J. Outrata: Impact of Boolean factorization on classification – improvement of classification

Department of Computer Science, Palacky University, Olomouc (17. listopadu 12, CZ–77146 Olomouc, Czech Republic)

Introduction

:: classification of Boolean data: objects described by Boolean (binary, yes-no) input attributes and assigned a (many-valued) class label

:: preprocessing of data to improve the quality of classification – wanted for Boolean data with a clear semantics (real-valued methods distort meaning of data)

:: in our previous papers: preprocessing based on Boolean matrix factorization (BMF) – replacing input attributes by factors computed from the attributes = feature extraction

:: factors = (some) formal concepts (from Formal Concept Analysis) of input data – clear meaning, easy interpretation

:: classification using a smaller number of factors as new attributes, yet with improvement of the quality

:: modification of the basic method which further improves the quality of classification

Boolean Matrix Factorization (BMF)

= decomposition of Boolean matrix I to Boolean matrices A and B s.t. I is (approximately) equal to $A \circ B$, $(A \circ B)_{ij} = \bigvee_{l=1}^{k} A_{il} \cdot B_{lj}$.

$$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \circ \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

:: I . . . object-attribute Boolean data, A . . . object-factor (usage) matrix, B . . . factor-attribute (basis vector) matrix

~ discovery of k factors (as new Boolean attributes) that (approximately) explain the data – factor l represented by lth column of A and lth row of B:

$A_{il} = 1$ . . . factor l applies to object i
$B_{lj} = 1$ . . . attribute j is a manifestation of factor l

:: $(A \circ B)_{ij}$ . . . "object i has attribute j iff there is a factor l such that l applies to i and j is a manifestation of l"

Problem 1, discrete basis problem (Miettinen et al., 2008):
For I and positive integer k, find A and B minimizing
$$||I - A \circ B|| = \Sigma_{i,j} |I_{ij} - (A \circ B)_{ij}|.$$
. . . the importance of the first k (most important) factors

Problem 2 (Belohlavek, Vychodil, 2010):
For I and positive integer ε, find A and B with k as small as possible such that $||I - A \circ B|| \leq \varepsilon$.
. . . the need to account for (and thus to explain) a prescribed portion of data

:: NP-hard optimization problems ⇒ approximation algorithms

BMF in preprocessing

Boolean data for classification $\langle X, Y, I, c\rangle$:
$X = \{1, \ldots, n\}$ . . . objects
$Y = \{1, \ldots, m\}$ . . . input Boolean attributes
$I \subseteq X \times Y$ . . . object-attribute relational Boolean data
$c : X \to C$ . . . class attribute, mapping assigning to each object $i \in X$ its class label $c(i) \in C$

Preprocessing

:: (approximately) decompose I to A and B, either as Problem 1 or Problem 2

:: $\langle X, F, A, c\rangle$ . . . new Boolean data for classification, $F = \{1, \ldots, k\}$ . . . factors = new Boolean attributes

Classification

:: use any classification model (e.g. a decision tree) developed for $\langle X, F, A, c\rangle$ to classify objects described by the original Boolean attributes from Y:

for object represented by vectors P and Q, apply the model to $Q = \mathbf{g}(\mathbf{P})$ using transformations $g : \{0,1\}^m \to \{0,1\}^k$ and $h : \{0,1\}^k \to \{0,1\}^m$ between the space of original attributes and the space of factors

$$(\mathbf{g}(\mathbf{P}))_\mathbf{l} = \bigwedge_{\mathbf{j}=1}^{\mathbf{m}} (\mathbf{B_{lj}} \to \mathbf{P_j}) \quad \text{and} \quad (h(Q))_j = \bigvee_{l=1}^{k} (Q_l \cdot B_{lj})$$

:: usually $k < m$ (⇒ reduction of dimensionality of data) and g is not an injective mapping

→ assigning to objects in $\langle X, F, A, c\rangle$ with equal $g(P)$ representations transformed from the objects in $\langle X, Y, I, c\rangle$ with different P representations and different assigned class labels the majority class label of those class labels

Improvement of classification quality

Outrata J.: Preprocessing input data for machine learning by FCA. Proc. CLA 2010, pp. 187–198.
Outrata J.: Boolean factor analysis for data preprocessing in machine learning. Proc. ICMLA 2010, pp. 899–902.

:: modified version of algorithm GreConD (Belohlavek, Vychodil) – columns of A and rows of B are objects and attributes of formal concepts of I greedily computed on demand to maximize the drop of $||I - A \circ B||$

:: employing entropy of class labels assigned to objects in the selection of factors – minimize

$$|A| \cdot \frac{E(\mathrm{class}|A)}{-\log_2 \frac{1}{|V(\mathrm{class}|A)|}} + |X \setminus A| \cdot \frac{E(\mathrm{class}|X \setminus A)}{-\log_2 \frac{1}{|V(\mathrm{class}|X \setminus A)|}}$$

$V(\mathrm{class}|A)$ . . . class labels assigned to objects A
$E(\mathrm{class}|A)$ . . . entropy of class labels of objects A, i.e.
$$E(\mathrm{class}|A) = -\Sigma_{l \in V(\mathrm{class}|A)}\, p(l|A) \cdot \log_2 p(l|A)$$

Experimental Evaluation

:: evaluating and comparing classification accuracy of models created by machine learning methods: decision trees ID.3 and C4.5, Nearest Neighbor (NN) and Naive Bayes (NB)

:: datasets selected from UCI ML Repository:

dataset	attrs (binary)	objects	class distribution
breast-cancer	9 (51)	277	196/81
car	6 (21)	1728	1210/384/69/65
kr-vs-kp	36 (74)	1598	835/763
mushroom	22 (125)	564	362/202
solar-flare 2	12 (45)	1066	147/211/239/95/43/331
tic-tac-toe	9 (27)	958	626/332
vote	16 (32)	232	124/108
zoo	15 (30)	101	41/20/5/13/4/8/10

:: binary attributes were obtained by plain nominal scaling

:: 10-fold stratified cross-validation test

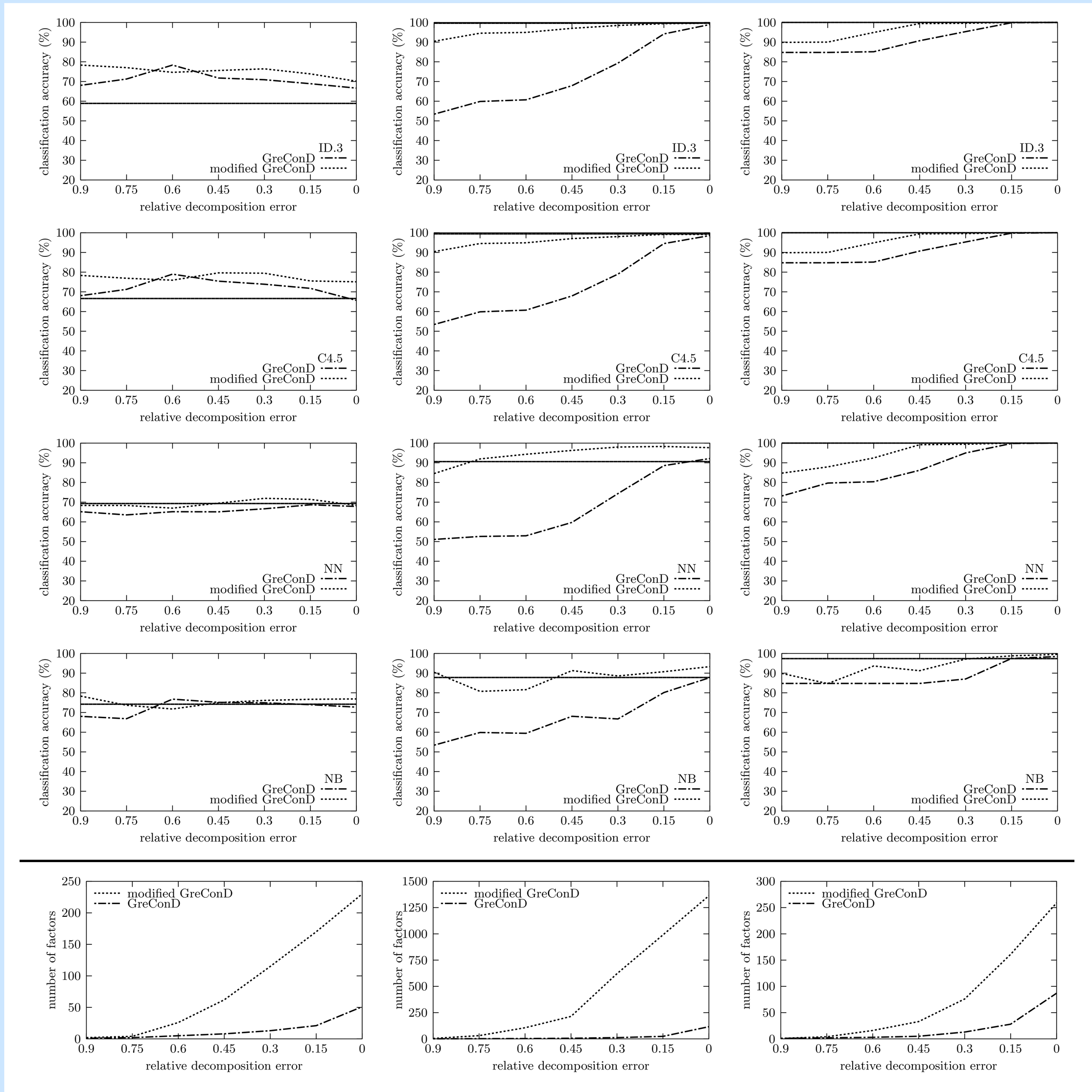
Results

:: for testing parts of the datasets

:: __ __ GreConD, ... modified GreConD, ____ original data

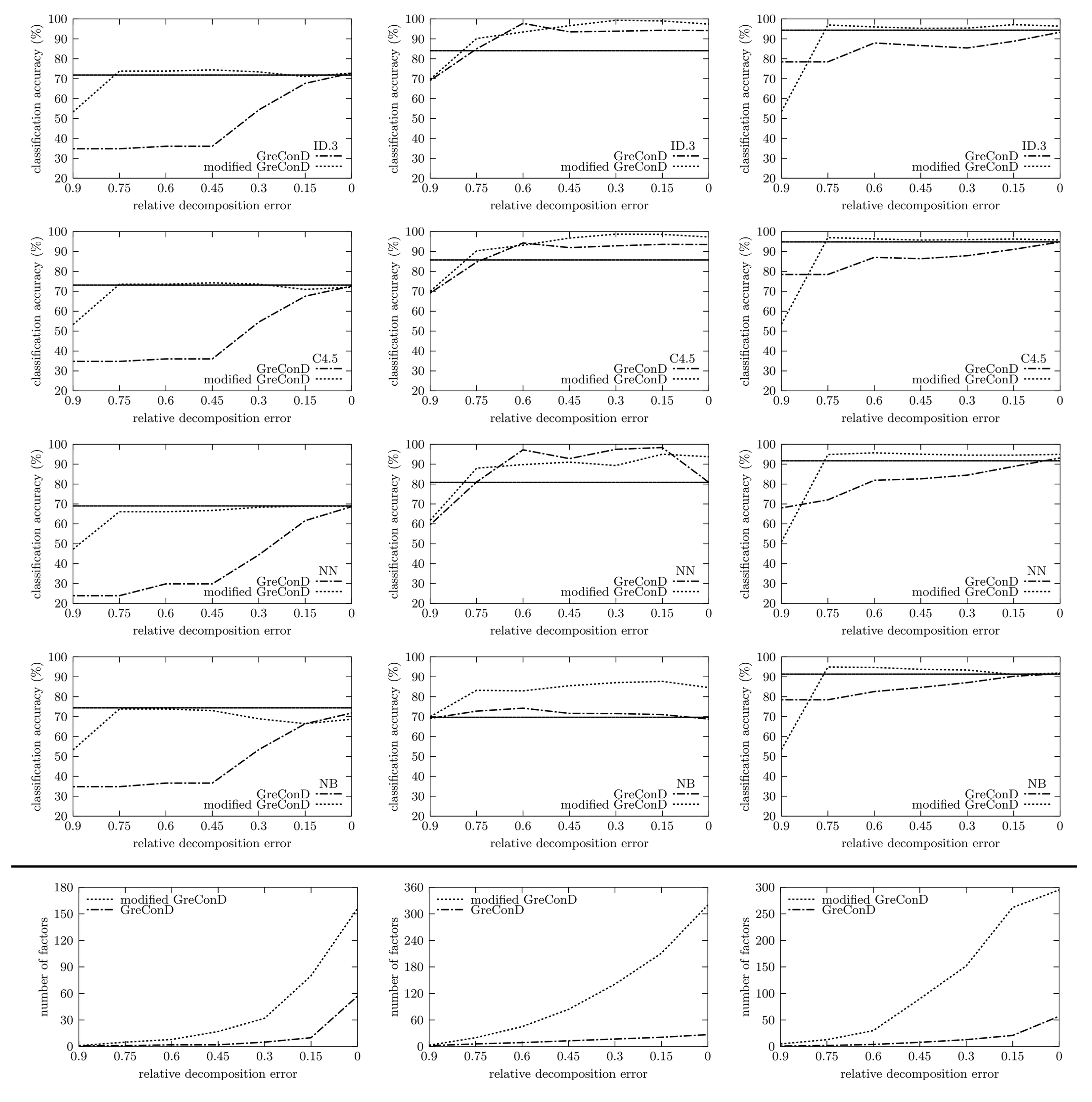
:: x-axis . . . relative decomposition error = ratio of 1s of input data left uncovered in factor decomposition

breast-cancer kr-vs-kp mushroom



Experimental Evaluation

solar-flare 2 tic-tac-toe vote



Observations

:: data preprocessed by modified GreConD is classified significantly better than data preprocessed by unmodified GreConD

:: data preprocessed by modified GreConD leads to a better classification than original data, already from a few factors covering just 25 % (0.75 on the x-axis) of input data, e.g. for solar-flare 2 or vote dataset – at that point, the number of the factors is considerably lower than the number of factors computed by the unmodified GreConD that cover 100 % of the input data and even lower than the number of original Boolean attributes

→ factors selected based on entropy of class labels appear to be very good new attributes for classification

:: breast-cancer and tic-tac-toe datasets: preprocessed data, either with modified or unmofified GreConD, are (much) better classified than original data, best with factors covering just 40 % (0.6 on the x-axis) of input data – many superfluous attributes or large noise in input data overcome by factors

Acknowledgment

Supported by the ESF project No. CZ.1.07/2.3.00/20.0060 and grant No. 202/10/P360 of the Czech Science Foundation.

International Center for Information and Uncertainty
SSIU 2014: International Spring School and Workshop "Information and Uncertainty"
http://mcin.upol.cz/SSWIU-2014/