International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Adaptive Condensed Nearest Neighbor for Imbalance Data Classification

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

[1]*Department of Computer Science and Engineering, Visvesvaraya Technological University, India*
* Corresponding author's Email: nijagunags@gmail.com

Abstract: Classification over numerous real-world datasets has a peculiar drawback called unstructured class problem. A dataset is said to be unstructured when the majority of the class has more samples insignificantly than the minor class. Such drawbacks result in an ineffective performance of data classification techniques. Classification is a supervised learning method which acquires a training dataset to form its model for classifying unseen examples. However, in unstructured data classification, the class boundary learned by standard machine learning algorithms can be severely skewed toward the target class. As a result, the false-negative rate can be excessively high. The researches focus on the unstructured data classification using uncertain Nearest Neighbor (NN) decision rule and also found the major issues face by k-Nearest Neighbor (k-NN). In any case, given a dataset, prediction of accuracy is a monotonous task to improve the execution of kNN by tuning k. Because of class imbalance, the performance of kNN decreases and this situation is represented by dissimilar characteristic from various classes. This paper addresses the issues faced by kNN by developing Adaptive-Condensed NN (Ada-CNN). The Ada-CNN classifier utilizes the distribution and density of test point's neighborhood and learn an appropriate point-explicit k by using artificial neural systems. Ada-CNN performed well compared to kNN and other well-known classifiers. The experimental results showed that Ada-CNN achieved nearly 94% accuracy for Diabetes dataset and 100% accuracy in pop-failure compared to kNN for imbalanced classification.

Keywords: Adaptive-condensed nearest neighbor, Decision rule, Imbalance data, k-Nearest neighbor, Unstructured data.

1. Introduction

A data set is imbalanced if the examples of one class outnumber those from the others. Generally, the issue of imbalance is regularly experienced by various machine-learning applications such as content arrangement, speech acknowledgment, programming imperfection detection [1 - 3], bioinformatics, and biomedical decision-making [4]. The skew class distribution of imbalance information prevents the execution of most standard classification algorithms that function on information collections with even class dissemination, namely Naive Bayes, IB1, C4.5 [5], Logistic Regression [6], Neural Networks and Support Vector Machines [7]. Subsequently, the imbalance information issue has attracted much consideration of the legitimate machine learning, data mining. So, various imbalance information-managing strategies proposed in the algorithmic level and data levels. In an imbalance issue, the most obvious characteristic is the skewed class distribution. By the way, a number of studies show that the skewed information distribution isn't the main factor that impacts the execution of an existing algorithm in distinguishing the rare events. At the same time, high dimensionality, small size sample and the issue of unpredictability will ruin the learning execution, because it is hard to assemble a good classification method over the high level of features with constrained examples. The imbalance data are small in size and high dimensionality especially on the minority class. The features lack of the discriminant ability and further lead to the poor performance in classification. But, the existing methods pay more consideration for re-adjusting the skewed class dispersion or algorithm adaption.

However, the method concentrates less on how to enhance the segregate capacity of features in the imbalance information indexes [8, 9]. In order to develop a superior classification model for imbalance learning issues, it is important to develop new features with high discriminant capacity rather than unique features.

The kNN classifier has been preferred for its systematic simplicity, nonparametric working standard and simplicity of usage. The kNN classifier includes a tuning of single parameter k (the quantity of NN to be considered). It is difficult to discover the estimation of k for which the algorithm performs ideally on an extensive variety of information collections (or for each point in similar informational index). Based on the informational index, decisions made rather than $k = 1$ might be more reasonable [10]. To optimize the performance, kNN keeps on running with various diverse k esteems. In this way, a few methods namely cross approval and probabilistic estimation [11, 12], might be utilized to pick the best k esteem among the tested k esteems. While probabilistic modeling-based algorithms usually depend on prior assumptions about the data set and hard to implement, the procedure of cross approval is computationally costly. In this paper, these facts motivated to choose an information point-specific k esteem utilizing a pointer of the density and class distributions of its local neighborhood. Besides the difficulty with the determination of k , kNN characterization rule additionally faces a challenge over the informational indexes with imbalance class, i.e., each class doesn’t have an identical number of representatives [13]. Subsequently, the paper also presents a class-explicit global weighting plan to handle the issue of class imbalance. **Novelty of the work**: The main contributions of this paper are to facilitate the choice of data-point-specific k, the proposed Ada-CNN method first finds each of the training points, a value of k for which CNN can accurately classify that point. Assuming this k value to approximate the local information about the neighborhood of a data point, this algorithm tries to estimate a suitable k value for each query point. The value of k gives rise to a nonlinear regression problem, which can be solved by a feedforward Multi-Layer Perceptron (MLP) using the Scaled Conjugate Gradient (SCG) learning algorithm to estimate the k-terrain. The proposed method is compared with existing methods, in that the Ada-CNN performed well on both small-and medium-scale data sets. A number of experiments performed on proposed method using UCI data sets having varying degrees of imbalance and compared the proposed algorithms with the existing methods for data imbalance classification. The proposed methods achieved competitive performance compared to the existing algorithms. In that Ada-CNN became the best performer in most cases.

The rest of this paper includes: Section 2, a brief but comprehensive review of the notable works done in the field is presented. Section 3 and 4, presented the problem and solutions of imbalance data classification. In Section 5, the paper described the proposed methodology for Ada-CNN to handle imbalance. Section 6 discussed about the selection of the data sets, experimental results and a comparative discussion of the proposed methods along with existing methods. Then, the paper presents the conclusion and the future work of this research in Section 7.

2. Literature review

In this section, the paper briefly described the notable past works by incorporating information about the neighborhood of a test point that make it suitable for handling class imbalance.

Y. Zhu, Z. Wang, and D. Gao [14] proposed new model based on the traditional Gravitational Fixed Radius NN (GFRNN) methodology. The GFRNN strategy was executed to work on the classification issues in imbalanced datasets. Especially, GFRNN does not require any manual set parameters in the entire procedure. By using FRNN rule, the GFRNN first chose the proper patterns out as the candidate, and afterward determine the gravitational energy between the query designs of every candidate. The experimental consequence of this paper presumed that the powerful and fundamentally basic NN learning proposed method dealt with the imbalanced acknowledgment assignments tasks more adequately and effectively. However, GFRNN spends more time when the size of the dataset becomes larger because the computation complexity of the basic FRNN is dataset-dependent.

S. Vluymans, I. Triguero, C. Cornelis, Y. Saeys, [15] built up a hybrid strategy to explicitly deal with class imbalance, called Evolutionary Prototype Reduction Based Ensemble for NN Imbalanced Data (EPRENNID). To prevent the method from the problem of overfitting of training set, this method performed an EPR that provides diverse solutions. The technique additionally enabled the procedure to lessen the under-represented class, which is the most widely recognized as preprocessing for classifying the class imbalance. The preprocessing step yielded various model sets that were later utilized in an ensemble, playing a weighted voting scheme with the NN classifier. The experimental results showed that the EPRENNID outperformed the existing strategies. The classification time of EPRENNID is marginally higher than that of other models, which was due to the target specific weight construction of the prototype sets in the ensemble.

K. Jiang, J. Lu, and K. Xia, [16] implemented a novel Genetic Algorithm-based Synthetic Minority Over-Sampling Technique (GASMOTE) method. The GASMOTE method utilized distinctive testing rates for various minority class instances and found the combination of ideal examining rates. This paper applied the GASMOTE method to a practical application like forecast of rockburst in the VCR rockburst datasets. The results showed that the GASMOTE method precisely anticipates the rockburst event and hence gave the direction to the design and development of safe deep mining engineering structures. The practicability and application scope of these proposed indexes is substantially limited as it is very difficult to accurately measure or calculate the strength of theory-based stress criteria, energy criterion.

Y. Xu, Y. Zhang, J. Zhao, Z. Yang, and X. Pan, [17] introduced a KNN based Maximum Margin and Minimum Volume Hyper-sphere machine (KNN-M3VHM). The two hyper-spheres were found by KNN-M3VHM, in that each hyper-sphere volumes were as small as possible and contained as many samples in one class. This strategy amplified the edge between two classes to pursue the maximum margin standard. KNN-M3VHM considered the characteristic of both margins that is within-class and between-class. The outcomes expressed that the proposed technique worked well other than the existing strategies for most of the cases. But, this technique gave poor accuracy if the imbalanced datasets were not preprocessed by testing strategies.

S. S. Mullick, S. Datta, and S. Das [18] developed a strategy called Adaptive KNN (Ada-KNN) for enhancing the execution of KNN in the presence of class imbalance. The Ada-kNN classifier utilized the thickness and dissemination of the test point neighborhood and learned a reasonable point-explicit k for classification with the assistance of artificial neural systems. The strategy further enhanced by replacing the neural system with a heuristic learning technique guided by an indicator of the neighborhood thickness of a test point. The proposed technique preserved the straightforwardness of kNN without causing serious computational weight, called as Ada-kNN2. Besides, this strategy proposed a class-based Global Weighting Scheme (Global Imbalance Handling Scheme or GIHS) to compensate the impact of imbalanced data. In this method, the distinctive distance measure influences the execution of Ada-KNN and Ada-KNN2 in the case of high dimensional datasets.

M. Ohsaki, P. Wang, K. Matsuda, S. Katagiri, H. Watanabe, and A. Ralescu, [19] developed a Confusion-Matrix-based Kernel LOGistic Regression (CM-KLOGR) to achieve a better classification performance by forming the dataset without task dependence and heuristics. 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. The ensemble methods and cost-sensitive were not used in this method, because those methods need task dependent and heuristic process.

To overcome the above issues of the existing works, this research work implemented the Ada-CNN to address the problem of class imbalance.

3. Problem of imbalance data

If class dispersion among classes in dataset is not uniform, it is said to be an imbalance dataset. In this condition, minority class represents at least one class in dataset, whereas majority class is described as rest of data in these dataset. The recent research showed that uneven distribution of class examples used in the learning process could leave these algorithms with performance bias. It implies that classifier gives high precision on the major share class. However, it gives poor accuracy on the minority class. This is because the conventional preparing criteria, for example, the general achievement can be incredibly affected by the larger number of cases from the dominant part class. In many real world problems, the minority classes play an important role for accurately classifying examples from this class. Scientists have recognized information imbalance issue into two fundamental composes: Binary class information imbalance and multi class information imbalance [20].

3.1 Binary class data imbalance dataset

A binary dataset consists of only two classes. If a class exists in the binary dataset, which is represented by only a few numbers of samples, then it is called binary class data imbalance problem. Zero class thresholds are used for separating the two classes in binary class dataset. Hence, the boundaries of classes are no need to identify in dataset.

3.2 Multi class data imbalance dataset

The dataset contains more than two classes called as multiclass dataset. The information imbalance issue makes extra overheads in multiclass dataset. Straightforward and proficient zero class edges can’t be utilized as a part of multiclass dataset. Complex techniques like Static Search Selection or Dynamic Search Selection should be utilized to overcome the imbalance problem. The multiclass issue should have been partitioned into numerous paired class issues for sometimes to order the dataset.

4. Solution for imbalance problem

The problem of imbalanced classification can be solved by using following four main types of solutions.

4.1 Sampling (solutions at the data level)

This type of arrangement comprises of adjusting the class conveyance by methods for a preprocessing procedure [21, 22]. Processes at information level are partitioned into 3 groups,

- **Under sampling methods:** It makes a subset of the original informational index by wiping out some portion of the instances of the majority class.
- **Oversampling methods:** It makes a superset of the first informational index by duplicating a portion of the instances of the minority class or making new minority cases, for example by insertion of unique instances.
- **Hybrid methods:** It consolidates two above techniques by decreasing the measure of the larger part of the class and expand the quantity of minority component.

The main advantage of the information level methodologies is that their utilization is independent of the classifier chosen [23].

4.2 Structure of explicit algorithms (solutions at the algorithmic dimension)

For this situation, a traditional classifier is adjusted to bargain specifically with the imbalance between the classes [24, 25], for instance, altering the expense per class [26] or changing the probability estimation in a decision tree to support the positive class.

4.3 Cost-sensitive solutions

These types of strategies fuse solutions at the information level, at algorithmic dimension, or at the two dimensions together, that endeavor to limit greater expense errors. Let $C(+,-)$ signify the expense of misclassifying a positive (minority class) occurrence as a negative (dominant part class) case and $C(-,+)$ the expense of the converse case. The method imposes $C(+,-) > C(-,+)$, i.e., the expense of misclassifying a positive occurrence should be higher than the expense of misclassifying a negative one [27 - 29].

4.4 Ensemble solutions

Ensemble strategies for imbalanced classification comprise a combination of ensemble learning algorithms and these methods are explicitly information level and cost-delicate. Through the expansion of an information level approach to deal with the ensemble learning technique, the new hybrid technique typically preprocesses the information before preparing every classifier. Rather than adjusting the base classifier in order to accept costs in the learning procedure, the cost-delicate gatherings manage the cost minimization by means of the ensemble learning algorithm [30].

5. Proposed methodology

Assigning different misclassification costs to incorrect class predictions [31] or developing improved training criteria that are more sensitive to the unbalanced class distributions are presented in common approaches compared to the standard overall error rate or accuracy. Enhanced learning criteria incorporate the normal classification accuracy of the minority and larger part classes. In wide-ranging approaches, only few strategies have been proposed by specialists. One of the well-known strategies which will look into the information imbalance issue is accomplished by utilizing an approach namely NN decision rule. Fig. 1 demonstrates the basic structure of the model of Ada-CNN.

5.1 Imbalance dataset

Information level methodologies work by resampling the training instances with a specific end goal to accomplish a more balanced dataset. This is finished by either over-sampling the minority classes' examples, under sampling the larger part classes' examples, or applying half of models which

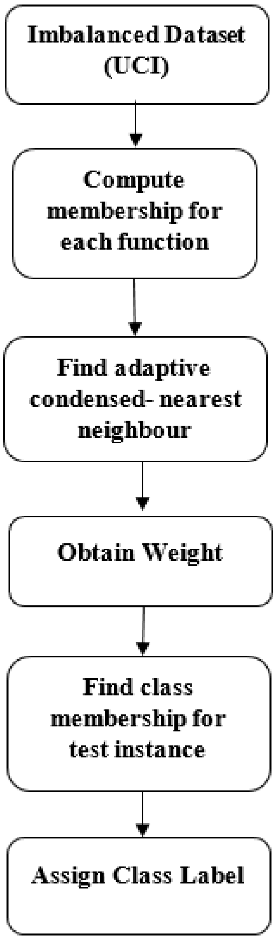
Imbalanced Dataset (UCI)
Compute membership for each function
Find adaptive condensed- nearest neighbour
Obtain Weight
Find class membership for test instance
Assign Class Label

Figure. 1 Basic structure of the proposed method

are a mix of over-testing and under sampling methods. Such strategies are considered as preprocessing approaches for managing the class imbalance issue. In conclusion, the effect of a resampling strategy on the classification task correspondingly relies on the dataset. Sometimes, picking the wrong resampling system may adversely influence the classification of unstructured data [32].

5.2 Computing membership function

While thinking about a classification issue, if the earlier probabilities and the state restrictive densities of all classes are known, the Bayesian decision hypothesis deliver the ideal outcomes which reduces the expected misclassification rate [33]. Under this situation, numerous non-Bayesian grouping methods such as clustering and discriminant analysis are based on the notion of the distance or similarity in the feature space that describes the observations.

5.3 Adaptive-condensed nearest neighbour

After computing the membership function, a semantic approach using Ada-CNN approaches are utilized for identifying imbalance data. Ada-CNN data reduction is used to summarize the training set, finding the most important observations of information, which will be used to classify any imbalance data. The selection of Ada-CNN data will reduce the number of comparisons, which will automatically classify a new observation with a slight reduction in accuracy.

The manner in which the algorithm works is to isolate the information focuses into 3 different types:

- **Outliers:** The points, which would not be perceived as the right kind whenever added to the database later.
- **Prototypes:** The minimum arrangement of points required in the training set for the various non-exception points to be effectively perceived.
- **Absorbed points**: The points which are not anomalies, and would be accurately perceived based on the arrangement of prototype points.

At that point, the strategy needs to contrast imbalance information with the prototype points. The algorithm to do this can be outlined as:

1. Go through the preparation set, removing each point, and checking whether it is perceived as the right class or not. If the point, set it back in the set. Or if it is an outlier, and should not be returned.
2. Make another database, and include an arbitrary point.
3. Pick any point from the original set, and check whether it is perceived as the right class dependent on the point in the new database, utilizing kNN with $k = 1$. If the point is in recognized class, then it is an assimilated point, and can be let alone for the new database. If not, it ought to be expelled from the original set, and added to the new database of models.
4. Proceed through the original set this way.
5. Repeat stages 3 and 4 until the point when no new models are included.

Since CNN has to keep repeating, this algorithm can take a long time to run as depicted in [34]. To improve the running time of this algorithm, the researchers solved this problem by using extended techniques. However, when it has been run, the kNN calculation will be much quicker. CNN is also influenced by noise in the preparation set. In order to examine this, three informational indexes were created, and for everyone, the quantity of noise point was expanded, and CNN runs and record the level of points appointed to each type. As it can easily expand, the number of arbitrary noise points influenced the results of the CNN algorithm in three principle ways:

- The level of points classed as exceptions expanded drastically.
- The level of points classed as retained diminished.
- The level of points classed as prototypes expanded slightly.

These ways might be normal. The level of exceptions increases because of the fact that there are ever more noise points in each bunch, which will lead them to be misarranged. The level of points regarded as prototype increase because the data collection presently has a substantially more unpredictable structure once the data incorporated all these irregular noise points. The level of absorbed points in this manner must decline as the other two types are expanding (and the three are fundamentally unrelated).

5.4 Handling imbalance in a dataset

The class-specific weighted CNN classifier can be proved to be efficient in situations plagued by class imbalance. The class-specific weights can be used to magnify the increments in the number of members from the minority classes in the neighborhood of a test point while diminishing those majority classes to compensate for their abundance.

One way to weigh the classes is to use a global class weighting scheme, which uses the same set of class weights for all test points. Since there should ideally be equal number of representatives from each class in P, the ideal probability of a point belonging to class $c \in C$ should be $r = (1/C)$. In order to have a better chance for each class in the neighborhood of a test point irrespective of class imbalance, the method assigns the ratio of the ideal and current probabilities for a class c as the global weight w_c, associated with that class which is given in the Eq. (1), i.e.,

$$w_c = \frac{r}{p_c} \qquad (1)$$

Where, p_c is the prior probability of class c.

5.5 Assigning class label

With the Ada-CNN calculation, class names of the k learning instances nearest to a testing instance, helps to decide the class name of the test example. Opposite separation weighting is to measure the vote of each neighbor as indicated by the backwardness of its separation from the test occurrence. By taking the weighted normal of the neighbors closest to the test instance smoothest out the effect of disconnected boisterous preparing occurrences. Besides, it lifts the heaviness of the instance.

6. Experimental result

In this section, the performance of Ada-CNN techniques is evaluated in terms of the standard indices using a collection of data sets having diverse nature.

6.1 Description of the data sets

The proposed method used standard data sets from the University of California at Irvine (UCI), (http://archive.ics.uci.edu/ml/datasets.html) machine learning repository. To get the imbalance dataset, randomly delete some negative points or positive points from the UCI datasets. Here, the method used some of the UCI datasets such as Haberman, Pop-Failure, Cancer and Diabets datasets. According to the two following criteria, the data were collected from the database.

6.1.1. Scale of data set

If a dataset contains more than 4000 data points and/or has a data dimension greater than 45, then data are called as large-scale which is used in this paper. All the other data sets are considered as small/medium-scale.

6.1.2. Degree of imbalance

This can be quantified in the form of Imbalance Ratio (IR). IR is defined as the ratio of the number of points in the majority class with the minority class for a two class data set. In the case of multi-class data sets, IR is taken to be the maximum values of IR calculated values between all the pairs of both minority and majority classes. Based on IR value, a data set can be either balanced (IR ≤ 1.15),

Table 1. Category of dataset

Degree of Imbalance	Scale of Datasets	
	Small/Medium Scaled Datasets	Large Scaled Datasets
Balanced Datasets	Category 1	Category 2
Mildly Imbalanced Datasets	Category 3	Category 4
Highly Imbalanced Datasets	Category 5	Category 6

mildly imbalanced (1.15 < IR ≤ 3.5), or highly imbalanced (IR > 3.5). A data set can be placed in any one of the six possible categories, based on the two above-mentioned criteria, as described in Table 1.

6.2 Performance of evaluation metrics

In this experimental analysis, the proposed Ada-CNN performance is compared with the other existing methods, in terms of Accuracy, Specificity (Spec), Positive Predictive Value (PPV), Negative Predictive Value (NPV), and Sensitivity (Sen). The estimation has been done for these parameters using TP, FP, FN, and TN values, where TP refers to True Positive, TN is True Negative, FP is False Positive and FN is False Negative. The calculation of parameters is described below.

6.2.1. Accuracy

Accuracy is the most instinctive execution measure and it is a proportion of effectively anticipated perception to the aggregate perceptions. The accuracy is directly proportional to true results by considering both true positives and true negatives between the aggregate number of cases investigated. The parameter of accuracy is calculated in Eq. (2),

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \times 100 \tag{2}$$

6.2.2. Specificity

This measures the proportion of negatives that are correctly identified which is described in Eq. (3).

$$Specificity = \frac{TN}{(TN+FP)} \tag{3}$$

6.2.3. Sensitivity

The sensitivity calculates the ratio of positives that are correctly recognized by samples. The mathematical equation of sensitivity is described in Eq. (4).

$$Sensitivity = \frac{TP}{(TP+FN)} \tag{4}$$

The general formula for PPV and NPV are explained in following Eq. (5) and (6)

$$PPV = \frac{TP}{TP+FP} \times 100 \tag{5}$$

$$NPV = \frac{TN}{TN+FP} \times 100 \tag{6}$$

The Table 2 represents the performance of the proposed method in terms of accuracy, sensitivity, specificity, PPV and NPV for different datasets such as Diabetes, Cancer, Haberman and Pop failure. The graphical representation of the performance of various parameters is shown in Fig. 2.

An inspection of the results of the small/medium scale data sets reveals that Ada-CNN achieved the best mean accuracy and also achieved better consistency and scalability. The proposed Ada-CNN achieved 96.581% PPV in Cancer dataset, but only 12.5% NPV in Habernan dataset. But, Ada-CNN achieved 100% accuracy, but achieved only 55.55% sensitivity in pop-failure dataset. On the other hand, Ada-CNN exhibits lower competence on the large scale data sets, which shows the low average accuracy. Ada-CNN failed to achieve good performance on the large-scale data sets, possibly due to insufficient learning of the k-terrain by the underlying MLP, leading to erroneous prediction of the k_{y_i} values.

6.3. Comparative analysis

In this area, the outcomes acquired by the Ada-CNN are compared with CMKLOGR [19], KLOGR [19], and SVM [19] discussed here. The target function of the existing techniques was the harmonic mean of different assessment criteria got from a confusion matrix, such evaluations are sensitivity, positive predictive esteem, and others for negatives. This target function and its optimization were reliably figured on the system of KLOGR, based on Minimum Classification Error and Generalized Probabilistic Descent (MCE/GPD) learning. The accuracy of the existing method can be calculated from the Harmonic Mean (HM) of sensitivity, specificity and PPV. The performance of the Ada-CNN for two datasets such as Huberman and pop-failure are discussed in Table 3.

Table 2. Performance of Ada-CNN method

Dataset	Accuracy (k=1)	Sensitivity	Specificity	PPV	NPV
Diabetes	93.2291	89.34426	45.714	74.149	71.11
Cancer	95.0	96.58	93.103	96.581	93.103
Haberman	89.6103	81.0810	66.66	98.36	12.5
Pop failure	100	55.555	99.206	83.333	96.899

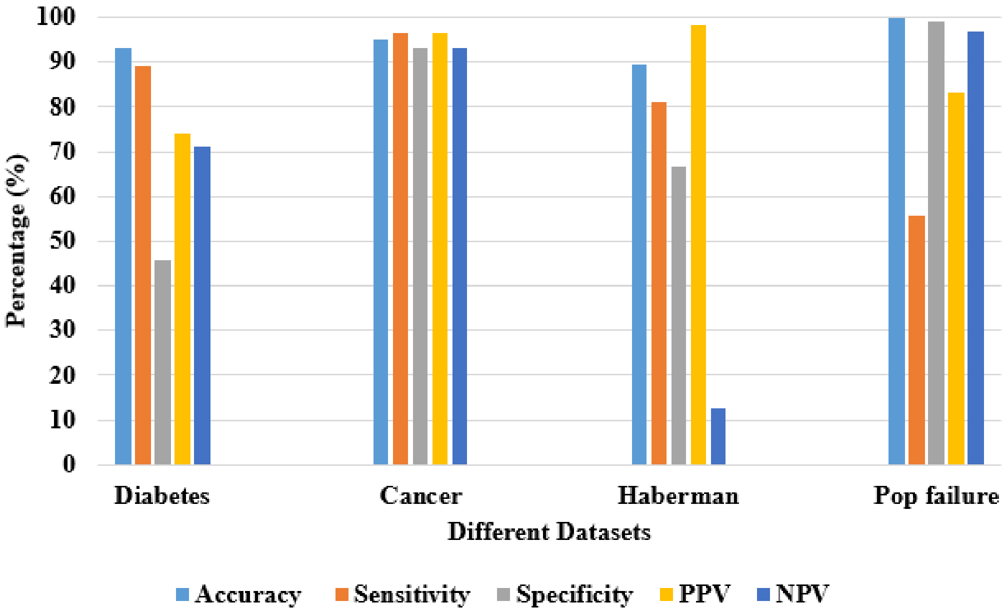

Figure. 2 Performance of Ada-CNN

Table 3. Performance of the proposed method

Datasets	Author	Methodology	Sensitivity	Specificity	PPV	HM
Huberman		CMKLOGR [19]	75	82.61	60	72.53
		KLOGR [19]	75	78.26	54.55	69.27
		SVM [19]	50	78.26	44.44	57.56
	Proposed	Ada-CNN	81.08	66.66	98.36	82.03
Pop-failure		CMKLOGR [19]	100	81.63	35.71	72.44
		KLOGR [19]	100	93.88	62.50	85.46
		SVM [19]	100	93.88	62.50	85.46
	Proposed	Ada-CNN	55.55	99.20	83.33	79.36

Ada-CNN ranked best, and this suggests that proposed method has a higher potential to maximize its performance than CMKLOGR, KLOGR and SVM. The HM of the proposed achieved nearly 83% for Huberman dataset, but the Ada-CNN achieved only 79.36% for pop-failure dataset. Compared to the existing methods like KLOGR and SVM for pop-failure dataset, the Ada-CNN method achieved less HM, because of its poor performance in sensitivity. From the above results it can be concluded that Ada-CNN outperformed its competitors (CMKLOGR, KLOGR and SVM with and without under/oversampling methods) in many conditions. Specifically, Ada-CNN is more effective to raise the harmonic mean of specificity, sensitivity and PPV. Considering the results, Ada-CNN worked effectively and it is flexible based on the prioritized evaluation criteria such as Sens, Spec, PPV, and NPV, or two of them.

7. Conclusion

In this section, the paper summarizes the different key discoveries of the proposed technique and present the concluding comments. The proposed adaptive decision of k values for the kNN classifier can turn out to be exceptionally helpful for information classification over the conventional global decisions, (for example, $k = 1$). This is because such versatile decision of k can represent the properties of neighborhood's test point, which may remain ignored by a single global decision of k. The advantage of this methodology is clear from the reliable execution of Ada-CNN, moreover, the execution of Ada-CNN is seen to be strong in IR. The experimental results showed that the proposed Ada-CNN achieved 93.22% accuracy and 89.34% sensitivity for diabetes datasets, whereas 100% accuracy and 55.55% sensitivity for pop failure dataset. The MLP-based Ada-CNN faced the adaptability issues of large scale dataset because of choosing k values adaptively. A future work of this paper might enhance the proposed heuristic learning strategy to limit the versatility issues by using index pointer. Another future direction of the proposed strategy is to enhance the weighting scheme for reducing time complexity.

References

[1] Y. Jiang, B. Cukic, and Y. Ma, "Techniques for evaluating fault prediction models", *Empirical Software Engineering*, Vol.13, No.5, pp.561-595, 2008.

[2] T. M. Khoshgoftaar and K. Gao, "Feature selection with imbalanced data for software defect prediction", In: *Proc. of IEEE International Conf. on. Machine Learning and Applications*, 2009.

[3] L. Pelayo and S. Dick, "Applying novel resampling strategies to software defect prediction", *Fuzzy Information Processing Society, IEEE Annual Meeting of the North American*, 2007.

[4] P.W. Novianti, K.C.B. Roes, and M.J.C. Eijkemans, "Evaluation of gene expression classification studies: factors associated with classification performance", *PloS One*, Vol.9, No.4, pp.e96063, 2014.

[5] D. Ryu, J. Jang, and J. Baik, "A hybrid instance selection using nearest-neighbor for cross-project defect prediction", *Journal of Computer Science and Technology*, Vol.30, No.5, pp. 969-980, 2015.

[6] J. V. Hulse, T. M. Khoshgoftaar, and A. Napolitano, "Experimental perspectives on learning from imbalanced data", In: *Proc. of the 24th International Conf. on Machine Learning*, 2007.

[7] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study", *Intelligent data analysis*, Vol.6, No.5, pp.429-449, 2002.

[8] X. Zhang, Q. Song, G. Wang, K. Zhang, L. He, and X. Jia, "A dissimilarity-based imbalance data classification algorithm", *Applied Intelligence*, Vol.42, No.3, pp.544-565, 2015.

[9] J. Vanhoeyveld and D. Martens, "Imbalanced classification in sparse and large behaviour datasets", *Data Mining and Knowledge Discovery*, Vol.32, No.1, pp.25-82, 2018.

[10] P. Hall, B. U. Park, and R. J. Samworth, "Choice of neighbor order in nearest-neighbor classification", *The Annals of Statistics*, Vol.36, No.5, pp.2135-2152, 2008.

[11] C.C. Holmes and N.M. Adams, "A probabilistic nearest neighbour method for statistical pattern recognition", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol.64, No.2, pp.295-306, 2002.

[12] A.K. Ghosh, "On optimum choice of k in nearest neighbor classification", *Computational Statistics & Data Analysis*, Vol.50, No.11, pp.3113-3123, 2006.

[13] Y. Sun, A.K.C. Wong, and M.S. Kamel, "Classification of imbalanced data: A review", *International Journal of Pattern Recognition and Artificial Intelligence*, Vol.23, No.4, pp.687-719, 2009.

[14] Y. Zhu, Z. Wang, and D. Gao, "Gravitational fixed radius nearest neighbor for imbalanced problem", *Knowledge-Based Systems*, Vol.90, pp.224-238, 2015.

[15] S. Vluymans, I. Triguero, C. Cornelis, and Y. Saeys, "EPRENNID: An evolutionary prototype reduction based ensemble for nearest neighbor classification of imbalanced data", *Neurocomputing*, Vol.216, pp.596-610, 2016.

[16] K. Jiang, J. Lu, and K. Xia, "A novel algorithm for imbalance data classification based on genetic algorithm improved SMOTE", *Arabian Journal for Science and Engineering*, Vol.41, No.8, pp.3255-3266, 2016.

[17] Y. Xu, Y. Zhang, J. Zhao, Z. Yang, and X. Pan, "KNN-based maximum margin and minimum volume hyper-sphere machine for imbalanced data classification", *International Journal of Machine Learning and Cybernetics*, pp.1-12, 2017.

[18] S.S. Mullick, S. Datta, and S. Das, "Adaptive Learning-Based k-Nearest Neighbor Classifiers with Resilience to Class Imbalance", *IEEE Transactions on Neural Networks and Learning Systems*, Vol.29, No.11, pp.5713-5725, 2018.

[19] M. Ohsaki, P. Wang, K. Matsuda, S. Katagiri, H. Watanabe, and A. Ralescu, "Confusion-matrix-based kernel logistic regression for imbalanced data classification", *IEEE Transactions on Knowledge and Data Engineering*, Vol.29, No.9, pp.1806-1819, 2017.

[20] A. Orriols-Puig and E. Bernadó-Mansilla, "Evolutionary rule-based systems for imbalanced data sets", *Soft Computing*, Vol.13, No.3, pp.213, 2009.

[21] N.V. Chawla, D. A. Cieslak, L.O. Hall, and A. Joshi, "Automatically countering imbalance and its empirical relationship to cost", *Data Mining and Knowledge Discovery*, Vol.17, No.2, pp.225-252, 2008.

[22] S. García, A. Fernández, J. Luengo, and F. Herrera, "A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and

interpretability”, *Soft Computing*, Vol.13, No.10, pp.959, 2009.

[23] E. Ramentol, Y. Caballero, R. Bello, and F. Herrera, “SMOTE-RSB*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced datasets using SMOTE and rough sets theory”, *Knowledge and information systems*, Vol.33, No.2, pp.245-265, 2012.

[24] Y.M. Huang, C.M. Hung, and H.C. Jiau, “Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem,” *Nonlinear Analysis: Real World Applications*, Vol.7, No.4, pp.720-747, 2006.

[25] D.A. Cieslak, T.R. Hoens, N.V. Chawla, and W. P. Kegelmeyer, “Hellinger distance decision trees are robust and skew-insensitive,” *Data Mining and Knowledge Discovery*, Vol.24, No.1, pp.136-158, 2012.

[26] J.W. Grzymala-Busse, J. Stefanowski, and S. Wilk, “A comparison of two approaches to data mining from imbalanced data”, *Journal of Intelligent Manufacturing*, Vol.16, No.6, pp.565-573, 2005.

[27] K.M. Ting, “An instance-weighting method to induce cost-sensitive trees”, *IEEE Transactions on Knowledge and Data Engineering*, Vol.14, No.3, pp.659-665, 2002.

[28] B. Zadrozny, J. Langford, and N. Abe, “Cost-sensitive learning by cost-proportionate example weighting”, In: *Proc. of the Third IEEE International Conf. on Data Mining*, 2003.

[29] Z.H. Zhou and X.Y. Liu, “On multi-class cost-sensitive learning”, *Computational Intelligence*, Vol.26, No.3, pp.232-257, 2010.

[30] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, “A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches”, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, Vol.42, No.4, pp.463-484, 2012.

[31] N.A. Le-Khac, M. O’Neill, M. Nicolau, and J. McDermott, “Improving fitness functions in genetic programming for classification on unbalanced credit card data,” In: *Proc. of European Conf. on the Applications of Evolutionary Computation*, pp.35-45, 2016.

[32] K.J. Wang, B. Makond, K.H. Chen, and K.M. Wang, “A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients”, *Application of Soft Computing*, Vol.20, pp.15–24, 2014.

[33] H. L. Chen, B. Yang, G. Wang, J. Liu, X. Xu, S. J. Wang, and D. Y. Liu, “A novel bankruptcy prediction model based on an adaptive fuzzy k-nearest neighbor method”, *Knowledge-Based Systems*, Vol. 24, No.8, pp.1348-1359, 2011.

[34] F. Angiulli, “Fast condensed nearest neighbor rule”, In: *Proc. of the 22nd international conf. on Machine learning, ACM*, 2005.