Confusion Diamond: Visualizing Performance of Classification Algorithms with Additional Re-Annotated Data

Megan K. Torkildson
University of Washington, HCDE, SCC Lab
mtorkild@uw.edu

The Challenge

The performance of machine learning (ML) classification algorithms with manual labels is difficult to assess because errors can exist both in the classification and the data. Overview visualizations, such as 2-dimensional confusion matrices, compare the automatically-generated output of a classification algorithm to ground truth data to characterize accuracy. An additional data set of manually re-annotated data can help tease apart classifier and data errors, but currently there is no effective way to visualize all three sets simultaneously. I developed the confusion diamond visualization to help expose these types of errors through three types of data: manual, automatic, and verifying.

Validation and Evaluation

I conducted a series of studies to validate the design of the confusion diamond. As part of **iterative design with a targeted group of expert scientists**, I conducted semi-structured interviews with 6 members of the ETC research group. This research group analyzes affect in chat logs and had no prior knowledge of the visualization. A subset of these students and additional ETC members, some with no previous knowledge of the visualization, were given an evaluation task with 15 confusion diamonds arranged in a canvas. These 7 participants had an overall accuracy of 76%.

Next, I conducted a **larger two part web deployment on Amazon Mechanical Turk for validation with novices**. The first part involved evaluation of an individual confusion diamond (N = 10). When prompted to provide a brief description of the problems that existed, crowd-workers made a total of 46 statements, of which 63% of were correct. The second part evaluated 15 confusion diamonds arranged in a canvas (N = 41), for an overall accuracy of 63%, with 94% corresponding justification statements being correct.

These results validate the confusion diamond as a promising visual aid for domain experts seeking to use ML for data analysis.

dub
Human Centered Design & Engineering
University of Washington

See Also

“Automating Large-Scale Annotation for Analysis of Social Media Content,” Katie Kuksenok, Michael Brooks, John J. Robinson, Daniel Perry, Megan K. Torkildson, Cecilia Aragon. 2nd Workshop on Interactive Visual Text Analytics, IEEE VisWeek (2012).

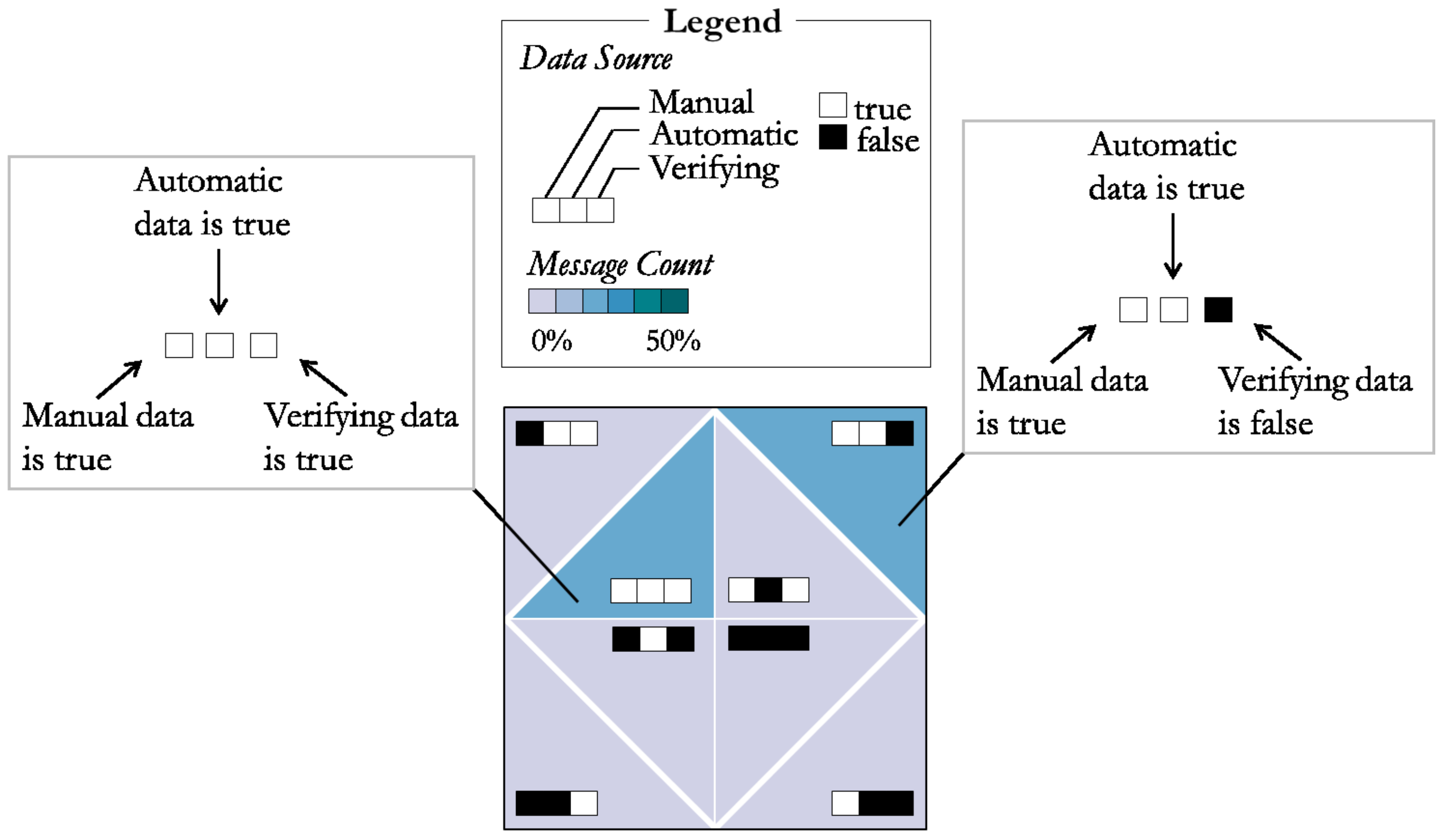
"A Tale of Two Online Communities: Fostering Collaboration and Creativity in Scientists and Children" (PDF), C. Aragon, S. Poon, A. Monroy-Hernandez, D. Aragon, ACM Conference on Creativity and Cognition, Berkeley, CA (2009).

“Statistical Affect Detection in Collaborative Chat,” Michael Brooks, Katie Kuksenok, Megan Torkildson, Daniel Perry, John Robinson, Paul Harris, Ona Anicello, Taylor Scott, Ariana Zukowski, Cecilia Aragon. CSCW (2013).

For more information, contact Megan Torkildson at mtorkild@.uw.edu or Cecilia Aragon at aragon@uw.edu or visit http://depts.washington.edu/sccl

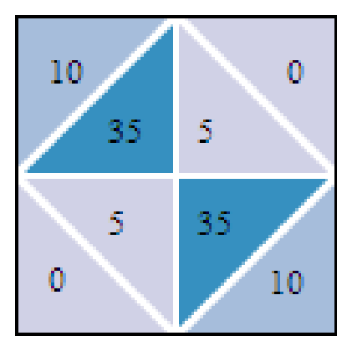
What is the confusion diamond?

Each triangle in the confusion diamond represents a combination of the three data types: manual, automatic, and verifying.

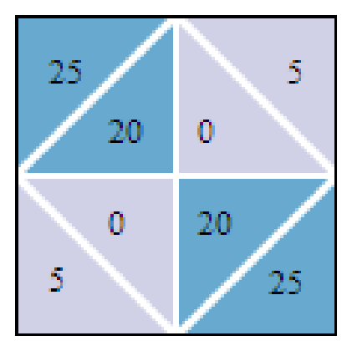


Five common types

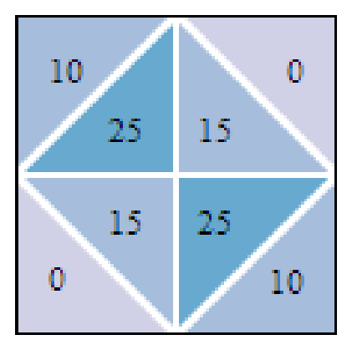
Each of the five types was based on accuracy and reliability combinations typically observed in the ETC Group data.



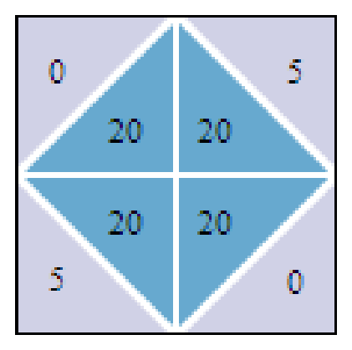
“Pretty good”
All three datasets have a high rate of agreement, with a low rate of disagreement with the manual labels.



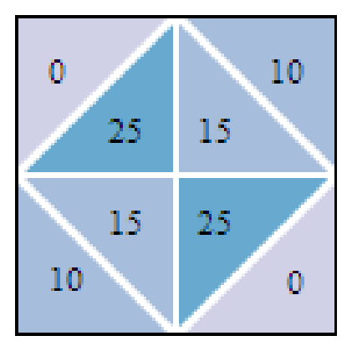
“Okay”
All three sets have a decent level of agreement with some disagreement with the manual labels, but the classifier is still accurate.



“Not Okay”
There is significant disagreement among the manual and verifying labels and also for the automatic and manual labels.

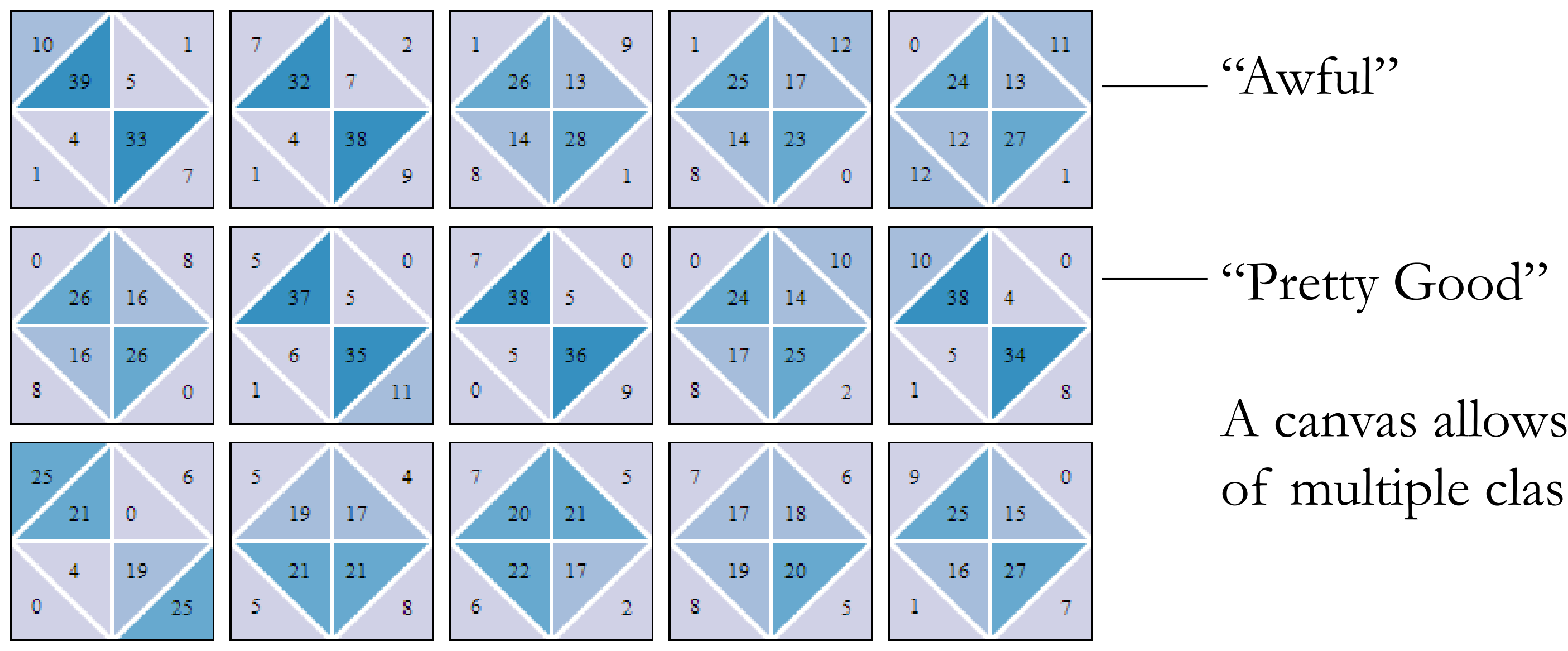


“Pretty Bad”
There is a medium rate of disagreement with the automatic data, but the manual and verifying data are consistent.



“Awful”
The classifier is very inaccurate for the data due to the high disagreement between the 3 datasets.

A canvas of multiple confusion diamonds



A canvas allows for comparison of multiple classifiers.