AUTOMATYKA • 2009 • Tom 13 • Zeszyt 3

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 = {t1, t2, …, tm} – the training set containing m objects t;
t = [a1, a2 … an]; an element of T set, a – a feature describing a point; n – a number of features
Z – the current reduced set;
∅ – a null set; key1 and key2 – working variables of logic type;
i00. START, A=∅, 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≠∅, go to i02;

* Computer Engineering Department, Technical University of Łódź

i08. If B=∅ and key2=true, B=A and A=∅ and go to i01;
i09. If B=∅ 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.

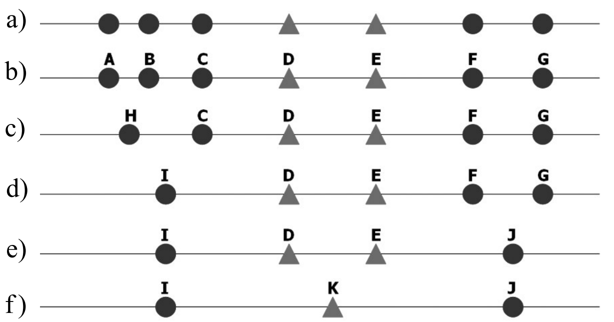

Fig. 1. Graphic interpretation of Chang's algorithm

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 selected 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.

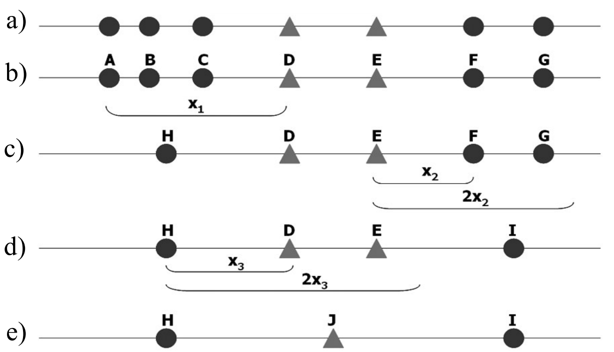

Fig. 2. Graphic illustration of modified Chang's algorithm

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;
∅ – a null set; key1 and key2 – working variables of logic type;
i00. START, A=∅, 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.

Table 1
Parameters of the sets used during the tests

Name of the set	Number of classes	Number of features	Number of samples	Size of separate classes in the set					
				1	2	3	4	5	6
PHONEME	2	5	5404	3818	1586	–	–	–	–
SATIMAGE	6	36	6435	1533	703	1358	626	707	1508
WAVEFORM	3	21	5000	1657	1647	1696	–	–	–

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.

a) b)

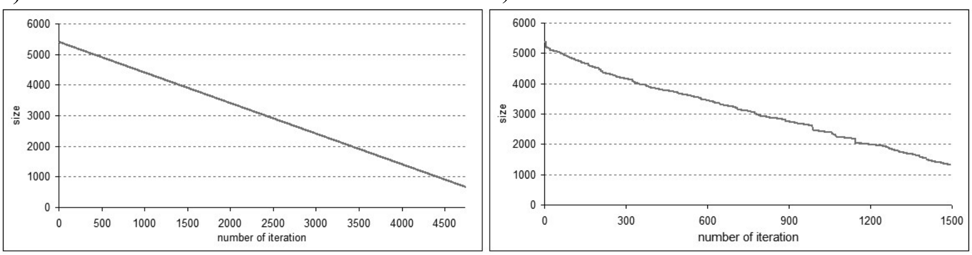

Fig. 3. Size of condensed reference sets for the *PHONEME* set related to the number of iterations: a) the original Chang's algorithm; b) the modified Chang's algorithm

a) b)

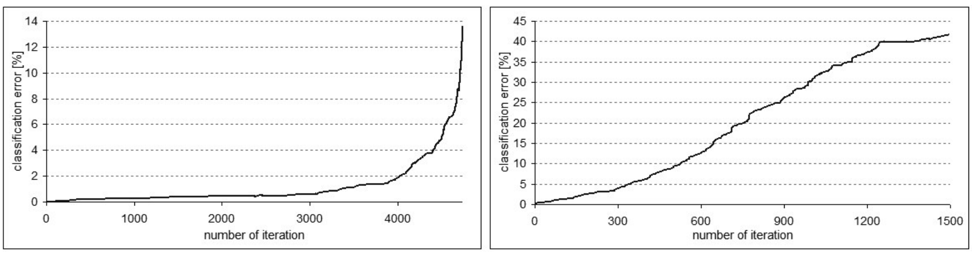

Fig. 4. The influence of the *PHONEME* set condensation on the quality of classification with the 1-NN method in case when the used condensation algorithm was: a) the original Chang's algorithm; b) the modified Chang's algorithm

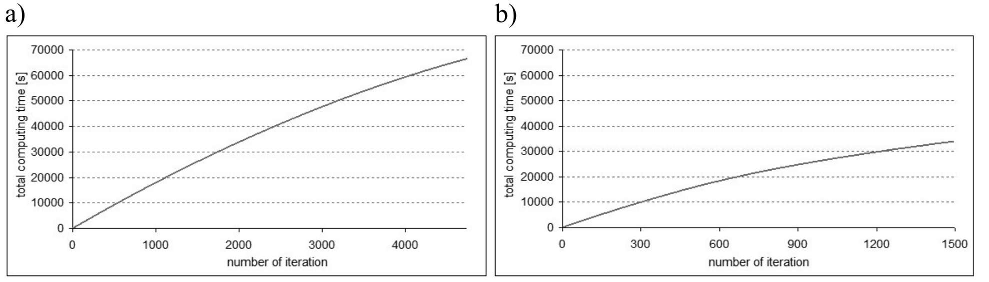

Fig. 5. Time of the *PHONEME* condensation: a) with the Chang's algorithm; b) with the modified Chang's algorithm

Table 2
The comparison of the original and the modified version of Chang's algorithm regarding to computation time and classification error rate for the *PHONEME* set

Size of the reduced set	The original algorithm			The modified algorithm			Difference of computation time*
	Iteration	Computation time [s]	Classification error [%]	Iteration	Computation time [s]	Classification error [%]	
5000	404	7496	0.20	58	945	0.46	7.93
4504	900	16245	0.26	200	6823	2.61	2.38
4011	1393	24437	0.35	343	11281	5.03	2.17
3484	1920	32638	0.43	575	17743	12.01	1.84
3001	2403	39648	0.46	777	22337	20.30	1.77
2471	2933	46796	0.57	988	26277	29.27	1.78

* the ratio of the computation time of the original method to the computation time of the modified method

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.

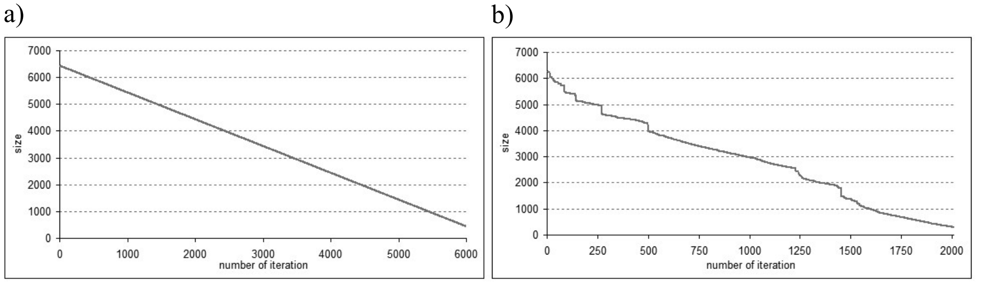

Fig. 6. Size of condensed reference sets for the *SATIMAGE* set related to the number of iterations: a) the original Chang's algorithm; b) the modified Chang's algorithm

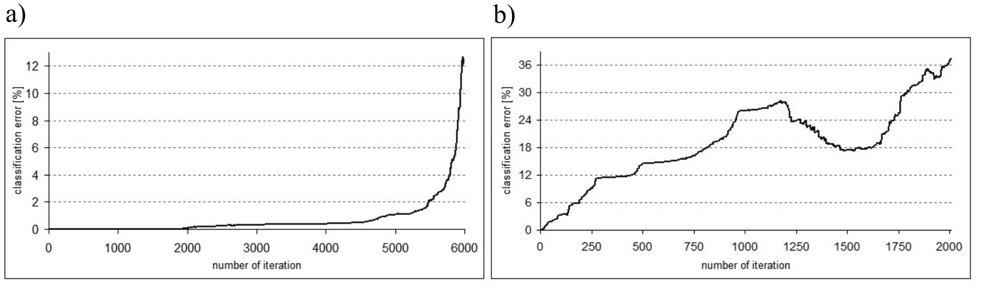

Fig. 7. The influence of the *SATIMAGE* set condensation on the quality of classification with the 1-NN method in case when the used condensation algorithm was: a) the original Chang's algorithm; b) the modified Chang's algorithm

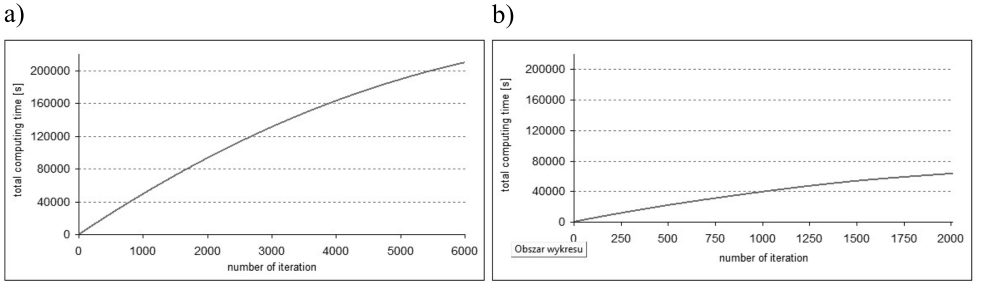

Fig. 8. Time of the *SATIMAGE* condensation: a) with the Chang's algorithm; b) with the modified Chang's algorithm

Table 3
The comparison of the original and the modified version of Chang's algorithm regarding computation time and classification error rate for the *SATIMAGE* set

Size of the reduced set	The original algorithm			The modified algorithm			Difference of computa-tion time*
	Itera-tion	Computation time [s]	Classification error [%]	Itera-tion	Computation time [s]	Classification error [%]	
6022	413	21237	0.00	26	1325	0.99	16.03
5000	1435	70288	0.00	231	10885	8.69	6.46
3961	2474	112202	0.26	499	21976	14.48	5.11
3000	3435	145984	0.34	981	38886	26.06	3.75
2000	4435	175343	0.48	1343	49802	21.49	3.52
955	5480	199968	2.05	1589	55648	17.73	3.59

* the ratio of the computation time of the original method to the computation time of the modified method

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.

Comparing the results of both algorithm's operation with *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.

a) b)

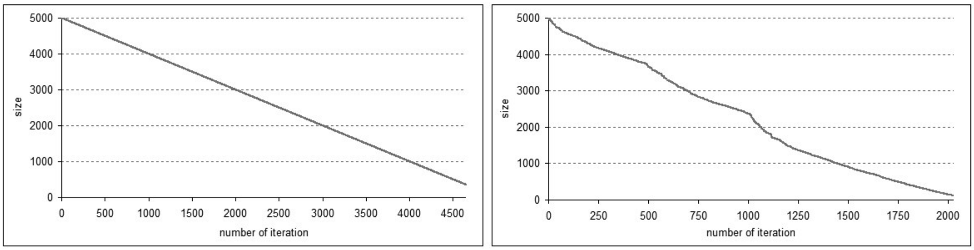

Fig. 9. Size of condensed reference sets for the *WAVEFORM* set related to the number of iterations: a) the original Chang's algorithm; b) the modified Chang's algorithm

a) b)

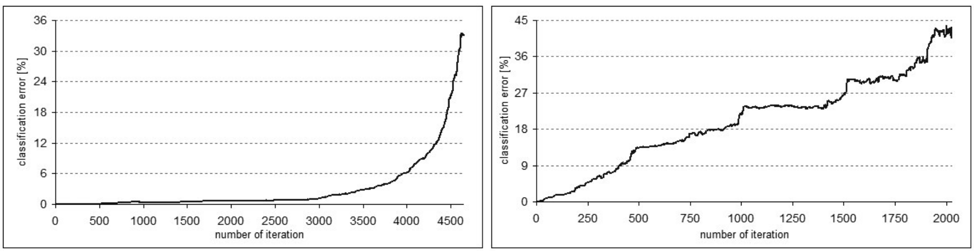

Fig. 10. The influence of the *WAVEFORM* set condensation on the quality of classification with the 1-NN method in case when the used condensation algorithm was: a) the original Chang's algorithm; b) the modified Chang's algorithm

a) b)

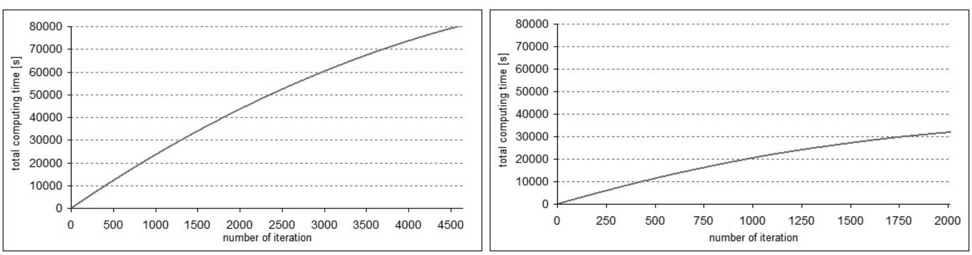

Fig. 11. Time of the *WAVEFORM* condensation: a) with the Chang's algorithm; b) with the modified Chang's algorithm

Table 4
The comparison of the original and the modified version of Chang's algorithm regarding computation time and classification error rate for the *WAVEFORM* set

Size of the reduced set	The original algorithm			The modified algorithm			Difference of computation time*
	Iteration	Computation time [s]	Classification error [%]	Iteration	Computation time [s]	Classification error [%]	
4501	499	12185	0.08	125	3048	1.78	4.00
3998	1002	23586	0.34	340	7939	6.90	2.97
3498	1502	34047	0.52	544	12274	13.56	2.77
2998	2002	43659	0.68	695	15181	14.92	2.88
2499	2501	52386	0.78	933	19355	18.82	2.71
1998	3002	60317	1.14	1061	21436	23.56	2.81

* the ratio of the computation time of the original method to the computation time of the modified method

In the first stage of the test, the increase in the error rate by 1.70% accompanied the reduction by 499 objects (almost 10% of the initial set) and 4-times acceleration, compared to the original version of the algorithm. Reduction by 1002 elements (app. 20% of the initial set) resulted in the increase in the error rate by 6.56%; however, the computations accelerated almost 3 times. This does not seem much; the error of the 1-NN method operating with the initial reference set for this set amounts to 14.67%.

5. Conclusions

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

[1] Chang C.L., *Finding Prototypes for Nearest Neighbor Classifiers*. IEEE Transactions on Computers, t. C-23, no. 11, 1974, 1179–1184.

[2] Merz C.H., Murphy P.M., *UCI repository of machine learning databases*. 1996, http://www.ics.uci.edu/~mlearn/MLRepository.html.