

University of Montana

## ScholarWorks at University of Montana

---

Graduate Student Theses, Dissertations, &  
Professional Papers

Graduate School

---

1991

### Classifying organic compounds using expert system and neural networks

Judit Ambro

*The University of Montana*

Follow this and additional works at: <https://scholarworks.umt.edu/etd>

**Let us know how access to this document benefits you.**

---

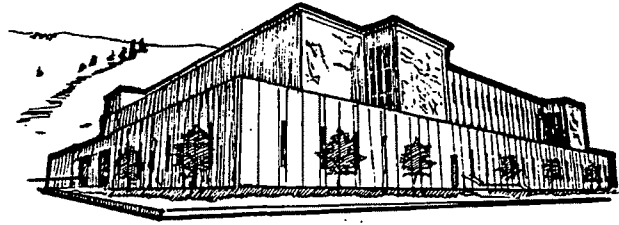
#### Recommended Citation

Ambro, Judit, "Classifying organic compounds using expert system and neural networks" (1991).

*Graduate Student Theses, Dissertations, & Professional Papers*. 5104.

<https://scholarworks.umt.edu/etd/5104>

This Thesis is brought to you for free and open access by the Graduate School at ScholarWorks at University of Montana. It has been accepted for inclusion in Graduate Student Theses, Dissertations, & Professional Papers by an authorized administrator of ScholarWorks at University of Montana. For more information, please contact [scholarworks@mso.umt.edu](mailto:scholarworks@mso.umt.edu).



Maureen and Mike  
**MANSFIELD LIBRARY**

---

Copying allowed as provided under provisions  
of the Fair Use Section of the U.S.

**COPYRIGHT LAW, 1976.**

Any copying for commercial purposes  
or financial gain may be undertaken only  
with the author's written consent.

---

University of  
**Montana**

CLASSIFYING ORGANIC COMPOUNDS  
USING EXPERT SYSTEM AND NEURAL NETWORKS

by

Judit Ambro

Presented in partial fulfillment of the requirements

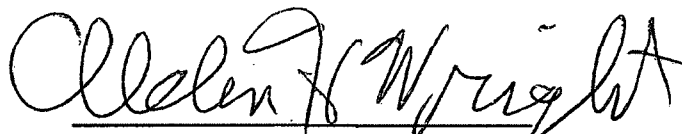
for the degree of

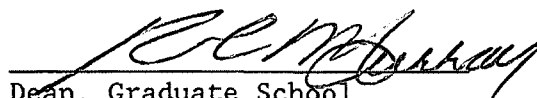
Master of Science


University of Montana

1991

Approved by

  
Chairman, Board of Examiners

  
Dean, Graduate School

  
Date

UMI Number: EP40568

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.

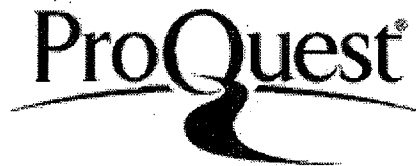


UMI EP40568

Published by ProQuest LLC (2014). Copyright in the Dissertation held by the Author.

Microform Edition © ProQuest LLC.

All rights reserved. This work is protected against unauthorized copying under Title 17, United States Code



ProQuest LLC.  
789 East Eisenhower Parkway  
P.O. Box 1346  
Ann Arbor, MI 48106 - 1346

Classifying Organic Compounds Using Expert System and Neural Networks

(69 pp.)

Director: Dr. Alden H. Wright



ABSTRACT

This work deals with the problem of classifying unknown organic compounds into classes based on their structure by using their infrared spectra.

The infrared spectrum of an organic compound consists of several peaks each characteristic to certain substructures (functional groups) of the molecule. The ambiguity of classifying peaks in an infrared spectrum by a human can be avoided if a neural network is trained to do the classification. The size of the pattern-recognition problem can be divided into several subproblems since only well defined parts of the spectrum contain information about the presence or absence of functional groups of interest.

A number of neural networks were trained so that each one can recognize the presence or absence of a particular functional group. A decision tree was then used to classify the compounds using the output of these networks.

The trained neural networks were able to identify fairly accurately the presence and the absence of functional groups. This method is compared to a pure neural network approach which used the same set of compounds. The final classification results were almost identical. When using the combined neural network - decision tree approach smaller networks can be used resulting in faster training, the outcome of a decision can be traced, and the modular structure of the system better accommodates changes in the overall classification goals. A disadvantage of the combined method is that structural differences and contamination in the organic compounds get more emphasis.

## Table of Contents

### 1. Introduction

#### 1.1 Infrared Spectroscopy

##### 1.1.1 Features of an Infrared Spectrum

##### 1.1.2 Interpretation of an Infrared Spectrum

#### 1.2 Decision Tree and Corresponding Rules

#### 1.3 Expert Systems

##### 1.3.1 Benefits of an Expert System

#### 1.4 Neural Networks

##### 1.4.1 Network Architecture

##### 1.4.2 Recurrent and Nonrecurrent Networks

##### 1.4.3 Training a Neural Network

##### 1.4.4 Backpropagation Algorithm

##### 1.4.5 Justification for the Backpropagation Algorithm

#### 1.5 Brief Review of Existing Spectral Interpreters

### 2. Objective

#### 2.1 A Comparison of the Conditions in the Two IR Interpretation Method

### 3. Methodology

#### 3.1 Data

#### 3.2 Preparation of Data

##### 3.2.1 Spectra

##### 3.2.2 Normalization

##### 3.2.3 Sampling

- 3.2.4 Hydroxyl Group in Alcohols and in Carboxylic Acids
      - 3.2.5 Phenyl
      - 3.2.6 Hydrocarbons
      - 3.2.7 Summary of the Input Data
    - 3.3 Neural Network Program
      - 3.3.1 Major Features of the Program
    - 3.4 Experimental Design
      - 3.4.1 Data
      - 3.4.2 Training Parameters
      - 3.4.3 Testing
    - 3.5 Decision Tree and Rules
  - 4. Results
    - 4.1 Bad Classifications
    - 4.2 Decision Tree
    - 4.3 Comparison with Fessenden and Györgyi's Result
  - 5. Conclusion
    - 5.1 Advantages and Disadvantages
    - 5.2 Comments
    - 5.3 Future Work
  - 6. References
    - 6.1 Appendix
    - 6.2 Bibliography

## List of Tables

- Table 3.1 Bonds and classes used in this study. (page 25)
- Table 3.2 Compounds used in this study. (page 26)
- Table 3.3 C=O and -OH absorption bands. (page 28)
- Table 3.4 Sampling frequency. (page 30)
- Table 3.5 Sampling frequency for phenyl. (page 33)
- Table 3.6 Absorption bands appearing in the classes studied. (page 34)
- Table 4.1 Summary of errors. (page 41)
- Table 4.2 Problematic compounds in the C-O-C absorption band. Each quotient indicates the ratio of the number of good, uncertain, or bad classifications of a certain testing compound over the total number that compound was in the testing set. These compounds were chosen randomly thus they were tested different number of times. (page 42)
- Table 4.3 Problematic compounds in the carboxylic acid -OH absorption band. (page 46)
- Table 4.4 Problematic compounds in the phenyl absorption bands. (page 48)
- Table 4.5 Output for AlCl<sub>3</sub> in T01. (page 51)
- Table 4.6 Output for Est12 in T05. (page 52)
- Table 4.7 Output for HCl in T01. (page 52)
- Table 4.8 Summary of errors when decision tree is used in the classification procedure. (page 53)



Table 4.9 Single coded results of Fessenden and Györgyi Acl12, Acl, Est1, Ket12, and Hcl. (page 54)

Table 4.10 My results for Acl12, Acl, Est1, Ket12, and Hcl. (page 55)

Table 4.11 Uncertain and bad classifications. (page 56)

Table 4.12 Total number of uncertain and bad classifications. (page 57)

## List of Illustrations

- Figure 1.1 Infrared spectrum of butyrophenone ( $C_6H_5COCH_2CH_2CH_3$ , Ket2).  
(page 4)
- Figure 1.2 An example for a binary decision tree. (page 7)
- Figure 1.3 Generic processing element (artificial neuron). (page 11)
- Figure 1.4 Schematics of a two-layer artificial neural network similar to the ones used in this study. (page 13)
- Figure 3.1 Wavenumber intervals used in this study to identify functional groups. The notation as follows: P1,P2 are the intervals for the identification of phenyl group; COC and CO are the intervals for the identification of carbon-oxygen single and double bond, respectively; OH-A and OH-C are the intervals for the identification of alcoholic and acidic O-H bond, respectively. (page 29)
- Figure 3.2 Typical -OH absorption bands in alcohols and carboxylic acids.  
(page 31)
- Figure 3.3 The binary decision trees used in this study to classify compounds. (page 40)
- Figure 4.1 Input data for the problematic compounds (A,B,C) in the study of the C-O-C absorption band that contain C-O-C bond. Est6 (D) was recognized correctly and is shown for comparison.  
(page 43)

Figure 4.2 Input data for the problematic compounds (A,B,C) in the study of the C-O-C absorption band that does not contain C-O-C bond. Ket7 (D) was recognized correctly and is shown for comparison. (page 45)

Figure 4.3 Input data for the problematic compounds (A,C) in the study of acidic -OH absorption band. Ac5 and Ac7 contain acidic -OH bond, while Hc11 and Hc6 do not. Ac7 (B) and Hc6 (D) do not. Ac7 (B) and Hc6 (D) were recognized correctly and are shown for comparison. (page 47)

Figure 4.4 Input data for the problematic compounds (A,C) in the study of the absorption bands of phenyl group that contain phenyl group. Est12 (B) and Hc6 (D) were recognized correctly and are shown for comparison. The break in the curves results from the use of the two separate wavenumber intervals. (page 49)

Figure 4.5 Input data for the problematic compound (A,C) in the study of the absorption bands of the phenyl group that do not contain phenyl group. Ac4 (B) and Alc2 (D) were recognized correctly and are shown for comparison. The break in the curves results from the use of the two separate wavenumber intervals. (page 50)

## 1. Introduction

Organic chemistry is the chemistry of carbon containing compounds. Since carbon atoms can be bonded in several different ways to each other and to many other atoms, the number of organic molecules is extremely large. There are several ways to study an organic compound. First the organic chemist determines the ratio of different elements in the compound this analysis is called elemental analysis. There are, however, many different compounds with the same ratio of elements. Thus it is necessary to gain structural information before the compound can be identified. Infrared spectroscopy is a technique used to get structural information about an unknown compound. The output of an infrared spectrophotometer is a graph that can be examined to learn about the compound. Since this process really is a pattern-recognition problem, an artificial intelligence approach could help humans.

In recent years a high interest in solving pattern-recognition problems has developed. Artificial neural networks, whose design have been influenced by biological neurons in the brain, perform superbly for a large class of pattern-recognition problems. Artificial neural networks are able to generalize from examples, learn from experience, and abstract important features from noisy data. Artificial neural networks are used in image processing, system controlling, mapping the human nervous system, etc. On the other hand, artificial neural networks are poor in computing with numbers, and thus can not be used for solving numerical problems.

### 1.1 Infrared Spectroscopy (Fessenden (1990 a))

Organic chemists use four major instrumental methods for structure determinations of unknown molecules : infrared (IR) spectroscopy, ultraviolet (UV) spectroscopy, nuclear magnetic resonance (NMR) spectroscopy, and mass spectrometry. IR and UV spectroscopy are based on the interactions of light with molecules. When light is passed through a sample the emergent light varies from compound to compound. Energy from the light is absorbed depending on, for example, the type of atoms, and/or type of bonds in the sample molecule. Molecules can be characterized by their absorption bands, which denote those wavelength (or frequency) regions where light absorption occurs. The shape of absorption bands is also important in most cases. In UV spectroscopy, absorption is due to changes in the electron configuration of the molecule, while infrared light is absorbed when vibrational modes of the molecule are changed.

### 1.1.1 Features of an Infrared Spectrum

The spectrum of a molecule is a graph of frequency or wavelength versus A or %T where:

$$A = \log(\text{original intensity}/\text{intensity})$$

$$\%T = (\text{intensity}/\text{original intensity}) * 100$$

(where intensity means the intensity of the light after the light passed through the sample, and original intensity is the intensity of the light before it is passed through the sample)

In this study %T and frequency were used. In this case the base line of the infrared spectrum is on the top of the graph where the sample did not absorb any light so the intensity equals the original intensity. An absorption peak or an absorption band can be seen on the infrared spectrum when the intensity of the radiation drops. This means the sample in the spectrophotometer absorbs the infrared light at that frequency. The position of an absorption band in the infrared spectrum can be expressed in  $\text{cm}^{-1}$ , which is called wavenumber, and is the usual frequency unit used in IR spectroscopy. The usual range of the spectrum is between  $4000 \text{ cm}^{-1}$  and  $625 \text{ cm}^{-1}$ . An example of an IR spectrum is shown in Figure 1.1.

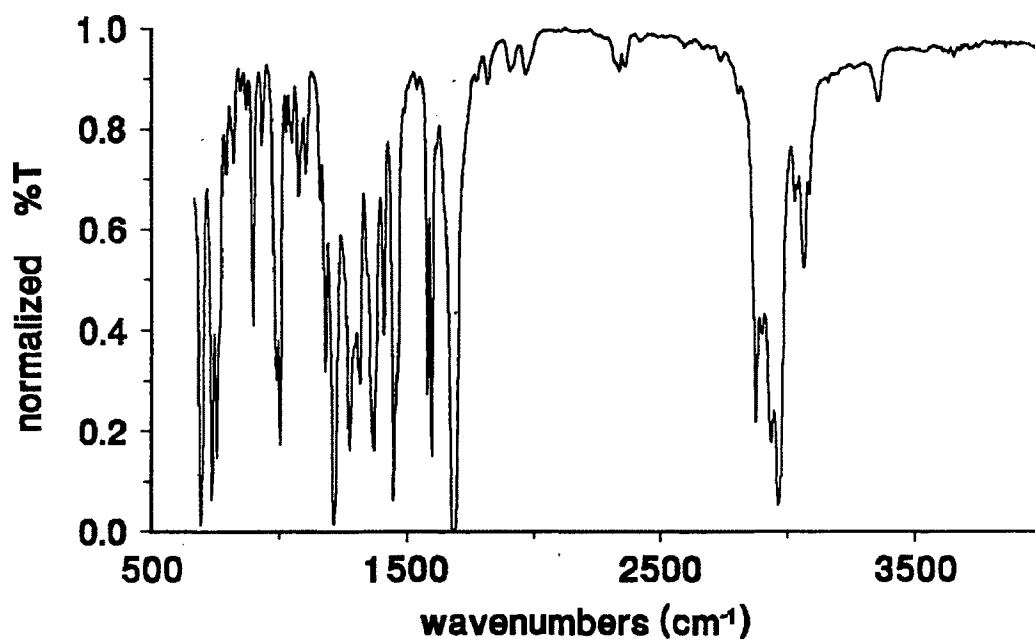


Figure 1.1. Infrared spectrum of butyrophenone ( $C_6H_5COCH_2CH_2CH_3$ , Ket2).

### 1.1.2 Interpretation of an Infrared Spectrum

Absorption bands in infrared spectra are characteristic of certain chemical bonds in the molecules of the sample. Some of the bonds may form reactive sites in the molecule, which are called functional groups. Experts in IR spectroscopy can structurally analyze an unknown molecule by just looking at its IR spectrum. These experts, because of years of experience, can recognize in an IR spectrum functional groups contained in the molecule because the presence of a functional group causes one or more peaks to appear in the spectrum. These peaks appear more or less at a given wavelength. There are a number of books available with correlation charts (Williams (1980)) to aid humans in interpreting IR spectra. Still, in case of a complicated molecule, peaks may overlap, the interactions of functional groups can cause a shift in the position of the peaks, other chemical effects can widen peaks, also the relative magnitude of the same peak, when compared to another, can vary from molecule to molecule. Furthermore, it is quite subjective how bands are defined and the books providing correlation charts do not satisfactorily agree in details.



## 1.2 Decision Tree (Weiss (1991)) and Corresponding Rules (Weiss (1991))

The decision tree technique can be very effectively used in partitioning samples into classes. A decision tree consists of nodes and branches just like an ordinary tree (see Figure 1.2). Each non-leaf node corresponds to a decision. Each branch below is labeled with a possible outcome of the decision. Depending on the possible outcomes of the decision, the tree will branch accordingly. In case of a binary decision tree, each non-terminal node has two branches, a false and a true branch. Depending on the decision, the tree will branch to the right or to the left. At each leaf of the tree a conclusion, i.e., a class assignment can be made. The characteristics of a binary decision tree are as follows: one branch enters each node except for the root, two branches leave each node, there are  $N-1$  non-terminal nodes, and  $N$  terminal nodes ( $N$  classifications can be made).

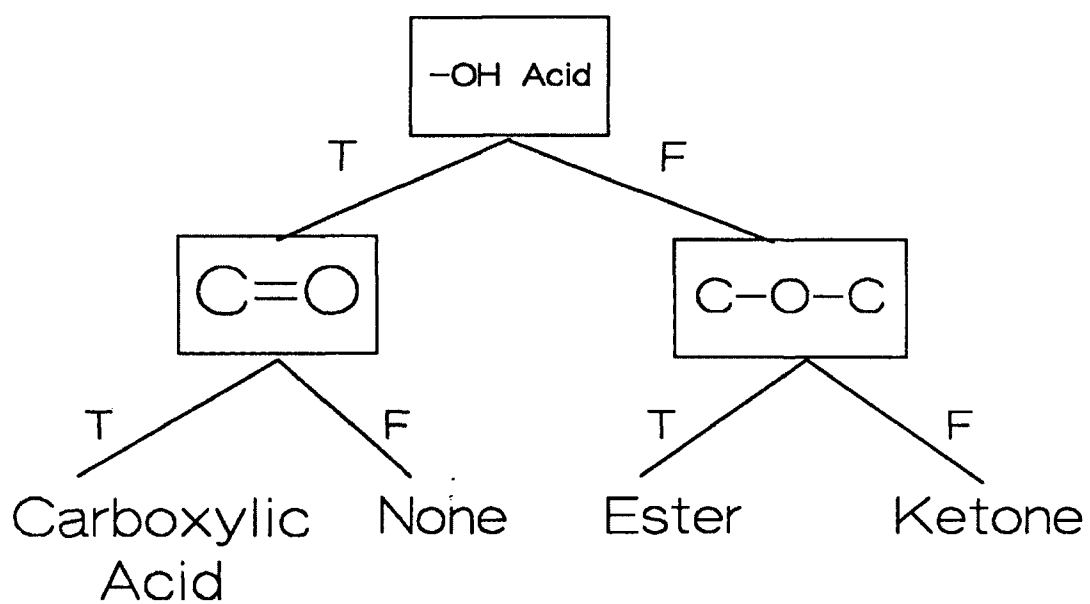


Figure 1.2. An example for a binary decision tree.

A decision tree can be transformed into an equivalent set of rules. The rule conditions are conjunctions of propositions each of which can be evaluated to be true or false.

The following is an example how to transform a classification done by the decision tree into an equivalent rule. First one has to follow the path of classification in the tree (each classification has a unique path), and use each node in the path as conjunctions of propositions in the condition of the equivalent rule. If the decision is to take the false branch in the tree the corresponding proposition in the condition is negated. The ester classification is done through the "-OH Acid" and "C-O-C" nodes. So the corresponding rule must contain "-OH Acid" and "C-O-C" as conjunctions of propositions in the condition of the rule. Since after the "C-O-C" proposition the false branch is taken, this proposition must be negated. The ester classification by a rule is then the following

```
IF NOT -OH Acid AND C-O-C THEN Ester.
```

In order to transform a decision tree into an equivalent set of rules one has to consider all possible classifications and for all classifications do the same process described in the example above. The following set of rules are equivalent to the decision tree in Figure 1.2.

```
IF -OH Acid AND C=O THEN Carboxylic Acid
```

```
IF -OH Acid AND NOT C=O THEN None
```

```
IF NOT -OH Acid AND C-O-C THEN Ester
```

```
IF NOT -OH Acid AND NOT C-O-C THEN Ketone
```

### 1.3 Expert Systems (Prerau (1990), Giarratano (1989))

Expert Systems (ES) are built so that they make use of specialized knowledge to solve problems at the level of a human expert. An ES consists of two major parts, a knowledge base and an inference engine. The knowledge base is usually built by knowledge engineers who, through intensive interview over a longer period of time, gain knowledge from a human expert. The knowledge base may not cover the problem domain completely so the result or the "expertise" of the ES may not give the correct answer to every possible problem. The inference engine, using the facts in the knowledge base, draws conclusions. The inference engine tries to copy the way human being solves problems. It works like a cognitive processor. Prerau (1990) defined ES as follows: "An advanced computer program that can, at a high level of competence, solve difficult problems requiring the use of expertise and experience; it accomplishes this by employing knowledge of the techniques, information, heuristic, and problem-solving processes that human experts use to solve such problems. Expert systems thus provide a way to store human knowledge, expertise, and experience in computers - that is, a way to clone human experts (at least to some degree)."

### 1.3.1 Benefits of an Expert System (Prerau (1990))

The advantages of using ES over using human experts are as follows:

- it provides expertise when it is scarce
  - it provides expertise when obtaining expertise is expensive
  - it provides expertise at times when experts are not available
  - it provides fast response
  - it provides steady solutions
  - it provides understandable and traceable solutions
  - it provides low cost availability
  - it provides permanent availability
- etc.

## 1.4 Neural Networks

Artificial neural networks are inspired by the neuron-level structure of the nervous system. Artificial neural networks exhibit brain like behaviors because they are able to learn and remember.

Hecht-Nielsen (1989) gave a definition of neural network as a computing system made up of a number of simple, highly interconnected processing elements, which processes information by its dynamic state response to external inputs.

Figure 1.3 shows an example for a generic processing element in a neural network.

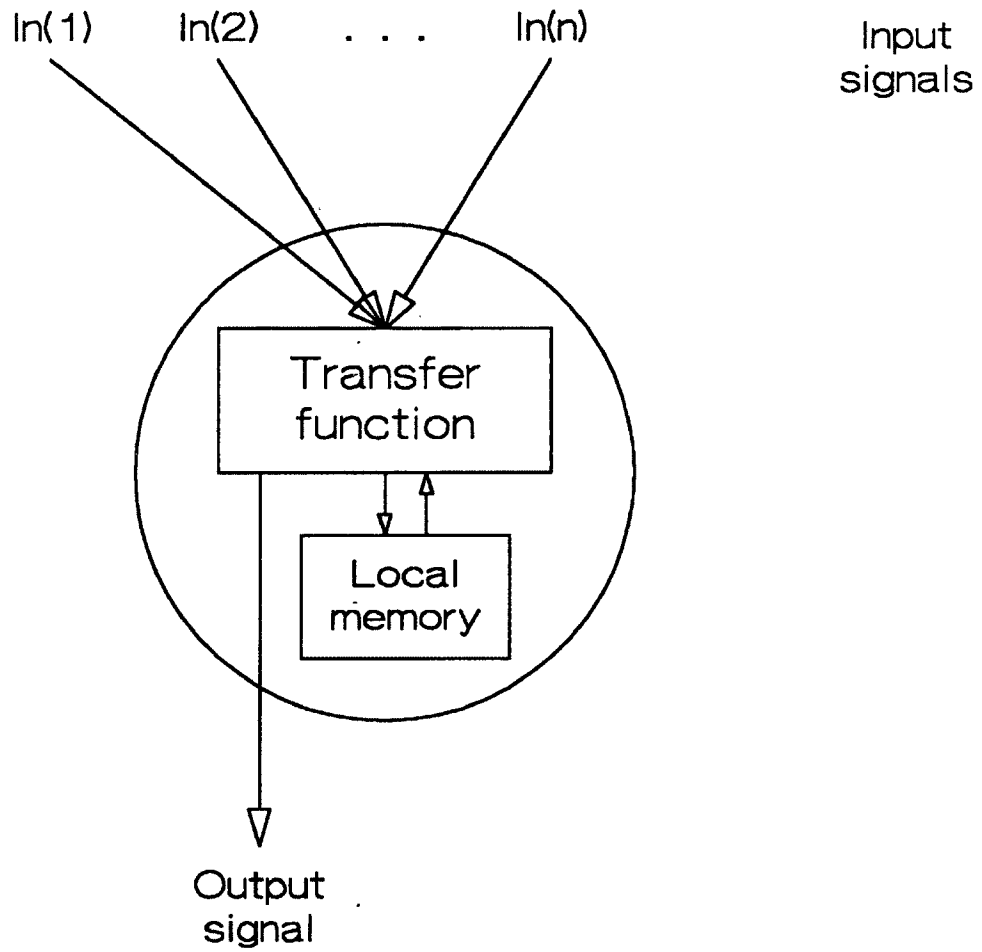


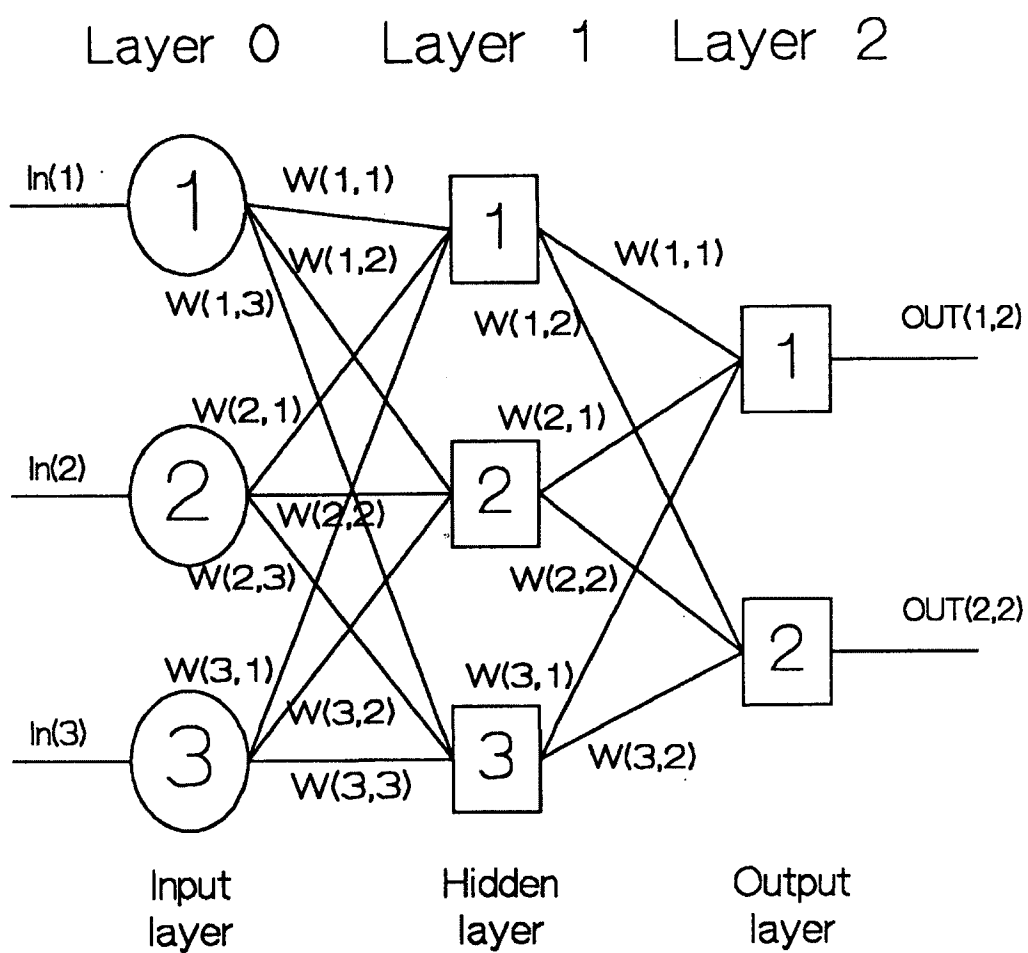
Figure 1.3. Generic processing element (artificial neuron).

A generic neural network processing element, an artificial neuron, has a number of input signals arriving via incoming connections. These input signals are then combined with the contents of the local memory of the neuron and the result is passed through the transfer function, characteristic to that processing element, to yield the output of the artificial neuron. There can be several input channels, but only one output for each processing element.

In this study the local memory is represented by the weights on the incoming connections. The incoming input signals are multiplied by these weights on the corresponding connections. Then these are summed and entered to the transfer function creating the output of a neuron.

#### 1.4.1 Network Architecture

A single-layer neural network consists of input and output neurons. The function of an input neuron is to distribute the input data. In the next layer there are the output neurons. From each input neuron there is a connection to each output neuron. Each connection is associated with a weight. The weights on the incoming connections constitute the local memory of that neuron.



**Figure 1.4.** Schematics of a two-layer artificial neural network similar to the ones used in this study.



Multilayer neural networks are formed by cascading single-layer neural networks after each other (see Figure 1.4). In this case the first layer is still called input layer, and the last one is called output layer while the middle ones are called hidden layers. This cascading typically is done so that each neuron at the  $k-1$  layer are connected with all neurons at layer  $k$ , and there is a weight assigned to each connection.

#### 1.4.2 Recurrent and Nonrecurrent Networks

Networks can also be categorized by the way information flows in the network. In case of nonrecurrent networks, or so called feedforward networks, there are no feedback connections. Connections only go from layer  $k$  to layer  $k+1$ , this way the output of the network is determined only by the current input and by the weights on the connections.

Recurrent networks are able to recirculate previous outputs to be inputs, so their output is determined by the current input and by the previous set of outputs.

#### 1.4.3 Training a Neural Network

The objective of training a network is to get the desired set of outputs for a given input. Algorithms to train a network can be categorized as supervised and unsupervised. Supervised training is done

by presenting a number of training pairs to the network. A training pair contains an input vector with a target (desired) set of output vector. For every input vector, the network calculates an output vector, this output vector is then compared to the target output vector. The difference is fed back and the weights on the connections are adjusted accordingly. This training process stops when the difference between the target and the calculated output is less than a given value or when a certain number of training runs have been completed.

Unsupervised training does not require a target output. Thus, it is usually considered to be a better model of learning of a biological system than supervised training. The training is conducted by presenting only input vectors to the network. This type of training algorithms change the weights so that for each similar training vector the output vector is the same. But there is usually no way to determine which output pattern will be produced for a certain input pattern set. The assignment of output pattern to input pattern can be done after training.

#### 1.4.4 Backpropagation Algorithm (Wasserman (1989))

The backpropagation learning algorithm belongs to the supervised training category. The goal of a training is to get the network to produce a desired set of outputs for given set of inputs. For every input data an output is calculated. Then this output is compared to a so called target or desired output and the differences fed back and the weights on the connections are adjusted accordingly. This training process stops

when the difference between the target output and the actual output becomes less than a given value or when a certain number of training cycles have been completed.

Following is an overview of the backpropagation learning algorithm:

Step 1. select the next training pair (containing an input and a desired output) from the training set; apply the input vector to the network input.

Step 2. calculate the output of the network

Step 3. calculate the error between the network output and the desired output

Step 4. adjust the weights of the network in a way that seeks to minimize this error

Step 5. repeat steps 1 through 4 for each vector in the training set until the error for the entire set is acceptably low.

A somewhat more detailed description of the backpropagation learning algorithm is as follows. Before the training starts the weights on the connections are initialized to small random numbers. Then the first input pattern is applied to the neural network. Using eq. (1) a NET value is calculated and then the output of a unit is calculated by applying a transfer function to the NET input, see eq. (2). The transfer function for backpropagation must be differentiable. The most common choice is the sigmoid function ( $F(x)=1/(1+\exp(x))$ ) as a transfer function. The output of the first layer is the input for the second layer. Thus eq. (1) and (2) are applied for every layer until the OUT value is calculated for the output layer. This concludes the forward pass in the neural network.

The goal of the reverse pass is to adjust the weights in the network

by propagating the error back through all of the layers. Eq. (3) shows the weight modifications to be add to the original weights in the network. The  $\delta$ 's are defined the following way. For neurons in the output layer the  $\delta$  is calculated by eq.(4) and for neurons in the hidden layer  $\delta$  is calculated by eq. (5). Notice that there is no target value for neurons in the hidden layer so  $\delta$  for a neuron in layer q-1 is calculated using the  $\delta$ 's from layer q weighted by the strength of the connection on which the error is propagated back. This procedure is repeated by applying new patterns to the network until the difference of the calculated output and the target output is less then a certain value.

The backpropagation algorithm is basically a gradient descent method for searching for the minimum of an appropriately defined error surface over the multidimensional weight-space.

### Backpropagation algorithm

Forward pass:

$$(1) \quad \text{net}_{pj} = \sum_i w_{ji} x_{pi}$$

$w_{ji}$  : weight between neuron j (in layer q+1) and neuron i (in layer q)

$x_{pi}$  : "input" associated with pattern p in neuron i

$\text{net}_{pj}$  : the net value associated with pattern p of neuron j

$$(2) \quad o_{pj} = F(\text{net}_{pj})$$

F : transfer function (  $F(x)=1/(1+e^{-x})$  )

$o_{pj}$  : the output of neuron j associated with pattern p

Reverse pass:

$$(3) \quad \Delta_p w_{ji} = \mu o_{pi} \delta_{p,j}$$

$\Delta_p w_{ji}$  : change in weight between neurons  $i$  and  $j$  where neuron  $i$  is in layer  $q$  and neuron  $j$  is in layer  $q+1$

$\mu$  : training constant (or learning rate) ( $0 \leq \mu \leq 1$ )

$o_{pi}$  : the output of neuron  $i$

$\delta_{p,j}$  : see below

$$(4) \quad \delta_{p,j} = (t_{p,j} - o_{p,j}) F'(net_{p,j}) \quad \text{for output neurons}$$

$\delta_{p,j}$  : delta value for neuron  $j$

$t_{p,j}$  : target value for output neuron  $j$  for pattern  $p$

$o_{p,j}$  : calculated value for output neuron  $j$  for pattern  $p$

$F'(net_{p,j})$  :  $F'$  denotes the derivative of  $F$  and

$$F'(net_{p,j}) = o_{p,j} (1 - o_{p,j})$$

$$(5) \quad \delta_{p,j} = \sum_k \delta_{p,k} w_{k,j} F'(net_{p,j}) \quad \text{for hidden neurons}$$

$\delta_{p,j}$  : delta value for neuron  $j$

$\delta_{p,k}$  : delta value for neuron  $k$

(neuron  $k$  is in layer  $q$  while neuron  $j$  is in layer  $q-1$ )

The momentum method (Wasserman (1989)) often enables the network to decrease the training time of the backpropagation algorithm. The momentum method adds a term to the weight adjustment of eq.(3) This term is proportional to the previous weight adjustment. The following equation is the modified eq. (3):

Momentum method:

$$(3') \quad \Delta_p w_{ji} = \mu o_{pi} \delta_{p,j} + \alpha (\text{previous } \Delta_p w_{ji})$$

$\alpha$  : momentum coefficient

#### 1.4.5 Justification for the Backpropagation Algorithm (McClelland (1988))

The backpropagation algorithm is a gradient descent method. The goal of the training is to minimize the error function:

$$\text{Error}(W,p) = \sum_j (0.5 (\text{target}(p,j) - o(p,j))^2)$$

W : represents all weights in the network

j : index of an output neuron

p : the current input pattern

This function, for each input pattern, defines a multidimensional surface (error surface) above the weight space. Each dimension of the weight space represents one weight, i.e., connection, in the network. Every point in this space is a possible state of the neural network. A point of the above defined error surface is the error for that particular state of the neural network at a given input pattern.

The backpropagation algorithm moves in the direction of the negative gradient of this surface at the current values of the weights, thus following the contour of the error surface always moving downhill in the direction of the steepest descent. In case of multilayer networks these error surfaces are quite complex and may have many minima. Some of the minima may constitute solutions to the problems in which the system reaches an errorless state. These minima are called global minima. The gradient descent method may fail to find a global minimum as it can get stuck in a local minimum from where every possible route is "uphill". It may also oscillate moving back-and-forth across a long and narrow "valley". Choosing an appropriately small learning rate helps to avoid

these oscillations. The momentum term is usually an improvement to the original backpropagation algorithm. It helps to filter out the high curvature and thus allows the effective weight steps to be bigger.

### 1.5 Brief Review of Existing Spectral Interpreters

The DENDRAL (Barr (1982)) is often considered to be the first expert system. It was built in the nineteen-sixties. This expert system tried to determine the molecular structure of an unknown organic compound by using its mass spectrum. The main idea was to create (by using a large set of rules) possible molecules from the results of the mass spectrum then to simulate the mass spectra of these possible molecules and compare the results to the spectrum of the unknown compound. Each rule in the knowledge base checked the "available" atoms and if there were enough each rule "created" different functional groups and they reduced the available atoms accordingly.

Fessenden and Györgyi (Fessenden (1990 b)) trained a neural network to identify functional groups in the infrared spectra of an unknown organic molecule. They used a two-layer neural network with the backpropagation learning algorithm. The whole spectrum of the molecule was presented to the neural network.

Robb and Munk (Robb (1990)) used a linear neural network for automated interpretation of infrared spectra. Their attempt was very similar to Fessenden (1990 b). They used a significantly larger set of compounds, but the neural network they used had no hidden layer and their

transfer function was linear. Their classification results were not satisfactory.

Meyer et al. (Meyer (1991)) used a neural network to identify complex organic compounds on the basis of their proton-NMR spectra. They used 13 compounds to train a network and they tested it with separately recorded spectra of the same compounds. In the lack of a larger training set they added fuzziness to their data to improve the robustness of their system.

Huixiao (1990) created ESSESA, an Expert System for Structure Euclidation by Spectral Analysis which contains a knowledge base of infrared spectra and an inference engine. The input to this expert system is given by a human, who, by looking at the spectrum, characterizes the appearance of peaks at certain places in the spectrum. This input is ambiguous. The inference engine of this expert system is a search tree which structurally classifies certain type of organic compounds.

There are several other earlier works on this subject. A most of them are listed in (Robb (1990)). Common characteristics of these systems that they use an inference engine to obtain structural knowledge from spectral information which was previously processed by a human expert.



## 2. Objective

There are a number of ways to interpret an infrared spectrum. Human experts use correlation charts. These charts are made by people who have a deep understanding of IR spectra. There are also computer programs and systems available that can interpret IR spectra. Most of these systems, like Huixiao (1990), need input from a human, who, by looking at the spectrum, characterizes the appearance of peaks at certain places in the spectrum. This input is ambiguous even when all precautions are taken.

Using neural networks, which are known to be good at pattern recognition and generalization, should be a good technique to avoid the possible ambiguity of this input.

The first goal of my thesis is to set up a prototype expert system for interpreting IR spectra by incorporating the human knowledge accumulated in this area and by using neural networks to do the otherwise ambiguous classification of peaks. The objective of the prototype expert system is to classify unknown molecules into classes. This approach is a combination of the expert system methodology and the neural network methodology. An evaluation of this approach can be done by comparing it to a method for the interpretation of infrared spectra which is based completely on a neural network methodology.

Fessenden and Györgyi used a two-layer neural network with the backpropagation learning algorithm to obtain structural information of unknown organic compounds. They presented the whole IR spectra of a selected group of organic compounds as input and their classifications as

a target output. Then they tested the trained neural network with unknown compounds (meaning that these compounds were not used in the training so they were unknown for the neural network).

My approach is to use neural network classification only on those portions of the spectrum that contain information about absorption bands of interest. Thus the knowledge in IR correlation charts can be utilized. A decision tree or a set of rules can then be used to combine and evaluate the information obtained from the individual neural networks.

The second goal of the thesis is to test the hypothesis that the combined approach described above is superior to the only-neural-network method if the correct intervals of the spectra are used for the identification of the absorption bands of interest. The accuracy of the two systems on the same compounds provided the basis for a comparison.

## 2.1 A Comparison of the Conditions in the Two IR Interpretation Methods

### Similarities

Data. I used the same forty-eight spectra as Fessenden and Györgyi did. A listing of the compounds can be found in Sec. 3.1. I used the same normalized input. Fessenden and Györgyi provided me with their training and testing sets.

Neural Network. In both cases two-layer feedforward neural networks with the backpropagation learning algorithm were used.

Evaluation of network output. The same criteria was used to evaluate the accuracy of the neural network performances (more details are

in Sec. 4.3).

### Differences

Data. While Fessenden and Györgyi presented the whole spectra to the neural network, for my approach each spectrum was cut into several intervals, and only those intervals where the correlation charts indicated the presence of a peak of interest were presented to the neural network.

Neural Network. The number of neurons used in the neural networks for the two approaches were substantially different. Fessenden and Györgyi used 250 input neurons, eighteen hidden neurons, and six output neurons. In my approach there was different number of input neurons for each interval studied, five neurons in the hidden layer, and one neuron in the output layer in all cases.

Functional group identification. Fessenden and Györgyi simply used the output of their trained network for this purpose. In my approach the output of many neural networks were combined and evaluated by a decision tree or by rules.

### 3. Methodology

#### 3.1 Data

Forty-eight different compounds were selected for this study. These compounds were chosen so that they belong to five different classes of compounds, namely alcohols, ketones, esters, hydrocarbons, and carboxylic acids. Every compound in the study had infrared active carbon-hydrogen bonds; therefore, I will not specifically mention this bond in the following discussion. Other bonds contained in the different compound classes are as follows:

alcohol	ketone	carboxylic acid	esters	hydrocarbon
-OH	C=O	-OH	C-O-C	none
		C=O	C=O	

Table 3.1 Bonds and classes used in this study

At least four out of ten compounds from the alcohols, esters, ketones, hydrocarbons contained a phenyl group. None of the carboxylic acids contained a phenyl group.

#### Alcohols

benzyl alcohol *	Alc1	$C_6H_5CH_2OH$
1-butanol	Alc8	$CH_3CH_2CH_2CH_2OH$
2-butanol	Alc2	$CH_3CH_2CH(OH)CH_3$
cyclohexanol	Alc5	$C_6H_{11}OH$
1-hexanol	Alc3	$CH_3(CH_2)_4CH_2OH$
isoamyl alcohol	Alc4	$(CH_3)_2CHCH_2CH_2OH$
isobutyl alcohol	Alc9	$(CH_3)_2CHCH_2OH$
<u>sec</u> -phenethyl alcohol *	Alc12	$C_6H_5CH(OH)CH_3$
2-phenoxyethanol *	Alc11	$C_6H_5OCH_2CH_2OH$
3-phenyl-1-propanol *	Alc10	$C_6H_5CH_2CH_2CH_2OH$

<b>Carboxylic Acids</b>		
acetic acid	Ac5	$\text{CH}_3\text{CO}_2\text{H}$
heptanoic acid	Ac4	$\text{CH}_3(\text{CH}_2)_5\text{CO}_2\text{H}$
hexanoic acid	Ac8	$\text{CH}_3(\text{CH}_2)_4\text{CO}_2\text{H}$
isobutyric acid	Ac1	$(\text{CH}_3)_2\text{CHCO}_2\text{H}$
octanoic acid	Ac6	$\text{CH}_3(\text{CH}_2)_6\text{CO}_2\text{H}$
oleic acid	Ac3	$\text{CH}_3(\text{CH}_2)_7\text{CH}=\text{CH}(\text{CH}_2)_7\text{CO}_2\text{H}$
pentanoic acid	Ac7	$\text{CH}_3(\text{CH}_2)_3\text{CO}_2\text{H}$
propanoic acid	Ac2	$\text{CH}_3\text{CH}_2\text{CO}_2\text{H}$
<b>Esters</b>		
benzyl benzoate *	Est12	$\text{C}_6\text{H}_5\text{CO}_2\text{CH}_2\text{C}_6\text{H}_6$
n-butyl acetate	Est1	$\text{CH}_3\text{CO}_2\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_3$
ethyl acetate	Est6	$\text{CH}_3\text{CO}_2\text{CH}_2\text{CH}_3$
ethyl benzoate *	Est3	$\text{C}_6\text{H}_5\text{CO}_2\text{CH}_2\text{CH}_3$
ethyl propionate	Est4	$\text{CH}_3\text{CH}_2\text{CO}_2\text{CH}_2\text{CH}_3$
isopentyl acetate	Est8	$\text{CH}_3\text{CO}_2\text{CH}_2\text{CH}_2\text{CH}(\text{CH}_3)_2$
isopropyl acetate	Est10	$\text{CH}_3\text{CO}_2\text{CH}(\text{CH}_3)_2$
methyl benzoate *	Est2	$\text{C}_6\text{H}_5\text{CO}_2\text{CH}_3$
phenylethyl acetate *	Est7	$\text{CH}_3\text{CO}_2\text{CH}_2\text{CH}_2\text{C}_6\text{H}_5$
n-propyl acetate	Est5	$\text{CH}_3\text{CO}_2\text{CH}_2\text{CH}_2\text{CH}_3$
<b>Hydrocarbons</b>		
cyclohexane	Hc2	$\text{C}_6\text{H}_{12}$
cyclopentane	Hc1	$\text{C}_5\text{H}_{10}$
ethylbenzene *	Hc5	$\text{C}_6\text{H}_5\text{CH}_2\text{CH}_3$
n-heptane	Hc9	$\text{CH}_3(\text{CH}_2)_5\text{CH}_3$
n-hexane	Hc10	$\text{CH}_3(\text{CH}_2)_4\text{CH}_3$
isopropylbenzene *	Hc6	$\text{C}_6\text{H}_5\text{CH}(\text{CH}_3)_2$
n-octane	Hc7	$\text{CH}_3(\text{CH}_2)_6\text{CH}_3$
petrolatum	Hc3	$\text{CH}_3(\text{CH}_2)_x\text{CH}_3$
polystyrene *	Hc11	$-\text{[CH}_2\text{CH}(\text{C}_6\text{H}_5)]_x-$
toluene *	Hc8	$\text{C}_6\text{H}_5\text{CH}_3$
<b>Ketones</b>		
acetophenone *	Ket3	$\text{C}_6\text{H}_5\text{COCH}_3$
acetone	Ket10	$\text{CH}_3\text{COCH}_3$
2-butanone	Ket1	$\text{CH}_3\text{COCH}_2\text{CH}_3$
butyrophenone *	Ket2	$\text{C}_6\text{H}_5\text{COCH}_2\text{CH}_2\text{CH}_3$
cyclohexanone	Ket5	$\text{C}_6\text{H}_{10}\text{O}$
cyclopentanone	Ket6	$\text{C}_5\text{H}_8\text{O}$
2-heptanone	Ket4	$\text{CH}_3\text{CO}(\text{CH}_2)_4\text{CH}_3$
2-octanone	Ket7	$\text{CH}_3\text{CO}(\text{CH}_2)_5\text{CH}_3$
phenylacetone *	Ket13	$\text{C}_6\text{H}_5\text{CH}_2\text{COCH}_3$
propiophenone *	Ket12	$\text{C}_6\text{H}_5\text{COCH}_2\text{CH}_3$

where \* denotes compounds containing a phenyl group.

Table 3.2 Compounds used in this study

## 3.2 Preparation of Data

### 3.2.1 Spectra

The forty-eight different sample compounds listed above were used to obtain the forty-eight spectra. A Perkin-Elmer 1600 series FT-IR instrument was used to obtain the spectra. No special care was taken either for purification or for obtaining the spectra. The spectrum of a sample contained 1668 data points, where each point was represented by a pair of values (cm-1, %T). The Perkin-Elmer Corporation supplied us with a BASIC program that allowed the conversion of the digitized IR spectra to ASCII format. Dr. Ralph Fessenden obtained the spectra and made the conversions.

### 3.2.2 Normalization

The absorption unit was normalized the following way:

$$\%T'(i) = (\%T(i) - \min \%T) / (\max \%T - \min \%T)$$

where

i	: i = 1..1668
%T'(i)	: normalized point
%T(i)	: data point
min %T	: minimum %T value
max %T	: maximum %T value.

### 3.2.3 Sampling

Books about IR spectroscopy usually contain charts specifying where to look for an absorption band for a given functional group. These charts indicate that a peak for an absorption band will be in a certain interval. They also suggest what kind of a peak to expect. For example, the H-bonded -OH peak between  $3600\text{-}3200\text{ cm}^{-1}$  is described as a "strong often broad but may be sharp" (Williams (1980)).

Looking at the graphs of the normalized %T value and considering the intervals for each functional group, it is clear that a different sampling frequency is needed for each absorption band, see Figure 3.1. For example, the C=O absorption band is found in the  $1725\text{-}1700\text{ cm}^{-1}$  interval and is described as a strong peak, while the -OH absorption band is in  $2700\text{-}2500\text{ cm}^{-1}$  interval and wide, see Figure 3.1.

Functional group	Band	Type
C=O	$1725\text{-}1700\text{ cm}^{-1}$	strong
-OH	$2700\text{-}2500\text{ cm}^{-1}$	wide

Table 3.3 C=O and -OH absorption bands

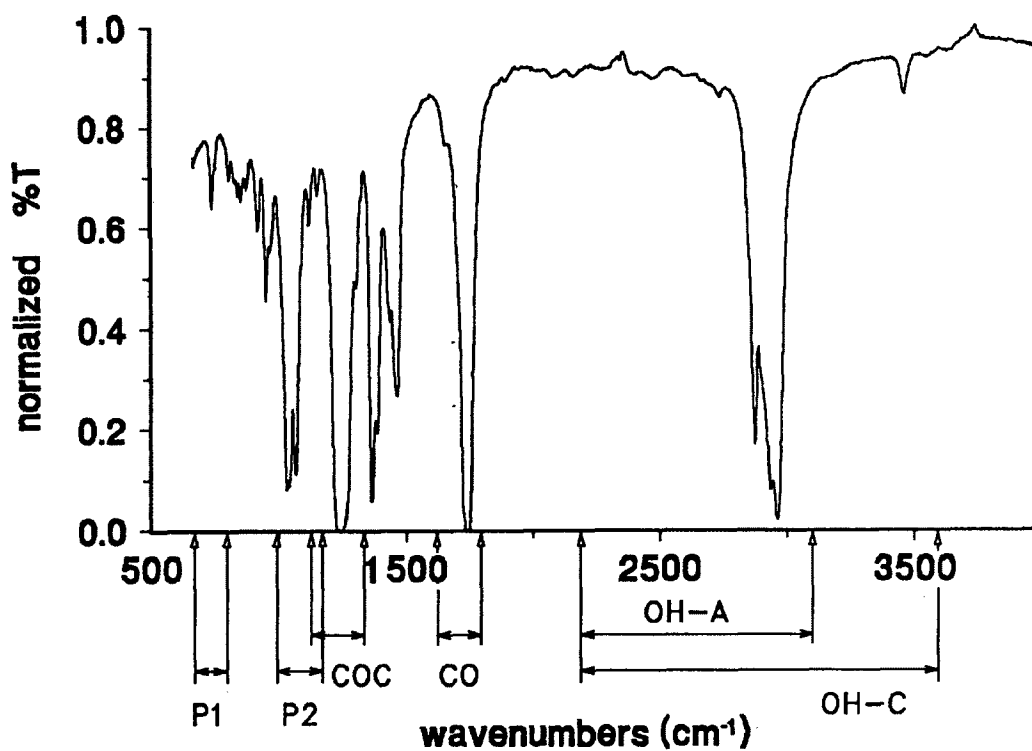


Figure 3.1. Wavenumber intervals used in this study to identify functional groups. The notation is as follows: P1,P2 are the intervals for the identification of phenyl group; COC and CO are the intervals for the identification of carbon-oxygen single and double bond, respectively; OH-A and OH-C are the intervals for the identification of alcoholic and acidic O-H bond, respectively.



The IR interpretation charts list those intervals that contain the tip of the absorption peak for a given chemical bond. While I was doing preliminary studies I found that these intervals are not sufficiently wide for training a neural network. Using graphics from a spreadsheet program, I determined the width of the interval and the number of points within the interval necessary to successfully represent an absorption band for the network. These parameters are listed below.

Functional group	Number of points	Interval (cm-1)
-OH (acid)*	28	2200-3600
-OH (alcohol)*	41	3100-3600
C=O	29	1625-1800
C-O-C	35	1340-1130

\*: see section 3.2.4

Table 3.4 Sampling frequency

### 3.2.4 Hydroxyl Group in Alcohols and in Carboxylic Acids

The width of the hydroxyl absorption band in alcohols and in carboxylic acids is different. The shape of these peaks in the spectra are also different, as it is shown on Figure 3.2. Both peaks are strong but the peak of carboxylic acid hydroxyl band is much wider.

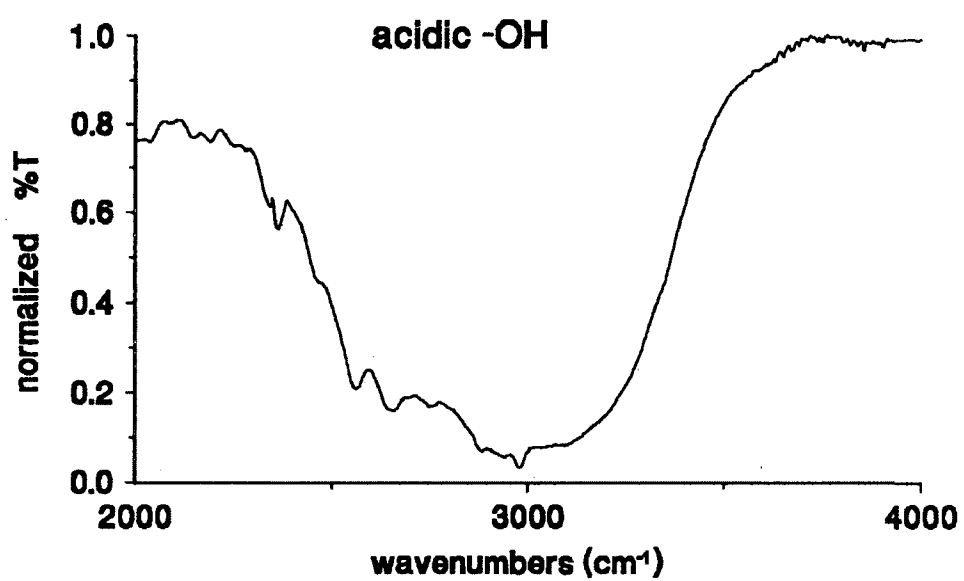
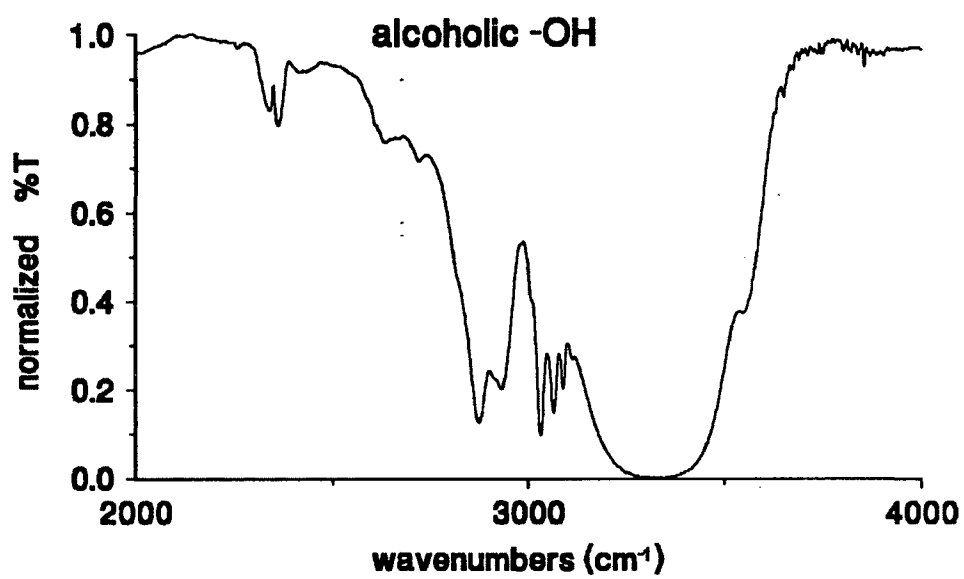


Figure 3.2. Typical -OH absorption bands in alcohols and carboxylic acids.

### 3.2.5 Phenyl

A benzene ring contains 6 carbon atoms, joined in a ring (Fessenden (1990 a)). When benzene is a substituent in an organic molecule, it is called phenyl. The phenyl group strictly speaking is not a functional group, but in this study it was used that way because it has characteristic IR absorption. In case of the phenyl, two major absorption intervals were considered. There is a problem with the phenyl group, in that both absorption bands fall into the so called fingerprint region of the spectrum. The fingerprint region is from  $1400\text{ cm}^{-1}$  to  $625\text{ cm}^{-1}$ . In this region correlations between absorption bands and functional groups can not be made unambiguously. The reason for this is that several vibrational modes give rise to absorption in the fingerprint region while in most part of the spectrum only one vibrational mode absorption is observed (Fessenden (1990 a)).

In the preliminary studies, I have found that training two neural networks separately on this two intervals produced poor results. The networks were not able to recognize phenyl with sufficient accuracy. When I changed the representation in such a way that the intervals were attached to one another, as indicated below, the results improved substantially.

Functional group	Number of points	Interval (cm <sup>-1</sup> )
phenyl-1	30	670-790
phenyl-2	31	990-1175
phenyl	61	670-790 and 990-1175

where:

phenyl-1 and phenyl-2 are the two major absorption band for the phenyl group, and phenyl is when phenyl-1 and phenyl-2 were combined.

Table 3.5 Sampling frequency for phenyl

### 3.2.6 Hydrocarbons

All compounds used in this study are hydrocarbons. However, following the notation used by Fessenden and Györgyi only those compounds which do not contain -OH, C-O-C, or C=O functional groups are named as "hydrocarbon" in this work. Because of this definition no separate absorption band testing is needed for "hydrocarbons".

### 3.2.7 Summary of the Input Data

The following table gives an overview of which class of samples contain what kind of infrared absorption bands:

	C=O	-OH (Alcohol)	C-O-C	-OH (Acid)	Phenyl
Alcohol	0	1	0	0	0/1
Carboxylic Acid	1	0	0	1	0
Ester	1	0	1	0	0/1
Ketone	1	0	0	0	0/1
Hydrocarbon	0	0	0	0	0/1

where 1 stands for the presence of a functional group; 0 stands for the absence of a functional group; 0/1 means that certain compounds contain and others do not contain that functional group from the same class of compounds

Table 3.6 Absorption bands appearing in the classes studied

### 3.3 Neural Network Program

A publicly available two-layer neural network using the backpropagation algorithm, written by Josiah Hoskins 1987 in the C language provided the basis of this work. This program originally was dedicated to one problem. It was extended by Kevin Lohn, and further extended by me.

#### 3.3.1 Major Features of the Program

The program can work with up to seventy input neurons, fifteen hidden neurons, and seven output neurons. The largest number of input patterns I tried was thirty-nine.

The current program uses an improved random number generator (Park (1988)) to initialize the weights on the connections in the neural

network.

The weights with the original random number generator were (0,1) interval. If the weights on the connections are positive, and in this study the input for all patterns are positive, the result of eq. (1) will be positive and also this result might be a large value especially if the initial weights were "big". If this happens, the OUT value of eq. (2) will be very close to one, which means that delta of eq.(3) and eq.(4) will be close to zero, and it means that the weights on the network will hardly change. This phenomenon is called network paralysis (Wasserman (1989)). In order to avoid network paralysis, the random numbers were generated in the [-0.5,0.5) interval. This ensures that the network does not become saturated with large values of weights (Wasserman (1989)).

The program allows the user to set the learning rate.

A tolerance can also be set by the user to test whether for a given pattern the difference between the target output and the calculated output is less than the tolerance.

An upper limit for the number of training cycles can also be given by the user. This limit, or the previously mentioned tolerance, can be used to stop the training process.

A new feature of the program I implemented is the so called epoch training (McClelland (1988)). Using epoch training, the weight adjustments are done after a certain number of training patterns have been cycled through the neural network. The delta values of eq. (3) and eq. (4) are calculated after each training cycle and they are accumulated. Eq.(7) and eq.(6) are calculated with the accumulated delta values at the end of the epoch. This feature of the program was not used in this study,

because, in the preliminary work that I did there was no indications of any kind of major improvement using any values larger than one.

### 3.4 Experimental Design

#### 3.4.1 Data

The data were separated into two parts: a training set and a testing set. It is important to separate the data into these two parts because if the training and the testing set is the same then we make no mistake in the testing procedure thus overfitting of the data can not be detected. Overfitting or overspecialization occurs in classifiers when they perform well for their training data but the performance dramatically worsens when new, previously unseen, data is presented to the system. The chances are small in real life that any new data is identical to one of those which were in the training set. In case of neural networks it desired that the neural network generalizes from the training set instead of memorizing it. If the testing set contains different data from the training set the user can get a better estimate of the true error rate.

There were fifteen random selections from the forty-eight compounds used for the training. In each selection thirty-nine out of the forty-eight were used for training a neural network. The remaining nine compounds were used for testing. The five absorption bands were handled separately. For each training set, three randomly selected seeds were used to initialize the weights on three neural networks. This means that

there were fifteen random selections of training data for the five absorption bands, and three neural networks were used for each training which made it to be a total of two hundred and twenty-five trained neural networks.

The selection of data was done so that each possible chemical group studied was represented in the training set and in the testing set for all fifteen selections. This means that among the thirty-nine training compounds there were always five alcohols, esters, hydrocarbons, and ketones without a phenyl; three alcohols, esters, hydrocarbons, and ketones with a phenyl; and seven carboxylic acids. In the testing set there were always one alcohol, ester, hydrocarbon and a ketone without a phenyl; one alcohol, ester, hydrocarbon, and a ketone with phenyl plus one carboxylic acid.

This selection of data was done exactly the same way as Fessenden and Györgyi did in their work.

#### 3.4.2 Training Parameters

The tolerance, that is the desired difference between the target output and the calculated output, was set to be 0.1 for all training patterns.

The learning rate was set to 0.5 for all training. This learning rate seemed to be the optimal because the convergence was not slow, and network paralysis (i.e. the weights on the connections became very large) occurred approximately ten times out of the two hundred and twenty-five



training. In other words, training was usually stopped when the error at the output neuron became less than 0.1. If the error did not become smaller than the desired tolerance after more than 2000 iterations were performed the training was halted, and a new training was started with a new seed for the random number generator which always solved the problem.

The momentum coefficient was always set to 0.9.

Weight adjustments were done after each training pattern was presented to the network (ie. the size of an epoch was set to one).

#### 3.4.3 Testing

In each random selection, the testing compounds were chosen so that they represented all possible types. Two test compounds were selected from alcohols, esters, hydrocarbons, and ketones, so that only one out of the two contained a phenyl group. There was only one selected from the carboxylic acids, since no carboxylic acid contained phenyl group. Each of the nine compound was tested for all absorption bands. The total number of tests was nine times two hundred and twenty-five, meaning that each neural network was tested with 9 compounds.

### 3.5 Decision Tree and Rules

After interviewing Dr. Fessenden, an organic chemist, about how to evaluate an infrared spectrum, I could set up Table 3.6, and then from this table the following decision tree can be built, see Figure 3.3. The phenyl decision is not included in the upper tree because it can be handled as a separate problem. When the classification of the organic compound is done it can be decided whether it contains a phenyl group or not.

These binary decision trees are equivalent to the following set of rules:

- (R1) IF C=O and OH (Acid) THEN Carboxylic Acid
- (R2) IF C=O and not OH (Acid) and C-O-C THEN Ester
- (R3) IF C=O and not OH (Acid) and not C-O-C THEN Ketone
- (R4) IF not C=O and OH (Alcohol) THEN Alcohol
- (R5) IF not C=O and OH (Alcohol) THEN Hydrocarbon
- (R6) IF Phenyl THEN Compound\_contains\_phenyl
- (R7) IF not Phenyl THEN Compound\_Does\_not\_contain\_phenyl

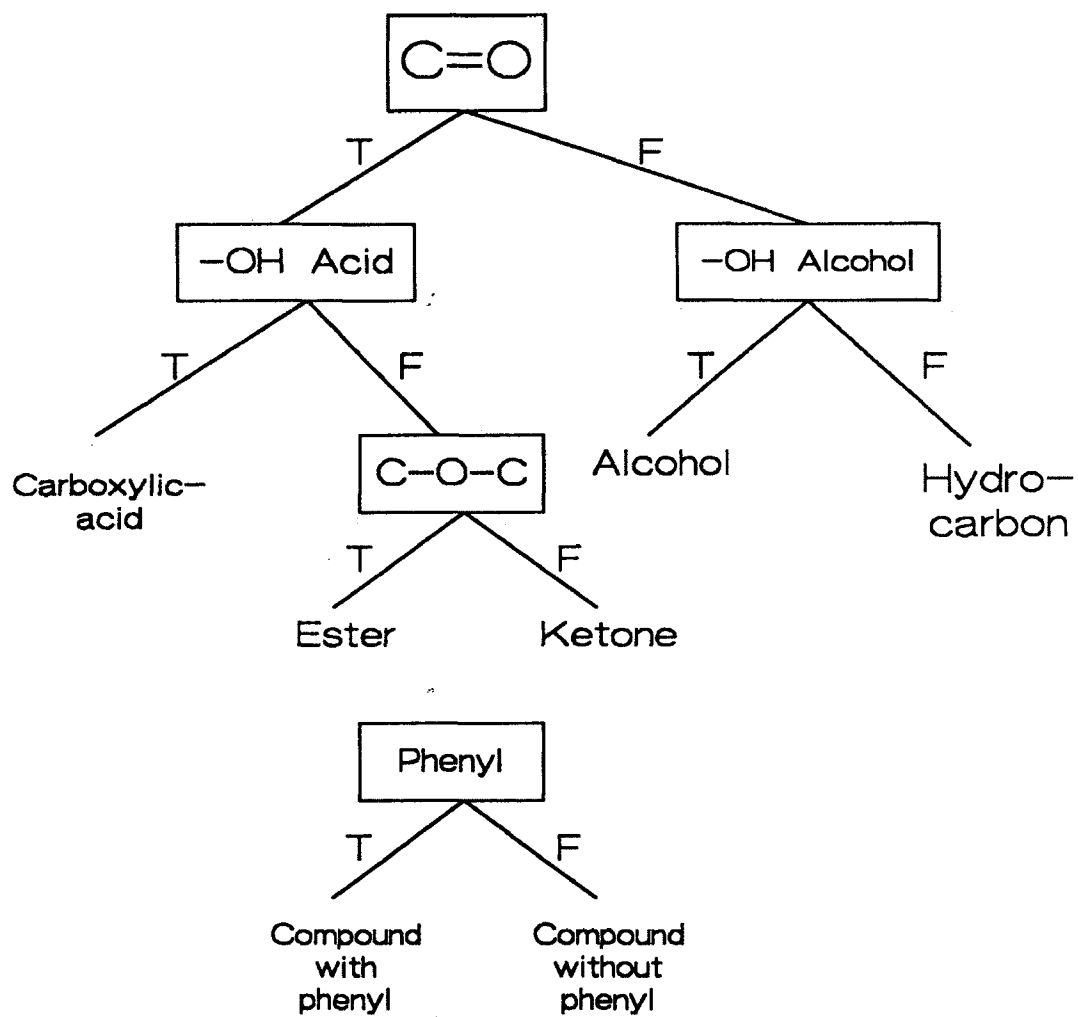


Figure 3.3. The binary decision trees used in this study to classify organic compounds.

#### 4. Results

The testing results were categorized by using the OUT value of the output neuron. If OUT was within 0.4 of the desired value it was classified as good. If the OUT value was between 0.6 and 0.4 it was classified as uncertain. In any other case it was classified as bad. See the Appendix for a detailed listing of the tests results.

Each absorption band was tested four hundred and five times (since there were nine test compounds, fifteen training set, and three networks). The -OH absorption band for alcohols, and the C=O were correctly recognized for all testing compounds. The -OH absorption band for the carboxylic acids were uncertain six times and bad once. The C-O-C absorption band was bad twenty-five times, and uncertain eight times. The phenyl absorption band was bad twenty-two times and uncertain seven times. Following is a table to summarize the errors:

	Good	Uncertain	Bad
C=O	405 (100%)	0 (0%)	0 (0%)
C-O-C	372 (92%)	8 (1.9%)	25 (6.1%)
OH(Alcohol)	405 (100%)	0 (0%)	0 (0%)
OH(Acid)	398 (98.4%)	6 (1.4%)	1 (0.2%)
Phenyl	376 (93%)	6 (1.4%)	23 (5.6%)
Sum	1956 (96.68%)	20 (0.94%)	49 (2.38%)

Table 4.1 Summary of errors

#### 4.1 Bad Classifications

C-O-C absorption band. Table 4.2 shows the performances of the neural networks on problematic compounds in the C-O-C absorption band.

Compound	Good	Uncertain	Bad
Est12	17/24	6/24	1/24
Est2	3/9	0	6/9
Est4	3/9	0	6/9
Ket3	0	0	9/9
Ket6	3/6	0	3/6
Ac1	1/3	2/3	0

**Table 4.2** Problematic compounds in the C-O-C absorption band. Each quotient indicates the ratio of the number of good, uncertain, or bad classifications of a certain testing compound over the total number that compound was in the testing set. These compounds were chosen randomly thus they were tested different number of times.

Figure 4.1 may provide some insight for the reason of these errors. Most esters in the C-O-C absorption band have very similar peaks to the peak Est6 has. This characteristic peak of the C-O-C absorption band is shifted to the left for Est12 and Est2. Est12, Est2 and Est3 (not shown) are structurally similar molecules (see Sec. 3.1). It might easily occur in the spectra as a shift in the absorption peak, most probably this is the case.

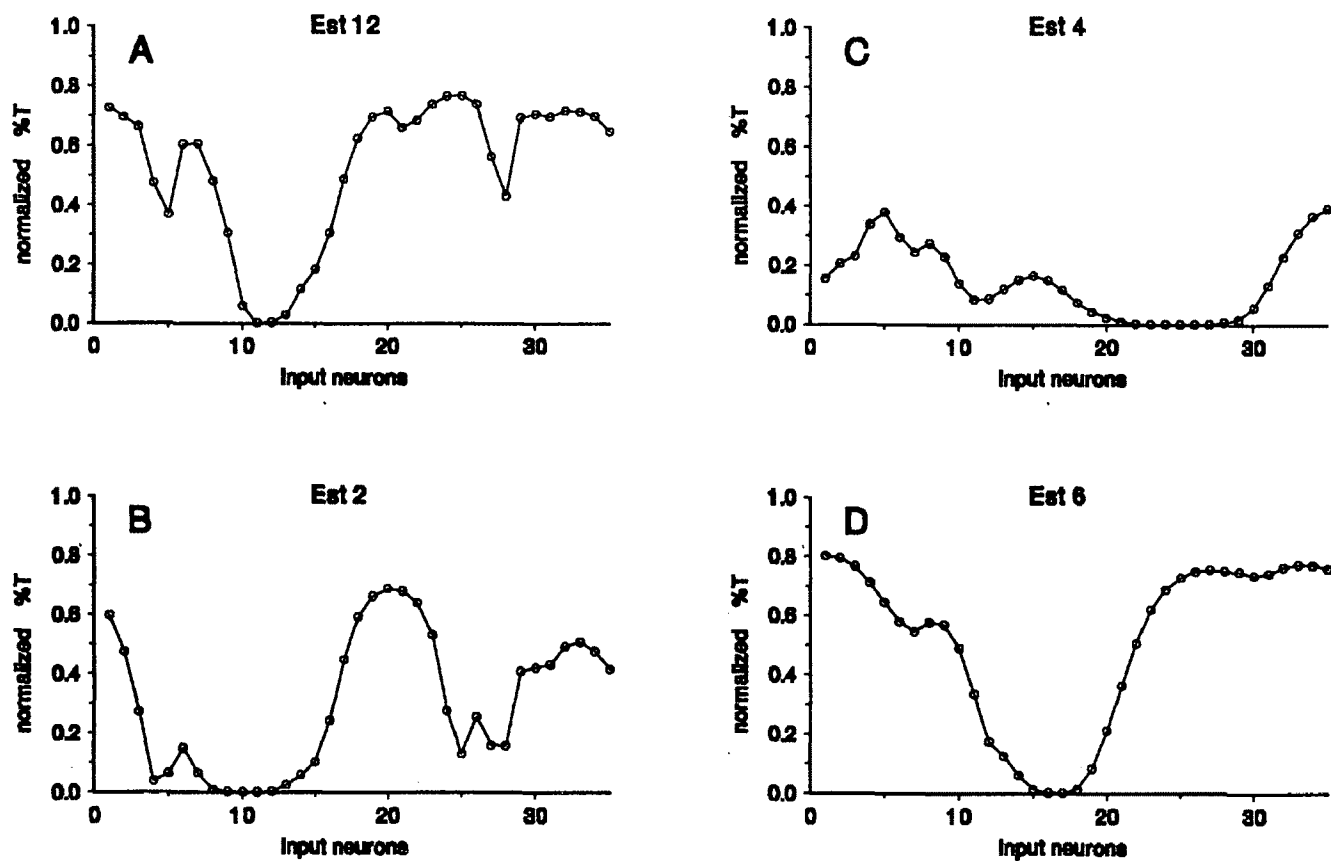


Figure 4.1. Input data for the problematic compounds (A,B,C) in the study of the C-O-C absorption band that contain C-O-C bond. Est6 (D) was recognized correctly and is shown for comparison.

One would say that Est4 is a "bad" spectrum. One reason why a spectrum like Est4 is called "bad" is as follows : if the concentration of the sample is high most of the IR light is absorbed, this usually results a spectrum which is flat and distorted meaning that the details are very poorly resolved. Besides Est4 is a structurally different compound from the rest in the training and in the testing set. It is difficult to say whether it is the structural difference or the high concentration of Est4 which caused this poor spectrum.

In Figure 4.2 I show some problematic compounds not containing C-O-C bond. Ket3 clearly has an ester contamination and this is why it has a peak in the C-O-C absorption band. Ket6 is tested two times, once in T10 and then in T12 (see Appendix). In case of the T10 training, Est2 and Est4 (already known as problematic compounds) were in the testing set which means that mainly "good" compounds were in the training so a slightly bad spectrum could pass the test. This was not the case in the T12 training. Est2 and Est4 were in the training set with "good" esters like Est6. It is hard to find what the neural networks picked as common features of these spectra. It is not possible to say anything about uncertain classifications of Acl.

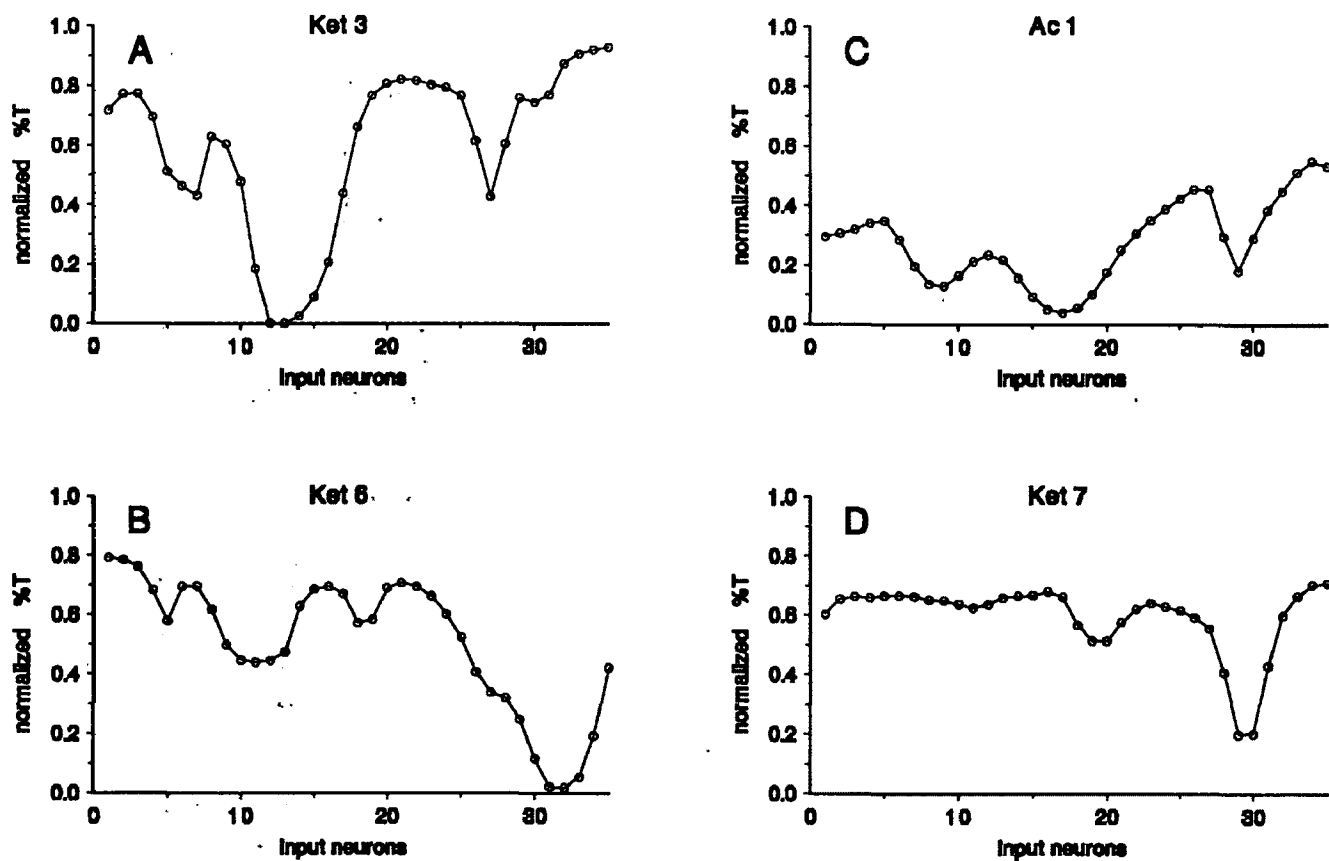


Figure 4.2. Input data for the problematic compounds (A,B,C) in the study of the C-O-C absorption band that does not contain C-O-C bond. Ket7 (D) was recognized correctly and is shown for comparison.



Carboxylic acid -OH absorption band. The following is a table to show the performances of the neural networks for the carboxylic acid -OH absorption band.

Compounds	Good	Uncertain	Bad
Hc11	11/12	0	1/12
Ac5	0	3/3	0

**Table 4.3** Problematic compounds in the carboxylic acid -OH absorption band

Figure 4.3 shows a comparison of the poorly classified compounds to two well classified compounds of similar structure. The problem arising of the low concentration of the sample when the spectrum is taken might be a reason for the uncertain classification of Ac5. Hc11 was classified as bad only once out of twelve. It is hard to find a reason for this wrong classification.

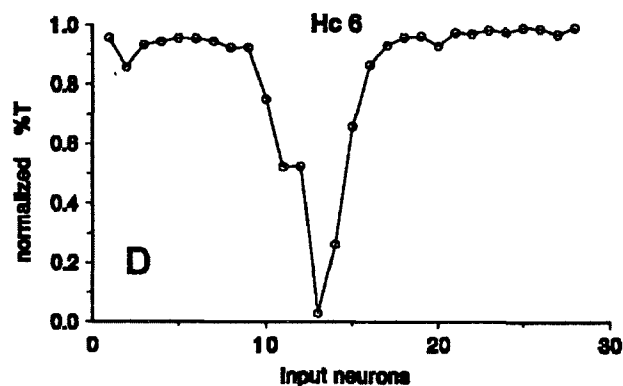
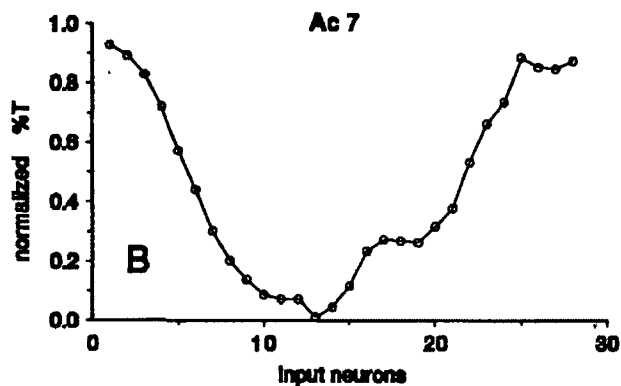
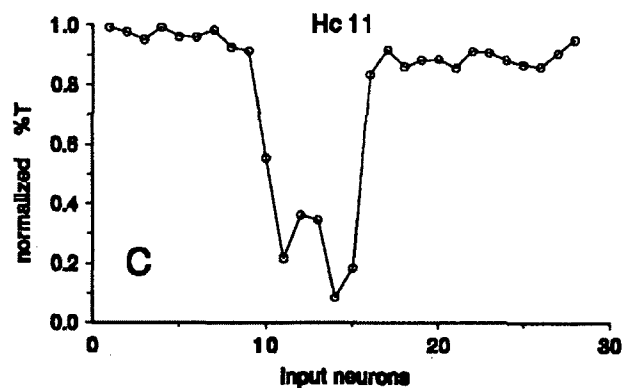
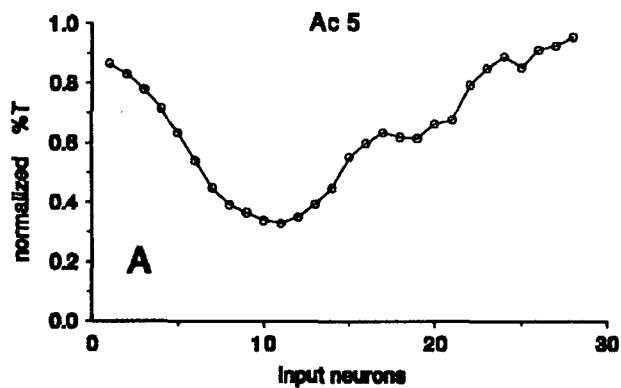


Figure 4.3. Input data for the problematic compounds (A,C) in the study of acidic -OH absorption band. Ac5 and Ac7 contain acidic -O-H bond, while Hc11 and Hc6 do not. Ac7 (B) and Hc6 (D) were recognized correctly and are shown for comparison.

Phenyl absorption band. The following is a table to show the performances of the neural networks on phenyl absorption bands.

Compounds	Good	Uncertain	Bad
Est3	6/12	1/12	5/12
Ac3	3/9	3/9	3/9
Hc8	0	1/15	14/15
Alc8	8/9	1/9	0
Alc11	8/9	0	1/9

**Table 4.4** Problematic compounds in the Phenyl absorption bands

As it can be seen in Figures 4.4 and 4.5 it is a very difficult problem to recognize a phenyl group. All "good" compounds are quite dissimilar. Charts about regions of absorption for phenyl suggests to look for a "very strong" and a "strong" peak in the 770-690 interval (in this study it was expanded to 790-670 interval) which is the 1 to 30 input neuron region in Figures 4.4 and 4.5, and for "all weak" peaks in the 1175-1000 interval (990-1175 in this study) which starts at 31 in the figures. Hc6 is the closest to the above description. Ac4 and Alc2 which were recognized well are very different from Hc6. At the same time Hc8 is much closer to Hc6 than Ac4 or Alc2 and still it was almost always classified as bad.

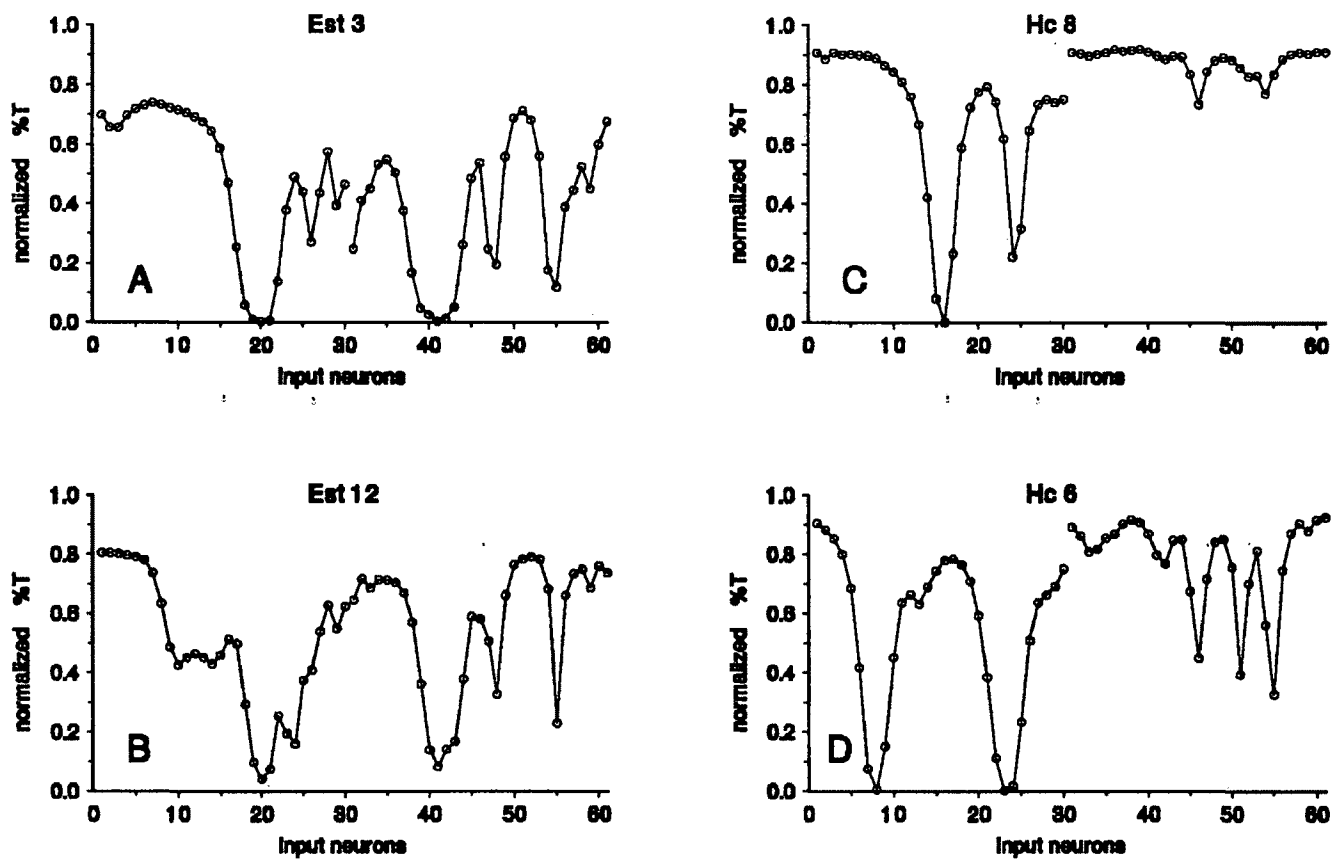


Figure 4.4. Input data for the problematic compounds (A,C) in the study of the absorption bands of phenyl group that contain phenyl group. Est12 (B) and Hc6 (D) were recognized correctly and are shown for comparison. The break in the curves results from the use of two separate wavenumber intervals.

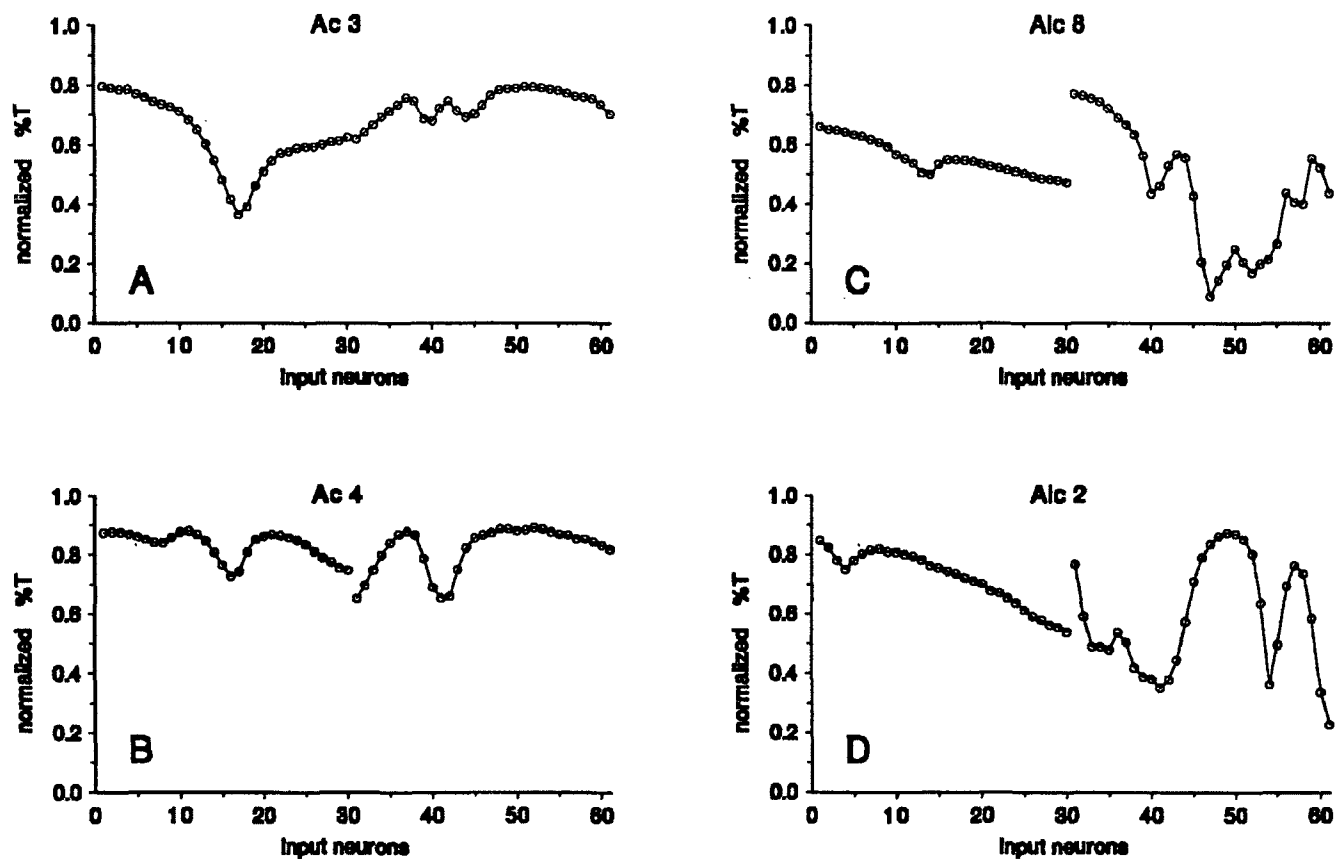


Figure 4.5. Input data for the problematic compounds (A,C) in the study of the absorption bands of phenyl group that do not contain phenyl group. Ac4 (B) and Alc2 (D) were recognized correctly and are shown for comparison. The break in the curves results from the use of two separate wavenumber intervals.

## 4.2 Decision Tree

For this study, a given set of compounds had to be classified into five different classes plus one more classification was needed, namely the compound does or does not contain a phenyl group.

In order to classify a compound as fast as possible a binary decision tree can be used. At each non-terminal node depending on the output of the corresponding neural network a decision can be made and then branch accordingly. The decision tree which was used to classify the compounds is in Figure 3.3.

The following is an example result which should be classified by the decision tree.

	-OH Acid	C-O-C	C=O	-OH Alcohol	Phenyl
Desired	0	0	0	1	1
Actual output	0.009	0.011	0.036	0.95	0.999

Table 4.5 Output for Alcl2 in T01

Since the output for C=O indicates that there is not C=O bond in the compound we take the False branch in the tree. Next question whether there is an alcohol -OH bond in the compound the answer is yes. The phenyl decision conclude that this compound contains phenyl. The classification concludes that the compound is an alcohol containing phenyl. The following is an example when there is a wrong output.

	-OH Acid	C-O-C	C=O	-OH Alcohol	Phenyl
Desired	0	1	1	0	1
Actual output	0.007	0.526	0.988	0.026	0.958

Table 4.6 Output for Est12 in T05

The C=O output indicates the presence of a C=O bond so the True branch is taken. Next the acid -OH output is examined and the False branch is taken. Then C-O-C output which is uncertain makes the whole classification procedure to fail and Est12 becomes uncertain. In case there is a bad output value the classification procedure can be halted similarly.

Using this decision tree the number of errors can be reduced. In case of the Ac1 (see Table 4.2) this compound would never be tested for the C-O-C absorption band if this decision tree is used. Also Hc11 (see Table 4.3) would never be tested for an acid -OH absorption band. To illustrate this the following figure shows the test results of Hc11.

	-OH Acid	C-O-C	C=O	-OH Alcohol	Phenyl
Desired	0	0	0	1	1
Actual output	0.602	0.011	0.036	0.024	0.999

Table 4.7 Output for Hc11 in T01

The C=O output for Hc11 indicates that there is not a C=O bond in the compound so we branch to the False. The next test is done to determine

whether there is an -OH for alcohol bond, the answer is no, so we branch to the False. The result of the phenyl decision is that this compound contains phenyl. The classification concludes that this compound is a hydrocarbon containing phenyl. The wrong classification of the -OH for acid was not used in this procedure.

The summary of errors, can be updated the following way.

	Good	Uncertain	Bad
C=O	405 (100%)	0 (0%)	0 (0%)
C-O-C	374 (92.5%)	6 (1.4%)	25 (6.1%)
OH(Alcohol)	405 (100%)	0 (0%)	0 (0%)
OH(Acid)	404 (99.8%)	0 (0%)	1 (0.2%)
Phenyl	376 (93%)	6 (1.4%)	23 (5.6%)
Sum	1964 (97.06%)	12 (0.56%)	49 (2.38%)

Table 4.8 Summary of errors when decision tree is used in the classification procedure

#### 4.3 Comparison with Fessenden and Györgyi's Result

The evaluations of my results were done by comparing it to Fessenden and Györgyi's single coded results. The following is an example of the single coded result of Fessenden and Györgyi.



	Hydro- carbon	Carboxylic Acid	Ester	Ketone	Alcohol	Phenyl
Desired	0	0	0	0	1	1
Actual	0.023	0.000	0.065	0.001	0.995	0.999
output						
Desired	0	1	0	0	0	0
Actual	0.015	0.919	0.016	0.013	0.009	0.000
output						
Desired	0	0	1	0	0	0
Actual	0.006	0.012	0.968	0.027	0.009	0.013
output						
Desired	0	0	0	1	0	1
Actual	0.169	0.001	0.165	0.822	0.001	0.963
output						
Desired	1	0	0	0	0	0
Actual	0.904	0.001	0.016	0.123	0.020	0.027
output						

Table 4.9 Single coded results of Fessenden and Györgyi Alcl2, Acl, Estl, Ketl2, and Hcl

The next table is an example result of the same testing in my experiment.

	-OH Acid	C-O-C	C=O	-OH Alcohol	Phenyl
Desired	0	0	0	1	1
Actual output	0.009	0.011	0.036	0.95	0.999
Desired	1	0	1	0	0
Actual output	0.974	0.311	0.998	0.051	0.016
Desired	0	1	1	0	0
Actual output	0.012	0.961	0.995	0.026	0.017
Desired	0	0	1	0	1
Actual output	0.010	0.012	0.770	0.255	0.991
Desired	0	0	0	0	0
Actual output	0.011	0.010	0.028	0.026	0.016

Table 4.10 My results for Alcl2, Acl, Estl, Ketl2, and Hcl.

As I mentioned earlier I used the same criteria that Fessenden and Györgyi introduced, for the classification of my testing results. To be classified as good, the output results from a test spectrum had to be within 0.4 of the correct values for the output neurons. To be classified as uncertain the output results had to be between 0.4 and 0.6. If the output did not fall into these categories then it was classified as a bad.

In order to compare my results to Fessenden and Györgyi's results Table 3.6 is needed which tells what kind of IR absorption bands occur in the different group of compounds. For example a tested compound is an alcohol with a phenyl if and only if -OH for carboxylic acids, C-O-C, and C=O tests results are in the [0,0.4] interval, while the output -OH for alcohol and for phenyl are in the [0.6,1] interval (see Table 4.10). In

case of Fessenden and Györgyi's results the tested compound is an alcohol with a phenyl if and only if the alcohol and phenyl outputs are in the [0.6,1] interval and the rest are in the [0,0.4] interval (see Table 4.9).

Common feature of the two methods, that if there was one wrong classification among the output results then the test for that compound failed and it was classified as uncertain or bad depending on the magnitude of the error. For example if there was one uncertain classification among the output results the compound was classified as uncertain. In case there were more than one wrong classifications the worst was used for classifying that compound. If I follow this evaluation I get the following result.

	Uncertain	Bad	$\Sigma$
Fessenden and Györgyi			
single coded results	24	29	53
My results without decision tree	20	49	69
My results with decision tree	12	49	61

Table 4.11 Uncertain and bad classifications

In my study one compound was tested five times for the five absorption bands while Fessenden and Györgyi tested each compound only once. Among my results there were not any compound with more than one wrong classifications in one test, while in Fessenden and Györgyi's tests there were a few compounds with more than one wrong classifications in one test. If I break down Fessenden and Györgyi's results to see the total number of wrong classifications I get the following table.

	Uncertain	Bad	$\Sigma$
Fessenden and Györgyi			
single coded results	35	34	69
My results without decision tree	20	49	69
My results with decision tree	12	46	61

**Table 4.12** Total number of uncertain and bad classifications

My second goal, to prove that my combined approach is superior to the only neural network approach has failed since these results, under the previously described circumstances, were the similar.

## 5. Conclusion

My study showed that a method which combines the human knowledge on IR spectroscopy with the pattern recognitions done by neural networks to identify functional groups in IR spectra provides comparable results with a pure neural network approach.

### 5.1. Advantages and Disadvantages

There are several advantages and disadvantages of the combined method when compared to the pure neural network approach. These are as follows:

#### Advantages of the combined approach

Existing knowledge of experts in IR spectroscopy on the absorption bands in an IR spectrum is used.

Depending on the form of the peak or peaks in a certain absorption band different sampling of data might be used. In case of a broad strong peak, like the -OH for carboxylic acid, less frequent sampling is enough. In case of several sharp peaks in a small interval, like the phenyl, all available data might be used.

This approach enables the user to trace a decision.

Instead of one big network several smaller networks are used which

could mean a shorter training time, and maybe better learning since in the combined approach the networks deal with simple patterns. There were no experiments made to test this statement.

The response time of the neural network is independent of the number of pattern it was trained on. This property of the neural networks make any system using them much faster than those which use conventional library search.

#### Disadvantages of the combined approach

An important interval for the identification of a functional group may not be considered since even the different books on IR analysis suggest different intervals for the recognition of certain functional groups.

Structural differences in compounds containing the same functional group get more emphasis than is desired.

Contamination in the sample may also be very critical. As seen in the result section the recognition may fail, or the networks may not train if one of the training compounds is contaminated. This disadvantage can be an advantage for example when the goal is to purify an organic compound.

## 5.2 Comments

Input data. First of all, much bigger variety of input data is needed. Each expected type of compound should be covered in the training. For example it is important to cover structural differences with plenty of examples in the training (see the problems with Est4 in sec. 4.1). It would also be beneficial if the compounds were purified before their spectra are taken (see the problems with Ket3 in sec 4.1). The concentration of the sample should be chosen carefully, meaning that the highest peak should be around 10% in percent transmittance. This would allow for most of the important details of the spectrum to be observed.

A solution for these problems could be the use of spectral libraries to train the system.

Performance of the neural networks. There are a number of ways to compare the performance of two neural networks. For example, it is possible to measure the time it takes the network to learn. This comparison was not possible in this study even for the same input since two different neural network programs were used on two different computers.

It is also possible to count the number of cycles it takes the neural network to reach a desired accuracy on the training set. Again in this study this comparison unfortunately can not be done. Fessenden and Györgyi used BrainMaker and the BrainMaker manual does not contain enough details about the way the program works. An example is the lack of information on the weight initialization in BrainMaker, which is known to

have an impact on the way the network learns see Sec. 3.3.1 thus it effects the number of training cycles.

Decision tree. In more complicated situations compounds can contain several functional groups (e.g. alcoholic -OH and C=O ) thus the tree must be extended and/or reorganized. There are automatic methods of inducing decision trees which may be useful in these situations.

### 5.3 Future Work

Uncertainty values. It may be possible to assign a quantity to each neural network used in the system which would show the expected accuracy of that unit. To do this some compounds should be separated from the training set for testing as was done in this work. After a thorough training is done a careful test could guide humans in evaluating real life problems. When the testing is finished each neural network performance is known, meaning that the number of uncertain and bad classifications are given. Weights can be assigned then to each unit based on the amount of bad and uncertain classifications. Using the following equation a so called uncertainty value can be assigned to each neural network.



$$W_i = 1 - (u_i w_u + b_i w_b) / n_i$$

where:

- $w_u$  : weight for the uncertain classifications. Suggested value: 1/2.  
 $w_b$  : weight for the bad classifications. Suggested value: 1.  
 $b_i$  : number of bad classifications by the  $i^{\text{th}}$  neural network  
 $u_i$  : number of uncertain classifications by the  $i^{\text{th}}$  neural network  
 $n_i$  : number of testing on the  $i^{\text{th}}$  neural network  
 $W_i$  : uncertainty factor of the  $i^{\text{th}}$  neural network

and

$$\begin{aligned}
 0 &\leq W_i \leq 1, \\
 0 &\leq u_i \leq n_i, \\
 0 &\leq b_i \leq n_i, \\
 u_i + b_i &\leq n_i
 \end{aligned}$$

Using the suggested values for  $w_u$  and  $w_b$ ;  $W$  will be 0 if all tests turned out to be bad, 0.5 if all were uncertain and 1 if all were good. If the path in the tree was through a number of neural networks and the compound was finally classified as being in class X, then it can be said that the compound was classified as X with the certainty of C where C is calculated as follows:

$$C = \pi_i W_i,$$

where  $W_i$  is the uncertainty value of the  $i^{\text{th}}$  neural network which was used to obtain the final classification.

## REFERENCES

### Appendix

The following tables are results of the fifteen training. The recognition of a functional group was categorized the following way. Good if the output of the neural network was in the [1,0.6) interval, uncertain [0.6,0.4), and bad [0.4,0]. In the tables only the uncertain (U), and the bad (B) classifications are shown, where the indices stand for in which training out of the three possible.

Training 01 (T01)	OH(Alc)	C=O	C-O-C	OH(Ac)	P
Est1					
Est2			B <sub>1,2,3</sub>		
Hc1					
Hc11				B <sub>1</sub>	
Ket10					
Ket12					
Alc9					
Alc12					
Ac1					

Training 02 (T02)	OH(Alc)	C=O	C-O-C	OH(Ac)	P
Est1					
Est12					
Hc7					
Hc11					
Ket3			B <sub>1,2,3</sub>		
Ket10					
Alc5					
Alc12					
Ac3					U <sub>1,2</sub> B <sub>3</sub>

Training 03 (T03)					
OH(Alc)	C=O	C-O-C	OH(Ac)	P	
Est3				B <sub>2</sub>	
Est6					
Hc6					
Hc9					
Ket5					
Ket13					
Alc8					U <sub>3</sub>
Alc12					
Ac2					
Training 04 (T04)					
OH(Alc)	C=O	C-O-C	OH(Ac)	P	
Est4					
Est12					
Hc3					
Hc8					B <sub>1,2,3</sub>
Ket2					
Ket7					
Alc4					
Alc11					
Ac7					
Training 05 (T05)					
OH(Alc)	C=O	C-O-C	OH(Ac)	P	
Est5					
Est12					
Hc6					
Hc10					
Ket2					
Ket7					
Alc1					
Alc5					
Ac3					B <sub>1,3</sub> U <sub>2</sub>
Training 06 (T06)					
OH(Alc)	C=O	C-O-C	OH(Ac)	P	
Est5					
Est12		U <sub>2</sub>			
Hc8					B <sub>1,2,3</sub>
Hc10					
Ket10					
Ket13					
Alc1					
Alc2					
Ac7					

Training 07 (T07)					
OH(Alc)	C=O	C-O-C	OH(Ac)	P	
Est2		B <sub>1,2,3</sub>			
Est10					
Hc5					
Hc7					
Ket7					
Ket13					
Alc1					
Alc8					
Ac4					
Training 08 (T08)					
OH(Alc)	C=O	C-O-C	OH(Ac)	P	
Est4		B <sub>1,2,3</sub>			
Est12					
Hc6					
Hc7					
Ket7					
Ket12					
Alc4					
Alc10					
Ac3					
Training 09 (T09)					
OH(Alc)	C=O	C-O-C	OH(Ac)	P	
Est3				B <sub>3</sub> U <sub>2</sub>	
Est6					
Hc2					
Hc8				B <sub>1,2,3</sub>	
Ket3		B <sub>1,2,3</sub>			
Ket7					
Alc8					
Alc10					
Ac6					
Training 10 (T10)					
OH(Alc)	C=O	C-O-C	OH(Ac)	P	
Est2					
Est4		B <sub>1,2,3</sub>			
Hc8				B <sub>1,3</sub> U <sub>2</sub>	
Hc10					
Ket3		B <sub>1,2,3</sub>			
Ket6					
Alc9					
Alc10					
Ac4					

Training 11 (T11)					
OH(Alc)	C=O	C-O-C	OH(Ac)	P	
Est10					
Est12		U <sub>1</sub> B <sub>3</sub>			
Hc6					
Hc7					
Ket10					
Ket13					
Alc3					
Alc10					
Ac5			U <sub>1,2,3</sub>		
Training 12 (T12)					
OH(Alc)	C=O	C-O-C	OH(Ac)	P	
Est6					
Est12		U <sub>1</sub>			
Hc10					
Hc11					
Ket6		B <sub>1,2,3</sub>			
Ket13					
Alc3					
Alc11					
Ac5			U <sub>1,2,3</sub>		
Training 13 (T13)					
OH(Alc)	C=O	C-O-C	OH(Ac)	P	
Est3				B <sub>1,2,3</sub>	
Est6					
Hc2					
Hc6					
Ket2					
Ket5					
Alc4					
Alc10					
Ac7					
Training 14 (T14)					
OH(Alc)	C=O	C-O-C	OH(Ac)	P	
Est1					
Est3					
Hc10					
Hc11					
Ket5					
Ket12					
Alc2					
Alc11				B <sub>3</sub>	
Ac1		U <sub>1,3</sub>			

Training 15 (T15)	OH(Alc)	C=O	C-O-C	OH(Ac)	P
Est5					
Est12			U <sub>2,3</sub>		
Hc1					
Hc8					B <sub>1,2,3</sub>
Ket7					
Ket12					
Alc5					
Alc12					
Ac4					

## Bibliography

- Barr (1989) Barr A., Feigenbaum E.A. The Handbook of Artificial Intelligence Vol. II. Addison-Wesley Publishing Company, Inc. Reading Massachusetts 1989.
- Caudill (1987) Caudill M. Neural Networks Primer. AI Expert, (Dec 1987), 46-52.
- Fessenden (1990 a) Fessenden R.J., Fessenden J.S. Organic Chemistry. Brooks/Cole Publishing Company, Pacific Grove, California, 1990.
- Fessenden (1990 b) Fessenden R.J., Györgyi L. Identifying Functional Groups in Infrared Spectra Using an Artificial Neural Network submitted.
- Giarratano (1989) Giarratano J., Riley G. Expert Systems Principles and Programming. PWS-KENT Publishing Company, Boston, 1989.
- Hecht-Nielsen (1989) Hecht-Nielsen R. Neurocomputing. Addison-Wesley Publishing Company, 1989.
- Huixiao (1990) Huixiao H., Xinquan X. J. Chem. Inf. Sci. 1990, 30, 203.
- Meyer (1991) Meyer B., Hansen T., Nute D., Albersheim P., Darvill A., York W., Sellers J. Identification of the  $^1\text{H}$ -NMR Spectra of Complex Oligosaccharides with Artificial Neural Networks. Science, Vol. 251 (Feb. 1991).
- McClelland (1988) McClelland J.L., Rumelhart D.E., Explorations in Parallel Distributed Processing, A Handbook of Models, Programs, and exercises. MIT Press, Cambridge, MA, 1988.
- Park (1988) Park S.K., Miller K.W. Comm.ACM. 1988 31,10.

- Prerau (1990) Prerau D.S. Developing and Managing Expert Systems. Addison-Wesley Publishing Company, Reading, Massachusetts, 1990.
- Robb (1990) Robb E.W., Munk M.E. Mikrochim. Acta 1990, I, 131.
- Wasserman (1989) Wasserman P. D. Neural Computing Theory and Practice. Van Nostrand Reinhold, 1989.
- Weiss (1991) Weiss S.M., Kulikowski C.A. Computer Systems That Learn. Morgan Kaufmann Publishers, Inc. San Mateo, California, 1991.
- Williams (1980) Williams D.H., Fleming I. Spectroscopic methods in organic chemistry. McGRAW-HILL Book Company (UK) Limited, London, 1980.