Evolving Systems (2017) 8:243–259
DOI 10.1007/s12530-016-9167-3

ORIGINAL PAPER

Evolving learners' behavior in data mining

Nitin Pise[1] · Parag Kulkarni[1]

Received: 29 September 2014 / Accepted: 1 September 2016 / Published online: 21 September 2016

Abstract An evaluation and choice of learning algorithms is a current research area in data mining, artificial intelligence and pattern recognition, etc. Supervised learning is one of the tasks most frequently used in data mining. There are several learning algorithms available in machine learning field and new algorithms are being added in machine learning literature. There is a need for selecting the best suitable learning algorithm for a given data. With the information explosion of different learning algorithms and the changing data scenarios, there is a need of smart learning system. The paper shows one approach where past experiences learned are used to suggest the best suitable learner using 3 meta-features namely simple, statistical and information theoretic features. The system tests 38 UCI benchmark datasets from various domains using nine classifiers from various categories. It is observed that for 29 datasets, i.e., 76 % of datasets, both the predicted and actual accuracies directly match. The proposed approach is found to be correct for algorithm selection of these datasets. New proposed equation of finding classifier accuracy based on meta-features is determined and validated. The study compares various supervised learning algorithms by performing tenfold cross-validation paired *t* test. The work helps in a critical step in data mining for selecting the suitable data mining algorithm.

Keyword Machine learning · Data mining techniques · Classification · Data characteristics · Learning algorithms · Intelligent data analysis

✉ Nitin Pise
nnpise@yahoo.com

[1] College of Engineering, Shivajinagar, Pune 411 005, India

1 Introduction

There is an immense amount of data available; a lot can be learned from this data. Learning manually is very time consuming. Many researchers have proposed methods to make machines learn from available data automatically. The purpose of learning in machine learning is to empower decision makers so that they can make better decisions. Similarly, machines should be empowered to make better decisions and improve their ability with value addition. In many real-life situations, the problem is not static. It can change with time and depend on the environment in which the problem is to be solved. The solution also can depend upon the decision context. The overall information is required to build the context.

Memorization and rudimentary learning are some of the examples of learning. The goal of learning is to help in better decision making. Learning gives intelligence and is centered on a goal.

There are three types of learning. Learning from a set of examples or historical data is supervised learning. It works on labelled data. This is a very common and the most frequently used forms of learning. It is mostly used for classification task in data mining. When labelled data is not available, it is difficult to use supervised learning. In these situations, learning without a teacher, i.e., unsupervised learning is used. Unsupervised learning is used for clustering applications. In practical situations, we need to learn from not only labelled data but also unlabeled data. This type of learning is called as semi-supervised learning.

Learning is a continuous process. Learning is not only knowledge acquision, but it involves different processes to gather knowledge, manage and augment knowledge. Learning needs prior knowledge. Learners can use the past

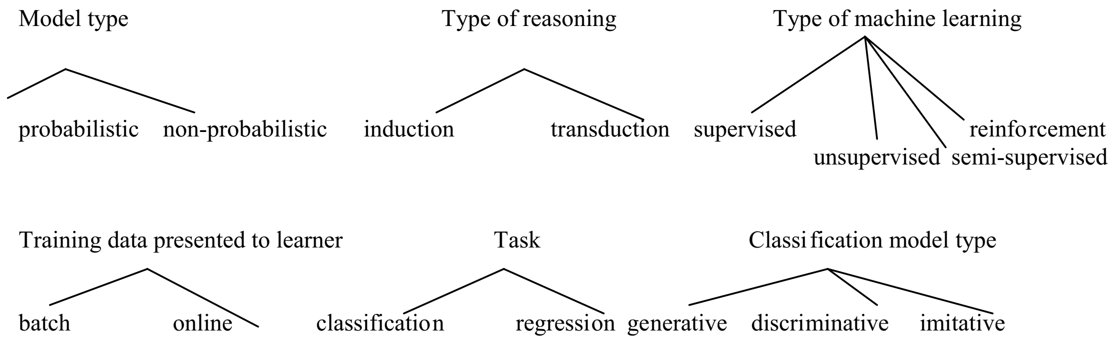


Fig. 1 Taxonomy of machine learning algorithms

knowledge to construct new understandings and make decisions on new data.

Different scenarios can be used to learn in supervised learning and expected outcomes can be further used as a learning sample. So if we are in a similar situation in the future, we can suggest the best possible decision available. This can be done if a new scenario is modelled to any of the previous scenarios.

Learning takes different forms, e.g., imitation, memorization, induction, deduction, inference, learning from examples and observation based learning, etc. There is another type of learning in which learning takes place based on the feedback. The feedback is in the form or reward or penalty. This is called reinforcement learning (Sutton and Barto 1998). In this type of learning, learners or software agents learn by interacting with the environment.

Learning can be based on data, different events and patterns. It can be system-based also. Adaptive machine learning algorithm (Kulkarni 2012) can be considered as a model where individuals need to respond and act in changing environments.

Imbalanced learning (Cai et al. 2014) is now a popular research topic for number of applications in data mining. Classification involving imbalanced class distributions poses a major problem in the performance of classification systems (Sun 2007). Many applications such as, network intrusion detection, fraud detection, medical diagnosis, etc., have to suffer due to the problem of imbalanced data. The paper assumes relatively balanced class distributions. So it doesn't depict different methods used to remove the effect of imbalance in data and special methods like SMOTE (Chawla et al. 2002).

The paper is prepared into six sections. After the introduction, Sect. 2 covers the background and related work on this topic. We discuss proposed work in detail in Sect. 3. Section 4 describes experimental set up, datasets used and empirical results found out. Section 5 depicts conclusions from the paper and how the work can be extended further. We conclude the paper listing contributions in the discussion section.

2 Background and related work

Kulkarni (2012) refers adaptive machine learning as learning with adapting to the environment, a learning task or a decision scenario. The learning can be based on past knowledge, experience from previous examples and expert advice. A particular method which is successful in one situation or for a specific task may not prove successful for all the learning types (Wolpert and Macready 1997). The learning process is closely associated with the learning problem. It also depends on what we are trying to learn and what are our learning goals. So, while selecting learning algorithms or methods, the problem is required to be understood. The learning problem needs to be analyzed and select the most suitable approach dynamically in adaptive learning. It is not just using more than one methods or moving from one method to other method. But it is selecting data intelligently and choosing the suitable learner.

Sewell (2009) explains a taxonomy of machine learning in which machine learning algorithms are categorized in six ways. Figure 1 shows taxonomy of machine learning algorithms. Model type decides whether the machine learing algorithm is probabilistic or non-probabilistic. The probabilistic model involves building a full or partial probability model. A discriminant or regression function is used in non-probabilistic model. Based on reasoning they can be classified as induction or transduction algorithms. Induction reasoning is learning from past training cases to general rules, that are further operated on the test cases.

Reasoning from observed, training examples to test examples is transduction reasoning.

Machine learning algorithms can be further categorized into batch or online depending upon how the learner receives training data. In batch learning, the learner is provided with all the data at the time of beginning. But this is not done in online learning. One example at a time is provided to the learner, which approximates the output, before receipt of the exact value in online learning. Each new example helps the learner for updating its current hypothesis and the total number of mistakes done during learning decides the quality of learning (Sewell 2009).

Depending upon the task which is to be carried out, machine learning algorithms are divided into classification or regression algorithms. Classification (Tan et al. 2013) is the assignment of objects to one of a number of existing classes. Classification is finding a function f mapping attribute set x to one of the existing classes y. It is a pervasive problem which encompasses diverse applications such as spam mail detection, analyzing MRI scans to categorize cells as malignant or benign, classifying millions of home loan applications into credit worthy and non credit worthy, and categorizing galaxies with their shapes, etc.

Regression (Tan et al. 2013) is a predictive modelling technique where the estimated target variable is continuous. Regression is learning a function f mapping attribute set x into a output y that is continuous-valued. Thus, regression finds a target function fitting the input data with minimum error. Examples of applications of regression include stock market prediction, projection of the total sales of a company by considering he amount spent for advertising, and so on.

Based on the classification model type, machine learning algorithms can be grouped into generative, discriminative or imitative algorithms (Kulkarni 2012). The class conditional density p(x | y) is modelled by some unsupervised learning procedure in generative algorithms (Chapelle et al. 2006). Bayes theorem (Mitchell 1997) is used to infer predictive density. Discriminative algorithms estimate p(y | x). Support vector machine (SVM) (Joachims 1999) is an example of discriminative algorithm.

There are different kinds of learning in machine learning. Four kinds of learning, namely supervised, unsupervised, semi-supervised learning and reinforcement learning are very important in machine learning. Supervised learning algorithms are classified by Hormozi et al. (2012). Supervised learning algorithms are compared empirically in (Caruana and Niculescu-Mizil 2006). Supervised learning algorithms are divided into the following methods:

- Decision trees
- Artificial Neural Networks
- Support Vector Machine (SVM) (Joachims 1999)
- Instance-based learners
- Bayesian networks
- Probably approximately Correct (PAC) learning (Valiant 1984)
- Inductive Logic Programming (ILP)
- Ensemble methods (Polikar 2006; Tan et al. 2013)

The back-propagation (BP) learning algorithm is a supervised neural network algorithm. It is used for multi-layered feed-forward neural architectures. (Curran et al. 2011) demonstrate that visual spectrum study and a back propagation neural network classifier can be used to discriminate the breadth patten in certain places of the body. Liu and Cao (2010) proposes application of recurrent neural network (RNN). They propose RNN to solve extended general variational inequalities based on the projection operator and a novel k-winners-take-all network (k-WTA) based on a one-neuron RNN. The advatntages of k-WTA are that it has a simple structure and finite time convergence. One more application of RNN that is based on the gradient method is proposed by Liu et al. (2010b) for solving linear programming problems. The proposed network globally converges to exact optimal solutions in finite time.

SVM (Joachims 1999) is a classification technique which is used by a number of researchers and is based on statistical learning theory. SVM is suitable for classifying high-dimensional data (Tan et al. 2013). Selecting the kernel function is probably the trickiest part in SVM. The kernel function is significant as it creates the kernel matrix, which plays the important role in summarizing all the data. Linear, polynomial and RBF kernels can be used in SVM.

Approaches to unsupervised learning are as follows:

- Clustering (Tan et al. 2013)
- Hidden Markov Models (HMMs)
- Principal Component Analysis (PCA)
- Independent Component Analysis
- Adaptive Resonance Theory (ART) (Yegnanarayana 2005)
- Singular Value Decomposition (SVD)
- Self Organizing Map (SOM) (Kohonen 2001)

Clustering techniques (Witten et al. 2005) are used when the instances are to be divided into natural groups. Clustering algorithms are classified as partitioning, hierarchical, density based and so on.

A detailed survey of semi-supervised learning algorithms is presented by Pise and Kulkarni (2008). Several semi-supervised classification algorithms are very popular which include Co-training, Expectation Maximization (EM) algorithm, transductive support vector machines (TSVMs). Self-training, graph based methods and multi-view learning are other important semi-supervised learning methods. Temporal Difference (TD) learning and Q-learning are important methods in Reinforcement learning which are explained in Kulkarni (2012).

There are number of algorithms in each of the category. Every algorithm cannot give better accuracy or does not perform well for all the datasets. When we want to work on a particular dataset, we have to evaluate a lot of algorithms for checking whether the particular algorithm is suitable for a given problem. This takes a lot of time by checking results of each of these algorithms on that particular

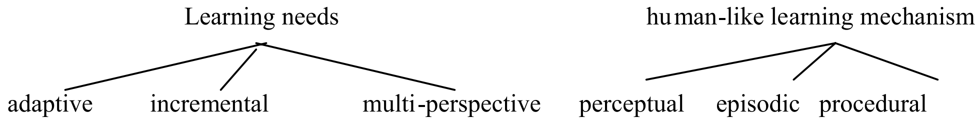


Fig. 2 Added taxonomy from Kulkarni (2012)

dataset. Instead of wasting the time required for evaluation of each of the algorithm on the particular dataset, we are using a classifier selection methodology based on dataset characteristics. Here three different data characteristics, namely simple, statistical and information theoretic measures are used. The focus of the paper is on supervised machine learning algorithms.

The added taxonomy is depicted in Fig. 2 (Kulkarni 2012). Based on the learning needs, machine learning algorithms are classified in adaptive, incremental or multi-perspective learning.

Adaptive machine learning (Kulkarni 2012) refers to learning that adapts with the environment or a learning problem. The learning uses the gathered information, experience, past knowledge, and expert advice. The learning process is closely associated with the learning problem or what are trying to learn. Hence the choice of learning methods demands an understanding of the learning problem. Adaptive learning involves the intelligent choice of the most appropriate method.

Incremental learning (IL) is proposed in Kulkarni (2012). The learning is done in stages, and during every stage the learning algorithm receives some new data for learning. So there is a need for incremental learning. IL effectively uses already created knowledge base during the next phase of learning and does not affect accuracy of decision making.

Multi-perspective learning (Kulkarni 2012) refers to learning that uses the knowledge and information acquired and is built from different perspectives. Multi-perspective decision making uses multi-perspective learning that includes methods for capturing perspectives and the captured data and knowledge perceived from different perspectives.

Further, they can be categorized perceptual, episodic or procedural based on human-like learning mechanism.

Ensemble learning (Polikar 2006) consists of more than one learner for the same problem. Kotsiantis et al. (2006) depicts a variety of classification algorithms and ensembles of classifiers that improve classifier accuracy. If we want to improve classification accuracy, it is hard to find a single classifier or a good committee of experts. There are advantages of ensemble methods, but they have three weakness. Ensembles require more storage because all classifiers need to be stored after training. The second weakness is increased computation; as all component classifiers must be processed. The third weakness is they are less comprehensible. As multiple learners or classifiers are involved in decision making, difficulty is faced by non-experts in perceiving the underlying reasoning process leading to decision.

An evaluation and selection of classification algorithm is a current research topic in data mining, artificial intelligence and pattern recognition (Kou and Wu 2014). Vilalta and Drissi (2002) reviews the different aspects of meta-learning. Meta-learning is learning about learners. It is learning at a meta-level. It works on the experience gathered from past data, i.e., it works on the past performance of different learners.

Smith-Miles (2008) presents the algorithm recomendation problem using meta-learning and explains its uses in classification, regression, prediction of time series, optimization and constraint satisfaction.

Several systems for algorithm selection are proposed in the literature. Sleenman and Rissakis (1995) presents an expert system called "Consultant" which finds out the characters of the application and the data. It questions users several times. This system does not test the data but considers the users' subjective experiences. Michie et al. (1994) describes STATLOG project in which various meta-features are extracted from registered datasets, and these meta-features are combined with the performance of the algorithms. Once a new dataset arrives, the system makes comparison of the meta-features of the new dataset and old datasets. This takes a lot of time. Alexandros and Melanie (2001) describes a system called Data Mining Advisor (DMA) having a set of algorithms and training datasets. K-NN algorithm (Cover and Hart 1967) is used to find a similar subset in the training dataset based on the performance of algoritms. It ranks the candidate algorithms and recommends based on the above subset.

The method which combines accuracy and execution time for comparing two algorithms' performance on the similar data set is called Adjusted Ratio of Ratios (ARR) is described in Brazdil and Soares (2000). Romero et al. (2013) applies meta-learning to recommend the best subset of classification algorithms for 32 Moodle datasets. Complexity, domain specific features and traditional statistical features are used in this paper. But the study is limited as educational datasets are used. Pinto et al. (2014) proposes a framework for decomposing and developing meta-features for meta-learning problems. Meta-features namely simple, statistical and information-theoretic metafeatures are decomposed using the framework.

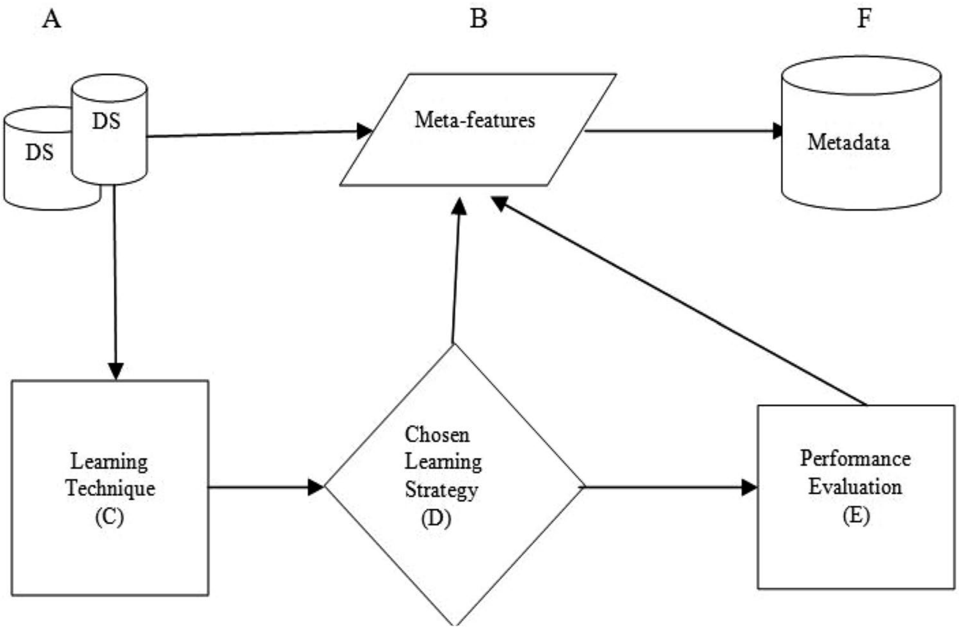


Fig. 3 Meta-learning: Knowledge acquisition adapted from Brazdil et al. (2008)

Fan and Lei (2006) explores a meta-learning approach that helps user to choose the most suitable algorithms. Selecting the suitable algorithm is crucial during data mining model building process.

Evolutionary computing (EC) is useful in fine-tuning hyper-parameters for the different learning algorithms. Genetic algorithms (GA), evolutionary programming (EP), etc. are important methodologies in EC. Oduguwa et al. (2005) bridges the gap between theory of EC and practice by taking case of manufacturing industry. Preitl et al. (2006) deal with not only theoretical but also application aspects concerning iterative feedback tuning (IFT) algorithms in the design of fuzzy control systems. Closed-loop data computes the likely gradient of the cost functions. IFT or other gradient-based search methods are useful for optimizing hyper-parameters of the learning algorithms. This helps to improve the learner's performance.

So we develop a method which uses supervised learning algorithms and ensembles. Based on the problem to be solved, the different learning algorithms are recommended by this system.

3 Proposed work

Our work is based on meta-learning (Brazdil et al. 2008). Meta-learning is learning about learners. Knowledge learned in previous experiences or experiments is used to manage new problems in a better way and is stored as metadata (F), particularly meta-features (B) and meta-target as shown in Fig. 3. The meta-features extracted from A to B depict the relation between the learners and the data used. The Meta-target is required to be extracted through C-D-E for further storing in F. The algorithm that works the best for a given dataset is represented by the meta-target.

Figure 4 shows the system flow for recommending suitable classifier. Data Characterization Tool (DCT) is implemented in Java for calculating dataset characteristics which are also referred as meta-features. New dataset characteristics are provided to k-NN algorithm and results are stored in the knowledge base that determines learning algorithm performance based on dataset characteristics. Similarity between historical datasets and a new dataset is used to recommend suitable algorithm.

How to define meta-features or data characteristics is the main issue in meta-learning. The state of the art shows that there are mainly three types of meta-features: (1) simple, statistical and information theoretic (Brazdil et al. 2003), (2) model-based (Peng et al. 2002), (3) landmarking (Pfahringer, Bensusan, and Giraud-Carrier 2000). In the first group we find out the number of instances, the number of attributes, kurtosis, skewness, correlation between numeric attributes or class entropy to name a few. Application of these meta-features provides knowledge about the problems. A model is generated by applying a learner to a problem or a dataset, i.e., the number of leaf nodes of a decision tree. Some characteristics of this model are captured by model-based meta-features. Land-marking meta-features are created by making a quick performance approximation of a learner in a particular dataset.

We have explored the first group of meta-features as shown in Table 1. Table 1 shows the various meta-features used and how they are denoted in the experimental work. Meta features numbered from 3–13 are simple

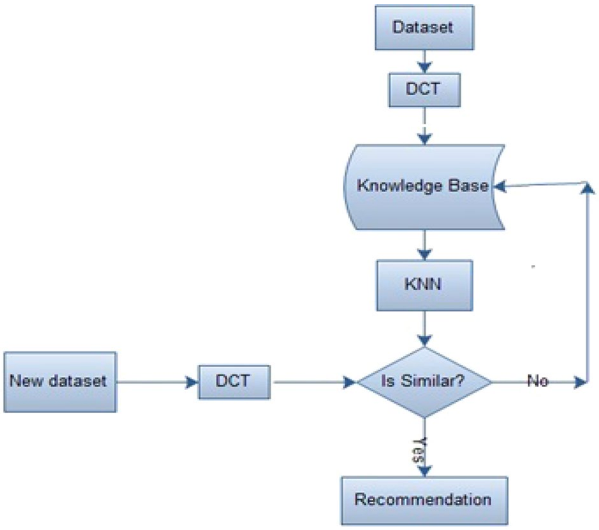


Fig. 4 System flow for recommending suitable classifier

Table 1 Meta- features used for experimentation, how they are referred as in database and corresponding database attribute

Ref. as	Meta-features	Database attribute
3	Number of attributes	Norm_attr
4	Number of instances	Norm_inst
5	Number of classes	Norm_class
6	Number of symbolic attributes	Norm_sym
7	Number of numeric attributes	Norm_num
8	Number of missing values	Norm_missing
9	Number of distinct values	Norm_dist
10	% Missing values	%missing
11	Dimension	Norm_dimension
12	Ratio of symbolic attributes	Norm_sym %
13	Ratio of numeric attributes	Norm_num %
14	Kurtosis	Kurtosis
15	Skewness	Skewness
16	Maximum probability	Maxprob
17	Entropy	Entropy

meta-features; kurtosis and skewness are statistical meta-features and entropy is information theoretic meta-feature.

Kurtosis is a measure of the flatness of the top of a symmetric distribution.

Distribution's degree of kurtosis,

$$\eta = \beta_2 - 3 \tag{1}$$

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

β_2 is often called "Pearson's kurtosis".

The third moment is used to define mean Skewness.

$$\gamma_1 = \frac{\sum(Y-\mu)^3}{n\sigma^3} \tag{2}$$

Skewness is negative if shape of distribution is skewed to the right.

Entropy is used for giving the amount of information in bits by a particular signal state.

$$\text{Entropy (S)} = -p_+ \text{log} p^+ - p_- \text{log} p^- \tag{3}$$

Where p^+ and p^- are positive and negative probabilities.

The proposed approach uses Euclidean distance which is computed as follows:

$$\text{d(p, q)} = \text{d(q, p)} = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}. \tag{4}$$

Initially the results were tested using other distance measures such as Manhattan distance, but the results were not satisfactory. Hence Euclidean distance measure is used in the system. An algorithm is recommended using similarity between a new dataset and historical datasets.

3.1 Methodology

The methodology used in this work consists of the following nine steps:

1. Datasets are collected from UCI machine learning repository.
2. Meta feature extraction is done using Data Characteristics Tool (DCT) for training.
3. A learning algorithm with performance measure is considered. Classification accuracy is used as a performance criterion.
4. Knowledge base is created by considering performance of learning algorithms and data characteristics or meta-features of datasets.
5. Extraction of meta-features is carried out for new unseen dataset using DCT.
6. Use k-NN for finding out k-similar datasets from the knowledge base.
7. Obtaining the algorithms for K-similar datasets.
8. Ranking of algorithms.

9. Algorithm recommendation which will help in decision making.

3.2 Learning algorithms or classifiers used

Learners or Classification algorithms (Nakamura et al. 2014; Han and Kamber 2011) are divided into several types such as function-based classifiers (e.g., Support Vector Machine (Joachims 1999) and neural network), tree-based classifiers (e.g., J48 (Quinlan 1993) and random forest (Leo 2001), distance-based classifiers (e.g., k-nearest neighbour (Cover and Hart 1967), and Bayesian classifiers. All available classifiers have advantages and disadvantages. For example, Support Vector Machine (SVM) (Joachims 1999) is a great classifier that gives the best performance for binary class problem. But it frequently performs poorly when applied to imbalanced datasets.

Mitchell (1997) describes cons of learning methods. Overfitting, caused by random noise in the training data is a significant practical difficulty in decision tree learning.

Ensembles are considered in our work due to their better accuracy over single classifiers (Polikar 2006). Experts consisting of a group of different classifiers offer corresponding information regarding the patterns which improves the efficacy of the overall classification method (Tiago et al. 2014). The tests are conducted on ten benchmark datasets from UCI machine leaning repository (Frank and Asuncion 2010) using ensemble techniques such as Bagging (Breiman 1996), Stacking (Dzeroski and Zenko 2004), AdaBoost (Polikar 2006), and LogitBoost (Friedman et al. 1998) which are available in WEKA (Mark et al. 2009). AdaBoost and LogitBoost represent two variations of boosting algorithm. LogitBoost is motivated by statistical view (Friedman et al. 1998). Figure 5 shows percentage classification accuracy of the above ensemble techniques. Stacking performs poorly in comparison with the rest of ensemble techniques used. So AdaBoost, Bagging and LogitBoost are considered further for selecting the suitable algorithm.

Table 2 shows the different classifiers, their categories and abbreviations used in the experimental study.

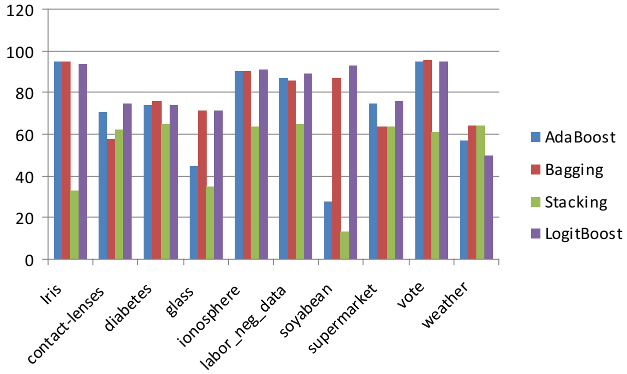


Fig. 5 Percentage classification accuracies of various ensembles used

Table 2 Classifiers from diffent types used in experimentation

Sr.	Classifier	Type	Abbrev.
1	Naïve Bayes	Bayes	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

3.3 Algorithm

The following algorithm shows steps in our approach.

Inputs:
K: the number of neighbours
d: data characteristics of new dataset
DC: data characteristics of historical datasets
Output:
Neighbours: the neighbour dataset for new dataset
Alg []: Set of algorithms
Algorithm:

```
i = 1;
For each D ∈ DC do
    Distance Table[i] = the distance between d and D i.e. |d-D|
    i = i+1;
Ordering Distance Table in ascending order
Neighbours = top K datasets of distance Table
j = 0;
For each j < K do
    Alg[j] = $D_j$'s Best Algorithm.
```

4 Experimental study and results

4.1 Classification measures

Confusion Matrix (Han and Kamber 2011) is used for analyzing performance of classifier. i.e., it indicates how accurate classification is performed. Table 3 shows confusion matrix that contains information about actual and predicted classifications for a classifier system.

Accuracy Accuracy indicates how a measured value is close to its actual or true value.

Table 3 Confusion matrix

Predicted class			
Actual class	Yes	No	Total
Yes	TP	FN	P
No	FP	TN	N
Total	P′	N′	P+N

TP true positive, *FN* false negative, *FP* false positive, *TN* true negative

Precision Precision indicates how close the measured values are to each other.

Recall Recall is used to measure fraction of relevant instances that are retrieved.

Evaluation measures are calculated as shown in Table 4 by using confusion matrix from Table 3. Accuracy, error rate, precison and recall are the commonly used evaluation measures in classification. Out of the above classification measures, the focus of this paper is on accuracy.

Learner or classifier algorithm selection is a multi-decision optimization problem. It is part of multi-objective model type selection problem (Rosales-Perez et al. 2014). Model selection involves both the selection of learning algorithms and choice of hyper-parameters for a given algorithm. Fine tuning of hyper-parameters can affect the generalization capability of learning algorithms. More than one classifier measures can play an important role in learner recommendation as the problem changes. EI-Hefnawy (2014) suggests a modified particle swarm optimizer (MPSO) to solve fuzzy bi-level single and multi-objective problems. In this approach the bi-level programming problem (BLPP) handles as fuzzy multi-objective problem. The present work has restriction that only classification accuracy is used. But this limitation will be removed in the extension of work, where the authors are working with other classification measures such as classifier testing time, complexity and comprehensibility of learning algorithm.

4.2 Datasets

Saitta and Neri (1998) have shown that supervised learning algorithms are used in various application domains. For the purpose of the present study, we have used 38 benchmark data sets from the University of California at Irvine Machine Learning Repository (Frank and Asuncion 2010). These datasets are from: medical diagnosis (breast-cancer, hypothyroid, etc.), pattern recognition (anneal, iris, etc.), image recognition (ionosphere, segment, etc.), commodity trading (credit-a, labor, etc.) and various control applications (balance).

Table 4 Evaluation measures

Measure	Formula
Accuracy, recognition rate	(TP + TN)/(P + N)
Error rate, misclassification rate	(FP + FN)/(P + N)
Sensitivity, true positive rate	TP/P
Specificity, true negative rate	TN/N
Precision	TP/(TP + FP)

Table 5 shows the datasets with the important data characteristics or meta-features. The datasets used in the experimental work are having instances of 10 to 3772. The number of classes for them varies from 2 to 24. The number of symbolic attributes varies from 0 to 69. The number of numeric attributes varies from 0 to 60. That means the total number of features or dimensionality of the datasets is from 4 to 69. Entropy in Table 5 is calculated using Eq. 3. Entropy is an information theoretic meta-feature.

4.3 Experimental set-up

The system is developed in Java language. All 38 datasets are having attribute relation file format (ARFF). Meta-features in Table 1 are calculated. We have used nine classifiers provided by Weka (Witten et al. 2005) as shown in Table 2. Weka is a set of machine learning algorithms used for data mining. In this work, the parameters of all the classifiers are kept as default.

4.4 Accuracy evaluation and paired *t* test

Table 6 describes the results evolved by the learner recommendation system. Actual accuracy is calculated using Weka (Witten et al. 2005). Pred1 accur. is the first predicted accuracy by the system. Similarly the second and the third recommendations are denoted as Pred2 accur. and Pred3 accur. Finally the best recommended accuracy is decided and its respective classifier is recommended. Difference is the term used to denote the difference between actual and recommended best classifier. It is used for further analysis.

Cross-validation (Hall et al. 2004) is "a model validation technique for evaluating how the results of a statistical analysis will generalize to an independent dataset". It is used in prediction and to assess how a predictive model works in practice.

Cross-validation involves partitioning of original dataset into training and testing datasets. A model developed in training phase is validated using testing dataset. There are different types of cross-validation, such as exhaustive cross-validation, leave-p-out cross-validation, k-fold cross validation, etc. K-fold cross-validation is used in the experimental work. Here k is 10. So it is called tenfold cross-validation (Hall et al. 2004). It works as follows:

- Data is divided into 10 sets of size n/10.

Table 5 Thirty eight datasets with different meta-features and best classifier

DATASET		Best classifier	No attr	No-instance	No_of_classes	No_symbolic attributes	No_numeric attributes	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	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	Lymphotherapy	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	Toplayornotoplay	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

Table 5 continued

	DATASET	Best classifier	No attr	No-instance	No_of_classes	No_symbolic attributes	No_numeric attributes	Missing_value	No_distinct	Entropy
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

- 9 datasets are used for training and testing is done on 1.
- The above process is repeated 10 times and a mean accuracy is taken.

Cross-validation is used to compare the performances of the different predictive modelling performances. Using cross-validation, the two different learners or classifiers are compared objectively. The performance measure used in the empirical work for the comparison is the classification accuracy.

A general method used for comparing supervised learning algorithms involves carrying out statistical comparisons of the accuracies of trained classifiers on specific datasets (Bouckaert 2003). Dietterich (1998) and Nadeau and Bengio (2003) explain several versions of the *t* test to solve this problem. So we use tenfold cross-validation paired *t* test for comparing the classifiers. Microsoft Excel is used to calculate paired *t* test.

Table 7 shows classifier accuracies of Bagging classifier, SMO classifier and their differences for 38 benchmark datasets from UCI machine learning repository. This table is used to calculate results as shown in Tables 8 and 9 for paired *t* test for statistical comparison of the classifiers.

A paired *t* test was performed to determine if the classifiers' accuracy was significantly different.

The mean of difference of classifiers' accuracy (M = −4.45, SD = 13.38, N = 38) was significantly less than zero, t(37) = −1.82, two-tail p = 0.047, providing evidence that the two classifiers are differing in accuracy. A 95 % C. I. about mean accuracy is (−6.88, 0.37). The above sample results are shown for comparing two classifiers namely Bagging & Sequential Minimal Optimization (SMO) classifier. SMO is the classifier available in WEKA (Mark et al. 2009) for support vector machine.

Figure 6 shows difference between actual and best of predicted 3 classifiers. Figure 7 describes the difference between actual and predicted best. As shown in Figs. 6 and 7, datasets 23 and 25 show more difference between actual and predicted accuracy and all others have almost best prediction.

The dataset 23 is Iris which is from the pattern recognition domain. The dataset 25 is Lympotherapy which is from medical diagnosis domain. They are having 150 and 148 number of instances, respectively. Also, they have 3 and 4 classes, respectively. These may be the reasons of more difference in predicted accuracies and actual accuracies for datasets 23 and 25. As shown in Fig. 7, there are very few datasets where difference is significant.

As shown in Fig. 8, the number of objects on the line and below line indicate predicted accuracy is equal or greater than actual accuracy. Many of the objects are on or close to line so our approach recommends algorithm almost accurately.

Table 6 Evaluation database consisting of actual and predicted accuracies

Sr. no	Dataset	Actual accuracy	Pred1 accur.	Pred2 accur.	Pred3 accur.	Best predicted accur.	Difference
1	Anneal	99.77	97.46	98.21	97.46	98.21	1.55
2	Audiology	84.95	80.77	84.95	78.76	84.95	0
3	Balance-scale	90.40	87.57	87.57	86.56	87.57	2.83
4	Breast-cancer	75.52	71.67	67.52	67.52	71.67	3.84
5	Breast-w	96.75	96.75	95.13	95.99	96.75	0
6	Bridges_version1	67.61	65.71	46.66	66.69	66.69	0.91
7	Bridges_version2	67.61	67.61	51.42	66.94	67.61	0
8	Car	95.77	93.62	93.46	93.62	93.62	2.14
9	Colic	85.16	82.66	85.16	82.66	85.16	0
10	Credit-a	85.72	85.57	84.88	84.88	85.57	0.15
11	Credit-g	75.40	72.5	70.80	70.50	72.5	2.90
12	Cylinder-bands	81.41	58.51	81.41	73.14	81.41	0
13	Diabetes	76.80	76.8	70.18	76.30	76.8	0
14	Ecoli	85.71	83.63	85.41	85.71	85.71	0
15	Flags	62.88	59.36	35.58	58.24	59.36	3.52
16	Glass	72.89	48.59	71.49	71.49	71.49	1.40
17	Heart-c	83.86	83.49	83.86	77.55	83.86	0
18	Heart-h	85.03	82.74	82.74	80.95	82.74	2.29
19	Heart-statlog	83.89	83.70	83.89	76.66	83.89	0
20	Hepatitis	85.77	85.77	83.87	83.87	85.77	0
21	Hypothyroid	99.57	99.57	99.36	95.28	99.57	0
22	Ionosphere	91.73	88.07	91.45	91.45	91.45	0.28
23	Iris	96.27	73.33	86.66	80.00	86.66	9.60
24	Labor	92.97	92.97	92.97	87.89	92.97	0
25	Lymphotherapy	96.82	86.48	86.48	77.02	86.48	10.34
26	Segment	97.66	97.14	97.66	80.21	97.66	0
27	Sick	98.80	98.80	98.30	92.60	98.8	0
28	Sonar	86.53	76.60	76.60	80.28	80.28	6.25
29	Soyabean	93.10	91.21	92.97	92.97	92.97	0.12
30	Sponge	95.63	95.63	93.42	92.10	95.63	0
31	Tic-tac-toe	98.95	98.33	98.33	72.12	98.33	0.61
32	Toplayornottoplay	71.42	50.00	57.14	66.00	66	5.41
33	Trains	90.00	90.00	70.00	90.00	90	0
34	Vehicle	77.06	44.79	71.74	72.45	72.45	4.61
35	Vote	96.32	95.77	90.11	95.77	95.77	0.55
36	Vowel	99.29	71.31	70.61	96.66	96.66	2.62
37	Weather	78.57	78.57	35.71	54.00	78.57	0
38	Zoo	96.08	95.04	95.04	89.10	95.04	1.04

4.5 Results of KNN approach

We have done many tests using different combinations of data characteristics which are listed as below:

1. KNN where K = 1 with different groupings
2. KNN where K = 3 with different groupings of meta-features
3. KNN where K = 3 with different groupings of meta-features and normalized values

In Fig. 9, first 8 entries are using K = 1 without normalization. First best 3 classifiers of the same dataset that are found similar for a new dataset are used for evaluation. 9 to 13 entries in the Fig. 9 are for K = 3 without normalization and 14 onwards entries are for K = 3 with normalization.

Table 7 Classifier accuracies of Bagging classifier, SMO classifier and their differences

BA	SMO	Difference
98.83	97.46	1.37
76.43	80.77	−4.34
83.40	87.57	−4.17
69.59	67.52	2.07
95.79	96.75	−0.96
41.95	66.69	−24.74
41.95	66.94	−24.99
91.98	93.62	−1.64
84.88	82.66	2.22
85.28	84.88	0.40
73.97	75.09	−1.12
60.80	81.41	−20.61
76.08	76.8	−0.72
83.58	83.48	0.10
35.58	59.36	−23.78
72.22	57.36	14.86
80.52	83.86	−3.34
79.86	82.74	−2.88
81.04	83.89	−2.85
80.98	85.77	−4.79
99.54	93.58	5.96
91.37	88.07	3.30
94.33	96.27	−1.94
82.66	92.97	−10.31
96.82	86.48	10.34
96.61	92.92	3.69
98.73	93.87	4.86
76.27	76.6	−0.33
87.00	93.1	−6.1
92.50	95.63	−3.13
91.33	98.33	−7.00
60.00	66.00	−6.00
36.00	70.00	−34
72.66	74.08	−1.42
95.74	95.77	−0.03
86.96	70.61	16.35
64.00	54.00	10.00
42.49	96.08	−53.59

Figure 10 shows different combinations of meta-features with normalization and predicted 1st best classifiers difference over 38 datasets. It is observed that combination of normalized meta-features NV-3-4-5-16-17 gives better recommendation; so we use this combination for recommendation of classifier.

Table 8 *T* test paired two samples for means

	Variable 1	Variable 2
Mean	77.887	82.34158
Variance	336.667	153.0374
Observations	38	38
Pearson Correlation	0.684	
Hypothesized Mean Difference	0	
df	37	
t Stat	**−2.05165**	
P(T <=t) one-tail	0.023666	
t Critical one-tail	1.687094	
P(T <=t) two-tail	**0.047332**	
t Critical two-tail	2.026192	

Prefix NV is used before different combinations of meta-features in the Figs. 9 and 10. It shows that normalized values of the meta-features are used. Min–max normalization is used for reducing impact of large values of some meta-features on accuracy of recommendation. E.g. There are two datasets namely Hypothyroid and Sick in UCI machine learning repository (Frank and Asuncion 2010) 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 value is required that helps to increase accuracy in recommendation. We have normalized values in the range of 0-1.

Thus the meta-features namely, the number of attributes, the number of instances, the number of classes, the maximum probability of class and the class entropy play a significant role in classifier accuracy and algorithm selection for 38 datasets and 9 classifiers used in our research work.

Locally weighted regression (Alpaydin 2010) is used to find the effect of the above meta-features on classifier accuracy. Locally weighted regression (Cleveland and Devlin 1988) is "a way of estimating a regression surface through a multivariate smoothing procedure, fitting a function of the independent variables locally and in a moving fashion analogous to how a moving average is computed for a time series". Figure 11 shows actual accuracy, predicted accuracy and the accuracy calculated using linear weighted regression as per the approximation function in Eq. 5 for 38 datasets.

$$\begin{aligned}\text{Accuracy} = {} & 90.84153 + 0.00436 \times \text{number of instances} \\ & + 0.4331 \times \text{number of classes} \\ & + 0.024926 \times \text{number of attributes} \\ & - 3.68516 \times \text{entropy} \\ & - 6.14672 \times \text{maximum probability of class}\end{aligned} \tag{5}$$

Table 9 Statistical results for comparison of two classifiers

Mean	−4.45421
Standard error	2.171043
Median	−1.53
Mode	#N/A
Standard deviation	13.38321
Sample variance	179.1103
Kurtosis	4.271084
Skewness	−1.74155
Range	69.94
Minimum	−53.59
Maximum	16.35
Sum	−169.26
Count	38
Confidence level (95.0 %)	4.398951

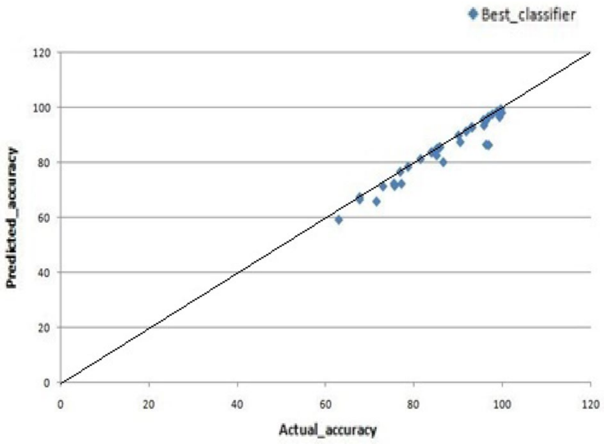


Fig. 8 Actual Vs best predicted classifier accuracy

Fig. 6 Dataset Vs actual and the first three predicted classifiers accuracy

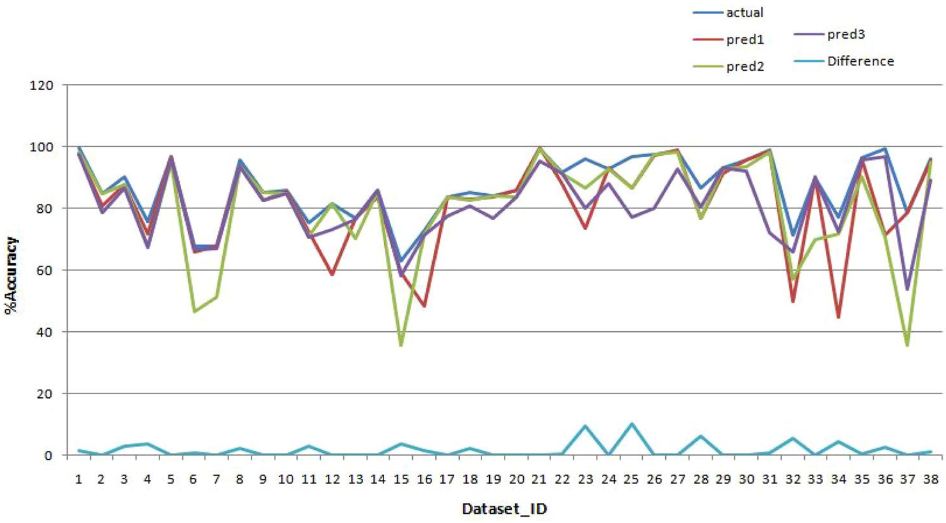
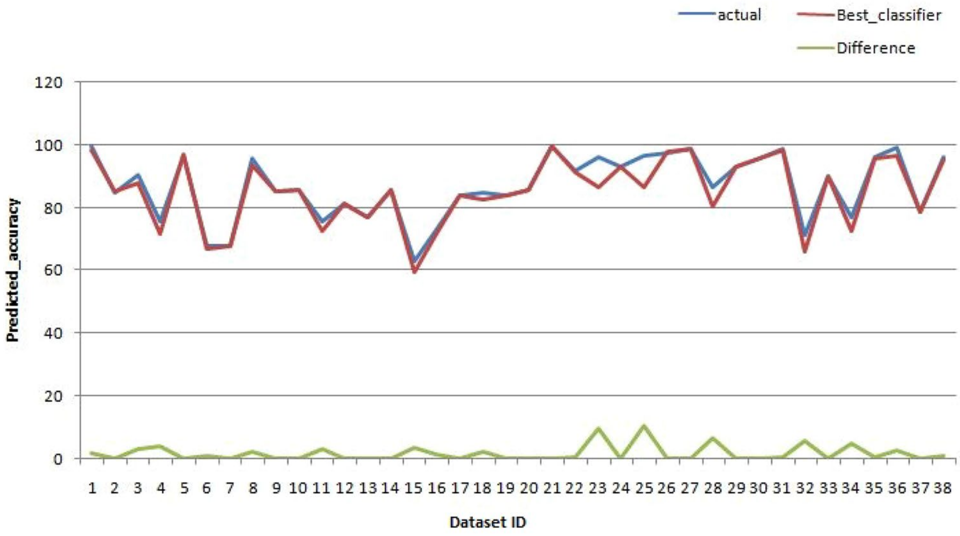


Fig. 7 Dataset Vs actual, the best predicted classifiers accuracy and difference of their accuracies



If we find out difference of accuracy using regression and actual accuracy for each dataset and take the average of 38 datasets it comes as −0.0016. This shows that our method correctly predicts the best accuracy on a new dataset.

5 Conclusion and future work

The paper proposes algorithm selection for classification problems in data mining. We find out meta-features of datasets and the performance of classifiers. K-similar datasets are returned based on the similiarity between a new dataset and historical datasets. We recommend the best classification algorithm for the given problem. Hence the user saves time for testing different learning algorithms, fine tuning the parameters for different algorithms.

k plays a significant role in the performance of the k-NN algorithm. In our system, k used is dynamic, so a non-expert can choose k.

Our algorithm selection method selects the best classifier based on accuracy as a performance measure with aim for helping non-experts in selecting algorithm. The experiment shows that predicted and actual accuracies match closely for 76 % of 38 benchmark datasets.

Three different categories of meta-features or data characteristics such as simple, statistical, and information theoretic are used and comparatively evaluated. Experiments on meta-features suggest the essential features such as the number of attributes,the number of instances, the number of classes, maximum probability of class, class entropy for classifier selection. These meta-features play a significant role in classifier accuracy and recommendation of the best algorithm for 38 benchmark datasets and 9 different classifiers used in our empirical work.

The empirical work uses 38 datasets from UCI machine learning repository. But still there is a need to extend the work to include more number of datasets as well as more number of algorithms. The framework which can suggest more number of meta-features is also required. The authors are extending the work to include another type of meta-features called as landmarking features to improve predictive accuracy. Rough sets is helpful for removing redundant meta-features and reducing time required for computation of meta-features. Also there is a demand for research in optimization of the parameters used for the different classifiers. There is necessity of more research for fine tuning of hyper-parameters of the different learners. Genetic algorithms and grid search techniques can play a major role in the above optimization.

Classification algorithms or learners may perform differently according to the context. Changing classification algorithm with dynamic interpretation of sensor data will be need of hour. The algorithm's quality in terms of accuracy and elapsed time can be enhanced by using the context-aware selection of classifier (Kwon and Sim 2013). Context-aware selection of classification algorithms is an important topic for pursuing further research. This method also provides core logics to expert systems which consider the characteristics of original dataset and current context for selecting optimal classification algorithms.

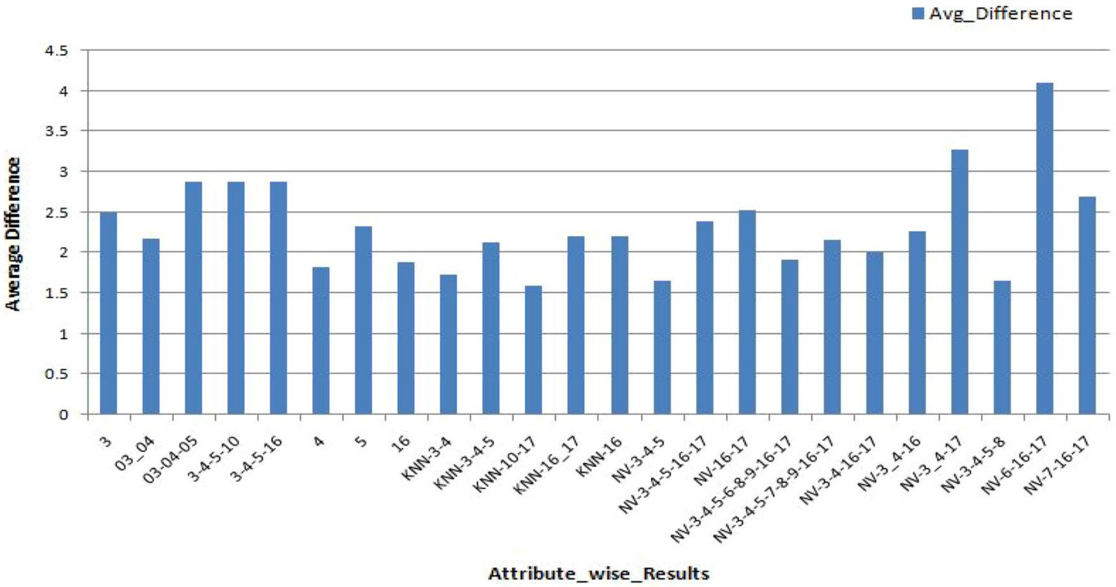


Fig. 9 Combination of meta-features Vs average difference

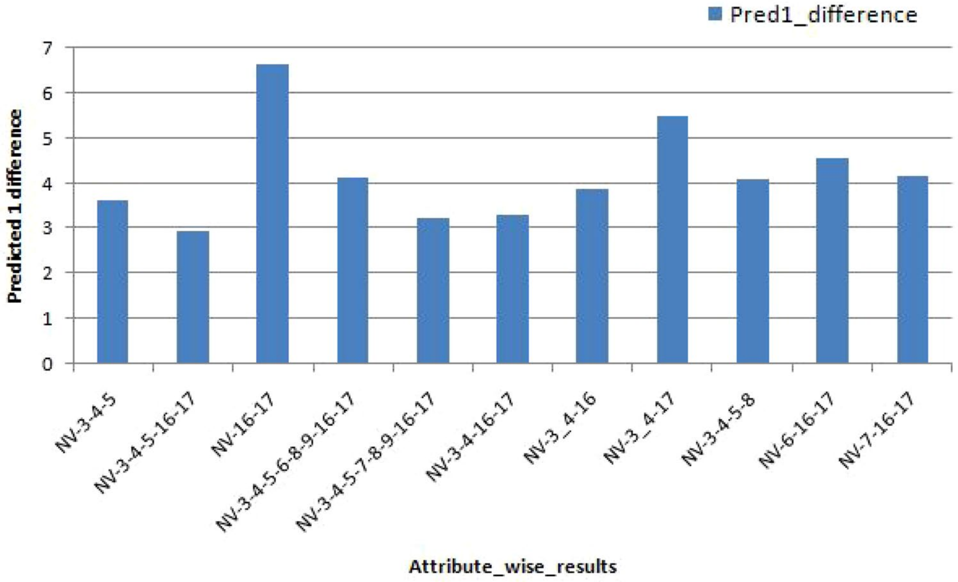


Fig. 10 Combination of normalized meta-features Vs average difference

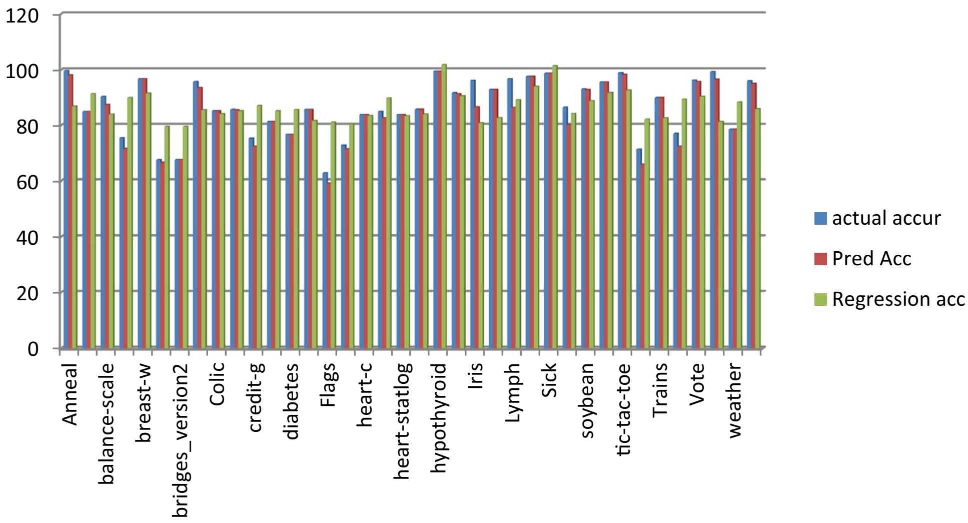


Fig. 11 Comparison of different accuracies on 38 datasets

This study does not consider data streams. So there is a scope for research in making decisions for extracting meta-features for dynamically changing data. Also intelligent data analysis of Big data in a business intelligence era and formulation of a meta-learning framework in the context of Big data are very hot research topics and require more research efforts.

6 Discussion

Our proposed work shows one approach for algorithm selection in data mining using meta-learning.

The major contributions of the work can be listed as below:

- The Data Characterization Tool for extracting simple, statistical and information-theoretic features is developed.
- Experiments are performed on 38 benchmark datasets, nine classifies are used from different types and three data characterization methods or meta-features are used for experimentation. The experimental work shows that for 76 % datasets, the predicted and actual accuracies closely match. Hence the algorithm selection or recommendation is correct for these datasets. One approach for adaptive learning is proposed and implemented.
- Min–max normalization is used for reducing impact of large values on accuracy of recommendation. Two datasets from UCI machine learning repository (Frank and Asuncion 2010) namely, hypothyroid and sick with

3772 instances each are used to explain the need of normalization. Meta-learning approach has not used normalization that improves accuracy of recommendation. But our approach uses normalization.

- Our work shows that the number of attributes, the number of instances, number of classes, maximum probabilityof class and class entropy are the major data characteristics or meta-features impacting classifier accuracy and algorithm selection. The average error of 38 datasets which is calculated using difference between accuracy from regression and actual accuracy is −0.0016. So it shows that our prediction about the above five meta-features affecting classification accuracy as a performance measure of learner is correct. We have contributed in formulating Eq. 5 which gives approximation function for modelling the classifier accuracy.

Thus new proposed equation of finding classifier accuracy based on meta-features is formulated and validated.

Acknowledgments The Authors wish to thank the Editors and the anonymous Reviewers for their detailed comments and suggestions which significantly contributed to the improvement of the manuscript. The authors acknowledge support and help by Suhas Gore, P.G. student during the work.

References

Alexandros K, Melanie H (2001) Model selection via meta learning: a comparitive study. Int J Artif Intell Tools 10(4):525–554

Alpaydin E (2010) Introduction to machine learning. PHI learning, New Delhi

Bouckaert R (2003) Choosing between two learning algorithms on calibrated tests. In: Proceedings of 20th international conference on machine learning. Morgan Kaufmann, pp 51–58

Brazdil P, Soares C (2000) A comparison of ranking methods for classification algorithm selection. In: de Mantaras R, Plaza E (eds) Machine learning: proceedings of the 11th European conference on machine learning ECML2000. Springer, Berlin, pp 63–74

Brazdil P, Soares C, Da Costa J (2003) Ranking learning algorithms: using ibl and meta-learning on accuracy and time results. Mach Learn 50(3):251–277

Brazdil P, Giraud Carrier C, Soares C, Vilalta R (2008) Metalearning: applications to data mining. Springer, Berlin

Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140

Cai Q, He H, Man H (2014) Imbalanced evolving self-organizing learning. Neurocomputing 133:258–270

Caruana R, Niculescu-Mizil A (2006) An Empirical comparison of supervised learning algorithms. In: Proceedings of the 23rd International conference on machine learning (ICML2006), pp 161–168

Chapelle O, Scholkopf B, Zien A (2006) Semi-Supervised Learning. MIT Press, Cambridge

Chawla N, Bowyer K, Hall L, Kegelmeyer W (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357

Cleveland W, Devlin S (1988) Locally weighted regression: an approach to regression analysis by local fitting. J Am Stat Assoc 403:596–610

Cover T, Hart P (1967) Nearest neighbor pattern classification. IEEE Trans Inf Theory 13(1):21–27

Curran K, Yuan P, Coyle D (2011) Using acoustic sensors to discriminate between nasal and mouth breathing. Int J Bioinform Res Appl 7(4):382–396

de Tiago PF, da Silva AJ, Ludermir TB, de Oliveira WR (2014) An automatic methodology for construction of multi-classifier systems based on the combination of selection and fusion. Prog Artif Intell 2:205–215

Dietterich TG (1998) Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput 10(7):1895–1924

Dzeroski S, Zenko B (2004) Is combining classifiers with stacking better than selecting the best one? Mach Learn 54:255–273

El-Hefnawy N (2014) Solving bi-level problems using modified particle swarm optimization algorithm. Int J Artif Intell 12(2):88–101

Fan L, Lei M (2006) Reducing cognitive overload by meta-learning assisted algorithm selection. In: Proceedings of 5th IEEE international conference on cognitive informatics, pp 120–125

Frank A, Asuncion A (2010) UCI machine learning repository (online). http://archive.ics.uci.edu/ml. Accessed 4 Aug 2012

Friedman J, Hastie T, Tibshirani R (1998) Additive logistic regression: a statistical view of boosting. Ann Stat 28(2):337–407

Hall P, Racine J, LI QL (2004) Cross-validation and the estimation of conditional probability densities. J Am Stat Assoc 99(468):1015–1026

Han J, Kamber M (2011) Data mining concepts and techniques. Morgan Kaufman Publishers, San Francisco

Hormozi H, Hormozi E, Nohooji HR (2012) The classification of the applicable machine learning methods in robots manipulators. Int J Machine Learn and Comput 2(5):560–563

Joachims T (1999) Making large-scale svm learning practical advances in kernel methods. In: Schölkopf B, Burges C, Smola A (eds) Support vector learning. MIT Press, Cambridge

Kohonen T (2001) Self-organizing maps. Springer, Berlin

Kotsiantis S, Zaharakis I, Pintelas P (2006) Machine learning: a review of classification and combining techniques. Artif Intell Rev 26:159–190

Kou G, Wu W (2014) An analytic hierarchy model for classification algorithms selection in credit risk analysis. Math probl Eng 2014:1–7. doi:10.1155/2014/297563

Kulkarni P (2012) Reinforcement and systemic machine learning for decision making, IEEE press series on systems science and engineering. Wiley, New Jersey

Kwon O, Sim JM (2013) Effects of data set features on the performances of classification algorithms. Expert Syst Appl 40:1847–1857

Leo B (2001) Random forests. Machine Learn 45(1):5–32

Liu Q, Cao J (2010) A recurrent neural network based on projection operator for extended general variational inqualities. IEEE Trans Syst Man Cybern-Part B Cybern 40(3):928–938

Liu Q, Dang C, Cao J (2010a) A novel recurrent neural network with one neuron and finite-time convergence for kwinners-take-all operation. IEEE Transactions on neural networks 21(7):1140–1148

Liu Q, Cao J, Chen G (2010b) A novel recurrent neural network with finite-time convergence for linear programming Neural Comput. 22(11):2962–2978

Mark H, Eibe F, Geoffrey H, Bernhard P, Peter R, Ian H (2009) The WEKA data mining software: an update. SIGKDD Explor 11(1):10–18

Michie D, Spiegelhalter DJ, Taylor CC (1994) Machine learning, neural and statistical classification. Ellis Horwood Series in Artifcial Intelligence. Ellis Horwood, Chichester

Mitchell T (1997) Machine learning. Burr Ridge, Mcgraw Hill

Nadeau C, Bengio Y (2003) Inference for the generalization error. Mach Learn 52:239–281

Nakamura M, Otsuka A, Kimura H (2014) Automatic selection of classification algorithms for non-experts using meta-features. China-USA Business Review. 13(3):199–205

Oduguwa V, Tiwari A, Roy R (2005) Evolutionary computing in manufacturing industry: an overview of recent applications. Applied soft computing 5(3):281–299

Peng W, Flach PA, Soares C, Brazdil P (2002) Improved data set characterisation for meta-learning. In: proceedings of the fifth international confernce on discovery science, LNAI 2534, pp 141–152

Pfahringer B, Bensusan H, 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

Pinto F, Soares C, Mendes-Moreira (2014) A framework to decompose and develop meta features. In: Proceedings of Meta-learning and algorithm selection workshop at 21st European conference on artificial intelligence, Prague, Czech Republic, 32–36

Pise N, Kulkarni P (2008) A survey of semi-supervised learning methods. In: Proceedings of international conference on computational intelligence and security, Suzhou, China, pp 30–34

Polikar R (2006) Ensemble based system in decision making. IEEE Circuit Syst Mag 6(3):21–45

Preitl S, Precup R, Fodor J, Bede B (2006) Iterative feedback tuning in fuzzy control systems. Theory Appl Acta Polytech Hung 3(3):81–96

Quinlan J (1993) C45 programs for machine learning. Morgan Kaufmann Publishers, San Francisco

Romero C, Olmo JL, Ventura S (2013) A meta-learning approach for recommending a subset of white-box classification algorithms for Moodle datasets. In: Proceedings of 6th international conference on educational data mining, Memphis, TN, USA, 268–271

Rosales-Pérez A, Gonzalez JA, Coello CAC, Escalante HJ, Reyes-Garcia CA (2014) Multi-objective model type selection. Neurocomputing 146:83–94. doi:10.1016/j.neucom.2014.05.077

Saitta L, Neri F (1998) Learning in the 'Real World'. Mach Learn 30(2–3):133–163

Sewell M (2009) Machine Learning, http://machinelearningmartin-sewell.com/machine-learning.pdf. Accessed 18 Sept 2014

Sleenman D, Rissakis M (1995) Consulatant-2: pre and post-processing of machine learning applications. Int J Hum Comput Stud 43(1):43–63

Smith-Miles K (2008) Cross-disciplinary perspectives on meta-learning for algorithm selection. ACM Comput Surv 4(1):6–25

Sun Y (2007) Cost-sensitive boosting for classification of imbalanced data. PhD thesis, department of electrical and computer engineering, University of Waterloo, Ontario, Canada

Sutton R, Barto A (1998) Reinforcement learning: an introduction. MIT Press, Cambridge

Tan P, Steinbach M, Kumar V (2013) Introduction to data mining, 2nd edn. Addison-Wesley, pp 792

Valiant LG (1984) A theory of the learnable. Commun ACM 27(11):1134–1142

Vilalta R, Drissi Y (2002) A perspective view and survey of meta-learning. J Artif Intell Rev 18(2):77–95

Witten IH, Frank E, Hall M (2005) Data mining: practical machine learning tools and techniques. Morgan Kaufmann series in data management systems, Morgan Kaufmann Publishers, CA

Wolpert D, Macready W (1997) No free lunch theorems for optimization. IEEE Trans Evolut Comput 1(1):67–82

Yegnanarayana B (2005) Artificial neural networks. New Delhi, PHI