Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 15 Issue 2 Version 1.0 Year 2015
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

P^{2}DM-RGCD: PPDM Centric Classification rule Generation Scheme

By S Kumara Swamy, Manjula S H, K R Venugopal & L M Patnaik
UVCE, Bangalore University, India

Abstract- In present day applications the approach of data mining and associated privacy preservation plays a significant role for ensuring optimal mining function. The approach of privacy preserving data mining (PPDM) emphasizes on ensuring security of private information of the participants. On the contrary majority of present mining applications employ the vertically partitioned data for mining utilities. In such scenario when the overall rule is divided among participants, some of the parties remain with fewer rules sets and thus the classification accuracy achieved by them always remain questionable. On the other hand, the consideration of private information associated with any part will violate the approach of PPDM. Therefore, in order to eliminate such situations and to provide a facility of rule regeneration in this paper, a highly robust and efficient rule regeneration scheme has been proposed ensures optimal classification accuracy without using any critical user information for rule generation. The proposed system developed a rule generation function called cumulative dot product (P2DM-RGCD) rule regeneration scheme. The developed algorithm generates two possible optimal rule generation and update functions based on cumulative updates and dot product. The proposed system has exhibited optimal response in terms of higher classification accuracy, minimum information loss and optimal training efficiency.

Keywords: *data mining, privacy preserving, vertical portioning, rule regeneration.*

GJCST-C Classification : *H.3.4*

Strictly as per the compliance and regulations of:

P²DM-RGCD: PPDM Centric Classification rule Generation Scheme

S Kumara Swamy[α], Manjula S H[σ], K R Venugopal[ρ] & L M Patnaik[ω]

Abstract- In present day applications the approach of data mining and associated privacy preservation plays a significant role for ensuring optimal mining function. The approach of privacy preserving data mining (PPDM) emphasizes on ensuring security of private information of the participants. On the contrary majority of present mining applications employ the vertically partitioned data for mining utilities. In such scenario when the overall rule is divided among participants, some of the parties remain with fewer rules sets and thus the classification accuracy achieved by them always remain questionable. On the other hand, the consideration of private information associated with any part will violate the approach of PPDM. Therefore, in order to eliminate such situations and to provide a facility of rule regeneration in this paper, a highly robust and efficient rule regeneration scheme has been proposed ensures optimal classification accuracy without using any critical user information for rule generation. The proposed system developed a rule generation function called cumulative dot product (P²DM-RGCD) rule regeneration scheme. The developed algorithm generates two possible optimal rule generation and update functions based on cumulative updates and dot product. The proposed system has exhibited optimal response in terms of higher classification accuracy, minimum information loss and optimal training efficiency.

Keywords: data mining, privacy preserving, vertical portioning, rule regeneration.

I. Introduction

In present day scenario the data mining techniques are playing very significant role for ensuring optimal data exploration, classification and further decision support systems (DSS). In numerous applications the process of data mining is having great significance such as search engines and DSS mechanisms for business houses, organizations and government agencies etc. On the other hand due to multi-party computation or communication scenario, the requirement of a robust privacy factor is realized. A number of researches are going on to ensure private data security in secure multiparty computation (SMC) scenario based mining facility. The newly proposed paradigm called Privacy preservation in data mining (PPDM) is one of the growing research sector where a number of approaches have been proposed and optimized for optimal and secure mining process. In order to achieve an optimal and secure mining facility, data distribution approaches such as vertical partitioning and horizontal data partitioning has been advocated. The systems based on vertically partitioned data are emerging due to its robust function and classification accuracy. On the other hand based on association rule mining a number of systems have been developed. In our previous research [1][2][3] we have already implemented numerous noble schemes to optimize data classification and performance efficiency and a robust privacy preserving data mining scheme using commutative RSA scheme. These all system has in fact exhibited optimal performance for classification efficiency and effective mining function. But taking into consideration of a scenario, where in vertically partitioned data the rules generated have to be divided among encompassing participants, there could be a possibility that some of the parties might have fewer rules.

When certain party possesses low rules count, the classification accuracy based on those confined rules might give lower accuracy. Therefore to ensure optimal classification accuracy and efficiency rules are required to be increased with enhanced information and classification attributes. On the contrary, in privacy preserving data mining (PPDM) scenario, no other party will like to share its critical, private information with other and if it takes place the PPDM itself will be violated. Therefore in such circumstances, the implementation of such approach which can ensure rule enhancement or rule regeneration without retrieving critical information of other participant will be required. In order to achieve this goal, here in this research paper, we have proposed a highly robust and efficient system model for rule regeneration which considers only some of the numerical attributes for rule regeneration and operates with two mathematical and logical operators. In this paper a rule regeneration scheme called cumulative dot product (P²DM-RGCD) has been proposed. The proposed scheme individually generates two distinct functions for rule regeneration on the basis of cumulative and dot product based rule updates. The, overall functions and rule regeneration schemes have been developed employing only some of the numerical attributes associated with other parties so as to perform rule regeneration. The considered numerical attributes even doesn't disclose the private or critical information related to parties. Thus, the proposed approaches of rule regeneration not only ensure the preservation of

Author α σ ρ : Department of Computer Science and Engineering University Visvesvaraya College of Engineering, Bangalore University, Bangalore. e-mail: kumar.aruna@gmail.com
Author ω : Indian Institute of Science, Bangalore, India.

privacy but also make the mining system optimized in terms of higher classification accuracy and efficiency. The proposed system has exhibited optimal response in terms of higher classification accuracy, minimum information loss and optimal training efficiency.

a) Motivation

Now days most of the researches are done for privacy preservation in mining but very few have made effort to ensure optimal performance with PPDM as most schemes employ the approach of vertically partitioned data for classification and in case of SMC scenario, the requirement of privacy preserving data mining is also inevitable. The situation also becomes complex in the scenario, where the overall rules are required to be split amongst allied participants. In this case, some of the parties remained back with low rule counts and therefore their classification accuracy is always under suspicion. Therefore, considering this requirement the development of certain optimized rule enhancement or even rule regeneration scheme can be a potential solution. A model based on PPDM is needed where an algorithm will regenerate the rules based on some numerical attributes using some operators which neither use critical information of other users not violates PPDM objective. These all motivate us to develop a robust rule regeneration scheme that can exhibit rule regeneration without causing any violation in PPDM. The proposed cumulative dot product (P²DM-RGCD) scheme can deliver these all expectations as it doesn't employ critical information of user, rather considers only some numerical attributes have not much information. The mathematical function (dot product and cumulative updates) makes the system more robust in function.

b) Contribution

The proposed rule regeneration approach, cumulative dot product (P²DM-RGCD) scheme possesses potential to optimize rule regeneration on the basis of some numerical attributes assocaited with any participants in MPC scenario. The proposed system performs well using two possible functions, in terms of cumulative rule generation updates and dot product based rule generation. Such combinations emerge out with enhanced rule generation efficiency, classification accuracy, minimum information loss and higher training efficiency. The developed system has been tested with varied datasets of varying sample size and the results obtained has exhibited that the proposed system can play a significant role for real time mining applications.

c) Organization

The remaining manuscript has been classified into certain sections where Section II represents related Work which are followed by research background. In Section III the proposed system has been discussed which is followed by results and analysis in Section IV. Section V presents conclusion. The references used are given at the last of presented manuscript.

II. Related Work

A number of researches have been done for PPDM oriented rule generation and performance optimization. Some of the work carried for PPDM and rule generation based mining enhancements are as follows:

Dehzangi, O. [4] Advocated on the application of fuzzy rule based systems and discusses the limitations in terms of rule-base generation and stated that in case of higher dimensional issues, not every possible rule can be generation correspondence with entire antecedent combinations. Ultimately authors proposed a rule generation approach using data mining and focused their system to accomplish rule-based generation with varied length. In [2] M.W. Kim et al. developed an effective fuzzy based rule generation scheme using fuzzy decision tree data mining approach and they combined the clarity of rules generated on the basis of decision tree approaches like ID3 and C4.5 enriched with presentative ability of fuzzy sets that facilitated better classification for varied patterns associated with non axis-parallel decision boundaries that is in fact intricate for implementation employing attribute-based classification scheme. Sabu, M.K. et al., [6] analyzed a recent scheme called Rough Set Theory (RST) and stated it as a system with ambiguity and insecurity. In fact RST is significant for various applications but cannot incorporate association rules that plays significant rule for data mining while ensuring association among varied attributes. To eliminate such limitation the author advocated a rough set based scheme for rule generation using an incoherent information model comprising preprocessed data and used LEM2 algorithm to perform rule generation. Ji Dan et al., [7] presented a data mining scheme called CA to enhance CURE and C4.5 and uses principle component analysis (PCA), parallel processing and grid partitioning to perform better feature and scale reduction for huge datasets. Trincă, D. et al., [8] emphasized on rule mining based PPDM and proposed an algebraic and recursive system based on two party protocols and focussed on collusion free mining still. In our last paper [9] we accomplished data mining while incorporating multiple parties and performed mining on vertically partitioned data and proposed a scheme called Key Distribution-Less Privacy Preserving Data Mining (KDLPPDM). To ensure security they employed Commutative RSA an advanced cryptosystem. Tran, D.H. et al., [10] proposed CRYPPAR scheme that facilitates a robust framework for privacy preserving association rule mining based on cryptosystem schemes. The authors employed secure scalar product algorithms for exhibiting efficient data mining with enhanced accuracy. Modi, C.N. et al., [11] proposed a noble heuristic scheme called decrease support of R.H.S. item of rule clusters that facilitates privacy for

perceptive rules at definite level while assuring optimal quality or mining efficiency for datasets. They performed clustering on the sensitive association rules based on defined conditions and perform rule hiding by means of some modifications. This is the matter of fact the some of the existing approaches have illustrated better results but unfortunately, no emphasis has been mode of system optimization using PPDM approach with rule regeneration without exploring critical information of associated parties. Some works either focuses on PPDM or classification accuracy, but for robust applications, the duo are needed to be enriched together.

III. Background Work

PPDM is one of the recent and most emerging technologies for data mining filed. This technology facilitates a novel framework for performing data extraction and classification with ensured security and preservation among various or multiple parties. In [1] a noble system model for PPDM has been developed for vertically partitioned data and authors have employed a robust cryptosystem to ensure data security in SMC environment. Commutative RSA scheme has been used for privacy preservation. Similarly in [9] the emphasis was made on classification accuracy. In uniqueness of this work was that this algorithm didn't employ any private data and in spite it came up with better association rule mining. This system came up with better efficiency in terms of rule generation, overhead minimization and classification efficiency.

IV. Proposed System

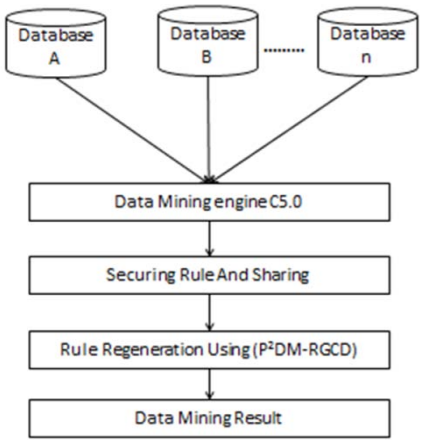

Figure 1 : PPDM-RGCD System Architecture

Taking into consideration of a highly robust and efficient system for privacy preserving data mining that employs multiple parties with vertically partitioned data and in which the rules associated with certain party defines the accuracy and performance, here in this research paper, an optimal solution of rule regeneration and performance optimization has been developed. The Prime objective of this research work is to develop a rule regeneration scheme without employing any critical information of allied parties. In this paper a noble cumulative dot product (P²DM-RGCD) scheme has been proposed for rule regeneration with the party having fewer rules, without retrieving any critical information of other parties can exhibit higher rule generation, resulting into better classification accuracy and efficiency. The proposed scheme has been discussed in this section.

a) PPDM Oriented Rule Regeneration Scheme

As in vertical partitioned data the overall rules are shared among the participating parties and thus it raises the probability where the rules available with certain party could be very few and on that basis the classification accuracy could not be optimal. On the other hand, sharing the private information about other parties to retrieve better classification for certain party is violation for privacy preservation. Therefore in such situation, in this paper rule regenerates have been done based on certain numerical attributes. In order to accomplish an optimized rule regeneration approach for PPDM applications, a classifier having mapping function $g(a)$ with $a \in A$into labels$b \in \{-1, 1\}$, are taken into consideration and is given by

$$g(a) = sgn[p(a)] \quad (1)$$

Where $p : A \rightarrow C$ states a real-valued predictor. The optimal predictor $p(a)$ represents the reduction factor for the classification problem. Mathematically,

$$C(p) = D_{A,B}\{M[bp(a)]\} \approx \sum_i M[b_i p(a_i)] \quad (2)$$

In above equation$\mathrm{D}(.)$represents the function for loss reflecting the upper bound as the error rate. The process of rule regeneration approach performs learning and approximating the best suitable predictor in terms of a linear combination of generic predictors$\mathrm{g\ n : A \rightarrow C}$. In fact this is stated as the party possessing lower classification rules and information to exhibit optimal classification. Thus, the reduction of error minimization function can be updated as

$$p(a) = \sum_n x_n hn(a) h \in H \quad (3)$$

Where $H = \{h_1, .., h_m\}$ refers for the combination or cluster of those all parties which don't have enough rules or information to perform classification. In order to achieve formulation consistency or regularity here we have considered that $h(a) = 0\ and h(a) = 1$are the parts of cluster or group of parties (H) having less rules. The process of rules regeneration performs learning of the linear combination by employing certain successive descent approach in functional space, in reference to the enhancement issues,

$$\begin{cases} min_{p(a)} C(p) \\ s.tp(a) \in \varphi_h \end{cases} \quad (4)$$

In the above presented expression, the variable H represents the cluster of parties possessing fewer rules for classification. Here it must be taken into consideration that Hrepresents a convex set and in case the parameter $C(.)$depicts certain convex function then in that situation the enhancement issues for equation (3) will always be a convex.Now, taking into consideration of the first and second derivatives of the loss function as employed with initial problem (Equation 2) is stated by

$$M' = \frac{\partial M(q)}{\partial q} \text{ and } M'' = \frac{\partial^2 M(q)}{\partial q^2} \quad (5)$$

Here it is assumed that, performing k iterations the value of the optimal predictor is given by a function$p^n(a)$.Now employing Taylor series expansion for $C(p^n + h)$towardsp^n, the resulting first and second derivatives can be retrieved by the following approach.

$$\delta C = \frac{\delta C(p^n + \gamma h)}{\partial \gamma}\Big|_{\gamma=0} \quad (6)$$

It provides the estimation of variations in C at instant p^n towards g.To make it simple, in many cases the argument $a$has been omitted all through the presented manuscript. The variations or warp of $C[p^n]$ towards $g$can be presented in terms of its second order derivative. Mathematically the curvature can be defined as

$$\delta^2 C(p^n; h) = \frac{\delta^2 C(p^n + \gamma h)}{\partial \gamma^2}\Big|_{\gamma=0} \quad (7)$$

Now, taking into consideration of these variables, the approximation of C can be accomplished using Taylor series expansion in the region of since$C[p^n]$, given below, solely depend on the illustrative value A through the training eventsa_i and there doesn't exist any loss while mapping the value of defined functions $p^n(a)$ into certain definite vector form $Q \in C^e$ of $[p^n(a_1), \ldots, p^n(a_e)]$, where $n$states for the size of the training set.

$$C(p^n + \beta g) = C(p^n) + \beta \delta C(p^n; h) + \frac{\beta^2}{2}\delta^2 C(p^n; h) + E(\beta^2) \quad (8)$$

In the defined vector Q the dot product presentation can be given by

$$\langle h_1, h_2 \rangle = \sum_i h_1(a_i), h_2(a_i) \quad (9)$$

Without causing any loss in generality, it is assumed that those parties who have fewer rule for classification$(h \in H)$, are in general processed for normalization to get$\langle h, h \rangle = 1$. Now, considering a function called unitary indication function given by$K(a) = 1$in case the variable $a$holds otherwise it possesses zero and the functional *gradient* of $C$is given by certain vectorized entities given by

$$\nabla C_{(p^n)}(a_i) = \frac{\partial}{\partial \gamma} C[p^n + \gamma K(a = a_i)]|_{\gamma=0} = \frac{\partial M[b_i, p^n(a_i) + \gamma]}{\partial \gamma}\Big|_{\gamma=0} \quad (11)$$

Meanwhile, the second order gradient vector is given as Hessian in the form of a matrix, mathematically it is presented as

$$\nabla^2_{C_{(p^n)}}(a_i, a_j) = \frac{\partial^2 C(p^n + \gamma_1 K(a = a_i) + \gamma_2 K(a = a_j))}{\partial \gamma_1 \partial \gamma_2}\Big|_{\gamma_1, \gamma_2 = 0} \quad (12)$$

$$= \begin{cases} \frac{\partial^2}{\partial \xi \gamma^2} M[b_i p^n(a_i) + \gamma]\Big|_{\gamma=0} & ipi = j \\ 0 & otherwise \end{cases}$$

To make is simple, it is presented further as

$$\nabla^2_{C_{(p^n)}}(a_i, a_j) as \nabla^2_{C_{(p)}}(a_i).C \quad (13)$$

$$\nabla^2_{C_{(p^n)}}(a_i) \geq 0, \forall p^n, \forall a_i$$

Now taking into consideration of the convex problem and parties to be enriched with rule regeneration, its projects comes out to be

$$\partial C(p^n; h) = \frac{\partial}{\partial \gamma} \sum_j M[b_j, p^n(a_j) + \gamma h(a_j)]\Big|_{\gamma=0} \quad (14)$$

$$= \sum_j h(a_j) \frac{\partial}{\partial \gamma} M[b_i, p^n(a_j) + \sigma)]|_{\sigma=0} = \langle \nabla_{C_{(p^n)}}, h \rangle$$

Ultimately, the first order derive has been obtained as following equation (Equation 15) and similarly the second derivate has been obtained in terms of Equation (Equation 17).

$$\sum_i b_i h(a_i) M'[b_i p^n(a_i)] \quad (15)$$

$$\delta^2 C(p^n; \gamma h) = \frac{\delta^2 C(p^n + \beta h)}{\partial \gamma^2}\Big|_{\gamma=0} \quad (16)$$

$$\sum_i g^2(x_i) L''[y_i f^k(x_i)] \quad (17)$$

Considering the above derived expressions, it can be found that the learner party is required to be enriched with the predictor or rule generator at its $n + 1$iteration and it is given by

$$h^* = \frac{\arg\min \delta C(p^n; h)}{h \in H} \quad (18)$$

In case of the consideration of the gradient factor, it can also be represented by

$$h^* = \frac{\arg max\, [\delta C(p^n; h)]^2}{h \in H \quad \delta^2 C(p^n; h)} \quad (19)$$

With a bet rule generator or party seeking rule regenerator $h*$, the best possible step size can be presented by

$$x^* = \frac{\arg\min C(p^{nk}; xh^*)}{x \in C} \quad (20)$$

Similarly, the values evaluated by predictor is updated for$n + 1$iteration events and it is given by

$$p^{n+1}(\mathrm{a}) = p^n(\mathrm{a}) + x^* h^*(a). \quad (21)$$

Thus, the ultimate predictor for rule regeneration will function as a linear combination of those parties who have fewer rules for classification.

a) Rule Regeneration: An Optimal Blend for new Features or Rules

This is the matter of fact that the rule regeneration selects some helpful features to perform classification; the genuine combination of parties seeking rule regeneration might not be so enriched for capturing every associated attributes of information to perform discrimination. As for illustration, it becomes necessary to employ certain conjunctions of the features for capturing few of the dimensions and in such conditions the linear combination becomes ineffective where even rule generation is not optimal solution to accomplish better classification. In this paper a mechanism of rule regeneration has been developed to ensure optimal privacy preservation and effective classification in terms of accuracy and optimal mining results. The proposed *cumulative dot product P²DM_RG-CD*scheme has been discussed as follows.

b) Cumulative Dot Product (P2dm_Rg-Cd) Based Rule Regeneration

Consider cluster of combination of parties having lower rules count$G$, the proposed noble scheme of rule generation called cumulative dot product (P²DM-RGCD)based rule generation emphasizes its function for solving the following problem

$$\begin{cases} min_{p(a)C(p)} \\ s.tp(a) \in \alpha_G^{CD} \end{cases} \quad (22)$$

Where α_H^{CD} represents the combination of all comprising set of achievable linear combinations of the dot product of parties, mathematically given as

$$\alpha_H^{CD} = \left\{ g(a) | g(a) = \sum_j \prod_l h_j k(a), h_j k \in H \right\} \quad (23)$$

Here it can be found that α_H^{CD} states for a convex combination and therefore for any similar functions$C(.)$, the optimization issues will be convex. Now, considering Taylor boost scheme as a comparative model, assume that after n iterations the predictor possesses m terms given by

$$p^n(a) \sum_{j=1}^m r_j^n(a) \quad (24)$$

The every such presentation will represent an unitary learner party and will be given by,

$$r_j^n(a) \prod_{k=1}^{t_j} h_j, k(a)\, h_j, k(a) \in H \quad (25)$$

At certain iteration $n + 1$ it is feasible to enhance $p^n(a)$possessing dual updates given by cumulative and dot product. A brief of the considered paradigms have been given as follows:

- *Cumulative Update:* In case of cumulative update paradigm it is considered that selecting or joining a learner party to the predictor will be like

$$p^{n+1}(a) = p^n(a) + h(a). \quad (26)$$

Here, the updates are done on the basis of rules regenerated with first and second derivatives factors given by $\delta C(p^n; h)$ and $\delta^2 C(p^n; h)$respectively. And the optimal partyh_0^*, can be achieved based on the selection of gradient descent approach. Here in terms of optimal step sizex_0^*, the newly generated rules or predictor is found to be with risk factor

$$\hat{C}_0 = C(p^n x_0^* h_0^*) \quad (27)$$

- *Dot product rule update:* In case of *P²DM_RG_DC)* , one of the available terms is processed for multiplication using a newer party given by

$$r_s^{n+1}(a) = r_s^n(x) \times g(a). \quad (28)$$

It can also be given by

$$p_s^{n+1}(a) = r_s^n(a)h(a) + \sum_{j \neq s} r_j^n(a) \quad (29)$$

$$= p^n(a) - r_s^n(a)] + r_s^n(a)h(a) \quad (30)$$

$$F_s^n(a) + r_s^n(a)h(a) \quad (31)$$

$$\text{With } F_s^n(a) = p^n(a) - r_s^n(a). \quad (32)$$

Now, taking into consideration of the above mentioned expressions a Taylor series expansion of $C(p^{n+1})$can be retrieved in the region of functional$F_s^n(a)$, and the first and second order variations for the risk factor wrt a dot product cum cumulative update of the s^{th} term in $p^n(a)$ is given by

$$\delta C(p^n; h, s) = \frac{\partial C[F_s^n + \omega r_s^n h]}{\partial \omega}\Big|_{\omega=0} \quad (33)$$

$$= \sum_i b_i h(a_i)\, r_s^n(a_i)]^2 M'' [b_i F_s^n(a_i)] \quad (34)$$

$$\delta^2 C(p^n; h, s) = \frac{\delta^2 C[F_s^n + \omega r_s^n h]}{\partial \omega^2}\Big|_{\omega=0} \quad (35)$$

$$= \sum_i [h(a_i)\, r_s^n(a_i)]^2 M''[b_i F_s^n(a_i)] \quad (36)$$

Pseudo Algo: Cumulative dot product (P²DM-RGCD) Based Rule Regeneration

Input: Sets for data training S_t, parties with lower rules $H = \{h_1, .., h_m\}$,
Iteration counts Z and a loss function $M(.)$.
Initialization: Select $n = 0, m = 0,$ $r_m^n(a) = 0$ and $p^n(a) = 0$
while $n < Z$ **do**
Estimate optimal Cumulative update $x_0^* h_0^*$
Select $\hat{C}_0 = C(p^n + x_0^* h_0^*)$
Initialize look
for $s = 1$ to m **do**
Estimate optimal update for s^{th} dot product term, $x_s^* h_s^*$
Update $\hat{C}_0 = C(p^n - r_s^n) + r_s^n x_s^* h_s^*]$
end for
Select $s^* = \arg min_s \hat{C}_s s = 0, \ldots, m,$
if $s^* = 0$ **then**

$$r_{m+1}^{n+1} = x_0^* h_0^*$$
$$m = m + 1$$

else

$$r_{s*}^{n+1} = r_{s*}^n \quad s \neq s*$$
$$p^{n+1}(a) = \sum_{s=1}^{m} r_s^{n+1}(a)$$
$$n = n + 1$$

end while
Output: Rule generated: sign[p^Z(a)]

The optimal party seeking rule regeneration given by g_r^* is retrieved with its optimal step size and it can be given as

$$x_s^* = \frac{\arg\min C(F_s^n + x h_s^*)}{x \in C} \quad (37)$$

Still, the updated rule generator or predictor does possesses the risk factor given by

$$\hat{C}_s = C(F_s^n x_s^* h_s^*) \quad (38)$$

The proposed and developed algorithm of cumulative dot product (P²DM-RGCD) based rule regeneration provides optimal new rules for those parties who don't have sufficient rules for classification due to vertically partitioned data and divided rules sets among other parties. Thus, implementing the above mentioned paradigms cumulative update and dot product cum cumulative update has exhibited higher rule generation without extracting any critical information assocaited with the other parties in the application scenario and thus it also preserves the privacy of participant. The enriched rules or regenerated rules make the system highly robust for optimal classification.

V. Results and Analysis

In a specific situation of SMC based PPDM where a number of participants do exhibit data mining without any disclosure of its private data or information with vertically partitioned data the splitting of rules for classification might cause a situation where some of the participants will have fewer numbers of rules that could result into inaccuracy, error prone and inefficient classification. In such cases even the other parties don't wish to share its information. In this paper a robust rule regeneration technique has been developed that exhibits rule regeneration using a noble scheme called cumulative dot product (P²DM-RGCD) without using significant information of other participants. C# and C++ programming languages was used for development. The model was implemented with GCC compiler on Linux platform and the system effectiveness has been analyzed in terms of its learning accuracy, testing accuracy, information loss etc. In order to exhibit the performance analysis with varied datasets or data count, various data samples like breast cancer data, diabetes datasets, satellite datasets etc have been considered. The results have been analyzed in terms of its specificity Vs sensitivity, performance rate, higher accuracy and minimum computational overheads information loss, classification accuracy and many more. Following figures represent the receiver operating characteristics (ROC) analysis for the developed research model. The results obtained in this paper have been compared with our previous work [1][2][3].Figure 1-3, illustrates the performance of the proposed system with employed breast cancer data of varied size. Here it can be found that the proposed system response is better as compared to existing vertically partitioned mining model with PPDM.

a) Performance Analysis for Breast Cancer datasets

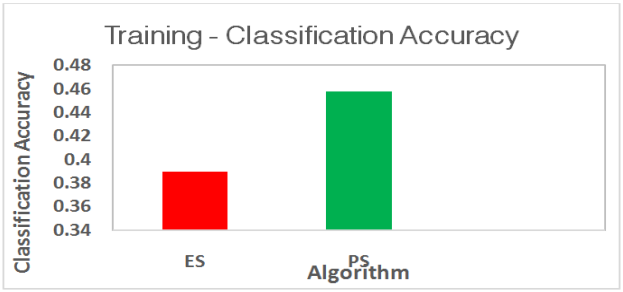

Figure 2 : Performance analysis for classification accuracy for Breast Cancer data

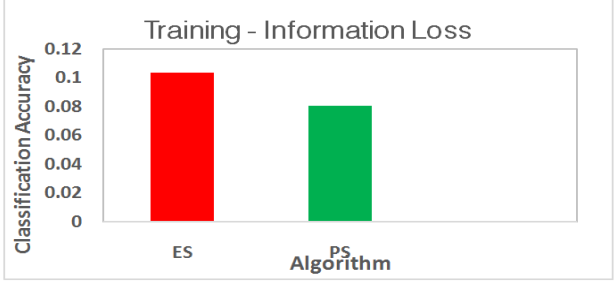

Figure 3 : Analysis for Information loss for Breast Cancer datasets

From Figure 2, Figure 4 and Figure 6, it can be found that the proposed system facilitates minimum information loss. The reason behind this achievement is that the proposed system does not employ the critical information associated with any participants. In order to exhibit rule regeneration, our proposed system has just employed some of the numeric values or parameters on basis of which processing with proposed cumulative dot product (P²DM-RGCD)scheme, the classification has been accomplished. Thus, the least utilization of critical information makes this system capable of delivering higher classification accuracy and performance (Figure1, Figure 4, and Figure 7) without causing much information loss as compared to existing systems.

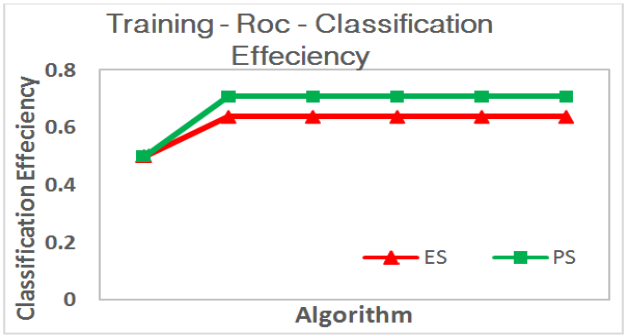

Figure 4 : ROC Analysis for Classification accuracy for Breast Cancer datasets

b) Performance Analysis for Diabetes datasets

The results obtained for diabetes datasets are as follows:

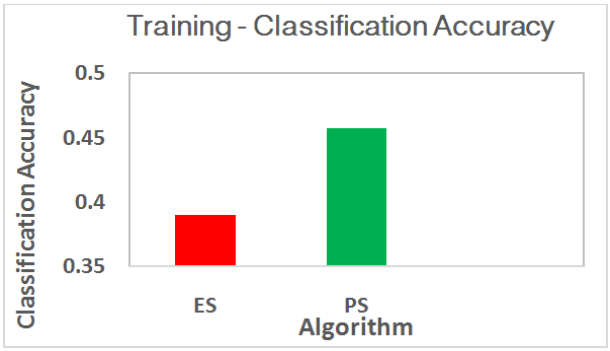

Figure 5 : Performance analysis for classification accuracy for Diabetic dataset

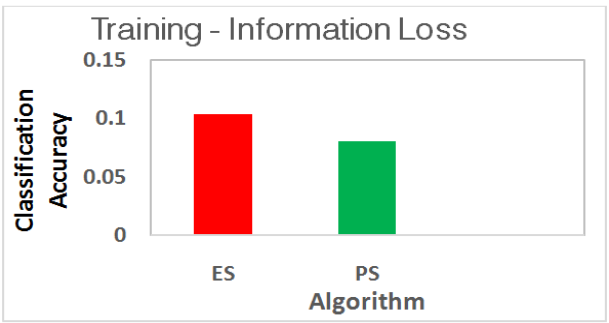

Figure 6 : Analysis for Information loss for Diabetes datasets

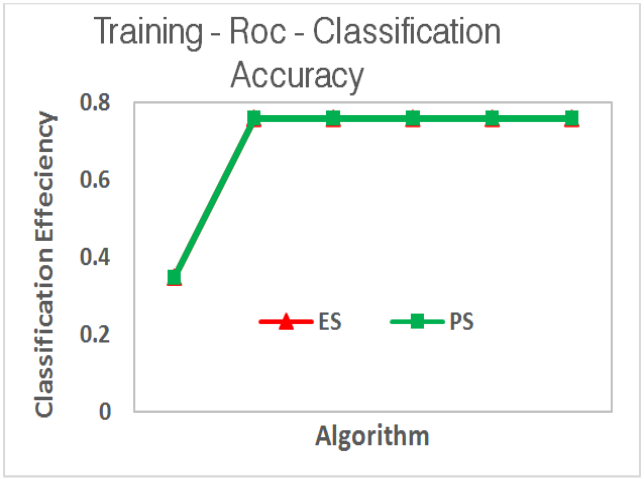

Figure 7 : ROC Analysis for Classification accuracy for Diabetes datasets

c) Performance Analysis for Satellite datasets

The results obtained for diabetes datasets are as follows:

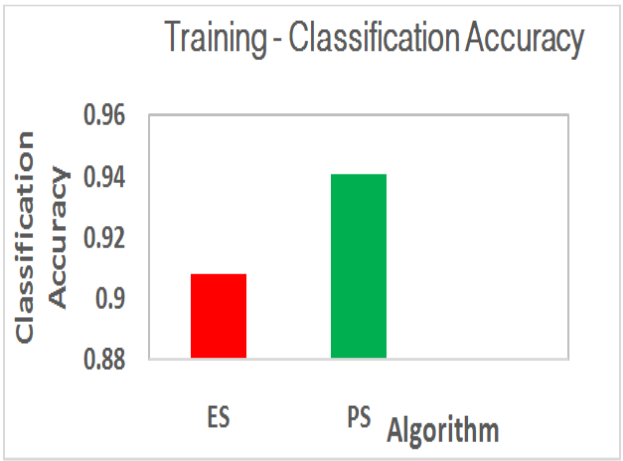

Figure 8 : Performance analysis for classification accuracy for Satellite datasets

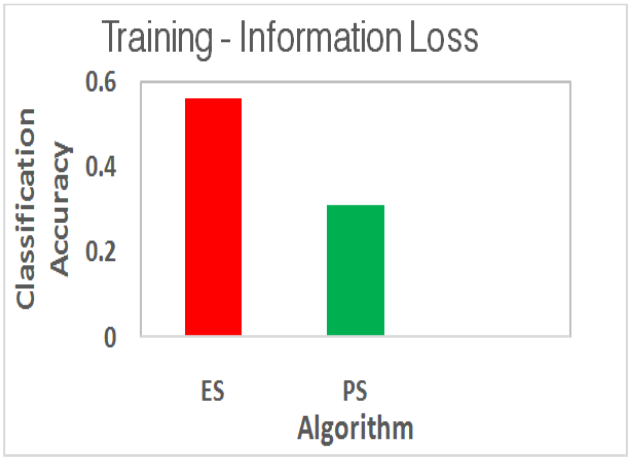

Figure 9 : Performance analysis for Information Loss for Satellite datasets

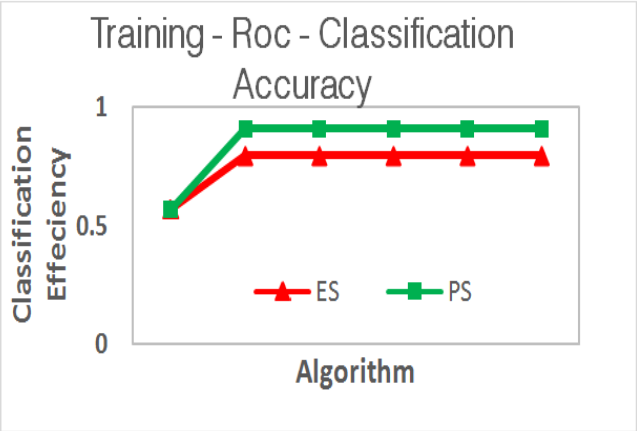

Figure 10 : ROC Analysis for Classification accuracy for Satellite datasets

The results illustrated depicts that the proposed system exhibits optimal classification with varied data size and sample space. The proposed system has exhibited better results in terms of higher classification accuracy in training, minimum information loss and optimized classification accuracy, which do satisfy every aspects of a robust privacy preserving data mining applications.

VI. Conclusion

The key requirement of an effective and robust data mining system is its security or data privacy with every participating users and optimal mining efficiency. In majority of mining applications vertically partitioned data are used predominantly. In case of vertically partitioned data along with the assurance of privacy preserving in data mining, creates a situation where the rules generated are divided among parties and then certain parties remain back with fewer rules. In such circumstances, on the basis of low classification rules, the accuracy and efficiency of mining is questionable. Considering this need to generate more rules in this paper a rule regeneration scheme was proposed which not only avoids the utilization of private information allied with other parties but also enhances the classification accuracy without any computational overheads.

The developed system *dot* cumulative dot product (P²DM-RGCD) has exhibited rule regeneration with two possible rule generation functions *called* cumulative rule updates and dot product rule update. Using the derived functions the rule regeneration has been accomplished that makes this system highly robust to generate accurate and precise outcomes and classification accuracy. The developed system has exhibited better results in terms of its training performance, optimal classification accuracy and minimum information loss. The performance of the developed system may ensure the optimal performance with real time mining applications which needs privacy preserving as well as optimal classification accuracy. The further evaluation and enhancement of the system can be done for Big Data applications and online web utilities.

References Références Referencias

1. S Kumara Swamy; Manjula S H; K R Venugopal; Iyengar S S; L M Patnaik; "A NOVEL PPDM PROTOCOL FOR DISTRIBUTED PEER TO PEER INFORMATION SOURCES"; International Journal of Computer Engineering and Technology (IJCET); Volume 4; Issue 4; July-August (2013).
2. S Kumara Swamy; Manjula S H; K R Venugopal; Iyengar S S; L M Patnaik; "Key Compromise Resilient Privacy Provisioning in Vertically Partitioned Data", In Cyber Times International Journal of Technology and Management,ISSN:2278-7518, vol. 6, no. 1, pp.18-29, March 2013.
3. S Kumara Swamy; Manjula S H; K R Venugopal; Iyengar S S; L M Patnaik; "A Data Mining Perspective in Privacy Preserving Data Mining Systems"; International Journal of Current Engineering and Technology, Vol.4, No.2 (April 2014).
4. Dehzangi, O.; Zolghadri, M.J.; Taheri, S.; Fakhrahmad, S. M., "Efficient Fuzzy Rule Generation: A New Approach Using Data Mining Principles and Rule Weighting," Fuzzy Systems and Knowledge Discovery, 2007. FSKD 2007. Fourth International Conference on , vol.2, no., pp.134,139, 24-27 Aug. 2007.
5. Myung-Won Kim; JoongGeun Lee; Changwoo Min, "Efficient fuzzy rule generation based on fuzzy decision tree for data mining," Fuzzy Systems Conference Proceedings, 1999. FUZZ-IEEE '99. 1999 IEEE International , vol.3, no., pp.1223,1228 vol.3, 22-25 Aug. 1999.
6. Sabu, M. K.; Raju, G., "Rule induction using Rough Set Theory — An application in agriculture," Computer, Communication and Electrical Technology (ICCCET), 2011 International Conference on , vol., no., pp.45,49, 18-19 March 2011.
7. Ji Dan; QiuJianlin; Gu Xiang; Chen Li; He Peng, "A Synthesized Data Mining Algorithm Based on Clustering and Decision Tree," Computer and Information Technology (CIT), 2010 IEEE 10th International Conference on , vol., no., pp.2722,2728, June 29 2010-July 1 2010.
8. Trincă, D.; Rajasekaran, S., "Towards a Collusion-Resistant Algebraic Multi-Party Protocol for Privacy-Preserving Association Rule Mining in Vertically Partitioned Data," Performance, Computing, and Communications Conference, 2007. IPCCC 2007. IEEE Internationa , vol., no., pp.402,409, 11-13 April 2007.

9. KumaraSwamy, S.; Manjula, S.H.; Venugopal, K.R.; Iyengar, S.S.; Patnaik, L.M., "Association rule sharing model for privacy preservation and collaborative data mining efficiency," Engineering and Computational Sciences (RAECS), 2014 Recent Advances in , vol., no., pp.1,6, 6-8 March 2014.
10. Tran, D.H.; Wee Keong Ng; Wei Zhao, "CRYPPAR: An efficient framework for privacy preserving association rule mining over vertically partitioned data," TENCON 2009 - 2009 IEEE Region 10 Conference , vol., no., pp.1,6, 23-26 Jan. 2009.
11. Modi, C.N.; Rao, U.P.; Patel, D.R., "Maintaining privacy and data quality in privacy preserving association rule mining," Computing Communication and Networking Technologies (ICCCNT), 2010 International Conference on , vol., no., pp.1,6, 29-31 July 2010.

This page is intentionally left blank