The Cosine Similarity Technique for Removing the Redundancy Sample

Worasak Rueangsirarak, *IEEE Member*
Computer and Communication
Engineering for Capacity Building
Research Unit
School of Information Technology
Mae Fah Luang University
Chiang Rai, Thailand
worasak.rue@mfu.ac.th

Teeravisit Laohapensaeng
Computer and Communication
Engineering for Capacity Building
Research Unit
School of Information Technology
Mae Fah Luang University
Chiang Rai, Thailand
teeravisit.lao@mfu.ac.th

Suppakarn Chansareewittaya
Computer and Communication
Engineering for Capacity Building
Research Unit
School of Information Technology
Mae Fah Luang University
Chiang Rai, Thailand
suppakarn.cha@mfu.ac.th

Anusorn Yodjaiphet
Department of Electrical Engineering
Faculty of Engineering
Rajamangala University of Technology
Lanna Chiang Rai
Chiang Rai, Thailand
anusorn@rmutl.ac.th

Abstract—— **The k-nearest neighbor algorithm is one of the basic and simple classification algorithms that share a common limitation of the algorithm which requires more computation cost when the size of training data is enlarged. To solve this problem, a new method applied to the cosine similarity for reducing the size of the training data set is proposed. This method reduces the data points that close to a decision boundary and retains the important points which affect classification accuracy. For the data far from the decision boundary and not affect the classification, these points will be removed from the training data set. The proposed method is evaluated its accuracy and reduction performance on the state of the art mechanisms, categorized as prototype selection algorithms. The 20 real-world data set are used to evaluate the proposed method. The experimental results are compared with 21 existing methods. As a result, our proposed method performs the best with 89.95% accuracy but has only a fair reduction ratio, when compared to other methods.**

Keywords—**prototype selection; data reduction; k-nearest neighbor; cosine similarity; classification**

I. Introduction

In the machine learning field of research, they are always experienced with the same problem of classification algorithm with high computation cost, such as the use of memory, computation times, when training data is increased. The consumption of resources depends on the concept or idea of algorithms. The simplest classification algorithm like k-nearest neighbor (k-NN) also has this kind of problem too [1]. The large training data are easy to adulterate with noise, outlier, and redundant data which cause a low generalization accuracy.

There is one simple option to solve the large training data problem, which is to reduce the data size. This method is the so-called prototype selection (PS) [2]. The idea of this method is to extract important data from the original training data set by using a prototype generation (PG) algorithm [3]. There are many previous pieces of research i.e. Fayed et al. [4] propose the idea which is based on defining the chain. This so-called chain is a sequence of nearest neighbors from alternating classes. This leads the chain to stand so close to the boundary of classification. Moreover, the chain is lined up on the cutoff value based on the patterns of the training data set. This algorithm is called template reduction for KNN (TRKNN). Olvera-Lopez et al. [5] propose the new method for large datasets finding. This method is a clustering algorithm that applied the border prototypes and interior prototypes for selection, which is called Prototype Selection by Clustering (PSC). Marchori [6] introduces the class conditional nearest neighbor (CCNN) which is a labeling method to improve the performance of the 1-NN rule. Nikolaidis et al. [7] introduce the Class Boundary Preserving (CBP) algorithm, which is a hybrid method between multi-stage methods for instance selection of training set. Verbiest et al. [8] also propose Fuzzy Rough Prototype Selection (FRPS) that uses the fuzzy rough set theory to express the quality of the instances and uses a wrapper approach to determine the right instances for pruning. Hamidzadeh et al. [9] develop the Large Margin Instance Reduction Algorithm (LMIRA) that can be used to remove the non-border instances and keep border ones. This method is worked based on keeping the hyperplane which separates a two-class data and provides large margin separation. Nikolaidis et al. [10] propose a Direct Weighted Pruning (DWP) algorithm that uses a set of weights to directly control which prototypes have to be discarded or survive. Leyva et al. [11] propose three instances selection method, the local set-based smoother (LSSm). The local set-based centroids selector (LSCo) and the local set border selector (LSBo) are proposed based on the local sets which are different. They complementary strategies are to reduce the number of instances in the training set without affecting the classification accuracy.

In this research, our main contribution is focusing on the prototype selection (PS) method. With one common algorithm of PS method, the condense the nearest neighbor (CNN) is adopted to this research. In [12] and [13] explain the important properties of the PS method into five parts distinguished to a direction of search, a type of selection, an evaluation of search, the other properties, and criteria to compare PS methods. The detail of these properties can be explained as follows:

978-1-7281-5419-0/19/$31.00 ©2019 IEEE

A. The Direction of Search

The searching criteria for extraction or reduction of the data point from the training set (*TR*) to reduced set (*S*) has five directions consist of incremental, decrement, batch, mixed and fixed.

B. Type of Selection

The PS algorithms have a different process then it retains/removes the border points, central points, or some other set of points in TR depend on the conditioned by the type of search. There are many types of selection such as consideration, edition, and, hybrid approach.

C. Evaluation of Search

The k-NN is a simple technique that can be applied for searching to find a direct search of a PS algorithm. The objective of this searching is to predict non-definitive selected data. Then, these selected data are compared to the other methods. This characteristic influences the quality criterion and it can be divided into filter and wrapper.

D. Other Properties

The other properties which are related to PS methods and dependent on the type of k-NN are a representation, distance function, and voting of the data.

E. Criteria to Compare PS Methods

The comparison of the performance of PS methods is generally measured with the value of storage reduction and generalization accuracy to show the advantages of each method.

II. Cosine Similarity

The cosine similarity is a measurement method to identify a similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them. Given two vectors of attributes, A and B, the cosine similarity, cos θ, is represented using a dot product and magnitude in (1) [14].

$$\text{Similar} = \cos\theta = \frac{\mathrm{A}\Box\mathrm{B}}{\|\mathrm{A}\|\|\mathrm{B}\|} \tag{1}$$

where; A and B are vectors explained in (2) and (3), respectively.

III. Methodology

The data reduction method begins with defined the training data set as ***T*** with ℓ samples, where ***T*** consists of the training sample $\boldsymbol{x}_i$ ($x_i \in \Box^N$ when $i = 1, 2, \ldots, \ell$) coupled with $\boldsymbol{y}_i$. These samples belong to one of C class. The reduced version of ***T*** is named $\boldsymbol{T}_{\boldsymbol{Reduce}}$, which has the same value of N and C. In addition, the number of samples is under this condition, $\ell_r \leq \ell$. The parameter ℓ_r represents the number of data points located near the decision boundary.

The data reduction is used to apply a cosine similarity technique consists of three steps: eliminating the noise and overlapping data, estimating the boundary of each class and reducing the redundant data by comparing the cosine similarity. The more details on each step are given as follows.

A. Removing Noise and Overlapping Data

The overlapping data and noises have affected the accuracy of the decision boundaries. In this first step, the fast and simple method, the 1-NN algorithm, is used to remove noise and overlapping data.

B. Class Boundary Approximation

In this step, the three nearest neighbor enemies of each data point for each class *i* of data in set ***T*** ($i = 1, 2, \ldots, C$) is identified. These enemies are stored in the *i*-th approximated class boundary set, denoted as $\boldsymbol{E}_i$. For example, if ***T*** has two classes, i.e. *C* = 2, then all three nearest neighbor enemies of class 1 are stored in $\boldsymbol{E}_1$ and all three nearest neighbor enemies of class 2are stored in $\boldsymbol{E}_2$.

C. Reduced the Class Boundary using Cosine Transform

In the third step, the duplicate samples in E_i are reduced. The process starts with selecting one sample of the class that wants to reduce (E_{x1}). Finding the nearest sample in the same class (E_{x2}) and the nearest sample in the opponent class (E_y) create two vectors (*A* and *B*) as in equations (2) and (3).

$$A = [E_y \; E_{x1}]^T \tag{2}$$

$$B = [E_{x2} \; E_{x1}]^T \tag{3}$$

After that, the similarity and direction of these two vectors are evaluated to find a similar direction by using the cosine similarity from equation (1).

IV. Evaluation

A. Evaluation Method

We used the criteria of testing accuracy (*Acc*), the reduction ratio (*RR*), and the product of *Acc* and *RR* (*Acc* RR*) to evaluate the performance of the proposed method. The proposed method was compared with 21 PS methods as shown in table I. We performed comparative experiments with a variety of methods to demonstrate the difference in the results of each method. We applied 10-fold cross-validation and averaged their 10 time-values of testing accuracy and reduction ratio. The results of these 21 PS methods were received from the KEEL software tool.

TABLE I. List of PS Methods

PS Method						
Condensation		Edition		Hybrid		
Filter	Batch	Filter	Batch	Filter	Batch	Wrapper
CNN	POP	MENN	AllKNN	C-Pruner	CCIS	CHC
FCNN	Reconsistent	RNGE		DROP3	HMNEI	GGA
MSS				IB3	ICF	SSMA
MCNN						RMHC
RNN						

B. *Real-World Data Set*

Our 10-fold cross-validation was evaluated on the data which were categorized into 2 groups by their size of data i.e. small data set and medium data set.

1) *Small Data Set*

The small data set consists of 10 sets of data where their data points are less than 2,000 samples. The details of each set are explained in table II, in the order of name, a number of sample points, a number of attributes, and the number of classes. This variety of attributes and classes could evaluate the generalization of the proposed method.

TABLE II. SMALL DATASET

Data Set	#Instance	#Attribute	#Class
Bupa	345	6	2
Hayes-roth	160	4	3
Iris	150	4	3
Led7digit	500	7	10
Monks	432	6	2
Sonar	208	60	2
Spectfheart	267	44	2
Vowel	990	5	11
Wisconsin	699	9	2
Zoo	101	16	7

2) *Medium Data Set*

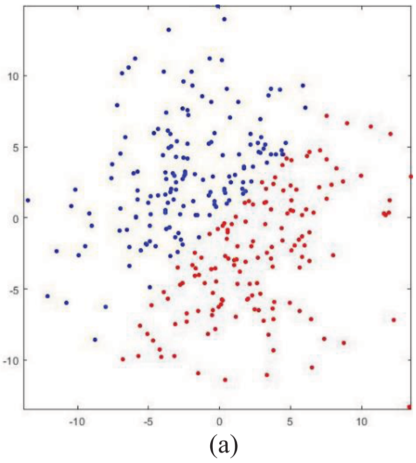
The medium data set consists of 10 sets of data where their data points less than 20,000 samples as shown in table III.

TABLE III. MEDIUM DATASET

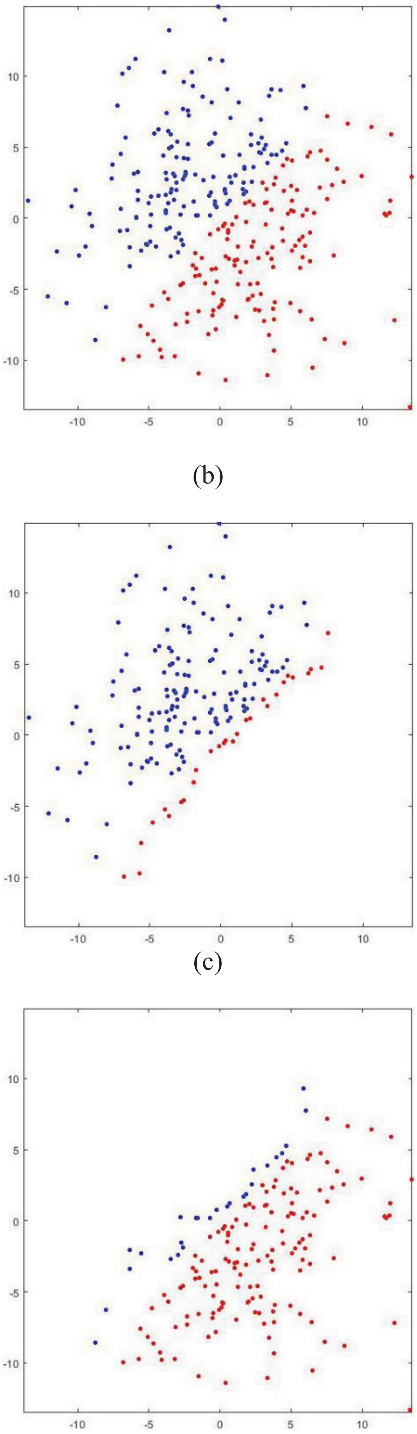
Data Set	#Instance	#Attribute	#Class
Coil2000	9,822	85	2
Magic	19,020	10	2
Page-blocks	5,472	10	5
Penbased	10,992	16	10
Phoneme	5,404	5	2
Satimage	6,435	36	7
Texture	5,500	40	11
Thyroid	7,200	21	3
Titanic	2,201	3	2
Twonorm	7,400	20	2

V. EXPERIMENTAL RESULTS

The preliminary result of the proposed method is shown in figure 1. The original data is on the top side and the result after reducing the process is on the bottom side. This method can apply to the multiclass data set. The result of multiclass data reduction is presented in figure 2.



(a)

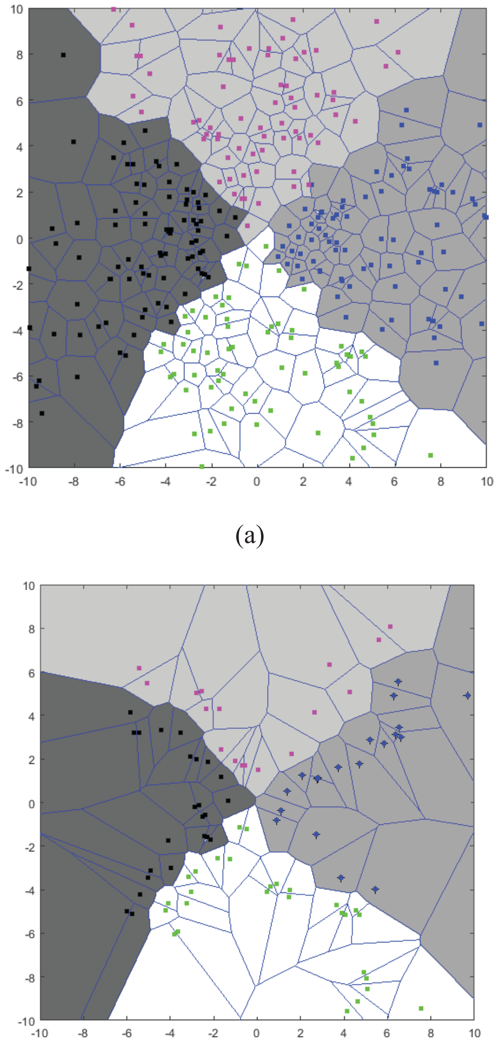


(b)

(c)

(d)

Fig. 1. Result of 2 class dataset (a) Original data (b) Romoving noise (c) Reduced class 1 (d) Reduced class 2

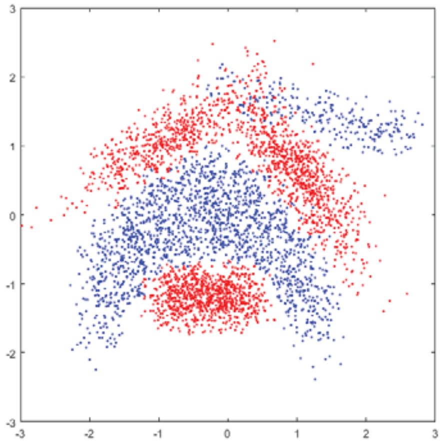


(a)

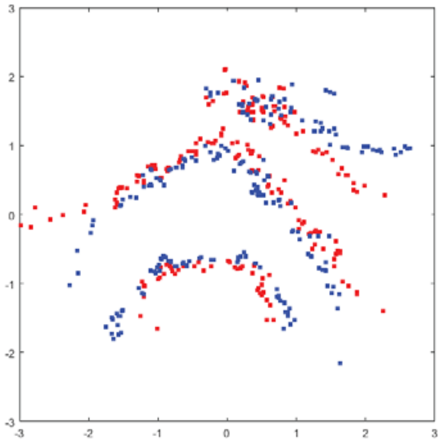
(b)

Fig. 2. Result of 4 class dataset (a) The Voronoi plot of the original data set (b) The reduced data set

Figure 2 shows the experimental result of 4 classes reduction. From the real-world data set names banana, the proposed method was applied to reduce the redundant and important samples in classification. Its result of the proposed method is shown in figure 3. The blue dot represents class 1 and the red dot represents class 2.



(a)



(b)

Fig. 3. (a) The banana data set (b) The reduced data set

The averaged results of testing accuracy, the reduction ratio, and the product of *Acc* and *RR* evaluated on 20 real-world data sets are shown in Table IV.

TABLE IV. THE AVERAGE OF THE 20 REAL-WORLD DATA SET

Methods	Accuracy (*Acc*)	Reduction Ratio (RR)	*Acc**RR
Proposed	**0.8995**	0.2388	0.2148
AllKNN	0.8299	0.2831	0.2349
CCIS	0.7339	0.9491	0.6965
CHC	0.8325	**0.9750**	**0.8116**
CNN	0.7907	0.6512	0.5149
C-Pruner	0.7460	0.8989	0.6706
DROP3	0.7660	0.8629	0.6610
FCNN	0.8028	0.7180	0.5764
GGA	0.8428	0.9150	0.7711
HMNEI	0.8122	0.5773	0.4689
IB3	0.8091	0.7205	0.5829
ICF	0.7014	0.7513	0.5270
MCNN	0.7512	0.9459	0.7106
MENN	0.7861	0.4197	0.3300
ModelCS	0.8340	0.0791	0.0660
MSS	0.8053	0.5193	0.4182
POP	0.8165	0.0897	0.0733
Reconsistent	0.7568	0.6121	0.4632
RMHC	0.8378	0.9009	0.7548
RNG	0.8292	0.1753	0.1454
RNN	0.8229	0.9301	0.7654
SSMA	0.8414	0.9591	0.8070

The scatter plot of the results in Table IV can be presented in Figure 4.

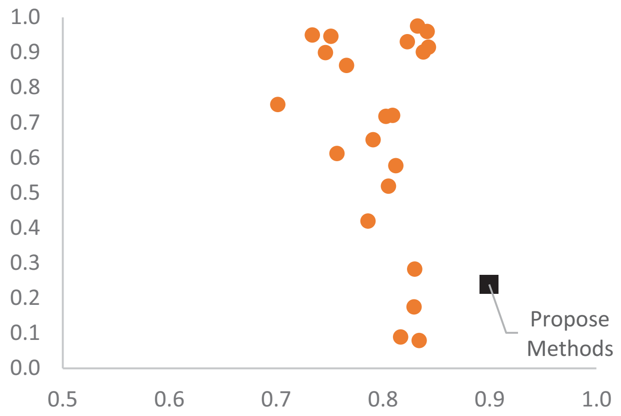


Fig. 4. The averaged results of 20 real-world data sets

VI. Conclusion

Due to a simple of the proposed methods, our processing time is quite quick. Based on the steps in the computation processes, it proves that the proposed method is categorized on the filter algorithm of the PS method that suitable for 2D data synthesis. Our proposed method was compared with the other 21 PS methods by applied to the 20 real-world multiclass data sets. The experimental results in figure 4 show that the proposed method can be used to preprocess the original data before the classification. The accuracy and its efficiency perform the best with applying on 3-kNN enemy identification. Even though the ratio of data reduction is not great, the results of our experiment which is shown in Figure 4 can be implied that our proposed method could reduce the number of real-world data for better classification.

Acknowledgment

This research is sponsored by Mae Fah Luang University.

References

[1] T.M. Cover and P.E. Hart, "Nearest Neighbor Pattern Classification," IEEE Transaction on Information Theory, vol. 13, no. 1, pp. 21-27, 1967.

[2] E. Pekalska, R.P.W. Duin, and P. Paclık, "Prototype Selection for Dissimilarity-Based Classifiers," Pattern Recognition, vol. 39, no. 2, pp. 189-208, 2006.

[3] W. Lam, C.K. Keung, and D. Liu, "Discovering useful concept prototypes for classification based on filtering and abstraction," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 8, pp. 1075 - 1090, 2002.

[4] Hatem A. Fayed and Amir F. Atiya, "A Novel Template Reduction Approach for the K-Nearest Neighbor Method," IEEE TRANSACTIONS ON NEURAL NETWORKS, vol. 20, no. 5, pp. 890-896, 2009.

[5] Olvera-López, J.A., Carrasco-Ochoa, J.A. and Martínez-Trinidad, J.F., "A new fast prototype selection method based on clustering," Pattern Analysis and Applications, vol. 13, no. 2, pp. 131-141, 2010.

[6] Elena Marchiori, "Class conditional nearest neighbor for large margin instance selection," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 2, pp. 364-370, 2010.

[7] K. Nikolaidis, J.Y. Goulermas, Q.H. Wu, "A class boundary preserving algorithm for data condensation," Pattern Recognition, vol. 44, no. 3, pp. 704-715, March 2011.

[8] N. Verbiest, C. Cornelis and F. Herrera, "FRPS: A Fuzzy Rough Prototype Selection method," Pattern Recognition, vol. 46, no. 10, pp. 2770-2782, 2013.

[9] Javad Hamidzadeh, Reza Monsef, Hadi Sadoghi Yazdi, "LMIRA: Large Margin Instance Reduction Algorithm," Neurocomputing, pp. 477-487, 2014.

[10] K. Nikolaidis, T. Mu and J.Y. Goulermas, "Prototype reduction based on Direct Weighted Pruning," Pattern Recognition Letters, vol. 36, no. 1, pp. 22-28, 2014.

[11] E. Leyva, A. González, R. Pérez, "Three new instance selection methods based on local sets: A comparative study with several approaches from a bi-objective perspective," Pattern Recognition, vol. 48, no. 4, pp. 1519-1533, 2015.

[12] D.R. Wilson and T.R. Martinez, "Reduction Techniques for Instance-Based Learning Algorithms," Machine Learning, vol. 38, no. 3, pp. 257-286, 2000.

[13] Garcı´a Salvador, "Prototype Selection for Nearest Neighbor Classification: Taxonomy and Empirical Study," IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, vol. 34, no. 3, pp. 417-435, 2012.

[14] S. Jayaprada, P. Krishna Prasad, R. Satya Prasad, “Enriched Semantic Similar Frequent Patterns using SSFPOA Neighborhood Ranking Algorithm,” International Journal for Modern Trends in Science and Technology, vol.03, no. 09, September 2017