Machine Learning Project Guideline

Dae-Ki Kang

April 23, 2008

1 Project Ideas

The following are the suggestions for the project in this semester. These are just suggestions, so the students can work on a different project if they want, but it has to be related with machine learning and is similarly difficult.

1. Taxonomy Guided Support Vector Machines
2. Naive Bayes Augmented Decision Tree
3. Conditional Random Field
4. Multi-Instance Learning
5. Multi-Relational Decision Tree
6. ROC Analysis
7. Transfer Learning

At the following sections, each item will be explained. All the references can be easily found on the Web with Google.

1.1 Taxonomy Guided Support Vector Machines

For this project, please read through Zhang et al.'s work (Zhang et al., 2006) first.

Remember that they generate Naive Bayes classifiers from the locally optimal cut. The optimality is calculated by conditional minimum description length (CMDL) or conditional log-likelihood (CLL).

Now, instead of Naive Bayes classifier, perform the experiments with Support Vector Machines (SVM). That is, (1) find a locally optimal cut from the CMDL or CLL, (2) generate instances corresponding to the cut, (3) and generate SVM.

Compare the results of the classifiers from AVT-SVM with those from regular SVM.

1.2 Naive Bayes Augmented Decision Tree

Carefully read Kang et al.'s paper (Kang et al., 2006) first.

Remember that if you generate Naive Bayes classifier h from a training set D, and classifiy each instace in D by h, you will get a training error e of h. Also, you will have a confusion matrix M with the error e. From the confusion matrix M, you can calculate an entropy which can be used for information gain.

That means one Naive Bayes classifier can be a test of a decision tree node. It will be interesting if you can combine Naive Bayes classifier based test to the current attribute based test of decision tree. Thus, generate a decision tree algorithm equipped with not only attribute-based test, but also Naive Bayes classifier based test.

1.3 Conditional Random Field

Read Lafferty et al. (Lafferty et al., 2001).

Also, consult Wikipedia's definition of Conditional Random Field (CRF) at `http://en.wikipedia.org/wiki/Conditional_random_field`. In the definition page, they have the pointers for the Java programs.

Implement CRF (in Java preferably) and compare its performance with SVM with n-grams (or SVM with spectrum/string kernel if you can) or other learning algorithms that can handle sequences. For the sequence data sets, please consult the instructor.

1.4 Decision Tree for Multi-Instance Learning

Read `http://www.cs.cmu.edu/~juny/MILL/mil_review.pdf`, (Chevaleyre & Zucker, 2001) and (Blockeel et al., 2005).

Implement decision tree algorithm for multi-instance learning according to one of the papers above.

Use Musk and Mutagenesis data sets to verify the algorithm. They are available at `http://www.cs.waikato.ac.nz/ml/proper/datasets.html`.

1.5 Multi-Relational Decision Tree

Read (Leiva, 2002) and implement decision tree algorithm for multi-relational data sets.

Use Mutagenesis data set and one more data set to verify the algorithm.

Data sets are at `http://www.cs.waikato.ac.nz/ml/proper/datasets.html`.

1.6 ROC Analysis

Read (Lachiche & Flach, 2003).

And read `http://www.cs.bris.ac.uk/~flach/ICML04tutorial/index.html`.

Improve the performance of Naive Bayes classifier using the idea described in the paper.

1.7 Transfer Learning

Read `http://www.cs.utexas.edu/~mooney/cs391L/hw2/` and (Dai et al., 2007). Implement TrAdaBoost.

2 Guideline

2.1 File Formats

For the format of project report, please use Lecture Notes in Computer Science (LNCS) format.

The file formats can be downloaded from
`http://www.springer.com/computer/lncs?SGWID=0-164-7-72376-0`.

2.2 LaTeX2e

For the final report, the instructor highly recommends the class to use LaTeX2e if possible.

References

Blockeel, H., Page, D., & Srinivasan, A. (2005). Multi-instance tree learning. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, (pp. 57–64)., New York, NY, USA. ACM.

Chevaleyre, Y. & Zucker, J.-D. (2001). Solving multiple-instance and multiple-part learning problems with decision trees and rule sets. application to the mutagenesis problem. In *AI '01: Proceedings of the 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence*, (pp. 204–214)., London, UK. Springer-Verlag.

Dai, W., Yang, Q., Xue, G.-R., & Yu, Y. (2007). Boosting for transfer learning. In *ICML '07: Proceedings of the 24th international conference on Machine learning*, (pp. 193–200)., New York, NY, USA. ACM.

Kang, D.-K., Silvescu, A., & Honavar, V. (2006). RNBL-MN: A recursive naive Bayes learner for sequence classification. In *10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2006)*, volume 3918 of *Lecture Notes in Artificial Intelligence*, Singapore. Springer Verlag.

Lachiche, N. & Flach, P. A. (2003). Improving accuracy and cost of two-class and multi-class probabilistic classifiers using roc curves. In *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003)*, (pp. 416–423). AAAI Press.

Lafferty, J. D., McCallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In

ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning, (pp. 282–289)., San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Leiva, H. A. (2002). *MRDTL: A multi-relational decision tree learning algorithm.* Masters thesis, Iowa State University.

Zhang, J., Kang, D.-K., Silvescu, A., & Honavar, V. (2006). Learning accurate and concise naive Bayes classifiers from attribute value taxonomies and data. *Knowledge and Information Systems*, *9*(2).