# Machine Learning for Intelligent Systems

SEM6420

Assignment One

Daniel McGuckin

[dam44@aber.ac.uk](mailto:dam44@aber.ac.uk)

Paper: “F Benevenuto, Gl Magno, T Rodrigues, and V Almeida. Detecting Spammers on Twitter. In Proceedings of the Seventh annual Collaboration, Electronic messaging, AntiAbuse and Spam Conference (CEAS 2010), Redmond, Washington, US, 2010”

The paper details an attempt to train a non-linear Support Vector Machine (SVM) using a Gaussian Kernel. They train the algorithm to classify spam versus legitimate data, firstly concerning Twitter user accounts and then individual tweets.

# The Problems

The paper identifies that Twitter is a target for spam due to the number of users who use the service and how they communicate with each other. Elements of the service like hashtags can be used by spammers to try to fool users into believing they are posting about relevant, popular topics when they are really trying to get users to buy products or even leading them to malicious or inappropriate websites. This is aided by their ability to obfuscate the URL’s in their tweets.

To stop this spam it must first be found, this is the problem that the paper attempts to solve. The authors declare that a supervised machine learning algorithm could be used to find either the spammer accounts themselves or the spam that they tweet.

# The Solutions

The authors first used data crawling software to gather information about users. Over 58 servers white-listed for them by Twitter they found almost 55 million users and 1.8 billion tweets. They got volunteers to label 8,207 of these Twitter accounts of which 355 were spam. They then randomly selected 710 legitimate accounts to even out the numbers. This resulted in 1,065 accounts used in their training process. The paper does not detail how many tweets were used in their experiments.

Next they decided on 62 attributes describing behaviour differentiating spam from legitimate accounts and tweets. These attributes fit into two different types:

* Content – attributes detailing information about user tweets such as the number of URL’s per words.
* User behaviour – attributes detailing information about the user account such as the number of followers.

Their experiments first looked into user account classification, this was their primary focus in the paper. A 5-fold cross-validation strategy was used, they split their data into 5 samples, 4 of these samples were used to train the SVM and the other one used to test it. They performed this training and testing 5 times, each time rotating their samples so that a different one was used to test. They then repeated this entire process 5 times resulting in 25 separate trails.

They declared that on average “Approximately, 70% of spammers and 96% of non-spammers were correctly classified.” The authors state that this algorithm can effectively identify spammer accounts, thus achieving to some degree the problem this paper was attempting to solve. However, they admit that although their SVM only misclassifies a small percentage of legitimate accounts, this “could not be suitable for detection mechanisms that apply some sort of automatic punishment”.

In an attempt to rectify this they used a cost-mechanism to give priority to one class over the other. They managed to tune their cost-mechanism so that they would only misclassify 0.3% of legitimate accounts, this was a 3.3% increase from 96.4%. However, this was at the cost of a reduction in spam classification, down from 70.1% to 43.7%, a reduction of 26.4%.

The paper then discussed attempting to reduce the number of attributes used. They attempted to find which attributes were most useful by applying the information gain and Chi Squared feature selection methods. Their findings showed that using either the top 10 or 20 attributes gave results of a similar quality as using all 62 attributes. They also show that content and user behaviour attributes have a similar weighting and as content attributes are much easier to disguise by the spammers, behaviour attributes could allow intelligent spammers to be detected more easily.

The second problem explored by the paper - classification of spam tweets - was then briefly delved into. They used a selection of the content attributes in the SVM. Their results showed that on average their SVM could detect around 79% of spam tweets and 93% of legitimate ones. Although the SVM performed better on detecting spam tweets than it did spam accounts, it performed almost twice as badly when classifying legitimate ones. The authors did not explore using their cost-mechanism for this problem but one could imagine that it would produce similar results to what was achieved in the first problem.

In summary, the paper explores the problem of detecting Twitter spammers. It divides this problem into the detection of user accounts and user tweets. A non-linear SVM is trained on labelled data and then tested. The SVM is proven to be able to classify 70% of spam and 96% of legitimate users with similar results for tweets, a valid attempt was therefore made to solve the problem at hand but it is arguable whether they have actually ‘solved’ the problem due to their error margin classifying legitimate accounts. They did managed to reduce this error percentage but with considerable loss to their ability to classify spam.

# The Issues

# The Alternative Approach

# The Limitations and Future Development