«Оптика атмосферы и океана», **19,** № 8 (2006)

УДК 539.194

А.Д. Быков, Л.Н. Чебакова

Разработка программного обеспечения для работы с колебательно-вращательными спектрами молекул

Институт оптики атмосферы СО РАН, г. Томск

Поступила в редакцию 23.05.2006 г.

Представлена экспертная система работы с линейчатыми спектрами: их объединение, пересечение, вычисление дополнений, т.е. использование основных понятий теории множеств. Система хранит и учитывает спектральную информацию, рассматривая в качестве свойств элементов множеств упорядоченный набор линий, характеризующийся центром линии, интенсивностью, полушириной, сдвигом. Во всех операциях со спектрами используется сравнение линий двух спектров, что является типичной задачей распознавания образов, так как известно, что экспериментальные спектры измеряются с некоторой погрешностью.

Введение

В настоящее время при исследовании молекулярных газов спектроскопическими методами регистрируется множество спектров при различных условиях: температуре, давлении и оптической толще. Такой подход дает возможность получить детальную информацию как о сильных, так и о слабых линиях. Современные спектрометры позволяют регистрировать спектры, содержащие большое количество спектральных элементов, в широких спектральных диапазонах. Например, спектр разряда в молекулярном водороде содержит более 40 тыс. спектральных элементов, банк данных HITRAN — более 1 млн линий; база рассчитанных данных Швенке по спектру водяного пара — более 300 млн линий.

Вследствие больших объемов информации, получаемых при измерениях или расчетах, возникает необходимость в разработке специальных программных средств, способных хранить, сравнивать и объединять полученную спектральную информацию. Такие программные средства должны учитывать спектроскопические моменты решаемой задачи, возможное наличие в спектрах линий различных газов, линии изотопических модификаций молекул, различные условия измерений спектров, наличие погрешностей в определении спектроскопических параметров линий.

Спектры можно рассматривать как множества, каждый элемент множества (отдельная линия) имеет некоторый набор характеристик. Это могут быть центры и интенсивности линий, коэффициенты уширения и сдвига, параметры, определяющие температурную зависимость. Некоторые спектроскопические задачи представляются в виде операций с множествами. Например, так называемая «тривиальная» интерпретация линий в спектрах, когда точно известны начальный и конечный уровни энергии некоторого перехода, представляется как поиск совпадений центров линий в расчетном и экспериментальном спектрах, что, очевидно, эквивалентно нахождению пересечения двух множеств.

Такая работа проводится в рутинном режиме, однако большие объемы спектроскопической информации затрудняют обработку, кроме того, наличие погрешностей измерений, различие в условиях измерений, требуют привлечения квалифицированных специалистов. Таким образом, создание специальных программных средств работы с «большими» спектрами, полученными при различных условиях, становится актуальным.

Цель данной работы заключается в создании программного обеспечения для манипуляции спектрами (объединение, поиск разностей и т.д.), использующего методы теории множеств и распознавания образов. Система должна хранить и учитывать спектральную информацию, рассматривая в качестве характеристик элементов множеств упорядоченный набор линий, где каждая линия характеризуется центром линии, интенсивностью, полушириной, сдвигом.

1. Операции со спектрами

Будем рассматривать спектры как множество элементов, имеющих спектральные характеристики. В качестве примера можно рассмотреть множество A — спектр H_2O в области 6000–10000 $см^{-1}$, зарегистрированный при температуре 600 К, и множество B — спектр смеси H_2O, HDO и D_2O в области 5000–12000 $см^{-1}$, зарегистрированный при $T = 296$ К. Для простоты будем считать, что A и B содержат только центры и интенсивности линий и точность определения этих параметров одинакова.

Как известно, для множеств можно определить операции объединения, пересечения и дополнения.

Объединением (или суммой) множеств A и B считается множество $C = A \cup B$, где $C = \{x | x \in A$ или $x \in B\}$. Для нашего случая спектр C есть банк данных для спектрального интервала 5000–12000 $см^{-1}$,

содержащий также линии изотопных модификаций и слабые линии водяного пара, соответствующие переходам на высокие вращательные уровни энергии, а также линии горячих полос (в области между 6000 и 10000 $см^{-1}$). Простое объединение двух списков линий может, очевидно, привести к ошибкам. Например, множество C может содержать два набора некоторых линий, присутствующих одновременно в A и B. Чтобы избежать повторения, необходимо предварительно установить соответствие между одними и теми же линиями в обоих спектрах. Поскольку температуры, при которых регистрировались спектры, отличаются, то необходим пересчет интенсивностей линий к одной температуре. Таким образом, простая операция объединения множеств требует некоторого предварительного анализа и расчета.

Пересечением множеств A и B называется множество $C = A \cap B$, удовлетворяющее условию $C = \{x \in A \text{ и } x \in B\}$. Для нашего примера это линии спектра только H_2O в области от 6000 до 10000 $см^{-1}$, соответствующие температуре 600 К (в предположении, что все линии спектра B проявляются при $T = 600$ К). Таким образом, пересечение есть линии спектра основной модификации в области 1,3 мкм, за исключением линий, проявляющихся при высокой температуре. Очевидно, что, как и в предыдущем случае, необходимы дополнительный анализ ситуации и детальное сравнение линий.

Дополнение. Пусть задано некоторое фиксированное множество S и $A \subset S$. Множество $A' = S \setminus A$ называется дополнением множества A в том смысле, что A' дополняет множество A до S. На нашем примере дополнение будет содержать только линии HDO и D_2O в области 6000–10000 $см^{-1}$, а также все линии спектра B в области 5000–6000 и 10000–12000 $см^{-1}$.

Все эти операции, являясь стандартными операциями со множествами, часто встречаются при работе со спектрами и спектральными банками данных. Однако при работе с линейчатыми спектрами важно определить и ряд дополнительных операций.

Во-первых, необходимо определить операцию выделения части спектра в соответствии с некоторыми условиями. Такая операция, очевидно, есть определение подмножества. Здесь полезно отметить, что выделение подмножества одних и тех же линий (определяемых в соответствии с некоторым условием, формально одинаковым для двух спектров) может приводить к отличающимся наборам линий. Поэтому до проведения отбора линий необходимо сначала установить соответствие между линиями двух спектров.

Во-вторых, спектры могут быть зарегистрированы при различных условиях, различной чувствительности, различном спектральном разрешении. Поэтому необходима процедура хотя бы приближенного пересчета параметров линий, т.е. получения нового множества.

Все эти операции предполагают распознавание одинаковых линий в двух различных спектрах, поэтому для решения задачи необходимо применять методы теории распознавания образов.

2. Задача распознавания образов

В настоящее время применение теории распознавания образов в спектроскопии ограничено несколькими конкретными задачами. В ряде работ методы теории распознавания образов применялись для решения некоторых задач спектроскопии молекул (см., например, [2–5])

В теории распознавания образов есть две основные разновидности методов классификации: с обучением и без обучения. Они различаются характеристиками решаемых ими задач. Первые решают задачу классификации при фиксированном числе классов, определяемых разработчиком. Вторые ставят целью выявление классов в имеющейся совокупности объектов.

В литературе задача распознавания образов формулируется в следующим виде. Пусть X – пространство описаний, характеристик объектов (так называемое «признаковое пространство»), в нашей задаче его составляют, например, центры и интенсивности линий; U – пространство решений. Каждый объект в пространстве X представляется точкой. Решить задачу распознавания образов означает построить отображение $u(x) : X \to U$, наилучшее в некотором заранее определенном смысле, например в смысле близости к отображению $u^*(x)$, задаваемому «учителем». Если известно полное вероятностное описание обоих пространств X и U, можно построить байесово решающее правило. В других случаях приходится проводить оценку либо совместных плотностей распределения, либо оценивать сразу решающее правило, которое есть не что иное, как разделяющая поверхность в пространстве описаний X. На рис. 1 в качестве примера приведены объекты двух классов (крестики и кружки). Если признаки выбраны удачно (в нашем случае это координаты x и y), то объекты разных классов будут находиться в различных областях признакового пространства. В этом случае для построения решающего правила достаточно определить разделяющую линию.

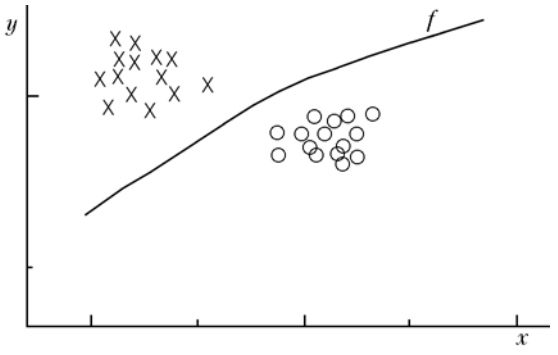

Рис. 1. Два класса распознаваемых объектов и разделяющей функции f

Работая со спектрами как с множествами, при выполнении любой операции необходимо сравнивать элементы двух множеств, устанавливать соответствие между ними. Поскольку параметры линий известны с некоторой, иногда не малой, погрешно-

стью, возникает задача распознавания пар соответственных линий. Такая задача является типичной при распознавании образов.

Рассмотрим далее метод Розенблатта, реализованный в данной работе для построения алгоритма распознавания. Метод был предложен Ф. Розенблаттом в 1959 г. для нейронных сетей (НС) [1]. Перцептрон Розенблатта (рис. 2) имеет пороговую функцию активации f.

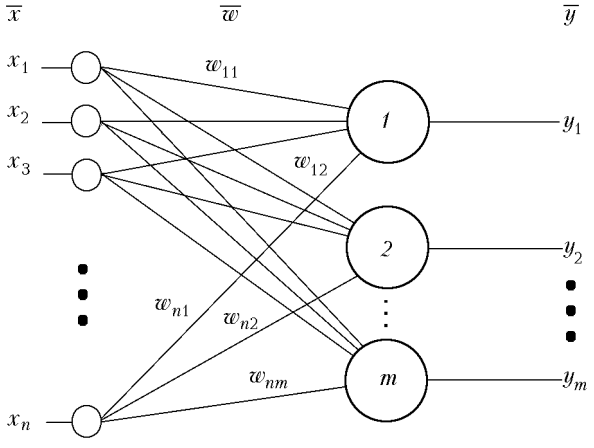

Рис. 2. Схема перцептрона Розенблатта

Процедуру настройки весов межнейронных (синаптических) связей при обучении однослойного перцептрона можно представить в виде итерационной схемы [4]:

$$w_{ij}(t+1) = w_{ij}(t) + \alpha x_i d_j, \qquad (1)$$

где x_i — сигнал на i-м входе в систему; d_j — желаемый (идеальный) результат на j-м выходе и параметр α ($0 < \alpha < 1$) есть весовый коэффициент (скорость обучения). Весовые коэффициенты меняются только в том случае, если реальное выходное значение не совпадает с идеальным выходным значением. Алгоритм обучения Розенблатта строится следующим образом:

1. Весовые коэффициенты НС инициализируются малыми случайными значениями.

2. На вход НС подается очередной учебный пример.

3. Если выход НС y_j не совпадает с идеальным выходом d_j, то происходит модификация весов по (1).

4. Осуществляется повтор вычислений начиная с п. 2, пока $\forall i : y_i = d_i$ или весовые коэффициенты не перестанут меняться

3. Реализация метода распознавания образов

В задаче сравнения двух спектров удобно использовать для построения алгоритма распознавания одинаковых линий в двух спектрах описанный выше перцептрон Розенблатта, а точнее его аналог — однослойную нейронную сеть (рис. 3).

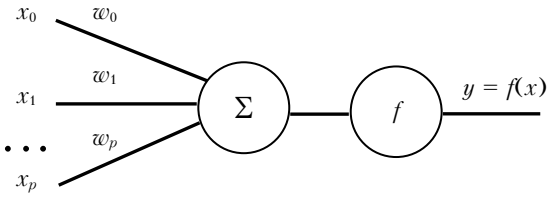

Рис. 3. Однослойный перцептрон Розенблатта (один выходной сигнал)

На вход сети подается p-мерный вектор признаков $\{x_i,\ i = 1, 2, \ldots, p\}$ (берутся такие характеристики линий первого и второго спектров, как центры линии и их интенсивности). Для определенности будем рассматривать случай, когда $p = 5$.

С учетом этих дополнений перцептрон теперь имеет вид, приведенный рис. 3, где p — размерность исходных данных (количество признаков используемых для классификации); x_i — компонента входного вектора признаков, $i = 1,\ldots,p$; w_i — весовые коэффициенты между входным и выходными слоями, $i = 0, 1, \ldots, p$; y — выходное значение нейрона сети (выход сети):

$$y = f\left(\sum_{i=1}^{p} w_i x_i + w_o\right) \equiv f\left(\sum_{i=0}^{p} w_i x_i\right)$$

— функция активации нейронов, в качестве которой на выходе НС предлагается использовать коэффициенты корреляции интенсивностей линий первого и второго спектров, а именно:

$$f(x,y) = \frac{\sum_i xy - \frac{\sum_i x \sum_i y}{n}}{\sqrt{\left[\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right]\left[\sum y^2 - \frac{\left(\sum y\right)^2}{n}\right]}}, \qquad (2)$$

и t-критерий Стьюдента

$$t_r = r\sqrt{\frac{n-2}{1-r^2}}. \qquad (3)$$

Для работы с НС желательно использовать исходные данные не в первоначальном виде, а после некоторой предварительной обработки. Мы будем ориентироваться при нормировке не на экстремальные значения, а на типичные, т.е. на статистические характеристики данных — среднее и дисперсия:

$$\tilde{x}_i = \frac{x_i - \overline{x}_i}{\sigma_i},$$

где

$$\overline{x}_i \equiv \frac{1}{p}\sum_{\alpha=1}^{p} x_i^{\alpha}; \quad \sigma_i^2 \equiv \frac{1}{p-1}\sum_{\alpha=1}^{p} (x_i^{\alpha} - \overline{x}_i)^2.$$

В этом случае основная масса данных будет иметь единичный масштаб, т.е. типичные значения всех переменных будут сравнимы. Однако теперь нормированные величины могут выходить за пределы единичного интервала, более того, максимальный

разброс значений $\tilde{x}_i$ заранее не известен. Для входных данных это может быть и несущественно, но выходные переменные будут использоваться в качестве эталонов для выходных сигналов нейронов. Мы рассмотрим случай, когда нейроны — сигмоидные, т.е. выходные сигналы могут принимать значения лишь в единичном диапазоне. Чтобы установить соответствие между обучающей выборкой и нейросетью, необходимо ограничить диапазон изменения переменных.

Линейное преобразование, представленное выше, не способно нормировать основную массу данных и одновременно ограничить диапазон возможных значений этих данных. Естественный выход из этой ситуации — использовать для предобработки данных функцию активации тех же нейронов. Например, нелинейное преобразование

$$\tilde{x}_i = F\left(\frac{x_i - \overline{x}_i}{\sigma_i}\right), \quad F(x) = \frac{1}{1 + e^{-x}}$$

нормирует основную массу данных, одновременно гарантируя, что $\tilde{x}_i \in [0,1]$.

Если через d обозначить желаемое значение сигнала на выходе сети (указание учителя), то ошибку системы для заданного входного сигнала (рассогласование реального и желаемого выходного сигнала) можно записать в следующем виде:

$$\varepsilon^k = y^k - d^k,$$

где k — номер обучающей пары в обучающей выборке, $k = 1, 2, \ldots, n_1 + n_2$, n_1 — количество векторов первого класса (правильных ответов сети в обучающей выборке), n_2 — число векторов второго класса (неверных ответов).

В качестве функционала оптимизации будем использовать критерий минимума среднеквадратической функции ошибки:

$$E = e^{-\mu\left|\nu_i^A - \nu_j^B\right|^2} \to \min,$$

где ν_i^A, ν_j^B — центры линий с номерами i (первого) и j (второго спектра); μ — некоторый коэффициент. В случае, когда функционал ошибки НС задан, главная задача обучения нейронных сетей заключается в его минимизации. Процедура обучения НС сводится к коррекции весов связей w_i. Перед началом обучения весовые коэффициенты НС устанавливаются произвольно, например обнулением.

На первом этапе на вход НС в определенном порядке подаются учебные примеры. На каждой итерации вычисляется ошибка для учебного примера E_L (ошибка обучения) и по определенному алгоритму производится коррекция весов НС. Целью процедуры коррекции весов является минимизация ошибки E_L.

На втором этапе обучения производится проверка правильности работы НС. На вход НС в определенном порядке подаются контрольные примеры. На каждой итерации вычисляется ошибка для контрольного примера E_G (ошибка обобщения). Если результат неудовлетворительный, то производятся модификация множества учебных примеров и повторение цикла обучения НС. После нескольких итераций алгоритма обучения ошибка обучения E_L падает почти до нуля, в то время как ошибка обобщения E_G вначале падает, а затем начинает возрастать. О такой ситуации говорят как об «эффекте переобучения». В этом случае обучение необходимо прекратить.

При однослойной сети алгоритм обучения с учителем простой. Желаемые выходные значения нейронов единственного слоя заведомо известны, и подстройка весов синаптических (межнейронных) связей идет в направлении, минимизирующем ошибку на выходе сети.

Описанные здесь алгоритмы распознавания были применены для создания экспертной системы работы с линейчатыми спектрами. Экспертная система позволяет проводить описанные выше операции со спектрами — объединять их, находить дополнения и объединение спектров. При этом осуществляется отождествление линий в двух спектрах, основанное на применении описанного выше алгоритма. В качестве возможного примера применения системы можно указать работу с расчетным спектром Швенке и экспериментальными спектрами водяного пара, что будет описано в последующих статьях.

Авторы выражают благодарность чл.-кор. РАН С.Д. Творогову за поддержку и внимание к данной работе.

Работа поддержана программой РАН «Оптическая спектроскопия и стандарты частоты» и грантом INTAS № 03-51-3394.

1. *Айзерман М.А., Браверман Э.И., Розоноэр Л.И.* Метод потенциальных функций в задачах обучения машин. М.: Наука, 1970. 384 с.
2. *Левин Л.Л.* Введение в теорию распознавания образов. Томск: Изд-во Том. ун-та, 1982. С. 10–28.
3. *Щербаков А.П.* Применение методов теории распознавания образов для идентификации линий в колебательно-вращательных спектрах // Оптика атмосф. и океана. 1997. Т. 10. № 8. С. 947–959.
4. *Быков А.Д., Синица Л.Н., Науменко О.В., Пшеничников А.М., Щербаков А.П.* Получение списка линий колебательно-вращательных спектров молекул // Оптика и спектроскопия. 2003. Т. 94. № 3. С. 528–537.
5. *Эляшберг М.Е., Грибов Л.А., Серов В.В.* Молекулярный спектральный анализ и ЭВМ. М.: Наука, 1980. 307 с.

A.D. Bykov, L.N. Chebakova. **Software for processing vibrational-rotational molecular spectra.**

The expert system for spectra processing is presented. In this system a spectrum can be treated as a set of elements (lines) characterized by several attributes: line position, line strength, etc. In processing spectra, one can define all operations usable in the set theory such as union, intersection, etc. All spectra manipulations include juxtaposition of spectral lines from two spectra. This is a typical problem of the pattern recognition theory.