Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 116 (2017) 107–112

Procedia Computer Science

www.elsevier.com/locate/procedia

2nd International Conference on Computer Science and Computational Intelligence 2017, ICCSCI 2017, 13-14 October 2017, Bali, Indonesia

Performance Comparison and Optimization of Text Document Classification using k-NN and Naïve Bayes Classification Techniques

Zulfany Erlisa Rasjid[a*], Reina Setiawan[a]

[a]*Computer Science Department, Doctor of Computer Science Department, Bina Nusantara University, Jl.KH Syahdan No. 9, Jakarta 11480, Indonesia*

Abstract

In the current era, information is available in several different formats, such as text, image, video, audio and others. Corpus is a collection of documents in a large volume. By using Information Retrieval (IR), it is possible to obtain an unstructured information and automatic summary, classification and clustering. This research is to focus on data classification using two out of the six approaches of data classification, which is k-NN (k-Nearest Neighbors) and Naïve Bayes. The text documents used is in XML format. The Corpus used in this research is downloaded from TREC Legal Track with a total of more than three thousand text documents and over twenty types of classifications. Out of the twenty types of classifications, six are chosen with the most number of text documents. The data is processed using RapidMiner software and the result shows that the optimum value for k in k-NN occurs at k=13. Using this value for k, the accruacy in average reached 55.17 percent, which is better than using Naïve Bayes which is 39.01 percent.

Peer-review under responsibility of the scientific committee of the 2nd International Conference on Computer Science and Computational Intelligence 2017.

Keywords: k-NN, Naïve Bayes, Text Document Classification, Information Retrieval.

1. Introduction

The rapid development of Information Communication and Technology (ICT) allows the information to be accessed easily and quickly. The information is available under several different formats, such as text, image, video, audio etc. Information Retrieval (IR) is a method to retrieve information (non-structured information) from a corpus as required [1]. IR has been implemented since 1960s and IR is related to uncertainty, context and relevance [23]. Corpus is a collection of documents with a large number of total documents.

**Corresponding Author:*
E-mail address:zulfany@binus.ac.id

1877-0509
Peer-review under responsibility of the scientific committee of the 2nd International Conference on Computer Science and Computational Intelligence 2017.
10.1016/j.procs.2017.10.017

IR application is not only limited to retrieve the information, but also capable to perform an automatic summary, classification and clustering. This research is to focused on basic text document classification in IR. There are six classification methods, which is Rocchio, k-Nearest Neighbors (k-NN), Regression Model, Naïve Bayes and Bayesian nets, Decision Trees and Decision Rules [45]. Previous research regarding classification focused on performance comparison of k-NN, Naïve Bayes and Decision Tree, using k=10 for k-NN [678910]. Most of those researches compare k-NN and Naïve Bayes method in terms of their performance but this research is focused not only on the performance of k-NN and Naïve Bayes. It is also to find the optimal value of k that would provide the best performance. The value of k plays an important role in affecting the performance of k-NN classification [678910].

The objective of this research is to compare two text document classification methods, which the k-Nearest Neighbor (k-NN) and Naïve Bayes and to find the optimal value for k in k-NN. The advantages of k-NN and Naïve Bayes classification methods are easy to understand and implement, computationally short time in training process and noise resistance [9].

The text document used is in the form of XML document. The corpus used in this research is downloaded from TREC Legal Track with a total number of more than three thousand text documents and more than twenty types of classification. Out of the twenty types of classifications, six are chosen with the most number of text documents. The expected result is to obtain a classification technique with the highest level of accuracy.

.

1. Literature Review

2.1 Classification Approach

Classification is analising data extraction using models that describes data classes. A model or classifier is constructed to predict categorical labels. There are several classification algorithms as can be seen in fig. 1 [11].

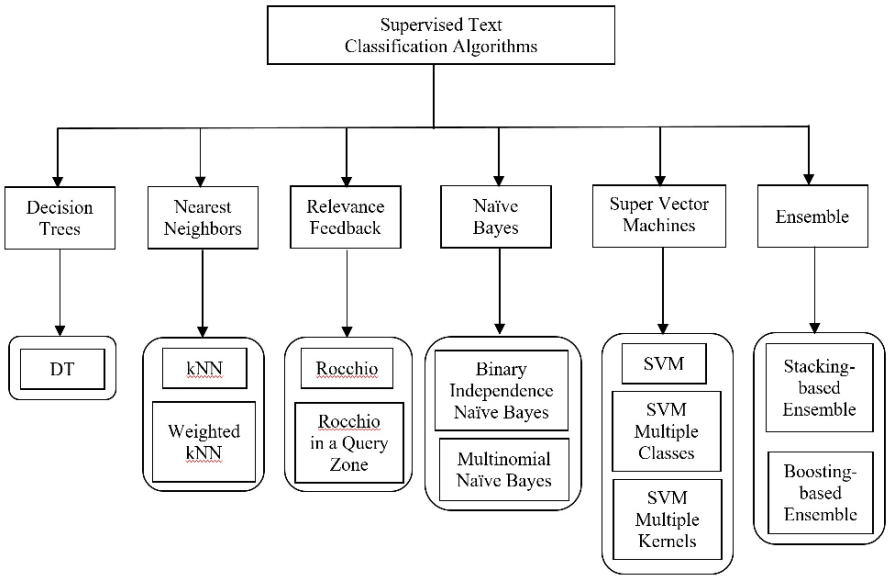

Fig.1 Classification Algorithm

The data label used in this corpus are Practice and Procedure, Trade Practices, Corporations, Migrations, Administrative Law and Bankruptcy. In this legal data, the classifier used is Practice and Procedure, Trade Practices, Corporations, Migration, Admistrative Law and Bankruptcy. Data classification involves two steps. The first step is the learning process. At this step the classification model is created. This is called the learning or training phase Data created from this step is called the “traning data”. The second step is clasification [12]. This step uses the model created in the first step to predict the class labels fo the data. The accuracy of the classifier the percentage of test data that are correctly identified [12]. This paper uses two classification methods, which is K-nearest neighbors (K-NN) and Naïve Bayes. Beside measuring the accuracy of the classification, this research also to provide the optimal value of k in k-NN.