Artur Sierszeń*

Reduction of Large Reference Sets with Modified Chang’s Algorithm

1. Introduction

The advantage of the Chang’s algorithm is a considerable reduction of the reference set. Its drawback is relatively low speed. The modification proposed by the author of this article aims at accelerating computations.

2. Chang’s algorithm

The original procedure of Chang’s reduction [1] is presented below in the form of a pseudocode.

**THE ORIGINAL CHANG’S ALGORITHM**

\( T = \{t_1, t_2, \ldots, t_m\} \) – the training set containing \( m \) objects \( t \);

\( t = \{a_1, a_2 \ldots a_n\} \); an element of \( T \) set, \( a \) – a feature describing a point; \( n \) – a number of features

\( Z \) – the current reduced set;

\( \emptyset \) – a null set; \( \text{key1 and key2} \) – working variables of logic type;

i00. START, \( A=\emptyset \), \( B=T \); \( A\)={a random object from \( B \)};

i01. \( \text{key1=false; key2=false; } B=B-A \);

i02. Find \( p \in A \) and \( q \in B \), so that the distance \( d(p,q) \) is minimum;

i03. If \( p \) and \( q \) are from the same class, determine a set \( Z=A \cup B \cup \{p^*\}-\{p,q\} \), where \( p^*=(p+q)/2 \);

i04. If \( p \) and \( q \) are from the same class and \( Z \) is the same as \( T \), \( \text{key1=true and key2=true} \);

i05. If \( \text{key1=true} \), \( A=A-\{p\} \cup \{p^*\} \) and \( B=B-\{q\} \);

i06. If \( \text{key1=false} \), \( A=A \cup \{q\} \) and \( B=B-\{q\} \);

i07. If \( B\neq\emptyset \), go to i02;

* Computer Engineering Department, Technical University of Łódź
i08. If $B = \emptyset$ and key2 = true, $B = A$ and $A = \emptyset$ and go to i01;

i09. If $B = \emptyset$ 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 recognise that no merger has taken place and that the algorithm should be ended.

Figure 1 presents the graphic illustration of the Chang’s algorithm for the set of seven objects (five circles and two triangles). The training set $T$ of objects is presented in Figure 1a. The algorithm starts its operation with the complete reference set, i.e. $Z = T$ (Fig. 1b), which is the same as the set presented in Figure 1a. The nearest neighbour method operating with the reference set presented in Figure 1b classifies correctly all objects from the training set (i.e. the objects from Fig. 1a). Since the objects $A$ and $B$ are nearest for each other and are from the same class, they are replaced with the new object $H$. The label of this object is the same as that of the objects $A$ and $B$. All objects (from Fig. 1a) are still correctly classified, therefore, by replacing $A$ and $B$ with $H$ prototype, a new reduced set of objects, $Z$, is obtained. It is presented in Figure 1c. By analogy, after having replaced $H$ and $C$ with the object $I$, the status, as presented in Figure 1d, is obtained. By merging $F$ and $G$ into the object $J$, the next reduced $Z$ set is obtained, which is presented in Figure 1e. Replacement of $D$ and $E$ with the object $K$ results in the status shown in Figure 1f. The reduced set $Z$ presented in Figure 1f still correctly classifies the initial set (Fig. 1a). If $I$ and $J$ were merged, some objects would be classified incorrectly; therefore the process of merging and replacing objects is finished. Artificially created objects shown in Figure 1f will be used as a reference set for the 1-NN rule.

![Fig. 1. Graphic interpretation of Chang’s algorithm](image)

3. Modified Chang’s algorithm

A modification proposed by the author of this study aims at accelerating computations by replacing a group of objects, rather than a pair of them, with a new object. For any object in the reference set it is possible to determine all objects from the same class which lie
within a shorter distance from it than any object from the opposite class. Then, all these objects are replaced with their gravity centre with the same label as that of the objects from which the gravity centre was computed.

The operation of the proposed modified Chang’s algorithm with an example analogical to that presented in Figure 1 was shown in Figure 2. In the first stage, a distance is found from a randomly selected point (in this case from A) to the nearest prototype representing another class (D) (Fig. 2b). This distance, marked as $x_1$, enables finding all prototypes lying closer than prototype D, that is B and C. After merging all points found in this way, a new point is obtained and marked as H (Fig. 2c). The new object H is found as a gravity centre of these merged points. The next randomly selected point can be F. Again, the distance to the nearest object representing another class (E) is determined and marked as $x_2$. Then, points are sought which are from the same class and which lie within the radius of $x_2$ (Fig. 2c). In this step, only F and G are merged and, as a result, object J is obtained. Analogically to the examples presented above, a new point is selected at random again (this time it is prototype D) (Fig. 2d). A distance to the nearest point from opposite class (prototype H) is determined and marked as $x_3$. Then, all objects from the same class and lying closer than H are found (only prototype E). The found prototype E is merged with D; they are replaced with prototype J.

In each case, after a new set has been determined, its reliability is checked by leave-one-out method applied to 1-NN rule. A detailed description of the modified Chang’s method is presented below.

**MODIFIED CHANG’S ALGORITHM**

$T$ – the training set containing $m$ objects $t$;
$Z$ – the current reduced set; $P$ – the working set;
$\emptyset$ – a null set; key1 and key2 – working variables of logic type;

$i00$. START, $A=\emptyset, B=T; A=$ \{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 for \( p \) object;

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 \( B \neq \emptyset \), go to i02;

i09. If \( B=\emptyset \) and key2=true, \( B=A \) and \( A=\emptyset \) and go to i01;

i10. If \( B=\emptyset \) 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 recognise that no merger has taken place and that the algorithm should be ended.

4. Verification of the algorithm

The algorithm of reference set condensation based on finding the mutually furthest points which are used to determine a cutting hyperplane was implemented in C++ in Microsoft Visual Studio .NET 2003 environment. This allowed the author to test the method in Windows environment with the use of a PC computer equipped with Intel Pentium processor 4 HT 3 GHz and 512 MB of operating memory.

The computation tests were conducted with the use of sets from the repository of the University of California in Irvine (Machine Learning Repository, University of California, Irvine) [2]. These tests are commonly used in literature. These are the following (Tab. 1):

- **PHONEME** – data set created as a result of an analysis of separate syllables pronunciation (e.g. pa, ta, pan etc.); what was taken into account in this analysis was the type of a vowel pronunciation – nasal or oral;

- **SATIMAGE** – this data set was generated basing on the analysis of satellite pictures supported with other methods of observation (radar data, topographic maps, data concerning agriculture). Classes determine a kind of soil or a type of cultivation;

- **WAVEFORM** – artificially generated data set, where each of the classes is created as a result of a combining 2 out of 3 sinusoids; for each attribute in a class noise is generated.

All tests were repeated 25 times; the presented results (time of the algorithm’s operation) were calculated as the average values obtained during these tests. Computing the error with the leave-one-out method was definitely most time-consuming. Therefore, the author decided that the error rate should be computed every second iteration. This enabled the considerable acceleration of computations.
4.1. PHONEME testing set

The charts present the results of the original Chang’s algorithm and the modified Chang’s algorithm with the use of the PHONEME testing set. The charts contain the speed of condensation as a function of the number of iterations (Fig. 3), the influence of condensation on the classification error rate (Fig. 4), and time required to perform the condensation of the set with both of the methods (Fig. 5). Additionally, the comparison was made (Tab. 2) between the exemplary results of the original and the modified Chang’s algorithm for the PHONEME set.

**Table 1**
Parameters of the sets used during the tests

<table>
<thead>
<tr>
<th>Name of the set</th>
<th>Number of classes</th>
<th>Number of features</th>
<th>Number of samples</th>
<th>Size of separate classes in the set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>PHONEME</td>
<td>2</td>
<td>5</td>
<td>5404</td>
<td>3818</td>
</tr>
<tr>
<td>SATIMAGE</td>
<td>6</td>
<td>36</td>
<td>6435</td>
<td>1533</td>
</tr>
<tr>
<td>WAVEFORM</td>
<td>3</td>
<td>21</td>
<td>5000</td>
<td>1657</td>
</tr>
</tbody>
</table>

**Fig. 3.** Size of condensed reference sets for the PHONEME set related to the number of iterations:

a) the original Chang’s algorithm; b) the modified Chang’s algorithm

**Fig. 4.** The influence of the PHONEME set condensation on the quality of classification with the 1-NN method in case when the used condensation algorithm was:

a) the original Chang’s algorithm; b) the modified Chang’s algorithm
By comparing the operation of Chang’s algorithm with the operation of the modified algorithm, it can be noticed that the number of iterations required to obtain the condensed reference set is reduced. Unfortunately, it is accompanied by a considerable increase in a level of incorrect decisions. Attention should be paid to the initial stage of the test, when the minimum increase of the error rate (0.26\% increase) was accompanied with reduction by 404 objects (which constitutes app. 7.5\% of the initial set); the operation of the condensation algorithm accelerated by almost 8 times. The reduction by 1393 elements (app. 25\% of the initial set) causes increasing of the error rate by 2.61\%, which does not seem a high price; the error rate of 1-NN method for this set amounts to 8.97\%.

### 4.2. SATIMAGE testing set

The charts present the results of the original Chang’s algorithm and the modified Chang’s algorithm with the use of the SATIMAGE testing set. The charts shows the speed of
condensation as a function of the number of iterations (Fig. 6), the influence of condensation on classification error rate (Fig. 7), and time required to perform the condensation of the set with both of the methods (Fig. 8). Additionally, the comparison was made (Tab. 3) between the exemplary results of the original and the modified Chang’s algorithm for the SATIMAGE set.

Fig. 6. Size of condensed reference sets for the SATIMAGE set related to the number of iterations:
   a) the original Chang’s algorithm; b) the modified Chang’s algorithm

Fig. 7. The influence of the SATIMAGE set condensation on the quality of classification with the 1-NN method in case when the used condensation algorithm was:
   a) the original Chang’s algorithm; b) the modified Chang’s algorithm

Fig. 8. Time of the SATIMAGE condensation: a) with the Chang’s algorithm;
   b) with the modified Chang’s algorithm
As in the case of the first testing set, by comparing the operation of the Chang’s algorithm and the modified algorithm, it can be noticed that the number of iterations required to obtain a condensed reference set is reduced; unfortunately, it is accompanied with an increase in the percentage of incorrect decisions. As opposed to the other testing sets, considerable fluctuations of the increase in classification error can be noted, what is caused by the specific nature of the set (the last and biggest class has large clusters of points).

At the beginning of the experiment, the reduction by 413 objects of the reference set (app. 6.4\% of the initial set) was observed; the computations accelerated by over 16 times. It was accompanied by an increase in the error rate of 0.99\%. The reduction by 1435 elements (app. 22\% of the initial set) causes increasing of the error rate by 8.69\% (the error of 1-NN method for this set amounts to 8.89\%); the operation of the algorithm accelerates by 6 times.

### 4.3. WAVEFORM testing set

The charts present the results of the original Chang’s algorithm and the modified Chang’s algorithm with the use of the WAVEFORM testing set. The charts contain the speed of condensation as a function of the number of iterations (Fig. 9), the influence of condensation on classification error rate (Fig. 10), and time required to perform the condensation of the set with both of the methods (Fig. 11). Additionally, the comparison was made (Tab. 4) between the exemplary results of the original and the modified Chang’s algorithm for the WAVEFORM set.

<table>
<thead>
<tr>
<th>Size of the reduced set</th>
<th>The original algorithm</th>
<th>The modified algorithm</th>
<th>Difference of computation time*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Iteration</td>
<td>Computation time [s]</td>
<td>Classification error [%]</td>
</tr>
<tr>
<td>6022</td>
<td>413</td>
<td>21237</td>
<td>0.00</td>
</tr>
<tr>
<td>5000</td>
<td>1435</td>
<td>70288</td>
<td>0.00</td>
</tr>
<tr>
<td>3961</td>
<td>2474</td>
<td>112202</td>
<td>0.26</td>
</tr>
<tr>
<td>3000</td>
<td>3435</td>
<td>145984</td>
<td>0.34</td>
</tr>
<tr>
<td>2000</td>
<td>4435</td>
<td>175343</td>
<td>0.48</td>
</tr>
<tr>
<td>955</td>
<td>5480</td>
<td>199968</td>
<td>2.05</td>
</tr>
</tbody>
</table>

* the ratio of the computation time of the original method to the computation time of the modified method
Comparing the results of both algorithm’s operation with \textit{WAVEFORM} set, it can be noticed that using the modified algorithm can result in a considerable reduction of a number of iterations required to obtain a condensed reference set. Unfortunately, like in the other sets, it is accompanied by an increase in the percentage of incorrect decisions.

\textbf{Fig. 9.} Size of condensed reference sets for the \textit{WAVEFORM} set related to the number of iterations:

a) the original Chang’s algorithm; b) the modified Chang’s algorithm

\textbf{Fig. 10.} The influence of the \textit{WAVEFORM} set condensation on the quality of classification with the 1-NN method in case when the used condensation algorithm was:

a) the original Chang’s algorithm; b) the modified Chang’s algorithm

\textbf{Fig. 11.} Time of the \textit{WAVEFORM} condensation: a) with the Chang’s algorithm; b) with the modified Chang’s algorithm
The main goal of the research concerning the modification of the Chang’s algorithm was to accelerate computations concerned condensation. What decided about selecting this algorithm was a possibility to steer between the quality and the speed of classification. The conducted tests indicate that the author’s modification of the Chang’s algorithm is significantly faster than the original algorithm; unfortunately, the deterioration of classification quality is considerable. However, most frequently it was possible to determine the level of condensation which was accompanied by an acceptable deterioration of classification quality. Deterioration of the classification quality is not always a monotonous function of the condensation level.

In the charts, the periods of more dynamic reduction can be noticed. This was the case when the algorithm found a homogenous area, where the reduction of the larger amount of points could be performed faster.
References
