Jednodimenzionalno l_1 grupiranje podataka na bazi traženja optimalnih centara

Rudolf Scitovski, Kristian Sabo
Odjel za matematiku, Sveučilište u Osijeku

Sažetak. Motivirani uglavnom radovima (Iyigun and Ben-Israel, 2010; Kogan, 2007; Teboulle, 2007) u ovom radu razmatramo jednodimenzionalni problem l_1 grupiranja podataka na bazi traženja optimalnih centara klastera. Konstruirana je vrlo efikasna iterativna procedura, na osnovi koje je moguće odrediti optimalnu particiju. Analizirana su osnovna svojstva i konvergencija iterativnog procesa, koji konvergira prema stacionarnoj tocki kriterijske funkcije cilja za proizvoljni izbor početne aproksimacije. Metoda je ilustrirana na više numeričkih primjera. Specijalno, problem je vizualiziran na traženju optimalne particije s dva klastera. Pri tome detektirane su sve stacionarne točke odgovarajućeg minimizirajućeg funkcionala. Također, naveden je odgovarajući algoritam, koji za izabranu početnu aproksimaciju u samo nekoliko koraka daje stacionarnu točku i pridruženu particiju.

* * * * * * * * * * *

Abstract. (One-dimensional center-based l_1-clustering method) In this paper we consider a one-dimensional center-based l_1-clustering problem and construct a very efficient iterative process, on the basis of which it is possible to determine optimal partition. We analyze the basic properties and convergence of our iterative process, which converges to a stationary point of the corresponding objective function for each choice of the initial approximation. The method is illustrated by several numerical examples, and in particular we visualize the problem of looking for an optimal partition with two clusters, where we check all stationary points of the corresponding minimizing functional. Given is also a corresponding algorithm, which for the given initial approximation in only few steps gives a stationary point and corresponding partition.

Key words: clustering, data mining, optimization, weighted median problem

MSC2010: 62H30, 68T10, 90C26, 90C27, 91C20, 47N10

References

A. Ben-Israel, C. Iyigun, *Probabilistic D-clustering*, Journal of Classification **25**(2007) DOI: 10.1007/s00357-007-0021-y

D. L. Boley, *Principal direction divisive partitioning*, Data Mining and Knowledge Discovery 2(1998), 325–344

D. L. Boyd, L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, 2004.

F. H. Clarke, *Generalized gradients an applications*, Transactions of the America Mathematica Society, **205**(1975), 247–262

F. H. Clarke, *Optimization and Nonsmooth Analysis*, SIAM, Philadelphia, 1990.

R. Cupec, R. Grbić, K. Sabo, R. Scitovski, *Three points method for searching the best least absolute deviations plane*, Applied Mathematics and Computation, **215**(2009), 983–994

I. S. Dhillon, S. Mallela, R. Kumar, *A divisive information-theoretic feature clustering algorithm for text classification*, Journal of Machine Learning Research **3**(2003), 1265–1287.

I. S. Dhillon, Y. Guan, B. Kulis, *Kernel k-means, spectral clustering and normalized cuts*, Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), August 22–25, 2004, Seattle, Washington, USA, 551–556, 2004

G. Gan, C Ma, J. Wu, *Data Clustering: Theory, Algorithms, and Applications*, SIAM, Philadelphia, 2007.

J. Kogan, *Introduction to Clustering Large and High-Dimensional Data*, Cambridge University Press, 2007.

J. Kogan, C. Nicholas, M. Wiacek, *Hybrid Clustering of large high dimensional data*, In M. Castellanos and M. W. Berry (Eds.), Proceedings of the Workshop on Text Mining, SIAM, 2007.

J. Kogan, M. Teboulle, *Scaling clustering algorithms with Bregman distances*. In: M. W. Berry and M. Castellanos (Eds.), Proceedings of the Workshop on Text Mining at the Sixth SIAM International Conference on Data Mining, 2006.

J. Kogan, C. Nicholas, M. Wiacek, *Hybrid clustering with divergences*. In: M. W. Berry and M. Castellanos (Eds.), Survey of Text Mining: Clustering, Classification, and Retrieval, Second Edition, Springer, 2007.

C. Iyigun, A. Ben-Israel, *A generalized Weiszfeld method for the multi-facility location problem*, Operations Research Letters **38**(2010) 207–214

D. Littau, D. L. Boley, *Clustering very large data sets with PDDP*. In J. Kogan, C. Nicholas, M. Teboulle (eds), *Grouping Multidimensional Data: Recent Advances in Clustering*, 99–126, Springer-Verlag, New York, 2006.

J. Petrić, S. Zlobec, *Nelinearno programiranje*, Naučna knjiga, Beograd, 1989.

K. Sabo, R. Scitovski, *The best least absolute deviations line – properties and two efficient methods*, ANZIAM Journal **50**(2008), 185–198

H. Späth, *Cluster-Formation und Analyse*, R. Oldenburg Verlag, München, 1983.

Z. Su, J. Kogan, C. Nicholas, *Constrained clustering with k-means type algorithms*, In M.W. Berry, J. Kogan (eds), *Text Mining Applications and Theory*, 81–103, Willey, Chichester, 2010.

M. Teboulle, *A unified continuous optimization framework for center-based clustering methods*, Journal of Machine Learning Research **8**(2007), 65–102