SN Computer Science (2020) 1:76
https://doi.org/10.1007/s42979-020-0085-x

ORIGINAL RESEARCH

Imbalance Data Classification Using Local Mahalanobis Distance Learning Based on Nearest Neighbor

Nijaguna Gollara Siddappa[1] · Thippeswamy Kampalappa[1]

Published online: 11 March 2020

Abstract
In the dataset, any one of its classes is normally outnumbered by other classes and is known as class imbalance data. Many standard learning algorithms face the classification problem in performance due to imbalance data. The issues can be solved by many existing conventional methods such as cost-sensitive, sampling or ensemble methods. But these methods alter the original data distribution, which leads to loss of useful information of the users and it may cause unexpected errors or increase the problem of overfitting. In this research, local Mahalanobis distance learning (LMDL) method is applied in the nearest neighbor (NN) for improving the performance of the classification in the imbalance dataset. The multiple distance metrics are used in the LMDL to investigate the data effectively and obtain the relevant features based on the analysis. The distance metric uses the original data for learning the prototype and support the NN. A number of experiments on various datasets are conducted for validating the quality as well as the efficiency of the proposed LMDL method. The experimental results stated that the proposed LMDL achieved nearly 82% in E-coli dataset, 94% in breast cancer dataset and 98% in Iris dataset for all metrics such as accuracy, precision, recall and *F*-measure.

Keywords Anomalies · Classification · Imbalance data · Information loss · Local Mahalanobis distance learning · Nearest neighbor · Sampling

Introduction

The events that occur very less when compared to frequently occurred events is referred as rare events or abnormal behavior which are difficult to detect, but often require responses from various management functions in a timely manner [1]. There are some of the rare events that include software defects, natural disasters, cancer gene expressions, fraudulent credit card transactions and telecommunications fraud [2]. Detecting events are a data classification problem or prediction problem in the field of data mining. Due to the infrequency and casualness, predicting the rare events is very difficult which leads to misclassify this abnormal behavior [3]. Within a dataset, one or number of the classes has a large number of sample data than other class is known as imbalanced [4]. The problem of classification with imbalanced data is the process of extracting the useful information from the datasets which creates the complications in extraction. This occurs because the number of instances in majority classes (negative class) is larger than the number of instances that belongs to the other classes (minority or positive class). In this situation, the learning algorithms focus on the minority class that needs to be correctly identified in these problems [5]. Big data are also affected when dealing with the problem of imbalanced datasets in uneven data distribution; moreover, standard classification algorithms are also failed to work appropriately with imbalance data. A mechanism neglects the rules that are associated with the minority class for using the global performance measures to construct the model [6].

In recent years, the most vital fields such as data mining and machine learning algorithms had an impact of class imbalance learning [7]. Data mining approaches are used to make the decision in commercial models using different classification models, but it is difficult to classify imbalanced data for these traditional classification models such as

This article is part of the topical collection “Advances in Computational Intelligence, Paradigms and Applications” guest edited by Young Lee and S. Meenakshi Sundaram.

✉ Nijaguna Gollara Siddappa
nijagunags@gmail.com

[1] Visvesvaraya Technological University, Belagavi, India

sampling approaches, cost-sensitive learning, active learning, kernel-based methods and ensemble learning [8–12]. When the data contain some important complexity, such as small sample size, high dimensionality and much noise to the dataset, the existing learning algorithms for imbalance data are generally becoming invalid [13–15]. The performance of these classification algorithms improves by computing the distance metrics and is learned with the help of the decision-making process in NN classifier. In many machine learning tasks, DM learning algorithm plays an important role to solve the POF [16]. Novelty of this work: Many conventional methods suffer from the unexpected errors or loss of information and change in the distribution of original class, but there is no change in original class distribution in proposed method which had all information in imbalance dataset. LMDL learns Mahalanobis distance metric (MDM) for a small set of samples (prototypes) according to its closely related objective function. The POF is reduced, and the objective function is minimized by adjusting the position of prototypes. The predictive performance of imbalance data improves by using LMDL method over the existing method which is validated by the number of different experimental evaluations. The contribution of this work can be as follows:

- The LMDL method uses the original data distribution by learning exactly one MDM for each prototype based on its objective functions which are closely related to the NN decision rule.
- To address the POF, the prototype position is adjusted.
- To make the proposed method more flexible and efficient, the LMDL method has been developed according to its distance concept.

The organization of this research draft is summarized as: “Literature Review” section provides a brief history of some related methods in imbalance data classification. “Problems in Imbalance Dataset” section presents the nature of the imbalance problem in details. “Proposed Methodology” section describes the proposed LMDL methodology in detail. “Experimental Analysis” section presents and discusses the experimental results. Finally, the conclusion of the paper with future work is given in “Conclusion” section.

Literature Review

Researcher developed many methodologies on imbalance data classification. In this subsection, a brief evaluation of a few essential contributions to the existing literature is presented.

Sun et al. [17] built a number of classifiers on multiple balanced data which were converted from an imbalanced dataset with the help of a specific classification algorithm. The new data were created by combining the specific ensemble rule with classification results which included these classifiers. To solve the highly imbalanced data problems, the experiments were conducted on 46 imbalanced datasets and experimental results stated that this proposed method was superior to the conventional imbalance data handling methods. This method delivered poor performance in classification results because of using a constant value in the denominator. In some cases, the classification performance was negligible due to the impact of these constant values.

Napierała and Stefanowski [18] developed the rule-based classifiers with the help of expert knowledge from class-imbalanced data to the learning process. The argument-based learning technique was used to adapt the learning rules from imbalanced data. ABMODLEM was proposed with specialized classification method which provided a rule induction algorithm. The experimental result shows that the ABMODLEM increases the recognition of minority class especially in the difficult data distribution. The possible limitation of the argument-based learning was less universal than fully automatic method to increase the classifiers since expected knowledge is not always available. The knowledge acquisition method was more time- and effect-consuming than automatic methods, and its applicability is limited to the environment that requires fast response from the learner.

Ohsaki et al. [19] achieved a better classification performance with the help of confusion matrix-based kernel logistic regression (CM-KLOGR) algorithm by forming a new dataset with task independence. Based on minimum classification error and generalized probabilistic descent (MCE/GPD) learning, the optimization and the objective function of CM-KLOGR were consistently formulated on KLOGR. The extensive experiments were conducted on benchmark-imbalanced datasets, and the results showed the effectiveness of CM-KLOGR when compared with existing technique. Because of heuristic process and task dependence, the ensemble methods and cost-sensitive were not used in this method.

Krawczyk et al. [20] developed a new ensemble method of cost-sensitive decision trees for classifying the imbalance dataset. According to the given cost matrix, base classifiers were implemented, but the diversity of ensemble members was ensured by training on random feature subspaces of imbalanced data. The committee member weights were assigned for fusion process, and selection of simultaneous classifier was done by the decision tree with the help of an evolutionary algorithm. The derivation of cost matrices was one of the major issues in cost-sensitive classification. The above limitation was addressed by ROC analysis and showed the correlation between the optimal cost matrix and dataset imbalance ratio. The proposed method provided poor performance when there was a more extreme imbalance ratio.

Park and Ghosh [21] implemented splitting and stopping criterion of decision tree ensemble methods for imbalanced data classification using the properties of α-divergence. When the α-divergence splitting criterion was applied to imbalanced data, the method tends to be less correlated by changing the value of α. During the growth of the tree, the base classifiers were used as a stopping criterion in the ensemble method. The effectiveness of this proposed ensemble method was proven by the experimental results on many class-imbalanced datasets over a wide range of imbalance data distributions. In LEAT framework, the overlaps between different α-tree rules were not studied to reduce the number of tree. While examining the logical relationships, the overlapping rules were not reduced that simplified the LEAT output.

Patel and Thakur [22] improved the fuzzy K-NN classification of imbalanced data with the help of adaptive K-NN, as this method tends to choose different values of K-based on its sizes. Compared to other simple fuzzy K-NN, acquired fuzzy memberships were more accurate for data instances in minority class using the adaptive K-NN. The experimental result shows that the adaptive K-NN has the better performance on imbalanced data. The adaptive K-NN has been designed for binary classification based on NN. However, the method provides poor performance while having the imbalance data in feature-based NN for multi-label applications.

To overcome the above-mentioned issues, the proposed method uses the learning-based algorithm for K-NN classification tasks. The objective function which is closely related to the NN decision rule is used in LMDL. To capture the local discriminative information, LMDL learns Mahalanobis metric for each prototype and uses the notion of this prototype to prevent the risk of overfitting. The proposed LMDL method automatically adjusts the position of prototypes for finding the best position.

Problems in Imbalance Dataset

Imbalance data can be described as the classes had an unequal conveyance in any datasets which provides poor performance in classification accuracy. The issue can be classified as intrinsic and extrinsic, imbalance due to rare instances and relative imbalance, dataset complexity and finally, imbalance with the small size dataset.

Intrinsic and Extrinsic

Due to different factors such as time and storage, the data can be considered as intrinsic imbalanced data that are directly related to the data space nature. In addition, extrinsic imbalance data are not directly related to the data space nature.

Relative Imbalance and Rare Instances

From the imbalance, the minority concept is learned accurately with little disturbance which is shown in a few studies for certain relative imbalanced datasets. When the target concept is rare, i.e., minority class is limited in a certain domain often referred to imbalance dataset due to the rare instances. In this circumstance, the absence of representative information will make learning troublesome with respect to between-class imbalance.

Dataset Complexity

The primary determining factor of classification deterioration is known as dataset complexity which can be enhanced by adding more number of relative imbalance data. The other issues such as lack of representative data, overlapping and small disjunctions are also included in the data complexity.

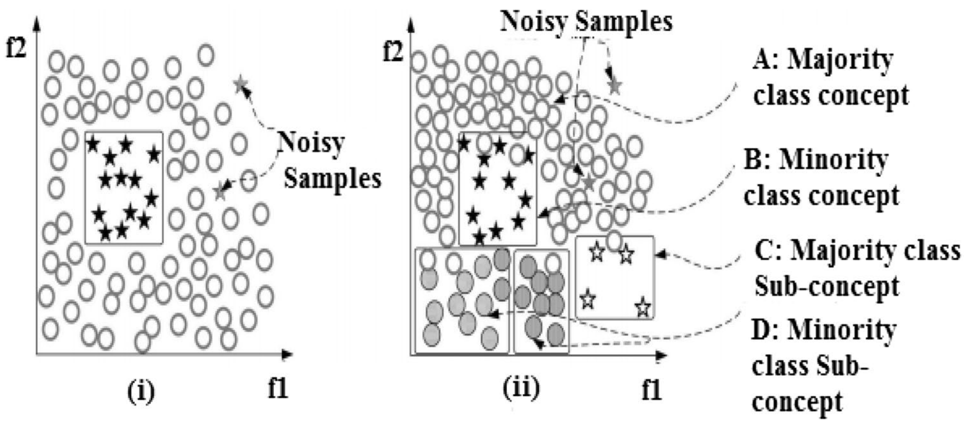
f2
Noisy Samples
(i)
f1
f2
Noisy Samples
A: Majority class concept
B: Minority class concept
C: Majority class Sub-concept
D: Minority class Sub-concept
(ii)
f1

Fig. 1 (i) A dataset with a between-class imbalance. (ii) A high complexity dataset

Consider Fig. 1 [23] as an example of imbalance dataset. The circles and stars in this figure represent the majority and minority classes, respectively.

In this paper, Fig. 1(i), (ii) exhibits relative imbalances which are shown above. However, there are no overlapping samples between the classes and has a single concept related to each class that is described in Fig. 1(i). In addition, Fig. 1(ii) has within- and between-class imbalances, multiple concepts and more overlapping. The sub-concepts of minority and majority classes are also described in Fig. 1(ii).

Imbalance with Small Size Dataset

The last issue is the combination of small size sample data with imbalanced data; moreover, the application of knowledge discovery and data analysis is often unavoidable to have data with high dimensionality. In the pattern recognition filed, the problem of small-sized samples is studied and overcome by the dimensionality reduction methods and various extension methods. This combined problem presents a new challenge to the community when the representative dataset exhibits the formation of imbalances like high-dimensionality data with small size sample data. In that situation, two issues have occurred simultaneously, which are described as follows:

- All of the issues are applicable which are related to absolute rarity and within-class imbalance, once the sample size is small.
- When the dataset presents with imbalance data, learning algorithms fail to generalize the inductive rules over the sample space.

When high-dimensionality data combined with the small size sample data, it delays the learning process due to the formation of conjunctions of limited samples with high-degree features. When the sample space is large enough, a general inductive rule can be characterized for the feature space. The POF arises when samples are limited and the formation of rules becomes too specific. Learning from such data requires more consideration in the community which is relatively a new research topic.

Proposed Methodology

The main aim of the DM learning method is to keep away the dissimilar points while keeping similar points close together in imbalance dataset. The performance of existing classification algorithms is improved by learning proper distance metric, and these distances are computed with the help of the decision-making process in the NN algorithm. Hence, learning a global DM called MDM gets more attention because of its simplicity and efficiency to solve the complicated problems. The main contribution of this research work is to learn the MDM for a small set of samples known as prototypes which are selected by LMDL. The objective function is minimized by adjusting the positions of prototype, and the POF can be reduced by selecting the suitable prototype while preserving the notion of locality.

Given a collection of M training points $x = \left\{\left(x^1, y^1\right), \dots \left(x^M, y^M\right)\right\}$, where $x^M \in \mathbb{R}^{d\times 1}$ and $y^M \in \{1, 2, \dots K\}$ define the corresponding class label, the ultimate goal is to learn a set of MDM, $W = \{W\}_{s=1}^{s}$ where $W^s \in \mathbb{R}_{+}^{d\times d}$ is a positive semi-definite matrix and corresponds to the s^{th} member of a set of randomly selected prototypes $P = \left\{\left(p^1, y^1\right), \dots, (p^s, y^s)\right\}$, $p^s \in \mathbb{R}^{d\times 1}$ and $S << M$. The uppercase or lowercase letters represent the scalars, whereas the boldface uppercase letters describe the matrices and boldface lowercase letters represent the vectors in this setting. Also, to have a compact representation of the parameters, suppose that $W \in \mathbb{R}^{(d\times d)\times s}$ is a matrix in which the sth column of W represents the vectorized form of W^s. Similarly, $P \in \mathbb{R}^{d\times s}$ is a matrix in which sth column of P holds p^s and the ith column of $X \in \mathbb{R}^{d\times M}$ is the ith point in set X.

Using the above notations, the squared MDM between sth prototype and ith point in the input space is given by Eq. 1.

$$d^2_{w^s}\left(x^i, p^s\right) = \left\|x^i - p^s\right\|^2_{w^2} = \left(x^i - p^s\right)^{\mathrm{T}} W^s \left(x^i - p^s\right) \tag{1}$$

where $W^s \in \mathbb{R}_{+}^{d\times d}$ is a symmetric positive semi-definite (PSD) matrix defined on the sth prototype. To minimize the error rate of the NN algorithm, the LMDL method uses the objective function which is a close approximation of the NN's error rate that is shown in Eq. 2.

$$J(W, P) = \frac{1}{M} \sum_{x^i \in X} \mathbb{S}_\beta\left(R\left(x^i\right)\right) \tag{2}$$

where $R\left(x^i\right) = \frac{d^2 W{=}\left(x^i, p^{=}\right)}{d^2 W{\neq}\left(x^i, p^{\neq}\right)}$, $\mathbb{S}_\beta(z) = \frac{1}{1 - \mathrm{e}^{\beta}(1-z)}$ is a sigmoid function and $P^{=}, P^{\neq} \in P$ are the nearest same- and different-class prototypes of x^i as given in Eqs. 3 and 4.

$$P^{=} = \begin{array}{c} \arg\min \\ p \in P \\ \text{class}(p) = \text{class}(x) \end{array} \quad d^2 W = (x, p) \tag{3}$$

$$P^{\neq} = \begin{array}{c} \arg\min \\ p \in P \\ \text{class}(p) \neq \text{class}(x) \end{array} \quad d^2 W \neq (x, p) \tag{4}$$

Accordingly, $W^{=}, W^{\neq} \in w$ are the corresponding Mahalanobis metrics of $P^{=}$ and $P^{\neq}$. The parameter β defines the slope of sigmoid function and if β is large, $S_\beta(z)$ acts like the step function more and more. Based on Eq. 2, the optimization problem can be written as follows in Eq. 5:

$$P^{=} = \underset{\substack{W \in \mathbb{R}^{(d\times d)\times s} \\ P \in \mathbb{R}^{d\times s}}}{\arg\min} \quad J(W,P) \tag{5}$$

The above equation is subject to $W \geq 0, \forall W \in w$ Eq. 5 and is a semi-definite programming with a non-convex objective function. Using the fact that $W \in w$ is a symmetric PSD matrix, it can be factorized as $W = \widetilde{W}\widetilde{W}^T$ where $\widetilde{W} \in \mathbb{R}^{d\times p}$ and $p \leq d$. Hence, Eq. 5 can be change to Eq. 6,

$$\left(\widetilde{W^*, P^*}\right) = \underset{\substack{W \in \mathbb{R}^{(d\times p)\times s} \\ P \in \mathbb{R}^{d\times s}}}{\arg\min} \quad J\left(\widetilde{W}, P\right) \tag{6}$$

where sth column of $\widetilde{W} \in \mathbb{R}^{(d\times p)\times s}$ is the vectorized form of the matrix $\widetilde{W^s} \in \mathbb{R}^{d\times p}$. Algorithm 1 summarizes the iterative gradient learning algorithm based on the above derivatives and similar to the learning procedure. As the algorithm shows, in each iteration, the algorithm visits $x \in X$ and updates those two metrics that have the highest impact in the prediction of sample x and represent by $\widetilde{W}^{=} and \widetilde{W}^{\neq}$. While $P^{\neq}$ gets away from x, the nearest prototypes of same and different class are modified as $P^{=}$ moves toward x.

Algorithm for LMDL

```
    //Input: $X, S$: number of prototypes, β: slope of sigmoid, $\varepsilon$: small constant
    //Output: $W, P$
    Intialize $W \& P$ randomly,
1. Set
2. $W^{new} = W, P^{new} = P, \lambda' = \infty, \lambda = (W, P)$
3. while $(|\lambda' - \lambda| > \varepsilon)$
3.1.   $\lambda' = \lambda$
3.2.   For $x \in X$
3.2.1.     $P^{=} = findNNSameClass(x, \mathrm{P})$
3.2.2.     $P^{\neq} = findNNDiffClass(x, P)$
3.2.3.     $R(x) = d^2 w^{=}(x, P^{=}) / d^2 w^{\neq}(x, \mathrm{P}^{\neq})$
3.2.4.     Calculate $\nabla_{w^{=}} J(w,P), \nabla_{w^{\neq}} J(w,\mathrm{P}), \nabla_{P^{=}} J(w,P), \nabla_{P^{\neq}} J(w,P)$
3.2.5.     $[\alpha_{W^{=}}, \alpha_{W^{\neq}}, \alpha_{P^{=}}, \alpha_{P^{\neq}}] = the\ Adadelta\ rule\ to\ find\ corresponding\ learning$
$rate$
3.2.6.     $w^{=,new} = w^{=} - \alpha_{W^{=}} \odot \nabla_{w^{=}} J(w,P)$
3.2.7.     $w^{\neq,new} = w^{\neq} - \alpha_{W^{\neq}} \odot \nabla_{w^{\neq}} J(w,P)$
3.2.8.     $P^{=,new} = P^{=} - \alpha_{P^{=}} \odot \nabla_{P^{=}} J(w,P)$
3.2.9.     $P^{\neq,new} = P^{\neq} - \alpha_{P^{\neq}} \odot \nabla_{P^{\neq}} J(w,P)$
3.3.   $W = W^{new} \& P^{new} = P$
3.4.   $\lambda = J(W,P)$
```

In order to update parameters for making the decision using NN rules, the method uses Adadelta rule, which is an extension of Adagrad rule and requires no manual tuning of the learning rate. Moreover, Adagrad rule appears robust to various data modalities, noisy gradient information and different model architecture choices. The good learning rate for each iteration of gradient descent is estimated with the help of heuristics algorithm in several attempts. To decrease the learning rate, the parameter updates slow down by using the Adadelta method. The accuracy of the proposed method is improved by the identification of rules by using NN from the DMs.

Experimental Analysis

This section presents the performance of LMDL techniques in terms of standard indices such as accuracy, precision, recall by using the collection of datasets having diverse nature. Moreover, this section presents the influence of other existing methods when compared with the proposed LMDL method on the classification of minority and majority classes.

Dataset Description

The LMDL used standard datasets from the University of California at Irvine (UCI), (http://archive.ics.uci.edu/ml/datasets.html) machine learning repository with various numbers of samples, classes and dimensions. To get the imbalance dataset, randomly delete some negative points or positive points from the UCI datasets. Table 1 provides a brief summary of these datasets.

According to the two following criteria such as degree of imbalance (DI) and scale of the dataset, the data were collected from the database. If the data dimension is greater than 45 or the data points are more than 4000, the dataset is known as a large-scale dataset. All other dataset are considered as small/medium-scale, and this paper used large-scale dataset for classifying the imbalance data. The imbalance ratio (IR) can be used to calculate the DI, which reveals that the ratio of number of points in minority class with majority class for a tow class dataset is defined as IR. In both minority and majority classes, the maximum values are considered based on the calculated IR values in the case of multi-class datasets. According to the IR values, the dataset can be either balanced ($\mathrm{IR} \le 1.15$), partially imbalanced ($1.15 < \mathrm{IR} \le 3.5$) or highly imbalanced ($\mathrm{IR} > 3.5$).

Table 1 Selected UCI dataset for proposed method

Dataset	Samples	Dimensions	Classes
Iris	150	4	3
Breast cancer	685	9	2
Wine	178	13	3
Diabetes	768	8	2
Glass	214	9	6
E-coli	336	7	5
Yeast	1484	8	3

Evaluation Measures

Accuracy is the common parameter for classification that shows the correctness of the function by using the majority class, but it seems insufficient for imbalanced datasets that neglect the minority classes. In this research work, specific metrics such as area under curve (AUC), *G*-mean and *F*-measure are used to evaluate the performance of proposed method which helps to measure in imbalance dataset.

Even though accuracy provides good results in classification by using majority classes, in imbalance dataset, it leads to poor performance because it neglects the minority classes. Hence, this research work uses some other metrics for evaluating the performance on imbalance dataset. Table 2 shows the confusion matrix for binary data which is used to evaluate the classifier.

Multi-class problems are converted into binary case by combining all the majority classes into a negative one. The precision, recall, TP rate and FP rate equation are shown in Eqs. 7–10.

$$\text{precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{7}$$

$$\text{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{8}$$

$$\mathrm{TP}_{\text{rate}} = \frac{\mathrm{TP}}{\text{Total } P} \tag{9}$$

$$\mathrm{FP}_{\text{rate}} = \frac{\mathrm{FP}}{\text{Total } N} \tag{10}$$

These measures are aggregated into *F*-measure which is described in Eq. 11:

$$\text{F-Measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{11}$$

Table 2 Confusion matrix

	Predicted positive	Predicted negative
Actual positive	True positive	False positive
Actual negative	False negative	True negative

Most of the studies on imbalanced data concentrate only on F-measure because of using minority class only. But, this research paper concentrates on G-mean which can show a trade-off between the recognition of both majority and minority classes. It is defined as Eq. 12.

$$\text{G-Mean} = \sqrt{\text{recall} \times \text{precision}} \tag{12}$$

The AUC and accuracy can be described in Eqs. 13 and 14

$$\text{AUC} = \frac{1 + \text{TP}_{\text{rate}} - \text{FP}_{\text{rate}}}{2} \tag{13}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{14}$$

Empirical Result

This is difficult to preciously measure the performance for the imbalanced data especially when the data are small. The distribution of instances varies between training, testing and validation sets due to the small number of instances and imbalance data. This leads to performance evaluation of improper setting and sometimes it provides poor performance. Hence, the following processes were divided and fed into the datasets for estimating better performance by setting the proper parameters. Table 3 represents the performance of the proposed method in terms of accuracy, F-measure, AUC, G-mean, precision and recall for different datasets. The graphical representation of the performance of various parameters is shown in Fig. 2.

From the above table, the experimental results stated that the LMDL method achieved 97% accuracy, 98% precision, 97% for both F-measure and recall, 97.36% G-mean and 90.62% AUC for Iris dataset, whereas the LMDL method provides poor performance in yeast when compared to all other datasets. For yeast, the LMDL achieved 56.87% accuracy, 57% of precision and recall, 55% F-measure, 50% AUC and 69.58% G-mean. The LMDL method achieved nearly 75% in three datasets like wine, diabetes and glass datasets for all parameters such as precision, recall, accuracy, F-measure, AUC, G-mean and accuracy.

Comparative Analysis

In this section, the results obtained by the LMDL are compared with existing methods such as random forest (RF) and synthetic minority over-sampling technique (SMO) in Sun et al. [17], LEAT [21] and fuzzy adaptive K-NN by Patel and Thakur [22] for three parameters such as F-measure, G-mean and AUC in E-coli, yeast, glass and wine datasets. The performance of the LMDL for four datasets in different parameters is discussed in Tables 4 and 5.

Table 3 Performance of various parameters of proposed method

Dataset	Accuracy	Precision	Recall	F-measure	G-mean	AUC
Iris	97.36	98	97	97	97.36	90.62
Breast cancer	92.30	93	92	92	90.82	91.064
Wine	71.11	72	71	71	71	67.35
Diabetes	76.04	76	76	75	69.11	71.38
Glass	70.37	67	70	68	73.31	100
E-coli	80.95	80	81	79	88.78	83.64
Yeast	56.87	57	57	55	69.58	50.0

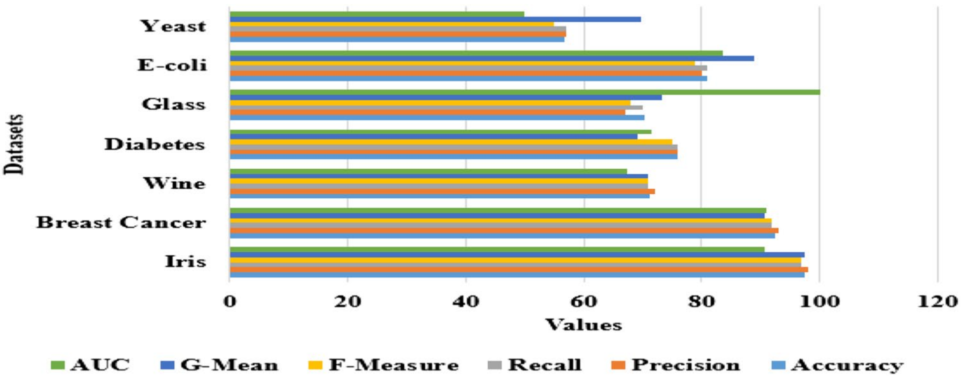

Fig. 2 Parameter performance of proposed method

Table 4 Comparison of performance of proposed method

Author	Datasets	Methodology	*F*-measure	*G*-mean	AUC
Patel and Thakur [22]	Yeast	Fuzzy adaptive	78.57	85.02	85.95
	E-coli	K-NN	68.42	74.10	76.83
Proposed methodology	Wine		75.0	23.17	51.95
	Glass	LMDL	68	73.31	100
	E-coli		79	88.78	83.64
	Wine		71	67.35	61.40
	Yeast		55	69.58	50.0

Table 5 Comparison of performance of proposed method for AUC

Dataset	Author	Methodology	AUC
	Park and Ghosh [21]	LEAT	80
E-coli	Proposed methodology	LMDL	83.64
	Sun et al. [17]	RF	75.94
	Park and Ghosh [21]	SMO	55.71
Glass	Proposed methodology	LEAT	98
		LMDL	100

From Tables 4 and 5, it is clear that the performance of proposed LMDL method achieved better results in all datasets. Though the proposed LMDL achieved 55% *F*-measure in yeast dataset, the LMDL method achieved 69.58% *G*-mean when compared with existing method by Patel and Thakur [22]. When compared to the existing methods, the LMDL method achieved 100% AUC in glass dataset, whereas it provides poor performance in yeast dataset because of using high nonlinear data. In E-coli dataset, the LMDL method achieved 79% *F*-measure, 88,078% *G*-mean and 83.64% AUC when compared with the existing method like fuzzy adaptive K-NN.

Conclusion

Many real-world applications are affected by the imbalance data, where the data distribution is uneven. In this work, a LMDL is used for enhancing the performance of the K-NN algorithm in which the similarity of local points is enlarged and local dissimilar points are reduced. The influence of neighborhood is considered by the LMDL, and local discrimination is increased by Mahalanobis metric distance for each prototype which is learned from the proposed LMDL method. In order to adjust the prototype's positions and metrics, the LMDL used an objective function that is closely related to NN error rate. A variety of experiments have been performed on both real-world and synthetic datasets, and the results demonstrate that the proposed LMDL method performed well when compared with the methods such as LEAT, RF, SMO and adaptive K-NN. The developed method is proven to solve the imbalance dataset problem and has the higher efficiency in the classification. The results showed that the proposed LMDL method achieved nearly 98% in Iris dataset, 93% in breast cancer dataset and 80% in E-coli dataset for all metrics used in this research work. The proposed LMDL provided poor classification performance in case of high nonlinear data. The future work of this method is that this can be extended to feature-based NN and can be applied to the multi-classification of nonlinear data.

References

1. Maalouf M, Trafalis TB. Robust weighted kernel logistic regression in imbalanced and rare events data. Comput Stat Data Anal. 2011;55(1):168–83.
2. Olszewski D. A probabilistic approach to fraud detection in telecommunications. Knowl-Based Syst. 2012;26:246–58.
3. Haixiang G, Shang YL, Mingyun J, Yuanyue GH, Bing G. Learning from class-imbalanced data: review of methods and applications. Expert Syst Appl. 2017;73:220–39.
4. Yijing L, Haixiang G, Xiao L, Yanan L, Jinling L. Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data. Knowl-Based Syst. 2016;94:88–104.
5. López V, Fernández A, García S, Palade V, Herrera F. An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. Inf Sci. 2013;250:113–41.

6. López V, del Río S, Benítez JM, Herrera F. Cost-sensitive linguistic fuzzy rule based classification systems under the MapReduce framework for imbalanced big data. Fuzzy Sets Syst. 2015;258:5–38.
7. Chawla NV, Japkowicz N, Kotcz A. Editorial: special issue on learning from imbalanced data sets. SIGKDD Explor Newsl. 2004;6:1–6.
8. Han H, Wang WY, Mao BH. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: International conference on intelligent computing. Springer, Berlin, Heidelberg, 2005.
9. Diamantini C, Potena D. Bayes vector quantizer for class-imbalance problem. IEEE Trans Knowl Data Eng. 2009;21(5):638–51.
10. Ertekin S, Huang J, Lee Giles C. Active learning for class imbalance problem. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, ACM (2007).
11. Hong X, Chen S, Harris CJ. A kernel-based two-class classifier for imbalanced data sets. IEEE Trans Neural Netw. 2007;18(1):28–41.
12. Oh S, Lee MS, Zhang BT. Ensemble learning with active example selection for imbalanced biomedical data classification. IEEE/ACM Trans Comput Biol Bioinf. 2011;8(2):316–25.
13. Lusa L. Class prediction for high-dimensional class-imbalanced data. BMC Bioinform. 2010;11(1):523.
14. Wasikowski M, Chen X. Combating the small sample class imbalance problem using feature selection. IEEE Trans Knowl Data Eng. 2010;22(10):1388–400.
15. Yu H, Ni J, Dan Y, Xu S. Mining and integrating reliable decision rules for imbalanced cancer gene expression data sets. Tsinghua Sci Technol. 2012;17(6):666–73.
16. Wang F, Sun J. Survey on distance metric learning and dimensionality reduction in data mining. Data Min Knowl Disc. 2015;29(2):534–64.
17. Sun Z, Song Q, Zhu X, Sun H, Xu B, Zhou Y. A novel ensemble method for classifying imbalanced data. Pattern Recognit. 2015;48(5):1623–37.
18. Napierała K, Stefanowski J. Addressing imbalanced data with argument based rule learning. Expert Syst Appl. 2015;42(24):9468–81.
19. Ohsaki M, Wang P, Matsuda K, Katagiri S, Watanabe H, Ralescu A. Confusion-matrix-based kernel logistic regression for imbalanced data classification. IEEE Trans Knowl Data Eng. 2017;29(9):1806–19.
20. Krawczyk B, Woźniak M, Schaefer G. Cost-sensitive decision tree ensembles for effective imbalanced classification. Appl Soft Comput. 2017;14:554–62.
21. Park Y, Ghosh J. Ensembles of α-trees for imbalanced classification problems. IEEE Trans Knowl Data Eng. 2014;26(1):131–43.
22. Patel H, Thakur GS. An improved fuzzy K-nearest neighbor algorithm for imbalanced data using adaptive approach. IETE J Res. 2019;65(6):1–10.
23. He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2009;21(9):1263–84.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.