Lecture 5

Semi supervised learning

Semi supervised learning is used for the same applications as supervised learning. But it uses both labeled and unlabeled data for training – typically a small amount of labeled data with a large amount of unlabeled data (because unlabeled data is less expensive and takes less effort to acquire). The basic procedure involved is that first, the programmer will cluster similar data using an unsupervised learning algorithm and then use the existing labeled data to label the rest of the unlabeled data. The typical use cases of such type of algorithm have a common property among them – The acquisition of unlabeled data is relatively cheap while labeling the said data is very expensive.

This type of learning can be used with methods such as classification, regression and prediction. Semi supervised learning is useful when the cost associated with labeling is too high to allow for a fully labeled training process. Early examples of this include identifying a person's face on a web cam.

Intuitively, one may imagine the three types of learning algorithms as Supervised learning where a student is under the supervision of a teacher at both home and school, Unsupervised learning where a student has to figure out a concept himself and Semi-Supervised learning where a teacher teaches a few concepts in class and gives questions as homework which are based on similar concepts.

Why bother?
Because people want better performance for free.
the traditional view

- unlabeled data is cheap
- labeled data can be hard to get
  -human annotation is boring
  -labels may require experts
  -labels may require special devices

Practical applications of Semi-Supervised Learning –

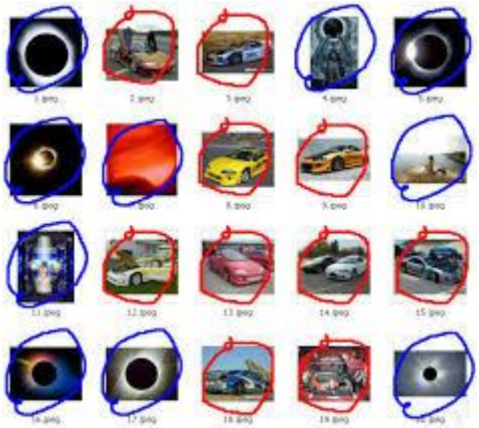
1. **Speech Analysis:** Since labeling of audio files is a very intensive task, Semi-Supervised learning is a very natural approach to solve this problem.

2. **Internet Content Classification:** Labeling each webpage is an impractical and unfeasible process and thus uses Semi-Supervised learning algorithms. Even the Google search algorithm uses a variant of Semi-Supervised learning to rank the relevance of a webpage for a given query.

3. **Protein Sequence Classification:** Since DNA strands are typically very large in size, the rise of Semi-Supervised learning has been imminent in this field.

Example of not-so-hard-to-get labels

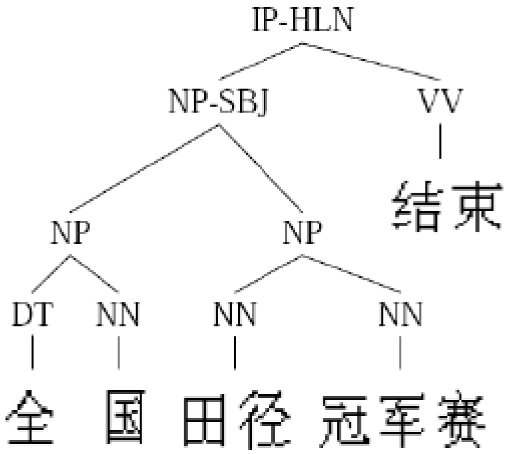
a little secret For some tasks, it may not be too difficult to label 1000+ instances.
Task: image categorization of "eclipse"



Example of Hard-to-get labels

Task: natural language parsing

- Penn Chinese Treebank
- 2 years for 4000 sentences



"The National Track and Field Championship has finished."