Geetha Peethambaran, Chandrakant Naikodi, Suresh Lakshmi Narasimha Setty

Balancing Privacy Vs Efficiency in Data Analytics using Nearest Neighbour Randomization

Abstract: ***The Digital era marked by the unrivalled growth of Internet and its services with day-to-day technological advancements has paved way for a data driven society. This digital explosion offers opportunities for extracting valuable information from collected data, which are used by organizations and research establishments for synergistic advantage. However, privacy of online divulged data is an issue that gets overlooked as a consequence of such large-scale analytics. Although, privacy and security practices conjointly determine the ethics of data collection and its use, personal data of individuals is largely at risk of disclosure. Considerable research has gone into privacy preserving analytics, in the light of Big Data and IoT boom, but scalable and efficient techniques, that do not compromise the usefulness of privacy constrained data, continues to be a challenging arena for research. The proposed work makes use of a distance-based perturbation method to group data and further randomizes data. The efficacy of perturbed data is evaluated for classification task that gives results on par with the non-perturbed counterpart. The relative performance of the algorithm is also evaluated on the parallel computing platform Spark. Results show that the technique does not hinder the use of data for holistic analysis while privacy is subjectively maintained.***

Keywords: Privacy Preserving, Analytics, Big Data, Perturbation, Performance, utility

I. INTRODUCTION

Privacy is an issue of concern in large scale analytics. Privacy preservation for healthy analytics has been a popular topic of research in the recent past, with pervasive technological advancements in the Internet of Things and Big Data arena. Analytics is widespread in a myriad of fields ranging from government, public sector, health, industry and research establishments. Data is collected from users and put to use for personalized recommendations and targeted marketing. Despite privacy guidelines issued by many of the organizations, there are no stringent regulations that realistically monitor the legal use of the collected data. It is therefore mandatory that privacy and security practices in an establishment sail together to avoid intentional misuse of data. Cloud services although advantageous, can be highly vulnerable to privacy breaches.

Techniques that impose privacy restrictions, typically affect data quality and consequently, worthwhile analytics. With data sizes and types growing rapidly, challenges in privacy preserving analytics is not restricted to data usefulness, but also has to cater to scale and efficiency.

Achieving an acceptable balance between these is demanding, since one trades the other. Therefore, defeating these competing goals of privacy preserving analytics requires leveraging the benefit of efficient parallel, faster processing that can handle large data with ease.

The remainder of the paper is organized as follows: Section II reviews the related work. Section III gives the overview of the problem with preliminary definitions, Section IV details the proposed methodology and workflow, Section V gives the functional description of the problem, with results followed by discussion and Section VI concludes the paper.

II. RELATED WORK

Privacy and analytics can work together, but the mining result of a privacy preserving design loses data quality. The privacy preserving techniques are classified into two broad categories, namely Privacy Preserving Data Publishing (PPDP) [1][2][3] and Privacy Preserving Data Mining (PPDM) [2][3][4]. The categories are based on privacy incorporation into the data life cycle process [5].A widespread review of these algorithms and modifications have been presented in [3].These papers discuss the privacy techniques from the perspective of utility on data of acceptable sizes, but scalability and efficiency are a matter of concern with Big Data. Sensitive data disclosure is more prominent with the increase in public cloud services for analytics.

Data anonymization [5] and data perturbation [5] are two major methods of modifications applied on data before it is published for. Data anonymization is a technique of hiding sensitive information of individuals before data is published [7]. k-anonymity[7], l-diversity[8],t-closeness[9] are variations of anonymization in which the aim is to ensure that the probability of an individual being identified from the released dataset is only 1/k. K- anonymity models follow different techniques such as generalization[2][7] or suppression[2][7] to create anonymous records. The research [7] shows that k-anonymity has the potential to mask large datasets, provided the value of k is intelligently chosen. Different works [10][11][12] on k-anonymity show that the proportion of information loss is directly related to the value of k. Hence optimal and heuristic techniques need to be adapted for improvising anonymity privacy models. But they are susceptible to different types of linkage attacks [2] and methods to overcome these should be catered to.

Revised Manuscript Received on October 30, 2019.
***** Correspondence Author

Geetha Peethambaran*, Department of CSE, Cambridge Institute of Technology, Bengaluru, India. Email: geetharaghuraj@gmail.com

Chandrakant Naikodi, Department of CSE, Cambridge Institute of Technology, Bengaluru, India Email: nadhachandra@gmail.com

Suresh L, Principal and Professor, Cambridge Institute of Technology, Bengaluru, India.. Email: suriaikls@gmail.com

Differential Privacy [13] is a technique of privacy preservation in which random noise is added to the dataset for the purpose of de-identification. Differential privacy is a privacy model that seeks to limit the impact of any individual subject's contribution on the outcome of the analysis [13]. It is of particular interest to big data since it can provide good privacy for large datasets. Differential privacy is a strong at protecting privacy but at the cost of degrading the utility of data.

While anonymization models focus on protecting the sensitive attributes, perturbation methods [3] are based on creating transformations on the dataset. Noise addition [14][15], randomization [15], noise multiplication [15], geometric data perturbation [16] are all different ways in which data can be distorted. Perturbed data can be synthetically modified [2], or transformations can be done that can individualistically change columns or certain attribute values. The resultant distortions help protect privacy without affecting the statistical properties of the data [16] which is required for holistic analysis. The aggregate statistics [17] help in reconstructing the data distribution [17] depending upon the kind of data. The reconstructed probability density function is helpful in analyzing the percentage of deviation that transformation has caused. This in turn can help to appropriately quantify metrics for privacy as well as utility. Utility is always traded as a result of privacy restriction. Condensation [18] and rotation [18] methods create homogenous groups [16][18] on which transformations can be applied to de-identify records or certain attributes. The work in [18] proposes an effective perturbation technique which implements privacy preservation of data streams and checks its efficiency with different types of classifiers. The proposed algorithm is tested on different types of datasets of varying dimensions. The authors also test the execution efficiency of the algorithm.

Another challenging part of privacy preserving analysis is to measure the privacy level. A number of privacy metrics have been studied and proposed [2][3][19], but identifying the correct metric for a particular scenario is demanding, given the size and variety of data. Privacy metrics are mainly classified into quality metrics and result metrics [2]. The former evaluates the level of privacy through suggested measures such as Discernability Metric [2] and Information Loss [2][3]. They can further be differentiated based on the kind of privacy technique, anonymization or perturbation. Perturbation privacy techniques are evaluated based on the normalized variance [19] in the data in comparison with the original data. The distortion introduced in the transformed data is measured through collective statistics and the differences are quantified. Greater the variance better is the privacy [19].Similarly with anonymization, generalisation and suppression is measured by calculating the amount of information loss. In the context of privacy preserving publishing, the amount of loss that occurs in the data directly affects mining results. Hence privacy preserved data's usefulness is measured by using it for standard mining tasks such as classification, rule mining, clustering etc...The results of these gauge the utility of data and hence called quality metrics [3]. In addition, with datasets growing, performance of these algorithms with increase in scale is also measured. Time and memory requirements are parameters that add to effectively quantifying any privacy technique.

With ever growing data, robust techniques are required to handle the competing challenges of privacy, data worth, scalability and performance in the current Big Data and IoT scenario. Many techniques have been studied in the recent past, parallel processing frameworks being one of them. Hadoop [20] and Spark [21] have been used for parallel computations using MapReduce. High Performance Computing has been receiving a lot of attention, recently, and can offer a number of opportunities for effective analytics, managing memory and speed requirements for various privacy techniques.

III. PROBLEM DEFINITION

The work proposed in the paper evaluates the efficacy of privacy preserved data. A grouping-based perturbation technique is used for imposing privacy. The perturbation remodels the data redistributing attribute values of data instances that have similar characteristics. We call this transformed data Remodeled data (RM) and the process as Data Remodeling. The proposed method is divided into two phases. The first phase associates each data sample with k instances that are closest to the case in point. The selection of k closest data instances is realized using a distance matrix based nearest neighbor method. The resultant grouping identifies similar data instances whose attribute values are then randomized to suitably transform data. The strength of transformation introduced in data is quantified by verifying the statistical properties of the remodeled data. The fundamental nature of any perturbation-based privacy preservation requires that the cumulative data remains useful for efficacious analytics, albeit a considerable decline in accuracy of results may be observed in comparison to the non-perturbed data. The execution efficiency of the proposed technique is then experimented on Apache Spark. [21]

A. Contributions of the paper

- We have proposed a blend of sampling and randomization techniques to distort data for privacy preservation. The random distortion is applied to sampled data that is grouped based on the similarity of data subjects. The intuition behind the idea is that an external entity who acquires access to data, tries to single out an individual based on this similarity. Needless to say, the size of the group greatly determines the extent of privacy. Our algorithm gives results at par with the contemporary results.

- Privacy and quality are both quantified, the former based on the percent of distortion and the latter based on the results of classification. Further performance is also measured with parallel execution.

- We have used Spark [21] based parallel computations for enhancing performance. We found that the advantage that it provides in terms of its use of memory and parallel processing can be greatly leveraged for Big Data.

B. Preliminary Definitions

Consider a set of N data instances in table T_o whose data elements are discrete in nature.

The discrete data can be represented as a set of random variables R_n,$R = \{R_1, R_2, R_3, R_4, \ldots\ldots R_k\}$, where each R_k can possibly have m outcomes. Let $O = \{O_1, O_2, O_3 \ldots\ldots O_m\}$ represent m different outcomes of any variable R_k. Let us assume that the table T_o containing the data instances are remodeled into a table T_d using the privacy technique. A multinomial distribution is used to generalize the categorical data attributes.

Statement 1:

The expected number of times any outcome O_i occurs over N instances of data is stated as follows:

$$E(O_i) = N\pi_i \qquad (1)$$

where π_i is the probability of occurrence of a particular outcome O_i

Statement 2:

The variance (SD) of an outcome O_i over N instances is stated as follows:

$$SD(O_i) = \sqrt{(N\pi_i(1 - \pi_i))} \qquad (2)$$

Statement 3:

Given a data table T_o containing N instances, that is transformed to a data table $T_{d,}$ the distortion between the two is measured from the variance between T_o and $Td_{.}$. The statistical component σ^2 is used to calculate normalized variance [19] which can used to quantify privacy introduced as a result of distortion. Privacy based on variance is defined as follows:

$$Priv_{NV} = \frac{T_o - T_d}{T_o} \qquad (3)$$

IV. METHODOLOGY

A. Data Remodeling Phase

The proposed workflow is split into two phases. This section explains the data remodeling phase. In the remodeling phase, a good representative of input data is generated using percentage sampling, drawing an ideal number of random instances. The sampling is initiated using a random seed value. Next, a distance matrix is calculated, that finds pair wise similarity between the data instances. Data is vectorized to facilitate distance measurement. The distance matrix is then used to choose k closest neighbors of every instance. The distance based categorization allows for any two dissimilar instances to be far from each other. Here, k is chosen arbitrarily and the attributes of chosen k neighbors are irregularly distributed and assigned amongst them. Characteristically, analogous groups are most often vulnerable to privacy breaches and hence randomization of attribute values minimizes the risk of privacy attacks on individual data subjects. However the aggregate properties and nature of data remains unaffected, favorable for wholesome analysis. In the rest of the paper, the proposed technique will be referred to as RNNR, Remodeling using Nearest Neighbor Randomization. Figure 1 describes the flow of the remodeling phase.

B. Algorithm RNNR

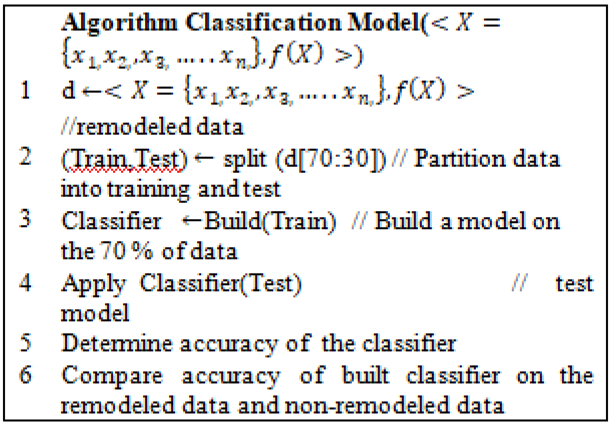
```
Algorithm
RNNR($< X = \{x_1, x_2, x_3, \ldots\ldots x_n\}, f(X) >$)
d ← load N instances of input  //Each instance is a vector of the form $<x_1, x_2, x_3, \ldots\ldots x_n, f(x) >$
pdata← preprocess(d)
sdata← generate a percentage sample from pdata
distMatrixdata ← Calculate DistanceMatrix(sdata, distMeasure)
choose an arbitrary k value
For each sample $x_q$ in sdata
  choose k instances$\{x_1, x_2, x_3, x_4, \ldots\ldots x_k\}$ from distMatrixdata that is closer to any instance $x_q$
  for every discrete attribute $a_i$ in A
    If $< a_1, a_2, a_3 \ldots\ldots a_v >$ is the set of values of attribute $a_i$ of the k selected neighbors
      generate a new vector $x_q'$ such that attribute values are assigned randomly from <$a_1$,$a_2$,$a_3$,....$a_v$>
  assign $x_q'$ to $x_q$
return sdata
```

The RM data in used in the second phase to perform a comparative study of classification accuracies relative to the non-remodeled data

C. Remodeled Data Analysis

This section details the analysis of the remodeled data for classification. ID3 based decision tree binary classifier is built on the RM data and the results are compared with that of the non-remodeled counterpart. A 70:30 split of RM<sdata> is done, the decision tree is trained, and the classifier is validated on test data. The process is repeated for non-remodeled data (NRM) and comparative results are studied. The algorithm was found to give credible results for classification accuracy.

D. Classification Algorithm



```
Algorithm Classification Model($< X = \{x_1, x_2, x_3, \ldots\ldots x_n\}, f(X) >$)
d ←$< X = \{x_1, x_2, x_3, \ldots\ldots x_n\}, f(X) >$ //remodeled data
(Train,Test) ← split (d[70:30]) // Partition data into training and test
Classifier ←Build(Train) // Build a model on the 70 % of data
Apply Classifier(Test) // test model
Determine accuracy of the classifier
Compare accuracy of built classifier on the remodeled data and non-remodeled data
```

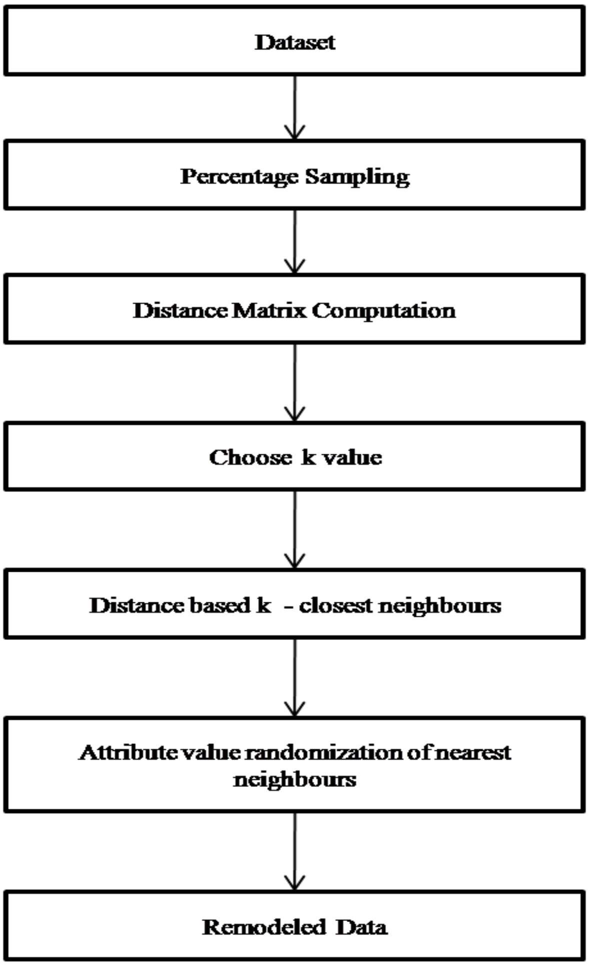

Dataset
Percentage Sampling
Distance Matrix Computation
Choose k value
Distance based k - closest neighbours
Attribute value randomization of nearest neighbours
Remodeled Data

Fig. 1 Data Remodeling

V. FUNCTIONAL DESCRIPTION

This section describes the results of the work from three perspectives. Firstly, we study the effectiveness of RNNR for privacy preservation. Secondly the utility of RM data is assessed and lastly, its performance is estimated in a parallel execution set up. MATLAB [22] and KNIME [23] are used for conducting the experiments. KNIME is used for testing the classical sequential execution of RNNR. Then we use the parallel computing framework Apache Spark to test the time efficiency of the algorithm.

A. Dataset Description

Two different datasets are used for the purpose of testing, the benchmark adult dataset and a clinical dataset from the UCI machine learning repository [24].The clinical data has about 16 attributes, with information about patients, symptoms of the disease, medications and the test results. The continuous attributes in the data is adaptively discretized. Same is the case with the adult dataset.

B. Quantification of RNNR

The efficiency of RNNR algorithm was assessed by calculating collective statistics of the data. This was measured through the relative error rate of remodeling .We define relative error rate(ER), as the percentage of distortion created in the attribute values for the variables, based on the number of instances relative to an attribute value in the data The relative error rate defines the privacy measure as discussed in Eqn. 3. Deviation in the remodeled data using RNNR for two variables, marital status and occupation has been shown in Figure 2 and Figure 3.

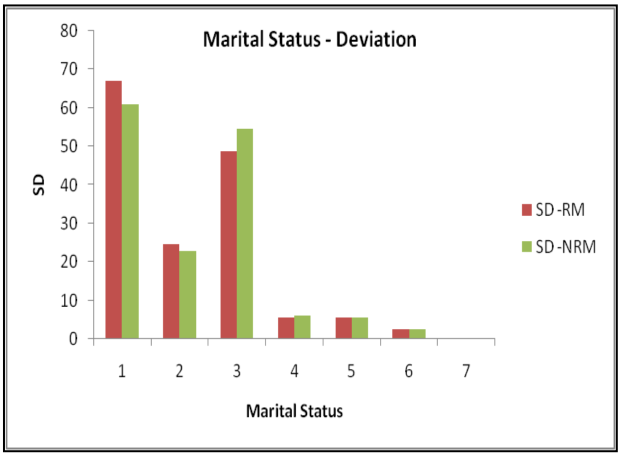


Fig 2 Deviation in Marital Status – RM vs NRM

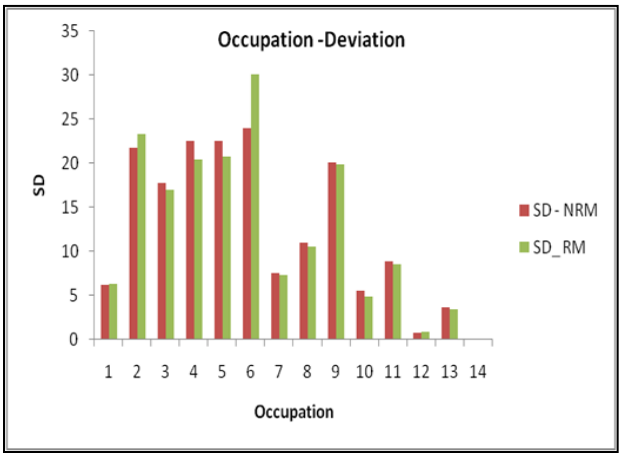


Fig 3 Deviation in Occupation – RM vs. NRM

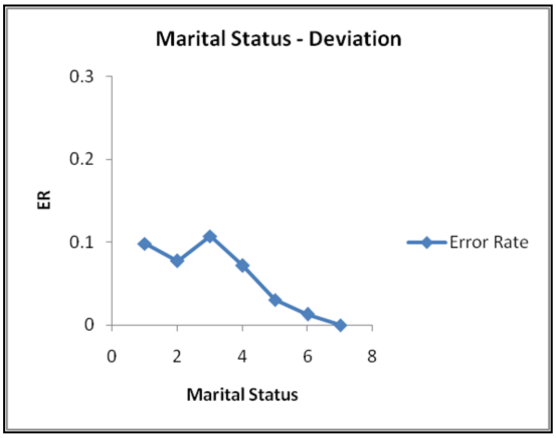


Fig 4 Error rate –Marital Status

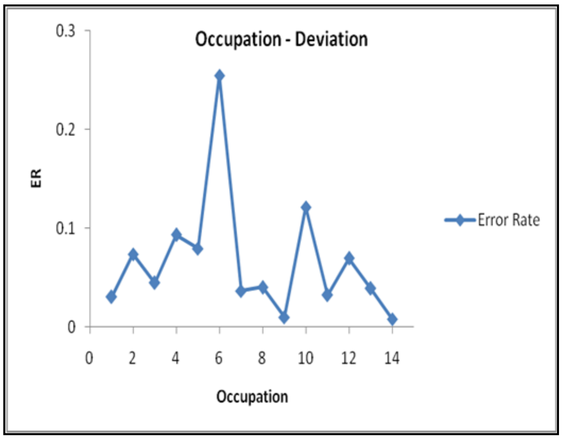


Fig 5 Error rate –Occupation

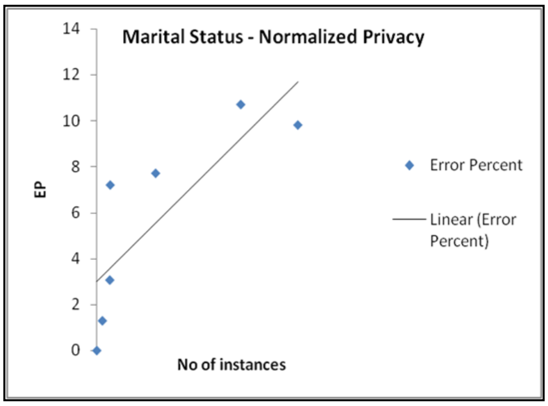


Fig 6 Error percent – Marital status

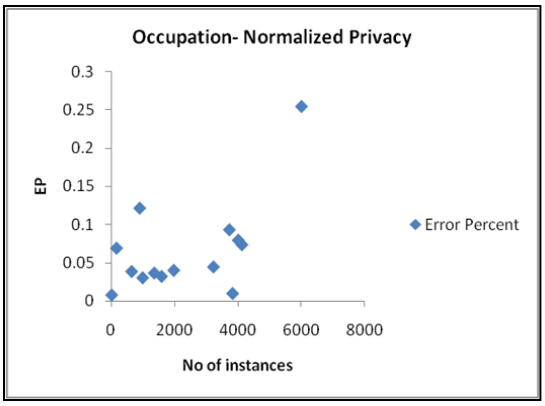


Fig 7 Error percent – Occupation

As discussed in Section B, the discrete variables were assumed to follow a multinomial distribution. Hence, an outcome < Oi> of an attribute value a_i's deviation was calculated using measures such as count of occurrence of O_i(C), its probability of occurrence(π_i), mean(E(O_i)) and standard deviation(SD(O_i))as discussed in Eqn.1 and Eqn. 2. The graphs in Fig 4 and Fig 5 show that there is modest deviation between the remodeled (SD -RM) and non-remodeled data (SD -NRM). Fig 6 and Fig 7 plots the error percent for the attributes marital status and occupation. The margin of error was found varying depending upon the number of instances. It could be seen that, more the number of instances for a particular outcome O_i, the greater the error rate. Such attributes were able to randomize better. Results showed that individually a data instance is sanitized, but aggregate results can still be obtained using RNNR.

C. Classification Performance

The utility of the RM data was evaluated based on relative accuracies of classifiers on RM and NRM data. Two classification algorithms were used for testing purpose, ID3 based Decision Tree and SVM. Results were compared for two different datasets with differing scale and dimensions. Though classification accuracies do show a variation between the two results, it does not limit the extent of usage of RM data for healthy analytics. The results are tested using KNIME shown in Fig 8 and Fig 9.

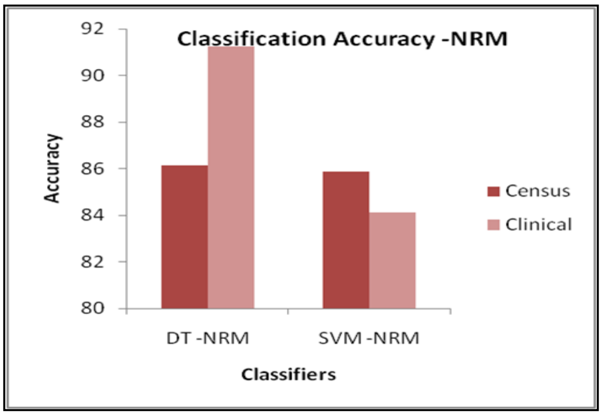


Fig 8 Classification Accuracy – NRM

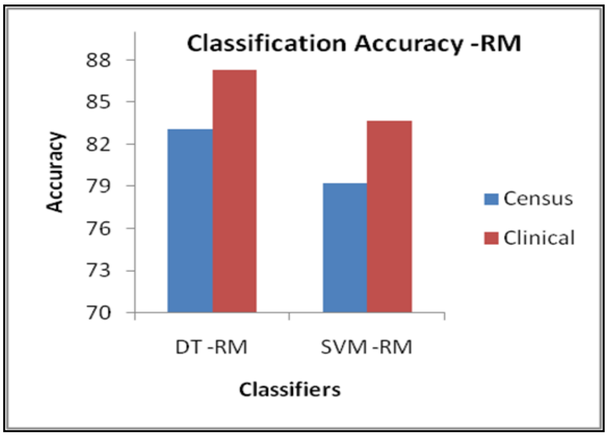


Fig 9 Classification Accuracy - RM

D. Execution Time

We tested the execution time of RNNR first on KNIME. The pairwise similarity calculation was initially tested on the benchmark data, using a representative sample. The data size was gradually increased .It was found that the time for execution increased due to the number crunching of distance matrix. The vectorization of data helped improve the running time to a great extent. Then we compared the results using Apache Spark as our experimental set up. The algorithm was slightly tuned to adapt to the current setting. We refer to it as modified RNNR(m-RNNR). Since the similarity join using RowMatrix [25][26] in Spark, implicitly finds similar vectors, time efficiency of RNNR was naturally improved.

The parallel computation showed performance enhancements relative to its sequential execution on KNIME. For relative results display, we call the runtime environments as Sequential Execution Environment (SEE) and Parallel Execution Environment (PEE**)**. The results are as shown in Fig 10.

The datasets were expanded and the algorithm was experimented on the PEE. The PEE gave better time efficiency for the datasets considered, yet the efficiency was found to decline with increase in scale. Although it can be argued that, analytics-generated observations are easy to replicate, techniques that can handle scale are better realizable by incorporating high performance techniques into analytics.

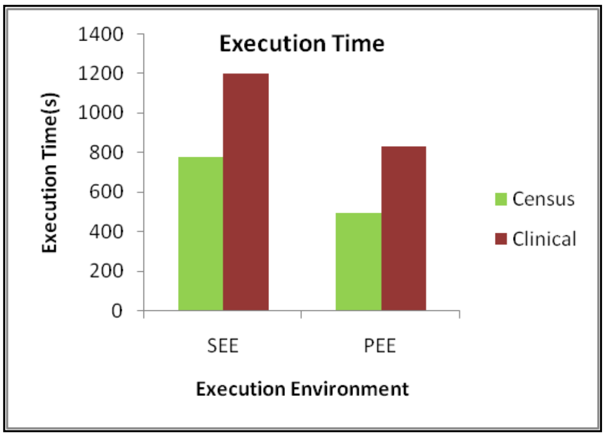


Fig 10 Execution Time - m-RNNR

VI. CONCLUSION

Big data technological era has created a far-fetched arena for building intelligent systems. Data collected from users is the winning bet for such decision support systems. These systems use collected data that may have sensitive information of users, which are at a risk of inadvertent use. Hence privacy and security are challenging domains of research in an era of Big Data.

The work in this paper, firstly proposes a privacy preservation technique based on the idea of obfuscating data that conceals the sensitive information of an individual, while allowing data to be used for productive wholesome analysis. Since privacy of an individual is compromised based on its relative existence with others in a published table, similar data subjects are identified for randomization. Sampling, similarity and randomization act together in the process of remodeling data. Secondly, classification accuracy is measured on the remodeled data and results are compared with the original data. The proposed technique is able to provide justifiable results proving the data's worth for holistic analysis. Thirdly, the work measures the performance efficiency of the algorithm using Apache Spark as the execution environment. Notwithstanding the fact that, any parallel processing would improve performance, the algorithm could scale better with larger processing power. Hence utilizing the power of GPU for parallel computations can greatly enhance performance which is the extension of the proposed work.

Summing up, efficient and scalable solutions for constructive data analytics in the Big data scenario, while respecting a user's privacy, demands a good synthesis of robust privacy techniques coupled with powerful processing environments.

REFERENCES

1. Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects Lei Xu, Chunxiao Jiang, Jian Wang, Jian Yuanand Yong Ren, Information Security in Big Data: Privacy and Data Mining, IEEE Transactions, volume 2,2014
2. Ricardo Mendes and Jo˜Ao P. Vilela, Privacy-Preserving Data Mining: Methods, Metrics, and Applications, IEEE Transactions, volume 5, 2017
3. Vennila . S, Priyadarshini . J,Scalable Privacy Preservation in Big Data
4. A Survey, Procedia Computer Science, 50, (2015), 369 – 373Gayatri Nayak, Swagatika Devi, "A survey on Privacy Preserving Data Mining: Approaches and Techniques", International Journal of Engineering Science and Technology, Vol. 3 No. 3, 2127-2133, 2011.
5. Y. Zhao, M. Du, J. Le, and Y. Luo, ''A survey on privacy preserving approaches in data publishing,'' in Proc. IEEE 1st Int. Workshop Database Technol. Appl., Apr. 2009, pp. 128–131
6. Latanya Sweeney "Achieving k-anonymity Privacy Protection UsingGeneralization and Suppression",May 2002, International Journal onUncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 571-588.
7. A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam,"`l-diversity:Privacybeyondk-anonymity,''AC MTrans.Knowl.Discovery Data, vol. 1, no. 1, p. 3, 2007.
8. Ninghui Li, Tiancheng Li, Suresh Venkatasubramanian, t-Closeness: Privacy Beyond k-Anonymity and ℓ-Diversity, Citiseer 2007
9. Mohammed Al-Zobbi, Seyed Shahrestani, Chun Ruan, Implementing A Framework for Big Data Anonymity and Analytics Access Control, 2017 IEEE Trustcom/BigDataSE/ICESS
10. C. Zhang, E. Chang, and R. H. C. Yap, "Tagged-MapReduce: A general framework for secure computing with mixed-sensitivity data on hybrid clouds," in 14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, CCGrid 2014
11. Liyue Fan and Hongxia Jin. 2015. A Practical Framework for Privacy-Preserving Data Analytics. In Proceedings of the 24th International Conference on World Wide Web (WWW '15). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 311-321
12. Shuo Wang, Richard O. Sinnott, Protecting personal trajectories of social media users through differential privacy, Computers & Security (2017)
13. R. Agrawal and R. Srikant. "Privacy Preserving Data Mining", ACM SIGMOD Conference on Management of Data, pp: 439-450, 2000 –data distribution
14. S. Sharma, K. Chen, A. Sheth, "Towards Practical Privacy-Preserving Analytics for IoT and Cloud Based Healthcare Systems" IEEE Internet Computing, March-April 2018. Towards Practical Privacy-Preserving Analytics for IoT and Cloud-Based Healthcare Systems
15. Chen, K. & Liu, L. Knowl Inf Syst (2011) 29: 657.," Geometirc data perturbation for privacy preserving outsourced data mining" https://doi.org/10.1007/s10115-010-0362-4
16. Machine Learning and Data Mining in Pattern Recognition, 5th International Conference, MDLM 2007, Leipzig, Germany, July 2007, Proceedings, Springer Publications
17. M.A.P. Chamikaraa,b,, P. Bertoka, D. Liub, S. Camtepeb, I. Khalila," Efficient Data Perturbation for Privacy Preserving and Accurate Data Stream Mining", "Journal of Pervasive Computing, April 18, 2018.
18. Isabel Wagner David Eckhoff, "Technical Privacy Metrics: A Systematic Survey", ACM Computing Surveys, Vol. 51, No. 3, Article 57.,June 2018
19. https://hadoop.apache.org
20. https://spark.apache.org

21. https://www.mathworks.com/products/matlab.html
22. https://www.knime.com/knime-software/knime-analytics-platform
23. https://archive.ics.uci.edu
24. https://stanford.edu/~rezab/slides/maryland_mlib.pd
25. https://spark.apache.org/docs/1.2.2/api/java/org/apache/spark//RowMatrix.html

AUTHORS PROFILE

Geetha P. is currently pursuing her Ph.D from Visveswaraya Technological University, Karnataka. She has received her M.Tech in Computer Science from VTU in 2011. Her areas of interest include Machine Learning, Data Mining and Big Data Analytics.She has presented and published papers in revered National and International conferences/journals She has a rich teaching experience of 12 years and is currently associated with Cambridge Institute of Technology, Bangalore, Karnataka.

Chandrakant Naikodi is working as a Principal Applications Engineer in a MNC, Bangalore, India. He has more than 14 years of teaching experience in Software Development industry. He has published various research papers in revered international journals and conferences. He has authored 16 technical text books published by Tata Mc-Graw Hill, SCHAND. His areas of interest include Computer Networks, MANETS, WSN, Big Data.He is a visiting faculty with Cambridge Institute of Technology.

Suresh L has received his Ph.D in the year 2010. He has a vast teaching experience of 29 years in academia and has been rendering his services in the education segment for the student community. Suresh has authored books on Data Structures, Algorithms, Java, Python and other programming languages. He has around 70 publications to his credit in reputed International and National journals. His areas of interest include Data Mining, Data Structures, Algorithms and Database Management Systems. He is currently the Principal in Cambridge Institute of Technology, Bangalore, Karnataka.