The Analysis of Attribution Reduction of \textit{K-Nearest Neighbor} (KNN) Algorithm by Using \textit{Chi-Square}

Muhammad Danil\textsuperscript{1}, Syahril Efendi\textsuperscript{2}, Rahmat Widia Sembiring\textsuperscript{3}

\textsuperscript{1,2,3}Department of Computer Science, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, 20155, Indonesia

* daniel.fikom@gmail.com

Abstract. Data reduction is one of the applicable techniques used to obtain the reduction representation from the data whose volume is much smaller, but still retains the original integrity of the data. Attribute reduction is a process to identify and eliminate an attribute with irrelevant or excessive values. In this study, attribute reduction was carried out using the Chi-Square Algorithm implemented in \textit{K-Nearest Neighbor} (KNN) for classifying the objects based on the closest data to the objects. The test was carried out on the Pima Indians dataset with the total of 768 data. The Chi-Square method was used to reduce the dimensions of large datasets and to improve the accuracy of the prediction of the closest \textit{K-Nearest Neighbor} results. \textit{K-Nearest Neighbor} using \textit{Chi-Square} as the choice of features proved to be accurate and effective in reducing data. The results of this study show that \textit{K-Nearest Neighbor} using \textit{Chi-Square} as the choice of features was accurate and effective in reducing data without reducing the integrity of the original data and reducing the quality of the information produced.

Keywords: \textit{K-Nearest Neighbor}, KNN, \textit{Chi-square}, classification, data reduction, Pima Indians dataset

1. Introduction

Analysis of complex and large amounts of data can take a long time to proceed and often leads to impractical or inappropriate analysis. Many researchers have conducted research on data reduction methods. Such technique is useful in reducing complex data without reducing the integrity of the original data and reducing the quality of the information produced. \cite{4}

This selection process is implemented using the \textit{K-Nearest Neighbor} algorithm. \textit{K-Nearest Neighbor} is a method for for classifying the objects based on the closest data to the objects. \textit{K-Nearest Neighbor} (KNN) can also be defined as a classifier used to classify data based on the comparison of the nearest K value. K parameter on \textit{K-Nearest Neighbor} (KNN) has a large influence on the final outcome of the prediction produced. \cite{5}

The method used for attribute selection in this study was the \textit{Chi-Square} method, one of the feature selection methods which include the filter method. \textit{Chi-Square} is a type of non-parametric comparative test performed on two variables, where the scale data of the two variables is nominal. If one of the two variables is variable with a nominal scale, \textit{Chi-Square} test must be carried out at the lowest degree. \cite{5}
2. Methodology
To obtain accurate, relevant, and valid data, the data was collected using the documentary research method. This method is a type of data collection examining various kinds of documents which are useful for analyzing the materials. The stages carried out in the implementation of this research are as follows:
a. Study of literature
In order to explore the previous studies discussing the process of data reduction in the *K-Nearest Neighbor* (KNN) algorithm by using *Chi-Square*, the study of literature was carried out. The data of this study were obtained from various books, journals, articles and several other references.
b. Research Analysis
The analysis carried out in this study was to evaluate the level of accuracy of the *K-Nearest Neighbor* algorithm by combining the *Chi-Square* algorithm in order to reduce the large data so that the accurate results can be gained.
Workflow diagram and Flowchart of Algorithms Work System undertaken in this study are illustrated in figure. 1 and figure. 2:

![Figure 1. Research Work Flow Chart](image1)

![Figure 2. Flowchart Work System Flow Algorithms](image2)

**Figure 1. Workflow diagram**

Start – Problem identification – Reviewing the literature- determining the purpose of the study- dataset collection- implementing the proposed algorithms- testing- analyzing the evaluation results- summarizing the result of the study- end.
Based on figure 1, it can be explained as follows:

a. Pre-processing Stages

Pre-processing is a preparation process that is carried out on the dataset so that it is ready to be processed. All documents including training data, test data, and new data will be subject to pre-processing stage first.

In the pre-processing stage, the data is uniformised and reduced. This stage is done by selecting, cleaning, and then transforming the data into the desired form so that the model can be prepared. The purpose of this stage is to prepare valid data before being processed in the next stage. Thus, cleansing, transformation, and attribute reduction are carried out in this stage. The data obtained is processed to gain relevant and appropriate attributes. However, in this study, pre-processing data stage was not carried out.

b. Attribute Reduction (Chi-Square)

In this study, Chi-Square was used to reduce the data. It checks the null hypothesis which states that there is no relationship between attributes. Based on this null hypothesis, the model for distributing data is categorized with the assumption that there is no relationship between attributes. This test is undertaken based on the comparison between the actual data distribution and the expected data distribution. Contingency tables for calculating Chi-Square are also desirable. The frequency observed for \( c_{ij} \) cell is \( n_{ij} \). The expected frequency for \( C_{ij} \) cell is \( e_{ij} \).

\[
e_{ij} = \frac{n_{ij}}{n} \tag{1}
\]

The Chi-Square test formula can be calculated as follows:

\[
(x^2) = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \tag{2}
\]

Information:
- \( n \): Expected frequency
- \( x^2 \): Chi-square value
- \( e \): Frequency obtained

c. Fold Cross Validation

In this stage, the data sharing stage is carried out to divide the training data and the testing data. The amount of training data is 90% while the test data is 10%. Such distribution is used based on the suggestion of k-fold cross validation method so that the performance of the classification model can be measured. 10-fold cross validation is the basic form of cross validation. In this study, the 10 fold cross validation was used because the data is quite large which is 768 datasets. It means that the testing data used in this study was 70 datasets for each fold.

In cross validation, the training and testing data must be crossed in a row so that each data has the opportunity to be validated. According to Refaeilzadeh et al. (2008), there are two main objectives using cross-validation:

1. To measure the performance on the provided training model using one algorithm. In another word, to check the general performance of the algorithm.
2. To compare the performance of two or more different algorithms and gain the best algorithm for the available data. Moreover, it also compares the performance of two or more variations of the parameter model.

d. Classification (K-Nearest Neighbor)

K-Nearest Neighbor (K-NN) is a classifier that is used to classify data based on the comparison of the nearest K value and the K parameter on KNN which has a large influence on the final outcome of the prediction. K-NN measures the similarity of data using distance measurements. In this study, Euclidean distance is used to calculate the similarity of two or more data.

\[
d(x, y) = \sqrt{\sum_{a=1}^{m} (x_a - y_a)^2} \tag{3}
\]
Variable d is the distance of both data. K-NN is known as a method that requires time and large computational costs because the testing data and all available training data need to be matched first. Thus, there is a need for a data reduction process in order to reduce attributes that are not needed in the data so that computing time can be accelerated.

e. Evaluation

The final step is to do an evaluation. This stage aims to determine the accuracy of the results from the datasets classification using the classification method of the test data. Evaluation is performed to find the optimal solutions resulting from various and complex classification methods that are carried out repeatedly. [13] In this study, evaluation was conducted to determine the level of accuracy of the results in the preprocessing data, attribute reduction, and classification of test data stages.

Novakovic et al. (2017) mention that there are several standard measurements used to measure classifications with two class labels: accuracy, true positive rate, false positive rate, true negative rate, false negative rate, and precision. [13] In this study, the accuracy standard was used because it is considered adequate. Accuracy can be formulated using the following equation:

\[
\text{Accuracy} = \frac{\text{jumlah nilai benar}}{\text{jumlah data keseluruhan}} \times 100\% \quad (4)
\]

f. Testing

The dataset that has been collected is then classified using the K-Nearest Neighbor (K-NN) algorithm and reduced by using Chi-Square. At this stage, the system that has been implemented is tested. After testing the system, the stage of measuring system performance was performed in order to draw conclusions on the results of the research conducted.

g. Analysis of the Evaluation Results and Conclusion of the Research Results

The next step is analyzing the results of the two algorithms in order to reach the conclusion of the study. From the results, it is expected that K-Nearest Neighbor and Chi-Square can predict a person's tendency to develop diabetes mellitus (DM). Furthermore, through the data reduction, it is also expected that the results and the accuracy of the study can be optimized.

3. Result and Discussion

The selection process in this study was implemented using the K-Nearest Neighbor algorithm. It is a method for classification on objects based on training data that is the closest to the object in terms of distance.

By carrying out the first test towards all attributes, there were 445 correct data out of 538 total data. The accuracy equation of the K-NN classification uses the following equation:

\[
\text{Accuracy} = \frac{\text{jumlah nilai benar}}{\text{jumlah data keseluruhan}} \times 100\% \quad (5)
\]

By using the above accuracy equation for k = 10, the result for the accuracy of the classification using KNN was 82.71%. The classification using the KNN method was carried out with 7-times-trial using different Ks for each experiment. The Ks used were 5, 10, 15, 20, 25, 30, and 35.

The results of the accuracy obtained are shown in table 1 below:

| The value of k used | Correct data | Accuracy Results |
|---------------------|--------------|-----------------|
| k-5                 | 430          | 79.93%          |
| k-10                | 445          | 82.71%          |
| k-15                | 408          | 75.84%          |
| k-20                | 428          | 79.55%          |
| k-25                | 405          | 75.28%          |
| k-30                | 413          | 76.77%          |
| k-35                | 404          | 75.09%          |
The results of the accuracy in the form of a diagram can be seen in Figure 3 below:

![Figure 3. Comparison of different k accuracy](image)

Used K value/ correct data/ accuracy results

The table above shows the results of accuracy of different k. From the table, it can be concluded that k = 10 has the highest accuracy of 82.71% with 445 true data out of 538 total data.

Then, the second test is carried out on all attributes resulting in 89 correct data out of 90 total data.

By using the above accuracy equation for k = 6, the results obtained from the accuracy of classification calculations using KNN is 98.89%. Classification using the KNN method is performed with 5 trials using different k for each experiment. The Ks used were 3, 5, 6, 10, and 15. The results of the accuracy obtained are shown in table 2 below:

| The value of k used | Correct data | Accuracy Results |
|---------------------|--------------|------------------|
| k-3                 | 87           | 96.67            |
| k-5                 | 88           | 97.78%           |
| k-6                 | 89           | 98.89%           |
| k-10                | 88           | 97.78%           |
| k-15                | 86           | 95.56%           |

Used K value/ correct data/ accuracy results

The table above shows the results of accuracy of different K. From the table, it can be concluded that the highest accuracy for k = 6 was 98.89%. The results of the accuracy in the form of a diagram can be seen in Figure 4 below:

![Figure 4. Comparison of different accuracy of k](image)

From figure 3, it can be seen that k-6 has the highest accuracy. Therefore, it was used for the next process to retrieve the value of accuracy.

After conducting the classification and normalization process using the K-Nearest Neighbor algorithm, attribute reduction was performed by using the Chi-Square method, one of the attribute
reduction methods including the filter method. It is also one type of non-parametric comparative test performed on two variables where the scale data of the two variables is nominal.

*K-Nearest Neighbor* can solve data class imbalance problems and the selection feature of *Chi-Square* is one way to reduce large dataset dimensions. Chi-Square has an important role in choosing the appropriate subset of the original feature sets since not all attributes are relevant to the problem. Hence, the use of *K-Nearest Neighbor* in this study can help reducing complex data without reducing the integrity of the original data and reducing the quality of the information produced. From the result of the study, it was proven the use of *K-Nearest Neighbor* and *Chi-Square* stabilized and even increased the accuracy.

4. Conclusion

Based on the testing and evaluation of the *K-Nearest Neighbor* (KNN) method in attribute reduction using the *Chi-Square* algorithm, it was concluded that the *K-Nearest Neighbor* algorithm showed stabil and increased accuracy from datasets with large data dimensions. The use of Chi-Square method can also reduce the dimensions of large datasets and help improving the results of *K-Nearest Neighbor* prediction accuracy. In this case, K-Nearest Neighbor utilizes attribute reduction function from *Chi-Square* to select data attributes with the characteristics of the data itself, and increases the accuracy of *K-Nearest Neighbor* predictions.

*Chi-Square* based on *K-Nearest Neighbor* only considers a small subset of all possible models so that the risk of skipping or losing the best model will increase, along with the addition of independent variables. This research can be further developed by using other Data Mining classification methods, the use of attribute reduction methods or other optimization methods that can overcome large dimensional data problems without reducing the integrity of the original data and reducing the quality of the information produced.

References

[1] Alpaydin, Ethem. 2010. Introduction to Machine Learning Second Edition.
[2] Ardy Wibowo Haryanto, Edy Kholid Mawardi, & Muljono, 2018. Influence of Word Normalization and Chi-squared Feature Selection on Support Vector Machine (SVM) Text Classification. IEEE International Seminar on Application for Technology of Information and Communication (iSemantic).
[3] D. Hand, H. Mannila & P. Smyth, 2001. Principles of Data Mining. MIT Press. Florin Gorunescu, Data Mining Concept, Models and Techniques, 12th ed., Prof. Lakhmi C. Jain Prof. Janusz Kacprzyk, Ed. Craiova, Romania: Springer, 2011.
[4] Han, J., & Micheline K., 2012. Data Mining: Concepts and Techniques Second Edition, Elsevier.
[5] Indah Listiowarni & Eka Rahayu Setyaningsih, 2018. Feature Selection Chi-Square dan K-NN pada Pengkategorian Soal Ujian Berdasarkan Cognitive Domain Taksonomi Bloom. Sekolah Tinggi Teknik Surabaya.
[6] Jatmiko Indriyanto, Purwanto, & Catur Supriyanto, 2014. Algoritma K-Nearest Neighbor Berbasis Chi-Squared Untuk Prediksi Nasabah Asuransi. Pascasarjana Teknik Informatika Universitas Dian Nuswantoro.
[7] Liu & Motoda, 199. Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic, 1998.
[8] Maimoon, Oded & Lior Rokach, 2010. Data Mining and Knowledge Discovery Handbook.
[9] M. Arinal Ihsan, 2018. Reduksi Atribut Pada Algoritma K-Nearest Neighbor (KNN) Dengan Menggunakan Algoritma Genetika. Tesis. Universitas Sumatera Utara.
[10] Miles, B. Mathew & Michael Huberman. 1992. Analisis Data Kualitatif Buku Sumber Tentang Metode-metode Baru. Jakarta: UIP
[11] Muhammad Ejazuddin Syed, 2014. Attribute Weighting In K-Nearest Neighbor Classification.
University of Tampere School of Information Sciences Computer Science

[12] Novaković J. 2011. “Toward Optimal Feature Selection Using Ranking Methods and Classification Algorithms” dalam: Yugoslav Journal of Operations Research 21(1).[Internet].[Diakses tanggal 19/03/2019 ]. http://yujor.fon.bg.ac.rs/index.php/yujor/article/download/364/255.

[13] Novaković J et al. 2017. “Evaluation of Classification Models in Machine Learning” dalam: Theory and Applications of Mathematics & Computer Science 1 (1), pp.39-46. [Internet]. [Diunduh tanggal 19/03/2019 ]. Dapat diunduh dari: http://uav.ro/stiinte_exacte/journal/index.php/TAMCS/article/viewFile/158/126.

[14] Nachirat Rachburee & Wattana Punlumjeak, 2015. A Comparison of Feature Selection Approach Between Greedy, IG-ratio, Chi-square, and mRMR in Educational Mining. IEEE International Conference on Information Technology and Electrical Engineering (ICITEE), Chiang Mai, Thailand.

[15] Refaeilzadeh P, Tang L, dan Liu H. 2009. Crossvalidation. Springer, pp. 532–538. [Internet]. [Diakses tanggal 19/03/2019]: http://leitang.net/papers/ency-crossvalidation.pdf.

[16] Runxiu Wu, 2010. Improved K-Modes Clustering Method Based on Chi-square Statistics. IEEE International Conference on Granular Computing.

[17] Said Bahassine, Abdellah Madani & Mohamed Kissi, 2016. An Improved Chi-Square Feature Selection For Arabic Text Classification Using Decision Tree. IEEE 11th International Conference on Intelligent Systems: Theories and Applications (SITA).

[18] Witten & Frank, 2005. Data Mining – Practical Machine Learning Tools and Techniques with Java Implementations, second edition, Morgan Kaufmann Publishers, USA, 2005.