Machine Learning Assists the Classification of Reports by Citizens on Disease-Carrying Mosquitoes

Antonio Rodriguez[1] Frederic Bartumeus[2,3,4] Ricard Gavaldà[1]

Universitat Politècnica de Catalunya, Barcelona (Spain)

Centre for Advanced Studies of Blanes (CEAB-CSIC), 17300 Girona (Spain)

CREAF, Cerdanyola del Vallès, 08193 Barcelona (Spain)

ICREA, Pg Lluís Companys 23, 08010 Barcelona (Spain)

Workshop on Data Science for Social Good, SoGood
September 2016

Overview

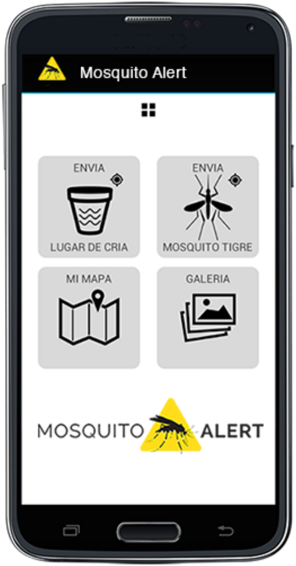
1. Introduction
2. Methodology
3. Project development
    - Exploratory data analysis
    - Data cleaning and pre-processing
    - Classifier training, evaluation and selection
    - Real-time classification system design
4. Discussion
5. Future work

Introduction - Mosquito Alert

MOSQUITO ALERT

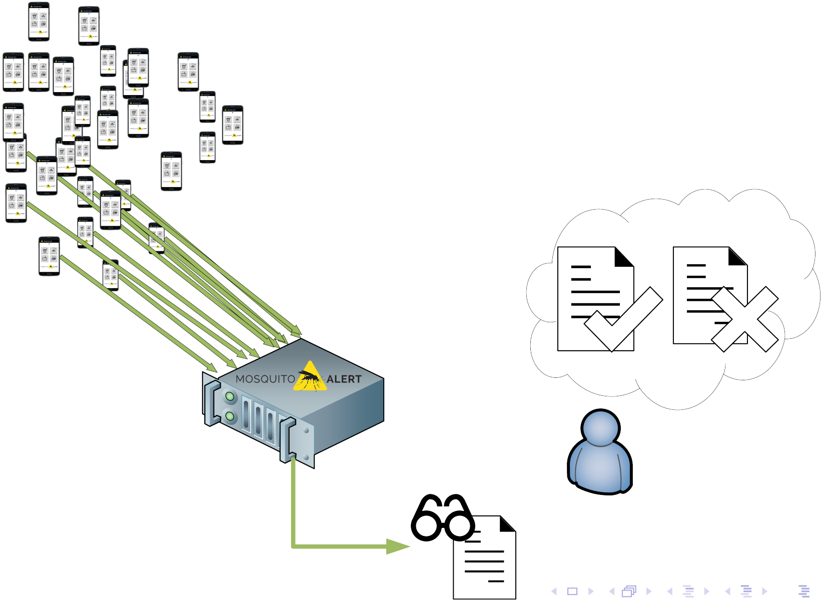
- Citizen Science Platform
- Mobile application
- Growing fast
  - Various mosquito species
  - Worldwide localizations

Introduction - Mobile App

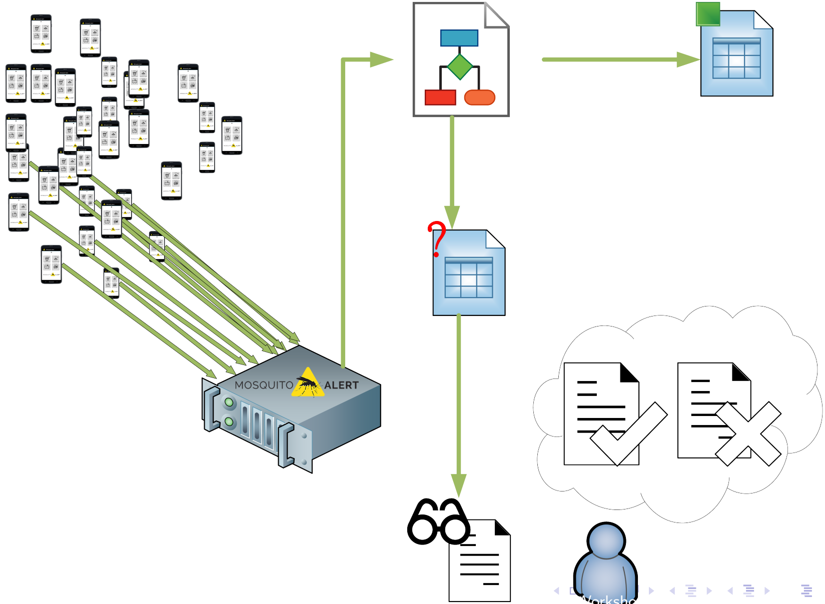


- Send breeding site
- Send specimen report
- Small questionnaire
- Geolocated!

Introduction - Mosquito Alert System



Introduction - Classification system



Methodology

1. Exploratory data analysis
2. Data cleaning and pre-processing
3. Classifiers
    - training
    - evaluation
    - selection
4. Real-time classification system design

Exploratory data analysis - Raw files

users 16967 observations of 10 variables

- userID
- userRegistTimeOriginal
- userRegistDatetime
- userRegistDate
- userRegistMonthNum
- userRegistMonthString
- userRegistWeekdayString
- userRegistWeekdayNum
- userSyst
- userDaysSystRelease

Exploratory data analysis - Raw files

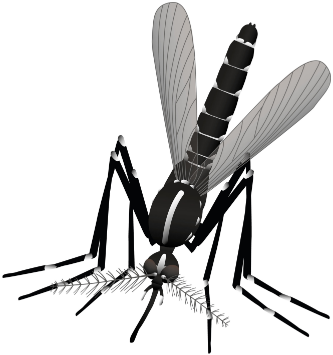
reports 10618 observations of 23+1 variables

- reportVersionID
- reportVersionNum
- userID
- reportID
- reportType
- reportNote
- os
- hide
- reportCreationDatetime
- reportCreationDate
- reportVersionDatetime
- reportVersionDate
- reportCreationMonthNum
- reportCreationMonthString
- reportCreationWeekdayString
- reportCreationWeekdayNum
- reportLong
- reportLat
- missionNum
- missionName
- tiger_q1_response
- tiger_q2_response
- tiger_q3_response
- **class**

Questionnaire variables

Questions

- Is small, black and has white stripes?
- Has a white stripe in both head and thorax?
- Has white stripes in both abdomen and legs?



Response values

-1	No
0	Not sure
1	Yes

The **class** variable

-2 The report is definitely not a valid specimen.

-1 The report doesn't seem to be a valid specimen. But it is not sure.

0 There isn't enough information to classify the report.

1 The report seems to be a valid specimen. But it is not sure.

2 The report is definitely a valid specimen.

Instance variables

Added

- reportNote
- reportTimeOfDay.
- newUser
- userNumReports
- userAccuracy
- userTimeForFirstReport
- userTimeSinceLastReport
- userMeanTimeBetweenReports
- userNumActionAreas
- userMobilityIndex
- reports1kmLast* (4)
- validReports1kmLast* (4)

Preserved

- os
- reportMonth
- reportQ*Answ (3)
- **class**

Generated instances

- 2094 instances from usable reports

Class	2	1	−1	−2
Frequency	47%	46%	2%	5%

- Class-imbalanced problem: positive instances over 7 times as frequent as negative ones.

Studied classifiers

- Naive Bayes
- k-nearest neighbors
- Decision trees
- Random Forests

Classifiers - Considerations

- Most classifiers have trouble dealing with imbalanced classes
- Merged “unsure” (-1,1) classes into “sure” ones (-2,2)
- Replication of minority class performed
- . . . but testing still on original proportion

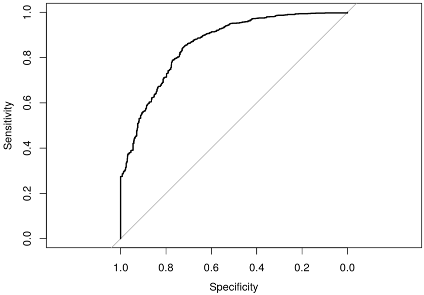
Classifiers - Selected classifier

	Positive	Negative
Accuracy	0,380	
Precision	0,983	0,086
Recall	0,344	0,912
F-measure (F1)	0,51	0,157

Table: Evaluation metrics, Naive Bayes

- Naive Bayes
- Training conditions:
  - Aggregated instances
  - Replicated (x10) negatives in training
- High positive *Precision*
- High negative *Recall*
- *can detect approximately 1 third of the valid reports with a precision near 98%*

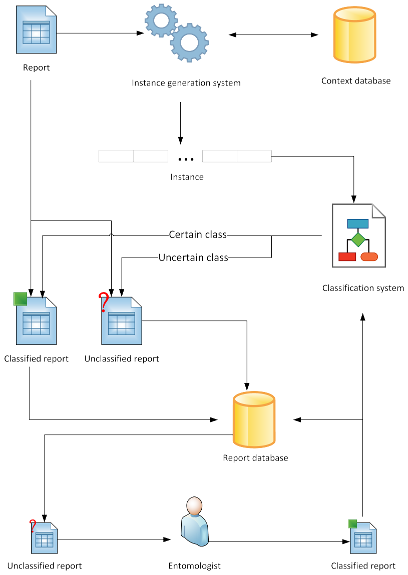
ROC curve and variable importance



Variable name	Importance
reportQ2Answ	0.7424
reportQ3Answ	0.7038
reports1kmLastMonth	0.6623
reportQ1Answ	0.6615
userNumReports	0.6405
userNumActionAreas	0.6348
validReports1kmLastMonth	0.6216
userTimeForFirstReport	0.6197
reports1kmLastWeek	0.6158
userAccuracy	0.6085

Table: Variable importance in the NB classifier. Numbers are the values of the model coefficients after standarization.

Real-time classification system design



- Two subsystems:
  - Instance generation system
    - Instance creation script
    - Environment
  - Classification system
    - Training script
    - Classifier
    - Classification script

Future work

Scalability

- Code modifications
- GIS enabled database
- Approximately the same computational resources

Improvements

- Classifier tuning
- Priority system
- Another classifier: Random Forest

Machine Learning Assists the Classification of Reports by Citizens on Disease-Carrying Mosquitoes

Antonio Rodriguez[1]   Frederic Bartumeus[2,3,4]   Ricard Gavaldà[1]

Universitat Politècnica de Catalunya, Barcelona (Spain)

Centre for Advanced Studies of Blanes (CEAB-CSIC), 17300 Girona (Spain)

CREAF, Cerdanyola del Vallès, 08193 Barcelona (Spain)

ICREA, Pg Lluís Companys 23, 08010 Barcelona (Spain)

Workshop on Data Science for Social Good, SoGood
September 2016