Chapter 4
Decision Trees and Random Forest for Privacy-Preserving Data Mining

Gábor Szűcs
Budapest University of Technology and Economics, Hungary

ABSTRACT

The objective of this chapter is to present brief literature and new results of research in privacy-preserving data mining as an important privacy issue in the e-business area. The chapter focuses on classification problems in business analytics, where the enterprises can gain large profit using predicted results by classification. The decision tree is a well-known classification technique, and its modification by the Randomized Response technique is described for privacy-preserving data mining. This algorithm is developed for all types of attributes. The largest contribution of this chapter is a new method, so called Random Response Forest, consisting of many decision trees and a randomization technique. Random Response Forest is similar to Random Forest, but it is able to solve privacy problems. This consists of many shallow trees, where a shallow tree is a special decision tree with a Randomized Response technique, and the precision of Random Response Forest is better than at a tree.

INTRODUCTION

The e-services collect large amounts of users' activity and evaluation data (Loukis, Pazalos, and Salagara, 2012). The data can be used in business analytics for prediction by classification algorithm, but the collected data contain users' sensitive information as well. Data mining can be used in more e-business areas, such as targeted marketing, customer information management, business intelligence, and e-banking; furthermore, there are many tasks where data mining like classification

DOI: 10.4018/978-1-4666-4181-5.ch004

techniques can help (e.g., credit rating, direct marketing, business process modeling, cross-selling, churn analysis, recommendation system).

E-mail is used as the primary tool for business communication and collaboration. An e-mail-driven business process method has appeared in interaction-centric business process modeling. This method identifies message threads from an e-mail archive, and constructs an interaction-centric process model based on the temporal order and similarity of the threads. A software tool named e-mail Interaction Miner (Stuit and Wortmann, 2012), has implemented this method by process-related information extracted from e-mail header fields. Considering another business problem, the enterprises should prioritize business e-mails over personal ones in their e-mail service. This task can be solved by content-based classification methods to categorize enterprise e-mails into business or personal correspondence (Wang, Tsai, Jheng, and Tang, 2012). But monitoring the details of e-mail contents may violate privacy rights that are under legal protection, requiring a careful balance of accurately classifying enterprise e-mails and protecting privacy rights.

Not only the e-mails, but a large part of corporate information is available in textual data formats. Text classification techniques are well known for managing on-line sources of electronic documents. The identification of key issues discussed within textual data and their classification into different classes could help decision makers or knowledge workers to manage their future activities better (e.g., future project management – Post Project Reviews (Ur-Rahman and Harding, 2012)) – and product or service quality improvement. These classification tasks use training data with sensitive information, but there is a demand for solving these by privacy-preserving algorithms.

The presentation of promotional product information to customers via the Web has become increasingly important for many companies. Customers can easily use the Internet to access information on various products from numerous companies. The information then influences their purchase decisions. Furthermore, companies can collect and analyze the customers' information in order to make better decisions in marketing policy through many types of information technology. The online channels are the most important media for customer–firm interactions, online customer centers deal with customer requirements. These online customer centers can gather knowledge about customers, full of customer reactions on existing products and customer expectations on new products. Furthermore, when analyzed together with Web-logs or customer personnel information, data from the online customer center can be an effective marketing tool (Park & Lee, 2011). Therefore developing customer models (called also profiles in the literature) is an important step for targeted marketing (Romdhane, Fadhel, & Ayeb, 2010). Analysis of customer interactions for this electronic customer relationship management (e-CRM) (Mahdavi, Movahednejad, & Adbesh, 2011) can be performed by using data mining, and the privacy problem can appear in these tasks as well.

One of the cornerstones in CRM is customer churn prediction, where a classifier tries to predict whether or not a customer will leave the company (Coussement, Benoit, & Poel, 2010). In order to accurately forecast and prevent customer churn in e-commerce, a customer churn forecasting framework is established through four steps. First, customer behavior data is collected and converted into a data warehouse by extract transform load (ETL). Second, the subject of data warehouse is established and some samples are extracted as train objects. Third, alternative predication algorithms are chosen to train selected samples. Finally, selected predication algorithm with extension is used to forecast other customers (Yu, Guo, Guo, & Huang, 2011).

E-commerce systems are able to recommend products to buyers as per their preferences (Mohanty & Passi, 2010). Not only Web users, but mobile customers are being tracked and profiled