# Problems in the INSPEC classification scheme

R.J. Henery
Statistics and Modelling Science
Strathclyde University
Glasgow G1 1XH
UK
bob@uk.ac.strath.stams

S. Greatrix & G.J. Meggs
Intelligent Systems Unit
BT Laboratories
Martlesham Heath
Suffolk IP5 7RE
UK
sgjg@info.bt.co.uk
gavin.meggs@bt-sys.bt.co.uk

February 12, 1997

**Abstract**

The INSPEC database is described, and the structure of the classification/keyword relationship investigated. Clusterings among the classifications and keywords are investigated. It appears that there are very few (as expected from an information-theoretic point of view). Thus classification and keyword schemes are essentially mutually exclusive. However, the relative rareness of classifications and keywords make the distinction between statistically independent and mutually exclusive a very fine one. Time-scales for changes in the structure are outlined. Classifications change over timescales of order 20 years, keywords with a time-scale of 5-10 words. Illustrative examples are discussed.

## Keywords

Hierarchical clustering; classification; INSPEC database; IEE classification; cluster validity; keyword similarity; adaptive; discount past; performance monitoring; bias; trend; Bayesian.

## 1 Introduction and Overview

The IEE maintains a large database, accessible remotely by telnet, on papers published in Physics, Electronics and Computing, known as the INSPEC database. A fuller description of the database is given in section **??**. The database covers over 4000 journals

and is currently adding more than 300,000 published papers and documents per year. Each record gives information on names of the authors, abstract, keywords, etc., so it may be difficult to decide the best way to scan the database for articles of interest. One potential use of the proposed procedure is as an aid to searches of the database.

We used the full INSPEC database for the years 1990-1996, containing more than a million records, for overall statistics relating to trends in classifications or keywords (rarely both). For more detailed work, we used a small subsample.

### 1.0.1 Small sample of 1000 papers

A small sample of 1000 consecutive papers, from around the middle of the full database, was examined in greater depth. The sample is described in section 1.2, and this sample was used to illustrate the Bayesian learning procedure in sections **??** and **??**, and also to investigate clustering of keywords and classifications in sections 1.2.3 and 1.2.2. At best, this small sample can only give a small snapshot of what was happening around the middle of 1993: at worst, the sample is completely unrepresentative of current trends. But this is precisely the problem at hand: how do we determine if what we see is a random fluctuation or part of a trend.

Some of the keywords relevant to "pollution and their frequencies over the years 1990-1996 are shown in table 1.

| Year | Key1 | Key2 | Key3 | Key4 | Total |
|---|---|---|---|---|---|
| 1990 | 1244 | | | 317 | 258348 |
| 1991 | 1577 | | | 465 | 259812 |
| 1992 | 1515 | | | 601 | 290960 |
| 1993 | 1246 | | | 548 | 250842 |
| 1994 | 1708 | 27 | 15 | 728 | 274097 |
| 1995 Jan-Jun | 871 | 230 | 152 | | 140441 |
| 1995 Jul-Dec | 879 | 176 | 207 | | 164542 |
| 1996 Jan-Jun | 1162 | 212 | 191 | | 158588 |

Table 1: Frequency of occurrence of three keywords: Key1 = "pollution"; Key2 = "pollution control", Key3 = "pollution measurement"; Key4 = "pollution detection and control". Key4 was in use until the end of 1993, when the term was replaced by the two keys Key2 and Key3.

Judging from the relative frequencies, it would appear that there has been an effective dichotomy of Key4 into either Key2 or Key3 (but not both). Perhaps surprisingly, in view of the trends noted later in section 2, namely those relating to the proportions of

papers in classification A8670 (Environmental Science), the number of papers quoting keyword "pollution" seems to have increased steadily until around 1994, from which time the proportion has remained constant.

A longer term aim would be to consider redesigning the IEE system in response to new topics or to changing emphasis within topics.

### 1.0.2 Clustering Classifications

Clustering of classifications may help in widening the search for relevant papers. On the one hand, the lack of apparent correlation between allocated classifications greatly simplifies the structure of the database, and reduces the number of items retrieved in any search, but on the other hand, the resulting search may well be much too narrow.

An alternative aim might be to merge classifications that are no longer fashionable, or are redundant (an almost-equivalent classification exists at same level of the hierarchy). An investigation of the small INSPEC sample of 1000 papers (in sections 1.2.2 and 1.2.3) shows that there is remarkably little association among the allocated classifications, probably due to a desire to avoid redundancy. See, for example, figure 2 for the clustering of classifications in the small sample. However, it would be misleading to conclude from figure 2 that there are no natural clusters of classifications. The true clustering of classifications only emerges if the hierarchy of the classifications is exploited.

### 1.0.3 Should we split an existing class?

When rapid expansion in a research area leads to what might be termed a new discipline, we need to create a new sub-classification. In its simplest terms this would involve splitting the class into two, so our work should certainly address this simplest of all problems. This raises the question of when should we split a class and, if so, how. This is an unsupervised learning problem. One possible solution is along semantic lines. Suppose, for instance, that a lot of work is being done on "atmospheric pollution measurement devices", and it is required to introduce an extra level into the hierarchy. This could be done on the basis of the "upper atmosphere" and "lower atmosphere", or perhaps on the nature of the measurement device, "laser" and "infra-red". This would raise the question of splitting other classifications using the same keywords, i.e. introducing a higher level in the hierarchy. If no other classifications are to be affected, it is the lowest level of the hierarchy that would be affected. Some kind of balance is required between the creation of new classifications, which might bring more order to current searches, and the preservation of old classifications, which simplifies searches over the entire database.

A secondary aim is to identify situations where a class may be too large and contain identifiable subclasses. Splitting this class may have greater conceptual use (e.g. by simplifying subsequent searches). For example the split (class + keyword) / (class - keyword) might be suggested on information-theoretic grounds. An extended example is considered in section 1.3.

## 1.1 Changes in database - Jan 1990 to Jun 1996

### 1.1.1 Classifications change in 20 year timescale

Definitions of classifications are rarely changed, but we can quote a few examples of changes to illustrate the time-scales involved. The broad category **A86** was introduced in 1979. Material under the topic Radiochemistry would be classified under A1930 (1969-72), A9260 (1973-1976), A8255 (1977-date). The dramatic developments in neural computing necessitated drastic changes in the C-section (Computing) in recent years.

### 1.1.2 Keywords change in 10 year timescale

The definitions of keywords, or more precisely, desciptors, is subject to change also. A particular example is noted above for the descriptor "pollution detection and control" which was used up to 1993, and two new keywords used in its place (see table 1). Other keywords change their meaning, or are replaced, on a time-scale of 5-10 years.

### 1.1.3 Popularity of classifications changes in 5 year timescale

It is of interest to know if classification C is predicted consistently, giving a homogeneous class, and is this class consistent with the definition? However, it is very difficult to assess this without referring to the original articles, or at the very least, the abstracts. We can at least look for consistency in classification frequencies. For example, if there is evidence of a trend, in that one particular class is becoming more frquent, is this evidence of changing fashions in research or a change in the effective definitions in the classification scheme? One source of confounding in the INSPEC scheme is the split of work from 1995, some classifications being performed in Karlsruhe, and some in London. Can we see any evidence of this? The answer is yes, possibly. Table 2 gives the frequency of occurrence of sub-classifications in the environmental science category A8670 for the years 1990-1996.

Figure 1 gives a correspondence analysis of the frequencies in table 2, showing how A8670 sub-classifications vary with year of publication. The first eigenscore represents linear variation by year of publication, so that the dominant trends might be said to be approximately linear in time. Clearly the variation of frequency in class A8670J is markedly different from the rest, denoting a substantial increase in relative popularity of

| a8670s | A8670 | A8670C | A8670E | A8670G | A8670J | A8670L | A8670Z |
|---|---|---|---|---|---|---|---|
| 1990 | 35 | 174 | 217 | 613 | 36 | 154 | 32 |
| 1991 | 80 | 215 | 249 | 703 | 23 | 210 | 46 |
| 1992 | 53 | 157 | 212 | 600 | 41 | 244 | 54 |
| 1993 | 42 | 184 | 152 | 390 | 26 | 196 | 25 |
| 1994 | 61 | 183 | 196 | 568 | 77 | 266 | 44 |
| 1995 Jan-Jun | 63 | 140 | 132 | 256 | 41 | 153 | 23 |
| 1995 Jul-Dec | 42 | 193 | 173 | 329 | 38 | 228 | 59 |
| 1996 Jan-Jun | 86 | 156 | 166 | 319 | 113 | 212 | 46 |

Table 2: Frequency of occurrence of classifications in the environmental science category A8670. Note the marked increase in the frequencies of classes A8670 and A8670J relative to the most frequent class A8670.

this class.

A fully Bayesian procedure would adjust automatically to such trends. However, it is important also to flag their existence, and to characterise their type, as they may be due to extraneous sources (in this case a potential divergence of operational procedures between London and Karlsruhe).

## 1.2 Reduced Sample - May 1994

The 1000 papers selected for the study consisted of 1000 consecutive records from the INSPEC database, so the papers from one issue of a journal occupied consecutive positions in the sample. This was very useful in providing a form of clustering by publication as some journals are very narrow in scope (e.g. *Chemical Physics Letters* would probably have IEE classifications **A33** or **A82**, while *Corrosion Science* would probably be confined to **A82**). However, the down side to using a systematic sample like this is that the sample is not representative of the complete range of journals. In future, it is intended to use a much larger random sample.

### 1.2.1 Sample Statistics

Overall, the 1000 papers involved

- an average of 3.827 IEE classifications per paper
- an average of 11.132 keywords per paper

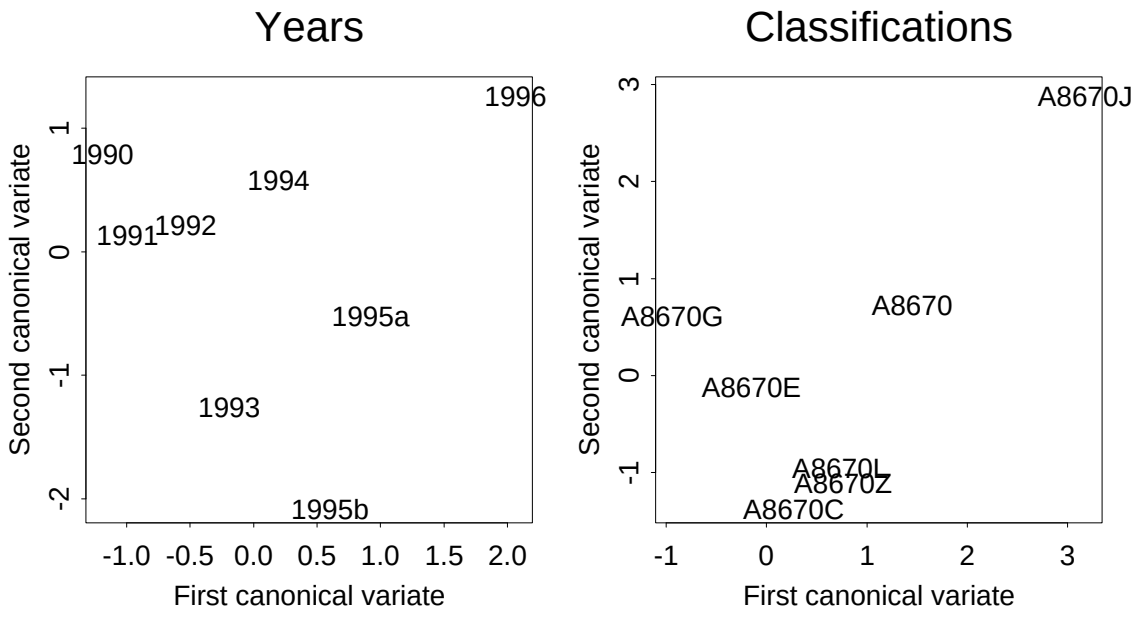


Figure 1: Correspondence between classifications and years. The first eigenscore accounts for two thirds of the variation in frequency, and is almost synonymous with date of publication. Classification A8670J stands out from the others in respect of its variation over this period.

- 86 different journals
- 89 issues of journals (i.e. some journals are represented more than once)
- 1061 different IEE classifications at level four (counting A8670Z and A8670G as different)
- 126 IEE classifications at level two (counting A82 and A81 as different).
- 2202 different keywords (so the vocabulary size is 2202 keywords).

### 1.2.2 Similarity and Structure of Classifications

From the small sample of 1000 records from 1993, we calculated the correlations between the occurrences of all classifications. In fact there were very few significantly large correlations, probably because the marginal frequencies were so small in this small sample, but also because the tendency would be not to assign a classification if a closely related classification had already been assigned to a paper. For example, if classifications A8670L has already been assigned to a paper, the addition of classification A8670G would not add much information about the paper, whereas a distant classification, say C3310G, would be much more informative. Figure 3 gives the clustering of classifications in the small INSPEC sample.

For whatever reason, there are very few pairs of highly correlated classifications, with correspondingly few clusters. In particular note that there is only one cluster with classifications at the same level of the hierarchy (A9710R and A9730E).

### 1.2.3 Similarity and Structure of Keywords

A very similar story is true for keywords. Figure 3 gives the clustering of keywords in the small INSPEC sample, based on correlations in the co-occurrence of keywords in the 1000 papers.

Once again, there are surprisingly few clusters of keywords, although it must be borne in mind that the correlations are based on a very small sample. There are a few definite clusters, however, and for the most part these are quite predictable, as they involve the same stem (e.g. "stellar models" and "stellar winds"). We use this lack of correlation when constructing the Bayesian model in section **??**, as it simplifies matters considerably to assume that predictors (keywords) are independent.

On a more speculative note, suppose we wish to establish a hierarchical classification scheme based only on the keywords used. There is little doubt that such a scheme would rely heavily on the semantics of the keywords, for instance whether the word “air” is used

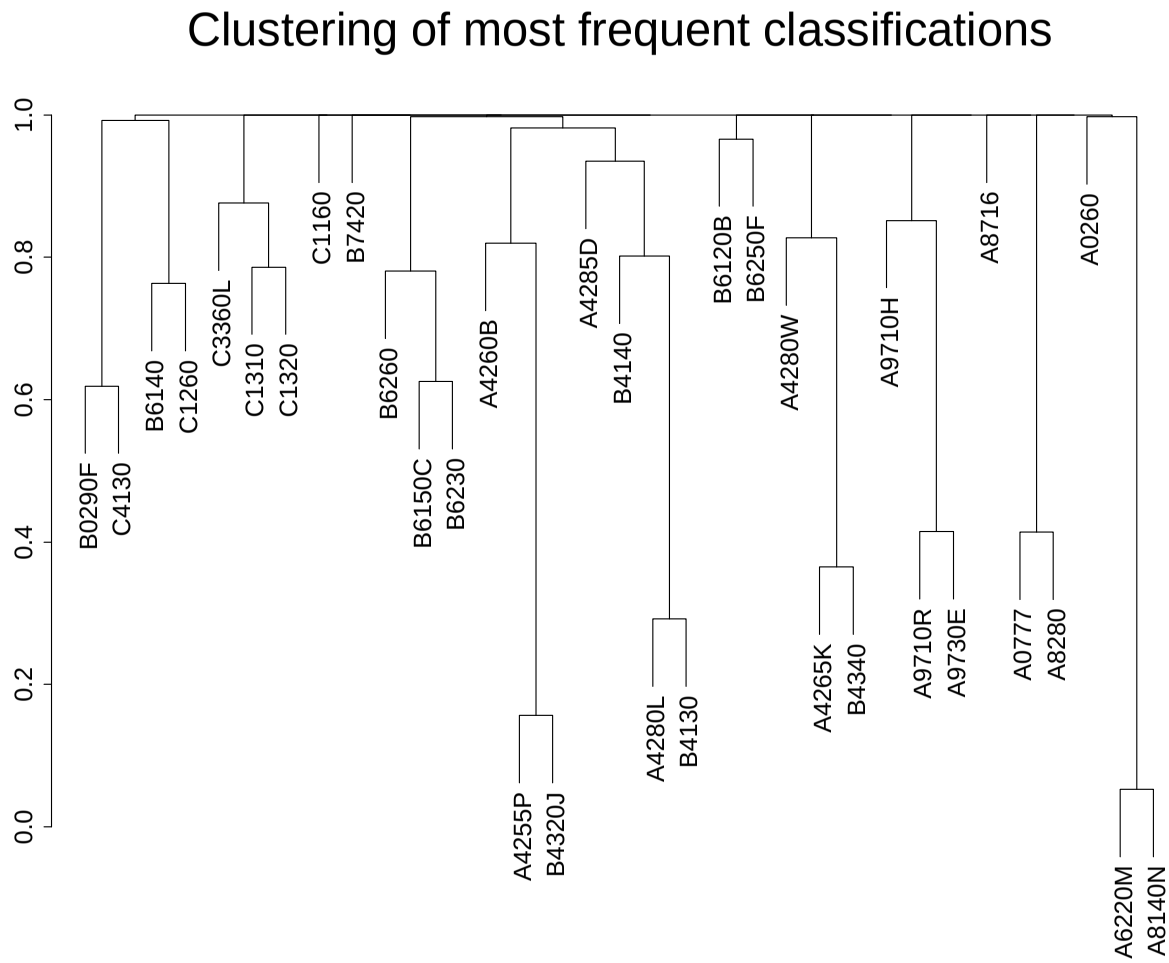


Figure 2: Clustering of classifications for small INSPEC sample. Note how few clusters relate to classifications at the same level of the hierarchy, the only real exception being (A9710R and A9730E).

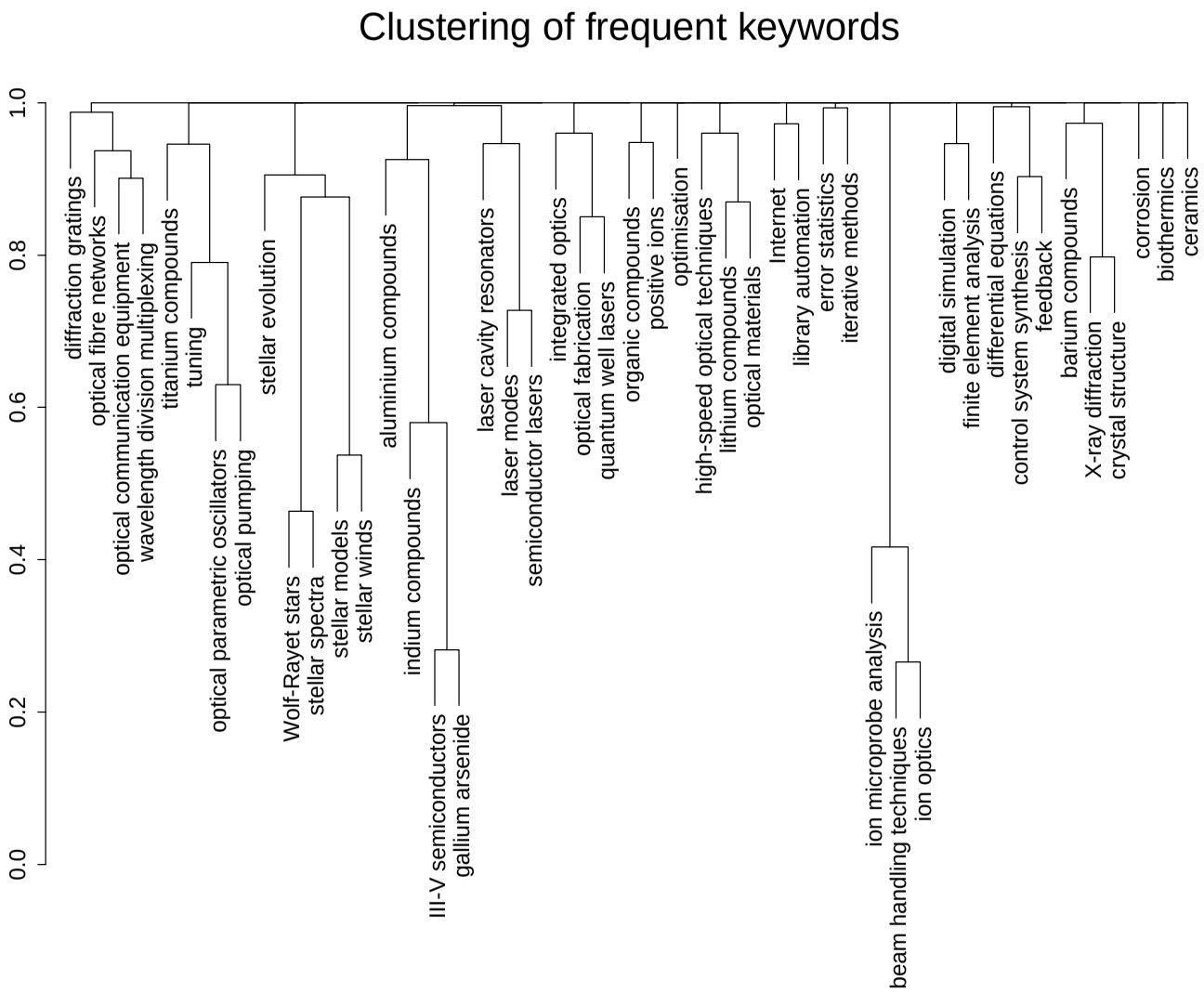


Figure 3: Clustering of keywords for small INSPEC sample. Again note how few clusters relate to keywords sharing the same stem, the two notable exceptions being ("stellar models" and "stellar winds") and ("ion microprobe analysis" and "ion optics")..

as noun, adjective or verb. We will probably look at keyphrases only, treating a phrase like 'atmospheric pollution measurement devices" either as a whole or as the simple sum of its four component words without regard to order. Even then, it will be necessary to establish a table of similarities for words and phrases (such as done, although very crudely, in search engines for the internet). This might be done on the basis of the number of letters in common (electron being similar to electric). However, if we are able to use the existing IEE scheme as basis, we can determine similarities on a more objective basis, namely how many times these terms appear together in the same list of keywords.

For example the word "electron" appears in 78 of the sample papers, and "electric" appears in 29. If these words were used independently, we would expect $78 * 29/1000 = 2.262$ papers with both words as part of the keywords. In fact only one paper (5140887) has both words, with keywords:

> *capacitance, carrier mobility, conduction bands, deep levels, electrical conductivity, electronic structure, Fermi level, gallium arsenide, III-V semiconductors, leakage currents, relaxation, Schottky barriers, Schottky diodes, semiconductor counters, space charge*

(note that we have counted "electrical" as an instance of "electric" and "electronic" as an instance of "electron"). we would conclude that the words "electron" and "electric" are **not similar** in their use as keywords. This is an illustration of the approximate independence of keywords that might have been supposed to be very similar *a priori*.

## 1.3 Extended Example - Pollution Cluster

To illustrate the ideas involved, consider the following simple example. Suppose that we have identified a potential cluster (defined in terms of the keywords only). We must compare this cluster with the IEE classification scheme to see if the cluster is valid (in IEE terms). This will involve identifying the nearest equivalent IEE classification or group of classifications, and then quantifying how well the proposed cluster performs relative to the IEE scheme.

### 1.3.1 Pollution Cluster

As a simple example, suppose we consider all papers which have the word "pollution" somewhere in the list of keywords as the **Pollution Cluster**. This cluster definition has the distinct merit of requiring only a single concept (pollution). The concept "pollution" is reasonably well defined in its everyday usage, and might be expected to produce a reasonably coherent cluster. There are 12 papers among the sample of 1000 papers chosen. These are listed in Table 3 together with their IEE classifications.

| INSPEC number | IEE Classifications |
|---|---|
| 5140000 | A8670Z |
| 5140001 | A8670Z |
| 5140002 | A8670Z |
| 5140003 | A8670Z |
| 5140004 | A8670Z |
| 5140005 | A8670Z |
| 5140006 | A8670Z A8670L |
| 5140008 | A8670Z |
| 5140009 | A8670Z |
| 5140034 | A9260S A8270R A8670G A9260M A9265V A9260T A9260J |
| 5140098 | A8160B A9260T |
| 5140189 | A8670L A3350D A8280D A3510B A3370F A8670G A9260K |

Table 3: Pollution cluster: papers that have the word **pollution** as part of their keywords, together with the IEE classifications.

## 1.4 IEE classifications involving Pollution

It is immediately apparent from Table 3 that the most relevant IEE classification is A8670Z, or, slightly more generally, any classification of the form A8670x, where **x** is a one-letter code such as **Z** or **G**. This suggests the following rule for IEE classification A8670:

> Classify the paper as A8670 if and only if the word **pollution** appears as one of the keywords.

This rule commits no errors of omission (all papers with classification A8670 are included in the membership) and only one error of commission (one paper is in class **Pollution Cluster** that is not given classification A8670). To see this, refer to Table 4, which gives all those papers with IEE classification A8670.

### 1.4.1 Should paper 5140098 be classified as A8670G?

A closer inspection of the data for paper 5140098 suggests that this paper is concerned with corrosion caused by pollution. The IEE have classified the paper as A8160B and A9260T. The former is relevant to the keyword "corrosion", but perhaps it should also be given a classification of A8670, as pollution plays an important role, and indeed "air pollution" is a keyword. As evidence for this, we quote the keywords and abstract. The keywords for paper 5140098 are:

> *air pollution, corrosion, environmental degradation, steel*

| INSPEC paper | Keywords |
|---|---|
| 5140000 | particle size, pollution |
| 5140001 | artificial satellites, pollution |
| 5140002 | artificial satellites, pollution |
| 5140003 | modelling, pollution |
| 5140004 | pollution control |
| 5140005 | artificial satellites, pollution control |
| 5140006 | pollution control |
| 5140008 | collections of physical data, pollution |
| 5140009 | pollution, pollution control, pollution measurement, remote sensing by radar |
| 5140034 | aerosols, air pollution, atmospheric precipitation, climatology |
| 5140189 | air pollution measurement, atmospheric temperature, fluorescence, iodine, isotope detection, optical radar, remote sensing by laser beam, spectral line intensity, spectrochemical analysis |

Table 4: Key words for the IEE classification **A8670**. Note that the key word **pollution** appears in every paper that has classification A8670, excepting only paper 5140098. However, comparing with Table 3 it seems probable that paper 5140098 should also be in this table as it too has the word **pollution** as part of the keywords (see section 1.4.1).

and the abstract is

> *A model of the influence of chloride deposition rate and $SO_2$ deposition rate on atmospheric corrosion of steel has been proposed. Accumulated corrosion and pollution data obtained since 1979 from different Cuban sites were statistically processed. A model is obtained when data from all corrosion stations are processed independently of the type of climatic territory. The influence of chloride ions is very significant in determining corrosion rate when there is already a corrosion products layer; however, when this layer is not completely formed the influence of time of wetness is the controlling factor. The existence of a competitive adsorption process between chloride and sulphur compounds present in airborne salinity is very possible.*

Referring again to the data of Table 4, the first 8 papers are classified as A8670Z and the last two as A8670G. Can we deduce, using only the keywords in Table 4, why this distinction is made? One possibility is to classify as A8670G if "air pollution" is a keyword. This problem is a sort of supervised learning problem, with the allocated IEE classifications as known classes and the keywords as predictors (regarded as relational attributes). Machine Learning algorithms capable of solving this problem in full generality would be ILP, CN2, AQ, CRS (see Sammut (1994) for a description of relational algorithms).

### 1.4.2 Should IEE classification A4265 be split?

Papers are very unequally shared between the four sections (A, B, C and D), of the IEE classification scheme. The most recently created section, D - Information Technology, accounts for only 53 of the 1000 papers in the sample. By contrast, a single classification A4265 (in the Physics section) accounts for 57 papers. On a purely numerical argument, it appears that there is a good case for splitting classification A4265 further, or perhaps modifying the hierarchical scheme higher up the hierarchy to spread the numbers over a greater range of classes. This class would form a good working example for unsupervised learning, with no existing structure to guide us, but with sufficient numbers to justify the formation of new classes.

On a separate tack, it appears that classification B4340 is very similar to classification A4265 since B4340 is given as an additional classification in 50 of the 57 papers classed as A4265. Of course the prefix letters A and B imply that the same topic is being approached from a Physics and an Electronics angle respectively. In this case, there would be an argument for making the numerical coding the same (for instance A4265 and B4265) as would probably be done in an automatic hierarchical scheme.

# References

Sammut, C. (1994). Knowledge representation. In *Machine Learning, Neural and Statistical Classification*, pages 228–245. Ellis Horwood.