Comment on "Clustering by Passing Messages Between Data Points"

Michael J. Brusco[1]* and Hans-Friedrich Köhn[2]

Frey and Dueck (Reports, 16 February 2007, p. 972) described an algorithm termed "affinity propagation" (AP) as a promising alternative to traditional data clustering procedures. We demonstrate that a well-established heuristic for the *p*-median problem often obtains clustering solutions with lower error than AP and produces these solutions in comparable computation time.

Frey and Dueck (*1*) described an algorithm for analyzing complex data sets termed "affinity propagation" (AP). The algorithm extracts a subset of representative objects or "exemplars" from the complete object set by exchanging real-valued messages between data points. Clusters are formed by assigning each data point to its most similar exemplar. The authors reported that "[a]ffinity propagation found clusters with much lower error than other methods, and it did so in less than one-hundredth the amount of time" (*1*). We demonstrate that an efficient implementation of a 40-year-old heuristic for the well-known *p*-median model (PMM) often provides lower-error solutions than AP in comparable central processing unit (CPU) time.

For consistency with AP in (*1*), we present the PMM as a sum of similarities maximization problem, while recognizing that this is equivalent to the more common form of minimizing the sum of dissimilarities (e.g., distances or costs). The PMM is a general mathematical problem that can be concisely stated as follows: Given an $m \times n$ similarity matrix, S, select p columns from S such that the sum of the maximum values within each row of the selected columns is maximized (*2*). Thus, each row is effectively assigned to its most similar selected column (exemplar) with the goal of maximizing overall similarity. One classic example of the PMM occurs in facility location planning: Locate p plants such that the total distance (or cost) required to serve m demand points is minimized. In data analysis applications where S is an $n \times n$ matrix of negative squared Euclidean distances between objects, clustering the n objects using the PMM corresponds to the selection of p exemplars to minimize error, which is defined as the sum of the squared Euclidean distances of each object to its nearest exemplar.

Lagrangian relaxation methods enable the exact solution of PMM instances with $n \leq 500$ objects (*3*, *4*). For larger problems, a vertex sub-stitution heuristic (VSH) developed in (*5*) has been the standard for comparison for nearly four decades. The VSH begins with the random selection of a subset of p exemplars, which is iteratively refined by evaluating the effects of substituting an unselected point for one of the selected exemplars. Frey and Dueck assert that this type of strategy "works well only when the number of clusters is small and chances are good that at least one random initialization is close to a good solution" (*1*). To the contrary, the VSH is remarkably effective and is often the engine for metaheuristics such as tabu search (*6*) and variable neighborhood search (*7*).

We compared AP to an efficient implementation of VSH (*7*) across eight data sets from the clustering literature: I: Hartigan's (*8*) birth/death rates data; II: Fisher's (*9*) iris data; III: Lin and Kernighan's (*10*) circuit-board data; IV: Reinelt's (*11*) circuit-board data; V, VI, and VII: Grötschel and Holland's (*12*) European city coordinates; and VIII: Olivetti face images as used in (*1*). The measure of similarity for all data sets was negative squared Euclidean distance. The AP preference vectors for I through VII were established based on median similarity, as recommended in (*1*). For data set VIII, we used the preference vector from (*1*).

Affinity propagation was applied to each test problem (*13*), and the selected number of exemplars (p), the observed error, and computation time were stored. Next, 20 restarts of the VSH were applied assuming the value of p selected by AP. Each restart used a different randomly selected set of p exemplars as the initial solution. The minimum error observed across the 20 restarts, as well as the computation time, were stored for each test problem. Finally, we obtained optimal solutions for problems I through VII, as well as a reasonably tight lower bound for problem VIII, using Lagrangian relaxation methods.

The results in Table 1 are discordant with the claims in (*1*) that AP is an appreciably more efficient algorithm that produces solutions with less error than standard iterative refinement heuristics. Affinity propagation and the VSH yielded the same error for problem I; however, the VSH obtained a better solution than AP for the other seven problems. Moreover, the AP error value exceeded the optimum by 3% or more for four of the eight problems, whereas the VSH error never exceeded the optimum by more than 0.27%. In terms of efficiency, the maximum ratio of VSH to AP computation time was 2.79:1, which is drastically less than the 100:1 ratio reported in (*1*).

We also applied VSH to the document summary and airline travel routing data sets from (*1*). These are asymmetric similarity matrices with violations of the triangle inequality and were included to refute the implication in (*1*) that the PMM is inapplicable for such conditions. The VSH accommodates the asymmetry and captures differential preferences as a "fixed charge" in the objective function. For the document summary data, AP produces a four-cluster solution with a net similarity index of –10,234.33. The four-cluster VSH solution yielded a better index of –10,216.63. For the travel routing data, the similarity index of –92,154 for the seven-cluster VSH solution was better than the corresponding figure of –92,460 for AP. The VSH was slightly faster than AP for both test problems.

In summary, using a classic heuristic for the PMM, we frequently obtained better solutions than AP in comparable computation time. Moreover, like AP, the PMM has sufficient flexibility to accommodate nonmetric forms of S. With respect to another purported advantage of AP, the selection of the number of exemplars, AP merely replaces the problem of choosing p with the problem of setting the preference vector. Thus, the problem of choosing p is not resolved by AP. In light of these issues, we contend that AP,

[1]College of Business, 307 Rovetta Business Building, Florida State University, Tallahassee, FL 32306–1110, USA. [2]Department of Psychological Sciences, 19 McAlester Hall, University of Missouri-Columbia, Columbia, MO 65211, USA.

*To whom correspondence should be addressed. E-mail: mbrusco@cob.fsu.edu

Table 1. A computational comparison of AP and VSH.

Test problem (label)	n (points)	p (exemplars)	f* = Optimal error (squared distance)	% > f* AP (min %)	% > f* VSH (min %)	Total CPU time VSH:AP (ratio)
I	70	6	863.41	0.00	0.00	0.46:1
II	150	6	43.98	4.50	0.00	0.59:1
III	318	11	44,882,137	0.22	0.00	1.20:1
IV	724	22	21,396,140.7195	1.88	0.25	2.70:1
V	202	17	2,065.5737	4.10	0.00	1.31:1
VI	431	16	45,316.3623	5.27	0.10	1.22:1
VII	666	17	127,447.545	3.53	0.21	1.20:1
VIII	900	62	9,665.84*	0.71	0.27	2.79:1

*The optimal solution for problem VIII could not be verified. The reported value is a lower bound.

although a welcome addition to the clustering literature, does not possess the relative advantages over existing procedures that were touted in (*1*). We invite analysts to compare AP to VSH on other data sets and render their own opinions (*14*).

References and Notes

1. B. J. Frey, D. Dueck, *Science* **315**, 972 (2007).
2. N. Mladenović, J. Brimberg, P. Hansen, J. A. Moreno-Pérez, *Eur. J. Operational Res.* **179**, 927 (2007).
3. G. Cornuejols, M. L. Fisher, G. L. Nemhauser, *Manage. Sci.* **23**, 789 (1977).
4. P. Hanjoul, D. Peeters, *Eur. J. Operational Res.* **20**, 387 (1985).
5. M. B. Teitz, P. Bart, *Oper. Res.* **16**, 955 (1968).
6. E. Rolland, D. A. Schilling, J. R. Current, *Eur. J. Operational Res.* **96**, 329 (1997).
7. P. Hansen, N. Mladenović, *Location Sci.* **5**, 207 (1997).
8. J. A. Hartigan, *Clustering Algorithms* (Wiley, New York, 1975).
9. R. A. Fisher, *Ann. Eugen.* **7**, 179 (1936).
10. S. Lin, B. W. Kernighan, *Oper. Res.* **21**, 498 (1973).
11. G. Reinelt, www.iwr.uni-heidelberg.de/groups/comopt/software.TSPLIB95/index.html (2001).
12. M. Grötschel, O. Holland, *Math. Prog.* **51**, 141 (1991).
13. Analysis was performed using Matlab function apcluster.m, downloaded 12 December 2007 from www.psi.toronto.edu/affinitypropagation.
14. All algorithms and data sets used to obtain results for this Technical Comment are available at http://garnet.acns.fsu.edu/~mbrusco.