МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Российская академия наук
Московский физико-технический институт
(государственный университет)
Российский фонд фундаментальных исследований
Федеральная целевая программа
«Научные и научно-педагогические кадры инновационной России»
на 2009–2013 годы
Фонд содействия развитию малых форм предприятий
в научно-технической сфере

ТРУДЫ
53-й НАУЧНОЙ КОНФЕРЕНЦИИ МФТИ

Современные проблемы фундаментальных и прикладных наук

Часть I
Радиотехника и кибернетика

Том 1

Москва–Долгопрудный
МФТИ
2010

УДК 004:51:621.3:537.8
ББК 32.97
Т78

Т78 **Труды 53-й научной конференции МФТИ «Современные проблемы фундаментальных и прикладных наук».** Часть I. Радиотехника и кибернетика. Том 1. — М.: МФТИ, 2010. — 178 с.
ISBN 978-5-7417-0322-9

В сборник включены результаты фундаментальных и прикладных исследований студентов, аспирантов, преподавателей и научных сотрудников МФТИ, а также ряда научных и учебных организаций. Они представляют интерес для специалистов, работающих в области вычислительных и инфокоммуникационных технологий, обработки и защиты информации, систем спутниковой связи, радиолокации.

УДК 004:51:621.3:537.8
ББК 32.97

ISBN 978-5-7417-0322-9

и классификатора. Использование такой схемы оказывается удобным по следующим причинам:

- рассматриваемая функция может иметь области с большой изменчивостью или разрывами. Выделение таких областей и построение на них локальных аппроксимаций позволяет существенно повысить точность модели;
- в конкретной задаче могут быть сформулированы разные требования к точности модели для разных областей. Например, может оказаться необходимым обеспечить высокую точность метамодели только на участке $A_\varepsilon = \{|F_M(\mathbf{X})| < \varepsilon,\ \mathbf{X} \in \mathrm{D} \subset \mathrm{R}^\mathrm{d}\}$. Выполнения этого требования можно добиться, если построить отдельную модель только для точек из области A_ε.

Таким образом, после того как определены требования к точности метамодели (определены области, где нужно обеспечить повышенную точность) и выделены области, где $F_M(\mathbf{X})$ имеет простую структуру, а также области с большой изменчивостью или разрывами, необходимо построить классификатор, который будет определять, к какой из областей отнести новую точку, а также свою аппроксимирующую функцию для каждой области. Для построения аппроксимирующих функций и классификатора нами были разработаны специальные алгоритмы.

Предложенная методология была успешно применена при решении задачи аппроксимации ограничений в задаче оптимизации формы обшивки пассажирского самолета.

Литература

1. *Burnaev E.V., Belyaev M.G., Prihodko P.V.* Approximation of multidimensional dependency based on an expansionn parametric functions from the dictionary // Proceedings of CDAM'2010 conference. — 2010.

2. *Burnaev E.V., Grihon S.* Construction of the metamodels in support of stiffened panel optimization // Proceedings of the conference Mathematical Methods in Reliability. — 2009. — P. 124–128.

УДК 519.688

О.А. Зверков, А.В. Селиверстов, В.А. Любецкий

zverkov@iitp.ru, slvstv@iitp.ru, lyubetsk@iitp.ru

Институт проблем передачи информации им. А.А. Харкевича РАН

Об одном алгоритме кластеризации белков

Предложен алгоритм кластеризации аминокислотных последовательностей белков, то есть слов различной длины в 20-буквенном алфавите, по их сходству. Алгоритм применён при исследовании пластомов водорослей и споровиков. Пусть задано семейство видов $\{S_i\}_{i=1\ldots N}$, в котором каждый вид представлен набором аминокислотных последовательностей его белков: $S_i = \{P_{ij}\}_{j=1\ldots M_i}$. Алгоритм вычисляет меру сходства $s_0(P_{ij},P_{kl})$ для всех пар белков из всех видов семейства как наименьший штраф за выравнивание пары белков (P_{ij},P_{kl}) и соответствующую нормированную меру $s(P_{ij},P_{kl})$ по формуле $2s_0(P_{ij},P_{kl})(s_0(P_{ij},P_{ij})+s_0(P_{kl},P_{kl}))^{-1}$. Рассмотрим полный неориентированный граф G_0 с множеством вершин $\{P_{ij}\}$ и сопоставим каждому ребру $e = \{P_{ij},P_{kl}\}$ значение $s(e) = s(P_{ij},P_{kl})$. Вместо полного графа можно использовать разреженный граф G, содержащий для каждого белка P_{ij} и каждого вида S_k ровно одно ребро $e_l = \{P_{ij},P_{kl}\}$ с максимальным значением величины $s(e_l)$, соответствующее ближайшему к P_{ij} в геноме S_k белку P_{kl}.

Алгоритм строит минимальный остовный лес F для графа G, используя в качестве веса ребра величину $w(e) = -s(e)$, то есть для каждой связной компоненты графа G строится покрывающее её дерево, такое, что сумма чисел, приписанных его рёбрам, максимальна. Для каждого дерева $T \in F$ выполняется следующая рекурсивная процедура разделения: если T не удовлетворяет сформулированному ниже критерию остановки, то из него удаляется наиболее слабое ребро, то есть ребро e_w с минимальным по дереву значением $s(e_w)$, а к каждому из двух полученных таким образом деревьев в свою очередь применяется процедура разделения. Критерий остановки для данного дерева $T = (V,E)$ следующий: (1) $|V| < pN$, где $|V|$ — число вершин дерева T, N — число всех видов семейства, а p — параметр алгоритма, выражающий максимально ожидаемую в исходных данных долю паралогов в кластерах (типично $1 < p < 2$); (2) $e_w = \{P_{ij},P_{kl}\}$, $i \neq k$, то есть самое слабое ребро e_w соединяет два белка различных видов; (3) любая пара вершин $\{P_{ij},P_{il}\}$ дерева T, соответствующих белкам

из одного вида, соединена в дереве T путём, состоящим только из вершин, соответствующих белкам этого вида. Полученный в результате лес представляет разбиение белков на кластеры, состоящие из вершин одного дерева.

Пример биологического результата — предсказание нового кластера у споровиков Piroplasmida. Соответствующие гены ups8 расположены между генами rpl14 и rps8, кодирующими рибосомные белки. Функциональная принадлежность белка, кодируемого ups8, неизвестна. Корректность выделения нового кластера белков подтверждается анализом 5'-лидерных областей генов ups8, где найдены консервативные сайты. Действительно, естественно ожидать, что консервативный сайт служит для регуляции экспрессии генов, а однотипная регуляция свидетельствует об общем функциональном значении соответствующих белков. Сайты имеют консенсус 5'kATAGAm3' и расположены на расстоянии от 170 до 100 нуклеотидов от инициирующего кодона ups8. У Babesia bovis и Babesia bigemina они располагаются внутри кодирующей области гена rpl14. Однако в окрестности сайта в белке произошла вставка (состава TSYSIDDRNRFKD у B. bovis), отсутствующая у ортологичных белков L14. У Theileria parva сайт не перекрыт кодирующими областями.

Работа выполнена при частичной финансовой поддержке Федерального агентства по образованию (Государственный контракт П2370).

УДК 681.327.12

В.В. Лихачев

vitaly.likhachev@phystech.edu

Московский физико-технический институт
(государственный университет)
Институт проблем передачи информации им. А.А. Харкевича РАН

Структурирование геоинформационных данных методами кластерного анализа

Одной из задач анализа геоинформационных данных является выделение в них однородных структур (кластеров). Широко применяемым методом кластеризации является Expectation–Maximization (EM) [1]. Обычно предполагается, что данные $\overrightarrow{x}$ представляют собой смесь многомерных нормально распределенных случайных величин, а число кластеров k задается [2].

В работе рассматриваются некоторые модификации алгоритма EM, которые учитывают расположение кластеров в географическом пространстве. На M-шаге алгоритма вычисляются новые центры кластеров $\overrightarrow{m_i}$ и их ковариационные матрицы S_i. Для элементов, входящих в кластер, определяется географический центр. На E-шаге вычисляются веса h_i^t, которые можно интерпретировать как вероятность принадлежности точки $\overrightarrow{x^t}$ i-тому кластеру: $h_i^t \sim |S_i|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\overrightarrow{x^t} - \overrightarrow{m_i})^T S_i^{-1}(\overrightarrow{x^t} - \overrightarrow{m_i})\right]$.

Можно предложить несколько вариантов учета географического расстояния: 1) географические координаты элементов кластера рассматриваются как дополнительный признак с некоторым коэффициентом λ: $\widetilde{h_i^t} \sim |S_i|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left\{\rho_f^2 + \lambda\frac{\rho_g^2}{\sigma_g^2}\right\}\right]$, где ρ_f — расстояние в признаковом пространстве, ρ_g — в географическом пространстве, $\sigma_g^2 = \sigma_x^2 + \sigma_y^2$, $\widetilde{h_i^t}$ — пересчитанное значение h_i^t; 2) сравнение с пороговым расстоянием ρ_t до центра кластера $\widetilde{h_i^t} = \begin{cases} h_i^t, \rho_g > \rho_t, \\ 0, \rho_g > \rho_t. \end{cases}$; 3) пороговое расстояние вводится до точек, принадлежащих этому кластеру на предыдущей итерации, тогда ρ_t играет роль максимального расстояния между соседними точками в кластере.