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].

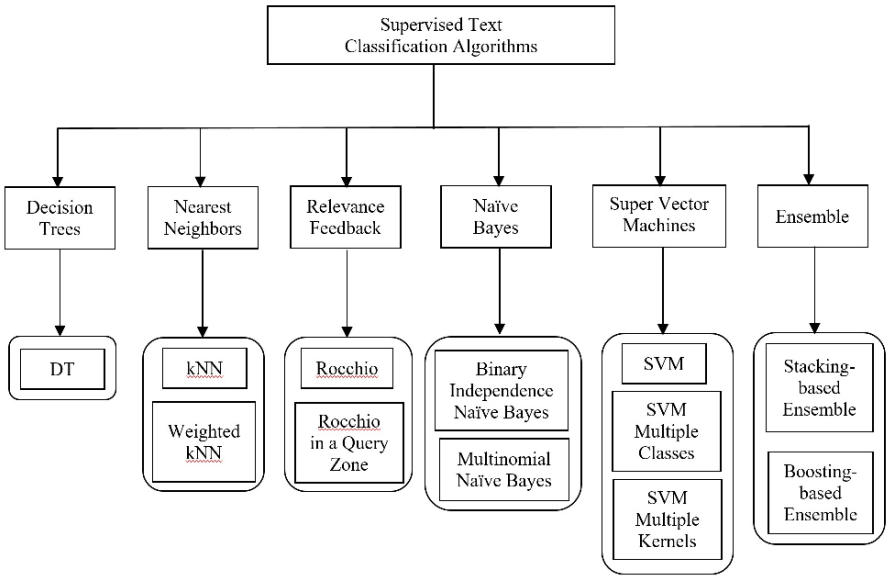


Fig.1 Classification Algorithm

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.

2.2 k-Nearest Neighbor

k-NN uses k-Nearest Neighbor classifier. This method is also known as 'lazy learners'. Introduced for the first time in 1950s and became popular in 1960s. This method compares the similarity of the training records that are nearest to it. The value k represent the number of neighbors being compared. The value of k shows the number of k nearest data that is compared. K-NN algorithm is as follows [13]:

start
select a positive integer k along with new samples
select the k entries in the database is closed to the sample
find the most common classification of these entries
this is the classification that will be given to the new sample
end

The similarity is calculated by distance, such as the Eucledean distance. The distance between two points, let's say X1 and X2, where X1=(x11,x12,x13,....x1n) and X2=(x21,x22,x23.....,x2n) is defined by

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n}(x_{1i} - x_{2i})^2}. \quad (1)$$

Min-max normalization is the applied, to transform a value v of a numeric attribute A to v' in the range of 0 This is defined by:

$$v' = \frac{v - min_A}{max_A - min_A} \quad (2)$$

To determine the best value of k, firstly use k=1 and keep incrementing k .

2.3 Naïve Bayes Classification

Naïve Bayes classification is a probabilistic classifier based on the Bayes Theorem. For each sample, the probability of a class is determined. It is assumed that all attribute are independent [1314]. The classifier choses the classification most similar to V_{nb} with the given attribute $a_1, a_2, a_3, \ldots, a_n$. The formula to calculate V_{nb} is as follows:

$$V_{nb} = argmax_{v_j \in V} P(v_j) \prod P(a_i|v_j) \quad (3)$$

The estimatimate $P(a_i|v_j)$ using m-estimate as follows [15]:

$$P(a_i|v_j) = \frac{n_c + mp}{n+m} \quad (4)$$

Explanation of symbols:
n = the number of training examples for which $v = v_j$
n_c = number of examples for which $v = v_j$ and $a = a_i$
p = a priori estimate for $P(a_i|v_j)$
m = an equivalent sample size
m = the equivalent sample size

2.4 Previous Works

In the last 5 years, the development related to classification coveres classification techniques such as Decision Tree, Naïve Bayes, and k-NN. In 2012 research on Training Algorithms using several methods which is Decision Tree, k-NN, SVM, Bayesian Network, Neural Network. The result shows that there are no algorithm that fulfills all of performance criteria such as accuracy, training speed, classification speed, level of tolerance and the handling of overfitting. Each methods shows a different result for each criteria. For example some criteria shows good result however in other criteria the result is average [10].

In 2013, a research on classification using Naïve Bayes on Big Data shows the level of accuracy of Naïve Bayes is high. The researcher is confident that Naïve Bayes is able to provide a better accuracy if an intelligenet filter is used [8].

In the same year there was a research on comparison of several classification method namely Naïve Bayes, Decision Tree, J48, and k-NN. The research focused on the comparison of accuracy performance. The result shows that J48 which is an algorithm for decision tree is more accurate compared to Naïve Bayes and k-NN [9].

The latest research focused on comparing the accuracy performance and processing speed on Naïve Bayes, k-NN dan Decision Tree. Naïve Bayes shows a better result in accuracy whereas Decision Tree has the best performance in speed[6].

3. Method

Data for the experiment is downloaded from TREC Legal Track with a toatal 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 six chosen topics are Corporation, Migration, Practice and Procedure, Trade Practice, Bankruptcy and Administrative Law. The documents are labeled by topics. The classification method used are k-NN and Naïve Bayes. Out of the data set used in this research, 30 percent is used as training data. Using RapidMiner Software, text classification is performed using Naive Bayes technique and k-NN, where the value of k is chosen from 1 to 25 (odd integers only). The reason of choosing odd numbers as to avoid situations where the result has equal number for positive and negative result, avoiding choosing the wrong class.

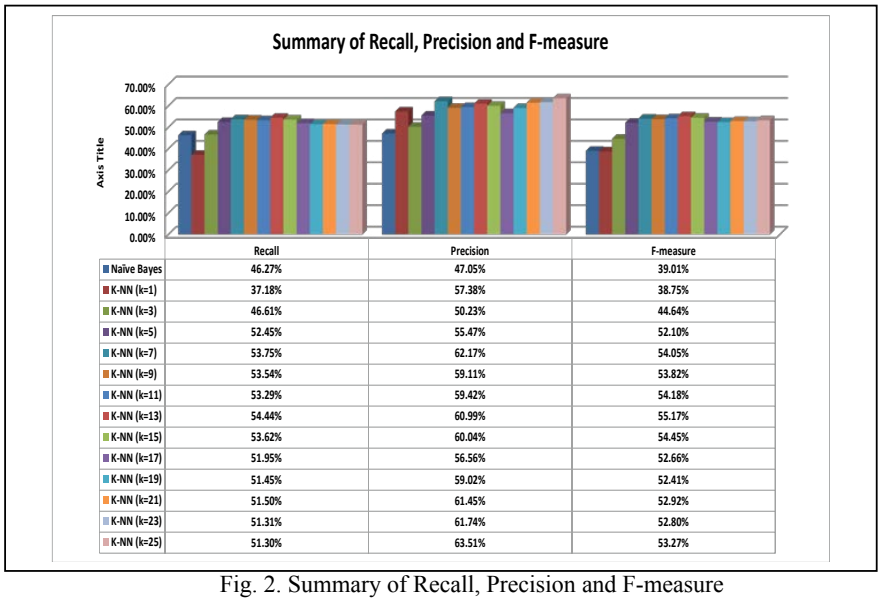
RapidMiner shows the result for Precision, Recall and Accuracy. Recall is the part of instances that was retrieved and relevant, compared to the total instances that are in the image that is relevant. Precision is the instances that are relevant compared to the total instance that was retrieved. In order to analyze the result, it is required to combine the value of precision and recall. In this case, the F-measure [11].

4. Result and Discussions

The performance evaluated in this experiment is Recall, Precision, F-measure and Accuracy. Recall shows the accuracy of classification based on the total number of documents. The result in fig. 2 shows that using Naïve Bayes classifier, the Recall performs better than k-NN with k=1. The result for Recall is at worst using k-NN clasifier with k=1 which is 37.18 percent and Naive Bayes 46.27 percent. This occurs because with k=1, there is only one neighbor selected and this is the main reason why the result is strongly dependent on the distance. With the increase of k, the Recall performance of k-NN becomes better, 46.61 percent for k=3, 52.45 percent for k=5 etc., where all of this values are better than Nive Bayes. From 13 experiments with odd k starting from 1 until k=25, Recall performance of k-NN shows the highest performance at k=13. For k=15, 17, ... the Recall performance becomes stable. This occurs because statistically the more sample used will represent a more realistic figure, in this case the sample is the number of neighbors, which is the value of k.

Precision shows the accuracy of classification based on the documents that was classified. In Precision, k-NN outperforms Naïve Bayes. It can be seen that Naive Bayes has the worst performance (47.05 percent) compared to k-NN with k=1 (57.38 percent) , k=3 (50,23 percent), k=5 (55.47 percent), and other values of k as shown in fig. 2.

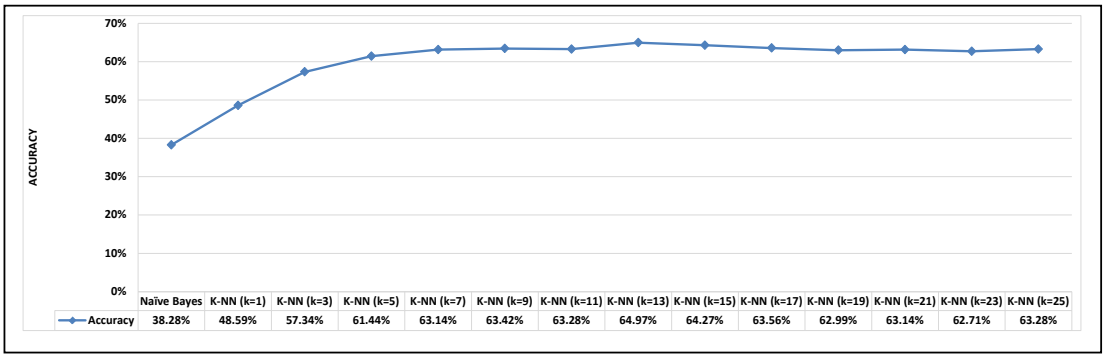
In order to get a combined result, the F-measure is calculated, and the result shows that Naive Bayes classification has the worst performance compare to all k-NN with different values of k. This can be seen in fig. 2 where the value of k=3 (44.64 percent), k=5 (52.10 percent) and the highest occurs at k=13 (55.17 percent). For Naive Bayes the F-measure is 39.01 percent. At k=1 for k-NN, the F-measure is 38.75 percent below the result of Naive Bayes. This condition is due to the fact that the Recall performance is low. Similar to Recall and Precission, starting from k=13, the F-measure becomes stable. The overall performance is quite low, since the contents of the text documents is not significantly specific, for example in the text document, Practice and Procedure label can be claimed as Trade Practices label.



	Recall	Precision	F-measure
Naïve Bayes	46.27%	47.05%	39.01%
K-NN (k=1)	37.18%	57.38%	38.75%
K-NN (k=3)	46.61%	50.23%	44.64%
K-NN (k=5)	52.45%	55.47%	52.10%
K-NN (k=7)	53.75%	62.17%	54.05%
K-NN (k=9)	53.54%	59.11%	53.82%
K-NN (k=11)	53.29%	59.42%	54.18%
K-NN (k=13)	54.44%	60.99%	55.17%
K-NN (k=15)	53.62%	60.04%	54.45%
K-NN (k=17)	51.95%	56.56%	52.66%
K-NN (k=19)	51.45%	59.02%	52.41%
K-NN (k=21)	51.50%	61.45%	52.92%
K-NN (k=23)	51.31%	61.74%	52.80%
K-NN (k=25)	51.30%	63.51%	53.27%

Fig. 2. Summary of Recall, Precision and F-measure

A graph is plotted for the accuracy to see whether an optimum value exists. Fig. 3 shows the graph is ascending at the beginning, and the stationary point is at k=13. At k greater than 13 the graph starts to stabelize. Therefore it can be concluded that the optimal value of k for this data set is k=13. From the F-measure and the accuracy, k=13 gives the most optimal value.



	Naïve Bayes	K-NN (k=1)	K-NN (k=3)	K-NN (k=5)	K-NN (k=7)	K-NN (k=9)	K-NN (k=11)	K-NN (k=13)	K-NN (k=15)	K-NN (k=17)	K-NN (k=19)	K-NN (k=21)	K-NN (k=23)	K-NN (k=25)
Accuracy	38.28%	48.59%	57.34%	61.44%	63.14%	63.42%	63.28%	64.97%	64.27%	63.56%	62.99%	63.14%	62.71%	63.28%

Fig. 3. Summary of Accuracy

5. Conclusion

The calculation of the performance is based on the F-measure. Based on the experiment, k-NN performs better than Naive Bayes with the exclusion of k=1. The F-measure shows 44.64 percent for k=3, 52.10 percent for k=5 and the highest is 55.17 percent for k=13. For Naive Bayes the F-measure is 39.01 percent. Similar to Recall and Precision, starting from k=13, the F-measure becomes stable. The optimal value for k is at k=13. It is recommended to use the value of k greater than 10. The overall performance is quite low, since the contents of the text documents is not significantly specific. RapidMiner software can be used for text classification using Naive Bayes and k-NN approach.

6. Future Works

Further research is necessary in order to assure that the optimal value of k-NN occurs at k greater than 10, regardless of the data set chosen. Different method of classification can also be used to obtain a better result, such using neural network or deep learning.

References

1. Manning, C.D., Raghavan, P., Schutze H. An Introduction to Information retrieval. Online edition of Cambridge UP; 2009.
2. Sanderson M, Croft WB. The history of information retrieval research. Proc IEEE. 2012;100(SPL CONTENT):1444–51.
3. Hyman H, Sincich T, Will R, Agrawal M, Padmanabhan B. A Process Model for Information Retrieval Context Learning and Knowledge Discovery. Artif Intell Law. 2015;
4. Bijalwan, V., Kumar, V., Kumar, Pascual J. KNN based Machine Learning Approach for Text and Document Mining. In International Journal of Database Theory and Application, vol. 7, no. 1, pp. 61-70, 2014;
5. Phyu TN. Survey of Classification Techniques in Data Mining. In: Proceedings of the International MultiConference of Engineers and Computer Scientists. Hong Kong; 2009.
6. Kalavathi KNSP. Performance Comparison between Naive Bayes, Decision Tree and k-Nearest Neighbor. 2015;9359(6):152–61.
7. Patil TR, Sherekar SS. Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. 2013;6(2).
8. Liu B, Blasch E, Chen Y, Shen D, Chen G. Scalable Sentiment Classification for Big Data Analysis Using Na ¨ ıve Bayes Classifier. 2013;99–104.
9. Jadhav SD, Channe HP. Comparative Study of K-NN , Naive Bayes and Decision Tree Classification Techniques. 2016;5(1):2014–7.
10. Bhavsar H, Ganatra A. A Comparative Study of Training Algorithms for Supervised Machine Learning. 2012;(4).
11. Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval: The Concept and Technology Behind Search. 2nd ed. Pearson; 2011.
12. Han J, Kamber M, Pei J. Data Mining Concepts and Techniques. 3rd ed. Elsevier; 2012.

13. Rish I. An empirical study of the naive Bayes classifier. IJCAI-2001 Work Empir Methods AI (also, IBM Tech Rep RC22230), pp 41--46 [Internet]. 2001;(February). Available at: https://sites.google.com/site/irinarish/publications/RC22230.pdf?attredirects=0&d=1&cm_mc_uid=80717800187114522513777&cm_mc_sid_50200000=1452251377
14. Ren J, Lee SD, Chen X, Kao B, Cheng R, Cheung D. Naive Bayes Classification of Uncertain Data. 2009;(60703110).
15. Meisner E. Naive Bayes Classifier example Car theft Example. Comput Linguist [Internet]. 2003;36(2):2–3. Available at: http://www.mendeley.com/research/naive-bayes-classifier-example-car-theft-example/