Конформная классификация в задачах прогнозирования физико-химических свойств молекул

Матвеев А. А.[1], Бурнаев Е. В.[2] и Тетко И. В.[3,4]

[1] Сколтех, ИППИ РАН
albert.matveev@skolkovotech.ru,
[2] e.burnaev@skoltech.ru
[3] Institute of Structural Biology, Helmholtz Zentrum München für Gesundheit und Umwelt (GmbH)
[4] BigChem GmbH
i.tetko@helmholtz-muenchen.de

Аннотация В современной химии все чаще применяются методы машинного обучения. Одной из задач, которые решаются такими методами, является прогнозирование свойств молекул. Это позволяет ускорить и удешевить разработку новых лекарственных средств. Отличительными особенностями такого рода задач являются большая размерность пространства признаков, шум в экспериментальных данных, неоднородность выборки. Эти особенности требуют дополнительного анализа и предложения новых подходов.
Кроме того, в таких задачах исследователь должен быть уверен в своем прогнозе, поэтому важно иметь оценку точности прогноза. Одним из методов, которые позволяют это сделать являются конформные предикторы.
В данной статье рассматриваются конформные предикторы для задачи многоклассовой классификации. Этот подход позволяет получать доверительные множества наперед заданной вероятности.
В результате исследования предложена новая конформная метрика для построения конформных предикторов в случае задачи классификации, разработана методика построения регрессии на основе классифицирующих алгоритмов, описанные методы применены к практической задаче и проведено численное моделирование на данных температуры плавления.

Ключевые слова: конформные предикторы, конформная классификация, температура плавления, химическая информатика

1 Введение

В настоящее время интеллектуальный анализ данных и машинное обучение все чаще применяются в качестве инструментов прогнозирования в естественных науках. Химическая информатика является разделом химии, в котором эти методы используются для обработки данных, полученных из экспериментов [1]. Такие задачи сопряжены с рядом сложностей, характерных

именно для химии: большая размерность пространства признаков, ошибки в записях данных, неустранимый шум в данных, неоднородность пространства признаков [2,3]. Эти сложности требуют разработки новых подходов и усовершенствования существующих.

Важнейшим применением химинформатики, на которое в настоящее время направлено значительное число научных исследований, является разработка новых эффективных лекарственных средств [4]. Исследования в данной области исторически сопряжены с большим количеством регуляторных ограничений, значительными временными затратами, и, как следствие, экономическими издержками [5,6].

Главным методом в этой области является построение количественных моделей структура-свойство (quantitative structure-activity (property) relationships, QSAR/QSPR) [5]. Идея этого метода заключается в осуществлении поиска соединений, обладающих необходимыми свойствами, используя математическое моделирование. Метод QSAR экономит ресурсы и и ускоряет процесс разработки новых соединений.

В данной статье рассматривается задача прогнозирования физико-химических свойств молекул и построения доверительных множеств для прогноза с помощью конформных предикторов. В области разработки новых лекарств эта задача появляется естественным образом при попытках оценить те или иные свойства молекул, которым должно удовлетворять новое лекарственное средство [7]: токсичность, растворимость в воде и т. д. В области оценки точности прогноза моделей QSAR использовались методы на основе domain adaptation [8]. Кроме того, в диссертации [9] проведен анализ точности прогноза на основе оценки области применимости. Применение метода ближайших соседей для тех же целей описано в [10]. Также оценку точности прогноза получали на основе конформных предикторов [11].

В данной работе предложена новая конформная метрика для построения конформных предикторов в случае задачи классификации, разработана методика построения регрессии на основе классифицирующих алгоритмов, описанные методы применены к практической задаче и проведено численное моделирование на данных температуры плавления (melting point, MP).

Статья организована следующим образом. В первом разделе описана постановка задачи, во втором разделе описаны конформные предикторы, в том числе для задачи классификации, третий раздел содержит описание данных для эксперимента, четвертый раздел посвящен описанию вычислительного эксперимента и результатам расчетов, в заключении приведены основные выводы.

2 Постановка задачи

Ставится задача прогнозирования физико-химических свойств вещества на основе структуры молекулы. Имеется обучающая выборка из N молекул: $S = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, вектор признаков (дескрипторов) $x_i \in \mathbb{R}^M$ для каждой молекулы содержит M элементов, каждой молекуле из обучающей

выборки присвоена метка $y_i \in \mathbb{R}$. Необходимо построить алгоритм, который на основе имеющегося вектора дескрипторов прогнозировал бы целевую переменную.

Рассмотрена постановка, приводящая задачу регрессии к многоклассовой классификации. Разобьем имеющийся набор действительнозначных меток на подмножества, и, представив их как классы, назначим новые метки. После этого применим алгоритмы многоклассовой классификации.

Разобьем выборку на классы по квантилям распределения выборки. Будем делить имеющуюся длину выборки на нужное число классов и брать в качестве границ классов значения целевой переменной, соответствующие порядковой статистике с указанным номером. Для данных MP разделим выборку на 10 классов.

3 Конформные предикторы

Во многих приложениях машинного обучения важны не только сами предсказания, но и доверительные интервалы для этих предсказаний. Например, в задачах медицинской диагностики необходимо иметь надежные предсказания и их доверительную вероятность. К таким приложениям также относится химическая информатика, когда рассматриваются задачи прогнозирования таких величин, как токсичность.

Подход, основанный на конформных предикторах [12,13], позволяет для заданной вероятности получить не только сам прогноз, но и его доверительный интервал (или доверительное множество в случае классификации). Еще одной особенностью этого метода является возможность настраивать уровень доверия, что приводит к изменению ширины интервала.

Конформные предикторы используют предыдущий опыт для определения величины доверительного множества для новых предсказаний. Пусть дана вероятность ε и алгоритм, который генерирует предсказание $\hat{y}$. Тогда описываемый метод позволяет получить множество, включающее в себя $\hat{y}$, которое также содержит истинное значение y с вероятностью $1 - \varepsilon$. Этот метод может применяться к любому алгоритму машинного обучения.

Главным требованием для работы этого подхода является взаимозаменяемость данных. Иначе говоря, данные должны быть независимы и одинаково распределены. Это достаточно сильное условие, которое, тем не менее, выполняется в большинстве возможных приложений.

Используемый для точечного прогноза алгоритм определяет меру неконформности, которая выражает, насколько нетипичным является предсказываемый объект по отношению к предыдущим объектам.

После задания меры неконформности и величины ε, вычисляется множество Γ^ε. Эти множества имеют важное свойство вложенности, характерное для доверительных интервалов: при $\varepsilon_1 > \varepsilon_2$ верно, что $\Gamma^{\varepsilon_1} \subset \Gamma^{\varepsilon_2}$. Назовем доверительное множество уровня ε валидным, если оно включает в себя истинную метку в $(1 - \varepsilon)\%$ случаев.

Назовем мультимножеством $\wr a_1, \ldots, a_n \int$ множество, полученное из последовательности $a_1, \ldots, a_n$, исключив информацию об упорядоченности элементов. Это понятие важно для определения меры неконформности. Назовем такой мерой следующее семейство отображений:

$$A_n : (\mathcal{Z})^{n-1} \times (\mathcal{Z}) \to \mathbb{R},$$

где $(\mathcal{Z})^{n-1}$ – мультимножество объектов (x, y) размера $n-1$. Мера определяет числа $\alpha_i = A_n(\wr z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n \int, z_i)$, выражающие, насколько объект z_i отличен от объектов в мультимножестве $\wr z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n \int$.

Сами по себе числа α_i не несут много информации, однако позволяют сравнивать объекты. Например, мы хотим получить такую информацию о новом предсказании $(x_{n+1}, \hat{y}_{n+1})$. Тогда мы можем посчитать величину

$$p(\hat{y}_{n+1}) = \frac{\#\{i = 1, \ldots n+1 : \alpha_i \geq \alpha_{n+1}\}}{n+1}$$

Эта величина принимает значения между $\frac{1}{n+1}$ и 1, назовем ее $p - value$. Важным свойством этой величины является следующее неравенство:

$$\mathbb{P}(\wr z_1, \ldots, z_{n-1} \int : p(y_{n+1}) \leq \delta) \leq \delta,$$

доказательство этого утверждения приведено в [14]. Таким образом, если наблюдается $p - value$ некоторой метки меньше наперед заданной малой величины, например, 0.05, это означает, что метка маловероятна. Следовательно, при заданной вероятности $1-\varepsilon$ конформные предикторы генерируют множество

$$\Gamma^\varepsilon = \{y : p(y) > \varepsilon\}.$$

3.1 Конформная классификация

Рассмотрим задачу многоклассовой классификации. Пусть у нас есть выборка $S = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, обозначим за z_i пару (x_i, y_i). $y_i \in \{1, \ldots, K\}$, где K – число классов. При многоклассовой классификации итоговое предсказание представляет собой K чисел, которые можно назвать вероятностями исходов для объекта x_i, обозначим их $o_i^1, \ldots o_i^K$. Для данной задачи выборка разбивается на тренировочную, калибровочную и тестовую. Далее определяется мера неконформности A, которая порождается алгоритмом классификации. На тренировочной выборке обучается базовый алгоритм, на калибровочной рассчитываются меры неконформности. При построении прогноза на тестовой выборке рассчитываются K значений меры неконформности для объекта для всех возможных исходов, затем для каждого из возможных исходов значения меры неконформности сравниваются со значениями на калибровочной выборке для получения $p - value$. В итоговое множество предсказаний входят классы, для которых $p - value$ больше уровня доверия. Точечный прогноз, производимый конформным предиктором – это класс, имеющий максимальное $p - value$. Поэтому качество классификации может отличаться от используемого базового алгоритма.

Обозначим за u истинный класс объекта x_i: $u = y_i$. Ожидается, что максимальное из чисел $o_i^1, \dots o_i^K$ соответствует истинному классу u. Таким образом, объект тем более конформен, чем больше соответствующее истинному классу число, то есть, чем увереннее он классифицируется. Наоборот, чем больше значения, соответствующие другим классам, тем менее конформен рассматриваемый объект. Тогда рассмотрим максимум среди чисел, которые соответствуют классам, отличным от истинного:

$$\max_{j=1,\dots,K, j\neq u} o_i^j.$$

После этого возможно определить меру неконформности [15]:

$$\alpha_i = \max_{j=1,\dots,K, j\neq u} o_i^j - o_i^u. \tag{1}$$

Эта мера может быть определена на произвольном алгоритме классификации. Например, для случайного леса эти вероятности определяются как доли деревьев, проголосовавших за данный класс [16]. Сами деревья решений генерируют такие вероятности в виде частот попаданий объектов из обучающей выборки в конкретный лист [17]. Для нейронных сетей вероятности получаются естественным образом после softmax слоя [15].

Однако, возможность произвольного выбора меры неконформности открывает широкие возможности для экспериментов. В данной работе предложена новая мера неконформности, которая в результате вычислительных экспериментов показывает лучшее качество по сравнению с приведенной выше стандартной мерой. Определим новую меру следующим образом:

$$\alpha_i = \frac{1}{o_i^u + \gamma} - \frac{1}{\max\limits_{j=1,\dots,K, j\neq u} o_i^j + \gamma},$$

где γ – некоторое малое число, вводимое во избежание деления на ноль.

В то же время, в случае рассматриваемой задачи, а именно при сведении регрессии к многоклассовой классификации, эту меру можно улучшить.

$$\alpha_i = \frac{|j-u|}{o_i^u + \gamma} - \frac{1}{\max\limits_{j=1,\dots,K, j\neq u} o_i^j + \gamma}, \tag{2}$$

Множитель $|j-u|$ появляется благодаря структуре рассматриваемой задачи. Поскольку номера классов упорядочены, и, вообще говоря, представляют собой возрастающие усредненные значения действительнозначных меток, возможно ввести расстояние между классами, и рассматривать этот множитель как своего рода пенализацию, то есть, объект тем более неконформен, чем дальше предсказанный класс расположен от истинного.

Существует несколько критериев для измерения качества конформных предсказаний [18]. Первый из них – точность конформного предиктора, то есть, число верных предсказаний, сгенерированных на основе максимального $p - value$.

Кроме того, измеряется средний размер конформного предсказания:

$$\frac{1}{k}\sum_{i=1}^{k}|\Gamma^{\varepsilon}(x_i)|,$$

где k – длина тестовой выборки. Желательно, чтобы этот показатель был как можно ближе к единице.

Еще одной метрикой качества является среднее значение $p - value$:

$$\frac{1}{k}\sum_{i=1}^{k}\sum_{y}p_i(y).$$

Значение этой метрики тем лучше, чем оно меньше.

4 Описание данных

Для проведения вычислительных экспериментов был выбран набор данных экспериментально измеренной температуры плавления веществ.

Датасет MP получен из описаний патентов химических соединений, в которые включаются экспериментально измеренные физико-химические величины. Этот набор данных состоит из $N = 271\,537$ молекул, которые получены из исходного датасета после исключения явных выбросов и ошибок данных. Его особенностью является высокий уровень шума в данных, то есть, при занесении записи в патенты, могут быть использованы не точные значения температуры, а интервалы, в которых лежит истинное значение, или может быть указано, что температура выше или ниже некоторого значения. Кроме того, измерение температуры плавления может быть сложно в экстремальных случаях, если температура значительно низкая (ниже 50 градусов) или значительно высокая (свыше 250 градусов). В результате этого, спрогнозировать эту величину со среднеквадратичной ошибкой меньше, чем 35 – это непростая задача [2].

Каждая молекула представлена записью SMARTS, которая является символьной записью структуры графа молекулы. На основе SMARTS рассчитываются молекулярные дескрипторы, которые используются как векторы признаков для алгоритмов машинного обучения. В химической информатике нет определенного подхода для выбора дескрипторов для конкретной задачи, для разных приложений лучший набор дескрипторов может быть разным. Потому чаще всего исследователи строят модели на разных наборах дескрипторов и выбирают тот набор, на котором алгоритмы показывают меньшую ошибку.

Для данной работы эта процедура предварительных расчетов была проведена на платформе OCHEM [19]. Были выбраны дескрипторы OEstate, при использовании которых были получены ошибки в среднем меньше, чем на других наборах данных. Сами значения векторов также были рассчитаны на данной платформе. При выгрузке значений проводится стандартная обработка данных, выбираются значения с коэффициентом корреляции менее

определенного порога, исключаются очень разреженные значения с большим количеством нулей. В результате размерность признаков для данных MP составляет $M = 349$.

Сами дескрипторы OEstate – индексы электро-топологического состояния, которые объединяют в себе информацию об электронах и топологии графа. Помимо таких индексов, дескрипторы включают в себя такие величины, как число акцепторов водородной связи, число атомов-галогенов, количество атомов водорода, кислорода и некоторых других в молекуле, polar surface area и другие.

5 Результаты

Поскольку шум в данных не позволяет прогнозировать рассматриваемое свойство с хорошей точностью, можно рассмотреть задачу как классификацию и строить прогноз в кусочно-постоянном виде.

В качестве базовых алгоритмов были выбраны логистическая регрессия, решающее дерево, случайный лес и многослойный перцептрон.

Для измерения качества предсказания будем использовать модифицированную метрику $RMSE^*$. Для ее расчета после проведения классификации будем назначать каждому классу среднее значение всех меток, принадлежащих данному классу, после этого будем считать значение $RMSE$ между средним значением предсказанном классе и реальным значением метки конкретной молекулы:

$$RMSE^* = \sqrt{\frac{1}{n}\sum_{j=1}^{K}\sum_{\hat{y}_i=j}(\overline{y}_j - y_i)^2},$$

где n - число объектов в обучающей выборке, K - число классов, $\hat{y}_i$ - предсказанная метка класса для y_i $\overline{y}_j$ - среднее значение класса j, j - метки классов.

Алгоритм	$RMSE^*$	CP $RMSE^*$	Среднее число классов в предсказании			Среднее $p-value$
			90%	95%	99%	
Лог. регрессия	58.109	58.115	6.99	8.09	9.22	3.15
Случайный лес	47.62	47.49	**5.68**	**6.83**	**8.18**	**2.52**
Решающее дерево	54.7	54.8	7.94	8.87	9.72	3.49
Перцептрон	**43.79**	**43.79**	6.9	8.28	9.49	2.84
Лог. регрессия	58.109	56.00	5.99	6.96	8.64	2.78
Случайный лес	47.62	46.53	5.46	6.51	8.13	2.47
Решающее дерево	54.7	54.27	6.71	8.04	8.88	3.02
Перцептрон	**43.79**	**43.63**	**5.32**	**6.34**	**8.07**	**2.42**

Таблица 1. Значения $RMSE^*$, среднего числа классов в предсказании и среднего $p-value$ для конформных классификаторов, верхняя часть – мера (1), нижняя – мера (2)

В таблице 1 представлены расчеты конформной классификации. Колонка $RMSE^*$ – ошибка базового алгоритма, CP $RMSE^*$ – ошибка точечного предсказания, сгенерированного конформным классификатором. Также представлены значения среднего числа классов в конформном предсказании и среднее значений $p - value$. Жирным шрифтом в обеих частях выделены лучшие результаты. При использовании меры (1) наименьшие значения среднего размера конформного предсказания и среднего значения $p - value$ наблюдаются у алгоритма случайного леса. Как видно из таблицы, предложенная мера (2) позволила улучшить качество прогноза. Особенно успешно она показала себя с базовым алгоритмом в виде нейронной сети, который в итоге демонстрирует лучшее качество.

6 Заключение

В работе рассмотрена задача прогнозирования физико-химических свойств молекул и представлена постановка задачи, сводящая регрессию к классификации. Для применения этого подхода была введена метрика качества прогноза и были построены базовые классификаторы. Затем к полученным прогнозам применялся подход конформных предикторов. Для этого в расчетах была использована стандартная мера неконформности, типично применяемая для такой задачи, и новая мера, предложенная в данной работе. Было продемонстрировано, что предложенная мера позволила улучшить все метрики качества конформной классификации.

Благодарности

Данное исследование проведено при поддержке грантов РФФИ 16-01-00576 А и 16-29-09649 офи_м.

Список литературы

1. J.B.O. Mitchell. Machine learning methods in chemoinformatics. *Wiley Interdiscip Rev Comput Mol Sci. 4(5): 468–481.*, 2014.
2. I.V. Tetko, D.M. Lowe, and A.J. Williams. The development of models to predict melting and pyrolysis point data associated with several hundred thousand compounds mined from patents. *J Cheminform; 8: 2.*, 2016.
3. M. Vogt and J. Bajorath. Chemoinformatics: a view of the field and current trends in method development. *Bioorg Med Chem. 20(18):5317-23*, 2012.
4. I.V. Tetko, O. Engkvist, and H. Chen. Does 'big data' exist in medicinal chemistry, and if so, how can it be harnessed? *Future Med Chem. 8(15):1801-1806.*, 2016.
5. A. Cherkasov et al. Qsar modeling: where have you been? where are you going to? *J Med Chem. 57(12): 4977-5010.*, 2014.
6. W. Warr. Some trends in chem(o)informatics. *Chemoinformatics and Computational Chemical Biology*, 2011.
7. J. Hodgson. Admet–turning chemicals into drugs. *Nat Biotechnol. 19(8):722-6.*, 2001.

8. U. Sahlin, N. Jeliazkova, and T. Oberg. Applicability domain dependent predictive uncertainty in qsar regressions. *Mol. Inf. 33, 26-35.*, 2014.
9. F. Sahigara, D. Ballabio, R. Todeschini, and V. Consonni. Defining a novel k-nearest neighbours approach to assess the applicability domain of a qsar model for reliable predictions. *J Cheminform. 5: 27.*, 2013.
10. I. Sushko. *Applicability domain of QSAR models.* Technical University of Munich., 2011.
11. U. Norinder, L. Carlsson, S. Boyer, and M. Eklund. Introducing conformal prediction in predictive modeling. a transparent and flexible alternative to applicability domain determination. *J Chem Inf Model. 54(6):1596-603.*, 2014.
12. G. Shafer and V. Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research 9 371-421.*, 2008.
13. V. Vovk, Gammerman A., and G. Shafer. *Algorithmic learning in a random world.* Springer Science and Business Media, 2005.
14. I. Nouretdinov, M.V. Vovk, V.and Vyugin, and A. Gammerman. Pattern recognition and density estimation under the general i.i.d. assumption. *Proceedings of the 14th Annual Conference on Computational Learning Theory and 5th European Conference on Computational Learning Theory, Vol. 2111 of Lecture Notes in Computer Science, pp. 337–353.*, 2001.
15. H. Papadopoulos, V. Vovk, and Gammerman A. Conformal prediction with neural networks. *IEEE International Conference on Tools with Artificial Intelligence, volume 2, pages 388–395.*, 2007.
16. D. Devetyarov and I. Nouretdinov. Prediction with confidence based on a random forest classifier. *IFIP International Conference on Artificial Intelligence Applications and Innovations, pages 37–44.*, 2010.
17. U. Johansson, R. Konig, T. Lofstrom, and H. Bostrom. Evolved decision trees as conformal predictors. *IEEE Congress on Evolutionary Computation, pages 1794–1801.*, 2013.
18. V. Vovk, V. Fedorova, and I. Gammerman A. Nouretdinov. Criteria of efficiency for conformal prediction. *Symposium on Conformal and Probabilistic Prediction with Applications, pages 23–39.*, 2016.
19. I. Sushko et al. Online chemical modeling environment (ochem): web platform for data storage, model development and publishing of chemical information. *J. Comput. Aided. Mol. Des. 25, 533-54*, 2011.