Research Article

International Journal of Advanced Technology and Engineering Exploration, Vol 8(75)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2020.762106

Comparison of affinity degree classification with four different classifiers in several data sets

Rosyazwani Mohd Rosdan*, Wan Suryani Wan Awang and Wan Aezwani Wan Abu Bakar
Faculty Informatics and Computing, University Sultan Zainal Abidin, Besut Campus, Terengganu, Malaysia

Received: 20-November-2020; Revised: 22-February-2021; Accepted: 25-February-2021

Abstract

The affinity notion has been widely used in research fields. Thus, in this research, affinity is employed to find the degree between two data sets and classify through prediction. But, as Affinity Degree (AD) classification is a new technique, the comparison with different classification types is needed to test the compatibility technique. Herein, this study compares various machine learning techniques and determines the most efficient classification technique based on the data set. Four different classification algorithms, K-Nearest Neighbour (KNN), Naive Bayes (NB), Decision Tree (J48), and Support Vector Machine (SVM), were used as other techniques to compare with AD classification. Three different data sets, breast cancer, acute inflammation, and iris plant, were used for experiment purposes. The results show J48 has the best rate in performance measures compare to the other four classifiers. However, the results of AD classification show the significance that more studies can improve it.

Keywords

Affinity degree (AD), K-nearest neighbour (KNN), Naive bayes (NB), Decision tree (J48), Support vector machine (SVM).

1.Introduction

Today, the affinity notion has been widely used in various fields such as biochemistry, business, and computing. In general, affinity is a natural linking of someone with something or attraction to a person, idea, or thing. Affinity can be different; the meaning depends on the situation. In chemistry, affinity was defined as the tendency of a molecule to bind with another molecule. It can also be an identifier for the immobilization or separation that can be carried [1]. In comparison, business affinity can be defined as consumer interest in the products [2]. Affinity may also be adequate for creating a composed service by linking two services or pointing out a similar service [3]. In this study, affinity was defined as the correlation, relationship, similarity, or tendency between two objects [4]. Adapted from peer-to-peer data replication [5], the term affinity degree (AD) refers to measuring the degree of correlation between two objects.

AD classification is a classifier that predicts based on similarity concept in the degree of correlation. More details about AD classification will elaborate later in section 3. Therefore, in machine learning classification, one technique used a similarity-based concept as a classifier predictor.

Machine learning classification has been used extensively in various fields. A technique that acts according to the If-Then rule aims to predict a variable based on other features known as predictors [6]. K-Nearest Neighbour (KNN) [7] is a classifier that predicts used distance measure depends not only on the closest distance sample data point but also on the value of K [8, 9]. Pointed out by Pelillo, the NN rule was first invented by Alhazen (Ibn al-Haytham), a very early scientist in the year 965 until 1040 in optics and perception. Therefore, the rule becomes widely famous after M. Cover, and P. E. Hart implied the function KNN where the prediction based on the nearest neighbor of samples by estimating K value [10, 11].

Naive Bayes (NB) [12] is a model that assumes that within each class, the measurements are independent. The term 'Bayes' also appears in the names because

*Author for correspondence

This work was supported in part by the Department of Informatics and Computing, University Sultan Zainal Abidin, Besut Campus, Terengganu, Malaysia.

of the Bayes theorem's use to deduce class membership probabilities. The model, referring to the class indicator's conditioning, is often defined as assuming conditional independence. The Decision Tree (DT) structure also includes root nodes, branches, and leaf nodes, much like an actual tree [13]. DT is a straightforward model that provides successful interpretations. A decision tree is a tree where a function, the attribute is shown by each node, a decision which is the rule is shown by each link, the branch and each leaf shows an outcome, the categorical or continuous value. The whole concept is building such a tree for the entire data and processing a single result on each leaf. First suggested by Durgesh and Lekha [14], the Support Vector Machine (SVM) is a binary category that belongs to a generalized linear classification family.

This research aims to see the compatibility of AD classification in prediction by comparing four different classification algorithms on the three different data sets. Thus, this study used KNN, NB, J48, and SVM using WEKA for comparison purposes. Besides, three different data sets with various objects have been used. There is early diagnosis of breast cancer using the Coimbra breast cancer data set [15], diagnosis of acute nephritis of the renal pelvis using the Acute Inflammations data set [16], and predict the class of iris plant using Iris Plants Database [17].

2.Literature review

This section summarizes various technical articles on the KNN technique applied to different predictions and related work on the affinity definition.

Nowadays, the affinity term is no longer strange in research fields. In 2018, Li et al. [1] studied Mobile Affinity Sorbent Chromatography (MASC). Affinity describes immobilized or portable separation in this analysis. This research aims to secrete interest analytes with a high degree of selectivity and affinity in particular. Besides, a small number of recurrent analysts can be defined and quantified in a large number of complex samples. When the analysis is selected based on specific 3D structural characteristics, the isocratic elution is separated with a mobile phase column. The analytes are collected and transited by columns of transport, which elucidate the volume in the column void.

In 2015, Bakhouya and Gaber [3] suggested adaptive systems and approaches. The research concentrates on the development of adaptive engineering systems through natural and biological systems. These programs' design helps organizations choose the best behavior approach based on current system status and environmental changes. Therefore, the affinity was described as the adequacy with which two services might commit themselves to build a composite service or pointing out similar services.

In 2017, Halim and Zulkarnain [18] research the market affinity and international image with actions and purchasing will. The relation between consumer and foreign culture, the emotions that the consumer possesses against the product, and changes in the trend, lifestyle, or personal interest of an item, may differ from those described in this study. This affinity can have tremendous effects on retail firms. The findings show that the effect of ethnocentricity on consumer behavior is dominant compared to affinity. Meanwhile, country goods play a role in encouraging the consumer's wish to buy the commodity.

Awang et al. [5] proposes using popularity and affinity as a strategy to optimize the maximum benefits from file replication, for selecting and accessing suitable models in distributed environments. Herein, the affinity degree defined the similarity between two or more correlated data. The affinity set is a set of any data that creates an affinity between files. Thus, sets A and B are the set consisting of the intersection of elements between A and B data and is not null. The cardinality of the affinity set A and B over A is the definition of affinity degree between data set A and data set B concerning A. An affinity degree's value was then categorized into five categories, adapted from Dancey and Reid [19].

Presented in 2021 by Assegie [20], value K can determine or produce better breast cancer detection accuracy. Also, the writer investigated the hyper-parameter tuning effect of breast cancer detection KNN in this study. The results of this research show that tuning the hyper-parameter affects the efficiency of the KNN model significantly. The hyper-parameter tuning effect is experimentally evaluated using a Wisconsin breast cancer data set from the Kaggle database. Lastly, the KNN has been compared to the tuned hyper-parameter and the regular hyper-parameter. The results of the model's output with the tests show that the optimized model is precisely 94.35%, and the default hyper-parameter for the KNN is 90.10%. The tests show that the model has 94.25% accuracy.

Comparison of machine learning algorithm on the spatial prediction of landslide hazards in the Hoa Binh province of Vietnam in a paper submitted by Dieu Tien Bui and colleagues [21]. First, the map of 118 landslides in constructed position inventory from various sources. The landslide was randomly divided into 30 percent for training the model and 70 percent for model validation. Second, ten landslide conditioning variables, such as the angle of slope, aspect of slope, amplitude of relief, and various other factors, were chosen. Using SVM, DT, and NB models, the landslide susceptibility indexes were determined. Finally, to verify and compare the landslide susceptibility maps, the researchers used the different landslide positions in the training process. The results of the validation show that the models derived using SVM have the highest capability for prediction. The model derived using DT has the lowest capability for prediction. The prediction capacity of the SVM models is marginally better compared to the logistic regression model. The prediction capacity of the models DT and NB are lower.

Pharswan and Singh [22] proposed a system based on machine learning to classify breast cancer and a comparative study of two SVM and KNN machine learning. The concern of classifier for selecting the region of interest (ROI) from the mammograms is still a difficult task to precisely detect and classify the cancer tumor. The trained classifier was used to classify the mammogram into either benign or malignant. Both classifiers' classification performance contrasts with accuracy, recall, precision, specificity, and F1 score. The findings show that even with better recall and F1 score, SVM achieved a higher efficiency of 94% than KNN.

In 2018, Thirunavukkarasu et al. [23] set the goal of developing a model capable of recognizing iris species using the K-Nearest neighbors' classification algorithm (KNN). The model is implemented through six simple machine learning phases. Data collection or data preparation, algorithm selection, model object creation, model formation, unexpected data or test data, and model evaluation prevail. The results show that of three classes, two classes are 100% accurate, but only one class is 96% correct because the result is misclassified.

3.Materials and methods

There are four based sections in classification: function, probability, similarity, and rule-based. Among these sections, similarity-based is the best classifier predicting the samples' class label [24].

3.1Affinity degree classification (AD)

AD is used to find the degree of the relationship between set A and set B. The degree than was used to predict the classification of the data set. The definition of affinity itself, which means it can include the similarity, relationship, dependency, and correlation between two or more objects in the similarity-based classification. For degree calculation, let A as a group of the population, B as a predicted class, and BT is the predicted class target.

$$AffinityDegree = \frac{|(A \cap B) + B_T|}{|A + B_T|} \tag{1}$$

Given showed the flow of AD in *Figure 1*. Start with data set, pre-processed is necessary if the data set were incomplete with missing value or the change of value from numeric to non-numeric or vice versa. Then, calculate the affinity degree between set A and B before predicted the attribute class. For classification purpose, there are some alternate step has been done. If the affinity degree in data replication is adapted from Dancey and Reid for the parameter to classify or rank the degree, but in this experiment, we used the highest value of affinity degree itself as a parameter to classify the predicted attribute. Given an example 1 for measurement and predicted class. Lastly, evaluate the classification results by comparing the results with the actual data set to define the TP, FN, FP, and TN. The implementation of this process will be detailed in the following sections.

Example 1

A new student entered school, and the teacher needs to assign the student to the department according to his/her previous subject’s record. Let set A as student previous subjects record and set B as departments. So, A = (x_1, x_2, x_3. x_4, x_5) and set B = (y_1, y_2), to predict which department is most suitable with the student, using the equation in (1), set A will be calculated with both (y_1) and (y_2).

$(A \cap B) + (y_1) / A + (y_1)$
= (50) + 50 / 150 + 50
= 100 / 200
= 0.5

$(A \cap B) + (y_2) / A + (y_2)$
= (30) + 50 / 150 + 50
= 80 / 200
= 0.4

From the calculation above, the student will be assigned in the department (y_1) as the value of the affinity degree higher than (y_2).

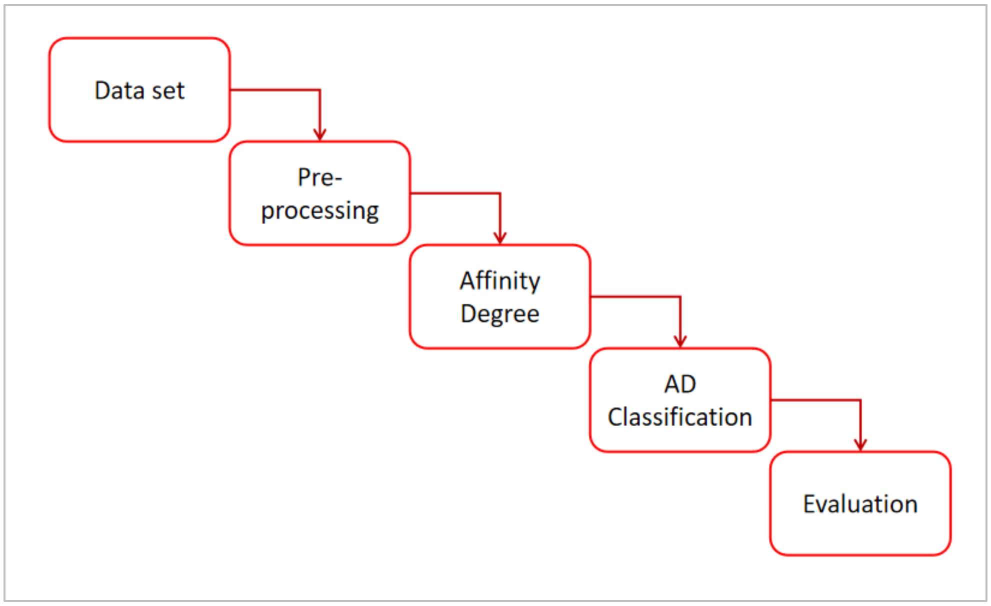

Figure 1 Affinity degree classification algorithm

3.2K-nearest neighbour (KNN)

KNN falls under similarity-based as the method is about measuring the similarity between objects or data and classifying based on the pairwise similarities. Even though KNN most widely used learning algorithm, in an attempt to predict a new sample point classification, KNN also classifies datasets through various classes. Nearest Neighbours or NN refers to a variable identical to other variables with measurements of the shortest distances. K refers to the number of immediate neighbors used to make the prediction. A threshold-based method is invariably used instead of the KNN-based approach, where all objects with similarity above a given value are taken when determining a prediction. For this study, KNN classification, the data was explorer using WEKA with K=3 nearest neighbor (s) for classification with 10-fold cross-validation and Euclidean distance function.

3.3Naive bayes (NB)

The algorithm for Naive Bayes [25] is a straightforward probability classifier that calculates a probability set by counting the frequency and value combinations in a given data set. The algorithm used in Bayes' theorem claims that all variables are independent given the class variable's value. In real-world applications, this conditional independence assumption is rarely valid, so it is defined as Naive. Still, in a variety of controlled classification problems, the algorithm tends to learn quickly. Bayes' theorem is a mathematical formula used to calculate conditional probability, named after Thomas Bayes, the British mathematician of the 18th century.

3.4Decision tree (J48)

J48 [26] was used for both classification and prediction operations in the decision tree. This study chose J48 or C4.5 for classification because this algorithm is one of the most used tools in Weka that provides consistency between the accuracy, speed, and interpretability of results. This algorithm also classifies knowledge in the form of a decision tree to easily classify weak learners.

3.5Support vector machine (SVM)

SVM was designed for numerical input variables, while nominal values are converted to numerical values automatically. Before being used, input data is often normalized. SVM works by finding a line that better divides the information into the two classes. The parameter of complexity controls shows how flexible the procedure can be for drawing the line to distinguish the classes. A value of 0 does not allow any margin violations, while the default value is 1. The Kernel type to use is a crucial parameter in the SVM. The simplest kernel is a linear kernel that uses a straight line or hyperplane to separate data. Therefore, this study uses a Polynomial Kernel.

3.6Data set

This study used three data sets for experiment purposes: breast cancer diagnosis, acute nephritises diagnosis, and iris plant class prediction. Later, they elaborated on these data sets more in the following subsection.

3.6.1Breast cancer

According to the World Cancer Research Fund, breast cancer is the most prevalent cancer in women globally. There were over 2 million new cases reported in 2018, whereby Belgium were the highest rates of breast cancer with 113.2 age-standardized rate1 per 100,000 [27]. Risk factors for breast cancer including increasing age, female, age at menopause, family history, race or ethnicity, age at first pregnancy, parity, lifestyle, and use of hormone replacement therapy [28–30]. The World Health Organization suggests that early diagnosis is critical to improving breast cancer outcomes and survival [31]. The breast cancer data set featured 116 patients 20-80 years old, 64 were diagnosed with breast cancer, and 52 were indicated as healthy controls patients. The data sets have ten predictor-dependent variables, indicating the form of "Patients" and "Healthy Controls" data given in *Table 1*.

Table 1 Description of breast cancer data set attribute

Attribute	Description
Age	Age of patient x_1 : age1 (<46) x_2 : age2 (>45 and <70) x_3 : age3 (>69)
BMI	Body mass index (kg/m^2) x_4 : bmi1 (<25) x_5 : bmi2 (>24 and <31) x_6 : bmi3 (>30)
Glucose	Energy source in living organism (kg/dL) x_7: g1 (<131) x_8: g2 (>130)
Insulin	2-hour serum insulin (µU/mL) x_9: i1 (< 21) x_{10}: i2 (>20 and <40) x_{11}: i3 (>39)
HOMA	Homeostasis model assessment x_{12} : homa1 (<13) x_{13} : homa2 (>12)
Leptin	Hormone produced by adipocytes cell (ng/mL) x_{14} : leptin1 (<33) x_{15} : leptin2 (>32 and <62) x_{16} : leptin3 (>61)
Adiponectin	Secreted from adipose tissue (µg/mL) x_{17} : a1 (<3.5) x_{18} : a2 (>3.4 and <22.6) x_{19} : a3 (>22.5)
Resistin	Adipose tissue specific secretory factor (ng/mL) x_{20} : r1 (<43) x_{21} : r2 (>42)
MCP.1	Monocyte Chemoattractant Preotein-1 (pg/dL) x_{22} : mcp1 (<597) x_{23} : mcp2 (>596 and <1148) x_{24} : mcp3 (>1147)
Classification	Presence or absence of breast cancer y_1 : Patients y_2 : Healthy Controls

3.6.2Acute inflammation

The acute nephritic syndrome occurs with some conditions that cause glomeruli swelling and inflammation that filter the urinary portion of the kidney and eliminate waste from the blood [32]. Family history, immune system disorder, improper use of antibiotics or medication, and recent urinary tract operations are the few risk factors for this disease [33]. The researcher used this data set to perform the presumptive diagnosis of acute nephritis, the urinary system's diseases. Each instance of data generated represents a potential patient. Acute renal pelvic nephritis occurs in women much more than in men. It starts with a sudden fever, often hitting over 40C. The fever comes with shudders and lumbar pains, either on one or both sides, which are often very intense. Active urinary bladder inflammation signs occur very regularly. Nausea and vomiting, and pain spreading across the abdomen are not infrequently present. They are comprised of 120 various symptoms patients, the data given in *Table 2*.

Table 2 Description of acute inflammation data set attribute

Attribute	**Description**
Temperature of patient	x_1: low (≤38C) x_2: high (>38C)
Occurrence of nausea	x_3: yes x_4: no
Lumbar pain	x_5: yes x_6: no
Urine pushing (continuous need for urination)	x_7: yes x_8: no
Micturition pains	x_9: yes x_{10}: no
Burning of the urethra, itch, swelling of urethra outlet	x_{11}: yes x_{12}: no
Inflammation of the urinary bladder	x_{13}: yes x_{14}: no
Nephritis of renal pelvis origin	y_1: yes y_2: no

3.6.3Iris plant

The data set was created by Ronald Fisher and then submitted to UCI Repository by Michael Marshall in 1988. The Iris plant's data collection contains 150 examples of each type of Iris plant with three groups of 50 instances. As a dependent variable, there are four different types of single domains. *Table 3* provides information about the attributes. This data set aims to determine the pattern by analyzing and predicting the iris plant's sepal and petal size. Thus the length and width of the sepal and petal are positively connected. It is easy to define this relationship by naked eyes or without any instruments and formulas. The sepals are often larger than the sepals, and the petals' length is usually more significant than the petals.

Table 3 Description of iris plant data set attribute

Attribute	**Description**
Sepal length	x_1 :1 (< 6cm) x_2 : 2 (≥6cm)
Sepal width	x_3 : 1 (<3cm) x_4 : 2 (≥ 3cm)
Petal length	x_5 : 1 (<3cm) x_6 : 2 (≥ 3cm and <6 cm) x_7 : 3 (≥ 6cm)
Petal width	x_8 : 1 (<1cm) x_9 : 2 (≥ 1cm)
Class	y_1 : Iris Setosa y_2 : Iris Versicolour y_3 : Iris Virginica

4.Results

This section presents the experimental results and analysis done for this study. Five classifiers, AD classification, KNN, NB, J48, and SVM, are conducted. For comparison purposes, an alternate technique in AD has been done. For the other four, the data was calculated using WEKA with 10-fold cross-validation. While in AD, the data calculation has been done following the AD classification process stated in section 3.1. A confusion matrix has evaluated the efficiency of the proposed solution.

The classification accuracy is typically summarized by performance measures such as accuracy, sensitiveness, and specificity [34]. Specificity or true negative rate is the correct negative value in data like a cat and non-cat category. True negative is the number of positive non-cat. Sensitivity, recall, or true positive rate is the correctly positive ratio or the number of positive cats. In contrast, the false-positive rate is contradicting to true positive rate. If true positive give the number of the positive cat, false-negative give the number of supposed to be a cat but were non-cat. Precision or positive predictive value is the probability that the subject or the cat being classify in the cat category. F1 score was defined as the harmonic mean of the model's precision and recall, and it is essential to evaluate the imbalanced class distribution. Matthews correlation coefficient or MCC is an alternate technique to evaluate the imbalanced class distribution with a more informative and truthful score [35]. Accuracy is the most intuitive one due to the correctly labeled subjects' ratio to the whole issue.

Other than that, this analysis also measures output for AUC, MAE, and RMSE in terms of a graph. A graph with two curve plots parameter, TP rate and FP rate, which shows the output at all classification thresholds, is ROC or Receiver Operating Characteristic. The curve plots by decreasing or rising to label more objects as positive at different classification thresholds, and the region under ROC was called AUC. The calculation of error between observation and prediction is Mean Absolute Error or MAE. Meanwhile, the square root of the distinction between observation and prediction is RMSE or Root Mean Square Error. *Table 4* shows the comparison of the detailed accuracy of five applied machine learning algorithms under each performance evaluation explained above.

Table 4 Comparison of detailed accuracy of five applied machine learning algorithms

Case	Classifier	Class	TP Rate	FP Rate	Precision	Specificity	F1 score	MCC	Accuracy	AUC	MAE	RMSE
Breast Cancer	AD	no	0.481	0.609	0.391	0.391	0.431	-0.129	0.431	0.436	0.568	0.754
		yes	0.391	0.519	0.481	0.481	0.431	-0.129	0.431	0.436	0.568	0.754
		weight average	0.436	0.564	0.436	0.436	0.431	-0.129	0.431	0.436	0.568	0.754
	KNN	non-Patient	0.766	0.577	0.620	0.423	0.685	0.201	0.612	0.628	0.442	0.506
		Patient	0.423	0.234	0.595	0.766	0.494	0.201	0.612	0.628	0.442	0.506
		weight average	0.612	0.423	0.609	0.577	0.600	0.201	0.612	0.628	0.442	0.506
	NB	non-Patient	0.594	0.462	0.613	0.538	0.603	0.123	0.569	0.563	0.489	0.523
		Patient	0.538	0.406	0.519	0.594	0.528	0.123	0.569	0.563	0.489	0.523
		weight average	0.569	0.437	0.571	0.563	0.569	0.123	0.569	0.563	0.489	0.523
	DT	non-Patient	0.656	0.442	0.646	0.558	0.651	0.214	0.612	0.590	0.448	0.507
		Patient	0.558	0.344	0.569	0.656	0.563	0.214	0.612	0.590	0.448	0.507
		weight average	0.612	0.398	0.611	0.602	0.612	0.214	0.612	0.590	0.448	0.507
	SVM	non-Patient	0.828	0.788	0.564	0.212	0.671	0.050	0.551	0.520	0.448	0.669
		Patient	0.212	0.172	0.500	0.828	0.297	0.050	0.551	0.520	0.448	0.669
		weight average	0.552	0.512	0.535	0.488	0.503	0.050	0.551	0.520	0.448	0.669
acute diagnosis	AD	no	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.000

Case	Classifier	Class	TP Rate	FP Rate	Precision	Specificity	F1 score	MCC	Accuracy	AUC	MAE	RMSE
		yes	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.000
		weight average	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.000
	KNN	no	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	0.001	0.001
		yes	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	0.001	0.001
		weight average	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	0.001	0.001
	NB	no	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	0.011	0.026
		yes	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	0.011	0.026
		weight average	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	0.011	0.026
	DT	no	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.000
		yes	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.000
		weight average	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.000
	SVM	no	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.000
		yes	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.000
		weight average	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.000
iris plant	AD	setosa	0.960	0.026	1.000	0.974	0.980	1.000	0.960	0.819	0.180	0.816
		color	0.980	0.013	0.653	0.987	0.797	0.679	0.662	0.819	0.180	0.816
		virginica	0.520	0.507	0.963	0.493	0.675	0.595	0.510	0.819	0.180	0.816
		weight average	0.820	0.182	0.872	0.818	0.817	0.758	0.711	0.819	0.180	0.816
	KNN	setosa	1.000	0.000	1.000	1.000	1.000	1.000	0.940	1.000	0.071	0.199
		color	0.960	0.070	0.873	0.930	0.914	0.871	0.940	0.933	0.071	0.199
		virginica	0.860	0.020	0.956	0.980	0.905	0.864	0.940	0.933	0.071	0.199
		weight average	0.940	0.030	0.943	0.970	0.940	0.912	0.940	0.955	0.071	0.199
	NB	setosa	1.000	0.000	1.000	1.000	1.000	1.000	0.946	1.000	0.057	0.183
		color	0.960	0.060	0.889	0.940	0.923	0.884	0.946	0.954	0.057	0.183
		virginica	0.880	0.020	0.957	0.980	0.917	0.879	0.946	0.954	0.057	0.183
		weight average	0.947	0.027	0.948	0.973	0.947	0.921	0.946	0.969	0.057	0.183
	DT	setosa	1.000	0.000	1.000	1.000	1.000	1.000	0.946	1.000	0.064	0.181
		color	0.960	0.060	0.889	0.940	0.923	0.884	0.946	0.941	0.064	0.181
		virginica	0.880	0.020	0.957	0.980	0.917	0.879	0.946	0.941	0.064	0.181
		weight average	0.947	0.027	0.948	0.973	0.947	0.921	0.946	0.961	0.064	0.181
	SVM	setosa	1.000	0.000	1.000	1.000	1.000	1.000	0.946	1.000	0.234	0.293

Case	Classifier	Class	TP Rate	FP Rate	Precision	Specificity	F1 score	MCC	Accuracy	AUC	MAE	RMSE
		color	0.960	0.060	0.889	0.940	0.923	0.884	0.946	0.950	0.234	0.293
		virginica	0.880	0.020	0.957	0.980	0.917	0.879	0.946	0.960	0.234	0.293
		weight average	0.947	0.027	0.948	0.973	0.947	0.921	0.946	0.970	0.234	0.293

5.Discussion

5.1Case1: breast cancer

None of the applied techniques in breast cancer cases can predict both the breast cancer patient's diagnosis and healthy control patient correctly. J48 and KNN had the same accuracy with 0.612 as both can predict 71 classes correctly. But, other than accuracy, all results were marked at a different point. Through the comparison, J48 holds the best results in all performance evaluations, followed by KNN, NB, and SVM. Meanwhile, AD classification only can predict 50 instant data correctly. Especially in MCC results, AD classification had unsatisfying results as they hit the negative number. There are 116 patients with 62 different types of breast cancer symptoms stated in the data set. So, the number of the patient against symptoms turn out to be imbalanced. Perhaps, the imbalance data in the breast cancer data set has influenced AD classification decision.

5.2Case2: acute nephritic syndrome

Next, the data set for the acute nephritic syndrome was about predicting nephritis of renal pelvis origin. All five techniques show excellent results in prediction as all classifiers predict the presence of nephritis of renal pelvis origin correctly, with 70 for the "no" class and 50 for the "yes" class. All classifiers have been hit the best rate in all categories of sensitivity, false-positive rate, specificity, precision, F1 score, MCC, and accuracy. Only KNN and NB had little different results in the last two MAE and MRSE, which the results, not an actual 0. Contradicting from the cancer data set, the number in acute nephritic syndrome data set has 120 patients with nine various symptoms. The number of patients against each symptom was more balanced, with 10 to 20 patients per symptom, which may be why AD classification can predict correctly.

5.3Case3: iris plant

Lastly, it has multi-class matrices for the iris plant data set as it predicts three types of iris plants. Like the first data set, all classifiers cannot predict the whole 115 instances correctly. Nevertheless, except for AD classification, the other classifier can classify Iris Setosa perfectly with 50 instances. Also, NB, J48, and SVM had the highest TP rate with 0.947, followed by KNN with 0.940 and AD classification with 0.820. For the other category, besides the KNN and AD classifier, all three classifiers hit the best rate but not the performance in terms of a graph. All classifiers hit the different mark in AUC, MAE, and RMSE. The results might influence the multi-class matrices for the predictive component used in this study, leading to such performances.

5.4Comparison

In this study, the weighted average was used to make the comparison between all five classifiers. By implemented KNN, NB, J48, SVM, and AD classification techniques in three separate cases with various conditions, this study can conclude that J48 gives the best results compared to the other four classifiers. J48 hit the best rate for all performance measures in all three cases. NB classifier gives the best performance in both breast cancer and iris plant data set. The same goes for the SVM classifier, which shows the best results in two data sets acute diagnosis and iris plant. KNN shows excellent performance with a stable average hit the mark in breast cancer and acute diagnosis data sets but a bit low in range compared to NB, J48, and SVM for iris data set. Meanwhile, AD classification shows an unsatisfying result when one of the performance measures MCC hit -0.1 and an average mark of 0.43 for the other components in a breast cancer case. However, AD had its best rate with 1.0 in acute nephritic syndrome cases and a stable rate with an average of 0.7 to 0.8 for iris plant cases in all the performance measures.

5.5Limitation

There is some limitation in this study as affinity degree was a new adapted algorithm, there are no tools to calculate the degree or classify the class automatically. Therefore, the work consumes time, mostly when the data was massive.

6.Conclusion

This study implemented KNN, NB, J48, SVM, and AD classification techniques on the three different UCI data set. Based on the results, J48 shows the best performance compares to the other four classifiers.

This study aimed to see the compatibility of affinity degree as a classifier technique. Nevertheless, various issues might have influenced these classification techniques, including the data set conditions and the aim of methods used. For instance, the affinity degree classification defines the correlation by calculating the affinity degree before classify accordingly based on the indicator. Here, as AD had different significance in contributing to predictive class, more future AD needs to be done—especially the affinity degree indicator, as there is room for improvement. Also, AD on the different data set compared with various classification methods needs to be done to define more significant AD classification values. Make a specific tool for calculating the degree of affinity also can be considered for next future work.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] Li Z, Kim J, Regnier FE. Mobile affinity sorbent chromatography. Analytical Chemistry. 2018; 90(3):1668-76.

[2] Asseraf Y, Shoham A. The "tug of war" model of foreign product purchases. European Journal of Marketing. 2016; 5(3-4):550-74.

[3] Bakhouya M, Gaber J. Approaches for engineering adaptive systems in ubiquitous and pervasive environments. Journal of Reliable Intelligent Environments. 2015; 1(2):75-86.

[4] Chen YW, Larbani M, Hsieh CY, Chen CW. Introduction of affinity set and its application in data-mining example of delayed diagnosis. Expert Systems with Applications. 2009; 36(8):10883-9.

[5] Awang WS, Deris MM, Rana OF, Zarina M, Rose AN. Affinity replica selection in distributed systems. In international conference on parallel computing technologies 2019 (pp. 385-99). Springer, Cham.

[6] Bost R, Popa RA, Tu S, Goldwasser S. Machine learning classification over encrypted data. In NDSS 2015 (pp. 1-14).

[7] Cover T, Hart P. Nearest neighbor pattern classification. IEEE Transactions on Information Theory. 1967; 13(1):21-7.

[8] Sonawane JM, Gaikwad SD, Prakash G. Microarray data classification using dual tree m-band wavelet features. International Journal of Advances in Signal and Image Sciences. 2017; 3(1):19-24.

[9] Prasatha VS, Alfeilate HA, Hassanate AB, Lasassmehe O, Tarawnehf AS, Alhasanatg MB, et al. Effects of distance measure choice on KNN classifier performance-a review. arXiv preprint arXiv:1708.04321. 2017.

[10] Nikam SS. A comparative study of classification techniques in data mining algorithms. Oriental Journal of Computer Science & Technology. 2015; 8(1):13-9.

[11] Pelillo M. Alhazen and the nearest neighbor rule. Pattern Recognition Letters. 2014; 38:34-7.

[12] Hand DJ, Yu K. Idiot's Bayes—not so stupid after all? International Statistical Review. 2001; 69(3):385-98.

[13] Patel HH, Prajapati P. Study and analysis of decision tree based classification algorithms. International Journal of Computer Sciences and Engineering. 2018; 6(10):74-8.

[14] Durgesh KS, Lekha B. Data classification using support vector machine. Journal of theoretical and applied information technology. 2010; 12(1):1-7.

[15] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra#. Accessed 15 February 2020.

[16] https://archive.ics.uci.edu/ml/datasets/Acute+Inflammations. Accessed 06 December 2020.

[17] http://archive.ics.uci.edu/ml/datasets/Iris/. Accessed 06 December 2020.

[18] Halim RE, Zulkarnain EA. The effect of consumer affinity and country image toward willingness to buy. The Journal of Distribution Science. 2017; 15(4):15-23.

[19] Dancey CP, Reidy J. Statistics without maths for psychology. Pearson Education; 2007.

[20] Assegie TA. An optimized K-Nearest neighbor based breast cancer detection. Journal of Robotics and Control. 2021; 2(3):115-8.

[21] Tien Bui D, Pradhan B, Lofman O, Revhaug I. Landslide susceptibility assessment in Vietnam using support vector machines, decision tree, and Naive Bayes Models. Mathematical problems in Engineering. 2012; 2(3):115-8.

[22] Pharswan R, Singh J. Performance analysis of SVM and KNN in breast cancer classification: a survey. In internet of things and big data applications 2020 (pp. 133-40). Springer, Cham.

[23] Thirunavukkarasu K, Singh AS, Rai P, Gupta S. Classification of IRIS dataset using classification based KNN algorithm in supervised learning. In international conference on computing communication and automation 2018 (pp. 1-4). IEEE.

[24] Mahdikhani L, Keyvanpour MR. Challenges of data mining classification techniques in mammograms. In 5th conference on knowledge based engineering and innovation (KBEI) (pp. 637-43). IEEE.

[25] Saritas MM. Yasar A. Performance analysis of ANN and naive bayes classification algorithm for data classification. International Journal of Intelligent Systems and Applications in Engineering. 2019; 7(2):88-91.

[26] Hamoud A. Hashim AS. Awadh WA. Predicting student performance in higher education institutions using decision tree analysis. International Journal of Interactive Multimedia and Artificial Intelligence. 2018; 5(2):26-31.

[27] https://www.wcrf.org/dietandcancer/cancer-trends/breast-cancer-statistics. Accessed 15 April 2020.

[28] Kuchenbaecker KB, Hopper JL, Barnes DR, Phillips KA, Mooij TM, Roos-Blom MJ, et al. Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers. Jama. 2017; 317(23):2402-16.

[29] Majoor BC, Boyce AM, Bovée JV, Smit VT, Collins MT, Cleton-Jansen AM, et al. Increased risk of breast cancer at a young age in women with fibrous dysplasia. Journal of Bone and Mineral Research. 2018; 33(1):84-90.

[30] Brinton LA, Brogan DR, Coates RJ, Swanson CA, Potischman N, Stanford JL. Breast cancer risk among women under 55 years of age by joint effects of usage of oral contraceptives and hormone replacement therapy. Menopause. 2018; 25(11):1195-200.

[31] https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/. Accessed 15 April 2020.

[32] https://medlineplus.gov/ency/article/000495.htm#:~:text=Acute%20nephritic%20syndrome%20is%20a,in%20the%20kidney%2C%20or%20glomerulonephritis. Accessed 25 January 2021.

[33] https://www.healthline.com/health/acute-nephritic-syndrome. Accessed 25 January 2021.

[34] Ruuska S, Hämäläinen W, Kajava S, Mughal M, Matilainen P, Mononen J. Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle. Behavioural Processes. 2018; 148:56-62.

[35] Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020; 21(1):1-3.

Rosyazwani Mohd Rosdan received her Bachelor’s Degree in Science Computer (Software Development) from Universiti Sultan Zainal Abidin (UniSZA). Currently a Master Degree student at the same university. Research focusing on adapting affinity degree from Peer-to-Peer networks into Machine Learning.
Email: wanirose2@gmail.com

Wan Suryani Wan Awang, obtained her PhD degree from Cardiff University, the United Kingdom in Computer Science. She is currently a Senior Lecturer in the Faculty of Informatics & Computing, UniSZA. Her main research interests are Parallel and Distributed Systems, Peer-To-Peer Networks, Cloud Computing, Big Data, Data Analysis and Data Mining, nd Machine Learning.
Email: suryani@unisza.edu.my

Wan Aezwani Bt Wan Abu Bakar received her PhD in Computer Science at Universiti Malaysia Terengganu (UMT) Terengganu in Nov 2016. She received her master's degree in Master of Science (Computer Science) from Universiti Teknologi Malaysia (UTM) Skudai, Johor in 2000 before finishing her study in Bachelor's degree also in the same stream from Universiti Putra Malaysia (UPM) Serdang, Selangor in 1998. Her master's research was on Fingerprint Image Segmentation in the stream of Image Processing. She was currently focusing her research towards association relationship infrequent/infrequent itemset mining, IoT and Big Data Analytics.
Email: wanaezwani@unisza.edu.my