1/2/7**		**5/2/3**		**3/1/6**		**2/1/7**		**4/2/4**	

Table 5: Accuracy of **NBC** on selected features for **UCI** data: Acc records 10-fold cross-validation accuracy rate (%) and p-Val records the probability associated with a paired two-tailed t-Test. The symbols "+" and "−" respectively identify statistically significant (at 0.1 level) wins or losses over $FCBF_{(log)}$.

Title	$FCBF_{(log)}$	$FCBF_{(0)}$		Full Set		ReliefF		CFS-SF		FOCUS-SF	
	Acc	Acc	p-Val	Acc	p-Val	Acc	p-Val	Acc	p-Val	Acc	p-Val
Lung-cancer	86.67	86.67	1	80.83	0.17	84.17	0.34	84.17	0.34	84.17	0.34
Promoters	80.18	80.18	1	78.09	0.42	82.36	0.55	80.18	1	81.36	0.67
Splice	94.01	94.14	0.64	93.98	0.89	90.53	**0.00**$^{-}$	93.39	**0.00**$^{-}$	93.79	0.11
USCensus90	98.12	98.12	1	98.19	0.39	98.12	1	97.99	**0.00**$^{-}$	98.21	0.11
CoIL2000	94.02	94.02	1	93.87	0.12	94.02	1	94.02	1	93.97	0.39
Chemical	95.41	95.41	1	94.13	**0.01**$^{-}$	95.94	0.14	95.94	0.14	95.31	0.86
Musk2	91.35	91.35	1	96.91	**0.00**$^{+}$	88.00	**0.00**$^{-}$	95.79	**0.00**$^{+}$	95.45	**0.00**$^{+}$
Arrhythmia	71.47	68.80	0.19	67.70	**0.04**$^{-}$	69.02	**0.07**$^{-}$	68.58	0.13	67.02	**0.04**$^{-}$
Isolet	49.17	75.77	**0.00**$^{+}$	79.87	**0.00**$^{+}$	59.10	**0.00**$^{+}$	81.35	**0.00**$^{+}$	68.84	**0.00**$^{+}$
Multi-feat	92.45	93.65	**0.04**$^{+}$	94.3	**0.01**$^{+}$	78.65	**0.00**$^{-}$	94.7	**0.00**$^{+}$	91.75	0.42
L/W/T	-	**0/2/8**		**2/3/5**		**4/1/5**		**2/3/5**		**1/2/7**	

Table 6: Accuracy of **C4.5** on selected features for **UCI** data: Acc records 10-fold cross-validation accuracy rate (%) and p-Val records the probability associated with a paired two-tailed t-Test. The symbols "+" and "−" respectively identify statistically significant (at 0.1 level) wins or losses over $FCBF_{(log)}$.

Эксперименты

NIPS Benchmark

- 3 набора с очень высокой размерностью и небольшим числом объектов
- Все наборы включают реальные и случайные признаки
- Ключевые выводы:
  - CFS-SF не завершает работу на некоторых наборах из-за нехватки памяти
  - FCBF работает в разы быстрее: напр., 1 минута vs 4.5 часа для FOCUS-SF
  - Все алгоритмы радикально уменьшают размерность
  - FCBF обеспечивает сравнимую или лучшую точность; максимум для NBC — на признаках FCBF

	$FCBF_{(log)}$	$FCBF_{(0)}$		Full Set		ReliefF		CFS-SF		FOCUS-SF	
Title	Acc	Acc	*p*-Val	Acc	*p*-Val	Acc	*p*-Val	Acc	*p*-Val	Acc	*p*-Val
Arcene	91.0	93.0	0.34	69.0	**0.00**$^{-}$	69.0	**0.02**$^{-}$	92.0	0.68	59.0	**0.00**$^{-}$
Dexter	90.0	90.0	1	88.0	0.26	73.00	**0.00**$^{-}$	N/A	N/A	90.0	1
Dorothea	97.5	98.38	**0.01**$^{+}$	90.25	**0.00**$^{-}$	94.38	**0.00**$^{-}$	N/A	N/A	95.25	**0.00**$^{-}$

Table 10: Accuracy of **NBC** on selected features for **NIPS** data: Acc records 10-fold cross-validation accuracy rate (%) and *p*-Val records the probability associated with a paired two-tailed t-Test. The symbols "+" and "−" respectively identify statistically significant (at 0.1 level) wins or losses over $FCBF_{(log)}$.

Эксперименты

Итоги эмпирического исследования

- FCBF позволяет:
    - Быстро и сильно сокращать размерность
    - Поддерживать или повышать точность модели
- Подходит для задач высокой размерности

- Итоговое подмножество признаков не зависит от алгоритма обучения

Проверка

Переход в Colab -> ссылка


ВЫСШАЯ • ШКОЛА • ЭКОНОМИКИ •