Acta Cybernetica **25** (2021) 401–419.

Zero Initialized Active Learning with Spectral Clustering using Hungarian Method*

David Papp[a]

Abstract

Active learning tries to reduce the labeling cost by allowing the learning system to iteratively select the data from which it learns. In special case of active learning, the process starts from zero initialized scenario, where the labeled training dataset is empty, and therefore only unsupervised methods can be performed. In this paper a novel query strategy framework is presented for this problem, called Clustering Based Balanced Sampling Framework (CBBSF), which aims to uniformly select the initial labeled training dataset. The proposed Spectral Clustering Based Sampling (SCBS) query strategy realizes the CBBSF framework, and therefore it is applicable in the special zero initialized situation. This selection approach uses ClusterGAN (Clustering using Generative Adversarial Networks) integrated in the spectral clustering algorithm and then it selects an unlabeled instance depending on the class membership probabilities. In order to derive class membership probablities from the clustering information SCBS uses the Hungarian algorithm. Experimental evaluation was conducted on balanced and imbalanced MNIST datasets, and the results showed that SCBS outperforms the state-of-the-art zero initialized active learning query strategies in terms of accuracy.

Keywords: active learning, zero initialization, query strategy, clustering, spectral clustering, hungarian method

1 Introduction

The main goal of classification applications is to make predictions with high accuracy. A crucial part of this process is the model creation, which is based on the labeled training data (where the labels are the ground truth categories); hence the gathering of labeled data is also an important component of supervised machine learning. One can collect large amount of inexpensive unlabeled data through

*The research was supported by the ÚNKP-19-3 New National Excellence Program of the Ministry of Human Capacities. The research has been supported by the European Union, cofinanced by the European Social Fund (EFOP-3.6.2-16-2017-00013, Thematic Fundamental Research Collaborations Grounding Innovation in Informatics and Info-communications).

[a]Dept. of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Budapest, Hungary, E-mail: pappd@tmit.bme.hu, ORCID: 0000-0002-8814-2745

DOI: 10.14232/actacyb.288006

real-world applications [16], however labels for this data can be expensive [23], time-consuming or difficult to obtain. For example accurate labeling of speech utterances requies trained linguists [31], pose labelling in videos is extremely time consuming [24], annotating gene and disease mentions for biomedical information extraction usually requires PhD-level biologists [4]. Consequently, in these cases it is recommended to limit the number of labeled data that used for training, while attempting to achieve high accuracy.

Let $U = \{u_i\}, i = 1...m$ denote the total amount of (unlabeled) data available for training; the goal is to select only a subset of this data and assign labels to them, thereby creating the $L = \{l_j\}, j = 1...n$ labeled dataset. The easiest technique is to randomly select L, this method is called passive learning, or random sampling; although the resulting labeled training dataset has a large variance due to the randomness. A more sophisticated approach would be to consider the informativeness of the unlabeled data and then select the most informative ones. This approach is called active learning [20], where the learning system is allowed to iteratively select unlabeled instances and ask for their label. The key idea is that carefully picked, informative data allow the learning algorithm to perform better with less training. A decisive part of an active learning system is how it estimates the informativeness of unlabeled instances; the procedure employed for this purpose is called query strategy.

Usually, active learning query strategies assume that the selection process already started and train a classification model based on L. In special zero initialized situation, the procedure starts with empty L, and therefore only unsupervised techniques (e.g. clustering) can be used. It is often observed, especially for imbalanced or multi-class data sets, that the active learning process does not select the same number of items from each category during the query iterations. This happens because traditional query strategies do not take sample distribution into account in the resulting labeled training dataset. However, the underrepresented classes contain small number of samples, and therefore some attributes are available to the learning system with only an incomplete set of values, thus they lead to sub-optimal models. In zero initialized active learning this is a critical problem, since the process starts with empty L, so in some cases, underrepresented categories contain no samples at all, consequently, the affected attributes are entirely missing. In other words, it is important to query several training items into each category at the start of a zero initialized active learning process.

The subject of this paper is the so-called pool-based unsupervised active learning (UAL) [21], where an instance can be selected from a pool of unlabeled instances (U), while there is not enough labeled data (L) to learn. The learning setup is a multiclass classification problem with k classes, although, the selection and the predictions are based on an unsupervised solution instead of a supervised machine learning method. This paper is concerned with the beginning of the unsupervised active learning, where the number of the labeled data not only a few but zero; i.e. zero initialized unsupervised active learning. Active learner starting from the initial training set selected by appropriate methods can reach higher accuracy faster than that starting from randomly generated initial training set [10]; and therefore, the

primary objective was to select a balanced initial training set (so the goal was to get almost the same number of instances of each class).

The main contributions of this paper are (i) the Clustering Based Balanced Sampling Framework (CBBSF) for zero initialized active learning, and (ii) the Spectral Clustering Based Sampling (SCBS) query strategy that realizes CBBSF. SCBS utilizes ClusterGAN (Clustering using Generative Adversarial Networks, [15]) integrated in spectral clustering [26] process to form the clusters in zero initialized environment. After that Hungarian method [12] is employed to connect class membership probabilites to cluster membership probabilities. The rest of this paper is structured in the following way: the next section contains the relevant related work in the literature, Section 3 delineates the proposed CBBSF framework, then Section 4 presents the SCBS selection strategy, and after that the experimental evaluation is presented, finally the conclusions are summarized in the last section.

2 Related work

There are some traditional query strategy frameworks in the literature, e.g. uncertainty sampling [6], query-by-committee (QBC) [25], expected model change [2], expected error reduction [14], or density-weighted method [1]. On the other hand, there are recently proposed query strategies, like uncertainty sampling with diversity maximization [29], Balanced Active Learning (BAL) method [17], extended margin and soft balanced strategy [18], Prototype Based Active Learning (PBAC) algorithm [3] and the hybrid, Expected Difference Change (EDC) [19]. However, these approaches expect the L to be not empty, because all of them applies some kind of supervised machine learning algorithm (e.g. decision tree, random forest [22]), where L is used as training data. Hence, they are not suitable for the special zero initialized active learning (where L is empty), moreover, in this situation most of them are even unable to be executed. The field of active zero-shot learning [28] [27] [7] is partially related to this subject, where the goal is to find a small number of informative seen classes to facilitate unseen class predictions. The setting of active zero-shot learning task contains seen and unseen categories, however in this paper a different (zero initialized) starting environment is examined, where only unseen classes are available.

Unsupervised learning techniques have been successfully used to select the intial training set for active learning. One method is called centroid based selection [11] [9], where unlabeled instances closest to the cluster centroids are selected as starting dataset. In the work [11] the selection happened in one step, while the proposed approach in this paper introduces information gain between the selection of two consequtive items, and therefore it is mandatory to select the items step-by-step. Another selection type is the border based selection [9] which selects the samples with small difference between their highest and the second-highest degrees of cluster membership confidence, i.e. the ones that are around the border between clusters. The combination of center-based selection and border-based selection is called by hybrid selection. Authors of [30] selected half of the instances with

center-based, another half with border-based selection, and they achieved this by alternating between the two methods. The centroid, border and hybrid selections were implemented and compared to SCBS during the experiments. The aim of CBBSF is not only to select the initial labeled training dataset, but to uniformly select the instances among the categories to get a balanced labeled training dataset.

3 Clustering Based Balanced Sampling Framework

In this section the Clustering Based Balanced Sampling Framework (CBBSF) active learning query strategy framework is presented. The aim of CBBSF is to select the initial labeled dataset in the special zero initialized situation, where the initial labeled training dataset L is empty. This condition designates a few guidlines: (i) only an unsupervised machine learning algorithm can be used, (ii) the balance of labeled items between the classes is important, (iii) the query strategy should select an item whose class label the learning system is assured of. Satisfying these criteria, CBBSF can be used as a selection strategy for both balanced and imbalanced datasets (see Section 5). After CBBSF selects the initial L set, the active learning could proceed by using another query strategy that is more focused on optimizing the accuracy, but CBBSF could also be used as an end-to-end strategy.

A CBBSF query strategy first performs a clustering algorithm on the unlabeled dataset U, then selects an unlabeled instance to be labeled by an oracle, as can be seen in Figure 1. The selected item should maintain the balance in L; however, in order to achieve this, the class membership probabilities are required so that an item that presumably belongs to the most underrepresented class could be selected. On the other hand, class membership probabilities can not be calculated explicitly because L is empty, and thus supervised machine learning techniques can not be performed.

The clustering algorithm used in CBBSF must return a cluster membership matrix Q, see Equation 1, where q_{ij} is the probability for the i^{th} item to belong to the j^{th} cluster.

$$\begin{gathered} Q = (q_{ij}) \in \mathbb{R}^{n \times k}, \\ 0 \leq q_{ij} \leq 1, \quad \sum_{j=1}^{k} q_{ij} = 1. \end{gathered} \tag{1}$$

Let P be the class membership probability matrix, see Equation 2, where p_{ij} is the probability for the i^{th} item to belong to the j^{th} class. It is important to note that $P \neq Q$, since cluster identifiers are not related to class identifiers. As it was mentioned above, determining P is essential to sustain balance in L, and the elements of P can be derived from matrix Q with an appropriate assignment solution between clusters and classes. During the active learning process, there is no true information about the connection scheme, but this can be estimated based on only the labeled items.

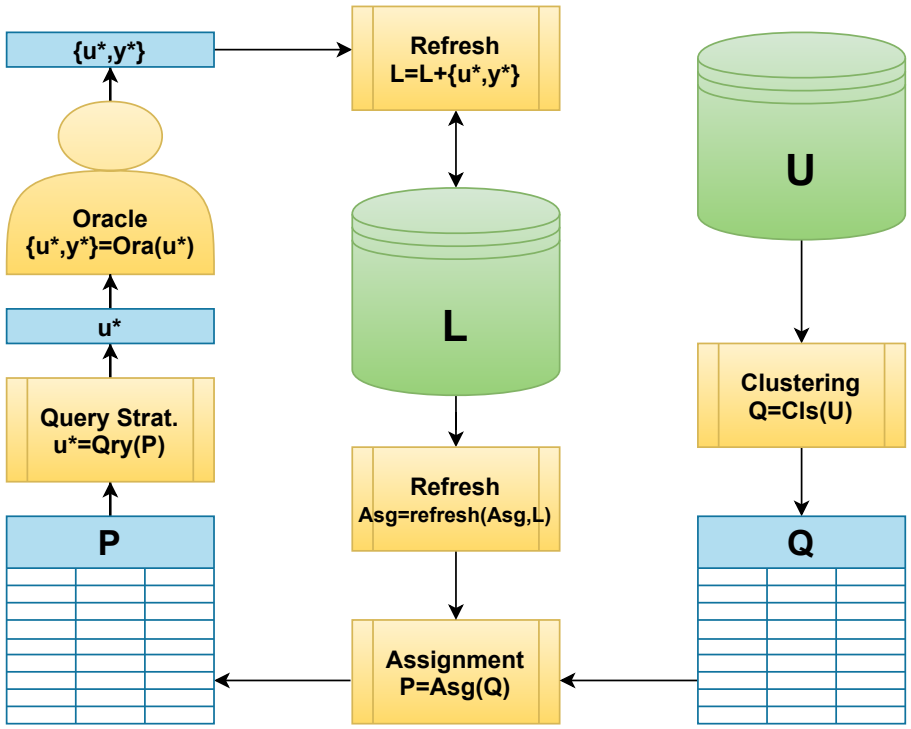

Figure 1: Process of the Clustering Based Balanced Sampling Framework

$$P = (p_{ij}) \in \mathbb{R}^{n \times k},$$
$$0 \leq p_{ij} \leq 1, \quad \sum_{j=1}^{k} p_{ij} = 1. \tag{2}$$

After P becomes available, the most informative unlabeled instance (denote it by u^*) can be selected, and then query its label y^* from an oracle (e.g. a human expert or an all knowing entity). Note that, in this case, most informative means that most likely to preserve the balance in the labeled dataset. The last step is to refresh L by adding the $\{u^*, y^*\}$ pair to it, and based on the new L refresh the assignment pattern as well. The process of CBBSF can be seen in Figure 1, where the datasets, sub-processes and matrices are represented by green, yellow and blue shapes, respectively. It is worth mentioning that U is excluded from the iterative part of this process, since the clustering algorithm is only performed at the beginning to get Q. The reason for this is that the more data is available for the clustering method to work with, the more accurately it can form the clusters.

Nevertheless, a fully iterative variant of this framework could also be used (where L influences the clustering of the remaining items in U), but in this paper such configuration is not examined.

4 Spectral Clustering Based Sampling

In this section, the Spectral Clustering Based Sampling (SCBS) active learning query strategy is presented, which belongs to CBBSF, and thus suitable to be performed in the special zero initialized environment. First, the spectral clustering [26] algorithm is briefly reviewed, and then the realization of CBBSF modules is discussed. Furthermore, Algorithm 1 shows a concise pseudocode for the CBBSF and SCBS based zero initialized active learning algorithm.

4.1 Clustering Module

Given a set of data points $x_1, ..., x_m$, pairwise similarities are calculated based on Euclidean distances, and then a similarity graph G is built to model local neighborhood relationship between the data points. Based on the constructed G graph, a similarity matrix $S = \{s_{ij}\}(i, j = 1...m)$ is derived, where s_{ij} corresponds to the weight of the edge between x_i and x_j in G (if those points are not connected by an edge in G, then $s_{ij} = 0$). Let D be a diagonal degree matrix with $D_{ii} = \sum_j s_{ij}$.

The fundamental step of spectral clustering is calculating the graph Laplacian matrix from the matrices S and D [8] For example, the unnormalized graph Laplacian matrix can be computed as expressed in Eq. 3, and this is the variant used in the SCBS algorithm. Another two popular Laplacians are the symmetric normalized and left normalized [5].

$$\Lambda = D - S \tag{3}$$

Let matrix V be defined as the matrix containing the first k eigenvectors $v_1, ..., v_k$ of Λ as columns. At this point, SCBS applies ClusterGAN [15] to form clusters $C_1, ..., C_k$. The input of ClusterGAN are the rows of V, so the spectral representation of the m datapoints. ClusterGAN is a relatively new clustering approach that performs clustering using generative adversarial networks (GAN). ClusterGAN uses a mixture of distributions (combination of discrete and continuous) to generate latent vectors and to identify different groups in the latent space (the space of latent variables). Besides, it uses a specific clustering error function to train the generator model. Once the data is transformed into latent space, they are clustered using the k-means algorithm. One advantage of using ClusterGAN is that it provides a probabilistic interpretation of the clustering. It outputs so-called cluster decision vectors $q_1, ..., q_m$ from which the cluster membership probability matrix Q can be built (i.e., $q_1, ..., q_m$ vectors are the rows of Q). This algorithm is performed on the initial unlabeled dataset U, and after that, items can be selected by the query strategy.

Algorithm 1 Zero initialized active learning with CBBSF using SCBS

input:
U: unlabeled image set
k: number of categories / number of clusters
$iter$: number of active learning iterations
initialize:
$C_1, ..., C_k \leftarrow$ Spectral ClusterGAN on U with k clusters
v_N: k-long zero vector
$L = \emptyset$
output: L initial labeled training dataset ($|L| = iter$)

for $N = 1...iter$ **do**
  if $L \neq \emptyset$ **then**
    Build the occurrence matrix A_o (Eq. 4)
    $A \leftarrow$ Hungarian algorithm based on A_o
    $\hat{P} = Q \times A$ (Eq. 5)
    $h = \text{argmax}(A[:, \text{argmin}(v_N)])$
  else
    $\hat{P} = Q$
    $h = \text{random}(1...k)$
  end if
  $bestValue = \infty$
  $bestIdx = 0$
  for $\forall u_i \in U$ **do**
    Calculate the informativeness value of $u_i \rightarrow \text{val}(u_i)$ (Eq. 7 or Eq. 8)
    if $(\text{val}(u_i) < bestValue)$ AND $(u_i \in C_h)^{\dagger}$ **then**
      $bestValue = \text{val}(u_i)$
      $bestIdx = i$
    end if
  end for
  $u^* = u_{bestIdx}$
  $y^* = \text{query}(u^*)$
  $v_N[y^*] \mathrel{+}= 1$
  $L = L \cup \{u^*, y^*\}$
  $U = U \setminus \{u^*\}$
end for

Note that L-SCBS uses the condition marked with † symbol, while G-SCBS considers only the condition before the AND operator.

4.2 Assignment Module

SCBS uses single-assignment procedure to implicitly calculate P, so that an appropriate unlabeled item can be selected that maintains even distribution in L.

The class identity and the cluster identity of the labeled items are known. This information can be structured in a table, based on which an occurrence matrix A_o is introduced, as can be seen in Equation 4, where a_{ij} is the number of the items that belong to class j while they are part of cluster i.

$$A_o = (a_{ij}) \in \mathbb{N}^{k \times k} \tag{4}$$

To find the best assignment in the matrix A_o, the Hungarian algorithm [12] is used, although in this case the sum of the entries in the assignment was maximized, instead of the minimization (as it originally happens in the Hungarian algorithm). The connection between Q and $\hat{P}$ is characterized by this best assignment A, which is actually a permutation matrix, thus Q is multiplied by A to get $\hat{P}$, where $\hat{P}$ is an approxiamtion of P; see Equation 5.

$$\hat{P} = Q \times A \tag{5}$$

4.3 Selection Module

Let $C_1, ..., C_k$ denote the k different clusters, and $Y_1, ..., Y_k$ denote the k different classes. Furthermore, introduce the vector $v_N = (N_1, ..., N_k)$ to contain the number of labeled items in the different categories, after N active learning iterations, where N_h is the number of items in Y_h $(h = 1, ..., k)$. The assignment module creates the bijection between C_g and Y_h $(g, h = 1, ..., k)$; hence the number of labeled items in the clusters are also known, at each step. Two variants of SCBS were developed: the Global SCBS (G-SCBS) and Local SCBS (L-SCBS); both of them essentially operates the same way. However, the former minimizes the informativeness metric over every element of U, while the latter examines only a reduced unlabeled set U_{C_g}, which contains the elements of a single cluster. Thus the local version of the algorithm aims to balance L directly by investigating only U_{C_g}, where C_g corresponds to the most underrepresented category in L, denoted by Y_h, as can be seen in Equation 6.

$$Y_h : h = \operatorname{argmin}(v_N) \tag{6}$$

In situations when v_N has multiple minimum values, one of them was randomly selected to designate Y_h.

In order to find the most informativeness unlabeled instance (u^*), two different techniques were used: (i) the first one maximizes the probability of the most probable class, and (ii) the second one minimizes the information entropy over all categories; as can be seen in Equation 7 and Equation 8, respectively.

$$u^* = \operatorname*{argmin}_i \left(1 - \hat{p}^*\right), \tag{7}$$

$$u^* = \operatorname*{argmin}_i \left(-\sum_{j=1}^{k} \hat{p}_{ij} \times \log \hat{p}_{ij} \right), \tag{8}$$

where $\hat{p}_{ij}$ is an element of $\hat{P}$ and $\hat{p}^*$ represents the probability of the most probable category. Despite that traditional active learning query strategies objective is usually to pick instances with maximum variance, the purpose of CBBSF is to evenly choose the instances from the classes. Consequently, SCBS must be confident that y^* ($\in \{Y_1, ..., Y_k\}$) is the true label of u^* ($\in U$), so that the assignment and the balancing could be feasible. This implies that the most representative unlabeled instance is the most informative for SCBS, i.e. the one which has minimal uncertainty about its true class label.

Table 1: Numer of items in balanced and imbalanced MNIST datasets.

	$\|Y_1\|$	$\|Y_2\|$	$\|Y_3\|$	$\|Y_4\|$	$\|Y_5\|$	$\|Y_6\|$	$\|Y_7\|$	$\|Y_8\|$	$\|Y_9\|$	$\|Y_{10}\|$	Sum
Balanced (B)	500	500	500	500	500	500	500	500	500	500	5000
Imbalanced1 (I1)	387	516	300	429	482	603	503	700	405	675	5000
Imbalanced2 (I2)	209	850	472	641	558	150	730	354	804	232	5000

5 Experimental evaluation

In this section the experiments are presented that were conducted on the MNIST [13] database of handwritten digits, which consist of 60,000 train and 10,000 test images. The train and test sets were combined into a 70K dataset, and then 5 Balanced (B), 5 Imbalanced1 (I1) and 5 Imbalanced2 (I2) subsets were randomly selected from this dataset, each of them contained 5,000 images (see Table 1). During the experiments, the following 4 SCBS method variants were tested:

- G-SCBS using minimal entropy (G-SCBS 1)
- G-SCBS using most confident (L-SCBS 1)
- L-SCBS using minimal entropy (G-SCBS 2)
- L-SCBS using most confident (L-SCBS 2)

Several additional methods proposed in the literature were also tested: the Centroid [11], the Border [30] and the Hybrid [30] active learning query strategies; furthermore, the Random sampling [9], which selects a random item at each iteration. The results of these competitor methods are compared to the results of the proposed SCBS based techniques.

The tests were performed in the special zero initialized situation, so at the start of the active learning process U contained the total 5,000 images of the test dataset and L was empty. At the testing of each dataset, the goal was to select the initial labeled image collection with a fix size: $|L| = 100$; therefore, in ideal situation each category should contain 10 labeled items. Consequently, only the first 100 active learning iterations were investigated, and in each iteration only one unlabeled instance was selected (i.e., the batch size was one).

In order to evaluate the balancedness in L, two new measures are introduced in this paper: the Average Cardinality Error (ACE, see Equation 9) and the Actual Balancedness (AB, see Equation 10). The latter expresses the amount of balance in L, at the actual active learning step. In case of perfect balance $AB = 1$, while in the worst case (when every item belongs to the same class) $AB = 0$. On the other hand, ACE can be calculated by taking the average of the deviation of actual state from the optimal one.

$$ACE = \sum_{j=1}^{k} \left(\frac{1}{k} \times \left| \left\lfloor \frac{N}{k} \right\rfloor - N_j \right| \right) \tag{9}$$

$$AB = \left(1 - \frac{1}{N} \times \left(\max_j \{N_j\} - \min_j \{N_j\} \right) \right) \tag{10}$$

where N_j is the cardinality number of class Y_j in L, and N is the number of active learning steps. After the evaluation of AB for each individual results got on MNIST datasets, the average of them were calculated, denoted by AAB, as can be seen in Table 2. Furthermore, the accuracy (ACC) was also measured at each iteration on the remaining items in U. ACC is the ratio of the correct decisions and all decisions, where the different types of decisions come from the confusion matrix: True Positive, False Positive, True Negative and False Negative. Note that since at zero initialized active learning there is not enough labeled items to perform supervised learning (i.e. classification), the predicted elements of the confusion matrix are derived from the clustering results by the assignment solution.

In Figures 2-4 the cardinality numbers (N_j) of the classes in L are presented, at iterations 20, 50 and 100, obtained on Balanced, Imbalanced1 and Imbalanced2 MNIST datasets, respectively. Figure 5 shows the average accuracy at each iterations, where SCBS methods are represented with dark, competitor methods are represented with gray lines; each strategy with different markers. The results show that L-SCBS 1 and L-SCBS 2 strategies could achieve higher accuracy than every other method, moreover, in case of balanced datasets both of them were able to perfectly balance L after 100 active learning steps. Regarding imbalanced datasets, L-SCBS 2 seems to perform slightly better, than L-SCBS 1. On the other hand, G-SCBS 1 and G-SCBS 2 could not balance $\{N_j\}$, and therefore it can be concluded that reducing U to only one cluster at a time by leveraging the assingment solution is advantageous. Competitor methods were also unable to reach equilibrium, although at balanced datasets Centroid seems to be promising, since it surpassed the global variants of SCBS. Border technique resulted the highest deviation in $\{N_j\}$, while it gave the highest accuracy on average, after 100 iterations (see Figure 5). This could be explained by analyzing the way it operates, Border method selects instances on the border of clusters, and thus it eliminates uncertain choices, which increases the accuracy. Other methods reached the same level of accuracy, L-SCBS 1 and L-SCBS 2 at around 10-11 steps, while for other approaches it took a longer time.

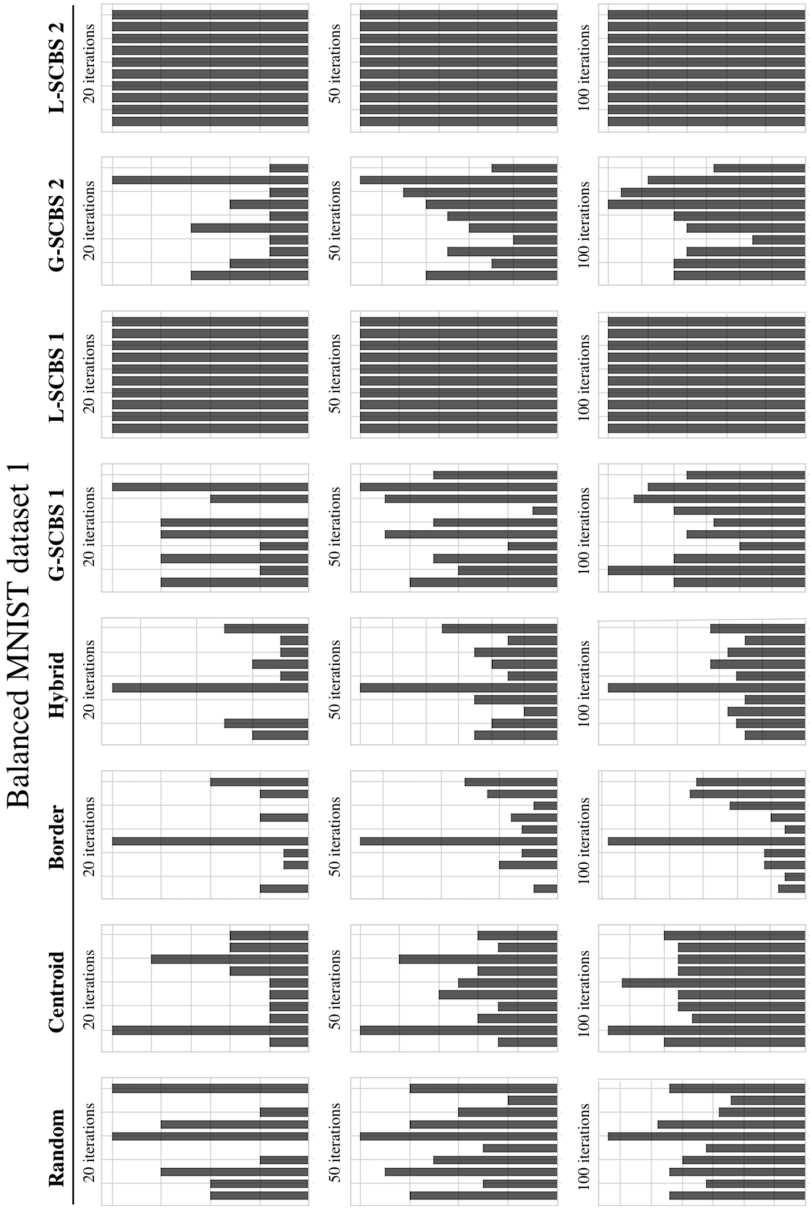

Figure 2: Distributions of the labeled instances among the categories got on one of the Balanced MNIST dataset at iterations 20, 50 and 100.

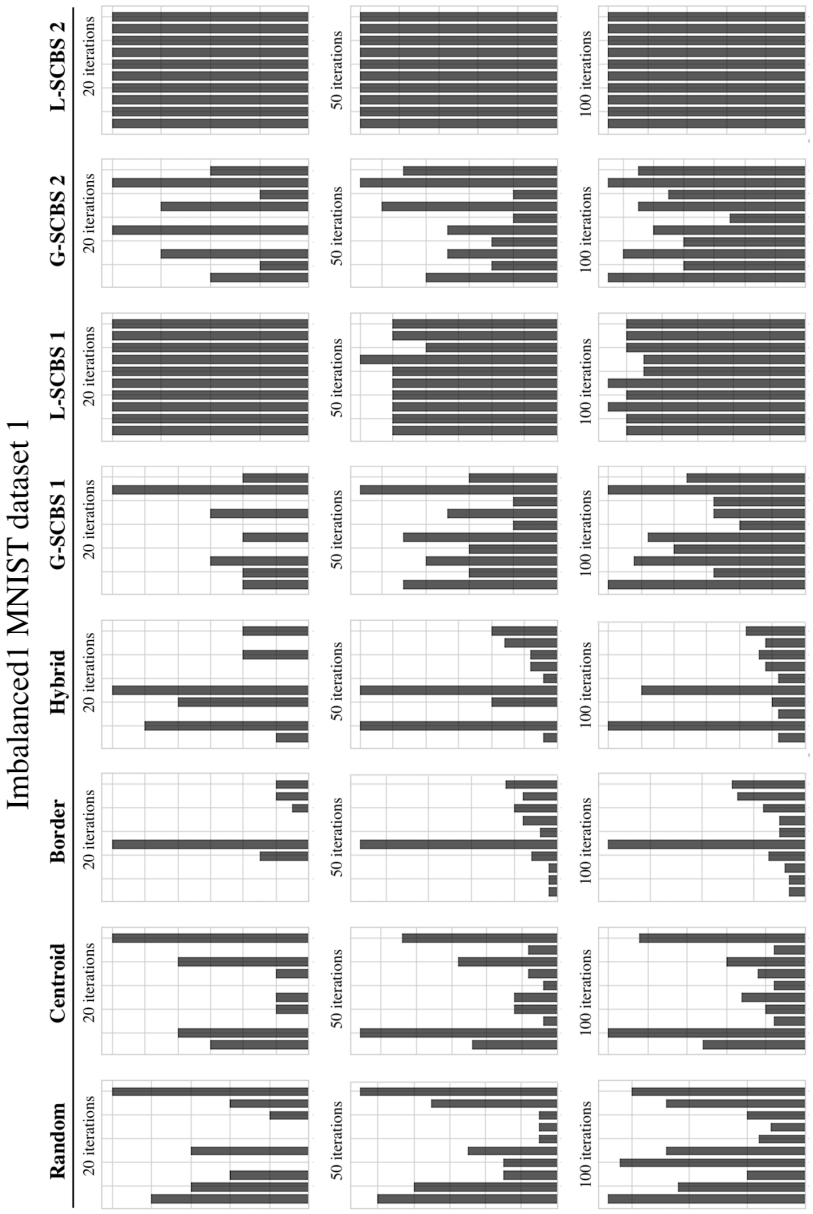

Figure 3: Distributions of the labeled instances among the categories got on one of the Imbalanced1 MNIST dataset at iterations 20, 50 and 100.

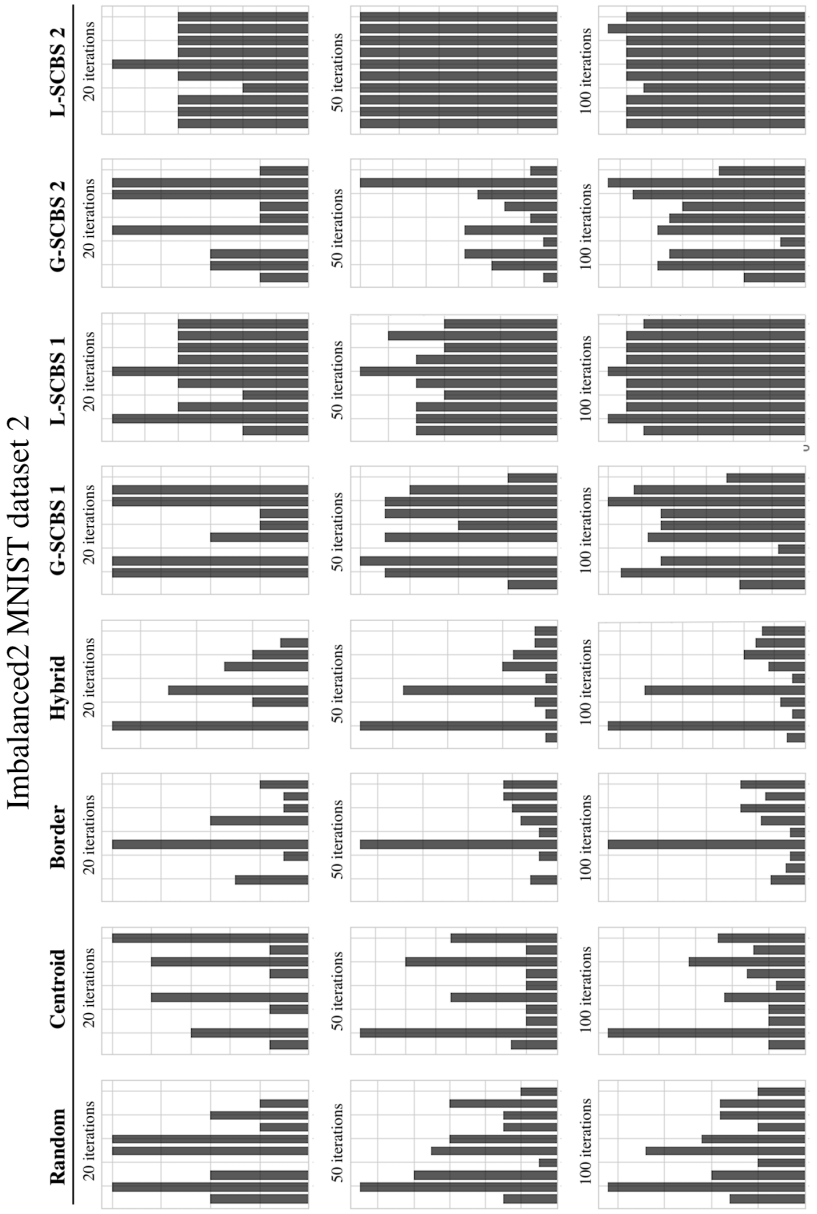

Figure 4: Distributions of the labeled instances among the categories got on one of the Imbalanced2 MNIST dataset at iterations 20, 50 and 100.

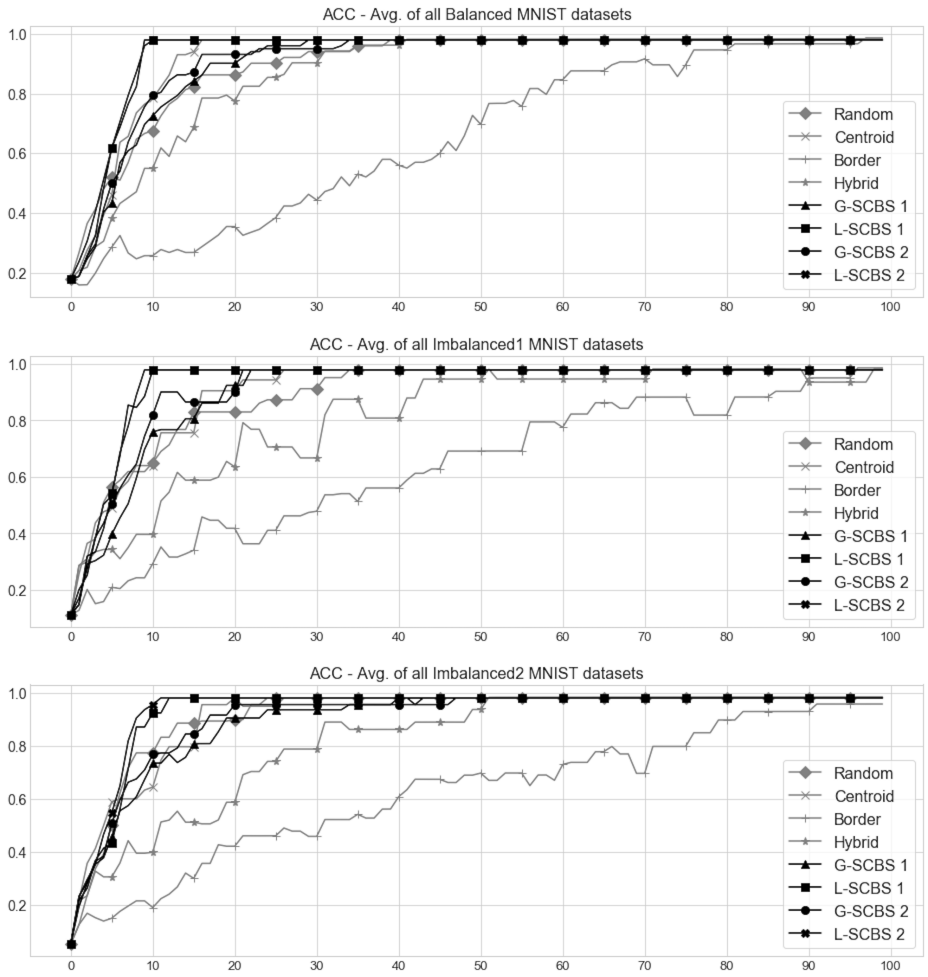

Figure 5: Average of accuracies got on the MNIST datasets at each active learning iteration; different query strategies are denoted by different markers, additionally, the darker lines correspond to the SCBS variants.

As can be seen on the figures, different datasets resulted different label distributions in L, however, taking the average of the different MNIST datasets regarding this aspect would be highly misleading and difficult to interpret. The reason for this is that the outcome of taking the average of low and high cardinality numbers could be around the perfect result, even though the difference between the individual results could be colossal. Consequently, each test were evaluated separately and deviations from the optimal cardinality number were calculated as errors. Table 2 summarizes the results, where the maximum and minimum $\{N_j\}$ are shown along

Table 2: Maximum, minimum cardinality numbers, and Average Cardinality Errors got on each MNIST dataset, additionally, last row of each block show the Average Actual Balancedness.

		Random	**Centroid**	**Border**	**Hybrid**	**G-SCBS 1**	**L-SCBS 1**	**G-SCBS 2**	**L-SCBS 2**
MNIST B 1	max	16	14	29	23	15	**10**	15	**10**
	min	6	8	3	7	5	**10**	4	**10**
	ACE	2.2	1.4	6.6	3	2	**0**	2.2	**0**
MNIST B 2	max	14	12	22	17	13	**10**	14	**10**
	min	8	8	4	6	7	**10**	6	**10**
	ACE	1.6	1	3.2	3.4	1.8	**0**	2.2	**0**
MNIST B 3	max	21	14	25	18	14	**10**	13	**10**
	min	4	7	4	7	6	**10**	7	**10**
	ACE	3.8	2	4.6	2.2	2.2	**0**	1.6	**0**
MNIST B 4	max	13	12	16	19	14	**10**	20	**10**
	min	4	9	5	5	6	**10**	5	**10**
	ACE	2.2	0.8	3.6	3.4	2	**0**	3	**0**
MNIST B 5	max	14	15	28	18	15	**10**	16	**10**
	min	1	8	5	7	6	**10**	5	**10**
	ACE	2.8	1.2	5	3	3	**0**	2.4	**0**
AAB		0.899	0.948	0.814	0.890	0.907	**1.000**	0.905	**1.000**
MNIST I1 1	max	17	25	38	30	15	11	13	**10**
	min	3	4	3	4	5	9	5	**10**
	ACE	4.6	5.8	7	7	3	0.4	2	**0**
MNIST I1 2	max	17	24	37	31	22	**11**	21	**11**
	min	5	4	2	4	5	**9**	5	**9**
	ACE	2.6	5	7.4	6.8	3.6	**0.2**	3.2	**0.2**
MNIST I1 3	max	16	27	24	31	21	11	20	**10**
	min	1	3	6	5	3	9	4	**10**
	ACE	3.6	6.2	3.4	5.6	4.2	0.4	3.8	**0**
MNIST I1 4	max	17	21	23	31	18	11	16	**10**
	min	3	5	2	3	2	9	2	**10**
	ACE	4	4.4	5.6	5.2	4	0.4	3.4	**0**
MNIST I1 5	max	15	24	21	30	15	**11**	14	**11**
	min	5	5	5	5	5	**9**	3	**9**
	ACE	2.6	4.8	3.6	4.8	3.2	**0.4**	2.4	**0.4**
AAB		0.870	0.800	0.750	0.736	0.858	0.980	0.870	**0.992**
MNIST I2 1	max	16	23	15	31	16	11	17	**10**
	min	5	4	6	5	2	9	2	**10**
	ACE	2.8	6.2	2.8	4.8	3.6	0.2	3.8	**0**
MNIST I2 2	max	21	27	40	32	15	**11**	16	**11**
	min	5	4	0	2	2	**9**	2	**9**
	ACE	3.8	5.2	7.2	7.6	3.4	0.4	3.2	**0.2**
MNIST I2 3	max	20	27	22	30	17	13	23	**10**
	min	1	4	5	3	3	9	2	**10**
	ACE	4.8	3.8	5	6.2	4	**0.6**	3.6	**0**
MNIST I2 4	max	23	23	43	31	18	11	18	**10**
	min	3	4	2	2	2	9	2	**10**
	ACE	5	5.4	8.6	7.6	5.2	0.4	5	**0**
MNIST I2 5	max	33	21	34	29	18	12	17	**10**
	min	3	4	3	4	4	9	3	**10**
	ACE	5.6	5.2	6.4	6.6	3.2	0.6	3.6	**0**
AAB		0.808	0.798	0.724	0.726	0.858	0.974	0.840	**0.996**

with the ACE for each MNIST dataset (indicated in the left column, where B, I1 and I2 refers to the type of MNIST dataset). Furthermore, the AAB measure was calculated for each query strategy, and presented in the last row of each block of Table 2. As can be seen in the table, L-SCBS 1 and L-SCBS 2 had zero deviation from the optimal distribution in all balanced cases, in addition, L-SCBS 2 could achieve perfect balance, even in imbalanced situations, while L-SCBS 1 performed marginally worse; as the values of the AAB metric shows. Therefore, L-SCBS 2 is the best method (among the tested ones) to employ for the zero initialized active learning task.

6 Conclusion

A novel active learning query strategy framework and an acitve learning query strategy that belongs to this framework were elaborated in this paper, the Clustering Based Balanced Sampling Framework (CBBSF) and the Spectral Clustering Based Sampling (SCBS), respectively. CBBSF focuses on the problem of zero initialized active learning, hence it selects the initial labeled training dataset and balances the items among the categories. The framework consists of three modules, (i) a clustering module, (ii) an assignment module and (iii) a selection module. SCBS realizes this framework, it utilizes ClusterGAN integrated in spectral clustering process to form the clusters and then Hungarian method is used during the assignment, after that it selects unlabeled items based on the class membership probabilities. Global and local variants of the SCBS method were developed, futhermore, two different techniques were applied to calculate the informativeness of the unlabeled instances, and thus four different SCBS approaches were examined. Average Cardinality Error (ACE) and Actual Balancedness (AB) new measures were introduced in the paper. During the experimental evaluation on MNIST datasets, ACE, AB and accuracy (ACC) were evaulated using each SCBS variant, moreover, state-of-the-art zero initialized active learning query strategies were also tested and compared to the results of SCBS, namely the Random, Centroid, Border and Hybrid approaches. The results showed that local versions of SCBS achieve high accuracy faster than every other method, and they are able to perfectly balance the labeled training dataset. In future work, the proposed approach will be extended with a solution that handles wrong clustering, i.e., when two categories are merged or one category is splitted. With this addition, the usability of the algorithm in real world scenarios could be improved significantly.

References

[1] B., Settles and M., Craven. An analysis of active learning strategies for sequence labeling tasks. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1069–1078, 2008. DOI: 10.3115/1613715.1613855.

[2] Cai, W., Zhang, Y., and Zhou, J. Maximizing expected model change for active learning in regression. *In: IEEE 13th International Conference on Data Mining*, pages 51–60, 2013. DOI: 10.1109/icdm.2013.104.

[3] Cebron, N. and Berthold, M. R. Active learning for object classification: from exploration to exploitation. *Data Mining and Knowledge Discovery*, 18(2):283–299, 2009. DOI: 10.1007/s10618-008-0115-0.

[4] Chowdhury, M. and Faisal, M. Disease mention recognition with specific features. In *In Proceedings of the 2010 workshop on biomedical natural language processing*, pages 83–90, 2010.

[5] Chung, F. R. and Graham, F. C. *Spectral graph theory*, volume 92. CBMS, Philadelphia, 1997. DOI: 10.1007/978-3-642-58058-1.

[6] D., Lewis and W., Gale. A sequential algorithm for training text classifiers. In *In Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12, 1994. DOI: 10.1007/978-1-4471-2099-5_1.

[7] Gavves, E., Mensink, T., Tommasi, T., Snoek, C. G., and Tuytelaars, T. Active transfer learning with zero-shot priors: Reusing past datasets for future tasks. *In Proceedings of the IEEE International Conference on Computer Vision*, pages 2731–2739, 2015. DOI: 10.1109/iccv.2015.313.

[8] HU, P. Spectral clustering survey. Technical report, The Chinese University of Hong Kong, 2012.

[9] Hu, R., Mac Namee, B., and Delany, S. J. Off to a good start: Using clustering to select the initial training set in active learning. *In Twenty-Third International Florida Artificial Intelligence Research Society Conference (FLAIRS 2010)*, pages 26–31, 2010.

[10] Kang, J., Ryu, K. R., and Kwon, H. C. Using cluster-based sampling to select initial training set for active learning in text classification. In *In Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 384–388. Springer, 2004.

[11] Kang, J., Ryu, K. R., and Kwon, H. C. Using cluster-based sampling to select initial training set for active learning in text classification. *In Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 384–388, 2004. DOI: 10.1007/978-3-540-24775-3_46.

[12] Kuhn, H. W. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1–2):83–97, 1955. DOI: 10.1002/nav.3800020109.

[13] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. DOI: 10.1109/5.726791.

[14] Mac Aodha, O., Campbell, N., Kautz, J., and Brostow, G. Hierarchical subquery evaluation for active learning on a graph. *In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 564–571, 2014. DOI: 10.1109/cvpr.2014.79.

[15] Mukherjee, S., Asnani, H., Lin, E., and Kannan, S. Clustergan: Latent space clustering in generative adversarial networks. *In Proceedings of the AAAI Conference on Artificial Intelligence*, 33:4610–4617, 2019.

[16] Panda, N., Goh, K. S., and Chang, E. Y. Active learning in very large databases. *Multimedia Tools and Applications*, 31(3):249–267, 2006. DOI: 10.1007/s11042-006-0043-1.

[17] Papp, D. and Szűcs, G. Balanced active learning method for image classification. *Acta Cybernetica*, 23(2):645–658, 2017. DOI: 10.14232/actacyb.23.2.2017.13.

[18] Papp, D. and Szűcs, G. Extended margin and soft balanced strategies in active learning. *In European Conference on Advances in Databases and Information Systems*, pages 69–81, 2018. DOI: 10.1007/978-3-319-98398-1_5.

[19] Papp, D., Szűcs, G., and Knoll, Zs. Difference based query strategies in active learning. *In Proceedings of the IEEE 17th International Symposium on Intelligent Systems and Informatics (SISY 2019)*, pages 35–39, 2019. DOI: 10.1109/sisy47553.2019.9111587.

[20] Settles, B. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

[21] Souza, V., Rossi, R. G., Batista, G. E., and Rezende, S. O. Unsupervised active learning techniques for labeling training sets: an experimental evaluation on sequential data. *Intelligent Data Analysis*, 21(5):1061–1095, 2017.

[22] Szűcs, G. Decision trees and random forest for privacy-preserving data mining. *In Research and Development in E-Business through Service-Oriented Solutions*, pages 71–90, 2013. DOI: 10.4018/978-1-4666-4181-5.ch004.

[23] Szűcs, G. and Henk, Z. Active clustering based classification for cost effective prediction in few labeled data problem. *Academy of Economic Studies. Economy Informatics*, 15(1):5–13, 2015.

[24] Szűcs, G. and Tamás, B. Body part extraction and pose estimation method in rowing videos. *Journal of computing and information technology*, 26(1):29–43, 2018. DOI: 10.20532/cit.2018.1003802.

[25] Tsai, Y.L., Tsai, R.T.H., Chueh, C.H., and Chang, S.C. Cross-domain opinion word identification with query-by-committee active learning. *In: Cheng, S.M., Day, M.Y. (eds.) TAAI 2014. LNCS*, 8916:334–343, 2014. DOI: 10.1007/978-3-319-13987-6_31.

[26] Von Luxburg, U. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007. DOI: 10.1007/s11222-007-9033-z.

[27] Xie, S. and Philip, S. Y. Active zero-shot learning: a novel approach to extreme multi-labeled classification. *International Journal of Data Science and Analytics*, 3(3):151–160, 2017. DOI: 10.1007/s41060-017-0042-5.

[28] Xie, S., Wang, S., and Yu, P. S. Active zero-shot learning. *In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 1889–1892, 2016. DOI: 10.1145/2983323.2983866.

[29] Yang, Y., Ma, Z., Nie, F., Chang, X., and Hauptmann, A.G. Multi-class active learning by uncertainty sampling with diversity maximization. *Int. J. Comput. Vis.*, 113(2):113–127, 2015. DOI: 10.1007/s11263-014-0781-x.

[30] Yuan, W., Han, Y., Guan, D., Lee, S., and Lee, Y. K. Initial training data selection for active learning. *In Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication*, page 5, 2011. DOI: 10.1145/1968613.1968619.

[31] Zhu, X. *Semi-Supervised Learning with Graphs*. PhD thesis, Carnegie Mellon University, 2005.