Characterising Complex Reaction Mechanisms using Machine Learning Clustering Techniques

Edward S. Blurock
blurock@risc.uni-linz.ac.at
Research Institute for Symbolic Computation
Johannes Kepler University
A-4040 LINZ
Austria

June 14, 1998

Abstract

A machine learning conceptual clustering method applied to reaction mechanisms provides an automatic and, hence, unbiased means to differentiate between reactive phases within a total reactive process. Similar reactive phases were defined by means of local reaction sensitivity values. The method was applied to the Hochgreb and Dryer aldehyde combustion mechanism of 36 reactions. Three major time ranges were found and characterised, an initial phase of aldehyde reaction, an intermediate phase where only a small amount of aldehyde is left and an end phase of reactions to final products. Further refinements of these phases into subtime intervals were found. All ranges found could be chemically justified. This method is meant as a supplement to existing methods of mechanism analysis and its main purpose is the automatic characterisation of existing mechanisms and can only indirectly (i.e., when supplemented with other methods) be used for mechanism reduction.

1 Introduction

As a reaction mechanism becomes more complex, the amount of information that has to be analysed to interpret the processes within the mechanism becomes complex. The complexity of a reaction mechanism is, in addition to (or as a consequence of) the number of reactions and species, characterised by the multitude of dynamic processes and interactions that are determined when the mechanism is allowed to 'react', i.e. the corresponding set of differential equations are solved.

Over time, the reaction process moves through different phases where different reactions and species change their significance in their contribution to the overall reactive process. With increasing complexity, the identification and characterisation of these phases becomes more difficult; The amount of information that one has to analyse can become prohibitively large. One way to overcome this expanse of data is to search for more automatic means to analyse or reduce the data. This paper demonstrates a technique to automatically characterise a reaction mechanism in terms of its reaction phases.

Throughout a zero-dimensional reaction process, i.e. where the temperature and species concentrations are changing over time, the mechanism can go through many phases of reaction. Between the different phases, the significance of different reactions and species changes. The method introduced in this paper demonstrates a machine learning method which identifies and classifies these phases automatically.

The key concept used in the classification is that of 'similarity'. Similar reactive states of the entire process are grouped together. The reaction process is described by a set of states where each state is defined by the temperature and a specific set of species concentrations at a given time. From this state, the significance of each reaction, defined by the normalised sensitivity constant, is calculated. Two states are similar if the calculated significance of each reaction is similar (i.e. the sensitivity values are close). After the classification, a set of time ranges defining the similar reaction phases and a corresponding description in terms of reaction significance is given.

2 Clustering and Similarity

Clustering provides an unsupervised (without human bias) method of determining sets of similar objects. Two objects are considered similar if they are in the same cluster. In these studies the 'similar objects' are temporal states or snapshots of the progress over time of a reaction mechanism. A 'state' is essentially the temperature, pressure and set of current species concentrations (and other information derived from these concentrations, such as reaction rates and sensitivities). Clustering is an 'unsupervised' method in that the decision of what objects should be clustered is made solely with respect to the data and is not biased by outside influences. In this study, time of reaction is not included in the state information. The time ranges fall out naturally from the non-temporal information.

The concept of similarity is not unique. The use, context and precision determine whether two objects may be considered similar or not. In terms of clustering, these factors translate into:

Representation How the objects are represented, i.e. what information is used to describe them, is the major factor in determining similarity.

Measure In clustering, the concept of similarity is translated into a measure which, given two objects, gives their measure of the similarity.

Hierarchy Given the measure of similarity, the decision has to be made as to how similar the objects must be to be clustered together.

Method There are several different principles of building clusters and this influences which objects are clustered together.

Interpretation Different clustering methods represent the clusters in different ways and hence yield different representations of the cluster.

These factors and how they are applied in the study of complex reaction mechanisms will be elaborated in the following sections.

2.1 Representation

A model is a simplified representation of reality in which only the 'essential' information is considered. This, naturally, will greatly influence the decision as to which objects are similar or not. The simplification of reality in this study is representing a complex mechanism and its progress over time as the results of numerical calculations. Each object is the 'state' of the mechanism at a given time.

The representation of these states influences which parts of the mechanism are considered similar. If the representation of the state is in terms of species concentrations, then two 'similar' states would be when the concentration profiles are about the same. In this study, similarity will be defined as the normalised sensitivity constants (see section 3.2). Thus, two states are similar if the individual reactions have similar significance in both states, i.e. each sensitivity value is similar. Furthermore, since the cluster method used (see section 3.5) works with yes/no questions (predicates), the set of sensitivity values will be divided into regions and the question will be whether the value lies between given boundaries. Thus similar states will have sensitivity values within the same ranges.

2.2 Measure

There are many ways to define similarity within a clustering method. In traditional vector clustering (often called 'pattern recognition' in the literature[1]), for example, the similarity between two objects (numeric vectors) is define as the 'vector distance', a function from n-dimensional vector space to a floating point number. But even here, there are many choices, of which the most common is the Euclidean distance. Each one can provide subtle differences in which objects are clustered together. In this study, the machine learning clustering method uses probabilities in a measure which tries to maximise the intracluster similarities and minimise the intercluster similarities. This roughly translates to trying to maximise the probability that the set of reaction conditions of the time states considered within the cluster fall within the same region. For example, for the time states at the beginning, the probability will be high that the initiation reactions have a high sensitivity value and termination reactions have a low sensitivity value. These would then be considered a similar condition and thus clustered together.

2.3 Hierarchy

Given the similarity measure, it still has to be determined whether the two objects are similar enough to be grouped into the same cluster. For example, how close must the sensitivity constants be before they are considered to be under similar reaction conditions. The typical way to handle this non-uniqueness is to create a hierarchial structure of similarities. At the top of the hierarchy is one 'cluster' representing all the objects. Moving down the hierarchy, the first 'roughly' similar similar objects are distinguished. For a reaction mechanism, this could be, for example, three clusters where the initiation processes, the intermediate processes and the termination processes dominate. At the very bottom of the hierarchy (if one would continually make distinctions between clusters) are 'clusters' consisting of exactly equivalent objects. The hierarchial structure can be thought of as a directed tree. The root of the tree is the entire set of objects. Each node of the tree represents a cluster. Emanating from each node (except for the bottom 'leaf' nodes) is a set of directed connections to the further cluster nodes representing 'more similar' objects. In the study of reaction mechanisms, moving down the hierarchy represents finer distinctions between processes occurring within the mechanism.

2.4 Method

How the similarity measure is used to build the hierarchy of cluster structures represents another variation in which clusters are formed. If all the objects are considered at once, as in most vector clustering methods, there are two classes of clustering methods, divisive and agglomerative. The divisive methods start with the entire set and iteratively split them apart into the individual clusters. The agglomerative methods start with set of clusters each having one object, and then slowly group them together. This can be done by each time pairing the most similar or by using similarity measure as a scale and clustering only those above a given threshold.

Another class of methods, which includes the machine learning method used in this study, are incremental; Each object is introduced into the hierarchy one at a time. At the beginning there are no objects. The first object forms a cluster. As each object is introduced, a decision is made whether it belongs to one of the current set of clusters, or if it belongs in a cluster of its own. If it is put in an existing cluster, then the same decision is made at the next level in the hierarchy.

2.5 Interpretation

The most obvious piece of information that the cluster offers is the set of clustered objects themselves. These can be manually examined, using knowledge of the system, to find patterns of similarity. In reaction mechanism analysis, the foremost piece of information that can be examined is the set of reaction times of the objects. If the times are contiguous, then the reaction time regions can be examined for plausibility.

Beyond this, each clustering method provides an internal description of the cluster. This information is what is used to form the cluster in the first place. In vector clustering methods, the representation is the 'center' or average element; All the objects within the cluster are close to this center. In this study, this would be the average sensitivity value for each reaction. Regardless of the representation, however, this average element can always be calculated.

The main component of the machine learning clustering representation is the set of probabilities that the sensitivity values are within given bounds. More specifically, the set of sensitivity values are ranked according to how the probabilities have changed with respect to the parent cluster. Those features which have changed the most due to the clustering will be listed first. This gives the most prevalent distinguishing features of the cluster. The most prevalent features which distinguish the cluster from the other clusters in the same level and from the parent cluster are listed. It is important to note that is it not necessarily true that the highest sensitivity values have the highest clustering significance. Those (regardless of whether they are high or low) which have changed the most are the most significant.

3 Mechanism Data and Representation

The data to be clustered stems from numerical calculations solving the set of differential equations associated with the reactions of the mechanism. The clusters are formed from transformed data derived from this 'raw' data. The entire procedure of transforming the data and classifying the mechanism involves the interaction of three different software systems, the REACTION system[2, 3] for managing the mechanism data (input and output), the CHEMKIN/SENKIN[4, 5] FORTRAN program for solving the differential equations and the ANALYSIS[6, 7] system for performing the machine learning clustering. The mechanism and its associated information is retrieved through the REACTION system. REACTION then sets up the input, remotely executes the CHEMKIN/SENKIN program and collects the numerical output, i.e. the time dependent temperature and concentrations of the species. From this data, REACTION computes the normalised sensitivity constants and sets up the input to the ANALYSIS program. The ANALYSIS program then creates the sensitivity value ranges, representing the mechanism, and computes the cluster hierarchy.

An outline of the procedure is summarised as follows:

CHEMKIN Setup Using the REACTION database, the mechanism reactions and their associated kinetic and thermodynamic constants are collected. The necessary files are set up for the CHEMKIN program.

SENKIN Integration With the specification of the initial temperature, pressure and concentrations of the species, the SENKIN program is called and the matrix of time versus temperature and species concentration data is collected.

ANALYSIS Setup For each time interval, the concentration data from the SENKIN output and the kinetic constants from the REACTION database are used to calculate the normalised sensitivity values for each reaction. This data is then put in a file read by the ANALYSIS system.

Cluster Representation The representation of the mechanism for cluster formation is a set of predicates, i.e. yes/no questions about intervals of sensitivity values. For each reaction, a set of boundaries are set up. These are (at least 10) values ranging from the minimum value to the maximum value for that reaction.

Cluster Formation The machine learning clustering procedure, in the ANALYSIS system, placed each object (mechanism time snapshots) incrementally into the cluster hierarchy. The time snapshots are selected randomly in order not to introduce the bias of sequential times. The output is the set of clusters and their representations.

Details about specific aspects of the data formation and manipulation are given in the following sections.

3.1 SENKIN Data

The CHEMKIN/SENKIN program is used to solve the set of differential equations associated with the mechanism. The mechanism used in this study is that suggested by Hochgreb and Dryer[8] for the combustion of aldehyde. The kinetic constants of the paper were used unless newer values were available in the compilation of Baulch et. al[9]. The SENKIN program is used to perform a zero dimensional constant pressure (1 atmosphere) 'experiment' with a given starting temperature (945K) and aldehyde (0.00356 atm), oxygen (0.00342 atm) and inert gas (0.99288 atm) concentrations. The output is the concentration and temperature evolution at given reaction times. Which reaction times are used is determined by the integration procedure. Essentially, when rapid changes are occuring, sequential times are close together. In regions of rapid change, the use of a regular time interval would result in a lose of information.

3.2 Normalised Sensitivity Values

To determine the relative importance of the reactions at each time interval, the normalised sensitivity values (based on the method of Turanyi[10]) are used. At each time state and for each reaction, the value is calculated as follows:

$$f_j = \Sigma_i^{n_r} |v_{ij}| R_i$$
$$N_i = \Sigma_j^{n_s} |v_{ij}|$$
$$s_i = \frac{R_i}{N_i} \Sigma_j^{n_s} \frac{|v_{ij}|}{f_j}$$

	n_r	=	The number of reactions
	n_s	=	The number of molecules
	v_{ij}	=	Change of species j in reaction i (stoichiometric coefficient)
where:	R_i	=	Rate of change of reaction i
	f_j	=	Absolute changes in concentration of species j
	N_i	=	Number of species reacting in reaction i
	s_i	=	Absolute sensitivity of reaction i

Since the factor v_{ij} can be either positive, production of the species as a product, or negative, the loss of the species as a reactant, the absolute value is taken, representing the absolute change in the species concentration . The term $\frac{R_i|v_{ij}|}{f_j}$ represents the relative contribution of reaction i to the changes in species j. The normalised sensitivity constant sums up all these relative changes giving the total relative change in species concentrations. It is divided by N_i, the number of species involved in the reaction, to insure that the value of the sensitivity constant does not exceed one. Thus the 'raw' numeric data representing the mechanism is a set of n_r (the number of reactions) dimensional vectors, one for each time interval.

3.3 Predicate Divisions

The representation used by the machine learning clustering method is a set of predicates, i.e. a set of yes/no (true/false) questions. The predicates are in the form of boundary ranges of the sensitivity values, $B_{ij}(s_j)$. For every sensitivity value, s_j, a set of bounds is given. If the sensitivity value is within the specified bounds, the value of the predicate is true, otherwise it is false. A complimentary predicate is also used, i.e. it is true when the sensitivity value is outside the given bounds. For this analysis, the boundaries were specified by hand. At least five boundaries were used for each sensitivity constant.

Each time state is represented by the same set of predicates. Thus each state is represented as a set of true/false values representing the answers to the corresponding predicates.

3.4 Similarity Measure

The similarity measure used with the clustering method is based on the probability that a given predicate, $B_{ij}(s_i)$, representing a sensitivity value range (Section 3.3), is true. This is given by:

$$P(B_{ij}(s_i)) = \frac{\#\ states\ where\ B_{ij}(s_i)\ is\ true}{total\ number\ of\ states}$$

In addition, a conditional probability is used, namely the probability that the predicate, $B_{ij}(s_i)$, is true given that the state is within cluster, C_k:

$$P(B_{ij}(s_i)|C_k) = \frac{\#states\ in\ C_k\ where\ B_{ij}(s_i)\ is\ true}{number\ of\ states\ in\ C_k}$$

In the derivation of the measure, a third probability is used, namely, $P(C_k|B_{ij}(s_i))$, the probability that it is in cluster C_k, given that the sensitivity value falls within the range.

In forming a 'good' cluster, a reasonable criteria is to maximise the intra-cluster similarities and the inter-cluster differences. A measure that accomplishes this is,

$$P(B_{ij}(s_i))P(B_{ij}(s_i)|C_k)P(C_k|B_{ij}(s_i))$$

The first factor weights the term by the importance of the particular parameter range. For example, the term

$$P(B_{ij}(s_i)|C_k)P(C_k|B_{ij}(s_i)),$$

for example, is one when $B_{ij}(s_i)$ is true only for one cluster.

Using Bayes rule for probabilities (see any standard statistics book[11] for further details), namely,

$$B_{ij}(s_i)\ P(C_k|B_{ij}(s_i)) = C_k\ P(B_{ij}(s_i)|C_k),$$

the measure can be rearranged to

$$P(C_k)P(B_{ij}(s_i)|C_k)^2$$

The final measure for clustering considers the change in probabilities due to clustering and after summing over all parameters, i, ranges, j, and clusters, k, and dividing by the number of clusters N_c, the measure becomes:

$$\frac{(\Sigma^k P(C_k)\Sigma^{ij} P(B_{ij}(s_i)|C_k)^2) - (\Sigma^{ij} P(B_{ij}(s_i))^2)}{N_c}$$

3.5 Clustering Method

The machine learning method used here is a (simplified version of) the COBWEB conceptual clustering method by Fisher[12, 13]. The clustering method itself uses the measure to incrementally, one state at a time, classify each state into a hierarchy of clusters. Since the states represent consecutive events, to eliminate the bias introduced by clustering them in the temporal order, a random state is chosen each time.

Using the measure on the state to be clustered, a decision is made whether to put the state in one of the existing clusters or to create a new cluster. Essentially, all possibilities are tried and that with the highest measure value is chosen. If the state is put in an existing cluster, the clustering algorithm continues iteratively to the next level. The states in the current cluster are then considered as the total set and the same decisions are made with the underlying clusters. This iteration through the levels continues until the bottom of the hierarchy.

In the full COBWEB method, there are operators which 'clean up' the cluster structure by merging and splitting clusters in a level. These operations are very expensive (and non-unique) and thus were left out of this implementation. To simulate simply the result of these operators, two techniques are used. First, the entire set of states is introduced to the algorithm more than once. This enables a given state to land in another cluster. Second, in the final result, those clusters with only a few states (for example, those containing only 5% of the total number of states) are ignored.

3.6 Cluster Descriptions

As explained in section 3.4, the measure of a particular cluster has two terms. The first has the form $P(B_{ij}(s_i)|C_k)$, the probability that the sensitivity s_i is within the value range B_{ij} given, i.e. within, cluster, C_k, and the second has the form $P(B_{ij}(s_i))$, the probability that the sensitivity s_i is within the value range B_{ij} for all the clusters (i.e. the probability in the parent cluster). The total measure tries to maximise the difference between these terms. The measure of how a particular predicate contributes to this measure can thus be given as (squaring each term, as given in the measure):

$$P(B_{ij}(s_i)|C_k)^2 - P(B_{ij}(s_i))^2,$$

the increase (or decrease) in the probability that a parameter is within the specified range. The ranges which have increased the most are considered the best descriptions of the clusters; these distinguish the cluster from the parent cluster the most. Furthermore, since the derivation of the total clustering measure includes maximising inter-cluster differences, the description also gives insight as to how each cluster differs from one another.

Each $B_{ij}(s_i)$ is a range of sensitivity values of a given reaction, for example, the range $S > 0.3$ for the reaction HCO+M=CO+H+M. $P(B_{ij}(s_i))$, the probability that the sensitivity is within the given range, would be zero if the reaction was never within this range and one if is was always within this range. $P(B_{ij}(s_i)|C_k)$ is the same type of probability, but restricted to a particular cluster, which represents a time range. The difference of these terms would reach a maximum, if a reaction within the cluster (i.e. time range) had sensitivity values always within the specified range. For example, if for the reaction HCO+M=CO+H+M, within the time range 0.0-0.098 secs., always had $S > 0.3$. The difference would even be higher if, for example, over the whole reaction time (from 0 to 0.16 secs) had values where $S > 0.3$ and $S < 0.3$, i.e. much different than within the cluster.

In forming the time ranges, the algorithm tries to maximise this probability difference. In the final answer (for example, in tables 1 through 4), this difference gives the 'importance' of the reaction sensitivity. In other words, what reactions and sensitivity ranges made this particular time phase unique compared to the others. The closer the importance is to one, the more significant contribution it has.

An important note is that the cluster description does not necessarily select the highest sensitivity values, i.e. the most significant reactions in the cluster. It is a measure of how all the sensitivity values changed, regardless of whether they are low sensitivities or high sensitivities. Thus, if a reaction with a low sensitivity value changed significantly, then it would be included in the description of the cluster. The more traditional description of listing only the reactions with high sensitivity values is derived separately.

4 Cluster Result Interpretation

In this section the results of a single clustering and its interpretation will be discussed in detail. The model, using the SENKIN software, is a constant pressure, 1 atmosphere, adiabatic experiment with a starting temperature of 945K. The output resulted in 987 time states ranging from 0.0 to .15 seconds. The sensitivity values calculated at these intervals were then clustered using the algorithm described. In the results shown here, the individual states will be either referred to by their order, ex. the first, second and third states are 0, 1 and 2, respectively, or by the reaction time they represent. Consecutive states represent ranges of time. For example, the states 0 through 549 represent the time interval from zero to .098 seconds reaction time.

The result of the analysis is a hierarchy of clusters. Each cluster gives essentially three types of information about the reaction process:

Time Intervals This is a list of time intervals making up the cluster. Due to the continuous nature of the sensitivity values versus time, the set of time intervals in a cluster are, for the most part, consecutive. Thus, a cluster can be assigned to a distinct phase, i.e. time range, of the reaction process.

Hierarchy of Intervals Each level in the hierarchy breaks down the previous 'parent' time interval of the previous level into finer time ranges. As one moves down the hierarchy, finer and finer time ranges are resolved. In this way finer distinctions in the reaction mechanism can be made.

Cluster Description The cluster description is in terms of the predicates described in the previous section, a set of sensitivity value ranges. With each predicate description is a measure corresponding to its 'importance' to the description.

4.1 Time Intervals

Although no information about time is present when the clusters are calculated, the clustering nevertheless results in consecutive time ranges. The exceptions are relatively minor. In some cases, only a single state is missing (as in the first cluster in table 1). In other cases, they are only isolated states (as in the second level cluster shown in Table 2).

The main reason for the grouping of consecutive time ranges can be attributed to the continuous nature of the combustion process. The value of the sensitivity constant changes only slightly from one time state to the next. This is also a direct consequence of the fact that the integration algorithm increases the density of time intervals in making its calculations when the reaction processes vary more. Thus, consecutive time samples are 'similar' and thus clustered together.

4.2 Hierarchy of Clusters

As explained earlier (Section 2.3), the output of the machine learning cluster calculation is a hierarchy of clusters, i.e. groupings of time ranges. Figure 1 shows a typical hierarchy. At the top of the hierarchy is the set of 987 states representing the full time range of 0.0 to 0.15 seconds. The next level broke down the reaction mechanism into three clusters (see also Table 1) representing the time intervals 0.0-0.098 seconds, the reaction of aldehyde phase, 0.098-0.117, when only small concentrations of aldehyde are remaining, and the end phase after the aldehyde is used up, the reaction of the final products.

Examination of sensitivity values shown in Figure 2 gives a rough understanding of the time ranges clustered. The first time range, from 0.0 through 0.098 seconds, is a phase of slow changes. The important reactions are those converting of aldehyde (through hydrogen and hydroxyl radicals) to formyl radicals and the further reaction of formyl radical to yield carbon monoxide. Less significant intermediate reactions such as peroxide and peroxyl radicals yielding

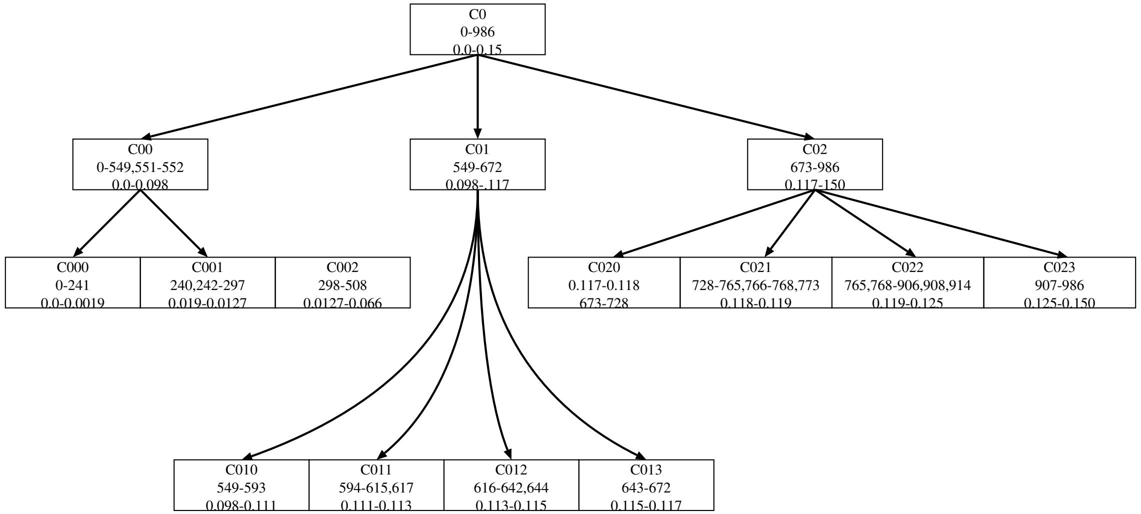

Figure 1: Cluster hierarchy showing the grouping of time intervals. The first line in the box is the cluster label, the second line are the list of states in the cluster and the third line is the derived time range the cluster represents.

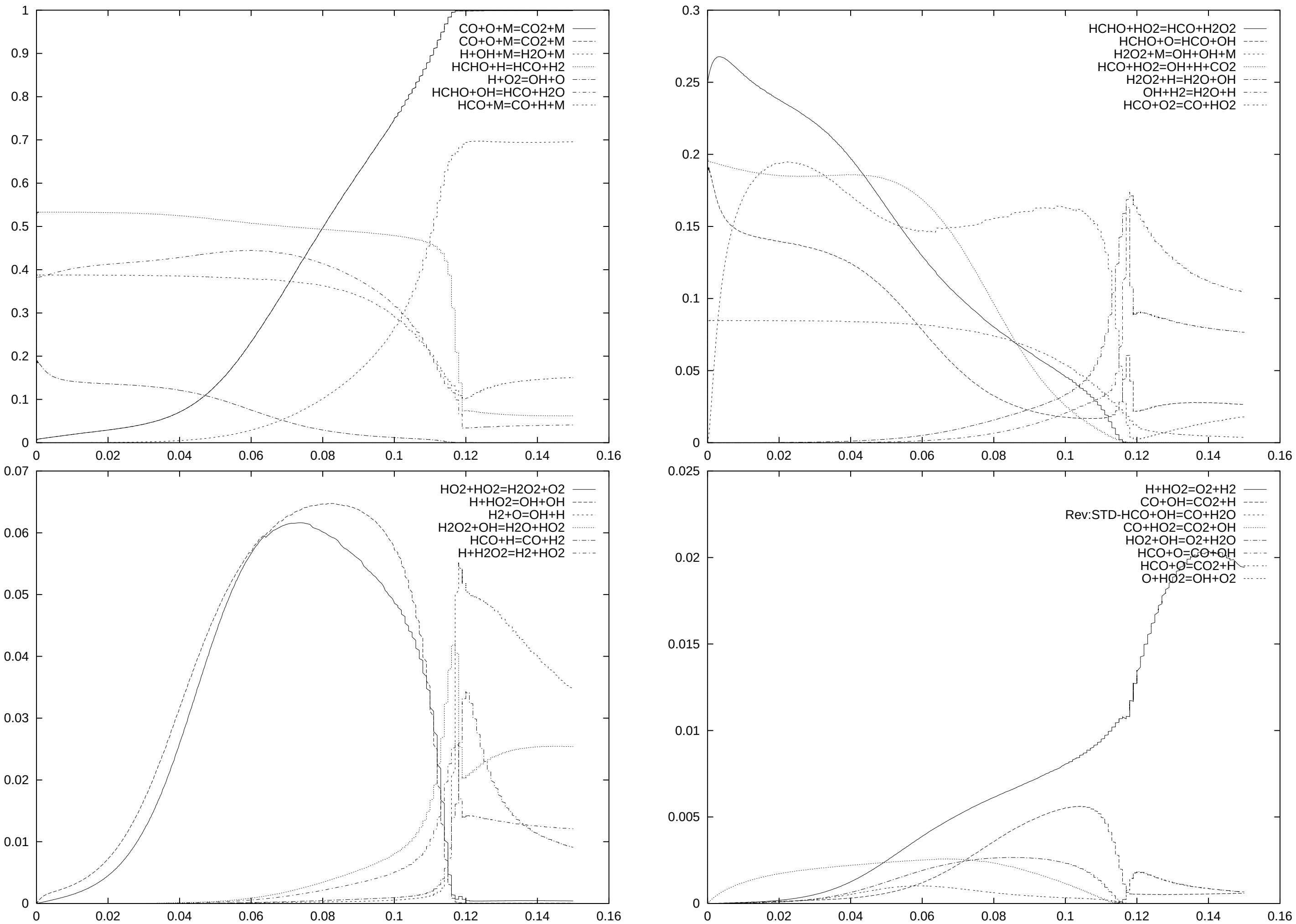

Figure 2: Sensitivity versus time for the time range 0.0 through 0.15

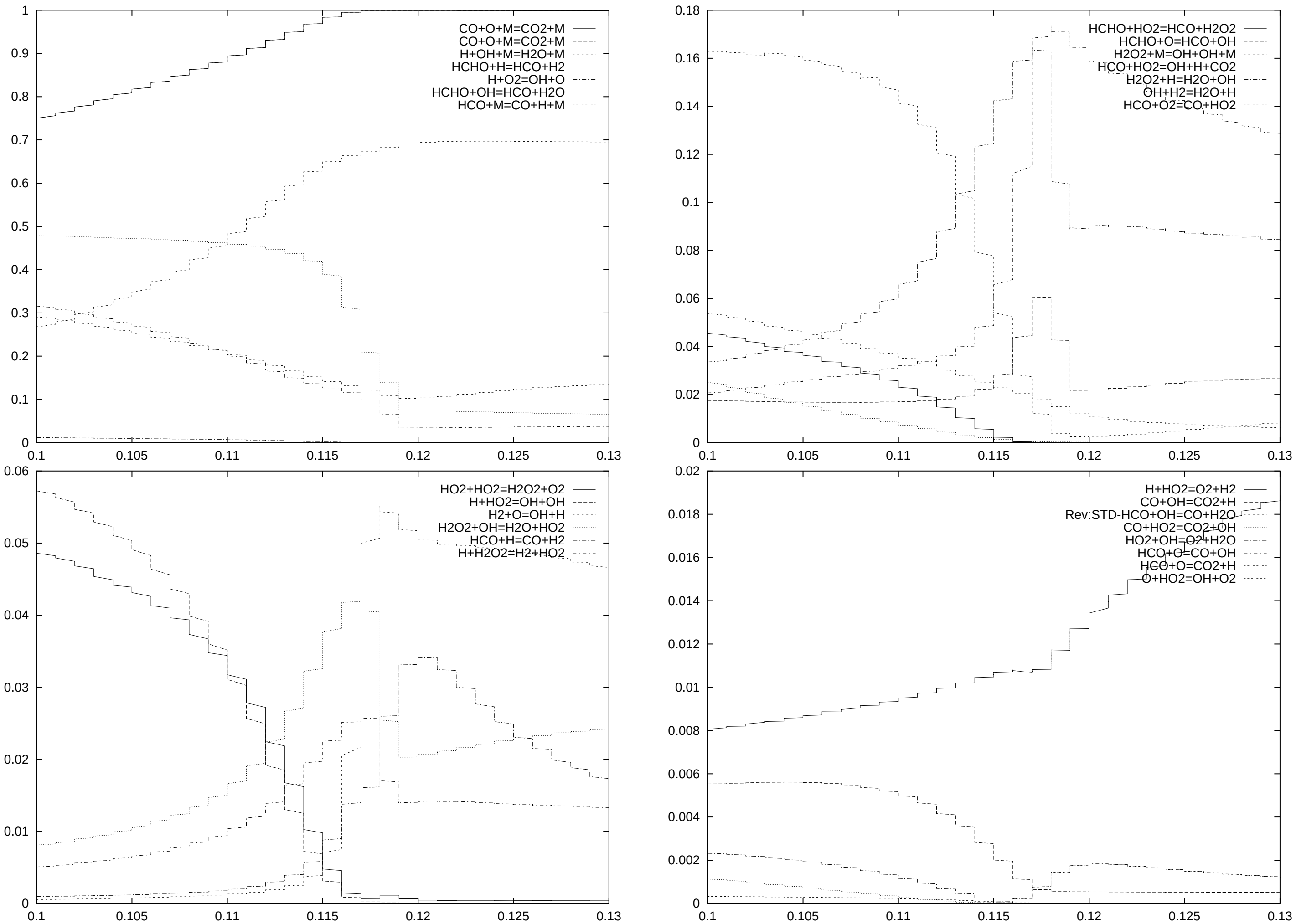

Figure 3: Sensitivity versus time for the time range 0.1 through 0.13 showing the regions of greater changes.

hydroxyl radicals also change slowly. Throughout this period, the importance of the conversion of carbon monoxide to carbon dioxide increases. In the next time range, from 0.098 to 0.117 seconds, the aldehyde concentration, and hence the importance of the corresponding reactions, goes to zero. The end of this time range (marking the beginning of the final time range) is characterised by the increase of the importance of the reactions of the by-products yielding finally carbon dioxide and water. The final time range, from 0.117 to 0.150 seconds, is dominated by these termination reactions to final products.

The first level represents a rough division of the reaction process. Each of these divisions have a finer structure established in the second level of the clustering hierarchy (shown in tables 2, 3 and 4).

The first time range, 0.0 through 0.098, is broken down into two short phases (0.0-0.0019 and 0.0019-0.012) characterising the initialisation reactions. These are followed by two longer ranges (0.012-0.066 and 0.066-0.098) characterised by the aldehyde (and resulting product) reactions.

The second time range, 0.098 through 0.117, is a fairly active region, i.e. the ordering of the significant reactions changes drastically. This range is broken down into two (relatively speaking) longer regions where the aldehyde to formyl reactions are still significant (especially that involving the hydrogen radical) and two smaller ranges where the aldehyde is negligent and by-product reactions (especially the carbonmonoxide to carbon dioxide conversion, which has steadily increased to a maximum) become significant.

The last time range, 0.117 through 0.150, is also broken down into two short ranges, the final stages of the transition to final product reactions and two longer ranges where the final product production reactions are the only significant reactions.

4.3 Cluster Descriptions

Table 1 lists the description of the first level clusters. The descriptions were ordered according to their clustering significance and only the most significant are shown. As explained previously (Section 3.6), these show the differences between the clusters. Tables 2, 3 and 4 show the second level cluster descriptions of the first, second and third main cluster, respectively.

Table 5 shows the most significant reactions, as defined by those with the sensitivity value throughout the given range. The significant reactions of the first range (0.0-0.098) correspond to the main reaction path from aldehyde to formyl radical to carbon monoxide. The second and third ranges are concerned with the production of final products of water and carbon dioxide.

In comparing the first cluster (0.0 - 0.098 seconds) and the last (0.117 - 0.150), there is no overlap in the ranges listed (Table 1), hence the descriptions provide a way to distinguish these ranges. The reactions not involving carbon (reactions of peroxide and peroxyl radical) have low values in a low range in the first cluster and a higher range in the third cluster. Conversely, the reaction involving carbon (the reaction from formyl radical to carbon monoxide) has a high value range for the first cluster and a low value range for the third cluster. This quantifies the observation that the first range involves the first set of reactions involving direct aldehyde products and the third range involves the reactions of the later products.

5 Discussion

Through the hierarchy of clusters, a hierarchy of dynamic reactive phases of a mechanism was obtained. These phases were determined by an unsupervised (unbiased) algorithm where the sole input was the sensitivity values of the reactions at a different times throughout the entire process. No explicit knowledge of time was used to determine these ranges and nevertheless, distinct contiguous phases were clustered together. Furthermore, the cluster determination was free from outside influences, i.e. they were determined automatically. Optimistically, this means that the resulting clusters are 'unbiased'; Pesimistically, this means that the result is not necessarily exactly what the chemist would expect. The results of this study shows, however, that the phases determined can be both quantitatively and qualitatively justified.

Time (secs)	Reaction	Sensitivity Region	Importance
0.0-0.098	HCO+M=CO+H+M	$S > 0.3$	0.66
	H+H2O2=H2+HO2	$0.005 > S$	0.63
	H2O2+H=H2O+OH	$0.06 > S$	0.59
	H2O2+OH=H2O+HO2	$0.015 > S$	0.58
	H+H2O2=H2+HO2	$0.01 > S$	0.58
	H+HO2=O2+H2	$0.01 > S$	0.57
0.098-.117	HCO+O2=CO+HO2	$0.02 < S < 0.06$	0.98
	HCHO+OH=HCO+H2O	$0.1 < S < 0.35$	0.95
	CO+OH=CO2+H	$S > 0.001$	0.91
	HCHO+H=HCO+H2	$0.2 < S < 0.5$	0.90
	OH+H2=H2O+H	$0.02 < S < 0.12$	0.87
	H+HO2=O2+H2	$0.005 < S < 0.015$	0.85
	HCO+M=CO+H+M	$0.03 < S < 0.3$	0.78
	H+H2O2=H2+HO2	$0.005 < S < 0.025$	0.66
	HCHO+O=HCO+OH	$S > 0.18$	0.57
	H+O2=OH+O	$0.1 > S$	0.55
0.117-.150	HCO+M=CO+H+M	$0.03 < S < 0.15$	0.84
	HCHO+O=HCO+OH	$0.02 < S < 0.08$	0.80
	H2O2+OH=H2O+HO2	$0.02 < S < 0.04$	0.75
	H+HO2=O2+H2	$0.01 < S < 0.02$	0.73
	H+H2O2=H2+HO2	$0.01 < S < 0.025$	0.71

Table 1: The descriptions of the first level of clustering. The 'Time' fields gives the reactive time range represented by the cluster. The 'Reaction' lists the reactions that are important in defining this time region (note that this is not necessarily reactions with high sensitivity values - see text). The 'Sensitivity Region' field gives the range of sensitivity values for the reaction. The 'Importance' field gives a relative importance of this reaction in defining the cluster.

Time (secs)	Reaction	Sensitivity Region	Importance
0.0-0.0019	CO+HO2=CO2+OH	$S < 0.0005$	0.81
	H+O2=OH+O	$S > 0.16$	0.73
	HCHO+HO2=HCO+H2O2	$S > 0.25$	0.71
	H2O2+M=OH+OH+M	$0.08 > S$	0.62
	HCHO+OH=HCO+H2O	$0.15 < S < 0.4$	0.59
	HO2+HO2=H2O2+O2	$0.01 > S$	0.52
0.0019-0.0127	HCHO+O2=HCO+HO2	$0.01 < S < 0.06$	0.98
	HCHO+O=HCO+OH	$0.14 < S < 0.18$	0.94
	H+O2=OH+O	$0.14 < S < 0.18$	0.92
	HCHO+HO2=HCO+H2O2	$S > 0.25$	0.71
	HCHO+O=HCO+OH	$S > 0.14$	0.68
	CO+HO2=CO2+OH	$S < 0.0015$	0.68
	H2O2+M=OH+OH+M	$0.08 < S < 0.18$	0.64
0.0127-0.066	HCHO+HO2=HCO+H2O2	$0.1 < S < 0.25$	0.84
	H+O2=OH+O	$0.06 < S < 0.14$	0.84
	HCHO+OH=HCO+H2O	$S > 0.4$	0.74
	CO+HO2=CO2+OH	$S > 0.001$	0.73
	H+O2=OH+O	$0.08 < S < 0.14$	0.72
	HCHO+O=HCO+OH	$0.06 < S < 0.16$	0.66
0.066-0.098	HCO+O2=CO+HO2	$0.03 < S < 0.08$	0.95
	CO+OH=CO2+H	$S > 0.001$	0.92
	CO+O+M=CO2+M	$0.3 < S < 0.8$	0.92
	HCHO+HO2=HCO+H2O2	$0.05 < S < 0.15$	0.87
	H+HO2=O2+H2	$0.005 < S < 0.01$	0.86
	HCHO+O=HCO+OH	$0.08 > S$	0.84

Table 2: The descriptions of the second level of clusters within the time range of 0.0 through 0.098. The fields are as those in Figure 1

Time (secs)	Reaction	Sensitivity Region	Importance
0.098-.111	HCHO+OH=HCO+H2O	$0.2 < S < 0.35$	0.88
	CO+O+M=CO2+M	$0.4 < S < 0.9$	0.88
	H+HO2=OH+OH	$0.03 < S < 0.06$	0.88
	H2O2+M=OH+OH+M	$0.14 < S < 0.18$	0.88
	HCO+M=CO+H+M	$S > 0.2$	0.86
0.111-.113	HCHO+OH=HCO+H2O	$0.15 < S < 0.2$	0.95
	H2O2+M=OH+OH+M	$0.1 < S < 0.14$	0.94
	H+HO2=OH+OH	$0.01 < S < 0.03$	0.92
	H2O2+H=H2O+OH	$0.06 < S < 0.1$	0.90
	H2O2+OH=H2O+HO2	$0.015 < S < 0.03$	0.89
	H+OH+M=H2O+M	$0.4 < S < 0.6$	0.87
0.113-0.115	H2O2+H=H2O+OH	$0.1 < S < 0.14$	0.91
	H2O2+OH=H2O+HO2	$0.025 < S < 0.04$	0.89
	HCO+HO2=OH+H+CO2	$0.001 < S < 0.004$	0.88
	H2O2+M=OH+OH+M	$0.06 < S < 0.12$	0.88
	OH+H2=H2O+H	$0.04 < S < 0.08$	0.84
	H+H2O2=H2+HO2	$0.015 < S < 0.025$	0.82
.115-.117	HCHO+H=HCO+H2	$0.1 < S < 0.4$	0.93
	OH+H2=H2O+H	$0.06 < S < 0.14$	0.92
	HCO+M=CO+H+M	$0.03 < S < 0.15$	0.90

Table 3: The descriptions of the second level of clusters within the time range of 0.098 through 0.117. The fields are as those in Figure 1.

Time (secs)	Reaction	Sensitivity Region	Importance
0.117-0.118	H2O2+H=H2O+OH	$0.02 > S$	0.95
	HCHO+O=HCO+OH	$0.04 < S < 0.080$	0.93
	H+H2O2=H2+HO2	$S > 0.015$	0.92
	HCO+H=CO+H2	$0.015 < S < 0.025$	0.91
	HCHO+OH=HCO+H2O	$0.05 < S < 0.15$	0.90
	HCHO+H=HCO+H2	$0.1 < S < 0.3$	0.89
0.118-0.119	HCHO+H=HCO+H2	$0.1 < S < 0.2$	0.91
	HCO+O=CO2+H	$S < 0.0016$	0.88
	HCO+H=CO+H2	$0.025 < S < 0.03$	0.86
	HCHO+OH=HCO+H2O	$0.05 < S < 0.1$	0.84
	HCO+OH=CO+H2O	$0.002 < S < 0.003$	0.82
	HCO+O=CO+OH	$0.0012 < S < 0.0016$	0.70
0.119-0.125	HCO+O=CO2+H	$0.0006 > S > 0.0016$	0.68
	HCO+O=CO+OH	$S > 0.0016$	0.66
	HCO+OH=CO+H2O	$0.0005 > S > 0.0025$	0.65
	HCO+H=CO+H2	$S > 0.025$	0.60
	H2O2+OH=H2O+HO2	$(0.005 < S < 0.025$	0.53
	H2O2+H=H2O+OH	$(0.08 < S < 0.1$	0.52
0.125-0.150	OH+H2=H2O+H	$0.04 < S < 0.14$	0.93
	H+HO2=O2+H2	$S > 0.015$	0.89
	HCO+O2=CO+HO2	$0.01 > S$	0.86
	H2+O=OH+H	$0.01 < S < 0.05$	0.82
	HCO+H=CO+H2	$0.005 < S < 0.025$	0.80
	HCO+OH=CO+H2O	$0.001 < S < 0.0025$	0.77

Table 4: The descriptions of the second level of clusters within the time range of 0.117 through 0.150. The fields are as those in Figure 1.

Time Range	Reactions
0.0-0.098	
	HCHO+H=HCO+H2
	HCHO+OH=HCO+H2O
	HCO+M=CO+H+M
	H+O2+M=HO2+M
	CO+O+M=CO2+M
0.098-0.117	
	H+OH+M=H2O+M
	CO+O+M=CO2+M
	HCHO+H=HCO+H2
0.117-0.150	
	H+OH+M=H2O+M
	H+O2+M=HO2+M
	CO+O+M=CO2+M

Table 5: The main reactions of the top clustered time ranges. The main reactions were determined by examining the sensitivity values throughout the region. If the reaction had a sensitivity value that was among the top eight throughout at least half of the region, the reaction was included in this list.

5.1 Usage

This method is proposed as a supplement to existing methods. It automatically provides another set of information, namely regions where the reactivity is similar. When used in conjuction with current methods, a more refined picture of the reactive process could be obtained. Instead of doing sensitivity analysis (for mechanism reduction or even coupled with reaction path analysis) on the entire mechanism at once, these analyses could be done within the time regions isolated. Mechanism reduction is refined because it adds an extra criteria that a reaction cannot be eliminated if it was deemed significant in any of the time regions isolated. Path analysis is refined because, first, within an entire region one is dealing with a (defined by the method) similar set of reacting conditions and, secondly, different regions are isolated which give different significant paths.

A sensitivity constant is used in this method, as in others, to give a measure of the importance of a reaction. However, in contrast, the purpose of this method is not primarily to reduce reaction mechanisms. Traditionally, those reactions above a certain sensitivity value are considered significant and are essential to the mechanism (those below are eliminated from the mechanism). In the method proposed, time regions are found where the same reactions have the same significance, i.e. throughout the region the same reactions are important. So in this method, the sensitivity constant is purely a measure used to characterise the reactive state.

In this study, the chemical 'experiment' was examining dynamic process over time. The time states, local snapshots of at given times of the sensitivity values, were clustered. The use of the cluster method is not, however, limited to just such processes. The only real requirement is that the states to be clustered represent a continuous change of parameters from one state to the next. For example, instead of using time states, one could use the x-coordinate in a one-dimensional experiment. Each state would be the conditions at each x-coordinate. The only practical obstacle could be the number of states that are needed. Though, mathematically, there is no lower bound to the number of states needed to form clusters (no singularities or other undefined situations will occur), to be statistically reasonable, a significant number of states are needed (relative to the number of uncoupled parameters).

Once the mechansim data has been generated by SENKIN, the complexity of the algorithm is relatively simple (as compared to the SENKIN run). The current implementation is in the **C** programming language (using GNU public domain software) within the more general ANALYSIS system under several UNIX systems.

5.2 Further Work

This work can be seen as establishing a systematic method calculating a particular characterisation of a given mechanism under one set of initial conditions. Future work will use this characterisation as a means of comparison between different mechanisms and initial conditions. The clustering structure would be compared. For example, the clustering structure could be examined as the set of initial conditions is changed. As the initial temperature is increased, it would be expected that the clustering structure would be the same until the pyrolysis reaction become more dominant. Another experiment, which could be relevant for this reaction is to slowly change the oxygen to aldehyde ratio. It is not certain at this time whether two different mechanisms that should be modeling the same process would have the same clustering structure (the reasons being due to the unsupervised or 'unbiased' nature of the clustering algorithms explained previously).

The clustering method used in this study was based on machine learning principles. It would be expected that other hierarchial clustering methods (as outlined in section 2) would also yield significant clustering results. The clustering method used in this study is prefered because there is more flexibility in how the time states can be represented. Future studies will be devoted to using other predicate representations. Among these will be the use of fuzzy logic predicates.

6 Conclusion

The clustering method applied to reaction mechanisms provides an automatic and, hence, unbiased means to differentiate between reactive phases within a total reactive process through various levels of refinement. Though the analysis uses a measure that is primarily for mechanism reduction, its purpose here is to gain further understanding of the dynamic behavior of existing mechanisms. Indirectly, however, the differentiation into distinct reactive phases promotes reaction reduction through the assurance that the most significant reactions in each phase are kept; Minor, but nevertheless significant phases or reactions would not be overshadowed by more dominant phases. Though this study only dealt with the analysis of a single mechanism with one set of initial conditions, further work is underway to use this method as a concrete comparision of many reactive conditions though the analysis of how the clustering characteristics change. In conclusion, the example shown in this paper provides a sysstematic reduction of the expanse of numeric data that is associated with a complex mechanism.

References

[1] J.T. Tou and R.C. Gonzalez. *Pattern Recognition Principles.* Addison-Wesley, 1974.

[2] Edward S. Blurock. Reaction: Symbolic methods in the generation and analysis of complex reaction systems. In Bertrand Braunschweig and Bernt Bremdal, editors, *Artificial Intelligence in the Petroleum Industry: Symbolic and Computational Applications II*, Paris, 1996. Editions Technip.

[3] Edward S. Blurock. Reaction: Symbolic modeling of complex reaction systems. In *10. CIC-Workshop Information und Wissen am Arbeitsplatz des Chemikers*, Hochfilzen, Austria, 1995.

[4] R.J. Kee, J.A. Miller, and T.H. Jefferson. CHEMKIN: A General-Purpose, Problem-Independent, Transportable, FORTRAN Chemical Kinetics Code Package. Technical Report SAND80-8003, Sandia National Laboratories, 1980.

[5] A.E. Lutz, R.J. Kee, and J.A. Miller. SENKIN: A Fortran Program for Predicting Homogeneous Gas Phase Chemical Kinetics With Sensitivity Analysis. Technical Report SAND87-8248, Sandia National Laboratories, 1994.

[6] Edward S. Blurock. Analysis 3.0: Implementation, Extensions and Use of the ID3 Algorithm. Technical report, RISC-Linz, Linz, Austria, 1992.

[7] Edward S. Blurock. Automatic learning of chemical concepts: Research octane number and molecular substructures. *Computers Chem.*, 19:91–99, 1995.

[8] S. Hochgreb and F.L. Dryer. A Comprehensive Study on CH_2O Oxidation Kinetics. *Combustion and Flame*, 91:257–284, 1992.

[9] D. L. Baulch, C. J. Cobos, R. A. Cox, C. Esser, P. Frank, J. A. Just, T. Kerr, M. J. Pilling, J. Troe, R. W. Walker, and J. Warnatz. Hydrocarbon combustion reaction constants. *J. Phys. Chem Ref. Data*, 21:411–429, 1992.

[10] S. Vajda, P. Valko, and T Turanyi. Mechanism reduction with principle component analysis. *Int. J. Chem. Kinet.*, 17:55–81, 1985.

[11] John E. Freund. *Modern Elementary Statistics*. Prentice-Hall, Inc., 1979.

[12] D. Fisher. Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, 2:139–172, 1987.

[13] John H. Gennari, Pat Langley, and Doug Fisher. Models of incremental concept formation. In J. G. Carbonell, editor, *Machine Learning: Pardigms and Methods*, pages 11–61. MIT Press, 1990.