Data Mining & Machine Learning Applications

David TJ Huang
Learning Outcomes

✓ **Understand** what is pattern and what is noise

✓ **Recognize** the *influence* of input data and preprocessing to the mining results

✓ **Connect** data mining and machine learning algorithms to real world problems

✓ **Know** how pattern & noise are defined differently in different problems
PART IV

Applications - Twitter & Social Media
Twitter & Social Media

Growing trend & application

• Unstructured
• Text / Images
• Highly noisy
• Fast changing
• Large volume
• High velocity
Twitter

General overview

• Social networking service
• Tweets - Send short messages (max 140 characters)
• “SMS of the internet”

• Founder Jack Dorsey’s intention for Twitter is that a message is “a short burst of inconsequential information.”
Twitter

Contents of Tweets?

- Pointless babble – 40%
- Conversational – 38%
- Pass-along value – 9%
- Self-promotion – 6%
- Spam – 4%
- News – 4%

Source – Pear Analytics (2009)
Different Types of Tweets

Justin Bieber
@justinbieber

CONGRATS to @McKaylaMaroney @jordynmarie2013 @kyla_ross96 @Aly_Raisman @gabrielledou on your GOLD MEDAL. #BeliebersWinGOLD #Proud ...

Barack Obama
@BarackObama

RT if you're on #TeamObama tonight.

Ellen DeGeneres
@TheEllenShow

If only Bradley's arm was longer. Best photo ever. #oscars pic.twitter.com/C9U5N0tGap

Will Ferrell

I want to have 3 kids and name them Ctrl, Alt and Delete. Then if they fuck up I will just hit them all at once.

3:05 AM Apr 5th via Hootsuite
Why Twitter??

What characteristics make Twitter data so fascinating?

• Real-time
• Large volume / streaming nature
• Captures a different set of information compared to search queries
  • Search queries: intent, desire, learning
  • Twitter: ideas, emotions, sentiments, actions
• Content is publicly available (vs. profile based services)
DM/ML and Twitter

What are people are working on…

• Opinion mining (sentiment analysis)
• Trend detection
• Prediction of an event based on Tweets
  • Natural events such as Earthquakes
  • Medical events such as outbreaks of disease and flu
• Hub detection (finding the most connected / influential person)
• Frequent graph mining
• Building models of recommendation
• etc.
Predicting Flu Trends

Similar to Google Flu Trends

• Goal is to predict the flu based on Twitter activity (the content of Tweets based on their IP/regions)

• Evaluated in the similar way as Google Flu Trends

• Compared against the CDC’s ILI cases and reports

• Results:
  • Highly correlated data with a Pearson correlation coefficient of 0.9864

• Have similar problems to the Google Flu Trends
  • While the goal is to fit a model, it may easily overfit
Event Detection

Because on Twitter people share what is on their mind in real-time...

• We can use this to our advantage and detect occurrences of real-time events – such as Earthquakes

• In addition to the keyword selection, a semantic analysis was performed on the collected Tweets

• “The earthquake yesterday was scary”
  vs.

• “Earthquake!! Shaking right now!!”

• Needed to distinguish the Tweets that had the real-time component from those that only describes an earthquake or a past event
Event Detection

How to distinguish between Tweets with the *real-time* aspect?

- Essentially a learning task by itself
Event Detection

How to distinguish between Tweets with the *real-time* aspect?

• Essentially a learning task by itself

• Consider various features:
  • Statistical features
    • The number of words in a tweet, the position of the query word
  • Keyword features
    • The actual words in the tweet
  • Word context features
    • The word before and after the query word

• SVM to classify positive vs. negative Tweets
Event Detection

The interesting part about event detection in social data…

• In addition to the temporal aspect of the Tweets…
• There is also the spatial aspect
  • IP address – location
Event Detection

The interesting part about event detection in social data…

• In addition to the temporal aspect of the Tweets…
• There is also the spatial aspect
  • IP address – location

• This means that you can build knowledge graphs or networks
  • The problem can be further extended and/or visualized as a Graph Mining task
Event Detection

The interesting part about event detection in social data…

- In addition to the temporal aspect of the Tweets…
- There is also the spatial aspect
  - IP address – location
Trend Detection

Twitter has a feature of trending topics / hashtags

• Goal is to predict or detect new trending topics
Trend Detection

Twitter has a feature of trending topics / hashtags

• Goal is to predict or detect new trending topics

• Some things to consider:
  • Should all terms carry the same weight?
  • Should all Tweets (of different users) carry the same weight?
  • What time frame? (how to keep representative data)
  • To what level of significance do we consider topic trending?
Sentiment Analysis

Twitter is filled with people’s ideas and thoughts they share publicly

- Naturally, we want to be able to understand what these thoughts represent and how they can be summarized

- Goal of sentiment analysis:
  - Extract subjective information from (text) sources
  - Often trying to determine whether your client / user feels *good* or *bad* about either a service, a good for sale, an event, etc.
Sentiment Analysis

What about change detection? Does it have any use here?

• Sentiments change over time
• Capturing changing sentiments can provide different info from purely just capturing sentiments

• Example: 2010 Toyota Crisis
  • Problem with accelerator pedals
  • Had to recall many of their cars
## Sentiment Analysis

<table>
<thead>
<tr>
<th>Term</th>
<th>Before</th>
<th>After</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>gas</td>
<td>0.122</td>
<td>0.484</td>
<td>0.363</td>
</tr>
<tr>
<td>pedals</td>
<td>0.129</td>
<td>0.438</td>
<td>0.309</td>
</tr>
<tr>
<td>wonder</td>
<td>0.017</td>
<td>0.214</td>
<td>0.198</td>
</tr>
<tr>
<td>problem</td>
<td>0.163</td>
<td>0.357</td>
<td>0.194</td>
</tr>
<tr>
<td>good</td>
<td>0.016</td>
<td>0.205</td>
<td>0.190</td>
</tr>
<tr>
<td>recalling</td>
<td>0.012</td>
<td>0.106</td>
<td>0.095</td>
</tr>
<tr>
<td>gm</td>
<td>0.011</td>
<td>0.089</td>
<td>0.077</td>
</tr>
<tr>
<td>#heard_on_the_street</td>
<td>0.040</td>
<td>0.113</td>
<td>0.072</td>
</tr>
<tr>
<td>social</td>
<td>0.031</td>
<td>0.099</td>
<td>0.068</td>
</tr>
<tr>
<td>sticking</td>
<td>0.070</td>
<td>0.125</td>
<td>0.055</td>
</tr>
<tr>
<td>fix</td>
<td>0.026</td>
<td>0.076</td>
<td>0.050</td>
</tr>
<tr>
<td>popularity</td>
<td>0.016</td>
<td>0.037</td>
<td>0.021</td>
</tr>
<tr>
<td>love</td>
<td>0.017</td>
<td>0.024</td>
<td>0.008</td>
</tr>
</tbody>
</table>
Sentiment Analysis
References & Further Reading

• Detecting Sentiment Change in Twitter Streaming Data
  • By Albert Bifet, Geoff Holmes, Bernhard Pfahringer

• Emerging Topic Detection on Twitter based on Temporal and Social Terms Evaluation
  • By Mario Cataldi, Luigi Di Caro, and Claudio Schifanella

• Earthquakes Shakes Twitter Users: Real-time Event Detection by Social Sensors
  • By Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo
Summary

• What is information
  • Pattern & Noise

• Preprocessing
  • Dealing with noise / how to get clean data

• Real-world applications
  • How everything connects / how to build models based on input

• Twitter & social media
  • Emerging area & problem / how to deal with the new issues
Summary

The goal is really to get you to think!

• So make sure you…
  • Think about what you are doing
  • Know the effects & bias
  • Justify it properly
Summary

The goal is really to get you to think!

- So make sure you...
  - Think about what you are doing
  - Know the effects & bias
  - Justify it properly

Do not feel limited!