Порядковая классификация с использованием частично упорядоченных наборов признаков

Папай Иван Дмитриевич

Московский физико-технический институт

Кафедра интеллектуального анализа данных ФПМИ МФТИ

Научный руководитель: д-р физ.-мат. наук В. В. Стрижов

2024

Цель исследования

Цель

Главным объектом исследования является задача ранговой классификации. Каждому объекту соответствует некоторый ранг из фиксированного и конечного диапазона. Этот диапазон тот же самый, которому принадлежат признаки этих объектов. Целью является разработка метода, корректно решающего эту задачу.

Задача

По имеющемуся набору объектов, каждый из которых задаётся множеством частично упорядоченных признаков, воссоздать ранг каждого из них. При чём сделать это нужно таким образом, чтобы избежать противоречий в аксиоматике частично упорядоченных множеств.

Постановка задачи

Определим $\mathfrak{D}$ как выборку состоящую из пар

$$\mathfrak{D} = \{(X_i, y_i)\}_{i=1}^{m},$$

где $X_i = [x_{i1}, ..., x_{ij}, ..., x_{in}]$ это объект, который требуется классифицировать, y_i это метка класса. Объект X_i является n-мерным вектором, каждый компонент которого принадлежит ч.у.м.-у X_j.

Частично упорядоченные множества

Другими словами, элемент j вектора X_i принадлежит множеству X_j, в свою очередь подчинённое некоторому заданному частичному порядку $\succeq$ со следующими свойствами:

- рефлексивность, $\forall a \in X \ (a \succeq a)$,
- антисимметричность, $\forall a, b \in X$, $(a \succeq b) \wedge (b \succeq a) \Rightarrow (a = b)$,
- транзитивность, $\forall a, b, c \in X \ (a \succeq b) \wedge (b \succeq c) \Rightarrow (a \succeq c)$.

Таким образом под объектом X имеется в виду декартово произведение ч.у.м.-ов $X_1, ..., X_n$,

$$X = X_1 \times X_2 \times \cdots \times X_n,$$

множества $X_1, ..., X_n$ являются множествами значений признаков.

Пример анализируемого датасета

Приведём пример рабочей выборки: Football Player List. В данном конкретном примере под объектами имеются в виду футболисты, популярность каждого из которых требуется оценить по шкале от 1-го до 5-го порядка(неизвестный и невероятно популярный соответственно).

Матрицы частичных порядков

Каждый частичный порядок $\succeq$, определённый на мн-ве X_j, описывается бинарной функцией $z_j(_{i,k})$

$$z_j(_{i,k}) = \begin{cases} 1, & \text{при} \quad x_{ij} \succeq x_{kj}, \\ 0, & \text{при} \quad x_{ij} \not\succeq x_{kj}, \end{cases}$$

Тогда определим матрицу частичного порядка $_j$ для выборки $\mathfrak{D}$ и каждого множества X_j таким образом, что матрица описывает бинарное отношение в каждой из пар объектов:

$$Z_j(i,k) = z_j(X_i, X_k),$$

где j индекс признака, а i,k индексы объекта; при этом Z_0 описывается по-другому:

$$Z_0(i,k) = \begin{cases} 1, & \text{при } y_i \succeq y_k, \\ 0, & \text{иначе.} \end{cases}$$

Конусы и генераторы

Выпуклый конус в $\mathbb{R}^m$ это такое множество $\mathcal{X}$, что

$$\mathcal{X} = \{\chi \mid A\chi \leq \mathbf{0}, \in \mathbb{R}^m\},$$

Theorem

Вектор χ, принадлежащий конусу $\mathcal{X}$ может быть единственным образом разложен в ЛК с неотрицательными коэффициентами с неотрицательными коэф-ами,

$$\chi = \sum_{k=1}^{m} \lambda_{k\,k}, \quad \lambda_k \geq 0,$$

где ζ_k это генератор конуса $\mathcal{X}$,

$$\zeta_k(i) = \begin{cases} 1, & \text{if} \quad x_i \succeq x_k, \\ 0, & \text{if} \quad x_i \not\succeq x_k, \end{cases}$$

Задача оптимизации

Поскольку целевой вектор $\hat{y} \in \sum_{i=1}^{n} \mathcal{X}_i$, можно применить теорему выше для декомпозиции вектора $\hat{y}$ в линейную комбинацию конусов $\mathcal{X}_1, ..., \mathcal{X}_n$,

$$\hat{y} = \sum_{j=1}^{n} \sum_{k=1}^{m} \lambda_{jk} \zeta_{jk}, \quad \lambda_{jk} \in \mathbb{R}_+,$$

Вектор ζ_{jk} также является k столбцом матрицы Z_j. Отсюда определим u для каждого объекта,

$$u(x) = \sum_{j=1}^{n} w_j \sum_{k=1}^{m} \lambda_{jk} z_j(x, x_k), \tag{1}$$

Таким образом мы свели задачу к минимизации лосса следующей функции по параметрам w и λ

$$f_{w,\lambda}(x_i) = \phi \left(\sum_{k=1}^{m} \lambda_k \Psi(x_i, x_k) \right) \tag{2}$$

Alt. Изотоническая регрессия

Задача изотонической регрессии ставится следующим образом:

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} \sum_{i=1}^{n} w_i (x_i - y_i)^2 \tag{3}$$
$$\text{s.t.} \quad Ax \leq 0$$

Где матрица определена следующим образом:

$$A = \begin{pmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & 1 & -1 \end{pmatrix} = \begin{pmatrix} a_1^\top \\ \vdots \\ a_{n-1}^\top \end{pmatrix} \in \mathbb{R}^{n-1 \times n} \ ,$$

Alt. Косые решающие деревья

Решающие деревья задаются последовательностью условий на ЛК на каждом из узлов. Более общно говоря, возьмем в качестве примера $X = x_1; x_2; \dots x_d;\ C_j$ где C_j метка класса и x_i вещественные признаки. Проверка условий на узлах имеет вид:

$$\sum_{i=1}^{d} a_i x_i + a_{d+1} > 0. \tag{4}$$

Здесь $a_1; \dots; a_{d+1}$ вещественные. Так как такие условия будут аналогичными тем же в обычных решающих деревьях при криволинейной замене координат, мы называем класс таких деревьев косыми.

Стохастическая модификация прежнего алгоритма

Заменим старые матрицы Z, элементы которых принимали бинарные значения на P, векторы которой будут лежать в вероятностных симплексах. Более формально

$$P_0 = \begin{pmatrix} P(x_1 \succeq x_1 | \mu_1 \succeq \mu_1) & P(x_1 \succeq x_2 | \mu_1 \succeq \mu_2) & & \\ & \ddots & P(x_i \succeq x_j | \mu_i \succeq \mu_j) & \cdots \\ & & & \ddots \\ & & & P(x_n \succeq x_n | \mu_n \succeq \mu_n) \end{pmatrix}$$

Оцениваем изначальное приближение матрицы по выборке(сэмплированием), а затем, меняя распределения бернуллиевских распределений компонент матриц, оптимизируем старый функционал, только уже по трём параметрам.

Вычислительный эксперимент

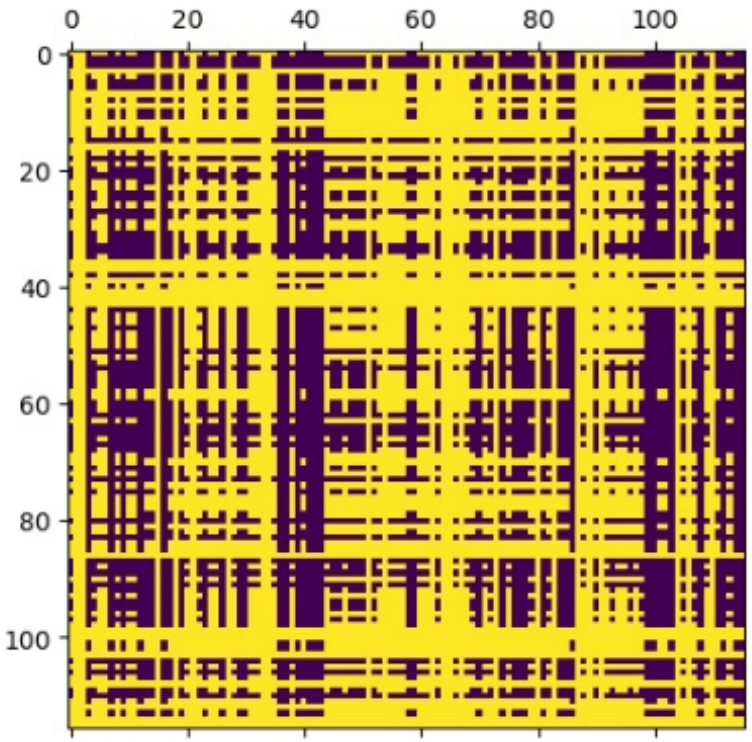

Рис.: изначальные отношения

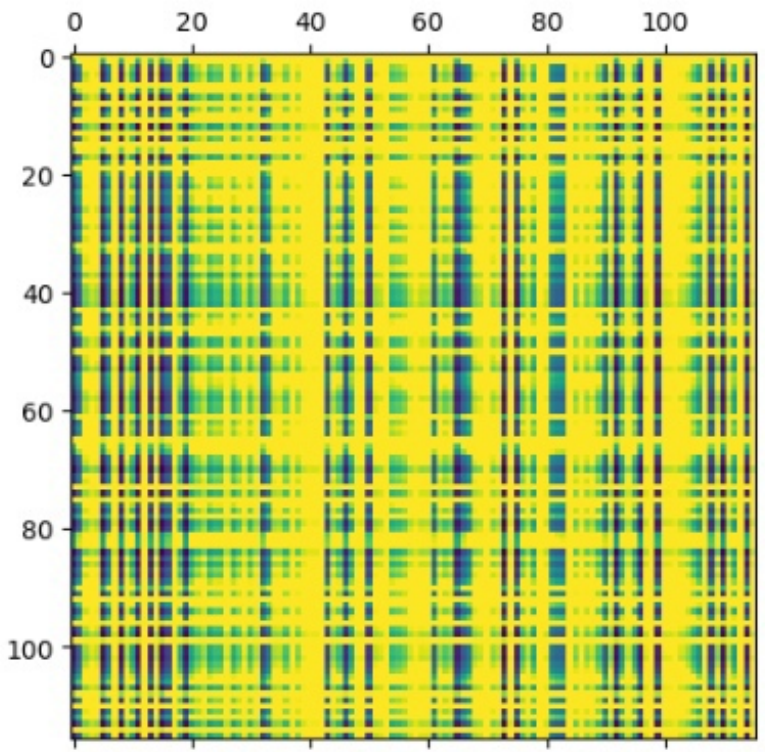

Рис.: приближение

Сравнение MAE

Таблица: Algorithm comparison on the Football Player dataset.

Algorithm	Learn error	Test error
Partial Orders	1.14 $\pm$0.05	1.69 $\pm$0.2
Isotonic Regression	0.98 $\pm$0.2	1.28 $\pm$0.4
Oblique Decision Trees	0.47 $\pm$0.5	1.06 $\pm$0.71
Stochastic Partial Orders	1.48 $\pm$0.14	1.32 $\pm$0.05

Сравнение моделей

Таблица: Comparative analysis of basic solution to my problem.

Solution	Strengths	Weakness
Partial Orders	no contradictions in the partial order	vulnerable to noise
Isotonic regression	working fast	inaccurately predictions
Oblique decision trees	learns non-linearity dependence	vulnerable to retraining
Stochastic Partial Orders	no vulnerability to noise	demand higher dimension of features

Заключение

- Как основной результат - было разработано два новых метода для решения задачи ранговой классификации. Был проведён их сравнительный анализ с другими методами, решающими данную задачу.
- Дальнейшим направлением исследования будет дальнейшее улучшение алгоритма, возможно в сторону прямого восстановления матрицы отношений между объектами через спектр графа его Лапласиана.