Метод построения случайного леса на основе отдаления друг от друга базовых моделей

A Preprint

Дмитриев Леонид Алексеевич
МГУ им. М.В. Ломоносова
ф-т ВМК, кафедра ММП
s02200542@gse.cs.msu.ru

д.ф-м.н. Сенько Олег Валентинович
МГУ им. М.В. Ломоносова
ф-т ВМК, кафедра ММП
senkoov@mail.ru

Abstract

В работе представлен новый метод случайного леса, который строится итеративно и в котором новое дерево обучается с учетом накопленного до него ансамбля. Данный метод хорошо проявляет себя на некоторых химических и медицинских данных. Исследование метода проводилось на датасете кристаллических решеток.

Keywords Случайный лес · Решающие деревья

1 Introduction

Случайный лес - широко известный алгоритм машинного обучения. Его способность хорошо аппроксимировать данные за счёт уменьшения разброса основана на предположении, что все деревья ансамбля независимы и различны. Но практическое уменьшение разброса значительно меньше теоретического, так как на деле получаемые деревья обучаются на объектах из одного и того же множества. Это проблема частично решается такими методами, как беггинг и бутстрап. Предлагается другой способ повышения разнообразия деревьев внутри леса. Вместо независимой и параллельной генерации деревьев, будем на каждом шагу добавлять дерево, сильно отличающееся от уже созданного ансамбля, с помощью специального функционала, учитывающего ответы предыдущих моделей. Выдвигается гипотеза, что данный метод повысит разнообразие деревьев в ансамбле и тем самым уменьшит разброс предсказаний. Помимо отличия от предыдущих моделей, следующему дереву можно придать свойства какого-либо другого метода, например градиентного бустинга. Это может сделать лес не только разнообразным, но и более качественным. В данной работе будет детально рассмотрено построение дерева на очередном шаге по вышеописанному методу и для тестирования будет исследовано его применение на данных о химических веществах.

2 Problem statement

2.1 Теория

Предположим, что у нас есть переменная Y, стохастически зависящая от вектора переменных $X_1, \ldots, X_n$, функции $G_1(x)$, $G_2(x)$, детерминировано зависящие от вектора переменных $X_1, \ldots, X_n$.

Имеется выборка $\tilde{S} = \{(y_j, x_j, G_1(x_j), G_2(x_j)), j = \overline{1,m}\}$, где

- y_j - значение переменной Y для объекта с номером j
- $x_j = (x_{j1}, \ldots, x_{jn})$ - вектор значений признаков $X_1, \ldots, X_n$ для объекта с номером j
- $G_1(x_j)$ - значение функции G_1 в точке x_j
- $G_2(x_j)$ - значение функции G_2 в точке x_j

Предлагается построить дерево $T(x)$, для которого достигается минимум функционала:

$$\Phi(\tilde{S}, T) = \Sigma_{j=1}^{m}\{\gamma_1[T(x_j) - y_j]^2 + \gamma_2[T(x_j) - G_2(x_j)]^2 - \mu[T(x_j) - G_1(x_j)]^2\}$$

где $\gamma_1 + \gamma_2 = 1; \quad \gamma_1, \gamma_2, \mu \in [0, 1]$

Как видно из структуры функционала Φ, дерево T, соответствующее его минимуму, будет аппроксимировать связь Y с переменными $X_1, \ldots, X_n$ при $\gamma_1 > 0$.

Одновременно дерево $T(x)$ будет удаляться от зависимости $G_2(x)$ при возрастании μ и приближаться к зависимости $G_1(x)$ при возрастании γ_2.

2.2 Реализация дерева

При построении дерева был использован "жадный" метод оптимизации целевого функционала: на каждом шагу к дереву добавляется узел, обеспечивающий наибольшее снижение используемого функционала Φ.

Предположим, что на каком-то шаге дерево T_k содержит k концевых узлов, которым соответствуют концевые выборки $S_1^k, \ldots, S_k^k$.

Новое дерево T_{k+1} строится через добавление к дереву T_k дополнительного узла u.

Узел u получается из некоторого концевого узла g с помощью порогового правила вида $X_u > \delta_u$, где X_u и δ_u признак и порог к нему соответственно.

Правило $X_u > \delta_u$ расщепляет выборку S_g^k на две подвыборки.

Признак X_u и порог δ_u ищутся из условия максимизации разности $\Phi(\tilde{S}, T_k) - \Phi(\tilde{S}, T_{k+1})$.

Процесс построения может быть прекращен при выполнении одного из перечисленных условий:

- На очередном шаге не удается уменьшить функционал
- На очередном шаге произошло изменение функционала, меньшее чем некоторое пороговое значение
- Кол-во объектов внутри узла меньше некоторого порогового значения

3 Headings: first level

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula. See Section 3.

3.1 Headings: second level

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

$$\xi_{ij}(t) = P(x_t = i, x_{t+1} = j|y, v, w; \theta) = \frac{\alpha_i(t) a_{ij}^{w_t} \beta_j(t+1) b_j^{v_{t+1}}(y_{t+1})}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i(t) a_{ij}^{w_t} \beta_j(t+1) b_j^{v_{t+1}}(y_{t+1})} \tag{1}$$

3.1.1 Headings: third level

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Paragraph Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

4 Examples of citations, figures, tables, references

4.1 Citations

Citations use `natbib`. The documentation may be found at

`http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf`

Here is an example usage of the two main commands (`citet` and `citep`): Some people thought a thing (? ?) but other people thought something else (?). Many people have speculated that if we knew exactly why ?) thought this...

4.2 Figures

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi. See Figure ??. Here is how you add footnotes. [1] Sed feugiat. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Ut pellentesque augue sed urna. Vestibulum diam eros, fringilla et, consectetuer eu, nonummy id, sapien. Nullam at lectus. In sagittis ultrices mauris. Curabitur malesuada erat sit amet massa. Fusce blandit. Aliquam erat volutpat. Aliquam euismod. Aenean vel lectus. Nunc imperdiet justo nec dolor.

4.3 Tables

See awesome Table 1.

The documentation for `booktabs` ('Publication quality tables in LaTeX') is available from:

`https://www.ctan.org/pkg/booktabs`

4.4 Lists

- Lorem ipsum dolor sit amet
- consectetur adipiscing elit.
- Aliquam dignissim blandit est, in dictum tortor gravida eget. In ac rutrum magna.

[1] Sample of the first footnote.

Таблица 1: Sample table title

Part		
Name	Description	Size (μm)
Dendrite	Input terminal	$\sim$100
Axon	Output terminal	$\sim$10
Soma	Cell body	up to 10^6

Список литературы

[1] O.V. Senko, A.A. Dokukin, N.N. Kiselyova, V.A. Dudarev , Yu.O. Kuznetsova - "New Two-Level Ensemble Method and Its Application to Chemical Compounds Properties Prediction"

[2] Liu, Y., Wang, Y., Zhang, J. (2012). New Machine Learning Algorithm: Random Forest. In: Liu, B., Ma, M., Chang, J. (eds) Information Computing and Applications. ICICA 2012. Lecture Notes in Computer Science, vol 7473. Springer, Berlin, Heidelberg.