
Summary: We study Hamming versions of two classical clustering problems. The Hamming Radius $p$-Clustering problem (HRC) for a set $S$ of $k$ binary strings, each of length $n$, is to find $p$ binary strings of length $n$ that minimize the maximum Hamming distance between a string in $S$ and the closest of the $p$ strings; this minimum value is termed the $p$-radius of $S$ and is denoted by $\varrho$. The related Hamming Diameter $p$-Clustering problem (HDC) is to split $S$ into $p$ groups so that the maximum Hamming group diameters is minimized; this latter value is called the $p$-diameter of $S$.

We provide an integer programming formulation of HRC which yields exact solutions in polynomial time whenever $k$ is constant. We also observe that HDC admits straightforward polynomial-time solutions when $k = O(\log n)$ and $p = O(1)$, or when $p = 2$. Next, by reduction from the corresponding geometric $p$-clustering problems in the plane under the $L_1$ metric, we show that neither HRC nor HDC can be approximated within any constant factor smaller than two unless P=NP. We also prove that for any $\varepsilon > 0$ it is NP-hard to split $S$ into at most $pk^{1/\varepsilon}$ clusters whose Hamming diameter does not exceed the $p$-diameter, and that solving HDC exactly is an NP-complete problem already for $p = 3$. Furthermore, we note that by adapting Gonzalez’ farthest-point clustering algorithm [T. Gonzalez, Theor. Comput. Sci. 38, 293–306 (1985; Zbl 0567.62048)], HRC and HDC can be approximated within a factor of two in time $O(pkn)$. Next, we describe a $2^{O(p/\varepsilon)}kO(p/\varepsilon)n^2$-time $(1 + \varepsilon)$-approximation algorithm for HRC. In particular, it runs in polynomial time when $p = O(1)$ and $\varrho = O(\log(k + n))$. Finally, we show how to find in $O((2 + kn)\log n + k^2\log n)(2^p k^2/\varepsilon)$ time a set $L$ of $O(p \log k)$ strings of length $n$ such that for each string in $S$ there is at least one string in $L$ within distance $(1 + \varepsilon)\varrho$, for any constant $0 < \varepsilon < 1$.

MSC:

68W25 Approximation algorithms
68Q25 Analysis of algorithms and problem complexity
62H30 Classification and discrimination; cluster analysis (statistical aspects)

Keywords:

Hamming distance; $p$-clustering problem; np-hardness; approximation algorithms; integer programming

Full Text: DOI

References:

[8] ()
[13] Li, M.; Ma, B.; Wang, L., Finding similar regions in many strings, (), 473-482 - Zbl 1346.68307

This reference list is based on information provided by the publisher or from digital mathematics libraries. Its items are heuristically matched to zbMATH identifiers and may contain data conversion errors. It attempts to reflect the references listed in the original paper as accurately as possible without claiming the completeness or perfect precision of the matching.