terjemah-format isemantic full rev2 2019 (1).doc

WORD COUNT 2578
TIME SUBMITTED 25-MAY-2019 11:02AM
PAPER ID 47350260

Classification of Ship-Based *Automatic Identification Systems* (AIS) Diversity Using K-Nearest Neighbour

Abstract— One of vessel monitoring systems which employs predetermined equipment to discover the movements and activities of fishing 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 the diversity of ships sailing in Indonesian waters. The algorithm is tested on real time data based on AIS data using k value of 1 to 10. The evaluation shows optimal accuracy at k value of 1 with an accuracy of 0.73

Keywords——Vessel; k-NN; AIS; Monitoring; Data Mining

I. INTRODUCTION

The monitoring system of fishing vessel has been used for long time, one of them is the use of Vessel Monitoring System. In Indonesia, the Ministry of Maritime Affairs and Fisheries has issued a regulation that every ship, especially fishing vessels,, exceeding 30 GT operating in the Republic of Indonesia Fisheries Region and on the high seas, shall install SPKP transmitters (*Sistem Pemantauan Kapal Perikanan* /Fisheries Ship Monitoring System) (Ministry of Maritime Affairs and Fisheries, 2018). Another monitoring system is AIS (Automatic Identification Systems). Vessel which is equipped with AIS devices can automatically transmit and receive various data information about the surrounding vessels, either in the form of a display on the radar screen, or electronic maps (Electronic Navigation Chart - ENC or Electronic Chart Display and Information System - ECDIS). In addition to sending and receiving data information, ships equipped with AIS can also monitor and track the movements of other vessels which are also equipped with AIS (at VHF range) (Vieira, 2016a). The use of AIS is required for vessels of 300GT and for cargo ships and passenger ships of 500GT sailing in national waters.

The intelligence system in navigation also uses AIS to improve maritime security because of lower costs compared to human resources. AIS is an effort to achieve maritime security and ship traffic systems that have been implemented by IMO (International Maritime Organization); AIS loads ship information which contains ship location, ship speed, ship heading, rate of turn, ship destination and estimated mooring vessels, and static data which includes ship name, ship size and sailing time. Such information is important for shipping purposes, for example the prediction of the ship's trajectory and the prevention of ship collisions (Mao, 2017a). Ships equipped with AIS are required to keep AIS operating without interruption, unless there is an international agreement regarding the rules or standards of navigation information services. A country where its flag is flown (Flag State), may give an exception for the vessels to be exempt of carrying AIS if the ships are not to be operated forever, two years after the enactment of AIS provisions (Capt. Hadi Supriyono, 2012). The development of technology enables public to be able to monitor shipping activities in real time. Various applications are developed as efforts to track the position of the ship through available site or application, one of them is www.marinetraffic.com. People can easily see pictures of ships in the world's marine waters through maps as shown below.

Fig. 1. The presentation of Ship Tracking on Marinetraffic

Based on the picture above, the data generated from a tracking site will be difficult to monitor at any time, considering that data is increasingly heterogeneous and complex as well as its volumes increase exponentially. As the result, the information becomes a 'big data' that describes very huge volume of data, whether structured or unstructured, which in the end requires the introduction of a separate pattern and data relation. The introduction of these patterns will be difficult if done manually and conventionally, so accessing data quickly and easily is needed to know the patterns and relations of a data. This method is called Data Mining. Several steps in a data mining process are exploration of data, identification of data patterns and dissemination of data (Ramageri, 2011). One of many techniques and algorithms that can be applied is *k-nearest neighbor*. This method is the most popular classification techniques in which additional data is not needed and classification rules are generated from training data ((Jabbar, 2013a).

In this study, a methodology is carried out to explore a number of data and to conduct a classifications based on data obtained from AIS data. By using k-nearest neighbor algorithm, a classification of ships sailing types in Indonesia will be accomplished based on 3 attributes, they are ship weight (DWT), ship length and ship width

II. LITERATURE REVIEW

A. *Automatic Identification System (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. 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.

Required electronic information system to be installed on ships exceeding 300 GT is AIS (Automatic Identification system). IMO (International Maritime Organization) also requires the use of AIS as a safety tool at sea. 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 (Mund, Ray A .; Campbell, 2005).

The emergence of AIS (Automatic Identification System) technology since 1980s as a communication system has functions to prevent collisions, vessel traffic services and search and rescue, and can also be used to monitor shipping in Indonesian Sea. Previous research mentioned the existence of an information technology for navigation services called MCCST (Monitoring & Control in Sea Transportation); it is information systems supporting AIS data to regulate shipping traffic in ports (Aulia Siti Aisjah et al., 2012). MCST algorithm testing on a laboratory scale is conducted by using a computer as a substitute for AIS on the ship. The testing indicates that MCST is able to monitor ship movements while providing navigation services on ships.

The data sent by AIS has two types, namely static data and dynamic data. Some of AIS's weaknesses become the background of several studies conducted by other researchers, one of them is carried out by Aisjah et al. (Aulia Siti Aisjah et al., 2012).

B. *Machine Learning*

Machine learning is a technique for inference to data with a mathematical approach. The core of machine learning is making a model (mathematically) that reflects data patterns (Putra, 2018). There are several algorithm methods in this machine learning, they are unsupervised learning and supervised learning. Unsupervised learning can be illustrated as a clustering process like Figure 2.1.

Fig. 2. Unsupervised learning methods framework

Unsupervised learning has no desired output (there is no teacher, no example). The search of the original distribution of data q (x) is based on several data samples. Learning is conducted by optimizing p (x | w) which optimizes w parameter. The difference between estimation and original function is called generalization loss. In another hand, in supervised learning, some data is used as learning so that the next data becomes test data. Such can be used as predictions for the upcoming data.

C. *K-NN*

The k-NN method works by searching for a number of data objects or patterns (of all existing training patterns) that are closest to the input pattern, then selecting the class with the highest number of patterns among k patterns. K-NN classifies patterns based on distance, similarity or dissimilarity, depending on attributes (Suyanto, 2017). The stages of K-Nearest Neighbor (K-NN) algorithm are explained as follows:

- Preparation of training data and test data .
- Determination of k value
- Calculation of the distance of test data to each training data as in (1)
- Determination of k value in training data that has the closest distance to test data
- Verification of labels from the nearest training data
- Determination of labels with the most frequency
- Entering test data into the class with the most frequency
- Classification of data.

$$d_{(x1,x0)} = \sqrt{(x_{i1} - x_{i1})^2 + (x_{i2} - x_{i2})^2 + \cdots \ldots \ldots + (x_{ip} - x_{ip})^2} \quad (1)$$

III. RESEARCH METHODS

Generally, the methodology used in this study can be described as in Figure 3:

Fig. 3. Research methods framework

A. *Properties of AIS Data*

The data sources used in this study are primary data collected in real time from www.marinetraffic.com starting on October 11, 2018 to January 2, 2019. These data are raw data emitted by AIS signal which installed on each ship of > 300GT. 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:

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

Fig. 4. Example of AIS Data

B. *Raw Data Cleaning and Selection*

The obtained data is raw data that still contains inconsistencies. To get good information, the data must be normalized by ignoring *tuple* or data that do not have attributes, which means empty attributes are not used.

The obtained AIS data is large both in terms of number and dimensions, so data selection is needed. Data selection is conducted so that the acquired information can be identified and utilized according to the research objectives. In this study, data selection is conducted based on 3 main attributes they are DWT, ship width and ship length and 1 secondary attribute called Type Name

	DWT	TYPE_NAME	WIDTH	LENGTH
0	179655	Cargo	45	291
1	49999	Tanker	32	183
2	682	Special Category	11	40
3	17948	Tanker	27	157
4	84484	Tanker	44	288
5	4771	Tanker	17	99
6	597	Special Category	12	45
7	179259	Cargo	45	292
8	208329	Cargo	50	299
9	107617	Tanker	42	243

Fig. 5. Sample of data selection

The table above is a data selection by choosing attributes that will be used in the classification process. AIS data contains at least 19 attributes of a set of ships as shown in table 1. Afterward, the selection process is carried out into the main attributes of DWT, TYPE_NAME, WIDTH and LENGTH with a population of 2,723 vessel data. Based on the data collected, there are 6 types of ships: 'Cargo Ship', 'Tanker' Ship, 'Special Category' Ship, 'Passenger' Ship, 'Wing in Ground' Ship and 'High-Speed Craft' Ship

C. *Processing*

After the data selection is done, the next process is to get a classification model. In this study, the classification model is established based on learning techniques automatically applied to a set of data that are expected to produce a classification model by dividing them as training data using DWT, WIDTH and LENGTH attributes and some of them become test data. The method used in this study is k_Nearest Neigbour (k_NN). The learning process is obtained from the nearest neighbor data which is learned from training data with a combination of 3 attributes (DWT, WIDTH and LENGTH).

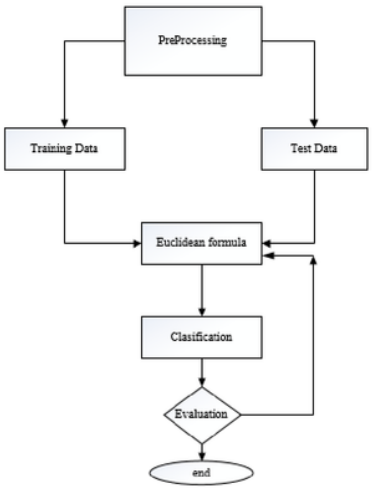


Fig. 6. Systems Flowchart

There are 2,723 ships that have been analyzed. 60% of the total data will be used as training data and the remainder is used as validation data on the k-NN algorithm. DWT as one of the employed initial attributes means Deadweight tonnage of ship. Furthermore, Width means wide measurement of the ship. And lastly, Length is the measurement lengthwise of the ship. Based on these 3 attributes, classification is done by k-NN algorithm. Classification settings are based on the calculation of nearest point / neighbor K, the closest points which form group according to the studied data. This classification uses Euclidean calculations which are used in training data. Next step is determining k value. K value is free, which means it can be determined at any value with the maximum limit of the amount of training data. This study

employs k value of 1 - 10 and the closest distance calculation uses Euclidean formula

D. Evaluation

Accuracy values are obtained by comparing training data with test data; the variables used in evaluating this are precision, recall, f1-measure. Recall is the success rate of recognizing a class that must be recognized. Precision is the level of accuracy of the classification results of all documents. F1-measure is a value that represents the overall performance of the system and is a combination of recall and precision values.

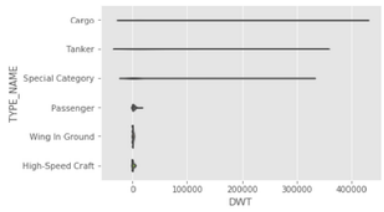
IV. RESULTS AND DISCUSSION

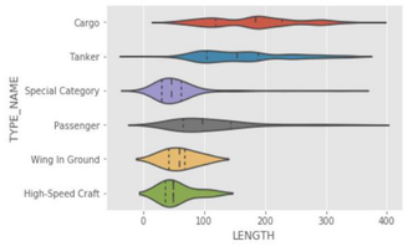
Based on preliminary data processing that has been carried out to obtain a number of information and patterns from AIS data mining, there are several types of vessels, namely 'Cargo Ship', 'Tanker' Ship, 'Special Category' Ship, 'Passenger' Ship, 'Wing in Ground' Ship and 'High-Speed Craft' Ship. The numbers of each type along with the distribution are displayed in the following table

	SHIPNAME	DWT	TYPE_NAME	WIDTH	LENGTH	SHIPTYPE
0	HL PIONEER	179655	Cargo	45	291	7
1	ALPINE MADELEINE	49999	Tanker	32	183	8
2	OSAM MANILA	682	Special Category	11	40	3
3	PASAMAN	17948	Tanker	27	157	8
4	TANGGUH SAGO	84484	Tanker	44	288	8
5	GAS ETHEREAL	4771	Tanker	17	99	8
6	EKA SAMUDRA 501	597	Special Category	12	45	3
7	SOLAR FRONTIER	179259	Cargo	45	292	7
8	PAN COSMOS	208329	Cargo	50	299	7
9	YAMATO SPIRIT	107617	Tanker	42	243	8

Fig. 7. Dataset selection

Table above is a dataset selection. There are several attributes including ShipName, DWT, Type_Name, Width, Length and ShipType with a total data of 2,723 ships. The applied classification process is variations in the attributes of DWT, Length and Width.





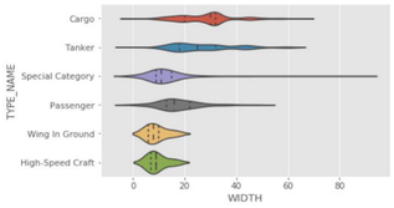


Fig. 8. Histogram of diversity of vessel

K-NN method works by finding a number of k data objects from all training data at which data is divided randomly into k sections. Then, each will be classified with 3 attributes needed, namely DWT, WIDTH and LENGTH. Based on these attributes, classification of vessel types will be carried out. As an example of classification, if DWT = 4.9999, length = 284, width = 32, the ship can be classified into the type of cargo ship. The selection of k is done from grades 1 to 10 which will then be calculated for accuracy. The amount of all data is 2,723 types of ships, 60% or 1,633 will be used as training data and the rest will be test data.

As an evaluation material, an algorithm will be shown by following the level of accuracy as follows:

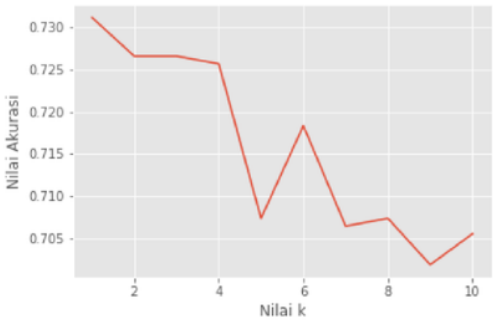
Table 1. Accuracy of each k value

K value	accuracy	K value	accuracy
k=1	0.73	k=6	0.70
k=2	0.71	k=7	0.69
k=3	0.72	k=8	0.68
k=4	0.71	k=9	0.68
k=5	0.70	k=10	0.68

The level of accuracy can be detected as follows (Rohman, 2015)

Accuracy 0.90 – 1.00 = Excellent classification
Accuracy 0.80 – 0.90 = Good classification
Accuracy 0.70 – 0.80 = Fair classification
Accuracy 0.60 – 0.70 = Poor classification
Accuracy 0.50 – 0.60 = Failure

Based on the level of accuracy, it is known that by using a low k value, the value of accuracy is higher. The graph below is a comparison of the employed k value and the value of accuracy.



This study presents the use of k-NN algorithm for the classification process. AIS-based real-time data obtained from a *trafficmarine* site obtained 2,723 ships that were crossing Indonesian waters during November-December 2018. The data then selected based on three attributes of ship weight (DWT), ship length and ship width. With k-NN method, 6 classifications of ship types have been obtained, namely 'Cargo Ship', 'Tanker' Ship, 'Special Category' Ship, 'Passenger' Ship, 'Wing in Ground' Ship and 'High-Speed Craft' Ship. By determining the k value of 1 to 10, it is found that the most optimal k value is 1 with data accuracy of 0.73.

VI. References

terjemah-format isemantic full rev2 2019 (1).doc

ORIGINALITY REPORT

4%

SIMILARITY INDEX

PRIMARY SOURCES

1	www.tandfonline.com Internet	36 words — 1%
2	Nur Azizah Vidya, Mohamad Ivan Fanany, Indra Budi. "Twitter Sentiment to Analyze Net Brand Reputation of Mobile Phone Providers", Procedia Computer Science, 2015 Crossref	35 words — 1%
3	parlinfo.aph.gov.au Internet	14 words — 1%
4	efficiensea.org Internet	11 words — < 1%
5	Kalita, . "An Overview of Machine Learning Methods", Network Anomaly Detection, 2013. Crossref	8 words — < 1%
6	www.myboatsgear.com Internet	8 words — < 1%

EXCLUDE QUOTES OFF
EXCLUDE MATCHES OFF
EXCLUDE BIBLIOGRAPHY OFF