ОБНАРУЖЕНИЕ ЭМПИРИЧЕСКИХ ЗАКОНОМЕРНОСТЕЙ С ПОМОЩЬЮ ЭВМ
(Вычислительные системы)
1984 год Выпуск 102

УДК 519.95:681.3.06

ПОЛИГОН ДЛЯ СРАВНЕНИЯ АЛГОРИТМОВ ТАКСОНОМИИ

Н.Г.Загоруйко, В.Н.Ёлкина, Г.Л.Полякова

I. Для решения задач автоматической классификации в настоящее время имеется большое количество различных алгоритмов таксономии. Естественно, возникает потребность в их сравнении и выборе алгоритма, в некотором смысле "лучшего". Алгоритмы можно сравнивать между собой по требуемым машинным ресурсам (памяти и времени), по применимости к трудным случаям (большие массивы информации, разнотипные признаки, наличие помех или пробелов в таблицах данных и т.п.).

Однако главное, что интересует пользователя - "качество" получаемых решений. Чтобы сформулировать критерий качества, по которому можно было бы сравнивать алгоритмы таксономии, напомним, что таксономия обычно делается не просто для компактной перекодировки, для замены некоторого набора объектов Z', взятых из множества Z , небольшим числом их представителей.Результаты таксономии - таксоны - используются в дальнейшем в качестве эталонов, по которым распознаются новые объекты генеральной совокупности Z.

Если по этим эталонам распознать все объекты Z, то будет ли получившееся разбиение по классам совпадать с разбиением, которое можно получить таксономией сразу всех объектов множества Z? Если да, то,значит,алгоритм таксономии удачно угадал структуру множества Z по случайной выборке Z'. Эта способность по малой выборке Z' правильно угадывать структурные закономерности генеральной совокупности Z и есть, по-видимому, основная характеристика (φ) качества алгоритма таксономии.

Сравнивать алгоритмы таксономии по свойству φ можно с помощью построенного нами программного испытательного комплекса (полигона "Таксон" [I]). Прежде, чем описать его работу, введем некоторые понятия и обозначения.

2. Пусть $A = \{a_1, \ldots, a_j, \ldots, a_l\}$ – множество алгоритмов таксономии; $Z = \{Z_1, \ldots, Z_i, \ldots, Z_m\}$ – исследуемое множество ("генеральная совокупность") объектов, представленных в виде векторов в n-мерном признаковом пространстве X, т.е. $Z_i = (x_{i1}, \ldots, x_{ip}, \ldots, x_{in})$, Z' – случайная выборка объема L из Z, $Z' \subset Z$.

Под базовой таксономической структурой S_j^0 множества Z будем понимать разбиение Z с помощью алгоритма таксономии a_j на подмножества $\{S_{jt}^0\}$, где $t = \overline{1,k}$, $\bigcup_{t=1}^{K} S_{jt}^0 = Z$, $S_{jt}^0 \cap S_{jf}^0 = \emptyset$ при $t \neq f$. "Разумность", "естественность" классификации S_j^0 зависит от разумности и естественности критерия F_j, используемого алгоритмом таксономии a_j.

Таксономическую структуру множества Z' обозначим через S_j', а таксономическую структуру множества Z, которая получается с помощью распознавания всех объектов Z по эталонам S_j', назовем восстановленной таксономической структурой и обозначим через $S_j^{0'}$.

Показателем устойчивости таксономической структуры множества Z будем считать меру расхождения $\varphi(S_j^0, S_j^{0'})$ как функцию от хэммингова расстояния между соответствующими бинарными матрицами смежности $|r_{\alpha\beta}|$ и $|r'_{\alpha\beta}|$, где

$$r_{\alpha\beta}\,(r'_{\alpha\beta}) = \begin{cases} 1, & \text{если при разбиении } S_j\,(S_j^{0'}) \\ & \text{объекты } \alpha \text{ и } \beta \text{ принадлежат разным таксонам;} \\ 0, & \text{если при разбиении } S_j(S_j^{0'}) \text{ объекты } \alpha \text{ и } \beta \\ & \text{принадлежат одному таксону;} \end{cases}$$

$$\varphi(S_j^0, S_j^{0'}) = \frac{\sum_{\alpha,\beta}^{m} (r_{\alpha\beta} - r'_{\alpha\beta})^2}{m(m-1)} .$$

Разбиения S_j^0 и $S_j^{0'}$ будем считать δ-эквивалентными ($\overset{\delta}{\sim}$), если мера расхождения $\varphi(S_j^0, S_j^{0'}) \leq \delta$.

3. На полигоне для каждого алгоритма $a_j \in A$ и разных множеств Z определяются условия сохранения устойчивости таксономической структуры множества Z в зависимости от объема L выборки Z', размерности n признакового пространства X и количества таксонов в S_j^0.

При создании полигона предусматривалось, что множество Z может представлять собою либо реальные данные практической задачи,

либо генеральную совокупность, характеризующуюся заданным законом распределения. В связи с этим в полигоне заложены возможности генерирования множества Z с разными законами распределения.

4. Работа полигона организована следующим образом.

I. Задается или генерируется множество Z .

2. Алгоритмом $a_j \in A$ делается таксономия S_j^0 генеральной совокупности Z . Количество таксонов K либо фиксировано, либо выбирается автоматически из заданного диапазона по экстремальному значению функционала качества таксономии. Считаем, что разбиение S_j^0 отражает базовую структуру множества Z.

3. С помощью датчика случайных чисел, равномерно распределенных на [0,I], формируется множество Z' как выборка объема L из множества Z.

4. Осуществляем таксономию множества Z' тем же алгоритмом a_j при тех же условиях, что и в п.2.

5. Разбиение $S_j^{0'}$ генеральной совокупности Z строим следующим образом. Принимаем решение о принадлежности объектов Z_i , не попавших в $Z'(Z_i \in \{ Z \setminus Z' \})$, к таксонам S_j' , полученным при разбиении выборки Z'. При этом используем таксономические решающие функции [3].

Объект $Z_i \in \{Z \setminus Z'\}$ считаем представителем того таксона, присоединение к которому дает экстремальное значение критерия качества таксономии F_j . Так, при исследовании алгоритма КРАБ2 [3] в качестве решающей функции был взят алгоритм ТРФ-2 [3].

6. Определяем меру расхождения $\varphi(S_j^0, S_j^{0'})$ между базовым и восстановленным разбиениями множества Z .

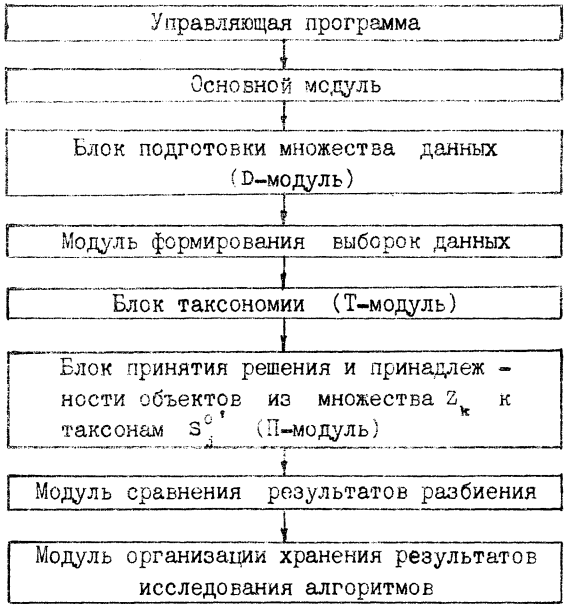
7. Для получения более объективной оценки $\varphi(S_j^0, S_j^{0'})$ при фиксированном объеме L повторяем пп. 3-6 заданное число (h) раз.

8. Устанавливаем зависимость меры расхождения $\varphi(S_j^0, S_j^{0'})$ от объема L выборки Z' путем повторения пп.3-7 при разном значении L .

9. Повторяем пп.3-8 при изменении размерности n признакового пространства. При этом определяем влияние размерности n признакового пространства X и объема L на устойчивость разбиения.

I0. Повторяем пп. 2-9 для разных множеств Z .

Блок-схема полигона представлена на рисунке. Группа D-модулей содержит модули генерирования множества данных с разными законами распределений и модуль, предназначенный для чтения и записи реальных данных. В настоящее время реализован модуль, генери -



рующий смесь n-мерных нормальных совокупностей, и модуль, предназначенный для работы с реальными данными. Реализована также возможность подключения программ формирования множества Z с разными законами распределения путем использования в D-модуле в качестве фактических аргументов имен этих подпрограмм.

Группа П-модулей содержит набор решающих правил, предназначенных для принятия решения о принадлежности $Z_i \in \{Z \setminus Z'\}$ к таксонам $S_j^{0'}$, полученным на выборке Z'. В настоящее время в этот набор включены алгоритмы группы ТРФ [3] для таксонов произвольной формы и таксонов, состоящих из набора гиперсфер.

Группа Т-модулей содержит наборы алгоритмов таксономии. В настоящее время в этот набор включены алгоритмы ФОРЭЛЬ, ФОРЭЛЬ-2 [3], SKAT, SKAT2 [4], КРАБ и КРАБ-2 [3].

Каждую группу модулей (D,Т,П) представляет соответствующий типовой модуль. Основной модуль состоит из "типовых" и постоянных модулей. К постоянным модулям относятся модули формирования выборок, сравнения и организации хранения результатов исследования алгоритмов.

При исследовании нового алгоритма таксономии программисту потребуется создать новый конкретный П-модуль для принятия решения о принадлежности объектов из множества Z к таксонам, полученным на выборке Z'. Согласование определений "типового" и конкретного модулей группы осуществляется путем задания соответствующих параметров в управляющей программе.

6. Результаты сравнения алгоритмов таксономии FOREL-2, KRAB-2, SKAT2.

Условия эксперимента.

Множество Z представляет собой смесь из K n-мерных нормальных совокупностей с неизвестными средними μ^t и ковариационными матрицами Σ_t, т.е. функция распределения имеет вид

$$f(Z) = \sum_{t=1}^{K} p_t N(Z, \mu^t, \Sigma_j), \qquad \sum_{t=1}^{K} p_t = 1,$$

p_t – априорная вероятность совокупности t.

Для моделирования множества Z было использовано предположение $\Sigma_t = \Sigma_f = E\sigma^2$, где t, f – номера таксонов, E – единичная матрица, σ^2 – дисперсия p-й компоненты вектора Z. Средние μ^t вычисляются в процессе моделирования множества Z при задании коэффициента пересечения C таксонов. Для нормальных распределений в n-мерном признаковом пространстве X коэффициент C определяется для каждой пары таксонов t, f с параметрами (μ^t, Σ_t) и (μ^f, Σ_f) уравнением вида $C = \dfrac{|\mu^t - \mu^f|}{\sigma^{(t)} + \sigma^{(f)}}$, где $|\mu^t - \mu^f|$ – евклидово расстояние между центрами таксонов, $2\sigma^t$, $2\sigma^f$ – длины принадлежащих таксонам t и f отрезков прямой, соединяющей центры этих таксонов. При заданном значении C вероятность линейной разделимости двух таксонов меньше $\exp(-C^2/2)$ [5]. При C = 2 таксоны генерируются изолированными и находятся на некотором расстоянии друг от друга, при C = 1 границы таксонов "касаются" друг друга.

В данном эксперименте количество m объектов в множестве Z было принято равным 1000. При моделировании было рассмотрено два варианта: 1) K = 2; $p_1 = 0{,}25$; $p_2 = 0{,}75$ – с коэффициентом пересечения C = 1 и C = 2; 2) K = 3; $p_1 = 0{,}25$; $p_2 = 0{,}25$; $p_3 = 0{,}5$ с коэффициентом пересечения также C = 1 и C = 2.

Объем L выборки Z' изменялся в диапазоне L = {10, 30, 50, 100, 300, 500}. Значения размерности признакового простанства вы-

бирались из множества n = {2, 3, 5, I0}. Для получения оценки $\varphi(S_j^0, S_j^{0'})$ при одном и том же множестве Z проводилось по I0 экспериментов (h = I0).

Для проведения таксономии на множестве Z и его подмножестве Z' используется одна и та же информация о числе таксонов (K). Значения K выбирались из множества {2,3,4,5,6,7}.

При испытании алгоритма FOREL-2 каждый вариант структуры генеральной совокупности Z разбивался последовательно на K таксонов (K = 2 - 7) . На это же число K таксонов разбивалась выборка $Z' \subset Z$. Затем, используя линейное решающее правило, распознавались объекты, не попавшие в Z' , и восстанавливалась структура $S_j^{0'}$ множества Z . Хорошее совпадение ($\varphi = 0$ при C = 2 и $\varphi \leq 0{,}06$ при C = I) базовой и восстановленной таксономических структур было получено, когда количество таксонов, задаваемых на Z и Z', совпадало. Вопрос автоматического выбора алгоритмом FOREL-2 предпочтительного числа таксонов не рассматривался.

Алгоритм KRAB-2 позволяет на основе критерия качества таксономии F выделить из заданного диапазона предпочтительное количество таксонов. Результаты испытания этого алгоритма на множествах Z , состоящих из 2 или 3 таксонов оказались следующими: I) при C = 2 мера расхождения между разбиениями S_j^0 и $S_j^{0'}$, $\varphi(S_j^0, S_j^{0'}) = 0$ и не зависит от объема L выборки Z' в диапазоне L = (0,0I-0,5)m; 2) при C = I для достижения меры расхождения $\varphi(S_j^0, S_j^{0'}) \leq 0{,}06$ необходимо, чтобы объем выборки L был не меньше 0,2 m.

Алгоритм SKAT2 правильно определяет количество таксонов из заданного диапазона значений при всех условиях генерирования множества Z и правильно восстанавливает структуру генеральной совокупности $\varphi(S_j^0, S_j^{0'}) \leq 0{,}06$ при L = (0,2-0,3)m и C = 1, а при C = 2, δ=0 для всех значений L , в диапазоне от 0,0Im до 0,5m.

В данных экспериментах не было обнаружено влияния на результаты размерности n признакового пространства X .

Литература

I. ЗАГОРУЙКО Н.Г., ЁЛКИНА В.Н., ПОЛЯКОВА Г.Л. Один из подходов к созданию полигона для сравнения алгоритмов таксономии (автоматической классификации).-Тез. докл. Всесоюз. конф. "Теория классификаций и анализ данных", 5-8 мая, Новосибирск, I98I.

2. ПОЛЯКОВА Г.Л. Структура программного обеспечения для сравнения алгоритмов таксономии (автоматической классификации). -Тез.

докл. региональной конф. "Вычислительная техника и дискретная математика", 22-23 апреля, Новосибирск, 1983.

3. ЗАГОРУЙКО Н.Г. Методы распознавания и их применение.-М.: Сов.радио, 1972.

4. ЁЛКИНА В.Н., ПОЛЯКОВА Г.Л. Процедура восстановления структуры множества данных.-Тез. докл. Второй Всесоюз. конф. "По применению математических методов и ЭВМ в почвоведении", 17-19 ноября, Пущино, 1983.

5. КРАМЕР Г. Математические методы статистики. -М.: Мир,1975.

Поступила в ред.-изд.отд.
29 июня 1984 года