Boolean reasoning in biclustering

Wnioskowanie boolowskie w zagadnieniu biklasteryzacji

Marcin Michalak, PhD.

Institute of Informatics, Silesian University of Technology

Special Interest Group of Machine Learning Seminary
AGH, Cracow, 26.04.2018

Plan of the presentation

Conjunction and Disjunction Normal Form of the boolean formula:

- CNF: the conjunction of alternatives of literals:

$$f(a, b, c) = (a \vee b) \wedge (b \vee c) \wedge (a \vee b \vee c)$$

- DNF: the alternative of conjunctions of literals:

$$f(a, b, c) = b \vee a \wedge c \quad \text{(from the formula above)}$$

Implicants, prime implicants

Let f is a boolean function of n variables:

$$f_n(a_1, a_2, \ldots, a_n)$$

The expression $I_f(a_I, a_{II}, \ldots, a_X)$, where $\{a_I, a_{II}, \ldots, a_X\} \subseteq \{a_1, a_2, \ldots, a_n\}$, is called the implicant of the function f iff from the fact that it has the value 1 it is implied that value of the fuction f is also 1:

$$I_f(a_I, a_{II}, \ldots, a_X) = 1 \Rightarrow f_n(a_1, a_2, \ldots, a_n) = 1$$

The implicant is the prime one when none of its literals can be removed without violating the implicant conditions.

Plan of the presentation

Boolean reasoning

The data analysis technique, where the initial problem is represented as the boolean formula, which is analyzed (transformed) and the results are interpreted as the solutions of the original problem.

Boolean reasoning — the most popular variant

The problem is coded as the boolean function in CNF (Conjunctive Normal Form) and its DNF (Disjunctive Normal Form) is requested. Implicants or prime implicants are then interpreted as solutions.

Typical applications:

- finding reducts in information systems,
- finding bireducts in information systems,
- and finally... finding biclusters ☺.

Plan of the presentation

clustering

(grouping, unsupervised classification, cluster analysis), a process of partitioning (in the term of the mathematical definition) of the set of objects due to their features values

biclustering

(co-clustering, two-mode clustering, two-dimensional clustering) a proces of finding submatrices of the given one, which elements are more similar to themselves (or even equal) than to other elements in the matrix

	f1	f2	f3	f4
o1	1	2	3	4
o2	2	3	4	5
o3	10	11	12	13
o4	11	12	13	14
o5	101	102	103	104

	f1	f2	f3	f4
o1	1	3	88	91
o2	1	4	87	95
o3	2	4	90	93
o4	2	5	91	94
o5	3	5	90	94

	f_1	f_2	f_3	f_4
o_1	1	1	2	2
o_2	1	1	8	19
o_3	1	32	31	32
o_4	12	18	32	34
o_5	20	27	35	36

	f_1	f_2	f_3	f_4
o_1	**1**	**1**	2	2
o_2	**1**	**1**	8	19
o_3	1	32	31	32
o_4	12	18	32	34
o_5	20	27	35	36

	f_1	f_2	f_3	f_4
o_1	**1**	**1**	2	2
o_2	**1**	**1**	8	19
o_3	1	32	31	32
o_4	12	18	32	34
o_5	20	27	35	36

Figure: Comparison of clustering and biclustering

similarities:

- finding similarities in the data,
- the input is the two-dimensional array,

dissimilarities

- the data are heterogeneous in clustering while they are homogeneous in biclustering,
- clustering introduces the partitioning while bislustering does not have to,
- the element of clustering result is the set of multidimensional objects while the element of the biclustering result is the ordered pair of subset of rows and subset of columns — the bicluster,

Fields of application:

- bioinformatic data,
- text mining,
- devices monitoring and diagnosing.

Conclusion (of biclustering)

Generally, the biclustering is applicable in the research where the data are homogeneous (are a matrix scalars) and it is possible to interprete the subset of rows and subset of columns simultaneously.

Plan of the presentation

A sample discrete matrix M:

	a	b	c
1	1	0	2
2	1	1	0
3	1	1	1

Let B be a set of row indices and X be a set of column indices.

The goal

To find the maximal (in the sense of inclusion — s.o.i.) exact biclusters BX (subsets of rows and columns which intersection provides cells with the same value).

Discernibility function definition

The discernibility function for biclusters in the discrete matrix M is the boolean function defined as follows:

$$f_M = \bigwedge (a \vee b \vee x) \wedge \bigwedge (c \vee y \vee z)$$

where $a, b, c \in B \quad x, y, z \in X$ such that:

$$\forall_{a,b,c,x,y,z} \ (a(x) \neq b(x) \wedge a \neq b) \vee (c(y) \neq c(z) \wedge y \neq z)$$

Boolean reasoning

Each prime implicant of the formula f_M codes a s.o.i. maximal exact bicluster in the data and vice verse.

Interpretation

Bicluster corresponding to the prime implicant is an ordered pairs of set of objects not represented by literals in the prime implicant (the complement set) and set of columns also not represented by literals in the prime implicant.

If BX is a bicluster then its complement in the terms of literals domain (the prime implicant) will be denoted as $B'X'$.

Example

(1, 2, 3 are labels not values)

	a	b	c
1	1	0	2
2	1	1	0
3	1	1	1

$$f_M = (1+a+b)(1+a+c)(1+b+c)(2+a+c)(2+b+c)(b+1+2)(b+1+3)(c+1+2)(c+1+3)(c+2+3)$$

$$f = 12 + 1c + bc + 13ab + 23ac + 23ab$$

Biclusters

$f = 12 + 1c + bc + 13ab + 23ac + 23ab$

	a	b	c
1	1	0	2
2	1	1	0
3	1	1	1

$bc : (\{1, 2, 3\}, \{a\})$

	a	b	c
1	1	0	2
2	1	1	0
3	1	1	1

$1c : (\{2, 3\}, \{a, b\})$

	a	b	c
1	1	0	2
2	1	1	0
3	1	1	1

$13ab : (\{2\}, \{c\})$

	a	b	c
1	1	0	2
2	1	1	0
3	1	1	1

$12 : (\{3\}, \{a, b, c\})$

	a	b	c
1	1	0	2
2	1	1	0
3	1	1	1

$23ac : (\{1\}, \{b\})$

	a	b	c
1	1	0	2
2	1	1	0
3	1	1	1

$23ab : (\{1\}, \{c\})$

Correctness of the approach

It is proved that for each implicant of the boolean function there exists an exact bicluster in the data. It is also proved that for each exact bicluster in the data there exists an implicant in the boolean formula.

Maximality of the approach

It is proved that for each **prime** implicant there exists s.o.i. maximal exact bicluster and vice verse.

Plan of the presentation

Binary matrix *Mb*

	a	b	c
1	1	0	1
2	1	0	0
3	1	1	1

Previous scheme coding:

$$f_{Mb} = (1+2+c)(1+3+b)(2+3+b)(2+3+c)(a+b+1)$$

$$(a+b+2)(a+c+2)(b+c+1)$$

$$f_{Mb} = 12 + 2b + bc + 13a + 3ac$$

Now, let us consider a different way of coding the differences in the matrix. Let $\square \in \{0, 1\}$. The function coding all cells of $\square$ in the matrix is defined as follows:

$$f_{Mb(\square)} = \bigwedge (a \vee x), \ a \in B, x \in X, a(x) = \square$$

The example

	a	b	c
1	1	0	1
2	1	0	0
3	1	1	1

$$f_{Mb(0)} = (1+b)(2+b)(2+c)$$

$$f_{Mb(0)} = 12 + 2b + bc$$

$$f_{Mb(1)} = (1+a)(1+c)(2+a)(3+a)(3+b)(3+c)$$

$$f_{Mb(1)} = 13a + 123 + 3ac + abc$$

Interpretation of prime implicants of $f_{Mb(0)}$ and $f_{Mb(1)}$

Prime implicants of $f_{Mb(0)}$ codes s.o.i. maximal exact biclusters of ones while prime implicants of $f_{Mb(1)}$ codes s.o.i. maximal exact biclusters of zeros.

The correctness and maximality of this approach is proved.

Plan of the presentation

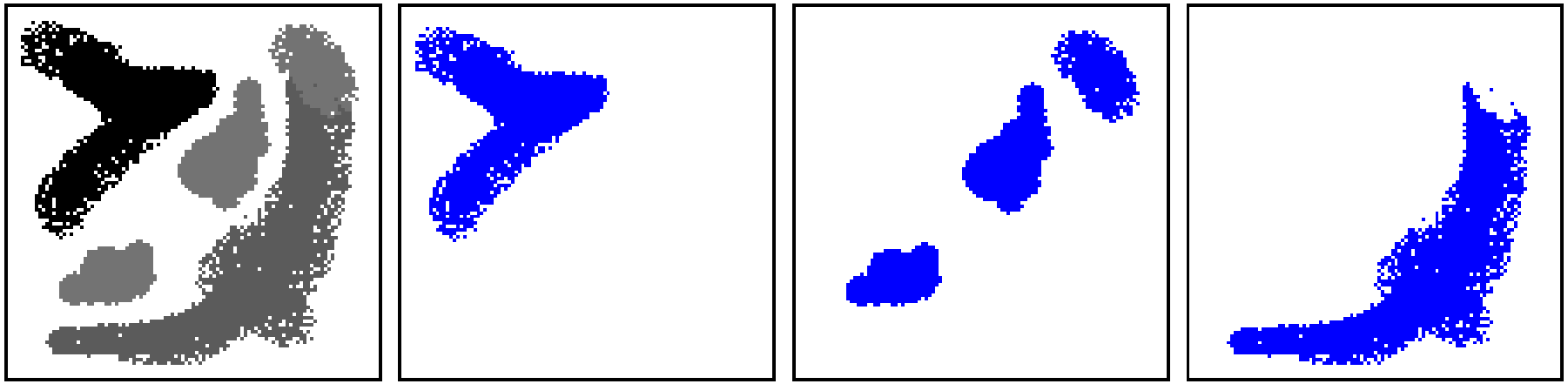

Figure: Original data 100x100 (left) and thre binarized matrices: # 0 (left center), # 77 (right center) and #237 (right).

Table: Biclustering results

value	number of		
	ones	clauses	biclusters
#0	1 415	2 256	5 463 (+2 empty)
#77	1 327	4 560	503 (+2 empty)
#237	2 148	5 267	30 194 (+2 empty)

Plan of the presentation

Conclusions

1. coherent, mathematically proved theory of boolean representation of biclustering problem:
   - exact biclustering of discrete data,
   - exact biclustering of binary data,
   - tolerance biclustering of continuous data (bicluster is defined as the s.o.i. maximal set of cells such that maximal difference between any two different of them **does not exceed** the assumed level T),
   - finding biclusters of chaos in continuous data (as above, but the difference between any two different of them **is at least on** the assumed level T),
   - exhaustive way of finding all possible ans sensible tolerance biclusters and biclusters of chaos (with the one formula);

Conclusions (cont.)

1. The proved quality, correctness and adequacy of raw biclustering results, they may became the basis for the further limitation and generalization.
2. Bringing the exact biclustering into the domain of prime implicant search gives the opportunity of application of well known strategies and heuristics of prime implicant approximation into the problem of biclustering.
3. Also the sequential covering strategies may be applied to limit the set of obtained bicluster to the minimal set of ones covering all the considered cells.

Further Works (but not limited to)

- development of heuristics of fast bicluster induction, covering all the binary data (such as Johnson's strategy),
- new bicluster definitions and corresponding CNFs definitions for new aspects of biclustering,
- application of these methods in biomedical data analysis,
- building the cases of biclustering application in the domain of machine diagnosis,
- software development and publishing,
- publications, publications, publications ☺.

Publications:

IF

- Michalak M., Ślęzak D.: Boolean Representation for Exact Biclustering, Fundamenta Informaticae 160:1–22, 2018
- Michalak M., Ślęzak D.: Boolean Representation for Continuous Biclustering, (finishing)

Web of Knowledge conferences

- Michalak M., Jaksik R., Ślęzak D.: Herustic Search of Exact Biclusters (pre-production ☺)

Thank You.