Synthetic Data Generator for Classification Rules Learning

Runzong Liu, Bin Fang
College of Computer Science
Chongqing University
Chongqing, China
Email: newmaybetrue@163.com

Yuan Yan Tang
Faculty of Science and Technology
University of Macau
Macau, China
Email: yytang@umac.mo

Patrick P.K. Chan
School of Computer Science & Engineering
South China University of Technology
Guangzhou, China
Email: patrickchan@ieee.org

***Abstract*—A standard data set is useful to empirically evaluate classification rules learning algorithms. However, there is still no standard data set which is common enough for various situations. Data sets from the real world are limited to specific applications. The sizes of attributes, the rules and samples of the real data are fixed. A data generator is proposed here to produce synthetic data set which can be as big as the experiments demand. The size of attributes, rules, and samples of the synthetic data sets can be easily changed to meet the demands of evaluation on different learning algorithms. In the generator, related attributes are created at first. And then, rules are created based on the attributes. Samples are produced following the rules. Three decision tree algorithms are evaluated used synthetic data sets produced by the proposed data generator.**

***Keywords*-Synthetic data; Automatic decision support; Data mining; Decision tree;**

I. Introduction

The recent improvement of both hardware and software of computer systems has greatly enhanced the ability of machine computing. Automatic data analysis and decision support become more and more important in our life, economy, social and so on. There is a strong market demand for efficient and valuable algorithms and systems to give people concise conclusions or advice from a big data set [1], [2], [3].

Decision trees, clustering methods, deep learning neural works, statistical and visualization tools are popular techniques used in automatic data analysis or data mining. Among them, decision tree algorithms learn classification rules from an input data set. The output rules tell people the relationship between feature attributes and targets. Decision trees are easy to visualize, therefore, people can understand how the rules are produced. When the final decision is made by human, this is an important virtue than other learning methods such as deep learning.

In practice, the value of learning machines lies on how accurate and how fast they can generate rules or knowledge to guide a system. This makes fundamental research in learning algorithms inherently empirical [4]. Therefore, researchers carry out extensive experimental studies to evaluate the performance of learning algorithms [5]. However, real world data is not controlled by researchers, for example, the data size is limited. The related feature attributes are also uncontrollable. Synthetic data sets are useful in these empirical studies, as they offer a controlled environment used by learning algorithm to evaluate the classifiers [6]. Besides, synthetic data set can be easily access. Therefore, synthetic data sets can be used with real world data sets to derive rigorous results for performance of learning algorithms.

There are a few researches in the field of synthetic data generation for the evaluation on classification algorithms. One is a data generator using five rules [7]. This work demonstrates synthetic data for a person database in which each person has the nine attributes given in Table 1. Five classification functions of increasing complexity that used the above attributes are also developed to classify people into different groups. According to these functions, a training set and a test set are generated for every experiment. Owing to lack of a classification benchmark, researchers used the above synthetic database to evaluate their algorithms [8]. However, since the amount of the attributes and the amount of the inherent rules or functions are fixed, there are limitations. Another work proposed a framework for generating synthetic data for multi-label learning [9]. It uses nine inputs of features, target labels, number of instance, radius of small hyperspheres and noise to generate synthetic data. However, this framework does not meet researchers' demand on the choice of the inherent rules generation. In this paper, we propose a novel rule-driven framework to generate synthetic data set for the evaluation on classification algorithms.

II. Rule-driven Synthetic data generator

The main idea of rule-driven synthetic data generator is to produce specified number of rules according to the specified attributes at first. And then, the generator produces specified amount of samples in compliance with the randomly generated rules.

A classical sample set of playing tennis is shown in Table I. A decision tree produced by the ID3 algorithm from samples of playing tennis in Table I is shown in Figure 1. Table II shows all the rules derived during the generation procedure of the ID3 decision tree algorithm.

From Table II, we can see that five rules are found by the algorithm. However, all these five rules do not consider

978-1-5090-3555-7/16 $31.00 © 2016 IEEE
DOI 10.1109/CCBD.2016.78

the temperature, which is an important factor for human to decide whether he or she will play tennis. Apparently, if the weather is too hot, nobody will go to play tennis even if it's sunny and the humidity is normal. Notice that there is no sample whose outlook is sunny, humidity is normal and temperature is hot in Table I. In fact, we can find two rules about temperature using the rule exhaustion decision tree algorithm [10]. These two rules are shown in Table III. This demonstrates that if an attribute has not been considered by the rule set, the rule set is not complete or the attribute is not necessary. Here, we does not consider that any attribute is not necessary when it is used to generate synthetic samples.

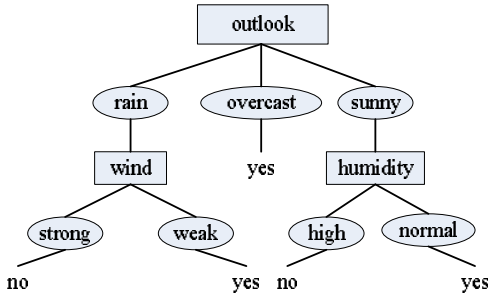

Figure 1. The ID3 decision tree of playing tennis.

Table I
SAMPLES OF PLAYING TENNIS.

day	Outlook	Temperature	Humidity	Wind	play (tennis)
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
3	overcast	hot	high	weak	yes
4	rain	mild	high	weak	yes
5	rain	cool	normal	weak	yes
6	rain	cool	normal	strong	no
7	overcast	cool	normal	strong	yes
8	sunny	mild	high	weak	no
9	sunny	cool	normal	weak	yes
10	rain	mild	normal	weak	yes
11	sunny	mild	normal	strong	yes
12	overcast	mild	high	strong	yes
13	overcast	hot	normal	weak	yes
14	rain	mild	high	strong	no

Table II
RULES FOUND BY ID3 FOR PLAYING TENNIS.

Attribute	Value	Attribute	Value	Target	Value
Outlook	overcast			play	yes
Outlook	sunny	Humidity	high	play	no
Outlook	sunny	Humidity	normal	play	yes
Outlook	rain	Wind	weak	play	yes
Outlook	rain	Wind	strong	play	no

The format of the synthetic sample set produced by the generator is shown in Table IV. The data of Table IV corresponds to the data in Table I. The format of the rule set produced by the generator is shown in Table V.

A. Generating attributes

A set of different attributes of size N should be input by users, and it is expressed as $[A_1, A_2, ..., A_n]$, A_i means that the attribute i has A_i discrete values. The last attribute is always the target attribute. This tuple needs to be input by users. The size of the tuple depends on applications. However, there are some limitations when users try to select attribute. Noticed that if no matter what values of a subset of attributes be, the target attribute is always the same, it means that the subset of attributes has no relations with the target attribute. Therefore, for attribute i in the rule set, there must be at least A_i rules which cover all the A_i values of attribute i and all the A_n values of the target attribute n. This also means that there should be at least $max(A_1, A_2, ..., A_{n-1}, A_n)$ rules to make the rule set complete. And it is vice versa. If no matter what values of the target attribute being considered, a subset of attributes is always the same value, it means that the relation between the target attribute and the subset of attributes is not clear. For example, no matter what color a cat is, it is our good friend if it can catch mice. Therefore, for the same value of a subset of attributes of different rules in the rule set, the values of the target attribute must be different.

Table III
RULES FOUND BY RULE EXHAUSTION DECISION TREE FOR PLAYING TENNIS.

Outlook	sunny	Temperature	hot	play	no
Outlook	sunny	Temperature	cool	play	yes

Table IV
FORMAT OF THE SAMPLE SETS IN TABLE 1.

day	Outlook	Temperature	Humidity	Wind	play tennis
1	0	0	0	0	0
2	0	0	0	1	0
3	1	0	0	1	1
4	2	1	0	0	1
5	2	2	1	0	1
6	2	2	1	1	0
7	1	2	1	1	1
8	0	1	0	0	0
9	0	2	1	0	1
10	2	1	1	0	1
11	0	1	1	1	1
12	1	1	0	1	1
13	1	0	1	0	1
14	2	1	0	1	0

Table V
FORMAT OF THE RULE SET IN TABLE 2.

Attribute	Value	Attribute	Value	Target	Value
1	1	0	0	5	1
1	0	3	0	5	0
1	0	3	1	5	1
1	2	4	0	5	1
1	2	4	1	5	0

B. Generating rules

When the attributes are assigned, M rules are generated according to the attributes. But there are some rules which are in conflict. For example, let the first rule be that if attribute i is 0, the target attribute is 1; Let the second rule be that if attribute i is 0, the target attribute is 0. Apparently, the two rules are in conflict. One strategy to solve this problem is that rules are create randomly, but before using them, a thorough check is carried on the whole rule set from the first to the last, and those rules which are in conflict with the former rules will be removed. Besides, when generating rules, if a rule needs to consider all the related attributes to determine the label of the target attribute, this rule is meaningless, because this rule merely corresponds to a sample. Generally, if rule A uses fewer attributes than rule B, rule A is better than rule B. Of course, there should not be a rule which considers no attributes but classifies the target attribute directly.

C. Generating samples

Finally, K samples are produced in compliance with the M rules. There are two ways to generate sample. The first way is the sequential way, samples are produced sequentially from the first rule to the last rule. The second way is the random way, samples are produced using rules in a random order. Since the synthetic data is produced sequentially by the engine rules and a front part of them are used as the training data. The training data used for learning rules can only cover a part of the entire data, the sequential way can not provide samples produced from the latter engine rules.

D. Relations between attribute and rule

From the above analysis, we can induce that there are some relationships between the size M of the rule set and the size of value sets of the attributes $[A_1, A_2, ..., A_n]$. First, no rules can consider all the related attributes. Second, for attribute i in the rule set, there must be at least Ai rules which cover all the Ai values of attribute i and these Ai rules can cover all the An values of the target attribute n. Finally, no rules should be in conflict. According to the above principles, we have the following formulas.

$$M \leq \sum_{i=1}^{n-1} \prod_{k=1,k\neq i}^{n-1} A_k$$

$$M \geq mp + np$$

$$mp = \arg\max(A_i, i = 1, 2, \ldots n-1)$$
$$np = \arg\min(A_i, i = 1, 2, \ldots n-1)$$

However, since some rules are in conflict, the amount of compliance rules is much smaller.

III. EXPERIMENTAL RESULTS OF THE EVALUATION ON THE PRECISION OF DECISION TREE ALGORITHMS

Experiments were carried out on a PC with Intel Pentium 1.50 GHz CPU and Windows XP operating system using Matlab 7.4.0.

In the experiments, we used the proposed generator to produce synthetic sample sets and rules in order to compare algorithms. We compared the classical ID3 algorithm [11] with the information energy decision tree (IET) [12] and the rule exhaustion decision tree (RDT) [10] of our recent work. Table VI, Table VII and Table VIII show the results about the classification precision.

Two attribute sets were used. The first set of 15 attributes was as follows,

$[2, 3, 2, 3, 2, 3, 2, 3, 2, 2, 3, 2, 3, 3, 2]$.

The second set of 20 attributes was as follows,

$[2, 3, 2, 3, 2, 3, 2, 3, 2, 2, 3, 2, 3, 3, 3, 2, 2, 3, 2, 2]$.

And the amount of rules equaled to the amount of attributes in all the experiments. Testing accuracy was used to measure the precision, $accuracy = correct/total$. Each sample set was divided into a training set and a testing set. In order to evaluate learning algorithms' ability to learn rules from a small and incomplete data set, both the random way and the sequential way to generate data were used. The results are shown in Table VI, Table VII and Table VIII. Each figure in a table corresponds to the statistics over 100 trials.

From the tables, we can see that the results of the algorithms are all above 96%, when the data is accessed randomly and the data size is above 10 times as the amount of the rules. On the other hand, when the data is accessed sequentially, the precision drops seriously. It's because that the training data can't cover some latter rules. However, for sequential data, the precision rate of the rule exhaustion decision tree is obviously better than ID3 and the information energy decision tree (IET), while IET is better than ID3.

Table VI
THE ACCURACY OF ID3 OR THE RULE EXHAUSTION DECISION TREE WHILE 25% OF THE SAMPLES ARE USED AS TRAINING SAMPLES. THE UNIT IS PERCENTAGE.'RND' MEANS 'RANDOM' HERE.

Rules Amount N	data access	Algorithm	Data size 10*N	100*N	1000*N
15	rnd	RDT	79.416	96.913	99.877
	rnd	ID3	81.336	97.732	99.910
	rnd	IET	80.814	97.254	99.594
20	rnd	RDT	81.720	96.149	99.758
	rnd	ID3	82.953	96.835	99.823
	rnd	IET	83.113	96.124	99.461
15	no rnd	RDT	64.584	69.962	68.593
	no rnd	ID3	61.212	67.533	66.263
	no rnd	IET	62.442	67.333	66.305
20	no rnd	RDT	65.013	71.333	71.604
	no rnd	ID3	65.400	68.984	70.760
	no rnd	IET	66.060	69.589	70.903

IV. CONCLUSION

A novel framework for generating synthetic data set is proposed for evaluation on classification rules learning algorithms. Researchers can apply the framework to produce experimental data sets on machine learning. The features of data set can be modified by researchers, such as, the data size, the amount of related attributes, the value set of the attributes, and the amount of rules. Relations among the rules and attributes when generating synthetic data have also been investigated. Three decision tree algorithms are evaluated using synthetic data generated under the proposed framework.

Table VII
THE ACCURACY OF ID3 OR THE RULE EXHAUSTION DECISION TREE WHILE 50% OF THE SAMPLES ARE USED AS TRAINING SAMPLES. THE UNIT IS PERCENTAGE.'RND' MEANS 'RANDOM' HERE.

Rules Amount N	data access	Algorithm	Data size 10*N	100*N	1000*N
15	rnd	RDT	86.307	98.529	99.961
	rnd	ID3	88.720	98.972	99.974
	rnd	IET	88.373	98.596	99.737
20	rnd	RDT	87.210	97.991	99.895
	rnd	ID3	88.580	98.501	99.934
	rnd	IET	87.800	97.776	99.664
15	no rnd	RDT	70.880	76.536	78.260
	no rnd	ID3	69.253	75.063	74.229
	no rnd	IET	70.187	75.239	74.022
20	no rnd	RDT	74.840	79.500	78.478
	no rnd	ID3	74.180	75.989	76.892
	no rnd	IET	74.150	78.114	77.656

Table VIII
THE ACCURACY OF ID3 OR THE RULE EXHAUSTION DECISION TREE WHILE 75% OF THE SAMPLES ARE USED AS TRAINING SAMPLES. THE UNIT IS PERCENTAGE.'RND' MEANS 'RANDOM' HERE.

Rules Amount N	data access	Algorithm	Data size 10*N	100*N	1000*N
15	rnd	RDT	90.026	99.179	99.978
	rnd	ID3	92.211	99.459	99.990
	rnd	IET	91.053	99.061	99.758
20	rnd	RDT	89.320	98.786	99.936
	rnd	ID3	91.260	98.980	99.965
	rnd	IET	90.340	98.336	99.717
15	no rnd	RDT	76.289	83.477	84.951
	no rnd	ID3	77.237	81.419	82.183
	no rnd	IET	77.763	83.061	83.113
20	no rnd	RDT	78.960	83.440	83.292
	no rnd	ID3	77.760	83.522	81.403
	no rnd	IET	78.440	83.694	82.404

ACKNOWLEDGMENT

This work is supported in part by the Fundamental Research Funds for the Central Universities (0903005203327), the Science and Technology Development Fund (FDCT) of Macau 019/2015/A, and Macau-China join Project 008-2014-AM, and in part by the National Natural Science Foundations of China (NSFC-61472053, NSFC-91420102).

REFERENCES

[1] S. Y. Kim and A. Upneja, "Predicting restaurant financial distress using decision tree and adaboosted decision tree models," *Economic Modelling*, vol. 36, no. 1, pp. 354–362, 2014.

[2] J. S. Lee and E. S. Lee, "Exploring the usefulness of a decision tree in predicting people's locations," *Procedia - Social and Behavioral Sciences*, vol. 140, p. 447C451, 2014.

[3] A. G. Malliaris and M. Malliaris, "What drives gold returns? a decision tree analysis," *Finance Research Letters*, vol. 13, pp. 45–53, 2015.

[4] T. G. Dietterich and T. G. Dietterich, "Exploratory research in machine learning." *Machine Learning*, vol. 5, pp. 5–9, 1990.

[5] P. Langley, "Crafting papers on machine learning," in *Seventeenth International Conference on Machine Learning*, 2000, pp. 1207–1212.

[6] V. Boln-Canedo, N. Snchez-Maro?o, and A. Alonso-Betanzos, "A review of feature selection methods on synthetic data," *Knowledge and Information Systems*, vol. 34, no. 3, pp. 483–519, 2012.

[7] R. Agrawal, S. Ghosh, T. Imielinski, B. Iyer, and A. Swami, "An interval classifier for database mining applications," *Proceedings of the 18th international conference on very large databases*, pp. 560–573, 1992.

[8] M. Wang, B. Iyer, and J. S. Vitter, "Scalable mining for classification rules in relational databases," in *International Symposium on Database Engineering and Applications*, 1998, pp. 58–67.

[9] J. T. Tomas, N. Spolaor, E. A. Cherman, and M. C. Monard, "A framework to generate synthetic multi-label datasets," *Electronic Notes in Theoretical Computer Science*, vol. 302, p. 155C176, 2014.

[10] R. Z. Liu, Y. Y. Tang, and B. Fang, "Automatic decision support by rule exhaustion decision tree algorithm," in *IEEE International Conference on 2016 Wavelet Ananlysis and Pattern Recognition*, 2016.

[11] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, pp. 81–106, 1986.

[12] R. Z. Liu, Y. Y. Tang, and B. Fang, "Automatic decision support by information energy decision tree algorithm," in *IEEE International Conference on Systems, Man and Cybernetics*, 2014, pp. 4047–4051.