Classification Reliability and Reject Option

Rejection in Pattern Recognition

- It has been a long time since Pattern Recognition researchers understood the importance of providing the classifier with a *reject option*
  - the possibility of refusing to assign the examined pattern to any class, possibly prompting for a further investigation by another system or by a human supervisor.
- The first work trying to cast the reject problem in a formal, theoretically sound framework is a paper by Chow published 50 years ago.

Open Issues

- Restrictive assumptions on the classifier outputs
  - in some cases (e.g. Chow's rule) it is required that the classifier provides a good estimate of the *a posteriori probability* of each class.
    - the output is a measure related to the *a posteriori probability*
  - in other cases it is at least required to have *type 3* classifiers (no crisp outputs)
- Lack of generality about the classification technique
  - several authors have proposed methods specifically tailored to a particular classifier
    - SVM
    - Binary classifiers
    - …

Chow's Rule

- The rationale of the Chow's rule relies on the **exact knowledge** of the *a posteriori* probabilities $P(C_i|x)$ for each sample x to be recognized.
- In this hypothesis, the class assigned to the generic sample is the one whose post-probability is the highest, provided that its value is higher than the threshold

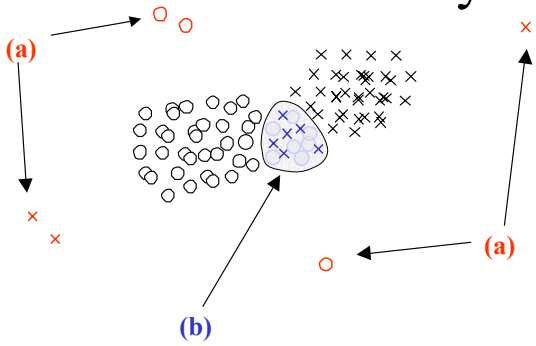
$$\frac{C_e - C_r}{C_e - C_c}$$

Otherwise, the sample is rejected.

A possible approach based on…

- The rationale lies on the characterization of the situations which may lead to unreliable classifications.
- The reliability of each classification act is evaluated on the basis of the output of the classifier
  - if it is greater than a threshold, the decision of the classifier is considered acceptable
  - otherwise the input sample is rejected.
- The optimal threshold value is determined by means of a training procedure which takes into account the specific recognition requirements through a function *P* which measures the classifier effectiveness in the considered application domain.
- No knowledge about probability distributions is needed, but we must have *type 3* classifiers.

… Reliability Evaluation



Case (a): the samples are significantly different from the ones of the TS ⇨ ψ_a

• Distance-based reject

Case (b): the samples lies in a confusion region of the feature space ⇨ ψ_b

• Confusion-based reject

$$\psi = \psi(\psi_a, \psi_b)$$

$$\psi_{min} = \min\{\psi_a, \psi_b\}$$

$$\psi_{med} = \frac{\psi_a + \psi_b}{2}$$

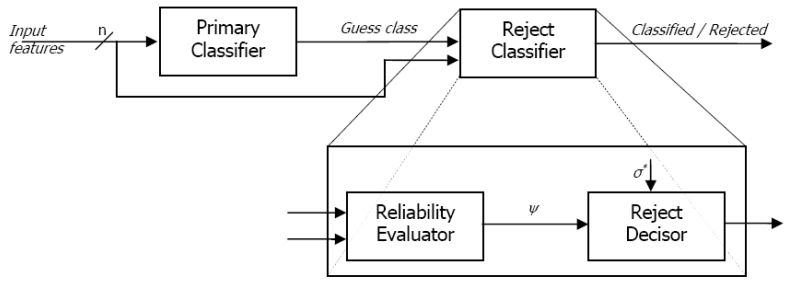
$$\psi_{max} = \max\{\psi_a, \psi_b\}$$

$$\psi_{sym} = \frac{\psi_a \psi_b}{1 - \psi_a - \psi_b + 2\psi_a \psi_b}$$

A more recent approach

- We propose to face the rejection problem as a new classification problem.
- We introduce a trainable classifier, that we call ***reject classifier*** to distinguish it from the classifier to which the reject option is applied (termed ***primary classifier***).
- This idea yields a reject option that is largely independent of the approach used for the primary classifier
  - it works also for systems providing as their only output the guess class.

The Proposed Architecture



Training the Reliability Evaluator

- For training the reliability evaluator we have chosen a supervised approach.
- The reliability classifier is trained on a *reliability-training set* for which the true class of the patterns is known.
- The desired output is set to
  - 0 for patterns to which the primary classifier assigns the wrong class,
  - 1 for correctly classified patterns.
- Our method relies on the **generalization** ability of the reliability evaluator to interpolate intermediate values for patterns that are across regions with good classification results and regions poorly classified.

Some considerations…

- The task of the reject classifier is itself a full classification task
  - it is expected to be simpler than the one faced by the primary classifier
- Both classifiers operate on the same input space, but:
  - the reject classifier has to face a two classes problem (reliable or unreliable patterns)
  - the primary classifier usually has to discriminate among two or more classes.
- So, even when the training set is not adequately representative for obtaining a good classifier or a good estimate of the *a posteriori* class probabilities, it is still possible that the patterns suffice to train a properly working reliability classifier.

Reliability Evaluator

– A Support Vector Machine (SVM).
  - since the goal of the reliability evaluator is to provide a real-valued measure, we have used the Support Vector Regression (ε-SVR).

The method for determining the optimal reject thresholds

- To determine the optimal values of the reject thresholds we define an **effectiveness function** *P,* which, taking into account the requirements of the particular application, evaluates the quality of the classification in terms of recognition, misclassification and reject rates.

$$P = \sum_{i=1}^{N} C_{ii}(R_{ii} - R_{ii}^0) - \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} C_{ij}(R_{ij} - R_{ij}^0) - \sum_{i=1}^{N} C_{i0} R_{i0}$$

- The **optimal reject threshold** value σ*, determining the best trade-off between reject rate and misclassification rate, is the one for which the effectiveness function *P* reaches its **absolute maximum**.
- The requirements of an application domain are specified by attributing **costs** to misclassifications, rejects and correct classifications.
  – The cost of an error can be a function of the guess and of the actual class

About the costs

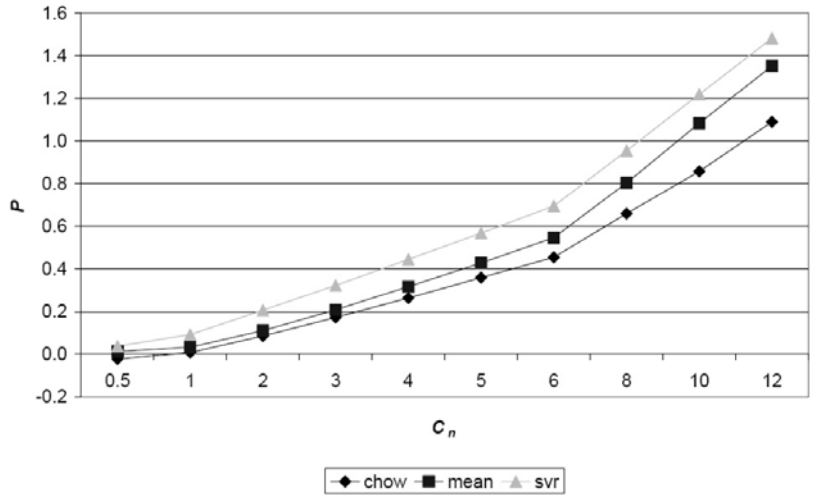
- If we assumed that the recognition task do not require to distinguish among the possible kinds of errors ($C_{ii} = C_c$; $C_{ij} = C_e$; $C_{i0} = C_r$).
- It is then possible to define a normalized cost as the ratio:

$$C_n = \frac{C_e - C_r}{C_c - C_r}$$

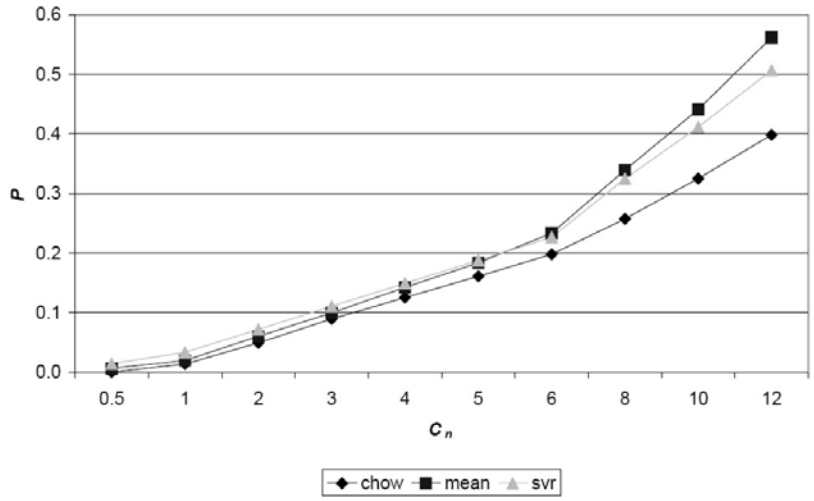
Some results

- Primary classifiers
  - A Nearest Neighbor (NN)
  - A Multi-Layer Perceptron (MLP)
    - 10 hidden units
    - 40 hidden units
    - 80 hidden units
- Three different databases coming from the UCI Machine Learning Repository
  - *letter* (contains the 26 capital letters in the English alphabet: the character images were based on 20 different fonts, randomly distorted).
  - *pendigits* (contains handwritten digits obtained by means of a digitizing tablet).
  - *spam* (refers to the problem of determining whether a given email is spam or not).

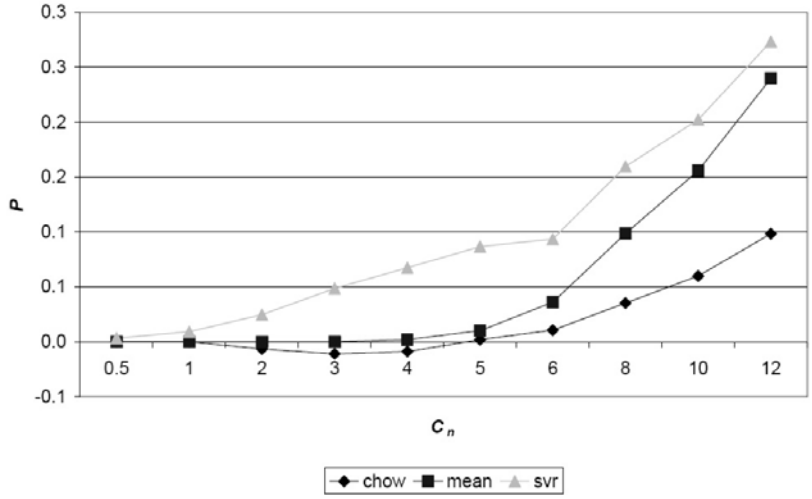
Letter dataset



Pendigits dataset



Spam Dataset



Conclusions

- The reject methods proposed so far make restrictive assumptions on the outputs provided by the classifier to which the reject option is applied. They also lack of generality regarding the choice of the classification technique.
- We developed a reject method based on the evaluation of the classification reliability that makes little or no assumption on the structure of the classifier and of its outputs
- It can adapt to some degree to the training information available in the application at hand.