Sentiment Knowledge Discovery in Twitter Streaming Data

International conference on Discovery science 2010, Albert Bifet and Eibe Frank

2016-10-11
Sentiment Knowledge Discovery in Twitter Streaming Data

Sentiment Knowledge Discovery in Twitter Streaming Data
International conference on Discovery science 2010, Albert Bifet and Eibe Frank

Contents

Contents
1. Background
2. Related Work
3. Theory
4. Evaluation
5. Results

- Background
- Related Work
- Theory
- Evaluation
- Result

Background

Background

- Twitter
- Data Stream Model
- Firehose

Contributions

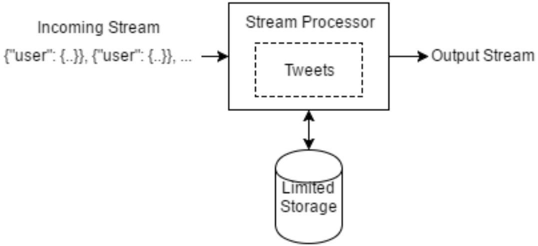
- Value of Twitter Streaming data
- Covering challenges of Twitter streaming data
- Sliding window Kappa Statistic
- Recommendation of a classifier

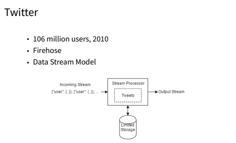
Contributions

- Value of Twitter Streaming data
- Covering challenges of Twitter streaming data
- Sliding window Kappa Statistic
- Recommendation of a classifier

Twitter

- 106 million users, 2010
- Firehose
- Data Stream Model





Twitter Streaming API

JSON

```
"user":{
   "statuses_count":3080,
   "favourites_count":22,
   "name":"Twitter API",
   "following":true,
   "description":"The Real Twitter API. I tweet about API
   changes, service issues and happily answer questions
   about Twitter and our API. Don't get an answer? It's on my
   website.",
   "location":"San Francisco, CA"
}
```


Twitter Streaming API
JSON
"user":{ "statuses_count":3080, "favourites_count":22, "name":"Twitter API", "following":true, "description":"The Real Twitter API. I tweet about API changes, service issues and happily answer questions about Twitter and our API. Don't get an answer? It's on my website.", "location":"San Francisco, CA" }

- The twitter API returns results in JSON that looks like this
- This has been shortened

Related Work

Related Work

- O'Connor et al - found surveys of consumer confidence and political opinion correlate with word frequencies in tweets.

Related Work with Data Mining

- Measuring user influence and dynamics of popularity
- Community Discovery and formation
- Social Information Diffusion

Related Work with Data Mining

- Measuring user influence and dynamics of popularity
- Community Discovery and formation
- Social Information Diffusion

- Measuring user influence and dynamics of popularity - Cha et al
- Community Discovery and formation - Java et al, Romero and Kleinberg
- Social Information Diffusion - De Choudhury et al

Related Work with Sentiment Analysis

- Data Mining for polling, O'Connor et al
- Implications of Micro-blogging as marketing strategy, Jansen et al
- Multinomial Naïve Bayes for Sentiment Analysis, Pak et al
- Comparison of various classifiers, Go et al

Related Work with Sentiment Analysis

- Data Mining for polling, O'Connor et al
- Implications of Micro-blogging as marketing strategy, Jansen et al
- Multinomial Naïve Bayes for Sentiment Analysis, Pak et al
- Comparison of various classifiers, Go et al

Theory

Theory

- Twitter Sentiment Analysis
- Kappa
- Multinomial Naive Bayes
- Stochastic Gradient Descent

Twitter Sentiment Analysis

Tweet Example

After a whole 5 hours away from work, I get to go back again, I'm so lucky!

- Need labeled data
- Emoticons as indicators of sentiment
  - Negative Sentiment - :(
  - Positive Sentiment - :)

Twitter Sentiment Analysis

Tweet Example

After a whole 5 hours away from work, I get to go back again, I'm so lucky!

- Need labeled data
- Emoticons as indicators of sentiment
  - Negative Sentiment - :(
  - Positive Sentiment - :)

Unbalanced Classes

	Predicted Class+	Predicted Class-	Total
Correct Class+	75	8	83
Correct Class-	7	10	17
Total	82	18	100

Confusion matrix for hypothetical classifier

	Predicted Class+	Predicted Class-	Total
Correct Class+	68.06	14.94	83
Correct Class-	13.94	3.06	17
Total	82	18	100

Confusion matrix for chance predictor

Unbalanced Classes

	Predicted Class+	Predicted Class-	Total
Correct Class+	75	8	83
Correct Class-	7	10	17
Total	82	18	100

Confusion matrix for hypothetical classifier

	Predicted Class+	Predicted Class-	Total
Correct Class+	68.06	14.94	83
Correct Class-	13.94	3.06	17
Total	82	18	100

Confusion matrix for chance predictor

$$(75+7)*(75+8)/100 = 68.06$$

Kappa statistic

- Kappa is 0 or less if there is no agreement between the classifiers other than chance
- Kappa is 1 when the classifiers are in complete agreement

Cohen's Kappa

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}$$

Observed Accuracy

$$p_o = \frac{\sum_{i=l}^{L} C_{ii}}{m}$$

Expected Accuracy

$$p_c = \sum_{i=l}^{L} \left(\sum_{j=l}^{L} \frac{C_{ij}}{m} \times \sum_{j=l}^{L} \frac{C_{ji}}{m} \right)$$

Kappa statistic
Kappa is 0 or less if there is no agreement between the classifiers other than chance
Kappa is 1 when the classifiers are in complete agreement
Cohen's Kappa
Observed Accuracy
Expected Accuracy

- observed proportionate agreement is p_o
- Normalizes the accuracy as a comparison of how much better the classifier is compared to a chance predictor

	Predicted Class+	Predicted Class-	Total
Correct Class+	75	8	83
Correct Class-	7	10	17
Total	82	18	100

Confusion matrix for hypothetical classifier

Observed Accuracy

$$p_o = \frac{\sum_{i=l}^{L} c_{ii}}{m}$$

$$p_o = \frac{75+10}{100} = 0.85$$

	Predicted Class+	Predicted Class-	Total
Correct Class+	75	8	83
Correct Class-	7	10	17
Total	82	18	100

Confusion matrix for hypothetical classifier

Class+ Accuracy

$$\frac{(75+7)\times(75+8)}{100} = 68.06$$

Class- Accuracy

$$\frac{(10+7)\times(8+10)}{100} = 3.06$$

Expected Accuracy

$$p_e = \frac{68.06+3.06}{100} = 0.7112$$

	Predicted Class+	Predicted Class-	Total
Correct Class+	75	8	83
Correct Class-	7	10	17
Total	82	18	100

Confusion matrix for hypothetical classifier

Class+ Accuracy

$$\frac{(75+7)\times(75+8)}{100} = 68.06$$

Class- Accuracy

$$\frac{(10+7)\times(8+10)}{100} = 3.06$$

Expected Accuracy

$$p_e = \frac{68.06+3.06}{100} = 0.7112$$

- $p_o = 0.85$
- $p_e = 0.7112$

Cohen's Kappa

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}$$

Kappa

$$\kappa = \frac{0.85 - 0.7112}{1 - 0.7112} \approx 0.48$$

2016-10-11

Sentiment Knowledge Discovery in Twitter Streaming Data
└─Theory

- $p_o = 0.85$
- $p_e = 0.7112$

Cohen's Kappa

$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}$

Kappa

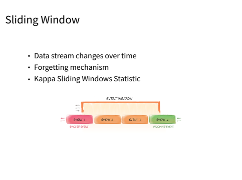
$\kappa = \frac{0.85 - 0.7112}{1 - 0.7112} \approx 0.48$

Sliding Window

- Data stream changes over time
- Forgetting mechanism
- Kappa Sliding Windows Statistic

2016-10-11

Sentiment Knowledge Discovery in Twitter Streaming Data
└─Theory
└─Sliding Window



Data Stream Mining Methods

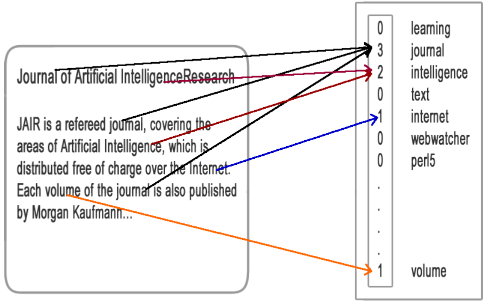
- Multinomial Naïve Bayes
- Stochastic Gradient Descent
- Hoeffding Tree

Data Stream Mining Methods

- Multinomial Naïve Bayes
- Stochastic Gradient Descent
- Hoeffding Tree

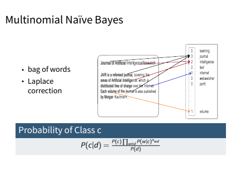
Multinomial Naïve Bayes

- bag of words
- Laplace correction



Probability of Class c

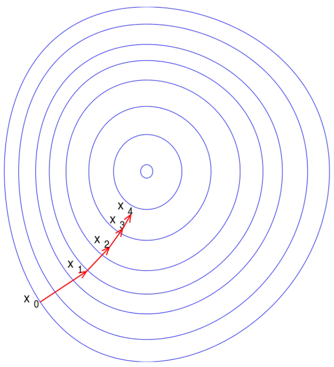
$$P(c|d) = \frac{P(c)\prod_{w \in d} P(w|c)^{n_{wd}}}{P(d)}$$



To avoid the zero-frequency problem, it is common to use the Laplace correction for all conditional probabilities involved, which means all counts are initialized to value one instead of zero.

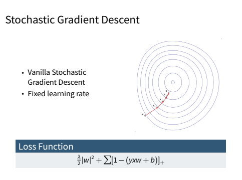
Stochastic Gradient Descent

- Vanilla Stochastic Gradient Descent
- Fixed learning rate



Loss Function

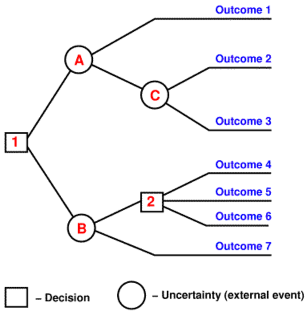
$$\frac{\lambda}{2}|w|^2 + \sum[1 - (yxw + b)]_+$$



- w - weight vector
- b - bias
- λ regularization parameter $= 0.0001$
- y - class label, -1 to 1
- learning rate per example = 0.1 - too sow and it wouldn't adapt to changes in stream

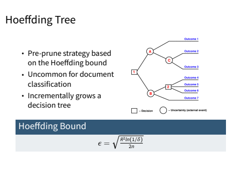
Hoeffding Tree

- Pre-prune strategy based on the Hoeffding bound
- Uncommon for document classification
- Incrementally grows a decision tree



Hoeffding Bound

$$\epsilon = \sqrt{\frac{R^2 ln(1/\delta)}{2n}}$$



- ϵ is the error on the node. R is the range random variable r. n is the number of observations.
- Each node tests an attribute
- Each branch is the outcome of that test
- Each leaf holds a class label
- For each training input - keep building till enough info, then split

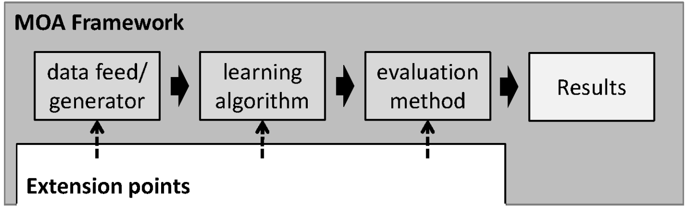
Evaluation

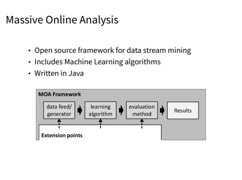
Evaluation

- We have evaluated specific parts and the overall system

Massive Online Analysis

- Open source framework for data stream mining
- Includes Machine Learning algorithms
- Written in Java





MOA is an open-source framework software that allows to build and run experiments of machine learning or data mining on evolving data streams.

Datasets

- twittersentiment.appspot.com
  - :), :-), :), :D, and =D - Positive
  - :(, :-(, and : (- Negative
  - Training set - 800.000 positive and negative tweets
  - Test set - 182 positive, and 177 negative
- Edinburgh corpus
  - 97 million tweets
  - Feature reduction
    - huuuuuungry → huungry
    - @ → USER token
    - URLs → URL token
  - Used English tweets with emoticons
    - Deleted after annotation
  - Reduced to a training set of 324,917 negative and 1,8m positive tweets

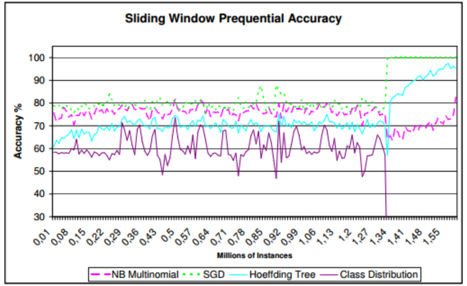
Datasets
- twittersentiment.appspot.com
  - :), :-), :), :D, and =D - Positive
  - :(, :-(, and : (- Negative
  - Training set - 800.000 positive and negative tweets
  - Test set - 182 positive, and 177 negative
- Edinburgh corpus
  - 97 million tweets
  - Feature reduction
    - huuuuuungry → huungry
    - @ → USER token
    - URLs → URL token
  - Used English tweets with emoticons
    - Deleted after annotation
  - Reduced to a training set of 324,917 negative and 1,8m positive tweets

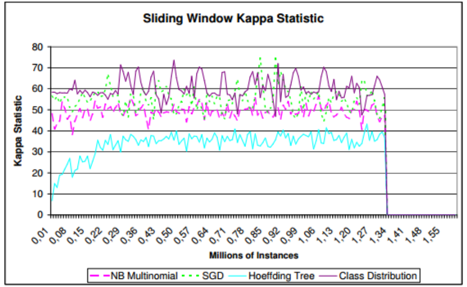
Results

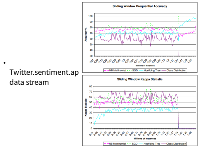
Results

- Two data stream experiments
- One classic train/test on each training set and then the test set

- Twitter.sentiment.ap data stream







	Accuracy	Kappa	Time
Multinomial Naïve Bayes	75.05%	50.10%	116.62 sec.
SGD	82.80%	62.60%	219.54 sec.
Hoeffding Tree	73.11%	46.23%	5525.51 sec.

Twittersentiment.appspot data stream

	Accuracy	Kappa
Multinomial Naïve Bayes	82.45%	64.89%
SGD	78.55%	57.23%
Hoeffding Tree	69.36%	38.73%

Twittersentiment.appspot test dataset

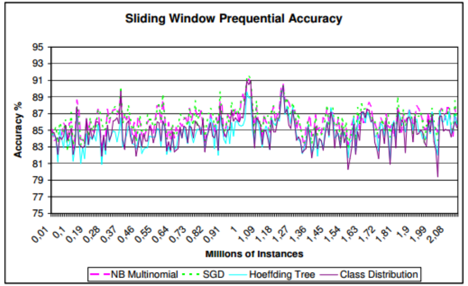
	Accuracy	Kappa	Time
Multinomial Naïve Bayes	75.05%	50.10%	116.62 sec.
SGD	82.80%	62.60%	219.54 sec.
Hoeffding Tree	73.11%	46.23%	5525.51 sec.

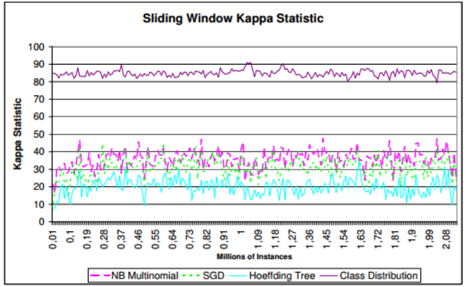
Twittersentiment.appspot data stream

	Accuracy	Kappa
Multinomial Naïve Bayes	82.45%	64.89%
SGD	78.55%	57.23%
Hoeffding Tree	69.36%	38.73%

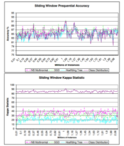
Twittersentiment.appspot test dataset

- Edinburgh corpus data stream





• Edinburgh corpus data stream



	Accuracy	Kappa
Multinomial Naïve Bayes	73.81%	47.28%
SGD	67.41%	34.23%
Hoeffding Tree	60.72%	20.59%

Accuracu and Kappa for twittersentiment.appspot.com using Edinburgh corpus as training data

	Accuracy	Kappa
Multinomial Naïve Bayes	73.81%	47.28%
SGD	67.41%	34.23%
Hoeffding Tree	60.72%	20.59%

Accuracu and Kappa for twittersentiment.appspot.com using Edinburgh corpus as training data

Conclusion

- SGD-based model is the recommended one for this data

Conclusion

- SGD-based model is the recommended one for this data

Future Work

- Real time analysis
- Geographical place
- Followers
- Number of friends

Future Work

- Real time analysis
- Geographical place
- Followers
- Number of friends

In future work, we would like to extend the results presented here by evaluating our methods in real time and using other features available in Twitter data streams, such as geographical place, the number of followers or the number of friends.