Extreme Value Protection Adjustment for Different Subpopulations in Complex Datasets

Anna Oganian[1]
Joint work with Mehtab Iqbal[2] and Goran Lesaja[3,4]

[1]National Center for Health Statistics, MD, USA
[2]Clemson University SC, USA
[3]Georgia Southern University, GA, USA
[4]US Naval Academy, MD, USA

Joint UNECE/Eurostat Expert Meeting on Statistical Data Confidentiality
December 2nd 2021

Disclaimer

Ideas, findings and conclusions in this presentation are those of the authors and do not necessarily represent the official position and neither current practice at the National Center for Health Statistics (USA), Centers for Disease Control and Prevention.

Introduction

- Public release of microdata sets with many variables of different types is very important for research, policymaking and training.
- To produce a good quality public data set, the data protector needs to account for the relationships between these variables.
- Data protectors are often concerned about disclosure risk associated with the extreme values of numerical variables.
- Thus, such observations are often top or bottom-coded or synthesized.

- A single rule for defining which records and values can be considered as extreme can lead to under-protection for some subpopulations and over-protection for others.
- This disparity in protection is evident when a certain subpopulation is different from other subpopulations in terms of a specific variable subject to protection/modification.

  Example: males and females of different race/ethnicity groups may have different definitions of "extreme" for variables such as height and weight.

In this presentation, we discuss varying definitions of   extreme values for different subpopulations.

Extremes of continuous variables

- The definition of “extreme” depends on the context and can vary among different subpopulations.

  Example: data protector synthesizes extreme values of income, where extremes are defined as income above $\$150{,}000$. Such a threshold may be inadequate for certain groups of individuals: those who work part-time, unemployed (part of the year), individuals with low level of education, and for certain occupation classes [1].

  Such individuals might not be adequately protected because the income of $\$150{,}000$ may be very rare within these groups, much rarer than in other groups, and hence might have a higher risk of re-identification.

- In [2], a procedure that can be used to determine variable thresholds for top-coding for different subpopulations is described.
- If synthesis is the preferred method for extreme value protection, various definitions of extremes can be used. One approach is to use Tukey's fences and synthesize the outliers:

$$Thresh_{sub} = Q3_{sub} + k * IQR_{sub},$$

where k can be 1.5 (possible outlier) or 3 (for extreme (see [3]).

In this presentation, we use $k = 3$ to illustrate our method. However, the data protector can choose $k = 1.5$ if a more conservative definition of an outlier is preferred in a certain scenario of data release.

Goals

- The main goal of our algorithm is to identify the subpopulations that may need their own thresholds defining extremes of numerical variables.
- While in some instances it may be intuitive and easy, such a task is not always trivial in datasets with large number of variables.
- Usage of effective data mining tools may be helpful for that.

Approach

- Group the variables in the data set into clusters around numerical variables T_i subject to protection. Distance between the variables may be measured by the squared canonical correlation, the higher the correlation - the closer the variables. The closest variables to T_i (at a distance less than some chosen threshold value h) are included into the corresponding cluster
- Form the rules defining extremes using techniques of Association Rule Mining (ARM) within the clusters of variables obtained on the previous step.
- Synthesize extreme values defined by the rules or apply other preferred method of disclosure limitation.

Algorithm

1. Compute $Thresh_{pop} = Q_{3_{pop}} + 3IQR_{pop}$ for T_i using all the records in the dataset. It will serve as a benchmark.
2. On the vertical partition of the data, $Clust_i$, mine the rules of the following type:

   $$X \to T_i < Thresh_{pop} - \Delta \qquad (1)$$

   Choose the rules with the confidence larger than

   $$supp\left(T_i < \ Thresh_{pop} - \Delta\right)$$

   Denote this set S.
3. Filter the set S: only those rules which refer to the subpopulations satisfying the following expression:

   $$Thresh_{sub} < Thresh_{pop} - \Delta \qquad (2)$$

   are retained.

Algorithm (cont-ed)

4. The LHS of the rules satisfying expression (2) describes the subpopulations that may need adjustment in the definitions of extreme values of T_i.
5. Compute $Thresh_{sub} = Q_{3_{sub}} + 3IQR_{sub}$ - actual thresholds for extremes of T_i for the identified subpopulations.
6. Synthesize the outliers above these thresholds using the preferred synthetic method or some other method of disclosure limitation.

Note:

The choice of the $\mathrm{Conf} = supp\left(T_i < \ Thresh_{pop} - \Delta\right)$ is based on the recommendations in the literature [4] to use the rules with $Lift \ > 1.$

For the association rules of the type $LHS \ -> \ RHS$, where LHS and RHS denote the antecedent and consequent of the rule correspondingly, $Lift$ is one of the measures of interest. It is defined as follows:

$$Lift(LHS \ \rightarrow RHS) =$$

$$\frac{Supp(LHS \ \cup RHS)}{Supp(LHS) * Supp(RHS)} = \frac{Conf(LHS \ \rightarrow RHS)}{Supp(RHS)}$$

If $Lift(LHS \rightarrow RHS) > 1$ then $Conf(LHS \rightarrow RHS) > Supp(RHS)$.

Thus, when mining rules of the type:

$$X \rightarrow T_i < Thresh_{pop} - \Delta$$

we choose the ones with $Conf > supp(T_i < Thresh_{pop} - \Delta)$.

Note, that $supp(T_i < Thresh_{pop} - \Delta)$ is computed on the entire dataset and thus, it can be done before mining process begins. It equals to the proportion of records in the entire dataset for which $(T_i < Thresh_{pop} - \Delta)$ is True.

Numerical Experiments

Census public data set [5].

This is a sample drawn from the Public Use Microdata Samples (PUMS) person 1990 US Census file.

1.2 million records and 66 variables of different types, numerical and categorical.

We illustrate our approach for variable $Income1$ - wages or salary earned by the individuals in 1989.

First, we clustered the variables in Census data around $Income1$.

Clustering Census data set produced a cluster for the variable $Income1$ with the following categorical and numerical variables:

Categorical variables:

$Class$ – class of worker (private for profit comp, private for non profit comp, local gov, state gov, etc.),
$Relat1$ – Relationship (Householder, Husband/wife, Son/Daughter, etc.),
$Occupclass$ – occupation class (Manag., Profes., Techn., Military, etc.),
$IndustryClass$ – industry class (Agriculture, Mining, Manufacturing, etc),
$Disable1$ - work limitation (yes, no),
$Rlabor$ – employment status (Civilian at work, Civilian not at work, Armed forces at work, etc.),
$Yearsch$ – Ed. Attainment (9th grade or less, 10th grade, 11th grade,..., PhD)

Numerical variables:

$Hour89$ - usual hrs. worked per week in 1989,
$Week89$ – weeks worked in 1989.

To make interpretation of the rules easier and to reduce the number of rules, we converted the variables above to categorical $W89$ and $H89$:

If $WEEK89 < 26$ then $W89 = 1$ and 0 otherwise.
If $HOUR89 < 30$ then $H89 = 1$ and 0 otherwise.

$W89 = 1$ represents those individuals who worked less than half of the year, including seasonal workers and those who were unemployed for at least one-half of 1989.

$H89 = 1$ can be thought as an indicator for part-time workers (similar definition is given in [1]).

Some examples of the rules learned

For the groups of individuals that fit the description that appears in the LHS of the rules, the threshold that defines outliers for $Income1$ according to the IQR rule appears on the RHS of the rules.

It is worth noting, that the threshold for outliers for $Income1$ computed on the entire dataset is 92,000.

$W89 = \text{Part-year worker} \rightarrow Income1 < \$16{,}400$

$H89 = \text{Part-time worker} \rightarrow Income1 < \$22{,}300$

$Relat1 = \text{Persons in group quarters} \rightarrow Income1 < \$29{,}300$

$Relat1 = \text{Other relat. of the householder} \rightarrow Income1 < \$51{,}200$

$Relat1 = \text{Son/daught. of the householder} \rightarrow Income1 < \$40{,}000$

Some examples of the rules learned (cont-ed)

$Occupclass$ = Service $\rightarrow Income1 < \$41{,}300$

$Occupclass = Farming \rightarrow Income1 < \$51{,}300$

$Relat1$ = Husb/wife of the housh. ∧
∧ $Yearsch$ = HS or less $\rightarrow Income1 < \$47{,}100$

$Class$ = Emp. of priv. not for prof comp ∧ $Yearsch$ = HS or less
$\rightarrow Income1 < \$50{,}000$

$Industryclass$ = Trade ∧ $Yearsch$ = HS or less
$\rightarrow Income1 < \$50{,}000$

$Industryclass$ = Agriculture ∧ $Yearsch$ = HS or less
$\rightarrow Income1 < \$41{,}200$

Synthesis of extreme values and assessment of the results

- After the rules were obtained, we synthesized the extreme values of $Income1$ according to the rules that we obtained.
- We used the R package Synthpop and a method based on Classification and Regression Trees (CART).
- Because we synthesize only the extreme values which are defined by the association rules described above, CART models were estimated using only the records with such extreme values, as it is recommended in [6], in order to maximize the utility of the resulting masked data.

- To assess the quality of the masked data, we computed several utility metrics and compared some of the univariate statistics of $Income1$ based on the original and masked data.
- Regarding univariate statistics, the mean of the masked $Income1$ is $\$21{,}644$ and the mean of the original data is $\$21{,}627$. The standard deviation for the masked data is $\$22{,}038$, and $\$22{,}007$ for the original data. Medians and interquartile ranges are exactly the same in the masked and the original data, being $\$17{,}000$ and $\$21{,}000$ respectively.
- Regarding utility metrics, we computed a confidence interval overlap measure [7] which represents a probability overlap in the confidence intervals for the regression coefficients computed on the masked and the original data.

- The regression model for this measure has $Income1$ as a predicted variable and all other numerical variables (from the same cluster) as its predictors. Average confidence interval overlap is 90% which is good because maximum probability overlap is 95%.
- We also computed a generic measure of data utility based on propensity scores [8]. The propensity score utility measure for this data is 0.05, which can be interpreted as a relatively good utility [8]. For reference, the possible range of values for this measure is 0 to 0.25, smaller values meaning better utility, with 0 corresponding to the case when original and masked data are identical.

- To assess the disclosure risk, we computed the absolute differences between the original and synthesized values of $Income1$. Average differences ranged between $\$14{,}000$ and $\$50{,}000$ among different subpopulations where extreme values were synthesized. We also computed the average percentage change:

$$\frac{|orig_i \; - \; syn_i|}{|orig\_i|} * 100\%$$

where $orig_i$ is the i-th original value of $Income1$ and syn_i - its corresponding synthetic value. The average percentage change was about 40% for this dataset.

Notes:

- Depending on the practice at a particular institution and the scenario of data release, data protectors can achieve different levels by varying method parameters and applying more or less intensely the protection method.
- For example, in the case of synthesis using CART, the data protector can specify a minimum size of a final node that a CART model can produce. This can help to decrease the disclosure risk even further - details can be found in [9].

Conclusions

- The analysis of the rules obtained by our procedure should be done by the data protector and subject area specialist for each particular data set and the scenario of data release.
- The association rules found by the proposed approach are meant to bring to the data protector’s attention particular combinations of the attributes that are rarely associated with the extreme values of the numerical variable that is subject to protection.
- This procedure may be used as an aid for the data collecting organizations in the disclosure review process as an alternative, or in addition, to their regular procedures.

- Identification of risky combinations of the attributes in big data sets, for example big surveys, is a complicated and lengthy process. Thus, an automatic or semi-automatic procedure can be very helpful. In this presentation we outlined an example of such procedure.
- In general, setting up particular thresholds for extreme values of numerical variables is a policy decision and depends on the practice of disclosure limitation at a particular agency.

References

[1] Ross,M., Bateman,N.: Meet the low-wage workforce. Tech.rep., Brookings (2019)

[2] Oganian,A., Iacob,I., Lesaja,G.: Multivariate Top-Coding for Statistical Disclosure Limitation. In: Domingo-Ferrer J, Muralidhar K. (ed.) Privacy in Statistical Databases, Lecture Notes in Computer Science. vol. 12276, pp. 136–148. Springer-Verlag (2020)

[3] Tukey,J.W.: Exploratory Data Analysis. Addison-Wesley (1977)

[4] Tufféry, S.: Data Mining and Statistics for Decision Making. Chichester, GB: John Wiley and Sons (2011)

[5] Census: US census (1990) data set. UCI Machine Learning Repository (2017), https://archive.ics.uci.edu/ml/datasets/US+Census+Data+%81990%29

[6] Reiter,J.: Using cart to generate partially synthetic public use microdata. Journal of Official Statistics 21(3), 441–462 (2005)

[7] Karr,A.F., Kohnen,C.N., Oganian,A., Reiter,J.P., Sanil,A.P.: A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality. The American Statistician 60(3), 224–232 (2006)

[8] Woo, M.J., Reiter, J.P., Oganyan, A., Karr, A.F.: Global measures of data utility for microdata masked for disclosure limitation. Journal of Privacy and Confidentiality 1(1), 111–124 (2009)

[9] Nowok,B., Raab,G., Dibben,C.: Synthpop: Bespoke creation of synthetic data in R. Journal of Statistical Software 74, 1–26 (2016), https://www.jstatsoft.org/article/view/v074i11