COMIDENCO

Constructing an Incidence Model for Dengue Fever

applied to Paraguayan communities

a CONACYT project by

CIMA ● FP-UNA ● UNC ● UNCA ● CEDIC

Paraguay

Presenter: Santiago Gómez-Guerrero

The Project: General Objectives

- Comidenco aims at constructing a predictive model, focused on (but not restricted to) incidence as response variable.
- The model would take local variables including anti-dengue actions, evaluate the probability of spread of the disease, and predict incidence.
- This will identify communities with greater danger of an increase in infection rate, helping to decide where to put resources into action.

The Project: Specific Goals

- Use historical and current data to identify preditor variables for certain neighborhoods in selected cities.
- Generalize to similar communities in Paraguay.
- Build software to show heat maps for the disease, and support decision making.

Preliminary Explorations

- Using only local data from the department (state) of Concepción, we are able to construct SVM models using R language packages.
- Also a few descriptive statistics in graphical and table form.

Measuring Multivariate Correlation

- Numerous variables involved in dengue modeling, many of them being ordinal or categorical.
- Correlation between variables needs to be evaluated for proper feature grouping and selection.
- Measures of correlation exist for numerical variables, but there is no reliable measure for bivariate or multivariate correlation among categorical/ordinal variables.
- The symmetrical uncertainty (SU) is a recently proposed entropy-based measure of correlation between 2 categorical variables. Exploring the behavior of SU evidenced the need for an *n*-variable measure.

Introducing MSU

Thus we extend the SU, introducing the *n*-dimensional or multivariate SU:

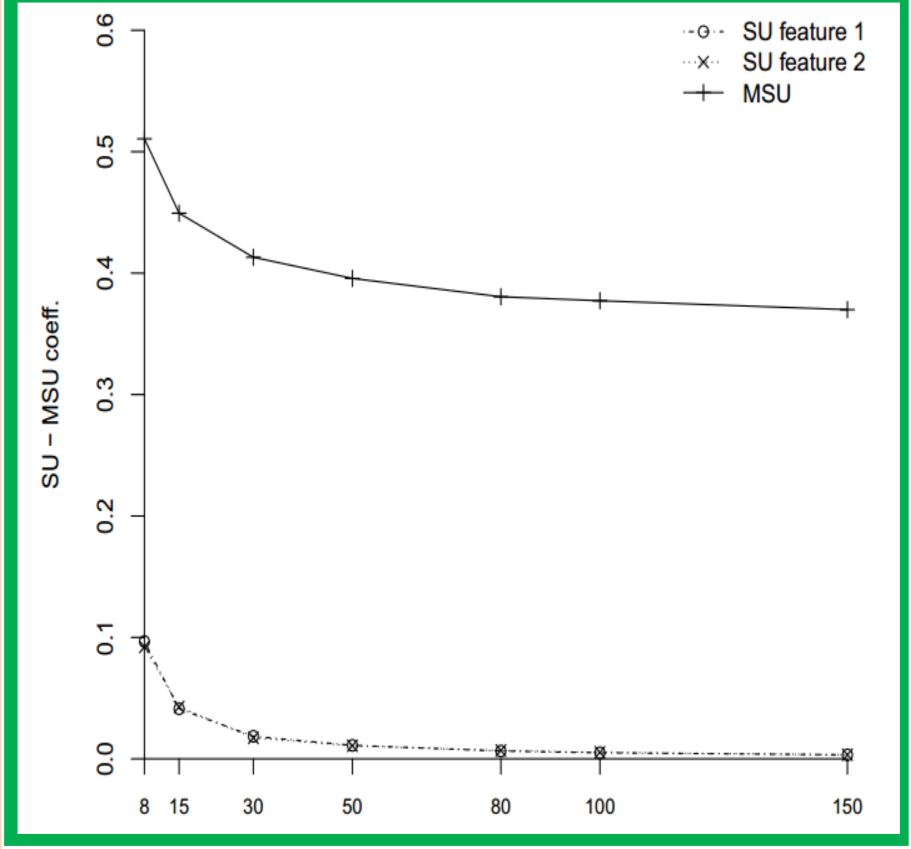
$$MSU(X_{1:n}) := \frac{n}{n-1}\left[1 - \frac{H(X_{1:n})}{\sum_{i=1}^{n} H(X_i)}.\right]$$

where *H*(*X*) is the entropy of random variable *X*.

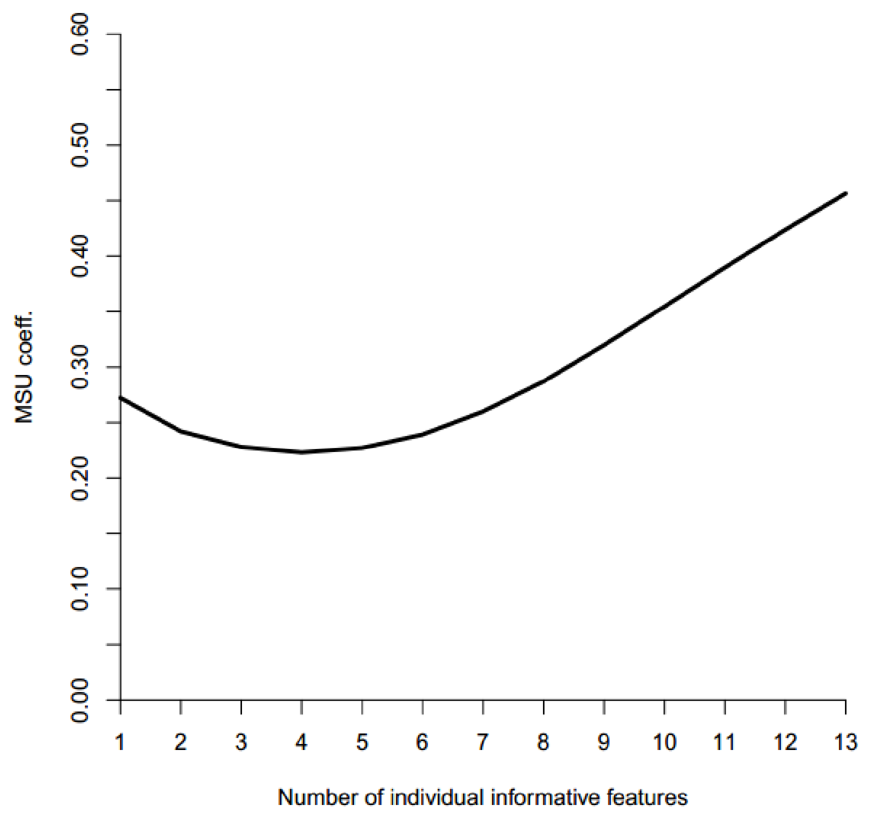
We evaluated the MSU over synthetic datasets, and verified that its behavior depends on: **number** of features, their **informativeness**, their **cardinalities** and the **sample size**.

MSU Behavior

Detection of Three-Way Collective Correlation



Bias because of Too Many Informative Features Considered



MSU Behavior

What we have found:

Property of MSU	How found	Consequence
Detects bivariate correlations	experiment	Nice property
Detects *n*-variable collective correlations (*n*-variable interactions)	experiment	Nice property
Values are high with informative features, low with non-informative ones	experiment	Nice property
Values become inflated when attributes have high cardinalities	experiment	Need for representativeness
If a category is missing in the sample, the MSU value is an under-estimation of the real MSU	proved	Need for representativeness
Converges to true value for large sample size *m*	experiment	What should *m* be?

Correcting Matters

Concept: A sample where each attribute has no missing category will be called a *totally representative* sample.

Calculating sample size m: Given attributes X_1, …, X_n the combinations of their labels can be seen as values of a multinomial variable.

Take conservative assumption of independent attributes, to yield maximum entropy. Using confidence intervals, we find that with probability $1 - \alpha$ a sample of size

$$m > z_{\alpha}^{2} \frac{1 - p_i}{p_i} \quad \forall i$$

is totally representative. The p_i are the probabilities in the multinomial distribution generated from all of the labels.

MSU, a Reliable Measure

- We have verified that the MSU measure properly detects bivariate and multivariate correlations in a set of *m* observations of *n* categorical or ordinal variables.
- Under the maximum entropy assumption, it is now possible to control the precision of the measure, by first determining a minimum sample size *m*.
- MSU can be used as a reliable piece in the upcoming feature selection process for our project.

Roads to Cooperation

- Some of our current needs …
  - Use data with city or community granularity.
  - Adapt MSU to handle categorical *and* real variables.
  - Strategies to incorporate the MSU measure into feature selection processes.
  - Use available experience in dengue modeling so as to correctly identify relevant variables.
- Possible collaboration with the InfoDengue team.

http://www.cimapy.org/es/investigacion/proyectos/comidenco
https://CRAN.R-project.org/package=msu

Questions

???

http://www.cimapy.org/es/investigacion/proyectos/comidenco
https://CRAN.R-project.org/package=msu
sgomezpy@gmail.com