Chapter VII
Business Collaboration by Privacy-Preserving Clustering

Stanley R. M. Oliveira
Embrapa Informática Agropecuária, Brazil

Osmar R. Zaïane
University of Alberta, Canada

ABSTRACT

The sharing of data is beneficial in data mining applications and widely acknowledged as advantageous in business. However, information sharing can become controversial and thwarted by privacy regulations and other privacy concerns. Rather than simply hindering data owners from sharing information for data analysis, a solution could be designed to meet privacy requirements and guarantee valid data clustering results. To achieve this dual goal, this chapter introduces a method for privacy-preserving clustering, called Dimensionality Reduction-Based Transformation (DRBT). This method relies on the intuition behind random projection to protect the underlying attribute values subjected to cluster analysis. It is shown analytically and empirically that transforming a dataset using DRBT, a data owner can achieve privacy preservation and get accurate clustering with little overhead of communication cost. The advantages of such a method are: it is independent of distance-based clustering algorithms; it has a sound mathematical foundation; and it does not require CPU-intensive operations.

INTRODUCTION

Data clustering is of capital importance in business and it fosters business collaboration as sharing data for clustering improves the prospects of identifying optimal customer targets, market more effectively and understand customer behaviour. Data Clustering maximizes return on investment supporting business collaboration (Lo, 2002; Berry & Linoff, 1997). Often combining different data sources provides better clustering analysis opportunities. Limiting the clustering on only some attributes of the data confines the correctness of the grouping, while benefiting from additional attributes could yield more accurate and actionable clusters. For example, it does not suffice to cluster

customers based on their purchasing history, but combining purchasing history, vital statistics and other demographic and financial information for clustering purposes can lead to better and more accurate customer behaviour analysis. More often than not, needed data sources are distributed, partitioned and owned by different parties insinuating a requirement for sharing data, often sensitive, between parties. Despite its benefits to support both modern business and social goals, clustering can also, in the absence of adequate safeguards, jeopardize individuals' privacy. The fundamental question addressed in this chapter is: how can data owners protect personal data shared for cluster analysis and meet their needs to support decision making or to promote social benefits? To address this problem, data owners must not only meet privacy requirements but also guarantee valid clustering results.

Achieving privacy preservation, when sharing data for clustering, poses challenges for novel uses of data mining technology. Each application poses a new set of challenges. Let us consider two real-life examples in which the sharing of data poses different constraints:

- Two organizations, an Internet marketing company and an on-line retail company, have datasets with different attributes for a common set of individuals. These organizations decide to share their data for clustering to find the optimal customer targets so as to maximize return on investments. How can these organizations learn about their clusters using each other's data without learning anything about the attribute values of each other?
- Suppose that a hospital shares some data for research purposes (e.g., to group patients who have a similar disease). The hospital's security administrator may suppress some identifiers (e.g., name, address, phone number, etc) from patient records to meet privacy requirements. However, the released data may not be fully protected. A patient record may contain other information that can be linked with other datasets to re-identify individuals or entities (Sweeney, 2002). How can we identify groups of patients with a similar pathology or characteristics without revealing the values of the attributes associated with them?

The above scenarios describe two different problems of privacy-preserving clustering (PPC). We refer to the former as PPC over centralized data and the latter as PPC over vertically partitioned data. To address these scenarios, we introduce a new PPC method called Dimensionality_Reduction-Based Transformation (DRBT). This method allows data owners to find a trade-off between privacy, accuracy, and communication cost. Communication cost is the cost (typically in size) of the data exchanged between parties in order to achieve secure clustering.

This chapter focuses on random projection, a powerful method for dimensionality reduction. The accuracy obtained after the dimensionality has been reduced, using random projection, is almost as good as the original accuracy (Kaski, 1999; Achlioptas, 2001; Bingham & Mannila, 2001). More formally, when a vector in d-dimensional space is projected onto a random k dimensional subspace, the distances between any pair of points are not distorted by more than a factor of $(1 \pm \varepsilon)$, for any $0 < \varepsilon < 1$, with probability $O(1/n^2)$, where n is the number of objects under analysis (Johnson & Lindenstrauss, 1984).

The motivation for exploring random projection is based on the following aspects. First, it is a general data reduction technique. In contrast to the other methods, such as PCA, random projection does not use any defined interestingness criterion to optimize the projection. Second, random projection has shown to have promising theoretical properties for high dimensional data clustering (Fern & Brodley, 2003; Bingham & Mannila, 2001). Third, despite its computational