Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 2 – No.6, August 2015 – www.caeaccess.org

Algorithm Selection based on Landmarking Meta-feature

Ashvini Balte
Post Graduate Student
MIT, Poud Road
Kothrud Pune

Nitin Pise
Associate Professor
MIT, Poud Road
Kothrud Pune

Ranjana Agrawal
Assistant Professor
MIT, Poud Road
Kothrud Pune

ABSTRACT
Knowledge discovery is the data mining task. Number of classification algorithms is present for knowledge discovery task in data mining. Each algorithm is differentiating with another based on their performance. No free lunch theorem [1] states that there no single prediction of algorithm is not possible for all kind of datasets. This implies that performance value of algorithm changes according to dataset characteristics. Non-expert can't understand which will be best classifier for his/her dataset. Meta-learning is one machine learning technique which supports non-expert users for selecting classifier. In meta learning dataset characteristics well know as meta-features. Based on these meta-features the prediction of well suitable classifier is done. In this paper, in the first experiment, the prediction classifier is done by landmarking meta-features with k-NN approach. In the second experiment in addition to first experiment Win/ draw/ loss of corresponding classifiers is calculated using recommendation method and based on that the best classifier is recommended. Here the simple linear regression value of classifiers is taken into consideration. In both the experiments performance measure is the accuracy of classifier.

General Terms
Algorithm Selection, Meta-learning, Data-mining, Performance Measure

Keywords
Landmarking meta-feature, No Free Lunch Theorem, Knowledge Base, Accuracy, k-NN, Recommendation

1. INTRODUCTION
Selecting best classifier on wide verity of situation it is challenging task. In addition there are large number of classifiers are present for data mining task. The no free lunch theorem[1] states that, no single classifier which work best on all available dataset. Studies show that the performances of classifiers are always depends on the characteristics of the dataset or in other term this characteristics of datasets play distinguish role while selecting classifier. Such characteristic is well known as features of the dataset. Meta-Learning is one machine learning technique which predicts the best classifier for given dataset. Meta-learning deals with meta-data of given dataset. Meta-data is data about data; the size of Meta-data is always less than the size of whole dataset. So, to processing meta-data is the time saving process compared to processing whole data set. The meta-learning technique extracts the meta-features from this meta-data. The prediction of the classifier depends on the meta-features. These features are used to train the meta-learning model. After word this trained model is applied on the meta-features of new dataset. The generated result is prediction of one or more classifier based on performance value. In last two decades different approaches present in meta-learning. The major focus of researchers has been on identifying what kinds of meta-features are suitable for characterizing a dataset. Landmarking [2] exploits the performance of very simple algorithms from different classes of learners and uses the accuracy as optimization criteria. Another approach of landmarking consists of using the performance of classifier on small samples of dataset to characterize the problem. The proposed work mainly focuses on landmarking meta-features that are extracted from the UCI repository dataset. The target classifier evaluated based on preprocessing and cross-validation of the classifier [3]. Meta-feature with their generated accuracy of classifier creates knowledge base. The regression technique is used to predicate the accuracy of classifier. This is training phase. In testing phase the landmarking techniques extract meta-features from the new data set and predicated model give the predictive accuracy of classifier. The main aim is to improve the accuracy of recommendation of classifier.

The remainder of this section is organized as follows; Section II contains the related work in algorithm selection problem. Section III is explains the basic terminology of techniques which will use in proposed work. Section IV explains the architectural work. Section V describes the experimental work and result analysis. Finally Section VI concludes the paper with future scope.

2. RELATED WORK
The selection of a meta-learning classifier directly depends on the problem and the task to be solved. Commonly, conventional classification algorithms are very favorable in meta-learning algorithm selection and can integrate meta-decision trees, SVM (Support vector machines), neural networks or any other classification algorithm, with the k-Nearest Neighbors existence another popular alternative [3]. Applying regression algorithm is less remarkable, even smaller is the number of feasible algorithms to learn rankings.

According to [3], meta-knowledge is derived in the course of conducting a learning system. A very conventional configuration of meta-knowledge is the performance of algorithms in absolute problem areas, which is to be chained with characteristics of the activity. Many possibilities for characterizing a problem domain exist. The most straightforward form of meta-knowledge extracted from the data involves statistical or information-theoretic features. For classification problems, [3] mention the number of classes and features, ratio of examples to features, degree of correlation between features and target concept and average class entropy. Rather than individual instance are labeled, in [4] define properties of dataset that specific to multi-instance setting in addition to that it extends concept of landmarkers to the multi-instance setting. In [5] two experiments are conducted, in 1st experiment characteristics of small sample of dataset is taken into consideration and try to predict the classifier performance best on the entire stream.

2nd experiment builds on meta-classifier that predicted based on measurable data characters.

The [6] intend measures for the difficulty of a classification problem that can be used as an input for meta-learning. They involve class variation, indicating the probability that, by means of a distance measure, any two neighboring data records have different class value and example cohesiveness, measuring the cohesion of the example separation in the training dataset. Optionally to looking at the dataset only, information of individual algorithms and how they solved the problem can be accounted, for example their predicted confidence interval. This can be obtained by using model that is fast to build and train and determining its properties. Another approach is landmarking as proposed in [7], using the performance of simple algorithms to clarify a problem and associated this information with the performance of more advanced learning algorithms. A list of landmarking algorithms can be determined in [8]. Landmarking algorithms can also be run on only a small sample of the data available, reducing the training time required.

Empirical evaluation of different categories of meta-features in the context of their suitability for predicting classification accuracies of a number of standard classifiers can be found in [9]. The authors distinguish 5 such categories of features i.e. simple, statistical, information-theoretic, landmarking and model-based, which corresponds to the general categorization evident from the literature. New approach DecT [10] for dataset characterization in meta-learning is proposed. The DecT is compared with DCT and landmarking both, but results are not better than DecT.

Research in the area of meta-learning is continuing in several directions. One area is the identification of meta-features. The vast majority of publications investigates extracting features from the dataset, mostly in the form of statistical or information theoretic measures. Landmarking is a different approach using simple base learning algorithms and their performance to describe the dataset at hand [11]. However, [3] argue that characteristics of learning algorithms and gaining a better understanding of their behavior would be a valuable research avenue with very few publications, for example [12], that exist in this area to date.

3. BASIC TERMINOLOGY

3.1 Landmarking Meta-feature

The Landamrking meta-features are extracted using C5.0 algorithm explained in [2]. C5.0 is decision tree classifier. Based on the information gain value the node of the tree is taken as the feature. The accurate meta-feature extraction is done using RapiMiner Tool[13].

3.1.1 Randomly Choose Node Learner

This result is based on randomly choose attribute. This node is used to test split the training set and classifies given examples.

3.1.2 Decision Node Learner or Best Node Learner

Based on information gain ratio, it shows how informative is an attribute with respect to classification task using its entropy. It chooses attribute which have highest information gain.

3.1.3 Average Node Learner

Calculates the average accuracy of single node decision tree where each node relates to one value.

3.1.4 One Nearest Node Learner

This landmark learner classifies how near the test point that belongs to same class.

3.1.5 Worst Node Learner

In this the information gain criteria uses the attribute which represents lowest selected value.

3.1.6 Naïve Bayes Learner

Training set uses bayes theorem to classify test cases.

3.2 Neighbor Recognition

The neighbor recognition is done by using K-NN approach [14]. In this approach the distance of new dataset with respect to old dataset is calculate. The k value indicates the how much nearest dataset is considered. In this paper the k=3 is taken, that indicate three nearest dataset are considered among 38 dataset.

3.3 Manhattan Distance

The distance between two points in a grid based on a strictly horizontal and vertical path (that is, along the grid lines), as opposed to the diagonal distance. The manhattans distance [15] is the simple sum of the horizontal and vertical components.

$$d_{i,j} = \sum_{h=1}^{p} | x_{i,h} - x_{j,h} |;$$

$$d(i,j) = | x_{i,1} - x_{j,1} | + | x_{i,2} - x_{j,3} | + \ldots\ldots\ldots + | x_{i,p} - x_{j,p} |$$

Where , d is distance, x=meta-features, i for new dataset and j for old dataset.

4. SYSTEM OUTLINE

The main architecture diagram for proposed system is as depicted in Figure 1. This system is for predict the best classifier for given data set. For such prediction, initial the system uses 38 dataset from standard UCI repository [16] known as training dataset. The landmarking meta-features extracted from the meta-data of these dataset. Also the Naive Bayes [17], IBK [18] , J48 [19], AdaBoost , LogitBoost [20], PART [21], RandomForest [22], Bagging [23] and SMO [24], classifier are apply on given dataset and accuracy of each classifier is calculated. This is knowledge base of system. These generated results are stored in backend database. This storage table contains the column namely dataset name, meta-features value and accuracy of each classifier. Also the best classifier for the each dataset is highlighted. After that Regression value [25] is calculated by applying simple linear regression formula i.e y=a+bx. A is sum of all meta-features, b is the variable which change in proportion with the accuracy of classifier. X is the accuracy of each classifier. The generated result and previous knowledge is input to the prediction model.

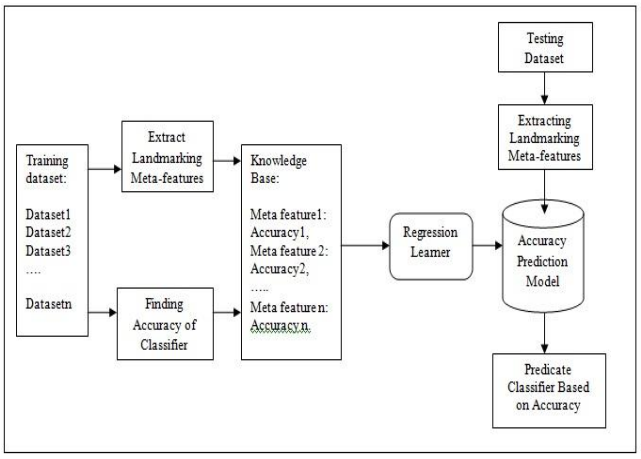

Figure 1: System Outline

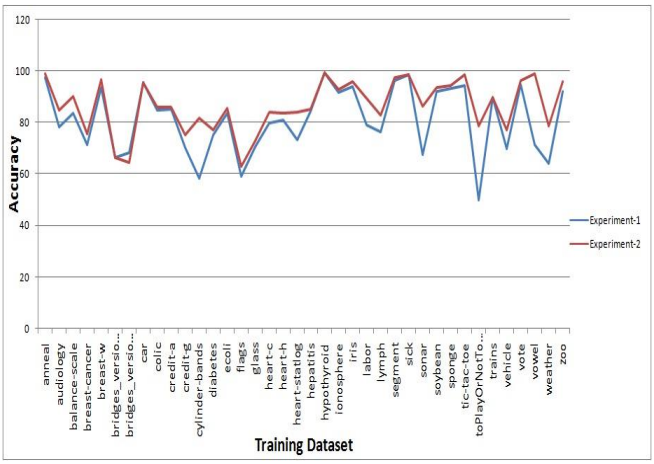

Figure 2: Difference: Experiment-1 vs. Experiment-2

4.1 Prediction Model

This prediction model implement for classifier prediction. This prediction is done by experiments. Two different experiment is conducted this are the combination of one or two algorithm or method. The details are explained in following.

4.1.1 Experiment-1

This technique is based on trial and error approach. Where, the neighbor selection [17] algorithm is used with traditional distance formula. The distance of new dataset from the old dataset is calculated by, distance = ∑(new meta-feature)-∑(old meta-feature). The generated distance is +ve distance or –ve distance. The lowest distance is considered to be nearest dataset.

4.1.2 Experiment-2

The Neighbor selection and recommendation [14] these algorithms are used in prediction model. The distance of new meta-features is calculated with respect to knowledge base. For that purpose manhattans distance formula is used in neighbor recognition. After this the nearest dataset is found out. Three classifiers of respective nearest dataset are found out by highest regression value. The receptive classifiers win, draw and loss is calculated [14] and in accuracy prediction model the win classifier is recommended as best classifier.

5. EXPERIMENTAL AND ANALYSIS

For performance checking same 38 training dataset from UCI repository [16] is taken as test dataset. The experiment-1 and experiment-2 predicted classifier for this dataset. By finding out accuracy of predicted classifier for respective dataset the generated graph is given in figure 2. This graph clearly gives the idea of the experiment-2 which gives the better recommendation of classifier compared to experiment-1. The experiment-2 recommended classifier is same as it in knowledge base.

The interesting fact is for many cases it having same accuracy or small difference of accuracy but, for toplayornottoplay dataset it gives large variation of accuracy value.

To find out how many times a particular classifier is recommended. For that purpose the count of each classifier is taken into consideration.

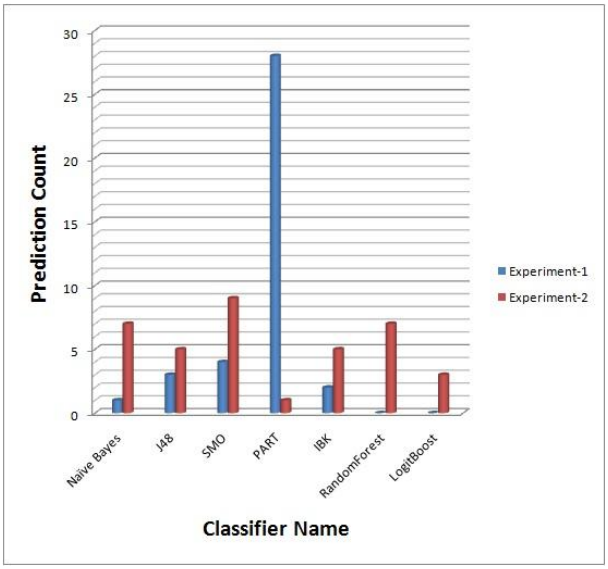

Figure 3: Count of Classifier: Experiment-1 vs. Experiment-2

The generated result graph is shown in figure 3. By observation it is find out RndomForest and LogitBoost are never recommended by experiment-1. Also the PART classifier count is 28. This implies that experiment-1 recommended PART for 28 datasets. This have a two main reason, 1st one the PART classifiers dataset is have smallest +ve distance and 2nd is respective dataset of PART classifier is at largest -ve distance. The sign are not change in experiment-1. These values are taken as it is.

The dataset tae, primarytumor, autos, arrhythmia and mushroom is used for testing, this are different dataset than training dataset. Using experiment-1 and experiment-2 classifier is predicted for this dataset. The accuracy of predicted classifier is mentioned in form of graph shown in figure 4. The difference of accuracy of experiment-1 and

experiment-2 is given in figure 4. By observing it found out experiment-2 gives better recommendation than experiment-1.

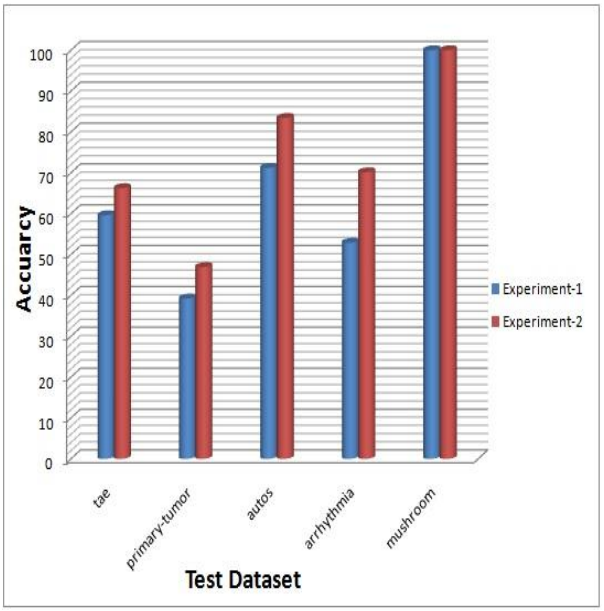

Figure 4: Accuracy of Classifier for Testing Dataset: Experiment-1 vs. Experiment-2

6. CONCLUSION AND FUTURESCOPE

The experiment-2 is depending on adaptive learning approach and the Experiment-1 depends on trial and error approach. The experiment-2 gives the better recommendation of classifier compared to experiment-1. The recommendation in experiment-2 is depend on regression value as well as accuracy of classifier. Also in few cases the experiment-1 also gives equal performance to experiment-2. The landmarking meta-features approach gives better prediction compared to other approach.

In future instead of recommending single classifier the more than one classifier or ranking of classifier is possible. The training time increase with respect to size of dataset. Also some classifier takes more time while training dataset. This can be reduced by taking small dataset and limited classifier. The performance of landmarking meta-features can be compared with other meta-feature.

7. REFERENCES

[1] Igel, C. and Toussaint, M. 2005. A No-Free-Lunch Theorem for Non-uniform Distributions of Target Functions. Journal of Mathematical Modelling and Algorithms, 3(4), 313-322.

[2] Fürnkranz, J. and Petrak, J. 2001 An Evaluation of Landmarking Variants. In: Working Notes of the ECML/PKDD 2000 Workshop on Integrating Aspects of Data Mining, Decision Support and Meta-Learning IDDM-2001), Freiburg, Germany, 57-68.

[3] Brazdil, P., Carrier, C. G., Soares, C. and Vilalta, R. 2008 . Metalearning: Applications to Data Mining. Springer Science & Business Media.

[4] Vanwinckelen, G. and Blockeel, H. 2014 A Meta-learning System for Multi-instance Classification. In Proceedings of the ECML-14 Workshop on Learning over Multiple Contexts, 1-14.

[5] Van Rijn, J. N., Holmes, G., Pfahringer, B. and Vanschoren, J. 2014 Algorithm Selection on Data Streams. In Discovery Science, Springer International Publishing, 325-336.

[6] Vilalta, R. and Drissi, Y. 2002 A Characterization of Difficult Problems in Classification. In: Proceedings of the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases, Helsinki, Finland (2002).

[7] Pfahringer, B., Bensusan, H. and Giraud-Carrier, C. 2000 Tell me who Can Learn You and I Can Tell You Who You Are: Landmarking various learning algorithms. In Proceedings of the 17th international conference on machine learning , 743-750.

[8] Vanschoren, J. 2010 Understanding Machine Learning Performance With Experiment Databases. Ph.D. thesis, Arenberg Doctoral School of Science, Engineering & Technology, Katholieke Universiteit Leuven.

[9] Reif, M., Shafait, F., Goldstein, M., Breuel, T., & Dengel, A. 2014 Automatic Classifier Selection for Non-experts. Pattern Analysis and Applications, 17(1), 83-96.

[10] Peng, Y., Flach, P., Soares, C. and Brazdil, P. 2002. Improved Dataset Characterization for Meta-learning. S. Lange, K. Satoh, C. Smith (eds.) Discovery Science, Lecture Notes in Computer Science, 193-208.

[11] Balte, A., Pise, N. and Kulkarni, P. 2014 Meta-Learning with Landmarking: A Survey. International Journal of Computer Applications, 105(8), 47-51.

[12] Vanschoren, J. and Blockeel, H. 2006 Towards Understanding Learning Behavior. In Proceedings of the Annual Machine Learning Conference of Belgium and The Netherlands, Benelearn, 89-96.

[13] Abdelmessih, S. D., Shafait, F., Reif, M. and Goldstein M. 2010. Landmarking for Meta-learning Using RapidMiner. RapidMiner Community Meeting and Conference.

[14] Song, Q., Wang, G. and Wang, C. 2012. Automatic Recommendation of Classification Algorithms Based on Data set Characteristics. Pattern recognition, 45(7), 2672-2689.

[15] Black E.P. 2006 Manhattan distance. in Dictionary of Algorithms and Data Structures [online], Vreda Pieterse and Paul E. Black, eds. 31 May 2006. (accessed TODAY) Available from: http://www.nist.gov/dads/HTML/manhattanDistance.htm

[16] Asuncion, A. and Newman, D. 2007 UCI machine learning repository. http://www.ics.uci.edu/~mlearn/MLRepository.html. University of California, Irvine, School of Information and Computer Sciences.

[17] Kohavi, R., Becker, B. and Sommerfield, D. 1997 Improving Simple Bayes.

[18] Aha, D. and Kibler D. 1991 Instance-based Learning Algorithms. Machine Learning. 6, 37-66.

[19] Quinlan, J. R. 1993 C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers.

[20] Friedman, J., Hastie T. and Tibshirani R. 1998 Additive Logistic Regression: a Statistical View of Boosting. Stanford University.

[21] Frank, E. and Witten, I. H. 1998 Generating Accurate Rule Sets Without Global Optimization. In: Fifteenth International Conference on Machine Learning, 144-151.

[22] Breiman, L. 2001 Random Forests. Machine Learning. 45(1), 5-32.

[23] Breiman, L. 1996 Bagging predictors. Machine Learning. 24(2), 123-140.

[24] Platt , J. 1998 Fast Training of Support Vector Machines using Sequential Minimal Optimization. In B. Schoelkopf and C. Burges and A. Smola, editors, Advances in Kernel Methods - Support Vector Learning, 3.

[25] Kenney, J. F. and Keeping, E. S. 1962 Linear Regression and Correlation. Ch. 15 in Mathematics of Statistics, Pt. 1, 3rd ed. Princeton, NJ: Van Nostrand, 252-285.