Research Statement

1 Review

Today, Big Data is ubiquitous. With the growing volume of data in the recent years, we have started to visualize problems that are large-scale. Different Machine Learning algorithms have been designed to deal with big data.

In our Research, we focus on approaches that are suitable for large-scale data and have the potential for parallel implementation. A problem is called large-scale if its training set can not be stored in modern computer's memory. A large training set poses a challenge for the computational complexity of a learning algorithm: in order for algorithms to be feasible on such datasets, they must scale at worst linearly with the number of examples. Conventional algorithms can not handle such problems as they no more have ready access to data stored in the memory. This mandates the need for the development of new algorithms and analysis of the challenges posed by them. Online algorithms are used in large-scale setting where infinite stream of training examples are presented one at a time. They are also used to solve batch problems which is desirable in the large-scale setting. The general approaches to solve large-scale batch learning problems are:

1. **Data Stream**: We can treat the training data as a stream and apply online algorithms.

2. **Parallelize**: The batch algorithm can be parallelized so that large learning problem can be split into multiple smaller problems(medium-scale).

3. **Random subset**: We can preprocess the training data and randomly sample a small subset of data to train on. By choosing the subset appropriately we can reduce the training size and make our training algorithm tractable.

Few algorithms that are used for large scale learning are Decision Trees, Perceptrons, Neural nets, Instance based learning and Support Vector Machines.

- The decision tree classifier that is designed to classify large training data is called SLIQ (**S**upervised **L**earning **I**n **Q**uest). SLIQ is capable of classifying disk-resident datasets, scalable for large datasets and uses pre-sorting techniques for efficiency. The disk resident data is too large to fit in memory but it still requires some information to stay memory-resident which

grows in proportion with the input records putting a limit on the size of the training data.

- Another decision tree based classification algorithm called SPRINT (**S**calable **P**a**R**allelizable **IN**duction of decision **T**rees) is a fast, scalable, no memory restrictions and easy to parallelize algorithm. It can handle larger datasets.

- Another classifier that learns decision trees from data streams is based on Incremental learning methods (Hoeffding Tree algorithm) is called VFDT (Very Fast Decision Tree Learner).

Our research is primarily focussed on Large-Scale Machine learning on OCR text. The related work in this domain used Winnow-based algorithm [3] achieving accuracy levels of 99% for 265 "confusion sets". However this approach was originally meant to correct predetermined words. It was unclear to what extent it could be scaled to correct all words in a sentence and moreover for allowing more than a small set of predetermined alternatives for each word. Moreover, another approach used was a baseline Hidden Markov Model (HMM) which was trained on wikipedia edits and then further augmented with perceptron reranking but it again caused performance bottleneck.

2 Problem Statement

The research problem is to develop an algorithm which when applied on a large-scale training data learns a model/classifier that would classify the future data. The challenge is to extend the current popular methods that are used for binary classification to multi-class classification. The large training data to be used is the online repository of historical newspapers articles in the holdings of cronicling america[1].

3 Research Focus

There are some limitations to the ability of the perceptron to classify some data.

1. The biggest problem is that sometimes data is inherently not separable by the hyperplane. The cure to this problem can be to find a function on the points that would transform them to another space enabling them to be linearly separable. But that would just be another case of overfitting, the situation where the classifier works very well with the training data as it has been carefully designed to handle each training example correctly. However, the classifier will not perform well with on the new data.

2. If classes are separable by one hyperplane then there can be many hyperplanes separating the points and not all of them are equally good. There lies a more equitable choice of separating hyperplane.

[1] http://chroniclingamerica.loc.gov/

3. Perceptron stops as soon as there are no misclassified points. As a result, the chosen hyperplane just manages to classify some of the data points correctly.

SVM is seen as an improvement on the perceptron that is designed to address the problems as mentioned above. It selects only one particular hyperplane out of all the possible valid hyperplanes such that it not only separates the two classes but also does so in the most efficient way. It selects the hyperplane that maximizes the margin (distance between the margin and the closest points of the training set.) There are SVM solvers like Pegasos which when presented with more training data, decrease the runtime in contrast to other solvers that increase superlinearly with the training data, like SVM^{perf}.

4 Scope

With the current trend in the growth of information, large-scale learning problems can not be overlooked. There is a need to develop algorithms that deal with these problems efficiently.
The objective of this research is to train a model on a large-scale data which can be used further to predict the errors in the unseen OCR text. This would help in building a clean online repository of historic newspapers providing rich source of information to historians, researchers, scholars and general users.

References

[1] Jeffrey D. Ullman Anand Rajaraman, Jure Leskovec. *Mining Of Massive Datasets*. Cambridge University Press, 2013.

[2] Andrew Carlson, Chad Cumby, Je Rosen, and Dan Roth. The snow learning architecture. Technical report, Technical Report UIUCDCS, 1999.

[3] AndrewR. Golding and Dan Roth. A winnow-based approach to context-sensitive spelling correction. *Machine Learning*, 34(1-3):107–130, 1999.

[4] ibm corporation. Bigdata: Why it matters to the midmarket. *Forward View: The magazine for mid-size business*, 2012.

[5] Thorsten Joachims. Advances in kernel methods. chapter Making large-scale support vector machine learning practical, pages 169–184. MIT Press, Cambridge, MA, USA, 1999.

[6] Aditya Krishna Menon. Large-scale support vector machines: algorithms and theory. *Research Exam, University of California, San Diego*, 2009.

[7] Elif Yamangil and Rani Nelken. Scalable lexical correction from wikipedia edits using perceptron reranking. *unpublished*, 2008.