УДК 621.391.234

Ю. П. ДРОБЫШЕВ, С. П. СОКОЛОВ

(*Новосибирск*)

МИНИМИЗАЦИЯ БОЛЬШИХ МАССИВОВ ДАННЫХ

В настоящее время наряду с такими проблемами, как автоматизация научных исследований, оптимизация научного эксперимента, особую остроту приобретает проблема оптимальной обработки больших массивов уже накопленной или получаемой в ходе того или иного эксперимента информации. Важным этапом последней проблемы является сокращение объема (описания) исходных информационных массивов. В общем случае минимизация описания сопряжена с неизбежными потерями информации, поэтому при решении конкретной практической задачи следует исходить из разумного компромисса между получаемой точностью воспроизведения исходной информации и достигаемой экономией. При этом следует учитывать и сложность расшифровывающего алгоритма. Существуют два основных подхода к проблеме минимизации.

1. Понижение мощности исходного множества G функционального пространства F путем аппроксимации подмножеств, достаточно близких по своим свойствам элементов, отдельными элементами. Если элемент множества G $f \in G_1 = \{Y/\rho_F(\psi_0, Y) \leqslant \varepsilon\} \subset G$, то он аппроксимируется элементом $\psi_0 \in G_2 \subset G$, имеющим мощность $N_{G2} < N_G$. Здесь $\rho_F(\psi_0, Y)$ — расстояние в пространстве F, ε — заданная точность аппроксимации. Таким образом, задача сводится к построению оптимальной (содержащей минимальное число элементов) ε-сети.

2. Понижение размерности пространства. Пусть задано конечное множество некоторого функционального пространства F или ε-сеть непрерывного компактного пространства. Задача заключается в отыскании такого подпространства фиксированной размерности $F_0 \subset F$ (соответствующего ортопроектора L_0 для элементов множества G) и соответствующего множества $G_0 \subset F_0$, элементы которого наилучшим образом (в смысле заданного критерия близости элементов в метрике пространства F) аппроксимируют элементы исходного множества G.

В том случае, когда фиксируется точность аппроксимации ε, искомое подпространство должно иметь минимальную размерность.

Оператор L_0 в зависимости от решаемой задачи может быть подчинен следующим условиям.

Прямая задача:

$$\max_{f \in G} Y^0(L_0 f, f) \leqslant \varepsilon \quad \text{или} \quad M[Y^0(L_0 f, f)] \leqslant \varepsilon'.$$

При этом размерности соответствующих подпространств минимальны.

Обратная задача:

$$\min_{L \in l} \left[\max_{f \in G} Y^0(Lf, f)\right] = \max_{f \in G} [Y^0(L_0 f, f)],$$

или

$$\min_{L \in l} \{M[Y^0(Lf, f)]\} = M[Y^0(L_0 f, f)], \quad f \in G,$$

при фиксированных размерностях подпространств. Здесь $Y^0(L, f, f)$ — заданный функционал, определяющий меру близости; M — символ операции усреднения по множеству G; ε и ε' — погрешности аппроксимации. Таким образом, задача минимизации исходного описания путем снижения размерности пространства сводится к отысканию оптимального базиса.

В зависимости от свойств множества G наилучший эффект будет достигаться за счет снижения его мощности или размерности исходного пространства F или того и другого вместе.

В общей постановке задача минимизации исходного описания сводится к отысканию такого подмножества G_0 в некотором подпространстве $F_0 \subset F$, что число запоминаемых при заданной точности аппроксимации числовых характеристик должно быть минимально:

$$N_0 = N_{G_0} n_{F_0} = \min_{G' \subset F' \subset F} \{N_{G'} n_{F'}\},$$

где N_{G_0} — мощность множества G_0, n_{F_0} — размерность подпространства F_0*.

В такой общей постановке решение задачи затруднительно. Целесообразно рассмотреть частные постановки.

Понижение мощности исходного множества G (синтез оптимальных ε-сетей). Задача построения оптимальных внутренних ε-сетей множества G функционального пространства F редуцируется к задаче отыскания минимальных (содержащих минимальное число элементов) покрытий этого множества элементами некоторого заданного множества G' (G'-покрытий).

Действительно, если в качестве i-го элемента множества G' выбрать сферу радиуса ε с центром в точке $\alpha_i \in G$, то задача построения внутренней ε-сети множества G становится эквивалентной задаче отыскания одного из минимальных G'-покрытий. Для отыскания покрытий удобно использовать матрицу C бинарных отношений между элементами множеств G и G'. Элемент C_{ij} этой матрицы равен 1 или 0 в зависимости от того, содержит i-й элемент множества G' j-й элемент множества G или нет. Каждому G'-покрытию множества G будет соответствовать некоторое подмножество строк матрицы C, объединение которых имеет единицы во всех столбцах.

Задача отыскания минимальных покрытий может решаться средствами линейного программирования, хотя это и сопряжено с известными трудностями, возникающими при отыскании целочисленных решений [1, 2].

Более эффективным представляется путь, основанный на прямом поиске методом сокращенного перебора, некоторые идеи которого были развиты в работах [3—5], а наиболее полно сформулированы А. Д. Закревским [6].

При реализации метода Закревского сначала отыскивается одно из G'-покрытий, а затем осуществляется поиск минимального из возможных покрытий. В основе метода лежат две теоремы. Первая из них в нашем случае позволяет перейти от матрицы C к матрице C_1 путем исключения тех строк, которые полностью входят в остающиеся. Вторая позволяет применять пошаговое понижение размерности исходной задачи путем разложения ее на последовательность аналогичных задач меньшей размерности. Указанные две операции образуют цикл алгоритма.

А. Отыскание одного из G'-покрытий.

Назовем строку и столбец матрицы, имеющие общую 1, связанными, а строки и столбцы, содержащие максимальное (минимальное) число единиц, соответственно максимальными (минимальными) строками и столбцами. Пусть со строкой l_i связано множество m_i столбцов. Если для какой-либо пары строк l_i и l_j $m_i \subset m_j$, то строку l_i назовем поглощаемой. Тогда процесс отыскания какого-либо из G'-покрытий может быть описан следующим образом.

---

* Заметим, что методы статистического оптимального кодирования здесь не рассматриваются, поскольку они предназначены для повышения средней скорости передачи сигналов по каналам связи, а не для минимизации описания ансамбля.

1. Из исходной матрицы C формируется матрица C_1 путем исключения всех поглощаемых строк.

2. Выбирается максимальная строка l_1, связанная с минимальным столбцом матрицы C_1.

3. От матрицы C_1 переходим к C_2 путем исключения строки l_1, всех связанных с ней столбцов и поглощаемых после этого строк.

4. Описанный выше процесс повторяется для матрицы C_2 и таким образом отыскивается второй элемент покрытия l_2.

5. Процесс продолжается до тех пор, пока не будут исчерпаны все столбцы (элементы множества G), т. е. получено покрытие, состоящее из строк $l_1, l_2, \ldots, l_s$.

Б. Отыскание минимального покрытия.

1. Из покрытия $\{l_1, l_2, \ldots, l_s\}$ исключаются три последние строки l_s, l_{s-1}, l_{s-2}. В соответствующей матрице C_{s-2} отыскиваются такие две строки, объединение которых содержит единицы во всех столбцах матрицы C_{s-2}. Если таких двух строк нет, то переходим к матрице C_{s-3}, в которой ищем уже объединение не более 3 строк, и т. д.

2. Если же в матрице C_{s-2} были найдены искомые строки l'_{s-2}, l'_{s-1}, то в матрице C_{s-3} ищем объединение опять только 2 строк. В случае отсутствия таковых переходим к матрице C_{s-4}, где ищем комбинацию не более 3 строк, связанную со всеми столбцами, и т. д.

3. Вообще, в каждом узле дерева перебора ведется поиск комбинации такого числа строк, которое, по крайней мере, на единицу меньше числа строк, исключенных из минимального среди построенных покрытий. Процесс заканчивается после того, как будет исследована начальная точка дерева. Последнее из найденных покрытий будет минимальным.

Описанный алгоритм был реализован на БЭСМ-6. Программа написана на языке АЛГОЛ-БЭСМ. Длина ее 4400_8 команд. При обработке больших информационных массивов (десятки тысяч элементов множества G) размеры матриц C_i не позволяют хранить их в оперативной памяти. Поэтому в реальных задачах приходится прибегать к оптимизации массива по частям. Разработанная программа была применена для сокращения массива электронных спектров органических соединений в одной из информационно-поисковых систем молекулярной спектроскопии. Минимизация проводилась циклами по 300 элементов. Результаты по одному циклу: сжатие исходного массива 3÷6 раз при 5%-й ошибке по критерию максимального абсолютного отклонения, время работы программы около 1,5 мин.

Понижение размерности пространства (синтез оптимального базиса). Минимизация исходного описания посредством понижения размерности пространства F сводится к построению в нем оптимального базиса, который должен:

а) обеспечивать минимальную ошибку аппроксимации при фиксированной размерности или минимальную размерность при заданном дефекте аппроксимации;

б) иметь, по возможности, инвариантную структуру, т. е. изменение требуемой величины ошибки аппроксимации должно сопровождаться добавлением или исключением из базиса нескольких последних компонентов без пересчета остальных.

Известно, что для среднеквадратичного критерия приближения оптимальный базис состоит из собственных функций интегрального оператора, ядром которого является корреляционная функция исходного описания, или определяется собственными векторами соответствующей матрицы.

Анализ показывает, что из всех вычислительных процедур, предназначенных для построения оптимального базиса, наиболее предпочтительной с точки зрения простоты и лаконичности вычислительного

процесса является процедура, предложенная А. М. Заездным и Г. В. Эйдукявичюсом [7].

Существо этой процедуры состоит в представлении элементов исходного множества функций G и функций искомого ортонормированного базиса A^* в некотором вспомогательном ортонормированном базисе A, предполагаемом конечным, и в определении собственных векторов матрицы, сформированной из комбинаций коэффициентов множества функций G в этом представлении, компоненты которых суть коэффициенты разложения функций оптимального базиса A^*.

Действительно, коэффициенты представления j-й функции в оптимальном базисе $\{a^*_{jk}\}$ и вспомогательном $\{a_{jk}\}$ связаны соотношением

$$a^*_{jk} = \sum_{n=1}^{N} X_{kn} a_{jn},$$

где X_{kn} — коэффициенты представления k-й функции базиса A^* в базисе A; N — мощность множества функций, образующих базис A.

Среднеквадратичная погрешность приближения j-й функции при фиксированном числе базисных функций $R<N$ есть

$$\delta_j^{(R)} = \frac{1}{T_j} \sum_{k=R+1}^{N} |a^*_{jk}|^2,$$

где T_j — длина интервала определения j-й функции.

Неравноценность погрешностей представления различных функций может быть учтена с помощью весовых коэффициентов $\{P_j\}$, $j=1, 2, \ldots, N_G$:

$$\delta^{(R)} = \sum_{j=1}^{N_G} \frac{P_j}{T_j} \sum_{k=R+1}^{N} |a^*_{jk}|^2 = \sum_{k=R+1}^{N} \sum_{m=1}^{N} \sum_{n=1}^{N} b_{mn} X_{km} X_{kn}, \qquad (1)$$

где

$$b_{mn} = \sum_{j=1}^{N_G} \frac{P_j}{T_j} a_{jm} a_{jn}; \quad \sum_{j=1}^{N_G} P_j = 1.$$

Оптимальный базис должен минимизировать $\delta^{(R)}$ при любом фиксированном числе членов аппроксимирующего ряда.

Каждое слагаемое внешней суммы выражения (1) — квадратичная форма, коэффициенты которой образуют симметричную матрицу с характеристическими числами $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_N$. Из теории квадратичных форм известно, что квадратичная форма $I = \sum b_{mn} X_{l_0 m} X_{l_0 n}$ при условии ортонормированности базиса $A^* \sum_{n=1}^{n} X_{km} X_{ln} = \delta_{kl}$ (δ_{kl} — символ Кронекера) достигает минимума λ_{l_0} на собственном векторе матрицы B, соответствующем собственному числу λ_{l_0}.

Таким образом, коэффициенты представления функций базиса A^* в базисе A суть компоненты соответствующих собственных векторов матрицы B, a:

$$\delta^{(R)}_{\min} = \sum_{k=R+1}^{N} \lambda_k.$$

Найденный, таким образом, базис характеризуется инвариантностью структуры и обеспечивает уменьшение величины λ_k с ростом номера k с наиболее возможной скоростью, и, следовательно, заданная погрешность аппроксимации достигается при минимальном числе членов аппроксимирующего ряда.

На основе описанной процедуры был разработан машинный алгоритм и одна из его возможных реализаций на языке АЛЬФА. Длина рабочей программы ~ 2000 команд.

Численный эксперимент проводился на модельном материале. Мощность исходного множества функций GN_6 была равна 10, а размерность вспомогательного базиса (соответствующая среднеквадратичной ошибке аппроксимации $\varepsilon = 0{,}005$) 30. В качестве вспомогательного базиса использовалась ортонормированная система тригонометрических функций. На БЭСМ-6 оптимальный базис был найден за 1 мин.

ЛИТЕРАТУРА

1. F. B. Pyne, E. F. Meclusky. On Essay on Primeimplicant Tables Fundamental Product Table of Combinations or Truthtable.— J. Soc. JAM, 1961, v. 9, № 4.
2. Т. Л. Майстрова. Линейное программирование и задачи минимизации нормальных форм булевских функций.— В кн.: Проблемы передачи информации. Вып. 12. М., Изд-во АН СССР, 1963.
3. В. Г. Новоселов. Нахождение кратчайших покрытий.— В кн.: Труды СФТИ. Вып. 48. Томск, 1966.
4. В. Д. Казаков. Структурная теория релейных устройств.— В кн.: Нахождение минимальных нормальных форм логической функции методом ограниченного перебора. М., Изд-во АН СССР, 1963.
5. А. Д. Закревский. О сокращении переборов при решении некоторых задач синтеза дискретных автоматов.— ИВУЗ. Сер. радиофизика, 1969, т. 7, № 1.
6. А. Д. Закревский. Оптимизация покрытий множеств.— В кн.: Логический язык для представления алгоритмов синтеза релейных устройств. М., «Наука», 1966.
7. А. М. Заездный, Г. В. Эйдукявичюс. Сокращенное представление сигналов с помощью систем ортогональных функций.— Радиотехника, 1963, № 11.

Поступила в редакцию 10 января 1974 г.

УДК 518.5 : 541.63+539.193

Б. С. ЖОРОВ

(Ленинград)

МОДЕЛИРОВАНИЕ НА ЭВМ ПРОСТРАНСТВЕННОЙ СТРУКТУРЫ ОРГАНИЧЕСКИХ СОЕДИНЕНИЙ

При решении многих задач теоретической химии и молекулярной биологии возникает потребность в определении устойчивых пространственных структур органических соединений. Установление энергетически выгодных конформаций молекул часто является необходимым этапом расчета молекулярных спектров. Важнейший, но весьма трудоемкий экспериментальный метод исследования конформаций молекул — рентгеноструктурный анализ — позволяет определить пространственную структуру молекул, находящихся только в кристаллическом состоянии. В тех же случаях, когда необходимо иметь представление о всех потенциальных конформационных возможностях соединения, достаточно полную информацию могут дать только расчетные методы.

В настоящее время методы теоретического конформационного анализа успешно применяются для изучения пространственной структуры различных химических соединений. Однако дело осложняется тем обстоятельством, что весьма дорогие программы для ЭВМ, реализующие эти методы, часто оказываются либо приспособленными к узкому классу соединений, либо требуют для расчета конкретной молекулы многочис-