Indian Journal of Science and Technology Vol. 5 No. 6 (June 2012) ISSN: 0974- 6846

Privacy preserving optimized rules mining from decision tables and decision trees

Ahmed Saeed Alzahrani and Muhammad Shuaib Qureshi

Department of Computer Science, Faculty of Computing & I.T, King Abdulaziz University, Jeddah 21589, Kingdom of Saudi Arabia.
asalzahrani@kau.edu.sa, qureshi.shuaib@gmail.com

Abstract

With the swelling amount of data mining projects and information sharing, preserving privacy is a challenging issue which needs higher priority to guarantee security. This paper is an extension of Qureshi *et al.* (2010) with the addition of security features to the existing classification rules to preserve privacy. Experimental results show secure, accurate, sanitized and optimized rules which we achieve by using the concept of genetic algorithm (GA).

Keywords: Genetic algorithm, Decision tree, Decision table, Knowledge base, Classification rules.

Introduction

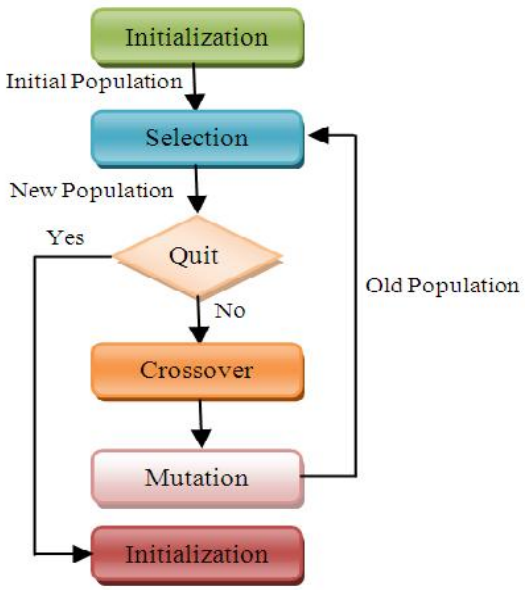
The pool of heuristics, stored facts and rules that are organized and can be used for problem solving is called knowledge base (KB). Various techniques are used to organize knowledge such as Semantic Nets, Frames, Decision Tables (DTables), Decision Trees (DTrees), etc (Wikipedia, 2009). Every knowledge representation scheme has certain inherent pros and cons. Among these DTables and DTrees are considered efficient and mostly used techniques.

A DTable encompasses rows and columns, separated by four quadrants: (a) The upper left quadrant indicates conditions (b) Upper right quadrant holds condition rules or alternatives (c) Lower left quadrant encloses certain actions to be executed for a specific combination of condition states and (d) The lower right quadrant contains action rules. In the context of knowledge based systems, in majority cases, the DTable techniques easily allow to check for ordinary validation and verification such as contradiction, inconsistencies, incompleteness, redundancy etc. in the problem specification (Nguyen *et al.,* 1987).

A DTree is a decision support tool that uses a graph or model of decisions and their possible consequences including chance event outcomes, resource costs and utility (Wikipedia, 2009). In DTree, nodes and links represent goal and decision respectively. DTree can easily be converted into other formats like set of (mutually exclusive) rules etc.

Literature explores variety of classification modules like Value Reduction, DTrees, DTables, NNs, Statistical Model, Bayesian Classifier etc, where DTrees, DTables and Classification Rules are the commonly used classification methods (Liu *et al.*, 2003). Similarly, for classification rules optimization different optimization techniques are used such as Genetic Algorithm (GA) (Fig.1), Ant Colony Optimization (ACO) (Fig.3), and Bionomic Algorithm (BA) etc. Among these, GA is considered best oriented for optimization purposes. It resists being trapping in local optima. It uses four different operators i.e., selection, crossover, mutation and reproduction for performing different operations. The basic building block of GA is chromosome that is an encoded form used for producing different individuals (Encyclopedia, 2012). Chromosome uses mutation process for generating new offsprings called as new population, with the assumption that the new population will have probably good outcomes than the previous one. For producing new population, offsprings are selected on the basis of fitness function (Encyclopedia, 2012). This process is repeated until a best solution is produced. Fig. 1 depicts GA operation model and Fig. 4 its detailed step-by-step procedure.

Fig. 1. Genetic algorithm process



System model and background

Knowledge is usually transformed into different layouts for decision making purposes. Every knowledge representation scheme is appropriate for specific situation. Hence, mapping knowledge onto different forms is necessary for granting earlier response and less computational amount. The DTable might be transformed into DTree and consequently DTree into set of human interpretable rules. Set of rules can be easily altered by a user as compared to DTable or DTree.

Indian Journal of Science and Technology Vol. 5 No. 6 (June 2012) ISSN: 0974- 6846

Fig. 2. Secure rules engineering model

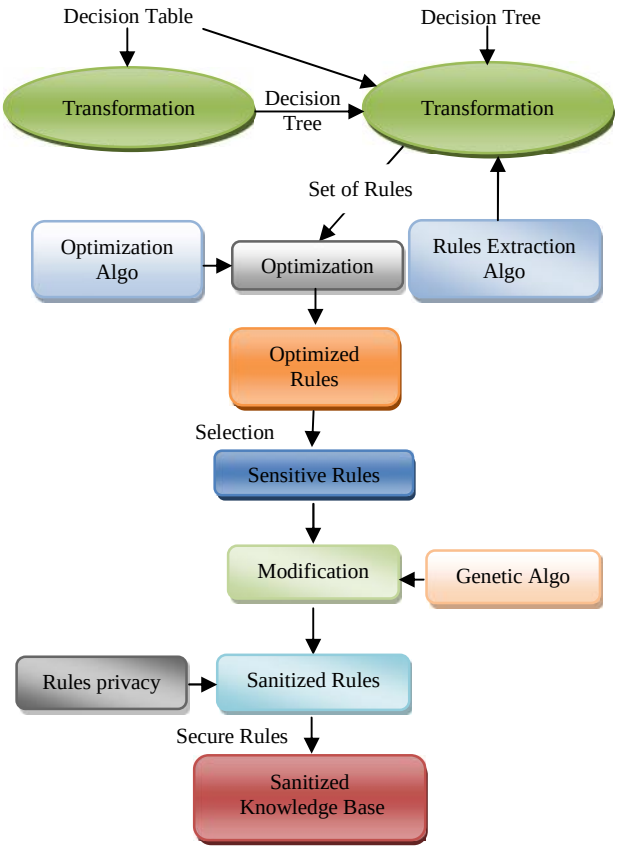


Fig. 3. Ant Miner Algorithm (Liu *et al.*, 2003)

Begin
TrainingDataset ⟵ All Examples
SetupRules_Array ⟵ Empty
While
TotalElements_TrainingDataset>Maximum_Uncovered_Example
Do
I ⟵ initializePheromones
BestRule ⟵ Empty
k ⟵ 1
Repeat
$Rule_k$ ⟵ GenerateRule ()
ComputeConsequent ($Rule_k$)
Prune ($Rule_k$)
UpdatePheromones (£, $Rule_k$)
If *R($Rule_k$) > R(BestRule)*
Then *BestRule ⟵ $Rule_k$*
End
k ⟵ k+1
Until *k ≥ Max|Rule|* ***OR***
Convergence
SetupRules_Array ⟵ SetupRules_Array U BestRule
TrainingDataset ⟵ TrainingDataset
CorrectlyCoverExamples (BestRule)
End
End

Fig. 4. Genetic Algorithm (Naderi et al., 2009)

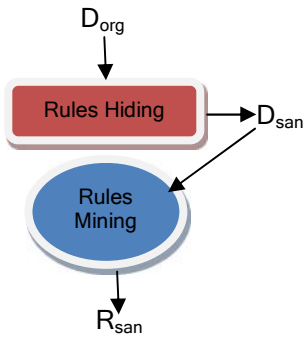
[Start] *Generate population of total n chromosomes*
[Fitness] *Calculate the fitness function of each chromosome exists in the population.*
[New Population] *Generate a new population by go through the steps until the new population is complete;*

1. ***[Selection]*** *Select two best parents from a population by evaluating their fitness.*
2. ***[Crossover]*** *Crossover the parents for new offspring. Without crossover there will be no difference in the parent and offspring.*
3. ***[Mutation]*** *Mutate a new offspring at each location in chromosome.*
4. ***[Accepting]*** *Give position to the new offspring in a new generated population.*

[Replace] *Process this new generated population for further run of algorithm*
[Test] *If the exit condition reaches,* ***Stop*** *and then return the best solution in the current population.*
[Loop] *Goto second step.*

Knowledge reduction is most important issue in the exploration of rough set theory (Chieh *et al.*, 2009). NP-hardness of optimal reduction of DTable is already proved. Thus people have been trying to search for more efficient heuristic algorithms to get an approximate reduction of DTables all the time (Decai Huang & Lingli Wang, 2006). Numerous phases of optimized tree production are NP-hard. The NP-completeness of constructing optimized DTree from DTable was proved (Hyafil & Ronald, 1976). NP-completeness of discovering the optimized parent node was discussed earlier (Cox *et al.*, 1989). NP-completeness of building optimal trees was proved earlier (Murphy & Mccraw, 1991; Naumov, 1991). The NP-hardness of building optimized tree structured vector quantizers (TSVQ) was discussed by (Lin *et al.*, 1994). Issues of space complexity and erroneous decisions are caused due to Knowledge redundancy (Decai Huang & Lingli Wang, 2006). The quality of DTree is determined by its complexity and accuracy. Optimal decision tree construction is an NP-hard problem (Liu *et al.*, 2003). Classification algorithm is a kind of important technology in Data Mining. Various classification modules exist like Value Reduction, DTree, DTable, Neural Network, Statistical Model, Bayesian Classifier, etc. The most commonly used methods are DTree, DTable and Rules (Liu *et al.*, 2003). "How to apply the decision logic" is a major issue during the construction of complete KB (Efraim & Jay, 2003). Production of fully optimized DTable and DTree are NP-hard problems. So "how to mine optimized, secure and sanitized rules from DTable and DTree? and how to ensure privacy of the optimized rules?"

Fig. 5. Rules hiding process



Indian Journal of Science and Technology Vol. 5 No. 6 (June 2012) ISSN: 0974- 6846

Fig. 6. Number of rules

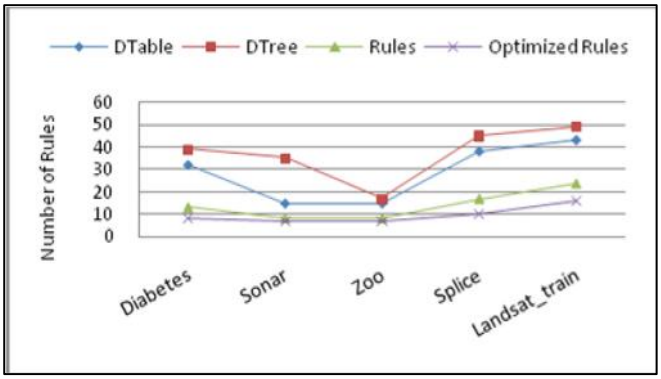


Fig. 7. Accuracy

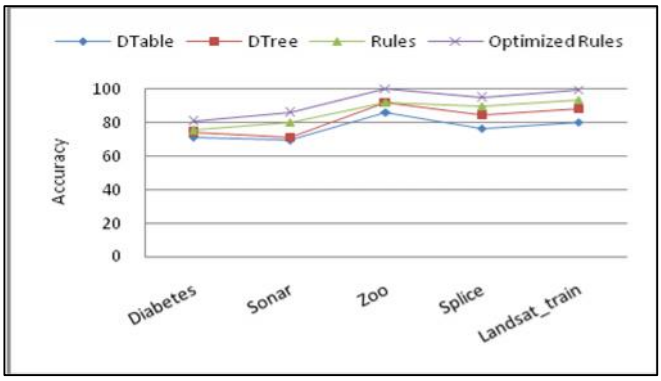


Fig. 8. Number of ghost rules

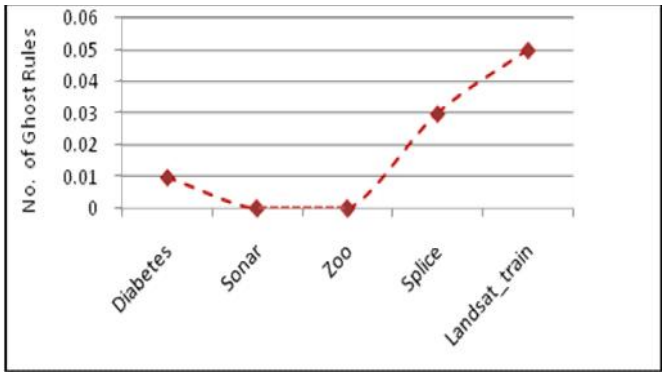


Fig. 9. Rules hiding failure

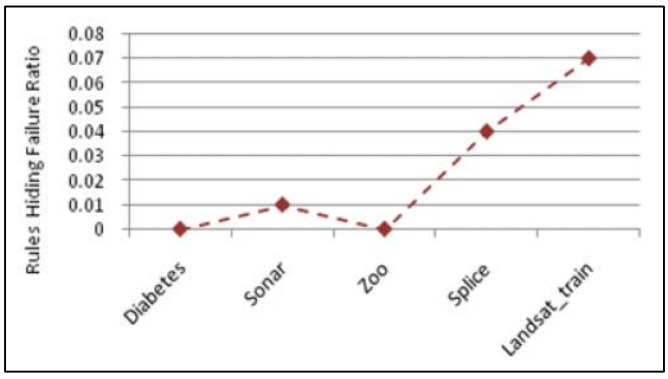
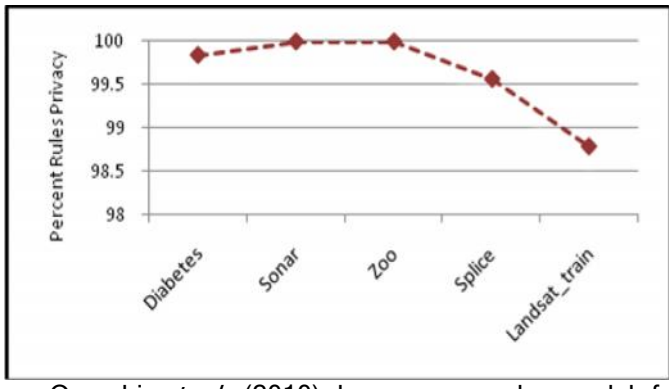


Fig. 10. Rules privacy



Qureshi *et al.* (2010) have proposed a model for optimal transformation of DTable and DTree into set of rules. With the swelling amount of data mining projects and information sharing, preserving privacy is a challenging issue which needs higher priority for security assurance. This paper is an extended version of Qureshi *et al.* (2010) with the introduction of privacy and security related features to the existing classification rules. After rules optimization, selection procedure is executed and sensitive rules are selected on the basis of its fitness from the optimized rules. The sensitive rules are then modified using genetic algorithm. As a result secure sanitized and optimized rules are achieved by enforcing rules privacy and then transformed into KB. Fig. 2 shows secure rules engineering procedure.

Rules hiding process

Let D_{org} is original dataset and D_{san} is a sanitized dataset after applying rules hiding procedure. This sanitized dataset D_{san} is mined using rules mining algorithm and hence sanitized rules R_{san} are extracted. Fig. 5 represents the process pictorially.

The key objective is to prevent sensitive rules from being exposed. R_{sen} is a set of sensitive rules, whereas R_{Total} and R_{NonSen} are sets of all classification and non sensitive rules respectively. R_{sen} is a sub-set of R_{Total}.

$$R_{Sen} \subseteq R_{Total}$$

$$R_{NonSen} = \left(R_{Total} - R_{Sen}\right)$$

Fitness strategy used for hiding sensitive rules

The applied fitness function is based on weighted sum function. Suppose W_1, W_2 are weights and C_{min} is the minimize cost function, then

$$C_{min} = (W_1 \times Rules\ Hiding\ Distances) + (W_2 \times No.of\ Modifications)$$

Where $(W_1 + W_2) = 1$

Rules Hiding Distances = $\sum_{k=1}^{|R_{Sen}|} R_k Hiding\ Distance$

Number of Modifications = $\sum_{i=1}^{|Critical\ Rules|} D_{Org_i} \oplus D_{San_i}$

Ghost Rules = $\frac{((|R_{San}| - |R_{Total} \cap R_{San}|))}{|R_{San}|}$

Rule Hiding Failure (RHF) = $\frac{|R_{San}|}{|R_{Sen}|}$

Indian Journal of Science and Technology Vol. 5 No. 6 (June 2012) ISSN: 0974- 6846

Table 1. Results of various datasets

Data Set	Rules and Accuracy	Decision Table	Decision Tree	Rules	Optimized Rules
Diabetes	No of Rules	32	39	13	8
	Accuracy (%)	71.22	73.82	75.26	80.51
Sonar	No of Rules	15	35	8	7
	Accuracy (%)	69.23	71.15	80.28	85.71
Zoo	No of Rules	15	17	8	7
	Accuracy (%)	86.13	92.07	92.07	100
Splice	No of Rules	38	45	17	10
	Accuracy (%)	76.21	84.56	89.47	94.62
Landsat_ train	No of Rules	43	24	24	16
	Accuracy (%)	79.85	93.69	93.69	99.12

Results and discussion

For mitigating the effectiveness of the framework, various experiments have been carried out using various data sets of varying sizes, instances and attributes accessed from the source (Mlearn databases, 2011). Results are presented in Table (1, 2, & 3) and Fig. (6, 7, 8, 9 & 10). The achieved results are very encouraging which clarify the attainment of privacy and hence, the sensitive rules have been hidden successfully.

Table 2. Properties of the datasets used

Data Set	No. of Instances	No. of Attributes
Diabetes	786	9
Sonar	208	61
Zoo	101	18
Splice	3190	62
Landsat_train	6435	36

Table 3. Experimental results

Dataset	No.of Ghost Rules (%)	Rules Hiding Failure Ratio (%)	Rules Privacy (%)
Diabetes	0.01	0.00	99.84
Sonar	0.00	0.01	100
Zoo	0.00	0.00	100
Splice	0.03	0.04	99.56
Landsat_train	0.05	0.07	98.79

Conclusion and future work

This paper is an extended version of our previous research (Qureshi *et al.,* 2010). We have preserved privacy during optimized rules extraction by using GA based approach for the security and sanitization of classification rules and encouraging results were achieved. Although much of the contribution have been made towards privacy preserving techniques, however due to the complex, important and sensitive nature of the problem, further research for the re-exploration and refinement of the existing techniques is needed. Combination of various security techniques for enhancing privacy of the proposed architecture is a choice for future work.

References

1. Liu B, Abbass HA and Mckay B (2003) Classification rule discovery with ant colony optimization. *Proc. IEEE/WIC Intl. Conf. Intelligent Agent Technol.* pp: 83-88.
2. Chieh (2009) Privacy preserving association rules by using greedy approach. *Proc. 2009 World Congress on Comput. Sci. & Information Engg.* IEEE. pp: 61-65.
3. Decai Huang and Lingli Wang (2006) Analysis on the drawbacks of the commonly used measures of the significance of attributes in decision table and a new measure. *Proc. IEEE 1st Intl. Multi-Sym. Comput. & Computational Sci. (IMSCCS).*
4. Decision tree, Wikipedia, http://en.wikipedia.org/ wiki/ Decision_tree, Retrieved Aug 02, 2009.
5. Efraim turban and Jay Eronson (2003) Decision support systems and intelligent systems. *6th Ed.,* New Dehli: *Prentice -Hall of India (P) Ltd*.
6. Genetic Algorithm (2012) http://encyclopedia2.thefree dictionary.com/ genetic algorithm, Accessed on Jan 06, 2012.
7. Naumov GE (1991) NP-Completeness of problems of construction of optimal decision trees. *Soviet Phys.* Doklady, 36(4), 270-271.
8. Hyafil and Ronald L Rivest (1976) Constructing optimal binary decision trees are NP-complete. *Info. Proces. Lett.* 5(1), 15-17.
9. Lin J and Storer LA (1994) Design and performance of tree structured vector quantizers. *Info. Proces. & Managnt.* 30(6), 851-862.
10. Cox LA and WK YQ (1989) Heuristic least-cost computation of discrete classification functions with uncertain argument values. *Annals of Operations Res.* 1(1), 1-30.
11. M. Shuaib Qureshi, M, I Saeed and S. M Saqlain (2010) Proposed architectural model for optimal transformation of decision table and decision tree into knowledge base. *Indian J. Sci. & Technol.* 3(3), 362-364.
12. Naderi M, (2009) A novel method for privacy preserving in association rule mining based on genetic algorithms. *J. Software.* 4(6), 555-562.
13. Nguyen, Perkins T, Laffey W and Pecora T (1987) Knowledge base verification. *AI Magazine*. pp: 69-75.
14. Murphy OJ and Mccraw RL (1991) Designing storage efficient decision trees. *IEEE Trans. Comput.* 40(3), 315-319.