ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

On Instance Weighted Clustering Ensembles

Paul Moggridge and Na Helian and Yi Sun and Mariana Lilley

School of Physics Engineering and Computer Science - Department of Computer Science
University of Hertfordshire, Hatfield - UK

Abstract. Ensemble clustering is a technique which combines multiple clustering results, and instance weighting is a technique which highlights important instances in a dataset. Both techniques are known to enhance clustering performance and robustness. In this research, ensembles and instance weighting are integrated with the spectral clustering algorithm. We believe this is the first attempt at creating diversity in the generative mechanism using density based instance weighting for a spectral ensemble. The proposed approach is empirically validated using synthetic datasets comparing against spectral and a spectral ensemble with random instance weighting. Results show that using the instance weighted sub-sampling approach as the generative mechanism for an ensemble of spectral clustering leads to improved clustering performance on datasets with imbalanced clusters.

1 Introduction

Ensemble clustering does have some disadvantages, such as computational complexity, sensitivity to the choice of the generative mechanism, and added difficultly when explaining the results. However, when given computational resources and applied effectively, ensemble clustering can be very powerful for improving the quality of clustering results [1][6][12][13].

When designing a clustering ensemble, there are three key design decisions to be made. The first decision is the *generative mechanism*. This is the choice of how to generate an ensemble of base clusterings. Typically, the goal of the generative mechanism is to create both diverse and high quality base clusterings [2]. The generative mechanism has significant impacts on performance, and there is a variety of approaches that can be applied, including: different algorithms, different parameters[2][4], different subsets of instances [3][6], different subsets of features [1] or a combination mechanisms [8]. The second decision is the *consensus function* which defines how to combine the outputs of the base clusterings. Typically, the goal of the consensus function is to combine the base clusterings into a single clustering that is higher quality than any of the base clusterings. The third decision is whether to use *bagging* or *boosting*. Bagging executes the base clusterings in parallel, whereas boosting executes the base clustering sequentially. The boosting approach has the advantage that information discovered by a base clustering is used to influence the next iteration of base clustering. But the bagging approach has the advantage that the base clusterings can run in parallel, which is a useful property given the ubiquity of multi-processor systems.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Ensembles can be weighted in a number of ways [13]. In the instance weighting approach, weights are applied to the rows of data. These weights can encode information that the clustering process can utilise to enhance the clustering performance. One such clustering algorithm that could benefit is the spectral clustering algorithm. This algorithm, is known to have reduced clustering performance when clusters are imbalanced [7]. By utilising the weights to emphasize the low density clusters this limitation could be overcome. The concept behind this approach is that the density based weighting scheme up-samples sparse clusters and down-samples dense clusters. Overall, the clusters would then present as more balanced to the spectral base clusterings. This could improve clustering performance. The aim of this paper is to research whether instance weighting can be applied within the generative mechanism of a spectral clustering ensemble to enhance robustness and clustering performance.

2 Proposed Methodology

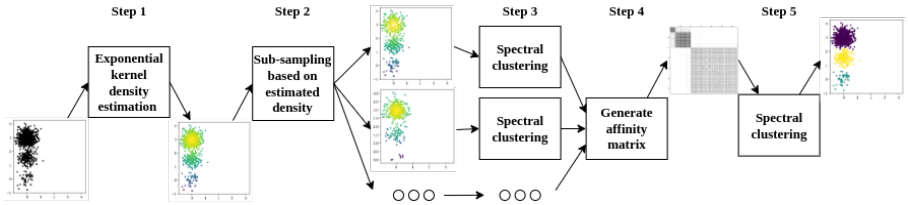

Fig. 1: Schematic illustration of the IWSE approach.

The first step calculates a weight for each instance, based on an exponential kernel density estimation function, the bandwidth value is chosen using the Silverman method [10]. The weights are min-max normalised between 0 and 1, and an additional inverted version of the weights is created where a high value indicates low density. Within Algorithm 1, the exponential kernel density estimation function can be seen, where n is the number of instances, y is an instance to calculate the weight of, x is each of the instances, and h represents the bandwidth parameter used to determine the weights. In the second step, the approach uses bagging. Bagging was chosen for its computational advantages. When using bagging, a choice between bootstrap and sub-sampling arises. Sub-sampling was chosen as it has been shown to perform favourably for clustering ensembles [6]. The algorithm generates instance weighted sub-samples of the dataset, randomly sized between 30% and 60%. These are done in parallel (given the appropriate hardware), corresponding to the number of bags parameter M. For example, $M = 32$ would indicate 32 sub-samples are created. For experimental purposes, there is three variations of the weighting scheme. In the "L" variation the low density instances are more likely to be selected. In the "H" variation the high density instances are more likely to be selected. Finally, the "U" variation uniformly randomly switches between the "L" and "H" weighting

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

schemes. In the third step, normalised spectral clustering [9] is utilised for the base clustering of the sub-samples. In the fourth step, once all spectral base clusterings have been executed, consensus takes place. The pair-wise similarities of the cluster assignments of the instances are accumulated into a co-association matrix. This approach is essentially Cluster-based Similarity Partitioning Algorithm (CSPA) [11]. However, instead of using METIS clustering algorithm, (typically used in the CSPA consensus function), spectral clustering is used; this substitution was made because both are similar (in that they are graph partitioning methods), but spectral is more readily available in well-tested libraries. The approach based on CSPA was chosen as it is simple and produces robust performance [5][6]. In the final step, this co-association matrix is treated as an affinity matrix to which spectral clustering is applied. This provides the final clustering result. Algorithm 1 provides a technical description of our approach.

Algorithm 1 Instance Weighted Spectral Ensemble (U)

Input: $\mathbb{X} = \{x_1, ..., x_n\}$, k, k^*, M

Calculate bandwidth value h using Silverman method.
Compute weights $\mathbb{W}$ using $\rho_K(y) = \sum_{i=1}^{n} exp\left(\frac{-dist(x_i,y)}{h}\right)$ for $\mathbb{X}$
Normalise weights $\mathbb{W}$ using equation $\frac{max(x)-min(x)}{x-min(x)}$
Compute inverted weights $\mathbb{W}^*$ using equation $x^* = |x - 1.0|$
for $m \leftarrow 1$ *to* M **do**
    Let $r \in \{0, 1\}$ with uniform probability (for switching weighting schemes)
    Let $s \in \{x \in \mathbb{R} | x \geq 0.3 \; and \; x \leq 0.6\}$ with uniform probability
    if $r = 1$ **then**
        Let $\mathbb{S} \subset \mathbb{X}$ be a sub-sample of size $n \times s$ using probability $\mathbb{W}$
    else
        Let $\mathbb{S} \subset \mathbb{X}$ be a sub-sample of size $n \times s$ using probability $\mathbb{W}^*$
    end if
    Partition $\mathbb{S}$ into $\mathbb{P} = \{C_1, ..., C_k\}$ using spectral with k and k^*
    Construct $n \times n$ co-association matrix $\mathbb{A}_m$ for $\mathbb{P}$
end for
Let $\mathbb{A}^* = \sum_{m=1}^{M} \mathbb{A}_m$
Partition $\mathbb{A}^*$ into $\mathbb{P}^* = \{C_1, ..., C_k\}$ using spectral with k and k^*

Output: $\mathbb{P}^* = \{C_1, ..., C_k\}$

3 Experiments

To empirically validate the approach, IWSE was compared with spectral (S), and a spectral ensemble with random sub-sampling (SER) on increasingly imbalanced datasets. For all algorithms, the k value was set to reflect the ground truth of the dataset and the k^* neighbours parameter was set to 9. Where applicable the bags parameter M was set to 32. Increasing M beyond 32 sees a diminishing return in terms of clustering performance for execution time spent.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Normalised Mutual Information (NMI) was used to evaluate the experiments. The 2D datasets are generated per run of the experiment to minimise effects from artefacts of the stochastic generation process. Each contains three clusters drawn from normal distributions positioned at (0,0), (0,1.5), and (0, 3), and each cluster has a variance of 0.1 in both dimensions. Cluster sizes were determined using a scaling factor x with values from 1 to 5 in increments of 0.25 where $|C_0| = 50, |C_1| = |C_0| \times x, |C_2| = |C_1| \times x$. A sample of the datasets can be seen in Figure 2.

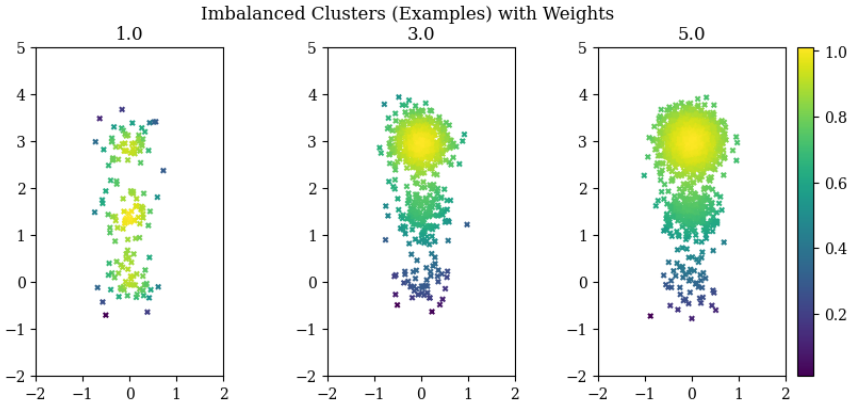

Fig. 2: A sample of the "imbalance" datasets, the colouration represents the weights. The title for each sub-plot shows the "imbalance ratio".

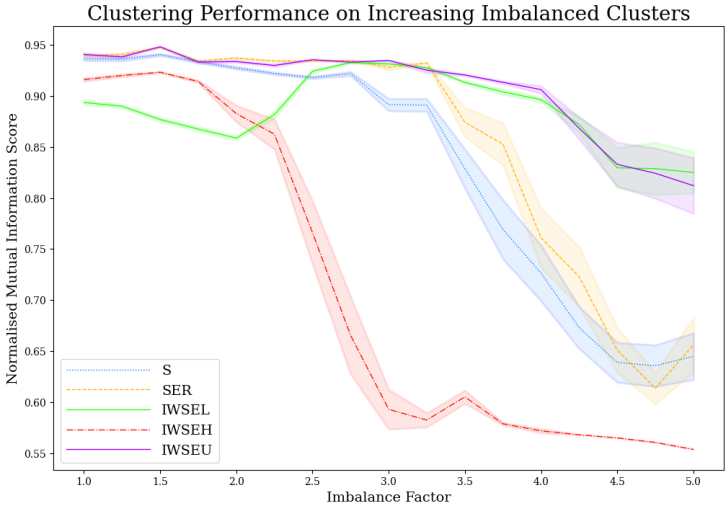

Fig. 3: IWSEU and IWSEL perform well despite imbalanced clusters.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Figure 3 shows that as the clusters become increasingly imbalanced, the performance of spectral clustering and SER drops significantly. When the imbalance ratio reaches 3.75, IWSEL and IWSEU offer superior performance over S or SER. As can be expected, IWSEH performs increasingly poorly as the imbalance increases (due to lack a sampling of the smallest cluster). However, interestingly IWSEU performs similarly to IWSEL and on occasion even better than IWSEL. This is despite incorporating degenerate "H" partitionings. It seems that this could be due to the consensus function benefiting from the degenerate partitionings, as has been observed in other research [2].

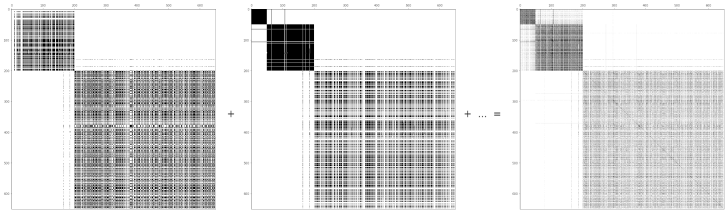

Fig. 4: Left: a co-association matrix generated by a base clustering in IWSE using mode "H". Centre: a co-association matrix generated by a base clustering in IWSE using mode "L". Right: The sum of the co-association matrices.

To illustrate this point, Figure 4 shows some of the partitionings created within the IWSEU approach and the resultant accumulated affinity matrix to be used for the final clustering. On the left of Figure 4, IWSEU randomly chose the "H"mode, in this weighting scheme the base clusterings confuse the small cluster with the medium cluster, but they do encode some information about the largest cluster that the "L" mode does not capture. In the centre of Figure 4, IWSEU randomly chose the "L" mode, in this weighting scheme the base clusterings consistently identify the smallest cluster and achieve good clustering performance. It seems the CSPA method of summing the pairwise similarity across the bags (Figure 4 right) then clustering, can handle some degenerate partitionings and even benefit from the information they provide.

4 Conclusion and Future Work

The experiments show that the instance weighted ensemble approach enhanced the ability of spectral clustering to handle imbalanced data. However, IWSE does have some drawbacks, most notably the consensus function is computationally expensive, although this could be replaced with the HyperGraph Partitioning Algorithm (HGPA), as recommended by [5] for greater efficiency. In summary, this work shows some promising initial research into how instance weights could be used to perturb sub-sampling to enhance an ensembles clustering performance. With regard to future work, it is suggested that combining multiple weighting schemes could enhance performance [5].

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

References

[1] Al-Razgan, M., Domeniconi, C.: Weighted clustering ensembles. In: Proceedings of the 2006 SIAM International Conference on Data Mining. pp. 258–269. SIAM (2006)

[2] Fern, X.Z., Lin, W.: Cluster ensemble selection. Statistical Analysis and Data Mining: The ASA Data Science Journal **1**(3), 128–141 (2008)

[3] Frossyniotis, D., Likas, A., Stafylopatis, A.: A clustering method based on boosting. Pattern Recognition Letters **25**(6), 641–654 (2004)

[4] Huang, D., Wang, C.D., Wu, J.S., Lai, J.H., Kwoh, C.K.: Ultra-scalable spectral clustering and ensemble clustering. IEEE Transactions on Knowledge and Data Engineering **32**(6), 1212–1226 (2019)

[5] Jia, J., Xiao, X., Liu, B., Jiao, L.: Bagging-based spectral clustering ensemble selection. Pattern Recognition Letters **32**(10), 1456–1467 (2011)

[6] Parvin, H., Minaei-Bidgoli, B., Alinejad-Rokny, H., Punch, W.F.: Data weighing mechanisms for clustering ensembles. Computers & Electrical Engineering **39**(5), 1433–1450 (2013)

[7] Qian, J., Saligrama, V.: Spectral clustering with imbalanced data. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3057–3061. IEEE (2014)

[8] Ren, Y., Domeniconi, C., Zhang, G., Yu, G.: Weighted-object ensemble clustering. In: 2013 IEEE 13th International Conference on Data Mining. pp. 627–636. IEEE (2013)

[9] Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on pattern analysis and machine intelligence **22**(8), 888–905 (2000)

[10] Silverman, B.W.: Density estimation for statistics and data analysis, vol. 26. CRC press (1986)

[11] Strehl, A., Ghosh, J.: Cluster ensembles—a knowledge reuse framework for combining multiple partitions. Journal of machine learning research **3**(Dec), 583–617 (2002)

[12] Topchy, A., Jain, A.K., Punch, W.: A mixture model for clustering ensembles. In: Proceedings of the 2004 SIAM international conference on data mining. pp. 379–390. SIAM (2004)

[13] Zhang, M.: Weighted clustering ensemble: A review. Pattern Recognition **124**, 108428 (2022)