УДК 51.76

ОЦЕНКА ВЗАИМНОЙ ИНФОРМАЦИИ ДЛЯ ОТБОРА ПРИЗНАКОВ ПРИ ПРОГНОЗЕ УСТОЙЧИВОСТИ ПЕНТАПЕПТИДОВ

И.В. Петров
Институт проблем управления им. В.А. Трапезникова РАН
Россия, 117997, Москва, Профсоюзная ул., 65
E-mail: ivan.petrov@phystech.edu

В.В. Цурко
Институт проблем управления им. В.А. Трапезникова РАН
Россия, 117997, Москва, Профсоюзная ул., 65
E-mail: v.tsurko@gmail.com

А.И. Михальский
Институт проблем управления им. В.А. Трапезникова РАН
Россия, 117997, Москва, Профсоюзная ул., 65
E-mail: ipuran@yandex.ru

А.А. Анашкина
Институт молекулярной биологии им. В.А.Энгельгардта РАН
Россия, 119991 ул. Москва, Вавилова, 32
E-mail: anastasya.anashkina@gmail.com

А.Н. Некрасов
Институт биоорганической химии
им. академиков М.М. Шемякина и Ю.А. Овчинникова РАН
Россия, ул. 117198 Миклухо-Маклая, Москва, 16/10
E-mail: alexei_nekrasov@mail.ru

Ключевые слова: взаимная информация, экспериментальная выборка, МДМ, информативные признаки, конформационно-устойчивые пентапептиды.

Аннотация: Рассматривается проблема отбора информативных признаков для прогноза устойчивости коротких фрагментов белков, состоящих из пяти аминокислотных остатков – пентапептидов. Ранее было показано, что белковые блоки такого характерного размера являются основными структурными элементами белковых молекул и играют важную роль в формировании их структуры. Отбор информативных признаков при классификации основан на оценке взаимной информации, содержащейся в паре признак-метка класса. Предложена непараметрическая оценка взаимной информации, вытекающая из теории эмпирических процессов. Приведены результаты классификации пентапептидов, устойчивость которых была определена ранее методом молекулярно-динамического моделирования.

1. Введение

Большие объемы экспериментальных данных, используемых в современных научных исследованиях и прикладных задачах, характеризуются неоднородностью, связанной с различием источников данных и методов их регистрации, зашумленностью и информационной избыточностью вследствие отсутствия модели, указывающей на факторы, однозначно связанные с целевыми изучаемыми показателями.

Подобные задачи характеризуются и высокой размерностью пространства признаков. Например, при изучении клетки или образца ткани при помощи методов секвенирования генома нового поколения возможно получение информации об экспрессии практически всех белок-кодирующих генов, а также коротких и длинных некодирующих РНК. Типичный размер набора таких данных ничтожно мал по сравнению с количеством признаков. Число признаков измеряется тысячами, а число образцов в лучшем случае сотнями.

Для повышения эффективности и надежности результатов анализа данных применяются методы отбора признаков, выделения закономерностей, присущих данным, кластеризации данных на однородные группы. В результате снижается размерность задачи, что эквивалентно увеличению числа наблюдений в данных, уменьшается влияние на результат случайных возмущений в данных и повышается статистическая надежность результата.

В качестве критерия отсеивания признаков часто используется величина корреляции признака с целевой переменной. Однако при наличии существенно нелинейной связи между значениями признаков и целевой переменной величина корреляции может оказаться близка к нулю, что приводит к неверному суждению о значимости признака и снижению качества работы алгоритма.

В докладе рассматриваются методы оценки взаимной информации по экспериментальным данным и применение этой оценки для отбора информативных признаков. Описываются результаты применения такого подхода в реальной задаче отбора признаков для прогноза устойчивости конформации пентапептидов.

2. Оценка взаимной информация по эмпирическим данным

Формально взаимная информация между случайными величинами X и Y, имеющих совместное распределение $P(x,y)$ с плотностью $p(x,y)$ определяется соотношением

$$I(X,Y)=\iint p(x,y)\ln\frac{p(x,y)}{p(x)p(y)}dxdy\,.$$

Из определения следует, что для независимых случайных величин взаимная информация равна нулю. Простой методом оценки взаимной информации состоит в замене интегрирования по распределению $P(x,y)$ усреднением по выборочным значениям экспериментальных данных $((x_1,y_1),\ldots,(x_n,y_n))$

$$\hat{I}(X,Y)=\frac{1}{km}\sum_{i=1}^{k}\sum_{j=1}^{m}p(x_i,y_j)\ln\frac{p(x_i,y_j)}{p(x_i)p(y_j)},$$

где $p(x_i)$, $p(y_j)$ и $p(x_i,y_j)$ – оценки плотностей по экспериментальным данным. В качестве таких оценок можно использовать гистограммные оценки на прямоугольной сетке размера k на m, либо непараметрические – ядерные оценки для плотностей $p(x_i)$, $p(y_j)$ и $p(x_i,y_j)$. Общим недостатком обоих подходов является необходимость использовать достаточно большой набор экспериментальных данных для получения достаточно точных оценок плотностей, в особенности плотности совместного распределения X и Y.

2.1. Оценка через решение интегрального уравнения

В случае, если случайная величина Y принимает два значения, как, например, при бинарной классификации, в [1] предложен способ оценки взаимной информации используя оценку отношения плотностей, которая получается как решение интегрального уравнения по эмпирической выборке $x^y{}_1,\ldots,x^y{}_{ly}$, y=0, 1. Оценка взаимной информации имеет вид

$$\hat{I}(X,Y)=\frac{1}{l_0+l_1}\sum_{t=\{0,1\}}p(y=t)\left(\sum_{i=1}^{l_0}r_t\left(x_i^0\right)\ln r_t\left(x_i^0\right)+\sum_{i=1}^{l_1}r_t\left(x_i^1\right)\ln r_t\left(x_i^1\right)\right)$$

где функция $r_t(x)=p_t(x)/p(x)$ удовлетворяет интегральному уравнению

(1) $$F_t(x)=\int I(x\ge u)r_t(u)dF(u).$$

Аппроксимация этого уравнения по эмпирическим данным имеет вид

$$\frac{1}{l_t}\sum_{i=1}^{l_t}I\left\{x\ge x_i^t\right\}=\frac{1}{l_0+l_1}\left(\sum_{i=1}^{l_0}I\left(x\ge x_i^0\right)r_t\left(x_i^0\right)+\sum_{i=1}^{l_1}I\left(x\ge x_i^1\right)r_t\left(x_i^1\right)\right),\ t=0,\ 1.$$

Уравнение (1) является плохо определенным интегральным уравнением первого рода. Оно решается путем регуляризации, в котором метрика определяется через специальную эмпирическую V-матрицу [1], адаптированную к выборке.

2.2. Оценка через минимизацию эмпирического риска

Введем функцию $w(x,y)=p(x,y)/(p(x)p(y))$. Взаимная информация запишется в виде

$$I(X,Y)=\iint p(x,y)\ln w(x,y)dxdy.$$

Оценку функции $w(x,y)$ будем искать путем минимизации функционала

$$J_0(\hat{w})=\frac{1}{2}\iint\left(w(x,y)-\hat{w}(x,y)\right)^2p(x)p(y)dxdy,$$

который имеет смысл среднего риска оценивания функции $w(x,y)$. Эквивалентно

$$J_0(\hat{w})=\frac{1}{2}\iint\hat{w}^2(x,y)p(x)p(y)dxdy-\iint\hat{w}(x,y)p(x,y)dxdy+C,$$

константа C не зависит от оценки $\hat{w}(x,y)$. Эмпирическая оценка среднего риска по выборке пар экспериментальных данных $(x_1,y_1,\ldots,x_n,y_n)$ – функционал эмпирического риска имеет вид

$$J_e(\hat{w})=\frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\hat{w}^2(x_i,y_j)-\frac{1}{n}\sum_{i=1}^{n}\hat{w}(x_i,y_i)+C.$$

Можно построить непараметрическую оценку $w(x,y)$, минимизируя в бесконечномерном функциональном пространстве регуляризованный функционал

(2) $$J_e(\hat{w},\lambda)=J_e(\hat{w})+\frac{\lambda}{2}\|\hat{w}\|_L^2,$$

где параметр λ>0, а $\|.\|_L$ обозначает норму в Гильбертовом пространстве L, в котором строится непараметрическая оценка. Добавление к функционалу эмпирического риска регуляризирующего члена обеспечивает единственность точки минимума и повышает устойчивость решения к случайным возмущениям в экспериментальным данным.

Если функционал (2) минимизируется в Гильбертовом пространстве с воспроизводящим ядром, то по теореме о представителе [2] приближение $\hat{w}$, минимизирующее функционал $J_e(\hat{w},\lambda)$ при фиксированном, λ представимо в виде

(3) $$\hat{w}(z)=\sum_{i=1}^{n}\alpha_i K(z,z_i),$$

где $z=(x,y)$, $z_i=(x_i,y_i)$, а неотрицательно определенная функция $K(z,t)$ – ядро, которое соответствует скалярному произведению в пространстве L, коэффициенты α_i определяются путем минимизации функционала (2). В качестве ядра можно использовать любую неотрицательно определенную функцию.

2.3. Подбор оптимальных параметров

Для подбора параметров λ и σ, необходимых для построения оценки (3), используется процедура скользящего контроля:

- экспериментальная выборка делится на K частей Z_k, $k = 1,\ldots, K$;
- по выборке Z_{-k}, не содержащей часть Z_k, вычисляется оценка $\hat{w}_k(x,y)$ путем минимизации функционала (2) и на остальной части данных вычисляется значение функционала
$$\hat{J}_k(\lambda,\sigma)=\frac{1}{2n_k^2}\sum_{x_i\in Z_k}\sum_{y_j\in Z_k}\hat{w}_k^2(x_i,y_j)-\frac{1}{n_k}\sum_{(x_i,y_i)\in Z_k}\hat{w}_k(x_i,y_i);$$
- достигнутое значение функционала среднего риска при заданных параметрах λ и σ оценивается как среднее арифметическое по вычисленным значениям $\hat{J}_k(\lambda,\sigma)$;
- значения параметров, при которых последнее выражение достигает наименьшего значения принимаются за оптимальные.

3. Отбор признаков при классификации пентапептидов

3.1. Задача оценки устойчивости пентапептидной конформации

Одной из ключевых задач современной биологии является изучение физико-химических и функциональных свойств белков, которые определяются их аминокислотной последовательностью (первичной структурой). Предсказание того, какую пространственную структуру примет белок в процессе его фолдинга, важно, в частности, для разработки лекарственных средств, влияющих на функционирование биологических систем.

В [3] показано, что трехмерная структура белка адекватно описывается короткими пентапептидами (информационными единицами). Конформационно-устойчивые пентапептиды детерминируют формирование белками нативной пространственной структуры. Методами молекулярно-динамического моделирования было исследовано 49745 пентапептидов. Конформация каждого пентапептида определяется торсионными углами (ϕ, ψ, ω, χ^1,... χ^n). Процедура молекулярного моделирования представляет собой решение уравнений движения атомов пентапептидов. Взаимодействие между атомами описывалось потенциальным полем [4, 5]. Шаг интегрирования составлял 0.001 пикосекунды, а через 1 пикосекунду МД моделирования записывалось полученное конформационное состояние пентапептида. Для каждого рассмотренного пентапептида было получено 10000 конформационных состояний, из которых анализировались последние 5000, чтобы избежать влияния начального конформационного состояния. Выделенные 5000 конформаций каждого пентапептида, были кластеризованы по значениям торсионных углов полипептидного остова (ϕ, ψ). Некоторые информационные единицы в процессе моделирования многократно образовывали схожие пространственные структуры, оказавшиеся объединенными в многочисленный кластер. Такие пентапетиды бы-

ли названы конформационно-устойчивыми. Пентапетиды, рассчитанные пространственные структуры которых не образовали кластеры с большим числом элементов, считались неустойчивыми.

3.2. Классификация устойчивости пентапептидной конформации

Расчеты МД моделирования занимают существенное время и требуют значительных вычислительных ресурсов. Поиск устойчивых пентапептидов можно проводить используя методы классификации. Классификация проводилась методом ближайших соседей. Экспериментальные данные представляли собой таблицу, содержащую пятибуквенное обозначение пентапептида и частоту его попадания в самый многочисленный кластер в процессе молекулярно-динамического моделирования. При частоте большей 80% пентапептид считался устойчивым и относился к классу 1, остальные пентапептиды относились к классу 0. Доля представителей класса 1 в данных составляла 3.5%. Для целей классификации каждый пентапептид кодировался бинарным вектором в 100-мерном признаковом пространстве, содержащем 5 единиц. Координаты 1 соответствовали одному из 20 аминокислотных остатков в одной из 5 позиций в пентапептиде. Данные разбивались случайным образом на обучающую и тестовую выборки. По обучающей выборке оценивалась взаимная информация между каждым признаком и частотой попадания в самый большой кластер. Отбор признаков заключался в фильтрации по жесткому порогу величины оценки взаимной информации. Признаки с оценкой взаимной информации ниже порога исключались из расчетов.

При расчетах использовалась реализация метода ближайших соседей на языке R из пакета fastknn с оценкой близости, обратно пропорциональной расстоянию соседа от искомой точки. В таблице 1 приведен результат классификации с порогом отсечки по взаимной информации и количеством соседей, показавшими наилучшее значение метрики F1 на тестовой выборке. Значение 0.445 точности определения устойчивых пентапептидов является высоким с учетом малости их представления в данных.

Таблица 1. Результат классификации методом ближайших соседей.

Набор признаков	**Точность**	**Полнота**	**F1**
Все признаки	0.245	0.212	0.227
Информативные признаки	0.445	0.309	0.365

4. Заключение

Описанный метод отбора информационных признаков по оценке взаимной информации показал эффективность при отборе признаков в задаче оценки устойчивости информационных единиц. В отличие от методов, основанных на оценивании корреляционных показателей, этот метод успешно обнаруживает не только линейные, но и нелинейные закономерности. В рассматриваемой задаче с его помощью была снижена размерность пространства признаков, увеличена точность и скорость работы алгоритма машинного обучения, а так же повышена интерпретируемость результата.

Список литературы

1. Vapnik V., Izmailov R. Statistical inference problems and their rigorous solutions // Statistical Learning and Data Sciences. Berlin: Springer, 2015. P. 33-75.

2. Scholkopf B., Herbrich R., Smola A.J. A Generalized Representer Theorem // Computational Learning Theory. Berlin: Springer, 2001. P. 416-426.
3. Nekrasov A.N., Anashkina A.A., Zinchenko A.A. A new paradigm of protein structural organization // Theoretical Approaches to BioInformation Systems. Belgrade: Institute of Physics, 2014. P. 1-24.
4. Ponder J.W., Case D.A. Force fields for protein simulations // Adv. Protein Chem. 2003. Vol. 66. P. 27-85.
5. Case D.A., Cheatham T., Darden T. Amber biomolecular simulation programs // The. J. Comput. Chem. 2005. Vol. 26, P. 1668-1688.