Unsupervised Ensemble Minority Clustering

Edgar Gonzàlez
TALP Research Center

Jordi Turmo
TALP Research Center

Abstract

Cluster analysis lies at the core of most unsupervised learning tasks. However, the majority of clustering algorithms depend on the all-in assumption, in which all objects belong to some cluster, and perform poorly on minority clustering tasks, in which a small fraction of signal data stands against a majority of noise.

The approaches proposed so far for minority clustering are supervised: they require the number and distribution of the foreground and background clusters. In supervised learning and all-in clustering, combination methods have been successfully applied to obtain distribution-free learners, even from the output of weak individual algorithms.

In this report, we present a novel ensemble minority clustering algorithm, Ewocs, suitable for weak clustering combination, and provide a theoretical proof of its properties under a loose set of constraints. The validity of the assumptions used in the proof is empirically assessed using a collection of synthetic datasets.

Keywords: Minority clustering, ensemble clustering, unsupervised clustering.

1 Introduction

The amount of data available in digital form is increasing every day. Given the expensive costs of human inspection (and annotation), unsupervised approaches to mining these data become more and more paramount.

Cluster analysis lies at the core of most unsupervised learning tasks. Jain et al. (1999) define clustering as "*the organization of a collection of patterns [...] into clusters based on similarity. Intuitively, patterns within a valid cluster are more similar to each other than they are to a pattern belonging to a different cluster*". In addition to *pattern*, each element to be clustered has also received the names of "*object, record, point, vector, [...] event, case, sample, observation, or entity*" (Tan et al., 2005, ch. 2). We will stick to the term *object* thorough this report.

In this the most common setting, it is assumed that all objects belong to some cluster. Even if several surveys have reviewed the vast literature on clustering methods (Dubes and Jain, 1980; Jain et al., 1999; Xu and Wunsch, 2005), so far they all have focused on this standard task, which can be named *all-in clustering*. Two of the most widely used methods to solve it are the distance-based k-means (MacQueen, 1967) and the probabilistic-model-based Expectation-Maximization (Dempster et al., 1977) algorithms.

However, there is a number of situations in which the data are known not to fit neatly within this all-in assumption. In such cases, we know there is a fraction of data which are neither similar to one another nor to the data within the clusters. Often, these data will correspond to a certain form of *noise* and should hence be separated from the sought regular clusters, which constitute the *signal*. Within this alternative setting, a number of different tasks can be identified according to the characteristics of the data and the aim of the task itself.

In one of these tasks, the all-in clustering goal is preserved, but the data are known to contain a small fraction of noise. This has been called the *robust clustering* task (Davé and Krishnapuram, 1997). To solve it, some authors have proposed changes to standard clustering methods to make them more robust to the presence of noise. The replacement of the centroid calculation in k-means by that of medoids in the k-medoids or partitioning about medoids (PAM; Kaufman and Rousseeuw, 2005, Ch. 2) algorithm, or the use of mixtures of Student t distributions instead of Gaussian ones (Peel and McLachlan, 2000) are examples of work in this direction.

In other approaches to the task, robustness is increased by explicitly incorporating a noise cluster, often with different properties from the regular signal clusters. For instance, distance-based

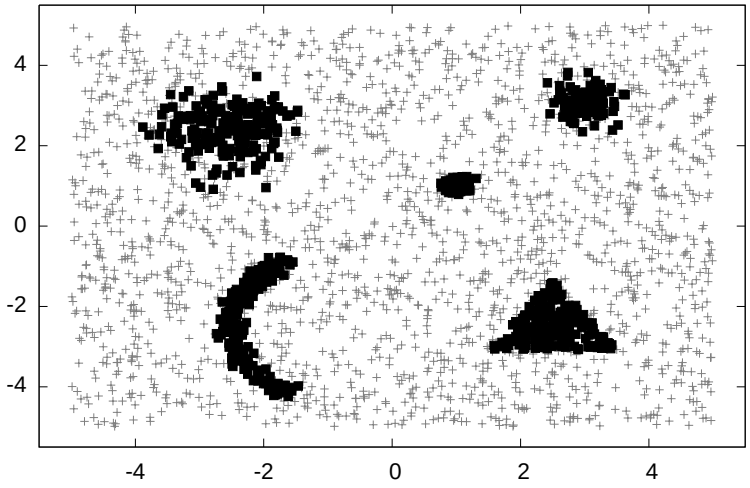

Figure 1: Sample Toy minority clustering dataset

methods have been extended to incorporate an ideal noise prototype, "*a universal entity such that it is always at the same distance from every point in the data-set*" (Davé, 1991); and model-based clustering methods have been proposed which incorporate, among a mixture of otherwise Gaussian components, an extra one with a Poisson (Banfield and Raftery, 1993) or uniform (Guillemaud and Brady, 1997; Biernacki et al., 2000, §4.2.3) distribution to account for noise.

A last family of approaches is that of algorithms specifically devised for robust clustering, such as BIRCH (Zhang et al., 1996) or DBSCAN (Ester et al., 1996).

It is worth noticing that there is a number of related tasks which share this setting, such as *one-class classification* or *learning* (Moya et al., 1993; Schölkopf et al., 2001; Tax and Duin, 2004) and *outlier detection* (Hodge and Austin, 2004; Chandola et al., 2009). In both cases, there is also a dataset containing both signal and a fraction of noise objects. However, the focus of these tasks shifts away from that of clustering, becoming the estimation of a model which covers the signal objects in the former, and the detection of the objects that significantly deviate from the rest in the latter.

Nevertheless, there is still another setting to be considered. In some cases, there will only be a minority of signal objects, standing against the majority of noise. Most often, the signal objects will be embedded within the noise ones, becoming respectively *foreground* and *background* objects, and the distinction between the former and the latter must be done on grounds of density criteria. In the literature, this task has been compared to "*clustering needles in a haystack*" (Ando, 2007), and has received names such as *one-class clustering* (Crammer and Chechik, 2004), *density-based clustering* (Gupta and Ghosh, 2006) or *minority detection* (Ando and Suzuki, 2006). As a catchall term, in this report we will refer to this setting and task as *minority clustering*.

Even if this new task is related to the previously presented ones, the reversal of the signal-to-noise ratio can make existing approaches unsuitable. For instance, Crammer and Chechik (2004) give insights into why existing one-class classification approaches, which are tailored to finding large-scale structures, may be unable to identify small and locally dense regions embedded in noise. Empirical comparisons have also stated the low performance exhibited by all-in and robust clustering methods in the minority clustering task (Gupta and Ghosh, 2006).

However, to the best of our knowledge, all the methods proposed so far require as an input the distribution of the foreground clusters or both the foreground clusters and the background noise, either in the form of a probability distribution or, equivalently, of a divergence metric[1]. This can become a significant issue when facing large amounts of data coming from a new and unexplored domain, whose distribution may be completely unknown.

[1] A Bregman divergence induces a probability distribution of the exponential family (Banerjee et al., 2005)

1.1 Ensemble Minority Clustering

In the context of supervised learning, combination methods have been successfully used to overcome the limitations of individual algorithms. They provide a way to obtain distribution-free learners able to perform competitively across a wide spectrum of learning problems, even from the combination of the outputs of weak learning algorithms (Freund and Schapire, 1995). More recently, a number of combination methods have appeared for all-in clustering (e.g., Strehl and Ghosh, 2002; Topchy et al., 2003, 2004; Gionis et al., 2005). Among them, Topchy et al. (2003) introduced the idea of using an ensemble of weak, almost random, clusterings to obtain a high-quality consensus clustering.

In this report, we present an unsupervised minority clustering approach, Ensemble Weak minOrity Cluster Scoring (EWOCS), based on weak-clustering combination. In it, a number of weak clusterings is generated, and the information coming from each one of them is combined to obtain a score for each object. A threshold separating foreground from background objects is then inferred from the distribution of these scores. We have been able to find a theoretical proof of the properties of the proposed method, and we have considered a number of criteria by which the threshold value can be determined. Finally, we have assessed the validity of the assumptions used in our proof empirically, using a collection of synthetic datasets.

The EWOCS algorithm has already been used in the real-world task of relation detection—which was reduced to a minority clustering problem (Gonzàlez and Turmo, 2009). However, we now provide a formalization of the approach, as a minority clustering algorithm by itself, and an study of its theoretical properties—which were both missing from our previous work.

The rest of the report is organized as follows. Section 2 gives an overview of related work in the fields of minority clustering and clustering combination. Next, Section 3 contains a description of the EWOCS approach, particularly the derivation of a minority clustering algorithm whose properties are theoretically proved under a set of conditions. The obtained algorithm has a number of components which allow different implementations: Section 4 briefly discusses the possibility of using weak clustering algorithms within EWOCS, whereas Section 5 sketches methods by which the threshold score can be determined. Section 6 contains the details and results of an empirical evaluation of the degree to which one particular weak clustering family satisfies the requirements to be used within EWOCS. Finally, Section 7 draws conclusions of our work.

Given that some readers might not be familiar with the terminology from fuzzy set theory, an Appendix contains, for reference, brief definitions for the concepts that are used in our work.

2 Related Work

One of the first works to identify the minority clustering task in opposition to that of one-class classification is that of Crammer and Chechik (2004). The authors formalize the problem in terms of the Information Bottleneck principle (IB; Tishby et al., 1999), and provide a sequential algorithm to solve this one-class IB problem. Given a Bregman divergence (Bregman, 1967) as a generalized measure of object discrepancy, and a fixed radius value, the OC-IB method outputs a centroid for a single dense cluster. The foreground cluster consists of the objects which fall inside the Bregmanian ball of given radius centered around the given centroid. More recently, Crammer et al. (2008) propose a different algorithm for the same model, based in rate-distortion theory and the Blahut-Arimoto algorithm (Blahut, 1972; Arimoto, 1972), and extend it to allow for more than one cluster.

In a different direction, Gupta and Ghosh (2005) reformulate the problem in terms of cost, defined as the sum of divergences from the cluster centroid to each sample within it, and extend the OC-IB method to avoid local minima. A triad of methods (HOCC, BBOCC and Hyper-BB) is proposed. However, the requirement of an a priori determination of the cluster radius (or equivalently, size) is not removed, and the output remains a single ball-shaped cluster.

To overcome this second limitation, Gupta and Ghosh (2006) propose Bregman Bubble Clustering (BBC), as a generalization of BBOCC to several clusters. However, the number of such clusters must still be given a priori, as well as the desired joint cluster size. The authors also propose a soft clustering version of BBC, as well as a unified framework between all-in Bregman clustering (Banerjee et al., 2005) and BBC, in all their hard and soft versions.

The work of Ando and Suzuki (2006) is similar to previous ones in that it also uses the Information Bottleneck principle as a criterion to identify a single minority cluster. However, the

method is more general in the sense that it allows arbitrary distributions, not only those induced by Bregman divergences, as foreground and background. Ando (2007) extends this last proposal, allowing multiple foreground clusters, and also provides a unifying framework of which not only the task of minority clustering, but also those of outlier detection and one-class learning, are particular cases.

Regarding clustering combination, the formalization of the most usual setting, *ensemble clustering*, is due to Strehl and Ghosh (2002). The authors define the ensemble clustering task as that of "*combining multiple partitionings of a set of objects without accessing the original features*". The authors also propose a set of algorithms for ensemble clustering (CSPA, HGPA and MCLA), all based on reduction to graph- and hypergraph-partitioning problems, as well as a criterion function to select the best one among the clusterings produced by them.

Following this same ensemble setting, the method of Gionis et al. (2005) starts by building a correlation matrix between all pairs of objects in the data. The value of each entry is the fraction of clusterings in the ensemble in which the two objects are clustered together. This reduces the problem of ensemble clustering to that of *correlation clustering*. The sought consensus clustering is the one which minimizes disagreement with respect to each clustering in the ensemble. Given that this is a hard combinatorial problem, a number of approximate correlation clustering algorithms (Balls, Agglomerative, Furthest) are proposed, even if the proof of approximation ratio is given for only the first one. Additionally, a local search procedure is devised which can improve the result obtained by the approximate algorithms.

Finally, Topchy et al. (2005) propose two more approaches to the same problem, based on mixture modelling and information theory, respectively. The former is solved using the EM algorithm, whereas for the latter k-means is used. However, this work is remarkable because it introduces the idea of *weak clustering ensemble*: instead of resorting to strong clustering algorithms to obtain the different individual clusterings to be later combined, the authors propose the use of a larger number of inexpensive weak clusterers, where a *weak clustering algorithm* is defined as one which "*produces a partition, which is only slightly better than a random partition of the data*". More specifically, a random hyperplane separator and a run of k-means on a random subspace of the data are used with this goal.

3 EWOCS

This section presents our Ensemble Weak minOrity Cluster Scoring (EWOCS) algorithm to solve the task of minority clustering.

First Section 3.1 defines our setting for the task of minority clustering. Section 3.2 presents, from a theoretical point of view, the scoring scheme that lies at the core of our method. Sections 3.3 and 3.4 then study the conditional probability distributions of the assigned scores: the first one on a single dataset; the second, across multiple dataset samplings. Next, Section 3.5 introduces the concept of *consistent clustering*, and shows how, when using clustering functions from a consistent family, an inequality on the score expectations for foreground and background objects can be established. This inequality will allow us to obtain as a corollary, in Section 3.6, a generic algorithmic procedure for minority clustering, based on the proposed scores. Finally, it is also possible to obtain a clustering model using this algorithm: its construction and application is described in the last Section 3.7.

3.1 Task Setting

Assume we have a finite set of $\hat{k}$ generative distributions or *sources* $\Psi = \{\psi_1 \dots \psi_{\hat{k}}\}$, with a priori probabilities $\{\alpha_1 \dots \alpha_{\hat{k}}\}$. Assume we also have a dataset $\mathcal{X} = \{x_1 \dots x_n\}$ of size n, which has been sampled from Ψ. Each object x_i will be generated by one of the sources in Ψ, and we can hence consider a set $\mathcal{Y}$ of hidden variables, with each $y_i \in \Psi$ containing the source which generated the corresponding x_i.

Without loss of generality, we will name the first of those sources, ψ_1, the *background source*; and the objects generated by it, the *background objects*. The rest of sources and objects shall be named the *foreground sources* (whose set will be denoted as Ψ^+) and the *foreground objects*, respectively.

In the setting we are interested in, we can make two assumptions which can be stated as follows:

Density Foreground sources are *dense*, i.e., objects generated by the same foreground source are more similar to each other than to those generated by the background source.

Locality Foreground sources are *local*, i.e., objects generated by different foreground sources are as similar to each other as they are to those generated by the background source

These assumptions are similar to those in other works, for instance, those of *atypicalness* and *local distribution* defined by Ando (2007).

We can then define:

Definition 1 (Hard partitional clustering)
*A **hard (partitional) clustering** Π of dataset $\mathcal{X}$ is a partition $\Pi = \{\pi_1 \dots \pi_k\}$ of $\mathcal{X}$ whose aim is to maximize a certain criterion function F. Each one of the subsets $\pi_c \in \Pi$ is a **hard cluster**.*

Definition 2 (Soft partitional clustering)
*A **soft (partitional) clustering** Π of dataset $\mathcal{X}$ is a fuzzy pseudopartition[2] $\Pi = \{\pi_1 \dots \pi_k\}$ of $\mathcal{X}$ whose aim is to maximize a certain criterion function F. Each one of the fuzzy subsets $\pi_c \in \Pi$ is a **soft cluster**.*

> Remark A hard clustering can be seen as a particular case of soft clustering where the grade of membership of a certain x_i to the π_c is zero for all but exactly one cluster, for which the grade is one.

We can also assume we have a (possibly infinite) family of *clustering functions* F. From them, a sequence of functions $(f_1 \dots)$ are independently drawn at random, with a certain probability density. When applied to the dataset, each f_r will produce a soft[3] clustering $\Pi_r = \{\pi_{r1} \dots \pi_{rk_r}\}$ with a number k_r of clusters.

3.2 Per-Clustering Scoring

After clustering function f_r is applied, the *cluster size* and *object scores* can be calculated from the output clustering Π_r.

Definition 3 (Cluster size)
*The **size** of cluster π_{rc} is the sum of the grade of membership to the cluster of all objects in the dataset:*

$$\text{size}(\pi_{rc}) = \sum_{x_i \in \mathcal{X}} \text{grade}(x_i, \pi_{rc}) \tag{1}$$

Definition 4 (Object score)
*The **score** of an object x_i by clustering function f_r is*

$$s_{ri} = \sum_{\pi_{rc} \in \Pi_r} \text{grade}(x_i, \pi_{rc}) \cdot \text{size}(\pi_{rc}) \tag{2}$$

i.e., the sum of the sizes of the output clusters, weighted by the grade of membership of x_i to each one of them.

An additional concept will turn out to be of much importance later.

Definition 5 (Co-occurrence vector)
*The **co-occurrence vector** for object x_i and clustering function f_r is $\vec{c}_{ri} = [c_{ri1} \dots c_{rin}]^T$, where each component c_{rij} is*

$$c_{rij} = \sum_{\pi_{rc} \in \Pi_r} \text{grade}(x_i, \pi_{rc}) \cdot \text{grade}(x_j, \pi_{rc}) \tag{3}$$

[2] Following the definitions of Bezdek (1981) and Klir and Yuan (1995) (see Definition B in the Appendix).
[3] The result is also valid for hard clustering families, being a particular case of soft clustering.

REMARK Using the co-occurrence vector, the score of object x_i by clustering function f_r can be written as

$$\begin{aligned} s_{ri} &= \sum_{\pi_{rc}\in\Pi_r} \text{grade}(x_i,\pi_{rc})\cdot\text{size}(\pi_{rc}) \\ &= \sum_{\pi_{rc}\in\Pi_r} \text{grade}(x_i,\pi_{rc})\cdot\sum_{x_j\in\mathcal{X}} \text{grade}(x_j,\pi_{rc}) \\ &= \sum_{\pi_{rc}\in\Pi_r}\sum_{x_j\in\mathcal{X}} \text{grade}(x_i,\pi_{rc})\cdot\text{grade}(x_j,\pi_{rc}) \\ &= \sum_{x_j\in\mathcal{X}}\sum_{\pi_{rc}\in\Pi_r} \text{grade}(x_i,\pi_{rc})\cdot\text{grade}(x_j,\pi_{rc}) \\ &= \sum_{x_j\in\mathcal{X}} c_{rij} \end{aligned}$$

From its definition, we can infer that the co-occurrence vector will satisfy the following property:

Proposition 6
The values of the entries c_{rij} in the co-occurrence vector belong to the interval $[0,1]$.

PROOF By the properties of fuzzy pseudopartitions[4], and hence of soft clusterings, we know that

$$\forall x_i : \sum_{\pi_{rc}\in\Pi_r} \text{grade}(x_i,\pi_{rc}) = 1$$

The product of two of these terms, which will also be equal to 1, can be expressed as

$$\begin{aligned} 1 &= \left(\sum_{\pi_{rc}\in\Pi_r} \text{grade}(x_i,\pi_{rc})\right)\cdot\left(\sum_{\pi_{rc}\in\Pi_r} \text{grade}(x_j,\pi_{rc})\right) \\ &= \sum_{\pi_{rc},\pi_{rc'}\in\Pi_r} \text{grade}(x_i,\pi_{rc})\cdot\text{grade}(x_j,\pi_{rc'}) \\ &= \sum_{\pi_{rc}\in\Pi_r} \text{grade}(x_i,\pi_{rc})\cdot\text{grade}(x_j,\pi_{rc}) + \sum_{\substack{\pi_{rc},\pi_{rc'}\in\Pi_r \\ \pi_{rc}\neq\pi_{rc'}}} \text{grade}(x_i,\pi_{rc})\cdot\text{grade}(x_j,\pi_{rc'}) \\ &= c_{rij} + \bigtriangledown c_{rij} \end{aligned}$$

Given that the grade of membership is by definition non-negative[5], all pairwise products of grades will also be non-negative—and, being sums of pairwise products, both c_{rij} and $\bigtriangledown c_{rij}$ will at their turn be non-negative too: $0 \le c_{rij}, \bigtriangledown c_{rij}$.

Finally, given that c_{rij} and $\bigtriangledown c_{rij}$ are two non-negative terms adding up to 1, it is clear that neither of them can exceed this value: $c_{rij}, \bigtriangledown c_{rij} \le 1$. Hence, as we wanted to prove, $0 \le c_{rij} \le 1$ ∎

Rather than considering a single application of one clustering function $f_r \in F$ on $\mathcal{X}$, we will mainly be concerned with aggregating the results over a number R of repetitions of the process. In this context, we can define:

Definition 7 (Average co-occurrence vector)
*The sequence of **average co-occurrence vectors** for object x_i is $(\vec{c}^{\star}_{1i}\ldots)$, where each component of $\vec{c}^{\star}_{Ri} = [c^{\star}_{Ri1}\ldots c^{\star}_{Rin}]^T$ is*

$$c^{\star}_{Rij} = \frac{1}{R}\sum_{r=1}^{R} c_{rij} \tag{4}$$

Definition 8 (Average score)
*The sequence of **average scores** of object x_i is $(s^{\star}_{1i}, s^{\star}_{2i}\ldots)$, where each $s^{\star}_{Ri}$ is*

$$s^{\star}_{Ri} = \frac{1}{R}\sum_{r=1}^{R} s_{ri} \tag{5}$$

[4]See Definition B in the Appendix.
[5]See Definition A in the Appendix.

REMARK Using average co-occurrence vectors, the average score of object x_i can be expressed as

$$s^\star_{Ri} = \frac{1}{R}\sum_{r=1}^{R} s_{ri} = \frac{1}{R}\sum_{r=1}^{R}\sum_{x_j\in\mathcal{X}} c_{rij} = \sum_{x_j\in\mathcal{X}}\frac{1}{R}\sum_{r=1}^{R} c_{rij} = \sum_{x_j\in\mathcal{X}} c^\star_{Rij}$$

It is interesting to note that

Proposition 9
The s_{ri} are linear transformations of $\vec{c}_{ri}$, and the $s^\star_{Ri}$ are linear transformations of $\vec{c}^\star_{Ri}$.

PROOF Using an all-ones vector,

$$\begin{aligned} s_{ri} &= \vec{1}^T\cdot\vec{c}_{ri} = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}\cdot\begin{bmatrix} c_{ri1} & c_{ri2} & \cdots & c_{rin} \end{bmatrix}^T = \sum_{x_j\in\mathcal{X}} c_{rij} \\ s^\star_{Ri} &= \vec{1}^T\cdot\vec{c}^\star_{Ri} = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}\cdot\begin{bmatrix} c^\star_{Ri1} & c^\star_{Ri2} & \cdots & c^\star_{Rin} \end{bmatrix}^T = \sum_{x_j\in\mathcal{X}} c^\star_{Rij} \end{aligned}$$

■

3.3 Dataset-Conditioned Distribution

The dataset $\mathcal{X}$ and clustering function f_r uniquely determine the values for the co-occurrence vectors $\vec{c}_{ri}$, and hence for all other values considered in the previous Section. However, as the selection of f_r is not deterministic, the c_{rij} can be regarded as random variables, and their conditional distribution across clustering functions, given a certain dataset $\mathcal{X}$, can be considered.

As the selection of each f_r is independent from the others, the values of the c_{rij} for different r will also be. The $\vec{c}_{ri}$ for different r will hence be independent and identically distributed random vectors, with a common expectation vector $\vec{\mu}_i$ and covariance matrix Σ_i. We will refer to each element, μ_{ij}, of $\vec{\mu}_i$ as the *affinity* of x_i and x_j.

Definition 10 (Object affinity)
*The **affinity** of objects x_i and x_j is the conditional expectation of c_{rij} given $\mathcal{X}$,*

$$\mu_{ij} = E[c_{rij} \mid \mathcal{X}] \tag{6}$$

REMARK Being the expectations of the c_{rij}, with $c_{rij} \in [0,1]$, the affinities μ_{ij} will also fall in the $[0,1]$ interval.

We can additionally define

Definition 11 (Object expected score)
*The **expected score** of object x_i is the conditional expectation of s_{ri} given $\mathcal{X}$,*

$$\mu_i = E[s_{ri} \mid \mathcal{X}] \tag{7}$$

It is then easy to successively prove that

Proposition 12
The value of the expected score μ_i of object x_i is

$$\mu_i = E[s_{ri} \mid \mathcal{X}] = \sum_{x_j\in\mathcal{X}} \mu_{ij} \tag{8}$$

PROOF As s_{ri} is the sum of the c_{rij}, its conditional expectation is

$$\mu_i = E[s_{ri} \mid \mathcal{X}] = E[\sum_{x_j\in\mathcal{X}} c_{rij} \mid \mathcal{X}] = \sum_{x_j\in\mathcal{X}} E[c_{rij} \mid \mathcal{X}] = \sum_{x_j\in\mathcal{X}} \mu_{ij}$$

REMARK Being the sum of $n = |\mathcal{X}|$ terms within the interval $[0,1]$, the value of μ_i will fall in the interval $[0,n]$. In order to make scores across differently-sized datasets comparable, we will also consider a **normalized expected score** $\bar{\mu}_i$, defined as $\bar{\mu}_i = \mu_i/n$.

Proposition 13
As the number of repetitions R increases, the conditional distributions of the average co-occurrence vectors $\vec{c}^{\star}_{Ri}$ approach a multivariate Gaussian distribution with expectation $\vec{\mu}_i$ and covariance matrix Σ_i/R.

PROOF As the c_{rij} are independent and identically distributed for different r, by the multivariate central limit theorem we know that the sequence

$$\sqrt{R}\left(\frac{1}{R}\sum_{r=1}^{R}\vec{c}_{ri} - \vec{\mu}_i\right) = \sqrt{R}\,(\vec{c}^{\star}_{Ri} - \vec{\mu}_i)$$

converges in distribution to a multivariate Gaussian distribution with expectation $\vec{\mu}_i$ and covariance matrix Σ_i. Hence, for large enough R,

$$\begin{aligned}\sqrt{R}\,(\vec{c}^{\star}_{Ri} - \vec{\mu}_i) &\approx \mathcal{N}(0, \Sigma_i)\\ \vec{c}^{\star}_{Ri} - \vec{\mu}_i &\approx \mathcal{N}(0, \Sigma_i/R)\\ \vec{c}^{\star}_{Ri} &\approx \mathcal{N}(\vec{\mu}_i, \Sigma_i/R)\end{aligned}$$

■

Proposition 14
As the number of repetitions R increases, the conditional distributions of the average scores $s^{\star}_{Ri}$ approach a Gaussian distribution with expectation μ_i.

PROOF Being linear transformations of random vectors $\vec{c}^{\star}_{Ri}$ approaching a multivariate Gaussian distribution, the $s^{\star}_{Ri}$ also approach a Gaussian distribution

$$s^{\star}_{Ri} = \vec{1}^{T}\cdot\vec{c}^{\star}_{Ri} \approx \mathcal{N}(\vec{1}^{T}\cdot\vec{\mu}_i, (\Sigma^{\star}_{Ri})^2)$$

■

with a certain variance $(\Sigma^{\star}_{Ri})^2$. The conditional expectation of these variables hence converges to

$$\lim_{R\to\infty} E[s^{\star}_{Ri} \mid \mathcal{X}] = \vec{1}^{T}\cdot\vec{\mu}_i = \sum_{x_j\in\mathcal{X}}\mu_{ij} = \mu_i$$

3.4 Sampling Distribution

We can now proceed to consider the distribution of the scores across multiple samplings of the dataset $\mathcal{X}$. In particular, we will first focus on the distribution of the affinity μ_{ij} between objects x_i and x_j, conditioned to their being respectively generated by a certain pair of sources ψ_s and ψ_t—a measure which we shall name the *affinity* of the two sources, ζ_{st}.

Definition 15 (Source affinity)
*The **affinity** of sources ψ_s and ψ_t is the conditional expectation of the object affinity μ_{ij}, given that $y_i = \psi_s$ and $y_j = \psi_t$, across all datasets $\mathcal{X}$ sampled from Ψ:*

$$\zeta_{st} = E[\mu_{ij} \mid y_i = \psi_s, y_j = \psi_t]$$

A particular case of affinity is that of $\psi_t = \psi_s$, which we shall name the *self-affinity* ζ_{ss} of source ψ_s.

We can now also consider the conditional expectation of the normalized expected scores $\bar{\mu}_i$ for objects from source ψ_s.

Definition 16 (Source normalized expected score)
*The **normalized expected score** of a source ψ_s is the conditional expectation of the normalized expected score $\bar{\mu}_i$ of objects x_i generated by ψ_s, across all datasets $\mathcal{X}$ sampled from Ψ:*

$$\zeta_s = E[\bar{\mu}_i \mid y_i = \psi_s]$$

This newly defined score satisfies that:

Proposition 17
The value of the normalized expected score ζ_s for a source ψ_s is

$$\zeta_s = \sum_{\psi_t \in \Psi} \alpha_t \cdot \zeta_{st}$$

PROOF The value of $\bar{\mu}_i$ is

$$\bar{\mu_i} = \frac{1}{n}\mu_i = \frac{1}{n}\sum_{x_j \in \mathcal{X}} \mu_{ij}$$

The conditional expectation of $\bar{\mu}_i$ across samplings of $\mathcal{X}$ for which $|\mathcal{X}| = n$ can then be found as

$$\begin{aligned} E[\bar{\mu_i} \mid y_i = \psi_s, |\mathcal{X}| = n] &= E\left[\frac{1}{n}\sum_{x_j \in \mathcal{X}} \mu_{ij} \,\middle|\, y_i = \psi_s, |\mathcal{X}| = n\right] \\ &= \frac{1}{n} E\left[\sum_{x_j \in \mathcal{X}} \mu_{ij} \,\middle|\, y_i = \psi_s, |\mathcal{X}| = n\right] \end{aligned}$$

Assuming the $x_j \in \mathcal{X}$ are independent and identically distributed, and using the law of total expectation, this can be expressed as

$$\begin{aligned} E[\bar{\mu_i} \mid y_i = \psi_s, |\mathcal{X}| = n] &= \frac{1}{n}\sum_{x_j \in \mathcal{X}} E\left[\mu_{ij} \mid y_i = \psi_s, |\mathcal{X}| = n\right] \\ &= \frac{1}{n}\sum_{x_j \in \mathcal{X}} \sum_{\psi_t \in \Psi} P(y_j = \psi_t) \cdot E\left[\mu_{ij} \mid y_i = \psi_s, y_j = \psi_t, |\mathcal{X}| = n\right] \\ &= \frac{1}{n}\sum_{x_j \in \mathcal{X}} \sum_{\psi_t \in \Psi} \alpha_t \cdot E[\mu_{ij} \mid y_i = \psi_s, y_j = \psi_t, |\mathcal{X}| = n] \\ &= \frac{1}{n}\sum_{\psi_t \in \Psi} \alpha_t \cdot E[\mu_{ij} \mid y_i = \psi_s, y_j = \psi_t, |\mathcal{X}| = n] \cdot \sum_{x_j \in \mathcal{X}} 1 \\ &= \frac{1}{n}\sum_{\psi_t \in \Psi} \alpha_t \cdot E[\mu_{ij} \mid y_i = \psi_s, y_j = \psi_t, |\mathcal{X}| = n] \cdot n \\ &= \sum_{\psi_t \in \Psi} \alpha_t \cdot E[\mu_{ij} \mid y_i = \psi_s, y_j = \psi_t, |\mathcal{X}| = n] \end{aligned}$$

Finally, assuming independence of normalized expected scores and source affinities with respect to dataset size n, and plugging the definition of the latter into the above formula, we obtain the desired result:

$$\begin{aligned} E[\bar{\mu_i} \mid y_i = \psi_s, |\mathcal{X}| = n] &= \sum_{\psi_t \in \Psi} \alpha_t \cdot E[\mu_{ij} \mid y_i = \psi_s, y_j = \psi_t, |\mathcal{X}| = n] \\ \zeta_s = E[\bar{\mu_i} \mid y_i = \psi_s] &= \sum_{\psi_t \in \Psi} \alpha_t \cdot E[\mu_{ij} \mid y_i = \psi_s, y_j = \psi_t] = \sum_{\psi_t \in \Psi} \alpha_t \cdot \zeta_{st} \end{aligned}$$

■

3.5 Consistent Clustering

We will now impose some conditions on the used clustering families, with respect to how they preserve the density and locality of the sources in Ψ. We will start by considering the *detectability* of a source by a clustering family:

Definition 18 (Source detectability)
Given a set of sources Ψ and a clustering family F, a foreground source $\psi_s \in \Psi^+$ is ***detectable by*** *F if and only if its normalized expected score ζ_s is larger than that ζ_1 of the background source ψ_1.*

Proposition 19 (Detectability criterion)
Given a set of sources Ψ and a clustering family F, a foreground source $\psi_s \in \Psi^+$ is detectable by F if and only if:

$$\alpha_s \cdot (\zeta_{ss} - \zeta_{1s}) > \alpha_1 \cdot (\zeta_{11} - \zeta_{s1}) + \sum_{\substack{\psi_t \in \Psi^+ \\ \psi_t \neq \psi_s}} \alpha_t \cdot (\zeta_{1t} - \zeta_{st})$$

PROOF From the definition of detectability and Proposition 17,

$$\begin{aligned}
\zeta_s &> \zeta_1 \\
\sum_{\psi_t \in \Psi} \alpha_t \cdot \zeta_{st} &> \sum_{\psi_t \in \Psi} \alpha_t \cdot \zeta_{1t} \\
\alpha_s \cdot \zeta_{ss} + \alpha_1 \cdot \zeta_{s1} + \sum_{\substack{\psi_t \in \Psi^+ \\ \psi_t \neq \psi_s}} \alpha_t \cdot \zeta_{st} &> \alpha_s \cdot \zeta_{1s} + \alpha_1 \cdot \zeta_{11} + \sum_{\substack{\psi_t \in \Psi^+ \\ \psi_t \neq \psi_s}} \alpha_t \cdot \zeta_{1t} \\
\alpha_s \cdot (\zeta_{ss} - \zeta_{1s}) &> \alpha_1 \cdot (\zeta_{11} - \zeta_{s1}) + \sum_{\substack{\psi_t \in \Psi^+ \\ \psi_t \neq \psi_s}} \alpha_t \cdot (\zeta_{1t} - \zeta_{st})
\end{aligned}$$

■

REMARK This arrangement of the terms in the difference $\zeta_s - \zeta_1$ is intended to capture the degree to which the clustering family captures the *density* and *locality* properties of the data in the minority clustering setting:

- For *dense* sources, self-affinity should be much larger than affinity to the background source. Therefore, the value of the left-side term should be large.
- For *local* sources, affinity to the background source and to other foreground sources should not be much different than their affinity to the background source itself. Therefore, the value of the right-side term should be small.

If a clustering family respects the density and locality of all foreground sources in a set, all of them will be detectable. In this case, the family is said to be consistent with the source set:

Definition 20 (Clustering family consistency)
*Given a set of sources Ψ, a clustering family F is **consistent with** Ψ if and only if all foreground sources $\psi_s \in \Psi^+$ are detectable by F.*

The importance of detectable sources and consistent families lies in the fact that:

Theorem 21
Given a dataset $\mathcal{X}$ sampled from a set of sources Ψ and a consistent clustering family F, for a sufficiently large number of repetitions R, the expected value of the average score s^*_{Ri} of objects x_i generated by a foreground source $\psi_s \in \Psi^+$ is larger than the expected value of the average scores s^*_{Rj} of objects x_j generated by the background source ψ_1.

PROOF Using $n = |\mathcal{X}|$, replacing the definitions of the different used quantities, and applying properties of the expectation, we know that, if ψ_s is detectable,

$$\begin{aligned}
\zeta_s &> \zeta_1 \\
n \cdot \zeta_s &> n \cdot \zeta_1 \\
n \cdot E[\bar{\mu}_i \mid y_i = \psi_s] &> n \cdot E[\bar{\mu}_j \mid y_j = \psi_1]
\end{aligned}$$

Assuming independence on the size of the dataset $\mathcal{X}$,

$$\begin{aligned}
n \cdot E[\bar{\mu}_i \mid y_i = \psi_s, |\mathcal{X}'| = n] &> n \cdot E[\bar{\mu}_j \mid y_j = \psi_1, |\mathcal{X}'| = n] \\
n \cdot E[\mu_i/n \mid y_i = \psi_s, |\mathcal{X}'| = n] &> n \cdot E[\mu_j/n \mid y_j = \psi_1, |\mathcal{X}'| = n] \\
n \cdot E[E[s^\star_{Ri} \mid y_i = \psi_s, \mathcal{X}', |\mathcal{X}'| = n]]/n &> n \cdot E[E[s^\star_{Rj} \mid y_j = \psi_1, \mathcal{X}', |\mathcal{X}'| = n]]/n \\
E[s^\star_{Ri} \mid y_i = \psi_s, \mathcal{X}', |\mathcal{X}'| = n] &> E[s^\star_{Rj} \mid y_j = \psi_1, \mathcal{X}', |\mathcal{X}'| = n]
\end{aligned}$$

which, assuming independence again, leads to

$$E[s^\star_{Ri} \mid y_i = \psi_s, \mathcal{X}] > E[s^\star_{Rj} \mid y_j = \psi_1, \mathcal{X}]$$

■

Algorithm 1 Ensemble Weak minOrity Cluster Scoring (EWOCS)

Input: A dataset $\mathcal{X}$
Input: A consistent clustering family F
Input: An ensemble size R
Output: A hard minority clustering Π of $\mathcal{X}$

1: Initialize the accumulated scores of all objects x_i to zero,

$$s_i^+ = 0$$

2: **For** $r = 1$ **to** R **do**
3: Draw a clustering function f_r at random from F,

$$f_r \in F$$

4: Apply f_r to obtain clustering Π_r,

$$\Pi_r = f_r(\mathcal{X})$$

5: Find cluster sizes,

$$\text{size}(\pi_{rc}) = \sum_{x_i \in \mathcal{X}} \text{grade}(x_i, \pi_{rc})$$

6: Update the accumulated scores of each object,

$$s_i^+ \leftarrow s_i^+ + s_{ri} = s_i^+ + \sum_{\pi_{rc} \in \Pi_r} \text{grade}(x_i, \pi_{rc}) \cdot \text{size}(\pi_{rc})$$

7: Find the final average scores of each object,

$$s_{Ri}^\star = \frac{s_i^+}{R}$$

8: Determine a threshold $s_{th}^\star$ separating the scores,

$$s_{th}^\star = \textbf{find\_threshold}(s_{R1}^\star \dots s_{Rn}^\star)$$

9: Create the foreground and background clusters, π_f and π_b,

$$\begin{aligned} \pi_f &= \{x_i \mid s_{Ri}^\star \geq s_{th}^\star\} \\ \pi_b &= \{x_i \mid s_{Ri}^\star < s_{th}^\star\} \end{aligned}$$

10: **Return** The minority clustering $\Pi = \{\pi_b, \pi_f\}$

3.6 Algorithm

A corollary of this last Theorem 21 is

Corollary 22
Given a dataset $\mathcal{X}$ sampled from a set of sources Ψ, and using a clustering family F which is consistent with Ψ, we can devise an algorithmic procedure to obtain a minority clustering of $\mathcal{X}$.

PROOF Given a dataset $\mathcal{X}$, we can apply a sequence of clustering functions f_r, drawn from F, and find the average score $s_{Ri}^\star$ for each object $x_i \in \mathcal{X}$. The expected value of the average scores of the background objects will be lower than that of the foreground ones. If a suitable threshold value is determined, we will be able to discriminate most foreground and background objects according to their score. ∎

The resulting algorithm, which we have named Ensemble Weak minOrity Cluster Scoring (EWOCS), is described in Algorithm 1.

The first step of EWOCS is the initialization of an auxiliary array, which will contain the accumulated scores s_i^+ of all objects, to zero (line 1). The main loop is then entered (lines 2–6). The

number of iterations of this loop, R, determines the ensemble size and is a user-supplied parameter. Larger values of R are expected to yield better results, but at the expense of a larger computational cost.

At each iteration, a clustering function f_r is drawn at random from family F (line 3) and is then applied to dataset $\mathcal{X}$ to obtain a clustering Π_r (line 4). The size of each cluster π_{rc} in clustering Π_r is then found (line 5), and then the score of each object, as defined in Equation 2, is found and added to the accumulated score s_i^+ (line 6).

When the main loop is over, the final average score of each object, $s_{Ri}^\star$ is found from the final accumulated score s_i^+ and the ensemble size R (line 7). From the distribution of these scores $s_{Ri}^\star$, a threshold value $s_{th}^\star$ which separates the scores of the foreground and the background objects is inferred (line 8). At this point, the only steps that remain are separating the objects according to their scores into a foreground and a background cluster (line 9) and returning the resulting clustering (line 10).

The obtained Ewocs algorithm has a number of components which allow different implementations: neither the consistent clustering function family F (line 3) nor the method for the determination of the threshold score separating foreground and background objects (line 8) are specified. As mentioned in the introduction, the following two sections, 4 and 5, give brief insights into each one of these two issues, respectively.

3.7 Clustering Model

Some algorithms are only devised to build a clustering of a input dataset, and do not provide any device to determine the hypothetical assignments of new objects to one of the obtained clusters. This is the case, for instance, of most hierarchical (including HAC) and ensemble clustering (such as Ghosh et al., 2002; Gionis et al., 2005) algorithms. However, most popular partitional methods—starting with k- and c-means, and continuing with all probabilistic mixture algorithms—provide, as a byproduct of the clustering process, a *clustering model* which may then be later used as a classification model for new data, after identifying the obtained clusters with classes.

In the case of Ewocs, if the functions in the used family F provide models together with the clusterings when applied to dataset $\mathcal{X}$, these individual models can be extended to obtain an aggregated minority clustering model.

More specifically, if the application of $f_r \in F$ to $\mathcal{X}$ produces clustering Π_r and clustering model $\mathcal{M}_r$, after Algorithm 1, an Ewocs minority clustering model $\mathcal{M}^E$ can be constructed, containing:

- the inner clustering models $\mathcal{M}_r$,
- the size of each cluster π_{rc} in the clusterings Π_r,
- and the threshold value $s_{th}^\star$ which separates foreground and background objects.

The process of classifying a new object x_x using the obtained model $\mathcal{M}$ is described in Algorithm 2. It follows the main steps of the previous Algorithm 1, but replacing the application of new clustering functions $f_r \in F$, by that of the previously obtained clustering models $\mathcal{M}_r$ (line 3). After all models have been applied, the average score of the object is found (line 5), and the object is deemed to belong to the foreground or background cluster according to whether its score exceeds the previously found threshold (line 6).

4 Weak Clustering

As stated in Section 3.5, the theoretical properties of the Ewocs algorithm depend only on the condition of the used clustering family being consistent. We believe that the requirements for being consistent, according to Definition 20, should be fairly loose—and that, hence, the Ewocs algorithm is suitable for use with weak clustering algorithms.

In this context, a clustering function family F is a clustering algorithm which includes elements of randomness. Each sequence of random values will determine a member function of the family. From a conceptual point of view, drawing a function f_r from the family F will hence correspond to drawing a sequence of random values to be later used by the algorithm. From a computational one, it can correspond, for instance, to choosing a seed value for the algorithm's internal random number generator.

Algorithm 2 Classification using an EWOCS clustering model

Input: An EWOCS minority clustering model $\mathcal{M}^E = (\{\mathcal{M}_r\}, \{\text{size}(\pi_{rc})\}, s^+_{th})$
Input: An object x_x
Output: The cluster $\pi_x \in \{\pi_b, \pi_f\}$ to which x_x would belong

1: Initialize the accumulated score of the object x_x to zero,

$$s^+_x = 0$$

2: **For** $r = 1$ **to** R **do**
3: Apply the clustering model $\mathcal{M}_r$ to obtain the grade of membership of x_x to each π_{rc}

$$(\text{grade}(x_x, \pi_{r1}) \dots \text{grade}(x_x, \pi_{rk_r})) = \mathcal{M}_r(x_x)$$

4: Update the accumulated score of the object

$$s^+_x \leftarrow s^+_x + s_{rx} = s^+_x + \sum_{\pi_{rc} \in \Pi_r} \text{grade}(x_x, \pi_{rc}) \cdot \text{size}(\pi_{rc})$$

5: Find the final average score of the object

$$s^\star_{Rz} = \frac{s^+_z}{R}$$

6: Assign the object to the foreground or background cluster, π_f or π_b, according to the relation between its average score and the separating threshold

$$\pi_z = \begin{cases} \pi_f & \text{if } s^\star_{Rz} \geq s^\star_{th} \\ \pi_b & \text{if } s^\star_{Rz} < s^\star_{th} \end{cases}$$

7: **Return** The object cluster π_z

Weak clustering algorithms include, for instance, *splitting by random hyperplanes* and *combination in random subspaces* as proposed by Topchy et al. (2003). In particular, Section 6 contains an estimation of the consistency of the former over a number of datasets. Its results shall provide an empirical assessment of the suitability of weak clustering families for use within EWOCS, and of the strength—or weakness—of the consistency requirement.

5 Threshold Determination

The last step of the EWOCS algorithm is that of determining, from the sequence of scores $s^\star_1 \dots s^\star_n$ found by the ensemble clustering process[6], a threshold value $s^\star_{th}$ which separates foreground and background objects.

One approach to this problem, appropriate to unsupervised minority clustering, was presented in our previous work on relation detection (Gonzàlez and Turmo, 2009). It uses a simple heuristic to determine the threshold value for scores generated by EWOCS, and arises from the observation of the distribution of the sorted sequence of scores of the clustered objects. An example of such distribution appears in Figure 2, for a run of EWOCS on the TOY data in Figure 1.

As observed in the figure, a small number of instances are assigned high scores whereas a large number are assigned low ones, presumably corresponding to foreground and background objects, respectively. The score sequence follows thus the shape of a decreasing convex function. This phenomenon was recurrent across most of the tested datasets.

The cutoff point should try to separate these two regions. Intuitively, this point will lie in the region of maximum convexity of the curve, and hence close to the lower left corner of the plot. This idea leads to the criterion to which we will refer as DIST, and which, as an approximate but efficient way to determine the threshold, minimizes the distance from the origin in a normalized plot of the

[6]For the sake of simplicity, we will be omitting in this section the R subindex from $s^\star_{Ri}$, as we believe there is no risk of confusion with other than the final scores.

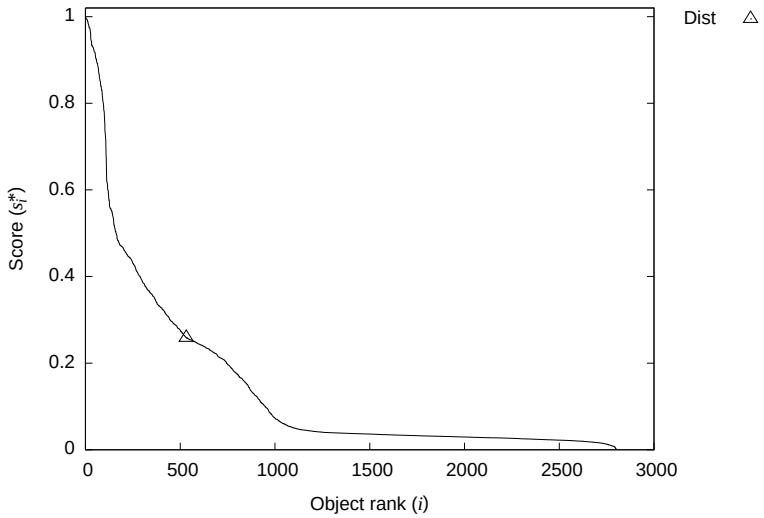

Figure 2: Accumulated score distribution (Toy data)

scores.

The first step in this criterion is hence sorting the objects $x_i \in \mathcal{X}$ by decreasing scores assigned to them by the Ewocs algorithm, so that, in the sequence $s_1^\star \dots s_n^\star$, $\forall i:\ s_i^\star \geq s_{i+1}^\star$. These scores are then linearly mapped to the range $[0 \dots 1]$, obtaining normalized versions $\bar{s}_i^\star$:

$$\bar{s}_i^\star = \frac{s_i^\star - \min s_j^\star}{\max s_j^\star - \min s_j^\star} \tag{9}$$

Then, the distance from the origin in the normalized plot is found for each object, and that at the minimum distance is selected as cutoff object x_{th}:

$$\mathbf{dist}(x_i) = \sqrt{(\bar{s}_i^\star)^2 + (i/\max i)^2} \tag{10}$$

$$x_{th} = \arg\min_{x_i \in \mathcal{X}} \mathbf{dist}(x_i) \tag{11}$$

This object is the one marked as Dist in Figure 2. Its score, $s_{th}^\star$, is the one returned as threshold value.

Even if this simple procedure was successfully used to obtain a threshold on the application of Ewocs to the task of relation detection (Gonzàlez and Turmo, 2009), we believe that the better understanding of Ewocs provided by its theoretical analysis can help in developing new methods—which may detect more accurate threshold scores. This thus remains an open line of research.

6 Evaluation

In order to estimate the degree to which the consistency requirement—imposed by the Ewocs algorithm on clustering families—is realistic, we have performed a small series of experiments on synthetic data. In particular, the consistency of one particular weak clustering algorithms has been empirically assessed.

Next sections give details about the evaluation procedure. Section 6.1 describes the used datasets, and Section 6.2 describes the evaluation protocol, including the considered metrics. Finally, Section 6.3 exposes and briefly discusses the obtained results.

6.1 Data

For our evaluation, we have prepared a number of synthetic datasets where foreground Gaussian sources are embedded within a set of uniformly distributed background objects. Several parameters, such as the number of sources, the number of foreground and background objects and the means

Number of dimensions	2, 3, 5, 8
Data range	$[-2.0 \ldots + 2.0]$
Number of background samples	5400 ... 12000
Number of foreground sources	3 ... 8
Number of foreground samples	700 ... 1800
Variance within foreground sources	0.125 ... 0.25
Minimum distance between foreground sources	0.75

Table 1: Parameter range for synthetic dataset generation

	Cons	M-Det	μ-Det		Cons	M-Det	μ-Det	
2 Dimensions	81.82	96.10	94.48	3 Dimensions	100.00	100.00	100.00	
5 Dimensions	100.00	100.00	100.00	8 Dimensions	100.00	100.00	100.00	

Table 2: Consistency of the proposed weak clustering algorithms (SYNTH data)

and variances of the Gaussian sources, were chosen at random. A summary of the ranges of these parameters can be found in Table 1.

In total, 160 such parameter settings have been generated. For each setting, 10 datasets using them were generated, and the whole 1600-dataset collection was used in the evaluation.

6.2 Protocol

For each dataset in the collection, 25 runs of the *splitting by random hyperplanes* weak clustering algorithm of Topchy et al. were performed, and the source affinities were estimated from the co-occurrence matrices of these 250 clusterings. We have then reported the fraction of datasets with which the considered methods are consistent (Cons), as well as, more precisely, the fraction of sources which are detectable by them—both macro- (M-Det) and micro-averaged (μ-Det) by dataset.

6.3 Results

Table 2 contains the values of consistency and averaged source detectability of the different weak clustering algorithms, estimated over all datasets. Given that more dimensional data will exhibit a larger degree of sparsity which may render the results not comparable with those of lower dimensional datasets, we have opted to present the results segregated by the number of dimensions in the datasets.

As seen in the table, our hypothesis that weak clustering algorithms are consistent with data generated by dense and local sources seems clearly corroborated by the empirical evidence coming from these experiments. We have found the property to hold in *all* tested datasets for 3-, 5- and 8-dimensional data. Only for 2-dimensional datasets, the algorithm fails to detect some of the sources—more specifically, a 5.52% of them. Overall, for these two methods full consistency is only achieved in over fourth fifths (81.83%) of the datasets. These results also confirm the intuition that 2-dimensional datasets, being less sparse, are harder to deal with.

However, even if perfect consistency is not achieved, the fact that, in the worst of the cases, more than 94% of the sources are detectable suggests that the consistency assumption is realistic—we are working with a very weak clustering algorithm—and also that the lack of full consistency may not necessarily hamper the actual performance of the EWOCS algorithm.

We find this results encouraging, and invite us to continue researching on minority clustering using EWOCS. In particular, we expect to be able to perform a systematic evaluation on a real clustering task briefly, using real-world data in addition to synthetic ones. And, as mentioned, our previous work (Gonzàlez and Turmo, 2009) already contains an evaluation of EWOCS on the task of relation detection, which prove the suitability of the approach for real-world problems.

7 Conclusions

In this report, we have considered the problem of minority clustering, contrasting it with regular all-in clustering. We have identified a key limitation of existing minority clustering algorithms—namely, we have seen how the approaches proposed so far for minority clustering are supervised, in the sense that they require the number and distribution of the foreground clusters, as well as the background distribution, as input.

The fact that, in supervised learning and all-in clustering tasks, combination methods have been successfully applied to obtain distribution-free learners, even from the output of weak individual algorithms, has led us to present a novel ensemble minority clustering algorithm, EWOCS, suitable for weak clustering combination.

After being used in previous work in relation detection, the EWOCS algorithms has now been formalized, and its properties have been theoretically proved under a set of weak constraints. The fact that these constraints are realistic, and that they can be fulfilled by weak clustering algorithms, has been empirically assessed using synthetic data. This confirmation open the doors to further work—in particular, to future full-fledged evaluations of EWOCS on minority clustering datasets, coming from synthetic and real-world sources.

At the light of the results, we believe that the EWOCS algorithm can be an effective method for ensemble minority clustering, and that it allows the building of competitive and unsupervised approaches to the task. It is appealing because of its simplicity, flexibility and theoretical well-foundedness, and can hence be taken into account for clustering on a diversity of domains, where unsupervised minority clustering tasks may be the rule and not the exception.

Acknowledgements

This work has been partially funded by the KNOW2 (TIN2009-14715-C04-04) project.

Appendix: Fuzzy Set Theory

As a reference for readers unfamiliar with them, this appendix contains a short definition of two key concepts of fuzzy set theory, which have been used in the report.

Definition A (Fuzzy set)
*A **fuzzy set** over an ordinary set $\mathcal{X}$ is a pair $\tilde{\mathcal{X}} = (\mathcal{X}, f_{\tilde{X}})$, where $f_{\tilde{X}} : \mathcal{X} \to [0,1]$ is the **membership function** (or **characteristic function**) of $\tilde{\mathcal{X}}$. For $x_i \in \mathcal{X}$, $f_{\tilde{X}}(x)$ expresses the **grade** of membership of x_i to $\tilde{\mathcal{X}}$, and will often be denoted as* $\mathrm{grade}(x_i, \tilde{\mathcal{X}})$

(Zadeh, 1965)

Definition B (Fuzzy c-partition)
*A **fuzzy c-partition** (or **fuzzy pseudopartition**) of an ordinary set $\mathcal{X}$ is a family of fuzzy sets $\Pi = \{\pi_1 \dots \pi_k\}$ over $\mathcal{X}$ such that*

$$\forall x \in \mathcal{X} : \sum_{\pi_c \in \Pi} f_{\pi_c}(x) = 1$$
$$\forall \pi_c \in \Pi : 0 < \sum_{x \in \mathcal{X}} f_{\pi_c}(x) < \|\mathcal{X}\|$$

(Bezdek, 1981; Klir and Yuan, 1995)

References

Shin Ando. Clustering needles in a haystack: An information theoretic analysis of minority and outlier detection. In *7th IEEE International Conference on Data Mining (ICDM)*, pages 13–22, 2007.

Shin Ando and Einoshin Suzuki. An information theoretic approach to detection of minority subsets in database. In *6th IEEE International Conference on Data Mining (ICDM)*, pages 11–20, 2006.

Suguru Arimoto. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, 18(1):14–20, 1972.

Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

Jeffrey D. Banfield and Adrian E. Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, 49(3):803–821, 1993.

Jim C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, 1981.

Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725, 2000.

Richard E. Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory*, 18(4):460–473, 1972.

Lev M. Bregman. The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7:200–217, 1967.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41:15:1–58, 2009.

Koby Crammer and Gal Chechik. A needle in a haystack: Local one-class optimization. In *21st International Conference on Machine Learning (ICML)*, pages 26–33, 2004.

Koby Crammer, Partha Pratim Talukdar, and Fernando C. Pereira. A rate-distortion one-class model and its applications to clustering. In *25th International Conference on Machine Learning (ICML)*, pages 184–191, 2008.

Rajesh N. Davé. Characterization and detection of noise in clustering. *Pattern Recognition Letters*, 12(11):657–664, 1991.

Rajesh N. Davé and Raghu Krishnapuram. Robust clustering methods: A unified view. *IEEE Transactions on Fuzzy Systems*, 5(2), 1997.

Arthur Pentland Dempster, Nan McKenzie Laird, and Donald Bruce Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Royal Statistical Society, Series B*, 39(1), 1977.

Richard Dubes and Anil Kumar Jain. Clustering methodologies in exploratory data analysis. In Marshall C. Yovits, editor, *Advances in Computers*, volume 19, pages 113–228. Elsevier, 1980.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *2nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 226–231, 1996.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *2nd European Conference on Computational Learning Theory (EuroCOLT)*, 1995.

Joydeep Ghosh, Alexander Strehl, and Srujana Merugu. A consensus framework for integrating distributed clusterings under limited knowledge sharing. In *NSF Workshop on Next Generation Data Mining*, pages 99–108, 2002.

Aristides Gionis, Heikki Mannila, and Panayiotis Tsaparas. Clustering aggregation. In *21st IEEE International Conference on Data Engineering (ICDE)*, pages 341–352, 2005.

Edgar Gonzàlez and Jordi Turmo. Unsupervised relation extraction by massive clustering. In *9th IEEE International Conference on Data Mining (ICDM)*, pages 782–787, 2009.

Régis Guillemaud and Michael Brady. Estimating the bias field of MR images. *IEEE Transactions on Medical Imaging*, 16(3):238–251, 1997.

Gunjan Gupta and Joydeep Ghosh. Robust one-class clustering using hybrid global and local search. In *22nd International Conference on Machine Learning (ICML)*, pages 273–280, 2005.

Gunjan Gupta and Joydeep Ghosh. Bregman bubble clustering: A robust, scalable framework for locating multiple, dense regions in data. In *6th IEEE International Conference on Data Mining (ICDM)*, pages 232–243, 2006.

Victoria Hodge and Jim Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22:85–126, 2004.

Anil Kumar Jain, M. Narasimha Murty, and Patrick Joseph Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, 1999.

Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley and Sons, 2005.

George J. Klir and Bo Yuan. *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Prentice Hall, 1995.

James B. MacQueen. Some methods for classification and anlysis of multivariate observations. In *5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

Mary M. Moya, Mark W. Koch, and Larry D. Hostetler. One-class classifier networks for target recognition applications. In *World Congress on Neural Networks*, pages 797–801, 1993.

David Peel and Geoffrey J. McLachlan. Robust mixture modelling using the t distribution. *Statistics and Computing*, 10:339–348, 2000.

Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alexander J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13 (7):1443–1471, 2001.

Alexander Strehl and Joydeep Ghosh. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2002.

Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Addison-Wesley, 2005.

David M.J. Tax and Robert P.W. Duin. Support vector data description. *Machine Learning*, 54 (1):45–66, 2004.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *37th Allerton Conference on Communication, Control, and Computing*, 1999.

Alexander Topchy, Anil Kumar Jain, and William Punch. Combining multiple weak clusterings. In *3rd IEEE International Conference on Data Mining (ICDM)*, pages 331–338, 2003.

Alexander Topchy, Anil Kumar Jain, and William Punch. A mixture model for clustering ensembles. In *SIAM International Conference on Data Mining (SDM)*, pages 379–390, 2004.

Alexander Topchy, Anil Kumar Jain, and William Punch. Clustering ensembles: Models of consensus and weak partitions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1866–1881, 2005.

Rui Xu and Donald C. Wunsch, II. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678, 2005.

Lotfi A. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338–353, 1965.

Tian Zhang, Raghu Ramakrishnan, and Miron Livny. BIRCH: an efficient data clustering method for very large databases. In *ACM SIGMOD International Conference on Management of Data*, pages 103–114, 1996.