Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 82 (2016) 12 – 19

Procedia Computer Science

Symposium on Data Mining Applications, SDMA2016, 30 March 2016, Riyadh, Saudi Arabia

DTW-Global Constraint Learning using Tabu search algorithm

Bilel Ben Ali[a,*], Youssef Masmoudi[b], Souhail Dhouib[c]

[a] *Faculty of Economics and Management of Sfax, University of Sfax, Tunisia*
[b] *Saudi Electronic University, Riyadh, Saudi Arabia*
[c] *Higher Institute of Industrial Management of Sfax, University of Sfax, Tunisia*

Abstract

Many methods have been proposed to measure the similarity between time series data sets, each with advantages and weaknesses. It is to choose the most appropriate similarity measure depending on the intended application domain and data considered. The performance of machine learning algorithms depends on the metric used to compare two objects. For time series, Dynamic Time Warping (DTW) is the most appropriate distance measure used. Many variants of DTW intended to accelerate the calculation of this distance are proposed. The distance learning is a subject already well studied. Indeed Data Mining tools, such as the algorithm of k-Means clustering, and K-Nearest Neighbor classification, require the use of a similarity/distance measure. This measure must be adapted to the application domain. For this reason, it is important to have and develop effective methods of computation and algorithms that can be applied to a large data set integrating the constraints of the specific field of study. In this paper a new hybrid approach to learn a global constraint of DTW distance is proposed. This approach is based on Large Margin Nearest Neighbors classification and Tabu Search algorithm. Experiments show the effectiveness of this approach to improve time series classification results.

Peer-review under responsibility of the Organizing Committee of SDMA2016

Keywords: Dynamic Time Warping; Tabu Search; Data Mining; KNN classification; Time series; DTW-Global Constraint Learning

1. Introduction

With the availability of enormous time series databases, such as bio-metrics and weather, there has been an explosion of interest in the exploration of these data. Many methods and algorithms have been proposed to classify, index, segment and discriminate time series.

Many classification methods, similarity measures and algorithms have been developed over the years, mostly for survey data. Unfortunately, most of these similarity measures can not be used directly on the time series data sets. New distances and new strategies have been identified, some of which are based on relatively recent tools or results of the analysis of time series: cepstrum coefficients, wavelet transform, hidden Markov models etc. Classification[1,2] is to group objects into classes whose have similar features and content. The objects of the same class are similar and objects from different classes are different. Each classification method is thus based on a ”similarity - dissimilarity”

* Corresponding author. Tel.: +216 97 496 468.
E-mail address: bilel_benali@yahoo.fr

Peer-review under responsibility of the Organizing Committee of SDMA2016
doi:10.1016/j.procs.2016.04.003

measure between objects, a measurement of "similarity - dissimilarity" between classes and aggregation strategy used to build the classes. Many classification methods are available in standard statistical software: partitioning methods (K-means, K-means clustering etc.) Self Organizing Maps and hierarchical methods etc.
Hundreds of distances have been proposed to classify time series, among which Euclidean distance is the most popular. But when it comes to classify time series using Euclidean distance, and any other Minkowski metric, can lead to very intuitive results. In particular, this distance is very sensitive to scale effects, the presence of atypical or missing items and does not take into account possible time lags. One way to solve these problems is to define new distances and similarity measures: Dynamic Time Warping is one of distance measure commonly used for time series data sets. We find that the use of DTW with KNN gave bad results for some instances such as "Swedish Leaf" given by works of Keogh[3]. To improve classification results we will consider to adapt the DTW distance to the studied case using distance learning[4]. The idea is to learn parameters of DTW using a hybridization of the Large Margin Nearest Neighbors and Tabu Search algorithms.
The paper is organized as follows: In the first section it is to present preliminary concepts: Time series and similarity measures, Large Marge Nearest Neighbors (LMNN) classification and Tabu Search (TS) algorithm. The second section is about the proposed approach: using TS algorithm and LMNN classification to learn a DTW Warping Window (DTWWW). A method to condense time series data set used in learning process is presented. The data condensing method minimize the the learning CPU Time. In the fourth section experiments are presented and finally the paper finish by a conclusion and future works.

2. Literature review and related works

2.1. Metric learning review

A lot of work on learning metrics and similarities is about learning the parameters of a Mahalanobis distance. The squared Mahalanobis distance, defined by $D^2_M(x_1, x_2) = (x_1 - x_2)^T M(x_1 - x_2)$, is parameterized by the Positive Semi-Definite(PSD) matrix M. The PSD constraint ensures that D_M is a (pseudo) metric, which allows acceleration of the k-NN classification based on the triangle inequality. Different literature methods differ primarily in the selection of the objective function and the regularization term. For example, in[5], authors forced examples of the same class to be closer than examples of different classes by some margin. In[6] the objective function is related to the error of the k-NN on the training set. Davis and Kulis[7] regulate with the divergence LogDet (which automatically imposes the PSD constraint) while Ying and Huang[8] use the norm (2.1) that promote the learning of a matrix M of low rank. There are also online learning methods, such as POLA[9] and LEGO[10]. The most costly aspect of many of these approaches is the satisfaction of the PSD constraint, although some methods are able to reduce the cost of computing by developing specific solvers.
Some research focuses on learning other types of distances. Qamar[11] optimizes a cosine similarity to treat information retrieval tasks. In the field of image recognition, Frome and Singer[12] learn a local distance for each example, while Chechik and Shalit[13] propose an online learning procedure for bi-linear similarity measure.
The information used in supervised metric learning is of two types: (i) constraints based on pairs of examples: x and y must be similar (or dissimilar), and (ii) the constraints based on examples triples: x must be more similar to y than z. Note that the two types of constraints can be built from labeled data. The objective is to find the metric or similarity that best satisfies these constraints. All methods presented above are generally used in the context of the Nearest Neighbors (NN) (and sometimes clustering). This is due to the fact that constraints based on pairs or triplets are easy to obtain and optimize the sense in the context of the k-NN or clustering algorithms that are based on local neighborhoods.

2.2. Related works

A common way to obtain a family of metrics on a vector space X is to consider the Euclidean distance after the linear transformation $x' = Lx$. These metrics calculate the square distances as given by equation (1).

$$d_L(x_i, x_j) = \|L(x_i - x_j)\|_2^2 \tag{1}$$

Where the linear transformation in equation (1) is set by the matrix L. Furthermore, it is often appropriate to express the squared distance in the equation (1) using the squared matrix $M = L^T L$. The squared distance $d_L(x_i, x_j)$ becomes as equation (2).

$$d_M(x_i, x_j) = (x_i - x_j)^T M(x_i - x_j) : \tag{2}$$

The pseudo-metrics of this form are called Mahalanobis metric. It has been shown that metric learning is very useful when combined with k-NN and other techniques depending on distance or similarity measures[6]. Different metrics learning methods have been considered by various authors in the literature. But the Mahalanobis distance remains by far the most used in practice.

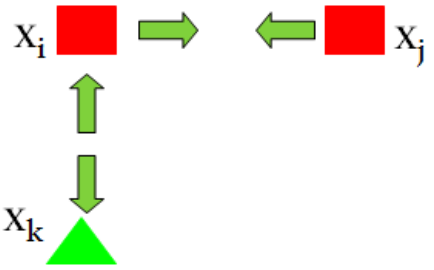
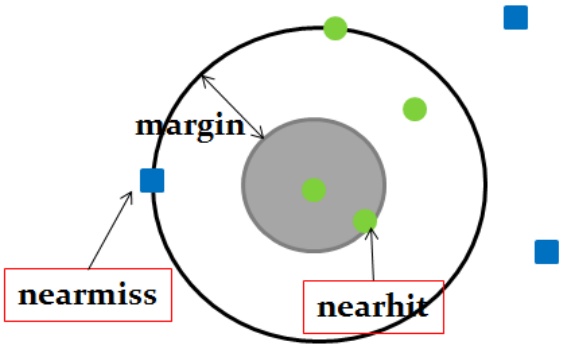


Fig. 1. Bring Target neighbors closer and examples of different classes farther

A good distance function should make the intra-class samples closer and the inter-class samples farther, so a large sample-margin can be obtained as shown in Figure (2). The margin is given by the equation (3).

$$margin = \|x - nearmiss\| + \|x - nearhit\| \tag{3}$$



$$margin = \|x - nearmiss\| + \|x - nearhit\|$$

Fig. 2. Large Margin Nearest Neighbors

A global constraint is represented by an array $S = [s_1, s_2, .., s_k]$ where s_i ($i \in [1 .. k]$) is the height above the diagonal in y axis and the width to the right of the diagonal in x axis. of time series length. Each constraint is evaluated using the evaluation function defined in equation (4)(R_{DTW_S} is the radius of the training set).

$$f(S) = \sum_{x \in T} \frac{DTW_S(x, nearmiss) - DTW_S(x, nearhit)}{R_{DTW_S}} \tag{4}$$

The aim is to find the constraint S that maximizes the 1-NN accuracy rate. It is a time-consuming process to search for nearhit and nearmiss under different constraint condition. In order to reduce the computational cost, a technique

is designed for prototype condensing [Section (5)]. In our work we use this technique to reduce the time to search nearhit and nearmiss of an instance from the training set.

3. Preliminary Concepts and Properties

3.1. Time series similarity measures

Several methods have been proposed to measure the degree of similarity between the time series. Away most used is the Euclidean distance. For two vectors C and Q of size N, the Euclidean distance is defined as equation (5).

$$d(x, y) = \sum_{i=1}^{N} (x_i - y_i)^2 \tag{5}$$

The Euclidean distance is commonly accepted as the simplest distance between time series. This distance considers that the dimensions are not structured. Also it mixes the order of acquisitions does not alter the result, since the addition is commutative. This distance requires two time series with same lengths. If the observed phenomena underwent temporal distortions, this distance is on-separated data. For example, two sequences $< 2, 2, 5, 2 >$ and $< 2, 5, 2, 2 >$ will be relatively distant from the point of view of the Euclidean distance, so that they represent similar trends. As the Euclidean distance has no upper limit and that its value increases with the number of features N, it is advisable to calculate the normalized Euclidean distance. Moreover, this distance ignores temporal dependencies between different sets of data. These two constraints do not allow to compare the shape of the signal, which is inconsistent with the purpose of classification in our case. To solve the problem of distortion in the time series, Sankoff and Kruskal[14] presented the Dynamic Time Warping (DTW) distance. DTW allows an elastic shifting of time axis.

DTW can be used to compare two time series of different dimension. The principle is to set the distance corresponding to sub-sequences which ”resemble” even if they do not correspond to the same time interval. The paired points of the two time series contributes to the calculation of the distance DTW. According to this principle, the DTW tends to explain variations in the Y axis by deforming the X axis. However, this may lead to undesirable alignments. To address this problem a DTW global constraint is applied that is well known: Sakoe-Chiba band[15]. This constraint is to define a band around the diagonal path.

3.2. Large Margin Nearest Neighbors Classification

In the proposed approach to learning a global constraint DTW, the wide margin nearest ’neighbors algorithm (LMNN) is used. It is also interesting to draw a parallel with the large margin metric learning approaches in the literature[6]. The method proposed by Weinberger and Saul is to learn a Mahalanobis metric that maximizes the margin between the examples of each class. However, there is an important difference with our work: learning focuses the covariance matrix, leading to a convex optimization problem. Then the covariance matrix learned in their approach is full. This allows great flexibility in modeling links between variable but also requires a lot of learning settings. In our work, the criterion of separation between the instances of different classes is used as an objective function (equation (4)) and as defined in[16].

3.3. Tabu Search algorithm

Tabu Search (TS) algorithm was proposed by Fred Glover in 1986[17,18]. Since then, the method has become very popular thanks to its successes to solve many problems. It is an heuristic of local search method used to solve complex and/or problems of very large size. Revolution of this method compared to the others overcomes the problem of local optima by the use of Tabu List (TL).
It is a method for adaptive memory:

- Short-term memory: diversification.
- Long-term memory: intensification.

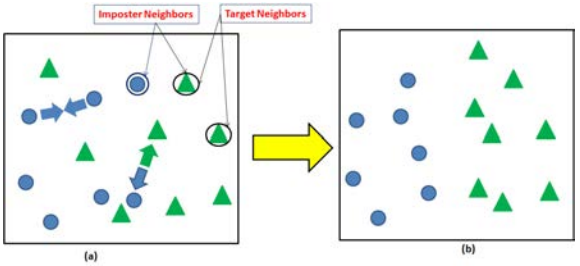


Fig. 3. Large Margin Nearest Neighbors

TS uses a Tabu List that contains movements that are temporarily banned. These movements are prohibited solutions. Its changing role in the resolution: diversification (exploration of the solution space) to intensification.
The neighborhood:

- Let S be the space of possible solutions, we call **movement**, the modification of a subset of components of a solution. A movement can explore the space of possible solutions from one solution to another,
- The neighborhood of a solution s noted $V(s)$ is the set of accessible solutions obtained with a single movement.

Algorithm:
The method is to move from one solution to another by observing the vicinity of the starting solution and define tabu transformations that are kept in memory. A tabu transformation shall not be applied to the current solution. In a first phase, the research method Tabou can be seen as a generalization of local improvement methods. Indeed, starting from a solution x belonging to the set of solutions S, we are moving towards a solution $S(x)$ belonging to the neighborhood $V(x)$ of x. The algorithm iteratively explores the set of solutions S. In order to choose the best neighbor solutions $S(x)$, the algorithm evaluates the objective function f at each point $V(x)$, and retains the neighbor that enhances the value of f.

4. Tabu Search Global Constraint Learning

The KNN algorithm is among the simplest classification algorithms. In a classification context of a new instance x, the basic idea is to vote the nearest neighbors of this instance. The class of x is determined by the majority class among the k nearest neighbors of this instance. This supervised non-parametric method is often effective. Furthermore, learning is quite simple, because it is of type learning by rote (we keep all learning samples). However, the prediction time is very long because it requires calculating the distance with all instances. Several studies have shown that the use of 1-NN with DTW further improves the classification results. Empirical evaluations on more than 40 data sets have showed that 1-NN classifier used with DTW outperforms most of other techniques used in time series classification.
As we use the DTW distance to calculate the distance between time series. The major inconvenient of DTW distance is that it has a quadratic complexity. The solution is to use a global constraint of DTW to accelerate its calculation. The well-known global constraint is Sakoe-Chiba band proposed by Sakoe and Chiba[15]. This constraint is to restrict the warping path to window around the diagonal path [Figure (**??**)].

5. Data condensing

It is a time consuming process to search for nearhit and nearmiss under different constraint condition. Condensing methods are developed to pick out a consistent subset of prototypes for a problem.
Given training set T, we need nearhit and nearmiss of x from $T - x$ to evaluate the constraint S. DTW distance is upper bounded by the Euclidean distance because the accumulated distance on the distance matrix is minimized by

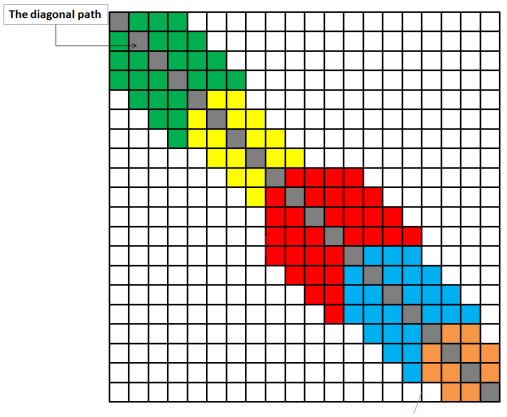


Fig. 4. The Warping Window

warping path. With a specific global path constraint S, DTW_S is lower or equal to DTW distance. For each element $x \in T$, eliminate $y \in T$ verifying:

$DTW(x, y) >$
$\max(D_{Euclidean}(x, neahit(x)), D_{Euclidean}(x, nearmiss(x))$ The rival set is used after condensing to evaluate the global constraint.

6. Experiments

In this section we present the results of classification in order to measure the accuracy of our approach. In order to prove the efficiency of the new approach, we present the 1-NN classification results for some instances presented in [?]. The distance used by 1-NN algorithm is DTW with Global Constraint learned by Tabu Search algortihm (DTW-GC TS) and results are compared with the approach used in [16]. First the effect of changing the different of parameters of TS algortihm are presented in order to fix the best parameters then in the section (6.4) a comparison between 1-NN with DTW-CG TS and LMNN-DTW presented in [16]. Results shows the effectiveness of our approach. 1-NN with the learned DTW gives best results for almost all data sets.

6.1. Varying the number of iterations

It is clear and evident that the second parameter of Tabu search algorithm (number of iterations without improvement of the solution) (Table (1)) is very effective. By increasing this parameter we obtain best and closest solutions to the optimal one.
Increasing the value of this parameter will increase the CPU time necessary to converge. As shown in the Table (1), lower approximation error rates were found with a small value of the iteration number. As example, for the instance "Adiac" the best precision is obtained with a number of iterations equal to 10 and for instance Gun_Point with an iteration number equal to 50.

6.2. Varying the size of Tabu List

The variation in the size of the Tabu List (TL) (Table (2)) affects the CPU time: if the size of the Tabu List is large, the number of exempted neighbors of a current solution is minimized which decline the CPU time. This change can also affect the cost of the obtained solution. The size of Tabu List is an important parameter affecting the execution time and the cost of the optimal solution. For our tests, we change the size of TL and with a number of iterations equal to 100. As shown in the table (2) good results are obtained with small size of Tabu List. This means that with tabu search and with small value of Tabu List (lower CPU time) accurate results can be reached. For example, for the instances "Gun_Point", "SwedishLeaf", "FaceFour", "ECG200" and "CBF", the accurate obtained results are with a

Table 1. Changing the number of iterations

Data sets/ItNb	10	50	100	200
Gun_Point	0.12	**0.02**	0.04	0.0267
SwedishLeaf	0.146	0.2064	**0.1392**	0.1456
50words	0.303	0.312	0.303	**0.283**
FaceFour	0.148	0.318	1477	0**.1363**
ECG200	0.11	**0.1**	**0.1**	**0.1**
CBF	0.042	**0.022**	0.0289	0.0289
Adiac	**0.339**	0.394	0.414	0.414

size of Tabu List equal to 2. For the instances "50words" and "Adiac" the best results are obtained with a size of Tabu List equal to 5.

Table 2. Changing the size of Tabu List

Data sets/TL size	2	5	10	20
Gun_Point	**0.02**	0.047	0.0267	0.0267
SwedishLeaf	**0.1392**	**0.1392**	0.1456	0.1648
50words	0.281	**0.280**	0.303	0.312
FaceFour	**0.148**	0.170	0.170	0.170
ECG200	**0.1**	0.11	0.11	0.11
CBF	**0.0156**	0.0178	0.0289	0.022
Adiac	0.412	**0.394**	**0.394**	**0.394**

6.3. Varying the number of global constraint segments

Varying the number of segment of the global constraint (Table (3)) can affect the accuracy of the classification results. For FaceFour instance, the higher accuracy is obtained for a global constraint composed by 8 segments. For Gun_Point instance, the higher accuracy is obtained by 4 segments.

Table 3. Changing the number of global constraint segments

Data sets/Segments number	4	5	8	10
Gun_Point	0.27	0.04	**0.0267**	**0.0267**
SwedishLeaf	**0.1392**	0.152	0.2288	0.2304
50words	0.290	0.286	0.266	**0.264**
FaceFour	0.330	0.193	**0.114**	0.125
ECG200	0.16	**0.11**	0.12	0.16
CBF	0.0289	0.061	0.0155	**0.012**
Adiac	0.417	0.412	**0.394**	0.407

6.4. Best results of Tabu Search

In this section, the best results obtained by Tabu search are presented. In the previous sections, we have presented the effect of each parameter of Tabu search on the classification results and the CPU time. Table (4) presents the best accuracy rate obtained by Tabu search (colunmn "DTW-GCL TS"). As shown in this table, using Tabu search to learn the size of the warping window gives accurate results. Compared with 1-NN using Euclidean distance, DTW, Best WWDTW and LMNN-DTW, 1NN with DTW-GCL TS gives the higher accuracy rate for five instances ("Gun_Point", "SwedishLeaf","FaceFour","ECG200" and "Adiac"). For the instance "50Words", the best result is given by Best Warping Window DTW. For the instance "CBF", the higher accuracy rate is given by the DTW distance.

Table 4. Best values

Data sets	LMNN-DTW (Yu, 2011)	DTW-GCL TS
Gun_Point	0.027	**0.02**
SwedishLeaf	0.152	**0.1392**
50words	0.292	**0.264**
FaceFour	0.180	**0.114**
ECG200	0.11	**0.1**
CBF	**0.05**	0.089
Adiac	0.396	**0.339**

7. Conclusion

It is a common case that two time series are out of phase, even they are from same class. Choosing an appropriate size of DTW global constraint improves the classification results. In this paper, we introduced a learning algorithm based on Tabu search algorithm. The optimal size of the warping window is determined using Tabu search algorithm. Then this optimal constraint is used in the classification task. Results shows the efficiency of the learning algorithm. The classification results are improved for almost all data sets. A condensing technique is used to minimize the CPU time. As a future work, we can try othe neighborhood structures that allows a journey of the solution space.

References

1. Saporta, G.. *Probabilités, analyse des données et statistique*. Editions Technip; 2006. ISBN 9782710808145. URL: `https://books.google.fr/books?id=rprNjztQYPAC`.
2. Rencher, A.C.. *Cluster Analysis*. John, Wiley and Sons, Inc.; 2003, p. 451–503. URL: `http://dx.doi.org/10.1002/0471271357.ch14`.
3. Chen, Y., Keogh, E., Hu, B., Begum, N., Bagnall, A., Mueen, A., et al. The ucr time series classification archive. 2015. `www.cs.ucr.edu/ eamonn/time_series_data/`.
4. Ratanamahatana, C.A., Keogh, E.. Making time-series classification more accurate using learned constraints. In: *In proc. of SDM Intl Conf*. 2004, p. 11–22.
5. Schultz, M., Joachims, T.. Learning a distance metric from relative comparisons. In: *In NIPS*. MIT Press; 2004, p. 1–44.
6. Weinberger, K., Saul, L.. Distance metric learning for large margin nearest neighbor classification. *The Journal of Machine Learning Research* 2009;**10**:207–244.
7. Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S.. Information-theoretic metric learning. In: *Proceedings of the 24th International Conference on Machine Learning*; ICML '07. New York, NY, USA; 2007, p. 209–216.
8. Ying, Y., Huang, K., Campbell, C.. Sparse metric learning via smooth optimization. In: Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., Culotta, A., editors. *Advances in Neural Information Processing Systems 22*. Curran Associates, Inc.; 2009, p. 2214–2222.
9. Shalev-shwartz, S., Singer, Y., Ng, A.Y.. Online and batch learning of pseudo-metrics. In: *In ICML*. ACM Press; 2004, p. 743–750.
10. Jain, P., Kulis, B., Dhillon, I.S., Grauman, K.. Online metric learning and fast similarity search. In: Koller, D., Schuurmans, D., Bengio, Y., Bottou, L., editors. *NIPS*. Curran Associates, Inc.; 2008, p. 761–768.
11. Qamar, A.M.. *Generalized Cosine and Similarity Metrics: A Supervised Learning Approach based on Nearest Neighbors*. Theses; Université de Grenoble; 2010.
12. Frome, A., Singer, Y., Sha, F., Malik, J.. Learning globally-consistent local distance functions for shape-based image retrieval and classification. In: *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. 2007, p. 1–8.
13. Chechik, G., Shalit, U., Sharma, V., Bengio, S.. An online algorithm for large scale image similarity learning. In: Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., Culotta, A., editors. *Advances in Neural Information Processing Systems 22*. Curran Associates, Inc.; 2009, p. 306–314.
14. Sankoff, D., Kruskal, J.B., editors. *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Reading, MA: Addison-Wesley; 1983.
15. Sakoe, H., Chiba, S.. Dynamic programming algorithm optimization for spoken word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on* 1978;**26**(1):43 – 49.
16. Yu, D., Yu, X., Hu, Q., Liu, J., Wu, A.. Dynamic time warping constraint learning for large margin nearest neighbor classification. *Information Sciences* 2011;**181**(13):2787 – 2796.
17. Glover, F.. Tabu Search - Part II. *ORSA Journal on Computing* 1990;**2**:4–32.
18. Glover, F.. Tabu search - part i. *ORSA Journal on Computing* 1989;**1**:190–206.