Survey of Nearest Neighbor Condensing Techniques

MILOUD-AOUIDATE Amal (corres. Author)

Computer Sciences Department
University of Sciences and Technology Houari
Boumediene, USTHB,
Algiers, ALGERIA

BABA-ALI Ahmed Riadh

Electronic Department
University of Sciences and Technology Houari
Boumediene, USTHB,
Algiers, ALGERIA

Abstract—**The nearest neighbor rule identifies the category of an unknown element according to its known nearest neighbors' categories. This technique is efficient in many fields as event recognition, text categorization and object recognition. Its prime advantage is its simplicity, but its main inconvenience is its computing complexity for large training sets.**
This drawback was dealt by the researchers' community as the problem of prototype selection.
Trying to solve this problem several techniques presented as condensing techniques were proposed. Condensing algorithms try to determine a significantly reduced set of prototypes keeping the performance of the 1-NN rule on this set close to the one reached on the complete training set. In this paper we present a survey of some condensing KNN techniques which are CNN, RNN, FCNN, Drop1-5, DEL, IKNN, TRKNN and CBP.
All these techniques can improve the efficiency in computation time. But these algorithms fail to prove the minimality of their resulting set. For this, one possibility is to hybridize them with other algorithms, called modern heuristics or metaheuristics, which, themselves, can improve the solution. The metaheuristics that have proven results in the selection of attributes are principally genetic algorithms and tabu search. We will also shed light in this paper on some recent techniques focusing on this template.

Keywords- Nearest neighbor (NN); kNN; Prototype selection; Condensed NN; Reduced NN; Condensing; Genetic algorithms; Tabu search.

I. INTRODUCTION

The K-nearest neighbor classification rule (KNN) proposed by T. M. Cover and P. E. Hart [4], is a powerful classification method that allows an almost infallible classification of an unknown prototype through a set of training prototypes. It is widely used in pattern recognition [20] [18], text categorization [10] [6], object recognition [8] and event recognition [23] applications.

An inevitable consequence of large sets of prototypes is the computational time implied by this research problem. The databases, used in some areas such as intrusion detection, are constantly and dynamically updated. This constitutes one of the main inconveniences of the KNN rule. Another important inconvenience comes from the fact that the training prototypes can contain noisy or mislabeled prototypes that may affect the results and distort them. The scientific community has tackled these problems and proposed a selection of prototypes which could modify an initial set of prototypes by reducing its size in order to improve the classification performance.

II. PROTOTYPE SELECTION

Prototype selection is the process of finding representative patterns from the data, which can help in reducing these data. This problem is classified as an NP-hard problem by many researchers [1] [25], because there is no polynomial algorithm allowing for the solution. The existing algorithms can just give acceptable solutions.

Like many other combinatorial problems, the prototype selection (PS) would require an exhaustive search to obtain optimal solutions in the general case. This has led some researchers to consider the problem of PS as a combinatorial optimization problem and use general techniques which are known for their good results in similar situations.

Heuristics, especially the nearest neighbor algorithm, and metaheuristics, especially genetic algorithms (GA) and tabu search (TS) have been proposed to solve this problem.

III. IMPROVING PROTOTYPE SELECTION

Condensing algorithms try to find a significant reduction of all prototypes so that the 1-NN classification gives results as close as possible to those obtained using all the original prototypes **[7]**. The problem that arises when this approach is used is that it cannot provide a proof about the resulting sets minimality. To correct this slight defect, the modern heuristics or metaheuristics came to complete them. The metaheuristics are strategies that guide the search towards an optimal solution. These techniques are designed to explore the search space efficiently in order to determine solutions (almost) optimal. They may contain mechanisms to avoid blocking in areas of space research.

Two types of metaheuristics have been successful in their hybridization with the traditional KNN: genetic algorithms and tabu search.

Genetic algorithms (GA) are an optimization technique guided by the principles of natural evolution and genetics, with a high presence of implicit parallelism. These algorithms perform a search in complex, large, and multimodal landscapes, and provide solutions for the quasi-optimal objective function.

The tabu search (TS) is a method of dynamic neighborhood, which selects, at each iteration, the best solution of the first local optimum by finding the best neighbor.

IV. HEURISTIC METHODS BASED ON THE CONDENSING K-NN RULE

In 1968 Hart [17] was the first to propose a method reducing the size of stored data for the nearest neighbor decision. This method is called "*The Condensed Nearest Neighbor Rule*" (CNN). The new idea about this rule compared to the traditional KNN is the process of selecting a subset TCNN from the initial training set TNN. TCNN must be as effective as TNN in the classification of unknown patterns. Actually the CNN rule minimizes the number of models stored, keeping only a subset of training data for classification, and employing a technique of low absorption. The basic idea is to look for very similar training models, and those that do not add additional information and eliminate them.

This rule presents a case where TCNN consistency is not achieved, when TNN is already the minimum set. In this case TCNN will be equal to TNN, if this happens, the algorithm will end if there are two equal models of different classes, but TCNN must classify patterns correctly.

The technique that corrects this case of inconsistency is the "*Reduced Nearest Neighbor rule*" (RNN) introduced by Gates [11]. This rule is an extension of the CNN rule, and like CNN, RNN reduces TNN.

The RNN algorithm starts at TRNN=TCNN and removes every instance from TRNN if this deletion does not cause a misclassification of another instance in TNN by the remaining instances in TRNN. From the perspective of computing, it is more expensive than the rule proposed by Hart, but it will always produce a subset of CNN, and thus will be less expensive in terms of computing and storage at the classification stage.

Angiulli introduced the "*Fast Condensed Nearest Neighbor rule*" (FCNN) [2] a scalable algorithm on large multidimensional data sets used to create subsets serving as consistent training sets based on the nearest neighbor decision rule. This algorithm allows selecting points very close to the border decision. It is independent of the order, and has low quadratic complexity.

The FCNN rule initializes the consistent subset S with a starting element from each class label of the training set. The starting elements that the rule uses are particularly the class's centroids in the training set. The algorithm is incremental. At each iteration the resulting set is increased till the stop condition is reached.

Wilson and Martinez suggested [21] a series of six algorithms for sets reduction based on the kNN algorithm where each algorithm improves the previous one. The first reduction technique presented was the DROP1 which constitutes the basic framework on which were built the five other techniques. The DROP1 represents an improvement of the RNN rule, which verifies the accuracy of the resulting set instead of the initial set T. This algorithm is based on the following rule: an instance P is removed only if at least some of its associates (neighbors from same class) in S can be classified correctly without P.

The first proposal causes a problem when the noisy instances, which are typically associated with a different class, cover only a small portion of the input space. DROP2 tries to solve this problem by considering the effect of removing an instance on all instances of the initial training T rather than S. For this purpose the rule of DROP1 has been improved to eliminate P if at least an acceptable number of its associates in T can be classified correctly without P.

DROP2 sorts S in an attempt to remove the central points before the border points (points which are near from the decision boundaries), however noisy instances can also be on borders, which can cause a change in the removal order. And even if a noisy instance is central attempting to remove it could eliminate border points that should be maintained. Hence DROP3 uses a noise filtering before sorting the instances of S. This is done using a rule that eliminates any instance misclassified by its k nearest neighbors. Except that DROP3 can sometimes remove an overly large number of instances.

DROP4 improves DROP3 rule and provides that an instance is removed only if:

1. It is misclassified by its k nearest neighbors, and
2. Its removal does not affect the classification of other instances

DROP5 upgrades DROP2 by proposing that the instances are considered beginning from the ones closest to the nearest enemy (an enemy is the nearest neighbor of an instance with a different class) and proceeding to outside.

The latest algorithm proposed by Wilson and Martinez [22] was the DEL which is similar to DROP3, except that it uses the length coding heuristic for deciding whether an instance can be removed or not. In DEL an instance is removed only if:

1. It is misclassified by its k nearest neighbors, and
2. The removal of the instance does not increase the cost of length encoding

Wu, Ianakiev and Govindraju proposed an "*Improved K-Nearest Neighbor Classification*" [24]. A solution to increase the speed of traditional kNN classification while maintaining its level of accuracy by suggesting two building techniques. The suggested IKNN algorithm is based on iterative elimination of models with high attraction capacity.

In the first technique called the model condensing, the authors suggested that all classes have the same probability, and that the training set Ω is initially created by the extraction of vector elements from a large set of images of interest, where each class has an equal representation. This set Ω is refined iteratively.

In the second technique, which is the *pre-processing*, an unknown pattern is compared to a prototype in two stages. In the first stage a rapid evaluation of the potential match is made. The prototypes that fail in the first match are not included in the second. Then, for a complete match, in the

second stage the norms difference of the prototype and the test pattern must be less than a predetermined threshold.

The observation that large sets of data had computational requirements which could be prohibitive to classify models using the kNN, has led Fayed and Atia [9] to propose the TRKNN, a way to lessen this problem through a condensation approach. The aim of their approach is to eliminate the reasons that makes load the calculation and does not contribute to improve the classification.

This approach consists in rejecting the prototypes that are far from the limits and have just a little influence on the KNN classification. To achieve this, the authors first introduced the concept of chain of nearest neighbors which is a sequence of the nearest neighbors from alternating classes.

This technique proposes that if the distances of a given model decrease in value, this model is considered to be probably an internal point and can be discarded, whereas if the distances do not decrease too much, then this point probably varies around the limit of classification and it is maintained.

In a more recent paper [14] the authors introduced a new approach *"The Class Boundary Preserving Algorithm"* (CBP), a multi-step method for pruning the training set. The proposed method aims at preserving instances that are close to the borders of classes. Because, according to the authors, these instances can provide most of the necessary information to properly describe the underlying distribution. On the other hand instances distant from limits are considered redundant by the authors.

The innovation in this approach is the procedure used to divide the training set into two subsets, X_B containing the instances near the surface of decision, and X_{NB} containing the internal samples. And because there is a noticeable difference in the importance of information held by these two sets, two different reduction processes were applied on both.

The algorithm considers an initial set of n instances related to a set of labels, and involves four main steps.

The first step deals with the smoothing of class boundaries, the second allows to distinguish between border and non-border instances. The third step has to do with the pruning of border instances and the last step is the clustering of non-border instances.

To implement this idea, the authors have used the *Mean Shift Clustering algorithm* (MSC) which converges to the points of maximum density to determine the cluster centers of the distribution. Then to make the maximum number of neighbors in obtained clusters, the authors applied a merger process which verifies the calculated clusters centers labels, and if the clusters of nearest neighbors share the same label, then the centers are merged.

A. Comparison

After examining some articles that have tackled the reduction of sets based on the condensing KNN, we consider useful to raise a comparative board grouping the idea, the advantages, and disadvantages of each algorithm in the cited articles.

TABLE I.   COMPARISON OF NEAREST NEIGHBOR CONDENSING TECHNIQUES

CNN	**Idea**	Remove Training models that are very similar and those that do not add additional information to the classification
	Advantages	- Improves search time and memory requirements - Reduces the size of training data.
	disadvantages	- CNN is order dependent, then it is unlikely that it removes border points - If the initial set is minimal, that causes an inconsistency in the resulting set when the program is stopped. - There is no guarantee that the resulting set is minimal.
RNN	**Idea**	Initially the resulting set is equal to the initial set, and then each instance that does not cause a wrong classification of another instance in the initial set is removed from the resulting set.
	Advantages	- Reduces the size of training data and eliminates models - Improves the search time and memory requirements.
	Disadvantages	- High computational cost - Time consuming - Its consistency depends on the consistency of the resulting set of CNN
FCNN	**Idea**	Select points very close to the decision boundary
	Advantages	- Average efficiency of 96.01% for an average number of iterations of about 69 iterations - Has a smaller complexity than CNN - Good rate of condensation - Independent of the order
	Disadvantages	- Requires a large number of iterations
DROP1	**Idea**	Initially the resulting set is equal to the initial set, and then an instance is removed only if at least some of his associates in the resulting set can be ordered without it.
	Advantages	- Reduces the size of the training data and eliminates instances - Constitutes a basis on which are built the rest of the DROP algorithms and DEL - Noise do not degrade accuracy
	Disadvantages	- Checks the consistency of the resulting set instead of the initial set - Low accuracy - Cannot use information from previously eliminated instances
DROP2	**Idea**	Improves DROP1 and eliminates an instance when at least a good number of its associates in the initial set can be classified without it.
	Advantages	- Checks the consistency of the initial set rather than the final set

		- Reduces the size of training data and eliminates instances - Reach a higher accuracy than KNN for noisy instances - Good storage reduction happens to 1 / 6 of the original set - Storage requirements lower than DROP 1, 4, 5
	Disadvantages	- Attempts to remove the central points before the borderer, which overlooks the border noisy instances - The attempt to eliminate a noisy central instance can remove border points that should be maintained. - For very large data sets DROP2 does not eliminate enough points
DROP3	**Idea**	Improves DROP2 by eliminating instances misclassified by their neighbors
	Advantages	- Is based on the classification of the instance itself to remove it - Reduces the size of training data and eliminates instances - Reach a higher accuracy than the traditional KNN in case of noisy instances - Very good reduction of storage that comes to 12% of the original set - Storage requirements lower than DROP1, 4, 5
	Disadvantages	- Can sometimes remove a too large number of instances
DEL	**Idea**	Improves DROP3 and eliminates an instance if: 1. It is misclassified by its k nearest neighbors, and 2. The removal of the instance does not increase the cost of length encoding
	Advantages	- Achieved higher accuracy than the traditional KNN in case of noisy instances - Reduces the size of training data and eliminates instances
	Disadvantages	- Accuracy less than that of DROP2-5 - Storage requirements higher than those of DROP2-5
IKNN	**Idea**	Remove iteratively models exhibitors of high capacity of attraction
	Advantages	- Reduces the data set by keeping the prototypes that are useful - The preprocessing allows a significant saving in computation time - The classifier shows a slight improvement in accuracy compared to traditional kNN - Reduces the size of the models maintaining the same level of accuracy
	Disadvantages	- Classes must have same probability in the training set - The work was built on intuition and not on a mathematical framework citing that the norm is an intrinsic characteristic - No theoretical proof that preprocessing guaranteed to filter the relevant prototypes or to maintain the accuracy - A test pattern is a distorted version of a prototype when the difference in standard is below a threshold associates to this prototype
TRKNN	**Idea**	Eliminate the patterns that are a burden on computing and does not contribute to improve the classification
	Advantages	- Reduces the size of the models without sacrificing accuracy - TRKNN is up to 3 times faster than IKNN and up to 4 times than DROP2
	Disadvantages	- Average condensation rate (35%) - Average level of accuracy
CBP	**Idea**	Preserve instances that are close to class boundaries
	Advantages	- Combine the selection and the abstraction - Applies to each case an appropriate procedure of condensation - The filtering Phase is based on the classification of the instance itself - Use the geometric characteristics of the distribution - Provides an overview of the distribution of samples - Reduces the size of the training set and the number of representatives - Good rate of condensation
	Disadvantages	- The filtering algorithm can sometimes remove a very large number of instances - Relatively high computation time

These approaches which are based on traditional KNN, select all the prototypes from training samples of initial set, by adding or cutting, with aiming at preserving the performance of classification with the use of heuristic methods. However, none of these methods can prove the minimality of the result set.

V. METAHEURISTICS: IMPROVING KNN

The basic concepts of metaheuristics can be described in the abstract, without requiring a specific problem [15]. These algorithms can therefore use heuristics, which in turn reflect the specificity of the problem treated, except that these heuristics are controlled by a higher level strategy.

In this article we present some improvements of classical kNN algorithms through the use of some metaheuristics such as genetic algorithms and tabu search.

R. Gil-Pita and X. Yao proposed [12] three improvements of the k-nearest neighbor using genetic algorithms: the use of an objective function based on mean square error, the implementation of a clustered crossover, and a fast smart mutation scheme.

In the first proposal they make use of a novel objective function based on the MSE function. They consider the kNN as a system with C outputs, so that each output is calculated using a defined equation. Therefore, the C outputs of the kNN

system are approximations of the posterior probabilities of the data.

In the second proposal the authors suggest a new crossover scheme for the GA, denominated *clustered crossover* (CC), in order to improve the determination of the best subset of the training set that minimizes the selected objective function by considering possible to determine the relationship between the different bits of the bit stream.

In the third proposal R. Gil-Pita and X. Yao describe the application of a mutation scheme that allows selecting the best gene or group of genes to be changed, taking into account the variations of the objective function with respect to each gene for a given set. They design a fast method for evaluating the error variation when each gene is changed, and they propose a mutation strategy based on these variations of the objective function. The authors denominate this mutation scheme as *fast smart mutation* (FSM) [13], as it allows increasing the effectiveness of the mutation stage in the genetic algorithm.

Compared to the use of a classic kNN classifier, obtained results demonstrate the good accuracy of the proposed GA-based technique.

The authors of "*A hybrid classification method of k-nearest neighbor, Bayesian methods and genetic algorithm*" [16] introduced a hybrid technique including KNN, genetic algorithms and Bayesian method. This approach consists of eight steps that begin with the application of the Bayesian algorithm and then the generation of new data using the genetic algorithm and the application of the K-nearest neighbor's method. And finally, several iterations of these algorithms occur orchestrated by the genetic mechanism. According to the authors, the method is useful on data sets that have a small amount of data. It generates an unlimited number of data that have similar characteristics to the original data, and improves these data according to the proposed algorithm.

This method shows better classification performance with respect to classic methods such as expectation maximization algorithm [5] used to develop it. It is intended, according to its authors, to hardware solutions with low cost based on clustering, to noisy data classification, and to classification in the data sets with little data.

To overcome the limitations of kNN, an improved version of KNN, "*Genetic KNN*" (GKNN), was proposed by Suguna and Thanushkodi [19]. A genetic algorithm is combined with the k-nearest neighbor algorithm (KNN). In the proposed method, using the genetic algorithm, k-number of samples are selected for each iteration and the accuracy of the classification is calculated as fitness. The greater accuracy is recorded each time. Thus, it is not necessary to calculate the similarities between all samples, or to consider the weight of the category.

The performance of GKNN classifier was compared with traditional KNN. The experiments and results show that the GKNN not only reduces the complexity of the KNN, but it improves the classification accuracy.

Ceveron and Ferri [3] proposed a method for the prototypes selection for the nearest neighbor rule which aims to obtain an optimal or close to optimal solution by presenting a new approach based on tabu search. The particularity of the proposed tabu search is that it uses the equations from objective genetic algorithms. In this approach all possible subsets of prototypes constitute the space of solutions. Possible moves from a particular subset consist of adding or removing each of the n initial prototypes. The attribute used to declare the movement taboo is the prototype that is added or deleted.

It is worth mentioning that the results obtained with TS consistently improved the classical condensing techniques previously published in the literature using the Iris database.

Wu's article [22] provides a method of selecting items based on the tabu search algorithm and KNN. First, the KNN algorithm is used to generate the initial solution needed for the tabu search. Then, the tabu algorithm is applied to obtain an optimal subset of items. The KNN algorithm uses the relevance of the elements to eliminate those redundant in large networks data, and the subset obtained is the initial solution of the tabu search algorithm.

This algorithm was tested via creating intrusion detection model. The tests show that by using the feature selection method proposed in this paper, the detection performance of the intrusion detection system is effectively improved without compromising detection accuracy, and the detection time and accuracy performance of the system is much better than that of the current feature selection methods. As a result the method proposed in this paper has proved to be effective and feasible.

VI. CONCLUSION

In this paper we tried to present and compare sets reduction techniques based on the principle of nearest neighbor. These techniques are of type "condensing". Both are improvements compared to basic KNN. These improvements have been proposed by the authors to reduce the training set to gain on speed and space efficiencies. To ensure the minimality of this training set we presented some recent proposals using metaheuristics to check the optimality of the resulting set of some KNN reduction techniques.

Note that each technique is very effective in a specific area and in special circumstances.

REFERENCES

[1] Chang Bao Rong, Naghibzadeh Mahmoud, Czarnowski Ireneusz, Danesh Malihe and Danesh Mohaddesh, "Data Clustering Based on an Efficient Hybrid of K-Harmonic Means, PSO and GA", Transactions of Computational Collective Intelligence, vol. IV, pp. 125-140, Springer, 2011.

[2] Fabrizio Angiulli, "Fast condensed nearest neighbor rule", Technical report, Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany, 2005.

[3] Vicente Cerveron and Francesc J. Ferri, "Another move toward the minimum consistent subset: A tabu search approach to the condensed nearest neighbor rule", IEEE Transactions on Systems, Man, and Cybernetics, vol. 31, pp. 408–413, 2010.

[4] T. M. Cover and P. E. Hart, "nearest neighbor pattern classification", IEEE Transaction on Information Theory, vol. 13, pp. 21–27, Jan 1967.

[5] F. Dallaert, "The expectation maximization algorithm", Technical report, College of Computing, Georgia Institute of Technology, February 2002.

[6] E. M. Elnahrawy, "Log based chat room monitoring using text categorization: A comparative study", University of Maryland.

[7] Pavel Paclik, Elzbieta Pekalska and Robert P.W. Duin, "Prototype selection for

dissimilarity-based classifiers", Pattern Recognition, vol. 39, pp. 189–208, 2006.

[8] F. Bajramovic, Frank Mattern, Nicholas Butko, and Joachim Denzler "A comparison of nearest neighbor search algorithms for generic object recognition", LNCS 4179, ACIVS 2006, pp. 1186–1197.

[9] 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 (5), pp. 890–896, May 2009.

[10] O. Kirmemis and G. Toker, "Text categorization using k nearest neighbor classification", Survey paper, Middle East Technical University.

[11] G.Gates, "The reduced nearest neighbor rule". IEEE Transactions on Information Theory, vol. 18, pp. 431–433, 1972.

[12] R. Gil-Pita and X. Yao, "Using a genetic algorithm for editing k-nearest neighbor classifiers", IDEAL 2007, LNCS, vol. 4881, pp. 1141-1150, 2007.

[13] X Gil-Pita, R. Yao, "Evolving edited k-nearest neighbor classifiers", International Journal of Neural Systems, vol. 18 (6), pp. 459–467, Dec 2008.

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

[15] Sean Luke, "Essentials of Metaheuristics". Lulu, 2009.

[16] Mutlu Avci Mehmet Aci and Cigdem Inan, "A hybrid classification method of k nearest neighbor, bayesian methods and genetic algorithm", Expert Systems with Applications, vol. 37, pp. 5061-5067, 2010.

[17] P. Hart, "The condensed nearest neighbor rule", IEEE Transactions on Information Theory, vol. 14, pp. 515–516, 1968.

[18] Y. Wu Shizen, "An algorithm for remote sensing image classification based on artificial immune b-cell network", The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XXXVII, Part B6b, pp. 107-111, Beijing 2008.

[19] Suguna and Dr. K. Thanushkodi, "An improved k-nearest neighbor classification using genetic algorithm", International Journal of Computer Science Issues, vol. 7, pp. 18–21, 2010.

[20] V.Vaidehi, S. Vasuhi, R.Kayalvizhi, K.Mariammal, Raghuraman.M.B, "Person authentication using face recognition", Proceedings of the World Congress on Engineering and Computer Science, 2008.

[21] Randall Wilson and Tony R. Martinez, "Reduction techniques for instance-based learning algorithms", Machine Learning, vol. 38 (3), pp. 257–286, 2000.

[22] Tao Ran, Wu Jian-guang and Li Zhi-Yong, "An improving tabu search algorithm for intrusion detection", IEEE Computer Society, pp. 435–439, 2011.

[23] Y. Yang and T. Ault, "Improving text categorization methods for event tracking", Carnegie Mellon University.

[24] Krasimir G. Iankiev Yingquan Wu and Venu Govindaraju, "Improved k-nearest neighbor classification", Pattern recognition, vol. 35, pp. 2311–2318, 2002.

[25] A. V. Zukhba, "Np-completeness of the problem of prototype selection in the nearest neighbor method", Pattern Recognition and Image Analysis, vol. 20, pp. 484-494, 2010.