ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2020.0290



# PEARSON CORRELATION COEFFICIENT K-NEAREST NEIGHBOR OUTLIER CLASSIFICATION ON REAL-TIME DATASETS

**D. Rajakumari**
*Department of Computer Science, Nandha Arts and Science College, India*

***Abstract***

***Detection and classification of data that do not meet the expected behavior (outliers) plays the major role in wide variety of applications such as military surveillance, intrusion detection in cyber security, fraud detection in online transactions. Nowadays, an accurate detection of outliers with high dimension is the major issue. The trade-off between the high-accuracy and low computational time is the major requirement in outlier prediction and classification. The presence of large size diverse features need the reduction mechanism prior to classification approach. To achieve this, the Distance-based Outlier Classification (DOC) is proposed in this paper. The proposed work utilizes the Pearson Correlation Coefficient (PCC) to measure the correlation between the data instances. The minimum instance learning through PCC estimation reduces the dimensionality. The proposed work is split up into two phases namely training and testing. During the training process, the labeling of most frequent samples isolates them from the infrequent reduce the data size effectively. The testing phase employs the k-Nearest Neighborhood (k-NN) scheme to classify the frequent samples effectively. The dimensionality and the k-value are inversely proportional to each other. In proposed work, the selection of large value of k offers the significant reduction in dimensionality. The combination of PCC-based instance learning and the high value of k reduces the dimensionality and noise respectively. The comparative analysis between the proposed PCC-k-NN with the conventional algorithms such as Decision Tree, Naïve Bayes, Instance-Based K-means (IBK), Triangular Boundary-based Classification (TBC) regarding sensitivity, specificity, accuracy, precision, and recall proves its effectiveness in OC. Besides, the experimental validation of proposed PCC-k-NN with the state-of art methods regarding the execution time assures trade-off between the low-time consumption and high-accuracy.***

***Keywords:***

***Data Mining, Distance-based Instance Learning, Outlier Detection, Outlier Classification, Pearson Correlation Coefficient, k-Nearest Neighbor***

## 1. INTRODUCTION

Hidden knowledge discovery from the large size spatiotemporal database includes the following processes: 1) periodic/frequent patterns detection, 2) Outlier Detection (OD) and 3) Outlier Classification (OC) [1]. Specifically, OD and OC play the major roles in health care applications, fault detection in industrial systems, fraud detection and abnormal region prediction in image processing, etc. The differentiation or isolation of outliers from the normal is the necessary stage in OD process and serves as the base for domain description and classification techniques [2]. With increasing the size of data, the differentiation among near and far end neighbors is poor and hence the proper reduction mechanism is required. The Fig.1 shows the major categories of OD and OC approaches. Among them, Distance-based OD and labeling-based OC are the most prominent research studies due to their dimensionality reduction capability [3]. The analysis of high dimensionality dataset consumes more time during the knowledge extraction phase [4]. The distance-based one class classifier called Prototype-based Domain Description (PDD) [5] that exploits the subset from the overall training set and selects the objects with distance basis.

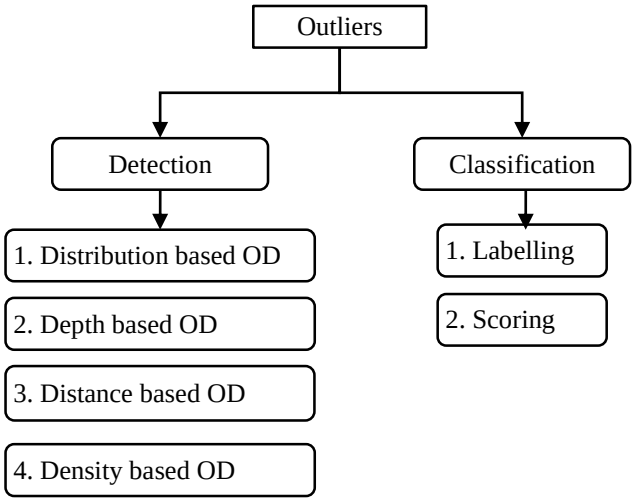


Fig.1. Outlier detection and classification approaches

High dimensionality is directly affects the detection quality of traditional OD techniques [6]. To improve the detection quality and reduce the data dimension of OD techniques, research studies introduce the isolation of data samples based on the distance measure called Mahalanobis Distance (MD) covariance matrix [7]-[11]. The absence of robust mean and variances from the MD covariance matrix causes the several problems in applications. Less overlap among the patterns is the major requirement for the formulation of the multiple MD measures into the new metric [12] [13].

Pearson Correlation Coefficient (PCC) is the mathematical measure to estimate the correlation or overlap among the patterns by the line slope. Several PCC measures such as absolute, non-centered and absolute non-centered are highlighted in research works to reduce the operational complexities in the classification performance. The PCC also governs the quantitative comparison of traditional and recent OC techniques. The employment of correlation measures efficiently reduce the dimensionality and thus it leads to minimum time consumption compared to MD approaches [14-17]. The performance of PCC-based OC depends on the frequency of label mismatches and distance measures.

The arrival of noise from the large size database affects OC in complex real-time applications. To alleviate these problems, *k*-nearest neighbor (*k*-NN) [18-22] identifies the pseudo nearest neighbors. The assigning of correctly classified objects to Meta or rejection class introduces the prediction errors that leads to deterioration quality of classification, excessive memory, resources and storage. The concept-evolution theory addresses the problems in the recurring class inclusion in the stream-based data

mining approaches [23]-[25]. Hence, the introduction section conveys that the high-data dimension and more computational steps reduce the classification performance. This paper performs the data reduction prior to *k*-NN application by using the PCC-correlation-based instance learning that reduces the computational steps that reduce the time consumption. The contributions of PCC-*k*-NN method are listed as follows:

- The correlation estimation through the PCC-based instance learning reduces the data dimensionality.
- The label assigning to the frequent/infrequent samples prior to *k*-NN classifier (with high value of *k*) reduces the time consumption and enhances the applicability to real-time datasets.
- The combination of PCC-*k*-NN instance learning and classification offer the better classification performance compared to the traditional classifiers.
- The reduction in a number of features provides the trade-off between the low-time and more accuracy that assures the effectiveness in Knowledge Discovery (KD).

This paper is organized as follows: section 2 describes the several research studies regarding the outlier detection and classification and their impact on the knowledge discovery. Section 3 illustrates the detailed description of the implementation of proposed Pearson Correlation Coefficient (PCC)-*k*-Nearest Neighbor (*k*-NN). Section 4 discusses the effect of PCC-*k*-NN on the performance measures and section 5 concludes the proposed implementation.

## 2. RELATED WORK

An appropriate pattern discovery and anomalous removal include the various Outlier Detection (OD) and Outlier Classification (OC) techniques. Sagade and Thakur summarized the techniques involved in OD and OC [1]. With an effective OD/OC techniques, the error identification and abnormal effect prediction in the data set are necessary for purification. Al-Khateeb et al. [2] addressed the problems in recurring class stream-based data mining. They proposed ensemble techniques to isolate the recurring classes from the novel. Higher data collection from the telecommunication devices turned the research studies into the knowledge discovery from the spatiotemporal data. Albanese et al. [3] proposed rough outlier set extraction of data on approximations basis. The presence of categorical and numerical values in the hybrid dataset initiate the new approach called fuzzy rough set approach. Qian et al. [4] proposed the accelerator that reduced the size of frequent samples and data together. But, the utilization of fuzzy rules consumed more steps and that leads to excessive time consumption. Angiulli [5] discussed the prototype-based domain description rule-based one-class classifier that performed the statistical tests necessary for generalization for OD. This scheme increased the computational burden and reduced the operational speed.

The reduction of outlier dimension to increase the processing speed depends on the variations in object angles. Pham and Pagh [6] investigated the angle-based outlier factor model to speed up the OD process. The complexity and the inaccurate detection of outliers were the major problems in classification process. An accurate outlier detection with the geometrical properties depends on the integration of hitting time with spectral clustering. The non-convex shaped isolation with the geometrical features of data was the necessary stage in clustering-based classifiers. Galluccio et al. [7] introduced a hitting time based distance measure for the construction of minimum spanning tree. The operation of weight assigning classifiers on the feature space in tree construction process initiates the multi-class classifiers deployment. Krawczyk et al. [8] presented the weighted classifier One-Class Support Vector Machine (OCSVM) that provided the solution to both single and multi-class classification problems. The combination of decomposition (in input stage) and the composition (in output stage) in OCSVM offered the stable recognition performance. The prediction of outliers on arbitrarily oriented spaces is an attractive research studies nowadays. Kriegel et al. [9] considered the attribute subset with local correlations to detect and classify the outliers. The explanation of level of outliers depends on the quantitative metric called outlier score in attribute-based approaches. Prediction of relevant features to increase the outlier score requires the spectral reflectance analysis or correlation estimation. The correlation estimation process employs the several distance measures such as Manhattan Distance (MD), Pearson Correlation Coefficient (PCC). Saeed et al. [10] discussed the palm oil quality prediction with the active optical sensor system on Fresh Fruit Bunch (FFB). They provided the quantitative discriminant analysis with the Mahalanobis Distance (MD) classifiers for high accuracy.

Zimek et al. [11] discussed the ensemble sub-sampling techniques for diversity introduction in OD. They provided the analytical judgment to the inclusion of ensemble techniques in distance-based outlier classifiers. Two metrics such as remoteness and isolation degree were derived from the resultant pairwise distance matrix to show the high reflectance of MD in real-time applications. Krishnan and Kerkhoff [12] discussed the qualitative and quantitative formulation of outlier screening and non-defective reliability respectively. The MD measure centered on data centroid provided the less isolation between the local and global samples. Todeschini et al. [13] derived the MD centered on each sample that offered the better isolation between the samples and features. Akila and Chandra [14] discussed the dynamic time wrapping models for spectral distance measurement. Through PCC measure, the global distance is minimized to find the slope of dynamic time wrapping models. Zeng et al. [15] extended the PCC measurement into the reconstruction of original images from moderate resolution imaging spectro-radiometer in land surface identification applications.

Xu et al. investigated the One-Class Partial Least squares (OCPLS) [16] with the correlation metric called PCC. They provided the improved OCPLS for untargeted detection models. Zawadzki et al. [17] proposed the density estimation models for bivariate dataset kernels including the probability distribution. The smallest probability cells regarded as outliers. The existence of noise, redundancy, irrelevant resources made frequent pattern extraction as the challenging one. Hund et al. [18] framed the new research method called subspace nearest neighbor search that supported the multiple queries dependent subspaces. Quality improvement and the anomalies removal were the challenging tasks in neighbor search-based OD. Liu and Zhang [19] proposed a learning algorithm called *k*-NN that predicted the unseen instance class labels called mutual neighbors rather than the nearest neighbors. The *k*-NN based unseen labels prediction

suffers from some difficulties during classification. Liu et al. [20] developed a probabilistic framework to address these difficulties on accurate classification and they termed as classification errors. The initialization of proper threshold values for either acceptance or rejections reduced the classification errors effectively.

The higher dimensionality made the similarity measurement as the difficult task in OC process. Tomas and Mladenic [21] proposed a hubness similarity measure under *k*-NN graph that effectively reduced the frequency of label mismatches. Zhong et al. [22] investigated the authentication problem by keystroke biometrics to meet the challenges of scale variations, feature interactions, and an outlier detection. Ahmed and Elaraby [22] predicted the final grade of students from the large size database by the Decision Tree (DT) algorithms. The representation of DT using logical rules efficiently supported the prediction of final grades of the student. An accurate prediction of student characteristics and their academic performance with less no. of features in Educational Data Mining (EDM) was the difficult task. Harb and Mousafa [24] applied feature selection algorithms that removed the irrelevant, noisy data and reduced the size of features. An automatic nearest neighbor prediction by the Instance-based *k*-NN improved the classification performance. Lam et al. [25] used the Naïve Bayesian classification based on Bayes theorem for meaningful OD in traffic management. They presented Gaussian Mixture Model (GMM) that exploited the confidence region in traffic datasets. The complexity, size and variants of dataset made the conventional techniques were unsuitable in OD and OC.

Bi-level and the multi-dimensionality methods were evolved in the research studies to improve the OD performance with the less dimension. Rajakumari and Pannirselvam [26] proposed the novel Triangular Boundary-based Classification (TBC) that contained the training and testing algorithms to overcome the challenges in conventional techniques. Rangarajan and Veerabhadrappa [27][28] discussed the multi-dimensionality reduction methods to improve the quality of OD. Sridevi et al. [29] discussed the modified correlation rough set feature selection that selected the features based on rough set in the initial level and from the reduced subset in second level. The incorrect classification instance rate was substantially reduced in traditional approaches. The research studies conveyed the data dimension reduction is the prerequisite step in OC.

# 3. PEARSON CORRELATION COEFFICIENT-K-NEAREST NEIGHBOR OUTLIER CLASSIFICATION

This section describes the implementation of proposed hybrid Pearson Correlation Coefficient (PCC) and *K*-Nearest Neighborhood (*k*-NN) distance-based outlier classification. The proposed work includes the two phases as follows: training and testing. The class label-based data instance splitting and nearest neighbor analysis in both training and testing validate the OC performance on Wisconsin Diagnostic Breast Cancer (WDBC) dataset. The utilization of PCC identifies the correlation between the data instances effectively. Moreover, the instance-based data splitting the data size and reduces the multi-dimensional vectors space. The flow of proposed PCC-*k*-NN as in Fig.2.

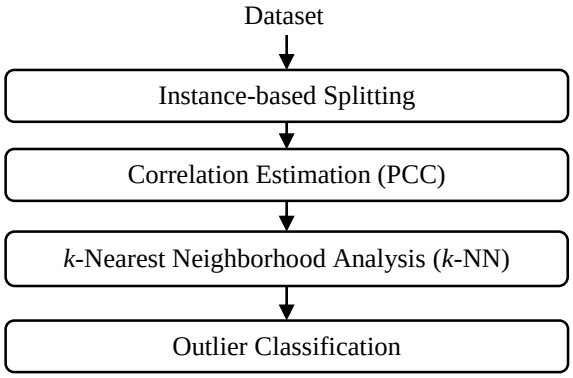


Fig.2. Overall flow of proposed PCC-*k*-NN

## 3.1 CORRELATION ESTIMATION

Discriminant analysis techniques (Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA)) provided the less guarantee if the prior assumptions were violated. Hence, the PCC estimation offers the solution to these problems. The numerical correlation coefficient value depicts the linear equation that describes the relationship between the two variables such as *X* and *Y*. The positive value (+1) states that the data points lying on the line for which *Y* increases as *X* increases. The negative value (-1) describes the data points lying on the line for which *Y* decreases as *X* increases. The lying of $X_i$ and $Y_i$ on the same line turns the $(X_i\text{-}X)(Y_i\text{-}Y)$ as positive. Hence, the correlation coefficient is positive for simultaneous greater and lesser means. The PCC measure is defined as the ration of the covariance of two variables to the product of standard deviations. The correlation estimation is defined as:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_x \sigma_y} = \frac{E\left[(X-\mu_X)(Y-\mu_Y)\right]}{\sigma_x \sigma_y} \quad (1)$$

Based on the correlation estimation, the following steps are used to reduce the dimension.

**Step 1:** Extract the number of attributes in dataset

**Step 2:** Assign the row number to the attributes

**Step 3:** Convert the database into the matrix form $L(i,j)$

**Step 4:** Assign the threshold value to reduce the dimension

**Step 5:** Check the element of row is less than the threshold value

**Step 6:** Construct the independent attribute matrix with the elements less than the threshold

**Step 7:** Compute the mean and standard deviation

**Step 8:** Calculate the correlation coefficient (R) based on the computed mean values.

$$R = \frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2}} \quad (2)$$

Based on the steps involved in correlation estimation process, the proposed work includes two processes namely training and testing. The training dataset for the class label *c* is defined as,

$$X^c = \begin{bmatrix} x^1 \\ x^2 \\ \vdots \\ x^n \end{bmatrix} \quad (3)$$

where the single $i^{th}$ instance $x^i$ represents the $m$ number of features is expressed as:

$$x^i = \left[ f_1^i, f_2^i, f_3^i \ldots f_m^i \right] \quad (4)$$

Consider the two random instances ($I_1$,$I_2$) related to the features defined in Eq.(3). The mean and standard deviations of the instances are $\left(\bar{I}_i, \bar{I}_{i+1}\right)$ and $\left(\sigma_{I_i}, \sigma_{I_{i+1}}\right)$ respectively. Then, the correlation estimation is redefined as follows:

$$R = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{I_i - \bar{I}_i}{\sigma_{I_i}}\right)\left(\frac{I_{i+1} - \bar{I}_{i+1}}{\sigma_{I_{i+1}}}\right) \quad (5)$$

$$d_p = 1\text{-}R \quad (6)$$

The Fig.3(a), Fig.3(b) and Fig.3(c) shows the dimensionality reduction by using the PCC measurement.

| Correlation Coefficient Estimation | | | | |
|---|---|---|---|---|
| *R* | 1 | 2 | … | *m* |
| 1 | - | $R_{12}$ | … | $R_{1m}$ |
| 2 | - | - | … | $R_{2m}$ |
| 3 | - | - | … | $R_{3m}$ |
| … | … | … | … | … |
| *m* | - | - | - | - |

(a)

| Correlation Coefficient Estimation | | | | |
|---|---|---|---|---|
| *R* | 1 | 2 | … | *m* |
| 1 | - | $R_{12}$ | … | $R_{1m}$ |
| 2 | - | - | … | $R_{2m}$ |
| 3 | - | - | … | $R_{3m}$ |
| … | … | … | … | … |
| *m* | - | - | - | - |

(b)

| Correlation Coefficient Estimation | | | | |
|---|---|---|---|---|
| *R* | 1 | 2 | … | *m* |
| 1 | - | $R_{12}$ | … | $R_{1m}$ |
| 2 | - | - | … | $R_{2m}$ |
| 3 | - | - | … | $R_{3m}$ |
| … | … | … | … | … |
| *m* | - | - | - | - |

(c)

Fig.3 (a) Correlation estimation, (b) Selection of independent attribute with $R$ < threshold (Green shaded cells) and (c) Elimination of attributes with $R$ > threshold (Brown shaded cells)

The correlation indicator predicts the relationship between the instances by the $R$ result. If $R$ is 0 or 1, then the indicator $d_p$ denotes that two vectors are correlated each other. Generally, this PCC measurement is used for grouping the instance. In our proposed approach, the PCC-based correlation measurement classifies the outlier. The major impact of the PCC estimation is to reduce the dimensionality on the basis of instance learning. The attributes in the green shaded cells are selected and brown shaded cells are removed. Thereby, the dimensionality is reduced. The upper and lower boundary layer for each class present in the training dataset are formulated by using class label-based instance splitting. These values helps to predict the accurate class for each instance. Overall mean ($\mu$) and deviation ($\sigma$) are computed by using the equations given below

$$\mu = \frac{1}{m}\sum_{i=1}^{m} R^{c,i} \quad (7)$$

$$\sigma \leftarrow \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}\left(R^{c,i} \mu\right)^2} \quad (8)$$

$$\text{Layer} \leftarrow \mu + \sigma * \alpha \quad (9)$$

where, $m$ is the maximum dimension and $R$ is the correlation coefficient

The dimensionality reduction prior to the $k$-NN algorithm efficiently avoids the curse of dimensionality (CD). The CD describes that the equidistance between the search query and the responses makes the Euclidean distance is not suitable. Hence, correlation coefficient based distance formulations are used to reduce the dimension.

## 3.2 K-NEAREST NEIGHBOR

In training algorithm, the dataset is initially split up into subsets with class label instance basis. The input features related to the instances in multi-dimensional space are initialized as training examples. The training phase includes the feature vectors storage with the class labels. The mean $\mu$ and $\sigma$ for the each data instance is calculated by using Eq.(7) and Eq.(8). To find the correlation between the instances, there is a need to compute the mean and deviation of the distance of all instances containing $c$ class labels using Eq.(5) respectively. The distribution of the distance is described by mean and deviation of the $c^{th}$ class. In the classification phase, the label assignment with the user defined constant ($k$) classifies the unlabeled vectors from the multi-dimensional space. The distance between the selected instances with the $k^{th}$ NN is considered as the local density estimate and outlier score in anomaly detection. If the distance to the $k$-NN is large, then the local density measure is low and the corresponding instance is regarded as an outlier.

**Training Algorithm**

**Input**: Training Data $D$

**Output**: $\mu$, $\sigma$, $PCC_r$, $d_p$ Classifier $C$

**Step 1:** Identify the Class Label

$M \leftarrow$ class label 1, $B \leftarrow$ class label 2 //Partition the dataset based on class attribute

**Step 2:** Estimate the coefficient

$$R = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{I_i - \bar{I}_i}{\sigma_{I_i}}\right)\left(\frac{I_{i+1} - \bar{I}_{i+1}}{\sigma_{I_{i+1}}}\right)$$

**Step 3:** For each instance $i \in D$

$$\mu = \frac{1}{m}\sum_{i=1}^{m} R^{c,i}$$

End For

**Step 4:** Apply *K*-Nearest Neighbor Classification

For each instance $i \in D$

Find the neighbor value based on correlation value

Build the Classifier

$C \leftarrow$ **IF** 0.1 < *n* < 0.5 **Then** class 1

$C \leftarrow$ **IF** 0.5 < *n* < 1.0 **Then** class 2

Return *C*;

End For

Depending on the correlation metric, the *k*-NN classifiers classifies the class 1 and 2 if the value is in between 0.1 and 0.5 and 0.5 between 1 by using the training algorithm.

**Testing Algorithm**

**Input**: Testing Data *D*.

**Output**: Outlier *T*

**Step 1:** Identify the class label

$M \leftarrow$ class label 1, $B \leftarrow$ class label 2 //Partition the dataset based on the class attribute

**Step 2:** Calculate Pearson Correlation Coefficient (PCC) between each instance

$$R = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{I_i - \bar{I}_i}{\sigma_{I_i}}\right)\left(\frac{I_{i+1} - \bar{I}_{i+1}}{\sigma_{I_{i+1}}}\right)$$

**Step 3:** Calculate the Mean value for correlation value of each instance

For each instance $i \in D$

$$\mu = \frac{1}{m}\sum_{i=1}^{m} R^{c,i}$$

End For

**Step 4:** For each instance $i \in D$

Find the neighbor value based on correlation value

Classify the Instance based on classifier build in Training

$C_1 < \{I_1, I_2, I_4\}$, $C_2 \rightarrow \{I_3, I_5, I_7\}$, $C_3 \leftarrow \{I_6, I_8, I_9\}$

//C1, C2, and C3 are the class labels

End For

**Step 5:** Regression Analysis using Root Mean Square Error

Calculate RMSD or RMSE Value

For each class $C_i \in D$

$RMSD(C_i) \leftarrow$ calculate_$RMSD(C_i)$;

End For

**Step 6:** Find outlier based on RMSD Value

For each class $C_i \in D$

$T \leftarrow \text{Max}(RMSD(C_i))$

End For

**Step 7:** Return *T*

In the testing phase, the Root Mean Standard Deviation (RMSD) or Root Mean Square Error (RMSE) utilization on testing datasets compute the outlier and identifies the maximum value for each class. A measure of difference between the estimated samples with the actual samples refer RMSD. The measured individual differences are regarded as prediction errors and it serve as the good metric for data aggregation. The RMSD estimation of class labels (*C*) is expressed as

$$RMSD = \sqrt{MSE(C)} = \sqrt{E\left(\left(\hat{C} - C\right)^2\right)} \quad (10)$$

The maximum RMSD value from the number of computations by Eq.(10) can be regarded as the outlier (*T*). The PCC, *k*-NN and the RMSD approaches efficiently detects the outlier with the less dimensionality.

## 4. PERFORMANCE ANALYSIS

This section discusses the performance of proposed hybrid Pearson Correlation Coefficient and *k*-Nearest Neighbor (*k*-NN) distance measurement on the data instances regarding accuracy, precision, recall, sensitivity, specificity, incorrect or correct classified instances. The comparative analysis of proposed PCC-*k*-NN with the following traditional methods: Decision Tree (DT), Naive Bayes (NB) and *k*-means Instance Based for *K*-Nearest neighbor (IBK), and Triangular Boundary-based Classification (TBC) [24] [26] show that prior data reduction improves the classification performance in Wisconsin Diagnostic Breast Cancer (WDBC) dataset. The execution time analysis of proposed PCC-*k*-NN also validated with the three additional datasets namely, Sonar, mines vs. rocks, ionosphere, wine recognition for fuzzy rough set approaches (Fuzzy-Positive Region (FPR), Forward Approximate-based Fuzzy Positive Region (FA-FPR), Fuzzy-features Selection with Conditional Entropy and FP-FCSE [4]. The Table.1 shows the dataset details to validate the performance of proposed and existing methods. The WEKA tool utilization validates the testing performance of existing methods. The 75% of features in the dataset are used for training and the remaining 25% used for testing processes.

Table.1. Dataset Description

| Datasets | Samples | Features | Classes |
|---|---|---|---|
| Sonar, Mines vs. Rocks | 208 | 60 | 2 |
| WDBC | 569 | 30 | 2 |
| Ionosphere | 351 | 34 | 2 |
| Wine Recognition | 178 | 13 | 3 |
| WBC | 683 | 9 | 2 |
| Corn Soya bean | 61 | 24 | 2 |

The Table.2 shows the comparative analysis of PCC-*k*-NN with the existing methods regarding TP, TN, FP, and FN. Among the existing methods, the predictions (TP, TN, FP, and FN) of TBC are 208, 347, 4, and 10 respectively. But, the utilization of Pearson Correlation Coefficient in the nearest neighbor analysis in proposed PCC-*k*-NN offers the predictions as 243, 315, 3, and

8 which is 16.82% improvement in outlier classification (TP) performance for WDBC dataset.

Table.2. Comparative analysis of TP, TN, FP and FN of the proposed approach and existing algorithms (WDBC)

| Algorithms | TP | TN | FP | FN |
|---|---|---|---|---|
| Decision tree-RF | 198 | 343 | 14 | 14 |
| Bayes-NB | 190 | 339 | 22 | 18 |
| *K*-Means IBK | 200 | 346 | 12 | 11 |
| TBC | 208 | 347 | 4 | 10 |
| PCC-*k*-NN | 243 | 315 | 3 | 8 |

The Table.3-Table.5 presents the comparative analysis of proposed PCC-*k*-NN with the multi-dimensionality reduction methods [27] regarding the average precision, recall and F-measure values for WDBC, WBC, WINE datasets.

Table.3. Comparison for WDBC dataset

| Methods | Average Precision | Average Recall | Average F-measure |
|---|---|---|---|
| PCA | 87.947 | 75.663 | 81.344 |
| Locality Preserving Projection (LPP) | 84.575 | 86.747 | 85.627 |
| Mutual Correlation (MC) | 89.523 | 78.538 | 83.638 |
| FQ Measure | 88.738 | 87.721 | 88.300 |
| MC+FQ+PCA | 93.339 | 89.483 | 91.370 |
| MC+FQ+LPP | 93.452 | 90.102 | 91.746 |
| FQ+MC+PCA | 91.614 | 91.796 | 91.705 |
| FQ+MC+LPP | 93.963 | 90.758 | 92.333 |
| PCC-*k*-NN | 98.125 | 98.525 | 98.625 |

Table.4. Comparison for WBC dataset

| Methods | Average Precision | Average Recall | Average F-measure |
|---|---|---|---|
| PCA | 89.454 | 78.114 | 83.400 |
| LPP | 88.782 | 82.121 | 85.322 |
| Mutual Correlation (MC) | 91.484 | 83.135 | 87.110 |
| FQ Measure | 92.564 | 85.452 | 88.866 |
| MC+FQ+PCA | 92.745 | 85.871 | 89.176 |
| MC+FQ+LPP | 96.500 | 96.766 | 96.633 |
| FQ+MC+PCA | 92.451 | 86.981 | 89.633 |
| FQ+MC+LPP | 96.354 | 96.959 | 96.655 |
| PCC-*k*-NN | 98.525 | 98.616 | 98.826 |

Table.5. Comparison for Wine Dataset

| Methods | Average Precision | Average Recall | Average F-measure |
|---|---|---|---|
| PCA | 89.845 | 88.749 | 89.294 |
| LPP | 91.095 | 91.608 | 91.351 |
| Mutual Correlation (MC) | 89.945 | 89.191 | 89.566 |
| FQ Measure | 93.172 | 93.984 | 93.577 |
| MC+FQ+PCA | 95.835 | 95.569 | 95.702 |
| MC+FQ+LPP | 95.756 | 95.767 | 95.761 |
| FQ+MC+PCA | 95.519 | 95.569 | 95.544 |
| FQ+MC+LPP | 96.018 | 96.489 | 96.253 |
| PCC-*k*-NN | 98.365 | 98.555 | 98.615 |

The comparison between the proposed and existing multi-dimensionality reduction methods shows that the PCC measurement followed by the *k*-NN method provides the 4.24%, 6.83% and 6.38% improvement in average precision, recall and F-measure for WDBC dataset. Similarly, the proposed PCC-*k*-NN offers 2.06%, 1.68% and 2.2% improvement in WBC dataset. Finally, the PCC-*k*-NN provides the 2.38%, 2.10% and 2.40% improvement for WINE dataset. The Table.6 presents the analysis of number of selected features for proposed PCC-*k*-NN with the bi-dimensionality reduction methods [28] for WDBC and WINE dataset.

Table.6. Analysis of Feature Extraction

| Methods | Features Extracted | |
|---|---|---|
| | WINE | WDBC |
| PCA | 10 | 16 |
| LPP | 10 | 18 |
| MC | 11-features selected | 20-features selected |
| MC+PCA | 11-features selected<br>6-features extracted | 20-features selected<br>6-features extracted |
| MC+LPP | 11-features selected<br>8-features extracted | 20-features selected<br>10-features extracted |
| PCC-*k*-NN | 13-features selected<br>11-features extracted | 28-features selected<br>22-features extracted |

The investigation of classification performance for proposed PCC-*k*-NN method and the existing methods on different datasets is discussed in the following sub-sections. The Table.7 describes the classification performance of proposed PCC-*k*-NN for original features and reduced features with several existing algorithms [29].

The comparative analysis shows that the proposed PCC-*k*-NN provides the 1.32% improvement with less dimensionality features. The application of conventional *k*-NN algorithm on WDBC provides the accuracy of 76.77%. But, the reduced features dimensionality in PCC-*k*-NN provides the 99.5612% accuracy which is 22.89% better than the *k*-NN method.

Table.7. Classification accuracy for WDBC dataset

| Algorithm | Accuracy (%) for all the features | Accuracy (%) for reduced features |
|---|---|---|
| Bayes Net | 94.7368 | 94.7368 |
| Naïve Bayesian | 90.3509 | 94.7368 |
| RBF | 92.9825 | 99.1228 |
| SMO | 97.3684 | 97.3684 |
| IBK | 95.6140 | 98.2456 |
| J48 | 92.9825 | 96.4912 |
| Simple Cart | 92.1053 | 94.7368 |
| PCC-*k*-NN | 97.154 | 99.5612 |

## 4.1 RMSE

The deviation of predicted and actual values defined by the metric called Root Mean Square Error (RMSE). The less deviation depicts the more correctly classified samples.

Since RMSE value depends on the deviation, the minimum values of difference causes the substantial reduction in RMSE. Hence, the algorithm effectiveness depends on the minimum RMSE values.

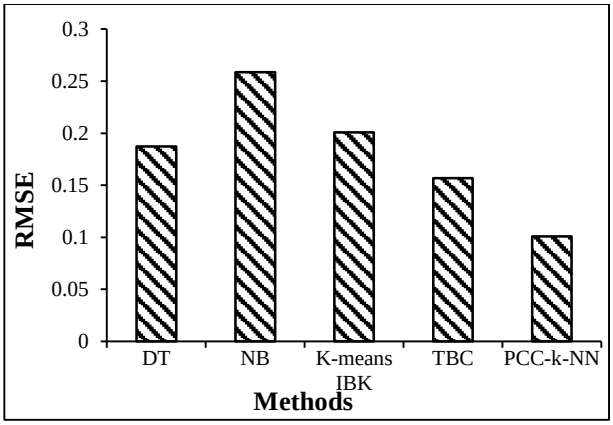


Fig.4. RMSE analysis

The Fig.4 shows the graphical illustration of RMSE variation for proposed PCC-*k*-NN and the existing methods of DT, NB, IBK, and TBC. Among the traditional methods, the TBC offers the minimum RMSE value 0.1568 compared to other methods. But, the data reduction prior to classification by using PCC offers less deviation between the actual and expected values which further reduces the RMSE (0.109) by 30% compared to TBC.

The Fig.5 shows the comparative analysis of correctly classified instance rate for proposed PCC-*k*-NN with the existing algorithms decision tree, NB, IBK, and TBC methods. The data reduction before classification effectively improves the correctly classified instances. The rate for TBC approach is 97.5395% which is DT (95.0791%), NB (92.9701%), and IBK (95.9578%). The correctly classified instance rate of proposed PCC-*k*-NN is 98.1135% which is 0.588% more than TBC.

## 4.2 CORRECTLY CLASSIFIED INSTANCE

The performance of the classification algorithms decided by a metric is called as a correctly classified instance. If the classified instance rate is higher, then the algorithm performance is better.

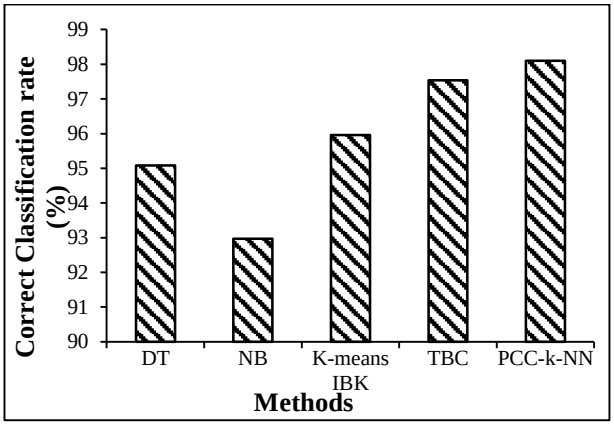


Fig.5. Correct classification rate analysis

## 4.3 TIME CONSUMPTION

The execution time for feature selection algorithm is low to state the effectiveness of the algorithm. The FA-FPR and FA-FSCE [4] reduce the time consumption of FPR and FSCE in traditional. The Table.8 shows the comparative analysis of proposed PCC-*k*-NN with the existing methods.

In the existing methods, the FA-FPR and FA-FSCE consume the less time and number of features extracted compared to FPR and FSCE. The instance-based learning and the splitting prior to classification in proposed work provide the trade-off between the low-time consumption and high-accuracy. The PCC-*k*-NN consume 43.127s, 41.258s, 161.28s and 1.4312s which are 13.75%, 6.06%, 6.16% and 3.16% less compared to FA-FSCE respectively for each dataset.

The comparative analysis of proposed PCC-*k*-NN and the existing methods of DT, NB, IBK and TBC illustrates that the instance-based data splitting, neighbor analysis, data reduction before classification in proposed PCC-*k*-NN provides the improved results of correct classification rate that shows the effectiveness in OC.

Table.8. Execution Time Analysis

| Datasets | Original Features | FPR | | FA-FPR | | FSCE | | FA-FSCE | | PCC-*k*-NN | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Features Selected | Time (s) | Features Selected | Time (s) | Features Selected | Time (s) | Features Selected | Time (s) | Features Selected | Time (s) |
| Sonar, Mines vs. Rocks | 60 | 20 | 135.4218 | 20 | 46.7187 | 41 | 300.5625 | 41 | 50 | 47 | 43.127 |
| Ionosphere | 34 | 24 | 213.2187 | 24 | 47.1250 | 24 | 137.0468 | 24 | 43.921 | 28 | 41.258 |
| WDBC | 30 | 22 | 313.7968 | 22 | 200.1875 | 27 | 228.9218 | 27 | 171.88 | 28 | 161.28 |
| Wine Recognition | 13 | 13 | 2.8906 | 13 | 1.7968 | 13 | 2.8906 | 13 | 2.0937 | 13 | 1.4312 |

## 5. CONCLUSION

The present paper focused the implementation of Distance Based Outlier Classification (DOC) with the reduced time and maximum accuracy. The problems addressed in the traditional classifiers are high-dimensionality, diverse features availability nearest neighborhood analysis. The present study performed the data reduction before classification for an accuracy improvement. The proposed work utilized the PCC to measure the correlation between the data instances. The extraction and labeling of most frequent samples from the training samples effectively classified the unlabeled instances. The PCC correlation-based instance learning and high value of *k* in training and testing phase reduced the dimensionality and noise respectively.

The frequent subset creation in accordance with the labels prior to *k*-NN improved OC. Moreover, this paper utilized the RMS in clusters cross-validation and optimal member extraction for effective outlier prediction. The comparative analysis between the proposed hybrid PCC-*k*-NN with the conventional algorithms regarding sensitivity, specificity, accuracy, precision, and recall proved the effectiveness of PCC-*k*-NN in OC. Besides, the experimental validation of proposed PCC-*k*-NN with the state of art methods dimensionality reduction methods regarding the execution time and classification accuracy provided the assurance of trade-off between the low-time consumption and high-accuracy.

## REFERENCES


[1] M.A.G. Sagade and R. Thakur, "Study of Outlier Detection Techniques for Low and High Dimensional Data", *International Journal of Scientific Engineering and Technology*, Vol. 3, No. 9, pp. 1-5, 2014.

[2] T. Al-Khateeb, M. M. Masud, L. Khan, C. Aggarwal, H. Jiawei and B. Thuraisingham, 'Stream Classification with Recurring and Novel Class Detection using Class-Based Ensemble", *Proceedings of 12th International Conference on Data Mining*, pp. 31-40, 2012.

[3] A. Albanese, S.K. Pal and A. Petrosino, "Rough Sets, Kernel Set, and Spatiotemporal Outlier Detection", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 26, pp. 194-207, 2014.

[4] Y. Qian, Q. Wang, H. Cheng, J. Liang and C. Dang, "Fuzzy-Rough Feature Selection Accelerator", *Fuzzy Sets and Systems*, Vol. 258, pp. 61-78, 2015.

[5] F. Angiulli, "Prototype-Based Domain Description for One Class Classification", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 34, pp. 1131-1144, 2012.

[6] N. Pham and R. Pagh, "A Near-Linear Time Approximation Algorithm for Angle-based Outlier Detection in High-Dimensional Data", *Proceedings of 18th ACM International Conference on Knowledge Discovery and Data Mining*, pp. 877-885, 2012.

[7] L. Galluccio, O. Michel, P. Comon, M. Kliger and A.O. Hero, "Clustering with a New Distance Measure Based on a Dual-Rooted Tree", *Information Sciences*, Vol. 251, pp. 96-113, 2013.

[8] B. Krawczyk, M. Wozniak and B. Cyganek, "Clustering-Based Ensembles for One-Class Classification", *Information Sciences*, Vol. 264, pp. 182-195, 2014.

[9] H. Kriegel, P. Kroger, E. Schubert and A. Zimek, "Outlier Detection in Arbitrarily Oriented Subspaces", *Proceedings of 12th IEEE International Conference on Data Mining*, pp. 379-388, 2012.

[10] O.M.B. Saeed, S. Sankaran, A.R.M. Shariff, H.Z.M. Shafri, R. Ehsani and M.S. Alfatni, "Classification of Oil Palm Fresh Fruit Bunches based on their Maturity using Portable Four-Band Sensor System", *Computers and Electronics in Agriculture*, Vol. 82, pp. 55-60, 2012

[11] A. Zimek, M. Gaudet, R.J. Campello and J. Sander, "Subsampling for Efficient and Effective Unsupervised Outlier Detection Ensembles", *Proceedings of 19th ACM International Conference on Knowledge Discovery and Data Mining*, pp. 428-436, 2013.

[12] S. Krishnan and H.G. Kerkhoff, "Exploiting Multiple Mahalanobis Distance Metrics to Screen Outliers From Analog Product Manufacturing Test Responses", *IEEE Design and Test*, Vol. 30, No. 3, pp. 18-24, 2013.

[13] R. Todeschini, D. Ballabio, V. Consonni, F. Sahigara and P. Filzmoser, "Locally Centred Mahalanobis Distance: A New Distance Measure with Salient Features Towards Outlier Detection", *Analytica Chimica Acta*, Vol. 787, pp. 1-9, 2013.

[14] A. Akila and E. Chandra, "Slope Finder-A Distance Measure for DTW based Isolated Word Speech Recognition", *International Journal of Engineering and Computer Science*, Vol. 2, No. 12, pp. 3411-3417, 2013.

[15] Z. Chao, S. Huanfeng, Z. Mingliang, Z. Liangpei and W. Penghai, "Reconstructing MODIS LST Based on Multitemporal Classification and Robust Regression", *IEEE Geoscience and Remote Sensing Letters*, Vol. 12, No. 3, pp. 512-516, 2015.

[16] L. Xu, S.-M. Yan, C.B. Cai and X.P. Yu, "One-Class Partial Least Squares (OCPLS) Classifier", *Chemometrics and Intelligent Laboratory Systems*, Vol. 126, pp. 1-5, 2013.

[17] K. Zawadzki, C. Feenders, M.P. Viana, M. Kaiser and L.D.F. Costa, "Morphological Homogeneity of Neurons: Searching for Outlier Neuronal Cells", *Neuroinformatics*, Vol. 10, No. 2, pp. 379-389, 2012.

[18] M. Hund, M. Behrisch, I. Farber, M. Sedlmair, T. Schreck and T. Seidl, "Subspace Nearest Neighbor Search-Problem Statement, Approaches, and Discussion", *Proceedings of International Conference on Similarity Search and Applications*, pp. 307-313, 2015.

[19] H. Liu and S. Zhang, "Noisy Data Elimination using Mutual K-Nearest Neighbor for Classification Mining", *Journal of Systems and Software*, Vol. 85, No. 2, pp. 1067-1074, 2012.

[20] Z.-G. Liu, Q. Pan and J. Dezert, "A New Belief-based K-Nearest Neighbor Classification Method", *Pattern Recognition*, Vol. 46, No. 3, pp. 834-844, 2013.

[21] N. Tomasev and D. Mladenic, "Hubness-Aware Shared Neighbor Distances for High-Dimensional K-Nearest Neighbor Classification", *Proceedings of International Conference on Hybrid Artificial Intelligence Systems*, pp. 116-124, 2014.

[22] Z. Yu, Y. Deng and A.K. Jain, "Keystroke Dynamics for User Authentication", *Proceedings of IEEE Computer*

*Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 117-123, 2012.

[23] A.B.E.D. Ahmed and I.S. Elaraby, "Data Mining: A Prediction for Students Performance using Classification Method", *World Journal of Computer Application and Technology*, Vol. 2, No. 1, pp. 43-47, 2014.

[24] H. M. Harb and M.A. Moustafa, "Selecting Optimal Subset of Features for Student Performance Model", *International Journal of Computer Science Issues*, Vol. 9, No. 5, pp. 253-262, 2012.

[25] P. Lam, L. Wang, H.Y. Ngan, N.H. Yung and A.G. Yeh, "Outlier Detection In Large-scale Traffic Data By Naive Bayes Method and Gaussian Mixture Model Method", *Proceedings of International Symposium on Electronic Imaging, Intelligent Robotics and Electronic Imaging, Intelligent Robotics and Industrial Applications using Computer Vision*, pp. 111-118, 2015.

[26] D. Rajakumari and S. Pannirselvam, "A Novel Triangular Boundary Based Classification Approach for Effective Prediction of Outliers", *International Journal of Applied Engineering Research*, Vol. 6, No. 1, pp. 322-328, 2015.

[27] Veerabhadrappa and L. Rangarajan, "Multi-Level Dimensionality Reduction Methods using Feature Selection and Feature Extraction", *International Journal of Computer Applications*, Vol. 4, No. 2, pp. 33-38, 2010.

[28] T. Sridevi and A. Murugan, "A Novel Feature Selection Method for Effective Breast Cancer Diagnosis and Prognosis", *International Journal of Computer Applications*, Vol. 88, No. 11, pp. 1-12, 2014.