Journal of Computer and System Sciences 82 (2016) 521–535

Contents lists available at ScienceDirect

Journal of Computer and System Sciences

www.elsevier.com/locate/jcss

# Exploiting hidden structure in selecting dimensions that distinguish vectors ☆

Vincent Froese *,[1], René van Bevern [2,3], Rolf Niedermeier, Manuel Sorge [2]

*Institut für Softwaretechnik und Theoretische Informatik, TU Berlin, Ernst-Reuter-Platz 7, 10587 Berlin, Germany*

ARTICLE INFO

*Article history:*
Received 18 July 2014
Accepted 15 November 2015
Available online 3 December 2015

*Keywords:*
NP-hardness
Fixed-parameter tractability
W-hardness
Machine learning
Combinatorial feature selection
Dimension reduction
Minimal reduct problem
Combinatorics of binary matrices
Δ-Systems

ABSTRACT

The NP-hard Distinct Vectors problem asks to delete as many columns as possible from a matrix such that all rows in the resulting matrix are still pairwise distinct. Our main result is that, for binary matrices, there is a complexity dichotomy for Distinct Vectors based on the maximum ($H$) and the minimum ($h$) pairwise Hamming distance between matrix rows: Distinct Vectors can be solved in polynomial time if $H \leq 2\lceil h/2 \rceil + 1$, and is NP-complete otherwise. Moreover, we explore connections of Distinct Vectors to hitting sets, thereby providing several fixed-parameter tractability and intractability results also for general matrices.



## 1. Introduction

Feature selection in a high-dimensional feature space means to choose a subset of features (that is, dimensions) such that some desirable data properties are preserved or achieved in the induced subspace. *Combinatorial* feature selection [24,7] is a well-motivated alternative to the more frequently studied affine feature selection. While *affine* feature selection combines features to reduce dimensionality, combinatorial feature selection simply discards some features. The advantage of the latter is that the resulting reduced feature space is easier to interpret. See Charikar et al. [7] for a more extensive discussion in favor of combinatorial feature selection. Unfortunately, combinatorial feature selection problems are typically computationally very hard to solve (NP-hard and also hard to approximate [7]), resulting in the use of heuristic approaches in practice [4,11,17,22].

☆ A preliminary version appeared under the title "A Parameterized Complexity Analysis of Combinatorial Feature Selection Problems" in the proceedings of the 38th International Symposium on Mathematical Foundations of Computer Science (MFCS'13), volume 8087 of Lecture Notes in Computer Science, pages 445–456, Springer, 2013 [19]. Parts of this work originate from the first author's master's thesis on combinatorial feature selection [18]. This article now exclusively focuses on the Distinct Vectors problem and provides all proofs in full detail. It additionally contains a new main result for Distinct Vectors regarding a computational complexity dichotomy for the parameters minimum and maximum pairwise Hamming distance of the data points.

* Corresponding author.
*E-mail addresses:* vincent.froese@tu-berlin.de (V. Froese), rvb@nsu.ru (R. van Bevern), rolf.niedermeier@tu-berlin.de (R. Niedermeier), manuel.sorge@tu-berlin.de (M. Sorge).

[1] Supported by Deutsche Forschungsgemeinschaft, project DAMM (NI 369/13).
[2] Supported by Deutsche Forschungsgemeinschaft, project DAPA (NI 369/12).

[3] Now at Novosibirsk State University, Novosibirsk, Russia.

http://dx.doi.org/10.1016/j.jcss.2015.11.011
0022-0000/

**Table 1**
Overview of our results.

| Result* | Reference |
|---|---|
| NP-hard for $\|\Sigma\| = 2$ and $H \geq 2\lceil h/2 \rceil + 2$ | Theorem 4 |
| poly-time for $\|\Sigma\| = 2$ and $H \leq 2\lceil h/2 \rceil + 1$ | Theorem 9 |
| W[1]-hard wrt. $t$ for $\|\Sigma\| = 2$ and $H \geq 4$ | Corollary 2 |
| W[2]-hard wrt. $k$ ($\|\Sigma\|$ unbounded) | Theorem 10 |
| FPT wrt. $(\|\Sigma\|, k)$ (no poly kernel wrt. $(n, \|\Sigma\|, k)$) | Theorem 12 |
| FPT wrt. $(H, k)$ (for arbitrary $\Sigma$) | Theorem 13 |

* $|\Sigma|$: alphabet size, $h$ ($H$): minimum (maximum) pairwise row Hamming distance of the input matrix, $t$: number of discarded columns, $k$: number of retained columns.

In this work, we adopt the fresh perspective of parameterized complexity analysis. We thus refine the known picture of the computational complexity landscape of a prominent and formally simple combinatorial feature selection problem called Distinct Vectors.

Distinct Vectors
**Input:** A matrix $S \in \Sigma^{n \times d}$ over a finite alphabet $\Sigma$ with $n$ distinct rows and $k \in \mathbb{N}$.
**Question:** Is there a subset $K \subseteq [d]$ of column indices with $|K| \leq k$ such that all $n$ rows in $S_{|K}$ are still distinct?

Here, $S_{|K}$ is the submatrix containing only the columns with indices in $K$. In the above formulation, the input data is considered to be a matrix where the row vectors correspond to the data points and the columns represent features (dimensions). Thus, Distinct Vectors constitutes the basic task to compress the data by discarding redundant or negligible dimensions without losing the essential information to tell apart all data points.

Intuitively speaking, the guiding principle of this work is to identify problem-specific parameters (quantities such as the number of dimensions to discard or the number of dimensions to keep) and to analyze how these quantities influence the computational complexity of Distinct Vectors. The point here is that in relevant applications these parameters can be small, which may allow for more efficient solvability. Hence, the central question is whether Distinct Vectors is computationally tractable in the case of small parameter values.

We are particularly interested in the complexity of Distinct Vectors if the range of differences between data points is small. This special case occurs if the input data is in some sense homogeneous. We measure the range of differences as the gap $H - h$ between the maximum $H$ and the minimum $h$ of pairwise Hamming distances of rows in the input matrix.[4] We initiate the study of this measure by completely classifying the classical complexity of Distinct Vectors with respect to constant values of $H - h$ on binary input matrices. For general matrices, we derive various tractability and intractability results with respect to the parameters alphabet size $|\Sigma|$, number of retained columns and number of discarded columns.

*Related work* Distinct Vectors is also known as the Minimal Reduct problem in rough set theory [28] and it was already early proven to be NP-hard by Skowron and Rauszer [29]. Later, Charikar et al. [7] investigated the computational complexity of several problems arising in the context of combinatorial feature selection, including Distinct Vectors. Seemingly unaware of Skowron and Rauszer's work, they showed that there exists a constant $c$ such that it is NP-hard to approximate Distinct Vectors in polynomial time within a factor of $c \log d$.

Another combinatorial feature selection problem called Minimum Feature Set is a variant of Distinct Vectors where not all pairs of rows have to be distinguished but only all pairs of rows from two specified subsets. This problem is known to be NP-complete for binary input data [12]. In addition, Cotta and Moscato [9] investigated the parameterized complexity of Minimum Feature Set and proved W[2]-completeness with respect to the number of selected columns even for binary matrices.

*Results and outline* Table 1 summarizes our results. We first focus on the case of input matrices over binary alphabets, that is $|\Sigma| = 2$, in Section 3. As our main result, we completely classify the classical computational complexity of (binary) Distinct Vectors according to the gap between $H$ and $h$. This yields the following dichotomy: If $H \leq 2\lceil h/2 \rceil + 1$, then Distinct Vectors is polynomial-time solvable, whereas it is NP-complete in all other cases. The corresponding NP-completeness proof also implies W[1]-hardness with respect to the parameter "number $t = d - k$ of columns to discard".

In Section 4 we consider general alphabets, that is, $|\Sigma| \geq 2$. We prove that, here, Distinct Vectors is W[2]-hard with respect to the number $k$ of retained columns if the alphabet size is unbounded. Moreover, Distinct Vectors cannot be solved in $d^{o(k)}(nd)^{O(1)}$ time, unless W[1] = FPT (which is strongly believed not to be the case [15]). In contrast to these hardness results, we develop polynomial-time data reduction algorithms and show fixed-parameter tractability by providing superexponential-size problem kernelizations with respect to the combined parameters $(|\Sigma|, k)$ and $(H, k)$. We also exclude polynomial-size problem kernels with respect to the parameter combination $(n, |\Sigma|, k)$ based on the hypothesis that NP $\not\subseteq$

[4] See Section 3 for a formal definition.