# On conservative approaches to learning for pattern recognition problems

Thesis summary

G. S. Ramasubramanian

There exist a number of situations in real life where the relationship among the variables that influence the behaviour of a system is not known or cannot be assumed. In such cases, one has to rely upon inferences based upon empirical observations. To this end, algorithms that generalize from examples, such as those used in training artificial neural networks, have been designed – these are known as learning algorithms. Learning theory is the analysis of such algorithms, and the principles governing them.

The bulk of learning theory has concentrated upon scenarios wherein the learning algorithm returns a prediction for every example it is presented with. However, this might not always be the case. An algorithm may, for instance, consider it prudent not to return a prediction on an example that falls in a region it considers ambiguous or has not seen before adequately. We shall call such learning algorithms *conservative*. This has also been referred to in the literature as learning with a reject option.

There are a number of problems that involve learning from examples, chief among them being pattern recognition (or classification), regression estimation and density estimation. Pattern recognition problems deal with the issue of learning the class labels applicable on different regions of a feature space. This thesis concentrates on conservative approaches to learning in pattern recognition problems. Pattern recognition is defined as the problem of assigning an appropriate class to

a given example, based on knowledge gleaned from a labeled sample known as a training set. Applications to management include bankruptcy prediction, credit rating and takeover target prediction.

The objective of this thesis is to examine in more detail, some issues pertaining to conservative approaches to learning in pattern recognition problems. The following questions have been dealt with:

1. Estimating bounds on the generalization ability of classifiers with a reject option, on the basis of the sample size and complexity.

2. Learning perceptrons with a reject option for two-class problems, by training both the network and the rejection scheme simultaneously.

3. Extending the existing rejection schemes for multiclass problems to account for the possibility of gradation among classes.

The contribution of this thesis to the existing body of research is summarized in figure 1 (contributions are italicized).

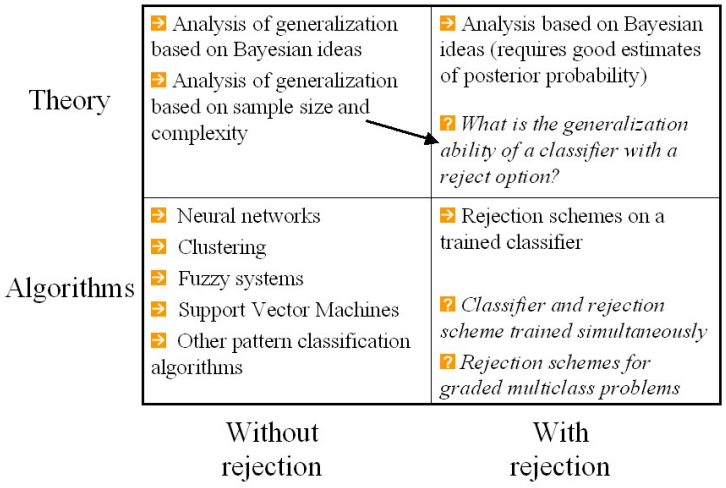


Figure 1: Contributions

# 1 Statistical learning theory perspective

Our interest is in the theoretical underpinnings of the problem of learning with a reject option, especially from the point of view of statistical learning theory. Research in learning theory has so far been focused on the analysis of algorithms without a reject option. The issue of predicting only in local regions of the example space has been dealt with; however, this formulation of the learning problem does not consider the issue of there being a penalty for rejection as well.

On the other hand, researchers in the pattern recognition domain have explored the issue of rejection and proposed various techniques to decide on which example to predict/reject. However, the generalization ability of these classifiers with a reject option as a function of the sample size and classifier complexity has not been studied so far, to the best of our knowledge.

One of the primary aims of this thesis is to extend the results in statistical learning theory to include learning with a reject option, wherein the relative cost of rejection vis-a-vis misclassification is used to bring out the rationale for rejection.

To this end, the problem of learning with a reject option is formulated, given the rationale that it may be more prudent to reject an example than run the risk of a costly potential misclassification. The fundamental inductive principle behind learning theory is that of empirical risk minimization (ERM), which states that the best classifier is the one that performs best on the available data. The ERM principle is applied to the various possible formulations of this problem, and the bounds on generalization ability of a classifier with a reject option are derived under these formulations.

Since the generalization ability of the classifier depends both on the complexity of the underlying zero-reject classifier and that of the rejection scheme applied, it is possible to consider two ways of learning:

**Decoupled rejection scheme** The more commonly used method is to train the zero-reject classifier first, and then train the rejection scheme given this classifier.

**Coupled rejection scheme** An alternative approach would be to train both the underlying classifier and the rejection scheme together, since it would allow us to converge to a better minimum of the overall risk. Proceeding further along these lines, it would be possible to envisage *tightly coupled* scheme wherein the example space is directly divided into regions for each class and regions where examples are rejected.

A comparative analysis of these two schemes is presented, in terms of the tradeoffs between quality of the optimal solution, computational complexity and generalization bounds. Examples for computation of the risk bounds for two-class problems are given, on the basis of VC dimension results for neural networks and some commonly used rejection schemes.

The issue of structural risk minimization and learning with maximum generalization is discussed for both rejection schemes. A nested support vector machine formulation for the decoupled case, and a constructive algorithm involving perceptrons with rejection are proposed to tackle this problem.

## 2 Perceptron learning with a coupled rejection scheme

The idea of using a coupled rejection scheme, i.e., training both the underlying classifier and the rejection method together, has not been explored much in the neural network literature. Therefore, methods of extending the perceptron learning rule to implement a coupled rejection scheme are explored.

The hypothesis being learnt is a hyperplane with a bandwidth on either side to allow for rejection. From the standpoint of division of the example space (tightly coupled case), this can be viewed as a set of two parallel hyperplanes, with the region in the middle being rejected.

The basic approach followed is to view the error in terms of the distance of an example from each hyperplane, weighted by the relative gain/loss with respect to that hyperplane. For instance, for a rejected example, the distance from one hyperplane is weighted by the loss in rejecting the example as opposed to correctly

classifying it, while the distance from the other hyperplane is weighted by the gain in rejecting the example as opposed to misclassifying it.

Three algorithms, namely CPLR0, CPLR1 and CPLR2 are suggested within the coupled rejection framework.

1. **CPLR0**: A tightly coupled algorithm, wherein two parallel hyperplanes are learnt.

2. **CPLR1**: A more loosely coupled algorithm, wherein the hypothesis is viewed as a hyperplane with two bandwidths, so the intercept term is learnt separately from the bandwidths.

3. **CPLR2**: An even more loosely coupled approach, wherein the weights and bandwidths are learnt using different error functions. In this case, while learning the bandwidths, we consider the trade-off between correctly classified, misclassified and rejected examples in the relevant region of the example space, rather than just considering the misclassified and rejected examples as in the other two algorithms.

The utility of a coupled approach to rejection is demonstrated through experiments on some datasets. Further extensions to multilayer perceptrons and constructive neural network algorithms are discussed.

## 3 Rejection in graded multiclass problems

The reject option in multiclass problems has so far been discussed in the literature without considering the possibility of there being a gradation among the classes. In this thesis, the issue of rejection in graded multiclass problems (GMP) is discussed.

The drawbacks of the existing schemes for rejection in general multiclass problems when dealing with GMP are highlighted. The problem of learning with rejection in GMP is formulated, with respect to the various performance metrics that could be considered in this case.

Extensions to the existing rejection schemes to handle the GMP case are explored, with illustrative examples from the problem of predicting the credit rating of debt instruments based on information pertaining to the financial status of the issuing firm. The rejection schemes have been applied to three different classifiers - feedforward neural networks, radial basis function networks and a fuzzy inference engine with a table lookup scheme.

The problem of computing risk bounds for multiclass classifiers is discussed, and some issues involved in computing this bound are explored in more detail. Extensions to a measure of complexity for multiclass classifiers known as the graph dimension are proposed, to cover various performance metrics that we might be interested in.

Further extensions to the tightly coupled case, and loss functions based on more than one performance metric are briefly discussed. Also, the case of classifiers assigning multiple class labels to an example has been explored.

# 4 Applications

The ideas and algorithms presented in this thesis have been applied to a number of prediction problems, both from benchmark datasets and live problems in management.

**Bankruptcy** Predict whether or not a firm would go bankrupt, given financial information about the firm's current position.

**Takeover targets** Predict whether or not a firm would be a target for a takeover bid, given information about the firm's financials.

**Bond rating** Predict the credit rating of a long term debt instrument, given financial information about the issuing firm.

**Pima Indians diabetes** Predict whether or not a patient has diabetes, given a set of numeric attributes.

**Liver disorders** Predict whether or not a patient has a liver disorder arising from alcohol consumption.

Apart from these, a synthetic dataset was also created for illustrative purposes and used in experiments on the CPLR versions.