Optimizing Privacy-Accuracy Tradeoff for Privacy Preserving Distance-Based Classification

Dongjin Kim, University of Maryland Baltimore County, USA
Zhiyuan Chen, University of Maryland Baltimore County, USA
Aryya Gangopadhyay, University of Maryland Baltimore County, USA

ABSTRACT
Privacy concerns often prevent organizations from sharing data for data mining purposes. There has been a rich literature on privacy preserving data mining techniques that can protect privacy and still allow accurate mining. Many such techniques have some parameters that need to be set correctly to achieve the desired balance between privacy protection and quality of mining results. However, there has been little research on how to tune these parameters effectively. This paper studies the problem of tuning the group size parameter for a popular privacy preserving distance-based mining technique: the condensation method. The contributions include: 1) a class-wise condensation method that selects an appropriate group size based on heuristics and avoids generating groups with mixed classes, 2) a rule-based approach that uses binary search and several rules to further optimize the setting for the group size parameter. The experimental results demonstrate the effectiveness of the authors’ approach.

Keywords: Data Mining, Data Security, Distance-Based Mining, Privacy Preserving Data Mining, Privacy Protection

INTRODUCTION
With the huge amount of data and its increasingly distributed sources across organizations, accurate, efficient, and fast analysis of the data for finding knowledge has become a major challenge. In many cases, these factors force companies or organizations to outsource their data mining tasks to a third party. In these circumstances, privacy of the outsourced data is a major concern because without proper protection, the data is subject to misuse.
For example, revealing identity information such as social security number, name, address, and date of birth may lead to identity theft. Another type of privacy risk is that revealing sensitive information such as preexisting medical conditions may cause negative impact such as denial of health insurance. Identity theft was the top concern among customers contacting the Federal Trade Commission (Federal Trade Commission, 2007). According to a Gartner
study (Gartner Inc., 2007), there were 15 million victims of identity theft in 2006. Another study showed that identity theft cost U.S. businesses and customers $56.6 billion in 2005 (MacVittie, 2007). Therefore, legislation such as the Health Insurance Portability and Accountability Act (HIPAA) and the Gramm–Leach–Bililey Act (also known as the Financial Services Modernization Act of 1999) requires that the privacy of medical and financial data be protected.

There has been a rich body of work on privacy preserving data mining (PPDM) techniques. Two excellent surveys can be found at (Aggarwal & Yu, 2008; Vaidya, Zhu, & Clifton, 2005). The goal of privacy preserving data mining is two-fold: to protect privacy of the original data and at the same time still preserve the utility of sanitized data (often measured in quality of data mining). Note that these two goals are conflicting to each other because most PPDM techniques distort the original data (e.g., by adding random noise or making data values less accurate) to provide privacy protection. Obviously, the more distortion introduced, the better the privacy protection, but the lower the utility of data. Most proposed PPDM techniques have some tunable parameters which will lead to different degree of privacy protection and data utility. Thus these parameters need to be set correctly to achieve the optimal privacy and utility tradeoff.

For example, K-anonymity is a very commonly used privacy protection model (Sweeney, 2002a) which makes K people in the data set indistinguishable such that their identities will not be revealed. A number of techniques have been proposed to implement this model (Bayardo & Agrawal, 2005; LeFevre, DeWitt, & Ramakrishnan, 2005, 2006a, 2006b; Samarati, 2001; Sweeney, 2002b; Xiao & Tao, 2006). However all these techniques must set the correct value of K. If K is too large, the data may be distorted too much such that the quality of mining may become very poor. If K is too small, the degree of privacy protection may not be sufficient. More recently researchers have proposed several privacy models such as L-diversity (Machanavajjhala, Kifer, Gehrke, & Venkitasubramaniam, 2007), t-closeness (Li, Li, & Venkitasubramaniam, 2007), and differential privacy (Dwork, 2006). All these models need to set some parameters, e.g., we need to set proper values for L in the L-diversity model, t in the t-closeness model, and ε (the degree of differential privacy) in the differential privacy model.

However, there has been little research on how to tune these parameters efficiently and effectively. Most existing research simply leaves the task of setting parameters to users. However, without proper guidelines, users often have trouble to set the correct parameter values. Another alternative is a brute-force approach. This approach tries many possible settings of parameters and examines the utility (often in terms of mining quality) and the degree of privacy protection of each setting. It then selects the setting with the best utility-privacy tradeoff. However, computing the utility and degree of privacy protection often requires two steps: 1) the privacy preserving technique being considered needs to be applied to the original data set to generate a sanitized data set; 2) the data mining algorithm needs to be executed on the sanitized data set to generate mining results. These two steps are both time consuming and the brute-force approach needs to repeat these two steps for every parameter setting. This is clearly inefficient in practice.

This paper studies the problem of optimizing parameters for a popular privacy preserving technique for distance-based classification: the condensation method (Aggarwal & Yu, 2004). The major benefit of the condensation method as compared to other methods is that it generates synthetic data so it is difficult to recover the identity of the original data. It also preserves the statistical properties of the original data so it works well for multiple distance-based classification algorithms. The condensation method works as follows. It divides data into clusters (groups) such that each cluster contains at least K points (individuals). Each group is then replaced with synthetic data by preserving statistics of the original group. However, the condensation method needs to set an appropriate
Related Content

Dynamical Systems and their Chaotic Properties
[www.irma-international.org/chapter/dynamical-systems-their-chaotic-properties/43270/](www.irma-international.org/chapter/dynamical-systems-their-chaotic-properties/43270/)

[www.irma-international.org/article/information-security-words-alone/34060/](www.irma-international.org/article/information-security-words-alone/34060/)

Information Systems Risk Management: An Audit and Control Approach
[www.irma-international.org/chapter/information-systems-risk-management/20641/](www.irma-international.org/chapter/information-systems-risk-management/20641/)

Privacy and Access to Electronic Health Records
[www.irma-international.org/chapter/privacy-access-electronic-health-records/13522/](www.irma-international.org/chapter/privacy-access-electronic-health-records/13522/)

Disassociations in Security Policy Lifecycles
[www.irma-international.org/article/disassociations-in-security-policy-lifecycles/145410/](www.irma-international.org/article/disassociations-in-security-policy-lifecycles/145410/)