Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2012) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

Complex Adaptive Systems, Publication 2
Cihan H. Dagli, Editor in Chief
Conference Organized by Missouri University of Science and Technology
2012- Washington D.C.

Towards A Differential Privacy and Utility Preserving Machine Learning Classifier

Kato Mivule[a]*, Claude Turner[b], and Soo-Yeon Ji[c]

[abc] Dept. of Computer Science, Bowie State University,14000 Jericho Park Road Bowie, MD 20715, USA

Abstract

Many organizations transact in large amounts of data often containing personal identifiable information (PII) and various confidential data. Such organizations are bound by state, federal, and international laws to ensure that the confidentiality of both individuals and sensitive data is not compromised. However, during the privacy preserving process, the utility of such datasets diminishes even while confidentiality is achieved--a problem that has been defined as NP-Hard. In this paper, we investigate a differential privacy machine learning ensemble classifier approach that seeks to preserve data privacy while maintaining an acceptable level of utility. The first step of the methodology applies a strong data privacy granting technique on a dataset using differential privacy. The resulting perturbed data is then passed through a machine learning ensemble classifier, which aims to reduce the classification error, or, equivalently, to increase utility. Then, the association between increasing the number of weak decision tree learners and data utility, which informs us as to whether the ensemble machine learner would classify more correctly is examined. As results, we found that a combined adjustment of the privacy granting noise parameters and an increase in the number of weak learners in the ensemble machine might lead to a lower classification error.

Keywords: Differential Privacy, Privacy Preserving Classification, Ensemble Machine Learning

1. Introduction

Organizations that transact in large amounts of data have to comply with state, federal, and international laws to guarantee that the privacy of individuals and other sensitive data is not compromised. However, during the privacy preserving process, when personal identifiable information (PII) is removed and privacy preserving algorithms such as differential privacy are applied, the utility of such datasets diminishes even while confidentiality is achieved--a problem that has been defined as NP-Hard [1-5]. Therefore, we investigate and present preliminary results of preserving differential privacy with machine learning ensemble classifier approach that maintains data privacy while maintaining a level of utility. In this study, we apply a strong data privacy granting technique on datasets using differential privacy. After which, we pass the perturbed data through a machine learning ensemble classifier. One of the aims of this study is to find a satisfactory threshold value that is measured by adjusting the differential privacy noise levels, with the aim of reducing the classification error. (It is important to note that a lower classification error tends to produce higher utility.) Additionally, the association between increased number of weak decision tree

learners and data utility to validate whether the proposed ensemble machine learner can classify differentially private data accurately is examined. The rest of this paper is organized as follows. Section 2 presents background and related work. Section 3 describes our methodology and experiment. Section 4 discusses results. Finally, Section 5 provides the conclusion.

2. Background and Related Work

2.1. Essential Terms

The following definitions will be essential in this paper in the context of privacy preserving classification. Privacy preservation in data mining and machine learning is the protection of private and sensitive data against disclosure during the data mining process [6, 7]. Ensemble classification is a machine learning process, in which a collection of several independently trained classifiers are merged so as to achieve better prediction. An example includes a collection of independently trained decision trees that are combined to make a more accurate prediction [8-10]. The classification error of a model M_j, is the summation of weights of each record in D_j that M_j classifies incorrectly, where err (X_i) is the classification error of record X_i, and d is the length of D_j. If the record was misclassified, then err (X_i) is 1 otherwise err (X_i) is 0 [27]. The classification error is computed as follows:

$$\text{Error}(\mathbf{M}_j) = \sum_{i=1}^{d} \mathbf{w}_{ji} * \text{err}(\mathbf{X}_i) \tag{1}$$

AdaBoost, also known as Adaptive Boosting, is a machine learning that utilizes several successive rounds by adding weak learners to create a powerful learner. At each successive round, a new weak learner is added to the ensemble classifier by adjusting weights with emphasis placed on misclassified data in earlier iterations [11-13]. The covariance of random two random variables X and Y, Cov(X, Y), is a measure of how the two random variables change jointly. If Cov(X, Y) is positive, then X and Y grow simultaneously. If the covariance is negative, then either X increases and Y decreases, or Y increases while X decreases. If the covariance is zero, the random variables are uncorrelated [26]. μ_X and μ_Y indicate the means of X and Y respectively, and the variances by σ_X and σ_Y, respectively. Then the covariance of X and Y is defined by Cov(X, Y) = E [(X-μ_X) (X-μ_Y)]. If we have a series of n measurements of X and Y written as row vectors x_j and y_k of length N, where j, k$\in${1, 2, ..., n}, then the sample covariance can be used to estimate the covariance Cov(X, Y). Let $\bar{y}$ denote the sample mean of the variable y. Then, the sample covariance is given by

$$c_{jk} = \frac{1}{N-1} \sum_{i=1}^{N} \left(x_{ij} - \bar{x}_j\right)\left(y_{ik} - \bar{y}_k\right) \tag{2}$$

The correlation measures the statistical dependence between two random variables X and Y. The most familiar measure of this dependence is Pearson's product-moment correlation coefficient (or Pearson's correlation), given by ρ_{xy} = Cov(X, Y) /($\sigma_X\sigma_Y$). If ρ_{xy} = +1, then there is a positive (increasing) linear relationship between X and Y [26]. If ρ_{xy} = -1, then there is a negative (decreasing) linear relationship between X and Y. Values of ρ_{xy} between 1 and -1 indicates the degree of the linear relationship between X and Y, with ρ_{xy} = 0 indicating that X and Y are statistically unrelated (or uncorrelated). If we have a series of n measurements of X and Y written as x_i and y_i, where i $\in${1, 2, ..., n}, with sample means then the sample correlation coefficient can be used to estimate the population Pearson correlation r_{xy} between X and Y. The sample correlation coefficient is written:

$$r_{xy} = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) \Big/ \sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \tag{3}$$

Data Utility versus Privacy. Data utility is the extent of how useful a published dataset is to the consumer of that publicized dataset. In the course of a data privacy process, original data will lose statistical value despite privacy guarantees. While attaining an optimal balance between data utility and privacy remains an NP-hard task, it is a balance that is highly desired and is continually pursued [1-5]. Differential Privacy, proposed by Dwork [14, 15], is a data privacy algorithm that works by adding Laplace noise to query answers from databases in such a way that a user of the database cannot determine if a data item has been altered. In such settings, it becomes very difficult, if not impossible, for an attacker to gain knowledge about any information in the schema. Therefore, a data privacy process, q_n, is said to achieve ε-differential privacy if the responses to any identical query ran on databases D_1 and D_2 are probabilistically alike and as long as those results satisfy the following requirement [14, 15]:

$$(P\,[q_n(D_1) \in R]) / (P\,[q_n(D_2) \in R]) \leq \exp(\varepsilon), \quad (4)$$

where D_1 and D_2 are the two schemas, $P[\cdot]$ denotes the probability, $q_n(D_i)$ is the privacy procedure on query results from the schema D_i, R is the perturbed query results from the schemas, and ε is a measure of the amount of noise.

2.2. Privacy and Utility

In recent years, there has been considerable research interest in privacy preserving classification, with much attention given to privacy preservation distributed data mining in which associates do data mining on private data held by other associates [18, 20, 22, 23]. For example, Gorai et al [21] have proposed utilizing bloom filters for privacy preserving distributed k-NN classification. Their experiments show that bloom filters do preserve privacy while conserving the correctness of classifications. However, there is a growing interest in investigating privacy preserving data mining solutions that provide a balance between privacy and utility [24]. Kifer and Gehrke [24] conducted an unprecedented, comprehensive study of data utility in privacy preserving data publishing by employing statistical approaches. In this statistical approach, they measured the probability distributions of the original data and anonymized datasets to enhance data utility. However, for the m-confidentiality algorithm, in which privatized datasets are made public with minimum information loss, Wong [1] described how achieving global optimal privacy while maintaining utility is an NP-hard problem. Yet still, researchers have continued to study possible tradeoffs between privacy and utility in which some sensitive data is either concealed or revealed, so as to grant both data privacy and information usefulness [2, 4, 5, 6, 24]. Yet, Krause and Horvitz [25] have noted that even such an endeavor of finding the tradeoffs between privacy and utility is still an NP-hard problem. Recently, researchers have also shown that while differential privacy has been known to provide strong privacy guarantees, the utility of the privatized datasets diminishes due to too much noise [16, 17]. Therefore, finding the optimal balance between privacy and utility still remains a challenge—even with differential privacy.

3. Methodology and Experiment

This section describes our proposed approach and experiment. In the first step of the proposed approach a strong data privacy granting technique using differential privacy is applied to a given dataset. Next, the perturbed data is passed through a machine learning ensemble classifier that performs a measure of the classification error. This procedure repeats until a satisfactory threshold is attained. If the classification error threshold is not attained, the differential privacy noise levels are re-adjusted to reduce the classification error. Since a higher classification error indicates lower utility, the intention of our approach is increasing utility by diminishing the classification error. At last, the proposed ensemble machine learner is evaluated by examining the association between increased number of weak decision tree learners and data utility. Thus, we can verify whether the proposed approach can classify data more correctly or equivalently to improve data utility.

Steps in the Differential privacy Machine Classification Approach

Step 1: Calculate the Max Difference.
Step 2: Generate, $\boldsymbol{b}$ in Laplace (0, $\boldsymbol{b}$), where

$\boldsymbol{b} = \Delta f / \varepsilon$, *The Max Difference / epsilon.*

Step 3: Assign a small ε, epsilon value of 0.01.
Step 4: Divide Max Difference by the ε, epsilon value: Max Difference/ ε.
Step 5: Generate Laplace Noise Distribution.
Step 6: Add Laplace noise to original data set.
Step 7: Send results to AdaBoost Fit Ensemble Machine Learner.
Step 8: Run AdaBoost Fit Ensemble on both Training and Testing data sets.
Step 9: Get the Classifier Error.
Step 10: If Classifier Error <= Threshold,
Step 11: Refine Differential Privacy by retuning the Laplace noise in step 5.
Step 12: Repeat steps 7 to 9.
Step 13: Stop when Classifier Error <= Threshold.

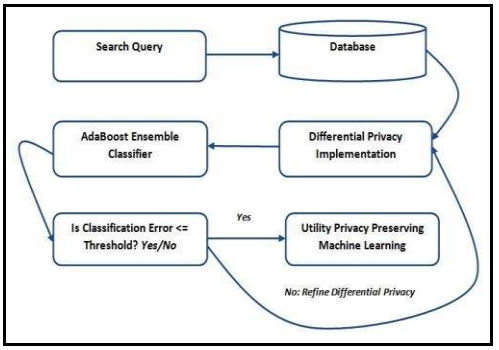

Fig. 1. (a) Steps in our differential privacy classifier approach; (b) A flow chart description our differential privacy utility classifier approach

Experiment: In our experiment, a Barack Obama 2008 campaign donations dataset made public by the Federal

Election Commission is used [26]. The data set contained 17,695 records of original unperturbed data and another 17,695 perturbed records of differentially private data. Two attributes, the donation amount and income status, are utilized to classify data into three groups. The three groups are low income, middle income, and high income. Low income indicates donations in the range $1 to $49, middle income for donations $50 to $80, and high income for donations greater than $81. The range of donations in this dataset was from $1 to $100. Validating our approach, the dataset is divided into two non-overlapping parts, which comprise the training and testing datasets. That is, 50 percent of the original dataset is used as training and the remainder is used for testing. The same procedure is used for the differentially private dataset. The dataset is then imported into the Oracle 11g database management system, after which it is accessed via MATLAB 2012a with MATLAB Oracle ODBC Connector. Next, a query is performed to the Oracle database via the MATLAB ODBC connector functionality. As a last step, MATLAB functionality is utilized to implement differential privacy on the query response using the pseudo code and corresponding block diagram illustrated in Fig 1 (a) and (b), respectively.

4. Results

In this paper, we applied differential privacy to the original dataset before the classification process, with the aim of investigating how the classifier would respond to a differentially privatized dataset. We found that essential statistical characteristics were kept the same for both original and differential privacy datasets a necessary requirement to publish privatized datasets. As depicted in Table 1, the mean, standard deviation, and variance of the original and differential privacy datasets remained the same. There is a strong positive covariance of 1060.8 between the two datasets, which means that they grow simultaneously, as illustrated in Fig. 2. (a). However, figure 2(b) shows that there is almost no correlation (the correlation was 0.0054) between the original and differentially privatized datasets. This indicates that there might be some privacy assurances given, as it becomes difficult for an attacker, presented only with the differential privatized dataset, to correctly infer any alterations.

After applying differential privacy, AdaBoost ensemble classifier is performed. In particular, AdaBoostM1 classifier is used since it is a better technique when few attributes are used, as is the case with our dataset. As can be seen in Table 2, the outcome of the donors' dataset was Low, Middle, and High income, for donations 0 to 50, 51 to 80, and 81 to 100, respectively. This same classification outcome is used for the perturbed dataset to investigate whether the classifier would categorize the perturbed dataset correctly. The training dataset from the original data (Fig. 3 (a)) showed that the classification error dropped from 0.25 to 0 when the number of weak decision tree learners increased. At the same time, the results changed with the training dataset on the differentially private data when the classification error dropped from 0.588 to 0.58 as the number of weak decision tree learners increased; however, this value remained constant with the increase in the number of weak decision tree learners. When the same procedure is applied to the test dataset of the original data using AdaBoost, as illustrated in Fig. 3 (b), the classification error dropped from 0.03 to 0 as the number of weak decision tree learners increased. However, when this procedure perform on the differentially private data, the error rate did not change even with increased number of weak decision tree learners in the AdaBoost ensemble, as depicted in Fig 3. (b).

In this study, we found that while differential privacy might guarantee strong confidentiality, providing data utility still remained a challenge. However, this study is instructive in a variety of ways. Specifically, it shows that;(1) the level of Laplace noise adapted in this experiment for differential privacy does affect the classification error (See Fig 2 and Fig 3), (2) increasing the number of weak decision tree learners did not have much of a significance in correctly classifying the perturbed dataset because the classification error almost remained the same at 0.58 for both the training and testing datasets of the differentially private data. Because of these issues, adjusting the differential private Laplace noise parameters, ε, is essential for our study.

5. Conclusion

While differential privacy in combination with AdaBoost ensemble machine learning techniques might offer strong confidentiality guarantees, our results show that providing data utility in this context still remains a challenge.

Table 1. Statistical properties before and after differential privacy.

Statistical Properties of Original and Differential Privacy Data 17695 Records Analyzed	
Statistical Property	**Results**
Mean of Original Data	54.71825148
Standard Deviation of Original Data	32.56985386
Variance of Original Data	1060.795381
Mean of Differential Privacy Data	54.48757564
Standard Deviation of Differential Privacy Data	32.44444794
Variance of Differential Privacy Data	1052.642202
Correlation	0.005590515
Covariance	1060.795381

Table 2. Expected classifications.

Original Donations	Classifier Before Privacy	Perturbed Donations	Classifier After Privacy
5	Low Earning	-2.245032757	Low Earning
5	Low Earning	30.64168207	Low Earning
50	Middle Earning	51.90167518	Middle Earning
50	Middle Earning	66.61995424	Middle Earning
100	High Earning	81.53029477	High Earning
100	High Earning	95.79161552	High Earning

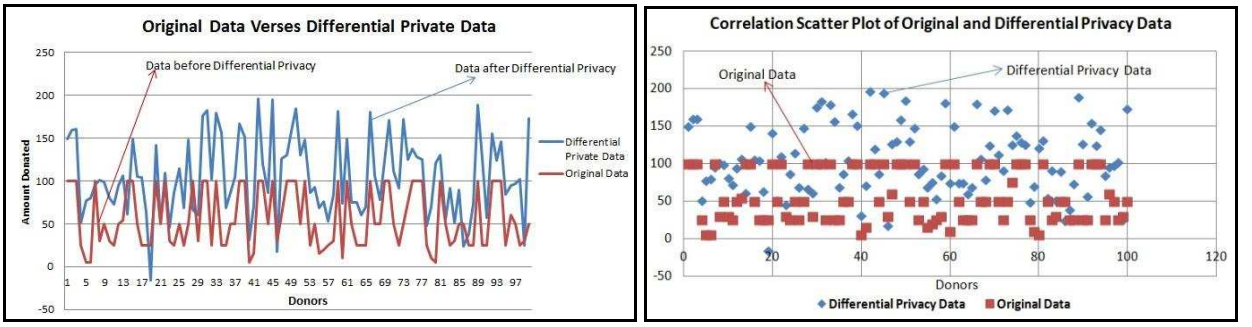

Fig. 2. (a) Original data verses Differential private data; (b) Correlation scatter plot of original and differential private data

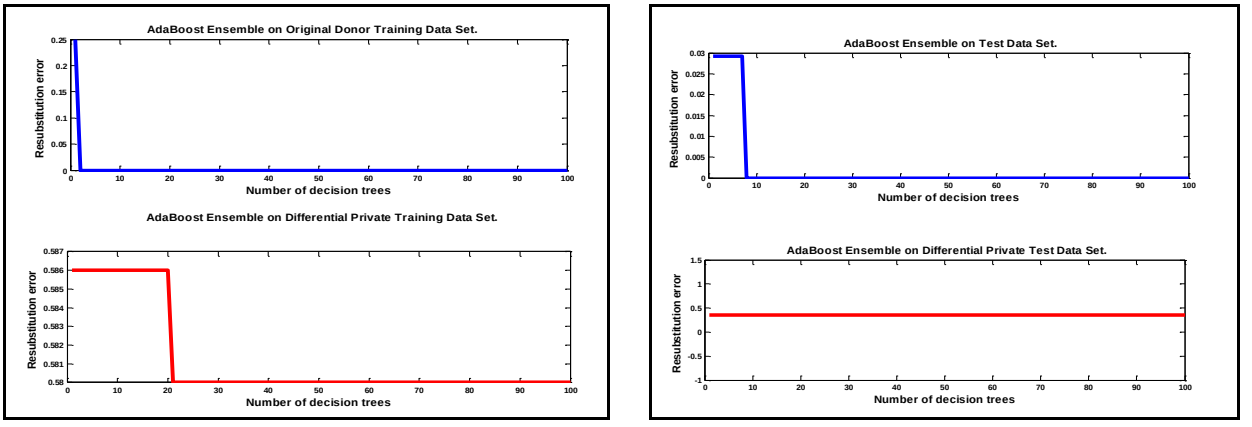

Fig.3. (a) Classification error and decision trees in Training dataset; (b) Classification error and decision trees in Testing dataset

A number of concerns arise. First, the level of Laplace noise parameter, ε, used in this experiment for differential privacy does have an impact on the classification error. Second, increasing the number of weak decision tree learners in the ensemble did not significantly affect the classification of the perturbed datasets (as showed in Fig. 3) with lower utility. The differentially private Laplace noise parameter, ε, might have to be adjusted to achieve higher performance of the perturbed dataset. This would require the perturbed data to be as close to the original data as possible for accurate classification. That is, high classification rate can guarantee better utility. However, since such accurate classification would come with loss of privacy, appropriate tradeoffs must be made between privacy and utility. Additionally, we found that the AdaBoost techniques used for this study can be implemented for both privacy preserving data mining and a data utility measure in differential privacy. However, achieving optimal utility by granting privacy still remains one of the most critical research topics. As a continuation of this study, we will develop an optimization algorithm for the differentially private noise parameter. Also, various ensemble classifiers, such as Bagging, will be employed and ten-fold cross validation will be used to validate the results.

Acknowledgements

This work is supported by the Department of Education's Historically Black Graduate Institution (HBGI) award.

References

1. Wong, R.C., et al, Minimality attack in privacy preserving data publishing, VLDB, pp.543-554, 2007.
2. Tiancheng Li and Ninghui Li, On the Tradeoff Between Privacy and Utility in Data Publishing KDD'09, pp. 517-526, 2009.
3. Daniel Kifer and Ashwin Machanavajjha, No Free Lunch in Data Privacy, SIGMOD'11, pp. 193-204, 2011.
4. Fienberg et al, Differential Privacy and the Risk-Utility Tradeoff for Multi-dimensional Contingency Tables, Privacy in Statistical Databases, LNCS 6344, pp. 187–199, 2010.
5. Gotz et al, Publishing Search Logs – A Comparative Study of Privacy Guarantees, IEEE Transactions on Knowledge and Data Engineering, Volume: 24, Issue: 3, Pages 520-532, 2012.
6. Sramka et al, A Practice-oriented Framework for Measuring Privacy and Utility in Data Sanitization Systems, ACM, EDBT, Article No. 27, 2010.
7. Charu C. Aggarwal, Philip S. Yu, Privacy-Preserving Data Mining: Models and Algorithms, Volume 34 of Advances in Database Systems, Springer, 2008, ISBN 9780387709918, Pages 24-25
8. D. Opitz and R. Maclin, "Popular ensemble methods: An empirical study," Journal of Artificial Intelligence Research, vol. 11, pp. 169-198, 1999.
9. MATLAB, "Ensemble Methods :: Nonparametric Supervised Learning (Statistics Toolbox™)." [Online]. Available: http://www.mathworks.com/help/toolbox/stats/bsvjye9.html#bsvjyi5. [Accessed: 10-Mar-2012].
10. T. G. Dietterich, "Ensemble methods in machine learning," Lecture Notes in Computer Science, vol. 1857, pp. 1-15, 2000.
11. Y. Freund and R. E. Schapire, "A Decision-Theoretic generalization of On-Line learning and an application to boosting," Journal of Computer and System Sciences, vol. 55, no. 1, pp. 119-139, Aug. 1997.
12. Y. Freund and R. E. Schapire. (1999) A short introduction to boosting. Journal of Japanese Society for Artificial Intelligence, No. 14. (1999), pp. 771-780
13. Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in International Conference on Machine Learning, 1996, pp. 148-156
14. Dwork, C., Differential Privacy, in ICALP, Springer, pp 1-12, 2006.
15. Dwork, C., Differential Privacy: A Survey of Results, In Theory and Applications of Models of Computation TAMC , pp. 1-19, 2008
16. Muralidhar, K., and Sarathy, R., Does Differential Privacy Protect Terry Gross' Privacy?, In PSD, Vol. 6344 (2011), pp. 200-209.
17. Muralidhar, K., and Sarathy, R., Some Additional Insights on Applying Differential Privacy for Numeric Data, In PSD, Vol. 6344 (2011), pp. 210-219.
18. Zhuojia Xu; Xun Yi; , "Classification of Privacy-preserving Distributed Data Mining protocols," Digital Information Management (ICDIM), 2011 Sixth International Conference on , vol., no., pp.337-342, 26-28 Sept. 2011
19. Nix, Robert; Kantarciouglu, Murat; , "Incentive Compatible Privacy-Preserving Distributed Classification," Dependable and Secure Computing, IEEE Transactions on , vol.9, no.4, pp.451-462, July-Aug. 2012
20. Gorai, M.R.; Sridharan, K.S.; Aditya, T.; Mukkamala, R.; Nukavarapu, S.; "Employing bloom filters for privacy preserving distributed collaborative kNN classification," WICT, 2011, vol., no., pp.495-500, 11-14 Dec. 2011.
21. F.Camara et al, Privacy Preserving RFE-SVM for Distributed Gene Selection, IJCSI, Vol. 9, Issue 1, No 2, Pages 154-159, 2012.
22. Barni, M.; Failla, P.; Lazzeretti, R.; Sadeghi, A.-R.; Schneider, T.; , "Privacy-Preserving ECG Classification With Branching Programs and Neural Networks," Information Forensics and Security, IEEE Transactions on , vol.6, no.2, pp.452-468, June 2011
23. Kifer D., Gehrke J.; Injecting utility into anonymized datasets, SIGMOD Conference, pages, 217-228, 2006.
24. Andreas Krause, Eric Horvitz; A Utility-Theoretic Approach to Privacy in Online Services, Journal of Artificial Intelligence Research, pages 633-662, 2010.
25. US Federal Election Commission, "Finance Disclosure Files about Candidates, Parties, and other Committees by Election Cycle." [Online]. Available: http://www.fec.gov/finance/disclosure/ftpdet.shtml. [Accessed: 10-Mar-2012].
26. Michael J. Crawley, Statistics: an introduction using R, John Wiley and Sons, 2005, ISBN 0470022973, Pages 93-95.
27. J. Han, M. Kamber, J. Pei; Data Mining: Concepts and Techniques, Elsevier, 2011, ISBN 9780123814791, Page 381.