Dynamic Algorithm Selection for Data Mining Classification

Suhas Gore, Prof. Nitin Pise

Abstract— Recommending appropriate classification algorithm for given new dataset is very important and useful task but also is full of challenges. According to NO-FREE-LUNCH theorem, there is no best classifier for different classification problems. It is difficult to predict which learning algorithm will work best for what type of data and domain. In this paper, a method of recommending classification algorithms is proposed. Meta learning tries to address the problem of algorithms selection by recommending promising classifiers based on meta-feature. Dynamic Algorithm Selection (DAS) with knowledge base, focus on the problem of algorithm selection, based on data characteristic. Algorithm selection will be better by using DAS in knowledge discovery process. In this paper we discuss the DAS architecture with knowledge base and Recommendation parameter measure. We present the architecture of DAS approach and Analysis of K-similar dataset produced by knowledge base.

Index Terms—Supervised Learning, Dynamic Algorithm Selection, Classification, Data Characteristic, Ensemble Learning,NO-Free-Lunch theorem, Knowledge Base,KNN.

———————————— ◆ ————————————

1 Introduction

OUR main approach is to develop an application that recommends most suitable algorithm based on the accuracy measure. Various data analysis, aspects and real time application suffer from this algorithm selection problem. Machine learning can have strong impact of algorithm selection problem.
Data Mining [21] is a process that extracts patterns from the large datasets. There are major research areas in Data Mining including association mining, clustering, classification, web mining, text mining, etc. Classification is one of the techniques in Data Mining that solves various problems like algorithm selection, model comparison, division of training and testing data, preprocessing. It is 2 step processes

Build classification model using training data. Every object of the data must be pre-classified. The model generated in the preceding step is tested by assigning class labels to data objects in a test dataset.

The test data is different from the training data. Every element of test data is also pre-classified in advance. The accuracy of the classification model is determined by comparing true class labels in the testing set with those assigned by the model.

Meta-learning [3] is to learn about training classifiers themselves, i.e. to predict the accuracy of algorithm on given dataset. This prediction is based on extracting meta-features; these are the feature that describes the dataset itself. These meta-features are used to train a meta-learning model on training data. Afterwards this training strategy is applied on meta-features of new dataset. The result is the classifier with high accuracy and performance. In last two decades, different approaches have been presented in the field of meta-learning. In data mining, concept of similarity and distance is crucial. We consider the problem of defining the distance between two different dataset by comparing statistics computed from the dataset. For distance calculation, we use Euclidean distance and Manhattan distance. Here we used 38 dataset from various domain and 9 algorithms of different class namely: IBK, SMO, Random Forest, Logit Boost, Naïve Bays, J48, Adaboost, PART, and Bagging.

Research Objective:
The objective of this work is to present a comprehensive empirical evaluation of algorithm selection technique in the context of supervised learning from diverse data. The main idea to recommend to the user an algorithm or set of algorithms based on the most similar dataset that are found in the knowledge base. Classification measure and basic architecture of our Dynamic algorithm selection (DAS) system is described in section 2 & 3 respectively.

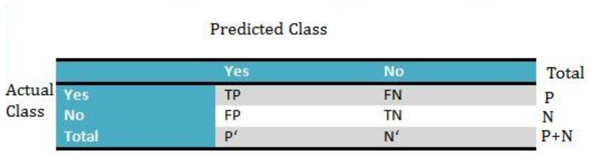
2 Related Work

There are several theoretical and practical reasons why we may refer an adaptive learning system. A survey on ensemble learning gives several advantages of adaptive learning. Following are some papers which describes some novel approach of algorithm selection.
Ensemble based [8] system may be more beneficial than their single classifier counterparts, different algorithms for generating ensemble components and various procedures through which the individual classifier can be combine. CBR [2] approach contains several advantages for algorithm selection, the user gets a recommendation of algorithms suitable or dataset as well as it gives an explanation for recommendation. CBR approach contains CASE that represents the knowledge about the execution of a special algorithm on a specific dataset. There are different classification measures, different approaches of algorithm selection. This is based on working of

- *Suhas Gore is currently pursuing masters degree program in Computer engineering in Pune University, India. E-mail:goresuhass@gmail.com*
- *Prof. Nitin Pise is currently pursuingPhD in Pune University,IndiaEmail: nnpise@yahoo.com*

Meta learning methods: Acquisition Mode and Advisory Mode. A recommendation method for classifier selection is presented in [3]; it is based on data set characteristic with an aim to assist people in algorithm selection process among a large number of candidates for a new classifier problem. In this method, the Data set feature is first extracted, the nearest neighbor of new dataset is then recognized and their applicable classifiers are identified.

Rule base classifier election approach is proposed, based on the prior knowledge of problem characteristic and the Experiments. The main aim is to assist in the algorithm selection of an appropriate classification algorithm without the need of trail-and-error testing on a vast array of available algorithms knowledge Base.



Predicted Class

		Yes	No	Total
Actual Class	Yes	TP	FN	P
	No	FP	TN	N
	Total	P'	N'	P+N

Fig. 1. Confusion Metrix

TABLE 1
CLASSIFICATION MEASURES

Measure	Formula
Accuracy, Recognition Rate	(TP+TN)/(P+N)
Error Rate, Misclassification Rate	(FP+FN)/(P+N)
Sensitivity, True Positive Rate	TP/P
Specificity, True Negative Rate	TN/N
Precision	TP/(TP+FP)

2.1 Classification Measure

Confusion Matrix (As shown in fig 1) is useful tool for analyzing how well your classifier can recognize tuple of different class i.e. it indicates how accurately classification is performed. Here TP: True Positive, FN: False Negative, FP: False Positive, TN: True Negative [21].

Accuracy: It is the degree of closeness of measurements of a quantity to that quantity's actual (true) value.

Precision: Precision is used to retrieved fraction of instances those are relevant

Recall: Recall is used to retrieve fraction of relevant instances that are retrieved. Recall is the fraction of the documents that are relevant to the query that are successfully retrieved.

Skewness: It refers to whether the distribution is symmetrical with respect to its dispersion from the mean.for univariate data.

Kurtosis: It refers to the weight of the tails of a distribution, where a large proportion of the scores are towards the extremes are said to be platykurtic.

Kappa-statistics: It is a statistical measure of inter-rater agreement or inter-annotator agreement for qualitative (categorical) items.

Entropy: Entropy is measure used in Decision tree algorithm for attribute selection. On the basis of selected attribute, classification is done for binary data.

Signal to noise ratio: Signal-to-noise ratio is defined as the power ratio between a signal (meaningful information) and the background noise (unwanted signal):

2.2 Data Characteristics

The data characteristic tool DCT[2] computes various meta data about a given data set. Subsequently, we characterize the relevant data characteristic. The data characteristic can be separated into three different parts:

1. Simple measurement or general data characteristics
2. Measurements of discriminant analysis and other measurement which can only be computed on numerical attribute (DC_numeric).
3. Information theoretical measurement and other measurement which can only be computed on symbolic attribute.(DC_symbolic)

The first part contains measurements which can be calculated for the whole dataset. The other group can only be computed for a subset of attribute in the dataset. The measurement of discriminant analysis and calculated only for numerical attribute whereas the information theoretical measurement are calculated for symbolic ones. All these measurements are calculated by our data characteristic tool (DCT)

2.2.1 Simple Measurement or General Data Characteristic

In algorithm selection problem the DCT tool determines the following general characteristic:

- NrRecords: number of records(n)
- NrAttr : number of attribute(m)
- NrBin: Number of binary attribute(nb)
- Sym: ratio of symbolic attribute (msym/m)
- NrClass: number of class(q)
- defError: default error rate defError=1-Accdef, (Accdef) probability of largest class or default accuracy.
- StdDev: standard deviation of the class distribution(σclass)
- rdefInst : Relative probability of defective records

  rdefInst= ndeftuple/n
  ndeftuple : number of record with missing values
- rmissVal : relative probability of missing values: rmissVal=hmissVal/(n*m)

  hmissVal: number of missing values.

Beside the normal simple measurement system selected relative measurements like the two last measurement. Such a ratio measurement contains more information and is more interpretable.

2.2.2 Discriminant Measurement:
Statistical DCT computes a discriminant analysis leading to the following measurement:

- Fract : Describes the relative importance of the largest eigen value as an indication for the importance of the 1st discriminant function.
- Cancor: Canonical correlation, which is an indicator for the degree of correlation between the most significant discriminant function and the class distribution. There is a strong correlation between the class and the 1st discriminant function if this measurement is close to unity.
- DiscFct : Number of discriminant function.
- Wlambda : Wilks lambda or U-statics, describes the significance of the r discriminant function.

- Standard Deviation Ratio
- Mean Absolute Correlation attribute
- Skewness of attribute
- Kurtosis of attribute

2.2.3 Information Theoretical Measurement:
Besides continuous (numerical) attribute, it is likely that symbolical attributes are used for describing a data space. Therefore, measures are needed to cover these (symbolic) dimensions as well. Again, the goal is primarily to investigate and deploy measures that are useful for the algorithm selection process. All these measurement are well known and based on the entropy of the attribute. Entropy measures have the common property that they deliver information on the information content of attribute.

- Class entropy (ClassEntr)
- Join entropy (JoinEntr)
- Average mutual information (AttrEntr)
- Average Mutual information(MutInf)
- Relevance-Measure(EqNrArrt)
- Signal Noise Ratio(NoiseRatio)

Additionally, we also use a measurement of range of occurance defined by
SpanSym=SymMax-SymMin
AvgSym which is the average number of symbolic values, such measurement is indicators of the complexity and the size of the hypothesis space for the problem.

3 DAS Architeure

3.1. Approaches for algorithm selection

There are different approaches to solve issue of algorithm selection. For different datasets, different classifiers are applied on each of them. Based on some parameter such as accuracy performance of each classifier can be analyzed so as to recommend best classifier for given dataset In deciding which classifier will work best for a given dataset, there are two options. The first is to put all the trust in an expert's opinion based on knowledge and experience. The second is to run through every possible classifier that could work on the dataset, identifying rationally the one which performs best. The latter option, while being the most thorough, would take time and require a significant amount of resources, especially with larger datasets, and as such is impractical. If the expert consistently chooses an ineffective classifier, the most effective classification rules will never be learned, and resources will be wasted. Neither methods, provides an effective solution and as a result it would be extremely helpful to both users and experts, if it were known explicitly which classifier, of the multitude available, is most effective for a particular type of the dataset.

1. Trial and Error Approach: Available Algorithms are applied for each dataset, Gives accurate recommendation of algorithm but too costly, If m algorithm are applied for n dataset then complexity is O(mn).
2. Random Selection: Randomly Selection of Classifier, Cost effective but Less Accuracy.
3. Expert advice: For each new dataset an expert advice is taken which is not always easy to acquire.
4. Case Base Reasoning: Classification is recommended based on previous case, i.e. for a dataset, some algorithms are tested on the basis of applicability test and for each new dataset an algorithm is recommended.
5. Heterogeneous meta decision tree(HDMT): HDMT [9] is induced in one domain and used in any other domain. In general HDMT clearly performs worse than MDT. But they are more generally applicable across different dataset.
6. Feature_vector [1] approach: This is also a part of proposed approach, multiple experiments are performed on this approach, and some data is lost while calculating average value of feature vector. This approach focus on only data of dataset instead of structure of dataset.
7. DAS Approach: This is proposed Framework, which is a part of adaptive learning. Performance of classification algorithms are evaluated on number of dataset with Knowledge base using KNN (Lazy Learners).

3.2. Architecute

Architecture of DAS system is shown in figure 2, as outlined in introduction, the problem of algorithm selection is based on these factors:

- New Dataset
- Historical Dataset
- K-NN
- DCT
- Knowledge Base

Dataset represents the training data as well as testing data, i.e. at the initial stage dataset are supplied to Data characteristic tool (DCT). DCT is a module used for the calculation of Data Characteristic such as accuracy, MIN, MAX, Standard Deviation etc, which will refer as META data. Furthermore knowledge bases contain the experience of known application as well, i.e. knowledge base represents knowledge about the execution of special algorithm on specific dataset. This knowledge includes training time, test time and error rate.

Nearest neighbor classifier are based on learning by analogy, i.e. by comparing a given test tuple with training tuple that are similar to it. The training tuple are described by n attribute. Each tuple represents a point in an n-dimensional space. In this way, all the training tuples are stored in an n-dimensional pattern space. When given an unknown tuple, a K-nearest-neighbor (KNN) [21] classifier searches the pattern space for the k training tuple that are closest to the unknown tuple. These k training tuples are the k “nearest neighbor” of the unknown tuple.

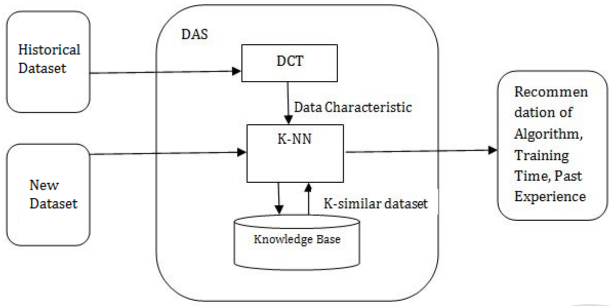


Fig. 2. Architecture OF DAS System

Flow:

A new approach is shown in fig. 2, first we have to calculate dataset characteristic for given dataset. A DCT (Data Characterization Tool) can be use to define dataset characteristic i.e. it computes various Meta data about given dataset. New dataset characteristic are provided to KNN (A lazy learner Algorithm) for analysis and then results are given to knowledge base. Knowledge base determines learning algorithm performance based on dataset characteristic. On the basis of similarity between predefined dataset and new dataset an algorithm is recommended. Ranking based on results from knowledge base are provided so as to recommend a proper and suitable algorithm. Further results are used for prediction and decision making.

The general work flow is that the user specifies his requirements and that the data characteristic for the given dataset is computed by DCT. These two groups of information define the problem description. Each case is defined by this problem description and a solution part, which specifies the applied algorithm as well as the experience gained from applying the algorithm. In our system we compute the most similar application problem description and offer the user also the results of the applied algorithm. Important is also to remark that problem description may be incomplete.

4 Experiments and Analysis

For the purpose of evaluating the performance and effectiveness of our proposed classification algorithm selection recommendation method, verifying whether or not the method is potentially useful in practice, and allowing other researchers to confirm our results, we set up our experimental study as follows.

1. 38 data sets from the UCI repository [20] are used in the experiments. Knowledge base represents the number of instances, the number of attributes by which each instance is described (not including the class label), and the number of classes for each data set. In order to facilitate the calculation of the feature vectors, for data sets containing continuous values, we basically focus on information theoretic measures. As described in knowledge base, Number of attributes, instances, classed, symbolic, numeric and entropy are considered as Meta-feature.

2. 8 different types of classification algorithms are selected to classify data sets. They are probability-based Naive Bayes (NB); tree- based J48 and Random Forest, rule-based PART; lazy learning algorithm IB1; and the support vector algorithm Sequential Minimal Optimization (SMO) . Besides these single learning algorithms, we also employ the ensemble classifier algorithms. Bagging is applied with the three simple classifiers J48, PART and Naive Bayes as the base classifier, respectively. At the same time, Adaboost is also employed with the same three simple classifiers as the base classifier, respectively.

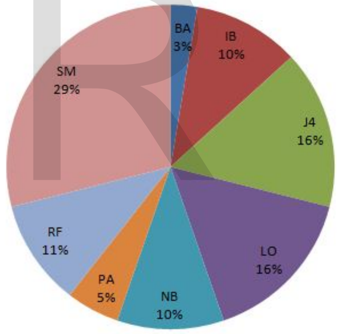
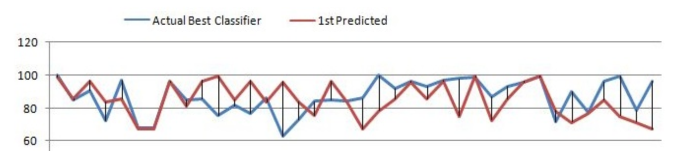


Fig. 3. Chart of % of Classifier over 38 datasets.

As shown in fig 3, it shows chart of percentage of accuracy of classifiers over 38 dataset. SMO is mostly best classifier occurs over 8 algorithms. Logit-Boost and J48 are also as best classifier over most of datasets, Most of these classifiers gives classification accuracy over 70%. Fig 4 represents best classifier for 38 dataset among these 8 classifiers, all accuracy are calculated through WEKA tool using 10-fold cross validation. Accuracy of all best classifiers is above 60%, so it indicates the algorithms chosen are average and over average. It describes the difference between actual best and predicted best, as shown in fig 5, there are very few dataset where difference need to be consider, for e.g. Dataset 38 having difference of about 28.As shown in number of objects on the line and below line indicated predicted accuracy is equal or lesser than actual accuracy. Dataset 5,20,21,26,28,33,35,36,38 are having less prediction accuracy than actual accuracy. So remaining 29 are well recommended. i.e. these gives 76% accurate results.



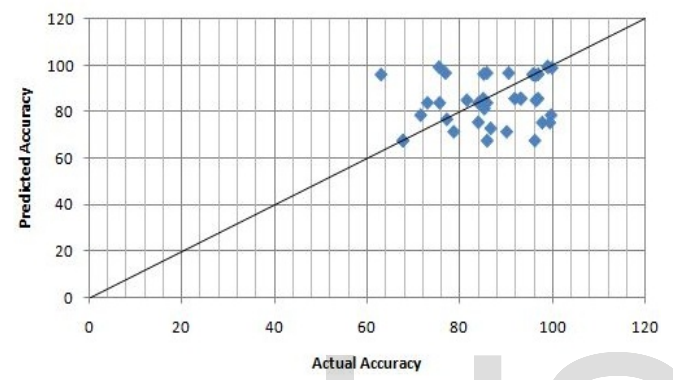


Fig. 5.Actual Best vs Predicted Best

5.CONCLUSION

According NFL, No single algorithm performs better for all type of dataset. There are different approaches to recommend algorithm from which Adaptive Learning (DAS) approach recommends approximate best classifier based on accuracy as performance measure with aim to assist non-experts in selecting algorithm. Three different categories of meta-features, namely simple, statistical, information theoretic were used and comparatively evaluated. After generation of knowledge base, Ranking is provided based on accuracy in algorithm selection task

REFERENCES

[1] Q. Song, G,. Wang, C. Wang, "Automatic Recommendation of Classification Algorithms Based on Dataset Characteristic", Pattern Recognition 45(2012) 2672-2689

[2] Guido Lindner and Rudi Studer, "AST: Support for algorithm selection with a CBR Approach".

[3] Nikita Bhatt, Amit Thakkar, Amit Ganatra, " A survey and current research challenges in meta learning Approaches based on dataset characterisstic", International Journal of soft computing and Engineering, ISSN: 2231-2307, Volume-2, Issue-1, March-2012

[4] Lior Rokach, Department of Information System Engineering, Ben-Gurion University of the Negev, Israel, Computational Statistics and Data Analysis 53 (2009), "Taxonomy for characterizing ensemble methods in classification tasks: A review and annotated bibliography"2001

[5] Geoffrey I. Webb (author for correspondence), School of Computer Science and Software Engineering Monash University, Clayton, Victoria, 3800, Australia," Multistrategy Ensemble Learning: Reducing Error by Combining Ensemble Learning Techniques" To appear in IEEE Transactions on Knowledge and Data Engineering.

[6] Nikolaj Tatti," Distance between dataset based on summary statistic", Journal of machine learning research 8(2007) 131-154

[7] Matthias Reif, M. Goldstein, A Dengel, "Automatic Classifier Selection for Non-expert",Pattern Analysis and application

[8] Robi Polikar, "Ensemble Based system in decision making " IEEE circuit and System Magazine 2006.

[9] B. Zenko,L Todorovski, S. Dzeroski, "Experiments with Heterogeneous Meta Decision Tree", IJS-DP 8638.

[10] Faliang Huang, Guoqing Xie and Ruliang Xiao Faculty of Software, Fujian Normal University, Fuzhou, China," Research on Ensemble Learning", 2009 International Conference on Artificial Intelligence and Computational Intelligence.

[11] Richard O. Duda, Peter E. Hart, David G. Stork,A book on " Pattern Classification" ,II nd edition.

[12] R. King, C. Feng, A. Suterland,"Stat-log: Comparison of classification algorithm on large real-world problem".

[13] L.I. Kuncheva," Combining Pattern Classifiers Methods and Algorithms".New York, NY: Wiley Interscience, 2005.

[14] F. Roli, G. Giacinto, and G. Vernazza, "Methods for designing multiple classifier systems," 2nd Int. Workshop on Multiple Classifier Systems,in Lecture Notes in Computer Science, J. Kittler and F. Roli, Eds., vol. 2096, pp. 78–87, 2001.

[15] Y. Lu, "Knowledge integration in a multiple classifier system," Applied Intelligence, vol. 6, no. 2, pp. 75–86, 1996.

[16] K. Tumer and J. Ghosh, "Analysis of decision boundaries in linearly combined neural classifiers," Pattern Recognition, vol. 29, no. 2, pp. 341–348, 1996.

[17] L. Kuncheva and C. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," Machine Learning, vol. 51, no. 2, pp. 181–207, 2003.

[18] R.O. Duda, P.E. Hart, and D. Stork, "Algorithm Independent Techniques," in Pattern Classification, 2 ed New York: Wiley, pp. 453–516,2001.

[19] U.Fayyad, K. Irani,"Multi-interval Discretization of Continuous-valued Attribute for Classification Learning".Machine Learning1022-1027

[20] C.J. Merz, and P.M. Murphy, Uci repository of machine learning database. Irvine, CA : University of California Department of information and Computer Science.

[21] J. Han, M. Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmannn Publishers, 2011.

[22] R. Valdovinos , J. Sánchez, Proceedings of the Fourth International Conference on Machine Learning and Applications, 2005.

[23] Nikunj Chandrakant Oza, "Online Ensemble Learning a dissertation" submitted in partial satisfaction of the Requirements for the degree of Doctor of Philosophy, UNIVERSITY of CALIFORNIA, BERKELEY.