Data Synthesis

for Precision Medicine

Gunjan Chandra

08.06.2021

Background

- Over 2.5 Quintilian bytes of data are created every day
- Benefits of data:
  - Solve problems
  - Maintain performances
  - Improve existing processes
  - Finding new knowledge
  - Verify previously made verdicts
- As of 2013 only 0.5% of the total data was analysed
  - No accessibility to data
    - Sensitive data

Data accesibility

- Sensitive information could be utilized for unethical activities
- Clinical data is required to be anonymized before leaving the hospital
  - Altering and removing explicit identifiers
    - A person can still be re-identified by data linking [1]
  - Use data aggregation techniques and induce random noise to the data
    - Noise can often be removed by averaging responses for carefully selected query sets
    - Distortion of the relationship between variables

Clinical Data

- Complex content and structure of modern healthcare databases
- Expense of producing and sustaining comprehensive databases
- Data anonymization techniques are not foolproof and hinder the opportunity of personalized evaluations
  - Patient's identity must be relinked to the data analytic results
  - Medical data cannot be fully and irreversibly anonymized

Synthpop

- Synthetic data set is created by replacing some or all observed values by sampling from an appropriate probability distribution, conditional on:
  - The variable to be synthesized,
  - The values from all previously synthesized columns of the original data set, and
    - The fitted parameters of the conditional distribution (simple synthesis)

      or

    - posterior predictive distribution of parameters (proper synthesis)
- while retaining the statistical properties of the original data set and relationships between the variables

Impacts of Data Synthesis

- Utility Measures of Data
  - General Utility
    - Overall similarities in the statistical properties and multivariate relationships
  - Specific Utility
    - Performance similarity of a fitted model

$$H_o : C^*\{t(D), t(S_i)\} \geqslant \alpha, \qquad \text{for all } t \in [0, \tau]$$

$$H_a : C^*\{t(D), t(S_i)\} < \alpha, \qquad \text{for any } t \in [0, \tau]$$

Let D denote an original data set, and Si denotes a synthetic data set where i indicates the index for synthetic data produced with the different synthesizing method. Let t denote a vector of tests which returns a statistic, and C^* be a comparison function which returns a $p - value$. Finally, comparing the output of C^* with α, a threshold value for the level of significance. The α is set to 0.05 for all tests.

Impacts of Data Synthesis

- Quality of Information content
  - Entropy

$$H(X) = -\sum_{i} P_X(x_i) log_b P_X(x_i)$$

    - If the system moves away from equally likely outcomes or introduces some predictability, the entropy goes down
  - Mutual Information

$$I(X;Y) = H(Y) - H(Y|X)$$

    - The amount of information or reduction in uncertainty that one random variable provides about the other

Type 1 Diabetes Prediction and Prevention data set (DIPP)

- Finland has the highest incidence of Type 1 Diabetes (T1D) in the world amongst young children. Approximately 72 in every 100,000 children under the age of 15 years
- The DIPP Study was established in 1994
- Population-based long-term clinical follow-up study that consists of screening newborns for increased genetic risk for diabetes
- Predict the probability of the positivity of autoantibodies before the age of 15 years

Synthesis of DIPP data set

- Synthesis using 5 different methods

Synthetic data	Method	Description
SynD1	"cart"	classification and regression tree
SynD2	"ctree"	classification tree
SynD3	"rf"	random forest
SynD4	"bag"	bagging
SynD5	"parametric"	parametric* method to each variable based on their data type

- Data set was divided into three splits before model fitting, 75.0% of data for training, 12.5% for validation, and 12.5% for testing

Specific Utility

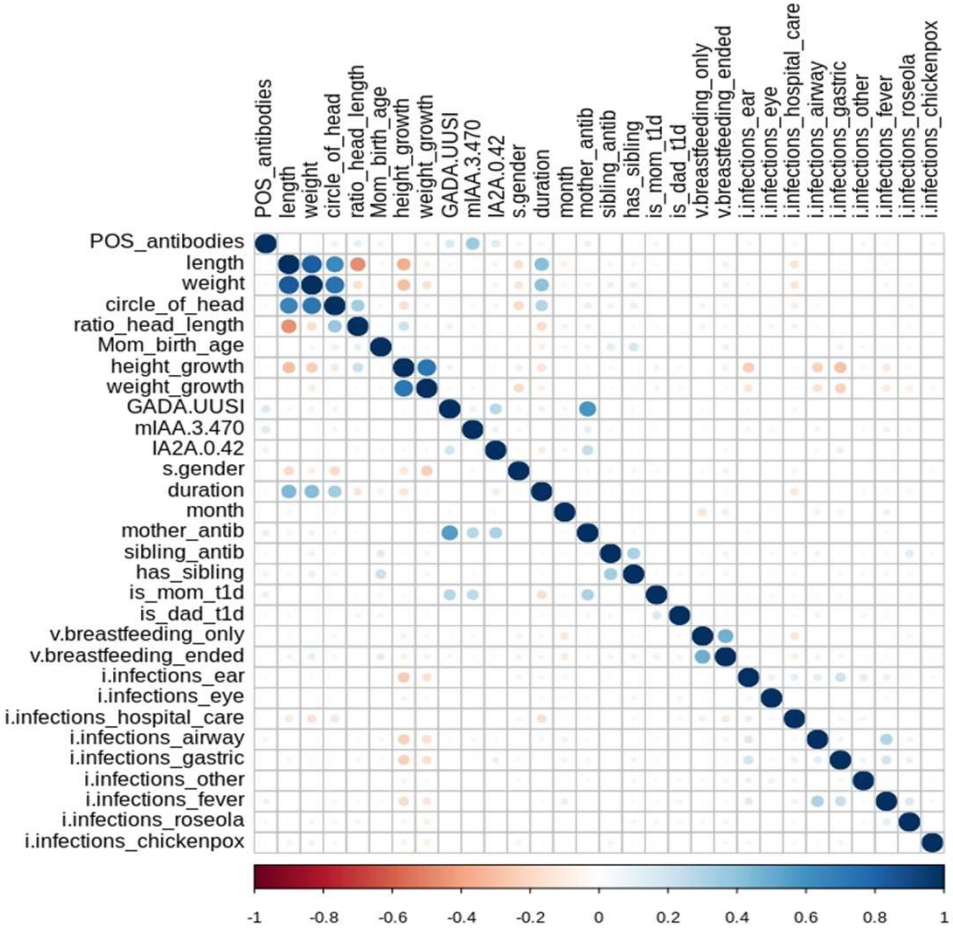
- Performance of synthetic and original data on Gradient Boosted regression Model
- CART performs best out of all 5 methods

Data set	P-value
SynD1	0.0965496
SynD2	0.0485093
SynD3	0.0026730
SynD4	0.0288157
SynD5	0.1755973

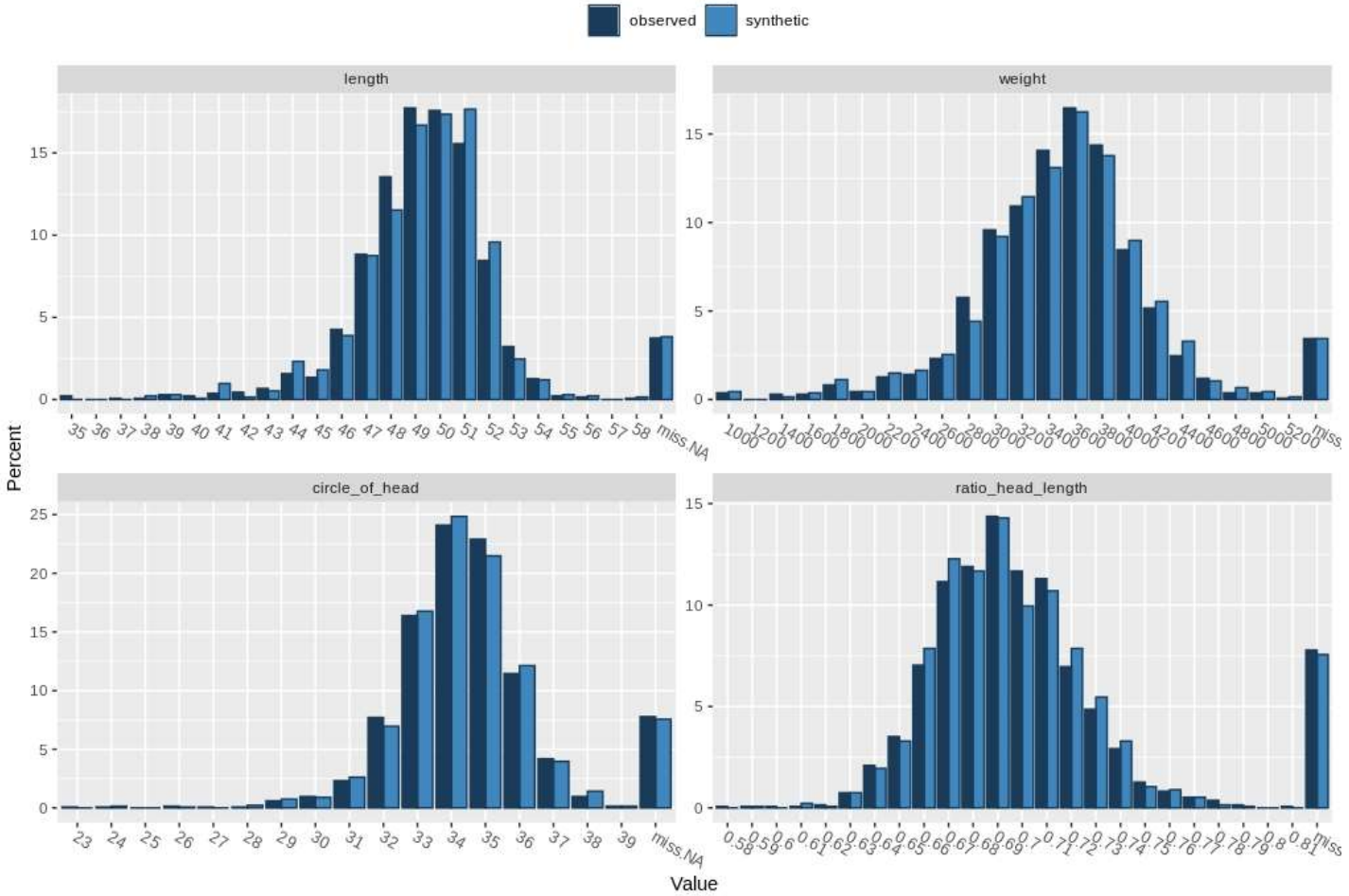
Data set	Confusion Matrix			Evaluation Parameter		Accuracy
Original Data		Predicted labels		F1 score	Area Under ROC	0.87
		Negative	Positive			
	Negative	89	16	0.85	0.95	
	Positive	5	56	0.82		
SynD1		Predicted labels		F1 score	Area Under ROC	0.88
		Negative	Positive			
	Negative	83	19	0.88	0.93	
	Positive	1	63	0.85		
SynD2		Predicted labels		F1 score	Area Under ROC	0.86
		Negative	Positive			
	Negative	82	20	0.87	0.93	
	Positive	3	61	0.82		
SynD3		Predicted labels		F1 score	Area Under ROC	0.90
		Negative	Positive			
	Negative	89	13	0.91	0.95	
	Positive	4	60	0.87		
SynD4		Predicted labels		F1 score	Area Under ROC	0.93
		Negative	Positive			
	Negative	90	12	0.92	0.97	
	Positive	0	64	0.89		
SynD5		Predicted labels		F1 score	Area Under ROC	0.88
		Negative	Positive			
	Negative	98	4	0.85	0.92	
	Positive	15	49	0.78		

General Utility

- Pearson Correlation
  - Original data on lower triangle and synthetic data on upper triangle
  - Correlation between POS_antibodies and IAA antibody is stronger in the SynD1 data set, ρ-value is 0.13 in original and 0.37 in synthetic data set

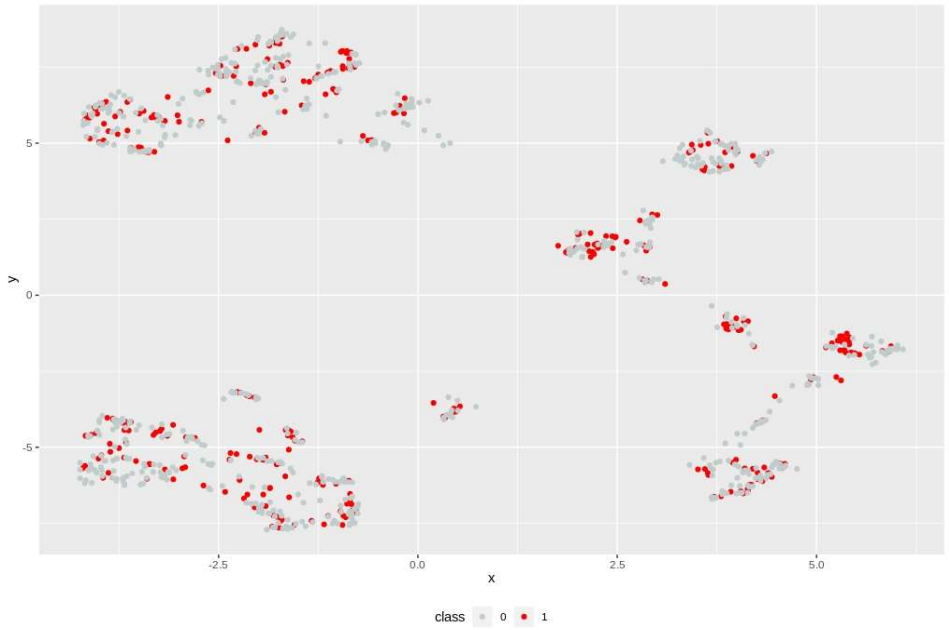


Compare: Relative Frequency Distribution

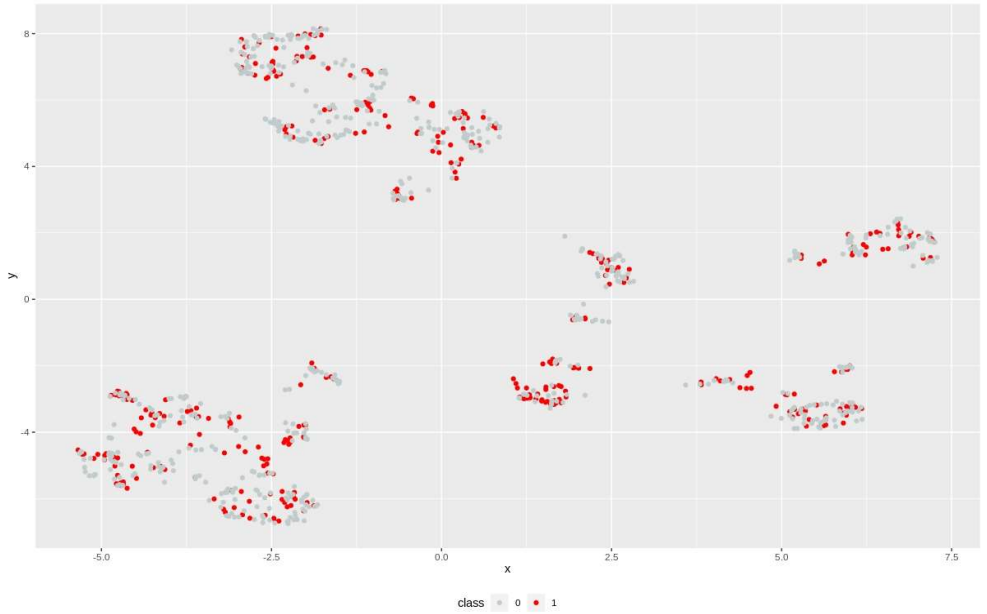


Uniform Maniford Approximation and Projection

Original

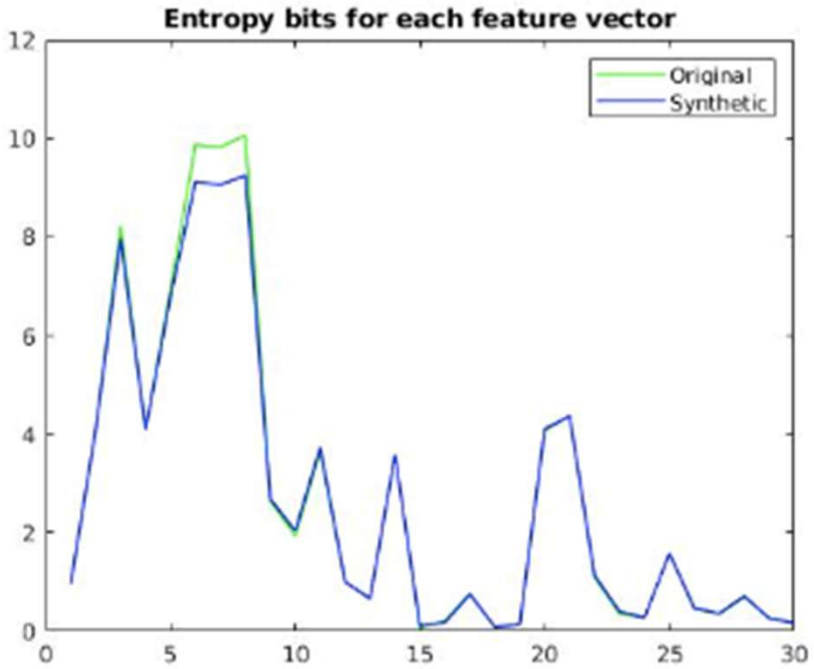


Synthetic



Quality of Information content

- Entropy
  - Entropy/Uncertainty is maximum when all outcomes are equally likely
  - Variables such as the age of mother at the time of birth, growth rate of height and weight had a decrease in entropy by approximately one bit
- Mutual Information
  - MI between original and synthetic datasets[2][3] remained unchanged



References

- [1] Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10, 05 (2002), 557–570. DOI: 10.1142/S0218488502001648.
- [2] Alexander Kraskov, Harald Stogbauer, Peter Grassberger, "Estimating mutual information", Phys. Rev. E, vol. 69, no. 066138, Jun 2004.
- [3] Ian Oliver, Yoan Miche, "On the Development of a Metric for Quality of Information Content over Anonymised Data-Sets", Quality of Information and Communications Technology (QUATIC) 2016 10th International Conference.

Thank you!


UNIVERSITY
OF OULU