Question Answering using Random Forests

Submitted as part of the requirments for:
CE888 Data Science and Decision Making

Name: Xie Chuan
Student Number: XIECH18204
Lecture: Spyros Samothrakis
Date: 23 February 2016

Abstract

The computer is difficult to read the human's language and answer the human's questions. This research focus on using the Random Forest and Word2vec method to deal with the computer reading and understanding the human's storied and questions problem. The major contribution of this research is that offer a basic artificial intelligent to handle the human's language problem.

Key words: Random Forests, Work2vec, Text Analyze, bABi Tasks, Classification

Introduction

This experiment is focus on using the Random Forests and Word2vec to build an artificial intelligence system to answer the question about a given story.

The data set is bAbi Tasks that has many stories and questions. There has 5 sets in the bAbi Tasks and they are en/, hn/, shuffled/, en-10k/ shuffled-10k/ hn-10k/ and en-valid/ en-vaild-10k/. Except en-10k shuffled-10k and hn-10k sets are 10000 training and others are 1000 training examples. The en/ and hn/ sets are in English and Hindi. And these can be readable by humans. The shuffled/ has shuffled letters so they are not readable by people. The train and valid portions are split in the en-valid/ and en-valid-10-k/ folds.

Random Forests and Word2vec are the main method in this experiment. Random Forests is used to classify the fixed length representations of vectors to build the forests and learn the question to find the answer from the random forests. Word2vec is used to convert tasks sentences.

Background

Random Forests algorithm is very popular in classification and machine learning. In the business, computer science, and many other domains. There is an example that is used Random Forests in optic field. Ham used the random forest to study the hyperspectral data. Because of the data features are high in dimension and represent multiple classes that are sometimes quite mixed the hyperspectral data is difficult to be analyzed. Using the random forest method based on the framework of classification and regression trees. The random forest classification improved the accuracies in this experiment. And using a forest rather than a single tree got the greater incremental benefit from ensemble [1].

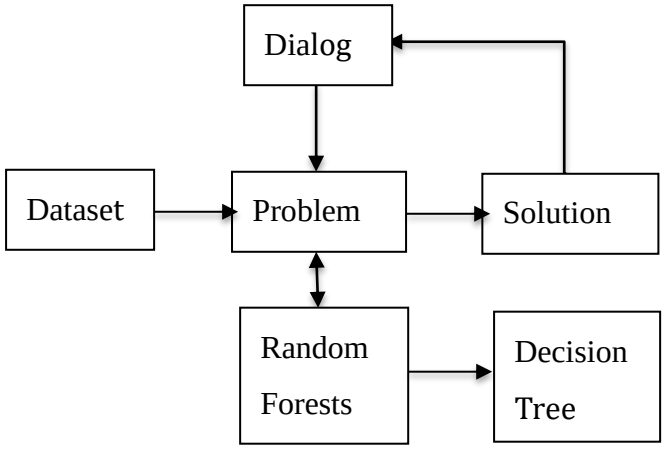
Methodology

Random Forests is a kind of machine learning algorithm. And a random forest is constructed by many classification trees [2]. In this experiment the decision trees are used to construct the random forests. Because of using decision trees to construct the random forests the forests are also called random decision forests. And using the random forests to regression estimate the goal class with the features given. The most important for random forests is growing the every tree [3]. The dataset that is used in this experiment is bAbi Task and it is from Weston's dataset construction method [4]. First, using the word2vec method to convert this tasks to the fixed length representations of vectors and feed them into random forests. Next using these vectors

to grow the decision trees. Choosing N cases randomly as a tree training set with replacement from the original dataset. And then determining a number m << M that is the number of features in original dataset [5]. Constructing m variables from M features and using this training set to growing a decision tree without pruning. Looping this process there are many decision trees build and these trees construct the random decision forests.

Experiments

There are 6 main classes in this experiment. They are Dataset, Problem, Solution, Random Forests, Decision Tree, Dialog classed. Dataset function is reading dataset from bAbi Tasks and convert tasks sentences. Problem function is feed the fixed length representations of vectors and user's input to random forests. Random function is create training set for decision trees, call decision tree function to create the random forests and based on the user's input to find the value of result and the process of inference. Decision Tree function is using the training set to grow the decision tree. Dialog function is creating a dialog to contact with user. Solution function is outputting the answer and process of inference.



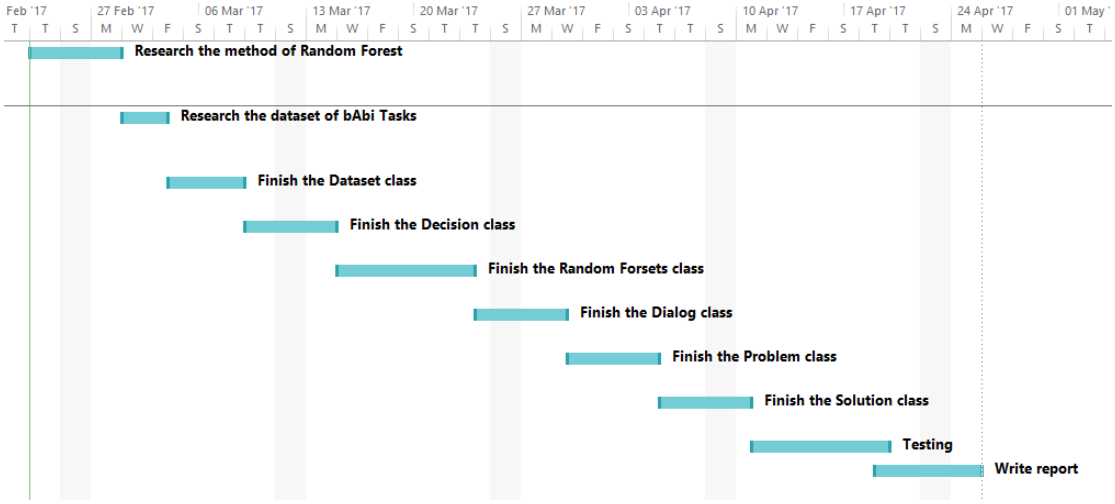
Discussion

Because of the random forests is constructed by random training set with decision trees the random forests are no need for cross-validation or a separate test. Just final function test is needed. Every tasks dataset has corresponding test set and using this test set without the accurate answers inputs into system to get the result and compare with the accurate answers to evaluate the system. But the Dataset, Dialog, Problem, and Solution classes have to be tested. Testing Dataset function is used a testing set that is a kind of story. And compare the accurate vector with the output vector to evaluate this function. Dialog class is tested the function of inputting and outputting. Problem function is tested with the whole system function. Solution function is tested the result is saved.

Conclusion

This experiment is aim to deal with the text analyze of answer the questions about stories. Final the function that is answer the questions in a given story can be achieved. And the system can talk with user like humans with this function in a dialog. The random forests can retrieve the answer in the story. And the inference process can show to user in the dialog. But there are some function cannot be achieve like the every new story is difficult to add in the random forests and learn the new words relationships from the new stories.

Plan



Reference

[1] J. Ham; Yangchi Chen; M. M. Crawford; J. Ghosh, "*Investigation of the random forest framework for classification of hyperspectral data*", IEEE Transactions on Geoscience and Remote Sensing, Volume: 43, Issue: 3, pp. 492 – 501, March 2005

[2] B. Leo, "*RANDOM FORESTS*" Statistics Department University of California Berkeley, CA 94720 January 2001

[3] B. Leo; C. Adele, "*Random Forests*", Link: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

[4] W. Jason, et al. "*Towards ai-complete question answering: A set of prerequisite toy tasks.*", arXiv preprint arXiv:1502.05698, 2015.

[5] S. Erwan; B. Gerard; V. Jean, "*CONSISTENCY OF RANDOM FORESTS*", The Annals of Statistics, Vol. 43, No. 4, pp. 1716-1741, 2015