Nikola Mrksic
Trinity College
nm480@cl.cam.ac.uk

Computer Science Tripos Part II Project Progress Report

# Semi-supervised Learning Methods for Data Augmentation

January 2013

**Project Supervisor:** Dr Sean Holden

**Directors of Studies:** Dr Arthur Norman and Dr Sean Holden

**Overseers:** Dr Simone Teufel and Prof Jon Crowcroft

# Work completed

As per the project proposal and further research, a thorough investigation of the following papers was conducted:

- *(Bayesian Sets, Ghahramani and Heller, 2006);*
- *(Learning to Classify Texts Using Positive and Unlabeled Data, Li and Liu, 2003);*
- *(Partially Supervised Classification of Text Documents, Liu et al, 2002);*
- *(Distributional Similarity vs. PU Learning for Entity Set Expansion, Li et al, 2010).*

The *Bayesian sets* and *Spy-EM* algorithms have been implemented (in *Matlab*) and tested on multiple data sets (*IMDB, Reuters, 20Newsgroups, KDD Cup 2001*). Results achieved with these datasets produce results similar to those obtained in the relevant papers(in terms of *F-scores* and *precision@N* values when applied to these well known data sets).

Through investigation of related work we identified another method applicable to data augmentation, called *Roc-SVM*. It is based on the same paradigm as *Spy-EM*: initially, it uses a *Rocchio classifier* and *k-means clustering* to separate the reliable negatives. Then, it iteratively trains a *support vector machine(SVM)* to obtain the final classifier. The *F-score* comparison with *Spy-EM* shows that *Roc-SVM* exhibits superior performance on many datasets, which means that it might be a better fit for one of our data sets as well. Hence, we have chosen to include *Roc-SVM* as the third algorithm in our data augmentation library. It has been implemented and tested on some data sets, but requires further optimization in order to become applicable to larger ones.

The first goal of the project was to create standard, generalized implementations of these *positive-unlabeled(PU)* learning algorithms. With the exception of the required *Roc-SVM* optimisations, we can say that this goal has been achieved.

The second goal is to create a framework that will achieve data augmentation in data-scarce applications, and subsequently analyse its performance in different application areas. A general platform for measuring the success of data augmentation has been implemented:

The training data is sampled to obtain smaller training sets, with the rest of the training set used as the *unlabeled set* for data augmentation. Standard *Support Vector Machines(SVM)* are trained using the reduced training data sets, and their performance on the test data is measured. Subsequently, we apply all three of our entity set expansion algorithms to obtain the augmented data sets. SVMs are retrained using these augmented training sets, and the change in the subsequent classifications' *F-scores* is measured in order to assess the quality of the data augmentation achieved.

The data sets we are using for this project are:

1. *Reuters21578* text categorisation collection(document classification setting);
2. *KDD Cup 2001, task 1, "Binding to Thrombin"*:(biomedical data, identifying active compounds);
3. The body of statements used for Dr Holden's supervised theorem prover.

Up to this point, we have examined the success of data augmentation via *Bayesian Sets* and *Spy-EM* with the *Reuters21578* and *KDDCup 2001* data sets. We still have to examine their performance with the theorem proving data. *Roc-SVM* needs to be optimized further to be applicable to the second data set.

# Further work

1. *Iterative Bayesian Sets:* implementing an augmented version of *Bayesian sets*, investigated in *(Entity Set Expansion in Opinion Documents, Zhang et al, 2011)*, that consists of bootstrapping *Bayesian Sets* so that each iteration uses the newly extracted positives to recompute the rankings and extract new positives.
2. Investigating the impact of a sampling bias between the testing and sampling data. This bias is present in our biomedical data set. *(Negative Training Data can be Harmful to Text Classification, Li et al, 2010)* investigates this phenomena and claims that positive-unlabeled learning outperforms traditional supervised learning on such datasets. Having implemented three *PU* algorithms, we can use them to investigate the validity of this claim.
3. *Final* ***potential*** *extension*: a computational learning theory proof of stability for Bayesian Sets is given in *(Growing a list, Letham et al, 2012)*. This proof and its implications justify the use of *Bayesian Sets* when the feature independence property is violated (Web as a prime example, but also interesting in the case of our data sets, especially the biomedical one). Thus, an ivestigation of this proof may give some theoretical insight, justification and deeper understanding of the results obtained using Bayesian Sets.

At the moment, we are approximately two weeks ahead of the proposed schedule. In addition to being close to fulfilling most of the success criteria(including the proposed extensions), I've further expanded the scope of the project to include the *Roc-SVM* algorithm, the *Iterative Bayesian Sets* algorithm, a discussion of the impact of the training set sampling bias on applicability of *PU* learning, and a *(potential)* examination of the theoretical grounding of Bayesian sets in computational learning theory.