TEST Engineering & Management

March - April 2020
ISSN: 0193-4120 Page No. 151 - 158

Fault Classification in Transmission Lines using K-Nearest Neighbour Approach

Qi Wei, Tan , Chee Yong, Lau and Jacqueline, Lukose
School of Engineering, Faculty of Computing, Engineering and Technology, Asia Pacific University of Technology & Innovation, Kuala Lumpur, Malaysia

Article Info
Volume 83
Page Number: 151 - 158
Publication Issue:
March - April 2020

Article History
Article Received*: 24 July 2019*
Revised*: 12 September 2019*
Accepted*: 15 February 2020*
Publication*: 12 March 2020*

Abstract
In the generation of pursuing to industrial revolution 4.0, there are major changes in most of the industrial sectors relating with technology and engineering to adapt the revolution. With these changes made, the utilities companies are being affected due to the increase in demand of electrical power. With the importance of electrical power, the reliability of the grid system is to be expected as ideal with minimal fault. In any case of fault occurrence, the grid system should be able to classify a fault efficiently to progress into protection coordination to minimize the effect of fault occurred. The issue is being studied and a machine learning model had been developed in respond to the issue. The model produces an accuracy of 93.9% in fault classification using k-Nearest Neighbor approach.

INTRODUCTION

Electrical energy has become a necessity of the public from a luxury of the riches since the last few decades with the evolution of technology. The usage of electrical energy had been increasing ever since due to the everchanging industries and growth of the population. With the addition of more power plants to generate more electrical energy to meet the demand of the growing population, reliability of the system is also another key factor to be heavily monitored to uphold the efficiency of the power grid system.

This paper focuses on the fault classification system using machine learning algorithm, k-Nearest Neighbour (kNN) approach for monitoring purposes for the reliability of the system which isdependent on a crucial protection coordination among the relay devices to minimize the impact of a fault.

LITERATURE REVIEW

Short Circuit faults (Shunt Faults)

Shunt faults which are also labelled as short circuit faults, are recorded to occur more frequently compared to series fault. Conventional precautions are taken for detection and circuit isolation purposes such as relaying systems. The characteristics of shunt faults are described as the surge in current, drop in frequency and voltage sag. (Gururajapathy, et al., 2014) Possible causes of shunt faults consist of lightning, overload and aging or faulty equipment, where the insulation decreases between phases and ground.

Shunt faults are described as any point of different phase is shorten to ground or other phases due to a failure of insulation between the two points, resulting an impedance much lesser than the line impedance. (Goh, et al., 2017) Shunt faults can be recognized in different manners

March - April 2020
ISSN: 0193-4120 Page No. 151 - 158

such as Single-Line to Ground (SLGF), Double-Line to Ground (DLGF), Line-to-Line (LLF) and Three-Phase to Ground and Three Phase itself. (Wani & Singh, 2016) According to researches, the most common shunt fault occurrence is the SLGF with an average percentage of 77% followed by LLF, DLGF and Three-phase fault. Figure 2.2 shows the different type of short circuit that may occur in the transmission lines. (Gururajapathy, et al., 2014)

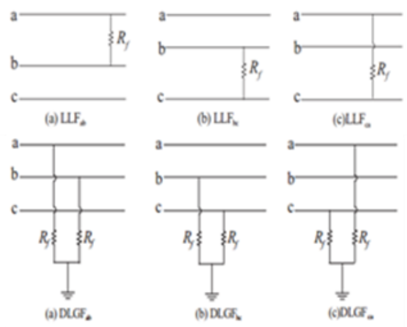

Figure 1: Types of Short Circuit Faults

Impedance Based Method

Impedance Based method is explained as focusing on the corresponding parameters during the incidental moment of fault, voltage and current. In this methodology, the output of the calculations is being output as impedance per unit length, in comparison with other methodologies, this technique is labelled as the least complex and less time taken for detection. (Goh, et al., 2017) Other than Goh et al., researchers Gurujapathy et al. (2017) generate similar concept in their review on fault detection stating that Impedance Based method is economical compared to others. Moreover, provided with a more detailed explanation of this system where the detectors are located as “nodes” in the system to monitor the instantaneous values and current to identify any abnormalities through impedance calculations. With the analysis of circuits, the line impedance is obtained using Nodal analysis: - (Gururajapathy, et al., 2014)

$$f_d = \frac{V_f}{I_f \times Z_l} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.1)$$

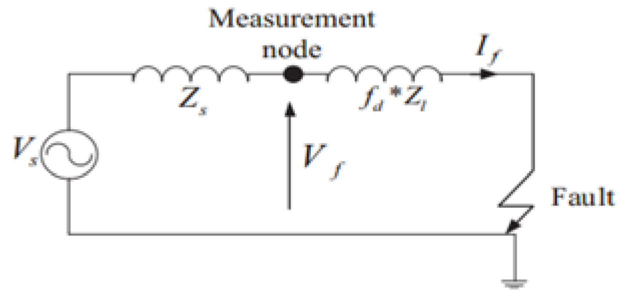

Figure 2: Impedance Based Method

Where the V_f&I_f are the responding fault voltage and current, Z_l is the line per unit length and $f_d$is the fault displacement from the node location. (Gururajapathy, et al., 2014)

Crucial limitation is stated by Gurujapathy et al. (2017) that Impedance Based Method does not considered the fault resistance, resulting measurement error in line parameters which caused such technique to be not practical and will cause errors which is also agreed by Goh et al. As a solution for this limitation, the recommended procedures to be taken is Takagi Algorithm which referred by both groups of researchers. Takagi method uses the frequency factor of both parameters’ voltage and current before and at the moment of fault occurred and run Thevenin equivalent on the parameters to locate a fault with an assumption of the fault current and the phase angle of the line current are equal. As tested by Gurujapathy et al., this algorithm managed to reach satisfactory level in terms of functionality. (Gururajapathy, et al., 2014)

k-Nearest Neighbour

The system is being projected using MATLAB software programming to achieve the objective. The main change made to achieve the project objectives such as the type of Machine Learning

March - April 2020
ISSN: 0193-4120 Page No. 151 - 158

utilized for the classification purpose. Initially, the system is subjected to SVM-based classification; which then to be replaced by k-Nearest Neighbour (kNN). kNN is a supervised machine learning with can be used for classification and regression. Classification function of kNN is approximated locally within MATLAB. In this proposed system, the kNN is utilized for classification of type of fault subjected into the system. By providing the input of the voltage and current numerical values of all three phases in the transmission, kNN produces an output of class membership of type of fault detected, either SLGF, DLGF, LLLGF or there was no fault occurred. Several settings of the kNN classification are being fine-tuned to achieve the optimum output accuracy of proposed system on different types of fault in transmission lines.

One of the impacting factors of the output accuracy is the number of neighbours also known as the neighbourhood size, k. In classification process using kNN, the input data is trained and subjected to voting of its neighbouring data, which is then categorised under a class label by the estimation by the k nearest neighbours. Variable k is a user-defined constant which lays as a positive integer. If k = 1, the input data is assigned to the class label of its nearest neighbour. The overall performance of the kNN depend heavily on the sensitivity of the selection of the neighbourhood size, k as the estimation radius is dependent on the value of k. For the proposed system of classification of type of faults in transmission required an optimal number of neighbours to achieve the desired accuracy. It has been found that the fixed optimal value of k for their test samples should be $k = \sqrt{n}$ under the implications of n> 100 where n is the number of samples. (ZHANG, et al., 2017) However, the number of samples exceeded the size of sample of this finding. Therefore, a range of number of neighbours of 5, 10, 50, 100, 500 and 1000 are assigned to the data samples to select the most optimize number of neighbours to achieve the best output of accuracy. If k small or large?

Other than neighbourhood size, k, another feature which outline major impact of the output accuracy of the proposed system is the Distance-Weights (DW). DW are defined as the relativity of closeness of each neighbour to the input sample data point. At some point of classification, the distance between the data point and its nearest neighbours are almost similar and required a method of differentiate nearest neighbours and increase the effect of the closest neighbour points to classification outcome new data points. The nearest neighbours are introduced a set of weights each to define the level of impact on the classification process of the data point. There are 3 options available for DW, Equal, Inverse and Squared Inverse.

Inverse distance weighting (IDW) uses a method of measuring the values of the neighbours, greater weights are organized for the neighbours closest to the data point and value of weights gradually reduces with a function relative to the increasing distance between the neighbour point and the data point. Equal distance weighting (EDW) is the opposite of the IDW. EDW is commonly labelled as classical kNN, where the smaller values of weights are being assigned to the closest neighbour and the values increases along with the function of distance. Other than DW, Distance metrics (DM) create major impacts on the accuracy outcome of the proposed system. DM are methods to acquire distance between new and existing data points by using the power function, p. If p = 0, it indicates that there is no decrement in distance as each weight are equal, if p increases the weights for further neighbours decrease rapidly under EDW and only the closest neighbours guide the new data points classification. 9 types of

DMs are being implemented in this system; Euclidean, Mahalanobis, City block, Minkowski, Chebychev, Cosine, Correlation, Hamming, Jaccard, and Spearman. (Chomboon, et al., 2015)

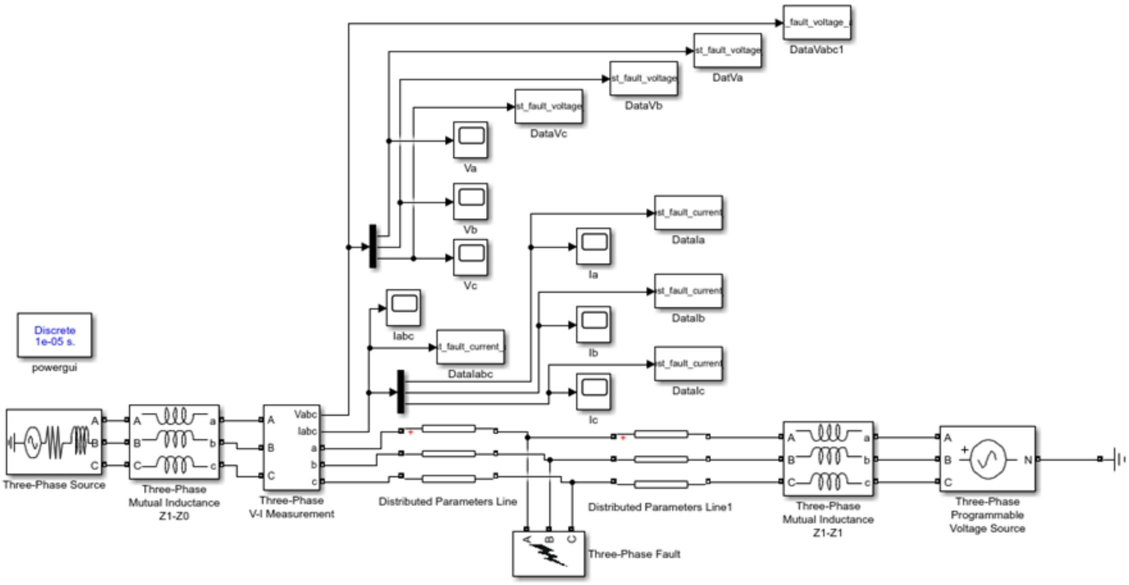

Figure 3: Simulink block for fault simulations

METHODOLOGY

The overall block diagram is shown in figure 4.1 for simulation of different type of fault in transmission line. A Three-Phase Source block is placed in as the incoming signal from the 3-phase generator for the system. From the output of the source, a block of phase inductance is allocated to simulate the inductance existed in real-time transmission line. The system is inclusive of the data reading using the 3-phase voltage and current measurement without causing any disruption to transmission system. Data extraction for voltage and current values of each phase is done via “To Workspace” block allocated for each phase voltage and current values. Multiple scope blocks are allocated to identify the successfulness of simulation of different type fault. The distributed parameters line is inclusive of selection of transmission lines length

The total period allocated for the simulation is set at 2 seconds and a sampling period of 1×10^{-5} s which accounts for 10kHz of sampling rate. A total of 200000 samples for each fault is recorded into the MATLAB workspace with variables of VA, VB, VC, Ia, Ib and Ic. The values are then imported into Microsoft Excel to compute a complete data set inclusive of all 6 variables in a single worksheet. Each type of fault is recorded separately from one another and a total of 4 types of fault are recorded; SLGF, DLGF, LLF and LLLGF. The data set is then being imported back into the Classification Learner to be trained and tested for fault classification.

The test carried in this subsection is to determine the effect of the neighbourhood size and different types of distance metrics on the classification accuracy in Single-Line to Ground fault (SLGF). A total of 9 given models of the distance metrics in classification learner to be tested their effect and a range from 5, 10, 50, 100, 500 and 1000 of neighbourhood size is allocated. The distance weight for this test was assigned as inverse weighting

RESULTS

Data Collection

TEST Engineering & Management

March - April 2020
ISSN: 0193-4120 Page No. 151 - 158

Classification learner (CL) programming in MATLAB software was being used to train the kNN technique for fault classification. By utilizing all 600000 data points of all 3-phase voltages and currents, different types of distance metric changes with the neighbourhood size from 5 until it reached the biggest size of 1000 neighbouring data points

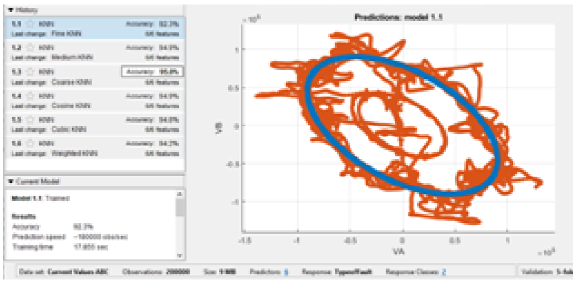

Figure 3: ***Fault Classification using kNN on SLGF***

Fig.3 provides an insight of theprediction model of Single-Line to Ground Fault (SLGF). The classified model attained an accuracy of 93.9% with a number of neighbours of 5 and achieved a higher percent of accuracy at 95% when the number of neighbours increases up to 10.

Table 1: Classification Accuracy using kNN

No. of Neighbour	5	10	50	100	500	1000
Type of distance metric						
Euclidean	93.9	94.7	95	95	95	95
City Block	94	94.8	95	95	95	95
Chebyshev	93.7	94.5	95	95	95	95
Minkowski (Cubic)	93.8	94.6	95	95	95	95
Cosine	92.9	93.5	93.9	94	94	94
Correlation	92.9	93.5	93.9	93.9	94	92.9
Spearman	94.8	94.5	91	73.9	81.4	95
Hamming	95	95	95	95	95	95
Jaccard	95	95	95	95	95	95

Data Analysis

From Table 1, the relationship between the neighbourhood size andclassification accuracy is analysed along with the impact of using different types of distance metrics on the accuracy. Figure 4 illustrate the accuracy for different neighbourhood size.

March - April 2020
ISSN: 0193-4120 Page No. 151 - 158

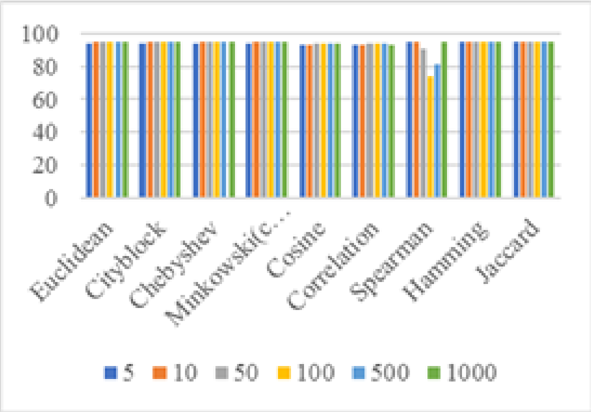

Fig 4: ***Comparison Chart of Fault Classificati on SLGF with different neighborhood size a***

As shown in figure 4, the average percentage of accuracy exceeded 90%. Among all different neighbourhood size, the highest averaged percentage of accuracy was when the neighbourhood size is 1000 and the lowest was neighbourhood size of 100. It can be observed that the averaged accuracy of neighbourhood size of 5 has a higher percentage compared to size of 100 and 500. This indicates that having a large size of neighbourhood doesn'trepresent a better classification accuracy. Among the distance metrics, Jaccard and Hamming performed the best in classification at an average of 95% and Spearman performed the lowest averaged accuracy of 88.4% due an anomaly at neighbourhood size of 100.

DISCUSSION

From the theoretical aspect, the model should be producing 100% accuracy. However, in the experiments conducted on the kNN model, the highest accuracy attained was at 95% for different types of faults and some of the distance weights produced lower percentage of accuracy.

From the observation made, this is due to kNN model data points reading which the neighbours from each data points are being monitored and weighted accordingly to their distance from the data point measured. With a smaller number of neighbours, the range of monitoring become smaller which caused the classification accuracy to be lower than expected.

Table 2:Comparison of fault classification accuracy with previous researches

No.	AI Model	Literature	Classification Accuracy from literature	Classification Accuracy from project
1	Support Vector Machine	Ray & Mishra, 2016	99.20%	
2	Artificial Neural Network	Bhattacharya & Sinha, 2017	94-95%	
3	Fuzzy Inference System	Chen, et al., 2016	95%	94%

Table 3: Fault Classification accuracy on different type of faults

Type of fault	Fault Classification Accuracy
SLGF	93.83%
DLGF	93.67%
LLF	93.92%
LLLGF	93.99%

However, if the number of neighbours is too large, the range of measurement from each data point become large, where undesired data points are being weighted accordingly, decreasing the accuracy. As the alternating current are in the sine and continuous waveform, the range of number of neighbours couldn't be too large as it oversteps the boundary to a normalized range of voltage values, making the classification measured as a pre-fault voltage signal.

March - April 2020
ISSN: 0193-4120 Page No. 151 - 158

3 errors are met while implementingthe project. The first error was involved inexperience while handling the classification learner programme. As such, initially, most of kNN models trained was not able to achieve over 90% and was solved by trial and error method of assigning a range of adjustments on the neighbourhood size, distance metrics and distance weightage. The second error was Classification learner was unable to directly defined the occurrence of fault and types of fault by using just only data points injected without and responses to be given. Therefore, the responses are being pre-defined in MS Excel before injecting as an input into classification learner.

The classification accuracy from the project is lower compared to Support Vector Machine developed by researchers Ray & Mishra, 2016. The researchers utilized the feature extraction of Discrete Wavelet Transform (DWT) which produce another variable called entropy. Based on entropy, Ray & Mishra uses the current values to identify the occurrence of fault which increases the input data compatibility with Support Vector Machine pushing the classifiction accuracy of SLGF to a higher percentage.

Other than researchers Ray & Mishra who developed Support Vector Machine, another team of researchers attempt an alternative solution for fault classification, Bhattacharya & Sinha. Bhattacharya & Sinha conduct an experiment using Artificial Neural Network (ANN) for fault classification and achieved a percentage between 94% and 95% on classification accuracy which is slightly higher than the accuracy from the project. This is achieved by Bhattacharya & Sinha using different approach with the input variables, their model consisting of an ANN model with 2 hidden layers of 60 and 40 neurons computing the maximum voltage deviations from the pre-fault voltage to classify the fault.

Moreover, a different method was being introduced by Chen, et al., 2016 using Fuzzy logic.and DWT feature extraction. The methodology consists of using the harmonics during fault to be inputted into the DWT at Daubechies 4 (Db4) mother wavelet to identify the occurrence of fault. The output from the DWT is then injected to non-linear Fuzzy Inference System (FIS) and further broken down to fifth-level of coefficient forfault classification to reduce the error of accuracy.

The solution for third error was using sampling rate of 100000Hz to extract data points for classification learner. The high sampling rate is to obtain high closeness to the continuous signal to achieve highest accuracy on fault classification.

CONCLUSION

The outcome fault classification accuracy from kNN model on different type of faults is being recorded and analysed for any error or anomalies. Several testings are done with different adjustments on the kNN model to determine the optimum settings for fault classification system. The overall classification accuracy attained is at 93.5%. For future work, an additional feature extraction can be computed on the data points to identify the presence of fault and type of fault via the current values as suggested in Ray & Mishra, 2016 with their SVM model.

REFERENCES

[1] CHOMBOON, K. ET Al., 2015. An empirical study of distance metrics for k-nearest neighbour algorithm. NAKHORN RATCHASIMA, SCHOOL OF COMPUTER ENGINEERING, INSTITUTE OF ENGINEERING, SURANAREE UNIVERSITY OF TECHNOLOGY.

[2] GOH, H. H. ET AL., 2017. Transmission Line Fault Detection: A Review. Indonesian Journal

March - April 2020
ISSN: 0193-4120 Page No. 151 - 158

of Electrical Engineering and Computer Science, 8(1), pp. 199-205.

[3] GURURAJAPATHY, S., MOKHLIS, H. & ILLIAS, H., 2014. Fault location and detection techniques in power distribution systems with distributed generation: A review. Renewable and Sustainable Energy Reviews, 74(1), pp. 949-958.

[4] RAY, P. & MISHRA, D. P., 2016. Support vector machine based fault classification and location. Engineering Science and Technology,, 19(3), pp. 1368-1380.

[5] WANI, N. S. & SINGH, D. R. P., 2016. Transmission Line Faults Detection- A Review. International Journal of Electrical Engineering & Technology (IJEET), 7(2), pp. 50-58.

[6] ZHANG, S. et al., 2017. Efficient kNN Classification with different numbers of nearest neighbors. IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, 1(1), pp. 1-11.