Data utility measures-a survey

Rajeshwari N O
Student (MTech)
Department of computer science and engineering
M S Ramaiah Institue of Technology

Sowmyarani C N
MTech, Assistant professor
Department of computer science and engineering
M S Ramaiah Institue of Technology

***Abstract*—the statistical data released to the public is normally generalized and anonymized to preserve the privacy of the respondents. In the course of preserving the privacy information loss will occur which affects the data analysis process. The balance between privacy and information loss is to be maintained in data mining. The quality of the data should be adequate to properly analyze the data. The data utility measures play a significant role before analyzing the data. This paper focus on survey of different data utility measures that are discussed in recent development.**

Keywords— Utility measures Cluster analysis Propensity score CDF measures kmeans cluster and veiled

I. INTRODUCTION

Data researchers need data for analyzing. The need for statistical data is increasing for analysis. Lot of statistical disclosure control techniques are applied on raw data to preserve the privacy of individuals. Thus released data is benefit to the society. The utility factor of the data is more important, in the other way it should also hide the privacy of individuals. These disclosure control techniques hide privacy part of the data, but this affects the utility of the data. The excess use of some of these techniques affect data utility. Many statistical disclosure methods have been developed which are applied to data depending on the intensity levels. If the intensity level is more, disclosure risk is less but it disturbs the value of the data. In categorical data, information loss is measured in three ways they are comparison of categorical value, comparison of contingency values, and entropy based measure. The difference between original and modified data measures the continuous data. Mean square error, mean absolute error and mean variation gives the measures of these differences.

II. GLOBAL DATA UTILITY MEASURES

The quality of data is measured from veiled data relative to the primary data. The data analyst designates a single regression model and computes the confidence intervals of original and masked data [9]. Global measures are functions of the differences between appropriateness of primary and veiled data summaries or instance of the masked and original data. Another is statistical dissemination of original and masked data. The data utility measures depend on analyst specifications and global measures. The analyst specifications depends on the analyst point of view, the view which is analyst's specific business aspect. Whereas the global measures are broad. They reflect the broad scale measures but affects the particular analysis. Global measures incorporate the nature of data swapping or adding additional noise to the data. Data utility can be measured in marginal and conditional distributions of values on variables in the original and veiled data sets. The good disclosure control method which preserve the univariate distributions of the veiled and primary data is assessed by experimental plots of the intensities and cumulative distributions.

Article [1] evaluates the four such measures, which obtains differences in the distributions of primary and veiled data. The propensity score measure evaluates the differences in disseminations of two sets of data. The second one is cluster analysis which separates the similar data with dissimilar data. The other two measures use Kolmogorov-smirov type statistics to examine the dissimilarities between experimental disseminations between primary and veiled data.

The likelihood of action assignment on observed value is the propensity score. This score lets to plan and examine an observed analysis so that it duplicates the some of the characteristics of a randomized controlled trial. Means the propensity score is a stabilizing score [2]. In this method the primary and veiled data are combined and adding variable equal to the one for all records from veiled data sets, equal to zero for all documents from original data set. In this the author calculates the probability of the veiled and original data which is the propensity score. Finally original and veiled data distributions are matched. If these distributions are similar then the data utility is high.

The propensity score is approximated via the logistic regression of veiled or primary data. Logistic regression is used to predict the chances of results based on input variables. The outcome variable of logistic regression is one that is true or false, yes or no. The similarity of the propensity score for the original and veiled observations can be assessed in various ways. The simple summary is to compute is equation (1)

$$UM_p = \frac{1}{M}\sum_{i=1}^{N}[\hat{n}_i - d]^2 \qquad (1)$$

Where M is the total number of documents in the combined data set,$\hat{n}$ is the estimated propensity score for unit i and d is the proportion of units with veiled data in the merged data set.in most of the cases the primary and veiled data have the size N_0 , In which case M=2 N_0 and c=$1/2$. The propensity scores for all units is equal to c when primary and veiled data have same distribution and UM_P is equal to zero. And if $\hat{n}$ is nearly zero for units from the primary data and one for units from veiled data the two data are distinguishable and UM_p is nearly equal to ¼.

The article [2] explains the propensity score as it allows to plan evaluate the observed study so that it impersonates some of the particular features. This article explains about the medical treatment and data collected from the treatments without changing the baseline characteristics. And also explains the propensity scores in two main steps they are estimation and application. The propensity score is itself is the estimation of the probability of observing a given value of

treatments, interventions, exposures on outcomes. The next step is application which involves the use of the approximated propensity score to make synthesized and non-synthesized data group. It also explains the four types of propensity scores. They are matching propensity score, stratification on propensity score, inverse probability of treatment weight on session is the true or false propensity score and covariate adjustment using propensity score. For the estimation of the effects of intrusions and disclosures on outcomes the randomized controlled trails are considered as the customary approach. Randomized experiments and observed studies both contains the propensity scores. The actual propensity score is known and is described by the design in randomized trails, where as in observed study the propensity score is not known. By using the study data the propensity score can be approximated. The regression model is the model in which treatment state is regressed with standard characteristics can also be used to approximate the propensity score. The article explains that the most common method for reckoning the propensity score is logistic regression.

The four propensity score method are used for removing the confusion when approximating the effects of data synthesis on results. The first one is propensity score matching which matches the one to one match of the propensity scores. It makes the pairs of synthesized and non-synthesized texts such that matched subjects have alike values of propensity scores. Then the synthesis effect can be approximated by directly matching results between synthesized and non-synthesized texts in matching sample. And if the result is continuous the effect of treatment can be approximated as the difference between the average outcome for synthesized texts and average outcome of non-synthesized data.

Stratification is one more type of propensity score checking in which the stratification on propensity score means based on the approximate propensity score mutually exclusive texts are classified. According to approximated scores the subjects are rated. The third method uses weights based on the propensity scores to create a mock sample in which the dissemination of measured covariates is independent of treatment assignment this method is called as inverse probability treatment weight using the propensity score. The fourth one is the covariate assignment using the propensity score using this approach the result variable is retreated on a pointer variable indicating the propensity score. This method considers the nature of relationship between result being sculpted and propensity scores.

The cluster analysis is the second type of global data utility measure which is explained in [1]. It is a form of unsupervised machine learning, which places records in to groups whose members have selected identical variables. The problem of locating veiled structure within untagged data is the unsupervised machine learning. The structure of data describe the interest things and which determines how leading to group the objects. A standard clustering method is K-Means clustering this is an analytical technique that for selected value of k, recognizes the k groups based on items vicinity to the middle of the group. Arithmetic average of the each cluster's dimensional vector of attributes is used to calculate the middle point or centroid.

The algorithm cluster analysis is, first chooses the value of k and k initial approximations of the centroid. And computes the distance from each data to each centroid. The distance between any two points is calculated in Cartesian coordinate form d $= \sqrt{(p_1 - p_2)^2 + (q_1 - q_2)^2}$. The centroid (x_c, y_c) can be calculated using equation (2)

$$(p_c, q_c) = \left(\frac{\sum_{i=1}^{m} p_i}{m}, \frac{\sum_{i=1}^{m} q_i}{m}\right) \quad (2)$$

Thus (p_c, q_c) is the pairs of arithmetic average of coordinates of the m points in a cluster. The above steps are repeated until the algorithm congregates to the nearest answer that is assigning each point to the nearest centroid calculated in step above, calculating the centroid of newly defined clusters and repeating till end answer [3].

Now the value for k to be chose is the main thing, it depends on the reasonable guess or predefined need. It would better to know how k-1 clusters or k+1 clusters better than the k. for this a heuristic method of using the within sum of squares (WSS) is defined in equation (3),

$$\text{WSS} = \sum_{I=1}^{M} dist(p_i - c^i)^2 \quad (3)$$

The term c^i indicates the closest centroid which is allied with the i^{th} point. WSS is very small if the points are very close. Thus the value of WSS do not prominently reduced by the k+1 clusters. But for measure of data utility the article [1] explains the cluster analysis as random separation of data set into group sizes X_a and X_b on average $X_a/(X_a + X_b)$ percent of observations in each cluster. Let primary data be OR_D and masked data is MS_d . Here the primary and veiled data sets and perform cluster analysis with fixed number of groups as the measure is given by equation (4),

$$UM_C = \frac{1}{G} \sum_{J=1}^{G} w_j \left[\frac{P_{JO}}{P_J} - C\right]^2 \quad (4)$$

Where the P_J is the number of observations in the collection P_{JO} is the number of observations on the primary data w_j is the weight of the pool. Selecting the value of G is the main issue of this measure.

The third measure is the experimental CDF measure. This measures the dissimilarities between the experimental dissemination functions of the primary and veiled data. Let P_X and P_y be the empirical distributions of the original data X and masked data. Then X has dimension as product of N_x and d. the expression is (5)

$$P_M(m_1, m_2 \ldots\ldots m_d) = \frac{1}{N_M} \sum_{j=1}^{N_M} J(m_{j1} \leq m_{j2} \ldots\ldots m_{jd} \leq m_d) \quad (5)$$

Where m_{jk} is the value of variable k for j^{th} observation and J () equals one when condition inside the parenthesis is true and equals zero otherwise. Similarly P_N is defined.

Let Q = (O, M) is the merged data having dimensions R×d $(R_O + R_M)$ ×d. the two measures of data utility from empirical distributions.

$$U_M = max|P_M(Z_i) - P_N(Z_i)| \text{ where } 1 \leq i \leq M \quad (6)$$

$$U_S = \frac{1}{N} \sum_{i=1}^{N} |P_M(Z_i) - P_N(Z_i)|^2 \quad (7)$$

U_M Is the maximum absolute distinction and U_S is the mean squared distinction or difference. This statistics have low power to detect the differences in the distributions.

III. SURVEY

The article [4] explains about the data collected from the sensor nodes which are set up in a section where the data related to that section is required for example the earth quake areas. The data is collected in time periods. By using the scheduling algorithm and joint coding scheme the utility of the data is increased. This paper emphases on general utility model that is documenting model. In this model the utility is calculated by

$$\text{MAX E [UM } (X_{[1,T]})] \tag{8}$$

Depending on the coding scheme selected the utility gain for one time period is calculated by

$$\text{E}b^{t-t_i} z_i(t) um_i(r_i(t)) \tag{9}$$

Where $b^{t-t_i} um_i(r_i(t))$ is the marginal utility gain, t_i is the at which first data is created. z_i Is the random variable dependent on r (t).The implementation of the scheduling algorithm is the main issue in this model, as each time, the sensors need to hoard the data.

The article [6] explains that the useful estimates can be discovered from the raw data. This article uses some experimental results on the clinical data. A different way of measuring the effectiveness of sanitized data is investigated in this article. The sanitized data is applied with a measures of gathering the information using some prediction based methods. Then it presents likelihood systems and illustrates how utility specified by such systems can be calculated.

The author uses the knowledge discovery system. It is a prediction based method. Conditions from predictions is the procedures to be discovered from predictions. The system uses three algorithms they are data mining algorithm, prediction algorithm and conflict resolution algorithm. The data mining algorithm extracts the outlines from the data and the rules become the results of data. The prediction algorithm takes the rules from the data mining algorithm and results some likelihoods of information to those rules. The third algorithm is a conflict resolution algorithm decides the best match from many likelihoods.

The formula $um_{a\to b}$ measures the utility of information discovered from which primary or sanitized learning data is applied to b. Here are the two expressions for computing utility

$$Utility_{\,orig\to orig} = \sum_{m\in T} w(m) \sum_{i\in I} E_i \left(CR\big(PA(DM(DB),m)\big),\overline{m}\right) \tag{10}$$

$$Utility_{\,orig\to san} = \sum_{m\in T} w(m) \sum_{i\in I} E_i \left(CR\left(PA\big(DM(DB),P(m)\big)\right),P(\overline{m})\right) \tag{11}$$

The results of experiments shows that the utility is increased to above 70% by applying the prediction rules compared to non-sanitized data.

Article [7] explains for the various data type the utility of the veiled data type can be evaluated. The model explained in this paper shows that the utility of veiled data can be evaluated effectively. This paper implements the data utility measure as homogeneous difference methods and statistical difference methods for continuous information. The direct comparison method and entropy comparison method is used for categorial information.

Euclidean distance is used to measure the distance of two records for continuous data. As in the global measures of data. For statistical data the normalized information vector $\overline{A_i}$ can be used that is,

$$\overline{A_I} = \frac{1}{m_i}\sum_{j=1}^{m_i} A_{ij} \tag{12}$$

It uses the sum of squares and sum of table squares. The data utility and information loss is measured as the proportion of sum of squared errors and sum of table squared errors. Information loss is measured as an overall view in homogeneous difference method. The statistical difference methods evaluates the information loss from micro view. The measure of information loss from the statistical difference between the veiled and primary data are mean square error, mean absolute error and mean variation.

The information loss measurement for categorial data is done in two methods. They are direct comparison and entropy based. In direct comparison, comparison of matrices require the distances for categorial data. Categorial data have two types nominal data and ordinal data. The nominal and ordinal data the distance functions are as follows,

$$\text{D}\quad (p_i, p_i') \;=\; \begin{cases} 0 \text{ if}(p_i = p) \\ 1 \;if\; (p_i = p_i') \end{cases} \qquad (c_i \text{ is nominal data}) \tag{13}$$

$$\text{D}\ (p_i, p_i') = \frac{|\{p_i'' \,|\min(p_i,p_i') \le p_i'' \le \max(p_i,p_i')\}|}{D(i)} \tag{14}$$

Where $p$ordinal data D (i) is the cardinality of is attribute i.

If p' be an unknown table of p, p and p' be tuples of P and P'respectively,p_i and p_i' be the attributes of p and p'. Then information loss can be defined as

$$IL_{DC} = \sum_{j=1}^{n} d(p_i, p_i') \quad \text{where} \quad \text{d}\ (t,t') = \sum_{i=1}^{p} d(p_i, p_i') \tag{15}$$

Entropy comparison, the entropy can be calculated as the information loss

$$H_J = -\sum_{I=1}^{L} p_i \log_2 p_i \tag{16}$$

p_i is the probability of the elements of order belonging to the particular category.

The information loss is calculated as

$$\text{ILR} = \frac{|Original\ entropy - new\ entropy|}{original\ entropy} \times 100 \tag{17}$$

Article [3] explains about the utility measures as the same as in article [7]. In anonymization literature, the sum of squared errors is the

measure used. The SSE is defined as the sum of squares of distances between primary record and veiled sets

$$\text{SSE} = \sum_{a_j \in A} (distance(a_j, a'_j))^2 \quad (18)$$

This paper proposes the empirical results which reduces the information loss by several orders of magnitude.

Article [1] explains the ambiguity as anonymization technique for privacy model. The data utility is maintained at satisfactory level in this paper. The author evaluates the information loss using the relative error.

Article [5] explains an online scheduling algorithm based on distributed correlated scheduling. The theoretical analysis shows that it accomplishes high average data utility. The article defines the data utility of the region as the data utility $p_i(t)$ produced by smartphones in the i^{th} region of interest in time slot t is defined as

$$p_i(\text{t}) = \widehat{p}_i(s_i(t), a_i(t)) \quad (19)$$

P_i is a constant. Such utility model is a special case of marginal effect. It also define the aggregate data utility of all regions as

$$P(t) = \sum_{i=1}^{M} log(1 + K_i p_i(t)) \quad (20)$$

The utility optimality is achieved by using these algorithms. Article [2] proposes a privacy preserving utility verification based in diff part. This proposal can measure the data utility based on the encrypted frequencies of the aggregated raw data instead of the plain values, which thus prevents privacy breach. The utility of set valued data published by diffpart is measured by average relative errors of counting queries. For a give item set a counting query Q over a dataset with respect to the result over raw dataset is defined as

$$\text{M (Q)} = \frac{|P(\bar{D}) - P(D)|}{MAX\{P(D), s\}} \quad (21)$$

This paper defines a new metric for data utility measurement

$$\text{M}(T_r) = \frac{1}{n} \sum_{i=1}^{n} \frac{(fr'_i - fr_i)^2}{(\max(fr'_i, s))^2} \quad (22)$$

fr_i fr'_i Are the original and the perturbed frequencies of the i-th record.

IV. CONCLUSION

In this paper the different utility measures are considered. Most of the articles uses the mean squared errors and distance variables as the utility measures. The recent article uses the encrypted frequencies of the aggregated raw data instead of plain values which prevents the privacy breach. Other articles use the different utility measures depending on the use cases and analysts particular choice of interest based on analysis perspective.

V. REFERENCES

[1] M. R. Haghjoo, "An improved Ambiguity+ anonymization technique with enhanced data utility," *Information and Knowledge Technology (IKT), 2015 7th Conference on, Urmia,* no. 10.1109/IKT.2015.7288743, pp. 1-7, 2015.

[2] A. F. Y. S. Z. S. Z. Hua J. Tang, ""Privacy-Preserving Utility Verification of the Data Published by Non-interactive Differentially Private Mechanisms,"," *in IEEE Transactions on Information Forensics and Security,* vol. PP, no. 10.1109/TIFS.2016.2532839, 2016.

[3] J. S.-C. J. D.-F. D. S. S. Martinez, "Improving the Utility of Differentially Private Data Releases via k-Anonymity," *Trust, Security and Privacy in Computing and Communications (TrustCom), 2013 12th IEEE International Conference on, Melbourne,* no. 10.1109/TrustCom.2013.47, pp. 372-379, 2013.

[4] X. J. L. Z. L. D. T. C. H., "Maximizing the Data Utility of a Data Archiving & Querying System through Joint Coding and Scheduling," *Information Processing in Sensor Networks, 2007. IPSN 2007. 6th International Symposium on, Cambridge, MA,* no. 10.1109/IPSN.2007.4379684, pp. 244-253., 2007.

[5] Y. Z. J. Y. Yang Han, "Utility-maximizing data collection in crowd sensing: An optimal scheduling approach,," *Sensing, Communication, and Networking (SECON), 2015 12th Annual IEEE International Conference on, Seattle, WA,* no. 10.1109/SAHCN.2015.7338334, pp. 345-353, 2015.

[6] R. S.-N. J. D. M. A. J. G. M. Sramka, "Utility of Knowledge Extracted from Unsanitized Data when Applied to Sanitized Data," *Privacy, Security and Trust, 2008. PST '08. Sixth Annual Conference on, Fredericton, NB,* no. 10.1109/PST.2008.30, pp. 227-231, 2008.

[7] H. J. W. Lixia, "Utility evaluation of K-anonymous data by microaggregation," *Computing, Communication, Control, and Management, 2009. CCCM 2009. ISECS International Colloquium on, Sanya,,* no. 10.1109/CCCM.2009.5270417, pp. 381-384, 2009.

[8] F. P. a. T. Menzies, "Privacy and Utility for Defect," *Proc. 34th Int'l Conf.,* pp. 189-199, Jun 2012.

[9] I. D. a. K. Nissim, "Revealing Information While Preserving," *Proc. 22nd ACM SIGMOD-SIGACT-SIGART Symp.,* pp. 202-210, 2003.