



Available at

[www.ElsevierComputerScience.com](http://www.ElsevierComputerScience.com)

POWERED BY SCIENCE @ DIRECT®

INTERNATIONAL JOURNAL OF  
APPROXIMATE  
REASONING

ELSEVIER International Journal of Approximate Reasoning xxx (2003) xxx–xxx

[www.elsevier.com/locate/ijar](http://www.elsevier.com/locate/ijar)

## Comparison of neural models for document clustering

Vicente P. Guerrero-Bote<sup>a,\*</sup>, Cristina López-Pujalte<sup>a</sup>,  
Félix de Moya-Anegón<sup>b</sup>, Victor Herrero-Solana<sup>b</sup>

<sup>a</sup> *Facultad de Biblioteconomía y Documentación, Universidad de Extremadura, 06071 Badajoz, Spain*

<sup>b</sup> *Facultad de Biblioteconomía y Documentación, Universidad de Granada, Campus Cartuja, Granada, Spain*

Received 1 January 2003; accepted 1 July 2003

---

### Abstract

12 We compared the application of different algorithms to document clustering. The  
13 algorithms studied were Fuzzy C-Means, Fuzzy ART, Fuzzy ART for Fuzzy Clusters,  
14 Fuzzy Max-Min, and the Kohonen neural network (only the first is not a neural net-  
15 work). We generated a testbed from LISA, using some of the descriptors corresponding  
16 to the different records for the comparison of the results. The best results were found  
17 with Kohonen's algorithm which also organizes the clusters topologically. We end by  
18 discussing in more detail the possibilities offered by Kohonen's algorithm.

19 © 2003 Published by Elsevier Inc.

20 *Keywords:* Artificial neural networks; Kohonen networks; Document clustering;  
21 Adaptive resonance theory

---

### 22 1. Introduction

23 Although one could say that sequential algorithm technology has vastly  
24 surpassed human capacities in certain tasks, such as performing mathematical

---

\* Corresponding author. Tel.: +34-924-289300; fax: +34-924-286401.

*E-mail addresses:* vicente@alcazaba.unex.es (V.P. Guerrero-Bote), clopez@alcazaba.unex.es (C. López-Pujalte), felix@goliat.ugr.es (F. de Moya-Anegón), victorhs@ugr.es (V. Herrero-Solana).

25 operations, other tasks that humans find easy are found to be very difficult for  
26 these classical methods to solve. Examples are optical character recognition,  
27 image processing, speech, etc.

28 Artificial neural networks arose in response to problems of this nature, and  
29 offer a way of attacking some otherwise unapproachable problems. The dif-  
30 ferent definitions that have been proposed for these networks all emphasize the  
31 great number of processing elements that they consist of [12], their massive  
32 interconnection [16], their arrangement in layers and their inspiration in other  
33 biological characteristics of the human brain [13], etc. They usually have an  
34 associated training procedure by means of which they adapt to the problem at  
35 hand, and also require long processing times. Nonetheless, the process is  
36 massively parallel and lends itself to being run on a computational structure  
37 that is itself a physical implementation of a neural network. One could  
38 therefore say that a neural network is at once a set of problem-solving algo-  
39 rithms and a computational structure.

40 Because of the variety of problems that they are capable of solving, some  
41 workers have seen them as a new paradigm of artificial intelligence. It has been  
42 found, however, that the field's current citation environment is quite distinct  
43 from that of artificial intelligence [32].

44 While there exist different kinds of neural network with quite varied func-  
45 tioning, they have a series of common characteristics in the way they process  
46 information that in the most part represent advantages and have led to their  
47 application in several areas [25]. Some of these characteristics are: adaptive  
48 learning (the capacity to learn to perform tasks on the basis of initial training  
49 or experience); self-organization (organization of the information that they  
50 receive during training in the weight structure of the network) which allows  
51 generalization (when they are presented with novel conditions or data they  
52 respond by generalizing what they had learnt before); fault tolerance (in the  
53 two senses of being able to respond to noisy data and of robustness against  
54 failure of a part of the network); and real-time operation (as they allow parallel  
55 processing, the processing speed can also be increased). In brief, each element  
56 performs thousands of information processing operations, and their sum gives  
57 rise to the intelligent overall behaviour of the network [5]. This form of pro-  
58 cessing is found to be best suited to tasks with a greater complexity, whereas it  
59 is poorly suited to traditional mathematical operations or similar tasks.

60 A neural network is capable of assigning multidimensional outputs to  
61 multidimensional inputs as a function of what was learned in the training  
62 phase, which is done offline in most networks. In other words, almost all  
63 network architectures consist of two phases, one of training and the other of  
64 production. This is known as the stability-plasticity dilemma. There also exist  
65 types of networks such as those corresponding to adaptive resonance theory  
66 (ART) that are capable of overcoming this stability-plasticity dilemma by  
67 continuing to learn when they are in production. The training too may be

68 either supervised or unsupervised. The difference is that in supervised networks  
 69 one uses a set of pairs formed by an input and its corresponding output to  
 70 adapt the network to the desired outputs on the basis of the mistakes that it  
 71 makes, while unsupervised networks, which are the cases that we have analysed,  
 72 cluster the training inputs. Thus, the first application that this type of  
 73 algorithm might have with document inputs is to perform a clustering operation  
 74 during this stage of training, and to find the cluster corresponding to a  
 75 document or an information query (if it is transformed into a document vector).  
 76 The most desirable networks for this purpose would therefore be those  
 77 that use unsupervised learning because they already perform clustering, and the  
 78 ARTs because they allow learning to continue into the production phase.

79 In the present study, we shall compare the following models of neural  
 80 networks for document clustering: Fuzzy ART [2,3], Fuzzy Max-Min [31],  
 81 Fuzzy ART for Fuzzy Clusters [28], and Kohonen's Model [15,17-20]. To  
 82 these models we have added an advanced non-neural clustering algorithm  
 83 which is widely used in various applications of artificial intelligence, expert  
 84 systems, etc. This is the Fuzzy C-Means method of Bezdek [1].

## 85 2. Material and methods/data and methods

86 In order to carry out the experiments, we generated a group of documents  
 87 extracted from the bibliographic database LISA on CD-ROM, whose documents  
 88 consist of literature references from articles belonging to the LIS. The  
 89 records all possess the same structure: a series of fields, one of which is an  
 90 abstract. We considered each of these abstracts to be an independent document.  
 91 Of the remaining fields, we ignored all except the descriptors, some of  
 92 which we used to test the results.

93 The purpose of the experiment was to compare the clustering capacities of  
 94 the five algorithms. We therefore created a small database of 64 documents to  
 95 allow us to test each algorithm's capacities quickly. We did this by retrieving  
 96 from the complete database those references that had at least one of the following  
 97 descriptors:

|   |              |
|---|--------------|
| Document management                           | 18 Documents |
| Information audits                            | 13 Documents |
| Oncology                                      | 12 Documents |
| Libraries and the future                      | 13 Documents |
| Internet primer for information professionals | 8 Documents  |
| Total   | 64 Documents |

4

*V.P. Guerrero-Bote et al. / Internat. J. Approx. Reason. xxx (2003) xxx-xxx*

98 As one sees from the table, the descriptors chosen gave a testbed database  
99 that was as varied as possible. The three first descriptors retrieved records that  
100 had quite uniformly extensive abstracts, suggesting that this type of document  
101 is broadly disseminated over the whole document space. The abstracts corre-  
102 sponding to the other two descriptors, however, were only two or three lines  
103 long, and also very similar to each other. Hence, we would be able to observe  
104 whether the algorithms possess the autoscaling property, adapting themselves  
105 to different document densities.

106 The resulting database contained these 64 documents and 1085 different  
107 words.

108 *Document vectorization/document representation:* The next step was to apply  
109 the vector space model to transform these documents into vectors that one  
110 could use as inputs for the algorithms. To this end, we first had to determine a  
111 series of terms that could serve to characterize the documents of the database,  
112 and then assign the corresponding weight to each term in each document.

113 As weighting scheme, we chose one that is very close to the classical *tf · idf*  
114 which, in the study of Noreault et al. [9,29], was one of those that yielded the  
115 best results. The weight of each component is determined by the expression:

$$a_{ij} = t_{ij} \cdot \log \frac{F}{f_j}$$

117 where  $a_{ij}$  = weight assigned the term  $t_j$  in the document  $D_i$ ,  $t_{ij}$  = number of  
118 times that the term  $t_j$  appears in the document  $D_i$ ,  $f_j$  = number of times that the  
119 term  $t_j$  appears in the entire database,  $F$  = total number of different words  
120 (tokens) in the database.

121 We decided to use this scheme instead of the classical IDF, because the latter  
122 assigns the greatest weights to the terms that appear in only a single document.  
123 Since we wish to cluster the documents, terms that only appear in a single  
124 document are translated into differences of the said document with the rest, and  
125 are of no use in clustering.

126 Nevertheless, there still remained a problem to solve. We have algorithms  
127 that expect fuzzy inputs, and the components of the document vectors that we  
128 have generated so far do not have to be between zero and one. Hence, given  
129 that in the aforementioned study [29] the best similarity measures were those  
130 which made angular comparisons between vectors, we opted to normalize the  
131 vectors by dividing them by their Euclidean norms. The final result is a set of  
132 vectors with many components of which by far the most are zero, and those  
133 that are not are quite small in value since the total norm is unity.

134 The total number of different terms extracted was 1085. Although in itself  
135 this is a manageable number, it would become far larger for a realistically sized  
136 document database, which would make these experiments inviable. We  
137 therefore set ourselves the question of how best to reduce the number of term.

138 The ideal way would be to select the terms that have the greatest discrimination  
139 values [8,9]. Due to the small size of our database and because the lower the  
140 frequency of the term the greater the weight assigned to it, many of the terms  
141 with the greatest discrimination values appear in a single document. As we  
142 mentioned above, these terms are of little use in clustering since they translate  
143 into distances to all the documents equally, i.e., they do not help to find sim-  
144 ilarities with some documents and differences with others which is what is  
145 needed to form clusters of documents. We therefore made the reduction on the  
146 basis of the frequency and the number of documents in which each term ap-  
147 peared. The procedure consisted of four phases:

148 • *Elimination of stopwords*: This first phase is designed to eliminate words of  
149 the language that have function but no meaning. To this end, since the lan-  
150 guage is English, we used the frequency dictionary of Kucera and Francis  
151 [21], generating a list of 200 stopwords corresponding to the words of great-  
152 est frequency of this language.

153 • *Elimination of words that they appear in a single document*: Given that our  
154 document representation is to be used to form clusters, the effect of eliminat-  
155 ing the words that appear in only one document is null. It was in this phase  
156 that the greatest number of words were eliminated.

157 • *Elimination of words with a very high frequency*: There are words that, while  
158 they can not be considered stopwords in general, do behave as such in this  
159 small database, so that we eliminated the words of the highest frequency.  
160 The threshold that we set was a maximum frequency of 30 in the 64 documents.

161 • *Stemming*: We performed a weak reduction by eliminating some prefixes and  
162 suffixes using the Porter Stemmer, a rule-based algorithm which is the most  
163 widely used in English [6,30].

164 With this reduction procedure, we were left with a set of 246 words which  
165 formed the final indexing terms. Hence, for each document a 246-component  
166 vector will be generated, each component being the weight of the corre-  
167 sponding term in the document.

168 In practice, the second and third phases were performed together, so that the  
169 procedure consisted of three steps. In the first, the stopwords were eliminated,  
170 with which the 1085 different words were reduced to 984. In the second, we  
171 eliminated those that appeared in a single document and those with a fre-  
172 quency greater than 30, thereby reducing the total to 286 words. And in the  
173 third, stemming yielded the definitive database with 246 terms. Fig. 1 shows  
174 how each of these three steps affected the frequency distribution as the dat-  
175 abase was reduced in size. The following features stand out:

176 • The elimination of the stopwords did not greatly alter the frequency charac-  
177 teristics. The only difference was the greater proportion of words of fre-  
178 quency one in the second. This is logical, since the terms just eliminated  
179 were those of highest frequency in the language, and their removal shifted  
180 the distribution to the lowest frequencies.

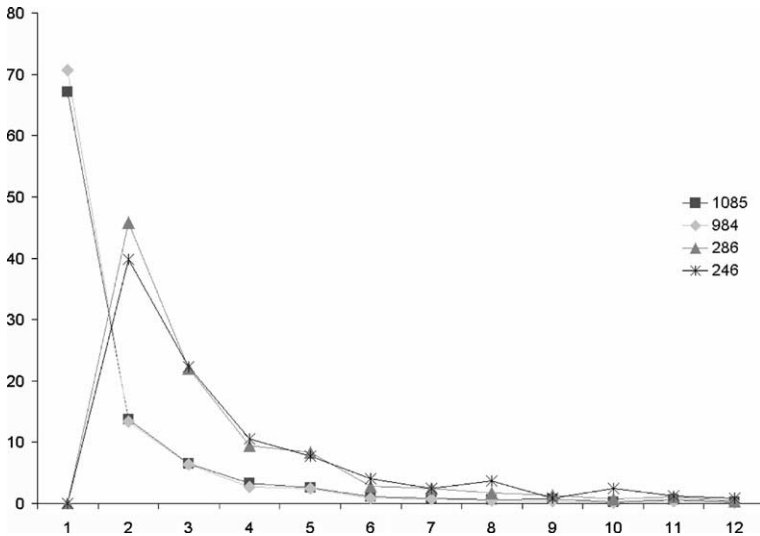


Fig. 1. The frequency distribution of terms at each reduction step according to their discrimination value.

- 181 • The other two curves, however, are very different due mainly to the drastic  
 182 elimination of the many words that appeared in a single document, with the  
 183 consequent sharp growth in the other frequencies. The more than threefold  
 184 reduction in number of terms was due to the elimination of these frequency-  
 185 one terms.
- 186 • Between the last two curves, there is a slight decline in low-frequency terms  
 187 and a corresponding rise in higher frequencies. This, of course, is because  
 188 stemming groups together words that possess the same root.

### 189 3. Results of the different techniques

#### 190 3.1. The fuzzy C-means method

191 This algorithm is the only one that requires the user to specify the number of  
 192 clusters into which the outputs are to be classified and the maximum variation  
 193 of the membership matrix before terminating. We therefore tried several val-  
 194 ues, generating random initial values of the membership matrix and using  
 195 Euclidean distances, all of which yielded similar classifications. An example is  
 196 the following, corresponding to 40 classes, superscript of the membership  
 197 matrix 2, and maximum variation 0.0000001 (1.0 E7):

| Cluster | Retrieval descriptor                          | Number of documents |
|---------|---|---------------------|
| 8       | Internet primer for information professionals | 8/8                 |
| 8       | Libraries and the future                      | 13/13               |
| 24      | Oncology                                      | 1/12                |
| 35      | Document management                           | 18/18               |
| 35      | Information audits                            | 13/13               |
| 35      | Oncology                                      | 11/12               |

198 This indicates that cluster 8 includes all the documents retrieved by the  
 199 descriptors *Internet primer for information professionals* and *Libraries and the fu-*  
 200 *ture*, and that cluster 35 includes all the documents retrieved by the descriptors  
 201 *Document management* and *Information audits* and most of those retrieved by  
 202 the *Oncology* descriptor. This means that the algorithm confuses the docu-  
 203 ments belonging to these descriptors, although it does differentiate between the  
 204 two broad classes of documents present in the database.

205 There was practically no variation in the result when the parameters of the  
 206 algorithm were varied, since the classifications remained very similar beginning  
 207 with only a very few clusters.

208 While this is not a good result, it has to be borne in mind that the database is  
 209 very small and that we are also granting it an absolute perfection in both the  
 210 elaboration of the abstracts and the assignment of the descriptors.

211 To all this one should add the fact that this is a fuzzy clustering algorithm,  
 212 i.e., its outputs represent a degree of membership to each cluster. However, in  
 213 order to compare it with the rest of the algorithms, we kept only the cluster  
 214 with the greatest degree of membership for each document, so that we were  
 215 making use of only a part of the classification.

216 In this case one might consider generating clusters that are not disjoint, but  
 217 rather to which all the documents surpassing a certain threshold belong.

### 218 3.2. Fuzzy ART

219 The parameters in this case are (i) the learning coefficient  $\alpha$  that regulates by  
 220 how much the weight vector of a node that has become resonant is modified,  
 221 and (ii)  $\rho$  which is the vigilance parameter. The first results that we obtained  
 222 were 7 clusters each of which included practically all the descriptors. To sepa-  
 223 rate them, we gradually increased the vigilance parameter to 0.26, setting  
 224  $a = 0.1$ . This gave a network that stabilized in a few iterations and whose result  
 225 is the following:

| Cluster | Retrieval descriptor                          | Number of documents |
|---------|---|---------------------|
| 0       | Document management                           | 1/18                |
| 1       | Internet primer for information professionals | 8/8                 |
| 1       | Libraries and the future                      | 2/13                |
| 2       | Document management                           | 2/18                |
| 3       | Information audits                            | 2/13                |
| 4       | Information audits                            | 1/13                |
| 5       | Information audits                            | 2/13                |
| 6       | Document management                           | 1/18                |
| 7       | Document management                           | 1/18                |
| 8       | Information audits                            | 1/13                |
| 9       | Information audits                            | 1/13                |
| 10      | Oncology                                      | 1/12                |
| 11      | Document management                           | 1/18                |
| 12      | Information audits                            | 2/13                |
| 13      | Information audits                            | 1/13                |
| 14      | Document management                           | 2/18                |
| 15      | Document management                           | 1/18                |
| 16      | Document management                           | 1/18                |
| 17      | Document management                           | 1/18                |
| 18      | Document management                           | 1/18                |
| 19      | Document management                           | 1/18                |
| 20      | Document management                           | 1/18                |
| 21      | Document management                           | 1/18                |
| 22      | Information audits                            | 1/13                |
| 23      | Oncology                                      | 1/12                |
| 24      | Document management                           | 1/18                |
| 25      | Oncology                                      | 1/12                |
| 26      | Oncology                                      | 2/12                |
| 27      | Oncology                                      | 1/12                |
| 28      | Oncology                                      | 1/12                |
| 29      | Oncology                                      | 1/12                |
| 30      | Oncology                                      | 1/12                |
| 31      | Information audits                            | 2/13                |
| 32      | Oncology                                      | 1/12                |
| 33      | Oncology                                      | 1/12                |
| 34      | Oncology                                      | 1/12                |
| 35      | Document management                           | 1/18                |
| 36      | Document management                           | 1/18                |
| 37      | Libraries and the future                      | 11/13               |

227 The number of clusters can be reduced by using a smaller value for the  
 228 vigilance parameter. A decrease in this parameter from 0.26 to 0.22 only led to  
 229 a reduction of 5 in the number of clusters. However, different descriptors began  
 230 to group together in the same clusters, as can be seen in the following:

| Cluster | Retrieval descriptor                          | Number of documents |
|---------|---|---------------------|
| 0       | Document management                           | 1/18                |
| 1       | Internet primer for information professionals | 8/8                 |
| 1       | Libraries and the future                      | 3/13                |
| 2       | Document management                           | 2/18                |
| 3       | Information audits                            | 2/13                |
| 4       | Information audits                            | 1/13                |
| 5       | Information audits                            | 2/13                |
| 6       | Document management                           | 1/18                |
| 7       | Document management                           | 1/18                |
| 7       | Oncology                                      | 1/12                |
| 8       | Information audits                            | 1/13                |
| 9       | Oncology                                      | 1/12                |
| 10      | Document management                           | 2/18                |
| 11      | Document management                           | 2/18                |
| 12      | Document management                           | 1/18                |
| 13      | Document management                           | 1/18                |
| 14      | Document management                           | 1/18                |
| 15      | Document management                           | 1/18                |
| 15      | Information audits                            | 1/13                |
| 16      | Document management                           | 1/18                |
| 17      | Information audits                            | 1/13                |
| 18      | Oncology                                      | 1/12                |
| 19      | Document management                           | 1/18                |
| 20      | Oncology                                      | 1/12                |
| 21      | Oncology                                      | 2/12                |
| 22      | Oncology                                      | 1/12                |
| 23      | Oncology                                      | 1/12                |
| 24      | Oncology                                      | 1/12                |
| 25      | Oncology                                      | 1/12                |
| 26      | Oncology                                      | 1/12                |
| 27      | Information audits                            | 2/13                |
| 28      | Document management                           | 2/18                |

(continued on next page)

| Cluster | Retrieval descriptor | Number of documents |
|---------|----------------------|---------------------|
| 29      | Document management  | 1/18                |
| 30      | Information audits   | 2/13                |
| 31      | Information audits   | 1/13                |
| 32      | Oncology             | 1/12                |

231 This is clearly worse than before since, while it spreads the documents out  
 232 over different clusters, it mixes those of *Internet primer for information pro-*  
 233 *fessionals* and *Libraries and the future*. Decreasing the vigilance parameter so as  
 234 to yield only 7 clusters resulted in completely mixing the documents of different  
 235 descriptors.

236 In order to understand this result, one must bear in mind that this algorithm  
 237 expects fuzzy inputs, whereas we have generated document vectors that lie on  
 238 the unit hypersphere with components between zero and one, and that this  
 239 restriction might not be the most appropriate. Perhaps some other represen-  
 240 tation needs to be investigated that is better suited to these fuzzy algorithms.

241 Unlike the previous case, there is no fuzzy classification.

### 242 3.3. Fuzzy ART for fuzzy clusters

243 In this case, the only operating parameter is the vigilance parameter  $\rho$ .  
 244 Whereas in the Fuzzy ART, an increase in this parameter decreased the size of  
 245 the clusters, it is now the contrary. The results with this algorithm were:

| Cluster | Retrieval descriptor                               | Number of documents |
|---------|--|---------------------|
| 0       | Internet primer for information profes-<br>sionals | 8/8                 |
| 0       | Libraries and the future                           | 13/13               |
| 11      | Document management                                | 3/18                |
| 12      | Document management                                | 1/18                |
| 12      | Oncology   | 4/12                |
| 13      | Oncology   | 1/12                |
| 14      | Document management                                | 1/18                |
| 14      | Information audits                                 | 6/13                |
| 14      | Oncology   | 1/12                |
| 15      | Document management                                | 2/18                |
| 16      | Document management                                | 5/18                |
| 17      | Document management                                | 4/18                |
| 17      | Oncology   | 1/12                |
| 18      | Document management                                | 1/18                |
| 18      | Information audits                                 | 2/13                |

|    |                     |      |
|----|---------------------|------|
| 19 | Oncology            | 3/12 |
| 20 | Document management | 1/18 |
| 20 | Information audits  | 1/13 |
| 20 | Oncology            | 2/12 |
| 21 | Information audits  | 4/13 |

246 This algorithm, which is not totally neural, yielded a classification very  
 247 similar to the previous case in that it did not manage to separate the documents  
 248 of the *Internet primer for information professionals* and *Libraries and the future*  
 249 descriptors. It did, however, form larger clusters of documents of a given de-  
 250 scriptor. For example, cluster 14 included 6 of the 13 *Information audits* doc-  
 251 uments.

252 Lastly, although this method is based on the C-means algorithm, the results  
 253 of the two algorithms were very different. Also, one must remember that this  
 254 algorithm has a fuzzy output, i.e., as also in the case of the C-means algorithm,  
 255 one is only making use of a part of the classification.

#### 256 3.4. Fuzzy max-min

257 In this algorithm the operating parameters are the size of the hyperboxes  
 258 and the slope of the membership function. For a size of 15, the results were the  
 259 following:

| Cluster | Retrieval descriptor                          | Number of documents |
|---------|---|---------------------|
| 0       | Document management                           | 2/18                |
| 0       | Internet primer for information professionals | 1/8                 |
| 0       | Libraries and the future                      | 4/13                |
| 1       | Information audits                            | 3/13                |
| 2       | Document management                           | 2/18                |
| 2       | Internet primer for information professionals | 6/8                 |
| 3       | Information audits                            | 1/13                |
| 3       | Internet primer for information professionals | 1/8                 |
| 3       | Libraries and the future                      | 9/13                |
| 4       | Information audits                            | 2/13                |
| 4       | Oncology                                      | 1/12                |
| 5       | Document management                           | 1/18                |
| 5       | Information audits                            | 2/13                |
| 6       | Document management                           | 3/18                |
| 6       | Information audits                            | 1/13                |

(continued on next page)

| Cluster | Retrieval descriptor | Number of documents |
|---------|----------------------|---------------------|
| 7       | Document management  | 3/18                |
| 8       | Document management  | 4/18                |
| 9       | Document management  | 2/18                |
| 9       | Information audits   | 1/13                |
| 10      | Information audits   | 1/13                |
| 10      | Oncology             | 2/12                |
| 11      | Document management  | 1/18                |
| 11      | Oncology             | 2/12                |
| 12      | Oncology             | 3/12                |
| 13      | Oncology             | 2/12                |
| 14      | Information audits   | 1/13                |
| 14      | Oncology             | 1/12                |
| 15      | Information audits   | 1/13                |
| 15      | Oncology             | 1/12                |

260 One sees that, with so much mixing, this is one of the worst results that we  
 261 obtained, although there are some clusters that recall the previous cases.

262 The results seem to confirm that the documents of *Internet primer for in-*  
 263 *formation professionals* and of *Libraries and the future* are very close in the  
 264 document space while the others are more separated. This is coherent with  
 265 what was seen during the generation of the database. As yet however, none of  
 266 the techniques has been capable of distinguishing between the two.

267 In this case the output is also fuzzy. The values are not excessively large since  
 268 all the degrees of membership are very high. The output is generated by aver-  
 269 aging the non-coincidences. Since the vectors have very few non-zero compo-  
 270 nents, there will be many coincidences and hence very large membership values.  
 271 To avoid this, we averaged only those components that were different from zero  
 272 in one of the two vectors concerned, thereby obtaining smaller values. In any  
 273 case, this may be one of the causes of the mixing, i.e., that the technique takes  
 274 great account of null coincidences. It might be interesting to test functions based  
 275 on angles or on the distance from the edge of the hypercube.

276 Although the *Fuzzy Max-Min* technique was designed for fuzzy vectors, its  
 277 operation is similar for non-fuzzy vectors. One has to bear in mind, however, that  
 278 this algorithm forms hyperboxes which, in our case, have to cover the hyper-  
 279 sphere of unit norm, a situation which does not seem to be the most appropriate.

### 280 3.5. Kohonen networks

281 For this test we used a two-dimensional map of  $4 \times 4$  nodes, resulting in 16  
 282 different clusters. The number of iterations was 1000, the initial neighbourhood

| Col./Row | 0                          | 1   | 2                          | 3                          |
|----------|----------------------------|---|----------------------------|----------------------------|
| 0        | Internet (6/8)             | Internet (2/8)<br>Libraries Future (2/13) | Libraries Future (6/13)    | Libraries Future (5/13)    |
| 1        | Oncology (1/12)            | Oncology (3/12)<br>Oncology (1/12)        | Oncology (4/12)            | Document Management (1/18) |
| 2        | Oncology (1/12)            | Document Management (2/18)                | Oncology (2/12)            | Information Audits (4/13)  |
| 3        | Document Management (3/18) | Document Management (1/18)                | Document Management (2/18) | Information Audits (4/13)  |
|          | Document Management (5/18) | Document Management (4/18)                | Information Audits (4/13)  | Information Audits (5/13)  |

Fig. 2. Classification obtained with a Kohonen neural network of dimension  $4 \times 4$ .

283 was  $4 \times 4$  (the entire network) decreased by 1 every 200 iterations, and the  
 284 learning rate was  $\alpha = 0.01$ . As well as these parameters, we used a network with  
 285 conscience which influences negatively the number of won competitions in  
 286 order to avoid the problem of stuck vectors. The result is the map of Fig. 2.

287 While, as in the previous results, there exists a certain degree of confusion,  
 288 we understand this technique to be more satisfactory than any of the others  
 289 since it also provides topological information.

290 Because of this topological organization, documents on similar topics are  
 291 clustered in the same zones of the map. For instance, the documents corre-  
 292 sponding to the descriptor *Information audits* occupy clusters near the lower  
 293 right-hand corner, those of *Libraries and the future* the upper right-hand cor-  
 294 ner, those of *Internet primer for information professionals* the upper left-hand  
 295 corner, and those of *Document management* the lower left-hand corner.  
 296 Lacking more corners, one sees that the remaining descriptor, which is spread  
 297 over the central zone, is the cause of most cases of confusion.

298 There is one detail of this technique that must be taken into account. To  
 299 judge the similarity of two documents, this algorithm calculates the difference  
 300 between the corresponding vectors. Since the vectors are normalized, it relies  
 301 on the angle between them, i.e., the smaller the angle, the more similar it  
 302 considers the vectors to be. As we noted above, this coincides with what was  
 303 found to be the best choice in studies of similarity measures between document  
 304 vectors [29].

305 This type of network is also capable of organizing documents in a two-di-  
 306 mensional space. For this, one has to define a number of hidden units that is  
 307 greater than the number of documents. We performed such a topological or-  
 308 ganization on a map of  $10 \times 10$  units, with an initial  $10 \times 10$  neighbourhood that  
 309 is reduced by 1 each 45 iterations while maintaining the rest of the parameters  
 310 fixed. The result is shown in Fig. 3.

311 One sees in the figure that the winning zones of each descriptor are fairly  
 312 well determined. There exist some cases of confusion, however, especially those  
 313 at the positions (3,4) and (4,6) that correspond to the *Oncology* descriptor but  
 314 are located in the zone of the *Document management* descriptor (the same  
 315 descriptor that it was mixed with in the previous case).

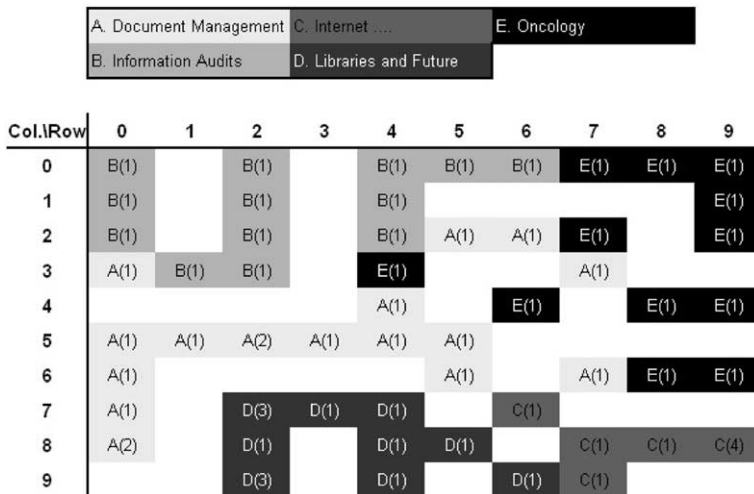


Fig. 3. Classification obtained with a Kohonen neural network of dimension  $10 \times 10$ .

316 Let us analyse in more detail one of these cases, looking at the content of the  
 317 documents present in the zone where the error occurred to see whether there  
 318 exists any resemblance. We shall choose the cluster corresponding to position  
 319 (3,4) and examine the following two documents that were classified in the  
 320 positions (3,4) and (4,4):

321 *Document corresponding to position (3,4):*

322 Database Title: Library and Information Science Abstracts

323 Accession Number: 894789

Title in English: Treatment of uncertainty in an oncology protocol by probabilistic and artificial intelligence approaches.

326 Author LN: de Rosis

327 Author FN: Fiorella

328 Author LN: Steve

329 Author FN: G.

330 Author LN: Biagini

331 Author FN: C.

332 Author LN: Maurizi-Enrici

333 Author FN: R.

334 Source: Methods of Information in Medicine

335 Source Info: 27 (1) Feb 88, 23-33. illus. tables. 22 refs.

Abstract: The decision process for diagnosis and treatment of Hodgkin's disease at the Institute of Radiology of Rome has been modelled integrating the guidelines of a protocol with uncer-

tainty aspects. Two models have been built, using a PROSPECTOR-like expert system shell for microcomputers: the first of them treats the uncertainty by the inferential engine of the shell, the second is a probabilistic model. The decisions suggested in a group of simulated and real cases by a second of the two models have been compared with an objective final diagnosis; this analysis showed that, in some cases, the two models give different suggestions and that approximations of the shell's inferential engine may induce wrong conclusions. A sensitivity analysis of the probabilistic model showed that the outputs are greatly influenced by variations of parameters, whose subjective estimation appears to be especially difficult. This experience gives the opportunity to consider the risks of building clinical decision models based on expert system shells, if the assumptions and approximations hidden in the shell have not been previously analysed in a careful and critical way. Original abstract

357 Classification: ZmVsRnM(616-006)

Feature Heading: Searching. Strategies. Data bases.  
Information services. Oncology. Expert systems

360 Language: English

361 Publication Year: 1988

362 Subject: Technical services

363 Information storage and retrieval

364 Information work

365 Subject indexing

366 Online information retrieval

367 Computerised information retrieval

368 Searching

369 Computerized information storage and retrieval

370 Strategies

371 Databases

372 Information services

373 Expert systems

374 Oncology

375 Cancer

376 *Document corresponding to position (4,4):*

377 Database Title: Library and Information Science Abstracts

378 Accession Number: 9502105

379 Title in English: Designing an expert system for classifying office documents.

16 V.P. Guerrero-Bote et al. / Internat. J. Approx. Reason. xxx (2003) xxx-xxx

Author LN: Savic

Author FN: D.

Source: Records Management Quarterly

Source Info: 28 (3) Jul 94, p. 20-9. refs.

Abstract: Can records management benefit from artificial intelligence technology, in particular from expert systems? Gives an answer to this question by showing an example of a small scale prototype project in automatic classification of office documents. Project methodology and basic elements of an expert system's approach are elaborated to give guidelines to potential users of this promising technology.

Language: English

Publication Year: 1994

Subject: Expert systems

Document management

Automatic classification

398 One sees that both documents are about expert systems: besides their cor-  
399 responding descriptors in our reduced testbed database (*Oncology* and *Docu-*  
400 *ment management*), they both had the *Expert systems* descriptor in the original  
401 database which was not picked up in our restricted subset of descriptors. It  
402 seems therefore that this network performs a more detailed analysis than is  
403 allowed for by the descriptors that we used.

#### 404 4. Conclusions

405 The algorithms that we have used can be classified into two groups. On the  
406 one hand there are those that expect a fuzzy input on which they operate using  
407 fuzzy operators, and on the other those that do not expect fuzzy inputs and  
408 perform the classification on the basis of the distances between the vectors,  
409 independently of whether or not the output is fuzzy. The Fuzzy ART and  
410 Fuzzy Max-Min belong to the first group, and the classifications using these  
411 algorithms were the worst of all since, although the input is between 0 and 1, its  
412 origin is in the IDF weights and not in degrees of membership. It would be  
413 advisable to test other different representations for these networks.

414 Better results were given by the algorithms that treat the inputs as vectors  
415 and perform the classification on the basis of the distance or (which comes  
416 down to the same thing) the angle between the vectors. As we mentioned be-  
417 fore, this corroborates the findings of studies carried out on similarity mea-  
418 sures. In this group, the best results were obtained with the Kohonen networks  
419 which, as well as the clustering, yield a topological organization and hence  
420 provide more information than the other techniques. This type of network is

421 currently being employed in *text data mining* [22] and to generate *topological*  
422 *maps* of document sets, even labeling each word or term's zone of influence  
423 [4,7,10,11,14,20,23,24,26,27].

424 We have to mention that the testbed database was selected to be difficult to  
425 classify. It contained documents of great size and others that were very small  
426 and differed very little from each other. This meant that while the documents  
427 corresponding to *Internet primer for information professionals* and *Libraries and*  
428 *the future* were very close together, those corresponding to the rest of the de-  
429 scriptors were spread out over the document space so that it was very difficult  
430 to separate the former without diversifying the latter. The network that was  
431 able to best adapt to this was Kohonen's model.

432 One must also take into account that our testbed database was classified on  
433 the basis of abstracts of original references, and that this classification was then  
434 tested against some of the descriptors that had been assigned not on the basis  
435 of the abstract, but of the entire reference.

436 A neural network is trained to learn the responses to certain inputs. In the  
437 case of unsupervised learning, this is done by clustering the inputs. These  
438 networks, however, were designed to respond to queries, i.e., to be capable of  
439 finding interactively the cluster nearest a given document vector which could  
440 belong to a query rather than a document. They not only perform the clus-  
441 tering therefore, but are subsequently able to calculate interactively which  
442 cluster is closest to a given query. The query simply has to be transformed into  
443 a document vector.

#### 444 Acknowledgements

445 This work was financed by the Junta de Extremadura-Consejería de Edu-  
446 cación Ciencia and Tecnología and the Fondo Social Europeo, as part of re-  
447 search project 2PR02A041.

#### 448 References

- 449 [1] J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum, NY,  
450 1981.
- 451 [2] G.A. Carpenter, S. Grossberg, D.B. Rosen, Fuzzy art: an adaptative resonance algorithm for  
452 rapid, stable classification of analog patterns, in: *International Joint Conference on Neural*  
453 *Networks Seattle II*, Piscataway, NJ, 1991, pp. 411–416.
- 454 [3] G.A. Carpenter, S. Grossberg, D.B. Rosen, Fuzzy art: fast stable learning an categorization of  
455 analog patterns by an adaptative resonance system, *Neural Networks* 4 (1991) 759–771.
- 456 [4] H. Chen, A. Houston, R. Sewell, B. Schatz, Internet browsing and searching: user evaluations  
457 of category map and concept space techniques, *J. Am. Soc. Inform. Sci.* 49 (7) (1998) 582–603.

- 458 [5] T.E. Doszko, J. Reggia, X. Lin, Connectionist models and information retrieval, in: M.E.  
459 Williams (Ed.), *Annual Review of Information Science and Technology*, vol. 25, Elsevier  
460 Science Publishers BV, Amsterdam, 1990, pp. 209–260.
- 461 [6] W.B. Frakes, Stemming algorithms, in: W.B. Frakes, R. Baeza-Yates (Eds.), *Information*  
462 *Retrieval: Data Structures and Algorithms*, Prentice Hall, Englewood Cliffs, NJ, 1992, pp.  
463 131–160.
- 464 [7] V.P. Guerrero-Bote, *Redes Neuronales aplicadas a las Técnicas de Recuperación Documental*,  
465 Ph.D. thesis, University of Granada, Spain, 1998.
- 466 [8] V.P. Guerrero-Bote, F. Moya-Anegón, Reduction of the dimension of a document space using  
467 the fuzzified output of a Kohonen network, *J. Am. Soc. Inform. Sci.* 52 (2001) 1234–1241.
- 468 [9] V.P. Guerrero-Bote, F. Moya-Anegón, V. Herrero-Solana, Document organization using  
469 Kohonen's algorithm, *Inform. Process. Manage.* 38 (2002) 79–89.
- 470 [10] V.P. Guerrero-Bote, M. Reyes-Barragán, F. Moya-Anegón, V. Herrero-Solana, Methods for  
471 the analysis of the uses of scientific information, *The Case of the University of Extremadura*  
472 (1996–7), *Libri*, 52 (2002), pp. 99–109.
- 473 [11] V.P. Guerrero-Bote, F. Moya-Anegón, V. Herrero-Solana, Automatic extraction of relation-  
474 ships between terms by means of Kohonen's algorithm, *Libr. Inform. Sci. Res.* 24 (2002) 235–  
475 250.
- 476 [12] R. Hecht Nielsen, Neurocomputing: picking the human brain, *IEEE Spectrum* 25 (1) (1988)  
477 36–41.
- 478 [13] J.R. Hiler, V.J. Martínez, *Redes neuronales artificiales, fundamentos, modelos y aplicaciones*,  
479 RAMA, Madrid, Spain, 1995.
- 480 [14] S. Kaski, Fast winner search for SOM-based monitoring and retrieval of high-dimensional  
481 data, in: *Proceedings of the Ninth International Conference on Artificial Neural Networks*  
482 (ICANN99), London, 1999, pp. 940–945.
- 483 [15] T. Kohonen, Self-organized formation of topologically correct feature maps, *Biol. Cyber.* 43  
484 (1) (1982) 59–69.
- 485 [16] T. Kohonen, An introduction to neural computing, *Neural Networks* 1 (1) (1988) 3–16.
- 486 [17] T. Kohonen, *Self-Organization and Associative Memory*, Third ed., Springer Verlag, Berlin,  
487 Germany, 1989.
- 488 [18] T. Kohonen, The self-organizing map, *Proc. IEEE* 78 (9) (1990) 1464–1480.
- 489 [19] T. Kohonen, *Self-Organization Maps*, Springer Verlag, Heidelberg, Berlin, Germany, 1995.
- 490 [20] T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, A. Saarela, Self  
491 organization of a massive text document collection, in: E. Oja, S. Kaski (Eds.), *Kohonen*  
492 *Maps*, Elsevier, Amsterdam, Holland, 1999, pp. 171–182.
- 493 [21] H. Kucera, N. Francis, *Computational Analysis of Present-Day American English*, Brown  
494 University Press, Providence, RD, 1967.
- 495 [22] K. Lagus, T. Honkela, S. Kaski, T. Kohonen, WEBSOM for textual data mining, *Artif. Intell.*  
496 *Rev.* 13 (5/6) (1999) 345–364.
- 497 [23] K. Lagus, S. Kaski, Keyword selection method for characterizing text document maps, in:  
498 *Proceedings of the Ninth International Conference on Artificial Neural Networks (ICANN99)*,  
499 London, 1999, pp. 371–376.
- 500 [24] X. Lin, Maps displays for information retrieval, *J. Am. Soc. Inform. Sci.* 48 (1) (1997) 40–54.
- 501 [25] A.J. Maren, C.T. Harston, R.M. Pap, *Handbook of Neural Computing Applications*,  
502 Academic Press, San Diego, 1990.
- 503 [26] F. Moya-Anegón, V. Herrero-Solana, V.P. Guerrero-Bote, Virtual reality interface for  
504 accessing electronic information, *Libr. Inform. Res. News* 22 (71) (1998) 34–39.
- 505 [27] F. Moya-Anegón, P. Moscoso, C. Olmeda, V. Ortiz-Repiso, V. Herrero-Solana, V.P.  
506 Guerrero-Bote, NeuroISOC: un modelo de red neuronal para la representación del  
507 conocimiento, in: M.J. López Huertas, J.C. Fenández Molina (Eds.), *La representación y la*  
508 *organización del conocimiento en sus distintas perspectivas: su influencia en la recuperación de*

- 509 la información, Actas del IV Congreso ISKO-España (EOCONSID'99), ISKO-España,  
510 Granada, 1999, pp. 151–156.
- 511 [28] S.C. Newton, S. Pemmaraju, S. Mitra, Adaptive fuzzy leader clustering of complex data sets  
512 in pattern recognition, *IEEE Trans. Neural Networks* 3 (5) (1992) 794–800.
- 513 [29] T. Noreault, M. McGill, M.B. Koll, A performance evaluation of similarity measures,  
514 document term weighting schemes and representations in a Boolean environment, in: R.N.  
515 Oddy, S.E. Robertson, C.J. Van Rijsbergen, P.W. Williams (Eds.), *Information Retrieval*  
516 *Research: Papers Given at the 1st Joint British Computer Society (BCS) and Association for*  
517 *Computing Machinery (ACM) Symposium: Research and Development in Information*  
518 *Retrieval*, Butterworths, London, England, pp. 57–76.
- 519 [30] M.F. Porter, An algorithm for suffix stripping, *Program* 14 (3) (1980) 130–137.
- 520 [31] P.K. Simpson, Fuzzy min-max neural networks. Part 2: Clustering, *IEEE Trans. Fuzzy Syst.* 1  
521 (1) (1993) 32–45.
- 522 [32] P. Van der Besselaar, L. Leydesdorff, Mapping change in scientific specialties: a scientometric  
523 reconstruction of the development of artificial intelligence, *J. Am. Soc. Inform. Sci.* 47 (6)  
524 (1996) 415–436.