Approximate Bicluster and Tricluster Boxes in the Analysis of Binary Data

Boris G. Mirkin\textsuperscript{1,2,*} and Andrey V. Kramarenko\textsuperscript{1}

\textsuperscript{1} National Research University – Higher School of Economics, Moscow, Russia
\textsuperscript{2} Department of Computer Science and Information Systems, Birkbeck University of London, UK
bmirkin@hse.ru, mirkin@dcs.bbk.ac.uk

Abstract. A disjunctive model of box bicluster and tricluster analysis is considered. A least-squares locally-optimal one cluster method is proposed, oriented towards the analysis of binary data. The method involves a parameter, the scale shift, and is proven to lead to “contrast” box bi- and tri-clusters. An experimental study of the method is reported.

Keywords: box, bicluster, tricluster.

1 Introduction

The concept of bicluster emerges when the data relate two different sets of objects to each other so that highly related pairs of subsets, partitions or even hierarchies can be distinguished in each of the sets. This cluster structure was first made explicit by J.Hartigan\textsuperscript{2,3} and dubbed as biclustering by B.Mirkin\textsuperscript{7}. The concept and corresponding methods gained popularity in several applied areas of which probably the most effective is bioinformatics (see, for example, Madeira and Oliveira\textsuperscript{5}, Prelic et al.\textsuperscript{10}). A somewhat more conservative and mathematically driven approach to establishing relations between a set of objects and a set of attributes was taken in developing the abstract formal model concept (Wille and Ganter\textsuperscript{1}). The notion of a formal concept was first developed for binary data matrix \( R = (r_{ij}), i \in I, j \in J \), where all \( r_{ij} \) are either 1 or 0, which is the case we consider here. A formal concept \( (V, W) \), where \( V \subseteq I, W \subseteq J \), corresponds to all \( r_{ij} = 1 \) for \( i \in V, j \in W \) in such a way that adding elements to either \( V \) or \( W \) would break the equation at least at one pair \((i, j)\). This notion is well justified in application to well developed “contexts” \( R \), but seems somewhat rigid when applied to real world datasets. This is why researchers have been trying to relax the notion of formal concepts by admitting some zeros inside the box \((V, W)\) and some unities outside it (see, for example, Pensa and Boulicaut\textsuperscript{9}). In this respect, the box clustering approach proposed by Mirkin et al.\textsuperscript{6} seems another form of relaxation of the notion of formal concept. Yet the box clustering algorithms proposed in Mirkin et al.\textsuperscript{6} are not applicable to the binary data. Therefore, we put a threefold goal for this paper:

* Corresponding author

© Springer-Verlag Berlin Heidelberg 2011
(1) To propose and explore a model and algorithm for biclustering boxes suitable for binary contexts;
(2) Extend it to triclustering of binary data involving three interrelated data objects;
(3) Apply both bi- and tri-clustering algorithms to real world datasets.

2 The Notions of Formal Concept and Box Clustering

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
\[
r_{ij} = 1 \text{ for all } (i, j) \in V \times W
\]
and neither \( V \) nor \( W \) can be increased without breaking the property (1). The cardinalities will be denoted by \( \#V = n \), \( \#W = m \).

The condition of all within-entries being non-zero can be too restrictive, especially with noisy data. There have been attempts at modifying both of these conditions by admitting a few zeros inside and most zeros outside (Pensa and Boulicaut, Rome and Haralick, Ignatov and Kuznetsov). The data recovery clustering can be utilized to address this as well.

A set of box clusters \( (\lambda_t, V_t, W_t) \), \( t = 1, \ldots, T \), forms a disjunctive box cluster model of data \( R \) if
\[
r_{ij} = \max_{t=1,\ldots,T} \lambda_t v_i w_j + \lambda_0 + e_{ij}
\]
where \( e_{ij} \) are sufficiently small, and \( \lambda_0, 0 < \lambda_0 < 1 \), plays the role of an intercept in linear data models. This model differs from those of additive bi-clustering since (2) involves the operation of maximization rather than summation. To fit (2) with a relatively small number of boxes, assume \( \lambda_0 \) to be constant and specified before the fitting of the model. Then the model in (2) can be rewritten by putting \( r'_{ij} = r_{ij} - \lambda_0 \) on the left, so that \( \lambda_0 \) becomes a similarity shift value rather than an intercept.

We apply here the one-by-one fitting strategy (Mirkin) so that each box cluster \( (\lambda_t, V_t, W_t) \) in (1) is found as a most deviant from the ”middle”, that is, minimizing the residuals in a single cluster model (with a constant \( \lambda_0 \))
\[
r'_{ij} = r_{ij} - \lambda_0 = \lambda v_i w_j + e_{ij}
\]
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 \( (i, j) \in V \times W \) like it is in a formal concept.

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
\[
L^2 = \sum_{ij} (r'_{ij} - \lambda v_i w_j)^2
\]