Transductive Confidence Machine for Active Learning

Shen-Shyang Ho and Harry Wechsler
Department of Computer Science
George Mason University
Fairfax, VA 22030-4444
Email: {sho, wechsler}@cs.gmu.edu

***Abstract*— This paper describes a novel active learning strategy using universal p-value measures of confidence based on algorithmic randomness, and transductive inference. The early stopping criteria for active learning is based on the bias-variance trade-off for classification. This corresponds to that learning instance when the boundary bias becomes positive, and requires one to switch from active to random selection of learning examples. The sign for the boundary bias and the increase in the classification error are two manifestations of the same phenomena, i.e., over-training. The experimental results presented show the feasibility and usefulness of our novel approach using a non-separable two-class classification problem. Our hybrid learning strategy achieves competitive performance against standard nearest neighbor methods using much fewer training examples.**

I. Introduction

The standard framework in machine learning in general, and pattern classification, in particular, presents the learner with a randomly sampled data set. There has been, however, a growing interest in active learning where one has the flexibility to choose the data points that seem most relevant for the learning task. According to [19], one analogy is that a standard passive learner is a student that sits and listens to a teacher, while an active learner is a student that asks the teacher questions, listens to answers and asks further questions based upon the teachers response. The active learner selects actions or makes queries that influence what data and in what order are added to the training set [4]. Many similar concepts are available for neural networks, e.g., conscience mechanisms [9]. The goals for active learning are two fold : (i) less computation (due to smaller training sets) without penalizing the performance of the classifier, and (ii) less (computational or manual) efforts and cost to label the training examples.

The major task for active learning is to decide how to select the membership of the training set and in what order. This is usually done by assessing to what extent the candidate samples are informative, e.g., quality of information and discrimination power. Toward that end many of the proposed methods draw from information theory, e.g., comparing how the entropy of the training ensemble would change after the inclusion of each of the candidate samples. The entropy by itself, however, cannot capture effectively the margin or separation of the classifier. To address those concerns, this paper introduces a novel framework for active learning that is based upon algorithmic randomness and transduction learning. The feasibility and usefulness of our novel approach is shown on a non-separable two-class classification problem.

II. Background

Active learning is relevant to any activity involving choice and uncertainty. The problem has been extensively studied but additional innovation and progress is still waiting. We briefly review some of the methods and their motivation. [18] suggests using a simple form of divide and conquer by selecting training examples that lie on or close to the separating hyper-plane. This approach draws from the emergence and growing interest in Statistical Learning Theory (SLT) [20] in general, and Support Vector Machines (SVM), in particular. Basic to SLT is to what extent the generalization ability is affected by the relationship between the capacity of the learning machine and its confidence (intervals) for future prediction. Another approach involves querying instances that split the current version space into two equal parts as much as possible. It takes advantage of the duality of parameter space and feature space [19], i.e., the geometry of the learning space, and is motivated by SLT and SVM.

The margin of separation referred to above is closely related to local learning and nearest neighbor classifiers. As an example, [8] have proposed making queries on the vertices of the induced Voronoi tessellation in the input space during active learning. [11] have developed a "look-ahead" selective sampling methodology, which rather than considering the uncertainty of the sample points in the input space, takes the effect of labeling of point on its neighborhood into account. Toward that end, one employs a random field model to estimate the probability of its possible labels and compute the expected information measure for the candidate sample.

Yet another form of learning, beyond induction, is transduction. Given an unlabeled validation test, in addition to the training set, the task now is to estimate the class for each unlabeled pattern in order to construct the best classifier rule for both the training and validation tests. The Overall Risk Minimization (ORM) strategy has been proposed to solve the transduction problem. It involves performing a comparative assessment of labeling the pattern iteratively as one of the available class assignments. ORM may improve on generalization and yield improvements when there is a significant deviation between the training and validation sets [2].

The constraints on the layout of the learning space and the search for improved margins are addressed in this paper using universal p-value measures of confidence based on algorithmic randomness[21], and transductive inference [15]. Our novel active learning methodology further requires some stopping

criteria to prevent over-training. This corresponds to choosing between using the active learning selection mode, a random selection mode, or stopping the generation of the learning set. Our proposed method for this purpose is based on the bias-variance trade-off for classification [6]. This corresponds to that learning instance when the boundary bias becomes positive, and requires one to switch from active to random selection of learning examples. The sign for the boundary bias and the increase in the classification error are two manifestations of the same phenomena, i.e., over-training. The experimental data presented later on validates our approach, and is based on KNN-TCM, which is an augmented Transductive Confidence Machine (TCM) using locality-based evidence, the K Nearest Neighbors (KNN).

III. Randomness and p-Values

Confidence measures suitable for active learning can be based upon universal tests for randomness, or their approximation. A Martin-Lof randomness deficiency [10] based on such tests is a universal version of the standard statistical notion of p-values.

A function $t : Z^n \to [0,1]$ is a *p-value function* with respect to any probability distribution P in Z if

1) for all $n \in N$ and $r \in [0,1]$,
$$P^n\{x \in Z^n : t(x) \leq r\} \leq r$$
2) t is semicomputable from above, i.e. there exists a computable sequence of computable functions $t_i : Z \to [0,1]$,$i = 1, 2, \cdots$ such that $t(z) = \inf_i t_i(z)$ for all $z \in Z$

Universal tests for randomness are not computable and hence one has to approximate the p-values using non-universal tests satisfying Item 1 in the above definition. In statistical significance testing, the p-value provides a measure on how well the data support or discredit the null hypothesis [22].

IV. Transduction and p-value Construction

We use the p-value construction in [7] and [15] to define a quality of information needed for data selection during active learning. In particular they proposed a new algorithm for pattern recognition that outputs some measures of reliability for every prediction made, in contrast to the current algorithms that output bare predictions only. The only assumption used is that data items are independent and produced by the same stochastic mechanism. Given a sequence of proximities (distances) between the given training set and an unknown sample probe, one quantifies to what extent the (classification) decision taken is reliable, i.e., non-random. Toward that end one defines the strangeness of the unknown sample probe i with putative label y in relation to the rest of the training set exemplars as:

$$\alpha_i = \frac{\sum_{j=1}^{K} D_{ij}^{y}}{\sum_{j=1}^{K} D_{ij}^{-y}}$$

where D_i^y is the sequence of distances of example i from other examples with the same classification y, sorted in ascending order such that D_{ij}^y stands for the jth shortest distance in the sequence. D_{ij}^{-y} is the jth shortest distance from sorted sequence of distances of example i from other examples with different classification from y. Clearly, this is the ratio of the sum of the K nearest distances from the same class to the sum of the K nearest distances from the other classes. The strangeness of an exemplar increases when the distance from the exemplars of the same class becomes larger and when the distance from the other classes becomes smaller. Note that the 1-nearest neighbor rule is asymptotically at most twice as bad as the Bayes error and that as K increases, the K-nearest neighbor gets progressively closer to the Bayes error, assuming asymptotic sampling [5]. For finite sampling, the K-nearest neighbor rule can be viewed as an (non-parametric density estimation) attempt to estimate the a posteriori probabilities from the data sample [3].

Define now a p-value function $t : Z^n \to [0,1]$ by

$$t(z_1, z_2, \cdots, z_n) = \frac{\#\{i = 1, \cdots, n : \alpha_i \geq \alpha_n\}}{n}$$

which satisfies

$$P^n\{(z_1, z_2, \cdots, z_n) : t(z_1, z_2, \cdots, z_n) \leq r\} \leq r$$

for any $r \in [0,1]$ and for any probability distribution P in Z, provided that the strangeness function returns the same value for each example independent of the input order of examples to the strangeness function [12]. Assuming that z_n is the testing example and α_n is the strangeness of z_n when it is assigned a possible classification, $t(z_1, z_2, \cdots, z_n)$ will be the p-value of z_n of that classification, given the training examples $z_1, z_2, \cdots, z_{n-1}$.

In analogy to to statistical significance testing, we test the null hypothesis "if we have an iid sequence of examples $z_1, z_2, \cdots, z_n$, then the distribution over all permutations of the sequence is uniform" against the alternative hypothesis "the last element (or the testing example) z_n is not generated by the same distribution as the rest of the sequence" [13]. Intuitively, z_n given classification c is not strange is our null hypothesis H_0 while the alternative hypothesis H_1 is z_n given classification c is strange. For significance testing, the p-value is often defined as the probability of observing a point in the sample space which can be considered as extreme as, or more extreme than, the observed samples. During a significance testing, the smaller the p-value, the greater the evidence against the null hypothesis.When the p-value is smaller than the significance level, the null hypothesis H_0 is rejected, and one accepts the alternative hypothesis H_1. Otherwise, the null hypothesis is neither rejected nor accepted, while the alternative hypothesis is rejected. Note that the hypothesis that we intend to prove is always assigned as the alternative hypothesis.

[15] predicts the class of a particular testing example for the largest "credibility" p-value from all possible (transductive) classifications, and assigning as its "confidence" value one minus the 2nd largest p-value. The confidence value indicates how improbable the classifications other than the predicted

classification are and the credibility value shows how suitable the training set is for the classification of that testing example.

V. Quality of Information

Let p_i be the p-values obtained for a particular example of the possible classification $i = 1, \cdots, n$ respectively. Sort the sequence of p-values in descending order so that the first two p-values, say p_j and p_k are the two highest p-values with classifications j and k respectively. Without loss of generality, we assume p_j to be the higher p-value between the two p-values.The predicted classification for the example is j with p-value p_j. This value defines the credibility of the predicted classification. If p_j is not high enough,the prediction is rejected [17]. The lower p-value, p_k, is used to calculate a confidence value on the predicted classification. Note that the smaller the confidence the larger the ambiguity regarding the top choice.

We consider four possible cases of p-values, p_j and p_k:

- Case 1: p_j high and p_k low. Prediction has high credibility and high confidence value.
- Case 2: p_j high and p_k high. Prediction has high credibility but low confidence value.
- Case 3: p_j low and p_k low. Prediction has low credibility but high confidence value.
- Case 4: p_j low and p_k "high". Prediction has low credibility and low confidence value.

In Case 4, since we know that $p_j > p_k$, p_k cannot be too high and may be close to p_j. Uncertainty in prediction occurs in Case 2, 3 and 4. Note also that uncertainty of prediction occurs if $p_j \approx p_k$. We define "closeness"

$$I(z_n) = |p_j - p_k|$$

which indicates the quality of information possessed by the testing example. As $I(z_n)$ approaches 0, the more uncertain we are about classifying the testing example. The addition of this example to the training data thus provides new information about the structure of the data-set.

During active learning, one specifies a threshold value ϵ for $I(z_n)$, and if $I(z_n) < \epsilon$, a decision is made to include z_n in the training set. The threshold value in our experiment is empirical.

VI. KNN-Transductive Confidence Machine (KNN-TCM)

In this section, we give the algorithm for KNN-TCM introduced in [15].

- for $i = 1$ to m (number of training examplars)
  - Find and store D_i^y and D_i^{-y}
- end for
- Calculate the alpha (strangeness) values, for all the training exemplars
- Calculate the similarity $dist$ vector as the distances of the new exemplar from all the training exemplars
- for $j = 1$ to C(number of classes) do
  - for every training exemplar t classified as j do
    - if $D_{tk}^j > dist(t)$, re-calculate the alpha value of exemplar t
  - end for
  - for every training exemplar t classified as non-j do
    - if $D_{tk}^{-j} > dist(t)$, re-calculate the alpha value of exemplar t
  - end for
  - Calculate alpha value for the new exemplar classified as j
  - Calculate p-value for the new exemplar classified as j
- end for
- Predict the class with the largest p-value
- Output as confidence one minus the 2nd largest p-value
- Output as credibility the largest p-value

VII. Bias-Variance Trade-off for Classification and Early Stopping

The question addressed in this section is when to stop being engaged in active learning. "There is a folklore that the generalization error decreases in an early period of learning, reaches a minimum and then increases as training goes on, while the training error monotonically decreases. Therefore, it is considered better to stop training at an adequate time, a technique often referred to as early stopping." [1]. They further note that "the asymptotic gain in the generalization error is small if we perform early stopping, even if we have access to the optimal stopping time". When the number of training examples is finite, the true risk function is different from the empirical risk function to be minimized. Methods to avoid over-fitting include cross-validation, regularization and model selection.

During active learning, there is a point in time that over-fitting phenomenon emerges when the classification error starts increasing. Early stopping is then required. Once the active learning process is terminated there is the option to continue learning using random selection and eventually coming to a complete stop. This leads to our novel and hybrid framework, which starts by (randomly) choosing the initial size and composition of the training set, then actively choosing how to augment the training set, early stopping, switching to random selection to eventually stop and freeze what has been learned. Early-stopping is addressed using the bias-variance trade-off for classification.

The mean squared prediction error (MSE) for the function estimate $\hat{f}$ over a training set D at x is

$$MSE(\hat{f}(x, D)) = bias^2(\hat{f}(x, D)) + var(\hat{f}(x, D))$$

$$bias\hat{f}(x, D) = f(x, D) - E(\hat{f}(x, D))$$

where the bias reflects the sensitivity of the training set D to the choice of the function estimate $\hat{f}$, while the variance reflects the sensitivity of the function estimate $\hat{f}$ to the training set D. It is thus desirable to have both low squared-bias and low variance in order to minimize the MSE. Due to the bias-variance trade-off, minimizing both components concurrently

is not realistic. The learner tries to gain in an efficient fashion as much information as possible, concerning the target function, from the training set D.

In function estimation, the estimation error is additive in bias^2 and variance. On the other hand, there is a nonlinear multiplicative relationship between bias and variance for classification problem. [6] argues that given a training set, low estimation variance is more important than low squared estimation bias. Simple but highly biased classification methods such as naive Bayes and nearest neighbor are apparently successful, using moderate to large training set, due to the fact that they are stable methods with low variance.

In active learning, one starts with a relatively small initial training set with the objective of gaining "maximum" information concerning the target function with the addition of some new examples from a larger set of available data. One aims at increasing the sensitivity of the function estimate with the addition of "informative" new examples. Our basic objective is thus to lower the estimation bias fast in terms of the number of new examples added to the training set and to stabilize the learner so that it has a low variance. When and how soon should one stop using our selection criteria which lowers the estimation bias?

Define the *boundary bias* as in [6]

$$b(f, E(\hat{f})) = sign(\frac{1}{2} - f)(E(\hat{f}) - \frac{1}{2})$$

The sign of boundary bias affects the effect of variance to the classification accuracy. As long as $b(f, E(\hat{f})) < 0$, one can decrease the variance to ensure an accurate classification. On the other hand, a non-negative boundary bias will result in a deterioration of accuracy given even a small variance. As a consequence, one needs to find that point in time, during active learning, shortly before the boundary bias becomes positive, and switch from active to random selection. The sign for the boundary bias and the increase in mis-classification are two manifestations of the same phenomena, i.e., over-training.

Although [6] suggests lowering variance as a more important factor in reducing classification error, he also pointed out that reducing estimation bias is a way to bring boundary bias down to a negative value.

Figure 2 and 3 display in a graphical form the fact that the classification error, while initially decreasing, starts to increase (around that point where one expects the boundary bias to change from negative to positive). As the number of "high information" examples increases in the training set, the structure of the training set is no longer uniform. It becomes over-biased. This is the point in time one needs to stop active learning and switch to random selection. For a fixed estimation bias, the variance generally decreases by increasing the size of the training set. The use of random selection is how one can decrease the variance and stabilize learning. During random selection the estimation bias is also reduced although not as fast as during active selection. Active selection thus ensures a speedy reduction in estimation bias, while random selection ensures a reduction in variance for stabilization purposes. The objective of using less data exemplars and time to attain comparable classification accuracy can now be achieved.

VIII. Active Learning (AL) Algorithm

Our new (AL) algorithm iteratively adds examples with high quality of information, i.e., low $I(z_n)$(see Sect. V), to the current training set T. The algorithm is initialized with some training set P, which consists of 10% of the whole training set TS, and quality of information threshold $\epsilon = 0.1$. During each iteration, one selects one example for each of the possible classifications to ensure a balanced training set. The classification performance is derived using the calibration set for that iteration. When the stopping criteria is reached, one switches to random selection of examples. The algorithm terminates when no further performance improvements can be obtained, i.e., the difference in performance falls below some threshold.

1) Split the original training TS set into the proper training set P and the calibration set C using the ratio $1:9$, such that $TS = P \cup C$ and $T = P$
2) Choose (randomly) c, $c \in C$, and remove c from C, i.e., $C = C \backslash \{c\}$
3) Use T to compute $I(c)$ (See Sect. VI).
4) if the quality of information $I(c) < \epsilon$, c augments the training set T used to build the new (KNN-TCM) classifier,i.e., $T = T \cup \{c\}$.
5) Repeat Step 2, 3 and 4 till (early) stopping criteria is reached.
6) Select examples from C randomly, disregarding their quality of information until performance cannot be further improved.

IX. Experimental Results

The classifier used during testing are KNN-TCM [15], which couples the k-nearest neighbors (KNN) and the transductive confidence machine, and KNN. Three different experiments are done using random selection, active selection and hybrid selection. During active selection, all the training example would iteratively augment the initial training set using the quality of information criteria. During hybrid selection, the early stopping criteria switches selection from active to random. The switching point is determined from the observed increase in the classification error, which is related to the "over-bias" curve. For each of the three type of selection experiments we performed 100 trials and report the average results.

The experimental results reported are based on Ripley's simulated data [16]. The noisy data-set comes from a non-separable (linearly and non-linearly) two-class classification problem (see Figure 1) where each population is an equal mixture of two bivariate normal distribution. Each class consists of 125 points and the testing set consists of 1000 points. It is reported [16] that the 5-nearest neighbor (5NN) method and the Bayes rule yield 13% and 8.0% classification errors, respectively, and that their standard deviation is around

1%. The 5NN is referred to in our comparative performance evaluation experiments as the “standard” method.

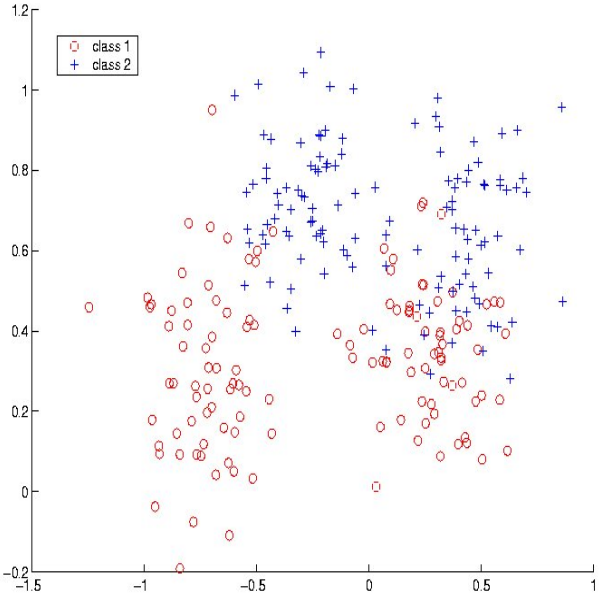

Fig. 1. Ripley’s two-class simulated data

The initial training set P, of size 24, consists of 12 examples randomly drawn from each class. The rest of the examples from the original training set make up the calibration set used for selection purposes. During the selection process, example is randomly chosen one at a time and using the selection criteria, a decision is made whether to add the example to the proper training set. During certain trials, the active and hybrid selection may exhaust the calibration set before the training set size limit which is set to 55.

From looking at the experimental data for active selection in Figure 2 and 3, one can see that there is an improvement in performance initially (steps 24-32), And then the performance becomes worse. This is due to the over-bias phenomenon, which changes the overall structure of the data-set as more new training examples accumulate over some specific regions in the input space. From Figure 2 and 3, one can also see that hybrid selection outperforms consistently random selection . The performance of hybrid selection using both 5NN-TCM and 5NN classifiers reaches the performance level of standard 5NN very quickly. For 5NN-TCM, by adding only 18 new examples to the initial training set, the classification error drops to the 13% classification level of standard 5NN. The size of the training set (42) is less than twice the size of the initial training set. Random selection requires 59 training examples, while standard 5NN uses all the 250 training examples for a classification error of 13%. Similarly, hybrid selection, using as its classifier 5NN, requires only around 49 training examples for a classification error of 13%. We also observed that during the early stage of the experiments the average classification error for random selection is only about as good as the worst classification error of active selection for both 5NN-TCM and 5NN (steps 28-38 and 28-32, respectively).

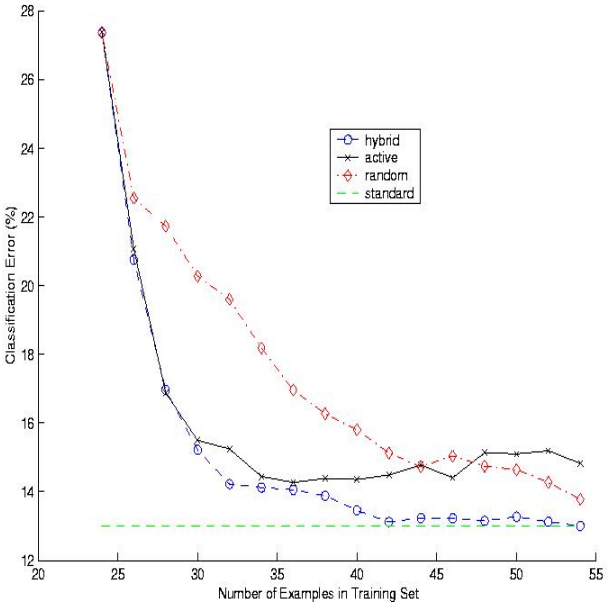

Fig. 2. 5NN-TCM as the classifier: Comparing Hybrid, Active, Random Selection of training set with Training using all examples

X. Conclusions

This paper describes a novel hybrid learning strategy that includes active and random selection of training examples. The active learning component is based upon universal p-value measures of confidence derived from the theory of algorithmic randomness, and transductive inference. The early stopping criteria for active learning is based on the bias-variance trade-off for classification. This corresponds to that learning instance when the boundary bias becomes positive, and requires one to switch from active to random selection of learning examples. The sign for the boundary bias and the increase in the classification error are two manifestations of the same phenomena, i.e., over-training. The experimental results presented show the feasibility and usefulness of our novel approach using a two-class classification problem that displays significant overlap. Our hybrid learning strategy achieves competitive performance against standard nearest neighbor methods using much fewer training examples.

Current research involves application of our hybrid active learning strategy on higher dimension data-sets and the estimation of boundary bias using re-sampling or bootstrap method which will be used to replace our empirical search for the stopping criteria.

Future research directions are related to expanding on the hybrid learning framework, additional fields of applications, more experiments and meta-analysis of results. In particular,

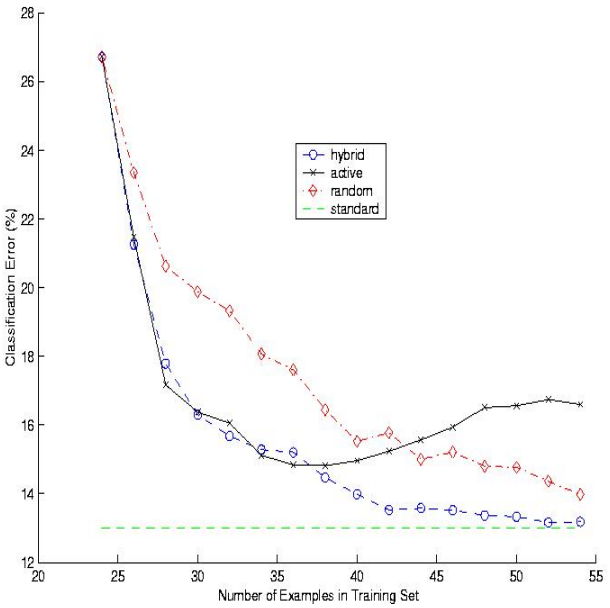

Fig. 3. 5NN as the classifier: Comparing Hybrid, Active, Random Selection of training set with Training using all examples

we plan to allow for the possibility that during the active learning stage, candidate examples from the calibration set whose quality of information is low and are thus eliminated, can be reevaluated later on for possibly augmenting the training set. As the landscape changes the quality of information can change too. Another possible extension is related to the choice of candidate examples and conscience mechanisms during active learning. The order of presentation is important and search diversity is important. As learning involves search, we plan to explore ways and means to enhance active learning using stochastic search, e.g., particle swarming and tabu search.

REFERENCES

[1] S. Amari, N. Murata, K. R. Muller, M. Finke and H. Yang, "Asymptotic statistical theory of overtraining and cross validation," METR 95-06, University of Tokyo, 1995.

[2] K. P. Bennett and A. Demirez, "Semi-supervised support vector machines," in *Neural Info. Processing Systems*, 11, MIT Press, 1998.

[3] C. M. Bishop, *Neural Network for Pattern Recognition*, Oxford University Press, 1995.

[4] D. Cohn, Z. Ghahramani and M. I. Jordan, "Active learning with statistical models," in *Journal of Artificial Intelligence Research*, 4: 129-145, 1996.

[5] R. Duda, P. Hart and D. Stork, *Pattern Classification*, 2nd ed, Wiley-Interscience Inc, 2001.

[6] J. Friedman, "On bias, variance, 0/1-Loss, and the curse-of-dimensionality," in *Data Mining and Knowledge Discovery*, 1: 55-77, 1997.

[7] A. Gammerman, V. Vovk, and V. Vapnik, "Learning by transduction," in *Uncertainty in Artificial Intelligence*, 148-155, 1998.

[8] M. Hasenjager and H. Ritter, "Active Learning with Local Models," in *Neural Processing Letters* 7:107-117, 1998

[9] R. Hecht-Nielsen, *Neurocomputing*. Addison-Wesley, 1989.

[10] M. Li and P. Vitanyi, *An Introduction to Kolmogorov Complexity and Its Applications*, 2nd ed. Springer-Verlag, 1997.

[11] M. Lindenbaum, S. Markovich and D. Rusakov, "Selective sampling for nearest neighbor classifiers," in *American Association for Artificial Intelligence*, 1999.

[12] T. Melluish, C. Saunders, I. Nouretdinov and V. Vovk, "Comparing the bayes and typicalness," in in *ECML 2001, LNAI 2167: 360-371*.

[13] I. Nouretdinov, T. Melluish and V. Vovk, "Ridge regression confidence machine," in *Proc. 18th Int. Conf. on Machine Learning*, 2001.

[14] M. Plutowski and H. White, "Selecting concise training sets from clean data," in *IEEE Transactions on Neural Networks*, 4: 305-318, 1993.

[15] K. Proedrou, I. Nouretdinov, V. Vovk and A. Gammerman, "Transductive confidence machines for pattern recognition," in *ECML 2002: 381:390*.

[16] B. D. Ripley. "Neural networks and related methods for classification (with discussion)," in *J. Roy. Statist. Soc. B*, 56: 409-456, 1994.

[17] C. Saunder, A. Gammerman and V. Vovk, "Transduction with confidence and credibility," in *Proc. of IJCAI '99*:722-726, 1999.

[18] G. Schohn and D. Cohn, "Less is more: active learning with support vector machines," in *Proc. 17th Int. Conf. on Machine Learning*:839-846, 2000.

[19] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," in *Proc 17th Int. Conf. on Machine Learning*:999-1006, 2000.

[20] V. Vapnik. *The Nature of Statistical Learning Theory*, 2nd ed. Springer-Verlag, 2000.

[21] V. Vovk, A. Gammerman and C. Saunders, "Machine-learning applications of algorithmic randomness," in *Proc. 16th Int. Conf. on Machine Learning*: 444-453, 1999.

[22] S. Weerahandi. *Exact Statistical Methods for Data Analysis*, Springer-Verlag, 1994.