О вычислительной сложности задач отбора объектов и признаков при построении монотонных классификаторов

Зухба А. В.

Светлогорск • 19–25 сентября 2015

Основные определения, обозначения, понятия

Дано множество **объектов** $\mathbb{X} = \{x_1, \ldots, x_\ell\}$, называемых **обучающей выборкой**, и $\mathbb{Y} = \{0, 1\}$ — множество **классов**.

Объекты описываются **признаками** $\mathbb{F} = \{f_1, \ldots, f_t\}$.

Каждый признак задает отображение $f_j : \mathbb{X} \to E_j$,
где E_j — линейно упорядоченное множество.

Любое непустое подмножество множества признаков $F \subseteq \mathbb{F}$ индуцирует отношение **частичного порядка** на $\mathbb{X}$:
$x \leqslant x'$ тогда и только тогда, когда $f(x) \leqslant f(x')$ для всех $f \in F$.

Основные определения, обозначения, понятия

Множество объектов $\mathbb{X}$ разбито на подмножества: $\mathbb{X} = \mathbb{A} \cup \mathbb{B}$
$\mathbb{A}$ — объекты класса **1** и $\mathbb{B}$ — объекты класса **0**.

Пара объектов $(a, b) \in \mathbb{A} \times \mathbb{B}$ называется **монотонной**, если $a > b$.
Множество всех монотонных пар обозначается через M.

Пара объектов $(a, b) \in \mathbb{A} \times \mathbb{B}$ называется **дефектной**, если $a < b$.
Множество всех дефектных пар обозначается через D.

Множество пар, монотонных по признаку f, будем обозначать M_f,
монотонных по совокупности признаков F — через M_F.

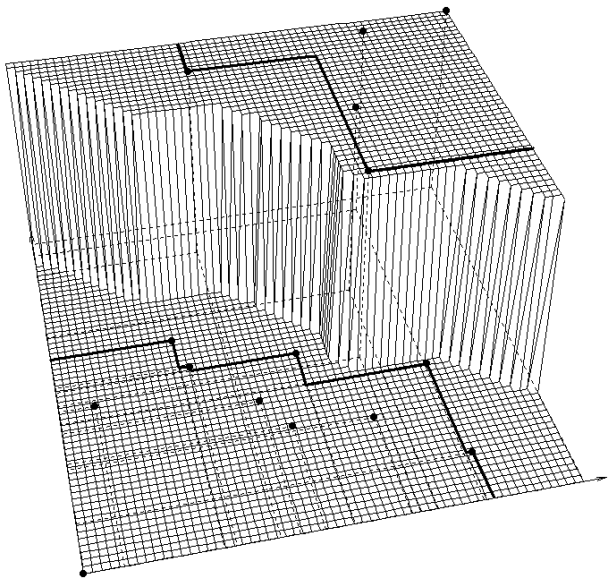
Множество пар, дефектных по признаку f, будем обозначать D_f,
дефектных по совокупности признаков F — через D_F.

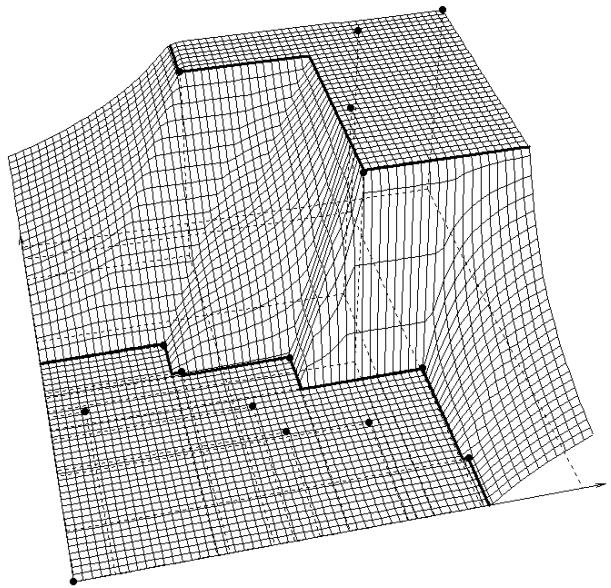
Задача построения монотонного классификатора

Функция $y\colon \mathbb{X} \to \mathbb{Y}$ монотонна, если для любых двух объектов $x, x' \in \mathbb{X}$ из $x < x'$ следует $y(x) \leqslant y(x')$.

Задача

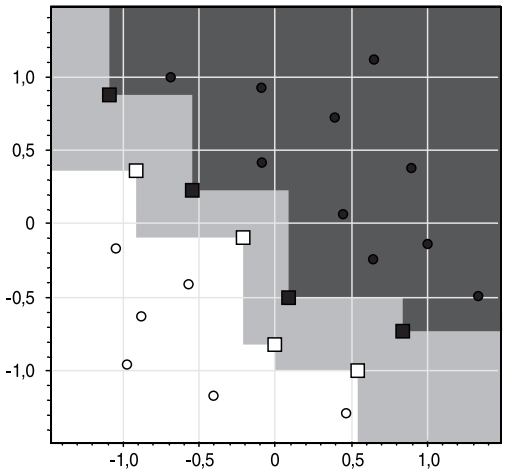
Приблизить неизвестную функцию $y^*\colon \mathbb{X} \to \mathbb{Y}$, заданную на объектах $\mathbb{X}$, монотонной функцией y.

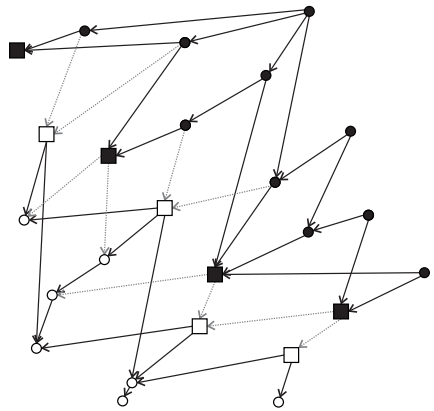




Задача построения монотонного классификатора

Построение монотонного классификатора по монотонной выборке:





Представим выборку в виде ориентированого графа частичного порядка с раскрашенными в два цвета вершинами. Цвета соответствуют классам, ребра направлены от большего объекта к меньшему.

Утверждение

Построение множества эталонных объектов можно совершить за $O(|X| + |E|)$, *где* E — *количество ребер.*

Монотонные алгоритмы классификации

- линейные модели с неотрицательными коэффициентами
- монотонный метод ближайшего соседа[1]
- монотонные решающие деревья[2]
- монотонные нейросети[3]

---

1. Воронцов К. В., Махина Г. А. Принцип максимизации зазора для монотонного классификатора ближайшего соседа. В: *15-я всероссийская конференция «Математические методы распознавания образов».* М.:МАКС Пресс, 2011. С.

2. Kamp R., Feelders A., Barile N. Isotonic classification trees. In: *Proceedings of the 8th International Symposium on Intelligent Data Analysis: Advances in Intelligent Data Analysis VIII.* Berlin, Heidelberg: Springer-Verlag, 2009. P. 405–416.

3. Sill J.Monotonic networks. In: *Advances in Neural Information Processing Systems.* Ed. Jordan M. I., Kearns M. J., Solla S. A. Cambridge: MIT Press, 1998. P. 661–667.

Задача монотонизации выборки

Замечание

Среди объектов выборки $\mathbb{X}$ могут присутствовать дефектные пары: $x < x'$, такие, что $y^*(x) > y^*(x')$, то есть $x \in \mathbb{A}$, $x' \in \mathbb{B}$.

Задача

Отобрать подмножества объектов $X \subseteq \mathbb{X}$ и признаков $F \subseteq \mathbb{F}$ так, чтобы монотонных пар M_F было как можно больше, а дефектных пар D_F — как можно меньше.

Функционалы качества

- **Степень монотонности**[4] (degree of monotonicity)

$$DgrMon = \frac{|M|}{|M| + |D|}$$

- **Эмпирический риск**[5]

$$\sum_{x \in \mathbb{A}} [y(x) = 0] + \sum_{x \in \mathbb{B}} [y(x) = 1]$$

То есть для минимизации эмпирического риска необходимо минимизировать количество объектов, на которых происходит ошибка. Переформулируем в виде задачи **отбора объектов**:

$$|D| = 0, |X| \to \max$$

считая, что алгоритм отбрасывает объекты, на которых ошибается.

4. Marina Velikova, Hennie Daniels. On Testing Monotonicity of Datasets. In: *Learning monotone models*. 2009. P. 11–22.
5. Гуз И. С. Минимизация эмпирического риска при построении монотонных композиций классификаторов. *Труды МФТИ.* 2011. Т. 3, № 3(11). С.115–121.

Параметры оптимизации

Утверждение

Для произвольного подмножества признаков $F \subseteq \mathbb{F}$

$$M_F = \bigcap_{f \in F} M_f, \quad D_F = \bigcap_{f \in F} D_f.$$

Следствие

Для любых подмножеств признаков F, $G \subseteq \mathbb{F}$

$$F \subseteq G \quad \Rightarrow \quad |M_G| \leqslant |M_F|, \quad |D_G| \leqslant |D_F|.$$

$|D_F| \to \min$ можно заменить $|F| \to \max$ или $|X| \to \min$,
$|M_F| \to \max$ можно заменить $|F| \to \min$ или $|X| \to \max$.

Систематизация задач монотонизации

Отбор признаков:

$|M| \to \max$, $|D| \to \min$,
$|F| \to \max$ (максимум информации),
$|F| \to \min$ (самая простая модель).

- $FS\big(|M| \geqslant m,\ |D| \leqslant d\big)$
- $FS\big(|M| \geqslant m,\ |F| \leqslant q\big)$
- $FS\big(|M| \geqslant m,\ |F| \geqslant q\big)$
- $FS\big(|F| \leqslant q,\ |D| \leqslant d\big)$
- $FS\big(|F| \geqslant q,\ |D| \leqslant d\big)$

Отбор объектов:

$|M| \to \max$, $|D| \to \min$,
$|X| \to \max$ (фильтрация выбросов),
$|X| \to \min$ (отбор эталонов).

- $PS\big(|M| \geqslant m,\ |D| \leqslant d\big)$
- $PS\big(|M| \geqslant m,\ |X| \leqslant n\big)$
- $PS\big(|M| \geqslant m,\ |X| \geqslant n\big)$
- $PS\big(|X| \leqslant n,\ |D| \leqslant d\big)$
- $PS\big(|X| \geqslant n,\ |D| \leqslant d\big)$

Всего 10 задач.

Систематизация задач монотонизации

Отбор признаков:

$|M| \to \max$, $|D| \to \min$,
$|F| \to \max$ (максимум информации),
$|F| \to \min$ (самая простая модель).

- $FS\big(|M| \geqslant m,\ |D| \leqslant d\big)$
- ~~$FS\big(|M| \geqslant m,\ |F| \leqslant q\big)$~~
- $FS\big(|M| \geqslant m,\ |F| \geqslant q\big)$
- $FS\big(|F| \leqslant q,\ |D| \leqslant d\big)$
- ~~$FS\big(|F| \geqslant q,\ |D| \leqslant d\big)$~~

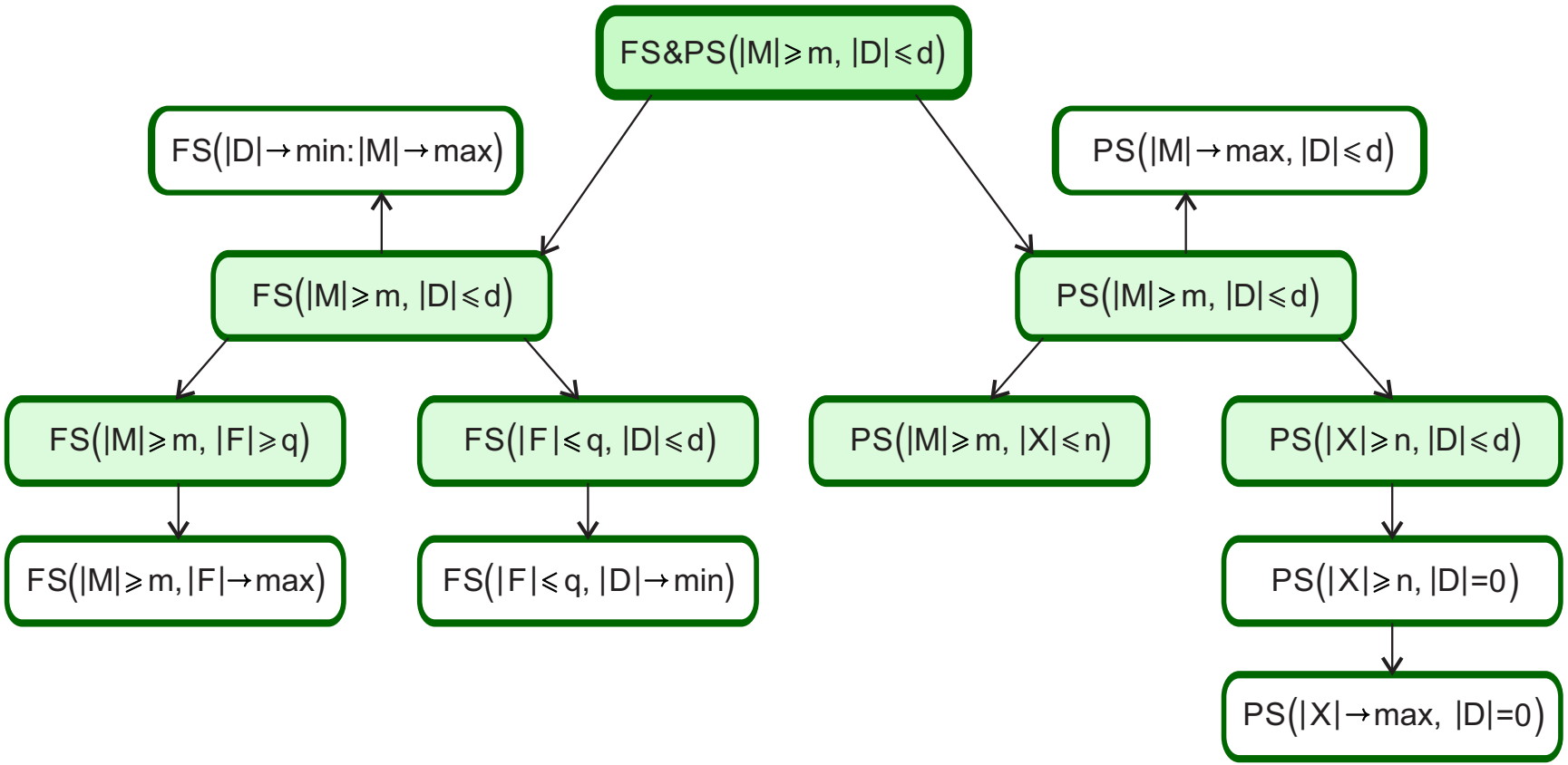
Отбор объектов:

$|M| \to \max$, $|D| \to \min$,
$|X| \to \max$ (фильтрация выбросов),
$|X| \to \min$ (отбор эталонов).

- $PS\big(|M| \geqslant m,\ |D| \leqslant d\big)$
- $PS\big(|M| \geqslant m,\ |X| \leqslant n\big)$
- ~~$PS\big(|M| \geqslant m,\ |X| \geqslant n\big)$~~
- ~~$PS\big(|X| \leqslant n,\ |D| \leqslant d\big)$~~
- $PS\big(|X| \geqslant n,\ |D| \leqslant d\big)$

Всего 10 задач. Из них 4 тривиальных и 6 содержательных.

Систематизация задач отбора объектов и признаков



Вычислительная сложность

Утверждение

Все предложенные постановки задач монотонизации кроме $PS\big(|X|\geqslant n,\ |D|\leqslant d\big)$, $PS\big(|X|\geqslant n,\ |D|=0\big)$, $PS\big(|X|\to\max,\ |D|=0\big)$ *являются NP-трудными.*

Для доказательства NP-трудности была построена полиномиальная сводимость к задачам монотонизаци таких известных NP-полных задач, как:

- *задача о рюкзаке,*
- *задача о биклике,*
- *задача о минимальном покрытии множества подмножествами.*

Вычислительная сложность

Утверждение

Решение задачи в постановках $PS\big(|X|\to\max,\ |D|{=}0\big)$ *и* $PS\big(|X|\geqslant n,\ |D|{=}0\big)$ *сводится к решению задачи поиска минимального вершинного покрытия для двудольного графа.*

Построен алгоритм, решающий задачу в постановках $PS\big(|X|\to\max,\ |D|{=}0\big)$ *и* $PS\big(|X|\geqslant n,\ |D|{=}0\big)$ *за время* $O(|D_0||X_0|^{0,5})$, *где* $|X_0|$ *— количество объектов, задействованных в дефектных парах, а* $|D_0|$ *— количество дефектных пар.*

Для постановки $PS(|D|\leqslant d, |X|\geqslant n)$ *построен алгоритм, решающий задачу за время* $O(|D_0||X_0|^{2|D_0|+0,5})$.

Основные результаты и выводы

- Рассмотрены различные функционалы качества монотонных классификаторов.
- Рассмотрены различные критерии качества монотонизации выборки.
- Предложена систематизация постановок задач монотонизации выборки как задач дискретной оптимизации.
- Сделаны оценки вычислительной сложности предложеных постановок задачи монотонизации выборки.
- Предложен алгоритм построения монотонного классификатора по монотонной выборке методом отбора эталонных объектов.

Литература


Воронцов К. В., Махина Г. А. Принцип максимизации зазора для монотонного классификатора ближайшего соседа. В: *15-я всероссийская конференция «Математические методы распознавания образов».* М.:МАКС Пресс, 2011. С.


Kamp R., Feelders A., Barile N. Isotonic classification trees. In: *Proceedings of the 8th International Symposium on Intelligent Data Analysis: Advances in Intelligent Data Analysis VIII.* Berlin, Heidelberg: Springer-Verlag, 2009. P. 405–416.


Sill J.Monotonic networks. In: *Advances in Neural Information Processing Systems.* Ed. Jordan M. I., Kearns M. J., Solla S. A. Cambridge: MIT Press, 1998. P. 661–667.


Marina Velikova, Hennie Daniels. On Testing Monotonicity of Datasets. In: *Learning monotone models*. 2009. P. 11–22.


Гуз И. С. Минимизация эмпирического риска при построении монотонных композиций классификаторов. *Труды МФТИ.* 2011. Т. 3, № 3(11). С.115–121.


Загоруйко Н. Г. *Прикладные методы анализа данных и знаний.* Новосибирск: ИМ СО РАН, 1999.


Кормен Т., Лейзерсон Ч., Ривест Р., Штайн К. *Алгоритмы: построение и анализ, 2-е издание.: Пер. с англ.* М.: Издательский дом «Вильямс», 2005.


Воронцов К. В. О проблемно-ориентированной оптимизации базисов задач распознавания. *ЖВМ и МФ.* 1998. Т. 38. № 5. С. 870–880.