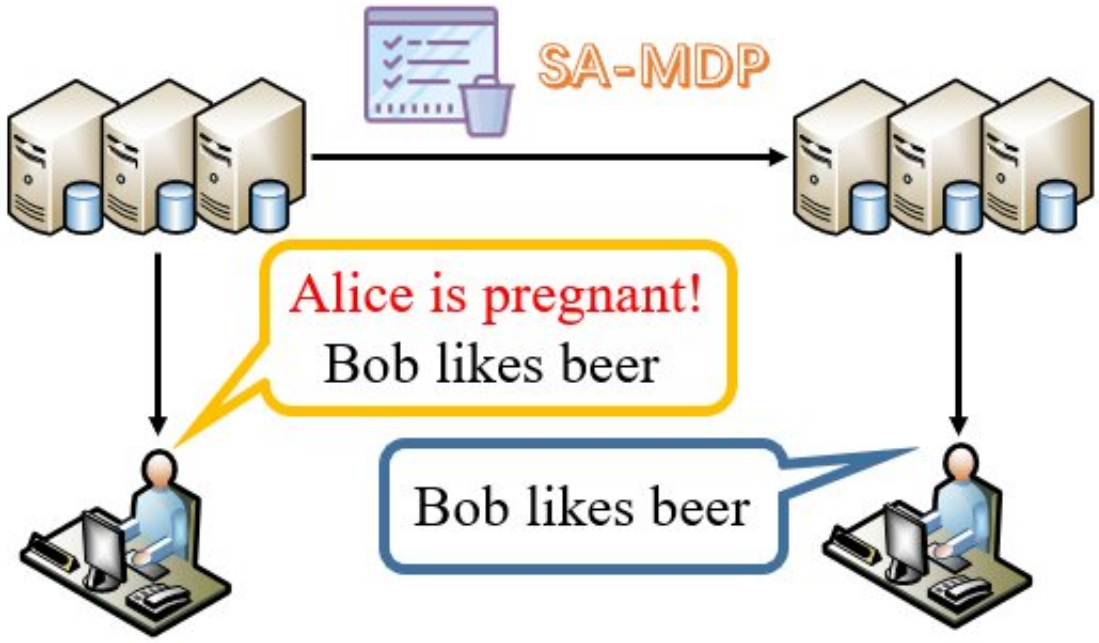
Researchers give privacy boost to sensitive proprietary patterns discovered in data mining

June 15 2022



Researchers reveal how to support association rules mining on published datasets while providing privacy protection for specific rules. Credit: Tsinghua University Press

Researchers have given a boost to privacy and protection of proprietary or other sensitive information during data mining, while not

compromising on the ability to discover useful patterns in huge datasets.

The technique, developed by a pair of computer scientists at Chongqing University, is described in an article published in the journal *Big Data Mining and Analytics*.

Data mining, the discovery of patterns in very large sets of data—often involving machine learning—and the sharing of that information for useful purposes frequently hits a roadblock when such data patterns are proprietary, undermine privacy, or compromise security. And yet such data sharing or publication enhances further discovery of useful patterns of benefit to the owners of those datasets and society at large.

Consider a very common data mining algorithm for discovering potentially useful relations between variables in large datasets: association rule mining. The classic, possibly fictional, example of association rule mining concerns a large dataset of supermarket sales, where it is discovered that male customers who buy diapers also tend to buy beer. The "rule" here is the association of beer, diapers and male customers. Based on this rule, a supermarket manager can offer a discount package for those buying beer and diapers together.

But were this "rule" to be discovered by competitors using a published dataset that the supermarket had shared to enhance further pattern discovery, they could steal customers from the original supermarket by providing the same discount strategy. The "diapers-means-beer" rule is thus commercially sensitive and would need to be protected before the supermarket would be comfortable in publishing its data for others to use.

Put another way, if greater data sharing is to be encouraged, there needs to be a way to allow data mining for non-sensitive association rules (NARs) while protecting data mining from discovering sensitive

association rules (SARS).

To solve the sensitive association rule problem, researchers in the past have proposed protecting the sensitive information by simply hiding it after discovery before any sharing of the dataset. This is achieved by decreasing the frequency of the appearance of any data in the dataset that suggest the association rule. This is however not very practical as only one such SAR can be protected at any one time, and the technique does not provide strong data privacy anyway.

Other researchers have tried to transform the SAR problem into a single objective optimization problem—finding the best solution for a specific criterion. This strengthens the data privacy but reduces the utility of the dataset. Another approach involves encrypting the data before performing any data mining on the dataset, but this can be very time-consuming, especially when implemented on particularly large datasets—the very ones with the greater potential to discover patterns of interest.

So the Chongqing researchers wanted to find a solution that decreases the potential for privacy leakage while also improving the data utility, and to do so while limiting the time such a technique would take.

Their solution, which they call "optimized sanitization approach for minable data publication," or simply SA-MDP, recognizes that any solution to the SAR problem needs to find an acceptable trade-off between data utility and data privacy, rather than solving for one or the other independently. This is a multi-objective optimization problem, rather than a single-objective optimization problem—where more than one objective must be optimized. While many fields, from logistics to engineering regularly face such problems, they are inherently thorny ones. A traveler wanting to find the cheapest plane ticket on a convenient day with the most comfortable seat while taking the shortest

journey with the fewest layovers is confronting a multi-objective optimization problem. The challenge lies in the fact that no one single solution exists that simultaneously optimizes each of these objectives; instead, there may be many, perhaps even an infinite number of optimal 'candidate' solutions that are equally good.

For SA-MDP, the researchers designed a customized "particle swarm optimization" (PSO) algorithm to efficiently solve this multi-objective optimization problem. The PSO method, a biologically inspired algorithm, was originally discovered in the 1990s by researchers aiming to simulate the social behavior of animals that swarmed such as flocks of birds or schools of fish. But the researchers found that their algorithm was in fact performing optimization calculations to solve problems for the swarm. Under PSO, a large group of candidate solutions are treated as particles like birds in a flock in the "search space"—the set through which the algorithm searches. Moving these particles within the search space according to some basic mathematical rules governing a particle's velocity and position is akin to imagining each individual bird helping the flock as a whole find the optimal solution.

To improve the exploration ability of SA-MDP, the technique also introduces the concept of particle splitting, which enables a particle to produce several "child particles."

And to speed up the process, the method involves a novel preprocessing mechanism that removes any irrelevant transactions so that the size of the search space can be decreased.

Having designed the new approach, the researchers then tested it on several publicly available datasets commonly used in such testing—a set of chess movements, a dataset of mushroom attributes used to classify them into edible or poisonous, and a series of clickstreams (the sequence of links clicked on) of visitors to websites. They found their technique

easily beat the competition.

"Our method provides the same privacy protection as the standard approach for hiding sensitive association rules, but with better data utility, all the while slashing running time," said Xiaofeng Liao, a computer scientist at Chongqing University and co-author of the paper with his doctoral student Fan Yang.

They compared these results to those of the cuckoo search optimization algorithm for hiding sensitive association rules, or COA4ARH, a common algorithm used to hide sensitive association rules (association rule hiding) when data mining.

They found that their approach delivered the same protective effect as COA4ARH's ability to hide sensitive rules, and beat it on ability to produce useful association rules, while cutting running time in half.

More information: Fan Yang et al, An Optimized Sanitization Approach for Minable Data Publication, *Big Data Mining and Analytics* (2022). DOI: 10.26599/BDMA.2022.9020007

Provided by Tsinghua University Press