

Enhanced ℓ – Diversity Algorithm for Privacy Preserving Data Mining

R. Praveena Priyadarsini^(✉), S. Sivakumari, and P. Amudha

Department of Computer Science and Engineering, Avinashilingam Institute
for Home Science and Higher Education for Women, Coimbatore, India

praveena.priya04@gmail.com,

prof.sivakumari@gmail.com, amudharul@gmail.com

Abstract. With the increase in use of e-technologies, large amount of digital data are available on-line. These data are used by both internal and external sources for analysis and research. This digital data contain sensitive and personal information about the entities on which the data are collected. Due to this sensitive nature of such information, it needs some privacy preservation procedure to be applied before releasing the data to third parties. The privacy preservation should be applied on the data such that its utility during data mining does not get reduced. ℓ -Diversity is an anonymization algorithm that can be applied on dataset with one sensitive attribute. Real life data contain numerous sensitive attributes that have to be privacy preserved before publishing it for research. This paper proposes an Enhanced ℓ -diversity algorithm that can diversify multiple sensitive attributes without partitioning the dataset. Two datasets namely, bench mark Adult dataset and Real life Medical dataset are used for experimentation in this work. The privacy preserved datasets using the proposed algorithm are compared for its utility with ℓ -diversified dataset for single sensitive attribute and original dataset. The results show that the proposed algorithm privacy preserved datasets have good utility on selected classification algorithms taken for study.

Keywords: ℓ -diversity · Enhanced ℓ -diversity · Multiple sensitive attributes · Privacy preservation

1 Introduction

Privacy-preserving data mining refers to the area of data mining that seeks to safeguard sensitive information from unofficial disclosure [14]. Government agencies and other organization frequently require to publish their data for investigate and other purposes. These micro-data contain records about individuals and organizations which contain sensitive and personal information about a person, a household, or society. Thus privacy preservation is needed to hide these sensitive information before they are used for research and publishing. The attributes in the micro- data can be divided into three types namely, sensitive attributes, non-sensitive attributes and Quasi-Identifiers from privacy preservation point of view. Sensitive attributes contain personal privacy information that must identify the exact information of individuals. Quasi-identifiers

are the set of attributes that can be linked with other datasets which is publically available to identify individual's private data. There are many techniques used for privacy preservation. The most commonly used technique for privacy preservation is anonymization technique. Here, the sensitive information is replaced by some random value based on probability distribution.

ℓ -Diversity proposed by Machanavajjhala et al. [1] is an extension of the k -anonymity model. This method overcomes homogeneous and background knowledge attack from k -anonymity technique. This requires every group of attributes should contain ℓ -diversified value to have ℓ -distinct values within the bucket. The major drawback of algorithm is that, it cannot handle multiple sensitive attributes. Most of the extensions in this algorithm have partitioned the dataset to help ℓ -diversity algorithm to accommodate multiple sensitive attributes. The proposed work proposes enhanced ℓ -diversity algorithm that can handle multiple sensitive attributes without partitioning the dataset. The proposed technique applies on three sensitive attributes within the dataset and the privacy preserved datasets from the proposed algorithm is evaluated based on its utility on classification function of data mining.

The paper is organized as follows: Sect. 1 gives the introduction. Section 2 presents literature survey. Section 3 gives the dataset description; Sect. 4 discusses about the proposed Enhanced ℓ -diversity algorithm. Section 5 presents experimental results and discussions and Sect. 6 gives the conclusions.

2 Literature Survey

Aggarwal et al. discussed about techniques like k -anonymity, ℓ -diversity, randomization, condensation and sanitization methods to achieve the privacy of sensitive data in a dataset [14]. Fung et al. discussed about techniques used to preserve the sensitive information before publishing [11].

Aggarwal and Yu [4] proposed a condensation approach for privacy preservation such that dataset are anonymized to a data that closely matches the characteristics of the original data and preserves the correlation among the attribute. Han et al. [5] proposed two anonymization methods to preserve privacy in mixed data. Both the methods MEGA and TSCKA affectively preserve the privacy of mixed data and also have good data quality. Han et al. proposed method called SLOMS where two sensitive attributes that have high chi-square correlation between them are clustered into separate tables. The quasi identifiers of these sensitive attributes are generalized [8].

Xiao and Tao [17] has proposed a method called Anatomy that releases the sensitive and quasi attributes in the dataset as a separate table. A grouping mechanism is performed that preserves privacy as well as captures correlation among the attributes in the dataset.

Machanavajjhala has discussed about the technique known as ℓ -diversity which overcomes the drawbacks of k -anonymity technique and provides highly efficient privacy preservation [1]. ℓ -Diversity requires the equivalence class to have at least ℓ well represented values for each sensitive attribute. But the most important limitation of ℓ -diversity is that it is not sufficient to prevent attribute disclosure attack. Also, when sensitive values of the group are semantically close there is a breach in privacy on a

ℓ -diversified dataset. This limitation of ℓ -diversity is addressed by t -closeness which defines that an equivalence class is t -close if the distance between the distribution of a sensitive attribute in this class and distribution of the attribute in the whole table is no more than a threshold ‘ t ’ [1]. David Kifer and Johannes Gehrka introduced anonymized marginal to increase the utility of the k -anonymous and ℓ -diverse tables [2]. Zak-erzadeh et al. proposed a generalized approach that uses inter-attribute correlates to decompose the data into vertical dataset. An anonymization process that complies with k -anonymity, ℓ -diversity and t -closeness models has been applied on the partitions for privacy preservation [3]. Sattar et al. [6] proposed a probabilistic d-linkable model to mitigate composition attack. This model uses partition based k -anonymization approach. Huiwang and Ruilin Liu proposed (d,l) inference model to anonymization micro-data with unsafe functional dependency with low information loss [7]. Das and Bhathacharrya proposed decomposition algorithm which does ℓ -diversity on a multiple sensitive attributes using partitioning method [9]. Aggarwal, discussed about maintaining ℓ -diversity across “ r ” sensitive attributes [10].

Grigorios Loukides et al. introduced a rule-based privacy model that allows data publishers to express fine-grained protection requirements for both identity and sensitive information disclosure using two different two different anonymization algorithms [12]. First algorithm is to recursively generalize data with low information loss. Second algorithm greatly improved scalability while maintaining low information loss.

Tamas S. Gal and Zhiyuan Chen proposed a technique which treats all the sensitive attributes uniformly. This method allows different degrees of diversity on different attributes at every horizontal partitioning of the data. This horizontal partitions are then anatomized separately [13].

3 Dataset Descriptions

Real time Medical Dataset and Adult Dataset from UCI Knowledge Discovery Archive database [15] are used for experimentation in this work. The adult dataset contains 15 attributes including one class attribute which has two categorical values, “>50 K” and “≤50 K”. The non-class attributes are age, workclass, fnlwgt, education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, and native-country. The description of Adult dataset is given in Table 1.

The Medical dataset consists of 150 records which contains patient details. There are 13 attributes including the one class attribute with categorical values “Breathing” and “Non-Breathing”. The non-class attributes are unique id, Age, Sex, Address,

Table 1. Adult dataset

Dataset	Description
Attribute characteristics	Categorical, integer
Number of instances	48842
Number of attributes	14
Missing values	Yes
No. of classes	2

Table 2. Medical dataset

Dataset	Description
Attribute characteristics:	Categorical, integer
Number of instances:	150
Number of attributes:	13
Missing values	No
No. of classes	2

Occupation, Complaints, Appetite, Mental generals, Built, Height, Weight, and Diagnosis. The description of Real life Medical dataset is given in Table 2.

4 Enhanced ℓ -Diversity Algorithm

The attributes in a dataset can be divided into three types namely, sensitive attributes, non-sensitive attributes and Quasi-Identifiers from a privacy preservation point of view. Sensitive attributes are those that contain personal privacy information to identify the exact information of individuals. Quasi-identifiers are the set of attributes that can be linked with other datasets which is publicly available to identify individual's private data. ℓ -Diversity technique is an extension of the k -anonymity model where a given record maps onto at least k other records in the data and results in loss of data. The ℓ -diversity requires every group of attributes should contain ℓ -diversified value to have ℓ -distinct values within the bucket. The major drawback of ℓ -diversity algorithm is that it cannot handle multiple sensitive attributes.

Tamas S. Gal and Zhiyuan Chen proposed an algorithm which partitions the data and applies anatomy to accommodate multiple sensitive attributes [13]. In this work the proposed algorithm tries to accommodate multiple sensitive attributes for ℓ -diversity by applying few conditions on setting the bucket size and accommodating the various attribute values of sensitive attributes within this bucket. In order to minimize the loss of information, if the ℓ -value for diversification is much lesser than the values present in the Sensitive attribute, generalization is applied. For categorical sensitive attributes where each SA has V_n different values, the ℓ value for diversity is set based on the occurrence of distinct values in QI attributes. The number of distinct values in the quasi identifiers of the dataset are recorded and compared with distinct value of SA attributes. The ℓ -diversity is applied for 1 attribute, 2 attributes and 3 attributes.

Let $\{q_1, q_2, \dots, q_n \in D\}$ be the set of QI attributes of the dataset D . The distinct values held by these attributes are $\{V_1, V_2, \dots, V_n \in D\}$. Let $\{SA_1, SA_2, \dots, SA_n \in D\}$ be the set of Sensitive Attributes (SA) in dataset D and the distinct value held by these SA be $\{Y_1, Y_2, \dots, Y_n \in SA\}$.

Let the minimum (V_n) = A, maximum (V_n) = B, minimum (Y_n) = C and maximum (Y_n) = D

The ℓ value and bucket size during diversification is set based on the following rules given in Eqs. (1, 2 and 3):

If $A = C$ then set bucket size (Z) and ℓ – value of the dataset as C (1)

If $C > A$ then set the values $Z = L = A$ (2)

If $C < A$ then set the values $Z = L = C$ (3)

Also, to decrease the information loss in SA attributes when Z value is set as $\min(V_n)$ generalization is applied on SA_n based on Eq. (4).

If $Y_n((SA_n)) > 2 \times A$ then generalize Y_n , $SA_n \rightarrow A$ (4)

Thus, each sensitive attributes in D will be $\min(V_n)$ diverse or $\min(Y_n)$ diverse.

For example, in the proposed work the QI attributes in the adult dataset have the distinct values $\{V_1, V_2, \dots, V_4\}$ as 7, 10, 13, 6. The distinct values of two SA attributes $\{Y_1, Y_n\}$ in the data set be 13, 6. Then, $\min(V_n) = 6$ and $\min(Y_n) = 6$.

Hence rule 1 applies and bucket size Z and ℓ - value for diversity for the dataset is set as 6. Since the SA attribute set has 13 distinct values inside it which is greater than two times $\min(V_n)$, then generalization is applied to reduce to $\min(V_n)$. Thus each bucket will be $\min(V_n)$ anonymized and all the equivalence class will have a well-defined $\min(V_n)$ diverse values.

As $Y_n(SA) > 2 \times \min(V_n)$ then sensitive attributes are generalized to $\min(V_n)$ value. Thus, each sensitive attributes is ℓ -diverse that helps in preventing membership disclosure attack.

```

Let  $\{q_1, q_2, \dots, q_n \in D\}$ ,  $QI \subseteq D$ 
And  $\{SA_1, SA_2, \dots, SA_n \in D\}$ ,  $SA \subseteq D$ 
 $SA_n \cap QI_n = \emptyset$ 
Let  $\{V_1, V_2, \dots, V_n \in q\}$  be the distant values of QI set
 $\{y_1, y_2, \dots, y_n \in SA\}$  be distinct values of SA set
Bucket size ( $Z$ ) and diversity value  $L$  is
If  $\min(y_n) = \min(v_n)$  then  $Z = L = \min(v_n)$ 
Else If  $\min(y_n) > \min(v_n)$  then  $Z = L = \min(v_n)$ 
Else If  $\min(y_n) < \min(v_n)$  then  $Z = L = \min(y_n)$ 
For  $SA_i$  where  $i = 1$  to  $n$ 
If  $SA_i(Y_n) > 2 \times \min(v_n)$  then
Generalized  $SA_i(Y_n) \rightarrow \min(V_n)$  diverse
 $\ell$ -diversity[SA] ;  $k$ -anonymity[QI] ;
Perturbed Dataset =  $\bar{D} = QI \cup SA \cup NSA$ 

```

Fig. 1. Enhanced ℓ -diversity algorithm

The proposed algorithm is given in Fig. 1. Using the proposed algorithm, the privacy preserved dataset will be having $\min(V_n)$ diversity or $\min(Y_n)$ on all SA

attributes of the dataset. Also all the quasi attributes will also be min (Y_n) or min (V_n) anonymized. Thus the proposed algorithm minimizes attribute value deletion during ℓ - diversity using generalization, and increases the utility of the privacy preserved dataset.

In Adult dataset, attributes described in Table 1, “Occupation”, “Relationship” and “Work class” are set as sensitive attributes. Suppose if third party knows individual’s occupation details then, he/she may easily identify other details about a person which violates their privacy. In the real time medical data set “Complaints”, “Diagnosis” and “Mental Generals” attributes are set as sensitive attributes. The attributes is set as sensitive based on Information gain value of the attribute on class attribute.

The privacy preserved datasets by applying the algorithm incrementally on more than one SA attributes are compared on its utility on classification algorithm using Weka simulator. The level privacy increases in the privacy preserved versions of datasets as the number of attributes subjected to ℓ - diversification increases. These versions can be distributed to users with different trust levels.

5 Experimental Results and Discussion

The two datasets taken for experimentation in this work are: Adult dataset and real world Medical dataset. The ℓ -diversity algorithm is coded using Java as front-end and MySQL as back-end. The various dataset obtained after applying the enhanced ℓ -diversity algorithm on both the datasets are given in Table 1.

Table 3. Privacy preserved adult and medical dataset versions

S.No	Privacy preserved Datasets	Abbreviations
1	ℓ -Diversity with 1 SA for adult dataset	L-D-1-SA- adult dataset
2	ℓ -Diversity with 2 SA for adult dataset	L-D-2-SA- adult dataset
3	ℓ -Diversity with 3 SA for adult dataset	L-D-3-SA -adult dataset
4	ℓ -Diversity with 1 SA for medical dataset	L-D-1-SA- medical dataset
5	ℓ -Diversity with 2 SA for medical dataset	L-D-2-SA- medical dataset
6	ℓ -Diversity with 3 SA for medical dataset	L-D-3-SA- medical dataset

These datasets are compared on their utility using classification accuracy and Receiver Operating Characteristic (ROC) metrics on classification algorithms like Naive Bayes, C4.5 and Ripper using Weka simulator [16].

The equation for calculating classification accuracy is given in Eq. (5),

$$\text{Classification accuracy} = \frac{\text{No. of correctly classified tuples}}{\text{Total No. of tuples in the dataset}} \quad (5)$$

ROC-Area returns the probability or ranking for the predicted class for each tuple present in the dataset. It is a plot of the true positive rate against the false positive rate. The Roc-Area values range 0 to 1.0. The closer the value of area to 0.5 the less accurate and value will be 1.0 for the perfect accuracy [18].

The classification accuracy of the three versions of privacy preserved adult and Medical datasets is measured using classification accuracy of three classifiers namely Naïve Bayes, C4.5 and Ripper algorithms are given in Figs. 2 and 3.

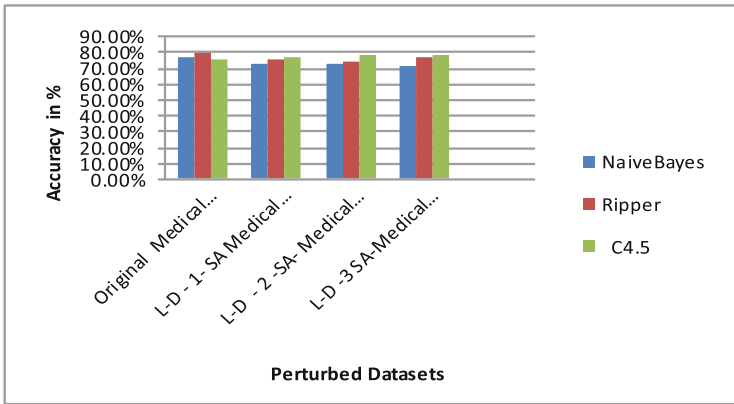


Fig. 2. Comparison of classification accuracy of privacy preserved adult datasets

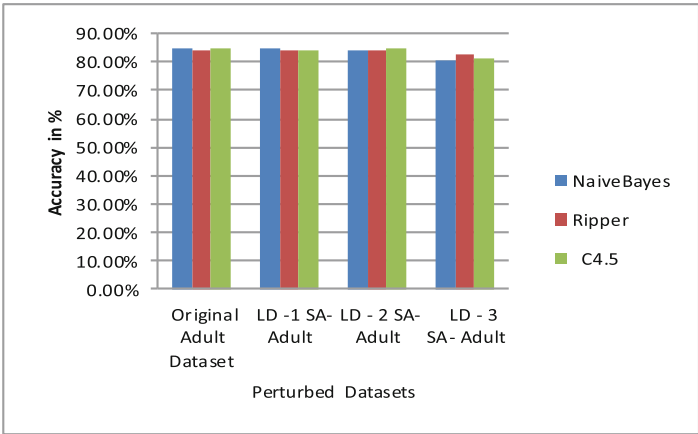


Fig. 3. Comparison of Classification Accuracy of Privacy Preserved Medical Datasets

Figure 2 shows that on all the privacy preserved datasets, Naïve Bayes classifier has a decrease of 4% in accuracy. On Ripper algorithm there is a decrease in 2% accuracy and on C4.5 algorithm there is a decrease in 1% accuracy. C4.5 classifiers have good accuracy when compared with all other classifiers on all the three privacy preserved datasets. When ℓ -diversity is extended to more number of SA, there is loss of utility for all other classifiers.

Figure 3 shows that the utility of ℓ -diversified medical dataset on multiple sensitive attributes is compared on all the three classifiers. The results show that in Naïve Bayes

classifier there is decrease of 4% accuracy. On Ripper algorithm there is a decrease in 2% accuracy and on C4.5 algorithm there is an increase in 2% accuracy. C4.5 classifiers have good accuracy when compared with all other classifiers for all the privacy preserved datasets. When ℓ -diversity is extended to more number of SA then there is a loss of utility (Fig. 3).

The ROC measurements of the privacy preserved versions of adult and medical datasets are tabulated in Tables 4 and 5.

Table 4. Comparing the ROC values of privacy preserved adult datasets

Dataset	ROC area for Naïve Bayes	ROC area for Ripper	ROC area for C4.5
OAD	0.892	0.799	0.791
ℓ -Diversity - 1 SA -adult dataset	0.888	0.796	0.762
ℓ -Diversity - 2 SA -adult dataset	0.881	0.776	0.764
ℓ -Diversity -3 SA - adult dataset	0.656	0.58	0.55

Table 5. Comparing the ROC values of privacy preserved medical dataset

Dataset	ROC area for Naïve Bayes	ROC area for Ripper	ROC area for C4.5
OMD	0.78	0.634	0.47
ℓ -Diversity -1 SA –medical dataset	0.428	0.454	0.468
ℓ -Diversity - 2 SA- medical dataset	0.356	0.464	0.468
ℓ -Diversity - 3 SA- medical dataset	0.343	0.51	0.468

ROC values on Adult ℓ -diversified datasets in Table 4 shows that ℓ -Diversity - 3 SA – Adult dataset has the least ROC values when compared to other L-diversified and original dataset. On medical dataset all the privacy preserved L-diversified datasets have lesser ROC values on Naïve Bayes and Ripper algorithm. ROC value of C4.5 algorithm remains the same as original dataset on all the privacy preserved versions on medical datasets.

Thus, the results indicate that there is a very little decrease in accuracy on the Enhanced L-diversified versions of privacy preserved datasets on both the datasets taken for study. Thus, Enhanced ℓ -diversity algorithm can produce anonymized versions of privacy preserved dataset for multiple SA attributes without vertically partitioning the datasets. Since all the Quasi Identifiers are k -anonymized and the SA attributes are L-diversified the proposed algorithm privacy preserved datasets can prevent homogeneous and background knowledge.

6 Conclusions

Real life datasets and high dimensional datasets contain number of sensitive attributes. In order to apply ℓ -diversity on these types of datasets which contain more than one sensitive attributes, this work proposes Enhanced ℓ -diversity algorithm. The proposed algorithm can diversify more than one sensitive attribute in a dataset without partitioning the dataset. Thus, the proposed algorithm tries to eliminate the drawback of ℓ -diversity algorithm by extending it to accommodate multiple attributes. When the proposed privacy preserved ℓ -diversified datasets are evaluated for their utility on selected classification algorithms they exhibit very less decrease in utility when compared with traditional ℓ -diversity algorithm and original dataset. Thus the proposed algorithm can produce ℓ -diversity on datasets with multiple sensitive attributes with negligible decrease in utility. As future extension feature reduction techniques can be used along with this method to perform privacy preservation on big data.

References

1. Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: l-diversity: privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data (TKDD)* **1**(1), 3 (2007)
2. Kifer, D., Gehrke, J.: Injecting utility into anonymized datasets. In: *ACM SIGMOD International Conference on Management of Data*, pp. 217–228 (2006)
3. Zakerzadeh, H., Aggarwal, C.C., Barker, K.: Towards breaking the curse of dimensionality for high-dimensional privacy: An Extended Version (2014). arXiv preprint [arXiv:1401.1174](https://arxiv.org/abs/1401.1174)
4. Aggarwal, C.C., Yu, P.S.: A condensation approach to privacy preserving data mining. In: Bertino, E., Christodoulakis, S., Plexousakis, D., Christophides, V., Koubarakis, M., Böhm, K., Ferrari, E. (eds.) *EDBT 2004. LNCS*, vol. 2992, pp. 183–199. Springer, Heidelberg (2004). doi:[10.1007/978-3-540-24741-8_12](https://doi.org/10.1007/978-3-540-24741-8_12)
5. Han, J., Yu, J., Mo, Y., Lu, J., Liu, H.: MAGE: A semantics retaining K-anonymization method for mixed data. *Knowl.-Based Syst.* **55**, 75–86 (2014)
6. Sattar, A.S., Li, J., Liu, J., Heatherly, R., Malin, B.: A probabilistic approach to mitigate composition attacks on privacy in non-coordinated environments. *Knowl.-Based Syst.* **67**, 361–372 (2014)
7. Wang, H., Liu, R.: Privacy-preserving publishing microdata with full functional dependencies. *Data Knowl. Eng.* **70**, 3 (2011)
8. Han, J., Luo, F., Lu, J., Peng, H.: SLOMS: A privacy preserving data publishing method for multiple sensitive attributes microdata. *J. Softw.* **8**(12), 3096–3104 (2013)
9. Das, D., Bhattacharyya, D.K.: Decomposition +: Improving ℓ -Diversity for Multiple Sensitive Attributes. In: *International Conference on Computer Science and Information Technology*, pp. 403–412. Springer, Heidelberg (2012)
10. Aggarwal, C.C.: Privacy and the dimensionality curse. In: Aggarwal, C.C., Yu, P.S. (eds.) *Privacy-Preserving Data Mining*, pp. 433–460. Springer US, New York (2008)
11. Fung, B., Wang, K., Chen, R., Yu, P.S.: Privacy-preserving data publishing: a survey of recent developments. *ACM Comput. Surv.* **42**(4), 14 (2010)
12. Loukides, G., Gkoulalas-Divanis, A., Shao, J.: Efficient and flexible anonymization of transaction data. *Knowl. Inf. Syst.* **36**(1), 153–210 (2013)

13. Gal, T.S., Chen, Z., Gangopadhyay, A.: A privacy protection model for patient data with multiple sensitive attributes. IGI Global, pp. 28–44 (2008)
14. Aggarwal, C.C., Philip, S.Y.: A general survey of privacy-preserving data mining models and algorithms. In: Aggarwal, C.C., Yu, P.S. (eds.) *privacy-preserving data mining*, pp. 11–52. Springer, US, New York (2008)
15. <http://archive.ics.uci.edu/ml>
16. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software:an update. *ACM SIGKDD Explor. Newsl.* **11**(1), 10–18 (2009)
17. Xiao, X., Tao, Y.: Anatomy: simple and effective privacy preservation. In: *32nd International Conference on Very Large Data Bases*, pp. 139–150. VLDB Endowment (2006)
18. Han, J., Pei, J., Kamber, M.: *Data Mining: Concepts and Techniques*. Elsevier, Amsterdam (2011)

Digital Connectivity – Social Impact

51st Annual Convention of the Computer Society of
India, CSI 2016, Coimbatore, India, December 8-9,
2016, Proceedings

Subramanian, S.; Nadarajan, R.; Rao, S.; Sheen, S.
(Eds.)

2016, XI, 301 p. 123 illus., Softcover

ISBN: 978-981-10-3273-8