Speeding Up Clustering-Based *k*-Anonymisation Algorithms with Pre-partitioning

Grigorios Loukides and Jianhua Shao

School of Computer Science
Cardiff University
Cardiff CF24 3AA, UK
{G.Loukides,J.Shao}@cs.cf.ac.uk

Abstract. K-anonymisation is a technique for protecting privacy contained within a dataset. Many k-anonymisation algorithms have been proposed, and one class of such algorithms are clustering-based. These algorithms can offer high quality solutions, but are rather inefficient to execute. In this paper, we propose a method that partitions a dataset into groups first and then clusters the data within each group for k-anonymisation. Our experiments show that combining partitioning with clustering can improve the performance of clustering-based k-anonymisation algorithms significantly while maintaining the quality of anonymisations they produce.

1 Introduction

A vast amount of data about individuals is being collected and stored worldwide. Such data can contain private information about individuals, for example, their credit card numbers, shopping preferences and medical records. When the data is released for studies such as lifestyle surveys, business analysis and healthcare research, privacy protection becomes a serious concern. Unfortunately, simply removing unique identifiers (e.g. credit card numbers) from data is not enough, as individuals can still be identified using a combination of non-unique attributes such as age and postcode [1].

K-anonymisation is a technique that has been proposed to address this issue. Assume that we have a table T consisting of m attributes $(a_1, \ldots, a_m)$. Without loss of generality we assume that the first q attributes are *quasi-identifiers* (QIDs) - they contain information that can potentially be used to identify individuals (e.g. age and postcode), and the remaining attributes are *sensitive attributes* (SAs) - they contain sensitive information about individuals (e.g. their shopping preferences or diagnosed diseases). K-anonymising T is to derive a view of T such that each tuple in the view is made identical (through some form of data generalisation) to at least $k-1$ other tuples with respect to QIDs [1]. It is easy to see that k-anonymised data helps prevent linking sensitive information to individuals, thereby providing privacy protection.

Many k-anonymisation algorithms have been proposed, employing different search strategies and optimality criteria [2,3,4,5,6]. Broadly speaking, they all

R. Cooper and J. Kennedy (Eds.): BNCOD 2007, LNCS 4587, pp. 203–214, 2007.

attempt to maximise data usefulness (by making as little change to a dataset as possible) and privacy protection (by making individual identification as difficult as possible). One class of such algorithms are clustering-based [4,6,5]. They derive k-anonymisations by first grouping data into clusters of at least k tuples using some quality measures, and then anonymising the data in each group separately using some form of data generalisation. These algorithms offer flexibility in k-anonymisation process and produce high quality anonymisations as result, but they can be rather inefficient to execute, making them not useful for large datasets.

In this paper, we propose a method that partitions a dataset before clustering it for k-anonymisation. Our method is based on the following observation: tuples in a cluster typically belong to a small subspace. Thus, instead of searching the whole dataset when clustering data, we can first find a partition of a dataset and then perform the clustering in each subspace separately. Our experiments show that combining partitioning with clustering can improve the performance of clustering-based k-anonymisation algorithms significantly while maintaining the quality of anonymisations they produce.

The paper is organised as follows. Section 2 describes a metric that we use to measure the quality of k-anonymisations. In Section 3 we introduce two representative clustering-based algorithms for illustrating our pre-partitioning approach. Our approach is presented in Section 4 and evaluated in Section 5. Finally, we conclude in Section 6.

2 Usefulness and Protection Measures

A k-anonymisation of a dataset is commonly derived through some form of data generalisation. Such a generalisation process can result in information loss. To see this, consider the generalisation of data in Table 1 to 4-anonymous data in Table 2, for example.

Table 1. Original data

Age	Height	Sal(K)
20	170	20
23	175	21
25	180	22
27	180	25
28	180	60
29	185	61
58	190	62
80	190	65

Table 2. A 4-anonymisation of Table 1

Age	Height	Sal(K)
[20-58]	[170-190]	20
[20-58]	[170-190]	21
[20-58]	[170-190]	60
[20-58]	[170-190]	62
[25-80]	[180-190]	22
[25-80]	[180-190]	25
[25-80]	[180-190]	61
[25-80]	[180-190]	65

Table 3. Another 4-anonymisation of Table 1

Age	Height	Sal(K)
[20-27]	[170-180]	20
[20-27]	[170-180]	21
[20-27]	[170-180]	22
[20-27]	[170-180]	25
[28-80]	[180-190]	60
[28-80]	[180-190]	61
[28-80]	[180-190]	62
[28-80]	[180-190]	65