# A finite time stochastic clustering algorithm

Andreea B. Dragut[1] and Codrin M. Nichitiu[2]

[1] LIF, Univ. Aix-Marseille II, Aix-en-Provence, France
[2] Long Island High Tech. Incubator, Stony Brook Univ., NY, USA

**Abstract.** We present a finite time local search $(1+\delta)$-approximation method finding the optimal solution with probability almost one with respect to a general measure of within group-dissimilarity. The algorithm is based on a finite-time Markov model of the simulated annealing. A dynamic cooling schedule, allows the control of the convergence. The algorithm uses as measure of within group dissimilarity a new generalized Ward index based on a set of well-scattered representative points, which deals with the major weaknesses of partitioning algorithms regarding the hyperspherical shaped clusters and the noise. We compare it with other clustering algorithms, such as CLIQUE and DBSCAN.

**Keywords:** Clustering, Finite-time Simulated Annealing, Approximation Scheme, Generalized Ward Index.

## 1 Introduction

It is generally acknowledged that there are two main families of clustering (unsupervised classification) methods: hierarchical and partitioning. The former ones create a tree structure splitting (reuniting) the initial set of objects in smaller and smaller subsets, all the way to singletons (and reverse), while the latter ones construct a partition of the initial set of objects into a certain number of classes, with the target number usually part of the input, along with the objects themselves. Most partitioning methods proposed for data mining [Jain *et al.*, 1999], [Gosh, 2003] can be divided into: discriminative (or similarity-based) approaches and generative (or model-based) approaches. In similarity-based approaches, one optimizes an objective function involving the pairwise data similarities, aiming to maximize the average similarities within clusters and minimize the average similarities between clusters. A fundamentally different approach is the model based approach which attempts to optimize the fit (global likelihood optimization) between the data and some mathematical model, and most researchers do not consider them as clustering methods. Similarity-based partitioning clustering is also closely related to a number of operations research problems such as facility location problems, which minimize some empirical loss function (performance measure). There are no efficient exact solutions known to any of these problems for general number of clusters $m$, and some formulations are NP-hard. Given the difficulty of exact solving, it is natural to consider approximation, either through polynomial-time approximation algorithms, which provide guarantees on the quality of their results, or heuristics, which make no guarantees. One of the most popular heuristics for the similarity-based partitioning problem is Lloyd's algorithm, often called the $m$-means algorithm. Define the neighborhood of a center point to be the set of data points for which this center is the closest. Thus, one can easily see that any locally minimal solution must be centroidal (i.e. each center lies at the centroid of its neighborhood). Unfortunately, $m$-means algorithm may converges to a local minimum that is arbitrarily bad compared to the optimal solution. Other heuristics with no proven approximation bounds are based on branch-and-bound searching, gradient descent, simulated annealing, nested partitioning, ant colony optimization, and genetic algorithms.

It is desirable to have some bounds on the quality of a heuristic. Given a constant $\delta \geq 0$, a $(1 + \delta)$-approximation algorithm (for a minimization problem) produces a solution that is at most a factor $(1 + \delta)$ larger than the optimal solution. With a tradeoff between approximation factors and running times, some clustering algorithms are able to produce solutions that are arbitrarily close to optimal. This includes $(1 + \delta)$-approximation algorithms for the Euclidean $m$-median problem by [Kolliopoulos and Rao, 1999] with a running time of $O(2^{1/\delta^s} n \log n \log m)$, assuming that the dimension $s$ is fixed. Another one is the $(1 + \delta)$-approximation algorithm for the Euclidean $m$-center problem given by [Agarwal and Procopiuc, 1998], which runs in $O(n \log m) + (m/\delta)^{O(m^{1-1/s})}$.

Another common approach in approximation algorithms is to develop much more practical, efficient algorithms having weaker, but still constant, approximation factors. These algorithms are based on local search, that is, by incrementally improving a feasible solution by swapping a small number of points in and out of the solution set. This includes the work of [Mettu and Plaxton, 2002] on the use of successive swapping for the metric $m$-means problem.

Unfortunately it is well known that $m$-means/medians/centers partitioning clustering algorithms have a tendency to partition the data into hyperspherical shaped clusters and do not adequately deal with outliers and noise.

The algorithm presented here is a local search $(1 + \delta)$-approximation method finding the optimal solution with probability almost one with respect to any general measure of within group-dissimilarity. It is actually a cooling schedule, obtained by stopping a simulated annealing algorithm in finite time, and it belongs to a family of approximation clustering algorithms of type $m$-median and $m$-means,

The algorithm addresses the weaknesses of partitioning algorithms in the way in which it constructs what we shall define as "critical" clusters, that are to be further expanded by the cooling schedule. As a measure of within group dissimilarity we introduce a new generalized Ward index based not on a single cluster representative i.e. centroid or median, but on a set of well-scattered representative points, which are shrunk toward the centroid. The idea of joining together points close to a set of representatives was introduced by [Guha *et al.*, 1998] to obtain a measure of inter-group dissimilarity in hierarchical clustering. Moreover, due to the particular choices of generation probabilities for the system of neighborhoods, the more dense a cluster is, the smaller the probability to have its elements reassigned to other clusters will be while trying to transform the current classification.

The rest of the paper is organized as follows. In Sections 2. and 3. we present the clustering problem as a combinatorial optimization problem and the general asymptotic convergence conditions for it Sections 4 and 5 describe and compare our algorithm with other clustering algorithms. Finally in Section 6. we present the conclusions and give directions for future research.

## 2 Setting

The general form of clustering problems considered is "given a set $X = \{1, 2, .., n\}$ of $n$ entities, to classify these entities means to partition the linear subspace $X$ into a number $m \leq n$ of clusters such that the $m-$partitioning is optimal according to a certain chosen criterion function defined on the set $\Pi_m$ of all $m-$partitions of the set $X$". Each element $i$ from the set $X$ has an input information vector $Y(i)$. There exists also a distance $d$ as a dissimilarity measure for every pairwise combination of entities to be clustered, and a function $\tau : P(X) \rightarrow R_+$ as a measure of within-group dissimilarities with the property that $\tau(A) = 0 \longleftrightarrow |A| = 1$. Let us consider the function $f : \Pi_m \rightarrow R$, $f(\pi_m) = \sum\limits_{i=1}^{m} \tau(A_i)$, where $\pi_m = (A_1, A_2, ..., A_m) \in \Pi_m$.

The class of clustering problems considered is (PC) $\min\limits_{\pi_m \in \Pi_m.} f(\pi_m)$. (PC) is a combinatorial optimization problem (see [Aarts *et al.*, 1997]) with a very large state space since the $|P(X)|$ given by the Bell number grows extremely rapidly with n; e.g., $B_{40} = 1.6 \times 10^{35}$ and $B_{100} = 4.8 \times 10^{115}$.

A first contribution of this work is the development of a stochastic search algorithm for finding $(1 + \delta)$-optimal partitions with a probability close to one. The basic idea is to construct a Metropolis-Hastings Markov chain via the simulated annealing algorithm.

A neighborhood function is a mapping $N : \Pi_m \rightarrow 2^{\Pi_m}$, which, for each classification $i \in \Pi_m$, defines a set $N(i) \subseteq \Pi_m$ of classifications that can be reached from $i$ by a single perturbation. At the beginning, an initial classification is given. The simulated annealing algorithm starts with it, and continuously tries to transform the current classification into one of its neighbors by applying the generation mechanism and an acceptance criterion. Better-cost neighbors are always accepted. To avoid being trapped in a local minimum, worst-cost neighbors are also accepted, but with a probability that is gradually decreased in the course of the algorithm execution. The lowering of the acceptance probability is controlled by a set of parameters whose values are determined by a cooling schedule.

As we mentioned in the introduction, the algorithm solves the (PC) problem for a general measure of within-group dissimilarities $\tau : P(X) \to R_+$ such that $\tau(A) = 0 \longleftrightarrow |A| = 1$. However to the best of our bibliographical knowledge, the already existent measures of within group dissimilarity constructed by the extension of a distance do not deal with arbitrarily shaped clusters, and are very sensitive to outliers. Among those indices, the most known are: Wilks index: $\tau(A) = \frac{1}{2|A|} \sum\limits_{x,y \in A} d^2(x,y)$, and Ward index $\tau(A) = \sum\limits_{x,y \in A} d^2(x, x_A)$, where $x_A$ is the centroid of $A$. The first index does not require $X$ to be a linear space and treats any point of the cluster as a cluster representative, which gives too much unfiltered information about the shape of the set to the clustering algorithm. Also, the (PC) optimization problem with this index leads to long shaped clusters. The second index treats the centroid as the unique cluster representative. This choice gives no information about the shape of the cluster and leads to the well known squared sum of errors criterion with his already discussed problems.

The new index we propose generalizes the Ward one considering multiple representatives for a cluster. We define the representatives index to be $\tau(A) = \sum\limits_{x \in A} \min\limits_{x_r \in R} d^2(x, x_r)$, where $R$ is the set of representatives. The idea of multiple representatives was introduced in hierarchical clustering by [Guha *et al.*, 1998]. They must be well spread across the whole cluster, and are thus obtained through an iterative selection: initially the farthest point from the centroid is picked, and then, up to $|R|$ (fixed in advance), the farthest point from the ones already picked is added. The distance from a candidate point to the set of already picked is the min of the pointwise distances from that point to each already picked. These representatives capture the geometry of the cluster, and upon a shrinking towards the centroid by a fixed factor, done after building $R$, the outliers get much closer to the centroid (moving more than average representatives within the bulk of the cluster).

## 3 The asymptotic convergence for the (PC) problem

*Notation 1.* $S$ : the set of solutions for the considered combinatorial optimization problem (here $S = \Pi_m$), and $S^*$ : the set of optimal solutions.

The simulated annealing can be mathematically modeled as a sequence of Markov chains. Each Markov chain has transition probabilities defined as

$$\forall\, i,j \in S : \quad P_{ij}(k) = \begin{cases} G_{ij}(c_k)\, A_{ij}(c_k) & \text{if } i \neq j \\ 1 - \sum\limits_{l \in S,\, l \neq i} G_{il}(c_k)\, A_{il}(c_k) & \text{if } i = j \end{cases} \tag{1}$$

where $G_{ij}(c_k)$ denotes the probability of generating a solution $j$ from a solution $i$, and $A_{ij}(c_k)$ the probability of accepting a solution $j$ that is generated from a solution $i$.

The matrix $P$ of equation (1) is stochastic and $G_{ij}\left(c_k\right)$ and $A_{ij}\left(c_k\right)$ are conditional probabilities. In the original version of simulated annealing, the acceptance probability is defined by:

$$\forall\, i,j \in S: \quad A_{ij}\left(c_k\right) = \exp\Big(- \ \left(f(j) - f(i)\right)^{+} / c_k\Big) \quad (2)$$

**Theorem 1 ([Aarts *et al.*, 1988])** *Let $(S,\, f)$ be an instance of a combinatorial optimization problem, $N$ a neighborhood function, and $P\,(k)$ the transition matrix defined by (1), with $c_k = c,\ \forall\ k = 0,1,....$ If we have (G1) $\forall c > 0,\ \forall i,j \in S,\ \exists p \geq 1,\ \exists l_0, l_1, ..., l_p \in S$ with $l_0 = i,\ l_p = j$ and $G_{l_k\ l_{k+1}}\left(c\right) > 0,\ k = 0,1,...,p-1$; (G2) $\forall c > 0,\ \forall i,j \in S:\ G_{ij}\left(c\right) = G_{ji}\left(c\right);$ (A1) $\forall c > 0,\ \forall i,j \in S:\ A_{ij}\left(c\right) = 1$ if $f(i) \geq f(j)$, and $A_{ij}\left(c\right) \in (0,1)$ if $f(i) < f(j)$; (A2) $\forall c > 0,\ \forall i,j,k \in S:\ A_{ij}\left(c\right) A_{jk}\left(c\right) A_{ki}\left(c\right) = A_{ik}\left(c\right) A_{kj}\left(c\right) A_{ji}\left(c\right);$ (A3) $\forall i,j \in S$ with $f(i) < f(j)$ $\lim_{c\to 0} A_{ij}\left(c\right) = 0.$ then the Markov chain has a unique stationary distribution $q(c)$, with*

$$q_i(c) = 1/\sum_{j\in S} \left(A_{ij}\left(c\right)/A_{ji}\left(c\right)\right)\ ,\ \forall i \in S, \quad (3)$$

**Remark 1** *For the following choice of the generation probabilities*

$$G_{ij} = \chi_{(N(i))}\left(j\right)/\left|N\left(i\right)\right|, \quad \forall i,j \in S, \quad (4)$$

*condition (G2) is no longer needed to guarantee asymptotic convergence, and the components of the stationary distribution are given by*

$$q_i(c) = \left|N\left(i\right)\right|/\sum_{j\in S} \left[\left(\left|N\left(j\right)\right| A_{ij}\left(c\right)\right)/A_{ji}\left(c\right)\right]\ \textit{for all}\ \forall i \in S, \quad (5)$$

We will consider this choice for the generation probability in order to solve the (PC) problem.

**Definition 1** *A cluster $A$ from $\omega \in \mathit{\Pi}_m$ is called critical for $\omega$ if*

$$\tau\left(A\right) = \max_{A_i\ \textit{cluster of}\ \omega} \tau\left(A_i\right).$$

*Notation 2.* $N'\left(\pi\right) = \{\omega = \left(A'_1, ..., A'_m\right) \in \mathit{\Pi}_m |\ A'_i$ are obtained from $A_i$ by a reassignment of up to $k$ elements from a critical cluster $A$, where $k = |A|\}$, for $\pi \in \mathit{\Pi}_m$. We say that $\left(N'\left(\pi\right)\right)_{\pi\in \mathit{\Pi}_m}$ is the set of critical neighborhoods.

**Proposition 1** *For the (PC) problem, the set of neighborhoods defined by $N'\left(\pi\right)$ satisfies the (G1) condition.*

*Proof.* It is a fact that $\forall i,\ j \in S,\ \exists\ p \geq 1$, and $l_0, l_1, ..., l_p \in S$ with $l_0 = i$, $l_p = j$ such that for any $k$ ,$l_k$ and $l_{k+1}$ are neighbors through a $n$-reassign system of neighborhoods. We shall prove that there also exists a path from $i \in S$ to $j \in S$ through a critical system of neighborhoods. Suppose that $u = 0, ..., p-1$ is the first step at which $l_u \in S$ and $l_{u+1} \notin N'\left(l_u\right)$. Let $A_u$ be a cluster in $l_u$ which has a maximal value for the within-group dissimilarity function $\tau$. Let $B_u$ be the cluster in $l_u$ from which $t$ elements are reassigned to some other clusters for obtaining $l_{u+1}$. Since $l_{u+1} \notin N'(l_u)$ then $\tau(A_u) > \tau(B_u)$. To get a path from $i \in S$ to $j \in S$ through a critical system of neighborhoods we will add a finite number of elements $l_u^{\backslash} \in N'(l_u)$ to the initial path. The procedure is the following: (1) We assign $k-1$ elements from $A_u$ to $B_u$, where $k = |A_u|$. The new classification $l_u^{\backslash}$ has only two modified clusters $A_u^{\backslash}$, $B_u^{\backslash}$, and $\tau\left(A_u^{\backslash}\right) = 0$ since $\left|A_u^{\backslash}\right| = 1$. (2) If $\tau\left(B_u^{\backslash}\right)$ has not the maximal value then $\exists\, A_{1u}$ such that $\tau(A_{1u})$ is maximal, and we will proceed as in the case of $A_u$ starting the construction of some $l_u^{\backslash\backslash} \in N'\left(l_u^{\backslash}\right)$. Since $1...n$ is a finite set after repeating for a finite number of times the procedure $\tau\left(B_u^{\backslash}\right)$ will be maximal. Now we construct a new classification $l_{u+1}^{\backslash} \in N'\left(l_u^{\backslash}\right)$ in the following way: from $B_u^{\backslash}$ the $t$ elements to other clusters as in the construction step from $l_u$ to $l_{u+1}$, and the elements belonging to the clusters $A_u$, $A_{1u}, ...$ are reassigned back to their clusters. We proceed in a similar way for all the steps $q$ at which $l_q \in S$ and $l_{q+1} \notin N'(l_q)$ preserving the other steps.

## 4 Finite-time model of simulated annealing

In practical applications, asymptoticity is never attained and thus convergence to an optimal solution is no longer guaranteed. Then we shall use the simulated annealing as an approximation algorithm, implementing a cooling schedule. The general idea of a cooling schedule is the following: start with an initial value $c_0$ for the control parameter and repeatedly generate a finite Markov chain for a finite number of decreasing values of $c$ until $c \simeq 0$. The parameters determining the cooling schedule are: the start value $c_0$ of the control parameter; the decreasing rule of the control parameter; the length $L_k$ of the individual Markov chains; the stop criterion of the algorithm. We will discuss the choice of those parameters for our problem such that the convergence towards near-optimal solutions will be ensured. Our cooling schedule follows the general ideas of the statistical cooling algorithm developed in [Aarts *et al.*, 1988] and designed for symmetric generation probabilities which lead to less complicate formulas for the stationary distributions.

### 4.1 The start value of the control parameter

This value should be large enough to ensure that initially all configurations occur with rather equal probabilities since $\lim_{c\to\infty} q_i(c) = |N_i'| / \sum\limits_{j\in S} \left|N_j'\right|$.

We distinguish two cases. In the first one, in which the set of system configurations corresponds to values of the cost function distributed over a number of distinct intervals whose mutual distances are large compared to their size, $c_0$ will be computed in the classical way as $\theta \cdot \max_{i,\,j\in S}\left[f\left(j\right)-f\left(i\right)\right]$, where $\theta \gg 1$. In the second case, the values for the cost function are sufficiently uniformly distributed. Thus, we can observe the behavior of the system before the actual optimization process takes place, and adjust the value of the control parameter such that the ratio $\chi$ of the system perturbations accepted over the total number of perturbations generated is kept close to the one given by $\lim_{c\to\infty} q_i\left(c\right)$. The initial value $c_0$ will be the final value of $c$ updated $m_1 + m_2$ times according to the relation:

$c = \underset{\Delta C_{i\,j}>0}{Average} \Delta f_{ij} / \ln\left[m_2 / \left(m_2\chi - \left(1-\chi\right)m_1\right)\right]$, where $\Delta f_{ij} = f\left(j\right) - f\left(i\right)$, and $m_2$, $m_1$ the numbers of rearrangements with $\Delta f_{ij} \leq 0, > 0$.

### 4.2 The decreasing rule of the control parameter

In the frame of the homogeneous Markov model for simulated annealing algorithm, the decreasing rule of the control parameter, as well as the lengths $L_k$ of the Markov chains are constructed in order to satisfy the following quasi-equilibrium condition: "$a\left(L_k, c_k\right)$ is close to $q\left(c_k\right)$", where $a\left(l, c_k\right)$ denotes the probability distribution of the classifications after $l$ transitions of the $k$-th Markov chain. The time behavior of the cooling schedule usually depends on the mathematical formulation of this condition. It is clear from an intuitive point of view that we will have larger differences between $q\left(c_k\right)$ and $q\left(c_{k+1}\right)$ if the decreasing rule of the control parameter allows large decrements of $c_k$, where we suppose we have reached the quasi-equilibrium. In this case it will be necessary to attempt more transitions at the new value $c_{k+1}$, for restoring the quasi-equilibrium. Thus, there is a trade-off between fast decrement of $c_k$ and small values for $L_k$. We will proceed as in [Aarts *et al.*, 1988] using small decrements in $c_k$ in order to avoid extremely long chains, and imposing for $\varepsilon, \delta$ small positive numbers: $\left\|q\left(c_k\right) - q\left(c_{k+1}\right)\right\| < \varepsilon \quad \approx \forall i \in S \quad 1/\left(1+\delta\right) < q_i\left(c_k\right)/q_i\left(c_{k+1}\right) < \left(1+\delta\right)$

**Remark 2** *For the components of the stationary distribution function from* (5) *we get* $q_i\left(c\right) = \left|N_i'\right| \cdot q_0\left(c\right) \cdot A_{i_0 i}\left(c\right)$, *where* $q_0\left(c\right) = \left[\sum_{j\in S}\left|N_j'\right| \cdot A_{i_0 j}\left(c\right)\right]^{-1}$, *and* $i_0 \in S^*$.

*Proof.* Let $i_0 \in S^* \implies f\left(j\right), f\left(i\right) \geq f\left(i_0\right)$. For $f\left(j\right) > f\left(i\right)$ we have $A_{ji} = 1$, and $A_{ij}\left(c\right) = \exp\left(-\Delta f_{ij}/c\right) = \exp\left(-\Delta f_{i_0\,j}/c\right) \cdot \exp\left(-\Delta f_{i\,i_0}/c\right) = A_{i_0\,j}\left(c\right) \cdot \exp\left(-\Delta f_{i\,i_0}/c\right)$. For $f(j) < f(i)$ we have $A_{ij}\left(c\right) = 1$. From the (A2) property of Theorem 1 we have $A_{i_0\,j}(c) \cdot A_{ji}(c) = A_{i_0 i}\left(c\right) = \exp\left(-\Delta f_{i_0\,i}/c\right) \Rightarrow A_{ji}\left(c\right) = \exp\left(-\Delta f_{i_0\,i}/c\right)/A_{i_0\,j}\left(c\right)$. So we get $A_{ij}(c)/A_{ji}\left(c\right) = A_{i_0\,j}(c)/A_{i_0\,i}(c)$. Then we have that $q_i\left(c\right) \overset{def}{=} \left|N_i'\right| / \left[\sum_{j\in S}\left|N_j'\right| \cdot A_{ij}\left(c\right)/A_{ji}\left(c\right)\right] = \left|N_i'\right| \cdot A_{i_0\,i}\left(c\right) / \left[\sum_{j\in S}\left|N_j'\right| \cdot A_{i_0\,j}\left(c\right)\right]$.

**Proposition 2** *If* $\forall i \in S, \forall k \in \mathbf{N}^*$ $c_k < c_{k+1}$*, and* $A_{i_0\,i}(c_k)/A_{i_0\,i}(c_{k+1}) < 1+\delta$*, where* $i_0 \in S^*$ *then the following inequalities are satisfied:* $1/(1+\delta) < q_i(c_k)/q_i(c_{k+1}) < (1+\delta)$.

*Proof.* Obviously $\sum\limits_{j\in S} A_{i_0 j}(c_{k+1}) < \sum\limits_{j\in S} A_{i_0 j}(c_k) < (1+\delta)\sum\limits_{j\in S} A_{i_0 j}(c_{k+1})$. Then we derive that $q_0(c_{k+1})/(1+\delta) < q_0(c_k) < q_0(c_{k+1})$, relation from which using the form of $q_i(c)$'s given by the previous remark we can obtain the desired inequality. Thus, using the hypothesis the second part of the desired inequality follows directly. The first part of the desired inequality is a result of introducing in the first part of the $q_0(c)$ 's inequality, the $q_i(c)$ 's expression, and the obvious relation: $A_{i_0\,i}(c_k) > A_{i_0\,i}(c_{k+1})$.

**Remark 3** *The relation given in the hypothesis of the previous proposition can be reformulated as:* $\forall i \in S, \forall k \in \mathbf{N}^* c_{k+1} > c_k/\left[1 + c_k \cdot \ln(1+\delta)/\Delta f_{i_0\,i}\right]$ *which is in fact a decreasing rule of the control parameter.*

To simplify the decreasing rule, we shall make an assumption often made in the literature, and supported by computational evidence (see [Aarts *et al.*, 1988] and [White, 1984]). What we really do is to restrict the decreasing rule to a set $S_{c_k}$ of configurations that occur with a greater probability during the generation of the $k$-th Markov chain. We will record the cost values of the classifications $X(1), ..., X(L_k) \in S = \Pi_m$ that occur during the generation of the $k$-th Markov chain, and we will assume that they are normally distributed with mean $\mu_k = \mu(c_k) = \left[\sum\limits_{j=1}^{L_k} f(X(j))\right]/L_k$, and variance $\sigma_k^2 = \sigma^2(c_k) = \left[\sum\limits_{j=1}^{L_k} f^2(X(j))\right]/L_k - \mu_k^2$. Thus, $\Pr\{\Delta f_{i_0\,i} \le \mu_k - f^* + 3\sigma_k\} \backsimeq 0.99$, where $f^*$ is the optimal value of the problem. Finally, we define $S_{c_k} = \{i \in S | \Delta f_{i_0\,i} \le \mu_k - f^* + 3\sigma_k\}$. Then $\Pr\{i \in S_{c_k}\} \backsimeq 0.99$, and we can replace the previous decreasing rule with a simpler one: $\forall i \in S_{c_k}, \forall k \in \mathbf{N}^*$ $c_{k+1} > c_k/\left[1 + c_k \cdot \ln(1+\delta)/\mu_k - f^* + 3\sigma_k\right]$. For us $f^*$ is not known but $\mu_k - f^* \ge 0$. Thus, the final decreasing rule of the control parameter is:

$$\forall i \in S_{c_k}, \forall k \in \mathbf{N}^* c_{k+1} > c_k/\left[1 + c_k \cdot \ln(1+\delta)/3\sigma_k\right] \quad (6).$$

### 4.3 The length $L_k$of the individual Markov chains

The length of a Markov chain is usually determined such that at each value $c_k$ a minimum number of transitions is accepted. Since transitions are accepted with decreasing probability, one would obtain $L_k \to \infty$ for $c_k \downarrow 0$. Therefore, $L_k$ is usually bounded by some constant $L_{\max}$ to avoid extremely long chains for small values of $c_k$. We take $L_{\max} = |X| - m \ge \max_{i\in S}|N'(i)|$.

### 4.4 The final value of the control parameter

This choice determines in fact the stopping criterion. We will follow the general idea of most of the dynamic cooling schedules (see [Aarts *et al.*, 1997]). Thus, the algorithm will stop at the $c_k$ value for which the cost function of the classification obtained in the last trial of a Markov chain remains unchanged for a number of $\rho$ consecutive chains. Schematically we have:

```
Compute($L_{\max}, c_0$) ; $c := c_0$; $\overline{f}[k] = MaxInt \; \forall k \in 0, ..., \rho$
repeat
for $i := 1$ to $L_{\max}$ do
   Generate($j \in N'(i)$)
   if $\Delta f_{ij} \leq 0$ then Accept($j$) =true
   else if $\exp(-\Delta f_{ij}/c)$ >randomize$[0, 1)$ then Accept($j$) =true;
   if Accept($j$) =true then $i := j$;
Compute($\sigma^2(c)$); Update($\overline{f}[0, ..., \rho]$) ; $c := \lceil c/[1 + c \cdot \ln(1+\delta)/3\sigma(c)] \rceil$ ;
until $\overline{f}[k_1] = \overline{f}[k_2] \; \forall k_1, k_2 \in 0, ..., \rho$
```

## 5 Comparison with other algorithms

The study is done comparing the speed and also the quality of the output classification, and using synthetic data generated in a setting constructed and acknowledged by several researchers, such as [Agrawal *et al.*, 1998] and [Zait and Messatfa, 1997]. In generating the data several parameters have been varied, such as size of the classes, their mutual distances, overlap factor, and also their local dimension, smaller than the one of the whole space where points where selected.

Our algorithm was compared to CLIQUE [Agrawal *et al.*, 1998] and DBSCAN, the latter being much less performant. For the algorithm presented here, we have noted a behavior of similar quality to the one of CLIQUE. However, CLIQUE reports overlapping classes in many cases (it has an approach based on density, varying the local dimensions in which it performs the search), and lower density zones in clusters are discarded as being outliers. Finally, CLIQUE requests the user to find appropriate values for some mandatory parameters controlling its behavior, which is a very difficult task in general. Finally, while both CLIQUE and our algorithm can end up making quite a number of passes over the data, the time required by our algorithm also depends on how fast the within-group dissimilarity $\tau$ can be computed, linear ones leading to faster algorithms. The building of the representative set $R$ takes $O(n|R|^2)$: $|R|$ steps, when each point of the current cluster is considered, and for each one, the minimum of the pointwise distance to each member of the increasing $R$, so another factor of $|R|$.

## 6 Conclusion

We have presented a finite time stochastic approximation clustering algorithm, which finds optimal solutions with probability almost one, and performs as well as good heuristic clustering algorithms, with a mathematical assessment of its properties, within the framework of the Markov chain analysis of simulated annealing. We have also introduced a new measure of within cluster dissimilarity improving the recognition of arbitrary shaped clusters and reducing the outliers effects.

Concerning outliers, CURE random sampling can filter out a majority of them. Chernoff bounds [Motwani and Raghavan, 1995] provide equations to analytically derive the random sample size required to have a low probability of missing clusters. Also for large databases making several passes over the whole database is undesirable, and clustering the random sample dramatically improves time complexity. Afterwards, the initial non-selected points are each assigned to the cluster of the closest among a fraction of randomly selected representatives for each cluster.

## References

[Aarts *et al.*, 1988]E. H. L. Aarts, J. H. M. Korst, and P. J. M. van Laarhoven. A quantitative analysis of the simulated annealing algorithm: A case study for the travelling salesman problem. *Journal of Stat. Phys.*, 50:187–206, 1988.

[Aarts *et al.*, 1997]E. H. L. Aarts, J. H. M. Korst, and P. J. M. van Laarhoven. Simulated annealing, Local search. In E. H. L. Aarts and J. K. Lenstra, editors, *Combinatorial Optimization.* John Wiley Interscience Series, New York, 1997.

[Agarwal and Procopiuc, 1998]P. K. Agarwal and C. M. Procopiuc. Exact and approximation algorithms for clustering. In *Procs. of the 9th AnnlACM-SIAM SODA*, pages 658–667, 1998.

[Agrawal *et al.*, 1998]R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *Procs. of the 1998 ACM-SIGMOD Int'l Conf. on Management of Data*, pages 94–105, 1998.

[Gosh, 2003]J. Gosh. *Handbook of Data Mining*, chapter Scalable Clustering Methods for Data Mining. Lawrence Erlbaum Assoc, 2003.

[Guha *et al.*, 1998]S. Guha, R. Rastogi, and K. Shim. CURE: an efficient clustering algorithm for large databases. In *ACM SIGMOD Int'l Conf. on Manag. of Data*, pages 73–84, 1998.

[Jain *et al.*, 1999]A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys (CSUR)*, 31(3):264–323, 1999.

[Kolliopoulos and Rao, 1999]S. Kolliopoulos and S. Rao. A nearly linear-time approximation scheme for the euclidean k-median problem. In J. Nesetril, editor, *Procs. of the 7th Annl. Euro. Symp. on Algs.*, volume 1643, pages 362–371. Springer Verlag, 1999.

[Mettu and Plaxton, 2002]R. R. Mettu and C. G. Plaxton. Optimal time bounds for approximate clustering. In *Procs. of the 8th Conf. on Uncertainty in Artif. Intell.*, pages 339–348, 2002.

[Motwani and Raghavan, 1995]R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.

[White, 1984]S. R. White. Concepts of scale in simulated annealing. In *Proceedings IEEE of the Int'l Conf. on Computer Design*, pages 646–651, 1984.

[Zait and Messatfa, 1997]M. Zait and H. Messatfa. A comparative study of clustering methods. *Future Generation Computer Systems*, 13:149–159, 1997.