

B. Mirkin  
Birkbeck University of London UK

## Deviant box and dual clusters for the analysis of conceptual contexts

Invited presentation at Fifth International Conference on Concept Lattices and Their Applications (24-26 October 2007, Montpellier France), available at: <http://www.dcs.bbk.ac.uk/~mirkin/papers/bd.pdf>

### 1. The notion of formal concept and box clustering model

A data matrix  $R=(r_{ij})$ , with  $i \in I$  (objects) and  $j \in J$  (attributes), such that either  $r_{ij} = 1$  or  $r_{ij} = 0$  is referred to as a **conceptual context**.

A **formal concept** is a pair of sets  $(V, W)$ , that is, a biset, such that  $V \subseteq I$ ,  $W \subseteq J$  and

$$r_{ij} = 1 \text{ for all } (i, j) \in V \times W \quad (1)$$

and neither  $V$  nor  $W$  can be increased without breaking the property (1). Set  $V$  is referred to as the **extent**, and  $W$  as the **intent** of the concept  $(V, W)$ . The cardinalities will be denoted by  $\#V = n$ ,  $\#W = m$ .

Formal concepts form a lattice over the set-theoretic inclusion of their extents. The literature on formal concepts and their applications in data analysis, knowledge discovery, software engineering is quite numerous: suffice to say that the number of references to a founding book by Ganter and Wille (1999) in Google Scholar is greater than 1100.

Some people feel that the notion of formal concept may be somewhat overly rigid in some fields of study and that the condition of all within-entries being non-zero can be too restrictive, especially, with noisy data. On the other hand, the condition that at least one outside row/column must be zero may seem too permissive, generating too many similar concepts sometimes. There have been attempts at modifying both of these conditions by admitting a few zeros inside and most zeros outside (Pensa and Boulicaut (2005), Rome and Haralick (2005)). I am going to show how the data recovery clustering can be utilized in this direction.

In clustering, analogous structures are referred to as biclusters, the term arguably coined by Mirkin (1996), see also Prelic and all (2006) for an up-to-date review. Here we focus on a special case referred to as box cluster in Mirkin, Arabie, Hubert 1995. A **box cluster** is a pair  $(V, W)$  along with its intensity  $\lambda$ .

A set of box clusters  $(\lambda_t, V_t, W_t)$ ,  $t=1, \dots, T$ , forms a **disjunctive box cluster** (data recovery) model of data  $R$  if

$$r_{ij} = \max_{t=1 \dots T} \lambda_t v_{it} w_{jt} + \lambda_0 + e_{ij} \quad (2)$$

where  $e_{ij}$  are sufficiently small, and  $\lambda_0$ ,  $0 < \lambda_0 < 1$ , plays the role of an intercept in linear data models. (Mirkin, Arabie, Hubert (1995) considered an additive model, not disjunctive one, in which summation stands instead of maximization.)

Obviously, model (2) can be fit with the trivial set of singleton box clusters corresponding to non-zero elements of  $R$  with all  $\lambda_t = 1$  and  $\lambda_0 = 0$ . This suggests that additional conditions are needed to avoid a trivial solution. We consider that our one-by-one extraction approach (Mirkin 1996) leading to most “deviant” clusters can serve as an additional condition.

Assume  $\lambda_0$  is constant and can be specified before the fitting of the model. Then the model can be rewritten as

$$r_{ij} - \lambda_0 = \max_{t=1\dots T} \lambda_t v_{it} w_{jt} + e_{ij} \quad (2')$$

so that  $\lambda_0$  becomes a similarity shift value rather than an intercept. The model on the right applies now to not the original 1/0 data but to the values  $r_{ij}' = r_{ij} - \lambda_0$  shifted by  $\lambda_0$ : what was 0 becomes a negative number:  $r_{ij}$  are not just “conceptual” entries but rather similarity scores.

Suggested “maximum deviance” approach: box clusters  $(\lambda_t, V_t, W_t)$  in (1) should be those most deviant from the “middle”, that is, those found by fitting a single cluster model (with a constant  $\lambda_0$ )

$$r_{ij}' = r_{ij} - \lambda_0 = \lambda v_i w_j + e_{ij} \quad (3)$$

with the least squares criterion. In this formulation,  $v=(v_i)$  and  $w=(w_j)$  are binary membership vectors of  $V$  and  $W$ , respectively, so that  $v_i w_j = 1$  if and only if  $(i,j) \in V \times W$ .

Let us initially assume  $\lambda_0 = 0$  so that  $r_{ij}' = r_{ij}$ . Box cluster  $(\lambda_t, V_t, W_t)$  minimizing the least squares criterion

$$L^2 = \sum_{ij} (r_{ij} - \lambda v_i w_j)^2 \quad (4)$$

over real  $\lambda$  and binary  $v_i, w_j$ , must approximate a pervasive concept: in the ideal case, only entries within the box are not zero while all the rest must be zero. The closer a box cluster to that, the better.

Elementary considerations show that, given sets  $V$  and  $W$ , optimal  $\lambda$  is equal to the within-box average:

$$\lambda = \sum_{i \in V, j \in W} r_{ij} / nm \quad (5)$$

which is the proportion of unities within the box, and, assuming that the  $\lambda$  is optimal, criterion  $L^2$  in (4) admits the following decomposition:

$$L^2 = \sum_{ij} r_{ij}^2 - \lambda^2 nm \quad (6)$$

At  $\lambda_0=0$ , this can be further simplified. Since the data entries are binary, the equation  $r_{ij}^2 = r_{ij}$  holds, and the left item on the right is just the total number of nonzero entries in R, denoted by  $r..$ , so that

$$L^2 = r.. - \lambda^2 nm, \tag{6'}$$

which means that minimizing criterion (4) is achieved by maximizing the product of the box's area  $nm$  and the squared proportion of its unity entries  $\lambda$  in (5):

$$g(V,W) = \lambda^2 nm \tag{7}$$

which is the proportion of the data scatter taken into account by the box. This criterion combines two contrasting criteria: (a) the largest area, (b) the largest proportion of within-box unities.

Only (locally) optimal box clusters should be admitted into the disjunctive model (2).

## 2. Optimal formal concepts

If we restrict ourselves with only those box clusters that correspond to formal concepts only, criterion (7) brings in a simple condition:

Statement 1.

*Only those formal concepts are optimal that have the maximum area of the corresponding rectangle.*

Indeed, criterion (7) expresses the area in this case, because  $\lambda = 1$  for each formal concept.

Given an object  $i$  or attribute  $j$ , let us refer to a formal concept as maximal if it has the maximum area among those containing the element. One could restrict the model (2) to only boxes corresponding to maximal formal concepts.

Consider an example from Siff and Reps (1999):

**Table 1.** A crude context for Mammals.

	A. Four-legged	B. Hair-covered	C. Intelligent	D. Marine	E. Thumbed
1. cat	+	+			
2. dog	+	+			
3. dolphin			+	+	
4. gibbon		+	+		+
5. human			+		+
6. whale			+	+	

The non-trivial formal concepts for this data set are in Table 2.

Concepts III, IV and VI form the minimum set of maximal concepts; they all have the maximum area size 4; they cover all the entities and removing any of them makes the model in (2) incomplete.

**Table 2.** Set of formal concepts for the context in Table 1.

Concept	Extent V	Intent W	Area
I	3, 4, 5, 6	C	4
II	1, 2, 4	B	3
III	4, 5	C, E	4
IV	3, 6	C, D	4
V	4	B, C, E	3
VI	1, 2	A, B	4

The question of finding a minimum covering set of maximal concepts seems rather difficult in the general setting.

### 3. Locally optimal box cluster

As the optimal intensity of a box  $(V, W)$  is fully determined by the summary entries within the box formed, we identify the box cluster with just sets  $V$  and  $W$ , that is, biset  $(V, W)$ . Given a biset  $(V, W)$ , let us define its neighbourhood  $N(V, W)$  as the set of all those box clusters that can be obtained from  $(V, W)$  by adding or removing just one item to/from either  $V$  or  $W$ . Let us take a  $(V', W')$  from  $N(V, W)$  and see the difference,  $\text{Diff}$ , in values of the criterion (7), assuming that  $\lambda$  is defined according to (5). With no loss of generality assume that it is  $V'$  that differs from  $V$ , by adding/removing an entity  $i^*$  :

$$\text{Diff}(i^*) = [z_{i^*} r^2(i^*, W) + 2z_{i^*} r(V, W) r(i^*, W) - r^2(V, W)/n] / (m(n+z)) \quad (8)$$

where  $z_{i^*}$  equals 1 if  $i^*$  is added to  $V$  and -1 if  $i^*$  is removed from  $V$ , and  $r(V, W)$  denotes the sum of all  $R$  entries within box  $V \times W$  and  $r(i^*, W)$  is the sum of all  $r_{i^*j}$  over  $j \in W$ . A symmetric expression holds for  $\text{Diff}(j^*)$  at  $j^* \in W$ .

A local search algorithm can be drawn starting from every entity  $i \in V$  (or  $j \in W$ ):

#### *Algorithm Box (i)*

0. Take  $V = \{i\}$  and  $W = \{j \mid r_{ij} = 1\}$  as the starting box.
1. Find  $\text{Diff}()$  in (8) for all elements of  $I$  and  $J$ , take that  $D^*$  which is maximum.
2. If  $D^*$  is not positive, halt. Otherwise, perform the operation of adding/removing for the corresponding entity and return to Step 1 with the update box.

The resulting cluster box is rather contrast, which can be expressed in terms of the average similarity  $\lambda$  within the cluster and that of an individual entity to it:

Statement 2.

If box cluster  $(V, W)$  is found with  $\text{Box}()$  algorithm then, for any entity outside the box, its average similarity to it is less than the half of the within-box similarity  $\lambda$ ; in contrast, for any entity belonging to the box, its average similarity to it is greater than or equal to the half of the within-box similarity  $\lambda$ .

To prove it, consider, for instance, the case of  $i^*$  outside of the found box, that is,  $i^* \notin V$ . Then  $\text{Diff}(i^*)$  in (8) is negative so that

$$r^2(i^*, W) + 2r(V, W)r(i^*, W) < r^2(V, W)/n$$

Then, even more so,  $2r(V, W)r(i^*, W) < r^2(V, W)/n$ , which implies that  $2r(i^*, W) < r(V, W)/n$  provided that  $r(V, W) > 0$  which is the case if  $\lambda_0 < 1$  as defined in the beginning. Dividing the latest inequality by  $2m$  proves the statement.

Fitting model (2) can be done by applying algorithm  $\text{Box}()$  starting from each of the entities and retaining only different solutions.

The set of different box clusters for our example is in Table 3.

**Table 3.** Set of different box clusters produced by the algorithm  $\text{Box}()$  for the context in Table 1.

Box cluster	Extent V	Intent W	Intensity	Contribution (to 13)
B1	3, 4, 5, 6	C, D, E	0.67	5.33
B2	1, 2, 4	A, B	0.83	4.17
B3	4, 5	B, C, E	0.83	4.17

Each of these box clusters contains zeros thus admitting some leaps over the data. The minimum set for model in (2) would consist just of two box clusters, B1 and B2.

#### 4. Changing the shift

Let us take the shift value in (2) to be equal to the mean value of entries in R:  $\lambda_0=0.43$ . This will affect the results of  $\text{Box}()$  algorithm as presented in Table 4.

**Table 4.** Set of different box clusters produced by  $\text{Box}()$  for the Table 1 shifted by  $\lambda_0=0.43$ . It coincides with the set of formal concepts in Table 2.

Box cluster	Extent V	Intent W	Intensity	Contribution (to data scatter =7.37)
B1	3, 4, 5, 6	C	0.57	1.28
B2	1, 2, 4	B	0.57	0.96
B3	4, 5	C, E	0.57	1.28
B4	3, 6	C, D	0.57	1.28
B5	4	B, C, E	0.57	0.96
B6	1, 2	A, B	0.57	1.28

The set of box clusters coincides with the set of all formal concepts. Their intensity, 0.57, reflects the subtracted  $\lambda_0=0.43$ . Those maximally contributing are expectedly those of maximum area. Why is such a change? Because the condition of  $\text{Diff}()>0$ , for adding/removing entities to/from the box being built, has changed. As proven above, it checks whether the similarity of an item to be added (or, removed) is greater (or, smaller) than the half of within-box similarity  $\lambda$  (5). By subtracting  $\lambda_0=0.43$  from all R entries, the average similarity of the item to the box changes by  $\lambda_0$  whereas the half within-box similarity changes by  $\lambda_0/2$  only, thus becoming relatively higher. That is, the threshold which the proportion of unities in a row or column is compared with in the Statement 2 **changes from  $\lambda/2$  to  $(\lambda+\lambda_0)/2$** ; the greater the  $\lambda_0$ , the greater the rise of the threshold. This implies the following.

Statement 3.

*Provided that the shift value  $\lambda_0$  subtracted from R entries is high enough, algorithm Box() will produce all the formal concepts.*

At intermediate shifts, box lists will differ.

What  $\lambda_0$  value should be taken? In Mirkin et al. (2007) we argue, in the context of clustering, that this should come from unrelated knowledge – in that case, the authors have been given sets of pairs of entities that **should** and those that **should not** belong in the same cluster, which led to a shift value chosen in between.

## 5. Dual clustering problems

There is an idea that good formal concepts should comprise similar objects and similar attributes. This intuition can be expressed mathematically within the box clustering formalism. Consider model (3) in the matrix format,  $R' = \lambda v w^T + E$ , and multiply it by its transposed version,  $R'^T = \lambda w v^T + E^T$ :

$$R'R'^T = \lambda^2 m v v^T + E E^T + 2\lambda E w v^T \quad (9)$$

The elements of this equation have reasonable interpretations.

The elements of the object-to-object matrix  $S=R'R'^T$  on the left side are inner products of R' columns. They can be considered **similarities** between objects and can be expressed as

$$s_{ii'} = N_{ii'} - \lambda_0 N_i - \lambda_0 N_{i'} + \lambda_0^2 \quad (10)$$

where  $N_i$ ,  $N_{i'}$ , and  $N_{ii'}$ , denote, respectively, the numbers of elements in the intent of  $i$ , intent of  $i'$ , and their overlap. This formula can be considered a linearization of the odds-ratio coefficient  $N_{ii'}/N_i N_{i'}$ , well known in marketing and bioinformatics research, in which  $\lambda_0$  serves as the penalty parameter. Also, this measure extends the popular match distance coefficient,  $d = N_i + N_{i'} - 2N_{ii'}$ , which expresses the size of the symmetric difference of the intents.

The elements of matrix  $F=EE^T$  are inner products of columns of the residual matrix E. The idea of minimizing them expresses the idea of minimizing correlations between the residuals

corresponding to individual features, which is a prominent criterion in factor analysis (Harman 1976) and independence analysis (Hyvärinen and Oja 2000).

The matrix  $\lambda^2 m v v^T$  has its elements  $\lambda^2 m$  within a square of  $(i, i') \in V \times V$  for the  $V$  yet to be determined; the unknown  $m$  can be hidden as part of  $\mu = \lambda^2 m$  which is also to be determined. A worry can be caused by the matrix  $E w v^T$  whose elements are averages of the residual rows in  $V$  over columns in  $W$ . One way to deal with it is to discard the matrix from the equation altogether, by imposing an underlying assumption that the residuals sum up to zero within  $V$  for all columns in  $W$ . The other way is to take into account a dual equation to (9), which is obtained by multiplication of the equation (3) on its transpose from the left, so that the decomposition involves column-to-column matrix  $C = R'^T R'$  rather than  $S = R' R'^T$ . We consider these ways in turn.

### 5.1. Approximate cluster and ADDI-S algorithm

After discarding the last item in equation (9), it can be rewritten as

$$s_{ii'} = \mu v_i v_{i'} + f_{ii'} \quad (11)$$

with the left part defined by formula (11) and real  $\mu$  and binary vector  $v$  to be determined by minimizing the least-squares criterion

$$M^2 = \sum_{ii'} (s_{ii'} - \mu v_i v_{i'})^2 \quad (12)$$

Mirkin (1987) treated this problem in two versions: with a pre-specified  $\mu$  (algorithm ADDI), or with an optimized  $\mu$  (algorithm ADDI-S). Only the latter case is of interest here:  $\mu$  is assumed to be optimal, that is, the average of similarities  $s_{ii'}$  over the current set  $V$ . This leads to a derived problem analogous to that considered above. To minimize  $M^2$ , one can equivalently maximize criterion

$$h^2(V) = \mu^2 n^2 \quad (7')$$

which expresses the contribution of cluster  $V$  to the quadratic scatter of similarities in  $S$ . And maximizing (7') can be done by maximizing its square root  $h(V) = \mu n$ , which expresses a compromise between the within-cluster similarity  $\mu$  and its size  $n$ : both should be as great as possible which cannot be done simultaneously. The algorithm ADDI-S is a local search algorithm, similar to Box(), for maximizing the latter criterion. It iteratively finds an  $i$  outside of  $V$  maximizing the change of  $h(V)$  if  $i$  is added to  $V$  and an  $i$  in  $V$  maximizing the change in  $h(V)$  if  $i$  is removed from  $V$ , and it takes the  $i$  giving the larger value. The iterations stop when this larger value is negative. The algorithm returns the resulting  $V$  and its contribution to the data scatter,  $h^2(V)$ . The change in  $h(V)$  is proven to coincide with the average similarity between  $i$  and  $V$  minus  $\mu/2$ . A similar algorithm utilizing an arbitrary threshold instead of  $\mu/2$  was proposed by Ben-Dor et al. (1999) under the title CAST and became popular in bioinformatics.

A property of the resulting cluster  $V$ , similar to that for the box cluster case, holds: *the average similarity between  $i$  and  $V$  is at least half the within-cluster average similarity if  $i \in V$ , and at most that value if  $i \notin V$ .*

The algorithm ADDI-S can be applied to clustering at any similarity data. Consider an example of applying it to the problem of representing a research organization over an ontology of the scientific field in question (Mirkin, Nascimento, Pereira 2007).

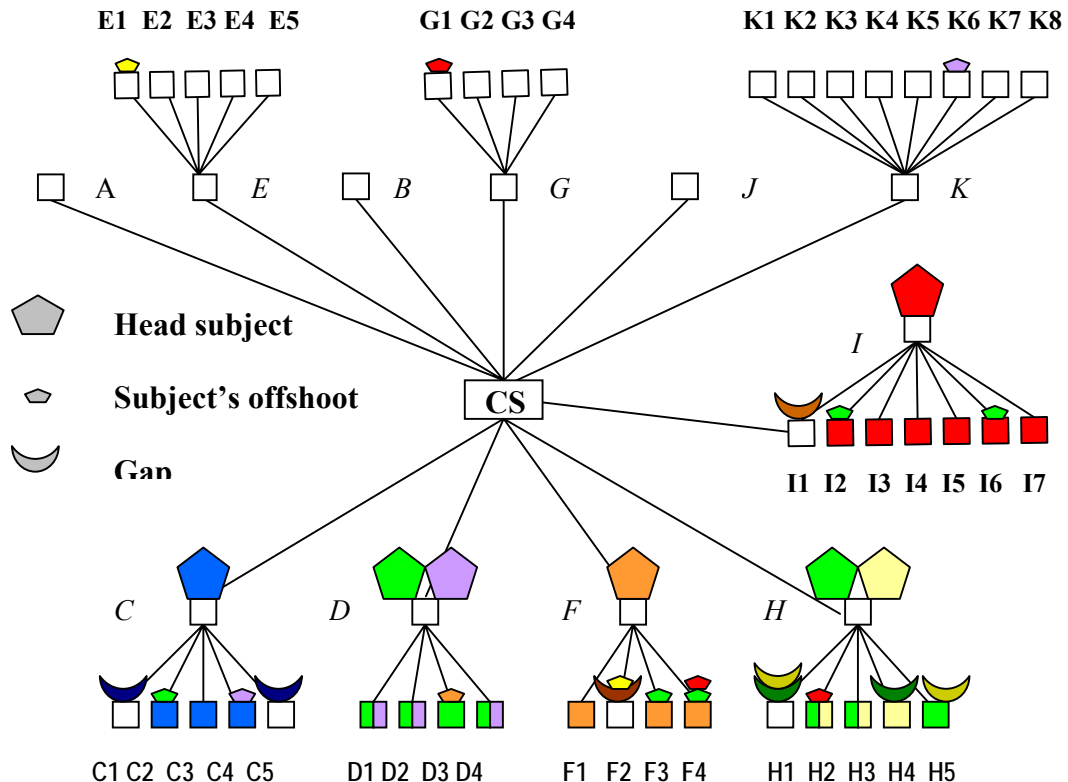
### 5.2. ADDI-S clusters of research topics

Consider

- (a) ACMC – Association for Computing Machinery Classification of Computing Subjects (a hierarchy);
- (b) Sets of ACMC topics being worked on by each member of an academic department;
- (c) Similarity measure  $S$  between ACMC topics based on (b);
- (d) ACMC topic clusters found with ADDI-S;
- (e) Parsimoniously lifting clusters within the ACMC.

These items constitute our method for representing a Computer Science research department over ACMC.

The total of (weighted) elements of the representation, such as Head subjects, gaps, and offshoots, constitute the minimized penalty. Parsimoniously lifting a cluster within a hierarchy can be done in a recursive manner involving two different assumptions of the “Parental behavior” (Mirkin et al. 2003).



**Figure 1.** Representing the major groupings of ACMC topics in the DI FCT UNL: 6 subject clusters have their head subjects shown using differently coloured pentagons. Subject items sharing two different head subjects are split coloured (viz. D1 or H2). The head subjects are:  
■ C. Computer Systems Organization    ■ D. Software & H. Information Systems    ■ F. Theory of Computation  
■ D. Software    ■ H. Information Systems    ■ I. Computing Methodologies

One can see that the clusters fit well within ACMC subjects, except for a “green” cluster that covers two head subjects, D. Software and H. Information Systems. This relates, in our view, to the emergence of Software Engineering as a major Computer Science activity, which is not recognized in the structure of ACMC yet.

### 5.3. Dual clusters of concept extents and intents

Consider dual forms of equation (9) both for similarities between objects and between attributes

$$R'R'^T = \lambda^2 m v v^T + E E^T + 2\lambda E w v^T, \quad R'^T R' = \lambda^2 n w w^T + E^T E + 2\lambda E^T v w^T \quad (9')$$

One can proceed from this with the following alternating process for finding binary  $v$ ,  $w$  and  $\lambda$  by minimizing each  $E E^T$  and  $E^T E$  in turn:

*Algorithm DUAL(obj, att)*

1. Given an object and attribute, find  $V(\text{obj})$  at  $R'R'^T$  and  $W(\text{att})$  at  $R'^T R'$  with ADDI-S.
2. Given  $V$  and  $W$ , find  $\lambda$  in the interval  $[0,1]$  minimizing the squared sum of all elements of  $E^T E$  and  $E E^T$  in the equations (9'). I do this with an evolutionary algorithm.
3. Given  $W$  and  $\lambda$ , find  $V$  minimizing the squared sum of elements of  $E E^T$  in the equation on the left of (9').
4. Given  $V$  and  $\lambda$ , find  $W$  minimizing the squared sum of elements of  $E^T E$  in the equation on the right of (9').
5. Check for the convergence and go to Step 1 if negative.
6. Output found  $V$ ,  $W$  and  $\lambda$  along with the value of the squared error.

**Table 5.** Sets of the best dual clusters produced by the algorithm Dual() for the Table 1 with and without scale shifting (A1 and B1-B3, respectively).

Box cluster	Scale shift	Extent V	Intent W	Intensity	Summary error
A1	None	1-6 (all)	A-E (all)	0.46	48.4
B1	Mean	1, 2	A, B	0.72	25.8
B2		3, 6	C, D	0.69	30.3
B3		4, 5	C, E	0.57	37.2

With the mean subtracted from the context  $R$  as  $\lambda_0$ , Dual() produces three best dual clusters that are by far better than the others. These are formal concepts. Are they any better than the others?

## 6. Experiments

To check how well the data recovery boxes and biclusters recover the data structure, an experiment by Pensa and Boulicaut (2005) has been extended as follows. A binary  $30 \times 15$  data

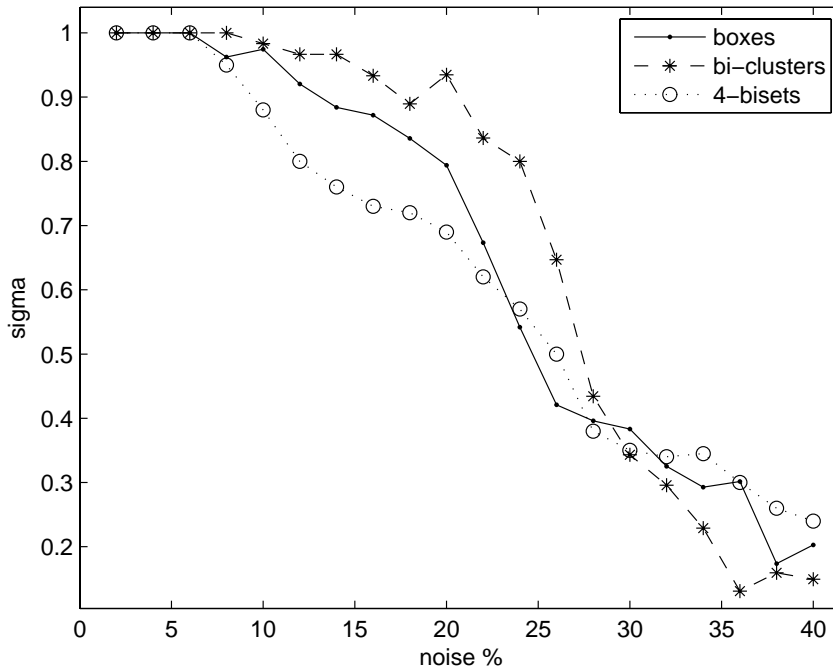
table R0 comprising three non-overlapping formal concepts has been considered. All R0's entries are zeros except for those within three boxes comprising, in respect, first 10 rows (from 1 to 10) and first 5 columns (from 1 to 5), second 10 rows (from 11 to 20) and second 5 columns (from 6 to 10), and third 10 rows (from 21 to 30) and third 5 columns (from 11 to 15), whose all entries are ones. Then this matrix is changed to a matrix Rp by randomly changing its every entry with the probability p%, p=1, 2, . . . , 40. Algorithms Box(.) and Dual(., .) have been applied to each Rp, with its mean subtracted as  $\lambda_0$ , at each  $i \in I$  and  $j \in J$ , and sets of differing results stored as Bp and Dp. To compare these results with the original concepts in R0, we utilised the extension of Jaccard coefficient described in Pensa and Boulicaut (2005).

Specifically, given two bisets, (V,W) and (V',W'), we find their intersection,  $(V \cap V', W \cap W')$ , and union,  $(V \cup V', W \cup W')$ . The ratio of the areas of the corresponding rectangles is taken as the measure of similarity between the bisets:

$$S((V,W), (V',W')) = \frac{|V \cap V'| |W \cap W'|}{(|V \cup V'| |W \cup W'|)}.$$

Then the similarity between two sets of bisets,  $B = \{(V_i, W_i)\}$  and  $B' = \{(V'_j, W'_j)\}$  is defined as the average similarity between a biset in B and its best match in B':

$$\sigma(B, B') = \frac{\sum_i \max_j S((V_i, W_i), (V'_j, W'_j))}{|B|} \quad (10)$$



**Figure 1.** Graphs of  $\sigma$  measure between the original three concepts and results of Box and Dual algorithms applied to the binary Rp matrix at different levels of random noise, p=1, 2, . . . , 40. The third graph represents the  $\sigma$  values at 4-bisets by Pensa and Boulicaut (2005).

The averaged results of runs of Box and Dual algorithms through several rounds of generated matrices  $R_p$  ( $p=1, 2, \dots, 40$ ) are summarised in Figure 1. We can see that Dual indeed slightly outperforms Box till the level of noise up to 25-30%. Moreover, the numbers of biclusters produced by Dual are much smaller up to the same order of noise being on the level of 9-10 whereas Box shoots to its highest levels of 40-45 from about 15% noise. However, the computation time of Dual, in its current version, is about 50-100 times greater than that for Box which takes just about 25-30 seconds on MatLab 7.1.0 in Fujitsu-Siemens Lifebook 1.2 Mh.

For comparison, sigma values supplied by Pensa and Boulicaut (2005) in the same experiment for their so-called delta-bisets, those admitting not more than delta zeros in every within biset column, are provided on the same figure at  $\delta=4$ , arguably their best performer. One can see that at the noise of 10-20% both Box and Dual outperform the 4-bisets, though at the noise of 30% and greater 4-bisets come slightly closer to the original concepts. These levels of sigma-similarity hardly matter. It should be mentioned that the average numbers of zeros in columns of boxes found with Box are very small, hardly greater than 1.

## 7. Conclusion

Data recovery biclustering and dual clustering approach:

- Suggests viable ways of modelling formal context data,
- Is attractive mathematically,
- Leads to effective computational procedures,
- Should be further explored as a supplement to formal concepts, including network analysis and associations.

## References

- A. Ben-Dor, R. Shamir, Z. Yakhini (1999) Clustering gene expression patterns, *Journal of Computational Biology*, 6, 281-297.
- B. Ganter and R. Wille (1999) *Formal Concept Analysis: Mathematical Foundations*, Springer.
- H. H. Harman (1976) *Modern Factor Analysis*, University of Chicago Press.
- A. Hyvärinen and E. Oja (2000) Independent component analysis: Algorithms and applications, *Neural Networks*, 13 (4-5), 411-430.
- B. Mirkin (1987) Additive clustering and qualitative factor analysis methods for similarity matrices, *Journal of Classification*, 4, 7-31; Erratum (1989), 6, 271-272.
- B. Mirkin, P. Arabie and L. Hubert (1995) Additive two-mode clustering: the error-variance approach revisited, *Journal of Classification*, 12, 243-263.
- B. Mirkin, T. Fenner, M. Galperin and E. Koonin (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes, *BMC Evolutionary Biology*, 2003, 3:2.

B. Mirkin, L.M. Pereira, S. Nascimento (2007) ACM Classification Can Be Used for Representing Research Organizations, *DIMACS Technical Report 2007-13*, 20 p.

R.G. Pensa and J.-F. Boulicaut (2005) Towards fault-tolerant formal concept analysis, in S. Bandini and S. Manzoni (Eds.) *AI\*IA 2005, LNAI 3673*, 212-223.

A. Prelić, S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L. Thiele and E. Zitzler (2006) A systematic comparison and evaluation of biclustering methods for gene expression data, *Bioinformatics*, 22(9), 1122-1129.

J.E. Rome and R.M. Haralick (2005) Towards a formal concept analysis approach to exploring communities on the World Wide Web, International Conference on Formal Concept Analysis, Lens, France.

M. Stiff and T. Reps (1999) Identifying modules via concept analysis, *IEEE Transactions on Software Engineering*, 25, no. 6, 749-768.