Journal of Computers Vol. 28, No. 5, 2017, pp. 26-38
doi:10.3966/199115992017102805003

Choosing Classification Algorithms and Its Optimum Parameters based on Data Set Characteristics

Yang Zhongguo[1], Li Hongqi[1*], Sikandar Ali[1] and Ao Yile[1]

[1] Department of Computer Science and Technology, China University of Petroleum, Beijing 102249
yangzhongguo@hotmail.com, hq.li@cup.edu.cn, hqsikandar@qq.com, 3180218892@qq.com

Received 17 June 2016; Revised 7 February 2017; Accepted 8 February 2017

Abstract. Choosing a correct classification algorithm for a given data set is an important task considering the existing multiple classifiers. A method of recommending a suitable algorithm and its optimum parameters for a given data set is proposed. Firstly, six different types of measures are computed for each data set to be representation of its characteristics. Then, the performance and optimum parameters for a given algorithm are computed by using grid search method. Afterwards, a model was built to predict the variance of classifiers for a given data set and another model was built to predict the best suitable algorithm. The proposed method tries to predict the optimum parameter for a certain algorithm based on knowledge learning from history data sets. To evaluate the performance of the proposed method, some extensive experiments for four different types of algorithms are conducted upon the UCI data sets. The results indicate that the proposed method is effective.

Keywords: algorithm automatic recommendation, algorithm performance, data set characteristics, optimum parameters

1 Introduction

Number of classification algorithms challenge practitioners of pattern recognition system to choose a proper algorithm for their problems. It is well known that different algorithm own its suitable field i.e., algorithm Alg1 could not defeat Alg2 in all data sets. As indicated by Weiss and Kapouleas [1], Back-Propagation neural networks achieve a higher accuracy than decision tree method on Iris and Appendicitis data but a lower accuracy on Breast cancer and Thyroid data. Many other examples strengthen this impression, such as Shavlik, Mooney and Towell, Duin, Ali and Smith, etc [2-4]. These examples show a brief account on the complex performance situation of the different classification algorithms. In conclusion, it reveals that no single algorithm can perform uniformly well over all data sets. In addition, it is consistent with Wolpert and Macready's well known No Free Lunch theorem [5]: "no single method may outperform others in all situations".

The necessity of choosing a proper algorithm for a new classification problem induces researchers to study algorithm recommendation method. The research efforts are currently focused on two fronts. The first aims to develop new algorithms to replace existing algorithms and the other line aims to choose a suitable algorithm by using empirical knowledge. The former method cannot guarantee the new algorithm could outperform other algorithms on all data sets. The later line gradually becomes the main line which aims to solve this problem and there are many study papers in this filed [19-33].

These algorithm recommendation methods are based on the assumption that data set characteristics are important factors which could affect the performance of a classification algorithm i.e., there exist some intrinsic relationship between classification algorithm performance and data set characteristics. A formalized version of an algorithm selection problem is proposed by Rice [6] and the content is as follows: for a given task in a problem space with features, finding the selection algorithm in algorithm space A, in the way that the selected algorithm maximizes the performance mapping in terms of a

[*] Corresponding Author

performance measure. This problem has been recognized as a learning task and was named meta-learning in the machine learning community.

Meta-knowledge or named meta-data and meta-target constitute the main part of a meta-learning system [7]. The meta-data are the characteristics or features extracted from the data sets and the meta-target is the target variable for the meta-learning system which could be performance of algorithms [8-9]. The meta-features of a data set could be rather simple, such as the number of samples, features, classes, types of features, class entropy, average feature entropy, average mutual information, noise-signal ratio, outlier measure of continuous features and number of features with outliers.

If we gather enough knowledge about different tasks and performance of distinct algorithms on them, we can rank algorithms by value of the forecasting error for each of those tasks. Based on the characteristics of a new task, algorithms can be ranked on the assumption that for tasks with similar characteristics the same algorithm will return the similar recognition error. More theoretical background and examples of meta-learning can be found in [7, 9].

In these works, how to characterize data set is the key point to the success of these methods and a number of different kinds of measures are employed to fulfill this task. In summary, they are statistical measures [8, 12], classification complexity measures [10], information-theoretic measures [11], land-marking measures [13-14], and model-based measures [15]. Some recent proposals can be found in [16-18]. In addition, the parameter setting for a classifier is an important factor which influences performance of an algorithm on a given data set and it is ignored in the process of evaluation of algorithm in different degree. This paper follows the line of algorithm recommendation method and proposes a method to choose a suitable classifier and its optimum parameter settings.

The proposed method consists of three main steps: (1) predict the variance of performance of different algorithms for a given data set (2) predict the best algorithm (3) predict the optimal parameters for algorithm. The reason for first step is that that if there is little variance among the performances of candidate algorithms and there is no need to do future recommendation. Also, the experiments conducted on UCI data sets demonstrated that C4.5, K-NN, SVM are the best algorithms among nine algorithms if they are executed under their own optimum parameter setting. So, the candidate algorithms could be limited to these three algorithms.

The contribution of this is as follows:

(1) The proposed divided these data sets into two groups considering their performance's sensitiveness to classification algorithm.

(2) The proposed method predicted the optimum parameter for C4.5 and k-NN algorithm.

The rest of this paper is organized as follows. Section 2 reviews the some related work on algorithm recommendation. Section 3 demonstrates a case study about the performance of classifiers. In section 4, a clear description of the proposed method is shown. Section 5 introduces the process of parameter turning. Section 6 shows the experimental study performed and the analysis of results. Finally, Section 7 enumerates some concluding remarks.

2 Related Works

Many meta-learning methods are proposed by using different characteristics of data sets. Rendell and Cho [19] developed rules based on simple meta-features such as number, error, "size", concept size and concentration to determine if a certain algorithm should be used for a problem instance or not.

Aha [20] presented multiple rules learning from case studies which are used to identify optimum algorithm. These rules were built in known performance of algorithms and database-characterization space.

Brodley [21] solved the automatic selection of learning algorithms problem by using a heuristic best first search method to conduct the search from the available algorithm space to find the best classifier automatically. This method captured the knowledge of domain experts concerning the applicability of certain classification algorithms.

Brazdil, Gama and Henery extracted meta-features for various data sets with statistical and information theoretic measures. And then, a lot of algorithms were executed on these data sets and algorithm applicability information was determined. Finally, a decision model was used to generate rules to give recommendations for a new data set [22]. This approach was later extended by using more features and a decision tree learner within the well-known StatLog project [23].

The case-based reasoning method was applied for meta-learning by Linder et al [24]. A multiple of known and solved problem instances were prepared as history data sets and meta-features were extracted as representation of data characteristics. The similarity of problems was determined by meta-features and the finally recommendation algorithm concerned more factors such as the interpretability of the produced model and training time.

Another type of meta-learning system is described by Gama and Brazdil [25] in which a linear regression model was used to recommend algorithms which aimed to capture information concerning applicability of algorithms. These models are generated automatically on the basis of data set characteristics. Köpf, Taylor and Keller [26] also presented results of using regression for meta-learning and M6 method was used as a meta-regression learner in the process of experiment. Experiments on artificial datasets were conducted and the error rates of three classification algorithms were predicted. Additionally, the regression method selected the classifier with the lowest predicted error as the best classifier and the experiments result was compared to the classification approach. The meta-features consist of information measures and statistical measures.

Ali and Smith [27] also proposed a rule-based classifier selection approach and their process consists of three steps: (1) meta-features on problem characteristics are extracted (2) empirical performance of eight classification algorithms are extracted (3) decision tree model was built to generate rules. The meta-features used here are based on Smith et al. [28, 29].

Brazdil et al. [8] presented a new strategy to recommend algorithm by using K-NN algorithm to identify the most similar history data sets. Similar performance of the candidate algorithms and relevant rank of them are expected based these history data sets. The similarity between data sets are computed by meta-features which consist of statistical information and entropy information.

Follow the idea of connecting data characteristics of data sets to algorithm performance, Kalousis and Gama [30] try to find functions that map data sets to algorithm performance. The key point is meta-features extracted to be representation of the characteristics of the learning tasks.

Bernado´-Mansilla and Ho [31] utilized the problem complexity measures to characterize the data sets and analyzed the relation between these measures and performance of classification algorithms. Then this relation was employed to recommend appropriate algorithms.

Ali et al. [4] used a rule learner based on meta-features to predict the most suitable classifier. For each target classifier, a rule was learned, determining if the classifier should be used. For the training data, the best classifier was determined by a combined measure of accuracy and time. This work is also one of the few papers until today taking the SVM classifier into account.

Be different from the previous work, Song [32] used structural and statistical information based feature vector instead of meta-features to characterize each data set and adopt the k-NN method to identify k most similar data sets. And algorithm recommendation is conducted on these similar data sets.

Matthias Reif [33] used five different categories of state-of-the-art meta-features to characterize data sets and built different regression model to connect data sets to each candidate algorithm. Finally, these models are used to predict performance of a new data set.

As is well known that some data sets is easy to hand by using any algorithm while other data sets are difficult to hand may due to their intrinsic difficulties. The proposed paper tried to divided data sets into two groups, one is easy to classify by using any algorithm and the other one is difficult to classify by using any algorithm. Based on their simplicity two different predict models are built to recommend different algorithm for them.

3 A Case Study for Performance of Classifiers

There are mainly four categories of classification algorithms, including tree-based (e.g., ID3 [34], C4.5 [35] and CART [36]), probability-based (e.g., Naive Bayes [37] and AODE [38]), rule-based (e.g., RIPPER [39], CN2 [40] and PART [41]), and association-based (e.g., CBA [42], CMAR [43], MCAR). Support vector machine [44] is another widely used method. In conclusion, nine algorithms were picked to study the characteristics of data set and they are C4.5, NB, 3NN, Decision Stump, LWT, KStar, OneR, PART and SVM.

A case study was conducted on a collection of data sets, which consists of 100 data sets from UCI Machine Learning Repository [45], covering a variety of application areas, such as engineering, physics, biology, medicine, and games. Fig. 1 shows the dimensionality of these data sets.

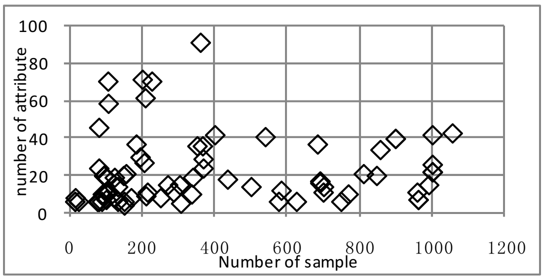

Fig. 1. The dimensionality of the data sets

The case study considers their performances on 84 UCI data sets under default parameters in Weka. Fig. 2 demonstrates the AUC value of ten algorithm by 10-fold cross validation under default parameters.

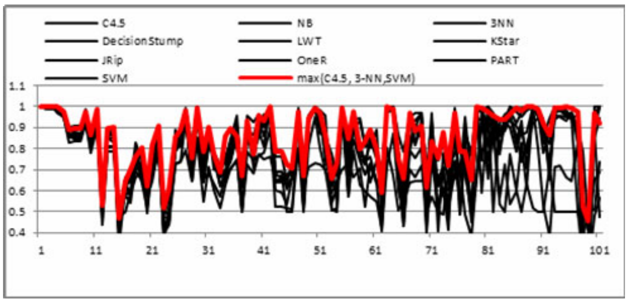

Fig. 2. AUC of data sets for ten algorithms

The red line is the maximum accuracy of C4.5, 3NN and SVM algorithm. As the experiment shows, the max (C4.5, 3-NN, SVM) achieves the best accuracy of all the nine algorithms. The result is reasonable considering the three algorithms' own principle; C4.5 is rule-based which suitable for globally well distributed data set while K-NN and SVM algorithm can take good advantage of data's local characteristics.

Fig. 3 demonstrates the variance of the accuracies of above-mentioned algorithms on each data set. As we can see, there are 90.5 percent of the data sets which own a small value of variance (<=2%) among the accuracy of nine algorithms, which means that the difference of the performances for these data sets for the nine classifiers is small. As indicated in [33], the accuracy of each algorithm is predictable by using linear regression model built on data set's characters. Naturally, the variance of the performances of these algorithms for each data set is predictable.

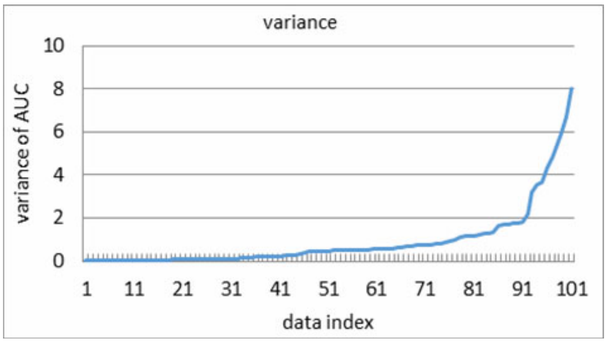

Fig. 3. Variance of AUC for data sets under ten algorithms

4 Method

4.1 Architecture

In order to provide a method based on the characteristics of the data which enables one to predict the best classifier and its parameter values. Fig. 4 shows the methodology.

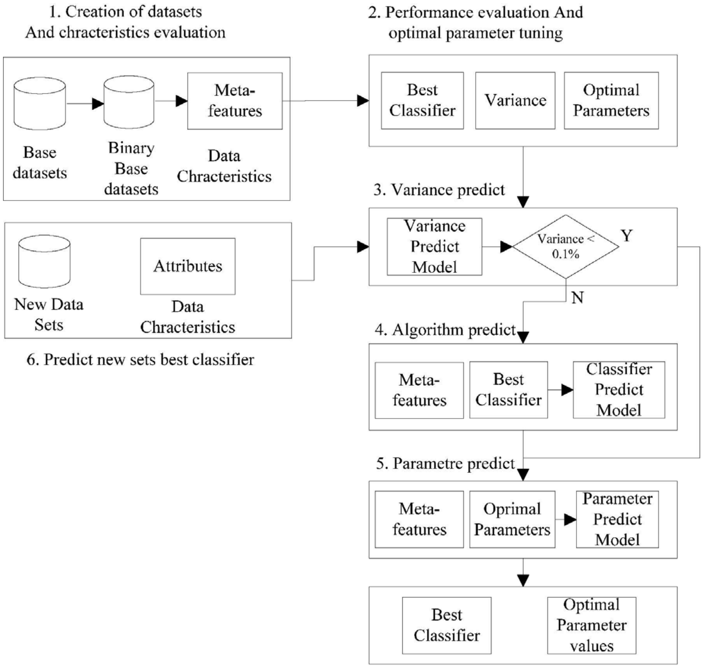

Fig. 4. Methodology to predict best classifier and parameter values for a given data set

The complete process is described as follows:

(1)84 different classification datasets are built and their meta-features are extracted. Multi-class datasets are used to create other binary datasets by means of the selection and/or combination of their classes. Only problems with two classes are considered as some data complexity measures are only well defined for binary problems. The meta-features used consist of six different types of data characteristic measures for each dataset. The details about these measures are demonstrated in appendix A.

(2)The test performances of nine classifiers on each of the 84 datasets are computed. The variance and the best classifier are computed and used as two labels to be predicted.

(3)Build model to predict the variance of performance for a given data.

(4)Build model to predict the suitable algorithm among C4.5, KNN and SVM classifiers.

(5)Build model to predict the optimal parameters for the recommended classifier.

(6)Use the model trained above to predict the variance, best classifier and optimal parameters for each data set.

4.2 The Proposed Measures for Describing Data Characteristics

Meta-attributes are common characteristics of real problems, which aim to identify structural similarities and differences among different problems [7, 46-47]. These characterization can be categorized into direct and indirect ones. The former consists of simple feature, statistic feature, information based features. A new characterization method named land-marking was proposed in [13], focusing on the usage of simple classification algorithms to extract the meta-attributes. For instance, the accuracy of the simple algorithms such as Naïve Bayes and decision trees are related to the intrinsic features of the problem and thus can be used to indicate problems with similar characteristics. In addition, the land-marking and model based characterization are known as indirect characterization because they are not directly related with the problems' attributes. Another new characterization method is based on information extracted from models built out of the data sets. Typically, a decision tree was constructed and some structural information about the tree was extracted, such as the number of nodes or leaves or the depth and width of the tree.

In conclusion, there are six different types of data characteristic measures including simple feature, statistic feature, information based features, model based features, land-marking measures and data complexity based measures. Appendix A shows more details of them. Some of these meta-attributes are only for continuous attribute, such as kurtosis and skewness of attributes. If a data set has only nominal attributes, then the value of these measures will receive zero values, and vice versa. For a deeper description of these data characteristic measures, the reader may consult [8, 10-11, 13, 48-50].

4.3 Predict Variance of Performance of Algorithms

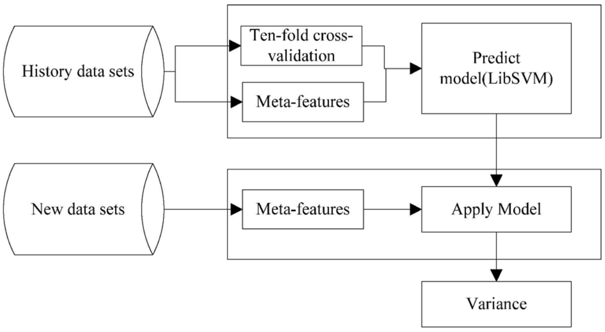

Fig. 5. Methodology to predict variance of data set

As mentioned in section 4.1, we have to build a predict model based on history data sets. Fig. 5 shows the process of predicting variance of a data set. The training data set have every history data set as its instance, the feature of training data consists of six different measures. LibSVM algorithm are applied as predict model because it performs well on small number of sample examples. The experiment is conducted as following steps:

Prepare history data. Public available UCI data sets are preprocessed to binary data sets with aim to suit for data complexity measures which are well designed for binary data set s.

Prepare training data. Every metric mentioned above are measured to be one of feature of the training data and the variance of ten algorithms' performance are computed as class label of the training data. The performance for each algorithm is computed under its default parameters setting. Each data set is one instance of the final training data sets.

Build prediction model and predict variance. LibSVM algorithm is applied as the prediction model and the ten-fold cross validation method are used. The result is shown in Table 1, the mean of MAE is 0.533 and the mean of MRE is 3.530 which show the effectiveness of this method. And the predicted variance and real variance of data set is shown in Fig. 6.

Table 1. The absolute error and the relative error of predicted variance of data sets

	Mean	Std
MAE	0.533	0.377
MRE	3.530	11.470

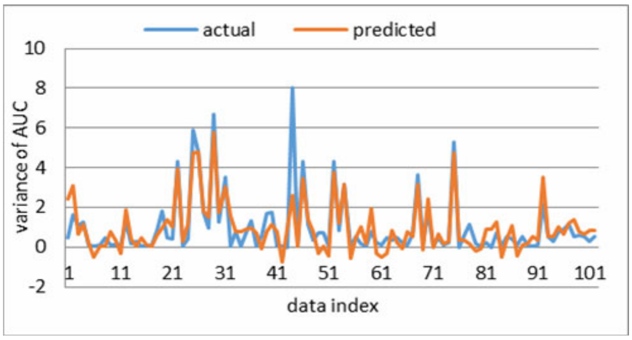

Fig. 6. Predicted variance for data sets

5 Parameters Tuning

Many academic papers and machine learning systems provide evidences that the experiments results may vary widely when different values for the parameters are employed [1, 6, 19, 21, 23]. When apply an algorithm to a date set, different parameters could result in different result. How to choose an appropriate parameter settings is a challenge task. Similarity, we use meta-learning method to predict the parameter for a certain algorithm. Given only C4.5, K-NN and SVM algorithms are final recommended algorithm, so, we have to predict parameters for each of them. Also, as the real experiments show that the radical function works well on most data set for SVM algorithm, we only need to predict parameters for C4.5 and K-NN.

5.1 The Best K for K-NN

K-NN algorithm is a widely used classifier given its simplicity. However, k is difficult to decide when facing a real data set. A little k as 1 means the algorithm would set the nearest neighbor's class to the new example while a large k as the number of examples means choosing the majority class's label to as the final result. Both of these two cases are not suitable and the best k is the one could achieve the best performance for a given data set. The performance is connected to data set's characteristics deeply and we could predict the best k by employing the right data characteristics.

A case study was conducted on UCI data sets following three steps as above mentioned architecture for predicting variance of performance of algorithms. Fig. 7 shows the process of predicting best K for K-NN for a data set. The label of the meta-data is the value of k instead of variance. Fig. 8 shows the AUC value of K-NN under predicted K and best K.

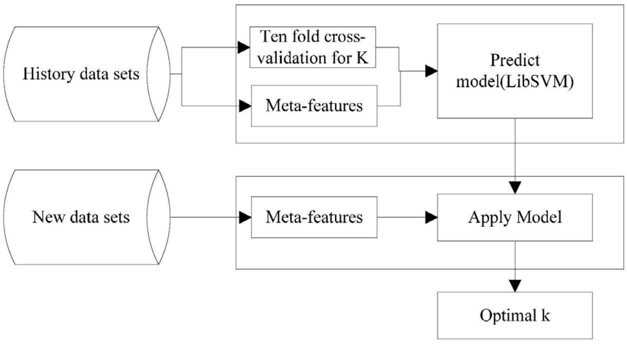

Fig. 7. The process of predicting k for K-NN

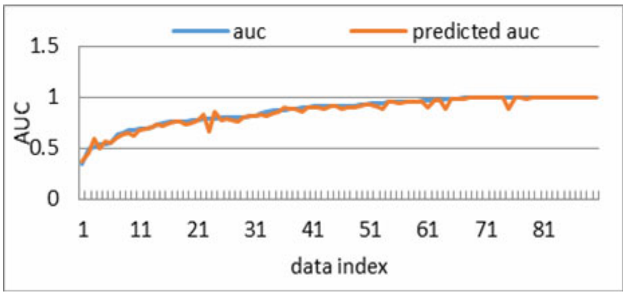

Fig. 8. The accuracy of K-NN under optimum parameters and predicted parameters

5.2 The Optimum Parameters for C4.5

Decision tree is another widely used classifier in data mining task. As indicated in [51], there are two parameters which influence the amount of pruning and the final performance of decision tree.

(1)confidenceFactor is the confidence factor for pruning, and it influences the size and predictability of the tree constructed. For each pruning operation, it defines the probability of error in the hypothesis that deteriorate on due to this operation is significant. The default value is 0.25. The lower this value, the more pruning operations allowed.

(2)minNumObj is the minimum number of instances per leaf. The default value is 2.

We choose the algorithm J48 in Weka for implementation of C4.5, and execute it using different parameters setting while stored the AUC value obtained in each execution as part of the meta-database. Each parameter combination is evaluated using 10-fold cross-validation and the AUC value is stored. The combination consists of confidence Factor(0.1, 0.25, and 0.5) and minNumObj(1, 2, and 10). In this way, each parameter combination is assigned with an AUC value of C4.5 and the optimum parameter combination is the corresponding result of the highest AUC value.

A case study was conducted on UCI data sets following the same process of predicting variance of data sets or the K value of K-NN. The difference is the label of training data which uses parameter combinations as the meta-target instead of variance or K values. The process is shown in Fig. 9. Fig. 10 shows the best AUC value of C4.5 under optimum parameter settings and predicted parameters for a give data set. The result indicates that we could predict the optimum parameter settings for C4.5 based on meta-data of history data sets.

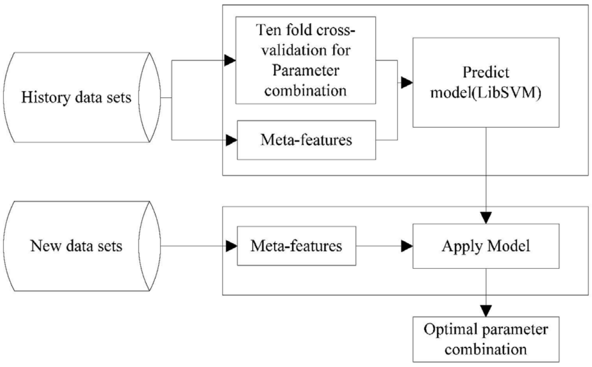

Fig. 9. The process of predicting optimum parameters of C4.5 for a data set

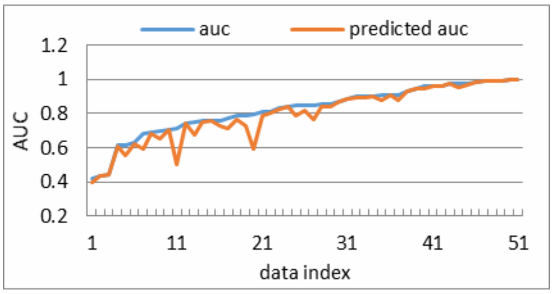

Fig. 10. The accuracy of C4.5 under optimum parameters and predicted parameters

6 Experimental Results and Analysis

6.1 Date Sets

All the data sets are come from publicly available UCI data sets and the performance of each classifier was computed by using their default parameters. As the recommendation algorithm is among the three algorithms: C4.5, K-NN, SVM, the training data owns a label which consists of C4.5, K-NN and SVM. The accuracy of each algorithm was evaluated under their default parameter settings using ten-fold cross-validation. Table 3 shows the process of predicting the best suitable algorithm.

6.2 Result and Analysis

As mentioned above, the proposed method is consists of three main steps and the first step is to predict the variance of one data set and the third step is to predict the parameter of a data set as illustrated in section 3 and section 4. This section focus on the second step to predict the classifier among C4.5, K-NN and SVM. Similar to the process of predicting variance or parameter of a data set, the process of predicting the best classifier is also based on training data sets. Firstly, a best classifier was assign to each data set. As a result, each data set owns a best classifier among three classifiers. The feature vector for the training data consists of six different types of measures of data characteristics. Secondly, we choose 5-NN as the predict model which performs well in this task.

Three measures are used to evaluate the proposed method and they are magnitude of absolute error (MAE), magnitude of relative error (MRE) and PRED (m). PRED (m) is the measure of ratio of the number of examples which owns a MRE smaller than m% to the total number of all examples.

To compute these measures, the highest accuracy of the all nine algorithms was compared with the classification accuracy of the recommendation algorithm under the optimal parameter setting. The result is shown in Fig. 11 and Table 2.

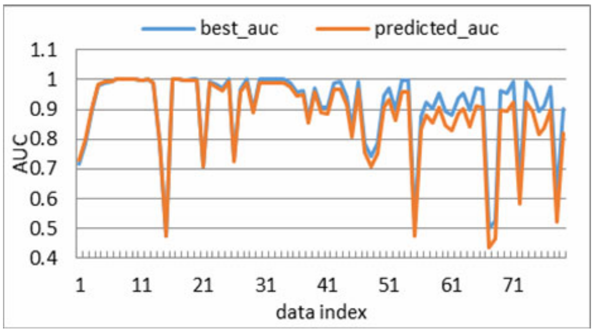

Fig. 11. Classification AUC of recommended algorithm under optimum parameter settings vs. highest classification accuracy of ten algorithms under their default parameter settings

Table 2. The statistical result of the AUC of the recommended algorithm

Data set name	Recomm_Algorithm	Recomm_parameter	Best_auc	Recomm_auc
kr-vs-kp	C45	0.25|3	0.999	0.998
BankNote	KNN	18	1	0.999
Soybean	C45	0.4|5	1	0.998
Transfusion	C45	0.25|3	0.709	0.707
soybean2	C45	0.4|5	0.992	0.989
house-votes-84	C45	0.4|5	0.983	0.977
statlog-vehicle-xaf	KNN	5	0.971	0.965
Lymphography	KNN	8	1	0.993
Abalone	KNN	64	0.734	0.727
Sick	C45	0.25|3	0.967	0.959
balance-scale2	SVM	default	1	0.991
Mammographic-Mass-Data	KNN	11	0.897	0.888
Iris	C45	0.4|5	1	0.99
Segment	C45	0.4|5	1	0.99
statlog-vehicle-xah	C45	0.4|5	0.955	0.892
balance-scale	SVM	default	0.992	0.925
au7	SVM	default	0.65	0.582
Anneal	C45	0.4|5	0.994	0.924
audiology.standardized	C45	0.4|8	0.964	0.892
Flags	C45	0.4|5	0.892	0.818
primary-tumor2	C45	0.4|5	0.911	0.836
breastTissue	KNN	4	0.975	0.899
au6_cd1	C45	0.4|5	0.607	0.523
cleveland-14-heart-disease	SVM	default	0.904	0.819
statlog-vehicle-xac	KNN	6	0.819	0.73
echocardiogram	SVM	default	0.826	0.666
horse-colic.ORIG	SVM	default	0.83	0.666
haberman	SVM	default	0.664	0.5
Shuttle	KNN	4	0.778	0.565

Table 3 shows that 77% of the data sets could achieve an AUC value which differs with the best accuracy in a small range (0.08). Also, the experiment conducted above shows the effectiveness of the proposed method to recommend a classifier for a given data set in the statistical view.

Table 3. The PRED (MAE) result of the proposed method

	6%	8%
MAE	67.28%	77.6%
MRE	62.12%	72.01%

7 Conclusions

In this paper, an algorithm recommendation method was presented based on data set characteristics which aims to assist people in choosing algorithms among a large number of classifiers for a new classification problem. In this method, the data set features are firstly extracted by using six different types of measures, the best algorithm and optimum parameters for these algorithms are computed by using ten-fold cross-validation. Then, several predict models were built which aim to predict variance, algorithm and parameter setting for a new data set.

In order to facilitate the recommendation, a whole process of the proposed method was presented, and it contains three main steps to predict the best algorithm: (1) predict the variance (2) predict the best algorithm among the three classifiers (3) predict the optimum parameter settings for classifier. The new process of recommend algorithm is quite different from the traditional meta-learning method.

With the aim to validate the proposed recommendation method, 100 data sets and 9 different algorithms are used in the experiment. The result shows that the simple algorithm such as K-NN and C4.5 could achieve a good accuracy if the optimum parameter settings are used. Also, the result shows that there exists some relationships between data sets' characteristics and algorithms' parameter settings and we could use them to help people to decide the parameter settings for a given algorithm.

The limitation of this paper is that we only studied nine algorithms which is a small part of classification algorithms. In addition, we did not recommend a proper parameter for SVM algorithm which is a difficult question.

For the future work, we plan to explore further the possible relationships between more data mining algorithm parameters and data set features.

References

[1] S.M. Weiss, I. Kapouleas, An empirical comparison of pattern recognition, in: Proc. of the Eleventh International Joint Conference on Artificial Intelligence, 1989.

[2] J.W. Shavlik, R.J. Mooney, G. Towell, Symbolic and neural learning algorithms: an experimental comparison, Machine Learning 6(2)(1991) 111-143.

[3] RP. Duin, A note on comparing classifiers, Pattern Recognition Letters 17(5)(1996) 529-536.

[4] S. Ali, K. Smith, On learning algorithm selection for classification, Applied Soft Computing 6(2)(2006) 119-138.

[5] D.H. Wolpert, W.G. Macready, No free lunch theorems for search, IEEE Transactions on Evolutionary Computation1 (1)(1997) 67-82.

[6] J.R. Rice, The algorithm selection problem, Advances in Computers 15(1976) 65-118.

[7] P. Brazdil, C.G. Carrier, C. Soares, R. Vilalta, Metalearning: Applications to Data Mining, Springer, Heidelberg, 2008.

[8] P. Brazdil, C. Soares, J.P. da Costa, Ranking learning algorithms: using IBL and meta-learning on accuracy and time results, Machine Learning 50(3)(2003) 251-277.

[9] C. Lemke, M. Budka, B. Gabrys, Metalearning: a survey of trends and technologies, Artificial Intelligence Review 44(1)(2015) 117-130.

[10] Y. Peng, P.A. Flach, C. Soares, Improved dataset characterisation for meta-learning, in: Proc. the 5th International Conference on Discovery Science, 2002.

[11] S. Segrera, J. Pinh, M. Moreno, Information-theoretic measures for meta-learning, in: E. Corchado, A. Abraham, W. Pedrycz (Eds.), Hybrid Artificial Intelligence Systems, Springer-Verlag, Berlin, 2008, pp. 458-465.

[12] D. Michie, D. Spiegelhalter, C. Taylor, Machine Learning: Neural & Statistical Classification, Ellis, New York, 1994.

[13] B. Pfahringer, H. Bensusan, C. Giraudcarrier, Meta-learning by landmarking various learning algorithms, in: Proc. International Conference on Machine Learning, 2000.

[14] H. Bensusan, C. Giraudcarrier, Discovering task neighbourhoods through landmark learning performances, in: Proc. European Conference on Principles of Data Mining and Knowledge Discovery, 2000.

[15] H. Bensusan, C. Giraudcarrier, C.J. Kennedy, A higher-order approach to meta-learning, in: Proc. the 9th Int. Workshop on Inductive Logic Programming, 2000.

[16] D. Elizondo, R. Birkenhead, M. Gamez, Linear separability and classification complexity, Expert Systems With Applications 39(9)(2012) 7796-7807.

[17] M. Matijas, J.A. Suykens, S. Krajcar, Load forecasting using a multivariate meta-learning system, Expert Systems with Applications 40(11)(2013) 4427-4437.

[18] M.A. Munoz, M. Kirley, S.K. Halgamuge, A meta-learning prediction model of algorithm performance for continuous optimization problems, in: C.A. Coello Coello, V. Cutello, K. Deb, S. Forrest, G. Nicosia, M. Pavone (Eds.), Parallel Problem Solving from Nature - PPSN XII, Springer-Verlag, Berlin, 2012, pp. 226-235.

[19] L. Rendell, H. Cho, Empirical Learning as a Function of Concept Character, Machine Learning 5(3)(1990) 267-298.

[20] D.W. Aha, Generalizing from case studies: a case study, in: Proc. the 9th International Conference on Machine Learning, 1992.

[21] C.E. Brodley, Addressing the selective superiority problem: automatic algorithm/model class selection, in: Proc. the 10th International Conference on Machine Learning, 1993.

[22] P. Brazdil, J. Gama, B. Henery, Characterizing the applicability of classification algorithms using meta-level learning, in: Proc. European Conference on Machine Learning, 1994.

[23] R.D. King, C. Feng, A. Sutherland, Statlog: comparison of classification algorithms on large real-world problems, Applied Artificial Intelligence 9(3)(1995) 289-333.

[24] G. Lindner, R. Studer, AST: support for algorithm selection with a CBR approach, in: Proc. the Third European Conference on Principles of Data Mining and Knowledge Discovery in Database, 1999.

[25] J. Gama, P. Brazdil, Characterization of classification algorithms, in: E.P. Ferreira, N. Mamede (Eds.), Progress in Artificial Intelligence. 7th Portuguese Conference on Artificial Intelligence (EPIA-95), Springer-Verlag, Berlin, 1995, pp. 189-200.

[26] C. Köpf, C. Taylor, J. Keller, Meta-analysis: from data characterization for meta-learning to meta-regression. in: Proc. the PKDD-00 Workshop on Data Mining, Decision Support, Meta-Learning and ILP, 2000.

[27] S. Ali, K. Smith, On learning algorithm selection for classification, Applied Soft Computing 6(2)(2006) 119-138.

[28] K.A. Smith, F. Woo, V. Ciesielski, R. Ibrahim, Modelling the relationship between problem characteristics and data mining algorithm performance using neural networks, Smart Engineering System Design: Neural Networks, Fuzzy Logic, Evolutionary Programming, Data Mining, and Complex Systems 1(11)(2001) 356-362.

[29] K.A. Smith, F. Woo, V. Ciesielski, R. Ibrahim, Matching data mining algorithm suitability to data characteristics using a self-organizing map, Hybrid Information Systems Advances in Soft Computing 14(1)(2002) 169-180.

[30] A. Kalousis, J. Gama, M. Hilario, .On data and algorithms: understanding inductive performance, Machine Learning 54(3)(2004) 275-312.

[31] E. Bernadomansilla, T.K. Ho, Domain of competence of XCS classifier system in complexity measurement space, IEEE Transactions on Evolutionary Computation 9(1)(2005) 82-104.

[32] Q. Song, G. Wang, C. Wang, Automatic recommendation of classification algorithms based on data set characteristics, Pattern Recognition 45(7)(2012) 2672-2689.

[33] M. Reif, F. Shafait, M. Goldstein, Automatic classifier selection for non-experts pattern, Pattern Analysis and Applications 17(1)(2014) 83-96.

[34] J.R. Quinlan, Discovering rules by induction from large collections of examples, in: D. Michie (Ed.), Expert Systems in the Micro-Electronic Age, Edinburgh University Press, Edinburgh, 1979, pp. 168-201.

[35] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA, 1993.

[36] L. Breiman, Classification and Regression Trees, Chapman & Hall/CRC, New York, 1984.

[37] A.W. Moore, D. Zuev, Internet traffic classification using Bayesian analysis techniques, in: Proc. the 2005 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, 2005.

[38] G.I. Webb, J.R. Boughton, Z. Wang, Not so naive Bayes: aggregating one dependence estimators, Machine Learning 58(1)(2005) 5-24.

[39] W.W. Cohen, Fast effective rule induction, in: Proc. the Twelfth International Conference on Machine Learning, 1995.

[40] P. Clark, T. Niblett, The CN2 Induction Algorithm, Machine Learning 3(4)(1989) 261-283.

[41] E. Frank, I.H. Witten, Generating accurate rule sets without global optimization, in: Proc. International Conference on Machine Learning, 1998.

[42] B. Liu, W. Hsu, Y. Ma, Integrating classification and association rule mining, in: Proc. Knowledge Discovery and Data Mining, 1998.

[43] W. Li, J. Han, J. Pei, CMAR: accurate and efficient classification based on multiple class-association rules, in: Proc. International Conference on Data Mining, 2001.

[44] F. Tseng, X. Chen, L. Chou, Support vector machine approach for virtual machine migration in cloud data center, Multimedia Tools and Applications 74(10)(2015) 3419-3440.

[45] A. Frank, A. Asuncion, UCI Machine Learning Repository, University of California, School of Information and Computer Science, Irvine, CA, 2010.

[46] K. Smithmiles, Cross-disciplinary perspectives on meta-learning for algorithm selection, ACM Computing Surveys 41(1)(2008) 1-25.

[47] K. Smithmiles, Towards insightful algorithm selection for optimisation using meta-learning concepts, in: Proc. International Symposium on Neural Networks, 2008.

[48] R. Engels, C. Theusinger, Using a data metric for preprocessing advice for data mining applications, in: Proc. the Thirteenth European Conference on Artificial Intelligence, 1998.

[49] S.Y. Sohn, Meta analysis of classification algorithms for pattern recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 21(11)(1999) 1137-1144.

[50] J. Cano, Analysis of data complexity measures for classification, Expert Systems with Applications 40(12)(2013) 4820-4831.

[51] M.M. Molina, J.M. Luna, C. Romero, meta-learning approach for automatic parameter tuning: a case of study with educational datasets, in: Proc. the 5th International Conference on Educational Data Mining, 2012.

Appendix A: Description of the Data Set Characteristic Measures

category	name	detail
Simple features	Num of examples	Number of examples
	Num of class	Number of class
	Num of features	Number of features
	Num of numeric features	Number of numeric features
	Num of nominal features	Number of nominal features
	Ratio of nominal features	Ratio of nominal features
	Ratio of numeric features	Ratio of numeric features
	Num of example of dimension	Number of example of dimension
Statistic features	kurtosis	mean kurtosis of features
	skewness	mean skewness of features
	CANCORI	First canonical correlation between a linear combination of class variables and a linear combination of features
	FRAC1	Proportion of total variation explained by the first canonical discriminate
	CORR	mean absolute correlation coefficients between two features
Information features	ClassEnt	the entropy of the class label
	AttrEnt	the entropy of all attributes
	MutualInf	the mutual information (entropy) of class and attributes
	JointEnt	the joint entropy
	EquivAttr	the equivalent number of attributes
	PropEquivAttr	proportion of the equivalent number of attributes
	NoiseSR	the noise signal ratio
	PropMV	the proportion of missing values
	PropExMV	proportion of number of examples with missing values
	StdDClass	a statistical measure that is the standard deviation of classes
Model based features	Tree width	the width of tree
	Tree height	the height of the tree
	NoNode	the number of nodes
	NoLeave	the number of leaves
	maxLevel	The maximum number of nodes at one level
	minLevel	The minimum number of nodes at one level
	meanLevel, devLevel	The mean and standard deviation of the number of nodes on levels.
	LongBranch, ShortBranc	The length of longest and shortest branches.
	meanBranch, Branch	The mean and standard deviation of the branch lengths.
	maxAtt, minAtt	The maximum and minimum occurrence of attributes.
Landmarking	Meta-learner errors	Meta-learner error rates for predicting nearest neighbor(k-NN), naive Bayes(NB), and boosted C5.0 suitability.
Data complexity measures	Measures of overlap in feature values from different classes	F1 Maximum Fisher's discriminate ratio
		F2 Overlap of the per-class bounding boxes
		F3 Maximum feature efficiency
		F4 Collective feature efficiency
	Class overlapping measure	L1 Minimized sum of the error distance of a linear classifier
		L2 Training error of a linear classifier
		N1 Fraction of points on the class boundary
		N2 Ratio of average intra/inter class nearest neighbor distance
		N3 Leave-one-out error rate of the one-nearest neighbor classifier
	Measures of geometry, topology, and density of manifolds	L3 Non-linearity of a linear classifier
		N4 Non-linearity of the one-nearest neighbor classifier
		T1 Fraction of maximum covering spheres
		T2 Average number of points per dimension