Exploiting Multiple Classifier Types with Active Learning

Zhenyu Lu
Department of Computer Science
University of Vermont
Burlington, VT 05401
zlu@uvm.edu

Josh Bongard
Department of Computer Science
University of Vermont
Burlington, VT 05401
jbongard@uvm.edu

ABSTRACT

Many approaches to active learning involve training one classifier by periodically choosing new data points about which the classifier has the least confidence, but designing a confidence measure without bias is nontrivial. An alternative approach is to train an ensemble of classifiers by periodically choosing data points that cause maximal disagreement among them. Many classifiers with different underlying structures could fit this framework, but some classifiers are more suitable for different data sets than others. The question then arises as to how to find the most suitable classifier for a given data set. In this work, an evolutionary algorithm is proposed to address this problem. The algorithm starts with a combination of artificial neural networks and decision trees, and iteratively adapts the ratio of the classifier types according to a replacement strategy. Experiments with synthetic and real data sets show that when the algorithm considers both fitness and classifier type for replacement, the population becomes saturated with accurate instantiations of the more suitable classifier type. This allows the algorithm to perform consistently well across data sets, without having to determine *a priori* a suitable classifier type.

Categories and Subject Descriptors

I.2.6 [**Learning**]: Knowledge acquisition

General Terms

Algorithms

Keywords

active learning, adaptive informative sampling

1. INTRODUCTION AND METHODS

In many real world applications, data sets are presented with only feature information and acquiring the labels is expensive. Active learning is a technique that selects a subset of data points for labeling and training. The subset of data points needs to be chosen carefully so that it is informative enough for learning, but small enough to keep the labeling cost manageable.

GECCO'09, July 8–12, 2009, Montréal, Québec, Canada.
ACM 978-1-60558-325-9/09/07.

A common approach to active learning is to iteratively train one classifier and select the data points using some confidence measure. Uncertainty sampling [5] uses c4.5 [7] as the classifier, and chooses the data points that the classifier is most uncertain about. This simple approach appears to be intuitive and simple, but designing confidence measures without bias is nontrivial. Another approach is to use an ensemble of classifiers and choose data points according to the uncertainty of the ensemble. Query by committee [8] suggests that given a set of diversified but partially accurate classifiers, the best data points are those that cause maximal disagreement among the class predictions of the classifiers when supplied without labels. However, the method for finding such classifiers or such disagreement-inducing data points is not specified. The estimation-exploration algorithm [2](EEA) uses multiple evolutionary algorithms to build on query by committee. An evolutionary algorithm is used to optimize a set of diversified classifiers, and another evolutionary algorithm is used to evolve desired data points that maximize disagreement among the classifiers.

Informative sampling [6] (henceforth referred to as IS) applies the EEA as an active learning method for classification. The algorithm works by iteratively evolving an ensemble of classifiers based on the current training set and scanning a portion of the whole data set to select the data point that causes maximal disagreement. Although artificial neural networks [4] were chosen as the classifier type in [6], informative sampling is a general algorithm that could work with a wide range of classifiers. Many types of classifiers exist in the literature, such as decision trees, ANNs, and support vector machines (SVM). Different types of classifiers have different underlying structures, making them differentially suitable for different data sets. The question then arises as to how to find the more suitable classifier type for a given data set.

Many approaches have been suggested in the field of ensemble learning [3] to construct ensembles of heterogenous classifiers, in which existing methods such as c4.5 and SVM were used to train base classifiers. These populations typically contain one instance of each type. The problem is that some methods such as c4.5 are deterministic, which means that only one classifier can be built on a given data set. In our case, different individuals from the same classifier type are expected to be developed such that disagreement among them indicates model uncertainty. In this paper, an algorithm that works with informative sampling is suggested: adaptive informative sampling (henceforth referred to as AIS). The algorithm starts with multiple classifiers

Table 1: Performance comparison between homogeneous and heterogeneous ensembles

	S1	S2	Spambase	Pendigit
30 ANNs	87.8% ± 2.9	92.6% ± 3.3	85.8% ± 2.7	44.7% ± 10.7
30 DTs	76.6% ± 2.6	99.5% ± 0.9	77.9% ± 3.9	69.7% ± 4.8
replacement#1	**90.6% ± 1.7**	**99.9% ± 0.1**	**87.6% ± 1.9**	80.5% ± 2.1
replacement#2	83.5% ± 1.8	**99.9% ± 0.1**	85.2% ± 4.0	**81.2% ±1.6**

drawn from different classifier types and proceeds through two stages. During each iteration, classifiers are trained and replaced based on a replacement strategy, the classifiers in the ensemble are optimized, and data points are chosen and added to the training set. The replacement aims to adapt the ratio of classifier types in the ensemble so that the more suitable model type will eventually saturate the ensemble. Two replacement strategies were studied. The first strategy considers both performance and classifier type by replacing the least accurate classifier with the best classifier from the other type. The second strategy considers only performance by replacing the least accurate classifier with the best classifier in the ensemble. To the best of our knowledge, this work is the first to adapt the ratio of classifier types in a heterogeneous ensemble.

In the current work, ANNs and decision trees were used as example classifier types. Experiments on two synthetic data sets and two real data sets from the UCI Machine Learning Repository [1] show that when working with informative sampling, decision trees outperform ANNs for some data sets, but do not work as well for others. Of the two synthetic data sets, S1 was generated to have a linear decision boundary, where decision trees are expected to perform well, and S2 was generated to have a highly non-linear decision boundary, where ANNs are expected to perform well. The two data sets from UCI are the Spambase Data Set and the Pen-Based Recognition of Handwritten Digits Data Set (henceforth referred to as Pendigit). For each data set, the suggested algorithm with the first replacement strategy performs no worse than using a homogeneous population of the better classifier type, because the better classifier type always saturates the population of classifiers automatically. Table 1 gives the performance comparison of homogeneous ensembles and the heterogeneous ensembles. Each unit in the table represents the mean and standard deviation of predictive accuracy across 30 independent runs.

2. CONCLUSIONS AND FUTURE WORK

In this paper, evidence was provided to demonstrate that different classifier types exhibit different performances when incorporated into the informative sampling algorithm, which indicates the requirement for a heterogeneous ensemble because no one classifier type does consistently well across the data sets tested. An extension of the informative sampling algorithm that adapts the ratio of classifier types in a heterogenous ensemble of classifiers was introduced, which is referred to as the adaptive informative sampling algorithm. This algorithm starts with a combination of multiple classifier types and updates the relative ratio of the classifier types in the population during each iteration. Of the two replacement strategies that are suggested, the one that takes not only the fitness but also the classifier type into account when performing selection and replacement is shown to have a better performance. This suggests that in addition to fitness, active learning techniques that maintain a population of multiple classifier types should explicitly consider classifier type when replacing members of the population. The adaptive informative sampling algorithm that made use of this strategy achieves performances no worse than the informative sampling algorithm when using the better classifier type on the data sets studied. This allows the algorithm to perform consistently well across data sets, without having to determine *a priori* a suitable classifier type.

Although ANN and decision trees were employed here, the adaptive informative sampling algorithm is a generalized algorithm that could work with a wide range of classifier types. Future work will include studying the behaviors of more classifiers with the adaptive informative sampling algorithm and applying the algorithm to more real world data sets.

3. ACKNOWLEDGMENTS

This work is supported in part by National Science Foundation grant EPS-0701410.

4. REFERENCES

[1] A. Asuncion and D. Newman. UCI machine learning repository, 2007.

[2] J. Bongard and H. Lipson. Automating genetic network inference with minimal physical experimentation using coevolution. In *Proceedings of the 2004 Genetic and Evolutionary Computation Conference*, pages 333–345. Springer, 2004.

[3] T. G. Dietterich. Ensemble methods in machine learning. *Lecture Notes in Computer Science*, 1857:1–15, 2000.

[4] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall, Upper Saddle River, NJ, USA.

[5] D. D. Lewis and J. Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Proceedings of the Eleventh International Conference on Machine Learning*, pages 148–156, 1994.

[6] Z. Lu, A. I. Rughani, B. I. Tranmer, and J. Bongard. Informative sampling for large unbalanced data sets. In *4th Workshop on Medical Applications of Genetic and Evolutionary Computation at GECCO 2008*, pages 2047–2054, 2008.

[7] J. R. Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[8] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the Fifth Workshop on Computational Learning Theory*, pages 287–294, 1992.