Comparison of Distance Measurement Methods on K-Nearest Neighbor Algorithm For Classification

Taca ROSA[1], Rifkie PRIMARTHA[1*], and Adi WIJAYA[2]

[1]Faculty of Computer Science, Sriwijaya University, Indonesia
[2]Informatics Engineering Department, Universitas MH Thamrin, Jakarta, Indonesia
*Corresponding author: rifkie77@gmail.com

ABSTRACT

K-Nearest Neighbor is a non-parametric classification algorithm that does not use training data and initial assumptions or models in the calculation process. The quality of the k-Nearest Neighbor classification results is very dependent on distance between object and value of k specified, so the selection for distance measurement method determines the results of classification. This study compares several distance measurement method, including Euclidean distance, Manhattan distance, Tchebychev distance and Cosine distance to see which distance measurement method can work optimally on the k-Nearest Neighbor algorithm. The selection of k values also determines the results of k-Nearest Neighbor classification algorithm, so determining the k value also needs to be considered. The data used in this study is a dataset of cervical cancer. The highest accuracy results obtained using the Cosine distance measurement method that is equal to 92.559% with a value of k = 9. Based on the accuracy values that have been compared, the most optimal distance measurement method is Cosine distance with the best k value obtained is k = 9 even though this distance measurement method has the highest computing time which is equal to 0.898 seconds.

Keywords: *Distance measurement, K-Nearest Neighbor, Euclidean Distance, Manhattan Distance, Tchebychev Distance,Cosine Distance*

INTRODUCTION

A lot of data is generated every day and along with the development of digital storage media which is increasing rapidly, causing a data explosion. The data includes various data in all fields, one of them is in the medical field. Diagnosis of cancer is the most discussed problem in medical field [1], so research to build cancer diagnosis technology with data mining methods is an interesting problem to develop.

Data mining is a method that performs process of extracting from a number of data in order to obtain a data pattern that will later become a new information or knowledge [2]. Classification technique is one of the main functions of data mining and one of the most widely used algorithms is k-Nearest Neighbor [3]. K-Nearest Neighbor is a non-parametric classification algorithm that does not use training data and assumptions or initial models in calculation process, rather it uses the directly hypothesis based on the training data provided [4, 5].

K-Nearest Neighbor algorithm uses a supervised learning approach where the data used is labeled data. In addition, this algorithm is simple and easily represented. Although simple, this algorithm has been tested in number of cases and produced quite high performance as research conducted on the breast cancer dataset can achieve an accuracy of 97.57% [6]. Medjahed, et al. [1] done research on the k-nearest neighbor algorithm by comparing several distance measurement functions. In that study, the method of measuring distances between Euclidean distance, Manhattan distance, Tchebychev distance, Cosine distance and Correlation distance were compared. This study provides the best accuracy results on 2 methods, euclidean distance and manhattan distance, where measurements with that method succeeded in providing an accuracy rate of 98.70% at k = 1.

The quality of k-Nearest Neighbor algorithm results very depends on proximity between objects and value of k specified [1]. The selection of methods for calculating distances is an important issue [7] because k-Nearest Neighbor method is very dependent on the calculation results of distances between objects.

Based on these, then this study discusses a comparison of distance measurement methods between Euclidean distance, Manhattan distance, Tchebychev distance and Cosine distance which can improve the performance of k-Nearest Neighbor algorithm for classification.

RELATED WORKS

Related research that used same dataset from UCI Machine Learning Repository were taken to compare the results that obtained from proposed method. In 2017, Ceylan and Pekel [8] investigated the efficacy of using multi-label classification techniques for diagnosing cervical cancer at early stage. Their compared four common learning algorithms such as Naïve Bayes, J48 Decision Tree, Sequential Minimal Optimization, and Random Forest. In this study, to handled multi-label classification, they used Problem Transformation (PT) methods such as Binary Relevance (BR), Classifier Chains (CC), and Conditional Dependency Networks (CDN), Label Combination (LC) on cervical cancer dataset. The dataset was randomly divided into two sets; training and test with ratio 70% training data (566) and 30% testing data (292). In order to evaluate the unbiased estimate of the four prediction models for comparing their performances the 10-fold cross-validation methods were used. the accuracy percent for examined algorithms were approximately over 80%, except for J48-BR and J48-CDN. The highest accuracy value for Naive Bayes method is obtained by Naive Bayes+LC at 84,48% while for J48 Decision Tree method is J48+CC and J48+LC which reached 88,7%. Then for Sequential Minimal Optimization method, the highest accuracy value reached by SMO+BR which is 87%. The last is Random Forest method that reached 89% of highest accuracy for Random Forest+CC and Random Forest+LC.

Other studies related to cervical cancer research has been done by Idris et al Ayyappan and SivaKumar [9]. They used Sequential Minimal Optimization method by applying various kernels such as Polykernel, Normalized Polykernel, Puk, and RBF Kernel. In this research work applied in weka 3.8.3 version for SMO classification to calculate predicting. The research results are Polykernel has 75.54% accuracy level, Normalized Polykernel has 68.43% accuracy level, Puk has 72.33% accuracy level, and RBF Kernel has 87.5% accuracy level. The experiment results show that SMO with RBF Kernel has 87.5% accuracy level achieved the best accuracy rates.

MATERIALS AND PROPOSED METHOD

Dataset used in this study is cervical cancer dataset that can be accessed on UCI Machining Learning Repository [9]. In this study, we developed a software to do classification process. Testing is done by using cervical cancer dataset which has 32 predictor attributes and 4 target labels (Hinselmann, Schiller, Cytology dan Biopsy) and has 858 data objects stored in file with extension .csv. The process that carried out in this study is data processing and classification process. At pre-processing stage, data cleaning is performed which is checking missing values and outliers data. Missing values were filled using the sample mean[10], for numerical data is filled by the average value of each attributes data and for nominal data is filled by dominant value, while for the data out of range or called outlier data is removed manually[11]. Then split the data by cutting from the system with a comparison of 80% training data and 20% testing data. Figure 1 shows a research diagram.

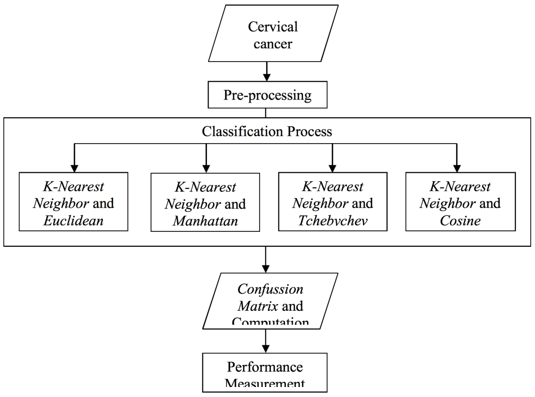
Cervical cancer
Pre-processing
Classification Process
K-Nearest Neighbor and *Euclidean*
K-Nearest Neighbor and *Manhattan*
K-Nearest Neighbor and *Tchebvchev*
K-Nearest Neighbor and *Cosine*
Confussion Matrix and *Computation*
Performance Measurement

Figure 1. Diagram of Research

Testing is carried out one time for all the k value inputted that are k = 1, k = 3, k = 5, k = 7, k = 9, k = 11 and k = 13. For testing computation time parameters, each k value is tested 5 times. Each k value inputted was tested with 4 distance measurement methods that are Euclidean, Manhattan, Tchebychev and Cosine. Then the software groups the test data by using the k-Nearest Neightbor method.

Confusion matrix is a method used to analyze and to compare classifiers in this paper. The predicted and true values of class membership can be cross–classified and counted in a confusion matrix [12]. For each data calculated the value of accuracy, sensitivity, specificity based on the resulting confusion matrix table and calculated computational time on the 4 data labels grouping results to evaluate the quality of classification model. After the classification process, an analysis of the results is carried out by looking at the average results of each data label and optimal k value for each research indicator.

EXPERIMENTAL RESULTS

The testing process uses 168 testing data and 672 training data. Tests carried out on different methods of measuring distances and k values. To analyze the results of research on comparative indicators used, the results of study are presented in graphical form which can be seen in Figure 2 through Figure 5. In addition, from the graph we can find out the optimal k value in the algorithm used for research.

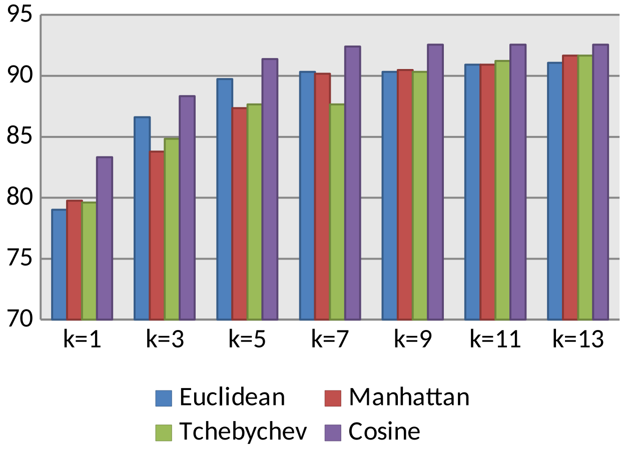

Figure 2. Comparison of Accuracy Value

Figure 2 shows the highest accuracy value in the distance measurement method is at k = 13 that is equal to 91.071% for the Euclidean distance measurement method, 91.666% for Manhattan and Tchebyshev. As for the Cosine method, the accuracy value obtained is the highest accuracy value among four methods that is equal to 92.559% at k = 9. Because the accuracy value obtained by the 3 distance measurement methods shows the maximum number at k = 13 then based on the level of accuracy, the optimal k value of the 3 methods Manhattan, Euclidean and Tchebychev is 13, while for the Cosine method, the optimal k value is 9.

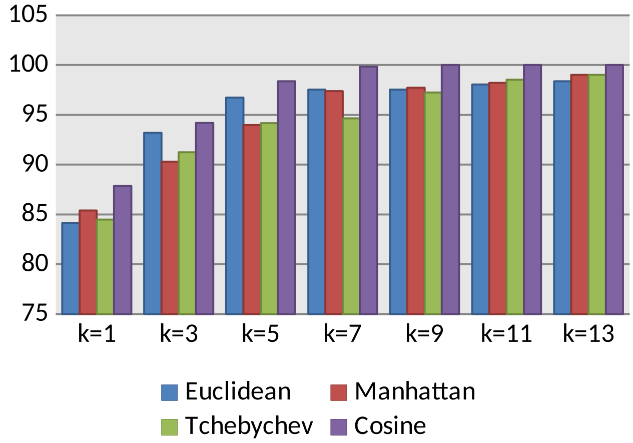

Figure 3. Comparison of Specificity Value

Graph in Figure 3 shows the greatest specificity value is at k = 13, with the Euclidean distance measurement method that is equal to 98,365%, Manhattan equal to 99,023%, and TChebyshev 99.021%. Whereas the Cosine method, when k = 9 has shown a maximum specificity value of 100%. Based on specificity value, the optimal k value is k = 9 on Cosine distance method and has the same result in the other 3 methods is k = 13.

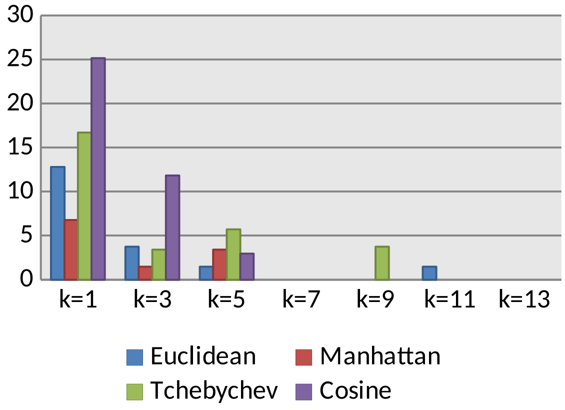

Figure 4. Comparison of Sensitivity Value

Figure 4 shows the greatest sensitivity value of k is 1. For the Cosine method it has the highest sensitivity value of 25.146%, Euclidean method of 12.803%, 6.78% on Manhattan method, and 16.702% on Tchebychev method.

Sensitivity or Recall is the metric that measures the accuracy on the positive instances, it can be defined as True Positive [13]. The sensitivity value obtained in the K-Nearest Neighbor algorithm calculation with this cervical cancer dataset has a very low value. This is caused by the condition of data that has unbalanced proportions of true and false classes [13], where the composition of true classes is less than false classes so with the greater k value, the K-Nearest Neighbor algorithm groups many data into false classes. This causes the increasing value of k inputted then less data are grouped in the true class. In addition, one of the contributing factors is condition of data with too many missing values.

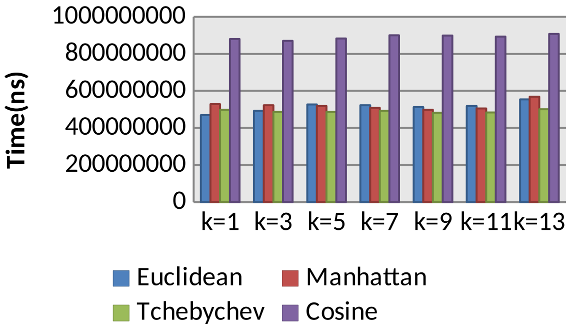

Figure 5. Computation time

From the graph above it can be seen that the Cosine distance measurement method has the longest time among four other distance measurement methods. If computing time is averaged, then Tchebychev method has the lowest computational time 490151464 ns or 0.49 s, while the Euclidean distance measurement method is 514006473 ns or 0.514 s, Manhattan method is 521358689.3 ns or 0.521 s and the Cosine method is 890057643 ns or 0.89 s.

From the results of graph analysis, it is concluded the results of the analysis in table 1. Table 1 is obtained from the analysis based on the optimal k value discussed earlier.

From the table 1, the highest accuracy value is obtained by the Cosine distance method of 92.559% at k = 9 and the lowest accuracy is obtained by the Euclidean method at 91.071% at k = 13. The highest specificity value in the Cosine method is 100% and the lowest is Euclidean that is equal to 98,365. While the sensitivity value obtained for all methods is 0% because of imbalanced features of the dataset. When viewed from the computational time, the lowest computation time is obtained by the Tchebychev distance measurement method which is 501553345.6 ns or 0.501 s and the Cosine method obtains the longest computational time which is 907129419.2ns or 0.9071 s.

Table 1. Comparison Performance Each Distance Measurement Methods

Distance measurement methods	Optimal *k* value	Accuracy	Sensitivity	Specificity	Computation time (*ns*)
Euclidean	*k*=13	91,071%	0%	98,365%	555021709,6
Manhattan	*k*=13	91,666%	0%	99,023%	568278994,8
Tchebychev	*k*=13	91,666%	0%	99,021%	501553345,6
Cosine	*k*=9	92,559%	0%	100%	898924107,8

CONCLUSION

The experimental result shows that the most compatible distance measurement method used in the cervical cancer dataset is Cosine distance method because it has the highest accuracy value of 92.559% at k = 9 while for the Manhattan distance measurement method of 91.666% with a value of k = 13, the Tchebychev distance measurement method is 91.666% with the value of k = 13 and the lowest accuracy value obtained by the Euclidean method that is equal to 91.071% at the value of k = 13. The Cosine distance method also has the highest specificity value of 100% even though it has the greatest computing time compared to other distance measurement methods.

In the future works, we expect the proposed method can be produce better accuracy so detecting cervical cancer will be more effective by trying to use others distance measurement method or be focuse on reducing features to improve computing efficiency.

REFERENCES

[1] S. A. Medjahed, T. A. Saadi, and A. Benyettou, "Breast cancer diagnosis by using k-nearest neighbor with different distances and classification rules," *International Journal of Computer Applications,* vol. 62, 2013.

[2] M. F. Kurniawan and I. Ivandari, "KOMPARASI ALGORITMA DATA MINING UNTUK KLASIFIKASI PENYAKIT KANKER PAYUDARA," *IC-Tech,* vol. 12, 2018.

[3] S. W. Binabar and I. Ivandari, "Optimasi Parameter K pada Algoritma KNN untuk Deteksi Penyakit Kanker Payudara," *IC-Tech,* vol. 13, 2018.

[4] R. Primartha, *Belajar Machine Learning Teori dan Praktik*, 2018.

[5] J. Walters-Williams and Y. Li, "Comparative study of distance functions for nearest neighbors," in *Advanced Techniques in Computing Sciences and Software Engineering*, ed: Springer, 2010, pp. 79-84.

[6] A. Goweda, M. Elmogy, and S. Barakat, "Blending Memetic Search Strategy with K-Nearest Neighbor Algorithm for Cancer Classification Problem," *Journal of Next Generation Information Technology,* vol. 8, 2017.

[7] G. Verdier and A. Ferreira, "Adaptive Mahalanobis Distance and $ k $-Nearest Neighbor Rule for Fault Detection in Semiconductor Manufacturing," *IEEE Transactions on Semiconductor Manufacturing,* vol. 24, pp. 59-68, 2010.

[8] Z. Ceylan and E. Pekel, "Comparison of multi-label classification methods for prediagnosis of cervical cancer," *International Journal of Intelligent Systems and Applications in Engineering,* vol. 5, pp. 232-236, 2017.

[9] D. Dua and E. Karra Taniskidou, "UCI Machine Learning Repository [http://archive. ics. uci. edu/ml]. Irvine, CA: University of California," *School of Information and Computer Science,* 2017.

[10] K. Fernandes, J. S. Cardoso, and J. Fernandes, "Transfer learning with partial observability applied to cervical cancer screening," in *Iberian conference on pattern recognition and image analysis*, 2017, pp. 243-250.

[11] T. V. Pollet and L. van der Meij, "To remove or not to remove: the impact of outlier handling on significance testing in testosterone data," *Adaptive Human Behavior and Physiology,* vol. 3, pp. 43-60, 2017.

[12] I. Düntsch and G. Gediga, "Confusion matrices and rough set data analysis," in *Journal of Physics: Conference Series*, 2019, p. 012055.

[13] A. Lazar and B. Shellito, "The Classification of Imbalanced Spatial Data," in *MAICS*, 2011, pp. 108-113.