Outlier Detection in Wellness Data using Probabilistic Mapped Mean-Shift Algorithms

Siriwan Phongsasiri[1] and Suwanna Rasmequan[2]

ABSTRACT: In this paper, the Probabilistic Mapped Mean-Shift Algorithm is proposed to detect anomalous data in public datasets and local hospital children's wellness clinic databases. The proposed framework consists of two main parts. First, the Probabilistic Mapping step consists of k-NN instance acquisition, data distribution calculation, and data point reposition. Truncated Gaussian Distribution (TGD) was used for controlling the boundary of the mapped points. Second, the Outlier Detection step consists of outlier score calculation and outlier selection. Experimental results show that the proposed algorithm outperformed the existing algorithms with real-world benchmark datasets and a Children's Wellness Clinic dataset (CWD). Outlier detection accuracy obtained from the proposed algorithm based on Wellness, Stamps, Arrhythmia, Pima, and Parkinson datasets was 93%, 94%, 80%, 75%, and 72%, respectively.

Keywords: Outlier detection, k-NN, Truncated Gaussian Distribution, Probabilistic Mapped, Mean shift

DOI: 10.37936/ecti-cit.2021152.244971

Article history: received February 17, 2021; revised July 30, 2021; accepted August 5, 2021; available online August 12, 2021

1. INTRODUCTION

Outlier detection was studied in order to enhance the classification ratio of the future data points. In general, an outlier detection problem refers to the distinction between faraway data points and common data points. However, if two classes are located in nearby areas and some parts of the first-class occur in the boundary of the second class, this is called an overlapping problem. In some cases, data points of the overlapping classes are placed in the same positions, but their outliers are marked in the area of the other class. Moreover, the Parkinson dataset contains outliers of about 75%, so it is much more difficult to identify the outlier positions in this dataset. Fig. 1 displays an examples of different characteristic outlier problems. In the example, there are two classes: an outlier and a normal class. The common outlier problem is illustrated in Fig. 1(a). Fig. 1(b) shows the normal class is larger than the outlier class but some parts of the outlier class occur at the boundary of the normal class. In contrast, in Fig. 1(c), the size of the outlier is larger than the normal class and they overlap. Note that in this study the relative sizes of different classes is not the main concern and has no effect on the accuracy of detection.

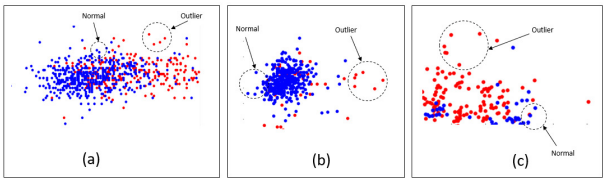

Fig. 1: *An example of outlier data (a) common outlier problem (b) normal class is larger than the outlier class (c) the size of the outlier is larger than the normal class.*

In Farag et al. [1][2][3], a novel data clustering algorithm based on gravity centre methodology was introduced. The strength of the gravity centre algorithm is that it does not need to specify parameters. But the disadvantage is that it cannot perform well

[1,2]The authors are with Faculty of Informatics Burapha University, Thailand., E-mail: 61910138@go.buu.ac.th and rasmequa@go.buu.ac.th

[2]Corresponding author: rasmequa@go.buu.ac.th

This work is funded by the faculty of Informatics, Burapha University, fiscal year 2019.

with a very small volume of data. The results of data clustering using Gravity Centre are compared with 3 other methods: K-means, K-medians, and K-medoids. The results indicate that Gravity Centre outperforms those techniques with the datasets used which were NNDSS's family, Health-infectious-disease-2001–2014, Unplanned Hospital Visits – Hospital, Diabetes, and Medicare National DMEPOS HCPCS.

Xiaokang et al. [4], proposed a density-weighted fuzzy outlier clustering approach for class imbalanced learning as a method for clustering fuzzy outlier clusters. This method considers the relationship of new ambiguous neighborhoods with local density data. When weighing the sample in the clustering process, it is mixed with the fuzzy outlier grouping method. In this way, the most representative samples are selected, while the anomalous samples are eliminated. The accuracy of this method shows a superior performance of 92% compared to other cluster sampling models. This indicates that the density-weighted fuzzy outlier grouping method can be used with imbalanced problems in real life. Blood-transfusion, Parkinson, Sick_numeric2, WDBC, and Wine datasets were used in this paper.

In Jiang et al. [5], a local-gravitation-based method for the detection of outliers and boundary points is presented. In it, each data point is viewed as an object with both mass and local resulting force (LRF) generated by its neighbor. When the number of neighbors increases, the LRF of outliers, boundary points, and internal scores change at different rates. The LRF rate of change of a lower density point has a higher scores. That is, the rate of change of the outlier is higher than that of the boundary and internal points. In other words, the highest-ranking score can be identified as an outlier. Likewise, the higher the rate of change of a point, the higher the LRF, and the greater the chance of it being a boundary point. The main advantage of the method is that it is independent of K-value selection. This will result in improving detection efficiency. Heart disease, Lymphography, Ionosphere, Breast cancer Wisconsin, Blood transfusion service center, and SPECTF heart datasets were used.

Aditya and Fitra [6] presented a method for Outlier Detection with a Supervised Learning Method. In their work, several popular classification methods, K-Nearest Neighbor, Centroid Classifier, and Naive Bayes, were compared as tools to handle outlier detection tasks. The results show that those methods achieved 81% of average sensitivity. They report that this is a reasonable starting value for future research to further modify the said methods to improve their performance. Elhossiny et al. [7] proposed using an enhanced K-Means++ to handle missing data and outliers. They worked on a diabetic classification system. They are looking for features that are related to classifying people into two groups: a diabetes group and a control group. The experimental result using RMSE is 17%. Vowels, Thyroid, Vertebral, Wine, Satellite, Breast cancer, and Ionosphere datasets were used.

Diego et al. [8] proposed the DBSCAN technique for large datasets. DBSCAN is a classic clustering method for identifying heterogeneous and isolated clusters with noise. There are a number of articles addressing DBSCAN scalability issues. Despite the scalability issues, the DBSCAN algorithm offers a reduction in execution time due to the reduction in the number of data formats. They also proposed a new heuristic called I-DBSCAN that can be modified and produced good results for the entire data set without any additional parameters. Abalone (Scale), Mushrooms, Pendigits, Letter, Cadata, Shuttle, Sensorless (Scale), SensIT (acoustic), SensIT (seismic), Skin-Nonskin and Poker datasets were used.

Paweł et al. [9] proposed the Fuzzy C-Means based Isolation Forest for outlier detection. In their study, they examined the feasibility of the proposed technique and analyzed it in detail for the different characteristics of data. For example, they considered database size, number of record attributes, and data type. FCM allows membership grade of the generated Isolation Forest nodes to the cluster based on these nodes to be generated. Therefore, this can lead to a fault score for a given element that may belong to a group of similar elements. To overcome this issue, a separate set of forest enrichment methods based on Fuzzy C-Means is proposed. The experimental results performed using 27 datasets indicated that FCM can play a key role in improving forest isolation approaches and increase the value of specific measures on the effectiveness of anomaly detection methods. Annthyroid, Arrhythmia, Breastw, Cardio, Cover, Glass, Ionosphere, Letter, Lympho, Mammography, Mnist, Musk, Optdigits, Pendigits, Pima, Satellite, Satimage-2, Shuttle, Speech, Thyroid, Vertebral, Vowels, Wine, Nad, and unsw0 datasets were used.

Paweł et al. [10] proposed a K-Means-based isolation forest to detect outliers. The K-Means-based Isolation Forest approach allows the creation of search maps. By employing many branches, as opposed to only two as considered in the traditional method, the k-mean grouping is used to estimate the number of divisions on each decision tree node. The proposed method is effective for information coming from different application areas including inter-model transport and spatial data. The advantage of this method is that information can be entered in the process of creating a decision tree. Moreover, it returns a more interesting anomaly score. Artificial sets, NYC Taxi, NYC Taxi (geographical positions), Ship transport, Ship transport, Train transport, Train transport, and Train transport datasets were used.

Patel and Kushwaha [11] reported in "Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model" that classification using GMM will enable discovering complex patterns and grouping them into cohesive, homogeneous components that are close representatives of real patterns within the dataset. In other words, it enables high cohesion among members of the same classes. That is, the dependency between the members is high. This can imply that it provides low coupling among different classes or the dependency between the classes is low.

In Yang et. al. [12], a mean-shift outlier detection and filtering technique is proposed to remove the bias caused from a large number of outliers before clustering without the need to know the number of outliers in advance. This method can also be applied with both numeric and string data. The experimental result of this method for the filtering task outperforms five other existing outlier removal methods: LOF, ODIN, NC, IFOREST and ABOD. This method also outperforms a number of the existing outlier detection methods: LOF, NC, KNN, ODIN, MCD, IFOREST, OCSVM, PCAD, and ABOD. This paper used Generated data, KDD-Cup99, Stamps, PageBlocks, Pima, Arrhythmia, and Parkinson datasets in the experiments.

In this work, the Probabilistic Mapped Mean-Shift Algorithm is proposed to detect outliers in public and children's wellness datasets. This method is based on its strong point that there is no need to know the number of outliers in advance. The difference from the classic mean-shift technique is the dispersion of the target position of the mapping depending on the joint probability density function of the mapping function. Unlike the mapping function proposed in [12], our proposed function will map the outliers to the proper point inside the boundary of each class. The mapped position is calculated using a Truncated Gaussian kernel. In other words, each outlier point has been mapped to a new position based on its probability value during the mapping process.

2. PROBLEM ANALYSIS

The framework proposed in this study is based on the conditions illustrated in Fig 1 (b) and (c). Given binary classes, outlier and normal points are occasionary placed in the same positions. We assume that the outlier class may overlap the normal class. There is no constraint imposed on the number of elements between the outlier class (n_o) and the normal class (n_n). This means that both $n_o > n_n$ and $n_n > n_o$ can be cases in this study. The only condition is the range of the output area of the mapping function. The mapping function proposed in [12] tried to map all outlier points into the mean or median point of each cluster as shown in Fig 2(a). In other words, this function will place all outlier points at the center of each cluster. This causes miss-classification in the testing step. Hence, the outlier detection framework proposed in this study focuses on the following issues:

1. How to increase the distribution of the range from the mapping function?
2. How to increase the coverage area of the mapped points obtained in issue 1, as shown in Fig. 2(b)?

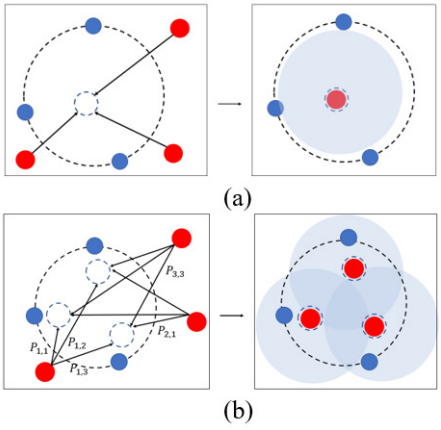

Fig.2: *Mapping functions: (a) Mean Shift Algorithm, (b) Probabilistic Mapped Mean-Shift Algorithm.*

In general, the mean shift algorithm will translate all outlier data points of each cluster to new positions as shown in Fig. 2(a). But in some scenarios, more distribution and more coverage are needed as shown in Fig. 2(b).

3. BACKGROUND

3.1 Mean-Shift Algorithm

The Mean Shift algorithm is a non-parametric clustering algorithm for finding high-density areas (a density function) of the input feature space. It is sometimes called the mode-seeking algorithm. The Mean Shift technique does not need to know the number of clusters in advance. In addition, it is not limited by the shape of the cluster. The algorithm is an iterative procedure. It will repeat until it reaches the optimal area having maxima mode. This algorithm consists of two main steps as follows:

step 1: Compute the mean shift vector: $m(X^t)$
step 2: Translate the input X^t to a new position: $X^t_{map} = X^t + m(X^t)$

3.2 Probabilistic Mapped Model

A probabilistic decision model is a decision method for predicting future events based on probability theory. In contrast, a deterministic decision determines an exact circumstance. The probabilistic mapped model is a mapping function for predicting a new position (Range) of input data (Domain) based on the

its properties. Generally, probabilistic mapped methods will work better than deterministic models. This is because the probabilistic method uses all properties or information contained in the input data. The combination of the probabilistic model with the mean shift method, which is an iterative mapping procedure, help improve the mapping process. That is, a better coverage area for each individual cluster is reached. Hence, outlier prediction accuracy will be higher. This is because the mapping of input data points from the current position is assigned to a new proper area within its class.

4. THE PROPOSED METHOD

In this work, the anomalous data, or outliers, in the large public benchmarks and Children's Wellness Clinic dataset were detected using our Probabilistic Mapped Mean-Shift Algorithm. The proposed method is based on the Mean Shift technique proposed by Yang et. al. [12] as mentioned in Section 1. The proposed framework is divided into two main sections. Section I: Probabilistic Mapping, consists of k-NN instance acquisition, data distribution calculation, and data point reposition. Section II: Outlier Detection, consists of outlier score calculation and outlier selection. A brief description of the proposed method is shown in Fig. 3.

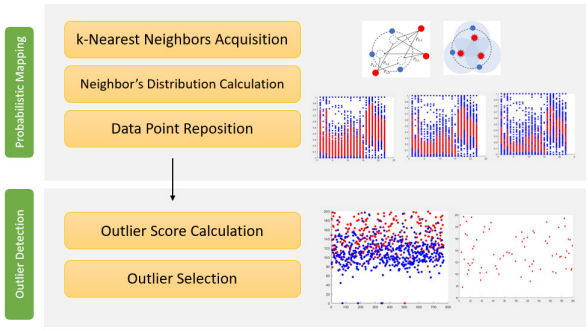

Fig.3: *The proposed framework.*

4.1 Probabilistic Mapping

In this section, mapping the input instance into a new location depends on its probalistic value is described. This procedure consists of three steps. In Step 1, the finding of k-nearest neighbors of each instance is carried out. In Step 2, the calculation of local Gaussian distribution is performed. In Step 3, the repositioning of the instance is achieved.

4.1.1 k-NN Acquisition

To calculate the local Gaussian distribution around the instance, a set of the nearest points is acquired by the k-nearest neighbor algorithm. The steps of k-NN can be described as follows:

step 1: Set value for $k, k \epsilon I^+$
step 2: For each point in the data do the following
→ **2.1** Calculate the distance between data point and each row of the dataset.
→ **2.2** Sort the data in ascending order based on the distance value.
→ **2.3** Select top k data point for the nearest group.
step 3: Repeat these steps until reaching the end of the dataset.

4.1.2 Neighbour's Distribution Calculation

After the data has been separated into groups by k-NN, each group of data is used to calculate local data distribution. To compute local probabilistic distribution, the Truncated Gaussian Distribution (TGD) model was applied. TGD can be described with Eq. (1).

$$(x; \mu, \sigma, a, b) = \frac{1}{\sigma} \frac{\phi\left(\frac{x-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)} \quad (1)$$

Let X be a normal distribution with mean μ and variance σ^2, within the interval (a,b). Then X conditional on $a \leq x \leq b$ has a Truncated Gaussian Distribution.

$$\phi(\xi) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\xi^2\right) \quad (2)$$

ϕ is the probability density function of the standard Gaussian Distribution.

$$\Phi(x) = \frac{1}{2}(1 + \text{erf}(x/\sqrt{2})) \quad (3)$$

Φ is its cumulative distribution function

The steps of TGD mapping 4 can be described as follows:

step 1: Calculate the lower boundary a and upper boundary b of probability in each k-NN set.

step 2: Calculate the Mean μ and standard deviation σ of each k-NN set.

step 3: Calculate distribution probability using Eq. (2).

step 4: Repeat until all k-NN sets have been processed as shown in Fig. 4.

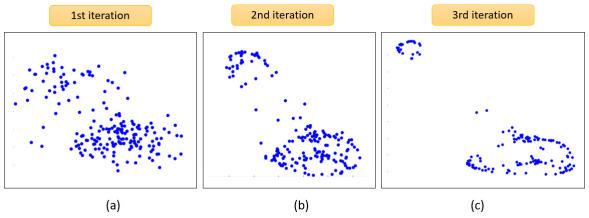

Fig.4: *Mapped data points using Probabilistic Mapped Mean-Shift Algorithms.*

4.1.3 Repositioning Data Points

To shift the data points to their new positions as shown in Fig. 5 for Synthetic data and Fig. 6 for Parkingson data, the highest probability value is selected to assign each new position. In Truncated Gaussian Repositioning there are three main steps:

step 1: Select the initial point from the k-NN set.
step 2: Repeat until converged:
→ **2.1** For each point, find weights encoding the probability of membership in each cluster
→ **2.2** For each cluster, update its location, normalization, and shape based on all data points, making use of the weights.
step 3: Select the highest probability weight.

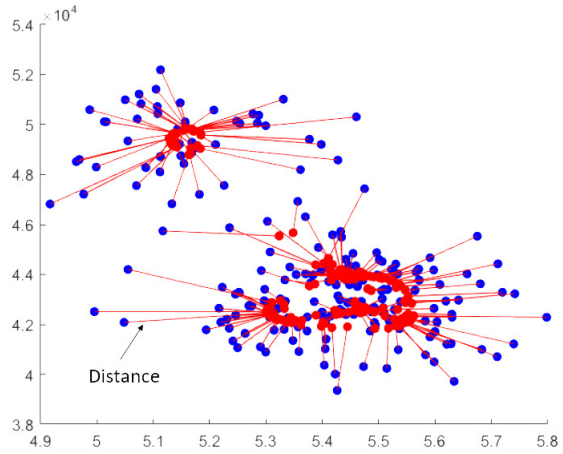

Fig.5: *Visualization of the Mapped points of each cluster in synthetic data. Blue points are original points and Red points are mapped points.*

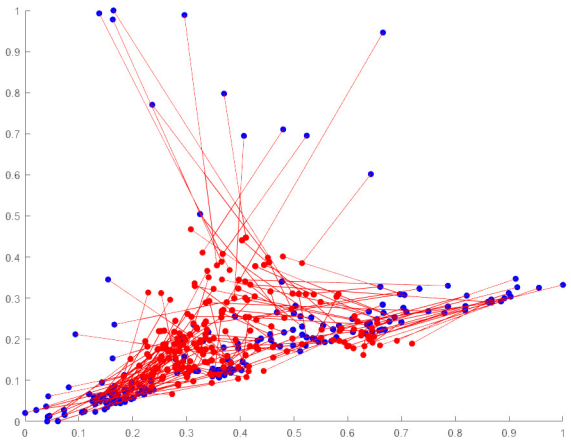

Fig.6: *Visualization of the Mapped points of each cluster in benchmark Parkinson data. Blue points are original points and Red points are mapped points.*

In Fig. 7, 3D space mapped points of Parkinson data are illustrated and the results show that the data was mapped into the new areas with the highest probability. This will increase the cohesion and uniformity between class members for each class as shown in Fig. 8.

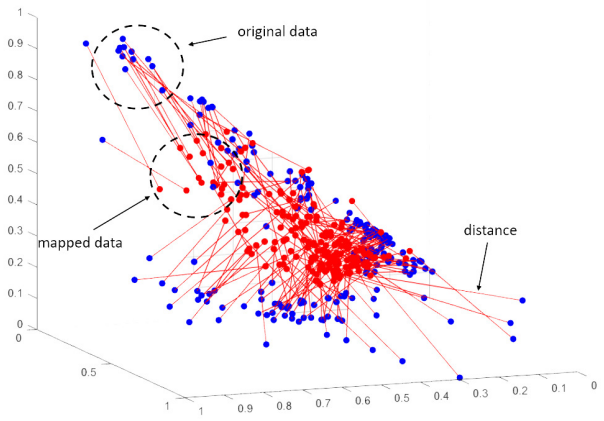

Fig.7: *Visualization in 3D space of the Mapped points of each cluster in benchmark Parkinson data. Blue points are original points and Red points are mapped points.*

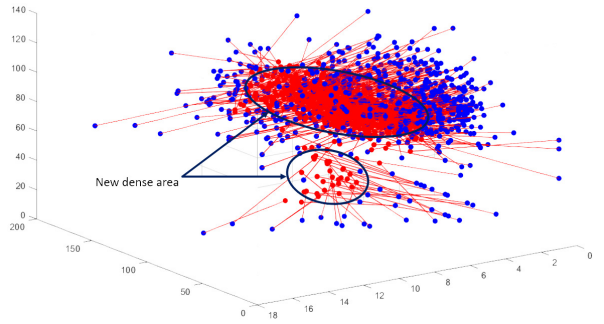

Fig.8: *Resulting Denser Area after Mapping.*

The result shows that the normal data (Blue vertices) was mapped into the denser area (Red vertices). The distribution of each class was normality tested using Shapiro-Wilk test [13]. The result shows the mapped data gained higher normality rates as shown in Eq. (4).

$$W = \frac{(\sum_{i=1}^{n} a_i x_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{4}$$

Let $x_{(i)}$ be the i^{th} order statistic, such as, the i^{th}-smallest number in the sample. The coefficients a_i are given by Eq.(5).

$$(a_1, \ldots, a_n) = \frac{m^T V^{-1}}{C} \tag{5}$$

Where C is the normalized vector given in Eq.(6).

$$C = \sqrt{(m^T V^{-1} V^{-1} m)}, m \in (m_1, \ldots, m_n)^T \quad (6)$$

4.2 Outlier Detection

In this section, to detect outliers among normal data, two steps are used. In Step 1, the outlier score is calculated. In Step 2, an outlier selection is performed.

4.2.1 Outlier Score Calculation

By using the local distribution of its k-NN, the current object was forcibly repositioned to the denser mean probability area. The length of the object's movement was computed as a piece of outlier evidence. To calculate the outlier score, instead of depending on only clustering or classification techniques, the distance between mapped instances and original data was used. The farther an instance was moved, the stronger the outlier score obtained. The distance between the objects can be calculated with Eq. (7).

$$d(X_i, X_{map(i)}) = \sum_{i=1}^{dn} \left| X_i - X_{map(i)} \right| \quad (7)$$

Let $d(X_i, X_{map(i)})$ be the distance between the original data and the mapped instance. X_i is data from the original set and $X_{map(i)}$is data from the mapped instance.

The mapped data contains more information from the data. In the case of a large number of outliers among the normal class, for more robustness, the variant of the distance set is used as an extended reference set for outlier score calculation. Thus, the outlier score can be computed by Eq.(8).

$$S = \sum_{i=1}^{n} \left| X_i - X_{map(i)} \right| + \frac{\sum (X_i - \mu)^2}{n-1} \quad (8)$$

Let S be an outlier score. X_i is data from the original set and $X_{map(i)}$ is data from the mapped instance. μ is the mean of the k-NN sets. The score is used in the outlier selection step. The steps for score calculation are:

step 1: Find the k-nearest neighbor for each instance for local variant calculation.
step 2: Compute the outlier score using Eq. (8)
step 3: Repeat until reaching the end of the dataset.

The distance between mapped points and original data points as depicted in Fig. 9 was used in outlier score calculation. Then, those scores were used in the outlier selection step.

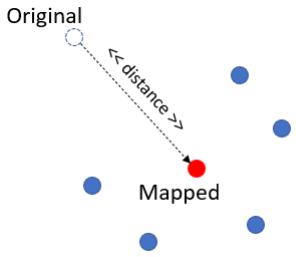

Fig. 9: *Example of distance acquisition between mapped data and original data.*

The scores from benchmark datasets, Stamp (Fig. 10 (a)), Pima (Fig.10 (b)), Arrh (Fig.10 (c)), and Parkinson (Fig.10 (d)) are illustrated in Fig. 10.

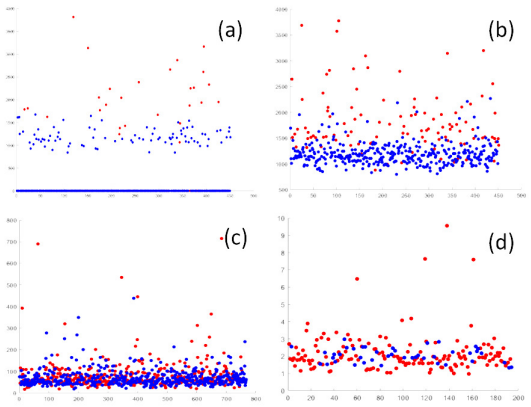

Fig. 10: *Visualization of outlier score calculation.*

4.2.2 Outlier Selection

To select outliers from the entire dataset, Yang et. al. [12] proposed a procedure to designate the Top-N outlier score. The number of chosen points can be determined using Eq. (9).

$$Top_n = \left\lfloor \frac{1}{2}(R_{out} N) \right\rfloor, Top_n \in I^+ \quad (9)$$

Top_n is the number of Top-N outlier scores. R_{out} is an outlier ratio or percentage of the datasets. N is the number of entire instances. For example, the Parkinson dataset contains 75.4% outliers. The Top-N of Parkinson dataset can be calculated by $\frac{1}{2}(0.75)(195) \approx 73$ points as shown in Table 1.

Table 1: *Top-N Selection Results.*

Dataset	Instances	Top-N
CWC	2533	50
Stamps	340	19
Arrhythmia	450	36
Pima	768	138
Parkinson	195	73

In Fig.11, Parkinson outlier Scores are shown. The red dots and blue dots are likely outliers and normal data respectively. In this step, the calculated outlier scores without sorting are shown in Fig. 11 (a). In order to select outlier candidates with the Top-N approach, the outlier scores must be sorted in descending order as shown in Fig. 11 (b). The outlier selection steps are:

step 1: Calculate Top-N for selecting outlier candidates using Eq. (9)

step 2: Sort the data points with outlier scores in descending order as shown in Fig.11 (b)

step 3: Select the first Top-N data points (from **step 1**) for outlier candidate selection.

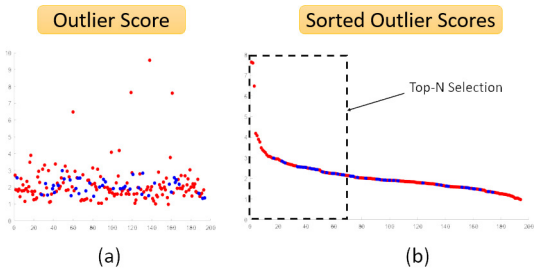

Fig.11: *Visualization of Parkinson Outlier Scores.*

The procedure of the Probabilistic Mapped Mean-Shift Outlier Detection (PMMS) process is summarised in Algorithm 1.

Algorithm 1 Probabilistic-Mapping Mean-Shift $\{PMMS(X,k,i)\}$

Input: Dataset X, Nearest Neighbor k, Iteration i

Output: Mapped Dataset X_{map}, Outlier Score S

REPEAT i **TIMES**:
  FOR $X_i \in X$:
    $K_{nn} \leftarrow kNN(X_i), K_{nn} \in X$ Find k-nearest neighbors of X_i
    COMPUTE the probabilistic distribution of K_{nn}
      $a_i \leftarrow$ **COMPUTE** lower boundary of K_{nn}
      $b_i \leftarrow$ **COMPUTE** upper boundary of K_{nn}
      $\sigma_i \leftarrow$ **COMPUTE** Standard Deviation of K_{nn}
      $\mu_i \leftarrow$ **COMPUTE** Mean of K_{nn}
      COMPUTE Truncated Gaussian Distribution
      FOR $k_j \in K_{nn}$:
        $P_i(k_j;\mu_i,\sigma_i,a_i,b_i) = \frac{1}{\sigma_i} \frac{\phi\left(\frac{k_j-\mu_i}{\sigma}\right)}{\Phi\left(\frac{b_i-\mu_i}{\sigma_i}\right)\Phi\left(\frac{a_i-\mu_i}{\sigma_i}\right)}$
    $X_{map} \leftarrow P_i$ **REPOSITION** X_i **WITH** P_i
  COMPUTE Outlier Score
    FOR $X_i \in X$:
    $K_{nn} \leftarrow kNN(X_i), K_{nn} \in X$ Find k-nearest neighbors of X_i
    $S_i \leftarrow \sum_{i=1}^{n} |X_i - X_{map(i)}| + \frac{\sum(X_i-\mu)^2}{n-1}$, $X_{map(i)} \in X_{map}$, $X_i \in X$

5. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, the dataset used in this experiment and a performance evaluation of the proposed methods are described.

5.1 Dataset

The proposed approach was tested on four public benchmark datasets and a real-world children's wellness clinic dataset.

5.1.1 Public Datasets

The proposed method was tested against the public datasets Stamp, Arrhythmia, Pima, and Parkinson as performed in Yang et. al. [12]. Those public datasets (Stamps, Arrh, Pima, and Parkinson) have outlier ratios of 9.1%, 15%, 34.9%, and 75.4%, respectively as shown in Table 2. The detection accuracy, as shown in Table 3, obtained from the existing methods is 89%, 73%, 74%, and 68%, respectively. These accuracy rates left room for improvement.

5.1.2 Children's Wellness Clinic dataset (CWC)

The real-world dataset was provided from Children's Wellness Clinic of Burapha University Hospital. It consists of 2533 instances with 9 dimensions. The outliers were labelled as 4% of entire dataset.

Table 2: *Benchmark and Real World dataset description.*

Dataset	Instances	Dimension	Outlier(%)
CWC	2533	9	4
Stamps	340	9	9.10
Arrhythmia	450	259	15.8
Pima	768	8	34.9
Parkinson	195	22	75.4

5.2 Experimental Results

The experimental results revealed in Table 3 indicate that these results address the first research issue concerning how the large dataset effects detection of outliers. For the second issue, the experimental results show that adding additional parameters helps improve the detection performance. That is, the proposed method with adjusted parameters outperforms the existing methods. In the performance evaluation, the accuracy is obtained with the confusion matrix measurement in Eq. (10).

$$Acc = \frac{TP}{TP + FP} \tag{10}$$

Let TP be the number of True positive cases (number of outliers were selected in Top-N method). FP is the number of False Negative cases (number of normal points were selected in Top-N method). The result shows the proposed method obtained 0.94, 0.80, 0.75, 0.72 accuracies in benchmark datasets, consists of Stamps, Arrhythmia, Pima, and Parkinson, respectively as shown in Table 3 and Fig. 12.

Table 3: *BResults of the proposed method compared with the existing methods for detecting anomalous data.*

Dataset ((Outliers)	CWC. (4%)	Stamps (9.1%)	Arrh (15%)	Pima (34.9%)	Parkinson (75.4%)	Average
PMMS (proposed)	**0.93**	**0.94**	**0.80**	**0.75**	**0.72**	**0.80**
DOD+	0.91	0.89	0.73	0.71	0.63	0.74
MOD+	0.89	0.91	0.72	0.73	0.66	0.76
DOD	0.92	0.81	0.73	0.68	0.68	0.73
MOD	0.92	0.78	0.74	0.65	0.60	0.69
LOF	-	0.53	0.73	0.60	0.56	0.61
ODIN	-	0.58	0.70	0.56	0.45	0.57
NC	-	0.50	0.60	0.52	0.46	0.52
KNN	-	0.89	0.74	0.72	0.54	0.72
ABOD	-	0.81	0.72	0.70	0.64	0.72

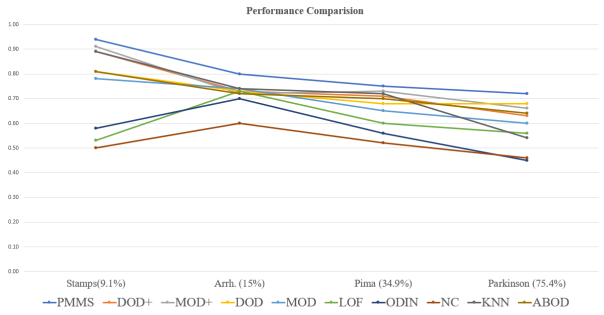

Fig. 12: *Performance Comparation.*

Likewise, the proposed method also outperforms the existing methods for the real world dataset Children's Wellness Clinic with an accuracy rate of 0.93. However, the proposed method needs to be improved further to reach higher accuracy in large datasets and datasets containing a large amount of outlier data.

5.3 Discussion

Although the proposed method outperforms the others (see Table 3), the accuracy is inverse to the outlier percentage as shown in Fig. 13. In the other words, the accuracy of the proposed approach tends to be lower in cases with a larger percentage of outliers. For the Parkinson dataset, which contains 75% outliers, the proposed approach obtains the lowest accuracy rate. To eliminate this weakness, an improved probability model and better outlier selection methods need to be found.

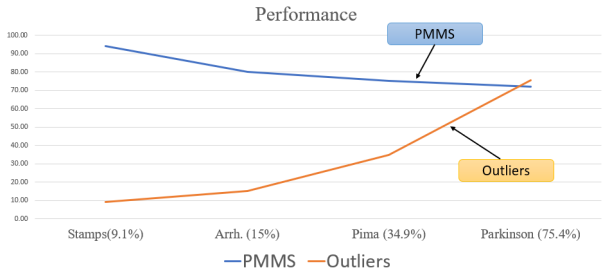

Fig. 13: *Relationship between proposed method and percentage of outliers.*

In addition, not only the datasets having high percentages of outliers decrease the accuracy, but the datasets having imbalanced class members do also. The latter case is shown in Fig. 14. Thus, in the future work, the detection of outliers for imbalanced classes will be pursued.

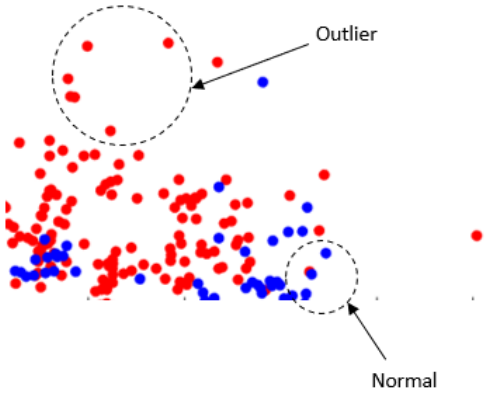

Fig. 14: *Visualization of fully overlapped problem.*

6. CONCLUSION

To enhance machine-learning or data-mining algorithms such as classification and clustering, outlier data must be eliminated. In this research, the Probabilistic Mapped Mean-Shift Algorithm is proposed to detect and filter outlier data in public benchmark and local hospital children's wellness clinic datasets. The experiment results indicated that the proposed method can address the first research issue by increasing the distribution of the range. The results also demonstrated that the proposed method can address the second issue by increasing the coverage area of the mapped points obtained in issue 1, so that each group of mapped points has its own normality. The normality of shifted instances was evaluated using the Shapiro-Wilk test. The test result demonstrated that the normality rate is increased. However, the distribution and the coverage still need to be improved further. In addition a wider variety of datasets should be used to test our method.

The proposed approach consists of two main phases: Phase I, Probabilistic Mapping, including k-NN instance acquisition, data distribution calculation, and data point reposition, and Phase II: Outlier Detection, including outlier score calculation, and outlier selection. The proposed approach was tested on benchmark datasets including Stamps, Arrhythmia, Pima, and Parkinson datasets, and also on a local hospital Children's Wellness Clinic (CWC) dataset. Experimental results indicated that the proposed algorithm achieves higher accuracy than the existing algorithms. The accuracy achieved by the proposed method is 94%, 80%, 75%, 72%, and 93% respectively. In future work, datasets with fully overlapped and imbalanced classes, as shown in Fig. 14, will be taken into consideration.

References

[1] A. Boukerche, L. Zheng and O. Alfandi, "Outlier Detection: Methods, Models, and Classification," *ACM Computing Surveys (CSUR)*, Vol.53, pp.1-37, 2020.

[2] J. Huang, Q. Zhu, L. Yang, D. D. Cheng and Q. Wu, "A novel outlier cluster detection algorithm without top-n parameter," *Knowledge-Based Systems*, Vol.121, pp.32-40, 2017.

[3] F. H. Kuwil, Ü. Atila, R. Abu-Issa and F. Murtagh,, "A novel data clustering algorithm based on gravity center methodology," *Expert Systems with Applications*, Vol. 156, 2020.

[4] X. Wang, H. Wang and Y. Wang, "A density weighted fuzzy outlier clustering approach for class imbalanced learning," *Springer Neural Computing and Applications*, 2020.

[5] J. Xie, Z. Xiong, Q. Dai, X. Wang and Y. Zhang, "A local-gravitation-based method for the detection of outliers and boundary points," *Knowledge-Based Systems*, Vol.192, 2020.

[6] A. H. Bawono and F. A. Bachtiar, "Outlier Detection with Supervised Learning Method," *2019 International Conference on Sustainable Information Engineering and Technology (SIET)*, pp.306-309, 2019.

[7] E. Ibrahim, M. A. Shouman, H. Torkey and A. El-Sayed, "Handling missing and outliers values by enhanced algorithms for an accurate diabetic classification system," *Springer Multimedia and Applications*, Vol.80, pp.20125-20147, 2021.

[8] D. Luchi, A. L. Rodrigues, F. M. Varejáo, "Sampling approaches for applying DBSCAN to large datasets," *Pattern Recognition Letters*, Vol. 117, pp.90-96, 2019.

[9] P. Karczmarek, A. Kiersztyn, W. Pedrycz and E. Al, "K-Means-based isolation forest," *Knowledge-Based Systems*, Vol. 195, 2020.

[10] P. Karczmarek, A. Kiersztyn, W. Pedrycz and D. Czerwiński,, "Fuzzy C-Means-based Isolation Forest," *Elsevier Applied Soft Computing*, Vol.106, 2021.

[11] E. Patel and D. S. Kushwaha, "Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model," *Procedia Computer Science*, Vol.171, pp.158-167, 2020.

[12] J. Yang, S. Rahardja and P. Fränti, "Mean-shift outlier detection and filtering," *Pattern Recognition*, Vol.115, 2021.

[13] E. González-Estrada and W. Cosmes, "Shapiro–Wilk test for skew normal distributions based on data transformations," *Journal of Statistical Computation and Simulation*, Vol.89, Issue 17, pp.3258-3272, 2019.

Siriwan Phongsasiri received the B.Sc. degree in Computer Science from Burapha University, Chonburi, Thailand, in 2017. Currently, she is a M.Sc. student in Informatics, at Faculty of Informatics, Burapha University, Chonburi, Thailand.

Suwanna Rasmequan received the B.B.A. degree in Finance and Banking and the M.Sc. degree in Computer Information System from Assumption University, Bangkok, Thailand, in1992 and 1994 respectively. She received Ph.D. degrees in Computer Science in 2002 from University of Warwick, Coventry, United Kingdom. She worked in the Business Sector from 1984 until 1997 in two major businesses namely Packaging Industry and Satellite Communication Provider. Her work responsibilities in those businesses was starting with the beginning post of Executive Secretary and ending with the post of Section Manager. She was a Lecturer at the Department of Computer Science, Burapha University, Chonburi, Thailand during 1997 – 2006. She has been worked as an Assistant Professor from year 2006 up to present at the Faculty of Informatics, Burapha University. Her major research interests include Empirical Modelling, Decision Support Systems, Machine Learning Applications.