Формулировка и решение задачи оптимизации, сочетающей классификацию и регрессию, для оценки энергии связывания белка и маленьких молекул

Анастасия Грачёва

Московский физико-технический институт

Курс: Численные методы обучения по прецедентам (практика, В. В. Стрижов)/Группа 694, весна 2019

Цель исследования

При разработке лекарства возникает задача поиска маленьких молекул - лигандов, наиболее сильно взаимодействующих с исследуемым белком, а значит являющихся основными кандидатами в лекарства.
Так как энергия связывания молекул в нативном положении достигает минимума, то, обучив классификатор, мы получаем возможность из сгенерированных положений выбирать одно наиболее близкое к нативному. Есть предположение, что качество предсказания может быть повышено, если использовать экспериментальные данные о свободной энергии связывания молекул и решать одновременно задачи регрессии и классификации. Проверке этого предположения посвящено исследование.

Постановка задачи

$$
\begin{aligned}
&\underset{\mathbf{w},\xi_{ij}}{\text{minimize:}} && \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{ij}\xi_{ij} + C_r\sum_{i}\left(\langle \mathbf{w}, \mathbf{x}_{i0}\rangle - s_i\right)^2 \\
&\text{subject to:} && y_{ij}[\langle \mathbf{w}, \mathbf{x}_{ij}\rangle - b_i] - 1 + \xi_{ij} \geq 0, \\
& && \xi_{ij} \geq 0, \\
& && i \in \{1, \ldots, P\} \\
& && j \in \{0, \ldots, D\}
\end{aligned}
\tag{1}
$$

Базовый алгоритм

Сведение к квадратичной задаче

$$\begin{aligned} &\frac{1}{2}x^\top Px + q^\top x \to min \\ &s.t. Gx + h \leq 0 \end{aligned} \tag{2}$$

$$\begin{aligned} &\mathbf{w}^\mathsf{T} \leftarrow (\mathbf{w}^\mathsf{T}, b_1, \dots, b_P, \varepsilon_{00}, \dots \varepsilon ij, \dots), \\ &\mathbf{x}_{10}^\mathsf{T} \leftarrow (\mathbf{x}_{1j}^\mathsf{T}, -1, 0, \dots, 0, 1, \dots 0), \\ &\mathbf{x}_{11}^\mathsf{T} \leftarrow (\mathbf{x}_{1j}^\mathsf{T}, -1, 0, \dots, 0, 0, 1, \dots, 0), \\ &\mathbf{x}_{2j}^\mathsf{T} \leftarrow (\mathbf{x}_{2j}^\mathsf{T}, 0, -1, 0, \dots, 0, \dots), \\ &\dots \\ &i \in \{0, \dots, P\}, j \in \{0, \dots, D\}. \end{aligned} \tag{3}$$

Продвинутый алгоритм

Сведение к стандартному SVM-виду, чтобы сохранить блочную структуру данных для повышения эффективности.

$$\begin{aligned} \underset{\lambda_{ij}}{\text{minimize:}} \quad & -\sum_{ij} \lambda_{ij} + \frac{1}{2} \sum_{(i,j),(p,q)} \lambda_{ij}\lambda_{pq}y_{ij}y_{pq}\langle \hat{\mathbf{x}}_{ij}, \hat{\mathbf{x}}_{pq} \rangle \\ \text{subject to:} \quad & 0 \geq \lambda_{ij} \geq C, \\ & i \in \{1, \ldots, P\}, \\ & j \in \{0, \ldots, D\}. \end{aligned} \tag{4}$$

Литература

Ближайшие работы

1) Maria Kadukova, Sergei Grudinin. Convex-PL: a novel knowledge-based potential for protein-ligand interactions deduced from structural databases using convex optimization. Journal of ComputerAided Molecular Design, Springer Verlag, 2017, 31 (10), pp.943-958.
2) Sergei Grudinin, Maria Kadukova, Andreas Eisenbarth, Simon Marillet, Frederic Cazals. Predicting binding poses and affinities for protein-ligand complexes in the 2015 D3R Grand Challenge using a physical model with a statistical parameter estimation. Journal of Computer-Aided Molecular Design, Springer Verlag, 2016, 30 (9), pp.791-804.

Заключение

Несмотря на то, что базовый алгоритм проще и стабильнее, он не эффективен на полном объёме данных, так как имеет квадратичную сложность. Поэтому следующим шагом будет привести задачу к стандартному SVM-виду, что позволит сохранить блочную структуру данных, т.е. не сравнивать между собой признаковые вектора, относящиеся к разным комплексам(блокам), и тем самым существенно повысить эффективность.