International Journal of Advanced Information Science and Technology (IJAIST) ISSN: 2319:2682
Vol.1, No.8, December 2012 DOI:10.15693/ijaist/2012.v1i8.1-6

An Improved learning strategy to classify complex and high dimensional datasets

BhanuPrakashBattula*,Dr. R. Satya Prasad**

**Assoc.Professor, Dept. of Computer Science & Engineering, AcharyaNagarjuna University, Guntur, A.P. India.
*Research Scholar, Dept. of Computer Science & Engineering, AcharyaNagarjuna University, Guntur, A.P., India.

***Abstract*—Most of the existing classification techniques concentrate on learning the datasets as a single similar unit, in spite of so many differentiating attributes and complexities involved. However, traditional classification techniques, require to analysis the dataset prior to learning and for not doing so they loss their performance in terms of accuracy and AUC. To this end, many of the machine learning problems can be very easily solved just by careful observing human learning and training nature and then mimic the same in the machine learning. To solve this dilemma, we propose a novel, simple and effective machine learning paradigm that explicitly exploits this important similar-to-different (S2D) human learning strategy, and implement it based on C4.5 efficiently. The framework not only analyzethe datasets prior to implementation, but also carefully allows classifier to have a systematic study so as to mimic the human training technique designed for efficient learning. Experimental results show that the method outperforms the state of art method (C4.5) in terms of learning capability and breaks through the gap between human and machine learning. In fact, the proposed method similar-to-different (S2D) strategy may also be useful in efficient learning of real-world complex and high dimensional datasets, especially which are very typical to learn with traditional classifiers.**

***Index Terms*— Classification, learning strategy, similar-to-different (S2D).**

1. Introduction

One of the research hotspots in the field of machine learning is classification. There are different types of classification models such as decision trees, SVM, neural networks, Bayesian belief networks, Genetic algorithm etc. The simple structure, the wide applicability on real time problems, the high efficiency and the high accuracy are the strengths for decision trees. In recent years, many authors proposed improvementsin decision trees learning strategy. A large number of classifiers build the model of dataset for classification by using the traditional learning strategies.

On the other hand, the traditional learning techniques are bottle necked the performance of the datasets. However, several investigations also suggest that there are other factors that contribute to such performance degradation, for example, size of the dataset, density of the dataset, and overall complexity of the dataset. This work focuses on the analysis of improved learning strategy for the open problems related to complex and high dimensional dataset. We focus on proposing an improved learning strategy for classification problems,for a wide range of benchmark datasets.

In this paper,we propose a novel, simple and effectivemachine learning paradigmthat explicitly exploits this important similar-to-different learning strategy, called *S2D*(Similar to Different). We explicitly exploit and implement this similar-to-different learning strategy, a ubiquitous human learning strategy, in the machine learning research. Its applications in human-oriented learning tasks, especially cognitive learning tasks, will be fruitful.S2D builds its model with similar examples first. More specifically, it selects those examples that are close toeach other (thus similar), and updates the model with them. S2D works iteratively in this way which is almost the same as the human learning from similar to different process.Experiment results show that our new learning paradigm S2D has several distinctive advantages overC4.5.

First of all, S2D does indeed take much less effort in building its model than C4.5. Second, minimal effort learning implies that the process of learning and the final learned model are more stable and reliable. This is certainly crucial for human learning, as well as for machine learning applications. Finally, even though S2D only locally updates the model with minimal effort, we show that it is as accurate as the global learner C4.5. One might thinkthat as S2D always takes the similar example to update its model locally and incrementally, it may not predict as accurately as the global learner C4.5 which builds its model on the whole

dataset. Our experiment results show that S2Dpredicts only slightly worse than C4.5. We perform our experimental study focusing on the measures such as accuracy, tree size, Area under the ROC curve (AUC) and error rate.

The rest of this paper is organized as follows. Several previous works related to different learning strategies are reviewed in Section 2. Section 3 describes a generic S2D paradigm.We discuss an efficient implementation and validation of S2D based on the decision tree learning algorithm in Section 4. Experiment results are shown in Section 5. We conclude our work in Section 6.

2. Related Work

Costantino Grana*et al.* [2] have proposed a novel algorithm to synthesize an optimal decisiontree from OR-decision tables, an extension of standard decision tables, complete with the formal proof of optimality and computational cost analysis. As many problems which require recognizing particular patterns can be modeled with this formalism, They select two common binary image processing algorithms, namely connected components labeling and thinning, to show how these can be represented with decision tables, and the benefits of their implementation as optimal decision trees in terms of reduced memory accesses.Joel E. Denny*et al.* [3] have demonstrate that a well-knownalgorithm described by David Pager and implemented in Menhir, the most robust minimal LR(1) implementation they have discovered that, it does not always achieve the full power ofcanonical LR(1) when the given grammar is non-LR(1) coupled with a specification forresolving conflicts. They also detail an original minimal LR(1) algorithm, IELR(1) (InadequacyElimination LR(1)), which they have implemented as an extension of GNU Bison and whichdoes not exhibit this deficiency.

Eileen A. Ni*et al.* [4] have proposed a novel, simple andeffective machine learning paradigm that explicitly exploits this important simple-to-complex (S2C) human learning strategy, and implement it based on C4.5 efficiently.Sanjay Kumar Shukla*et al.* [5] havedeveloped a novelmethodology, genetically optimized cluster oriented soft decision trees (GCSDT), to glean vital information imbedded in the large databases.In contrast to the standard C-fuzzy decision trees, where granules are developed through fuzzy (soft) clustering, in the proposedarchitecture granules are developed by means of genetically optimized soft clustering. In the GCSDT architecture, GA ameliorates thedifficulty of choosing an initialization for the fuzzy clustering algorithm and always avoids degenerate partitions. This provides an effectivemeans for the optimization of clustering criterion, where an objective function can be illustrated in terms of cluster's center. Growth of the GCSDT is realized by expanding nodes of the tree, characterized by the highest inconsistency index of the information granules.

Sanjay Jain*et al.* [6] havea present study aims at insights into the nature of incremental learning in the context of Gold's model of identificationin the limit. With a focus on natural requirements such as consistency and conservativeness, incremental learning is analyzed both for learning from positive examples and for learning from positive and negative examples.In [7] authors introduced a novel form of decision tables, namely OR-Decision Tables, which allow including the representation of equivalent actions for a singlerule. An heuristic to derive a decision tree for such decision tables was given,without guarantees on how good the derived tree was.In [8], authors presented a preliminary version of a bottom-up dynamic programming proposed by Schumacher et al. [9] which guarantees to find the optimal decision tree given an expandedlimited entry (binary) decision table, in which each row contains only one non zero value.

3. Similar to Different (S2D) Learning Strategy

In this section, we follow a design decomposition approach to systematically analyze the learning strategy. We first briefly introduce the design decomposition methodology adopted for new proposed approach.

The C4.5 [1] algorithm is a global learning algorithm i.e. its learns from all the dataset to generate the model and it selects instances randomly to build the model. This peculiar behavior of C4.5 leads to a limitation, where we can make a proper strategy to make the learning more effective. The best strategy for any system is only a natural strategy, where as humans follow for learning new things.

The strategy which humans follow tolearning new things is learning set of similar components and then going for another set of similar components which are different from earlier set and then learning recursively in the above fashion. We can name this strategy as similar to different (S2D) learning strategy. This strategy may also be known as curriculum learning strategy, since all academic curriculum learning will follow the same strategy.

We implemented this S2D strategy in C4.5 learning process. As according to our expectationC4.5 has performed better in almost all the cases. We modified the learning process of C4.5 to suit with our S2D

strategy. In the first phase, C4.5 is allowed to learn on similar set of instances in the dataset, the similarity in the instances can be measured by using metrics relating attributes of the instances. In the next phase of the approach, we need to expose C4.5 algorithm to learn on to the other similar set of instances, which are different from the earlier sets. This new learning strategy is applied to a base algorithm; in this case C4.5 is used to obtain measures such as AUC and accuracy.

4. Datasets and measures

We considered six benchmark real-world imbalanceddataset from the UCI machine learning repository [10] to validateour proposed method. Table 1 summarizes the details of these datasets. This contains the name of the dataset, the total number of examples (Instances), attribute, the number of target classes for each dataset and the missing values.

Table 1 Summary of benchmark imbalanced datasets

Datasets	Instances	Attributes	Class	Missing
Arrhythmia	452	280	16	Y
Credit-g	1,000	21	2	N
Glass	214	10	7	N
Hepatitis	155	20	2	Y
Ionosphere	351	35	2	N
Waveform	5,000	41	3	N

Evaluation Criteria:

In this paper, we use accuracy, tree size, AUC and error rate as performance evaluation measures.

Let us define a few well known and widely used measures:

The most commonly used empirical measure; accuracy is computed by using the below equation (1),

$$ACC = \frac{TP + TN}{TP + FN + FP + FN} \quad \text{---------- (1)}$$

Another measure for performance evaluation is AUC. A quantitative representation of a ROC curve is the area under it, which is known as AUC. When only one run is available from a classifier, the AUC can be computed as the arithmetic mean (macro-average) of TPrate and TNrate.

The Area under Curve (AUC) measure is computed by equation (2),

$$AUC = \frac{1 + TP_{RATE} - FP_{RATE}}{2} \quad \text{---------- (2)}$$

In these experiments, the size of the tree is calculated by the depth of the tree using number of nodes and leaves.Testing errorsis computed as the number of errors produced when separate training and testing set is used for training and testing.

5. Experimental Settings

5.1 Algorithms and Parameters

In first place, we need to define a baseline classifier which we use for our proposed learning strategy. With this goal, we have used C4.5 decision tree generating algorithm [1]. Furthermore, it has been widely used to deal with imbalanced data-sets [11]–[13], and C4.5 has also been included as one of the top-ten data-mining algorithms [14]. Because of these facts, we have chosen it as the most appropriate base learner. C4.5 learning algorithm constructs the decision tree top-down by the usage of the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is the one used to make the decision.

5.2Evaluations on Six Real-World Datasets:

We evaluate theS2D model on sixreal-world datasets obtained from the University of California at Irvine machine learningrepository [10].

For every data set, we perform a tenfold stratified cross validation. Within each fold, the classification method is repeated ten times considering that the sampling of subsets introduces randomness. The accuracy, tree size, AUC and error rate of thiscross-validation process are averaged from these ten runs. The whole cross-validation process is repeated for ten times, and the final values from this method are the averages of these ten cross-validation runs.

6. Experimental Results

We performed theimplementation using Weka on Windows XP with 2Duo CPU runningon 3.16 GHz PC with 3.25 GB RAM.We have analyzed the performance of our proposed algorithmS2D on the following six real-world datasets.The results of the tenfold cross validation with standard deviation are shown in Table2 to 5.

6.1 Test Results on Accuracy:

From table 2, one can observe the results of accuracy of S2D against C4.5 algorithms. The bold dot '●' indicates a win of S2D method on C4.5 and a '○' indicates a loss of S2D method on C4.5. The accuracy results of S2Don all the datasets are better than C4.5.Fig 1 gives the results of S2D method against C4.5 on all the datasets in terms of accuracy.

Table 2. Summary of results on Accuracy

System / Dataset	C4.5	S2D
1. Arrhythmia	65.648±5.860	67.201±7.053●
2. Credit-g	71.250±3.170	73.163±4.388●
3. Glass	67.626±9.321	70.088±13.956●
4. Hepatitis	79.221±9.567	82.377±11.266●
5. Ionosphere	89.744±4.38	89.781±05.32●
6. Waveform	75.252±1.90	76.798±01.89●

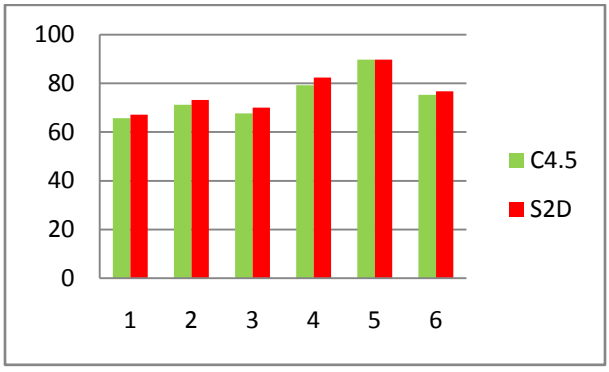

Fig. 1 Test results on accuracy between the C4.5 and S2D for all the datasets.

6.2 Test Results on Tree Size:

From the results of Table 3, we can conclude that the datasets arrhythmia, credit-g, glass, hepatitis, ionosphere and waveform have shown their performance up to the expectation and had registeredwins against C4.5. Fig 2 gives the results of S2D method against C4.5 on all the datasets in terms of tree size.

Table 3. Summary of results on Tree size

System / Dataset	C4.5	S2D
1. Arrhythmia	80.620±5.945	60.66±6.137●
2. Credit-g	126.850±20.664	20.48±4.577●
3. Glass	46.160±4.576	22.693±2.747●
4. Hepatitis	17.660±4.749	7.18±2.299●
5. Ionosphere	26.740±3.894	18.470± 4.168●
6. Waveform	591.940±24.390	417.266±28.754●

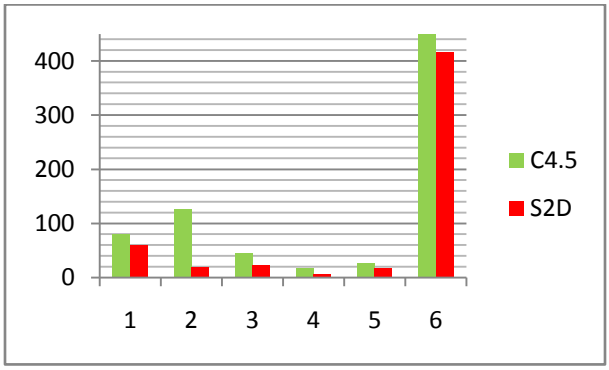

Fig. 2 Test results on tree size between the C4.5 and S2D for all the datasets.

6.3 Test Results on AUC:

From Table 4, we can see the results of S2D against C4.5, in terms of AUC. The datasets arrhythmia, credit-g, ionosphere and waveform have performed well, by registering good number of wins against C4.5. The datasets glass and hepatitis have not performed well when compared to C4.5.In overall, the results with respect to AUC are satisfactory. One the Reason for the underperformance of some datasets may be their unique properties such as size, irrelevant attributes present in the dataset. Fig 3 gives the results of S2D method against C4.5 on all the datasets in terms of AUC.

Table 4. Summary of results on AUC

System / Dataset	C4.5	S2D
1. Arrhythmia	0.776±0.07	0.802±0.08●
2. Credit-g	0.647±0.062	0.680±0.067●
3. Glass	0.794±0.104	0.776±0.180●
4. Hepatitis	0.668±0.184	0.640±0.176●
5. Ionosphere	0.891±0.06	0.892±0.070●
6. Waveform	0.810±0.02	0.847±0.03●

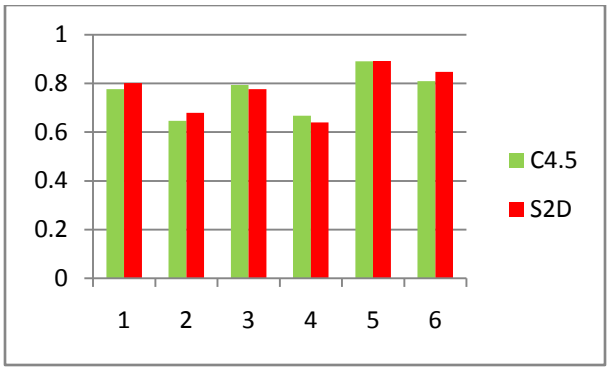

Fig. 3 Test results on AUC between the C4.5 and S2D for all the datasets.

6.4 Test Results on Error Rate:

From Table 5, one can observe the results of error rate. In terms of error rate also all the datasets have performed better than C4.5 when compared S2D method.Fig 4 gives the results of S2D method against C4.5 on all the datasets in terms of error rate.

Table 5. Summary of results on Error Rate

System / Dataset	C4.5	S2D
1. Arrhythmia	34.342±5.862	32.798±7.051●
2. Credit-g	28.750±3.170	26.837±4.385●
3. Glass	32.374±9.312	29.912±13.956●
4. Hepatitis	20.779±9.567	17.619±11.266●
5. Ionosphere	10.256±4.384	10.218±5.321●
6. Waveform	24.748±1.900	23.201±1.979●

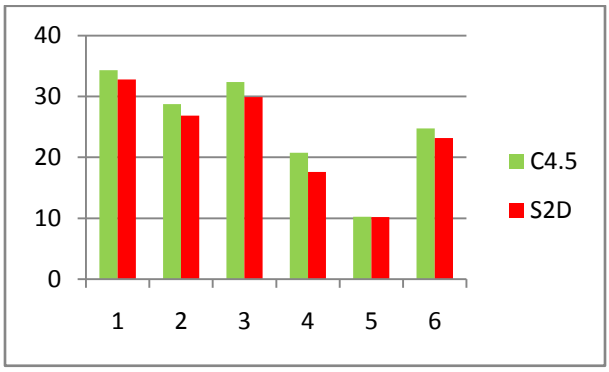

Fig. 4 Test results on error rate between the C4.5 and S2D for all the datasets.

In overall, from all the tables we can conclude that our proposed method S2D have given good results when compared to C4.5. The unique properties of datasets such as size of the dataset, majority, minority ratio and the number of attributes will also effect on the results of our S2D method. The above given results are enough to project the validity of our approach and more deep analysis should be done for further analysis.

7. Conclusion:

Traditional classification techniques build the model for the datasets by following traditional and old strategy. New and novel learning strategies which mimic human learning can of great use to improve the process of model building for the datasets. In this paper we present a novel, simple and effective machine learning strategy, similar-to-different (S2D) and implemented it based on C4.5. Experimental results show thatour proposed S2D method performed well in the case of a wide range of selected datasets. In our future work, we will apply our proposed methodfor learning wide range of tasks, especially for high dimensional feature learning tasks.

References:

1. J. R. Quinlan, *C4.5: Programs for Machine Learning*, 1st ed. San Mateo, CA: Morgan Kaufmann Publishers, 1993.
2. Grana, C., Montangero, M., Borghesani, D., Optimal Decision Trees for Local Image Processing Algorithms, Pattern Recognition Letters (2012), doi: http://dx.doi.org/10.1016/j.patrec.2012.08.015.
3. Joel E. Denny_, Brian A. Malloy, "The IELR(1) algorithm for generating minimal LR(1) parser tables forb non-LR(1) grammars with conflict resolution", Science of Computer Programming 75 (2010) 943_979.
4. Eileen A. Ni and Charles X. Ling, "Supervised Learning with Minimal Effort", M.J. Zaki et al. (Eds.): PAKDD 2010, Part II, LNAI 6119, pp. 476–487, 2010.
5. Sanjay Kumar Shukla a, M.K. Tiwari," Soft decision trees: A genetically optimized cluster oriented approach", Expert Systems with Applications 36 (2009) 551–563.
6. Sanjay Jain a,1, Steffen Lange b, Sandra Zilles, "Some natural conditions on incremental learning", Information and Computation 205 (2007) 1671–1684.
7. C. Grana, D. Borghesani, R. Cucchiara, Optimized Block-based Connected Components Labeling with Decision Trees, IEEE T Image Process 19 (2010) 1596–1609.
8. C. Grana, M. Montangero, D. Borghesani, R. Cucchiara, Optimal decision trees generation from or-decision tables, in: Image Analysis and Processing - ICIAP 2011, volume 6978, Ravenna, Italy, pp. 443–452.

9. H. Schumacher, K. C. Sevcik, The Synthetic Approach to Decision TableConversion, Commun ACM 19 (1976) 343–351.
10. A. Asuncion D. Newman. (2007). *UCI Repository of Machine Learning Database* (School of Information and Computer Science), Irvine, CA: Univ. of California [Online]. Available: http://www.ics.uci.edu/∼mlearn/MLRepository.htm
11. C.-T. Su and Y.-H. Hsiao, "An evaluation of the robustness of MTS for imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 10, pp. 1321–1332, Oct. 2007.
12. D. Drown, T. Khoshgoftaar, and N. Seliya, "Evolutionary sampling and software quality modeling of high-assurance systems," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans.*, vol. 39, no. 5, pp. 1097–1107, Sep. 2009.
13. S. Garc´ıa, A. Fern´andez, and F. Herrera, "Enhancing the effectiveness and interpretability of decision tree and rule induction classifiers with evolutionary training set selection over imbalanced problems," *Appl. Soft Comput.*, vol. 9, no. 4, pp. 1304–1314, 2009.
14. X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, pp. 1–37, 2007.