IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 8, Issue 5, May 2019

Multi - Class Active Learning to Classify and Map Ecological Zones across the Dune - Beach Interface

Amani M. Alattas

King Abdul-Aziz University, Computing & Information Systems College, Jeddah, Saudi Arabia

Abstract: The main goal of Active Learning (AL), is to empower the learning process by reducing the cost of obtaining the labels with a limited training samples and by selecting the most informative samples from the unlabelled set. AL was implemented to solve a wide range of problems in all fields [1]. In this paper, we will go forward to utilize a multi-class active learning model over an ecological zones dataset in order to classify the dune-beach interface to a divided sub-environments category. Pool, ranked & stream-based sampling were used as an active learning frames where three query strategies were tested with each frame to achieve the best performance. The performance metrics values would be illustrated in three different comparative statements, in term of different framework implemented by different query strategies, in term of incremental learning process behaviour in a pool-based sampling frame work by selecting 20 different query quires from the unlabelled set to present the incremental learning process behaviour with each query strategy (random, entropy & margin) and finally in term of incremental learning process behaviour above a three different AL frame works and three different query strategies in term of accuracy metric.

Keywords: Multi-classification, Active Learning, Query Strategy, Active Learning Frame Works

I. INTRODUCTION

Active Learning (AL) aims to achieve a high level of performance by a small limited labelled set as much as possible. Generally, the process of labelling (annotating) the data is the most expensive, exhausting part of any active learning process. A lot of solutions have been proposed in order to reduce the cost of obtaining the labels. Such solutions were about implementing a functions that's works by selecting the most informative label [1].
Multi-class Active Learning, is a sensitive scenario of classification by an active learning setting where the framework works by multinomial labels rather than binary labels [2]. Thus, the multi-class active learning model is affected by the diversity and performance of the queried queries while it is one of many targets. As long as it was an informative query the labelling process is useful in enhancing the model's performance [3].
Multi-class Active Learning has a wide range of applications were such models prove a high performance behaviour in address AL issues. In this paper we aim to test a multi-class active learning approach to map ecological zones across the dune beach interface using an empirical dataset. Pool, ranked & stream-based sampling were used as an active learning frame works where three different query strategies (random, entropy & margin) were implemented with random forest classifier. The paper was arranged in different sections as follows: section 1 presents all the aspects of this research, section 2 discusses the take the related work in this filed and section 3 encompasses the details of the experimental setting and the results, then the paper will be concluded with a conclusion and a section about future works.

II. RELATED WORKS

Researchers in their studies have focused to enhance the AL model performance by addressing the challenges, issues and propose some solutions then evaluate them. These issues and challenges are well discussed in previous studies [4]. On the other hand, many studies were conducted to solve some domain problems. In this section, we will view the related studies in two sub-sections namely: the multi-class active learning applications and studies of computational challenges researchers and the second sub section will cover the applications of multi-class active learning models in a wide range of domains.

- Multi-class active learning studies.

Active Learning (AL) in general as we mentioned before works to minimize the cost of obtaining labels by enhancing the quality of the labelled set [5], or enhance the process of annotating the labels [6].AL implemented was in multiple frame works, each frame-work have its own mechanism such as the stream-based sampling where the obtaining an unlabelled instance is free , so it can first be sampled from the actual unlabelled set and then the learner can decided

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 8, Issue 5, May 2019

whether or not to request its label [7], ranked [8] or the pool-based sampling [9]. Pool based sampling works by dividing the dataset into two sets where the smallest set is the training set and the larger set is the pool and the query queried in a greedy fashion. Different query strategies control the model based on its own functionality such as random [10], entropy [11] and margin [12].

- Multi-class active learning applications.

Multi-class active learning has a wide range of applications [13]. In this section we will mention some of these applications. In 2014, a multi-class active learning study have been implemented for analysis the sentiment in the financial domain [14], where a series of experiments have been proposed to prove a good affect and impact. Another interesting search was in sound classification field [15], were the researchers take in mind a goal of minimizing the need of human annotation by make an efficient combination of confidence-based active learning and set training. Ecosystem filed also take a lot of focus to implement the AL learning strategies over in order to classify and process its dataset. such studies we could find in [16-18].

III. EXPERIMENTAL SETTING AND RESTULS

This is the main part of the paper where the explanation of the experiments is illustrated in 4 sections as follows: multi-class active learning model, experimental setting & finally experimental results. Each sections of theses sub-sections are followed by a graphs, tables and an analysis.

A. Multi-class Active Learning Model.

AL model is work as integrated framework, where sub-sections work together to run the learning process. The first part we will start discussing it is the dataset $\boldsymbol{D}$ preprocessing. At this stage the model will prepare the dataset to be enrolled to the second level. The dataset $\boldsymbol{D}$ will be lunched at once to the roller and according to the active learning mechanism we will divide the $\boldsymbol{D}$ into *labelled set* $\boldsymbol{D}_L$ & *unlabelled set* $\boldsymbol{D}_{Un}$. The main goal of active learning is to minimize the time of obtaining labels, thus, the labelled set will always be smaller than the unlabelled set $| \boldsymbol{D}_L | < | \boldsymbol{D}_{Un} |$.

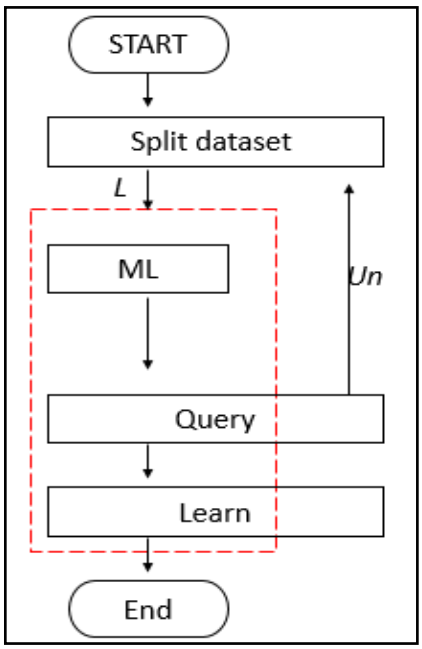


Figure 1: General AL framework

In another terms, it is noted that $\boldsymbol{D}_L$ is the *training set* and the $\boldsymbol{D}_{Un}$ is the *pool set*. Let us take a deeper look into one computational point that will surely would affect the model performance over all its stages and tasks. The smaller set affects the whole performance of the integrated model. The more the labelled set $\boldsymbol{D}_L$ was reliable and accurate, the higher the performance brought by the model.

The pre-processing stage is terminated with different output parameters $\boldsymbol{D}_L$ & $\boldsymbol{D}_{Un}$, thus, the model will now accept the inputs in order to activate the second stage.

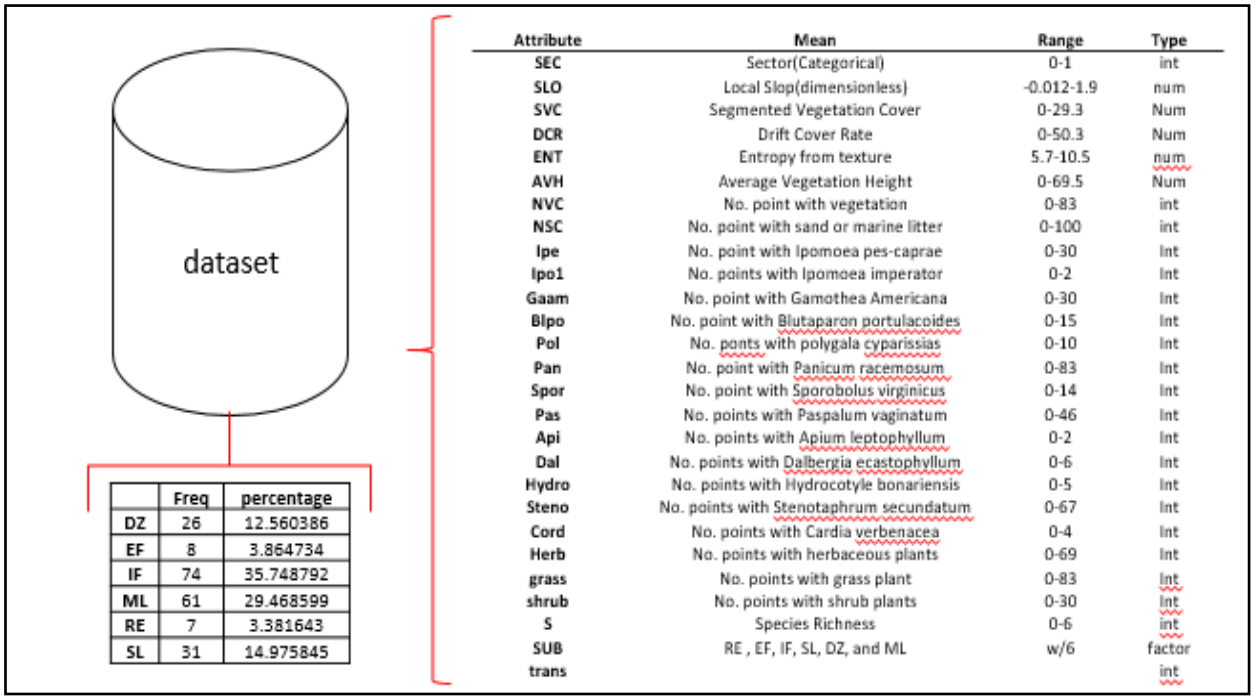
At this point the *Learner* would be activating by accepting the training labelled set $\boldsymbol{D}_L$. the *Learner* will direct $\boldsymbol{D}_L$ to the machine learning classifier $\boldsymbol{ML}$ in order to train the model over a small set of labelled instances. Every applied label instances $\boldsymbol{x_i}$ was predicted over the set of targets $\boldsymbol{t}$, which are 6 deferent targets with the following sections we will take

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 8, Issue 5, May 2019

it in details in the following sections. The model knowledge was a useful knowledge as long as these labelled data were reliable and have high accuracy. With each trained instance ***x,*** the model will learn the query results.
Currently, we are still in the learner phase but shifting between its subsection. In the previous section we discussed the $\boldsymbol{D}_L$ training part using ***ML*** to activate the classification, prediction and learning process. Next, we proceed, to the main core of active learning frame work. we will take this step with our variables which are the $\boldsymbol{D}_{Un}$ and *trained model*.
Afterwards, the larger/pool set and the unlabelled set enrolled into the model. Each different active learning frame work has its own mechanism in handling the $\boldsymbol{D}_{Un}$, however, all these different types of frame work will be fetching the most informative query. In our model, we set the query to fetch about 20 queries according the decided query strategy. When the 20 queries have been queried the results would be recorded. In case of the query met the condition it would be removed from the $\boldsymbol{D}_{Un}$ set and added to the $\boldsymbol{D}_L$ set. See figure 1

B. Multi-class Active Learning Experimental Setting.

Here we will use an empirical dataset [19] that was collected in 2011. The used data set contain 207 records each record contains 24 features. These features have a type of num , int and factor . The model will classify the labels to be predicted into 6 different targets (sub-environments) as the follows [DZ, EF, IF, ML, RE and SL]. figure 2



	Freq	percentage
DZ	26	12.560386
EF	8	3.864734
IF	74	35.748792
ML	61	29.468599
RE	7	3.381643
SL	31	14.975845

Attribute	Mean	Range	Type
SEC	Sector(Categorical)	0-1	int
SLO	Local Slop(dimensionless)	-0.012-1.9	num
SVC	Segmented Vegetation Cover	0-29.3	Num
DCR	Drift Cover Rate	0-50.3	Num
ENT	Entropy from texture	5.7-10.5	num
AVH	Average Vegetation Height	0-69.5	Num
NVC	No. point with vegetation	0-83	int
NSC	No. point with sand or marine litter	0-100	int
Ipe	No. point with Ipomoea pes-caprae	0-30	Int
Ipo1	No. points with Ipomoea imperator	0-2	Int
Gaam	No. point with Gamothea Americana	0-30	Int
Blpo	No. point with Blutaparon portulacoides	0-15	Int
Pol	No. ponts with polygala cyparissias	0-10	Int
Pan	No. point with Panicum racemosum	0-83	Int
Spor	No. point with Sporobolus virginicus	0-14	Int
Pas	No. points with Paspalum vaginatum	0-46	Int
Api	No. points with Apium leptophyllum	0-2	Int
Dal	No. points with Dalbergia ecastophyllum	0-6	Int
Hydro	No. points with Hydrocotyle bonariensis	0-5	Int
Steno	No. points with Stenotaphrum secundatum	0-67	Int
Cord	No. points with Cardia verbenacea	0-4	Int
Herb	No. points with herbaceous plants	0-69	Int
grass	No. points with grass plant	0-83	Int
shrub	No. points with shrub plants	0-30	Int
S	Species Richness	0-6	int
SUB	RE , EF, IF, SL, DZ, and ML	w/6	factor
trans			int

Another setting was set up such as the framework, machine learning and query strategies that will presented in table 1.

Figure 2 : Data set Attuributes & Targets

Table 1 : Parameters Setting .

Parameters	**Notes**
Frame Work	Pool , Stream & Ranked Based Sampling
Machine Learning	Random Forest Classifier
Query Strategy	Random , Margin & Entropy.

C. Experimental Results of Multi-class Active Learning.

In this part, we are going to illustrate the model performance results. The model performance was illustrated by more than one comparative statements, i.e. by present the performance behaviour with the three frameworks and the three query strategies with each one of the frameworks. second comparative statements are about the incremental learning process behaviour in a pool-based sampling frame work by selecting 20 different query quires from the unlabelled set to present the incremental learning process behaviour with each query strategy (random, entropy & margin) and finally in term of incremental learning process behaviour above a three different AL frame works and three different query

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 8, Issue 5, May 2019

strategies in term of accuracy metric. So we will start now by the first comparative statements by present the performance behaviour with the three frameworks and the three query strategies with each one of the frameworks in term of accuracy values.

Table 2:Accuracy Metrics over the 3 Different AL frameworks / Query Strategies.

	Random	Entropy	Margin
Pool (AL Frame work)	0.7826	0.7295	0.7971
Stream (AL Frame work)	0.7729	0.8019	0.8019
Ranked (AL Frame work)	0.7439	0.8309	0.8467

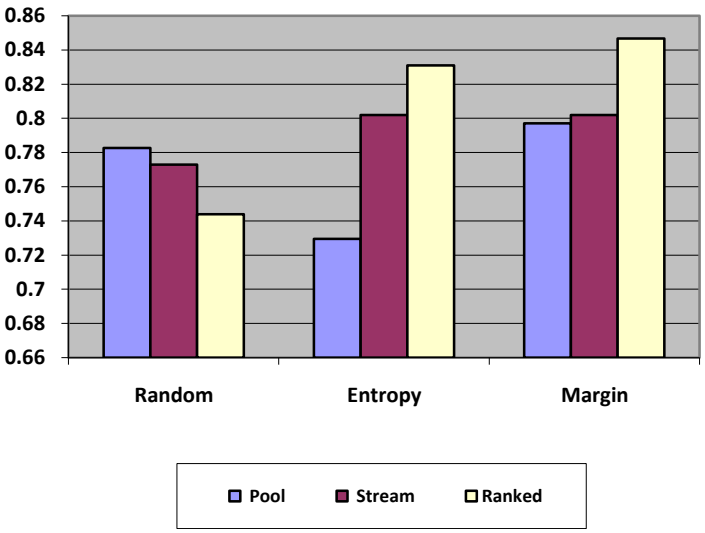


Figure 3 : The Accuracy values over the three ueriy strategy / AL frame works

From table 2, we could notice the best accuracy value was with margin query strategy by 0.8467 in case of ranked active learning frame work, where the worst values were with Entropy query strategy by 0.7295 in case of pool based sampling frame work. Generally, over all types of frame works the margin query strategy present the best accuracy level over other kind of query strategy by values in range of [0.7971-0.8467].

Afterward, in term of accuracy metrics we will present the accuracy of incremental learning process behaviour in a pool-based sampling frame work by selecting 20 different query quires from the unlabelled set to present the incremental learning process behaviour with each query strategy (random, entropy & margin). As its clear in table 3 the accuracy values changes in incremental manner in case of random query strategy the values start from 0.7778 till 0.7826 where with entropy the accuracy value start from 0.7633 till 0.7295 and finally the margin query strategy start from 0.7585 till 0.7971.At the margin query strategy the accuracy incremental learning process following smoothly by a closet steps between each iteration while random & entropy we can notice some drops in accuracy values with some iteration ,i.e. the 15th iteration in entropy .

Table 3:Accuracy Metrics over 20 queriy iterations & 3 diffent query strategies .

	Random	Entropy	Margin
1	0.7778	0.7633	0.7585
2	0.715	0.744	0.7826
3	0.7246	0.7633	0.7729
4	0.7246	0.7681	0.7681
5	0.6715	0.7729	0.744
6	0.657	0.7536	0.715

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 8, Issue 5, May 2019

7	0.6425	0.7729	0.7681
8	0.6522	0.7729	0.7729
9	0.6667	0.7488	0.7778
10	0.6763	0.7688	0.7778
11	0.7391	0.7536	0.7826
12	0.6522	0.7295	0.7874
13	0.657	0.7343	0.7874
14	0.6618	0.7536	0.7971
15	0.6425	0.6957	0.7923
16	0.657	0.7874	0.7971
17	0.7874	0.7971	0.7874
18	0.7729	0.7729	0.7874
19	0.726	0.7343	0.7971
20	0.7826	0.7295	0.7971

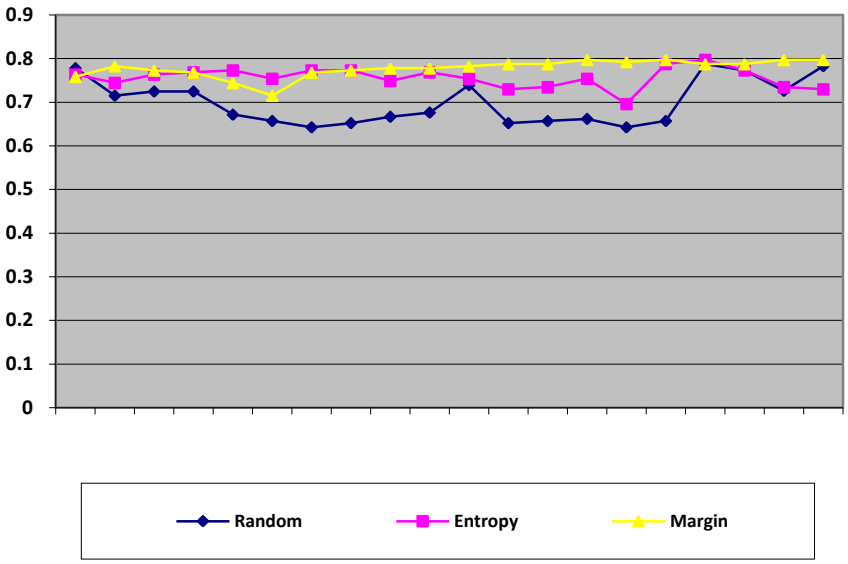


Figure 4 :The Accuracy Values over the three different query strategy

It time to scope out our comparative statements to cover more scenarios, while each one would affect model performance behaviour with the following part we will take three active learning frame work (stream, pool & ranked based sampling) over 20 different queries and 3 different query strategies. The first section of this part we will present the major accuracy values over all scenarios as its in table 4, where its clear that's the ranked based sampling frame work prove the best performance value with margin query strategy by 0.8647 and the worst case was at the pool based sampling with entropy query strategy by 0.7295.

Table 4 : Accuracy over all

Pool Based Sampling			Ranked Based Sampling			Stream Based Sampling		
Random	Entropy	Margin	Random	Entropy	Margin	Random	Entropy	Margin
0.7826	0.7295	0.7971	0.7439	0.8309	0.8647	0.7729	0.8019	0.8019

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 8, Issue 5, May 2019

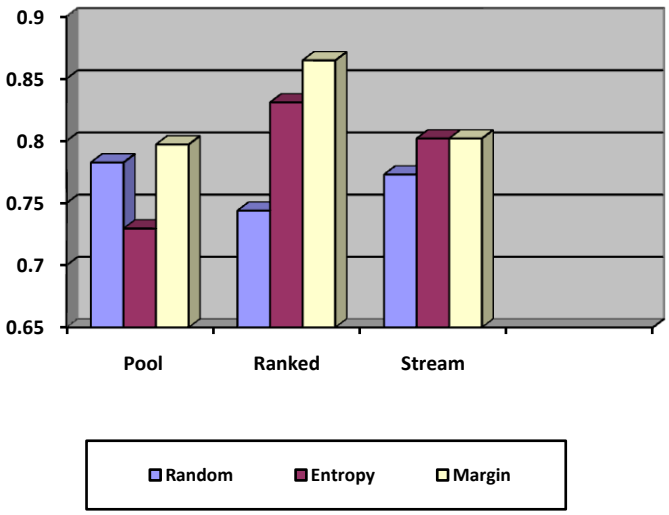


Figure 5 : Accuracy over all

Table 5: Accuracy over all with 20 quieries.

	Pool Based Sampling			Ranked Based Sampling			Stream Based Sampling		
	Random	Entropy	Margin	Random	Entropy	Margin	Random	Entropy	Margin
1	0.7778	0.7633	0.7585	0.7487	0.7487	0.7487	0.7246	0.7342	0.7246
2	0.715	0.744	0.7826	0.7101	0.7536	0.7777	0.7487	0.7536	0.7729
3	0.7246	0.7633	0.7729	0.4685	0.826	0.7246	0.7777	0.7198	0.7729
4	0.7246	0.7681	0.7681	0.4734	0.826	0.7922	0.7632	0.7681	0.7729
5	0.6715	0.7729	0.744	0.4927	0.8019	0.8019	0.7777	0.7342	0.7729
6	0.657	0.7536	0.715	0.4879	0.8212	0.7971	0.7584	0.7536	0.7681
7	0.6425	0.7729	0.7681	0.6956	0.8405	0.7971	0.7584	0.7487	0.7729
8	0.6522	0.7729	0.7729	0.7294	0.8260	0.8115	0.7536	0.541	0.7342
9	0.6667	0.7488	0.7778	0.7342	0.797	0.826	0.7632	0.7632	0.7487
10	0.6763	0.7688	0.7778	0.7584	0.8019	0.855	0.76811	0.7536	0.7632
11	0.7391	0.7536	0.7826	0.7729	0.826	0.8647	0.7342	0.7681	0.7729
12	0.6522	0.7295	0.7874	0.7391	0.8212	0.8743	0.7536	0.7632	0.7681
13	0.657	0.7343	0.7874	0.7342	0.8309	0.8647	0.7632	0.7777	0.7439
14	0.6618	0.7536	0.7971	0.7294	0.7971	0.8695	0.7584	0.7729	0.7681
15	0.6425	0.6957	0.7923	0.7294	0.8309	0.8309	0.7777	0.3816	0.7439
16	0.657	0.7874	0.7971	0.7391	0.8019	0.8792	0.7826	0.7777	0.7729
17	0.7874	0.7971	0.7874	0.7391	0.8067	0.8888	0.4589	0.7874	0.7777
18	0.7729	0.7729	0.7874	0.7391	0.8115	0.8792	0.7777	0.8115	0.7681
19	0.726	0.7343	0.7971	0.7439	0.7826	0.8888	0.7681	0.7777	0.7729
20	0.7826	0.7295	0.7971	0.7439	0.8309	0.8647	0.7729	0.8019	0.8019

DOI 10.17148/IJARCCE.2019.8501

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 8, Issue 5, May 2019

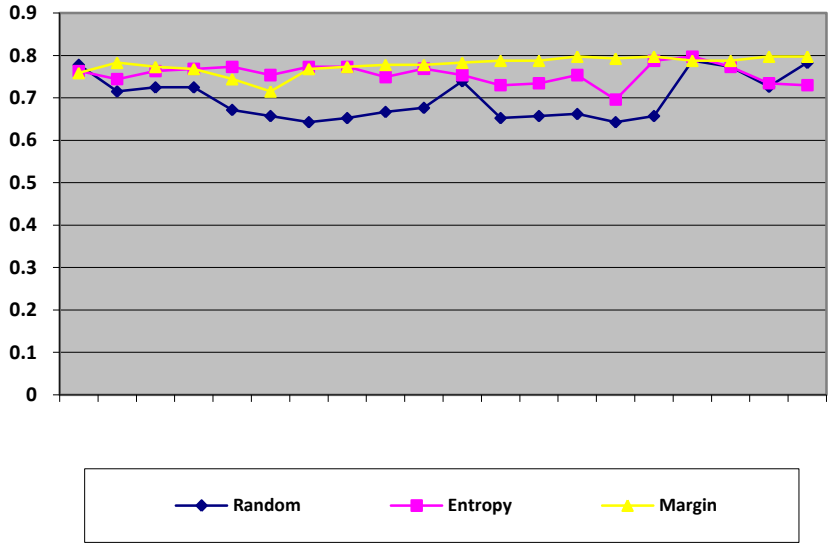


Figure 6: Accuracy with Pool Active Learning Frame work

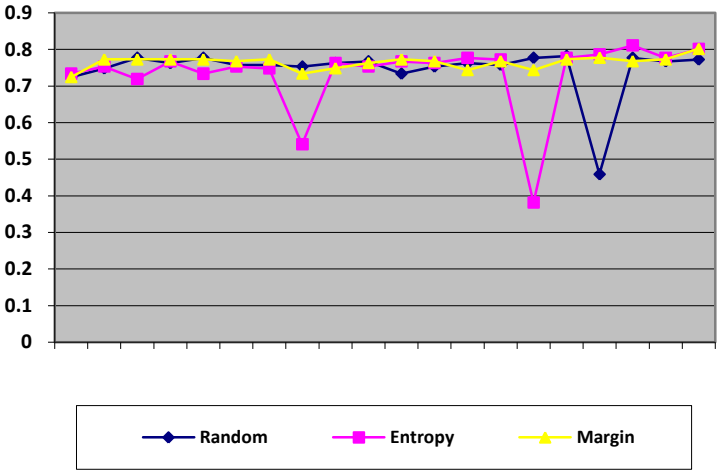


Figure 7 : Accuracy with Stream Active Learning Frame work

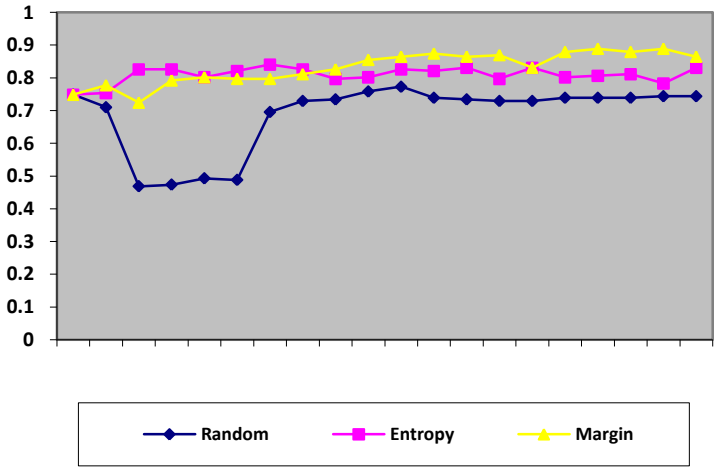


Figure 8 : Accuracy with Ranked Active Learning Frame work

IV. CONCLUSION

Selecting the most informative samples from the unlabelled set is one of the techniques that's is used in active learning strategy to empower the learning process. Active learning has a set of verify frame works i.e. (pool, stream & ranked) also it has a different types of query strategies i.e. (random, margin & entropy). With this paper we have been utilize all these setting of active learning strategy by implemented a multi-class active learning model to classify and map ecological zones across the dune-beach interfaces. Each AL framework presented a different and unique performance behaviour with such dataset. Ranked AL framework achieve the highest value by 0.8467 in term of accuracy metric, the other two frameworks (pool & stream) present the following values in sequence [0.7971-0.8041]. On another hand, by term of query strategy the margin query strategy presents the highest values in term of accuracy metric over all types of query strategy also with all types of used frameworks by values in range of [0.7971 – 0.8467].

REFERENCES

[1]. Burr Settles, " Active Learning," in *Active Learning*, Morgan & Claypool,2012, pp.
[2]. C. Aggarwal,""Data Classification Algorithms and Applications"", CRC,2015.
[3]. Haibo He; Yunqian Ma, " Class Imbalance and Active Learning," in *Imbalanced Learning: Foundations, Algorithms, and Applications*, IEEE, 2013, pp.
[4]. Masashi Sugiyama; Motoaki Kawanabe," Introduction," in *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation*, MITP ,2012, pp.
[5]. A. H. Shah , G. Gopalakrishnan , A. Rajendran and U. Liebel,"Data Mining and Sharing tool for high content screening large scale biological image data , " 2014 IEEE International Conference on Big Data(Big Data), Washington, DC,2014,pp. 1068-1076.
[6]. Liu Jun, Wang Gang, Liu Shuai and Zhou zhihua,"Labeling optimization of differential unitary space-time modulation,"*2010 2nd International Asia Conference on Informatics in Control, Automation and Robotics (CAR 2010),* Wuhan, 2010, pp.219-222.
[7]. F. Carcillo , Y,L. Borgne, O. Caelen and G. Bontempi,"An Assessment of Streaming Active Learning Strategies for Real-Life Credit Card Fraud Detection," *2017 IEEE International Conference on Data Science and Advanced Analytics(DSAA),Tokyo,2017,pp.631-639.*
[8]. M.Babaee , S. Tsoukalas,M. Babaee and M. Datcu,"Active learning using a low-rank classifier,"2015 23rd Irania Confernce on Electrical Engineering , Tehran,2015,pp.561-566.
[9]. S. Wang, J. Wang, X. Gao and X. Wang, "Pool-based active learning based on incremental decision tree" *2010 International Conference on Machine Learning and Cybernetics*, Qingdao, 2010, pp. 274-278.
[10]. T. Tsutaoka and K. Shinoda , " Acoustic model training using committee-based active and semi-supervised learning for speech recognition , " *Proceedings of The 2012 Asia Signal and Information Processing Association Annual Summit and Conference* , Hollywood , CA,2012,pp. 1-4.
[11]. S. Jie, F. Xin and S. Wen , " Active Learning for Semi-supervised Classification Based on Information Entropy , " *2009 International Forum on Information Technology and Applications* , Chnegdu , 2009 , pp, 591-595.
[12]. D. Tuia , F. Ratle , F. Pacifict , A. Pozdnoukhov , M. Kaneski , F. Del Frate , D. Soliminni , W. J. Emery ,"Active Learning of Very-high Resolution Optical Imagery with SVM : Entropy vs Margin Sampling" , IGARSS 2008-2008 IEEE International Geoscience and Remote Sensing Symposium, Boston , MA,2008 , pp. IV-73-IV-76.
[13]. Masashi Sugiyama ; Motoaki Kawanabe , " Applications of Active Learning ," in *Machine Learning in Non-Stationary Environments : Introduction to Covariate Shift Adaptation* , , MITP,2012,pp.
[14]. Jasmina Smailvoic , Miha Grcar , Nada Lavrac , Martin Znidarsic , "Stream based active learning for sentiment analysis in the financial domain " , in *information Science journal 2014* , pp.181-203.
[15]. Z. Shuyang , T. Heittola and T.Virtanen , " An active Learning Method Using Clustering and Committee-Based Sample Selection for Sound Event Classification , " *2018 16th International Workshop on Acoustic Signal*
[16]. *Enhancement (IWAENC),* Tokyo , 2018 , pp. 116-120..
[17]. Pablo D.B Guilherem , Carlos A. Borzone , Andre A. Padial , Linda R. Harris , " A semi-automated approach to classify and map ecological zones across the dune-beach interface", In *Estuarine ,Coastal and Shelf Science 2018* , pp. 61-69.
[18]. F. B. J. R. Dallaqua , F. A. Faria and A. L. Fazenda , " Active Learning Approaches for Defrosted Area Classification , " *2018 3 1st SIBGRAPI Conference on Graphics , Patterns and Images (SIBGRAPI),* Parana , 2018, pp. 48-55.
[19]. C. Persello , M. Dapllonte . T. Gobakken and E. Naesset, " Optimizing the ground sample collection with cost-sensitive active learning of tree species classification using hyperspectral images, " *2013 IEEE International Geoscience and Remote Sensing Symposium* – IGARSS, Melbourne, VIC, 2013, pp.2091-2094.
[20]. Beach_sub-environments, https://doi.org/10.5281/zendodo.1042519