Constrained Intelligent K-Means: Improving Results with Limited Previous Knowledge

Renato Cordeiro de Amorim
Birkbeck, University of London
renato@dcs.bbk.ac.uk

Abstract

It is here presented a new method for clustering that uses very limited amount of labeled data, employees two pairwise rules, namely must link and cannot link and a singlewise one, cannot cluster. It is demonstrated that the incorporation of these rules in the intelligent k-means algorithm may increase the accuracy of results, this is proven with experiments where the real number of clusters in the data is unknown to the method.

1. Introduction

Nowadays it is common for companies to possess vast amounts of data and some can argue that thanks to new or improved technologies, acquire such an amount of data ceased to be the real problem, the real issue could be defined as extracting a meaning out of this data.

There are different algorithms that attempt to extract information from data, those can be normally classified as unsupervised, when no previous knowledge of the data is used in the clustering process and the opposite, the supervised approach. As it has been argued [3] these should not be seen as competitors but as complementary approaches to the task of data classification.

The fact that a detached view of them cannot be applied to all cases gave birth to semi supervised learning where only part of the data is labeled; this approach uses a limited and often very limited amount of knowledge of the data.

In this paper, firstly the classical K-means will be reviewed together with two modern variants of it, namely the intelligent K-means algorithm, introduced by [1] and the COP-KMeans, a constrained K-means introduced by [2]. After that the need for a compact and unique approach for clustering, a constrained K-means approach where K is unknown, will be assessed and the Constrained Intelligent K-Means algorithm will be presented together with experimental results.

2. K-means and iK-means

The K-means method has one of the most well know algorithms for unsupervised clustering, which can be defined as to partition a finite amount of data into a number of clusters by understanding underlying structure [6]; And although it has some stimulating advantages such as being computationally easy, fast and memory efficient [1] it is difficult to know how good the results are relative to the best possible clustering of the data [5] also, as these results depend on the initial seeds they tend to be unstable.

The algorithm can be defined as follows:

1. Define the number of clusters, and also define its initial seeds, which are normally chosen randomly.
2. Determine the contents of each cluster based in the distance between the seeds and the entities.
3. Recalculate the position of the seed based in the mean of the elements of its cluster, stop if there is no change in the contents of a cluster, or else go to 2.

Even so unsupervised approaches tend to be less powerful than supervised ones, as they are prone to fail if no distinct cluster structure is present in the data [3] the K-means algorithm has been extensively and successfully used in the most diverse problems, for instance in [7].

The Intelligent K-Means algorithm addresses possibly the major drawback of K-means, which is the definition of the number of clusters in the data.

The Algorithm establish the seeds using anomalous patterns, in other words, after the data is normalized

the non clustered entities that are the farthest away from the center become, one by one, tentative seeds. Their cluster determined through the distances between the entities and the only other seed, the center itself.
At the end, small clusters are removed according to a threshold and the final seeds are then used as initial seeds in the K-Means Algorithm, iK-Means has also been successfully applied to a number of different comparative experiments, such as in [12]

The importance of finding good seeds has been object of much research, algorithms such as the k-means++, introduced by [11] have already attempted this problem, but this for instance does not find the number of clusters, just better seeds for the given number of clusters.
X-means, introduced by [8] is an example of an algorithm that in fact tries to determine the right number of clusters, but in this algorithm a range where the true number of clusters is has to be provided.
The algorithm calculates the score of all possible number of clusters, provided in the range, and calculate their score using Bayesian Information Criterion (BIC) Akaike Information Criteria (AIC), and the one with the best score is output.
There are in fact other k-means variants that also attempt to find the right number of clusters by follow up analysis, but it has been experimentally proven in [12] that those do not provide results as consistent as iK-Means.
Evidently the fact that iK-Means provides better results than other algorithms does not mean that it always provides good results.

3. Constrained Intelligent K-Means

Progress in digital data acquisition and storage technology has resulted in the growth of huge databases [5] and extensive hand-labeling would be costly and time-consuming enough to make standard supervised learning algorithms unfeasible [4].

Clustering is usually implemented in an unsupervised fashion which means that in K-means apart from the number of clusters no other previous knowledge is used. However, in some cases information about the problem domain is available [2] surely the amount of labeled data may not be very vast which leaves the problem somewhere in between been supervised possible with the use of the limited previous knowledge and unsupervised achievable since this knowledge may not be enough.

Since semi supervised classification is said to be a fruitful approach when dealing with a small rate of labeled data [3] it is the type of classification that should be used in such problems.

The COP-KMeans is a semi supervised version of K-means [2] it uses pairwise rules which define what entities must be linked together in the same cluster and what entities cannot be linked.

Its algorithm can be defined as:

1. Initialize the seeds
2. Assign each entity to the closest seed making sure it follows the must-link, cannot-link rules
3. Update the values of the seeds based in their cluster entities mean
4. Iterate between (2) and (3) until convergence
5. Return the Clusters

Other constrained algorithms have also made use of the same pairwise rules [3][4] and it is argued [9] that their use has led to improved performance on several real-world applications.

But it has to be noted that in a partially labeled database, the labels should be able to indicate if an entity has not got the properties one is interested in cluster, for instance due to inaccuracy of the values in the entity, so here a new rule “Cannot Cluster” is introduced.

By consequence the algorithm here presented not only combines iK-means and Constrained K-means but also expand the possible types of rules in order to better deal with noise in the data.

The Constrained Intelligent K-Means uses normalized data and works as follows:

1. Seeds are chosen from entities, one by one based on the distance from the center, the most far away are chosen first.
2. Entities are clustered to the seeds if
   - They are closer to the seed than to the center or any of the entities in the Cannot Cluster Rule
   - They do not break a Cannot Link Rule

   Also the entities in the Must Link rule are included if that is the case
3. The value of the seeds are updated based in the mean of the entities in its cluster
4. Steps 2 and 3 are repeated until convergence
5. Small clusters are removed using a threshold
6 The Algorithm repeats itself with the full data but using the obtained seeds as initial seeds in instead of the ones obtained with anomalous pattern there is no

step 5

7 Return the Clusters

It has been argued that semi supervised approaches not only address the important issue of dealing with data that has a very small training set [4] but are also said [3] to be more robust than both unsupervised and supervised approaches as it combines two fundamentally different sources of information.

In practice the CiK-Means is a good alternative to be tested when the number of clusters is unknown and other cluster approaches such as iK-Means have failed to provide good results.

4. Evaluation Method

It has be argued [4] that the Rand index is natural evaluation criteria for clustering pairwise constraints, and even so the Constrained iK-Means is not fully pairwise it still makes sense to use it.

Rand Index = #correct free decisions / #total free decisions

The uniqueness of the Constrained iK-Means algorithm makes comparisons difficult to be done, for instance a comparison with the constrained K-Means would not be practical as in the later the true value of k was previously located and used in the experiments [2].

In order to analyze the CiK-means results, 2 datasets were used. To facilitate the understanding of the results both datasets have the same amount of entities: 150 divided into 3 clusters of 50 plus 3 entities made with abnormal values in order to represent noise. One of the datasets (D1) has been artificially generated and the second (D2) is the standard Anderson-Fisher Iris database plus 3 entities representing noise.

There are two important observations to be made, in D1 although the clusters are really well separated the center (0,0) is inside one of them, and in D2, 2 of the clusters are not well separated.

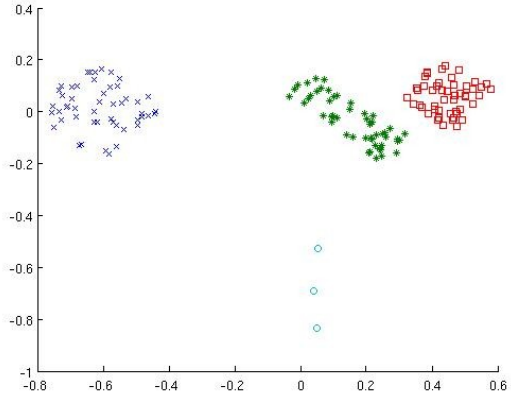

Fig1. D1: Artificially generated Dataset

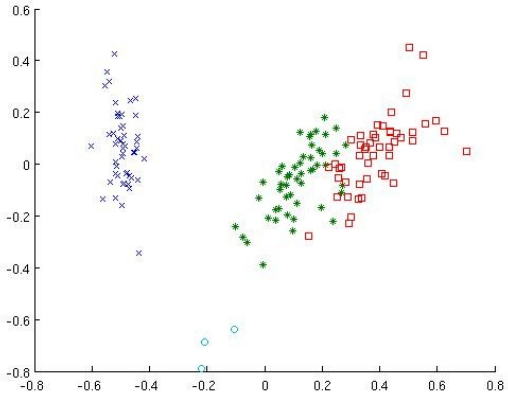

Fig2. D2: Avison-Fisher Iris standardized Database

5. Experimental Results

Starting with the Dataset D1, the average accuracy of K-means using 4 as the number of clusters, being those the ordinary 3 plus the noise, random seeds and 100 trials was found to be 0.8612.

In terms of iK-Means performance it is important to observe that the used threshold has a direct impact in the number of clusters found and by consequence in the accuracy of results. Using an optimized proportional threshold of 3% which provides 4 clusters the found index, which ceases to be variable, was 0.8497.

In the same environment the Constrained Intelligent K-Means, using an optimized threshold of 6% had simply perfect results which an average index of 1.0.

Different thresholds were used because for comparison purposes, iK-Means has to find 4 clusters, and the CiK-Means has to find only 3 as it automatically finds

the noise cluster.

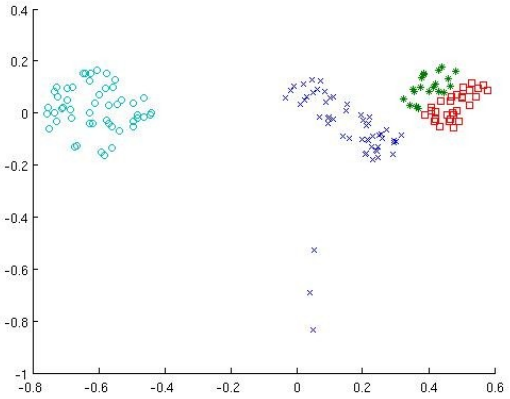

Fig3. D1: Example of K-Means Clustering

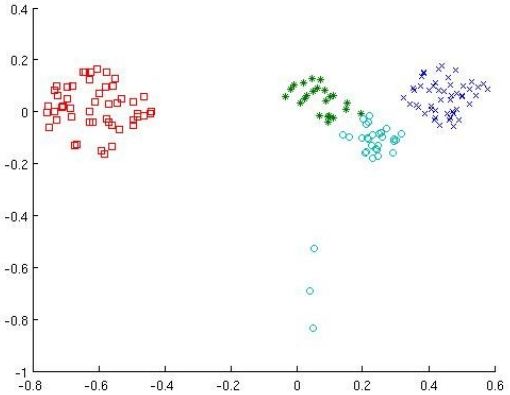

Fig4. D1: iK-Means

In the other hand, experiments with D2 were not favorable to CiK-Means. The average index for 100 trials provided by K-Means using 4 as the number of clusters was found to be 0.8435.

The iK-Means algorithm had a better performance, finding a stable index of 0.8693 using an optimized proportional threshold of 5%.

As stated CiK-Means did not have great results, the average index for 100 trials was found to be 0.8539 having a minimum of 0.8448, which is close to the result provided by K-Means, and a maximum of 0.8784.

It is interesting to note that while iK-Means always provides the same results, those provided by CiK-Means will depend on the rules and the entities present on those.

6. Conclusions and Future Research

There is still some level of discussion regarding the benefits of using a constrained version of k-means, at the same time that there are claims [9] that these algorithms can produce significantly worse results that not using constraints at all, others point out [10] that constrained variances of k-means have shown they are effective in experimental results.

Specifically in terms of COP-Kmeans, it has been argued [4] that the algorithm has failed to show marked improvements when using very few constraints. The results here presented demonstrate that although in some scenarios the results obtained by CiK-Means may not be as good as the results obtained with iK-Means, in others the difference of result is much more favorable to the CiK-Means, producing 100% correct answers using only around 3% of labeled data, which proves the algorithm may be very useful in certain situations.

Of course there is still room for improvement and future research will address basically three issues:

- How to increase the number of situations where the algorithm has superior results.
- There is a need to improve the reliability of the algorithm when a poor threshold is provided, in this cases the number of clusters found diverge from the real one and by consequence the accuracy of results is compromised.

One possible solution would be to readjust the CiK-Means to calculate the discarding threshold based on the Hartigan-adjusted iK-Means introduced by [12].

- Also, as it has already been pointed out, constrained algorithms based in hard constraints may be too sensitive to small labeling errors [2][3], a possible solution for that is the integration of a fuzzy clustering algorithm. These are already known for having better results than K-means in certain scenarios, such as in [6].

10. References

[1] Mirkin, B., *Clustering for Data Mining: A Data Discovery Approach*, Chapman and Hall/CRC, Boca Raton Fl. USA, 2005

[2] Wagstaff, K., C. Cardie, S. Rogers and S. Schroedl, Constrained K-means Clustering with Background Knowledge, Proceedings of the 18th International Conference on Machine Learning, 2001, pp.

577-584
[3] Handl, J, J. Knowles, On Semi-Supervised Clustering via Multiobjectve Optimization, Proceedings of the 8th annual conference on genetic and evolutionary computation, ACM, NY USA, 2006.
[4] Klein, D., A. D. Kamvar, C. D. Manning, From Instance-level Constraints to Space-level Constraints: Making the Most of Prior Knowledge in Data Clustering, The 19th International Conference on Machine Learning, 2002.
[5] Hand, D, H. Mannila, P. Smyth, Principles of Data Mining, MIT Press, Cambridge MA, 2001
[6] Goktepe, A.B., S. Altun, and A. Sezer, Soil clustering by fuzzy c-means algorithm, Advances in *Engineering Software*, Elsevier Science Ltd., Oxford UK, 2005, pp 691-698
[7] Castilho, W. F., H. A. Prado, and M. Ladeira, Informed K-Means: A Clustering Process biased by prior knowledge – A Case study in the dactyloscopic domain, Proceedings of the 6th International Conference on Enterprise Information Systems, Portugal, 2004
[8] Pelleg, D., A. Moore, X-means: Extending K-means with Efficient Estimation of the Number of Clusters, Proc. 17th International Conference on Machine Learning, Morgan Kaufmann, 2000, pp. 727-734.
[9] Wagstaff, K. L., S. Basu, and I. Davidson, When is Constrained Clustering Benefical and why?, The 21st National Conference on Artificial Intelligence and the 18th Innovative Applications of Artificial Intelligence Conference, AAAI Press, Boston USA, 2006, pp.
[10] Ng, M. K., A Note on Constrained K-Means Algorithms, Pattern Recognition, Elsevier Science Ltd., 2000 pp. 515-519
[11] Arthur, D., S. Vassilvitskii, K-means++: The Advantages of Careful Seeding, SODA '07: Proceedings of the 18th annual ACM-SIAM symposium on Discrete Algorithms, Society for Industrial and Applied Mathematics, Philadelphia USA, 2007
[12] Chiang, M.M., B. Mirking, Experiments for the number of clusters in K-Means, HIS '05 5th International Conference on Hybrid Intelligent Systems, 2005, pp. 6+