International Journal of Future Generation Communication and Networking
Vol. 13, No. 4, (2020), pp. 2587–2597

Asset Mapping Using K-NN To Evaluate The Distance Measure Between Assets

Sree Divya. K [1], P. Bhargavi 2 , S. Jyothi [3]

[1] Research Scholar, [2]Assistant professor, [3]Professor
Department of Computer Science, Sri Padmavati MahilaVisvavidyalayam, Tirupati
[1,3]{divya.kpn,jyothi.spmvv}@gmail.com ; [2]pbhargavi18@yahoo.co.in

Abstract

This paper discusses about future challenges in terms of bigdata and new technologies. There are several utilities for collecting large amount of data, but they are hardly utilized because they are huge in amount and also there is uncertainty associated with it. Oversee monitoring of assets collects large amount of data during periodic operations. The main query raises are "how to gather information from large amount of data?". But big data analytics will handle enormous amount to data with onset of Machine learning techniques. Along with the technological advancements like QGIS (Quantum Geographic Information System), Big data analytics plays a major role for mapping of assets in the community. In this paper, remonstrance is solved by different boulevard and ground rule to make the current asset mapping and management practices smarter for the future smart cities, towns and villages. Bigdata with QGIS framework which provides for a simple-to-use asset classification system, management guidelines based on the relationship between importance and fragility of the asset, and a set of indicators based on the pressure–state–response model for monitoring the progress.

1. Introduction

Asset mapping portrays a form of deliberate assessment of positive elements in a very community. Participating during this endeavor produces a summary of things that tend to constitute sure classes, as well as people, organizations, physical elements, and social elements. whereas there's a huge array of styles of assets that may be found in a given context, and diverse approaches to reason them, communities typically establish six sorts: (1) individuals, (2) native teams or associations, (3) organizations and institutions, (4) physical house and infrastructure, (5) economic characteristics and (6) culture. Separating these types additional and giving a more in-depth interpretation of each can make it a lot of clear what individuals look for after they embrace plus mapping. Hypothetically and stormily the thought of assets is extremely powerful thanks to its positive approach and it feeds into current thinking of authorization and self-sufficiency. All but, however, changes from want based mostly approach to strengths-based one isn't easy. In this paper we have a tendency to principally concentrate on desires assessments and community deficiencies. perhaps, government organizations and aerial communities need some applicable tools to form the concepts in to one in a good way, which might result in increased understanding of their surroundings and higher call making. When plus mapping is employed as a central part in action research, each tangible and intangible results are generated. Tangible results typically take the shape of specific community building or economic development activities that emerge out of the enhanced awareness of residents and organizations regarding their own capability to act effectively. These results could embrace things cherish organizing residents for campaigns on native issues, creating employers attentive to the abilities of residents as potential employees, registering voters and serving to individuals participate within the choice method etc., Intangible results are harder to quantify, and that they occur in the process of participating individuals and making connections and linkages among them. As people interact with their neighbors in productive activities, trust and social capital are expanded, and a few of the barriers to participation are removed. To boost and predict the tangible ends up in a selected community, the aerial specification exploitation QGIS and fashionable machine learning algorithms are used for the meliorate of community development. To address the requirement of growing for complete inventories, several government and personal agencies are proactively looked into Geographical data Systems (GIS). There are many on-line services collect street-level wide assets information on a really huge scale like, Google Earth, Google street-view, Microsoft street aspect etc., as shown in Fig1. the provision of those databases

ISSN: 2233-7857 IJFGCN

offers the likelihood of playacting automatic survey of college assets and address the present issues. In particular, exploitation google earth positioning of location will scale back the amount of redundant enterprise system that collect and manage the college assets. Applying completely different machine learning algorithms to those large collections of information has produce the mandatory inventories a lot of efficiently.

The changes in illumination, clutter, variable positions and orientation, the intra-class variability will challenge the task of plus mapping and classification. Using these trends and technology, oft updated google earth, this paper presents an end-to-end system to notice and classify school assets and map their locations -together with kind – on google maps. The projected system has 2 key components: 1) an API(Application Programming Interface) that extracts location data exploitation google Earth 2) a machine learning algorithms are capable of sleuthing and classifying multiple categories of college assets. In easy terms, the system out-source the task of information assortment and reciprocally provides associate in surfing correct geo-spatial localization of school plus in conjunction with the data cherish street number, city, state, zip-code and sort of asset by visualizing then on the google map. It conjointly provides automatic inventory queries permitting professionals to pay less time sorting out assets, rather concentrate on a lot of necessary task of observance existing conditions. within the following connected work for asset mapping of school inventory management is concisely reviewed. Next, the algorithms for predicting school asset patterns and characteristic heat map are conferred in detail.

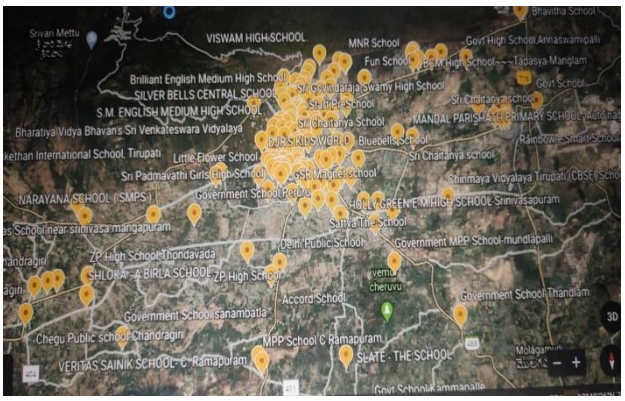


Fig 1: Asset Mapping using Google Earth

2. Preliminaries and Notations

Asset mapping consists of facts compendium, situation espial, asset audits, analysis and culpable. It could be very crucial to consciousness on accumulating accurate facts and enhancing facts great facts. Conditional tracking structures have become commercially appealing and might even include integrated withinside the destiny property bought via way of means of the utilities.

Creating a shape file from KML file:

ISSN: 2233-7857 IJFGCN

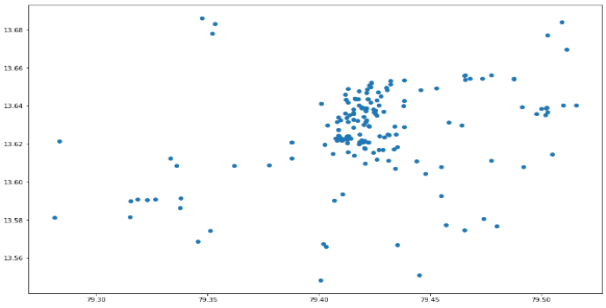


Fig 2: Outcome of shapefile in python

First, create a KML file in google earth and this file could be default saved in google drive. The geographical coordinates of the assets could be mapped primarily based totally on the latitude and longitude of the community assets. This KML file allows creating a shapefile for python-primarily based totally illustration as Fig 2. The QGIS application converts the KML file into a shapefile via way of means of the use of the ESRI shapefile layout. The final results could be a .shp file and this shapefile could be an entry for python-primarily based totally illustration. Geospatially permit your current big data cluster with spatial analytics tools, machine learning algorithms, and artificial intelligence strategies that permit you to reveal patterns, relationships, and incidents in massive quantities of data, no matter layout and source. Classification could be done at the spatial data generated through QGIS.

Classification is a supervised machine learning technique that maps input records into groups or classes [1]. The major circumstance for making use of a classification method is that each one data items should be assigned to lessons, and that every of the records items ought to be assigned to simplest one elegance. Classifiers may be binary classifier or Multi-elegance classifiers. Binary classifiers are Classification with simplest 2 wonderful lessons or with 2 viable outcomes. Multi-Class classifiers are Classification with more than distinct classes.

One, not unusual place class method primarily based totally on using distance measures is k-Nearest Neighbors (k-NN) [3]. The conventional k-NN classification algorithm reveals the k-nearest neighbors(s) and classifies numerical data facts via way of means of calculating the distance among the test the pattern and all training samples the usage of the Euclidian distance [4]. The primary recognition of the k-NN classifier has been on data units with natural numerical features [5]. However, k-NN also can be implemented to a different form of data consists of express data [6]. Several investigations had been finished to reveals a right express degree for such records K-NN categorized an item via way of means of a majority vote of the item's neighbors, withinside the area of input parameter. The item is assigned to the elegance that is maximum not unusual place amongst its k- an integer specific by a human nearest neighbor. It is a non-parametric algorithm because it does now no longer make any assumption on records distribution, the records do now no longer ought to be generally distributed. It is lazy because it does now no longer actually examine any version and make generalization of the records. It does now no longer teach a few parameters of a few characteristic in which enter X offers output y.

Definition 1: A distance measured ꞉ $X \times X \rightarrow R$ is a characteristic Definition 1known as metric if it satisfies the subsequent requirements [16] $\forall x, y, z \in X$:

1. $0 \leq d(x, y)$(Non-negative);

2. $d(x, y) = 1$, if and simplest if $x = y$ (Identity);

3. $d(x, y) = d(y, x)$ (Symmetry);

4. $d(x, z) \leq d(x, y) + d(y, z)$ (Triangle inequality).

However, the similarity dimension suggests extra debates, because it gives a few flexibilities withinside the identity of the way near records objects could be. A similarity degree is generally perceived as complementary to a distance measure.

ISSN: 2233-7857 IJFGCN

Definition 2: similarity measure $S : X \times X \rightarrow R$ is a function that satisfies the following requirements $\forall x, y \in X$:

1. $0 \leq S(x, y)$(Non-negative);

2. $S(x, y) = 1$,if and only if $x = y$ (Identity);

3. $S(x, y) = S(y, x)$ (Symmetry)

It additionally aims to supply a primary try of steerage on the simplest combination of the many functions which will be used with k-NN for various knowledge classification. during this a part of work, the performance of k-NN classification on numerical dataset victimisation two varieties of measures. The well-known (Euclidean and Manhattan) distances and therefore the combination of similarity measures that are fashioned by fusing existing numerical distances with binary data distances.

3. Related Work

Plenty of studies investigated , analyzed , and evaluated the performance of k NN on pure numerical and pure categorical data sets .Regarding applying k NN to heterogeneous data described by numerical and categorical features , the most widely used method is to treat the data before feeding it to the classifier .This can be done by converting non numerical features into numerical features using different techniques , and then the traditional k NN can be applied with any numerical distance. A study presented by Hu et al evaluated the performance of k NN on three types of school data sets, pure numerical, pure categorical, and mixed data using different numeric measures.

Distance-based classification algorithms are techniques used for classifying data objects by computing the distance between the test sample and all training samples using a distance function. Distance-based algorithms though were originally proposed to deal with one type of data using distance-based measurements to determine the similarity between data objects. These algorithms were subsequently developed to enable handling of heterogeneous data as real-world data sets are often diverse in types, format, content and quality, particularly when they are gathered from different sources. In general, when classifying heterogeneous data using distance-based algorithms, there are two categories of methods. The first category converts values from one data type to another and then, distance based algorithms can be used with an appropriate measurement to classify the data .However, this method is not effective as the similarity measure of the transformed data does not necessarily represent consistently the similarity of the original heterogeneous data, especially when the transformation is not fully reversible.

The second category extends distance-based algorithms to match the heterogeneous data. This can be done using distance measures that can handle heterogeneous data. One common classification technique based on the use of distance measures is k nearest neighbors (K- NN). The traditional k NN classification algorithm finds the k nearest neighbors and classifies numerical data records by calculating the distance between the test sample and all training samples using the Euclidian distance. The primary focus of the k NN classifier has been on data sets with pure numerical features . However, k NN can also be applied to other types of data includes categorical data. Several investigations have been done to find a proper categorical measure for such data. On the other hand, studies have used the combination approach for classifying heterogeneous data using K- NN. Such a study presented by Pereira et al. 28 has proposed a new measure for computing the distance between heterogeneous data objects and used this measure with k NN. This distance is called the Heterogeneous Centered Distance Measure (HCDM). It is based on a combination of two techniques Nearest Neighbors Classifier CNND distance for numerical features and Value Difference Metric VDM with k NN for classifying heterogeneous data sets, described by two different features type; numerical and categorical.

The combination measures include Heterogeneous Euclidean Overlap Metric HEOM, which uses the overlap metric for categorical features and the normalized Euclidean distance for numerical features; Heterogeneous Manhattan Overlap Metric HMOM, which uses the overlap metric for categorical features and Manhattan distance for numerical features; Heterogeneous Distance Function HVDM which uses the Value Difference Metric VDM for categorical features and the normalized Euclidean distance for numerical features .

ISSN: 2233-7857 IJFGCN

Generally, the most commonly used approaches for classifying heterogeneous data by k NN classifier can be described as a mixture of numerical and categorical features which included is:

1. The conversion approach a method of converting the data set into a single data type, and then applying appropriate distance measures to the transformed data.
2. Unified approach a method to integrate two or more different measures to infer the overall value.

Methodology

In this study, we will investigate the performance of K-NN for classifying heterogeneous data by using different locations of school assets.

We have chosen the most representative measures from these school assets, as they have been applied with K-NN in different studies for classifying the data and represent good references for critical comparisons of result. The five chosen measures belong to the following families:

1. Lp Minkowski family it is also known as the p-norm distance. The chosen measures from this family include:

(i) Manhattan distance is defned by:

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

(ii) Euclidean distance is defned by:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} |xi - yi^{2}|}$$

2. Inner product family distance measures belonging to this family are calculated by some products of pair wise values from both vectors. Two measures have been selected from this family:

(i) Cosine similarity measure is defined by:

$$S(x,y) = \frac{X.Y}{||X||Y||} = \frac{\sum_{i=1}^{n} xi.yi}{\sqrt{\sum_{i=1}^{n} xi^{2}}.\sqrt{\sum_{i=1}^{n} Yi^{2}}}$$

(ii) Jaccard distance is defined by:

$$d(x,y) = \sum_{i=1}^{n} \frac{xi - yi^{2}}{(xi)2 + (yi)2 - [(xi)(yi)]}$$

1. L1distance family the distances in this family are calculated based on fnding the absolute diference. Only one measure has been chosen from this family:

$$d(x,y) = \sum_{i=1}^{n} \frac{|xi - yi|}{|xi| + |yi|}$$

i) Canberra distance is defined by: As we mentioned in this section, the chosen measures have been widely applied with k-NN for classifying the datasets.

Most the equations are confirmed metrics: Euclidean, Manhattan, Canberra according to [34, 35], and Jaccard according to [36], satisfy the Cosine measure is not metric. It does not satisfy condition 4 in Definition 1.

Generally, categorical data is classified as a type of qualitative data [37]. Such data corresponds to a possible representation for nominal, binary, ordinal, and interval instances. For the sake of simplicity,

ISSN: 2233-7857 IJFGCN

in this work, we will focus on only condition in one type of categorical data which is binary data. The set of measures developed for dealing with binary data is known as matching coefficients [38]. They calculate the distance between two data objects x and y defined as x = {x1, x2,…, xp}, and y = {y1, y2,…, yp}, where p represents the number of binary features in The strategy behind these methods is that the two data objects are viewed as similar to the degree that they share a common pattern of feature values among the binary variables. The matching coefficient values range between 0 for not similar at all and 1 for completely similar [39]. Figure 3 shows the main four quantities of binary features. Any binary feature has only one of two cases: 0 means that the feature is absent and 1 means that the feature is present, this is called symmetric binary features [39]. Those are listed below:

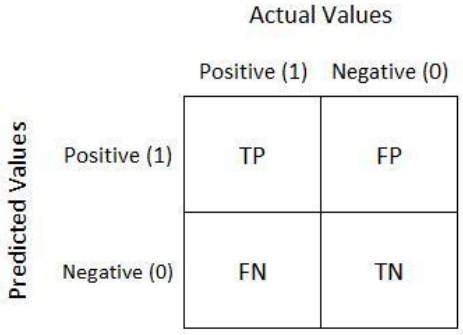
(i) TP represents the total number of features in both x and y have a value of 1.

(ii) FN represents the total number of features where the feature of x is 0 and y is 1.

(iii) FP represents the total number of features where x is 1 and y is 0.

(iv) TN represents the total number of features in both x and y have a value of 0.

Each feature in data objects must belong to one of these four categories TP FN, FP, and TN, and TP+ FN +FP +TN = P, where P is the total number of binary features.



There are many methodologies applied to overlap measure with k-NN for both classification and regression tasks. They used overlap measure for comparing categorical (nominal/ binary) data However, the main limitation of this measure is that this measure only determines whether the features are match to one another (TP and TN), and does not make full use of the rest of the classification information. Therefore, in this study, Jaccard coefficient similarity measure is adopted to deal with binary data and is defined as:

$$S(x, y) = \frac{TP + TN}{P}$$

It should be noted that the Jaccard coefficient similarity measure excludes TN from consideration which represents joint absences for both features. According to [43], the TN value does not necessarily represent a resemblance between data objects, since a large proportion of the binary dimensions in two data objects are more likely to have negative On the other hand, the study presented by Faith et al. [44] considered TN value in the calculation of comparing binary data. However, these studies showed that positive matches as more considerable, therefore they give the former less weight comparing to the negative matches.

ISSN: 2233-7857 IJFGCN

4. Experimental Results

This section evaluates the effectiveness of both traditional k-NN, and k-NN with the combination of similarity measurements over two heterogeneous data sets from different domains. The data sets are described by mixtures of numerical and binary features only.Every dataset should satisfy the following conditions:

1. Data set should contain numerical and binary features only.

2. The data should not contain more than 3% of missing values.

3. The number of features for each type of data should be enough for calculating the similarity (not less than 2).

4. The number of classes should be small.

Before running the experiments, all datasets were preprocessed by removing irrelevant features (ID), and data objects with missing values. Numerical features were normalized to fall between 0 and 1. Each data set was split randomly into 80% for training and 20% for the testing sets.

Five k values were evaluated: 1, 3, 5, 7 and 9 neighbors. We investigated the implementation of k-NN with different categories of measures; among them the first category includes Euclidean and Manhattan measures. It should be noted that we applied normalized Euclidean and normalized Manhattan distances to numerical datasets. Therefore, all the obtained results fall between 0 and 1. Because the similarity is complementing of the distance, in this study the similarity is computed based on: All the measures are used with the k-NN classifier individually with three different weights, and these measures are applied with k-NN to the same training and test samples each time. For evaluating the performance of k-NN we have used both accuracy (A) and F-score (F) metric.

It should be noted that:

1. The values of w1 and w2 are set by default as following:

(i) When the numerical features are most impotent than the binary features, we set w1 = 0.8 and w2 = 0.6.

(ii) When the binary features are most impotent than the numerical features, we set w1 = 0.6 and w2 = 0.8.

(iii) When the numerical and binary features have the same degree of importance, we set w1 = 0.5 and w2 = 0.5. 2.

The values w1 = 0 and w2 = 1 or w1 = 1 and w2 = 0 are not suggested for heterogeneous data because this leads to using a single measure, negating the advantages of a combined measures.

The implementation of classifying heterogeneous data can be summarized in the following steps:

1. For each data, set the value of k, w1 and w2.

ISSN: 2233-7857 IJFGCN

2. Split the data randomly into 80% for training and 20% for the test sample.

3. Apply k-NN with the measures Euclidean, Manhattan independently to the data set.

4. Repeat steps 2 and 3 for a number of times (3 times).

5. Calculate the average of both accuracy and F-score values.

As it can be seen from the experiments, for traditional k-NN, the results showed that k-NN with Manhattan distance produces better results compared to the classifier with Euclidean distance for all data sets and all k values. The experiments showed that k-NN with the combination of similarity measures performs well for classifying the six heterogeneous data sets, and outperforms k-NN with Euclidean distance. The two combination of similarity measures are efficient in handling both numerical and binary features together.

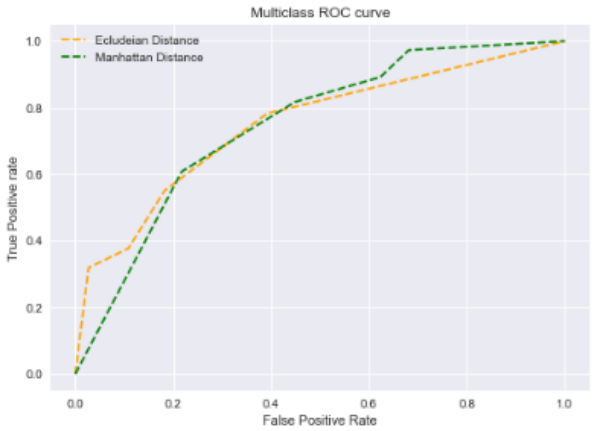
Table 1: k-NN classification is based on Ecludiean Distance

	precision	recall	F1-score	support
0	0.76	0.46	0.58	69
1	0.79	0.93	0.85	148
accuracy			0.78	217
macro avg	0.78	0.7	0.72	217
weighted avg	0.78	0.78	0.77	217

	precision	Recall	F1-score	support
0	0.78	050	0.64	72
1	0.82	0.95	0.90	162
accuracy			0.82	230
macro avg	0.80	0.74	0.77	230
weighted avg	0.80	0.80	0.9	230

Table 2: k-NN classification is based on Manhattan Distance

ISSN: 2233-7857 IJFGCN



The Table 1 and Table 2 shows that the accuracy, macro average and weighted average for precision, recall and F1-Score for the data set with locations latitude and longitude. With this result, the distance between each and every nearest school is calculated with both Euclidean distance and Manhattan distance but, the overall accuracy of this distance with accuracy is more in K-NN with Manhattan distance that Euclidean distance.

Conclusion

Since the k-NN classification is based on measuring the distance between the test sample and each of the training samples, the chosen distance function plays a vital role in determining the final classification output. The major objective of this study was to investigate the performance of k-NN, using several measures includes single measures (Euclidean and Manhattan) and a number of combinations of similarity measures, for computing the similarity between data objects described by numerical and binary features. Experimental results were carried out on two heterogeneous data sets from different domains. The overall results of our experiments showed that Euclidean distance is not an appropriate measure that can be used with k-NN for classifying a heterogeneous data set of numerical and binary features. Furthermore, our results showed that combining the results of numerical and binary

References

1. Han J, Pei J, Kamber M (2011) Data mining: concepts and techniques. Elsevier, Amsterdam
2. Shavlik JW, Dietterich T, Dietterich TG (1990) Readings in machine learning. Morgan Kaufmann, Los Altos
3. Cover TM, Hart P (1967) Nearest neighbor pattern classifcation. IEEE Trans Inf Theory 13(1):21–27
4. Tan P-N (2018) Introduction to data mining. Pearson Education, Chennai
5. Wettschereck D (1994) A study of distance-based machine learning algorithms
6. Bramer M (2007) Principles of data mining, vol 180. Springer, Berlin
7. Hu L-Y, Huang M-W, Ke S-W, Tsai C-F (2016) The distance function efect on k-nearest neighbor classifcation for medical datasets. SpringerPlus 5(1):1304
8. Singh A, Halgamuge MN, Lakshmiganthan R (2017) Impact of diferent data types on classifer performance of random forest, naive Bayes, and k-nearest neighbors algorithms. Int J Adv Comput Sci Appl 8:1
9. Sentas P, Angelis L (2006) Categorical missing data imputation for software cost estimation by multinomial logistic regression. J Syst Softw 79(3):404–414
10. Todeschini R, Ballabio D, Consonni V, Grisoni F (2016) A new concept of higher-order similarity and the role of distance/ similarity measures in local classification methods. Chemom Intell Lab Syst 157:50–57

ISSN: 2233-7857 IJFGCN

11. Jiang L, Cai Z, Wang D, Jiang S (2007) Survey of improving k-nearest-neighbor for classification. In: Fourth international conference on fuzzy systems and knowledge discovery (FSKD 2007), vol 1. IEEE, pp 679–683
12. Liu C, Cao L, Philip SY (2014) Coupled fuzzy k-nearest neighbors classification of imbalanced non-IID categorical data. In: 2014 international joint conference on neural networks (IJCNN). IEEE, pp 1122–1129
13. Walters-Williams J, Li Y (2010) Comparative study of distance functions for nearest neighbors. In: Elleithy K (ed) Advanced techniques in computing sciences and software engineering. Springer, Berlin, pp 79–84
14. Deza MM, Deza E (2014) Encyclopedia of distances. Springer, Berlin ISBN 9783662443422
15. Jajuga K, Sokolowski A, Bock H-H (2012) Classification, clustering, and data analysis: recent advances and applications. Springer, Berlin
16. Deza MM, Deza E (2009) Encyclopedia of distances. Springer, Berlin, pp 1–583
17. Evelyn F, Hodges JL Jr (1951) Discriminatory analysis-nonparametric discrimination: consistency properties. Technical report, California University, Berkeley
18. Mohammed M, Khan MB, Bashier EBM (2016) Machine learning: algorithms and applications. CRC Press, Boca Raton
19. Larose DT (2015) Data mining and predictive analytics. Wiley, New York
20. Larose DT, Larose CD (2014) Discovering knowledge in data: an introduction to data mining. Wiley, New York
21. Weinshall D, Jacobs DW, Gdalyahu Y (1999) Classification in non-metric spaces. In: Advances in neural information processing systems, pp 838–846
22. Chomboon K, Chujai P, Teerarassamee P, Kerdprasop K, Kerdprasop N (2015) An empirical study of distance metrics for k-nearest neighbor algorithm. In: Proceedings of the 3rd international conference on industrial application engineering, pp 1–6
23. Prasath VB, Alfeilat HAA, Lasassmeh O, Hassanat A, Tarawneh AS (2017) Distance and similarity measures effect on the performance of k-nearest neighbor classifier—a review. arXiv preprint arXiv:1708.04321
24. Cunningham P, Delany SJ (2007) k-nearest neighbour classifiers. Mult Classif Syst 34(8):1–17
25. Todeschini R, Ballabio D, Consonni V (2006) Distances and other dissimilarity measures in chemometrics. In: Meyer RA (ed) Encyclopedia of analytical chemistry: applications, theory and instrumentation. Wiley, New York, pp 1–34
26. Lopes N, Ribeiro B (2016) On the impact of distance metrics in instance-based learning algorithms. In: Iberian conference on pattern recognition and image analysis. Springer, Berlin, pp 48–56
27. Ali N, Rado O, Sani HM, Idris A, Neagu D (2019) Performance analysis of feature selection methods for classification of healthcare datasets. In: Intelligent computing-proceedings of the computing conference. Springer, Berlin, pp 929–938
28. Pereira CL, Cavalcanti GDC, Ren TI (2010) A new heterogeneous dissimilarity measure for data classification. In: 2010 22nd IEEE international conference on tools with artificial intelligence, vol 2. IEEE, pp 373–374
29. Deekshatulu BL, Chandra P (2013) Classification of heart disease using k-nearest neighbor and genetic algorithm. Procedia Technol. 10:85–94
30. Cha S-H (2007) Comprehensive survey on distance/similarity measures between probability density functions. City 1(2):1
31. Liu H, Zhang S (2012) Noisy data elimination using mutual k-nearest neighbor for classification mining. J Syst Softw 85(5):1067–1074
32. Batista G, Silva DF et al (2009) How k-nearest neighbor parameters affect its performance. In: Argentine symposium on artificial intelligence, pp 1–12
33. Peterson MR, Doom TE, Raymer ML (2005) Ga-facilitated KNN classifier optimization with varying similarity measures. In: 2005 IEEE congress on evolutionary computation, vol 3. IEEE, pp 2514–2521
34. Akila A, Chandra E (2013) Slope finder—a distance measure for DTW based isolated word speech recognition. Int J Eng Comput Sci 2(12):3411–3417

ISSN: 2233-7857 IJFGCN

35. Yang K, Shahabi C (2004) A PCA-based similarity measure for multivariate time series. In: Proceedings of the 2nd ACM international workshop on multimedia databases. ACM, pp 65–74
36. Cesare S, Xiang Y (2012) Software similarity and classification. Springer, Berlin
37. Silverman D (2006) Interpreting qualitative data: methods for analyzing talk, text and interaction. Sage, Beverly Hills
38. Dillon WR, Goldstein M (1984) Multivariate analysis methods and applications. Number 519.535 D5
39. Finch H (2005) Comparison of distance measures in cluster analysis with dichotomous data. J Data Sci 3(1):85–100
40. Choi S-S, Cha S-H, Tappert CC (2010) A survey of binary similarity and distance measures. J Syst Cybern Inform 8(1):43–48
41. Spencer MS, Prins SCB, Beckom MS et al (2010) Heterogeneous distance measures and nearest-neighbor classification in an ecological setting. Mo J Math Sci 22(2):108–123
42. Salvador-Meneses J, Ruiz-Chavez Z, Garcia-Rodriguez J (2019) Compressed KNN: K-nearest neighbors with data compression. Entropy 21(3):234
43. Sokal R, Sneath PHA (1963) Principles of numerical taxonomy. W.H. Freeman, San Francisco
44. Faith DP, Minchin PR, Belbin L (1987) Compositional dissimilarity as a robust measure of ecological distance. Vegetatio 69(1–3):57–68

ISSN: 2233-7857 IJFGCN