Question Answering using Random Forests

Yi Peng
The University of Essex,UK
Module Supervisor: *Spyros Samothrakis*
Yp16181@essex.ac.uk

Abstract

Random-Forest algorithm of decision trees has advantage of short training time, low complex , fast forecast . this article will use the advantages of this algorithm, to design a question answering system based on Random-Forest.

1 Introduction

Word2vec is an efficient tools for converting a word representation to real value vector, open-source by Google in mid - 2013. Using thought of deep learning to simplify text content into k dimensional vector,and similarity of the vector space can be used to represent text semantic similarity[1].

Group of Tomas Mikolov also showed several original skip - gramm extension of the model that can be used in several orders of magnitude[2].

The Generation and development of random forest algorithm was designed by Leo Breiman, Adele Cutler, Ho Tin notes, Dietterich, Amit and Geman. Leo Breiman and Adele Cutler first puts forward the key opinion to perform random forest algorithm, Characteristics and immediately choose thoughts was brought into this algorithm by Amit, Gemen and Ho Tim, and then used the ideas of Breiman of "set" to build the control variance of decision tree. And in the end the optimization of the node was introduced to further improve random forest[3].
The data that I will use to train the Random-Forest is bAbi task,there will be 20 tasks

in this experiment,so 20 Random-Forest should be built. The purpose of the work is to answer the question by Random-Forest that will be built.

2 Background

Combination model (for example a Boosting, Bagging, etc.) related to the decision tree algorithm is much more popular than before, the final results of these algorithms is to generate hundreds of tree (or above), so that we can greatly reduce the trouble which are brought by single decision tree , although the hundreds of each tree in the decision tree are very simple, but they will be strong when they complete combination On paper in recent years, such as ICCV meeting, there are many articles are related to Boosting and random forest. there are two more fundamental algorithm about Model combination + related Decision Tree algorithm, that is random forests and GBDT (Gradient Boost Decision Tree), the other new model combination + the Decision Tree algorithm are from the extension of the two algorithms.

For example Random forest algorithm can be used in application of human recognition[4],Random-Forests can turn gesture recognition problem into the object recognition problem based on the determination of space vector from different parts of the body

3 Methodology

In this exercise,in order to answer the question by learning information in the short stories, the data from the bAbi task will be used. And at first these tasks sentences will be converted to vectors, a proper model “Continuous Skip-gram Model (Skip-gram)”showed by Tomas’ team will be chosen to fix length representations of vectors, and then use these victors to train a Random Forest. By using the correlation of the vectors, the most probable answer can be found. Because in random forests, we will generate a lot of decision trees compared with only one tree in the CART model.

When a new object criterion is classificated based on some properties. every tree in the Random Forest will give their classification and "vote" the output of the whole forest classification result ; In the regression problem, the output of the random forest will be the average output of all trees.

4 Experiments

First of all ,the out put of word2vec should be considered carefully because these vectors will contain the characteristics in the data, and the correlation of these characteristics are the key information to build the learning system. Then the input(question in vector way) is judged to decided which Random-Forest should be used In next part . And how to train Random-Forest and what the Decision-tree should be like are the key step. Because this experiment should consider the amount of calculation, the number of Decision-tree and how many samples that are used to trained should also be considered. All of these provide the theoretically feasibility.

5 Discussion

In the output part ,the result should be tested several times in order to do the chi-square test and ensure a good system we got. And another test should be done is reduce or increase the quantity of training. It can verify the accuracy of this experiment. And the judgment in first step should be explored in some different ways ,because the the logic between these tasks can be stronger.

6 Conclusion

There some advantages and disadvantages by using Random-Forest,

Advantages:

1、In order to get a proper model, they don't need to do a lot of adjustments. Just use a lot of trees, the model will not produce a lot of deviation.

2、For random forest, the model and the algorithm itself is very acceptable.

Disadvantages:

1、The main disadvantage of Random-Forest algorithm is the model size.Model is easily to spend hundreds of megabytes of memory.

2、The model prediction is not very friendly to question out of the database.

7 Plan

From now to 10th of March design the structure of the program

Until 25th of March finish the Word2vec section.

Finish the main body of experiment before 10th of April

Do the tests and improvement between 10th and 17th of April

Finish and check report during 17th to 26th

8 References

[1]Tomas M; Kai C; Greg C; and Jeffrey D; *Efficient Estimation of Word Representations in Vector Space*.2013

[2]Tomas M; Kai C; Greg C; and Jeffrey D;..*Distributed Representations of Words and Phrases and their Compositionality,*In Proceedings of Workshop at ICLR, 2013

[3]*A Complete Tutorial on Tree Based Modeling from Scratch (in R & Python)*,2016,website

https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/

[4]Nan ,C .*Random forest algorithm in the application of human recognition*
http://bbs.pinggu.org/thread-3607444-1-1.html

Github link https://github.com/yp16181/ce888labs/upload/master