Active Learning as a Way of Increasing Accuracy

Hamza Osman Ilhan and Mehmet Fatih Amasyalı

***Abstract*—In machine-learning areas, number of the data for training process alters the success of models. More samples in training give more success. However obtaining data with label information is costly and long-lasting process. Active learning algorithms are emerged to overcome this problem. It can be used with any machine learning algorithms. Active learning algorithms try to maintain same success resulted by regular machine learning methods with fewer samples. In this study, a modified active learning algorithm tested on six datasets with different machine learning methods. Comparative results presented with charts in result. Algorithm are not only providing same success but also slightly increasing total success with smarter training process.**

***Index Terms*—Active learning, random forest, single vector machines, k-nearest neighbor, naïve bayes, machine learning.**

I. Introduction

Data classification is emerged because of the necessity for getting more meaningful information derived from datasets. Classification process is the main step in all machine-learning methods. Different machine-learning methods come out to literature based on algorithms used in classification process. Many algorithms for classification are presented in literature [1]. The key idea of the classification is using some samples from dataset to form a classification model and make probabilistic calculations by using model on rest of the data or some separated data which refers to test set and validation set respectively [2]. Classification success will arise with the quality of the data used in forming model. This process is named as training process of the model. Not only the quality, but also the sample size selected in training step plays important role on the results. More selected samples in training generate more accurate classification models hence, logically; the result success should be increased as well [3]. This theory can't be generalized over all datasets but mostly data set classification models give more successful separation with more training samples.

The requirement for more samples using in training process reveals some deficiencies in daily life. Increasing number of samples necessity requires samples' label information at the same time. However gathering label information is not an easy way. Most datasets have no label information or wrong (untrusted) label information. Acquiring process of the labels is costly and time-consuming process. Studies over the machine learning methods turned into another new area called active learning to overcome this problem [4].

Active learning methods aim to make logical queries to select more informative samples for training of the classification models. Many active learning methods are available in literature [4]. In general term, Active learning process depends on the probability values of the samples related classes. Queries are formed to select more informative samples based on probability values. Different active learning algorithms make differ queries used in informative sample selection.

In the scope of this paper, some current active learning algorithms using over some classification methods in literature are briefly explained in Section II. Another active sample selection algorithm that is used in this study and differences from cited active learning algorithms in Section II are presented in Section III in details. Section IV includes used datasets information and process steps of application. Figures and tables about the results are given in Section V with explanations. Conclusion and the future works will be mentioned in Section VI.

II. Related Works

Success rate of a classifier is directly related with the sample selection step for training process. It is a general phrase in literature that more samples in training gives more successful classification result [3]. But studies also focus on cost and time effect of classification process [4], [5]. More samples in training require label information of the corresponding data as well. As it is mentioned in [4], obtaining process of samples is a time consuming and costly process. Moreover, consistency and reliability for label information of the datasets is another research area that has to be done in computer science before using all label information.

Some studies have done in literature to overcome mentioned problems [5]-[7]. These studies prove that models can be formed with more quality and logically selection of few samples instead of more samples without any connection. Logically selected samples, in other words, more informative samples are extracted by the algorithms which named as queries in literature [4]-[7]. Process of using queries in machine learning at sample selection step called as active learning or query learning process [4]. Different queries reveal different active learning algorithms. All studies aim to define an optimal sample set using in training for the best classification success of the model.

The simplest and most common algorithm in literature is uncertainty-sampling algorithm [6]. Algorithm generates query from the state of the samples in test set depends on the response of the first model trained by minimal size of randomly selected samples. If the sample in test set is already classified into any label by the model, uncertainty-sampling algorithm ignores selecting it for creating new training model in next iteration. In an adverse situation of the sample, in

Manuscript received October 28, 2013; revised March 19, 2014. This work was supported in part by the Turkey, Yildiz Technical University.

The authors are with Yıldız Technical University, Istanbul, 34220, TR (e-mail: hoilhan@yildiz.edu.tr, mfatih@ce.yildiz.edu.tr).

other words sample has many probabilities for labels, algorithm defines the sample as uncertainty. Uncertainty samples are used for next training process to form new classifier model. Uncertainty-sampling algorithm is powerful on datasets having two-classes. Query loss its efficiency on datasets with great number of classes because of the different probability values of samples over labels. More classes results more probability values belonging to samples for different classes, however algorithm works on the strictly separation of the probabilities which can't be defined in many classes structure as strictly. In order to overcome the problem, another active learning algorithm is studied in literature named as margin sampling [7]. Algorithm takes most informative two possibilities of samples and query for their label information. New model trained with the new selected samples by margin sampling algorithm. Margin sample algorithm is also not efficient on multi-class and high dimensional datasets because it is also ignores the distribution of the probabilities belonging to samples similar to uncertainty sampling algorithm. Entropy based algorithms emerged to involve the effect of probabilities' distribution into the sample selection step [7].

In this study, another algorithm is presented based on margin sample algorithm. It provides to involve probabilities distribution effects on margin sampling algorithm instead of using entropy based algorithms. In this study, not only the closest two probability values of samples selected like regular margin sampling algorithm, but also all probabilities distribution is merged in algorithm. Details of the algorithm will be given in next section.

III. Methodology

This study aims to show effects of using an active sample selection algorithm on training step of some learning methods. In this sense, learning methods with random sample selection (RSS) and active sample selection by the query are compared. Active sample selection is also named as informative sample selection in some other studies [4], [5]. A modified active learning algorithm is used within the study. Margin sampling algorithm is already being used in literature [7], but as it is cited, algorithm does not use all probabilities of the samples. All probabilities of the samples are taken into account in algorithm presented in this study, in contrast to margin sampling algorithm. Algorithm will be referred as margin distance sampling (MDS) within this paper.

MDS algorithm forms a query which interrogates for the minimum differences between all probabilities of the samples belonging to potential classes after first iteration of the model on test or validation set. Selection probability (*SP*) of a sample is calculated by (1) where *r* represents the class, *k* is the sample number and $P(x_k)^r$ is the probability of *k*th sample *x* for class *r*. *N* and *T* nominated as total number of class and sample, respectively.

$$r = 1,2 \ldots N \quad k = 1,2 \ldots T \qquad SP(x_k) = \max(P(x_k)^r) - \min(P(x_k)^r) \qquad (1)$$

Algorithm's main steps can be summarized as in Fig. 1. *E* represents classification algorithms, which are bunch of machine and ensemble learning algorithms. Labeled data set is abbreviated as *L*. *U* is donated as data pool.

- Generate first train set with 1% of all data randomly selected

Repeat n times the following steps :

- Generate classification model

C = Classification Model (E, L)

- $\forall x_k \in$ U calculate Selection Probabilities (SP) for each sample

$SP(x_k)$ k = 1, 2, … T

SP = sort (SP) (descending)

- S = Select first 1% sample from SP
- Obtain labels of S, from U
- Delete S from U and add to L

Fig. 1. Margin distance sampling algorithm.

IV. Application

Some machine and ensemble learning methods with active and regular random sample selection is separately tested in MATLAB programing interface [8]. Datasets information and used classification algorithms in tests will be explained briefly in this section.

A. Data Sets

Four data sets from UC Irvine Machine Learning Repository [9] are selected to demonstrate the effect of using active sample selection algorithm. Data sets with the relatively class information and sample sizes are listed in Table I. Datasets are selected regarding to their size and dimensional properties. Generally, active learning process's effect can be seen well on large-scale datasets, in other words, large sample size or many attributes included datasets. Furthermore, large dimensional datasets are also selected to present the success rate of MDS algorithm on multi-class datasets.

TABLE I: Datasets Information Used in Application

Name	Number of Instances	Number of Attributes	Number of Class
d159	7182	33	2
letter	20000	16	26
waveform	5000	41	3
ringnorm	7400	20	2

B. Classification Methods

Classification algorithms are chosen from literature based on the success of the models with regular RSS. Models classification success varies on the datasets because of the distribution of the data and structure of the algorithms.

Single Vector Machines (SVM) is selected for testing MDS effects on a model which already gives successful classification result for the datasets in literature [10]. This study proves MDS improves the results by the means of active learning.

Algorithm is also applied on an ensemble learning method; Random Forest. In ensemble methods, many classifiers are formed and final decision is made upon all classifiers results in the manner of some algorithms. Different algorithms reveal different ensemble methods such as Bagging, Adaboost and Random Forest. RF results adequate classification as much as SVM [11].

K-Nearest Neighborhood (KNN) is used in many classification processes in literature because of easy to implement on problems [12]. However, KNN is not successful as much as SVM or RF.MDS algorithm is also tested on KNN classifier to present the impact of active sample selection on regular KNN. K values are fixed with the

number of the total class information related to datasets.

Naïve Bayes (NB) is simple algorithm based on Bayes theorem. Generally, Naïve Bayes gives successful result on independent data sets [13]. Dependence on data sets increases with more features. Hence, Naïve Bayes results inefficient success on high dimension and large sample size datasets. Moreover, not only the size of datasets but also noises play effective role on success. Naïve Bayes impose by noises more than other methods in this study. MDSis also applied on Naïve Bayes classification to enhance the success of regular Naïve Bayes in this study.

C. Implementation

Dataset divided into three groups with 3-fold cross validation: Train, Test and Validation Set. Validation set is arranged with a fixed number of samples, however train and test set is combined into one set, which named as data pool. Then new train set is arranged with 1% of data pool with randomly selection as initial point to train first classifiers. First classifier is used for both sample selection strategies: random and active sample selection. Rest of the data pool considered as test set. Train set samples are used for training classifier while samples in validation set are selected for testing classifier success. At the same time, test set is employed for getting the selection probabilities (SP) of the samples that will be used in active learning process. Probability values of the samples belong to each classes resulted by model on test set and success rates of classification concluded by same model for validation set registered separately. At the next iteration, another 1% of data pool is selected with defined relevant sample selection algorithm: In regular sample selection, this part selected randomly from train set which referred as RSS within the paper. In contrast to regular method, samples are selected with margin distance sampling (MDS) algorithm to form new classifier model in active learning. Following to each iteration, training sample size is increased by 1% of all data pool based on probability values in active sample selection and randomin RSS respectively.

V. Experimental Results

MDS algorithm's effect tested on four datasets with four classification methods. Results will be given with figures depending on separately each individual classification methods. Furthermore tables are listed with the most successful results for datasets within all tests in the end of this section.

SVM gives full success classification rate (100%) with RSS on D159 dataset in this study as it seems on Fig. 2a. However, this success rate is reached with more samples in training. Using MDS algorithm on SVM provides the same success with few samples.

RF gives satisfactory result with the long term of training process. Classification results get higher with the size of the randomly selected train samples. Unlike the random selection, same success can be provided at the earlier step of training process with MDS algorithm similar to SVM. It can be seen clearly on Fig. 2b.

The effect of the informative sample selection obviously can be seen on Fig. 2c. NB is originally not successful method on large-scale, less class datasets such as D159 dataset because of the high samples' feature dependencies to each other. Classification results is around 80% in NB, while SVM and RF over than 95%.Moreover, noises disrupt the results more than other methods in NB. Other classification methods provide much better classify on the dataset. However, success rate of the early steps in training with the more informative independent samples selection almost the same with other methods. Furthermore, it is more than KNN result. But noise effect can be seen with incremental sample size used in training. Classifications are affected by noise data. Results stayed in low rates with all samples used in training.

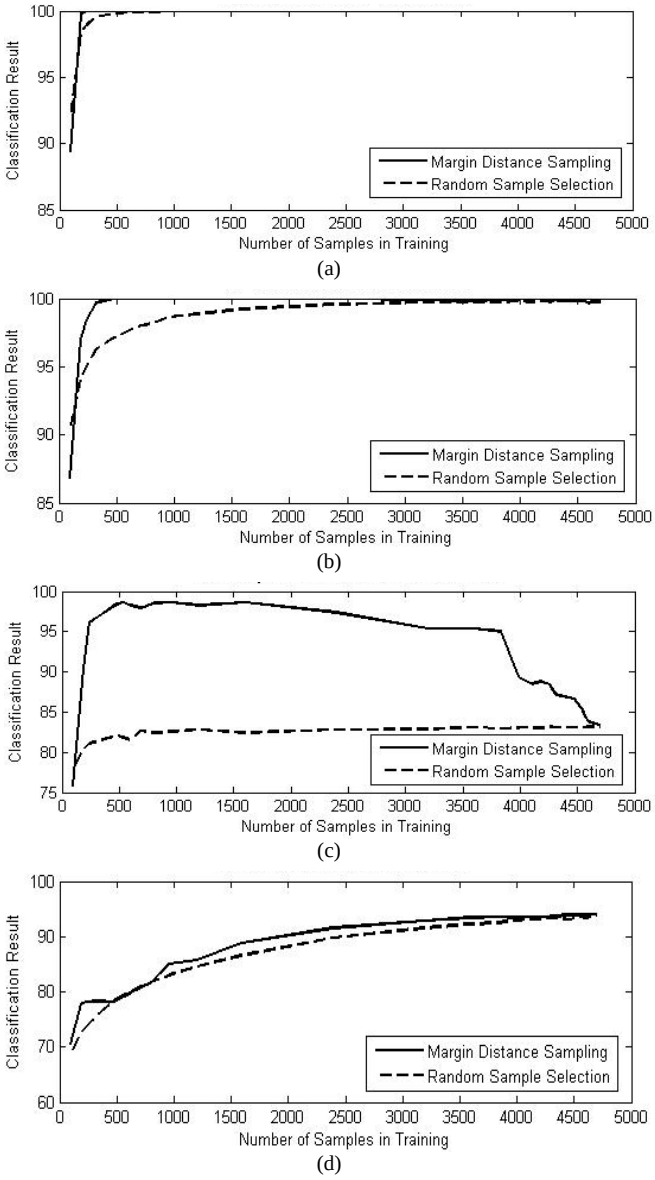

Fig. 2. SVM (a), RF (b), NB (c) and KNN (d) classification result for D159 Dataset with RSS and MDS algorithm.

The last test performed on KNN classification method. Generally, KNN works around a *K* value. In this study, K defined with a fixed number depending on the related dataset total class number. MDS algorithm gave fairly successful result over regular method as it is shown in Fig. 2d.

Another dataset is Letter from UCI. Letter dataset containsthe largest sample size within all datasets in the tests. Consequently, active sample selection is a necessity for

large-scale datasets to reduce time-consuming effects of the classification process. Classification process can be done in the earlier steps with the help of active learning algorithms. By the purpose of this, MDS is also applied on Letter dataset. Results obtained by SVM and RF methods shows that MDS algorithm provides better classification with few samples as it is demonstrated in Fig. 3a and 3b.

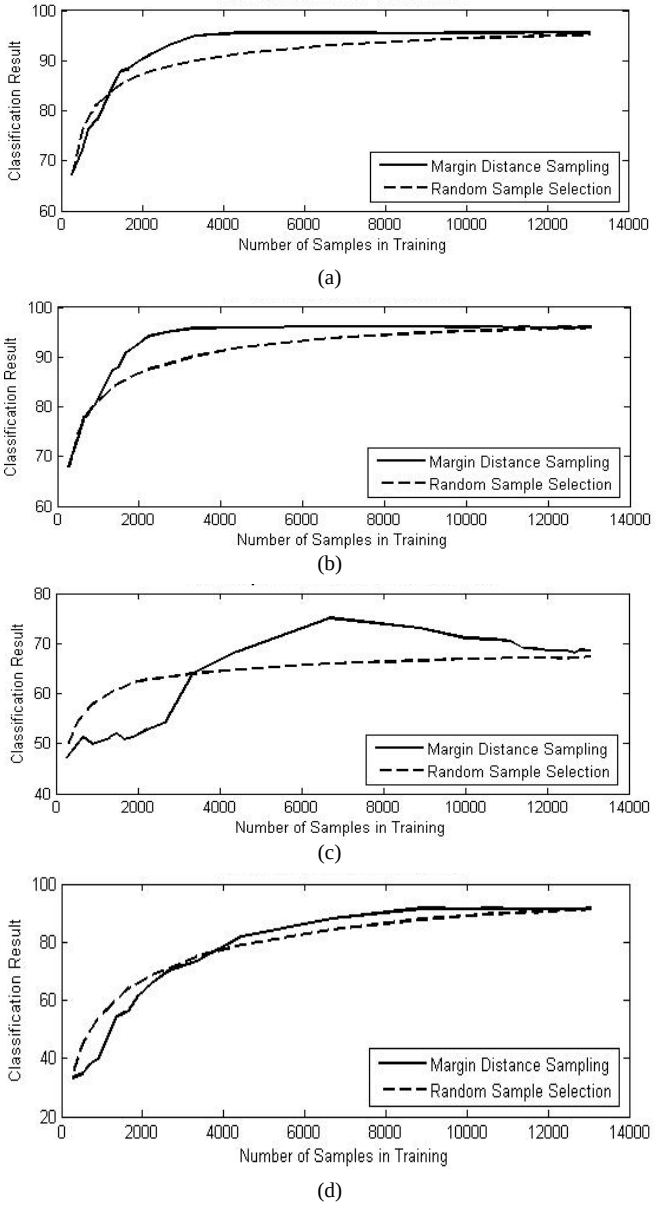

Fig. 3. SVM (a), RF (b), NB (c) and KNN (d) classification result for Letter Dataset with RSS and MDS algorithm.

Another example starts with less success for MDS algorithm when compare to regular method RSS on NB and KNN classification, but it increasesatfuture iterations with the more logically selection by algorithm.

NB method results well on strictly separated samples.MDS algorithm maintains a strict division at a point where the most informative samples selected. Fig. 3c marks that point with the highest success rate.

The same as the previous two datasets classification results, SVM and RF also classified Rignorm dataset in an efficient way. Moreover, MDS algorithm also changed the result in more successful classification when compare to RSS as in Fig. 4a and 4b. This improvement can be seen more brightly on RF model classification. Few samples selected by active learning algorithm in trainingcaused more successful classification than classification trained by full samples. Process should be stop when the success rate moves in decrement. Time consuming process can be prevented by that manner because there is no meaning to use all samples. This study confirms to stop iteration in a point instead of using all labeled data in training. Same conclusion also can be derived from Fig. 4d. KNN classification with MDS algorithm shows more successful classification with the fewer samples which is more informative about the dataset since incremental sample size makes classification success low because of non-meaning samples usage in training.

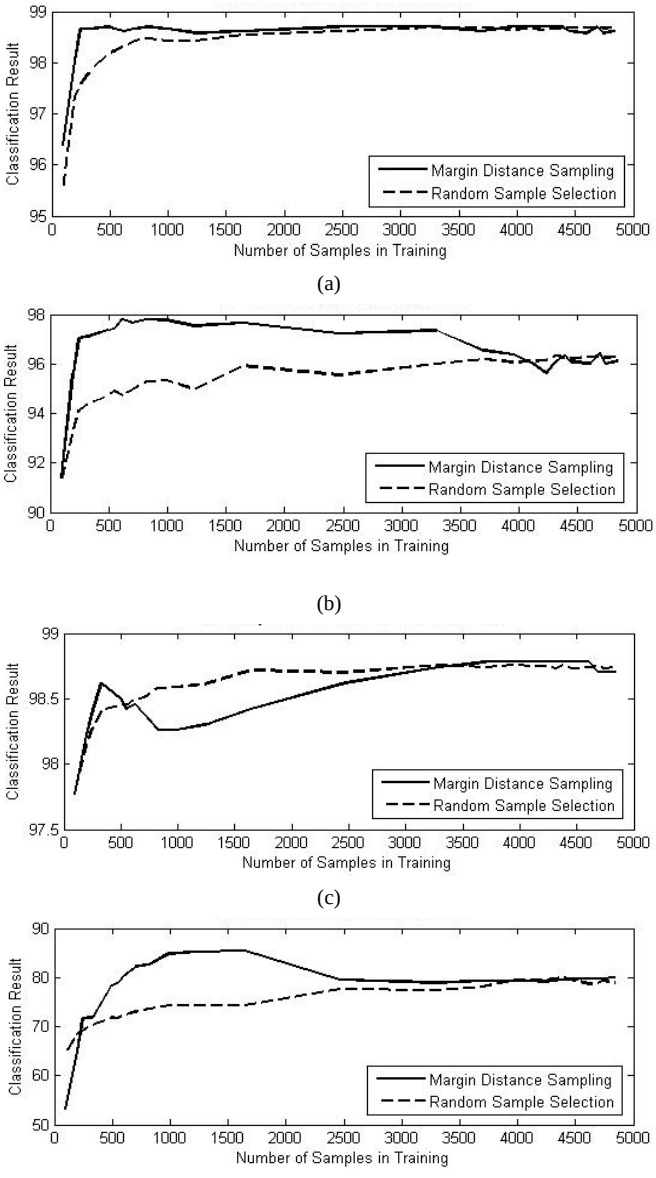

Fig. 4. SVM (a), RF (b), NB (c) and KNN (d) classification result for Rignorm Dataset with RSS and MDS algorithm.

Waveform dataset distribution is moreunstable than other datasets presented in this paper. SVM and RF total success on Waveform is not as much as other datasets as well. Anyway, active sample selection also plays important role on success of the model (Fig. 5a and 5b).

NB as mentioned before decides based on the dependency of the features and noise effects. Therefore, it is normal to see

peak changes in the results as in Fig. 5c and also other figurescorresponding datasets. Informative selected samples can increase related feature dependency with others. In other meaning, using high sample size in training can diminish classification results of NB model which can be seen on all figures in this study. Not only dependencies but also noises affect the results depending on incremental sample size.

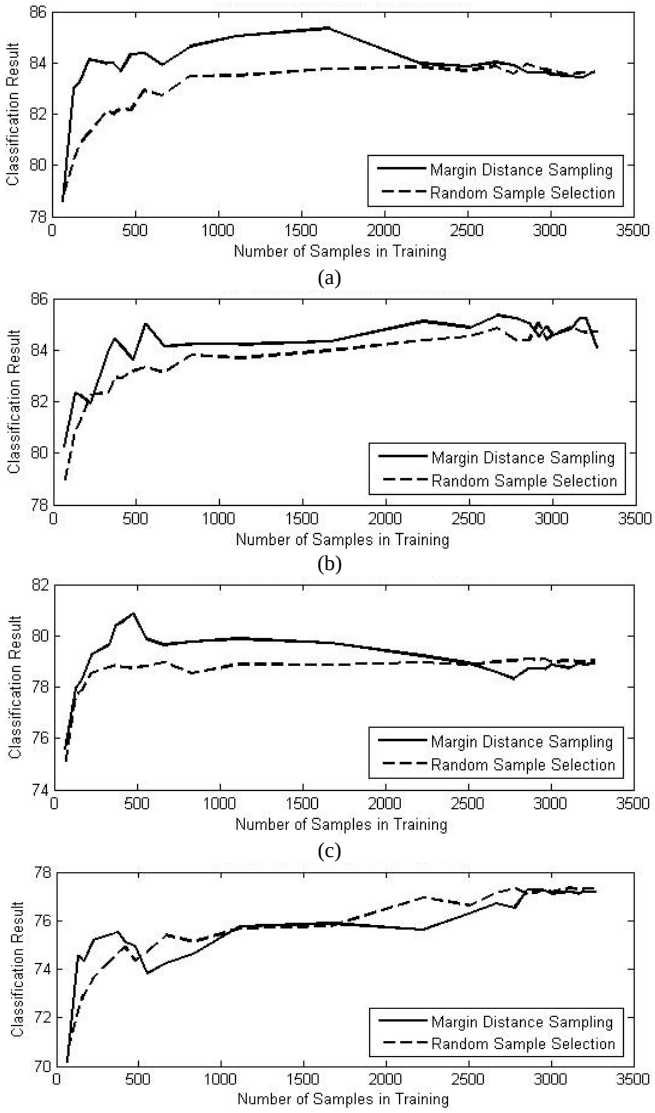

Fig. 5. SVM (a), RF (b), NB (c) and KNN (d) classification result for Waveform Dataset with RSS and MDS algorithm.

The most successful classifierand the best results corresponding datasets are listed in Table II. MDS represents the Margin Distance Selection as cited before and Random sample selection is abbreviated as RSS in the tables. Success rates are listed with related datasets and methods. Moreover, Table II also presents the sample size used for reaching the best classification success in the study.

TABLE II: BEST RESULTS AND CLASSİFİERS FOR DATASETS

	D159	Letter	Rignorm	Waveform
BestResult with MDS (%)	100	96,21	98,70	85,35
BestResult with RSS (%)	100	95,86	98,69	85,04
Classifier for BestResult(MDS)	SVM	RF	SVM	RF
Classifier for Best Result(RSS)	SVM	RF	SVM	RF
Sample size for Best Result(MDS) (% of data pool)	**4.47**	**43.51**	**9,94**	**61,52**
Sample size for Best Result (RSS) (% of data pool)	25.51	100	68,23	91,82

Table II shows the benefits of using an active sample selection on training process of the models clearly. Same or fractional successful results achieved with fewer samples with MDS. As it can be seen in Table II, numbers written as bold indicates the MDS success over RSS on corresponding success rates and classifier.

In addition to Table II, training process divided into 5 sections with a proportional as %10, %25, %50, %75 and %100 of data pool. Proportions are listed respectively in Table III. Relevant parts show the success rates of the classifier which gives best results on corresponding dataset mentioned in Table II at the defined sample size.

TABLE III: CLASSIFICATION SUCCESS AT DEFINED POINTS

%	D159		Letter		Rignorm		Waveform	
	ASS	RSS	ASS	RSS	ASS	RSS	ASS	RSS
Part I	**100**	99,82	**87,13**	83,72	**98,70**	98,20	**84,03**	82,37
Part II	**100**	99,91	**95,78**	89,42	**98,63**	98,43	**84,21**	83,82
Part III	100	100	**96,16**	93,75	**98,70**	98,62	**84,33**	83,97
Part IV	100	100	**96,13**	95,36	98,67	98,69	**85,11**	84,53
Part V	100	100	**95,98**	95,86	98,62	98,69	84,09	84,71

As it can be observed in Table III, datasets that are classified with MDS reach better classification result in early step. Results are listed as bold numbers in the table within this context. MDS maintains a peak starting point over dataset classification.

VI. CONCLUSION

In this study, an active sample selection algorithm is tested on four machine-learning methods about classification of some UCI datasets. Active sample selection algorithms generally focus on selection more informative samples in datasets. Many algorithms about sample selection in the manner of logical are located in literature. One of the active learning algorithms based on margin distances of samples is studied. Results obtained by methods are presented in this paper. Study contains a comparison result between randomly sample selected classification and margin distance based active sample selected classification. This paper presents the benefits of selecting more informative samples in training step of the classifiers. Besides the random selection, active sample selection provides same or, sometimes, slightly more success with few sample size in training. In addition, paper shows that sample size can be reduce in training process with the help of active sample selection.

In contrast to success of this paper, some initial point of the active sample selection tests gave inefficient results as it can be seen on figures. Main reason of this is RSS is applied at initial point of both tests with 1% of data pool. That can be overcome with another study about clustering the first data. In the future works, some clustering methods will be used for initial sample selection.

REFERENCES

[1] L. Haoyong and T. Hengyao, "Machine learning methods and their application research," in *Proc. 2nd International Symposium on Intelligence Information Processing and Trusted Computing (IPTC)*, 22-23 Oct. 2011, pp. 108-110.
[2] E. Alpaydın, *Introduction to Machine Learning*, 2nd ed. London, England: The MIT Press, 2010.
[3] C. Beleites, U. Neugebauer, T. Bocklitz, C. Krafft, and J. Popp, "Sample size planning for classification models," *Analytica Chimica Acta*, vol. 760, pp. 25-33, 14 January 2013.

[4] B. Settles, "Active learning literature survey," Technical Report 1648, University of Wisconsin – Madison.
[5] C. Agan and M. F. Amasyali, "Active learning with committees and the selection of starting sets," in *Proc. Signal Processing and Communications Applications Conference (SIU)*, 24-26 April, 2013, pp. 1-4
[6] H. Xu, X. Wang, Y. Liao, and C. Zheng, "An uncertainty sampling-based active learning approach for support vector machines," in *Proc. AICI '09. International Conference on Artificial Intelligence and Computational Intelligence*, 7-8 Nov., 2009, vol. 3, no. 213, p. 208.
[7] D. Tuia, F. Ratle, F. Pacifici, A. Pozdnoukhov, M. Kanevski, F. Del Frate, D. Solimini, and W. J. Emery, "Active learning of very-high resolution optical imagery with SVM: entropy vs margin sampling," in *Proc. IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2008*, 7-11 July, 2008, vol. 4, pp. 73- 76.
[8] *MATLAB and Statistics Toolbox Release*, The Math Works, Inc., Natick, Massachusetts, United States, 2012b.
[9] UC irvine machine learning repository. [Online]. Available: http://archive.ics.uci.edu
[10] D. Thanh-Nghi and J. Fekete, "Large scale classification with support vector machine algorithms," in *Proc. Sixth International Conference on Machine Learning and Applications (ICMLA) 2007*, 13-15 Dec., 2007, pp. 7-12.
[11] S. Wan and H. Yang, "Comparison among methods of ensemble learning," in *Proc. International Symposium on Biometrics and Security Technologies (ISBAST)*, 2-5 July, 2013, pp. 286-290.
[12] Y. Zhiyong and X. Congfu, "Combining KNN algorithm and other classifiers," in *Proc. 9th IEEE International Conference on Cognitive Informatics (ICCI)*, 7-9 July, 2010, pp. 800-805.
[13] J.-L. Xiao and G. Y. Yuan, "Learning lazy naive Bayesian classifiers for ranking," in *Proc. 17th IEEE International Conference on Tools with Artificial Intelligence, ICTAI*, 16 Nov., 2005, pp. 416-421.

Hamza O. İlhan received the B.Sc. degree in electronics and computer science from Marmara University, Istanbul, Turkey in 2010. His M.Sc. degree was received in computer engineering from Yalova University, Yalova, Turkey in 2012. He is currently a Ph.D. student in Yıldız Technical Universiy (YTU), Istanbul, Turkey. He was appointed to Yıldız Technical University as a research assistant in 2011. His research interests are in the areas of autonomous robots, image and signal processing, machine learning and pattern recognition with applications to biomedical engineering.

M. Fatih Amasyalı received the MSc degree from the Yildiz Technical University, Turkey, in 2003, and the Ph.D. degree from the same university in 2008. Dr. Amasyali is currently an assistant professor at the Computer Engineering Department, Yildiz Technical University. His interests include machine learning, natural language processing and autonomous robotics. He has published several scientific papers.