ZESZYTY NAUKOWE POLITECHNIKI ŁÓDZKIEJ
Nr 1045 ELEKTRYKA, z. 117 2009

ARTUR SIERSZEŃ
Technical University of Łodz
Computer Engineering Department

METHODS OF REFERENCE SET CONDENSATION FOR DECISION RULES BASED ON DISTANCE FUNCTIONS

Reviewer: **Adam Jóźwik, Ph.D., D.Sc.**

Manuscript received 28.01.2009

The article presents four original algorithms of the reference set condensation to control the compromise between the speed and the quality of classification based on the obtained condensed reference set. The results obtained confirm the usefulness of the proposed algorithms, particularly in the case of very large training sets.

1. INTRODUCTION

The article refers to one of the biggest problems of pattern recognition, which is a compromise between the time of making a decision and its correctness. The speed of executing an algorithm is a key factor in the case of making a decision in the constrained time interval, e.g. in industrial processes, where the continuity of production cannot be broken because of too slow decision-making. In other cases (e.g. in biomedical data analysis), certainty that classification is correct is more important, i.e. whether the reliability of classification is acceptable. The reduction of incorrect decisions is usually possible by using a more complex recognition algorithm; however, this results in extending the time of a decision-making process.

This article presents a group of 'flexible' algorithms which enable one to control the compromise between the speed and the quality of classification within possibly the widest range. They were created according to the author's own concepts of the reference set size reduction for the decision rule based on the nearest neighbourhood. These are the following:

- The condensation of the reference set by the method of cutting hyperplanes.
- Modified Chang's algorithm.
- Bubble condensation algorithm.
- Cascade algorithm.

2. CONDENSATION OF THE REFERENCE SET BY THE METHOD OF CUTTING HYPERPLANES

The presented method of condensation is based on assigning precisely one pair of the mutually furthest points (from different classes) to each point from the learning set. In the case when the furthest points are located at equal distances, the one with a lower index is selected. Because the same pair of the mutually furthest points or the pair in which a particular point coincides with another may correspond to many objects, a pair with a lower number (i.e. the one which was found faster) is selected. The problem of the coincidence of points belonging to the same classes and possessing precisely the same characteristics was solved by omitting them (removing them from the test sets). This did not have a negative effect on the level of the classification error, when classification was performed with the use of the obtained condensed sets. The pair of the mutually furthest points is used to determine the hyperplane dividing the selected subset of the reference set. The hyperplane passes through the centre of the segment connecting these points and is orthogonal to it. As the first set, the whole learning set, which plays a role of the reference set, is divided. The next subset to be divided is automatically determined by the algorithm. New subsets obtained as a result of the division are replaced with the gravity centers and their assignment to a specific class are performed according to the majority criterion, i.e. they are assigned to the classes most heavily represented in each of the two obtained parts.

For the needs of the described condensation method, the algorithm of finding the mutually furthest points was introduced (Jóźwik and Kieś 2005), the outline of which is presented by means of a pseudo code.
T – the set of all testing objects, t – an element of T set;
$t = [a_1, a_2 \dots a_n]$; a – a feature describing the point; n – the number of features

i00. START;
i01. choose $t_k = t_0$ (t_0 = the first element of T set);
i02. $t_z = t_k$;
i03. find t_x element so that $\| t_k , t_x \| = \max$, if $t_x = t_z$ go to **i05**;
i04. $t_z = t_x$, $t_k = t_x$, go to **i03**.;
i05. END

Below, the pseudo code of the algorithm for the reference set condensation, based on determining the mutually furthest objects, is given.

T – the learning set containing m objects; Z – the condensed set;

T^j – subsets of the learning set, j=1,2,...,i, determined after performing i^{th} iteration;

$T = \{t_1, t_2, \ldots, t_m\}$; x, k – indexes of elements from the set T;

$t = [a_1, a_2 \ldots a_n]$; a – a feature describing an object; n – the number of features;

i00. START; i=1; $T^i = T$; $Z=\Theta$ {i.e. empty set};
i01. Find a pair of the mutually furthest objects t_j and t_k in the set T^i;
i02. Construct a cutting hyperplane of g(t)=0 equation basing on points t_j and t_k;
i03. $T_{iA}= \{t \in T^i: g(t)>=0\}$; find a centre of gravity z_{iA} of the set T_{iA};
i04. $T_{iB}= \{t \in T^i: g(t)<0\}$; find a centre of gravity z_{iB} of the set T_{iB};
i05. Delete T^i, store T_{iA} as T^i and T_{iB} as T^{i+1}; next i=i+1;
i06 Delete the gravity center of T^i;
i07. $Z= Z \cup \{z_{iA}, z_{iB}\}$;
i08. Estimate a classification error for 1-NN rule working with the condensed set Z and store it;
i09. Arrange T^j sets, j=1,2,...i, so that T^i is the largest set;
i10. If T^i contains more than one objects, go to **i01**;
i11. END

After a new reduced set has been determined (i.e. after each iteration), the classification error for 1-NN rule applied to the present condensed set is computed with the use of the leave-one-out method.

The application of the author's algorithm, particularly when computing the error every second iteration, allowed one to obtain a better result in a shorter time than the standard 1-NN method in the case of medium sized sets (5000-6000 elements). Basing on the computations performed for various sets, it was noticed that in each case the construction of the reference set enables one to find the best, at a particular moment, condensed set which may replace the original reference set; such a condensed set gives a lower classification error.

3. MODIFIED CHANG'S ALGORITHM

The advantage of Chang's algorithm is a considerable reduction of the reference set. Its drawback is a relatively low speed. The original procedure of Chang's reduction is presented below in the form of a pseudo code:

ORIGINAL CHANG'S ALGORITHM
T = {t_1, t_2, …, t_m} – the learning set containing m objects t;
t = [a_1, a_2 … a_n]; an element of T set, a – a feature describing a point; n – the number of properties
Z – the current reduced set;
Θ – an empty set; key1 and key2 – working variables of logic type;

i00. START, A=Θ, B=T; A={a random object from B};
i01. key1=false; key2=false; B=B-A;
i02. Find p∈ A and q∈ B, so that the distance d(p,q) is minimum;
i03. If p and q are from the same class, determine a set Z=A∪B∪{p*}- {p,q}, where p*=(p+q)/2;
i04. If p and q are from the same class and Z is the same as T, key1=true and key2=true;
i05. If key1=true, A=A-{p}∪{p*} and B=B-{q};
i06. If key1=false, A=A∪{q} and B=B-{q};
i07. If B≠Θ, go to **i02;**
i08. If B=Θ and key2=true, B=A and A=Θ and go to **i01**;
i09. If B=Θ and key2=false, Z=A and END.

'key1' is to register that two objects p and q have been replaced with one object p*, 'key2' is to recognize that no merger has taken place and that the algorithm should be ended.

The modification proposed by the author of this thesis aims at accelerating computations by replacing a larger number of objects, not only a pair of them, with one object. For any object, it is possible to determine all objects from the same class which are located at a shorter distance to it than any other object from the opposite class. Then, all those objects are replaced with their gravity centre with the same label as objects on the basis of which it was computed. In each case, after a new set has been determined, its reliability was checked with the use of the 1-NN rule and the computation of the error by the leave-one-out method. Modified Chang's method is presented below.

MODIFIED CHANG'S ALGORITHM
T – the learning set containing m objects t;
Z – the current reduced set; P – the working set;
Θ – an empty set; key1 and key2 – working variables of logic type;

i00. START, A=Θ, B=T; A={a random object from B};
i01. key1=false; key2=false; B=B-A;

i02. Find $p \in A$ and $q \in B$ from other class than p object, so that the distance d(p,q) is minimum;
i03. Determine a set $P=\{t \in B: d(t,q)<d(p,q)\}$ and its centre of gravity p* with the same label as the label of the object p;
i04. Determine a set $Z=A \cup B \cup \{p^*\}-P$;
i05. If Z does not worsen the classification of T set, key1=true and key2=true;
i06. If key1=true, $A=A \cup \{p^*\}$ and B=B-P;
i07. If key1=false, $A=A \cup P$ and B=B-P;
i08. If $B \neq \Theta$, go to **i02;**
i09. If $B=\Theta$ and key2=true, B=A and $A=\Theta$ and go to **i01**;
i10. If $B=\Theta$ and key2=false, Z=A and END.

'key1' is to register that two objects p and q have been replaced with one object p*, 'key2' is to recognize that no merger has taken place and that the algorithm should be ended.

The original Chang's method does not perform further reductions of the reference set if the classification error increases with the use of the currently obtained condensed set. In the tests, the author allowed the original procedure to perform further computations in order to examine how much Chang's algorithm may condense the original reference set, that is the learning set. It was necessary to compare the original Chang's algorithm with its modification proposed by the author.

The performed tests indicate that the author's modification of Chang's method is considerably faster than the original algorithm; however, the quality is, unfortunately, much worse. Most frequently, however, it was possible to determine such a degree of condensation which was accompanied by the acceptable lowering of the classification quality. The lowering of the classification quality is not always a monotone function of the condensation degree.

4. CASCADES ALGORITHM

The aforementioned algorithms of the reference set condensation, one of which is based on finding the mutually furthest points and the other is the modification of Chang's algorithm, were incremental and eliminative, respectively, i.e. the size of the condensed set increased or was reduced as a result of a subsequent iteration. Therefore, the question arises whether a combination of both aforementioned types of condensation, i.e. the cascade algorithm of condensation, is more effective than each of these algorithms executed separately.

The two algorithms presented above are based on the following two different types of reduction:

- incremental reduction – in which the condensed set is constructed beginning from an empty set which is then successively increased;
- eliminative reduction – in which the reduced set is initially complete and then its elements are successively deleted until the stop criterion is met.

The performed tests have not explicitly answered the question which of these approaches is more effective.

The analysis of very large sets indicated that the incremental reduction leads to a high level of errors at the initial stage. Much time is required to decrease those errors. The eliminative reduction usually results in a low level of errors at the initial stage of computations; however, much time is needed to achieve a considerable reference set size reduction. The combination of both these condensation types may consist in their sequential application. Firstly, the classifier based on the method of finding the mutually furthest points was used and during the second stage the reduction with the use of the modified Chang's method was applied. A significant issue to be solved was the transition criterion from the first component algorithm to the second one. The misclassification rate was adopted as a basis for its determination. To this aim a plot of dependence between the error rate and the condensation degree (i.e. iteration number) was used. The following two characteristic features of the classification error graph were analyzed:

- intervals of constant (unchanging) error rate – the sequential condensation steps do not considerably lower the classification quality;
- local minima – indicating the specific iteration of the reduction algorithm that offers better classification quality than the condensed set of the previous iteration as well as the condensed set of the following iteration.

5. BUBBLE CONDENSATION ALGORITHM

The proposed algorithm of the reference set condensation (reduction) requires a new decision rule. Therefore, the condensation itself may be considered as the learning of the classifier. However, since it requires the modification of the classification rule, the learning phase as well as the classification phase deserve a detailed description. In the classification phase, the standard version of the 1-NN rule is not used any more.

During the learning phase, the set of hyperspheres, each containing objects from the learning set which belong to one class only, is constructed. These hyperspheres are disjunctive and cover all objects of the learning set. Such hyperspheres are called homogenous hyperspheres. Each of these homogenous hyperspheres is characterized by the following parameters:

- the base point – the point of the learning set which is the centre of a hypersphere;

- the radius – the distance from the base point to the nearest point belonging to the opposite class or to the nearest surface of another hypersphere, constructed earlier;
- the number of points inside the hypersphere.

The algorithm of the learning phase, i.e. the bubble algorithm, selects a random point from the learning set (at the beginning from the complete learning set) and marks it as the base point. Then, the distance from this base point to the nearest point of another class or to the surface of another homogenous hypersphere is sought. This distance is used to determine the range of a homogenous hypersphere. The algorithm searches for the points lying inside this hypersphere, stores their number and then deletes these points. The point from another class which was used to determine the size of the area is deleted as well. The learning set is covered with hyperspheres.

This algorithm is presented below in the form of a pseudo code.

T – the learning set containing m objects;
T^i – sets reduced in individual iterations;
$K^i = (t_i,r_i,n_i)$ – a combination: the centre, a radius, the number of points inside a hypersphere without the base point;
Z^i – a set of points from T set located inside K^i sphere;

00. START; i =2; $T^1 = T$;
01. Choose a random point t_1 from the set T^1, find the distance r_1 to the nearest point y_1 from another class and the number n_1 of points from the T located in the hypersphere $K^1=(t_1,r_1,n_1)$;
02. $T^i = T^{i-1}–Z^{i-1}–\{y_{i-1}\}$, if y_{i-1} does not exist, assume that $\{y_{i-1}\}$ is an empty set;
03. If T^i is empty, go to 06;
04. Choose a random point t_i from the set T^i and determine $K^i = (t_i,r_i,n_i)$:
- distance d_1 from t_i to the nearest point y_i from another class,
- distance d_2 to the nearest, already existing, hypersphere K^j, j = 1,2,..,i (it is computed as a distance to the centre of the sphere minus its radius);
- mark the shorter of these two distances as r_i;
- determine the number of points n_i from the set T^i located in the hypersphere with t_i centre and r_i radius;

05. i = i+1; go to 02;
06. END

6. CONCLUSIONS

The thesis presents new methods of the reference set condensation. The most desired property of each of the methods presented was the possibility of controlling the compromise between the speed and the quality of classification

based on the obtained condensed reference set. Obviously, a good solution would improve the quality of classification and accelerate the computations as well, while reducing the size of the reference set. However, one should bear in mind that classification may be one of the stages of a larger task connected with processing information. In some applications (e.g. quality control), classification must be performed in the real time of the production process, i.e. the classifier must make the decision within the determined time. It is necessary to accept that the recognition system, the component task of which is classification, may increase the percentage of incorrect decisions; it is the time limit for completing the task that is more important. However, the time of making a decision is not the only key issue. There are cases where the classification quality is of the highest priority. In the case of a medical data analysis, it is considerably more important that the system verifies the data precisely than that it makes it fast.

As it was already mentioned, steering between the speed and the quality of classification is of great importance in many cases. This thesis may constitute a valuable contribution towards solving this problem.

REFERENCES

[1] **Białynicka-Birula I.**: Modelowanie rzeczywistości. Prószyński i S-ka S.A., Warszawa 2002, p. 122.

[2] **Chang C.L.**: Finding Prototypes for Nearest Neighbor Classifiers. IEEE Transactions on Computers 1974, tom. C-23, No. 11, pp. 1179-1184.

[3] **Sánchez J.S., Pla F., Ferri F.J.**:) On the use of neighbourhood-based non-parametric classifiers. Pattern Recognition Letters 1997, Vol. 18, No. 11-13, pp. 1179-1186.

[4] **Sánchez J.S., Pla F., Ferri F.J.:** Prototype selection for the nearest neighbour rule through proximity graphs. Pattern Recognition Letters 1997a, Vol. 18, No. 6, pp. 507-513.

[5] **Sanchez J.S., Pla F., Ferri F.J.**: Improving the k-NCN classification rule through heuristic modifications. Pattern Recognition Letters 1998, Vol. 19, No. 13, pp. 1165-1170.

[6] **Sánchez J.S., Barandela R., Marqués A.I., Alejo R., Badenas J.**: Analysis of new techniques to obtain quality training sets. Pattern Recognition Letters 2001, Vol. 24, pp. 1015-1022.

[7] **Lozano M., Jos´E S., Sánchez J.S., Pla F.:** Reducing Training Sets by NCN-Based Exploratory Procedures. F.J. Perales et al. (Eds.): IbPRIA 2003, LNCS 2652, pp. 453-461, Springer-Verlag, Heidelberg, New York.

[8] **Moghaddam B., Pentland A.:** An automatic system for model based coding of faces. IEEE International Conference on Image Processing 1995, Washington DC, USA.

[9] **Skarbka W.**: Multimedia – Algorytmy i Standardy Kompresji AOWPLJ, Warszawa 1998.

METODY KONDENSACJI ZBIORU ODNIESIENIA DLA REGUŁ DECYZYJNYCH OPARTYCH NA FUNKCJI ODLEGŁOŚCI

Streszczenie

W artykule przedstawiono cztery autorskie algorytmy kondensacji zbioru odniesienia, charakteryzujące się możliwością sterowania pomiędzy szybkością a jakością klasyfikacji, opartej na uzyskanym skondensowanym zbiorze odniesienia. Przeprowadzone testy dowodzą, że zaproponowane algorytmy umożliwiają znaczącą redukcję wielkości zbioru odniesienia dla reguły typu najbliższy sąsiad przy jednoczesnym zachowaniu jakości klasyfikacji bliskiej tej, jaką uzyskuje się z zastosowaniem pełnego zbioru uczącego użytego w roli zbioru odniesienia.

Promotor dr hab. inż. Adam Jóźwik
Recenzenci pracy doktorskiej:
prof. dr hab. Zygmunt Ciota, Politechnika Łódzka
dr hab. inż. Leszek Chmielewski, IPPT PAN, Warszawa