Received: December 29, 2016

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Improved Fuzzy-Optimally Weighted Nearest Neighbor Strategy to Classify Imbalanced Data

Harshita Patel[1]* **Ghanshyam Singh Thakur[1]**

[1] *Department of Mathematics and Computer Applications, Maulana Azad National Institute of Technology, Bhopal-462003, India*
* Corresponding author's Email: hpatel.sati@gmail.com

Abstract: Learning from imbalanced data is one of the burning issues of the era. Traditional classification methods exhibit degradation in their performances while dealing with imbalanced data sets due to skewed distribution of data into classes. Among various suggested solutions, instance based weighted approaches secured the space in such cases. In this paper, we are proposing a new fuzzy weighted nearest neighbor method that optimally handle the imbalance issue of data. Use of optimal weights improve the performance of fuzzy nearest neighbor algorithm for default balanced distribution of data, for the classification of imbalanced data concept of adaptive K is merged with it that apply large K, number of nearest neighbors for large class and small K for small class. We deploy this combination to classify imbalanced data with better accuracy for different evaluation measures. Experimental results affirm that our proposed method perform well than the traditional fuzzy nearest neighbor classification for these type of data sets.

Keywords: Fuzzy classification, Imbalanced datasets, K-nearest neighbors, Optimal solution.

1. Introduction

Data mining is a very popular and continuously growing research field with its various application areas. Ever increasing data of modern developed scientific era make the data mining more significant to find new milestones. Core data mining techniques and their applications provided the efficient solutions for decision making to the scientific society and continuously facilitating with their analytical power [1, 2]. New developments in technology uncover new challenges for the data mining researchers. Learning from imbalanced data is one of the most important real world classification issues refer to distribution of data. Unequal distribution of data into classes severely affects the performances of traditional classifiers because these classifiers are designed for balanced distribution of data by default, resulting in misclassification and classification accuracy is biased towards the larger class having more data even when the smaller class is of more interest [3]. In concerned with imbalanced data, class with large quantity of data examples is called majority class and other with few examples known as minority class. The issue of imbalance is discussed in detail by He et al. [3], they provided good literature and various aspects and possibilities for learning from imbalanced data.

The issue of classification of imbalanced data is comes under the top ten challenging issues of data mining [4] and also considered as new and frequently exploring trend of data mining [5]. Real world data applications are usually suffered with imbalance due to natural distribution of their data and for classification, obviously data is not in the required form that traditional classification algorithms look for. Medical Diagnosis [6, 7], Oil-spill Detection [8], Credit Card Fraud Detection [9], Culture Modeling [10], Network Intrusion, Text Categorization, Helicopter Gearbox Fault Monitoring [3] etc. are some examples of imbalanced data from real world need to be treated

in special manner to find accuracy of classification in correct way to protect results to be hazardous.

Worldwide research is going on to treat the sensitive issue of data imbalance that suggest various solutions like data balancing technique by re-sampling methods, alteration in traditional classification algorithms, analysis based on cost associated with data and ensemble techniques [3]. Alteration in traditional classification algorithms is one good solution while dealing with such data where classic algorithms are altered to the level to find accuracy of minority classes too with majority classes and overall accuracy improved with both class accuracies. This is good to classify data for its natural distribution without information replication or loss as happened sometimes in re-sampling techniques. Also algorithm modification is simply applicable where associated costs are not known. Almost all traditional classification algorithms have been re-proposed with their modified versions for imbalanced datasets and successfully running from yester years in both crisp and fuzzy manners and improvements are going on. Instance based learning is one of them. In this paper we will discuss our proposed nearest neighbor approach with fuzzy weights. In instance based learning such as nearest neighbor approaches no classifier is prepared in advance. Nearest neighbor algorithm is properly discussed by [11, 12]. Fuzzy logic makes the algorithm more efficient and weights put a stronger hold. For imbalanced data optimal weights with fuzzy concept provide the opportunity to find better classification for both classes. Default nearest neighbor considers equal weights for all data instances so may lead to misclassification, weighted concept deal with it. The performance of nearest neighbor classification is dependent on the opted weighing strategy that how we choose weights. In optimally weighted fuzzy nearest neighbor this becomes more reliable and least biased because here weighing is based on kriging that is a best linear estimator [29]. However skewed distribution of data of imbalanced datasets does not provide expected results with optimally weighted fuzzy nearest neighbor as other traditional classification approaches. Adaptive concept of K, i.e. large K for large classes and small K for small classes proposed by [28] deals with imbalanced text data very well. Our proposed method combines these two beneficial approaches to refine the classification of imbalanced data. In the experiments section it will be shown with the comparison of other methods that our proposed methodology is good enough for different evaluation measures.

The paper is organized as follows. In section 2, related works is given. Section 3 provides fundamental basics. Section 4 is dedicated to our proposed algorithm that is followed by experimental and result discussion of section 5. Conclusion and result discussion is given in section 6 and references are arranged at last.

2. Related Work

Traditional classification algorithms are designed for the default balanced distribution of data into classes. So when these algorithms face skewed or unevenly distributed data i.e. large quantity of data in one class while others have just few data elements than accuracy bias towards majority class. From solutions of this issue re-sampling is used in wide manner though information loss occurs in under-sampling and data redundancy increases in oversampling applications. Similarly cost sensitive approaches are applicable obviously when costs are given. Modification of algorithm is popular in classification of imbalanced data because these methods do not alter the original distribution of data. These approaches perform well in various ways. Nearest neighbors algorithms are the simplest to find out the class of an unknown instance and so for imbalanced data too with appropriate modifications.

Prati et al. [13] evaluated the performances of classifiers for different degrees of imbalance on their proposed evaluation model. They proposed confidence interval method to inspect classifier performance statistics and concluded that high degree of imbalance results in high misclassification and vice versa.

This section contains literature review on the work done for issue of imbalance with imbalanced approaches. Various modified nearest neighbor algorithms have been proposed in crisp and fuzzy manner with and without weights, some are discussed here. CCNND, a single class algorithm to minimize the classification cost is proposed by Kriminger et al. [14], they applied local geometric structure in data for this purpose. This algorithm is applicable on multiclass data too. Tomasev et al. [15] argued about hubness effect related to nearest neighbor that minority class instances are responsible for misclassification in high dimensional data unlike the fact that majority classes are mostly the reason of misclassification in low and medium dimensional data. Ryu et al. [16] proposed HISNN, an instance based hybrid selection using nearest neighbor for cross project defect prediction. In this class imbalance is existed in source and target projects. This method is worked in the way that

local learning is done by nearest neighbor and naïve bayes is used for global learning. Resultantly it is very efficient in finding high performance in software defect prediction.

Weighted approaches give good results for the purpose. Class based weighted nearest neighbor approach is proposed by Dubey et al. [17]. Distribution of nearest neighbor of test instances becomes the base for the weight calculation. Patel et al. [18] proposed hybrid neighbor weighted approach by merging adaptive concept with neighbor weighted strategy i.e. large weights for small classes and small weights for large classes with different K. Convex optimization technique was proposed by Ando [19] to find weights with a strong mathematical base improve non-linear performance measure for training data. Liu et al. [20] proposed class confidence weights for imbalanced learning. Weight prototypes were based on posterior probabilities gained from attribute probabilities, for this purpose Mixture Modelling and Bayesian Network were used.

In fuzzy scenario a few algorithms were proposed for imbalanced data with nearest neighbor concept. Ramentol et. al. [21] proposed a fuzzy rough ordered weighted average nearest neighbor method for binary classification with six weight vectors blended with some indiscernibility relations. Fernandez et al. [22] analyzed the fuzzy rule based classification systems for imbalanced data sets. For better classification results adaptive parametric conjunction operators were applied for different imbalanced ratios. Han et al. [23] proposed a nearest neighbor approach that was based on fuzzy-rough properties to minimize the biasness occur due to majority class. Liu et al. [24] proposed coupled fuzzy K-nearest neighbor approach for unevenly distributed categorical data instances, where strong bonds exist among class, attributes and other examples. Patel et al. [25] proposed an improved weighted algorithm. Large weight assignment to small classes and small weight assignment to larger classes got efficient when it merged with fuzzy logic.

3. Basic Milestones

Before moving to the proposed thought, it is required to be acquainted with fundamental terminologies and their usage. The proposed algorithm is a systematic enhancement of fuzzy K nearest neighbor concept with optimal weights in adaptive K manner. So we need to know about K nearest neighbor, fuzzy K nearest neighbor, adaptive K strategy and optimal weights.

Nearest neighbor approach is a classification method comes under lazy learning [1] in which no model of classifier is prepared initially like other eager learning algorithms. Whole training data is kept for classification of test instances and it is done by distance measuring of a test instance to the all training objects to find certain number of nearest neighbors say K. We mean by nearest neighbors are the training instances which have minimum distances with the test instance to be classified. After finding the K nearest neighbors of a test instance, class with maximum number of nearest neighbors is assigned to this instance. Distance measure and number of K may vary as per the researcher's requirement. More discussion is done by [12, 26].

Fuzzy K nearest neighbor algorithm is a fuzzy complement of it crisp version and deal with soft boundaries of data distribution. Instead of identifying the particular class of a test instance, this method finds the memberships of instances into classes that how much a instances belong to a class that helps in improving accuracy. Fuzzy K nearest neighbor algorithm was proposed by Keller et al. [27].

The perception of adaptive K is given by Baoli et al. [28] that K will be larger for larger class and smaller for smaller class for better categorization of imbalanced text data as common K for all classes is not suitable where one class having large quantity of data and have just few instances.

Optimally weighted fuzzy k nearest neighbor was proposed by Pham [29] based on kriging. Optimal weights are calculated for the traditionally found nearest neighbor of test instance. It was an excellent weighted improvement in fuzzy K nearest neighbor algorithm.

Combination of optimally weighted fuzzy K nearest neighbor with adaptive K strategy yields better results on imbalanced data.

4. Proposed Approach

We proposed a weighted strategy to classify imbalanced data with fuzzy optimal weights for adaptive K nearest neighbor. This is a good combination of adaptive approach for imbalanced data that says K should be different with different classes i.e. large K for large classes and small K for small classes. Optimally weighted fuzzy nearest neighbor was an improved nearest neighbor approach become specific while combine with adaptive strategy to deal with imbalanced data now.

Proposed Algorithm

Step 1. Find K_{Ci} for each class using

$$K_{Ca} = \min\left(\lambda + \left\lceil \frac{K \times I(Ca)}{\max\{I(Ca) \mid i = 1,2\}} \right\rceil, K, I(Ca)\right)$$

Step 2. Find memberships of training data into each class using

Let a training instance $y \in Ca$,*Then*

$$\mu_{Cn}(y) = \begin{cases} 0.51 + (m_{Cn} / K_{Ca}) \times 0.49 & If \ \ m = a \\ (m_{Cn} / K_{Ca}) \times 0.49 & othrerwise \end{cases}$$

$$While\ taking \ \sum \mu_{Ca}(y) = 1$$

Step 3. For test instance *t*, find a set of nearest neighbors X for any *K* from training dataset

Where X= (x_1, x_2….x_p), for *K*= p (some integer)

Step 4. Get covariance matrix C_t between nearest neighbors of *t*

Step 5. Get covariance matrix C_{tx} between *t* and its nearest neighbors

Step 6. Calculate weight matrix using

$$W = C_t^{-1} C_{tx}$$

Step7. Normalize negative weights to positive

$$w_{new} = \frac{w_b + \gamma}{\sum_{b=1}^{K} w_b + \gamma}, \forall b$$

$$Where \ \gamma = -\min w_b$$

Step 8. Find membership of test instance *u* using

$$\mu_{Ca}(t) = \frac{\sum_{b=1}^{Kca} w_b \times \mu_{Cab}}{\sum_{b=1}^{Kca} w_b}$$

Step 9. Assign class label to test instance *u* by

$$Ca(t) = \begin{cases} Ca & if\ \mu_c(t) \geq 0.51 \\ Random\ Assignment & Otherwise \end{cases}$$

Description of terms used in proposed algorithm

All terms could be understood in the given manner:

K = An integer input represents number of nearest neighbors to be found.

K_{Ca}= Different *K* for different classes calculated using equation given in step 1. Our purpose is to find large number of nearest neighbor for large classes and vice versa.

$I(C_a)$ = Number of instances in each Class C_a, Where *a*= 1 and 2 (for binary classification).

λ = Constant integer value to prevent very small results.

C_n = Nearest Neighbors of training instance *y* from class *C*.

$\mu_{Cn}(y)$ = Membership of *y* into class *C*.

X= Number of Nearest Neighbors of test instance.

C_t = Covariance matrix of nearest neighbors of *t*.

C_{tx} =Covariance matrix of nearest neighbors between t and its nearest neighbors.

W= Weight Matrix.

w_{new} = Modified weights to avoid negative weight values.

$\mu_{Ca}(t)$ = Membership of test instance in class *a* where *a*= 1 or 2.

$C_a(t)$ = Class assigned to test instance.

5. Experiments and Results

Proposed algorithm is experimented on five datasets. These datasets are taken from standard UCI [30] and KEEL [31] repositories. These are the benchmarked datasets already explored well for the same purpose of research in imbalanced datasets by many researcher of the field. Data sets used for experimental purpose are briefly explained in following Table 1:

Table 1. Dataset Description

Data- sets	Source	# Instance	Class (1/0)	#Attributes	IR
Ionosphere	UCI	351	Bad/Good Radar Returns	34	1.79
Glass0	KEEL	214	Positive/ Negative	9	2.05
Vertebral	UCI	310	AB/No	6	2.1
Ecoli1	KEEL	336	Positive/ Negative	7	3.36
Spectfheart	UCI	267	Abnormal/ Normal	44	3.85

As we know calculation of accuracy is not enough for imbalanced data classification because it is biased towards majority classes. F-Measure and G-Mean are popular accuracy measures to evaluate such cases of imbalanced data and so taken here. The evaluation measures F-Measure and G-Mean could be understood on the basis of confusion matrix given in Table 2:

Table 2. Confusion Metric for Binary Classification

Calculated / Actual Outcomes	Calculated Positive	Calculated Negative
Actual Positive	True Positive	False Positive
Actual Negative	False Negative	True Negative

True positive (TP) are actual positive instances which are correctly classified as positive. False positive (FP) are actual positive instances, incorrectly classified as negative. True negatives (TN) are actual negative instances also predicted correctly as negative and False negatives (FN) are actual negative instances that are incorrectly classified as positive.

On the basis of these predictions, evaluation measure F-Measure and G-Mean work as follows:

$$F-Measure=\frac{2\times \Pr ecision\times \mathrm{Re} call}{\Pr ecision+\mathrm{Re} call} \quad (1)$$

$$G-Mean=\sqrt{\frac{TP}{TP+FN}\times\frac{TN}{TN+FP}} \quad (2)$$

Where

$$\Pr esicion=\frac{TP}{TP+FP} \quad (3)$$

and

$$\mathrm{Re} call=\frac{TP}{TP+FN} \quad (4)$$

Experiments are done for these two evaluation measures are performed for all above five datasets. Also comparison of proposed algorithm is done with the fuzzy neighbor weighted nearest neighbor algorithm [25], a good fuzzy-weighted approach for imbalanced classification in nearest neighbor scenario that implies the better performance of our proposed approach. Five K values were taken for these evaluations (i.e. 5, 10, 15, 20, and 25) and table 3 and 4 are showing average results for all K.

Fuzzy neighbor weighted nearest neighbor algorithm (Fuzzy-NWKNN) and our proposed weighted fuzzy adaptive K nearest neighbor (Weighted FAKNN) are compared in Table 3 and Table 4 for experimental results of F-Measure and G-Mean and performance of proposed Weighted FAKNN found better than the Fuzzy-NWKNN.

Graphical representations of performances of both algorithms are shown in Figure 1 and Figure 2 in form of bar charts that can easily illustrate the better performance of Weighted FAKNN.

Table 3. Results for F-Measure of Fuzzy-NWKNN and Weighted FAKNN

Datasets	Fuzzy-NWKNN	Weighted-FAKNN
Ionosphere	0.5161	0.7033
Glass0	0.649	0.7696
Vertebral	0.439	0.7717
Ecoli1	0.3934	0.7392
Spectfheart	0.2639	0.3761

Table 4. Results for G-Mean of Fuzzy NWKNN and Weighted FAKNN

Datasets	Fuzzy-NWKNN	Weighted-FAKNN
Ionosphere	0.4289	0.7447
Glass0	0.6806	0.8029
Vertebral	0.4588	0.7984
Ecoli1	0.5445	0.8001
Spectfheart	0.4578	0.6109

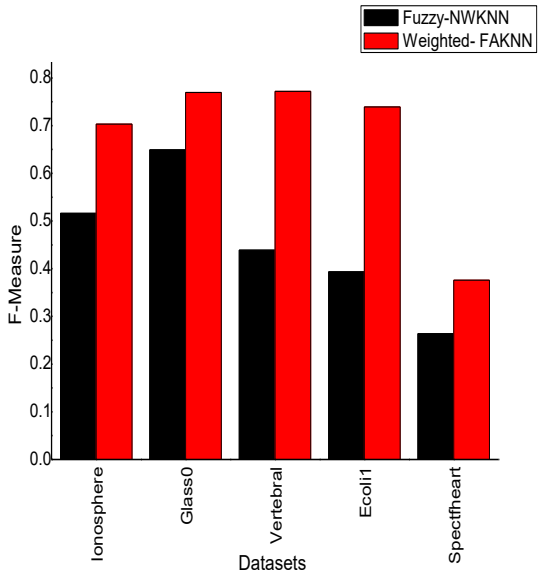

Figure 1. F-Measure for Fuzzy-NWKNN and Weighted FAKNN

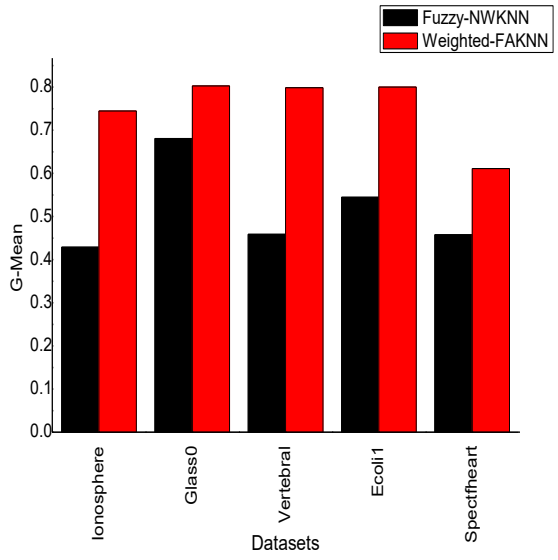

Figure 2. G-Mean for Fuzzy-NWKNN and Weighted FAKNN

6. Conclusion and Future Scope

In this paper, we proposed a fuzzy weighted adaptive nearest neighbor approach for better classification of imbalanced data. By using adaptive strategy, we choose different K for different classes according to their sizes. This results in comfortable classification of imbalanced data of different sized classes with very large or small quantity of data into them. Uniform K is mostly not good for such cases. This concept is merged with optimally weighted fuzzy K nearest neighbor and yielding better outcomes for imbalanced data. This combination is dealing better now with imbalanced data. Experiments are done on the data sets of different imbalance ratios for well known evaluation measures F-measure and G-mean. For this paper we consider all features as necessary and binary classification was our main intension. In future the proposed algorithm can be extended for multiclass classification with possible feature selection.

References

[1] J. Han and M. Kamber, *Data Mining, Concepts and Techniques*, 3rd ed., Morgan, Kaufmann, 2006.

[2] H. Patel and D.S. Rajput, "Data Mining Applications in Present Scenario: A review", *International Journal of Soft Computing*, Vol. 6, No. 4, pp. 136-142, 2011.

[3] H. He and E. A. Garcia, "Learning from Imbalanced Data", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 21, No. 9, pp. 1263-1284, 2009.

[4] Q. Yang and X. Wu, "10 challenging problems in data mining research", *International Journal of Information Technology and Decision Making*, Vol. 5, No. 4, pp. 597-604, 2006.

[5] Editorial, Special issue on "New trends in data mining", NTDM. *Knowledge Based Systems*, Elsevier, pp. 1-2, 2012.

[6] R. Pavón, R. Laza, M. Reboiro-Jato and F. Fdez-Riverola, "Assessing the impact of class-imbalanced data for classifying relevant/irrelevant medline documents", *Advances in Intelligent and Soft Computing*, Vol. 93, pp. 345–353, 2011.

[7] R. B. Rao, S. Krishanan and R.S. Niculscu, "Data Mining for Improved Cardiac Care", *ACM SIGKDD Exploration Newsletter*, Vol. 8, No. 1, pp. 3–10, 2006.

[8] M. Kubat, R. C. Holte and S. Matwin, "Machine Learning for the Detection of Oil Spills in Satellite Images", *Machine Learning*, Vol. 30, No. 2, pp. 195–215, 1998.

[9] P. Chan and S. Stolfo, "Toward Scalable Learning with Non-Uniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection", *In Proceedings of Knowledge Discovery and Data Mining*, pp. 164–168, 1998.

[10] X. C. Li, W. J. Mao, D. Zeng, P. Su and F. Y. Wang. "Performance Evaluation of Machine Learning Methods in Cultural Modeling", *Journal of Computer Science and Technology*, Vol. 24, No. 6, pp. 1010-1017, 2009.

[11] G. Loizou and S. J. Maybank, "The Nearest Neighbor and the Bayes Error Rates", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 9, No. 2, pp. 254-262, 1987.

[12] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification", *IEEE Transactions on Information Theory*, Vol. 13, No. 1, pp. 21-27, 1967.

[13] R. C. Prati, G. E. A. P. A. Batista and D. F. Silva, "Class imbalance revisited: a new experimental setup to assess the performance of treatment methods", *Knowledge and Information Systems*, Vol. 45, No. 1, pp. 247-270, 2015.

[14] E. Kriminger and C. Lakshminarayan. "Nearest Neighbor Distributions for Imbalanced Classification", In: *Proc. of WCCI 2012 IEEE World Congress on Computational Intelligence*, Brisbane, 2012, pp. 10-15.

[15] N. Tomašev and D. Mladenic. "Class Imbalance and the Curse of Minority Hubs", *Knowledge-Based Systems*, Vol. 53, pp. 157–172, 2013.

[16] 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.

[17] H. Dubey and V. Pudi. "Class based weighted k nearest neighbor over imbalanced dataset", *PAKDD 2013, Part II, LANI, 7819*, pp. 305-316, 2013.

[18] H. Patel and G.S. Thakur, "A Hybrid Weighted Nearest Neighbor Approach to Mine Imbalanced Data", *In: Proc. of the 12th International Conference on Data Mining (DMIN'16)*, pp. 106-110, 2016.

[19] S. Ando, "Classifying imbalanced data in distance-based feature space", *Knowledge and Information Systems*, vol. 46, No. 3, pp. 707–730, 2016.

[20] W. Liu and S. Chawla. "Class Confidence Weighted kNN Algorithms for Imbalanced Data Sets", *PAKDD 2011, Part II, LANI 6635*, pp. 345-356, 2011.

[21] E. Ramentol, S. Vluymans, N. Verbiest, Y. Caballero, R. Bello, C. Cornelis, and F. Herrera, "IFROWANN: Imbalanced Fuzzy-Rough Ordered Weighted Average Nearest Neighbor Classification", *IEEE Transactions on Fuzzy Systems*. 2014.

[22] A. Fernandez, M. J. Jesus and F. Herrera, "On the Influence of an Adaptive Inference System in Fuzzy Rule Based Classification Systems for Imbalanced Data-Sets", *Expert Systems with Applications*, Vol. 36, No. 6, pp. 9805-9812, 2009.

[23] H. Han and B. Mao, "Fuzzy-Rough k-Nearest Neighbor Algorithm for Imbalanced Data Sets Learning", In: *Proc. of FSKD 2010-Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, IEEE circuits and systems society, China, pp. 1286-1290, 2010.

[24] C. Liu, L. Cao and P.S. Yu, "Coupled Fuzzy *K*-Nearest Neighbors Classification of Imbalanced Non-IID Categorical Data", In: *Proc. of IJCNN - International Joint Conference on Neural Networks*, IEEE, Beijing, 2014, pp. 1122-1129.

[25] H. Patel and G. S. Thakur, "Classification of Imbalanced Data using a Modified Fuzzy-Neighbor Weighted Approach", *International Journal of Intelligent Engineering and Systems*, Vol. 10, No. 1, pp. 56-64, 2017.

[26] E. Fix and J. L. Hodges, "Discriminatory Analysis-Nonparametric Discrimination: Consistency Properties", *International Statistical Review*, Vol. 57, pp. 238–247, 1989.

[27] J. M. Keller, M. R. Grey and J. A. Givens Jr., "A Fuzzy k- Nearest Neighbor Algorithm", *IEEE Transactions on System, Man and Cybernetics*, Vol. 4, pp. 580-585, 1985.

[28] L. Baoli, L. Qin and Y. Shiwen "An adaptive k-nearest neighbor text categorization strategy" *ACM Transactions on Asian Language Information Processing*, Vol. 3, No. 4, pp. 215-226, 2004.

[29] T. D. Pham, "An optimally weighted fuzzy k-NN algorithm", In: *proc of International Conference on Pattern Recognition and Image Analysis*, U.K. 2005, pp. 239-247.

[30] A. Asuncion and D. J. Newman, UCI machine learning repository.
University of California, School of Information and Computer Science, Irvine, 2007.
http://www.ics.uci.edu/~mlearn/MLRepository.html

[31] KEEL: *Knowledge Extraction based on Evolutionary Learning*.
http://sci2s.ugr.es/keel/imbalanced.php