ISBN 978-1-7281-3832-9

isemantic
International Seminar on Application for Technology of Information and Communication
2019

PROCEEDING

INDUSTRY 4.0; RETROSPECT PROSPECT AND CHALLENGES

21–22 September 2019

Universitas Dian Nuswantoro Semarang

PROCEEDING

isemantic
International Seminar on Application for Technology of Information and Communication
2019

ISBN 978-1-7281-3832-9

Technically Supported by:

IEEE

IEEE Universitas Dian Nuswantoro Student Branch

Listed in ifory

Organized by:

LPPMUDINUS

Computer Science Faculty

Engineering Faculty

PROCEEDINGS

2019 International Seminar on Application for Technology of Information and Communication

(iSemantic)

Industry 4.0: Retrospect, Prospect, and Challenges

September 21st – 22nd, 2019
Universitas Dian Nuswantoro
Semarang, Indonesia

ISBN: 978-1-7281-3832-9
IEEE Catalog Number: CFP19CUE-ART

TABLE OF CONTENT

Classification of Ship-Based *Automatic Identification Systems* Using K-Nearest Neighbors

1st Natalia Damastuti
Engineering Physic Department
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
natalia.damastuti@narotama.ac.id

2nd Aulia Siti Aisjah
Engineering Physic Department
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
auliasa@ep.its.ac.id

3rd Agoes A. Masroeri
Marine Engineering Department
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
masroeri@its.ac.id

Abstract— One of vessel monitoring systems which employs predetermined equipment to discover the movements and activities of vessels is AIS (Automatic Identification System). AIS contains the ship data either static (ship name, ship size, sailing time) or dynamic data (ship speed, rate of turn, ship heading). The ship tracking information system can be accessed by public, but manual monitoring will be difficult to do, given that data is increasingly heterogeneous and complex as well as its volumes increase exponentially. As a result, a more efficient method of data mining and processing are needed. In this study, k-NN algorithm is applied with the aim of classifying ships sailing in Indonesian waters. The algorithm is tested on real time AIS database using k-NN and the neighborhood component analysis (NCA). The result shows that NCA,KNN has higher accuracy than using k-NN on original classifier.

Keywords——Vessel; k-NN; AIS; Neighborhood component analysis; Data Mining

I. INTRODUCTION

Recently, ship monitoring systems has been improving; one of them is the application of Automatic Information System (AIS). This System technology provides information in real time which will greatly help to monitor the ship position, ship activities and maritime supervision. AIS transmits and receives GPS signals, information on ship speed, time and destination. AIS proves its numerous advantages by also enabling the improvement of safety, armada management and navigation[1]. Information emitted by AIS includes kinematic information in the form of ship location, speed, heading, rate of turn, destination and estimated ship arrival and static information in the form of ship code, ship name, ship size and sailing time which will be useful for ship safety and security systems[2].

Data obtained from AIS is increasing every day, for example; the retrieved information of marine traffic sites www.marinetraffic.com [3]shows that every day more than 500 ships sailing in Indonesian waters and such number will increase as represented in the figure 1.

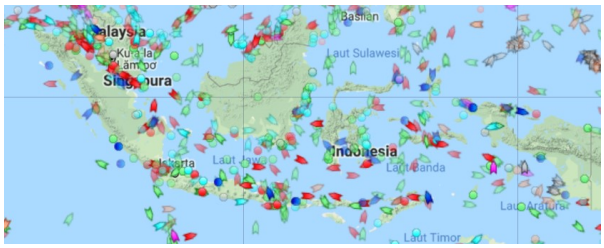


Fig. 1. The presentation of Ship on Marinetraffic

AIS Data which is accepted by the receiving station will be sent to a data center connected by networks and will certainly experience redundant and data intermittency; thus, such error should be reduced. Many studies perform preprocessing of these errors before processing later [4].

The vast number of ships with increasing activities results in the increasing and complicated amount of data received by the server. The management of available databases is done to obtain certain patterns which enable the analysis of each ship characteristics. Consequently, a prediction of the database capabilities will be structured which will improve sailing safety. Increasingly heterogeneous and complex data as well as its volumes which increase exponentially turn into a 'big data.' It describes data volume that is very large, whether it is structured or unstructured, and requires the autonomous introduction of pattern and data relations. The introduction of these patterns will be difficult if done manually and conventionally so that the employment of machine learning is needed.

Only few researches employ AIS database especially machine learning method. K-NN algorithm is a simple algorithm for classification. This paper shows how to learn K-NN classification with and without measurement of neighborhood component analysis. In addition, those will also be compared.

978-1-7281-3832-9/19/$31.00 ©2019 IEEE

II. PROPERTIES OF AIS

AIS is an electronic equipment as a navigation system for sea transportation. The competences of AIS are the capacity to identify the location of the ship's sailing, and the ability to exchange data electronically, such as the position, activity, condition or speed of the ship, with other nearby vessels and the Vessel Traffic Services (VTS) station as seen on figure 1[1]. This is also a communication system used on ships and VTS or shipping ship traffic . International Maritime Organization (IMO) and International Convention for Safety of Life at Sea (SOLAS) require the use of AIS on international shipping ships of ≥ 300 GT, and passenger ships of all sizes. AIS is an autonomous communication tool between ships. The principle works of it is a ship sends data to another ship which is equipped with AIS within VHF range[5][6]. AIS as communication system, between ships, or ships to land stations and vice versa, will be able to provide information and identification of ship movements throughout its voyage [6]. Figure 1 describes the data recorded in AIS which consists of statistical information, dynamic information as well as the route of a ship information.

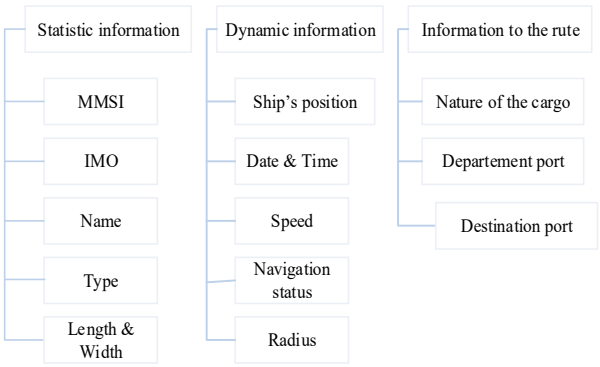


Fig. 2. AIS attributes

III. METHODOLOGY

A. Dataset

The data sources used in this study are primary data collected in real time from www.marinetraffic.com starting on October, 2018 to January, 2019. These data are raw data emitted by AIS signal. Real time data collection is carried out every once an hour. The chosen research location is ships that cross or are sailing in Indonesian waters with latitude N06 ° 47'43.92 - S14 ° 03'04.79 and longitude E094 ° 42'07.73 - E143 ° 39'26.01. Obtained information from AIS data includes the following:

TABLE 1. SAMPLE OF RAW AIS DATA

SHIPNAME: [SAT-AIS]	ELAPSED": "639	LAT	0
SHIP_ID": "TWpNd016QTJNak13TXpBMk1qT	COURSE": "334	SPEED	112
LON: 106.6437	LEGEND": "1"	HEADING	334
STATUS_NAME": Underway using Engine	SHIPTYPE : 7	WIDTH	58

In general, AIS dataset that has been recorded is 2,723 stored data in the form of raw data files. Each raw data file contains lon, speed, heading, width, lat, ship_id. Table 1 is illustrates a portion of raw data.

B. Data Selection

Data obtained from AIS is raw data that still uses many attributes such as static information and dynamic information. The normalizatin of the data is intended to change 'json' form into 'csv' form to make it easier to process. As the purpose of this study, not all attributes are used. In this study several attributes will be chosen with the intention of the type ship classification. Table 2 informs the AIS data that has been normalized and selected into 6 columns consisting of 6 attributes, namely shipname, DWT, Type-Name, Width, Lenght, Ship_Type.

TABLE 2. SAMPLE OF SELECTION DATA

Shipname	DWT	Type Name	Width	Length	Ship type
HL PIONEER	179655	Cargo	45	291	7
ALPINE MADELEINE	49999	Tanker	32	183	8
OSAM MANILA	682	Special Category	11	40	3
PASAMAN	17948	Tanker	27	157	8
TANGGUH SAGO GAS	84484	Tanker	44	288	8
ETHEREAL EKA	4771	Tanker	17	99	8
SAMUDRA 501	597	Special Category	12	45	3
SOLAR FRONTIER	179259	Cargo	45	292	7
PAN COSMOS	208329	Cargo	50	299	7
YAMATO SPIRIT	107617	Tanker	42	243	8

Afterward, selection of attributes that will be used in the classification is carried out; attributes that are in accordance with table 3. Data with DWT, WIDTH and Length attributes will be analyzed through a learning process.

In this study, machine learning was used by using k_NN and NCA-kNN.

TABLE 3. SAMPLE OF SECOND SELECTION DATA

DWT	Type_Name	Width	Length
179655	Cargo	45	291
49999	Tanker	32	183
682	Special Category	11	40

17948	Tanker	27	157
84484	Tanker	44	288
4771	Tanker	17	99
597	Special Category	12	45
179259	Cargo	45	292
208329	Cargo	50	299
107617	Tanker	42	243

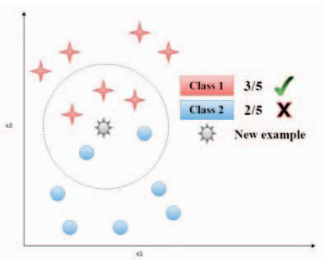
The obtained datasets from that area cover 2,723 ships. Figure 5 shows the distribution of data based on the types of vessels obtained in real time AIS data over a period of 4 months. There are 6 types of vessels that can be detected, namely 'Cargo', 'Tanker', 'Special category', 'Passenger', 'Wing in ground' and 'High-Speed Craft'. The classification of the types of vessels is grouped based on 3 attributes, namely DWT, WIDTH and LENGTH. These attributes will be processed using Nearest Neighbor learning. Datasets that have been grouped according to the types of vessels show the amount of Cargo is 1752, Tanker is 692, Special Category is 177, Passenger is 92, Wing in Ground is 5 and High-Speed Craft is 5.

C. K-Nearest Neigbor and Neigborhood Component Analysis

K-Nearest neighbor is a machine learning method that does not require many training parameters. Beside, the complexity of the computer for this method is low but its performance is satisfactory [7]. For this reason, this paper employs this algorithm concept. The K-NN algorithm is quite simple by applying 2 steps only; the first step is looking for the value of the nearest neighbor, and the second step is classifying the data points into certain classes [8]. Equation (1) is equation for finding neighbors by measuring the distance.

$$Euclidean\ distance\ (x,y) = \sqrt{\sum_i (x_i - y_i)^2} \quad (1)$$

Fig. 3 show the architecture of k-NN clasifier. There are 2 classes, namely clas 1 (red) and class 2 (blue) and the k-value is 5. Among these 5K-value, there are 3 data that resemble class 1 and 2 data that resembles class 2. Classification is carried out on new sample in the class which can be identified as class 1[8].



Fiq. 3. Illustration of the k-NN

We also employs neighborhood component analysis (NCA) calculation to compare the two processes. The later process is a symmetric positive semi-definite matrix and is estimated using inverse square roots [9]

NCA process is a symmetric positive semi-definite matrices and estimated using inverse square roots[10]. The purpose of NCA is to find optimum results from linear transformations matrix n_component, n_feature which maximizes the sum of all samples i from the probability of pi classified i correctly following Eq.2 and eq.3 with N is the number of n-samples and pi is the probability of samples i according to the nearest rule neighbor.

$$\arg\max_L \sum_{i=0}^{N-1} p_i \quad (2)$$

$$p_i = \sum_{j \in C_i} p_{ij} \quad (3)$$

where Ci is the set of points in the same class as sample i, and pij is the normalized exponenitial function to Euclidean distances in the embedded space:

$$p_{ij} = \frac{\exp\left(-\|A_{xi} - A_{xj}\|^2\right)}{\sum_{k \neq 1} \exp\left(-\|A_{xi} - A_{xk}\|^2\right)} \quad (4)$$

IV. EXPERIMENT AND EVALUATION

This chapter demonstrates the experiment of database classification using attributes selection namely dwt, width and length. Experiment is carried out with classification learning using K-nearest neighbor calculation.

The Euclidean distance measures the distance between training data and testing data. This selects sample k from the training data, and then chooses the highest number data of their class where the selection of k will put off them from ambiguous values [9].

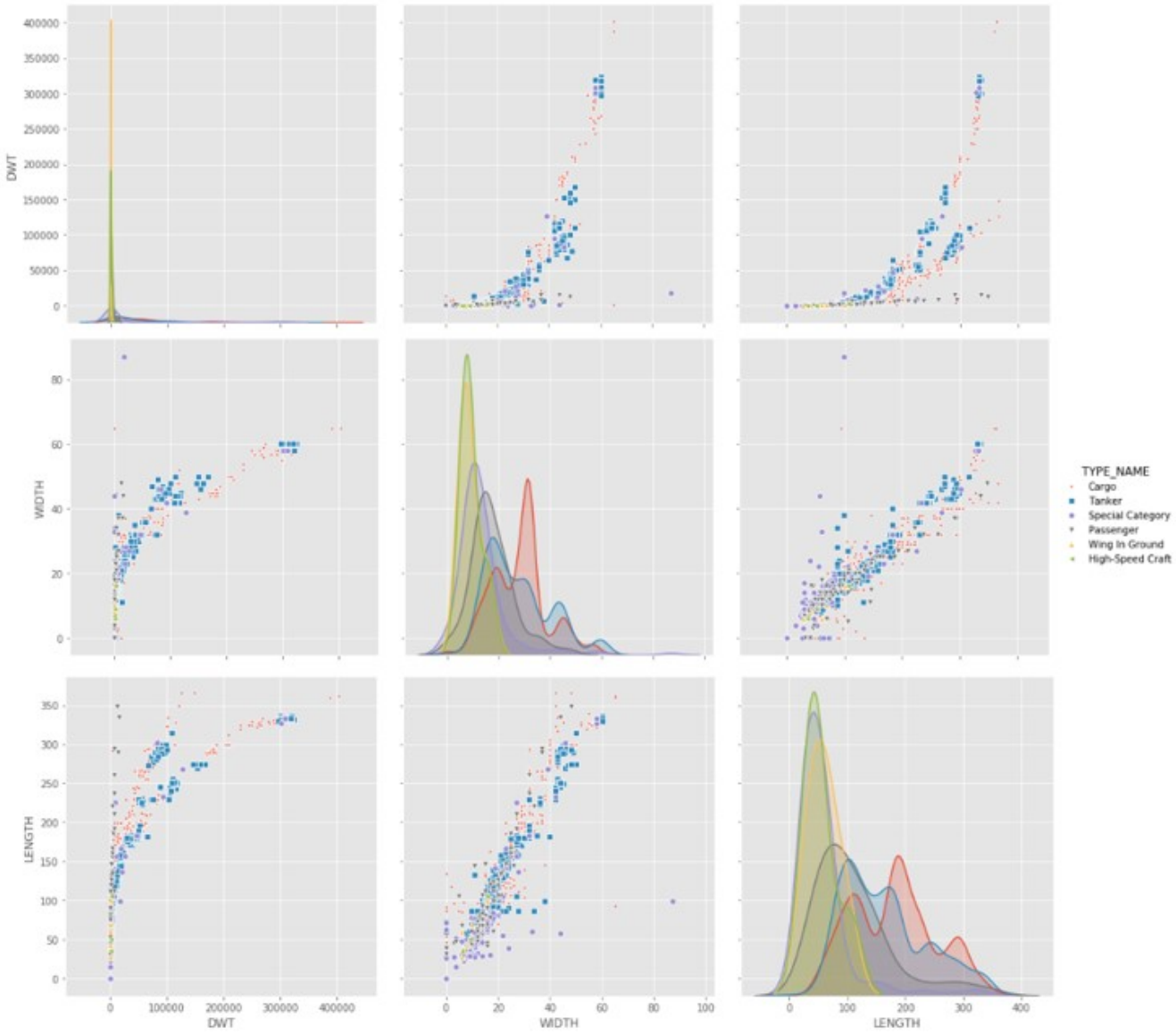
In this experiment, the used dataset is 2,723 ships. Fiq 5 show the distribution and histogram of vessel type from the data selected. Ship distribution based of DWT atribut, Weight atribut and Length atribute. Tables 4 and 5 are examples of datasets in real time in which time series used by the dataset divided into 2, Data for training as much as 60% or equal to 1633 ships while the remaining 40% becomes testing data.

TABLE 4. SAMPLE OF TRAINING DATA (60%)

DWT	WIDTH	LENGTH
172	9	29
14348	21	134
15370	44	335
13517	24	141
18793	32	199
4200	18	103
17828	32	179

DWT	WIDTH	LENGTH
5705	18	116
2368	10	74
76318	32	225
6792	16	114
37085	28	177

TABLE 5. SAMPLE OF TESTING DATA (40%)

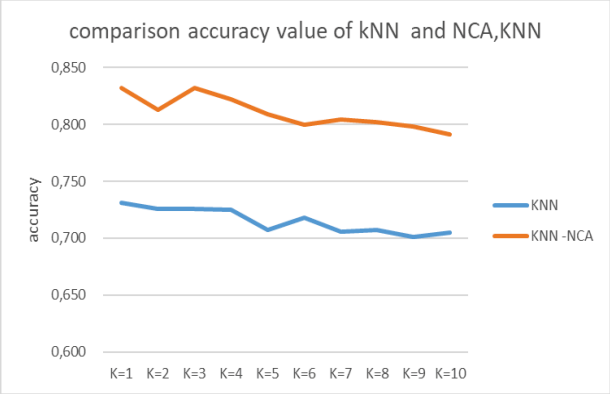


Fiq. 5. Scaterplot to vizualize the distribution of label point in 2D graph using DWT, WIDTH and LENGTH

k-NN algoritms is used by eq. 1, where y is the testing data and x is training data. Tables 4 and 5 are examples of datasets for training data and testing data. Data for training as much as 60% or equal to 1633 ships while the remaining 40% becomes testing data. This research applies k = 1 to k = 10. K value is used in K-NN on original and NCA,KNN.

TABLE 6. ACCURACY RESULT FOR EACH K-VALUE

k-Value	k-NN	k-NN, NCA
1	0.731	0,832
2	0,726	0,813
3	0,726	0,832
4	0,725	0,822
5	0,707	0,809
6	0,718	0,800
7	0,706	0,804
8	0,707	0,802
9	0,701	0,798
10	0,705	0,791



Fiq. 6 the comparison the result of accuracy between kNN and NCAKNN

Table 6 shows a comparison of the results of accuracy for each calculation using k-NN on original and NCAKNN algorithms. After making a comparison, a conclusion is obtained. it shows that k-NN with neigborhood component analysis (NCA) has a higher accuracy result than using k-NN on original. With the selection of k = 1 to k = 10, the accuracy result is respectively 83.2%, 81.3%, 83.2%, 82.2%, 80.9%, 80%, 80.4%,. 80.2%, 79.8% and 79.1%. The accuracy result by employing lower k-NN is respectively 73.1%, 72.6%, 72.6%, 72.4%, 70.7%, 71.8%, 70.6%, 70,7%, 78.7% and 70.5%.

V. CONCLUSION

This paper uses AIS database which has been obtained in real time from October 2018 until January 2019. The data is organized based on 3 attributes, namely DWT, WIDTH and LENGTH to classify the vessel types. This research experiment provides a comparison between the sole use of k-NN algorithm and the employment of k-NN along with NCA as an alternative for calculating the nearest neighbor. By employing k value of k = 1 to k = 10, the experiment shows the accuracy result for each k of NCA,KNN algorithm has a higher value than using the k-NN on original algorithm .

ACKNOWLEDGEMENT

This research is conducted as part of manifestation of the authors' passion to share the knowledge to society. So, it is fully funded by the authors. No other parties are involved in funding this experiment.

VI. REFERENCES

[1] A. Serry, "(AIS): A DATA SOURCE FOR STUDYING To cite this version : HAL Id : hal-01724104 CONFERENCE Technological , Innovation and Research," 2018.

[2] S. Mao, E. Tu, G. Zhang, L. Rachmawati, E. Rajabally, and G.-B. Huang, "An Automatic Identification System (AIS) Database for Maritime Trajectory Prediction and Data Mining," 2017.

[3] M. Trrafic, "The Marine Traffic Website," *http://www.marinetraffic.com/*, 2019. .

[4] G. Pallotta, M. Vespe, and K. Bryan, "Vessel pattern knowledge discovery from AIS data: A framework for anomaly detection and route prediction," *Entropy*, 2013.

[5] Shilavadra Bhattacharjee, "Automatic Identification System (AIS): Integrating and Identifying Marine Communication Channels." [Online]. Available: https://www.marineinsight.com/marine-navigation/automatic-identification-system-ais-integrating-and-identifying-marine-communication-channels/.

[6] B. J. Tetreault, "Use of the Automatic Identification System (AIS) for Maritime Domain Awareness (MDA)," *Proc. MTS/IEEE Ocean. 2005*, vol. 2005, 2005.

[7] Y. Cai, D. Ji, and D. Cai, "A KNN Research Paper Classification Method Based on Shared Nearest Neighbor," *NTCIR-8 Work. Meet.*, pp. 336–340, 2010.

[8] A. Bablani, D. R. Edla, and S. Dodia, "Classification of EEG data using k-nearest neighbor approach for concealed information test," *Procedia Comput. Sci.*, vol. 143, pp. 242–249, 2018.

[9] V. Rodriguez, "Breast Cancer Prediction with K-Nearest Neighbor Algorithm using Different Distance Measurements by," no. December 2018, pp. 0–36, 2019.

[10] J. Manit and P. Youngkong, "Neighborhood components analysis in sEMG signal dimensionality reduction for gait phase pattern recognition," *IB2COM 2011 - 6th Int. Conf. Broadband Commun. Biomed. Appl. Progr.*, pp. 86–90, 2011.

COPYRIGHT

IEEE Catalog Number: CFP19CUE-ART
ISBN: 978-1-7281-3832-9

Additional copies of this publication are available from
Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
+1 845 758 0400
+1 845 758 2633 (FAX)
email: curran@proceedings.com