Survey of New Approaches on Prototype Selection and Generation

Joaquín Derrac, Isaac Triguero, Salvador García and Francisco Herrera

Abstract

Prototype Selection and Prototype Generation are two lively research fields. As time goes by, new methods are developed following the basic guidelines that any Prototype Reduction algorithm should accomplish. These new methods often offer to the comunity different viewpoints to the problem, as well as better and more efficient approaches.

This technical report provides a survey of the newest Prototype Reduction approaches, highlighting their main characteristics. These approaches are given a place in the already proposed taxonomies, depending of their specific behavior. With their inclusion, both Prototype Selection and Prototype Generation taxonomies will be fulfilled, representing properly all the Prototype Reduction methods proposed in the literature.

Index Terms

Prototype selection, prototype generation, nearest neighbor, taxonomy.

I. Introduction

The nearest neighbors classifier is one of the most used and known techniques for performing recognition tasks. It has also demonstrated to be one of the most interesting algorithms in the data mining field in spite of its simplicity. However, the nearest neighbors classifier suffers from several drawbacks; such as high storage requirements, low efficiency in classification response and low noise tolerance. These weaknesses have been the subject of study of many researchers and many solutions have been proposed.

This technical report is provided as a complement of the Prototype Selection and Prototype Generation taxonomies already established in [1] and [2]. Recent methods, published after both reviews, are reviewed and categorized within the proposed taxonomies. Section II deals with the new Prototype Selection approaches, whereas Section III surveys the Prototype Generation proposals related.

J. Derrac, I. Triguero and F. Herrera are with the Department of Computer Science and Artificial Intelligence, CITIC-UGR (Research Center on Information and Communications Technology), University of Granada, 18071 Granada, Spain.
E-mails: jderrac@decsai.ugr.es, triguero@decsai.ugr.es, herrera@decsai.ugr.es

S. García is with the Department of Computer Science, University of Jaén, 23071, Jaén, Spain.
E-mails: sglopez@ujaen.es

II. New Prototype Selection Approaches

A. Condensation - Incremental

- **Complete Cross Validation Functional Algorithm 2 (CCV-2) [3]**:
  This method begins by adding certain $K + 1$ instances into the reference set. Then the algorithm begins successive addition of instances with respect to their Complete Cross Validation (CCV) measure (a measure which characterizes the generalization ability of a classifier). If, at a certain step, it becomes impossible to add any instance without reducing the functional of the CCV measure , the algorithm finishes.

B. Condensation - Decremental

- **Complete Cross Validation Functional Algorithm 1 (CCV-1) [3]**: In a similar way that CCV-2, the CCV-1 method makes use of the CCV measure for evaluating the quality of a given reference set.
  In this case, the algorithm includes 2 removing stages. The first one aims to exclude outlaiers from the training set. After all outliers are removed, the second stage removes periphery instances, that is, those instances lying far from the boundary and surrounded by objects of the same class. After removing noise and periphery objects, the algorithm finishes.
- **Instance Selection by using Polar Grids (ISPG) [4]**: This methods works by mapping the original training set into a Polar coordinate system, divide the data space into a certain number of polar grids. In each grid, noisy and inner instances are removed, whereas boundary instances are selected for becoming part of the reference set.

C. Edition - Decremental

- **Instances Selection algorithm based on Classification Contribution Function (ISCC) [5]**: This method inherits the *Reachable* and *Coverage* concepts of the ICF methods, extending it through the use of two new measures denoted *DReachable* and *DCoverage*, which represents the local competence of each instance regarding the different classes of the problem. Instances whose *DReachable* and *DCoverage* measures drop below a given threshold are edited from the reference set.
- **Instance Selection based on Local SVM (LSVM) [6]**: In this method, Local Support Vector Machines (LSVM) are used to evaluate the quality of each training instance. For each training example an SVM is trained on its neighborhood and if the SVM classification for the central example disagrees with its actual class, the instance is removed.

D. Edition - Batch

- **RewardPunishment Editing (RP-Editing) [14]**: In this method, all the instances of the training set are ranked regarding three different measures: **WR(i)**, which estimates how many instances are positively affected by the instance (that is, is a nearest neighbor from the same class); **WP(i)**, which estimates how many instances are

negatively affected by the instance (that is, is a nearest neighbor from a different class); and **WPR(i)**, which denotes the number of times the instance is correctly classified applying the KNN to the prototypes selected from a clustering algorithm over the training set. Those instances whose rank falls down of a given threshold are removed.

E. Hybrid - Mixed+Wrapper

- **Instinctive Mating Genetic Algorithm with (Correct My Wrongs distance/Hamming distance/Multiple populations) (IM-CMW, IM-H, IM-MP) [7]**: These are three versions of a modified Genetic Algorithm for Prototype Selection Algorithm. The Instinctive Mating procedure of the Genetic Algorithm relates to the way in which the parents are chosen before applying the crossover operator. In this way, three approaches were developed using the Correct-My-Wrongs distance (IM-CMW), the Hamming distance (IM-H) and a multiple population version (IM-MP).
- **Supervised and Nonparametric Evaluation of Sets of Instances (Eva) [8]**: A Bayesian framework based method which uses a Variable Neighborhood Search strategy to choose the best possible reference subset according to a probabilistic criterion.
- **Local Search with Tabu List for Instance Selection (LS+TL) [9]**: Four different clustering methods (based on a similarity coefficient, a stratification strategy, a modified stratification strategy or a k-means clustering) are use d to split the training set into different clusters. Then, a Local Search with Tabu List is applied to each cluster for selecting the best training instances in each one. Finally, all solutions are merged to produce the final reference set.
- **Sequential Reduction Algorithm (SeqRA) [10]**: A random selection technique which iteratively tries to add and subtract instances from the reference set until no further improvement is found. The instances are checked sequentially regarding their class.

F. Hybrid - Fixed+Wrapper

- **A Genetic Algorithm for Prototype Selection with Dissimilarity Representation (GA+LDA) [11]**: A classic Genetic Algorithm for Prototype Selection developed for training sets stored using a Dissimilarity based representation. This algorithm requires to fix the number of training instances which will be selected.

III. New Prototype Generation Approaches

A. Positioning Adjustment - Condensation - Fixed

- **Weighted LVQ (WLVQ) [12]**: A Context Dependent Clustering algorithm is employed as a preprocessing step in this method. The output is a set of weights which are applied to modify the cost function of the LVQ algorithm, improving its performance.

B. Positioning Adjustment - Edition - Mixed - Semi-Wrapper

- **Editing based on a Fast Two-String Median Computation (JJWilson) [13]**: This method is a modification of the Wilson's editing algorithm (ENN). Instances marked as noise by the original algorithm are deleted only if none of its nearest neighbors are from its same class. If a nearest neighbor of the same class can be found, a new example is generated at the median point between both instances.

C. Centroids - Hybrid - Mixed

- **Class Boundary Preserving Algorithm (CBP) [15]**: This Prototype Generation method performs four steps:
  - Step 1: The ENN method is applied over the training set.
  - Step 2: The instances of the reference set are divided into boundary and inner instances.
  - Step 3: The set of boundary instances is reduced, maintaining only those pair of mutual enemies with minimum distance between them.
  - Step 4: The Mean Shift Clustering algorithm is employed over the set of inner instances, generating a new set of prototypes.

  The sets obtained through the application of the steps 3 and 4 are merged to create the final reference set.

D. Space Splitting - Hybrid

- **Instance Seriation for Prototype Abstraction (IPSA) [16]**: This algorithm starts by applying the ENN method over the training set. Then, a Visual Assessment of Cluster Tendency is employed to serialize instances into a minimum spanning tree. This tree is iteratively merged, finally generating the instances of the reference set.

REFERENCES

[1] S. García, J. Derrac, J. Cano, and F. Herrera, "Prototype selection for nearest neighbor classification: Taxonomy and empirical study," *IEEE Transactions on Pattern Analysis and Machine Intelligence, In press*.

[2] I. Triguero, J. Derrac, S. García, and F. Herrera, "A taxonomy and experimental study on prototype generation for nearest neighbor classification," *IEEE Transactions on Systems, Man, and Cybernetics–Part C: Applications and Reviews, In press*.

[3] M. N. Ivanov, "Prototype sample selection based on minimization of the complete cross validation functional," *Pattern Recognition and Image Analysis*, vol. 20, no. 4, pp. 427–437, 2010.

[4] Y. Sang and Z. Yi, "Instance selection by using polar grids," in *Third International Coriference on Advanced Computer Theory and Engineering*, vol. 3, 2010, pp. 344–348.

[5] Y. Cai, B. Wu, Y. He, and Y. Zhang, "A new instance selection algorithm based on contribution for nearest neighbour classification," in *Ninth International Conference on Machine Learning and Cybernetics*, 2010, pp. 155–160.

[6] N. Segata, E. Blanzieri, S. J. Delany, and P. Cunningham, "Noise reduction for instance-based learning with a local maximal margin approach," *Journal of Intelligent Information Systems*, vol. 35, pp. 301–331, 2010.

[7] A. Franco, D. Maltoni, and L. Nanni, "Data pre-processing through rewardpunishment editing," *Pattern Analysis and Applications*, vol. 13, pp. 367–381, 2010.

[8] T. Quirino, M. Kubat, and N. J. Bryan, "Instinct-based mating in genetic algorithms applied to the tuning of 1-nn classifiers," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 12, pp. 1724–1737, 2010.

[9] S. Ferrandiz and M. Boullé, "Bayesian instance selection for the nearest neighbor rule," *Machine Learning*, vol. 81, pp. 229–256, 2010.

[10] I. Czarnowski, "Cluster-based instance selection for machine classification," *Knowledge and Information Systems, In press*.

[11] M. Raniszewski, "Sequential reduction algorithm for nearest neighbor rule," in *International Conference on Computer Vision and Graphics, Part II*, vol. 6375. Lecture Notes in Computer Science, 2010, pp. 219–226.

[12] Y. Plasencia-Calaña, E. García-Reyes, M. Orozco-Alzate, and R. P. Duin, "Prototype selection for dissimilarity representation by a genetic algorithm," in *International Conference on Pattern Recognition*, 2010, pp. 177–180.

[13] M. Blachnik and W. Duch, "Improving accuracy of lvq algorithm by instance weighting," in *International Conference on Artificial Neural Networks*, vol. 6354. Lecture Notes in Computer Science, 2010, pp. 257–266.

[14] J. I. Abreu and J. R. Rico-Juan, "A new editing scheme based on a fast two-string median computation applied to ocr," in *Structural, Syntactic, and Statistical Pattern Recognition*, vol. 6218. Lecture Notes in Computer Science, 2010, pp. 748–756.

[15] K. Nikolaidis, J. Y. Goulermas, and Q. H. Wu, "A class boundary preserving algorithm for data condensation," *Pattern Recognition*, vol. 44, pp. 704–715, 2011.

[16] K. Nikolaidis, E. Rodriguez, J. Y. Goulermas, and Q. H. Wu, "Instance seriation for prototype abstraction," in *Fifth International Conference on Bio-Inspired Computing: Theories and Applications*, 2010, pp. 1351–1355.