Performance of Machine Learning Algorithms and Diversity in Data

Hyontai SUG[1*]

[1]Division of Computer Engineering, Dongseo University, 47 Jurye-ro, Sasang-gu, Busan 47011, Korea

Abstract. Recent world events in go games between human and artificial intelligence called AlphaGo showed the big advancement in machine learning technologies. While AlphaGo was trained using real world data, AlphaGo Zero was trained using massive random data, and the fact that AlphaGo Zero won AlphaGo completely revealed that diversity and size in training data is important for better performance for the machine learning algorithms, especially in deep learning algorithms of neural networks. On the other hand, artificial neural networks and decision trees are widely accepted machine learning algorithms because of their robustness in errors and comprehensibility respectively. In this paper in order to prove that diversity and size in data are important factors for better performance of machine learning algorithms empirically, the two representative algorithms are used for experiment. A real world data set called breast tissue was chosen, because the data set consists of real numbers that is very good property for artificial random data generation. The result of the experiment proved the fact that the diversity and size of data are very important factors for better performance.

1 Introduction

AlphaGo and AlphaGo Zero are two successful applications of machine learning algorithms called deep learning of neural network [1] that are based on very two different sources of data. While AlphaGo used the records of sixteen hundred thousand go games for its training, AlphaGo Zero used only the rules of go game and massive random generation of data [2]. It has been reported that AlphaGo Zero defeated AlphaGo with 100:0. Even though the hardware of AlphaGo Zero is slightly better and efficient, the main reason for the winning of AlphaGo Zero against AlphaGo is that the characteristics of data themselves that were used for the training. The supplied data to AlphaGo is real go game data records of sixteen hundred thousand that were played between human go players. But, because the number of intersection in checkerboard of go game is 19 x 19 = 316, the data records cover only small fraction of 361!, even though we might think that the size of data is very large. On the other hand, AlphaGo Zero used massive random data for training, so that it was exposed to more novel data abundantly than AlphaGo. In other words, the spectrum of training data is better. That's the reason why some human expert in go admired the novelty of AlphaGo Zero's moves in go games [3]. When a machine learning algorithm is trained with more diverse values of its training data, the trained model could cover more future unseen cases, so that the performance of the model could become better like AlphaGo Zero's case.

In the application area of machine learning algorithms each problem domain is similar to the problem domain of go game with some difference in the complexity of the problem. One advantageous point to solve the go game problem is that we have the rules of movement so that illegal movement can be checked easily, when we generate random data. But, when we try to apply machine learning algorithms in other domains, there are almost no domains that have such rules to check whether the generated random data can belong to the domain, and this is one of the main reasons that make the data mining task hard.

There are many machine learning algorithms proposed. Among them, neural networks and decision trees can be two most preferred machine learning algorithms in data mining [4]. Neural network based algorithms are relatively stable, because the effect of training data is distributed evenly in the structure of the network so that their performances are affected less by the perturbation of training data than other less stable algorithms like decision trees. One the other hand, because decision tree algorithms try to divide training data decisively in the structure, they are more sensitive to the quality of the data [5]. In this paper we will see how the diversity and size of data could affect the performance of the two different machine learning algorithms experimentally. Because we need more data for our experiment, we'll try to generate the diverse data artificially using a random and linear interpolation method of real data.

[*] Corresponding author: sht@gdsu.dongseo.ac.kr

In the following section 2 the method used for our experiment will be described, and conclusions will be provided in section 3.

2 Method and experiment

We need randomness and diversity in training data, so an artificial data generation algorithm like SMOTE [6] could be good for our purpose. SMOTE algorithm was invented as a method of over-sampling for a minority class, because the minority class usually does not have enough training instances for accurate classifiers. Success was reported for several machine learning algorithms including decision trees and rule generation algorithms. SMOTE generates new data instances based on randomization in the linear interpolation of nearest neighbours in existing instances, so the algorithm satisfies our two purposes - novelty by the randomization and increase of training data set size by the new artificial data generation.

Therefore, we apply SMOTE to generate large number of artificial instances for our experiment. But, we apply the algorithm differently. While the original algorithm generates artificial instances of a minority class, we generate new artificial instances for all classes in the data set to get the diversity. For our experiment, a data set called ‘Breast Tissue’ from UCI machine learning depository [7] is used. The data set was chosen because all of its attributes are continuous attributes so that it’s advantageous for the linear interpolation. Breast tissue data has 9 conditional attributes and one decision attribute. The decision attribute has 6 different class values which classify breast tissue. The total number of instances is 106. Table 1 shows the meaning of each attribute.

Table 1. The attributes of data set ‘breast tissue’.

No.	**attribute**	**domain**
1	I0	Impedivity at zero frequency, real
2	PA500	Phase angle at 500KHz, real
3	HFS	High frequency slope of phase angle, real
4	DA	Impedance distance between spectral ends, real
5	AREA	Area under spectrum, real
6	A/DA	Area normalized by DA, real
7	MAX IP	Maximum of the spectrum, real
8	DR	Distance between I0 and real part of the maximum frequency point, real
9	P	Lenght of the spectral curve, real
10	Class	car(carcinoma), fad(fibro-adenoma), mas(mastopathy), gla(glandular), con(connective), adi(adipose)

As deep learning neural network and decision tree algorithm, multilayer perceptron (MLP) [8] and C4.5 [9] were used respectively.

Table 2 shows the accuracy of the two algorithms for the data set. In the experiment 10 fold cross-validation was used. For MLP epochs of 4000 and the learning rate of 0.3 and the number of hidden layers of 5 were used. Default parameters were applied for C4.5.

Table 2. The result of MLP and C4.5

algorithm	**Accuracy (%)**	**Tree size**
MLP	69.8113	N/A
C4.5	66.0377	29

The accuracy achieved by MLP is comparable to the result by Norte’s SVM which is 70.598% in 3 fold cross-validation [10]. SVM is known for its ability to get high accuracy in classification task [11]. The following is the generated tree by C4.5.

```
IO <= 551.879287
| area <= 1664.674076
| | DA <= 53.5996
| | | MaxIP <= 18.131014: fad (4.0)
| | | MaxIP > 18.131014
| | | | PA500 <= 0.165806
| | | | | DA <= 35.780061: gla (11.0)
| | | | | DA > 35.780061
| | | | | | DA <= 38.940168: mas (3.0)
| | | | | | DA > 38.940168: gla (5.0/1.0)
| | | | PA500 > 0.165806: fad (4.0/2.0)
| | DA > 53.5996
| | | IO <= 355
| | | | area <= 346.091312: mas (3.0)
| | | | area > 346.091312
| | | | | PA500 <= 0.127409: fad (8.0)
| | | | | PA500 > 0.127409
| | | | | | HFS <= 0.08238: car (2.0/1.0)
| | | | | | HFS > 0.08238: mas (3.0)
| | | IO > 355: mas (4.0)
| area > 1664.674076
| | MaxIP <= 49.327862
| | | MaxIP <= 44.740154: car (4.0)
| | | MaxIP > 44.740154: mas (3.0/1.0)
| | MaxIP > 49.327862: car (16.0)
IO > 551.879287
| P <= 1524.609204: con (15.0/1.0)
| P > 1524.609204: adi (21.0)
```

SMOTE based over-sampling was applied to add more new data for each class. For each class over-sampling rate of 200% and 400% are applied with the parameter of 5 nearest neighbours. Table 3 shows the change of the number of data instances for each class after the over-sampling.

Table 3. The result of over-sampling for all classes

Over-sampling rate	car	fad	mas	gla	con	adi	Total # of insta-nces
0%	21	15	18	16	14	22	106
200%	63	45	54	48	42	66	318
400%	105	75	90	80	70	110	530

Table 4 shows the trained result for over-sampling rate of 200%.

Table 4. The result of MLP and C4.5 for over-sampling rate of 200%

algorithm	**Accuracy (%)**	**Tree size**
MLP	80.8176	N/A
C4.5	85.5346	45

Table 5 shows the trained result for over-sampling rate of 400%.

Table 5. The result of MLP and C4.5 for over-sampling rate of 400%

algorithm	**Accuracy (%)**	**Tree size**
MLP	84.717	N/A
C4.5	88.6712	57

The following is resulting decision tree of over-sampling rate of 400%.

```
IO <= 551.879287
| area <= 1370.838068
| | DA <= 37.452537
| | | MaxIP <= 18.131014: fad (10.0)
| | | MaxIP > 18.131014
| | | | PA500 <= 0.166368
| | | | | DA <= 34.21955: gla (49.0/1.0)
| | | | | DA > 34.21955
| | | | | | HFS <= 0.09436: gla (6.0)
| | | | | | HFS > 0.09436: fad (3.0/1.0)
| | | | PA500 > 0.166368: fad (3.0/1.0)
| | DA > 37.452537
| | | aDA <= 4.942034: mas (17.0)
| | | aDA > 4.942034
| | | | HFS <= 0.121332
| | | | | IO <= 359.673081
| | | | | | DR <= 48.512974
| | | | | | | MaxIP <= 22.02108
| | | | | | | | P <= 185.184278: mas (3.0)
| | | | | | | | P > 185.184278: fad (13.0/1.0)
| | | | | | | MaxIP > 22.02108
| | | | | | | | P <= 204.940569: fad (4.0)
| | | | | | | | P > 204.940569
| | | | | | | | | DA <= 53.5996
| | | | | | | | | | area <= 479.381335: gla (17.0)
| | | | | | | | | | area > 479.381335: mas (3.0/1.0)
| | | | | | | | | DA > 53.5996
| | | | | | | | | | PA500 <= 0.135279: mas (3.0)
| | | | | | | | | | PA500 > 0.135279: fad (2.0)
| | | | | | DR > 48.512974: fad (40.0/2.0)
| | | | | IO > 359.673081
| | | | | | DA <= 58.817611: gla (3.0)
| | | | | | DA > 58.817611: mas (20.0/1.0)
| | | | HFS > 0.121332
| | | | | IO <= 177.700232: fad (3.0)
| | | | | IO > 177.700232: mas (29.0/1.0)
| area > 1370.838068
| | PA500 <= 0.142761
| | | HFS <= 0.07462: gla (4.0)
| | | HFS > 0.07462: mas (6.0)
| | PA500 > 0.142761
| | | MaxIP <= 49.327862
| | | | MaxIP <= 45.119064
| | | | | HFS <= 0.369083: car (15.0)
| | | | | HFS > 0.369083
| | | | | | DA <= 157.884181: mas (2.0)
| | | | | | DA > 157.884181: car (2.0)
| | | | MaxIP > 45.119064
| | | | | IO <= 318.564551: mas (6.0)
| | | | | IO > 318.564551: car (2.0)
| | | MaxIP > 49.327862: car (85.0)
IO > 551.879287
| P <= 1524.609204
| | area <= 11888.39183: con (70.0)
| | area > 11888.39183: adi (3.0)
| P > 1524.609204: adi (107.0)
```

From the generated decision trees we can see that the complexity of the tree increases as we supply more data. The following graph in fig. 1 summarizes the change of accuracy for the two machine learning algorithms, MLP and C4.5, as the size of training data increases.

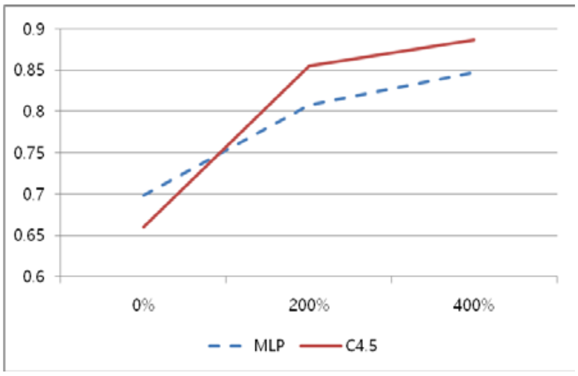

Fig. 1 The change of accuracy of MLP and C4.5 as the size of training data increases

As we can see in the graph, when the size of data set is relatively smaller, the accuracy of MLP is better, but, as the data size becomes bigger, the accuracy of decision tree becomes better. This graph reflects the property of the two machine algorithms well. In other words, because decision tree reflects the property of data more frankly than neural networks, the accuracy of decision tree improves better with the increasing data. All in all, based on the two very different and representative machine learning algorithms, we can confirm that fact that more diverse and bigger data could generate more accurate classifiers empirically.

3 Conclusions

Recent events in go games between human go players and a machine go player called AlphaGo revealed the amazing progress of machine learning technologies. And, moreover, AlphaGo Zero that was developed after AlphaGo showed that diversity and size of data are very important for better performance for the machine learning algorithms especially in deep learning algorithms of neural networks. While AlphaGo used hundreds of thousands records of go games played by human for training, AlhpaGo Zero used massive random moves that were checked by the rules of go game only for training, resulting in total victory against AlphaGo. The random moves supplied more novel training data set to the algorithms, so that it implies the fact that the diversity and size in data are very important in the success of the machine learning algorithms.

Neural networks and decision trees are widely accepted machine learning algorithms because of their robustness in errors and comprehensibility respectively. In order to confirm the diversity and size in data are important factors in the performance of machine learning algorithms experimentally, the two representative algorithms are selected and used. A real world data set called breast tissue which consists of continuous attributes only was used. In order to prepare more diverse data over-sampling algorithm called SMOTE was applied for all classes in the data set. Experiment showed that as the size and diversity of data increase, the accuracy of the machine learning algorithms improves very much. Further research task could be how to check the correctness of the new artificial instances from SMOTE like AlhhaGo Zero has rules of go game to check the random data, because the randomization in the linear interpolation of nearest neighbours does not guarantee the correctness of the new instances 100%.

References

1. J. Schmidhuber, Neural Networks, **61**, 85-117 (2015)
2. D. Silver, A. Huang, C.J. Maddison, A. Guez, L. Sifre, G.V.D. Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, Nature, **529** (7587), 484-489, doi :10.1038/nature16961 (2016)
3. T. Chouard, Nature, doi :10.1038/nature.2016.19533 (2016)
4. P. Tan, M. Steinbach, A. Karpatne, V. Kumar, *Introduction to Data Mining*, 2nd ed., Pearson (2018)
5. W. Sun, *Stability of machine learning algorithms*, PhD thesis, Purdue University (2015)
6. N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, J. of Artificial Intelligence Research, **16**, 321-357(2002)
7. A. Frank, A. Suncion, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Sciences (2010)
8. M. Popescu, V.E. Balas, L. Perescu-popescu, N. Mastorakis, WSEAS Transactions on Circuits and Systems, **8**, 7, 579-588 (2009)
9. J.R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufman Publishers (1993)
10. M.W. Nonte, *Classification of Breast Tissue Based on Electrical Impedance Spectroscopy Data*, ECE/CS /ME 539 Introduction to Artificial Neural Networks and Fuzzy System-Fall 2013 Semester Class Projects, University of Wisconsin (2013)
11. C.J.C. Burges, Data Mining and Knowledge Discovery, **2**, 2, 121-167 (1998)