ABSTRACT

In 1990s, the evolution of information technology enabled many enterprises to collect and store huge amounts of data in their databases. Since a lot of business intelligence is hidden in the data, companies developed data mining tools that performed variety of analysis based on statistics in order to find patterns and regularities in the data. Association rule mining, a data mining technique, is very useful for product promotion and helps online-shops in personalizing their website and cross-sell their products by making recommendations. This task of association rule mining requires the organizations to share their data. In several dealings the need for sharing data occurs while outsourcing data mining to solve a particular business problem. Most of this sharing takes place with little secrecy which often results in the organizations taking unnecessary risks while sharing their data and also increases the legal responsibility of the parties involved in that process.

Association rule mining imposes threat to sharing of data as it is possible to infer sensitive information from non sensitive data. The challenge is to defend tactical decisions, while not dropping the benefit of association rule mining. An association rule is characterized as sensitive when its confidence is above disclosure threshold. These sensitive rules should be made uninteresting before releasing the dataset publicly. This is done by modifying the data such that the confidence of these sensitive rules is reduced below disclosure threshold. However, it is essential to maintain a suitable balance between privacy protection and knowledge discovery. So research on hiding sensitive association rules becomes essential.

A lot of techniques have been proposed in literature for hiding sensitive association rules. These techniques are categorized into Exact, Border Based, Cryptographic, Reconstruction and Heuristic Techniques. Most of these techniques hide sensitive rules mined from binary datasets. These techniques also generate higher side effects in terms of lost rules, ghost rules, hiding failure and amount of distortion to the original dataset. But real life applications consist of quantitative data, which cannot be mined using classical mining technique. The common way of mining rules from quantitative database is to divide the quantity of items into intervals and to mine rules from these intervals. But sharp boundary problem, in which the members of the set that are near the boundary of the intervals are either ignored or overemphasized, is a serious setback of this method. So fuzzy concept is used to mine quantitative association rules.

This research proposes techniques to hide sensitive fuzzy association rules mined from quantitative database with minimal side effects while maintaining an optimal balance between knowledge mined and privacy attained. Initially, the performance of the existing technique called Decreasing the Support of Right Hand Side of the Rule (DSR) is analyzed. Various techniques proposed in the research include Weighted Item Grouping Approach (WIGA), Rank Based Correlated Item Hiding Approach (R-CA), Genetic Algorithm (GA) based Fuzzy Sensitive Rule Hiding (GA-FSH), PSO based Fuzzy Sensitive Rule Hiding (PSO-FSH), Hybrid PSO-GA Fuzzy Sensitive Rule Hiding (PSO-GA-FSH) and Differential Evolution based Fuzzy Sensitive Rule Hiding (DE-FSH). The proposed techniques hide the sensitive rule by decreasing the support of the items occurring in the sensitive rule.

Novel objective measures namely Association Measure (AM) and Transaction Sensitivity Support (TSS) are proposed to reduce the side effects. The performance of the objective measures like confidence, mutual information, association measure and transaction sensitivity support were analyzed in terms of side effects produced. Finally, a parallel approach for fuzzy association rule mining and hiding using map reduce programming model is proposed and tested on a Hadoop cluster.

The effectiveness of hiding is measured using various metrics like number of lost rules, number of ghost rules, hiding failure, percentage of modification and computation time. The proposed techniques are tested with the breast cancer dataset from UCI machine learning repository, University of California and the traffic accidents dataset obtained from National Institute of Statistics (NIS) in the region of Flanders (Belgium) for the period 1991-2000.

Experimental results show that the number of non sensitive rules lost as a side effect of hiding sensitive rules is less in GA-FSH technique and DE - FSH technique using mutual information as objective measure. PSO and DE based techniques does not generate any ghost rules while other techniques generate more number of ghost rules. The WIGA and R-CA techniques that implement both Increasing the Support of Left Hand Side item of the Rule (ISL) and DSR to cope with transitivity of items in sensitive association rules result in generation of more ghost rules.

The amount of data distortion or the percentage of entries in original dataset perturbed to hide sensitive rules is less in both the techniques GA-FSH and DE - FSH using mutual information as objective measure, while WIGA and R-CA records higher modification. It is found that WIGA and R-CA has a hiding failure of 33% and GA technique has a hiding failure of 22% while the DSR, PSO, hybrid PSO-GA and DE based techniques have no hiding failure.

Techniques like DSR, WIGA and R-CA takes less time when compared with the evolutionary algorithms like GA, PSO and DE for hiding sensitive association rules. Among the proposed evolutionary approaches DE-FSH takes less time for hiding. The results are more evident when the size of the dataset is increased.

The parallel implementation reduces the lost rules by an average of 3.5% and percentage of modification by 4.5 % when compared to the serial approach. The parallel implementation also shows 70% improvement in speed.

From the experiments, it is concluded that Differential Evolution algorithm implemented with Mutual Information as objective function is efficient in hiding sensitive rules as it is not prone to hiding failure and ghost rule generation. It also reduces data modification which results in the reduction of non sensitive rules that are falsely hidden.