# A Survey On Instance Reduction algorithm without k parameter based on Natural Neighborhood graph classifier

*Kavita K. Kadam[1], Prof S. V. Chobe[2]*
*Computer Engineering*
*Dr. D. Y. Patil Institute of Technology*
*Pune, India*
*(E-mail: kavitakadam2013@gmail.com, sanchobe@yahoo.com)*

***Abstract* —** k-Nearest Neighbor based classification methods and high-performance pattern recognition techniques. This methods are vulnerable to parameter choice. In KNN algorithm selection of parameter k for the classification of data remains the same. The proposed work introduces a new supervised classification algorithm, Instance Reduction algorithm without k parameter based on Natural Neighborhood graph classifier.

New algorithm provides a reduction of the complexity of Natural Neighborhood Based Classification Algorithm (NNBCA). The KNN based algorithm, depends on a prior k, NNBCA predicts different k for different samples. Therefore, Instance Reduction algorithm without k parameter based on Natural Neighborhood graph classifier is able to learn more from flexible neighbor information.

***Keyword***: *Instance-Based Learning, Instance Reduction, Natural Neighbor, Natural Neighborhood Graph, classification, Particle swarm optimization.*

## I. INTRODUCTION

The goal of data mining also known as Knowledge discovery in the database is to find interesting information from massive databases in automatic ways. For industrial areas, various methods using data mining approaches such as classification and pattern mining have been developed in order to analyze complex industrial data and diagnose given problems. In particular, pattern mining has been widely used to analyze various data and find valuable pattern information.

Due to the continuous expansion of data availability in the area of science and engineering, finding patterns from vast amounts of data and identifying members of a predefined class, which is called classification, it becomes very critical tasks. Therefore, classification is a one of the fundamental problem, especially in pattern recognition and data mining. Many effective algorithms have been successfully applied in many real-world applications. In setting a classification, two kinds of classifiers exists, are parametric and nonparametric classifiers. k-Nearest Neighbor classifier, is the type of nonparametric classifiers, are basic classifiers that are used to classify a query object into the category as its nearest example.

Many other algorithms have been proposed to improve neighborhood based classifiers. Tangand He [2] proposed an Extended Nearest Neighbor (ENN) method for classification, which uses the different way of communication style. The existing classic KNN algorithm depends on the nearest neighbors of a test sample to make a classification decision, ENN algorithm not depend on but also depend on the test sample their nearest neighbors. This different way of communication style is both an advantage and disadvantage of the ENN method. The advantage is that the classification decision of ENN method depends on all variable training data. The disadvantage is that the problem of parameter selection of k in KNN still exists. ENN has a parameter of k, on the basis of different size and different shape of the graph can be modified. This feature makes additional optimization possible. However, the problem of parameter selection continues to persist [5].

Currently, data pre-processing has become an increasingly important step in knowledge discovery in databases with an increasing number of data collections. In instance-based machine learning algorithms, all instances of a training set are used to construct inference structures. If the chosen data set contains too many instances and features, excessive consumption of storage and large computation time in the stage of classification may result. KNN has been widely applied in pattern recognition, image processing, and data mining, but it still faces many problems, such as large storage requirements, low prediction performance, and sensitivity to noisy and redundant patterns. Instance reduction (IR) is developed to reduce the size of a training set [4]. IR aims to remove noisy and redundant instances in the dataset, which may lower the prediction accuracy.

The previous system, construct classification algorithm for nearest neighbor classification, with the help of NaN methods[5]. This NNBCA consists of the training stage and the testing stage. In the training stage, the algorithm deals with training data, and the NaN method uses parameter k in the generate class-wise statistic calculation step. NaNG can then be used to calculate the distance for a given test sample and natural neighbor efficiently for a given test sample. Therefore, in the generate class-wise statistic calculation step, the algorithm stores NaNG and transfers it to the next stage. The time complexity is O (m-log (m), m is the size of the training data set. In the testing stage, the algorithm is to determine to which class the sample in the testing data set. Taking into account the diversity of samples, we use the number of samples' natural neighbors as the neighborhood parameters to calculate the generate class-wise statistic. The time complexity of is O(m-n), m is the size of the training data set, and n is the size of the testing data set[5].

Particle swarm optimization used population based search strategy which finds an optimum by stochastically "flying" a set of particles through a search space. Particles iteratively search a region between and beyond their own individual prior best position and the position of their most successful neighbor. In doing so, fitter positions may be found. In each, updating their individual best positions, the particles change their search directions to explore these new fitter positions. Through this process found the maximum/minimum of particles converges. In PSO a set of particles find an optimum iterative process in which particles sample a search space and then adjust their search directions to sample near to their fitter neighbors. Neighbors are those particles which can share information. The set of neighbor - connections between all of the particles from the swarm's topology or sociometry and affects the swarm's exploitation and exploration behavior [8]. In standard PSO topologies, there is no spatial significance between neighboring particles as neighbors are random in terms of their relative positions.

Delaunay triangulation spatially sub-divides a set of points into triangles in 2D where the endpoints of the simplex (an n-dimensional equivalent of a triangle) edge lie on the circumference of the circumcircle. Standard PSO used for the fast and simple high dimensional optimizer. Delaunay triangulation (DT) subdividing a set of points in expected near linear time in low dimensions 2D and 3D. Delaunay triangulation achieves a spatial topology, by computing the closest surrounding neighbors for each particle.

## II. LITERATURE SURVEY

Hatem A. Fayed et al [1] proposed the nearest neighbor (KNN) rule is one of the most widely used pattern classification algorithms. The nearest neighbor rule is one of the most widely used pattern classification algorithms. For large data sets, the computational demands for classifying patterns using KNN can be prohibitive [1]. Proposed approaches improve the level of classification accuracy as the traditional KNN. Moreover, it is a simple and fast condensing algorithm. In given system selection parameter k still exists.

B. Tang and H. B. He [2] Proposed ENN is able to learn from the global distribution to improve pattern recognition performance. In KNN algorithm considers the nearest neighbors of a test sample to make a classification decision. The advantages of the existing KNN approach are retained in our ENN classifier, such as easy implementation classification performance. ENN classifier makes a prediction by not considering who are the nearest neighbors of the test sample, but also consider the test sample as their nearest neighbors. In this system selection parameter k still exists.

Q. S. Zhu et al [3] proposed a parameter-free classification method based on extended nearest neighbor method, the parameter k selecting problems in training stage with respect to the testing stage are solved by the natural neighbor method in uncorrelated ways. ENaN algorithm achieves the highest accuracy and stability. Proposed algorithm increases the accuracy of classification result as well as its adaption to different kind of data sets and avoids the neighborhood choosing problem. ENN method uses a different way of communication style for predicting result it does not consider only who the nearest neighbors of the test sample are, but also consider the test sample as their nearest neighbors. ENN is improving pattern recognition performance and providing a powerful technique for a wide range of data analysis applications.

Lijun Yang et al [4] In Graph-based instance reduction algorithm natural neighborhood graph (NaNG) is automatically constructed by the natural neighbor search algorithm. The algorithm uses the concept of a natural neighborhood graph to remove the noisy and redundant instances. The natural neighbor was originally proposed to solve the parameter selection problem for a k-nearest neighbor. Natural neighbor is a scale-free neighbor form, and its whole computation procedure is automatically fulfilled without any parameters.

The main advantages of NNGIR are it does not require any user defined parameters, by using NaNG, it increases the reduction rate greatly, while maintaining or even improving accuracy, its fluctuation of reduction rates is small for different types of datasets. Experiments demonstrate that the proposed algorithm can reduce training sets greatly while maintaining or even improving the classification accuracy.

Ji Feng et al [5] Proposed Parameter-free classification algorithm called NNBCA, and the problem of parameter k selection in the training and testing stages is solved perfectly by using the NaN method. The efficiency of NN classification mostly depends on the type of distance measure, especially database is in a large-scale and high-dimensional. Traditional KNN algorithm accept a fixed k for all query samples regardless of their geometric location and related specialties. The geometrical placement may be more important than the actual distance to predict a query sample's neighborhood. This algorithm increases the accuracy of classification results, adapts well to different kinds of data sets, and solves the neighborhood selection problem, which means that the value of parameter k is no longer needed. But the complexity of NNBCA is large.

James Lane at [8] proposed Particle Swarm Optimization with Spatially Meaningful Neighbours. In the proposed system designed which uses heuristics to leverage the natural neighbors computed with Delaunay triangulation. Delaunay neighbors as a spatial topology for PSO. The nature of this topology does facilitate meaningful spatial heuristics which modulate the connections to accomplish local searching.

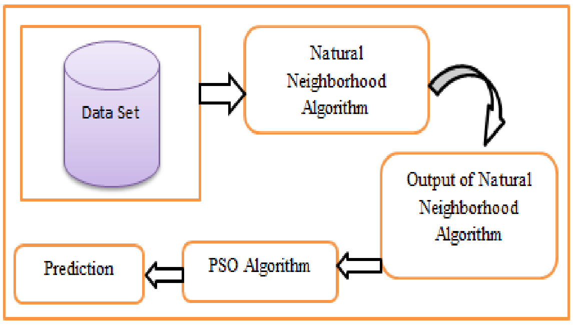
Data Set
Natural Neighborhood Algorithm
Output of Natural Neighborhood Algorithm
PSO Algorithm
Prediction

Figure 1: Proposed system architecture

## III. CONCLUSIONS

This paper proposes Instance Reduction algorithm without k parameter based on Natural Neighborhood graph classifier, and the problem of parameter k selection in the training and testing stages is solved perfectly by using the NaN method, which means that the value of parameter k is no longer needed. NaN method is a new concept of the nearest neighbor method. NaN method used in clustering, outlier detection, prototype reduction, and classification. The proposed system uses particle swarm optimization algorithm to increases the accuracy of classification results.

## REFERENCES

[1] Hatem A. Fayed and Amir F. Atiya," A Novel Template Reduction approach for the -Nearest Neighbor Method ", IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 20, NO. 5, May 2009.

[2] B. Tang and H. B. He,"ENN Extended nearest neighbor method for pattern recognition", IEEE Computer. Intel. Mag., vol. 10, no. 3, pp. 52–60, 2015

[3] Q. S. Zhu, J. Feng, and J. L, "Huang, Natural neighbor: A self-adaptive neighborhood method without parameter K, Pattern Recognit. Lett", vol. 80, pp. 30–36, 2016.

[4] Lijun Yang, Qingsheng Zhu, Jinlong Huang, Dongdong Cheng, Quanwang Wu, Xiaolu Hong, "Natural Neighborhood Graph-based Instance Reduction Algorithm without Parameters", Applied Soft Computing Journal 2018.

[5] Ji Feng, Yan Wei, and Qingsheng Zhu," Natural Neighborhood- Based Classification Algorithm Without Parameter k", IEEE TRANSACTIONS ON BIG DATA MINING AND ANALYTICS ISSN 2096-0654 01/06 pp257–265 Volume 1,2018.

[6] S. H. Wang, Q. M. Huang, S. Q. Jiang, Q. Tian, and L. Qin, "Nearest-neighbor method using multiple neighborhood similarities for social media data mining", Neurocomputing, vol. 95, pp. 105–116, 2012.

[7] D. Lunga and O. Ersoy, "Spherical nearest neighbor classification Application to hyperspectral data, in Machine Learning and Data Mining in Pattern Recognition", P. Perner, ed. Springer, 2011.

[8] James Lane, Andries Engel Brecht and James Gain, "Particle Swarm Optimization with Spatially Meaningful Neighbours", IEEE Swarm Intelligence Symposium 2008.

[9] T. Inkaya, S. Kayalrgil, and N. E. Ozdemirel, An adaptive neighbourhood construction algorithm based on density and connectivity, Pattern Recognit. Lett, vol. 52, pp. 17– 24, 2015.

[10] K. Kozak, M. Kozak, and K. Stapor, Weighted K-nearest neighbor techniques for high throughput screening data, Int. J. Biomed. Sci, vol. 1, no. 3, p. 155, 2006.

[11] P. Mitra, C.A. Murthy and S.K. Pal, Unsupervised feature selection using feature similarity, IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 3, pp. 301–312, 2002.