Многомодальный метод релевантных векторов в задаче распознавания вторичной структуры бкъелка

Сунгуров Д.С.
Московский физико-технический институт
Кафедра «Интеллектуальные системы»

Научный руководитель
Д.т.н., профессор В.В. Моттль

2013

Представление белка

1. Первичная структура $\omega = \left(\alpha_t, t = 1,..., M_\omega\right)$ - конечная аминокислотная последовательность длины M_ω, где $\alpha_t \in A = (\alpha^1,..., \alpha^m), m = 20$ - символы алфавита аминокислот
2. Вторичная структура: $y = (y_t, t = 1,..., M_\omega), y_t \in Y = \{h, s, l\}$, где h, s, l - три возможных типа элементов вторичной структуры

Задача предсказания вторичной структуры белка

По обучающей выборке $\{(\omega_l, y_l) = l = 1,..., N\}$ для всякой новой аминокислотной последовательности $\omega = \left(\alpha_t, t = 1,..., M_\omega\right)$ научиться предсказывать вторичную структуру соответствующего белка $\hat{y} = (\hat{y}_t, \mathrm{t} = 1,..., M_\omega)$.

Существующие методы предсказания вторичной струкутры

1. Общепринятый принцип прогнозирования вторичной структуры белка (**H**elix, **S**trand, **C**oil) заключается в оценивании структуры по локальному фрагменту аминокислотной цепи (симметричному окну).
2. Искомая классификация аминокислотного окна неизбежно должна быть результатом его сравнения с множеством «эталонных» окон из обучающей совокупности.
3. Известные методы прогнозирования вторичной структуры:
   – базируются на целом ряде чисто биологических предположений,
   – используют эталонный набор из десятков тысяч фрагментов.

Существующие методы предсказания вторичной струкутры

4. Общепринятый принцип прогнозирования вторичной структуры белка (**H**elix, **S**trand, **C**oil) заключается в оценивании структуры по локальному фрагменту аминокислотной цепи (симметричному окну).
5. Искомая классификация аминокислотного окна неизбежно должна быть результатом его сравнения с множеством «эталонных» окон из обучающей совокупности.
6. Известные методы прогнозирования вторичной структуры:
   – базируются на целом ряде чисто биологических предположений,
   – используют эталонный набор из десятков тысяч фрагментов.

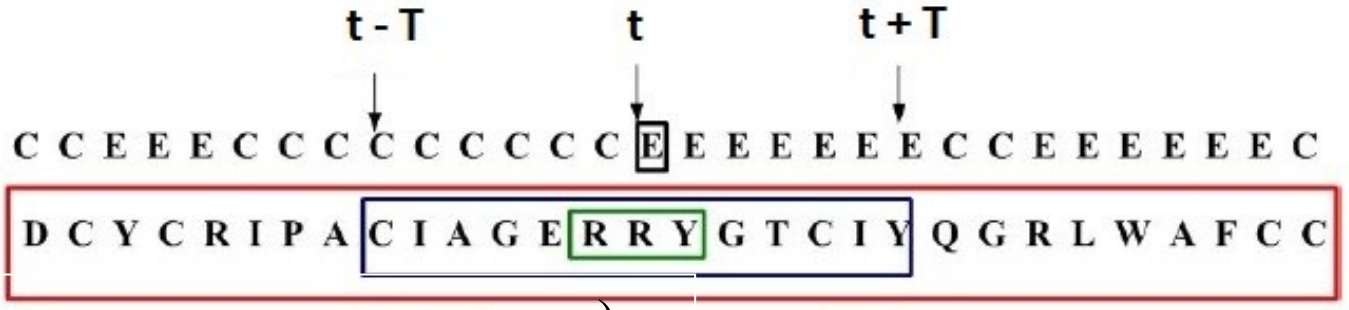
Основная идея предлагаемого подхода

1. Рассматривать задачу классификации аминокислотного фрагмента как задачу распознавания без привлечения биологических предположений.
2. Предлагаемый аппарат обучения опирается на большое множество «пробных» функций парного сравнения аминокислотных окон, из которого автоматически отбирается небольшое число «релевантных» функций сравнения.
3. Разработана новая версия метода релевантных векторов (RVM), названная многомодальным методом релевантных векторов (Multimodal RVM).

Метод скользящего окна

Решение о классе элемента вторичной структуры на позиции t делается на основе симметричного интервала $\omega = (\alpha_t, t - T \le t \le t + T)$ вырезанного из всей аминокислотной последовательности $\omega = (\alpha_t, t = 1, ..., M_\omega)$.

Мы ограничиваемся здесь распознаванием двух классов структуры
$(S, \bar{S}) = (1, -1)$ – Strand, Non-strand



(исходная аминокислотная последовательность)

Исходная задача сократилась до последовательности независимых задач определения класса вторичной структуры (1,-1) аминокислоты в очередном окне.

Принцип прогнозирования структуры аминокислотного окна

Аминокислотное окно: $\omega = \left(\alpha_{t-T} \cdots \qquad \cdots \right.$

Вторичная структура: $\boxed{y_t} \in \{-1, 1\}$

Обучающая совокупность аминокислотных окон: $\left\{ (\omega_j, y_j),\ j = 1,..., N \right\}$

Искомое решающее правило для нового окна:

$$d(\omega \mid \mathbf{a}, b) = \sum_{j=1}^{N} a_j \underbrace{S(\quad)}_{\text{функция парного сравнения}} - b \begin{cases} > 0 \to \hat{y}(\omega) = 1 \\ < 0 \to \hat{y}(\omega) = -1 \end{cases}$$

неизвестные коэффициенты

Классическая идея релевантных векторов (Relevance Vector Machine – RVM): Найти a_j, оставив ненулевыми $a_j \neq 0$ только коэффициенты при «нужных» (релевантных) объектах обучающей совокупности.

Многомодальный метод релевантных векторов: множество функций парного сравнения

$$d(\omega \mid \mathbf{a}, b) = \sum_{i \in I} \sum_{j=1}^{N} a_{ij} S_i(\omega_j, \omega) + b \begin{cases} > 0 \to \hat{y}(\omega) = 1 \\ < 0 \to \hat{y}(\omega) = -1 \end{cases}$$

множество объектов обучающей совокупности и множество функций парного сравнения

Найти a_j, оставив $a_j \neq 0$ только коэффициенты при релевантных объектах и релевантных функциях (**модальностях**) сравнения

Многомодальный метод релевантных векторов

$$\begin{cases} \sum_{i=1}^{n}\sum_{l=1}^{N}\left[(1-\mu)a_{il}^2+\mu|a_{il}|\right]+C\sum_{j=1}^{N}\delta_j\to\min(a_{il},b,\delta_j), \\ y_j\left(\sum_{i=1}^{n}\sum_{l=1}^{N}a_{il}S_i(\omega_l,\omega_j)+b\right)\geq 1-\delta_j,\ \delta_j\geq 0,\ j=1,\ldots,N. \end{cases}$$

Параметр селективности $\mu > 0$. Чем больше μ, тем больше число нулевых коэффициентов $a_{ij}=0$

Итоговая дискриминантная гиперплоскость:

$$d(\omega)=\sum_{ij\in\hat{F}}a_{ij}S_i(\omega_j,\omega)+b\gtrless 0,$$

$\hat{F}$ – подмножество релевантных вторичных признаков; F – полное множество вторичных признаков.

$\hat{J}=\{j:\exists i(a_{ij}\neq 0)\}\subseteq J=\{1,\ldots,N\}$	подмножество релевантных аминокислотных окон в обучающей совокупности
$\hat{I}=\{i:\exists j(a_{ij}\neq 0)\}\subseteq I=\{1,\ldots,n\}$	подмножество релевантных функций парного сравнения окон

В нашей задаче прогнозирования вторичной структуры белка: $n=6$ шесть функций парного сравнения аминокислотных окон

$S_1(\omega',\omega'')$ $S_2(\omega',\omega'')$ $S_3(\omega',\omega'')$	классический позиционный метод сравнения аминокислотных окон	$S_4(\omega',\omega'')$ $S_5(\omega',\omega'')$ $S_6(\omega',\omega'')$	вновь разработанный метод на основе пребразования Фурье

Математическая структура функций парного сравнения

Еще раз шесть функций парного сравнения:

$S_1(\omega',\omega'')$ $S_2(\omega',\omega'')$ $S_3(\omega',\omega'')$	классический позиционный метод сравнения аминокислотных окон	$S_4(\omega',\omega'')$ $S_5(\omega',\omega'')$ $S_6(\omega',\omega'')$	вновь разработанный метод на основе пребразования Фурье

Сравниваемые окна ω' и ω''

Векторы признаков $\mathbf{z}(\omega')$ и $\mathbf{z}(\omega'')$ – разные в позиционном и тригономет. сравнениях

Математические способы сравнения векторов признаков

$$\left.\begin{matrix} S_1(\omega',\omega'') \\ S_4(\omega',\omega'') \end{matrix}\right\} = \mathbf{z}^T(\omega')\mathbf{z}(\omega'') \text{ – классическое скалярное произведение}$$

$$\left.\begin{matrix} S_2(\omega',\omega'') \\ S_5(\omega',\omega'') \end{matrix}\right\} = \exp\left(-\gamma\left\|\mathbf{z}(\omega')-\mathbf{z}(\omega'')\right\|_2^2\right) \text{ – радиальная потенц. функция с нормой } l_2$$

$$\left.\begin{matrix} S_3(\omega',\omega'') \\ S_6(\omega',\omega'') \end{matrix}\right\} = \left\|\mathbf{z}(\omega')-\mathbf{z}(\omega'')\right\|_1 \text{ – норма } l_1$$

Функции парного сравнения аминокислотных окон на основе преобразования Фурье

Способ основан на представлении каждой из 20-ти аминокислот
$A = \{\alpha_1, ..., \alpha_{20}\}$
20-мерным вектором ее частотных признаков согласно теории PAM – Point Accepted Mutation (Margaret Dayhoff, В.В. Сулимова).
$\alpha^k \to \mathbf{a}^k \in \mathbb{R}$

Аминокислотное окно:

$$\begin{array}{c} \\ 1 \\ k \downarrow \\ 20 \end{array} \begin{array}{c} \begin{array}{cccc} -T & \to & s & \to \quad T \end{array} \\ \left[\begin{array}{cccc} a^1_{t-T} & \cdots & & \cdots \\ \vdots & & \vdots & \vdots \\ a^{20}_{t-T} & \cdots & & \cdots \end{array}\right. \\ \begin{array}{ccc} \mathbf{a}_{t-T} & \mathbf{a}_t & \mathbf{a}_{t+T} \end{array} \end{array}$$

Каждая k–я строка рассматривается как скалярный сигнал $a^k_s, s = -T, \ldots \quad \ldots$
Этот преобразуется в амплитудный спектр Фурье с пятью гармониками $[z^k_i, i = 0, \ldots$

Игнорирование фазовых компонент оставляет спектральное представление окна сильно-чувствительным к последовательности аминокислот в нем $(\alpha_{t-T} \cdots \quad \cdots \qquad$, но делает его малочувствительным к сдвигам окна t.

Общий вектор признаков окна: $\mathbf{z}(\omega) = \left[z_{il}(\omega), i = 1, ..., 20, l = 0, ..., 4\right] \in \mathbb{R} \qquad \mathbb{R}$

Численное решение задачи обучения

Еще раз исходная задача многомодального метода релевантных векторов:

$$\begin{cases} \sum_{i=1}^{n}\sum_{l=1}^{N}\left[(1-\mu)a_{il}^2+\mu|a_{il}|\right]+C\sum_{j=1}^{N}\delta_j\to\min(a_{il},b,\delta_j), \\ y_j\left(\sum_{i=1}^{n}\sum_{l=1}^{N}a_{il}S_i(\omega_l,\omega_j)+b\right)\ge 1-\delta_j,\ \delta_j\ge 0,\ j=1,\ldots,N. \end{cases}$$

Задача выпуклого программирования

Двойственная задача выпуклого программирования:

$$\begin{cases} W(\lambda_1,\ldots,\lambda_N|\mu)=\sum_{j=1}^{N}\lambda_j- \\ \dfrac{1}{4(1-\mu)}\sum_{i=1}^{n}\sum_{l=1}^{N}\left\{\min\begin{bmatrix}\mu+\sum_{j=1}^{N}y_j\lambda_j x_{il,j}\\ 0\\ \mu-\sum_{j=1}^{N}y_j\lambda_j x_{il,j}\end{bmatrix}\right\}^2\to\max, \\ \sum_{j=1}^{N}y_j\lambda_j=0,\ 0\le\lambda_j\le C,\ j=1,\ldots,N. \end{cases}$$

$$\begin{cases} \hat{a}_{il}=\dfrac{1}{2(1-\mu)}\left(\sum_{j:\hat{\lambda}_j>0}y_j\hat{\lambda}_j x_{il,j}+\mu\right), \\ \qquad\qquad \sum_{j:\hat{\lambda}_j>0}y_j\hat{\lambda}_j x_{il,j}<-\mu, \\ \hat{a}_{il}=0,\ -\mu\le\sum_{j:\hat{\lambda}_j>0}y_j\hat{\lambda}_j x_{il,j}\le\mu, \\ \hat{a}_{il}=\dfrac{1}{2(1-\mu)}\left(\sum_{j:\hat{\lambda}_j>0}y_j\hat{\lambda}_j x_{il,j}-\mu\right), \\ \qquad\qquad \sum_{j:\hat{\lambda}_j>0}y_j\hat{\lambda}_j x_{il,j}>\mu, \end{cases}$$

Число переменных равно числу аминокислотн. окон в обучающей совокупности

Результат обучения :

$$d(\omega)=\underbrace{\sum \ldots a_{il}S_i(\omega_l,\omega)}_{\text{только релевантные окна и функции сравнения}}+b\begin{cases}strand & >0\\ не\ strand & <0\end{cases}$$

Требования к алгоритму

- Алгоритм поиска должен легко параллелиться
- Годятся алгоритмы делающие небольшое количество итераций
- Каждая итерация должна легко параллелиться
- По этим причинам мы решили использовать метод внутренней точки для поиска максимума исследуемой функции

Требования к алгоритму

- Алгоритм поиска должен легко параллелиться
- Годятся алгоритмы делающие небольшое количество итераций
- Каждая итерация должна легко параллелиться
- По этим причинам мы решили использовать метод внутренней точки для поиска максимума исследуемой функции

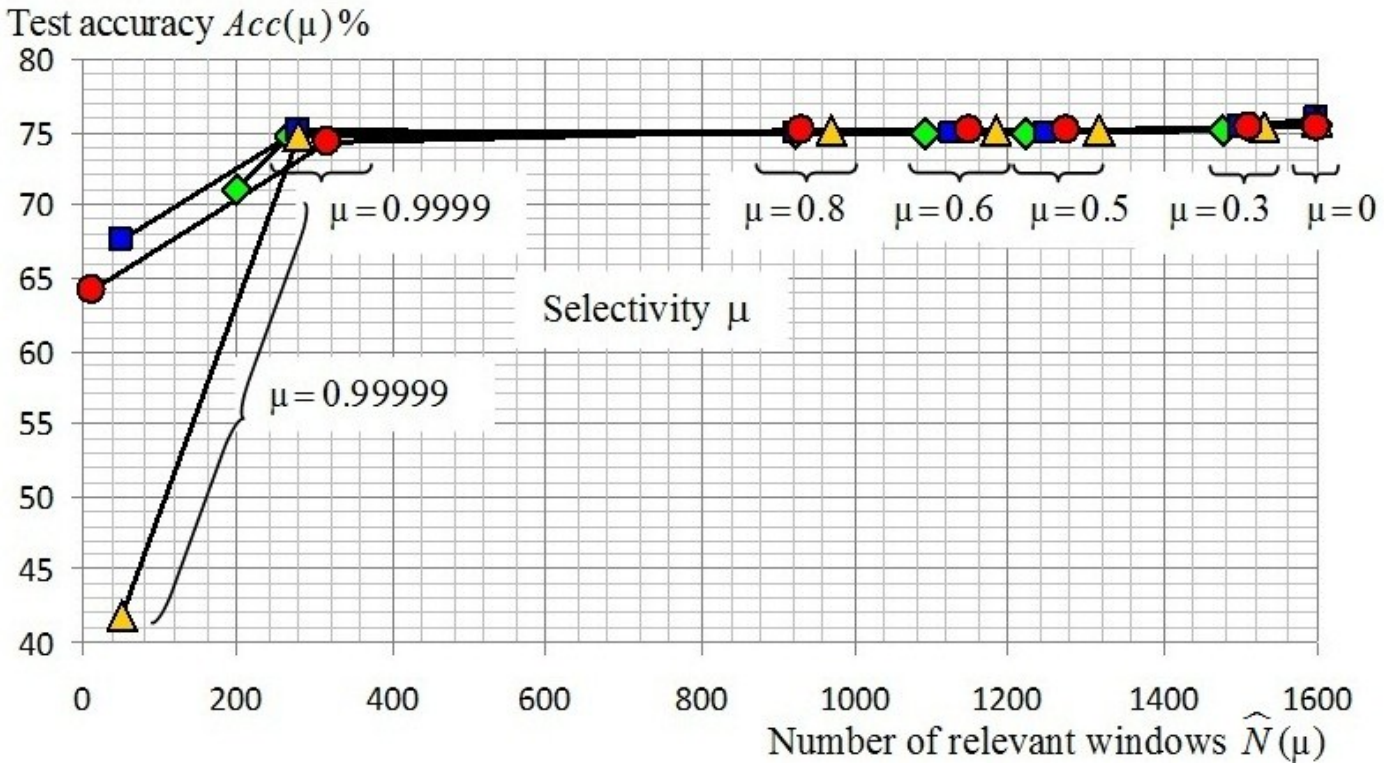
Особенности использования метода внутренней точки

- Две тяжелые операции: перемножение матриц и решение системы линейных уравнений.
- Мы распараллелили умножение матриц на CUDA.
- Решение системы линейных уравнений распараллелить не удалось.
- Небольшое количество итераций: 40 – 50.
- Метод внутренней точки является методом второго порядка, т.е требует держать в памяти матрицы$(N \times N)$, где N это количество переменных в двойственной задаче.

Эксперимент

- Мы использовали набор данных RS126 который содержит 126 белков представленных первичной и вторичной структурой. Белки в данном наборе имеют схожесть менее 25% для фрагментов более 80 аминокислот.
- Средняя длина белка в наборе 300-400 аминокислот.
- Проблема распознавания стрэндов во вторичной структуре решалась как задача двух классовой классификации.
- Мы использовали длину окна в $2T+1=35$. При этой длине окна из данных было получено 19075 аминокислотных фрагментов.
- Использовалось 6 различных функций парного сравнения.
- Данные разбивались случайным образом на обучение размера $|\Omega_{tr}|=1600$ и контроль $|\Omega_{test}|=17475$.
- Обучение проводилось для 7 различных значений параметра селективности: $\mu=0$, $\mu=0.3$, $\mu=0.5$, $\mu=0.6$, $\mu=0.8$, $\mu=0.9999$ and $\mu=0.99999$.
- Всякий раз вычислялась доля ошибок прогнозирования вторичной структуры на контрольной совокупности.

Результаты экспериментов



- Точность определения стрендов остается на одном уровне в 75% для всех значений селективности от $\mu \geq 0$ до $\mu = 0.9999$.
- С селективностью $\mu = 0.9999$, только $\hat{n} = 3$ функции сравнения и $\hat{N} = 300$ из 1600 объектов обучающей выборки были признаны релевантными.
- Дальнейшее увеличение селективности $\mu = 0.99999$ влечет сильную потерю в точности.

Заключение

- В этой работе был применен многомодальный метод релевантных векторов, показавший свою эффективность.
- Количество эталонных объектов используемых в решающем правиле в 5 раз меньше обучающей совокупности.
- Только 3 из 6 функций парного сравнения используются в решающем правиле.
- Средняя точность распознавания стрендов оказалась равной примерно 75%, что находится на уровне уже существующих алгоритмов
- Особенно интересно то, что не было замечено переобучения, несмотря на то, что размерность векторов вторичных признаков в несколько раз выше размера обучающей совокупности.