Merging CLU and PSR Methods for Accurate Prototype Selection

Shih-Wen Ke*, Dewi Wulandari
Dept. of Information and Computer Engineering, Chung Yuan Christian University, 200 Chung Pei Road, Chung Li District, Taoyuan City, Taiwan.

* Corresponding author. Tel.: +886 3 265 4729; email: swgke@ice.cycu.edu.tw
Manuscript submitted February 20, 2016, accepted April 15, 2016.
doi: 10.17706/jsw.11.7.638-645

Abstract: Data reduction in data mining is very important when dealing with large dataset. Through data reduction one can increase storage efficiency and reduce the run time of data mining process. One of the methods to reduce the volume of data is to selectively retain a subset of the dataset as the representation of the original dataset. This methods is known as prototype selection. Prototype selection aims to discard the superfluous instances in training set, because superfluous instances will influence the result in data mining. In this study we proposed a hybrid prototype selection method using clustering algorithm and selected the most relevant subset of cluster members. The results showed that the hybrid approach performed better than the original methods used individually.

Key words: Data mining, data reduction, fuzzy c-means clustering, prototype selection methods

1. Introduction

Mining huge datasets can raise many problems, such as high cost, superfluous data, high run time mining process, etc. Data reduction can resolve and handle this problem, because reducing the size of dataset can result in [1]:

- Increased capabilities and generalization properties of the classification model;
- Reduction in space complexity of the classification problem;
- Reduction in computational time;
- Diminishing the size of formulas obtained by an induction algorithm on the reduced data sets.

Traditionally, the concept of data reduction has received several names, e.g. editing, condensing, filtering, thinning, etc. depending on the object of the reduction. One of the goals of data reduction is to reduce the quantity of instances. The process of reducing the quantity of instances in dataset, particularly in training set, is known as prototype selection.

Therefore through prototype selection process one can selectively reduce the size of the dataset. Prototype selection method is to obtain a subset of training set, which aims to discard the superfluous instances [2]. There are many prototype selection methods proposed by researchers with various algorithms. Based on the strategy used for selecting instances, prototype selection can be divided into two approaches [2], [3]:

- Wrapper: The methods are based on classifier algorithm.
- Filter: The methods are not based on classifier algorithm.

Section 2describes some prototype selection methods, our proposed method and the experimental

setups will be described in Sections 3 and 4, and the results of experiment will be shown in Section 5 followed by the conclusion in Section 6.

2. Related Work

One of the earliest wrapper prototype selection methods is the Condensed Nearest Neighbor (CNN) [4] CNN starts with an empty subset S() of training set (T), and by randomly picking one instance from the training set to S. Iterating through each instance in training set using the nearest neighbor algorithm [5] to find the nearest instance in S. If the instance classified no correct the transfer instance to S. This iteration will stop until no instance is moved to S. Selective Nearest Neighbor (SNN) [5] is an extension of CNN with improved qualities, such that it produces a more consistent S, in which the distance between any instance and its nearest selective neighbor is less than distance from the instance of the other class, and the S is the smallest possible. Another variant of CNN is the Generalized Condensed Nearest Neighbor (GCNN) [6], which is identical to CNN but GCNN includes in S instances that satisfy an absorption criterion according to a threshold.

Another early wrapper prototype selection methods is the Edited Nearest Neighbor (ENN) [7]. ENN starts with S equal to T and for each instances in S find its k-NN with k=3. If a given instance is not same as the maximum class, the instance is removed from S. Another variant of ENN is the all k-NN algorithm [8] which the k-NN algorithm applied after ENN algorithm and repeated ENN algorithm until no removed instances pass through iteration all instances in S. After finished repeating the ENN algorithm, each remaining instance in S is flagged to 1. Then find the k-NN of instance, if the instance classify incorrectly change flag to 0. And the end removed all instances with flag 0.

Other wrapper methods related to k-NN are decremental Reduction Optimization Procedure (DROP) [9]. DROP algorithm employ k-NN algorithm by maintaining list of k+1 and based on the concept of associate. The DROP1 is the basic of DROP algorithm and develop the variant of DROP1 with name DROP2, DROP3, DROP4 and DROP5. DROP1 will remove instance (p) if at least as many of its associates in S would be classified correctly without p. DROP2 is similar to DROP1, but in DROP2 will removed p in S only if at least as many of its associate including those instance that may have already been removed from S are classifies correctly without p. DROP2 also changes the order of removal of instance. It initially sorts the instances in S by the distance to their nearest enemy, and removed the instance begin from the furthest its nearest enemy, with purpose can increase the chance of retaining border points. DROP3 uses a noise filtering pass before sorting the instance in S in DROP2. This noise filtering is done using a rule similar to ENN: any instances misclassified by its k nearest neighbor is removed. DROP4 is identical to DROP3 except instead of applying ENN, the noise filtering pass removing each instance only if it is misclassified by its k nearest neighbors and it does not hurt the classification of other instances. The last variant DROP5 is modifies DROP2, if in DROP2 removal from the furthest its nearest enemy, but DROP5 beginning with instance nearest from its nearest enemy. By removing instance with the nearest it will be smooth the decision boundary.

Prototype selection based on clustering (CLU) [10] is one of the example of filter methods based on clustering. The purpose using clustering algorithm for finding subgroups of similar instances. CLU employ Fuzzy c-means algorithm [11] divided the training set into k-clusters and stored the cluster's centroid to S.

The other prototype selection filter methods are Prototype selection via relevance (PSR) [12]. This method come from the paradigm there are some instances which are more similar than others in the same class. The most similar instances could be more representative or relevant than the less similar ones, then it makes sense of prototype selection to retain the most relevant instance. The relevance of each instances is calculate by average similarity from the other instances in the some class. The relevance of an instance p is given in terms of the average similarity A_N which is computed as below:

$$A_N(p) = \frac{\sum_{p' \epsilon C, p \neq p'} S(p, p')}{|C| - 1}$$

where:

- C is the set of training instances belonging to the same class with p.
- $S(p, p')$ is a similarity function for comparing prototypes.

PSR algorithm computes the relevance of each instance and retains the most relevance ones denoted by r for each class in T. Additionally, in order to preserve the discrimination regions between classes, PSR also retain border instances which are found through the most relevance instances. To observation this PSR algorithm Olvera et al using ten small datasets and three medium-large datasets which are taken from UCI Repository. For the similarity function they are used Heterogeneous Value Difference Metric (HVDM) [13] with the reason HVDM works over numeric, non-numeric and missing features.

3. Proposed Methods

Our proposed prototype selection algorithm is from the CLU algorithm and PSR algorithm and we called it CLU-R. The CLU-R begins with dividing the training set into k-clusters by employing the FCM algorithm. Although the runtime for the FCM algorithm is usually longer than the k-means algorithm, the FCM algorithm can produce good clusters [14],[15]. After dividing the training set into k clusters, from each cluster we find the total number r of member instances for representing the cluster. These instance are the most relevant ones kept while the other instances in the cluster are discarded. The relevance value of each instance in the k^{th}-cluster is given by their average of weight similarity toward the all member in the same cluster, which is borrowed from the idea of the PSR algorithm [12]. For the similarity function, Euclidean distance was chosen. The illustration of CLU-R is shown in Fig. 1.

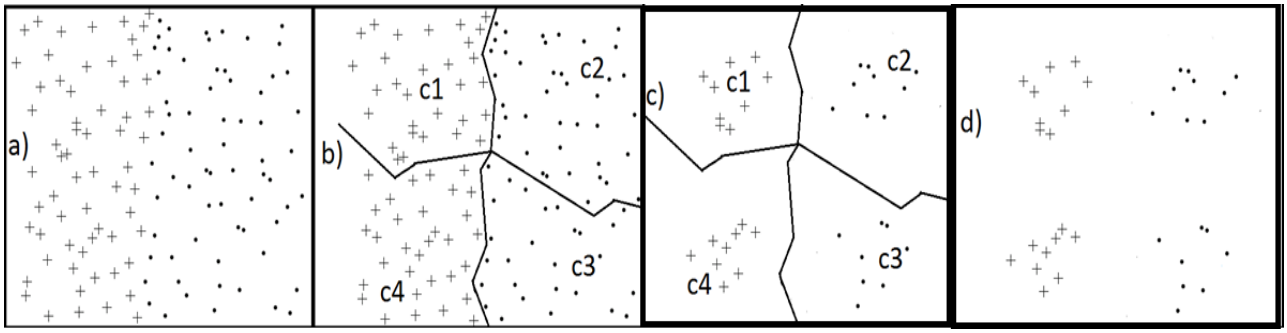

Fig. 1. Illustration of CLU-R.

4. Experimental Setups

4.1. Dataset

Table 1. Detailed Information of the Datasets Used in This Study

Dataset	# of instances (training/test)	# of attributes	# of classes	Missing values	Attribute type
Iris	105/45	4	3	No	Numerical
Car evaluation	1211/517	6	4	No	Categorical
Adult	32,561/16,281	14	2	No	Mixed
Localization data for person activity dataset	115,406/49,454	8	11	No	Numerical
Skin Segmentation	171,541/73,516	4	2	No	Numerical

Datasets are taken from the UCI dataset repository. Iris, car evaluation, adult, localization data for person

activity, and skin segmentation dataset have been chosen. The datasets chosen by considering of the total number of instance, number of attribute, and number of classes. The detailed information of the datasets is given in Table 1.

4.2. Software and Equipment

R was chosen for building the experiment. R is a language and environment for statistical computing and graphics. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, etc.) and graphical techniques, and is highly extensible. The advantages of R are as follow:

1) Open source code
2) Available for Windows, Linux and Mac
3) Extensive support extensions
4) APIs with other language, such as C

R tool was installed in Windows 7 Professional operating system, with processor Pentium(R) Dual-Core, RAM 4GB, 32-bit.

The result recorded with value k-clusters are 4c, 8c, 10c, 12c (where c is the number of classes in the given dataset) and for each value k-clusters recorded for different percentage total instances most relevance with r is 10%, 20%, 30%, 40% and 50%. Each pair of parameters k and r was repeated for five times. FCM algorithm with parameter fuzzification and maximum iteration was 500.

4.3. Pre-processing

Data preprocessing in data mining is process where the data are prepared for mining [16]. In this preprocessing data used for understanding data, then manage missing value if occurs. To handle missing values by ignored or removed object is the most common for resolved missing values [17, 18]. Because this method very simple and easiest, also not the focus on this research, therefore this method will be used in my experiment. The next step converting attribute categorical to numerical, and the last if the dataset no has test data, test data come from separate the dataset for 70% training set and 30% test set for each classes by random (in this case except adult dataset because from the original already separated between training set and test set). Whereas prototype selection methods (CLU, PSR and CLU-R) will be applied in training set to produce new subset training set, then the test set will be used for evaluation and measure the accuracy of the new subset training set. Table 1 shows the total number of instances of training set and test set for each dataset.

5. Experimental Results

Table 2. Accuracy of CLU-R, k=4c, r=10% ~ 50%

Dataset	r=10%	r=20%	r=30%	r=40%	r=50%
Iris	0.9556	0.9556	0.9778	0.9556	0.9333
Car evaluation	0.7234	0.7582	0.824	0.8569	0.8704
Adult	0.7141	0.7737	0.7614	0.7719	0.7665
Localization data for person activity dataset	0.5296	0.5458	0.5561	0.5775	0.5896
Skin Segmentation	0.9514	0.9511	0.9532	0.9537	0.9547
Average:	0.7748	0.7969	0.8146	0.8231	0.8229

Table 2 to Table 4 are the results of the accuracy of the CLU-R algorithm, and show that pair of parameter k=12c and r=50% give the average accuracy better than others, therefore this pair of parameter value's

result will be compared with CLU and PSR algorithm as a baseline algorithm. CLU algorithm result is shown in Table VI, and PSR algorithm result shown on Table VII. From the CLU accuracy result that k=10c give the best accuracy than other, therefor CLU k=10c taken as comparable. Whereas the PSR accuracy using r=40%. The comparison table between three methods CLU, PSR and CLU-R shown in Table 3.

Table 3. Accuracy of CLU-R, k=8c, r=10% ~ 50%

Dataset	r=10%	r=20%	r=30%	r=40%	r=50%
Iris	0.8889	0.9778	0.9556	0.9556	0.9556
Car evaluation	0.7369	0.7698	0.795	0.8356	0.8704
Adult	0.7208	0.7294	0.7522	0.7278	0.7258
Localization data for person activity dataset	0.5536	0.5664	0.5804	0.5927	0.6034
Skin Segmentation	0.9643	0.9634	0.9639	0.9698	0.9733
Average:	0.7729	0.8014	0.8094	0.8163	0.8257

Table 4. Accuracy of CLU-R, k=10c, r=10%50%

Dataset	r=10%	r=20%	r=30%	r=40%	r=50%
Iris	1	1	0.9333	0.9556	0.9556
Car evaluation	0.7079	0.7621	0.8027	0.8472	0.8665
Adult	0.7288	0.7536	0.7318	0.7387	0.7537
Localization data for person activity dataset	0.5637	0.5757	0.5852	0.5955	0.609
Skin Segmentation	0.9594	0.9591	0.9629	0.9622	0.969
Average:	0.7920	0.8101	0.8032	0.8198	0.8308

Table 5. Accuracy of CLU-R, k=12c, r=10% ~ 50%

Dataset	r=10%	r=20%	r=30%	r=40%	r=50%
Iris	0.8889	0.9556	0.9333	0.9778	0.9556
Car evaluation	0.7195	0.7563	0.8143	0.8414	0.8743
Adult	0.7441	0.7434	0.7543	0.7345	0.7565
Localization data for person activity dataset	0.5655	0.5775	0.5867	0.5997	0.6099
Skin Segmentation	0.9701	0.9747	0.9783	0.9778	0.9805
Average:	0.7776	0.8015	0.8134	0.8263	0.8354

Table 6.Accuracy of CLU, k=4c, 6c, 8c, 10c, 12c, r=50%

Dataset	k=4c	k=6c	k=8c	k=10c	k=12c
Iris	0.9778	0.9556	0.9778	0.9778	0.9778
Car evaluation	0.6867	0.7021	0.6925	0.6963	0.6673
Adult	0.7637	0.7637	0.7637	0.7637	0.7637
Localization data for person activitydataset	0.4498	0.5156	0.5067	0.5312	0.5443
Skin Segmentation	0.8468	0.9276	0.9423	0.9403	0.9391
Average:	0.7450	0.7729	0.7766	0.7819	0.7784

Table 7. Accuracy of PSR, k=4c, 6c, 8c, 10c, 12c, r=50%

Dataset	k=4c	k=6c	k=8c	k=10c	k=12c
Iris	0.9111	0.8889	0.9333	0.9556	0.9556
Car evaluation	0.735	0.7621	0.795	0.8066	0.8201
Adult	0.4946	0.5175	0.5469	0.774	0.6073
Localization data for person activity dataset	0.2723	0.331	0.4554	0.4882	0.4915
Skin Segmentation	0.8926	0.8958	0.8984	0.9055	0.9146
Average:	0.6611	0.6791	0.7258	0.7860	0.7578

As reported in Table 2 to Table 4, the average of the accuracy CLU-R outperformed CLU and PSR algorithms. Although in this experiment the three of prototype selection algorithms still cannot outperform the dataset without prototype selection process. But in here for a huge dataset still recommended for applied data reduction or prototype selection methods, because from our analysis in the runtime process for total of instances more than 245,057 instances such as skin segmentation dataset can reduce the runtime mining process by more than two folds as seen in Fig. 2.

Table 8. Accuracy of PSR, k=4c, 6c, 8c, 10c, 12c, r=50%

Dataset	Original (w/o prototype selection)	CLU	PSR	CLU-R
		k=10c	r=40%	k=12c, r=50%
Iris	0.9111	0.9778	0.9556	0.9556
Car evaluation	0.9342	0.6963	0.8066	0.8743
Adult	0.7595	0.7637	0.774	0.7565
Localization data for person activity dataset	0.6655	0.5312	0.4882	0.6099
Skin Segmentation	0.9994	0.9403	0.9055	0.9805
Average:	0.85394	0.78186	0.78598	0.83536

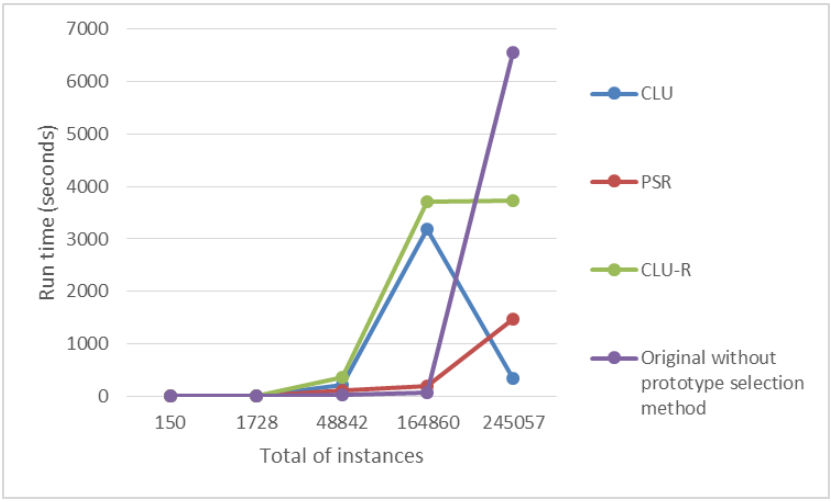

Fig. 2. Time taken to classify all test data between CLU, PSR, CLU-R and original.

6. Conclusions

Prototype selection based on clustering can produce smooth distribution, such as our CLU-R and CLU algorithms. But CLU stored the subset T with centroid of each cluster, this has the drawback with big probability the new subset T will dismiss some classes because the new subset has very few members. PSR algorithm does not dismiss some classes in new subset T because PSR algorithm takes the most relevant instances for each classes. However, the PSR does not produce smooth distribution like the CLU algorithm does. Therefore the CLU-R algorithm can produce new subset T with smooth distribution and is less likely to dismiss some classes. And CLU-R through this experiment with five datasets proven with merging CLU (clustering) and PSR (most relevance) can increase the accuracy.

Acknowledgment

This study is partly funded by the Ministry of Science and Technology in Taiwan, grant number: 104-2221-E-033-043.

References

[1] Czarnowski I., & Jedrzejowicz P. (2008).Data reduction algorithm for machine learning and data mining. 276-285.
[2] Lopez,J.A. O. (2010). Prototype selection methods.*Computation System*, *13(4),* 449-462.
[3] Lopez,J.A. O., Ochoa,J.A. C., TrinidadJ.F. M.,& Kittler, J.(2010). A review of instance selection methods. *Artif Intell Rev, 34,* 133-143.
[4] Hart, P.E. (1968). The condensed nearest neighbor rule.*IEEE Trans Inf Theory, 14,* 515-516.
[5] Ritter, G. L., WoodruffH. B., Lowry,S.R., & Isenhour, T. L. (1975). An algorithm for a selective nearest neighbor decision rule.*IEEE Trans Inf Theory, 21(6),* 665-669.
[6] Chien-Hsing,C., Bo-Han, K.,& Fu, C.(2006). The generalized condensed nearest neighbor rule as a data reduction method. *Proceedings of the 18th International Conference on Pattern Recognition* (pp. 556-559). IEE Computer Society, Hong-Kong.
[7] Wilson, D. L. (1972). Asymptotic properties of nearest neighbor rules using edited data.*IEEE Trans Systems Man Cybernetics, 2,* 408-421.
[8] Tomek,I. (1976). An experiment with the edited nearest neighbor rule.*IEEE Trans Syst Man Cybern,* 6-6, 448-452.
[9] Wilson, D. R., & Martinez,T. R. (2000). Reduction techniques for instance-based learning algorithms. *Mach Learn, 38,* 257-286.
[10] Lumini, A., & Nanni, L. (2006). A clustering methods for automatic biometric template selection. *Pattern Recognit, 39,* 495-497.
[11] Bezdek,J.C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithm.* Plenum Press, New York.
[12] Lopez, J.A. O., Ochoa,J.A. C, & Trinidad, J.F. M. (2008). *Prototype Selection via Prototype Relevance.* Habana, Cuba, 153-160.
[13] Wilson, D.R., & Martinez, T.R. (1997). Improved heterogeneous distance functions.*Journal of Artificial Intelligence Research*, *6(1),* 1-34.
[14] Gayathri, A.,& Mohanavalli, S. (May 2011). Enchanced customer relationship management using fuzzy clustering. *IJCSET, 1(4),* 163-167.
[15] Shaaban,H. R. M., Obaid,F. A.,& Habib, A. A. (2015). Performance evaluation of k-mean and fuzzy c-mean image segmentation based clustering classifier.*International Journal of Advanced Computer Science and Applications, 6(12).*
[16] Han, J., Kambe,r M.,Pei,J. (2012). *Data Mining Concepts and Techniques* (3rd ed.). USA: Elsevier Inc.
[17] Pigott, T. D. (2001). A review of methods for missing data.*Education Research and Evaluation*, *Educational Research and Evaluation, 7(4),* 353-383.
[18] Shalabi,L.A., Najjar, M.,& Kayed,A. A. (2006). A framework to deal with missing data in data sets. *Journal of Computer Science, 2(9),* 740-745.

Shih-Wen Ke currently is an assistant professor in the Dept. of Information and Computer Engineering at Chung Yuan Christian University in Taiwan. Dr. Ke received his PhD from University of Sunderland, UK in 2008. He has been working as an IT development consultant at the Faculty of Medicine, University of Southampton, UK since then.

He also worked as a technology consultant for AES Digital Solutions Ltd., which is a UK-based company specialized in providing ERP, CRM and multimedia solutions for various industries. The areas of his expertise include multimedia information systems, database-driven IT systems, web programming, eLearning applications and IT project management.

Dr. Ke's latest research interests include multimedia applications, mobile learning with social network technologies, information retrieval, machine learning, natural language processing, serious games and human computer interaction.

Dewi Wulandari currently is a postgraduate student studying at the Dept. of Information and Computer Engineering, Chung Yuan Christian University. Her research interests include machine learning, statistical analysis and data mining.