Query Based Hybrid Learning Models for Adaptively Adjusting Locality

Yuanchun Zhu, Guyue Mi, and Ying Tan, *Senior Member, IEEE*
Key Laboratory of Machine Perception (Ministry of Education),
Department of Machine Intelligence, School of Electronics Engineering and Computer Science,
Peking University, Beijing, 100871 P.R. China
{ychzhu, miguyue, ytan}@pku.edu.cn

***Abstract*—Local learning employs locality adjusting mechanisms to give local function estimation for each query, while global learning tries to capture the global distribution characteristics of the entire training set. When fitting well with local characteristics of each individual region, the locality parameter may help local learning to improve performance. However, the real data distribution is impossible to get for a real-world problem, and thus an optimal locality is hard to get for each query. In addition, it is quite time-consuming to build an independent local model for each query. To solve these problems, we present strategies for estimating and tuning locality according to local distribution. Based on local distribution estimation, global learning and local learning are combined to achieve a good compromise between capacity and locality. In addition, multi-objective learning principles for the combination are also given. In implementation, a unique global model is first built on the entire training set based on empirical minimization principle. For each query, it is measured that whether the global model can well fit the vicinity space of the query. When an uneven local distribution is found, the locality of the model is tuned, and a specific local model will be built on the local region. To investigate the performance of hybrid models, we apply them to a typical learning problem—spam filtering, in which data are always found to be unevenly distributed. Experiments were conducted on five real-world corpora, namely PU1, PU2, PU3, PUA, and TREC07. It is shown that the hybrid models can achieve a better compromise between capacity and locality, and hybrid models outperform both global learning and local learning.**

I. Introduction

In learning theory, there generally exist two kinds of statements [1] about the learning problem, namely global learning and local learning. The main difference between them lies in whether to fit a model to entire or partial training set. According to global learning, a unique model would be built to minimize empirical error on the entire training set, and it is assumed all the data may come from a fixed unknown distribution. On the contrast, local learning [1], [2] assumes that data are unevenly distributed in the input space, and models (functions) are estimated in the vicinity of a query. For analyzing learning models, Vapnik [1] presented two important learning principles, namely Empirical Risk Minimization (ERM) and Structural Risk Minimization. Based on these two principles, learning process was formalized as a tradeoff between capacity and locality.

When building global models, capacity of models may be tuned so as to match the numbers of examples in training set. In literature, many global algorithms, including Naive Bayes (NB) [3], Support Vector Machine (SVM) [4], Artificial Neural Network (ANN) [5], decision trees [6], [7], etc., have been proposed and applied to many real-world problems. Although these algorithms build models based on different induction rules and assumptions, they all try to minimize empirical error on the entire training set, and tune their capacity through control parameters and the model structure.

Local learning shares some similarities with global learning, and also has mechanisms for adjusting model capacity. Nevertheless, compared to global learning, local learning has one more free parameter for tuning locality. Besides some classical local algorithms, such as k-Nearest Neighbors (k-NN) and Radial Basis Function networks (RBF) [8], many novel local algorithms have been proposed in recent years. Most of these novel algorithms can be viewed as hybrid algorithms of k-NN and global algorithms. Some prevalent local algorithms [2], [9]–[12] are reviewed in Section II.

The locality parameter generally differentiates local algorithms from global algorithms. With the parameter, it is theoretically possible to achieve a better compromise between capacity and locality. Nevertheless, it is not that easy to find an optimal locality for each query for a real-world problem. With a bad locality, local learning may perform worse than global learning. In addition, finding an optimal locality and building independent local models for each query would be quite time-consuming.

In this paper, we propose hybrid strategies for combining global learning with local learning, and give multi-objective risk minimization principles for the combination. Our original intuition is to adaptively utilize global learning or local learning according to local characteristics of data distribution. When the data distribution of a local region is approximately identical to that of the whole training set, a global model is utilized, since it fits global distribution better than local models do. However, local models are necessarily built when uneven distribution is found. The basic idea is to adaptively tune locality based on distribution of the vicinity around a query. To investigate the performance of the hybrid strategies, they are applied to a typical learning problem—spam filtering.

The remainder of this paper is organized as follows. Section II introduces learning principles for local and global learning, and local algorithms are reviewed. In Section III,

multi-objective learning principles are presented, and hybrid strategies for combining global and local models are designed and discussed. In Section IV, we analyze the experimental results on five spam corpora. Finally, conclusion and future direction are given in Section V.

II. Related works

A. learning principles

Learning is a process of choosing the optimal approximation to the supervisor's response. To find an optimal learning model, many learning principles have been proposed, which help in improving performance of classifiers. In [1], [13], learning was defined as a problem of Risk Minimization (RM). It was pointed out that the goal of learning was to minimize a risk function.

$$R(w) = \int L(y, f(x, w))dP(x, y), \tag{1}$$

$$R(w, b, x_0) = \int L(y, f(x, w)) \frac{K(x, x_0, b)}{\parallel K(x_0, b) \parallel} dP(x, y), \tag{2}$$

where w represents parameters of a learning model, b controls the degree of locality, and $L(y, f(x, w))$ measures the difference between the estimation of a model and the supervisor's response. $R(w)$ denotes global RM, while $R(w, b, x_0)$ denotes local RM.

Risk minimization was a natural way of measuring models' effectiveness. Nevertheless, it was unable to get the joint probability distribution $P(x, y)$ in reality. As information about training data was available, empirical risk minimization (ERM) principle was proposed [1] as an alternative to RM. The empirical risk function was defined on training set.

$$E(w) = \frac{1}{l} \sum_{i=1}^{l} L(y_i, f(x_i, w)). \tag{3}$$

$$E(w, b, x_0) = \frac{1}{l} \sum_{i=1}^{l} L(y_i, f(x_i, w)) \frac{K(x, x_0, b)}{\parallel K(x_0, b) \parallel}. \tag{4}$$

, where $E(w)$ denotes global ERM, and $E(w, b, x_0)$ denotes local ERM.

To ensure that ERM can converge uniformly to the RM and provide a guaranteed bound for RM, Vapnik-Chervonenkis-dimension (VC-dimension, which is a measure of capacity) and locality need to be controlled. The controlling principle, namely Structural Risk Minimization (SRM) [1], [14], was then proposed by introducing a nested structure on learning models.

$$S_1 \subset S_2 \subset, \cdots, \subset S_n. \tag{5}$$

The VC-dimension h_p of each subset S_p satisfies

$$h_1 < h_2 < \cdots < h_n. \tag{6}$$

Both VC-dimension of $f(x_i, w)$ and $K(x, x_0, b)$ may affect the capacity of a learning model. According to SRM, the parameters w and b, and an optimal element S^* are selected, so as to minimize the guaranteed bound (guaranteed risk) for RM.

B. local learning algorithms

In [1], vapnik first proposed a statement of local learning problem. Instead of building a unique model over the full train set (i.e. input space), multiple local models were built up according to characteristics of separate regions. Two reasons of using local models were given and analyzed: (1) There may not exist a unique predictor (model) for the full input space, but several local predictors for specified regions. Local learning may help when global learning cannot achieve good performance. (2) Local learning provided a locality parameter b, which may help in finding deeper minima of the guaranteed risk.

To demonstrate the performance of local learning, [2] presented a local network for character recognition. According to the algorithm, the fifth layer of network was implemented by using a local algorithm, where k closest training data were utilized in building local learning model for each specific testing pattern. In essence, the algorithm was a combiner of k-Nearest Neighbor (k-NN) and Neural Network. It was shown that the local learning algorithm achieved better performance than global network, which resulted from a better control between locality and capacity.

Based on the idea of local learning, [9] proposed a local Support Vector Machine, namely SVM-KNN, for visual category recognition. The algorithm consisted of two main phases. First,k nearest neighbors of a query (testing pattern) were selected according to distances. Then the k nearest neighbors were utilized for building local SVM, which was used for labeling the query. As finding k nearest neighbors was quite time-consuming, the algorithm was speeded up by using two strategies: pruning partial neighbors by computing a "crude" distance and caching pairwise distance. The performance of SVM-KNN was investigated on large multi-class data sets, on which it outperformed both SVM and nearest neighbor. In [12], Kecman et al. proposed a Locally Linear SVM (LLSVM), which shared some similarities with local learning and SVM-KNN. According to LLSVM, linear models were built for separate regions, which differentiated it from other local learning algorithms. It was shown that locally linear model averagely performed better than non-linear ones.

One major problem of local learning lies in its high computational complexity. To overcome the problem, cheng et al. [10], [11] presented a framework of localized SVM and an efficient algorithm, namely Profile SVM (PSVM), by using clustering techniques. It was demonstrated that both LSVM and PSVM outperformed global SVM, and PSVM worked faster than LSVM. Nevertheless, it could be seen that PSVM was still quite slower than global SVM.

III. Hybrid models for combining global learning and local learning

A. The necessity of building hybrid models

Locality control parameter may help local learning find optimal match between capacity and the numbers of examples. However, this mechanism causes new problems to local

learning. To tune locality according to data, many local models need to be built for separate regions, which is very time-consuming. In addition, it is hard to choose an optimal locality for a specific problem. Bad selection of a locality may lead to over-fitting problem, and small locality may cause the local model sensitive to the outliers. To solve these problems, it is necessary to study strategies of adaptively tuning locality, and thus to combine global learning with local learning.

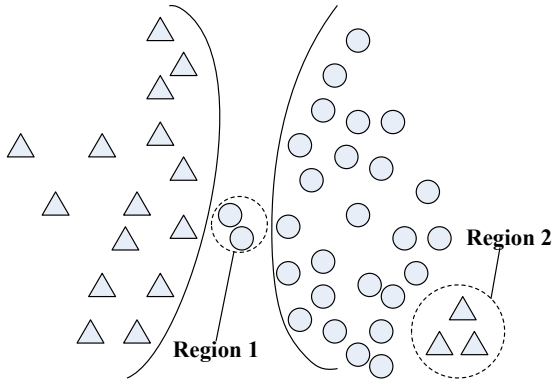

Fig. 1. Example of pattern space with uneven distribution of data

Figure 1 illustrates a pattern space with uneven distribution of data, in which a global decision boundary is also given. From this example, it can be seen that the general data distribution is well described by the global model. However, data distribution in region 1 and region 2 are quite different. Two specific local models need to be built for these two regions. In building the two models, their locality may necessarily be tuned according to the separate data distribution.

B. multi-objective learning principles

In literature, learning principles formulate learning problem as a process of function selection, where a single optimal approximation function is selected for a specific region (locally or globally). By using an additional locality parameter, local learning seems to be superior in minimizing empirical error on training data. Nevertheless, as the true distribution of data is unknown for a real-world problem, it is impossible to decide whether a limited locality may help reduce generalization error for a specific problem. In addition, no practical criteria is available in find an optimal value for the locality parameter. To provide mechanisms for combining learning models with varied locality, we put forward a Multi-Objective Risk Minimization (MORM) principle by combining global RM and multiple local RMs.

$$R(w,b,x_0)=\begin{cases} Minimize & R_1(w,b_1,x_0), \\ Minimize & R_2(w,b_2,x_0), \\ \vdots & \vdots \\ Minimize & R_n(w,b_n,x_0), \end{cases}, \quad (7)$$
$$Subject\ to:\ 0<b_1<b_2<\ldots<b_n,$$

where each R_p denotes an objective with regard to b_p,

$$R(w,b_p,x_0)=\int L(y,f(x,w))\frac{K(x,x_0,b_p)}{\parallel K(x_0,b_p)\parallel}dP(x,y), \quad (8)$$

and R_n denotes the global RM, in which $b_n\to\infty$.

Accordingly, Multi-Objective Empirical Risk Minimization (MOERM) is given as an approximation to MORM, since the true distribution of P(x,y) cannot be got.

$$E(w,b,x_0)=\begin{cases} Minimize & E_1(w,b_1,x_0), \\ Minimize & E_2(w,b_2,x_0), \\ \vdots & \vdots \\ Minimize & E_n(w,b_n,x_0), \end{cases}, \quad (9)$$
$$Subject\ to:\ 0<b_1<b_2<\ldots<b_n,$$

where each E_p denotes an objective with regard to b_p,

$$E(w,b_p,x_0)=\frac{1}{l}\sum_{i=1}^{l}L(y_i,f(x_i,w))\frac{K(x,x_0,b_p)}{\parallel K(x_0,b_p)\parallel}, \quad (10)$$

and E_n denotes the global ERM, in which $b_n\to\infty$.

As the parameter b does not affect capacity but only locality, the existing SRM principle can be directly applied to find an optimal element S^*, except that MOERM should be utilized in minimizing each element S_p of a structure.

C. Strategies for combining global learning and local learning

According to MOERM principle, multiple models with varied locality are built, and they are combined to achieve an optimal compromise between capacity and locality. Fig. 2 depicts a natural and direct strategy for combining global learning and local learning. Using the strategy, a global model and multiple local models with varied locality are built simultaneously. These models are then combined by evaluating how well are their capacity and locality matched.

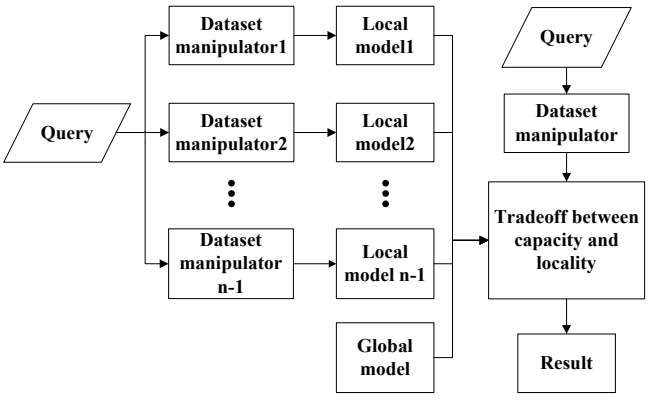

Fig. 2. A simple strategy for combing models with varied locality

The strategy is useful in minimizing empirical error. However, the strategy is quite time-consuming, as multiple models need to be built for each query (testing instance). A cascade combination strategy is shown in fig. 3. By using the cascade strategy, learning models with varied locality are successively

built. A successor model would be built only when the predecessor models cannot achieve a good compromise between capacity and locality. To reduce computational complexity, global model is built at the first stage and preferably utilized for a query, since a uniform global model can be used for all the queries (test instances). When the capacity of the global model cannot well match number of examples, local models with smaller locality are necessarily built. Locality of models are adaptively adjusted based on the data distribution around a query.

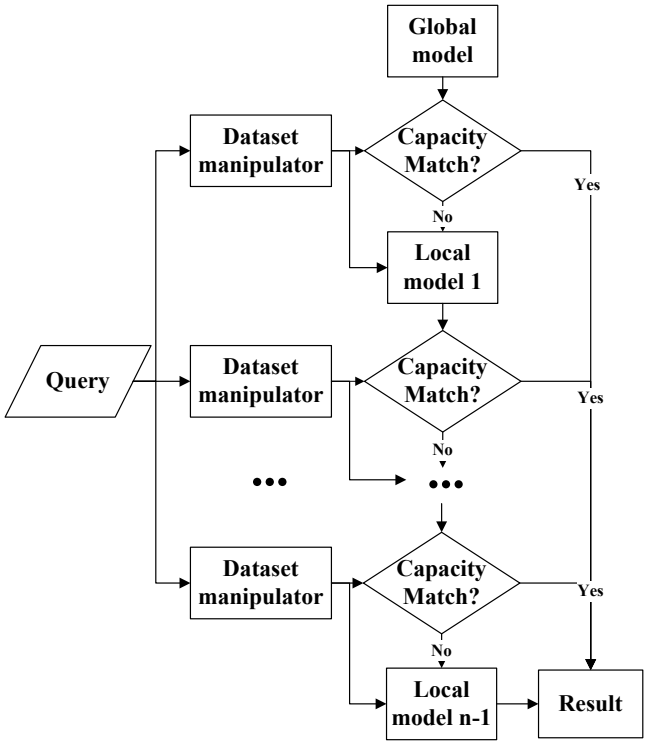

Fig. 3. Query based cascade strategy for adaptively tuning locality of models

D. Local tradeoff between capacity and locality

In implementing the cascade strategy, effective evaluation mechanisms need to be adopted, so as to justify the necessity of tuning locality of models. The main purpose of tuning locality is to minimize generalization error (actual risk) of a given model, and thus generalization error is an ideal evaluation mechanism. However, it is unable to get generalization error for real-world problems. By taking inspiration from danger theory, we estimate the match degree between capacity and locality based on the characteristics of local data distribution around a query.

In [15], Matzinger proposed a novel biological immune paradigm—Danger theory(DT), which well explained how an immune recognition was triggered. Based on this theory, many artificial immune based learning models [16], [17] were designed by mimicking the recognition (learning) process between antibody and antigen. One important principle of DT is the usage of a danger zone, which provides a way of recognition confirmation. From the perspective of machine learning, a query represents an antigen [18], [19], a learning model corresponds to antibodies, and function of danger zone amounts to local estimation of model's performance. Therefore, it is logical to design an artificial danger zone, and utilize it to estimate whether model's capacity matches locality of data. In addition, local estimation mechanisms are not only found in danger theory based models but also in traditional learning approaches. In [20], local estimation was utilized for computing weights of classifiers in building multi-classifier systems.

Based on these analyses, we apply local estimation mechanism in maintaining tradeoff between capacity and locality. Two ways of local estimation may be considered. One strategy is defining a danger zone for each query on the training set. As the data in a danger zone have already been seen by a learning model, the corresponding estimation may be inaccurate and over-optimistic. To solve the problem, the other one is to divide labeled data into a training set and an independent validation set, and define danger zones on the validation set. Preliminary results show that the latter strategy performs much better.

E. Hybrid models for combining models with varied locality

Algorithm 1 demonstrates a cascade strategy for combining global learning and local learning. The labeled data are divided into two independent sets: a training set and a validation set. The training set is utilized to build a global model and multiple local models. The independent validation set is utilized to evaluate the performance of current models, and determine the necessity of tuning locality and building up suitable models. After segmentation of data set, a unique global model is built on the whole training set. In classifying test instances, the global model is first applied, and its performance is evaluated on validation set. In performance evaluation, k_t nearest neighbors to a test instance are found on the validation set, and they are utilized to estimate whether a model well fits to the local distribution around the test instance. When the global model cannot fit a local distribution well, the locality parameter is gradually decreased and local models are built.

According to the algorithm, the locality parameter is gradually adjusted with a step-width of Δb. In implementation, the parameter Δb can be either set to a fixed value or adaptively estimated based on local data around a query. Once an optimal locality is found, the test instance is classified using the optimal model or a combination of existing models. In combining a global model and multiple local models, various strategies, including best model selection, majority voting and weighted voting, can be adopted.

In essence, Algorithm 1 presents a multi-stage strategy for gradually tuning locality. Although it helps achieve an optimal compromise between capacity and locality, the tuning process of the algorithm may cost much time, especially when the step-width Δb is small. One alternative is to simplify this multi-phase strategy into a two-phase strategy, which is shown in Algorithm 2. When a global model cannot work well, multiple

Algorithm 1 A cascade strategy for combining global learning and local learning

Divide the labeled data into two set:
  training set $T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, and
  validation set $V = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$;
Initialize the locality parameter: $b = n$;
Build a global learning model $f_g(x, w)$ on the whole training set T, by minimizing $E(w)$;
for an instance x_0 in the testing set **do**
  Select k_t samples from V with minimal $K(x, x_0)$;
  Use the k_t samples to evaluate the performance of $f_g(x, w)$: $p = 1 - \frac{1}{l}\sum_{i=1}^{k_t} L(y_i, f(x_i, w))\frac{K(x,x_0)}{\|K(x_0)\|}$;
  while $p < \theta$ *and* $b - \Delta b > 0$ **do**
    $b = b - \Delta b$;
    Build a local learning model $f_l(x_i, w, b)$ on T, by minimizing $E(w, b, x_0)$;
    Use the k_t samples to evaluate the performance of existing models:
    $p = 1 - \frac{1}{l}\sum_{i=1}^{k_t} L(y_i, \sum \alpha_j f_j(x_i, w, b_j))\frac{K(x,x_0)}{\|K(x_0)\|}$,
    where α_j is the weight of $f_j(x_i, w, b_j)$;
  end while
  Label x_0 using existing models:
  $y_0 = \sum \alpha_j f_j(x_i, w, b_j)$;
end for

Algorithm 2 A two-phase strategy for combining global learning and local learning

Divide the labeled data into two set:
  training set $T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, and
  validation set $V = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$;
Build a global learning model $f_g(x, w)$ on the whole training set T, by minimizing $E(w)$;
for an instance x_0 in the testing set **do**
  Select k_t samples from V with minimal $K(x, x_0)$;
  Use the k_t samples to evaluate the performance of $f_g(x, w)$: $p = 1 - \frac{1}{l}\sum_{i=1}^{k_t} L(y_i, f(x_i, w))\frac{K(x,x_0)}{\|K(x_0)\|}$;
  if $p < \theta$ **then**
    Estimate a set of suitable locality parameter $B = \{b_1, b_2, \ldots, b_l\}$, according to the global model and its current performance;
    for each $b_p \in B$ **do**
      Build a local learning model $f_l(x_i, w, b_p)$ on T, by minimizing $E(w, b_p, x_0)$;
    end for
  end if
  Label x_0 using existing models:
  $y_0 = \sum \alpha_j f_j(x_i, w, b_j)$;
end for

complementary local models are built simultaneously. Locality tuning is achieved by assigning weights for these models.

F. Relation to multiple classifier combination

In literature, we don't find similar hybrid approaches, except that the Multiple Classifier System (MCS) [20] may share few similar points with the proposed models. However, their principles are quite different in essence. In MCS [20], multiple global classifiers are built on the whole training set. On the contrast, in Combination of Global learning and Local learning (CGL), a global model and multiple local models are built according to distribution characteristics of separate regions. Their difference lie in the following aspects.

- **induction principles:** In most MCS, individual classifiers are built based on different induction principles. However, a unique induction principle is utilized for building global and local models in CGL.
- **capacity and locality:** MCS pays more attention to combining models with varied capacity, while CGL focuses on adaptively tuning locality according to local data distribution.

IV. Experiments.

A. Spam corpus

Spam denotes those unsolicited bulk e-mail (UBE), which cause many problems to our life. To solve spam problems, many learning approaches have been applied to spam filtering, and some benchmark spam corpora have been collected from real-world email streams. To investigate the performance of hybrid models, we conducted experiments on five benchmark corpora PU1, PU2, PU3, PUA, and TREC07[1] using 10-fold cross validation. Each PU corpus has already been partitioned into 10 stratified folds, which facilitates our experiments and evaluation. For TREC07, ten thousand instances were randomly selected from the origin corpus, and were randomly partitioned into 10 stratified folds.

B. Experimental Setup

All the experiments were conducted on a PC with CPU of Intel T7250 and 2G RAM. WEKA toolkit [21] was utilized in implementation of learning models.

C. Investigation of parameters

The effects of parameters were investigated on a small independent corpus, where two thousand instances were randomly selected from TREC07. The small corpus was partitioned into 10 stratified folds, and 10-fold cross validation was utilized. In experiments, Id3 tree was utilized as basic inducer for building models.

1) Danger zone size: In hybrid models, match degree between capacity and locality is estimated on danger zones, which are defined as the vicinity of a query on the validation set. In addition, the parameter controls the compromise between global learning and local learning.

Figure 4 depicts the performance of the hybrid model under different size of danger zone. Four performance measures are utilized to evaluate the model performance. Among the

[1]The four PU corpora can be downloaded from the web site: http://www.aueb.gr/users/ion/publications.html.
The TREC07 corpus can be downloaded from the web site: http://plg.uwaterloo.ca/∼gvcormac/spam/

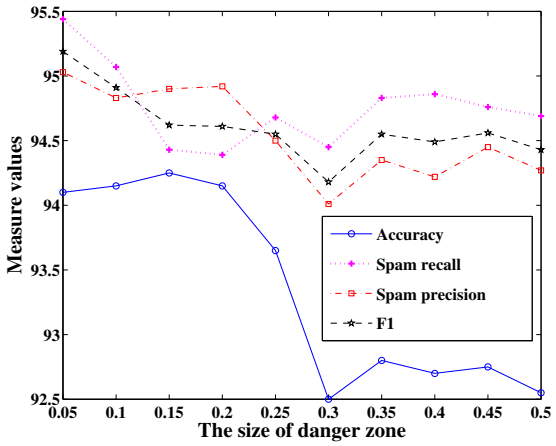

Fig. 4. Performance of the hybrid model with varied danger zone size

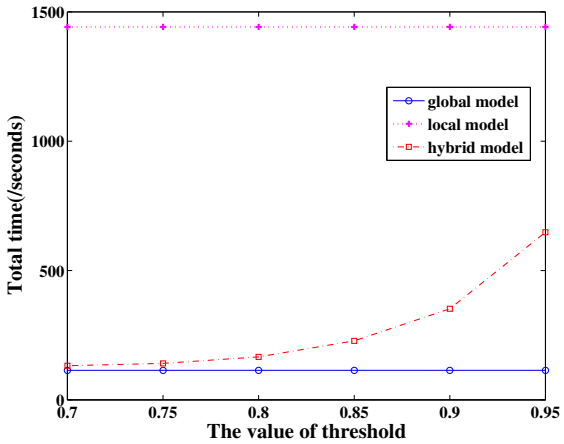

Fig. 5. Complexity of the hybrid model with varied threshold θ

measures, F_1 measure [22] is a combination of spam recall and spam precision. The size value denotes the ratio of validation set, which is utilized in estimation. It can be seen that the model performs better with small danger zones, and the performance declines slightly when the danger zone becomes large. As the main function of danger zone is to estimate local distribution around a query, it is reasonable to use small dangers zone in applications.

2) Effects of threshold: The threshold θ works as a knob for combining global and local models. With larger θ, local models are built with higher probability, and more instances will be classified using combination of multi-locality. Fig. 5 shows the complexity of the hybrid model on different threshold θ. With the growth of θ, the computational complexity gradually increases. However, even with a large θ, the hybrid model still performs faster than local learning. A preliminary conclusion can be drawn that the complexity of the hybrid models lies between the complexity of global learning and local learning. In real-world applications, the threshold may be tuned according to limitation of complexity, and it endows the hybrid models with high scalability. In Section IV-D, experimental results show that the hybrid models achieve better accuracy than both global and local models.

D. Comparison among local models, global models, and hybrid models

The hybrid models were compared to global and local models on five corpora using 10-fold cross validation, and Naive bayes, C4.5, ID3, and SVM were utilized as basic inducers. In the experiments, performance of models with varied locality was fully investigated, so as to show that optimal locality may depend on data distribution of a specific corpus.

Table I shows the accuracy of global and local models on the five spam corpora. In the table, the parameter b controls the locality, and n denotes the number of examples on the training set. From the results, an optimal locality depends on both data distribution and capacity of models. Good performance is achieved for Naive bayes with small locality b, medium locality b is suitable for decision tree C4.5 and Id3, while large locality should be chosen for SVM. As data distribution are different in each corpus, optimal locality are also different on these corpora. For instance, Naive bayes with $b = 0.1n$ performs best on PU3 and TREC, whereas Naive bayes with $b = 0.4n$ achieves highest accuracy on the other three corpora. It is impossible to find a unique optimal locality for all the corpora and classifiers. Thus, locality should be tuned according to local data distribution and the match degree between locality and capacity.

In table II, accuracy of hybrid models on the corpora are given, and the average performance of models in table I is taken as baseline. Two general strategies of combination are fully investigated. One is to combine a global model with a single local model (global+uni-local), and the other is to combine a global model with multiple local models (global+multi-local). In global+multi-local 1, a global model is combined with multiple models, and the locality parameters of these local models are respectively 0.7n,0.4n,and 0.1n. In global+multi-local 2, the locality parameters of those local models are respectively 0.7n, 0.5n, 0.3n.

The experimental results in table II well demonstrate that hybrid models can effectively achieve better performance, compared to global and local models. In most cases, global+multi-local strategy outperforms the global+uni-local strategy. For SVM, global+uni-local with $b = 0.7n$ gives the best performance. In general, both the strategies can improve performance of models, and the improvements result from the mechanism of adaptively tuning locality. Better performance can be expected, when the two strategies are unified in a cascade framework. In addition, the results also verify the reasonability of multi-objective learning principles.

TABLE I
PERFORMANCE OF GLOBAL AND LOCAL MODELS UNDER SETTINGS OF DIFFERENT LOCALITY ON SPAM CORPORA

(a) Naive bayes

Approach \ Corpus	PU1	PU2	PU3	PUA	TREC
global model ($b = n$)	91.47	90.70	87.72	94.65	86.15
local model ($b = 0.7n$)	96.61	89.44	93.61	94.82	86.05
local model ($b = 0.4n$)	97.61	93.38	94.50	95.53	84.12
local model ($b = 0.1n$)	97.34	93.10	95.08	94.30	90.03
Average	95.76	91.66	92.73	94.83	86.59

(b) C4.5

Approach \ Corpus	PU1	PU2	PU3	PUA	TREC
global model ($b = n$)	91.28	89.72	92.25	88.68	95.74
local model ($b = 0.7n$)	91.28	90.42	92.42	87.89	95.52
local model ($b = 0.4n$)	91.38	88.31	92.47	86.93	95.57
local model ($b = 0.1n$)	87.80	88.31	91.26	87.81	95.14
Average	90.44	89.19	92.10	87.83	95.49

(c) Id3

Approach \ Corpus	PU1	PU2	PU3	PUA	TREC
global model ($b = n$)	92.39	88.73	92.18	88.16	94.84
local model ($b = 0.7n$)	92.94	88.87	92.03	88.42	95.27
local model ($b = 0.4n$)	91.65	90.00	92.20	88.86	95.59
local model ($b = 0.1n$)	89.36	86.76	91.50	89.56	95.70
Average	91.59	88.59	91.98	88.75	95.35

(d) SVM

Approach \ Corpus	PU1	PU2	PU3	PUA	TREC
global model ($b = n$)	96.33	94.08	95.33	92.63	96.90
local model ($b = 0.7n$)	96.33	93.24	95.42	92.02	96.89
local model ($b = 0.4n$)	95.96	92.82	95.59	92.11	96.73
local model ($b = 0.1n$)	94.59	91.83	94.04	91.58	95.62
Average	95.80	92.99	95.10	92.09	96.54

V. CONCLUSION

In this paper, learning is formalized as a multi-objective risk minimization problem, and it is pointed out that risk should be minimized with regards to multiple locality. By combining global and local models, we propose hybrid models for adaptively tuning locality according to local distribution around a query. The different essence between hybrid models and MCS are well demonstrated. In the experiments, the performance of hybrid models was investigated on five real-world spam corpora, and results show that hybrid models outperform both global and local models. The success of the hybrid models lies in a good compromise between capacity and locality. In future, we intend to combine multiple global models with multiple local models, and try to adaptively tune both capacity and locality based on local distribution around a query.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (NSFC), under grant number 61170057 and 60875080, and partially supported by the National High Technology Research and Development Program of China (863 Program), with grant number 2007AA01Z453.

Prof. Ying Tan is the corresponding author.

REFERENCES

[1] V. Vapnik, "Principles of risk minimization for learning theory," *Advances in neural information processing systems*, vol. 4, pp. 831–838, 1992.

[2] L. Bottou and V. Vapnik, "Local learning algorithms," *Neural computation*, vol. 4, no. 6, pp. 888–900, 1992.

[3] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A bayesian approach to filtering junk e-mail," in *Learning for Text Categorization: Papers from the 1998 workshop*, vol. 62. Madison, Wisconsin: AAAI Technical Report WS-98-05, 1998, pp. 98–105.

[4] H. Drucker, D. Wu, and V. Vapnik, "Support vector machines for spam categorization," *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 1048–1054, 1999.

[5] J. Clark, I. Koprinska, and J. Poon, "A neural network based approach to automated e-mail classification," in *Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on*. IEEE, 2003, pp. 702–705.

[6] J. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[7] ——, *C4. 5: programs for machine learning*. Morgan kaufmann, 1993.

[8] Y. Lee, "Handwritten digit recognition using k nearest-neighbor, radial-basis function, and backpropagation neural networks," *Neural computation*, vol. 3, no. 3, pp. 440–449, 1991.

[9] H. Zhang, A. Berg, M. Maire, and J. Malik, "Svm-knn: Discriminative nearest neighbor classification for visual category recognition," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2. IEEE, 2006, pp. 2126–2136.

[10] H. Cheng, P. Tan, and R. Jin, "Localized support vector machine and its efficient algorithm," in *Proc. SIAM IntlConf. Data Mining*. Citeseer, 2007.

TABLE II
PERFORMANCE OF HYBRID MODELS THAT COMBINE GLOBAL LEARNING AND LOCAL LEARNING ON SPAM CORPORA

(a) Naive bayes

Approach \ Corpus	PU1	PU2	PU3	PUA	TREC
baseline	95.76	91.66	92.73	94.83	86.59
global+uni-local(b=0.7n)	98.35	94.65	93.75	95.26	88.23
global+uni-local(b=0.4n)	99.08	94.51	94.87	96.84	88.02
global+uni-local(b=0.1n)	98.44	93.66	95.57	94.91	90.08
global+multi-local 1	99.36	94.51	95.01	97.02	88.88
global+multi-local 2	99.08	95.07	94.87	97.19	87.17

(b) C4.5

Approach \ Corpus	PU1	PU2	PU3	PUA	TREC
baseline	90.44	89.19	92.10	87.83	95.49
global+uni-local(b=0.7n)	93.30	92.96	92.78	87.72	95.61
global+uni-local(b=0.4n)	93.58	91.83	93.08	89.30	95.82
global+uni-local(b=0.1n)	93.12	92.54	92.93	88.51	95.59
global+multi-local 1	93.03	92.68	93.20	88.86	95.87
global+multi-local 2	93.76	93.24	93.39	88.33	96.01

(c) Id3

Approach \ Corpus	PU1	PU2	PU3	PUA	TREC
baseline	91.59	88.59	91.98	88.75	95.35
global+uni-local(b=0.7n)	93.85	90.70	92.54	89.12	95.87
global+uni-local(b=0.4n)	94.77	91.83	92.91	89.47	95.79
global+uni-local(b=0.1n)	93.67	91.41	92.06	88.77	95.89
global+multi-local 1	94.40	92.25	93.08	90.79	96.48
global+multi-local 2	94.50	92.39	93.39	90.79	96.25

(d) SVM

Approach \ Corpus	PU1	PU2	PU3	PUA	TREC
baseline	95.80	92.99	95.10	92.09	96.54
global+uni-local(b=0.7n)	96.88	93.52	96.08	93.42	96.87
global+uni-local(b=0.4n)	96.70	93.52	95.81	93.60	96.83
global+uni-local(b=0.1n)	96.33	93.52	95.54	92.63	96.77
global+multi-local 1	96.97	88.73	94.77	93.51	95.79
global+multi-local 2	96.97	88.73	94.70	93.25	95.81

[11] ——, "Efficient algorithm for localized support vector machine," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 4, pp. 537–549, 2010.

[12] V. Kecman and J. Brooks, "Locally linear support vector machines and other local models," in *Neural Networks (IJCNN), The 2010 International Joint Conference on*. IEEE, 2010, pp. 1–6.

[13] V. Vapnik and L. Bottou, "Local algorithms for pattern recognition and dependencies estimation," *Neural Computation*, vol. 5, no. 6, pp. 893–909, 1993.

[14] J. Guyon, V. Vapnik, B. Boser, L. Bottou, and S. Solla, "Structural risk minimization for character recognition," *Advances in neural information processing systems*, pp. 471–471, 1993.

[15] P. Matzinger, "The danger model: a renewed sense of self," *Science*, vol. 296, no. 5566, p. 301, 2002.

[16] A. Secker, A. Freitas, and J. Timmis, "A danger theory inspired approach to web mining," *Artificial Immune Systems*, pp. 156–167, 2003.

[17] Y. Zhu and Y. Tan, "A danger theory inspired learning model and its application to spam detection," *Advances in Swarm Intelligence*, pp. 382–389, 2011.

[18] Y. Tan, C. Deng, and G. Ruan, "Concentration based feature construction approach for spam detection," in *Neural Networks (IJCNN 2009), International Joint Conference on*. IEEE, 2009, pp. 3088–3093.

[19] G. Ruan and Y. Tan, "A three-layer back-propagation neural network for spam detection using artificial immune concentration," *Soft Computing-A Fusion of Foundations, Methodologies and Applications*, vol. 14, no. 2, pp. 139–150, 2010.

[20] K. Woods, W. Kegelmeyer Jr, and K. Bowyer, "Combination of multiple classifiers using local accuracy estimates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 4, pp. 405–410, 1997.

[21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The weka data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[22] T. Guzella and W. Caminhas, "A review of machine learning approaches to spam filtering," *Expert Systems with Applications*, vol. 36, no. 7, pp. 10 206–10 222, 2009.