NRMCS : NOISE REMOVING BASED ON THE MCS

XI-ZHAO WANG, BO WU, YU-LIN HE, XIANG-HAO PEI

Key Lab. of Machine Learning and Computational Intelligence, College of Mathematics and Computer Science, Hebei University, Baoding, 071002, Hebei, China
E-MAIL: wbfeixue@163.com

Abstract:

MCS (Minimal Consistent Set) is one of the classical algorithms for minimal consistent subset selection problem. However, when noisy samples are present classification accuracy can suffer. In addition, noise affect the size of minimal consistent set. Therefore, removing noise is an important issue before sample selection. In this paper, an improvement approach based on MCS to select the representative samples is proposed. Compared with other algorithms which remove the noise by Wilson Editing in advance for the representative samples selection, this algorithm performs the processes of noise removing and samples selection simultaneously. According to this method, most noise can be deleted and the most representative samples can be identified and retained. The experiments show that the proposed method can greatly remove the redundant samples and noise as well as increase the accuracy of solutions when it is used for classification tasks.

Keywords:

MCS; Sample Selection; ICF; Representative Subset; Noise; Wilson Editing

1. Introduction

Nearest Neighbor (NN) classification is one of the important non-parametric classification methods. It is a simple supervised concept learning scheme which classifies unclassified samples by finding the closest previously observed sample. It becomes widely used in data mining and pattern recognition applications. Nevertheless, The main limitation on their usage in practice has been their computational demands and storage in the operational phase.

In order to reduce the computational demands and memory storage, one may find a representative subset of the training data so as to reduce the times of feature space distance computations. Another approach advocated over the years may appropriately organize the given data and use efficient search algorithm.

The very first study of this kind was probably Hart [2] who presented the "Condensed Nearest Neighbor Rule (CNN)" to obtain reduced and consistent subsets to be used with the 1-NN rule in 1968,. In 1972, the "Reduced Nearest Neighbor Rule (RNN)" of Gates [3], took an opposite track to the CNN approach. Instead of growing the condensed set from a null set , the reduced set is derived by iteratively contracting the given set. The resultant RNN subset is always smaller than the CNN subset and hence in the operational phase, the RNN based classifier tends to be computationally more efficient than the CNN based one. The Selective Nearest Neighbour Rule (SNN) devised by Rittet[4] et al improved on the CNN and RNN by ensuring that a minimal consistent subset is found. The selection criteria is strict by enforcing the following rule: all samples in the training set must be closer to a sample in the selected set than any sample of a different class found in the training set. Ritter reported improved prediction accuracy when compared to the CNN. Additionally, there are some similar works (Swonger[5],1972; Ullmann[6],1974; Tomek[7],1976; Smyth and Keane owda and Krishna[8],1979, Zhang, J., Yim, Y.-S.and Yang, J, 1997[12] ; Wilson and Martinez, 2000[13]). All of these algorithms aim to reduce the size of the representative subset. Meanwhile there are some other approaches. In 1974 ,Chang[14] offers a novel approach to remove samples by repeatedly attempting to merge two existing samples into a new sample. In 1995, the Footprint Deletion policy was introduced by Smyth and Keane[15] which is a filtering scheme designed for use within the paradigm of Case-Based Reasoning. In 2002, Tabu Search was presented by Zhang[16]. He used Tabu search to select the near optimal reference subset for the nearest neighbor classification. In 2005, Angiulli[17] offered a new approach (FCNN) to remove samples.

In 1994, Dasarathy[1] presented a condensing algorithm (Minimal Consistent Set: MCS) for selecting an optimal consistent subset based on the concept of the nearest unlike neighbor subset (NUNS). The algorithm introduced a voting mechanism to select the minimal consistent set (MCS) based on the samples representative significance. At each iteration, every sample casts a vote for the same class which is closer than its NUN. The more

978-1-4244-2096-4/08/$25.00 ©2008 IEEE

votes a sample receives, the higher its representative Samples are selected according to this ranking until consistency is achieved. The algorithm starts with the whole set and iterates while the selected subset size decreases. In this way, it generally obtains a high quality subset with lower cardinality than the previous approaches regardless of the initial ordering of dataset. Dasarathy's algorithm is the best known algorithm in terms of consistent subset size and the selected samples'representative nature. However, his conjecture of the minimality of obtained MCS later is proved not to be true by Kuncheva and Bezdek [9] and Cerveron and Fuertes [10] for the popular IRIS data set.

Brighton and Mellish presented Iterative Case Filtering Algorithm (ICF)[11]. The ICF algorithm uses the concept of reachable and coverage sets that we can liken to the neighbourhood and associate sets. The deletion rule is simple: we remove samples which have a reachable set size greater than the coverage set size. The author illuminated that the ICF algorithm achieved the highest degree of instance set reduction as well as the retention of classification accuracy.

In this paper , we will introduce an improvement of MCS algorithm called NRMCS for selective sampling that is suitable for nearest neighbor classification. It does not remove the noise samples in advance, but handling them with the voting simultaneously. The experiments show that the proposed method can greatly remove the noise samples as well as enhance the testing accuracy of solutions to some extent . Especially, it costs fewer time than ICF and MCS. This article is organized as follows.: In Section 2, we analyze the drawbacks of the Wilson Editing that is a popular method of removing the noise samples and presenting a new way of handling the noise samples. The NR-MCS algorithm procedure is also described in this Section . This is followed by the experimental results and analysis in Section 3. The conclusion is provided in Section 4.

2. The improvement of MCS algorithm

2.1. Noise and Wilson Editing

Generally speaking, the samples which affect the classified accuracy are called noise. So we should remove the noise to increase the classified accuracy. Meanwhile,we should consider the distribution of all samples when the samples are removed as noise. Wilson Editing[18] is one kind of classical algorithms which attempt to remove noise by making a pass through all the samples in the training set and removing those which satisfy an editing rule. The rule is simple: all samples which are incorrectly classified by their nearest neighbours are assumed to be noisy instances. The samples which satisfy such a rule will be those that belong to different classes to their neighbours: (1)These samples appear as exceptions within regions containing samples of the same class. For example, the original set is shown in Figure.1(a). The result applying the Repeated Wilson Editing algorithm is shown in Figure.1(b). (2)Other candidates fulfilling this rule could be the odd samples lying on a border between two different classes. It can remove some samples on the decision boundary, which will lead to the decrease of the classification accuracy. Especially, The quality of the ICF algorithm is poor when the database has many classes. Therefore, our aim is to delete the samples in (1) and to reserve the ones in (2).

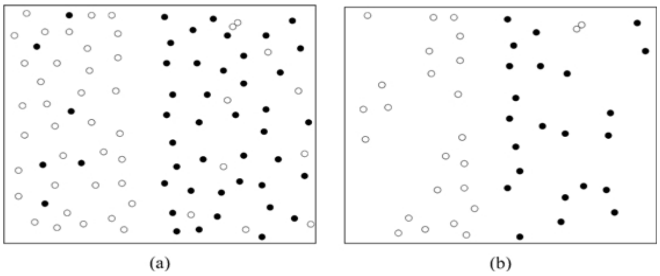
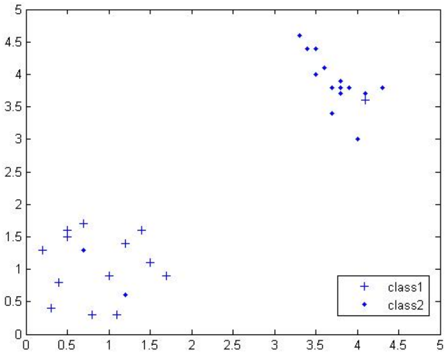


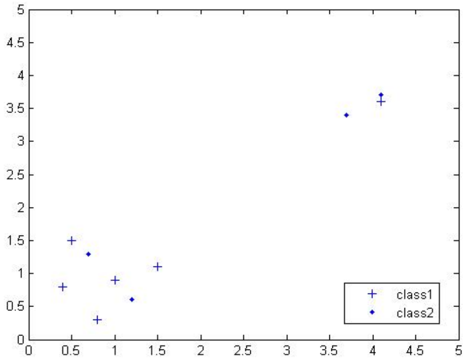
Figure 1. (a)Original set (b) Repeated Wilson Editing

2.2. The drawbacks of MCS

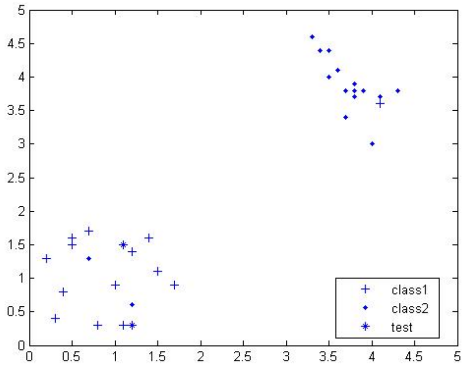
MCS selection strategy is based on the concept of NUNS（the Nearest Unlike Neighbor Subset）which can be looked upon as an descriptor of the inter-class boundaries. Based on this concept, we can see that for every given sample, the sufficient condition for consistency is the presence within MCS a sample from its own class that is closer than its NUN distance. over-fitting of the training data is the drawback of the MCS and It may lack generalization ability. In order to illustrate the behavior of the MCS algorithm, an experiment using the artificial data was carried out. A sets of 30 float-valued two-dimensional (2-D) vectors were generated and then labeled that are shown in Figure.2(a). The result applying the MCS algorithm is shown in Figure. 2(b). It is worth noting that two negative samples among the positive samples and one positive sample among the negative samples can not be removed by the MCS algorithm, because each sample has a singleton NUN coverage set. This property protects the sample from removal. So some unlike neighbor samples are also selected for correct classification. If there are so many such kind of samples in the training set, it is difficult to find a small representative subset. On the other hand, it can be observed that the test samples (positive sample) which are classified (1-NN) correctly by original set in Figure. 2(c) and incorrectly classified by MCS in Figure. 2(d). These samples are regarded as noise by Wilson Editing Figure.1(a). In addition, the noise samples may hurt the classification accuracy. If we want to obtain high classification accuracy and compression ratio, the noise samples which are mentioned above should be removed. In a word, MCS pays more attention to the consistency and ignores the data structure.



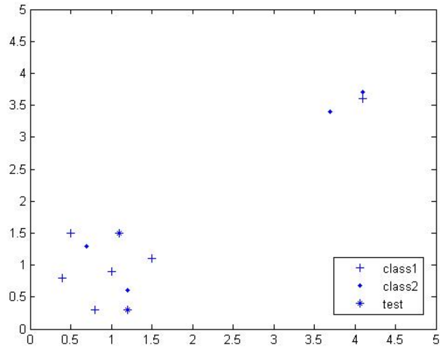
(a)



(b)



(c)



(d)

Figure 2. (a)Original set (b) MCS
(c) Test samples in original set (d) Test samples in MCS

2.3. The NRMCS algorithm

In order to remove the noise samples, many algorithms for prototype selection adopt Wilson Editing method so far, such as ICF. However, almost all of the algorithms remove the noise samples in advance, then select the representative subset from the training set. The two steps are separated. If we can combine the two steps, that is to say, perform the processes of removing the noise samples and selecting the representative samples simultaneously, then it is able to reduce the computational demands and memories.

The MCS algorithm offers a voting strategy to select the representative subset. Every sample casts a vote for the same class samples which are closer than its NUN distance. So the noisy sample which is surrounded by different class es samples has a singleton NUN. For each sample except itself, it is neither the voter nor the voted sample. The samples being signed noise are just the ones appearing as exceptions within regions containing samples of the same class in Wilson Editing method. Meanwhile, the drawbacks of removing the samples on the decision boundary brought by Wilson Editing can be avoided.

The NRMCS algorithm is as follows:

Step1: For each sample, identify all the neighboring samples from its own class in the given data set which are closer than its NUN distance and cast an approval vote to each of these samples in the given set.

Step2: Create a potential candidate consistent set consisting of all samples in the given set which are either (a) already present in the current consistent set or (b) whose inclusion will not create an inconsistency. In the first iteration, the entire consistent set remains as the candidate consistent set as all samples satisfy condition(a).

Step3: Delete the samples which neither vote nor are voted except itself from potential candidate consistent set.

Step4: Identify the most voted sample in this candidate , and designate it as a member of a newly selected consistent set and identify all of its contributing voters.

Step5: Delete these voters wherein they currently appear and correspondingly decrement the appropriate vote counts.

Step6: Repeat Step 2 through 4 till all the voters have been accounted for by the selected consistent set.

Step7: Repeat Step 1 through 5 using all the samples in the given data set to identify a new consistent set.

3. Experimental results and analyses

3.1. Databases

In this section we experimentally evaluate the algorithm proposed in Sec.2. To further verify the advantages of NRMCS, we select eight databases from machine learning repository (UCI) and HBU(Table 1). We compare the performance with the MCS, ICF and the NR-MCS on the selected eight databases.

Table 2 shows our experimental results which are the 5-cross-validation of ten times experimenting on the eight databases: the average number statistics over 10 independent runs. A random 20% of samples where taken to form a testing set. Different testing sets where taken for each run. 1-Nearest Neighbour rule was used for classification. Euclidean distance is applied to measure the distance.

The experimental results are summarized in Table 2.

Table 1. The databases used in experiments

Database	Attribute number	Class number	Instance number	source
Glass	9	6	214	UCI
Ionosphere	34	2	351	UCI
Iris	4	3	150	UCI
Haberman	3	2	299	UCI
RenRu	32	2	148	HBU
Pima	8	2	768	UCI
Zoo	17	7	101	UCI
Wine	8	2	178	UCI

3.2. Results and discussions

Table 2. NRMCS algorithm's performance compare with MCS on eight databases

Database	Mcs on testing set	ICF on testing set	NR-MCS on testing set	Number of selected representative subset (MCS)	Number of selected representative subset(ICF)	Number of selected representative subset(NR-MCS)	Time rate (MCS:ICF: NR-MCS)
Glass	0.6860	0.6964	0.7012	71.8	30.3	32.4	100:72:55
Ionosphere	0.8840	0.8766	0.9142	37.6	4	4	100:78:63
Iris	0.9333	0.9333	0.9667	11.6	3	3	100:75:55
Haberman	0.6812	0.7016	0.7464	110.8	29.1	36.6	100:67:51
RenRu	0.8109	0.8666	0.9126	30.2	8.4	10	100:63:46
Pima	0.6333	0.6917	0.7114	140.6	101.7	115.4	100:72:58
Zoo	0.9088	0.9242	0.9426	45.6	14.6	18.2	100:66:54
Wine	0.6861	0.8462	0.7918	51.2	20.6	22.8	100:59:52

From Table 2, we can analyze the effectiveness of NRMCS algorithm from three aspects: Testing accuracy, Cardinality of selected representative subset, Operating time.

(1)Testing accuracy: It is easy to see that the testing accuracy of NRMCS is obviously higher than that of MCS from table 2. Meanwhile, the testing accuracy of NRMCS is also higher than that of ICF except the wine database.

(2)Cardinality of selected representative subset: comparing with the cardinality of selected representative subset which MCS algorithm gets, NR-MCS and ICF can get smaller ones. The cardinality of selected representative subset which NR-MCS gets is a little bit more than that which ICF gets.

(3)Operating time: From table 2, we can see that MCS is the most time-consuming. The time NRMCS costs is less than ICF.

We can see from (1) and (2), NR-MCS and ICF is obviously superior to MCS. Comparing with ICF, NRMCS can get higher testing accuracy but more number in the selected representative subset. That is just because the noise of which NRMCS removes only a part of that which ICF

978-1-4244-2096-4/08/$25.00 ©2008 IEEE

removes, it removes those samples appearing as exceptions within regions containing samples of the same class and retains the samples on the decision boundary. All of these have been analyzed in section 2.

From (3) to know, operating time of NR-MCS is the least. That is because the deleting of noise makes each selected sample get more votes, which is presumed that the selected samples more representative. When deleting voters , compared with MCS , NRMCS can traverse the entire set rapidly and get a smaller consistent set , thus reduces the time needed for each iteration. At the same time, Because of each iteration to get the smaller set, NR-MCS can be quickly set the convergence , bringing the number of interactive times reduction.

4. Conclusions

In this paper, an improvement approach based on the MCS for representative subset is proposed. Using this method, most noise can be deleted and the most representative samples can be identified and retained. The experiments show that the proposed method can greatly remove the redundant samples and noise as well as increase the accuracy of solutions when it is used for classification tasks.

Acknowledgments

This research is supported by the Natural Science Foundation of Hebei Province (F2008000635), by the key project foundation of applied fundamental research of Hebei Province (08963522D), by the plan of 100 excellent innovative scientists of the first group in Education Department of Hebei Province, and by the Scientific Research Foundation of Hebei Province (06213548).

References

[1] B. V. Dasarathy.Minimal consistent set (MCS) identification for optimal nearest neighbor decision systems design [J] . IEEE. Trans Syst.man Cybern. March ,1994 ,24 (3) :511 - 517.

[2] P.E.Hart.The condensed nearest neighbor rule [J].IEEETrans.InformationTheory ,May.1968 ,IT214(3) :515 - 516.

[3] G.W. Gates. The reduced nearest neighbor rule [J] . IEEE Trans. Information Theory,May 1972 ,IT218(3) :431 - 433.

[4] G.L. Ritter, H.B. Woodruff, S.R. Lowry and T.L. Isenhour, An algorithm for a selective nearest neighbour rule, IEEE Trans. Inform. Theory IT11-IT21 (1975), 665–669.

[5] C.W.Swonger,Sample set condensation for a condensednearest neighbor decision rule for pattern recognition,in:S.Watanade (Ed.), Frontiers ofPattern Recognition,Academic Press, New York, 1972, pp. 511–519.

[6] J. R. Ullmann, "Automatic selection of reference data for use in a nearest neighbor method of pattern classification," ZEEE Trans. Information Theory, vol. IT-20, no. 4, pp. 541-543, July 1974.

[7] I. Tomek, "Two modifications of CNN," IEEE Trans. Syst., Man,Cybern., vol. SMC-6, no. 11, pp. 769-772, Nov. 1976.

[8] K. C. Gowda and G. Krishna, "The condensed nearest neighbor rule using the concept of mutual nearest neighborhood," IEEE Trans. Information Theory, vol. IT-25, no. 4, pp. 488490, July 1979.

[9] L.I. Kuncheva, J.C. Bezdek, Nearest prototype classification: clustering, genetic algorithms, or random search,IEEE Trans. Syst. Man and Cybern. 28 (1) (1998)160–164

[10] V. Cerveron, A. Fuertes, Parallel random search and Tabu search for the minimal consistent subset selection problem, Lecture Notes in Computer Science, Vol. 1518, Springer,Berlin, 1998, pp. 248–259.

[11] Henry Brighton,"Advances in Instance Selection for Instance-Based Learning Algorithms" Data Mining and Knowledge Discovery,6,153-172,2002.

[12] Zhang, J.,Yim,Y.-S.,& Yang,J.(1997). Intelligent selection of instances for prediction functions in lazy learning algorithms. Artificial Intelligence Review, 11:1–5, 175–191.

[13] Wilson,D.R.,Martinez,T. R.(2000).Reduction techniques for instance-based learning algorithms. Machine Learning, 38:3, 257–286.

[14] C. L. Chang. Finding prototypes for nearest neighbor classifiers [J] .IEEE Trans. Computers ,Nov. 1974 ,c223(11) :1179 - 1184.

[15] Smyth,B.andKeane, M.T. 1995. Remembering to forget. In IJCAI-95, Proceedings of the Fourteenth International Conference on Artificial Intelligence, C.S. Mellish (Ed.), Vol. 1., pp. 377–382.

[16] H B Zhang, G Y Sun. Optimal reference subset selection for nearest neighbor classification by tabu search [J]. Pattern Recognition, 2002, 35:1481-1490

[17] F Angiulli. Fast Condensed Nearest Neighbor Rule [C]. In: Proc of the 22nd International Conference on Machine Learning, Bonn: ACM Press, 2005. 25-32

[18] Wilson, D.L. 1972. Asymptotic properties of nearest neighbor rules using edited data. IEEE Transactions on Systems, Man, and Cybernetics, SMC-2(3):408–421.