KNN Enhancement with Chi-Square and Manhattan Distance Formula Applied to Fake Website Detection

Ivan C. Capili[1], Richard Vincent B. Ferrer[2], Vivien A. Agustin[3], Jonathan C. Morano[4], and Mark Christopher R. Blanco[5]

[1,2,3,4,5]Computer Science Department, Pamantasan ng Lungsod ng Maynila, Manila, Philippines

Abstract**—** With the increase in data usage, measures are taken to combat the associated problems. One problem is the "Curse of Dimensionality," which refers to the problem related to high dimensional data. K-Nearest Neighbor is a widely used machine learning algorithm, and this study aims to improve performance by implementing Chi-Square as an attribute reduction process and Manhattan Distance Formula for the nearest neighbor search in the KNN process. Cross-validation was added to the algorithm to help in K value selection. The algorithm was tested on a dataset containing phishing and legitimate websites and its attributes. The results showed that the enhancements added effectively improved the algorithm's performance scores in terms of high dimensional data.

Keywords**—** Chi-Square, Cross-validation, Curse of Dimensionality, High Dimensional Data, K-Nearest Neighbor, Manhattan Distance Formula.

I. INTRODUCTION

A straightforward supervised machine learning approach called K-nearest neighbors (KNN) classifies new instances based on how closely their attributes resemble those of cases in a training dataset. Due to its popularity and simplicity, the technique has been utilized in statistical estimates and pattern recognition since the 1970s [1].

However, one drawback of KNN is that it is susceptible to the phenomenon known as the "curse of dimensionality," which occurs when the dataset's number of dimensions rises. The space between points widens, making it harder for the algorithm to recognize patterns and relationships [2]. As a result, it might be harder for machine learning algorithms, like the KNN method, to find patterns and relationships in the data, resulting in a decline in performance and accuracy [3]. Because many qualities could be used to represent websites, such as website content, website structure, website layout, website linkages, and other factors, the number of dimensions (features) in the case of fake website detection may be huge.

For this reason, the proponents would like to implement enhancements that could improve the performance of KNN. The first enhancement is the use of Chi-Square, a statistical test that could be used to measure the independence of two variables. Chi-Square as attribute reduction is a simple yet effective way of attribute reduction in datasets. Chi-Square is best suited for categorical variables, hence a wide application in textual data which can also be found on fake websites [4]. The following enhancement that the proponents would be adding is the use of Manhattan Distance which is a way to measure the distance between two points by counting the number of blocks (horizontally or vertically) needed to travel between the points. The proponents would also add a cross-validation process for an optimum k selection to boost the performance accuracy of the algorithm further. Setting a suitable K for a given training dataset is a crucial step in KNN classification. With the research done in 2022, most data analysts employing KNN classification assume that the users will provide the K for their datasets. It is the most common way to provide a K. However, it is challenging for users to establish effective K values this way. It can result in either a large K used or a small K value used. This, in return, can affect the accuracy of the results provided by the algorithm [5]. With these enhancements, the proponents would like to answer the following statement of the problem.

1. KNN Suffers from the Curse of Dimensionality.
2. A wrong choice of K in the algorithm will degrade the result.

With the stated problem, the proponents would like to accomplish the following objective.

1. To implement the combination of the Chi-square attribute reduction method and the Manhattan distance metric in improving the accuracy of the k-nearest neighbors (KNN) classifier.
2. To determine whether cross-validation can assist in enhancing the output of the KNN algorithm and figuring out the appropriate value for k.

By achieving the objective, the study aims to provide insights into the effectiveness of attribute reduction using chi-square for improving the accuracy and performance of KNN in high-dimensional datasets. The study also seeks to add the cross-validation process for the k selection. Internet users are at serious risk from fake websites because the internet can be used to transmit malware, phishing for personal information, and engage in other forms of online fraud. Therefore, a significant issue that needs to be solved to safeguard internet users from these dangers is an effective fake website detection.

II. REVIEW OF RELATED LITERATURE

KNN or K-Nearest Neighbor is one of the popular Machine Learning Algorithms out there. According to a research paper, it is considered an effortless but productive machine learning algorithm effective for classification and regression that is used widely in the application of classification prediction [6].

In KNN, distance metrics play a big role in the process. They are used to find the nearest neighbor of K and summarize the difference between two objects in a specific domain. There are several types of distance measuring techniques with the most used is Euclidean distance. This distance is the default metric that the SKlearn library of Python uses for KNN. Euclidean distance is a measure of the true straight-line distance between two points in Euclidean space [7]. Although it is widely used, one of the problems associated with the Euclidean Distance Formula is that it suffers from the curse of dimensionality [8]. According to research, most real-world problems operate in a high-dimensional data space, hence, Euclidean distance is generally not a desirable metric for high-dimensional data mining applications [9].

Another distance formula is Manhattan distance wherein it calculates the absolute value between two points by calculating the distance exactly as the original path did not take any diagonal or shortest path. This is the simplest way or technique to calculate the distance between two points, often called Taxicab distance or City Block distance [10]. According to research, it is said that Manhattan distance is better when dealing with high dimensional data due to its nature of finding the nearest neighbors [11].

The curse of dimensionality refers to problems associated with high dimensional data. [12]. This can affect the algorithm's performance such as in analyzing or visualizing the data to identify patterns, and some manifest while training machine learning models [13]. According to an article, this curse of mining high-dimensional data arises when concepts like proximity, distance, or the nearest neighbor become less meaningful with the increasing dimensionality of data sets [14].

It is said that attribute reduction refers to a process that identifies an attribute with irrelevant or excessive values [15]. Processing big data requires serious computing resources. According to an article, Chi-Square as attribute reduction is a simple, yet effective way of attribute reduction in datasets. This is best suited for categorical variables, hence a wide application in textual data which can also be found on fake websites [16].

According to an article, cross-validation is used to properly find the right K for the dataset. In cross-validation, instead of splitting the data into two parts, they are split into three: Training data, cross-validation data, and test data. Training data is used for finding nearest neighbors, cross-validation data is used to find the best value of "K" and finally to test the model on totally unseen test data [17]. It is present in Sklearn and is very helpful in selecting the correct Model and Model parameters. By using that, one can see the effect of different Models or parameters on structural accuracy [18].

III. METHODOLOGY

This section of the paper discusses the methodologies and procedures that will be involved in formulating the said alterations for KNN.

A. Equations

$$d(x,y) = \sum_{i=1}^{n} |x_i - y_i|$$

Where,

d = the distance
x and y = the coordinates in a 2D space
i = the ith element in the space

The equation above is the Manhattan distance metric. The equation would be used in line with the nearest neighbor search of the KNN.

$$x = \sum \frac{(O - E)^2}{E}$$

Where,

x = the chi-square test statistic
O = the observed frequency

E = the expected frequency

The equation stated above is Pearson's chi-square test is a statistical test for categorical data. It is used to determine whether your data significantly differs from what you expected (Turney, 2022). This equation would be used as an attribute reduction for data classification.

B. Dataset Discussion

The dataset used in the thesis is "dataset.csv" sourced from Kaggle. The dataset contains various features of a phishing website that the algorithm can use as a learning curve for its detection. The dataset contains 32 columns that signify an attribute of a website, and 11055 rows each with a corresponding index. The algorithm and the attribute reduction process would use all of the columns. With the provided dataset, it would be loaded to the machine learning algorithm to test the enhancements done. The dataset would be split into a training dataset consisting of 80 percent and a test dataset containing the latter 20 percent.

C. Proposed Enhancement

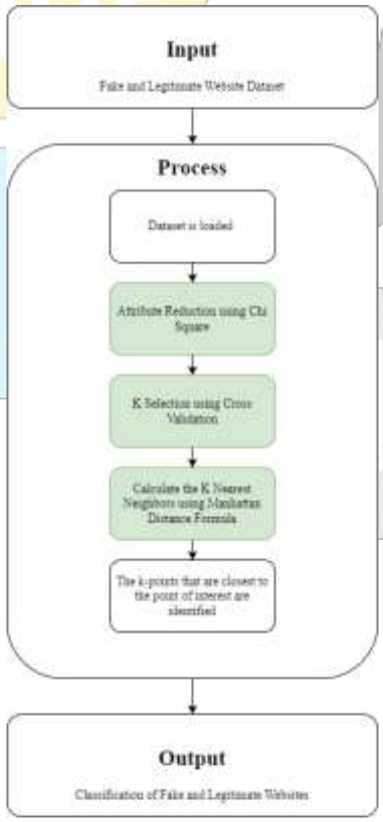
Input
Fake and Legitimate Website Dataset
Process
Dataset is loaded
Attribute Reduction using Chi Square
K Selection using Cross Validation
Calculate the K Nearest Neighbors using Manhattan Distance Formula
The k-points that are closest to the point of interest are identified
Output
Classification of Fake and Legitimate Websites

***Figure I**. Enhanced KNN Conceptual Framework*

Fig. I showcase the KNN process with added enhancement highlighted in green. The input for this process is a dataset of labeled websites. The dataset is then loaded into the algorithm to perform splitting with training and testing datasets. Next is applying the attribute reduction using The Chi-Square attribute reduction technique. The target of the process is to select the most relevant features from the dataset that would be beneficial for identifying fake websites, thus reducing the dimensionality and improving accuracy. After selecting the features that would be used, a Cross-validation process will be loaded to select the best and appropriate value of K. The chosen K would be used throughout the. For the KNN internal process, the standard Euclidean distance formula would be replaced by the Manhattan distance formula that would calculate the distance of K to its closest relative feature. The chosen value of k is then used to find the k-nearest neighbors from the training dataset being loaded. With the ongoing process, the proponents' target is to produce classified data with a high accuracy rate, thus making the classification more reliable. After the loading of the training dataset, the algorithm will learn the appropriate classification information that will be tested for the test dataset. As a result of this procedure, the target output is the classification's improved accuracy and effectiveness made possible by the Chi-Square and Manhattan Distance used in the KNN algorithm.

D. Benchmarking Tool

For the benchmarking tool, Performance benchmarking will be used to evaluate the result of the enhanced KNN in comparison to the traditional KNN. The benchmarking will include its accuracy score, precision, recall, and f1-score.

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN}$$

Where,

TN = True Negative
TP = True Positive
FN = False Negative
FP = False Positive

The equation above is the accuracy score that states the correctness of the classified data. It does show how well the algorithm predicted the correct classified data. It starts with the total predicted data over the total number of instances.

$$Precision = \frac{TP}{TP + FP}$$

Where,

TP = True Positive

FN = False Negative
FP = False Positive

The precision score is defined as the true positives over the total true positives and false positives.

This formula primarily focuses on the ability of the algorithm to minimize the false positive results, thus making the algorithm capable of producing a low rate of false positive instances from the data.

$$Recall = \frac{TN + TP}{TN + FP + TP + FN}$$

Where,
TN = True Negative
TP = True Positive
FN = False Negative
FP = False Positive

Recall, on the other hand, is the opposite of Precision score as this formula provides the algorithm's effectiveness to provide a low rate of false negative instances from the data.

$$F1\ Score = \frac{Precision * Recall}{Precision + Recall}$$

With the F1 score, the results in this formula are directly correlated to the precision and recall scores.

With an F1 score, it will consider both results to provide a balanced result.

VI. RESULTS AND DISCUSSION

In this section of the paper, the proponents discuss the enhanced KNN with Chi-Square as attribute reduction and Manhattan as its distance formula, along with the help of the cross-validation process for k selection.

It is then applied to a dataset that contains phishing websites with various features. The analysis of the results will be on a problem basis according to the stated statement of the problem.

Table I. *KNN Enhancement Performance Result*

	Performance Evaluation			
	Accuracy	Precision	Recall	F1-score
Traditional KNN with Euclidean Distance	60.87%	61%	61%	61%
Traditional KNN with Manhattan Distance	75.30%	75%	75%	75%
Enhanced KNN with Manhattan Distance and Chi-Square	94.12%	94%	94%	94%

Table I discusses the performance scores of the various algorithms that are tested. Traditional KNN with Euclidean distance is first tested in the testing and evaluation. Traditional KNN with the Manhattan distance formula is also added to provide a comparison and then the proposed enhancement with Manhattan distance and Chi-Square.

The higher scores in the tests done, the better performance in the evaluation metrics used for the measurements in the performance metric. For the first algorithm, which is the traditional KNN with Euclidean distance, an accuracy of 60.87 percent accuracy was gathered with 61 percent for the precision, recall, and f1-score.

The second test is done for the traditional KNN with the Manhattan distance formula, and the evaluation metrics showed a 75.30 percent accuracy, with 75 percent for the precision, recall, and f1-score. The next algorithm was added with the improvement of Chi-Square for the attribute reduction and Manhattan distance formula. The results from the evaluation came out with a 94.12 percent accuracy and 94% precision, recall, and f1-score.

With the results gathered, there is a difference between traditional KNN with Euclidean and Manhattan distance formula in terms of the performance evaluation, with the Manhattan distance formula getting a better performance score percentage than Euclidean.

The results imply that Manhattan distance does have a better performance handling than Euclidean distance with high dimensional data.

In contrast to the enhanced algorithm, the margin between the scores compared to the traditional KNN had a significantly increased accuracy, precision, recall, and f1-score.

This signifies that the enhanced algorithm reduced the aspects of false positives and false negatives while improving the accuracy of the classification process from the test dataset.

***Table II.** KNN Cross Validation Enhancement Performance Result*

	Performance Evaluation			
	Accuracy	Precision	Recall	F1-score
Enhanced KNN with Manhattan Distance and Chi-Square	94.12%	94%	94%	94%
Enhanced KNN with Manhattan Distance, Chi-Square, and Cross-Validation	95.57%	96%	96%	96%

Table II discusses the cross-validation improvement results. For the cross-validation, the same Performance evaluation metric would be used to test and check the effectiveness of cross-validation for the k-value choice. The first row showcases the performance result of the enhanced KNN with the Manhattan distance formula and Chi-Square. The performance scores have a high number, with accuracy sitting at a 94.12 percent mark and 94 percent for its precision, recall, and f1-score. The next row discusses the same enhanced algorithm but with an addition of cross-validation before the commencement of the KNN process. Cross-validation will be the one that will pick the appropriate K for the process, and with the addition of it, the results showed slightly better performance scores, sitting in the 95.57 percent mark for accuracy and 96 percent mark for the precision, recall, and f1 score.

VII. CONCLUSIONS AND RECOMMENDATIONS

In this section of the paper, the proponents discussed the study's conclusion in line with the statement of the problem and objective. The conclusions would be written according to the tests done and the results that it got. This section would also include the proponents' recommendations for the said paper's future.

A. Conclusion

KNN or K-Nearest Neighbor Algorithm, is one of the most used and studied algorithms in the computer science field. Due to its simplicity and high reliability, it is highly utilized in different areas of computing and learning. Although it is highly used, it also has its problems, including the "curse of dimensionality". This is a term used to describe problems associated with high dimensional data – data with a large number of features. With this, the proponents have proposed an enhancement for the traditional KNN with an implementation of Chi-Square as an attribute reduction and Manhattan as a Distance formula. The proponents also added a cross-validation process for K selection. The said algorithm would be applied to fake website detection using a dataset containing fake and legitimate websites and their attributes.

With the gathered tests from the performance evaluation metric, the proponents conclude that implementing Chi-Square as an attribute reduction along with Manhattan as a distance formula can significantly improve the performance of the KNN algorithm in terms of its accuracy, precision, recall, and f1-score. This is compared to the traditional KNN with Euclidean and even Manhattan as a distance formula. The proponents also conclude that cross-validation for K selection can further help improve the performance of the KNN algorithm in terms of its accuracy. This is added to the already enhanced KNN with Chi-square and Manhattan Distance.

B. Recommendations

For future researchers and future researchers, the proponents recommend testing the said enhancements to other available datasets to test its capability from different perspectives. One can also compare the enhanced algorithm to other algorithms for future studies. The study can also be extended by letting the results undergo a security integrity check to ensure that the performance enhancements would not affect the integrity of the fake website detector. The proponents would also like to recommend using the algorithm in real-world scenarios for more extensive testing.

REFERENCES

[1] O. Harrison, “Machine learning basics with the k-nearest neighbors algorithm,” July 14, 2019. Available on https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

[2] IBM, “What is the k-nearest neighbors algorithm,” Available on https://www.ibm.com/ph-en/topics/knn

[3] T. Yiu, “The curse of dimensionality,” September 29, 2019. Available on https://towardsdatascience.com/the-curse-of-dimensionality-50dc6e49aa1e

[4] S. Goswami, “Using the CHI-squared test for feature selection with implementation,” November 13, 2020. Available on

https://towardsdatascience.com/using-the-chi-squared-test-for-feature-selection-with-implementation-b15a4dad93f1

[5] S. Zhang, “Challenges in KNN Classification,” 2022. IEEE Transactions on Knowledge and Data Engineering, 34(10).

[6] K. Taunk, S. Verma and A. Swetapadma, "A brief review of nearest neighbor algorithm for learning and classification," 2019 International Conference on Intelligent Computing and Control Systems (ICCS).

[7] S. Gokte, “Most Popular Distance Metrics Used in KNN and When to Use Them,” n.d. Available on https://www.kdnuggets.com/2020/11/most-popular-distance-metrics-knn.html#comments

[8] Slater, “Is euclidean distance meaningful for high dimensional data,” September 29, 2021. Available on https://indicodata.ai/blog/is-euclidean-distance-meaningful-for-high-dimensional-data/

[9] C. C. Aggarwal, A. Hinneburg and D. A. Keim, “On the surprising behavior of distance metrics in high dimensional space. Database Theory — ICDT 2001, 420-434.

[10] M. Badole, “How KNN uses distance measures,” August 6, 2021. Available on https://www.analyticsvidhya.com/blog/2021/08/how-knn-uses-distance-measures/

[11] K. Gohrani, “Different types of distance metrics used in machine learning,” November 10, 2019. Available on https://medium.com/@kunal_gohrani/different-types-of-distance-metrics-used-in-machine-learning-e9928c5e26c7

[12] K. Arun, “Understanding curse of dimensionality,” May 30, 2022. Available on https://www.mygreatlearning.com/blog/understanding-curse-of-dimensionality/

[13] V. Katara, “Dimensionality Reduction — PCA vs LDA vs t-SNE,” Decenmber 30, 2020. Available on https://medium.com/analytics-vidhya/dimensionality-reduction-pca-vs-lda-vs-t-sne-681636bc686

[14] H. Kaur, “Why should euclidean distance not be the default distance measure,” October 19, 2022. Available on https://towardsai.net/p/l/why-should-euclidean-distance-not-be-the-default-distance-measure

[15] M. Danil, S. Efendi and R. Widia Sembiring, “The analysis of attribution reduction of k-nearest neighbor (KNN) algorithm by using CHI-square,” 2019. Journal of Physics: Conference Series, 1424(1), 012004.

[16] S. Goswami, “Using the CHI-squared test for feature selection with implementation,” November 13, 2020. Available on https://towardsdatascience.com/using-the-chi-squared-test-for-feature-selection-with-implementation-b15a4dad93f1

[17] S. Jain, “When would one use Manhattan distance as opposed to Euclidean distance,” August 30, 2019. Available on https://datascience.stackexchange.com/questions/20075/when-would-one-use-manhattan-distance-as-opposed-to-euclidean-distance

[18] Q. Sun, “How to deal with cross-validation based on KNN algorithm, compute AUC based on naive Bayes Algorithm,” May 18, 2018. Available on https://medium.com/@svanillasun/how-to-deal-with-cross-validation-based-on-knn-algorithm-compute-auc-based-on-naive-bayes-ff4b8284cff4