Algorithm Selection for Classification Problems

Nitin Pise, Parag Kulkarni
Department of Computer Engineering and Information Technology
College of Engineering
Pune, India

***Abstract*—A number of algorithms are available in the areas of data mining, machine learning and pattern recognition for solving the same kind of problem. But there is a little guidance for suggesting algorithm to use which gives best results for the problem at hand. This paper shows an approach for solving this problem using meta-learning. The paper uses three types of data characteristics. Simple, information theoretic, and statistical data characteristics are used. Results are generated using nine different algorithms on thirty eight benchmark datasets from UCI repository. The proposed approach uses K-nearest neighbor algorithm for suggesting the suitable algorithm. Classifier accuracy is taken as a basis for recommending the algorithm. By using meta-learning, accurate method can be recommended as per the given data, and cognitive overload for applying each method, comparing with other methods and then selecting the suitable method for use can be reduced. Thus it helps in adaptive learning methods. The experimentation shows that predicted accuracies are matching with the actual accuracies for more than 90 % of the benchmark datasets used. Thus it is concluded that the number of attributes, the number of instances, the number of classes, maximum probability of class and class entropy are playing a major role in classifier accuracy and algorithm selection for thirty eight datasets used for experimentation.**

Keywords*—machine learning; data mining; meta-learning; algorithm selection; classifiers

I. Introduction

The model and algorithm selection has recently gained a lot of interest in areas of machine learning, data mining and pattern recognition. Selection of algorithm from a set of algorithms and tuning of parameters are increasingly relevant now a day. Researchers and practitioners working in the technology and science are having an ample choice of algorithms in the areas of data mining and machine learning. But there is a little guidance for suggesting algorithm to use which gives best results for the problem at hand. There are plenty of algorithms and systems for solving the same kind of problem in the literature. Each of these is behaving differently with respect to the performance for solving a particular problem. Recent research is focusing on creation of algorithm portfolios. The algorithm portfolios contain a set of different algorithms [1].

The algorithm selection problem was mentioned first in literature in 1976 by Rice [2]. Then several machine learning systems have been created so far. The systems performing algorithm selection justify their choice of a machine learning methodology. It can be single learner or a combination of different learners. The systems do this by comparing performance with one of the algorithms selected from. But they do not critically assess the real performance.

Meta-learning [3] assists to solve significant problems in the areas of data mining and machine learning. It helps in classification and regression. The end users are finding difficulty in choosing the models which can solve the problem and combining them if more than one model is required. The users don't possess the required expertise while choosing a suitable model. Also there are many options available for trials. This problem is solved by creating the meta-learning framework. It can guide the users by mapping a particular problem or a task at hand to a suitable algorithm. Sometimes instead of a single algorithm, combination of algorithms can be suggested.

As there are a number of machine learning or classification algorithms available, there is a challenge to the data mining users to find and use a suitable algorithm for their task or problem at hand. In general, it is unclear which particular classifier should be used for a particular data to achieve good results. Some classifiers such as Support Vector Machines (SVM) [4] provide highly accurate results for a set of datasets. However, as per the no-free lunch theorem [5] there is no learning algorithm available which provides better results for all problem instances. Hence, no general recommendation is made for arbitrary data. The characteristics of the data or meta-features of datasets always affect the actual performance of a classifier.

The paper doesn't recommend a set of classifiers that are the best in general for algorithm selection and be used in all cases. But it guides novice users or researchers having little experience in algorithm selection.

II. Related Work

This work discusses an example of algorithm selection or recommendation problem. Performance of a number of algorithms is given for training problems. The best candidate for a particular problem instance is required to be chosen. The performance is predicted on unseen problems. An algorithm portfolio [6] consists of a number of algorithms. A subset is selected and applied sequentially or in parallel to a problem instance, according to some schedule.

Many systems are using algorithm portfolios in some form which are developed over the years. Smith-Miles [7] presents a cross-disciplinary survey of many different approaches. [7] Describes the applications of algorithm selection in sorting, time series forecasting, regression, optimization and constraint satisfaction apart from classification.

978-1-4673-8460-5/16/$31.00 ©2016 IEEE

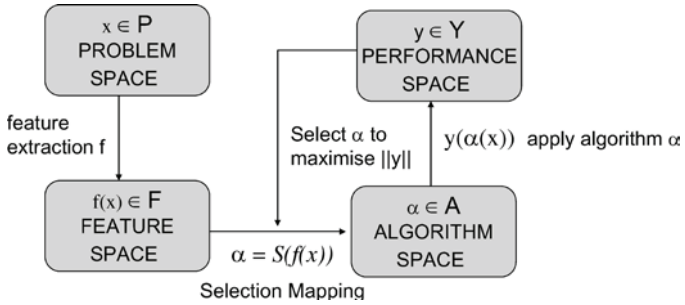

Fig. 1. Rice's [1976] algorithm selection problem model taken from [2]

Rice [2] first modeled meta-learning abstractly. Fig. 1 shows the algorithm selection model by Rice [2]. It consists of different spaces such as feature space, problem space, algorithm space, and performance space. The goal in the above approach is to choose the selection mapping S returning the best algorithm α from algorithm space A. For a particular problem, a mapping function selects the algorithm maximizing the performance. Meta-features affect the mapping function. Given a problem there exists number of algorithms to solve it.

Predicting the performance of classifiers or machine learning algorithms, ranking of learning algorithms, adaptively selecting a suitable learner for a given task are now research topics in data mining and machine learning field [8, 9, 10]. The METAL project [11] focused on finding new and significant data characteristics. It used meta-learning for finding a ranking among number of classifiers. The project created the Data Mining Advisor (DMA). DMA is a web- based solution which automatically selects classification algorithms. The DMA proposed two ranking mechanisms, one uses the ratio of accuracy and training time [8] and the other ranking technique based on the concept of data envelopment analysis [12,13].

Literature uses different meta-features which are classified into five groups. Meta-features or data characteristics are grouped as simple, information-theoretic, statistical, model-based, and land marking meta-features [14]. Simple meta-features are obtained directly from the data. The number of attributes, the number of samples and the number of classes are some examples of meta-features. Statistical features depict statistical properties of the data, e.g. kurtosis [15].

Information-theoretic data characteristics use entropy measures [16]. Landmarking and model-based features are proposed more recently. Simple and fast computable classifiers are used in the landmarking approach.

A method which uses evaluation and selection of learning algorithm is introduced in [29]. It involves comparisons of decision tree algorithm C4.5, neural network, and SVM [4] with some other rule-based and statistical classifiers. Different required techniques for building meta-learning systems are discussed in [32]. It concludes by showing that meta-learning plays as assistant tool for selection of models. [33] Evaluates scenarios observing data characteristics impacting most on the classifier's performance.

After reviewing the papers in meta-learning literature, the following limitations are found: the most of papers except one paper have not calculated all five meta-features. These papers have carried only limited feature selection. SVM classifier was used in very few papers. Parameter tuning of the target classifiers was not done mostly although this can vary classification results notably. Recent experiments [34] suggest that parameter optimization may affect classifier accuracy.

III. Proposed Work

The proposed work is based on meta-learning [3]. Algorithm selection problem is to find a learning model fitting the data. In meta-learning, experience of previous machine learning (ML) experiments is used to learn to improve automatic learning. Here meta-learning is used to gain insight into learning behavior which helps to improve existing algorithms. The most promising learning techniques are chosen after analysis of new learning tasks.

A. System architecture

The system architecture used in the current approach is described in [27]. A dataset consists of both the training data as well as testing data. Data characteristics tool (DCT) is a module which computes data characteristics such as number of attributes, number of symbolic attributes, number of classes, mean skewness, mean kurtosis etc, These are referred as meta-features. Further knowledgebase represents knowledge about the performance of many different algorithms on specific dataset. This knowledge may involve training time, test time, error rate and some more parameters. But only classifier accuracy is used in the proposed approach.

Although there are other possible choices, the most instance- based learners use Euclidean distance. So the proposed approach uses Euclidean distance which can be calculated as follows:

$$d(p,q) = d(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}. \quad (1)$$

Relationship between a historical dataset and new dataset is used to select an algorithm for a given task or problem at hand.

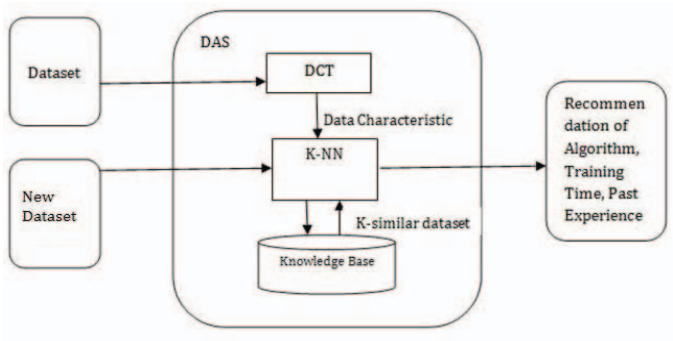

Fig. 2. System architecture for algorithm recommendation

B. Methodology

1) *Dataset Collection*
2) *Meta-features are extracted for training from DCT*

978-1-4673-8460-5/16/$31.00 ©2016 IEEE

3) Learning algorithm with performance measure such as accuracy

4) Generate Knowledge Base

5) Meta-features extraction of a new dataset using DCT.

6) Find K-similar datasets from knowledge base.

7) Ranking of algorithms

8) Recommendation of algorithm for prediction.

C. Classifiers Used

Classification algorithms or classifiers [19] [20] are divided into several categories such as function-based classifiers (e.g., support vector machine [4] and neural network), tree- based classifiers (e.g., J48 [21] and random forest [22]), distance-based classifiers (e.g., k-nearest neighbor and k- star algorithm), and Bayesian classifiers. All existing classifiers have pros and cons. For example, SVM [4] is a powerful classifier for binary class problem; it often shows poor classification performance on class imbalanced problem.

Ensembles [35] are chosen because they give better accuracy over single classifiers. Experts, which are a pool of different classifiers, can offer matching information about the patterns to be classified. This increases the value of the overall classification process [30].The experiments are done on ten real world datasets from UCI repository [24] using ensembles such as AdaBoost [35], Bagging [36], Stacking [37] and LogitBoost [38]. AdaBoost and LogitBoost are two variants of boosting algorithm. LogitBoost is motivated by statistical view [38]. Fig. 3 shows percentage accuracies of these ensembles. It is clear that Stacking does not give more accuracy as compared to the other three ensembles. So AdaBoost, Bagging and LogitBoost are included in the further experimentation for algorithm selection.

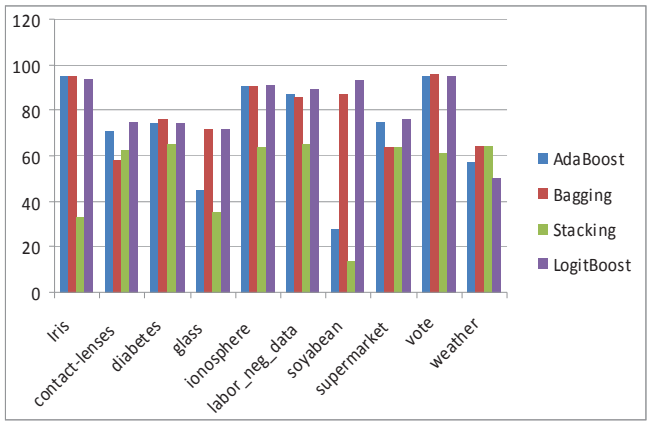

Fig. 3. Percentage accuracies of different ensemble methods

Bagging is used with three base classifiers Naïve Bayes, J48, and PART. The above same three base classifiers are also used in another ensemble Adaboost.

TABLE I. CLASSIFIERS FROM VARIOUS CATEGORIES USED FOR EXPERIMENTAL WORK

Sr.	Classifier	Category	Abbreviation used
1	Naïve Bays	Bays	NB
2	IBK	Lazy	IB
3	J48	Tree	J4
4	Adaboost	Meta	AD
5	Logitboost	Meta	LO
6	PART	Rules	PA
7	Random Forest	Tree	RF
8	Bagging	Meta	BA
9	SMO	Functions	SM

D. Data Characteristics

Data characteristics or meta-features determine the performance of classifier when it works on that dataset. A complete description of the characteristics is available in [23]. Table II shows data characteristics used for the experimentation. Sr. 1 to 11 characteristics are representing simple meta-features, 12-13 represent statistical meta-features and 14-15 represent information theoretic meta-features.

TABLE II. DATA CHARACTERISTICS USED IN EXPERIMENTATION

Sr.	Dataset Characteristics
1	Number of attributes
2	Number of instances
3	Number of Classes
4	Number of Symbolic attributes
5	Number of Numeric
6	Number of missing values
7	Number of distinct values
8	% missing values
9	Dimension
10	Ratio of Symbolic attributes
11	Ratio of Numeric attributes
12	Kurtosis
13	Skewness
14	Maximum Probability
15	Entropy

a) Statistical meta-features

Kurtosis and skewness are statistical meta-features, which are explained below:

Kurtosis measures peakedness or how much asymmetric the distribution is. Kurtosis is calculated as:

978-1-4673-8460-5/16/$31.00 ©2016 IEEE

$$\beta_2 = \frac{\sum(Y-\mu)^4}{n\sigma^4} \quad (2)$$

β_2 is called as Pearson's kurtosis.

Skewness is defined with respect to the third moment about the mean.

$$\gamma_1 = \frac{\sum(Y-\mu)^3}{n\sigma^3} \quad (3)$$

Skewness value is negative when the shape of the distribution appears skewed to the right.

b) Information-theoretic features

Entropy denotes the amount of information in bits given by a particular signal state. The amount of information contained in the complete distribution of signal states (for example, a database column of values) can be calculated as:

$$\text{Entropy(S)} = -p_+ \log p^+ - p_- \log p^- \quad (4)$$

Where p is probability of signal.

E. Algorithm

The best algorithm is found by using the algorithm. Algorithm used is as below:

1. ii=1
2. For each D ∈ DC do
3. DistanceTable [ii] = the distance between d and D i.e. |d-D|
4. ii= ii+1
5. Arrange Distance Table in ascending order
6. Neighbors = top K datasets of distance table
7. jj=0
8. For each jj < K do
9. Alg[jj] = D_j's Best Algorithm

IV. Experiments and Results

Thirty eight benchmark datasets are obtained from the University of California Irvine Machine Learning Repository [24] and used in experiments. The system is developed in java language. Nine classifiers from the different categories are used. Table I shows nine classifiers. In this work, the parameters for each classifier are set as default. Cross-validation [26] has been used in this work for obtaining accuracy. The current approach uses 10 fold cross validation, where data is decomposed randomly into 10 parts. The algorithm is executed on the rest nine parts. Its classification accuracy is calculated on the holdout set. Finally, average of 10 accuracy values gives an overall accuracy.

Many experiments are performed on different combinations of data characteristics which are listed as below:

1) KNN where K=1 with different attribute combinations
2) KNN where K=3 with different attribute combinations
3) KNN where K=3 with different attribute combinations and normalized values

Here min-max normalization has been used as some of the meta-features e.g. number of instances are affecting more as it varies from a few to a few thousands.

Here comparison is done between actual and predicted values of accuracy for features: number of attributes, number of instances, number of class labels. As shown in fig. 4, data sets 23 and 25 show more difference between actual and predicted accuracy and all others have almost best prediction.

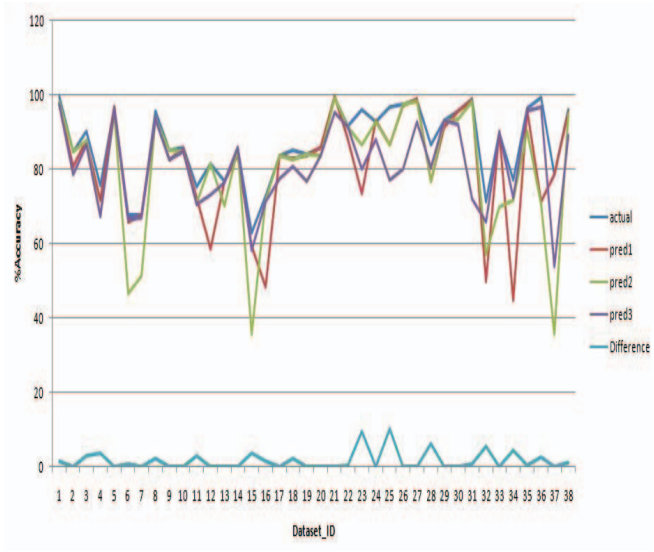

Fig. 4. Graph of dataset Vs actual and 3 predicted classifiers accuracy

The graph in fig. 4 shows difference between actual and best of predicted 3 classifiers.

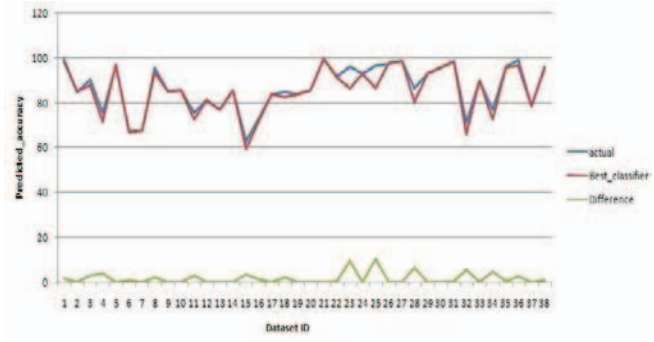

Fig. 5. Graph of dataset Vs actual and best predicted classifiers accuracy

The difference between actual best and predicted best classifiers is shown in fig. 5. Difference need to be considered for very few datasets.

978-1-4673-8460-5/16/$31.00 ©2016 IEEE

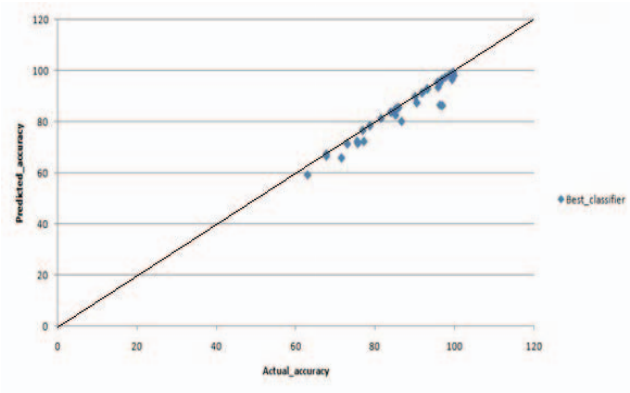

Fig. 6. Graph of actual Vs 3 predicted classifiers accuracy

Fig. 6 shows that predicted accuracy is equal or greater than actual accuracy. Many of the objects are on or close to line so the recommendation is almost accurate.

A. Results of KNN approach

Table IV shows the different data characteristics and how they are referred in the experimental work.

TABLE IV. DATABASE SCHEMA OF DATASET CHARACTERISTICS

Data Characteristics	**Referred as**	**Database attribute**
Number of attributes	3	Norm_attr
Number of instances	4	Norm_inst
Number of classes	5	Norm_class
Number of symbolic attributes	6	Norm_sym
Number of numeric attributes	7	Norm_num
Number of missing values	8	Norm_missing
Number of distinct values	9	Norm_dist
% missing values	10	%missing
Dimension	11	Norm_dimension
Ratio of symbolic attributes	12	Norm_sym%
Ratio of numeric attributes	13	Norm_num%
Kurtosis	14	Kurtosis
Skewness	15	skewness
Maximum Probability	16	Maxprob
Entropy	17	entropy

TABLE V. DIFFERENT COMBINATIONS OF DATA CHARACTERISTICS

Sr	**Attribute_wise_results**	**Best_recommended**	**Avg_Difference**
1	3	17	2.5
2	3_4	19	2.18
3	3-4-5	20	2.87
4	3-4-5-10	20	2.87
5	3-4-5-16	20	2.87
6	3-4-5-17	20	2.87
7	4	22	1.827
8	5	12	2.33
9	16	18	1.88
10	KNN-3-4	16	1.73
11	KNN-3-4-5	16	2.13
12	KNN-10-17	12	1.59
13	KNN-16_17	16	2.21
14	KNN-16	17	2.21
15	NV-3-4-5	16	1.659
16	NV-3-4-5-16-17	18	2.39
17	NV-16-17	14	2.53
18	NV-3-4-5-6-8-9-16-17	16	1.91
19	NV-3-4-5-7-8-9-16-17	18	2.16
20	NV-3-4-16-17	16	2
21	NV-3_4-16	18	2.26
22	NV-3_4-17	13	3.28
23	NV-3-4-5-8	17	1.66
24	NV-6-16-17	14	4.1
25	NV-7-16-17	13	2.7

978-1-4673-8460-5/16/$31.00 ©2016 IEEE

In table V, first 9 entries of table are using K=1 without normalized values. First best 3 classifiers of the same dataset that are found similar for a new dataset are used for evaluation. 10 to 14 entries in table V are for K=3 without normalized values and 15 onwards entries are for K=3 with normalized values.

Fig. 7 shows different combinations of dataset characteristics and predicted 1st best classifier's average difference over all thirty eight datasets.

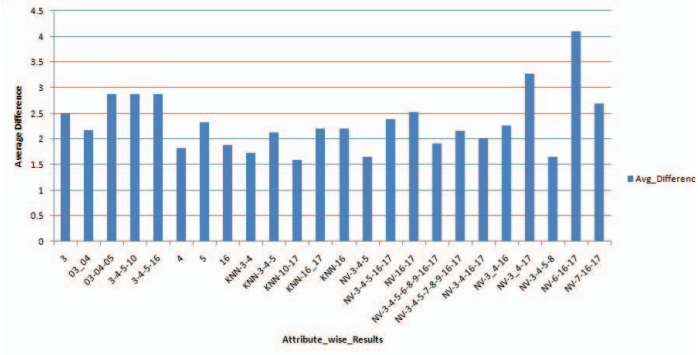

Fig. 7. Combination of attributes Vs average difference

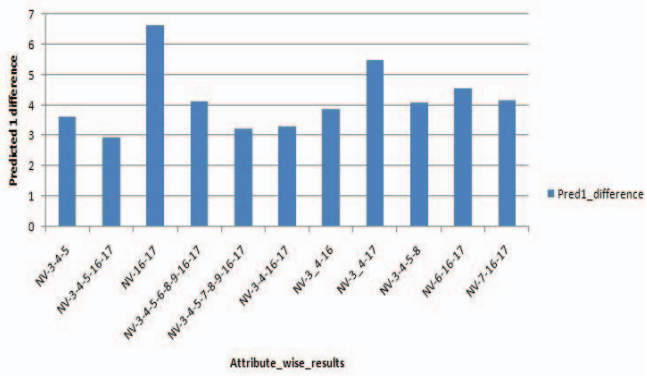

Fig. 8. Combination of normalized attributes Vs average difference

Fig. 8 shows different combinations of data characteristics with normalized values and predicted 1st best classifier difference over all thirty eight datasets. From the above graph, NV-3-4-5-16-17 gives better recommendation so this combination is used for recommendation of classifier.

Prefix NV is used before different combinations of data characteristics in fig. 8. It shows that normalized values of the data characteristics are used. Min-max normalization is used which reduces impact of large values on accuracy of recommendation. E.g. There are two datasets namely hypothyroid and sick datasets in UCI repository [24] having 3772 instances each. If we use the number of instances as a meta-feature then it will dominate the other meta-features. Hence normalization of such meta-feature values is required. It helps to increase accuracy in recommendation. We have normalized values in the range of 0-1.

Thus the number of attributes, the number of instances, number of classes, maximum probability of class and class entropy are playing a major role on classifier accuracy and algorithm selection.

As shown in the fig. 9, a linear equation is formulated on the basis of how objects are plotted on two dimensional planes i.e. A graph of actual and predicted accuracies are plotted. But in this graph the object distance from mean line is above 6.46.

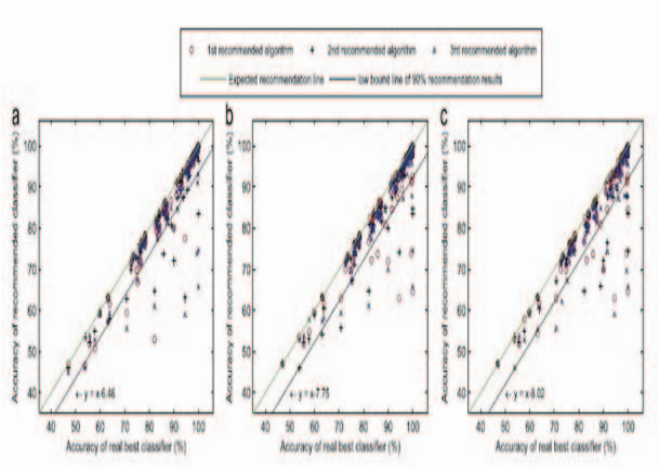

Fig. 9. Accuracy Comparison in approach [28]

In the research work, the distance factor is reduced to 2.953 as shown in fig. 10. This means accuracies of actual and predicted are close enough as compared to previous results. If different combinations are tried then it gives maximum 5.47 distances from mean. It shows that results from proposed work are better than in [28].

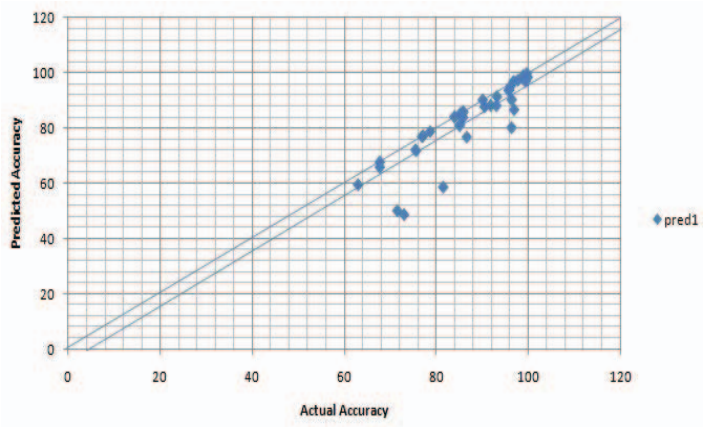

Fig. 10. Graph of actual Vs predicted accuracy for best combination

V. Conclusion and Future Work

In this paper, algorithm selection is proposed for classification problems in data mining. The characteristics of datasets and the performance of classification algorithms are found out. Then based on the problem of classification, or the new problem at hand, mapping is done between the datasets and the benchmark performance of different classifiers. K-similar datasets are returned. Then ranking of classification algorithms is performed and based on the highest rank, the classification algorithm is recommended for the problem at hand. Hence the user doesn't need to waste time for working on different data mining algorithms, fine tuning the parameters for different algorithms. The algorithm is directly recommended for his problem.

978-1-4673-8460-5/16/$31.00 ©2016 IEEE

As per No Free Lunch Theorem, no single algorithm gives better performance for all the available types of data. Different approaches are used in literature to recommend algorithm. The proposed algorithm selection approach recommends approximate best classifier using classification accuracy measure. It helps non-experts in deciding algorithm. The experimentation shows predicted accuracies are matching with the actual accuracies for more than 90 % of the benchmark datasets used for experimentation.

Empirical work uses three different types of meta-features, namely simple, information theoretic and statistical meta-features. Experiments on feature selections suggests the essential features such as number of attributes, number of instances, number of class labels, maximum probability of class and class entropy for classifier selection for thirty eight benchmark datasets and nine classifiers used.

Improvements in algorithm selection can be made by using weighted K-NN algorithm and rough sets. There is a need to work on other meta-features. Further there is requirement to optimize meta-features selection. Here rough sets can be used for meta-features reduction and optimal selection. This will reduce burden on calculation of all the different meta-features and the time will be saved while the algorithm is selected and cognitive overload will be reduced further for non-experts.

The learning parameters used for training classifiers affects the performance of the classifiers. So the further work is required to discover the suitable values for a given dataset. Grid search and evolutionary algorithms can be used for this task. This will be helpful in model selection which includes not only algorithm selection but also parameter optimization.

VI. Discussion

The major contributions of the work are as follows:

- Experiments are performed on thirty eight real world datasets, nine classifies have been used from different categories and three data characterization methods or meta-features have been used for experimentation. The experimental work shows that for 90 % datasets, the predicted and actual accuracies closely match. Hence the algorithm selection or recommendation is correct for these datasets.
- Min-max normalization is used which reduces impact of large values on accuracy of recommendation. This is not used so far in meta- learning approach. There are two datasets namely hypothyroid and sick datasets in UCI repository [24] having total 3772 instances. If we use the number of instances as a meta-feature then it dominates the other meta-features. Hence normalization of such meta-feature values is required. It helps to increase accuracy in recommendation. We have normalized values in the range of 0-1.

Our work shows that the number of attributes, the number of instances, number of classes, maximum probability of class and class entropy are the major data characteristics which affect classifier accuracy and algorithm selection.

One important research direction in meta-learning is to find out other meta-features which will be helpful for characterizing datasets. Meta-learning can be used as an assistant tool for model selection and combination tasks. Classification and regression are the most common tasks and widely used in a daily business practice. Also they are used in a number of application domains. Hence, a decision support is offered by a meta-learning system. It will have a strong impact in future applications in data mining and other fields. Expert knowledge is always costly and not readily available. Also there can be personal preferences. In these situations, meta-learning can strongly serve as useful complement. This is possible by automatic collecting and exploiting meta-knowledge.

There is a need to develop an adaptive system which will be intelligent enough to change data characteristics dynamically and then change or improve classifier adaptively. There is a lot of scope for research in selection of algorithms based on context. This technique can be used in expert systems for optimally selecting classifier based on original characteristics

978-1-4673-8460-5/16/$31.00 ©2016 IEEE

TABLE III. META-FEATURES FOR THIRTY EIGHT DATASETS

ID	DATASET	Best_classifier	NO_attr	No-instance	no_of_class	no_symbolic	no_numeric	missing_value	no_distinct	entropy
1	Anneal	RF	38	798	5	32	6	0	83	1.18
2	Audiology	LO	69	226	24	69	0	317	178	3.42
3	balance-scale	NB	4	625	3	0	4	0	3	1.31
4	breast-cancer	NB	9	286	2	9	0	9	43	0.87
5	breast-w	SM	9	699	2	0	9	16	2	0.92
6	bridges_version1	NB	12	107	6	10	2	73	134	2.30
7	bridges_version2	LO	12	107	6	11	1	73	141	2.30
8	Car	PA	6	113	2	5	1	0	14	0.17
9	Colic	BA	22	368	2	16	6	1927	57	0.95
10	credit-a	J4	15	690	2	9	6	67	42	0.99
11	credit-g	NB	20	1000	2	13	7	0	56	0.88
12	cylinder-bands	SM	39	540	2	22	17	999	533	0.98
13	Diabetes	SM	8	768	2	0	8	0	2	0.93
14	Ecoli	LO	7	336	8	0	7	0	8	2.18
15	Flags	LO	29	194	8	3	26	0	217	2.32
16	Glass	RF	9	214	6	0	9	0	6	2.17
17	heart-c	LO	13	303	2	8	5	7	21	0.99
18	heart-h	NB	13	294	2	8	5	782	21	0.94
19	heart-statlog	SM	13	270	2	0	13	0	2	0.99
20	Hepatitis	SM	19	155	2	13	6	167	28	0.73
21	Hypothyroid	LO	29	3772	4	23	6	6064	50	0.46
22	Ionosphere	PA	34	351	2	0	34	0	2	0.94
23	Iris	SM	4	150	3	0	4	0	3	1.58
24	Labor	SM	16	57	2	9	7	326	23	0.93
25	Lymph	BA	18	148	4	15	3	0	48	1.22
26	Segment	RF	19	2310	7	0	19	0	7	2.80
27	Sick	J4	29	3772	2	23	6	6064	48	0.33
28	Sonar	IB	60	208	2	0	60	0	2	0.99
29	Soybean	SM	35	683	19	35	0	1	151	3.83
30	Sponge	SM	45	76	3	42	3	22	230	0.47
31	tic-tac-toe	IB	9	958	2	9	0	0	29	0.93
32	ToPlayOrNotToPlay	LO	4	14	2	4	0	0	12	0.94
33	Trains	J4	32	10	2	16	16	51	40	1.00
34	Vehicle	RF	18	846	4	0	18	0	4	1.99
35	Vote	RF	16	435	2	16	0	392	34	0.96
36	Vowel	RF	13	990	11	3	10	0	30	3.45
37	Weather	IB	4	14	2	2	2	0	7	0.94
38	Zoo	SM	17	101	7	16	1	0	137	2.39

REFERENCES

[1] Lars Kotthoff, Ian P. Gent, Ian Miguel: An evaluation of machine learning in algorithm selection for search problems. AI Communications. 25(3), pp.257–270, 2012.

[2] J. Rice: The algorithm selection problem. Advances in Computing. 15, pp.65-118, 1976.

[3] Ricardo Vilalta, Christophe Giraud-Carrier, Pavel Brazdil, Carlos Soares: Using meta-learning to support data Mining. IJCSA. 1(1), pp.31-45, 2004.

[4] T. Joachims: Making Large-Scale SVM Learning Practical. Advances in Kernel Methods – Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT Press, 1999.

[5] D.H. Wolpert, W.G. Macready: No free lunch theorem for search, Technical Report SFI-TR-05-010, Santa Fe Institute, Santa Fe, NM,

978-1-4673-8460-5/16/$31.00 ©2016 IEEE

1995.

[6] Carl P. Gomes, Bart Selman: Algorithm Portfolios. Artificial Intelligence, 126, pp.43- 62, 2001.

[7] K. A. Smith-Miles: Cross-disciplinary perspectives on meta-learning for algorithm selection. ACM Computing Surveys. 41(1), 2008.

[8] Pavel Brazdil , Carlos Soares, J. Da Costa: Ranking learning algorithms: Using IBL and meta-Learning on accuracy and time results. Machine Learning. 50(3), pp.251-277, 2003.

[9] Carlos. Soares, Pavel Brazdil: Zoomed ranking: selection of classification algorithms based on relevant performance information, In: Proceedings of 4th European Conference on Principles of Data Mining and Knowledge Discovery, pp. 126-135, 2000.

[10] R. Vilalta, Y. Drissi: A perspective view and survey of meta-learning. Artificial Intelligence Review, 18, pp.77-95, 2002.

[11] Christophe Giraud-Carrier: Meta-learning tutorial. Technical report, Brigham Young University, 2008.

[12] P. Andersen, N. C. Petersen: Procedure for ranking efficient units in data envelopment analysis. Management Science. 39(10), pp.1261-1264,1993.

[13] H. Berrer, I. Paterson, J. Keller: Evaluation of machine learning algorithm Ranking advisors, In: Proceedings of the PKDD Workshop on Data-Mining, Decision Support, Meta-Learning and ILP: Forum for Practical Problem Presentation and Prospective Solutions, 2000.

[14] Matthias Reif , Faisal Shafait, Markus Goldstein, Thomas Breuel, Andreas Dengel: Automatic classifier selection for non-experts. Pattern Analysis and Applications. 17(1),pp. 83-96, 2014.

[15] R. Engels, C. Theusinger: Using a data metric for pre-processing advice for data mining applications", In: Proceedings of the European Conference on Artificial Intelligence, pp. 430-434, 1998.

[16] S. Segrera, J. Pinho, M. Moreno: Information-theoretic measures for meta-learning, In E. Corchado, A. Abraham, W. Pedrycz (eds.) Hybrid Artificial Intelligence Systems, Lecture Notes in Computer Science, 5271, pp.458-465, Springer Berlin / Heidelberg , 2008.

[17] B. Pfahringer, H. Bensusan, C. Giraud-Carrier: Meta-learning by landmarking various learning algorithms, In: Proceedings of the Seventeenth International Conference on Machine Learning, pp. 743-750 2000.

[18] T. Cover, P. Hart: Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), pp. 21–27, 1967.

[19] Munehiro Nakamura, Atsushi Otsuka, Haruhiko Kimura: Automatic selection of classification algorithms for non-experts using meta-features. China-USA Business Review. 13(3), pp. 199-205, 2014.

[20] J. Han, M. Kamber: Data Mining Concepts and Techniques, Morgan Kaufman Publishers, 2011.

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

[22] B. Leo: Random forests. Machine Learning, 45(1), pp.5-32, 2001.

[23] Michie, D., Spiegelhalter, D.J., Taylor, C.C. Machine Learning, Neural and Statistical Classification, Ellis Horwood Series in Artifcial Intelligence, Chichester, 1994.

[24] K. Bache, , M. Lichman: UCI machine learning repository, University of California, School of Information and Computer Science, Irvine, CA, 2013. http://archive.ics.uci.edu/ml/

[25] H. Mark, F. Eibe, H. Geoffrey, P. Bernhard, R. Peter, H. W. Ian: The WEKA data mining software: An update. SIGKDD Explorations. 11(1), pp.10-18, 2009.

[26] P. Hall, J. Racine, Li Q: Cross-validation and the estimation of conditional probability densities. Journal of the American Statistical Association. 99(468), pp.1015- 1026 , 2004.

[27] S. Gore, N. Pise: Dynamic algorithm selection for data Mining classification. International Journal of Scientific & Engineering Research. 4(12), pp. 2029-2033, 2013.

[28] Q. Song, G, Wang, C. Wang: Automatic recommendation of classification algorithms based on dataset Characteristics. Pattern Recognition. 45(7), pp.2672-2689, 2012.

[29] A. Shawkat, A. S. Kate: On learning algorithm selection for classification. Applied Soft Computing. 6(2), pp.119-138, 2006.

[30] Tiago P. F. de Lima, Adenilton J. da Silva, Teresa B. Ludermir, Wilson R. de Oliveira: An automatic methodology for construction of multi-classifier systems based on the combination of selection and fusion. Progress in Artificial Intelligence. 2, pp. 205–215, 2014.

[31] Yonghong Peng, Peter A. Flach, Carlos Soares, Pavel Brazdil: Improved dataset characterisation for meta-learning. Lecture in Notes in Computer Science, 2534, pp.193-208, Springer Berlin Heidelberg , 2002.

[32] Pavel Brazdil, Christophe Giraud-Carrier, Carlos Soares, Ricardo Vilalta: Meta learning: Applications to Data Mining, Springer Publishing Company, 2008.

[33] Ohbyung Kwon, Jae Mun Sim: Effects of data set features on the performances of classification algorithms. Expert Systems with Applications. 40, pp.1847–1857, 2013.

[34] Parker Ridd, Christophe Giraud-Carrier: Using meta-learning to predict when parameter optimization is likely to improve classification accuracy, :In Proceedings of 21st European Conference on Artificial Intelligence, pp.18-23, 2014.

[35] Robi Polikar: Ensemble based system in decision making. IEEE Circuit and System Magazine. 6(3), pp.21-45, 2006.

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

[37] S. Dzeroski, B. Zenko: Is combining classifiers with stacking better than selecting the best one?". Machine Learning. 54, pp.255–273, 2004.

[38] J. Friedman, T. Hastie, R. Tibshirani: Additive logistic regression: a statistical view of Boosting. Annals of Statistics. 28(2), pp.337-407, 1998.

978-1-4673-8460-5/16/$31.00 ©2016 IEEE