Chapter 31
A Clustering Approach Using Fractional Calculus-Bacterial Foraging Optimization Algorithm for k-Anonymization in Privacy Preserving Data Mining

Pawan R. Bhaladhare
SNJB's College of Engineering, India

Devesh C. Jinwala
Sardar Vallabhbhai National Institute of Technology Surat, India

ABSTRACT

A tremendous amount of personal data of an individual is being collected and analyzed using data mining techniques. Such collected data, however, may also contain sensitive data about an individual. Thus, when analyzing such data, individual privacy can be breached. Therefore, to preserve individual privacy, one can find numerous approaches proposed for the same in the literature. One of the solutions proposed in the literature is k-anonymity which is used along with the clustering approach. During the investigation, the authors observed that the k-anonymization based clustering approaches all the times result in the loss of information. This paper presents a fractional calculus-based bacterial foraging optimization algorithm (FC-BFO) to generate an optimal cluster. In addition to this, the authors utilize the concept of fractional calculus (FC) in the chemotaxis step of a bacterial foraging optimization (BFO) algorithm. The main objective is to improve the optimization ability of the BFO algorithm. The authors also evaluate their proposed FC-BFO algorithm, empirically, focusing on information loss and execution time as a vital metric. The experimental evaluations show that our proposed FC-BFO algorithm generates an optimal cluster with lesser information loss as compared with the existing clustering approaches.

DOI: 10.4018/978-1-5225-7113-1.ch031

1. INTRODUCTION

With the rapid growth of the database technologies, an immense amount of personal data of individuals has been collected for the analysis purpose by the various organizations. Data mining techniques have been used to find out the useful information from the collected data. The collected data could be associated with the medical database, voter database, and census database. However, the collected data might contain sensitive personal data. Once mining such data, the individual privacy could be in danger and would disclose his/her personal sensitive data. Therefore, sensitive data needs to be protected before conducting the data mining. For this reason, the privacy preserving data mining becomes an important issue in recent years (Agrawal & Srikant, 2000; Lindell & Pinkas, 2003).

Generally, two main approaches have been discussed in the literature for preserving the privacy. First approach supports cryptographic techniques (Zhan, 2007; Upmanyu et al, 2010; Jagannathan et al, 2010) and the second support non-cryptographic techniques (Sweeney, 2002; Samarati,1998; Byun et al, 2007; LeFevre et al, 2006; Loukides & Shao, 2007; Chiu & Tsai, 2007; Lin & Wei, 2008; Kabir et al, 2011; Gionis & Tassa, 2009; Machanavajjhala et al, 2006; Li & Li, 2007; Wong et al, 2006; Goldberger & Tassa, 2010; LeFevre et al, 2005; Moon et al, 2001; Meyerson & Williams, 2004; Aggrawal et al, 2005; Iyenger, 2002; Nergiz & Clifton, 2006; Ghinita, 2009; Bayardo & Agrawal, 2005; Kiffer & Gehrke, 2006; Samarati, 2001; Gionis et al, 2008). However, our focus here is on non-cryptographic approaches, owing to the lesser computation cost of the same as compared to their cryptographic counterpart (Zhan, 2007; Upmanyu et al, 2010; Jagannathan et al, 2010).

One of the methods amongst the non-cryptographic approach is the *k-anonymity model* (Sweeney, 2002; Samarati, 1998; Samarati, 2001). The k-anonymity model protects sensitive data from identification using any combination of data *generalization* and/or *suppression* (Sweeney, 2002). The k-anonymity model partitions the records into several groups in such a way that each group contains at least k similar records. Such a group of similar records represents a cluster.

For example, consider a medical database and their corresponding 3-anonymized medical database is as shown in Table 1 and Table 2, respectively. The database consists of three kinds of attributes such as *identifier*, *quasi-identifier* and *sensitive attribute*. In Table 2, we explicitly remove the attribute *Name* from the database and replaced it with the attribute *Row number*. This is necessary to protect the identity and preserve the privacy of an individual. The generalization and/or suppression techniques have been perform on the quasi-identifier such as *Age, Gender* and *Zipcode* to preserve the privacy of an individual in the database. The generated table is considered as an anonymized database. In Table 2, we shows the use of k-anonymity approach for the privacy preservation. As there are three records in each group, therefore, the generated Table 2 is called as 3-anonymized database. However, the sensitive attribute is kept as it is. Thus, the database after processing is given to the data miner for further analysis.

However, in doing so, we clearly loose the specific information about the quasi-identifier attributes viz. *Age, Gender and Zipcode* of an individual. Evidently, the information loss increases with increasing the level of generalization and/or suppression.

In the recent literature, a state of the art clustering approaches (Sweeney, 2002; Samarati,1998; Byun et al, 2007; LeFevre et al, 2006; Loukides & Shao, 2007; Chiu & Tsai, 2007; Lin & Wei, 2008; Kabir et al, 2011; Gionis & Tassa, 2009; Machanavajjhala et al, 2006; Li & Li, 2007; Wong et al, 2006; Gold-