Knowledge-Based Systems 62 (2014) 38–46

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Privacy-preserving kriging interpolation on partitioned data

Bulent Tugrul [a], Huseyin Polat [b,*]

[a] Computer Engineering Department, Ankara University, 06100 Ankara, Turkey
[b] Computer Engineering Department, Anadolu University, 26470 Eskisehir, Turkey

ARTICLE INFO

Article history:
Received 21 January 2013
Received in revised form 17 February 2014
Accepted 25 February 2014
Available online 15 March 2014

Keywords:
Privacy
Kriging
Partitioned data
Prediction
Geo-statistics

ABSTRACT

Kriging is well-known, frequently applied method in geo-statistics. Its success primarily depends on the total number of measurements for some sample points. If there are sufficient sample points with measurements, kriging will reflect the surface accurately. Obtaining a sufficient number of measurements can be costly and time-consuming. Thus, different companies might obtain a limited number of measurements of the same region and want to offer predictions collaboratively. However, due to privacy concerns, they might hesitate to cooperate with each other.

In this paper, we propose a protocol to estimate kriging-based predictions using partitioned data from two parties while preserving their confidentiality. Our protocol also protects a client's privacy. The proposed method helps two servers create models based on split data without divulging private data and provide predictions to their clients while preserving the client's confidentiality. We analyze the scheme with respect to privacy, performance, and accuracy. Our theoretical analysis shows that it achieves privacy. Although it causes some additional costs, they are not critical to overall performance. Our real data-based empirical outcomes show that our method is able to offer accurate predictions even if there are accuracy losses due to privacy measures.

1. Introduction

Using different techniques to estimate an unknown measurement from known data has received increasing attention. Given a set of known measurements, prediction algorithms compute an unknown measurement using various approaches. Because collecting measurements for every location or every data item might not be practical, it is preferable to predict the unknown measurements from the known ones. Although it is difficult to predict the observed values, such values can be estimated with decent accuracy without performing expensive and time-consuming measurements. Thus, accuracy losses can be balanced by the time and money savings.

Geo-statistics makes a prediction for a specific point based on the measurements for some other locations in the same region [13]. The goal is to estimate an unknown measurement from known ones without too much deviation from the correct value. Geo-statistics usually utilizes kriging and inverse distance weighted (IDW) interpolations for prediction purposes. Geo-statistics has been very popular since the work conducted by Matheron [21], where the author describes the details of the kriging method. It is assumed that the effects of known measurements are inversely proportional to the distance. Kriging has two phases [13]. The first phase is to investigate the gathered data to create a semi-variogram model. The second phase is to make predictions for unobserved coordinates. The concept of kriging was first introduced by a mining engineer, Krige [18]. Kriging formulates Tobler's first law of geography [30]. Tobler's law assumes that things closer to each other are more alike than things farther apart. A short summary of the kriging method is given by Rojas-Avellaneda and Silván-Cárdenas [28]. The basic assumptions and formulas of kriging are presented by Kleijnen [17].

In a traditional kriging interpolation, there are two participating parties. One of them is referred to as the server, which holds measurements for a specific region to make predictions for some locations in the same region. The second party is called the client. Unlike the server, it does not hold measurements; it looks for predictions. The client may need predictions to make a commercial decision in the same region as the server's measurements. Basically, the client sends the coordinates of a sample point to the server. Then, the server calculates the prediction based on its data using kriging interpolation and returns it to the client.

Two different companies might have measurements for the same region. It might be costly to collect enough measurements in a specific region; collecting sufficient measurements for kriging

[*] Corresponding author. Tel.: +90 222 321 3550.
E-mail address: polath@anadolu.edu.tr (H. Polat).

http://dx.doi.org/10.1016/j.knosys.2014.02.017
0950-7051/© 2014 Elsevier B.V. All rights reserved.

usually requires considerable time and money. Due to limited time and budgets, such parties might decide to provide kriging interpolations based on their integrated data instead of collecting all of the measurements individually. Moreover, the accuracy and dependability of geo-statistics models, in general, depend on the total number of available measurements. Hence, two parties, even competing ones, may create a model based on their integrated data and offer predictions collaboratively. However, they are hesitant to share their data with each other because the collected measurements are considered confidential. Additionally, the client considers the target location and the estimated prediction private. Thus, the parties need to perform kriging interpolations while preserving their privacy and the clients' confidentiality.

We propose a privacy-preserving kriging interpolation method to estimate predictions on data partitioned between two parties while preserving their privacy and the clients' privacy. Our method helps such data owners and the clients perform kriging interpolations on partitioned data without violating their confidentiality. The proposed method prevents the data owners from deriving information about each other's private data. It also prevents them from learning the client's confidential data. At the same time, the client cannot learn the data owners' private data. Due to privacy measures, our solution might introduce extra costs, such as storage, communication, and computation. However, such costs should not prevent the servers from providing predictions efficiently. Moreover, accuracy losses are inevitable due to the utilized privacy measures. On the one hand, we hypothesize that accuracy improves if two servers decide to collaborate. On the other hand, privacy measures cause accuracy losses. However, the overall gains due to collaboration should compensate for the losses due to privacy concerns.

The remainder of the paper is organized as follows: In Section 2, we survey related studies in the literature and briefly present the differences between our work and previous studies. After summarizing kriging interpolation in Section 3, we describe our proposed scheme in detail in Section 4. We investigate our method with respect to supplementary storage, computation, communication costs, and privacy in Section 5. In Section 6, we discuss our real data-based experiments and their outcomes to analyze our solution in terms of accuracy based on the empirical outcomes. Finally, we conclude our paper and provide some future directions for research in Section 7.

2. Related work

With the spread of data mining methods, general privacy-preserving data mining (PPDM) and PPDM on partitioned data are becoming very popular. These methods allow us to apply data mining methods without disclosing confidential data. Horak et al. [9] and Branković et al. [6] find a largest family of a range of queries that does not lead to compromising of a database and determine the usability of the database. The authors prevent query owners from learning the input values of the database. Dwork [7] introduce differential privacy, which ensures that the removal or addition of a single item does not affect the outcome of the analysis.

Data collected for various data mining purposes might be partitioned among two or more companies. There are various data partitioning scenarios, including horizontal, vertical, or arbitrary partitioning. In horizontal partitioning, different rows are combined to conduct data mining methods. Different columns are combined in vertical partitioning. In addition to these methods, arbitrary partitioning assumes that different parties hold different portions of the database [12].

Secure two-party computation is a special case of the multi-party case, which was first defined by Yao [38]. Goldreich et al. [8] extend the two-party case to a multi-party case. Lindell and Pinkas [20] present a scenario, where two parties with private databases wish to cooperate by calculating a data mining algorithm on the union of their databases. To mine association rules over horizontally partitioned data (HPD) between two parties while preserving privacy, Kantarcioglu and Clifton [14] propose a method based on encryption. Kantarcioglu and Vaidya [15] propose privacy-preserving methods for learning a naïve Bayesian classifier (NBC) from HPD. Yi and Zhang [39] propose a two-party protocol and a multi-party protocol to achieve an NBC on HPD without violating the data owners' privacy. The authors in [11] discuss how to create a dissimilarity matrix from HPD while preserving confidentiality. In another study, Kaya et al. [16] examine how to perform clustering in a distributive manner based on HPD with privacy. Their scheme is based on an efficient homomorphic additive secret sharing method. A secure calculation of Hamming and Euclidean distance calculations for two parties are studied by Rane et al. [27]. They apply their solution to a private biometric authentication problem.

Vaidya et al. [35] introduce a generalized privacy-preserving variant of the ID3 algorithm for vertically partitioned data (VPD) over two or more parties. Skillicorn and McConnell [29] propose a simpler privacy-preserving prediction approach for VPD, called attribute ensembles. Vaidya and Clifton [34] investigate how to estimate an NBC from VPD while preserving the data holders' confidentiality.

Privacy-preserving recommendation approaches for partitioned data are also introduced in the literature to provide predictions to customers. Polat and Du [24] present a privacy-preserving scheme to provide predictions based on VPD without jeopardizing the confidential data of the online vendors. In another study [25], the authors investigate how to estimate top-N recommendation lists from HPD between two online vendors while preserving their privacy. Yakut and Polat [36] study how to estimate singular value decomposition-based predictions based on HPD or VPD while preserving the data holders' privacy. Basu et al. [4] propose a privacy-preserving item-based prediction scheme, which can be applied to both HPD and VPD. Their scheme is based on an additively homomorphic public-key cryptosystem. Basu et al. [3] show the feasibility of privacy-preserving prediction services on a cloud platform.

Jagannathan and Wright [12] present a privacy-preserving protocol for k-means clustering based on arbitrarily partitioned data (APD). To cluster APD while preserving privacy, Prasad and Rangan [26] develop a privacy-preserving BIRCH algorithm, where they introduce secure protocols for distance metrics. Hu et al. [10] present a privacy-preserving support vector machine classifier solution on APD. Their scheme does not divulge any confidential data held by data holders. To perform neural network learning from APD without violating privacy, Bansal et al. [2] present a privacy-preserving algorithm. They show that their algorithm leaks no knowledge about data holders' data. Li et al. [19] offer a privacy-preserving distance-based outlier detection protocol on APD. In another study [37], the authors discuss how to provide numeric ratings-based predictions from APD while preserving confidentiality.

Privacy-preserving geo-statistics has also been receiving increasing attention lately. Tugrul and Polat [31] propose a privacy-preserving scheme to provide kriging-based predictions from data held by a single server while preserving confidentiality. Their method helps a server and a client perform kriging interpolations while hiding their confidential data from each other. In another study [32], the authors investigate how to estimate IDW-based predictions without jeopardizing the server's and the client's privacy. It is assumed that data collected for interpolation purposes is held by a single server. Therefore, these two studies are considered as central server-based schemes because the data are held by a single party. Unlike their methods, we study how to estimate kriging-based predictions from data partitioned between two parties