Technical updates on “Detecting Concept Drift with Support Vector Machines (2000)”

The method used in the paper uses SVM as its learning model. The objective of the paper is to recognize and handle concept changes with support vector machines (SVM). The method maintains a window on the training data. The key idea is to automatically adjust the window size so that the estimated generalization error is minimized.

A window adjustment algorithm has to solve the following trade-o  
. A large window provides the learner with much training data, allowing it to generalize well given that the concept did not change. On the other hand, a large window can contain old data that is no longer relevant (or even confusing) for the current target concept. Finding the right size means trading of the quality against the number of training examples.

The experiments are performed in an information filtering domain, a typical application area for learning drifting concept. Text documents are represented as attribute-value vectors (bag of words model), where each distinct word corresponds to a feature whose value is the frequency of that word in that document. Words occurring less than three times in the training data or occurring in a given list of stop words are not considered. Each document feature vector is normalized to unit length to abstract from different document lengths.

The performance of a classifier is measured by the three metrics prediction error, recall, and precision. Recall is the probability, that the classifier recognizes a relevant document as relevant. Precision is the probability, that a document classified as relevant actually is relevant. All reported results are estimates averaged over ten runs.

The paper includes an experiment which shows the efficiency of this method on real world concept drift scenario of text data. The experiments use a subset of 2600 documents of the data set of TREC consisting of English text. The texts are randomly split into 20 batches of equal size containing 130 documents each. Each text is assigned some categories from 1 to 6.

Updates that can be practiced

The model works on the assumption that the next batch of the data will be similar to the immediate previous batch. If this is not the case, the model will fail deliver accuracy and will give false prediction. However the real world data is also similar to the assumption. But more work can be done in this regard so that the model may work under no assumptions.

The model just uses the average error rate and not the true error rate. So we can say that the model will work fine for text classification task, but the accuracy of the model is under question for any other task where the bias is large enough for acceptance. So further updates that are possible includes a model which take into account the actual error value and then predict the label and not rely on average error.

We can work on a model which can sense a pattern of concept drift and make a generalised model for this. We can test in under extreme condition to check if it is a perfect generally accepted model. We can learn to predict concept drift. As this model has peaked the task of predicting concept drift. Instead of handling concept drift, now we can work on predicting concept drift. So these were some of the possible updates in this regard.

Prajjaval Verma

B19ME054