Development of a software for processing the vibrational-rotational spectra of molecules

A.D. Bykov and L.N. Chebakova

Institute of Atmospheric Optics,
Siberian Branch of the Russian Academy of Sciences, Tomsk

Received May 23, 2006

We present an expert system for processing line spectra by use of operations like union, intersection, computation of relative complements, i.e., using the basis of set theory. The system stores and takes into account the spectroscopic information, considering the ordered set of lines characterized by the line center, its intensity, half-width, and shift as properties of set elements. All operations applied to spectra, use the comparison of lines of two spectra what makes a typical problem of pattern recognition, because it is a known that experimental spectra are measured with a certain error.

Introduction

At present investigations of molecular gases by means of spectroscopic techniques are being conducted while recording a lot of spectra under different conditions, i.e., varying temperature, pressure and optical thickness. Such an approach enables one to obtain detailed information on both strong and weak lines. Modern spectrometers are capable of recording the spectra containing large number of spectral elements within wide spectral ranges. For example, the spectrum of electric discharge in molecular hydrogen contains more than 40 thousand spectral elements; the HITRAN databank contains information on more than 1 million lines; Schwenke database on water vapor spectrum contains more than 300 million lines.

Owing to that huge bulk of information, obtained in measurements or calculated, there is a necessity in developing specialized software, capable of storing, making comparison and uniting the spectroscopic information obtained. Such software should take into account many spectroscopic aspects of the problem to be solved, including the presence of lines of different gases in the spectra, lines of the molecular isotopic modifications, various conditions at measurements of the spectra, as well as errors in determination of spectral line parameters.

Spectra can be considered as sets, each element of the set (single line) being characterized by a certain set of characteristics. It can be the centers and intensities of spectral lines, coefficients of broadening and shift, parameters determining temperature dependence. Some spectroscopic problems are presented as operations with sets. For example, the so-called "trivial" interpretation of spectral lines, when the initial and final energy levels of some transition are known exactly, is in seeking lines whose centers in calculated and experimental spectra coincide that, obviously, is equivalent to operation of finding the intersection between the two sets.

Such an operation is carried out in a routine regime; however, large bulks of spectroscopic information make the processing difficult, besides, the measurement errors and difference in measurement conditions, demand engaging highly skilled experts. Therefore, the problem of generating special software codes for processing "large" spectra, obtained under various conditions, becomes very urgent.

The aim of this study was to generate the software code for processing the spectra (union, seeking differences, and so on) using the methods of set theory and pattern recognition. The system should provide for storage and account of the spectroscopic information, considering the ordered set of lines as the characteristics of set elements, where each line is defined by the line center, intensity, half-width, and shift.

1. Operations applied to spectra

Let us consider spectra as sets of elements with spectral characteristics. As an example, let us consider the A set to be the H_2O spectrum in the region from 6000–10000 cm^{-1}, recorded at 600 K, and the B set being the spectrum of H_2O, HDO and D_2O mixture in the region from 5000–12000 cm^{-1}, recorded at $T = 296$ K. For simplicity, we shall consider that A and B sets contain only the centers and intensities of lines determined with the same accuracy in both sets.

As known, it is possible to define the operations of union, intersection and relative complements for the sets.

Union of the A and B sets is the set $C = A \cup B$, where $C = \{x | x \in A \text{ or } x \in B\}$. In our case, the C spectrum is the data bank for the spectral range from 5000–12000 cm^{-1} containing the lines of isotopic modifications and weak lines of water vapor corresponding to transitions to high rotational levels, and the lines of hot bands (in the range between 6000 and 10000 cm^{-1}). Simple union of two lists of

0235-6880/06/08 658-04 $02.00

lines can, obviously, lead to errors. For example, the C set can contain two sets of certain lines that belong to A and B sets simultaneously. In order to avoid the replication it is necessary, first, to establish the correspondence between the same lines in both spectra. Since the temperatures at which spectra were recorded differ, the recalculation of line intensities to same temperature is necessary. Therefore, the operation of simple union of the sets requires some preliminary analysis and pre-calculation.

Intersection of the A and B sets is the set $C = A \cap B$ satisfying the condition $C = \{x \in A$ and $x \in B\}$. In our case, this involves the spectral lines of only H_2O in the region from 6000 to 10000 cm^{-1} corresponding to the temperature of 600 K (assuming that at $T = 600$ K all lines of the B spectrum are seen). Thus, intersection involves the spectral lines of basic modification in the region of 1.3 μm, except for the lines that are seen at a high temperature. It is obvious that additional analysis of the situation and detailed comparison of lines are necessary, as in the previous case.

Relative complement. Let a fixed set S be given and $A \subset S$. The set $A' = S \setminus A$ is called a relative complement of A set in the sense that A' completes the set A to S. In our example, the relative complement will contain only the HDO and D_2O lines in the region from 6000 to 10000 cm^{-1}, and all lines of the B spectrum in the region from 5000 to 6000 and from 10000 to 12000 cm^{-1}.

All these operations being the standard operations with sets are often applied in processing spectra and spectroscopic data banks. However, in the case with line spectra, it is important to determine also some additional operations.

First, it is necessary to determine the operation of isolating a portion of the spectrum according to certain conditions. Such an operation is, obviously, the operation of identifying a subset. It is important to note here that isolating a subset of the same lines (determined according to a certain condition, formally identical to two spectra) can lead to different sets of lines. Therefore, before doing the line selection, it is necessary to establish the correspondence between the lines of the two spectra.

Second, it may happen that the spectra could have been recorded under different conditions, like different sensitivity, and different spectral resolution. Therefore, a procedure is needed for recalculating, though only approximately, the line parameters, that is for generating a new set.

All these operations assume the recognition of identical lines in two different spectra; therefore, it is necessary to apply methods of the pattern recognition theory for solving this problem.

2. The task of pattern recognition

At present, application of the pattern recognition theory to spectroscopy is limited by some particular problems. In a number of studies, methods of the pattern recognition theory were applied to solving some problems in molecular spectroscopy (see, for example, Refs. 2 to 5).

There are two basic types of classification methods in the pattern recognition theory: with training and without training. They differ by the characteristics of problems to be solved. The methods of one type solve a classification problem at a fixed number of classes set by an application designer. The methods of the other type are aimed at revealing the classes in the set of objects available.

In the literature, a problem of pattern recognition is stated as follows. Let X be the space of descriptions, i.e., of the characteristics of objects (the so-called attribute space). In our problem, centers and intensities of lines do compose this space. Let U be the space of solutions. Every object from the space X is presented as a point. To solve the problem of pattern recognition means to construct a representation $u(x): X \to U$, which is the best in a certain sense, for example, in the sense of proximity to the $u^*(x)$ representation set by a "teacher." If the full probabilistic description of both X and U spaces is known, it is possible to construct the Bayes estimator. In other cases, we have to estimate either simultaneous distribution densities, or directly the decision rule. This rule is nothing but the separation surface in the description space X. Figure 1 presents the objects of two classes as an example (circles and crosses). If the attributes have been chosen successfully (in our case, it is the x and y coordinates), the objects of different classes will be in different domains of the attribute space. In this case, to formulate the decision rule it is sufficient to determine the separation line.

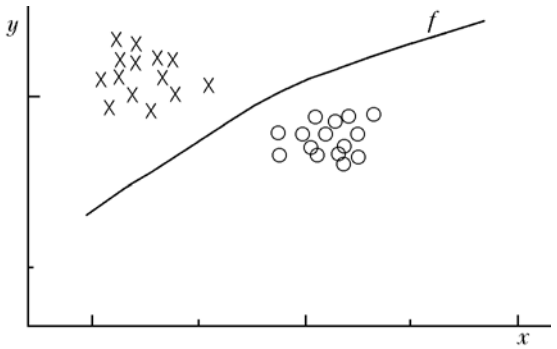

Fig. 1. Two classes of recognizable objects and separation function f.

In processing spectra as sets, it is necessary, while doing any operation, to compare the elements of two sets and to establish the correspondence between them. Since the line parameters are known with a certain sometimes not small error the problem arises on recognizing the corresponding line pairs. Such a problem is typical in the pattern recognition.

Let us consider the Rosenblatt method realized in this study while constructing the recognition algorithm. The method was proposed by

F. Rosenblatt[1] in 1959 for the neural networks (NN). The Rosenblatt perceptron (Fig. 2) has the threshold activation function f.

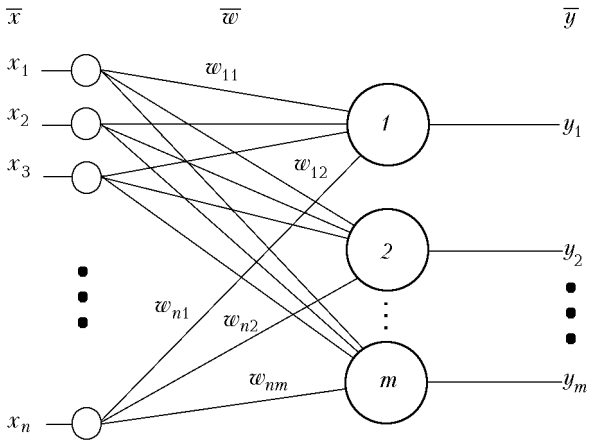

Fig. 2. Scheme of the Rosenblatt perceptron.

The procedure of adjusting the weights of transneuronal (synaptic) connections at training the single-layer perceptron can be presented by an iterative scheme[4]:

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

where x_i is the signal at the ith input of the system; d_j is the desirable (ideal) result at the jth output, and α $(0 < \alpha < 1)$ parameter is the weighting coefficient (learning velocity). The weighting coefficients change only in case, when the real output value does not coincide with the ideal output value. Rosenblatt's algorithm is constructed in the following way:

1. Weighting coefficients of NN are initialized by small random values.

2. The next learning example is applied to the NN input.

3. If the NN output y_j does not coincide with ideal output d_j, the modification of the weights is being done following Eq. (1).

4. The calculations are reiterated starting from Point 2, until $\forall i : y_i = d_i$ or unless the weighting coefficients stop changing.

3. Realization of pattern recognition method

In the problem on comparing two spectra, it is convenient to use the above-described Rosenblatt perceptron or, what is more correct, its analog, the single-layer neural network (Fig. 3), for constructing the algorithm of recognizing the lines which are identical in two spectra.

The p-dimensional vector of characteristics $\{x_i, i = 1, 2, \ldots, p\}$ is applied to the network input (such line characteristics of the first and second spectra as line centers and their intensities are taken). For certainty, we shall consider the case, when $p = 5$.

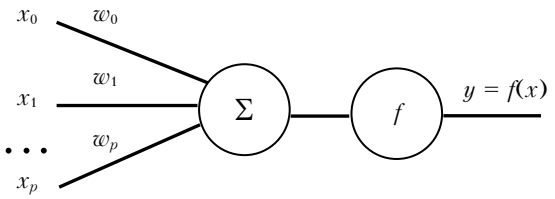

Fig. 3. The single-layer Rosenblatt perceptron (single output signal).

Having these additions in mind, the perceptron now takes the form presented in Fig. 3, where p is the dimensionality of the initial data (the number of characteristics used for classification); x_i is the component of input vector of characteristics, $i = 1, \ldots, p$; w_i are the weighting coefficients between input and output layers, $i = 0, 1, \ldots, p$; y is the output value of the network neuron (network output):

$$y = f\left(\sum_{i=1}^{p} w_i x_i + w_0\right) \equiv f\left(\sum_{i=0}^{p} w_i x_i\right)$$

is the neuron activation function. It is proposed to use, at the NN output, the coefficients of correlation between line intensities of first and second spectra, as such an activation function, namely

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

and t-student criterion

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

To make use of operation with NN it is desirable that the initial data are used not in the original form but after a certain preprocessing. We shall normalize the data by the typical values, that is statistical mean and the variance, rather than the extreme ones:

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

where

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

In this case, the major bulk of data will have same scale that means that the typical values of all variables will be comparable. However, now the standardized values can leave the unit interval,

moreover, the maximum spread of the $\tilde{x}_i$ values is not known beforehand. It may be insignificant for the input data, but the output variables will be used as standards for the output neural signals. We shall consider the case, when the neurons are sigmoids, that is the output signals take values only within a unity interval. In order to establish the correspondence between the learning sample and the neural network, it is necessary to limit the range within which the variables may vary.

The linear transformation presented above does not allow one to normalize the major bulk of data and to limit, simultaneously, the range of values allowed for these data. A natural way to resolve this situation is to use the activation function of the same neuron for the data preprocessing. For example, the non-linear transformation

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

normalizes the major bulk of data and simultaneously guarantees that $\tilde{x}_i \in [0,1]$.

If we denote the desirable signal value at the network output as d (teacher's instruction), the system error for the given input signal (mismatch between real and desirable output signals) can be written as follows:

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

where k is the number of a pair in the learning sample, $k = 1, 2, \ldots, n_1 + n_2$, n_1 is the number of vectors of the first class (right answers of the network in the learning sample), n_2 is the number of vectors of the second class (wrong answers).

We shall use the minimum criterion of the root-mean-square-error function as an optimization functional:

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

where ν_i^A, ν_j^B are the centers of lines with the numbers i (in the first) and j (in the second spectrum); μ is some coefficient. In the case when the NN error functional is set, the main problem in teaching neural networks is its minimization. The NN teaching procedure is reduced to correcting the weights w_i of the connections. Before training, the NN weight coefficients are set arbitrarily, for example, by resetting to zero.

At the first stage, the learning samples are applied to the NN input in a certain order. Using the learning sample E_L an error (learning error) is calculated at each iteration and the NN weights are corrected following some algorithm. The aim of the weight correction procedure is minimization of the error E_L.

At the second stage of teaching, the control of the NN operation is carried out. Control samples are applied to the NN input in a certain order. Using the control sample the error E_G (error of generalization) is calculated at each iteration. If the result is poor, modification of a great number of teaching samples is done and the NN learning cycle is repeated. After some iterations of the learning algorithm, E_L falls almost down to zero, while E_G first falls, but then starts to increase. This situation is called the "overtraining effect." In this case, the learning should be ceased.

In the case of a single-layer network, training algorithm with the participation of a teacher is quite simple. Desirable output neural values of a single layer are certainly known and the weight adjustment of the synaptic (interneural) connections is performed in the way to minimize the error at the network output.

Recognition algorithms described here were applied to building up an expert system for processing the line spectra. The expert system allows one to make the operations with spectra described above, that is, to sum those up, to find the relative complements and unions of spectra. Thus, the line identification in two spectra is carried out using the above-described algorithm. As a possible example of the system use, we can mention its application to processing the calculated Schwenke spectrum and the experimentally measured water vapor spectra that will be described in the subsequent papers.

Acknowledgments

Acknowledgment is made to Corresponding Member of RAS S.D. Tvorogov for his support and attention to this work.

This work was supported by the Program of the Russian Academy of Sciences "Optical spectroscopy and frequency standards" and by the INTAS Grant No. 03–51–3394.

References

1. M.A. Aizerman, E.I. Bravermann, L.I. Rozonoer, et al., *Method of Potential Functions in Problems of Computer Training* (Nauka, Moscow, 1970), 384 pp.
2. L.L. Levin, *Introduction to the Pattern Recognition Theory* (Izd. Tomsk Univ., Tomsk, 1982), 18 pp.
3. A.P. Shcherbakov, Atmos. Oceanic Opt. **10**, No. 8, 591–597 (1997).
4. A.D. Bykov, L.N. Sinitsa, O.V. Naumenko, et al., Opt. Spektrosk. **94**, No. 3, 528–537 (2003).
5. M.E. Elyashberg, L.A. Gribov, and V.V. Serov, *Molecular Spectral Analysis and ECM* (Nauka, Moscow, 1980), 307 pp.