Tweet Classification Problem Statement

Michael Morris

December 9, 2019

To estimate (influenza like illness) ILI rates from Twitter data, tweets are sampled based on keywords such as flu, cold, sneezing etc. Frequencies of tweets containing these words are regressed to the true ILI rate and this is used to estimate future ILI rates. An assumption in this method is that every tweet containing a keyword about flu is useful for prediction of ILI rates, regardless of whether it is from a patient, advertising, or otherwise. It would be expected that the best predictor of flu rates would be from people who actually have the flu. In this case it would be beneficial to be able to classify tweets on whether they are from somebody who has the flu -a patient, or something else -not a patient. Which would give frequencies of tweets in these categories.

The probability of a flu tweet being from a patient is $P_f = P(tweetflu|havingflu)$, and the probability of it being from somebody else is: $P_{nf} = P(tweetflu|nothavingflu)$. If $P_f \simeq \alpha P_{nf}$ then a classifier will be of little use. The proportion of tweets from patients and non patients will be roughly constant and the classifier will just produce a scaled tweet frequency. For a linear regression model using tweet frequency X as an input to estimate ILI rate: $Y = mX + b$, the same output would be given from tweet frequency of people with the flu X_f from $Y = \frac{m(1+\alpha)}{\alpha} X_f + b$. If the proportions P_f and P_{nf} are independent (i.e α is continually varying) then this relationship is not true and a more accurate model may be possible.

Culotta [1] created a simple classifier for tweets which improved flu estimator performance during false alarm events by up to 50% for twitter data. Although these experiments were run on simulated data by adding erroneous tweets to existing data. These events are those where some other factor dramatically changes the distribution of tweets for a short period of time. Outside of these events the estimator's performance was unchanged. A more accurate classification method requiring less data would be to use semi-supervised-learning to create a classifier which does not require large amounts of unlabelled data. The problems are:

- Does α vary continually making a classifier worthwhile?
- Would a good classifier improve the performance of an estimator?
- If the first two are true, then what is the best way to classify tweets?

References

[1] Aron Culotta. Lightweight methods to estimate influenza rates and alcohol sales volume from twitter messages. *Language Resources and Evaluation*, 47, 03 2012.