Proposed decades ago, k-means is still the most popular algorithm for clustering. Despite the drawbacks of k-means, its advantages make it most attractive. Several researches have been conducted to alleviate the problems of k-means. We suggest here some simple modifications to optimize k-means for scalability without much sacrifice in the precision. Current shift in emphasis of data mining towards Big Data requires fast algorithms that can scale well. We propose an idea how time-tested techniques can be adapted to changing needs. The implementation results demonstrate the impact simple modifications can bring.

References

- X. Wu et al. Top 10 algorithms in data mining. Knowledge and Information Systems,
Optimizing k-means for Scalability


Index Terms

Computer Science

Applied Mathematics

Keywords

Data mining Big Data k-means