Deterministic Annealing and Robust Scalable Data Mining for the Data Deluge

Geoffrey Fox
School of Informatics and Computing
Indiana University
Bloomington IN 47408, USA
gcf@.indiana.edu

ABSTRACT
We describe data analytics on large systems using a suite of robust parallel algorithms running on both clouds and HPC systems. We apply this to cases where the data is defined in a vector space and when only pairwise distances between points are defined. We introduce new O(N logN) algorithms for pairwise cases, where direct algorithms are $O(N^2)$ for N points. We show the value of visualization using dimension reduction for steering complex analytics and illustrate the value of deterministic annealing for relatively fast robust algorithms. We apply methods to metagenomics applications.

Categories and Subject Descriptors
G.4 MATHEMATICAL SOFTWARE: Algorithm design and analysis; Parallel and vector implementations; Reliability and robustness; Efficiency

General Terms
Algorithms, Experimentation, Performance, Reliability

Keywords
Deterministic Annealing, Clustering, Multidimensional Scaling, Clouds, Parallel Computing

1. DETERMINISTIC ANNEALING
Deterministic annealing[1] is motivated by the same key concept as the more familiar simulated annealing, which is well understood from physics. We are considering optimization problems and want to follow nature's approach that finds true minima of energy functions rather than local minima with dislocations of some sort. At high temperatures systems equilibrate easily as there is no roughness in the energy (objective) function. If one lowers the temperature on an equilibrated system, then it is a short safe path between minima at current temperature and that a higher temperature. Thus systems which are equilibrated iteratively at gradually lowered temperature, tend to avoid local minima. We cannot put our delicate mathematical equations in a real forge and the Monte Carlo approach of simulated annealing is often too slow, so we perform integrals analytically using a variety of approximations within the well-known mean field approximation. In this sense deterministic annealing is related to variational Bayes or variational inference methods while Markov chain Monte Carlo (MCMC) methods are roughly single temperature simulated annealing. In the basic case we have a Hamiltonian $H(\chi)$ which is to be minimized. In annealing, we introduce the Gibbs Distribution at Temperature T

$$P(\chi) = \exp(- H(\chi)/T) / \int d\chi \exp(- H(\chi)/T) \quad (1)$$
$$\text{or } P(\chi) = \exp(- H(\chi)/T + F/T) \quad (2)$$

PDAC'11, November 14, 2011, Seattle, Washington, USA.
ACM 978-1-4503-1130-4/11/11.

and minimize the Free Energy F combining Objective Function and Entropy

$$F = < H - T\, S(P) > = \int d\chi\, [P(\chi)H + T\, P(\chi)\, \ln P(\chi)] \quad (3)$$

as a function of χ, which are (a subset of) parameters to be minimized. The temperature is lowered slowly – say by a factor 0.95 to 0.99 at each iteration. For some cases such as vector clustering and Mixture Models, one can do integrals analytically but usually that will be impossible. So we introduce a new Hamiltonian $H_0(\chi, \varepsilon)$ which by choice of ε can be made similar to real Hamiltonian $H(\chi)$ and which has tractable integrals. Then we use the approximate Gibbs distribution

$$P_0(\chi) = \exp(- H_0(\chi)/T + F_0/T) \quad (4)$$

And calculate

$$F(P_0) = < H - T\, S_0(P_0) >|_0 = < H - H_0 > |_0 + F_0(P_0) \quad (5)$$

where $<\ldots>|_0$ denotes averaging with Gibbs $P_0(\chi)$ i.e. $\int d\chi\, P_0(\chi)$. Note that the true Free Energy satisfies the Gibb's inequality

$$F(P) \leq F(P_0) \quad (6)$$

which is also related to Kullback-Leibler divergence. This leads to an expectation maximization method where in first expectation step E, one finds the values of ε minimizing $F(P_0)$ in equation (5) and follow with M step of Expectation Maximization setting

$$\chi = <\chi> |_0 = \int d\chi\, \chi\, P_0(\chi) \quad (7)$$

Note there are three types of variables in the general case. The set ε are used to approximate the real Hamiltonian $H(\chi)$ by $H_0(\chi, \varepsilon)$; the set χ are subject to annealing while one can follow the determination of χ by finding yet other parameters (the third set) optimized by traditional methods. Note one iterates over temperature decreasing it, but also at fixed temperature until the EM step converges. Now we turn to some examples.

2. CLUSTERING
Here the annealed variables $\underline{\chi}$ are $M_i(k)$, which is the probability that point *i* belongs to cluster *k*. In either central clustering (where points are in a metric space) or pairwise clustering

$$H_0 = \sum_{i=1}^{N} \sum_{k=1}^{K} M_i(k)\, \varepsilon_i(k) \quad (8)$$

which is linear in $M_i(k)$ allowing integrals in equations (5) and (7) to be done analytically. The metric space central clustering has $\varepsilon_i(k) = (\underline{X}(i) - \underline{Y}(k))^2$ where points have position $\underline{X}(i)$ and cluster centers are $\underline{Y}(k)$, so:

$$H_{Central} = H_0 = \sum_{i=1}^{N} \sum_{k=1}^{K} M_i(k)\, (\underline{X}(i) - \underline{Y}(k))^2 \quad (9)$$

and the expectation step gives:

$$<M_i(k)> \;=\; \exp(-\varepsilon_i(k)/T) / \sum_{k=1}^{K} \exp(-\varepsilon_i(k)/T) \quad (10)$$

and the cluster centers $\underline{Y}(k)$ are easily determined in M step.

In the non-metric space pairwise clustering case[2] the Hamiltonian is given by a form nonlinear in $M_i(k)$,

$$H_{PWC} = 0.5 \sum_{i=1}^{N} \sum_{j=1}^{N} \delta(i, j) \sum_{k=1}^{K} M_i(k)\, M_j(k) / C(k) \quad (11)$$

where $\delta(i, j)$ is pairwise distance between points i and j and $C(k) = \sum_{i=1}^{N} M_i(k)$ is the number of points in Cluster k. One takes the same form $H_0 = \sum_{i=1}^{N} \sum_{k=1}^{K} M_i(k)\, \varepsilon_i(k)$ as for central clustering so the linear (in $M_i(k)$) H_0 and quadratic H_{PWC} are different. The parameters $\varepsilon_i(k)$ are determined to minimize

$$F_{PWC}(P_0) = < H_{PWC} - T\, S_0(P_0) >|_0 \qquad (12)$$

where integrals can be easily done to find $\varepsilon_i(k)$ and one gets at M step that $<M_i(k)>$ is given by equation (10) again.

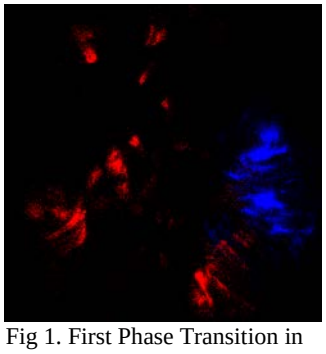

Fig 1. First Phase Transition in Metagenomics Clustering.
2 Clusters appear at T=0.0475

In many problems, decreasing temperature is a classic multiscale step with finer resolution being used as temperature T decreases. Note from equations (9) and (10) that we have factors like $(\underline{X}(i) - \underline{Y}(k))^2 / T$ and $\sqrt{T}$ acts as a distance scale. In clustering there is just one cluster at infinite temperature (the starting point) at the mean position over all points. One can start with just one cluster and decrease temperature. The system becomes unstable and additional clusters "pop" out in what is a phase transition in the physics interpretation of the system. The transition can be detected by either examining second derivative matrix or by placing multiple centers at each cluster and perturbing them. At phase transitions, the perturbed points will separate. This is illustrated in figures 1 and 2 showing a pairwise Metagenomics clustering visualized using dimension reduction with two early phase transitions revealing structure at finer scales. This sample has around 125 clusters in all so the structure is only starting to be revealed!

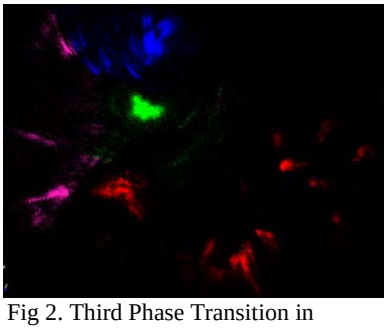

Fig 2. Third Phase Transition in Metagenomics Clustering.
4 Clusters appear at T=0.0388

3. DIMENSION REDUCTION

We have investigated deterministic annealing for two well-known dimension reduction algorithms, whose use is illustrated by figs 1 and 2 which map sequences in a high dimensional space (they are up to 600 base pairs in length in the figs) to 2 dimensions for visualization. There is GTM Generative Topographic Mapping applicable for points in a metric space and Multidimensional Scaling MDS, which only needs pairwise distances. DA-GTM is an example of the application of annealing to mixture models, which can also be applied to Gaussian mixture models and topic classification of documents. Here we just discuss DA-MDS[3] where the objection function called STRESS or H_{MDS}

$$H_{MDS} = \Sigma_{i<j=1}^{N} \text{weight}(i,j)\, (\delta(i, j) - d(\underline{X}(i), \underline{X}(j)))^2 \qquad (13)$$

Where $\delta(i, j)$ are observed dissimilarities, weight(i,j) is an arbitrary weight function and each point i is to be mapped to a point $\underline{X}(i)$ in a Euclidean space with metric $d(\underline{X}(i), \underline{X}(j))$. H_{MDS} is quartic or involves square roots, so we cannot use H_{MDS} in the DA Gibbs function but need the idea of an approximate Hamiltonian H_0. Previously we used linear Hamiltonians in equation (8) but here H_0 is a factorized quadratic form.

$$H_0 = \Sigma_{i=1}^{N} (\underline{X}(i) - \underline{\mu}(i))^2 / 2 \qquad (14)$$

where $\underline{\mu}(i)$ are the parameters ε of the general formula and the mapped vectors $\underline{X}(i)$ are the χ. In this case one will find the simple M step

$$\underline{X}(i) = \underline{\mu}(i) \qquad (15)$$

While the E step minimizes

$$\Sigma_{i<j=1}^{N} \text{weight}(i,j)\, (\delta(i, j) - \text{constant}.T - (\underline{\mu}(i) - \underline{\mu}(j))^2)^2 \qquad (16)$$

Here the solution has $\underline{\mu}(i) = 0$ at large temperature and the points emerge from the origin as temperature is lowered

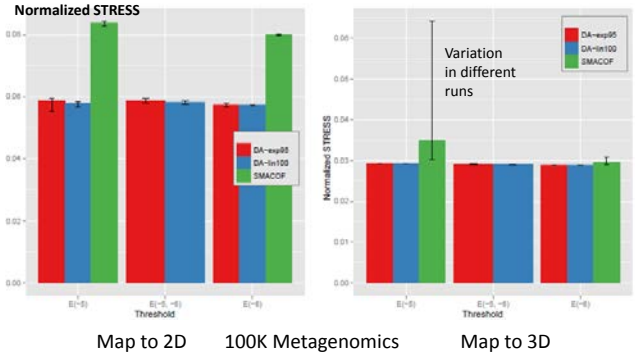

Fig 3. Comparison of MDS with (red, blue) and without (green, SMACOF) deterministic annealing

Fig 3 illustrates that deterministic annealing improves the quality of the fit both in terms lowered STRSS and lowered variation in value between different starting points. Dimension reduction and clustering can be combined[4] to give a robust approach that clusters and visualizes allowing user steering. For large systems the $O(N^2)$ computational complexity can be addressed[5] by using similar ideas to those used to reduce the $O(N^2)$ particle dynamics algorithms to O(N) or O(NlogN). One uses MDS on a sample of the data to map points to 3D and then an oct-tree decomposition provides a rigorous way to implement a hierarchical version of the algorithms. We have implemented this approach on clouds and HPC infrastructure.

4. ACKNOWLEDGMENTS

Our work is partially supported by Microsoft and by the NSF under the FutureGrid Grant No. 0910812.

5. REFERENCES

[1] Ken Rose, *Deterministic Annealing for Clustering, Compression, Classification, Regression, and Related Optimization Problems*. Proc. IEEE, 1998. 86: p. 2210--2239.
[2] Hofmann, T. and J.M. Buhmann, *Pairwise Data Clustering by Deterministic Annealing*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997. 19(1): p. 1-14.
[3] Hansjorg Klock and Joachim M. Buhmann, *Data visualization by multidimensional scaling: a deterministic annealing approach*. Pattern Recognition, April, 2000. 33(4): p. 651-669.
[4] G. Fox, S-H. Bae, J. Ekanayake, X. Qiu, and H. Yuan, *Parallel Data Mining from Multicore to Cloudy Grids*. chapter of High Speed and Large Scale Scientific Computing. 2009: IOS Press, Amsterdam.
[5] S-H. Bae, J. Y. Choi, J. Qiu, and G. Fox, *Dimension reduction and visualization of large high-dimensional data via interpolation*, in Procs of the 19th ACM International Symposium on High Performance Distributed Computing. 2010, ACM. Chicago, Illinois. pages. 203-214.