Adaptive Learning for Algorithm Selection in Classification

Nitin Pise
Research Scholar
Department of Computer Engg. & IT
College of Engineering, Pune, India

Parag Kulkarni
Phd, Adjunct Professor
Department of Computer Engg. & IT
College of Engineering, Pune, India

ABSTRACT
No learner is generally better than another learner. If a learner performs better than another learner on some learning situations, then the first learner usually performs worse than the second learner on other situations. In other words, no single learning algorithm can perform well and uniformly outperform other algorithms over all learning or data mining tasks. There is an increasing number of algorithms and practices that can be used for the very same application. With the explosion of available learning algorithms, a method for helping user selecting the most appropriate algorithm or combination of algorithms to solve a problem is becoming increasingly important. In this paper we are using meta-learning to relate the performance of machine learning algorithms on the different datasets. The paper concludes by proposing the system which can learn dynamically as per the given data.

General Terms
Machine Learning, Pattern Classification

Keywords
Learning algorithms, Dataset characteristics, algorithm selection

1. INTRODUCTION
The knowledge discovery [3] is an iterative process. The analyst must select the right model for the task he is going to perform, and within it, the right model or algorithm, where the special morphological characteristics of the problem must always be considered. The algorithm is then invoked and its output is evaluated. If the evaluations results are poor, the process is repeated with new selections. A plethora of commercial and prototype systems with a variety of models and algorithms exist at the analyst's disposal. However, the selection among them is left to the analyst. The machine learning field has been evolving for a long time and has given us a variety of models and algorithms to perform the classification, e.g. decision trees, neural networks, support vector machines [4], rule inducers, nearest neighbor etc. The analyst must select among them the ones that better match the morphology and the special characteristics of the problem at hand. This selection is one of the most difficult problems since there is no model or algorithm that performs better than all others independently of the particular problem characteristics. A wrong choice of model can have a more severe impact: A hypothesis appropriate for the problem at hand might be ignored because it is not contained in the model's search space.

There is an increasing number of algorithms and practices that can be used for the very same application. Extensive research has been performed to develop appropriate machine learning techniques for different data mining tasks, and has led to a proliferation of different learning algorithms. However, previous work has shown that no learner is generally better than another learner. If a learner performs better than another learner on some learning situations, then the first learner usually performs worse than the second learner on other situations [5]. In other words, no single learning algorithm can perform well compared to the other algorithms and outperform other algorithms over all classification tasks. This has been confirmed by the "no free lunch theorems" [6]. The major reasons are that a learning algorithm has different performances in processing different datasets and that different variety of 'inductive bias' [7]. In real-world applications, the users need to select an appropriate learning algorithm according to the classification task that is to be performed [8],[9]. If we select the algorithm inappropriately, it results in a slow convergence or may lead to a sub-optimal local minimum. Meta-learning has been proposed to deal with the issues of algorithm selection [10]. One of the aims of meta-learning is to help or assist the user to determine the most suitable learning algorithm(s) for the problem at hand. The task of meta-learning is to find functions that map datasets to predicted data mining performance (e.g., predictive accuracies, execution time, etc.). To this end meta-learning uses a set of attributes, called meta-attributes, to represent the characteristics of classification tasks, and search for the correlations between these attributes and the performance of learning algorithms. Instead of executing all learning algorithms to obtain the optimal one, meta-learning is performed on the meta-data characterizing the data mining tasks. The effectiveness of meta-learning is largely dependent on the description of tasks (i.e., meta-attributes).

Ensemble methods are learning algorithms that construct a set of classifiers and then classify new data points by taking a vote of their predictions. Combining classifiers or studying methods for constructing good ensembles of classifiers to achieve higher accuracy is an important research topic [1] [2].

The drawback of ensemble learning is that in order for ensemble learning to be computationally efficient, approximation of posterior needs to have a simple factorial structure. This means that most dependence between various parameters cannot be estimated. It is difficult to measure correlation between classifiers from different types of learners. Also there are learning time and memory constraints. Learned concept is difficult to understand.

So we are trying to propose adaptive learning. We need to propose algorithm for selection of methods for classification task. The datasets are identified and we are trying to map to learning algorithms or methods. We need to generate adaptive

function. Adaptive learning will be built on the top of ensemble methods.

2. RELATED WORKS

Several algorithm selection systems and strategies have been proposed previously [3][10][11][12]. STATLOG [14] extracts various characteristics from a set of datasets. Then it combines these characteristics with the performance of the algorithms. Rules are generated to guide inducer selection based on the dataset characteristics. This method is based on the morphological similarity between the new dataset and existing collection of datasets. When a new dataset is presented, it compares the characteristics of the new dataset to the collection of the old datasets. This costs a lot of time. Predictive clustering trees for ranking are proposed in [15]. It uses relational descriptions of the tasks. The relative performance of the algorithms on a given dataset is predicted for a given relational dataset description. Results are not very good, with most relative errors over 1.0 which are worse than default prediction. Data Mining Advisor (DMA) [16] is a system that already has a set of algorithms and a collection of training datasets. The performance of the algorithms for every subset in the training datasets is known. When the user presents a new dataset, DMA first finds a similar subset in the training datasets. Then it retrieves information about the performance of algorithms and ranks the algorithms and gives the appropriate recommendation. Our approach is inspired by the above method used in [16].

Most work in this area is aimed at relating properties of data to the effect of learning algorithms, including several large scale studies such as the STATLOG (Michie et al., 1994) and METAL (METAL-consortium, 2003) projects. We will use this term in a broader sense, referring both to 'manual' analysis of learner performance, by querying, and automatic model building, by applying learning algorithms over large collections of meta-data. An instance based learning algorithm (K-nearest neighbor) was used to determine which training datasets are closest to a test dataset based on similarity of features, and then to predict the ranking of each algorithm based on the performance of the neighboring datasets.

3. LEARNING ALGORITHMS AND DATASET CHARACTERISTICS

In general there are two families of algorithms, the statistical, which are best implemented by an experienced analyst since they require a lot of technical skills and specific assumptions and the data mining tools, which do not require much model specification but they offer little diagnostic tools. Each family has reliable and well-tested algorithms that can be used for prediction. In the case of the classification task [11], the most frequent encountered algorithms are logistic regression (LR), decision tree and decision rules, neural network (NN) and discriminant analysis (DA). In the case of regression, multiple linear regression (MLR), classification & regression trees (CART) and neural networks have been used extensively.

In the classification task the error rate is defined straightforwardly as the percentage of the misclassified cases in the observed versus predicted contingency table. When NNs are used to predict a scalar quantity, the square of the correlation for the predicted outcome with the target response is analogous to the r-square measure of MLR. Therefore the error rate can be defined in the prediction task as:
Error rate = 1 - correlation^2 (observed, predicted)

In both tasks, error rate varies from zero to one, with one indicating bad performance of the model and zero the best possible performance.

The dataset characteristics are related with the type of problem. In the case of the classification task the number of classes, the entropy of the classes and the percent of the mode category of the class can be used as useful indicators. The relevant ones for the regression task might be the mean value of the dependent variable, the median, the mode, the standard deviation, skewness and kurtosis. Some database measures include the number of the records, the percent of the original dataset used for training and for testing, the number of missing values and the percent of incomplete records. Also useful information lies on the total number of variables. For the categorical variables of the database, the number of dimensions in homogeneity analysis and the average gain of the first and second Eigen values of homogeneity analysis as well as the average attribute entropy are the corresponding statistics. For the continuous variables, the average mean value, the average 5% trimmed mean, the median, the variance, the standard deviation, the range, the inter-quartile range, skewness, kurtosis and the Huber's M-estimator are some of the useful statistics that can be applied to capture the information on the data set.

The determinant of the correlation matrix is an indicator of the interdependency of the attributes on the data set. The average correlation, as it is captured by Crobach-α reliability coefficient, may be still an important statistic. By applying principal component analysis on the numerical variables of the data set, the first and second largest Eigen values can be observed.

If the data set for a classification task has categorical explanatory variables, then the average information gain and the noise to signal ratio are two useful information measures, while the average Goodman and Kruskal tau and the average chi-square significance value are two statistical indicators. Also in the case of continuous explanatory variables, Wilks' lambda and the canonical correlation of the first discrimination function may be measures for the discriminating power within the data set.

By comparing a numeric with a nominal variable with the student's t-test, two important statistics are produced to indicate the degree of their relation, namely Eta squared and the Significance of the F-test.

Table 1. DCT dataset properties [17]

Nr_Attributes	Nr_num_attributes
Nr_sym_attributes	Nr_examples
Nr_classes	MissingValues_Total
MissingValues_relative	Mean_Absolute_Skew
MStatistic	MeanKurtosis
NumAttrsWithOutliers	MstatDF
MstatChiSq	SDRatio
WiksLambda	Fract
Cancor	BarlettStatistic
Class Entropy	Mutual Information
Joint Entropy	Eqivalent_nr_of_attrs
Entropy Attributes	NoiseSignalRatio

4. PROPOSED METHOD

Here we are considering properties of scenarios. We need to classify learning scenario. We are extracting features of input data or datasets. We are using the concept of meta-learning. Meta-learning relates algorithms to their area of expertise using specific problem characteristics. The idea of meta-learning is to learn about classifiers or learning algorithms, in terms of the kind of data for which they actually perform well. Using dataset characteristics, which are called meta-features; one predicts the performance results of individual learning algorithms. These features are divided into several categories:

- Sample or general features: Here we need to find out the number of classes, the number of attributes, the number of categorical attributes, the number of samples or instances etc.
- Statistical features: Here we require to find canonical discriminant, correlations, skew, kurtosis etc.
- Information theoretic features: Here we need to extract class entropy, signal to noise ratio etc.

We are proposing adaptive methodology. Different thoughts can be considered, e.g. parameters such as the input data, learning methods, learning policies, learning methods combination. Here there can be a single learner or multiple learners. Also we can use simple voting or averaging while combining the performance of the different learners.

5. EXPERIMENTS

5.1 Experimental Descriptions

Here we need to map the dataset's characteristics to the performance of the algorithm. We are capturing the knowledge about the algorithms' from experiments. Here we are calculating the algorithms' accuracy on each dataset. After the experiments, accuracy of each algorithm corresponding to every dataset is saved in the knowledge base for the future use. The Ranking procedure is shown in Figure 1.

Given a new dataset, we use k-NN [7] to find out the most similar dataset in the knowledge base with the new one. K-Nearest Neighbor learning is the most basic instance-based method. The nearest neighbors of an instance are defined in terms of the standard Euclidean distance. Let an arbitrary instance x be described by the feature vector

$<a_1(x), a_2(x), \text{---} a_n(x)>$

Where $a_r(x)$ denotes the value of the r^{th} attribute of instance x. Then the distance between two instances x_i and x_j is defined to be $d(x_i - x_j)$,

$$d(x_i - x_j) = \sqrt{(\sum(a_r(x_{i)} - a_r(x_j))^2}$$

Here r varies from 1 to n in summation. 24 characteristics are used to compare the two dataset's similarities. A distance function that based on the characteristics of the two datasets is used to find the most similar neighbors, whose performance is expected to be similar or relevant to the new dataset. The recommended ranking of the new dataset is built by aggregating the learning algorithms' performance on the similar datasets. The knowledge base KB stores the dataset's characteristics and the learning algorithms' performance corresponding to each dataset.

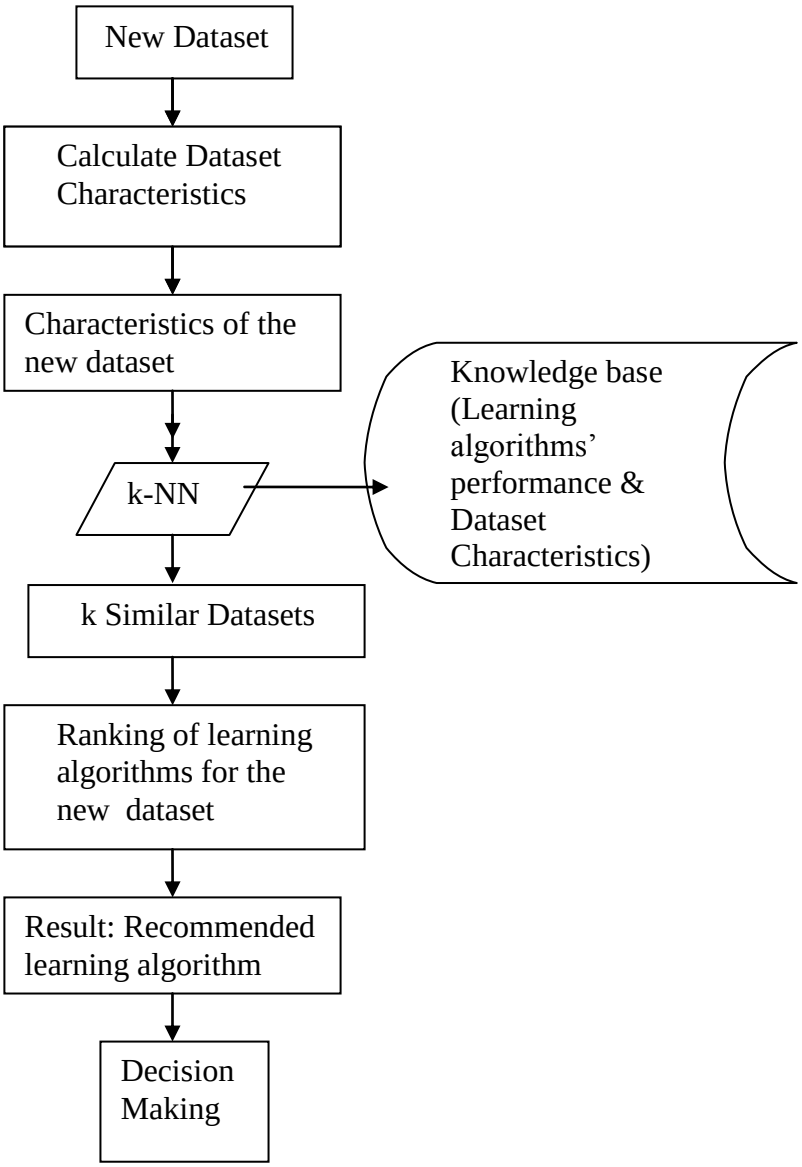

Fig 1: The Ranking of Learning Algorithms

6. RESULTS AND DISCUSSIONS

Here we have used Adult Dataset [13]. The dataset Adult has following features:

- 48842 instances
- 14 attributes (6 continuous, 8 nominal)
- Contains information on adults such as age, gender, ethnicity, martial status, education, native country, etc.
- The instances are classified into either "Salary >50K" or "Salary <= 50K"

Table 2 shows the ranking of eight algorithms used on Adult Dataset from UCI Repository. The table shows highest rank to LogitBoost algorithm, then to J48, oneR and finally lowest rank is given to ZeroR algorithm.

Table 2. Ranking of different algorithms on Adult Dataset

Algorithm	Rank
LogitBoost	1
J48	2
OneR	3
DecisionStump	4
IB1	5
IBK	6
NaiveBayes	7
ZeroR	8

Table 3. Correctly & Incorrectly Classified Instances for Adult Dataset

Algorithm	% of Correct classified instances	% of Incorrect classified instances
LogitBoost	84.68	15.32
ZeroR	76.07	23.93

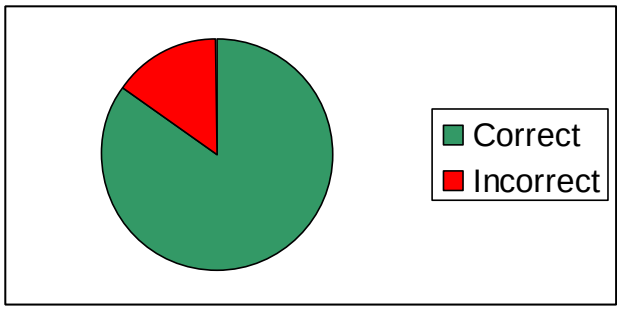

Fig. 2: % Classified instances with top ranked algorithm LogitBoost on Adult Dataset

Figure 2 shows percentage of classified instances with the top ranked algorithm called LogitBoost on Adult Dataset. Here 84.68 % instances are correctly classified.

Figure 3 shows percentage of classified instances with the lowest ranked algorithm called ZeroR on Adult Dataset. Here 76.07 % instances are correctly classified.

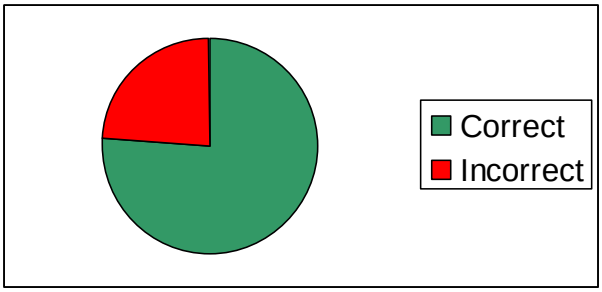

Fig. 3: % Classified instances with lowest ranked algorithm ZeroR on Adult Dataset

7. CONCLUSIONS AND FUTURE WORK

In this paper, we present our preliminary work on using meta-learning method for helping user effectively to select the most appropriate learning algorithms and give the ranking recommendation automatically. It will assist both novice and expert users. Ranking system can reduce the searching space, give him/her the recommendation and guide the user to select the most suited algorithms. Thus the system will assist to learn adaptively using the experiences from the past data. In the future work, we will investigate more on our proposed method and test extensively on other datasets. Meta Learning helps improve results over the basic algorithms. Using Meta Characteristics on the Adult dataset to determine an appropriate algorithm, almost 85% correct classification is achieved for LogitBoost algorithm. So out of eight algorithms LogitBoost algorithm is recommended to the user.

8. ACKNOWLEDGMENTS

Our thanks to the experts who have contributed towards development of the different algorithms and made them available to the users.

9. REFERENCES

[1] Kuncheva, L, Bezdek J., and Duin, R. 2001 Decision Templates for Multiple Classifier Fusion: An Experimental Comparison, Pattern Recognition. 34, (2), pp.299-314, 2001.

[2] Dietterich, T. 2002 Ensemble Methods in Machine Learning 1st Int. Workshop on Multiple Classifier Systems, in Lecture Notes in Computer Science, F. Roli and J. Kittler, Eds. Vol. 1857, pp.1-15, 2002.

[3] Alexmandros, K. and Melanie, H. J. 2001 Model Selection via Meta-Learning: A Comparative Study. International Journal on Artificial Intelligence Tools. Vol. 10, No. 4 (2001).

[4] Joachims, T. 1998 Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Proceedings of the European Conference on Machine Learning, Springer.

[5] Schaffer, C. 1994 Cross-validation, stacking and bi- level stacking: Meta-methods for classification learning, In Cheeseman, P. and Oldford R.W.(eds) Selecting Models from Data: Artificial Intelligence and IV, 51-59.

[6] Wolpert, D. 1996 The lack of a Priori Distinctions between Learning Algorithms, Neural Computation, 8, 1996, 1341-1420.

[7] Mitchell, T. 1997 Machine Learning, McGraw Hill.

[8] Brodley, C. E. J.1995 Recursive automatic bias selection for classifier construction, Machine Learning, 20, 63-94.

[9] Schaffer, C. J. 1993 Selecting a Classification Methods by Cross Validation, Machine Learning, 13, 135-143.

[10] Kalousis, A. and Hilario, M. 2000 Model Selection via Meta-learning: a Comparative study, Proceedings of the 12th International IEEE Conference on Tools with AI, Canada, 214-220.

[11] Koliastasis, D. and Despotis, D. J. 2004 Rules for Comparing Predictive Data Mining Algorithms by Error Rate, OPSEARCH, VOL. 41, No. 3.

[12] Fan, L., Lei M. 2006 Reducing Cognitive Overload by Meta-Learning Assisted Algorithm Selection, Proceedings of the 5th IEEE International Conference on Cognitive Informatics, pp. 120-125, 2006.

[13] Frank, A. and Asuncion, A. 2010. UCI machine learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[14] Michie, D. and Spicgelhater, D. 1994 Machine Learning, Neural and Statistical Classification. Elis Horwood Series in Artificial Intelligence, 1994.

[15] Todorvoski, L. and Blockeel, H. 2002 Ranking with Predictive Clustering Trees, Efficient Multi-Relational Data Mining, 2002.

[16] Alexandros, K. and Melanie, H. J. 2001 Model Selection

[17] Peng, Y., Flach., P., Soarces C. and Brazdil, P., 2002 Improved Dataset Characterization for Meta-learning, Springer LNCS 2534, pp. 141-152, 2002.