Journal of Computer Science

Original Research Paper

A Comparison Study between Different Sampling Strategies for Intrusion Detection System of Active Learning Model

Ghofran Mohammad Alqaralleh, Mohammad Aref Alshraideh and Ali Alrodan

Department of Computer Science, the University of Jordan, Amman, Jordan

Article history
Received: 7-05-2018
Revised: 08-06-2018
Accepted: 17-08-2018

Corresponding Author:
Mohammad Aref Alshraideh
Department of Computer Science, the University of Jordan, Amman, Jordan
Email: mshridah@ju.edu.jo

Abstract: Active learning aims to train an accurate model with minimum cost by labeling the most informative instances without compromising the model performance. So, choosing an efficient criterion for instance selection is the most important step. Sampling stage is the main issue in active learning for many problems such as intrusion detection system. There are many methods for sampling stage to select the informative instances, but what the method should be used to provide the most accurate to the Intrusion Detection System (IDS). So, we made a comparison between three of these methods, uncertainty sampling, Query By Committee (QBC) and expected model change. The contribution of this study is analyzing and examining three of common strategies that used to select the most informative instances to determine the best one of them. The experimental result showed that the expected model change method achieved the highest accuracy compared with uncertainty sampling and query by committee methods.

Keywords: Active Learning, Expected Model Change, Uncertainty Sampling, Query by Committee

Introduction

Machine learning is a type of artificial intelligence that provides computers with the ability to learn without being explicitly programmed. Interest in machine learning is due to several factors such as growing volumes and varieties of available data, computational processing that is cheaper and more powerful and affordable data storage. Based on these factors, it became easy to quickly and automatically produce models that have the ability to analyze bigger, more complex data and receive faster, more accurate results. Therefore, by making accurate models institutions have a better chance of specifying profitable chances of avoiding unknown risks especially in competitive and adversarial environments. The most common type of machine learning is supervised learning. In supervised learning training data includes both the inputs and the desired outputs. The correct outputs (targets) are known and are given to the model during the learning process (Salah *et al*., 2011; Qatawneh *et al*., 20017; Farhan *et al*., 2015). This type of learning is usually fast and accurate, while this approach is not applied in active learning. In active learning, we use the initial labelled samples and among the unlabelled samples, we try to find out labelling which small number of them will get much better performance. But, in supervised learning to produce an accurate model big data should be available, these labeled data require time-consuming, high cost. Large of sensitive institutions require large amounts of labeled data to obtain an accurate model such as network intrusion detection system. To solve such of these problems active learning is used. Active learning is a subfield of machine learning and the kind of learning. The principle of work for this framework, the learner has the freedom and influence to select which instances will be added to its training set (Roy and McCallum, 2001; Cohn *et al*., 1994).

Active learning aims to train an accurate prediction model with minimum cost by labeling the most informative instances without significantly compromising the model performance (Fu *et al*., 2013). Active learning aims at reducing the number of training examples to be label by automatically processing the unlabeled examples then selecting the most informative ones to label. The problem of active learning is to find the best selection strategy to quickly reach to high classification accuracy (Zhao *et al*., 2016). So, choosing an efficient criterion for instance selection is the most important step in active learning. In active learning, there is a small number of labeled data (training data) and a large number of unlabeled data (rest data). Model is produced

Science Publications

by using the training data (initial data) in the learning process and uses this classifier to select the most informative instances from the rest data to label and add these instances to the training data, then remove these instances from the rest. After that, the model is learned from the updated training set. This process is repeated to obtain the information size of training data. The standard supervised learning includes the training phase and testing phase, but in active learning, there is a sampling phase before the training and testing phases (Zhao *et al.*, 2012). The stage of selection of the sample which will be added to the training set is the most important stage in active learning; this stage is distinguishing active learning from supervised learning. In supervised learning, there are two phases learning phase and testing phase, but in active learning, there are three phases sampling phase, learning phase and testing phase. In case we used the supervised learning to improve the accuracy for sensitive applications, a large amount of labeled data must be provided. This is impractical, time-consuming and costly. So, the solution will be using active learning. Since the sampling is the most important stage of active learning, the accuracy and success of the model will depend on the effectiveness and success of this phase. The broad development of active learning has led to the use of many strategies such as query-by-committee (Gilad-Bachrach *et al.*, 2006; Iglesias *et al.*, 2011; Bloodgood, 2018), uncertainty sampling (Joshi *et al.*, 2009; Settles, 2010; Tong and Koller, 2001; Yang *et al.*, 2015), expected model change (Sznitman and Jedynak, 2010; Vezhnevets *et al.*, 2012; Long *et al.*, 2015). Among these strategies is what meets certain applications such as visual recognition (Long *et al.*, 2015; Luo *et al.*, 2005), foreground-background segmentation (Konyushkova *et al.*, 2015), natural language processing (Olsson, 2009; Tong and Koller, 2001), preference learning (Maystre and Grossglauser, 2015; Singla *et al.*, 2016). In addition to many applications.

Sampling Strategies

Uncertainty Sampling

Uncertainty sampling is one of the public strategies for measuring the most informative instance (Lewis and Gale, 1994). Principle of its work, the most informative instance is the instance where the model not uncertain how to label it. This framework uses the probabilistic models to evaluate the information of instances; the predicted results of the instance are represented by a vector, whose elements are the posterior probability with respect to each class label. For a binary classification, the most uncertain instance is the one whose posterior probability of being positive is the nearest 0.5 (Lewis and Gale, 1994). But, for problems with three or more class labels; there are three methods according to the number of posterior probabilities to select the most informative instance the Least Confidence (LC), margin and entropy.

Query by Committee

One of the major active learning strategies was proposed in (Seung *et al.*, 1992). This framework uses a classifier committee constructed from the training set, each member of the committee makes a vote on the class label of the instance and then the majority vote of the committee members is the final prediction. The instance with the most disagreement in the prediction is the most informative instance. In this strategy, multiple learners are generated and then select the instance where the learners disagree about label it. For example, suppose there are five learners among which three learners predict positive and two learners predict negative for an instance (xi), while four learners predict positive and one learner predict negative for the instance (xj) then these learners disagree more on (xi) than on (xj) and therefore (xi) will be selected for the query rather than (xj). Two points must be taken into consideration to implement query by committee: Construct a committee of hypothesis representing the different fields of a version space and design a measure to evaluate the disagreements between committee members.

According to construct the committee of classifiers, there are two methods to do this, the Query by Bagging (QBBagging) and Query by Boosting (QBBoosting) (Mamitsuka, 1998). The second point to implement query by committee strategy, we must design a measure to evaluate the disagreements between committee members. According to this point, there are two methods to do that vote entropy and average Kullback-Leibler (KL) divergence.

Expected Model Change

Another common active learning strategy is the expected model change; it works to choose the instance that will lead to a significant change in the current model if it is a label was known. An example query on this strategy is the “Expected Gradient Length” (EGL) approach for discriminating probabilistic model classes. This approach was proposed by (Settles *et al.*, 2007) for active learning in the multiple-instance setting. The EGL strategy utilizing in any learning problem where gradient based training is used. This strategy works the formation of a committee of models using samples of data labels. On the contrary of the QBC, unlabeled data are scored on the basis of the difference between the outputs of the committee on the one hand and expected outputs of the model built on the labeled dataset on the other hand. It is measured this disagreement through the absolute variance between the current model of the hand and the aggregated output of the committee on the other (O'Neill *et al.*, 2016). Which characterizes this strategy is that it is able to assess the classification score change for whole instances and then choose the instance with the highest effect. The main idea of this strategy is the instances that have the maximize change in the output are the most probable to enhance the model's accuracy.

Case Study

The dataset used in our study is KDD CUP 1999 off-line ID project from the University of California Irvine (UCI) machine learning repository. However, in our study we used only 5,000 records of the actual dataset size which is contained five million of records. 800 records of them are tested and the 4200 records form the whole data. 100 records form the labeled data and the remaining form the unlabeled data. Each record of the dataset contains 41 of attributes and the label set contains 23 various labels 22 attack types and 1 normal. But we have converted the dataset to two classes for simplification. These attributes represent a midst two network hosts and the categorical features are encoded using numerical values. These attributes fall under three distinct types: Content, traffic and intrinsic. The content attributes take the content of the packet in consideration to describe the network behavior. The traffic attributes measure the number of network events on a number of different ports. The intrinsic attributes consist of information about the network packet-level.

Methodology

In this study, the dependent variable is the class of the network traffic instance, normal or attack. The normal instances in the dataset are labeled as 0, while the attack instances are labeled as 1. The neural network model is trained to predict for the dependent variable a real number between 0 and 1. The instance in network traffic for IDS will be classified as an attack if its value greater than 0.5, but if its value less than 0.5 it is classified as normal. Active learning procedure is shown in Fig. 1.

The whole process of the active learning algorithm can be shown in algorithm 1.

Algorithm 1. Active learning procedure

```
Input: U-unlabeled instance, L-labeled instances, $N_i$-number of instances to be selected per iteration, $N_{itr}$: number of iterations, C-Classifier
ITR = 0
For ITR<$N_{itr}$
    Learn classifier C from L
    Query a set instances $N_i\{x^*\} \in U$ according to sampling strategy and label it $\{(x^*,y^*)\}$
    $L \leftarrow L \cup Ni\ \{(x^*,y^*)\}$
    $U \leftarrow U \setminus N_i\ \{x^*\}$
End
```

Uncertainty Sampling

According to this method and since there are two class labels for a binary classification; the most uncertain instance is the one whose posterior probability of being positive is the nearest 0.5 regardless of the use of any method of uncertainty strategy whether it was less confident, margin or entropy (Tong and Koller, 2001). The uncertainty sampling method scenario is shown in Fig. 2.

In this method, we consider the first 100 instances are the labeled set; the unlabeled data are the rest of the whole data (4100). Firstly, will be training the network based on the labeled set to obtain the classifier through giving random weights. Then will be passing the rest data onto the base classifier. After that, the absolute value of the difference between 0.5 and the value of output for each instance will be found. Then sort the instances in the unlabeled set in ascending order according to this difference and take the top (N_i) instances and their outputs according to the threshold. Add the top (N_i) instances to the labeled set and remove it from the unlabeled data.

Query by Committee

In this technique, a committee (two or more) of different classifiers trained on the initial labeled data will be built. In this study, it is sufficient to build two hypotheses (NNs). The normal instance in the dataset is labeled as 0, while the attack instance is labeled as 1. Figure 3 shows the scenario for this method as the following:

1. We will build committee consists of two classifiers (NN); each classifier will be trained on the parts of labeled data, the first classifier trained on the first 50 instances and the second classifier trained on the second 50 instances
2. Pass the rest of data onto the two classifiers and take the difference between the outputs from the two classifiers for each instance
3. Sort the instances in descending order based on the difference and take the top N_i instances
4. Add these instances with its true label to the labeled set and remove it from the rest data
5. The first N_i instances will be used to build a base classifier and in each iteration will be retrained the committee of classifier and the base classifier on the updated labeled data Nitr times

Expected Model Change

Which characterized this strategy from the previous strategy; we will be building the base classifier (NN) by the entire labeled data on the hand and create the committee of classifiers based on a sample of labeled data on the other hand. The scenario of this method is shown in Fig. 4.

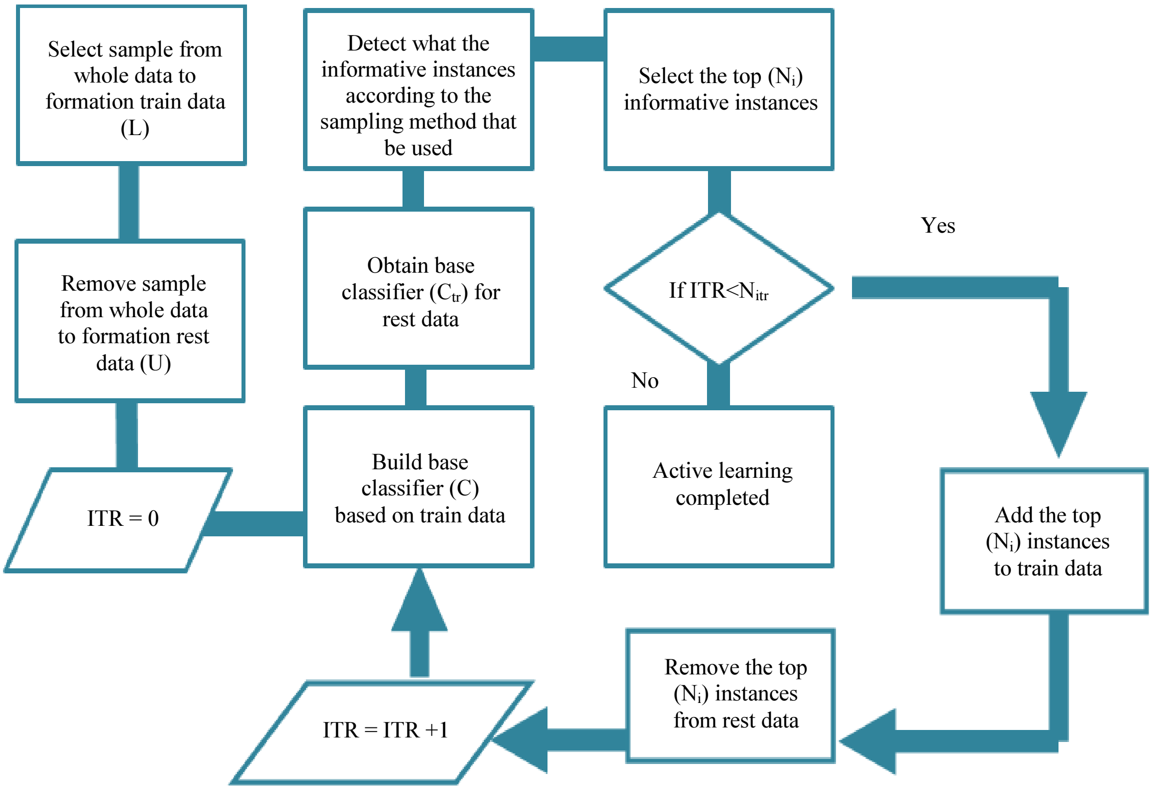
Select sample from whole data to formation train data (L)
Detect what the informative instances according to the sampling method that be used
Select the top (N_i) informative instances
Remove sample from whole data to formation rest data (U)
Obtain base classifier (C_{tr}) for rest data
If ITR<N_{itr}
Yes
No
ITR = 0
Build base classifier (C) based on train data
Active learning completed
Add the top (N_i) instances to train data
Remove the top (N_i) instances from rest data
ITR = ITR +1

Fig. 1: Active learning procedure

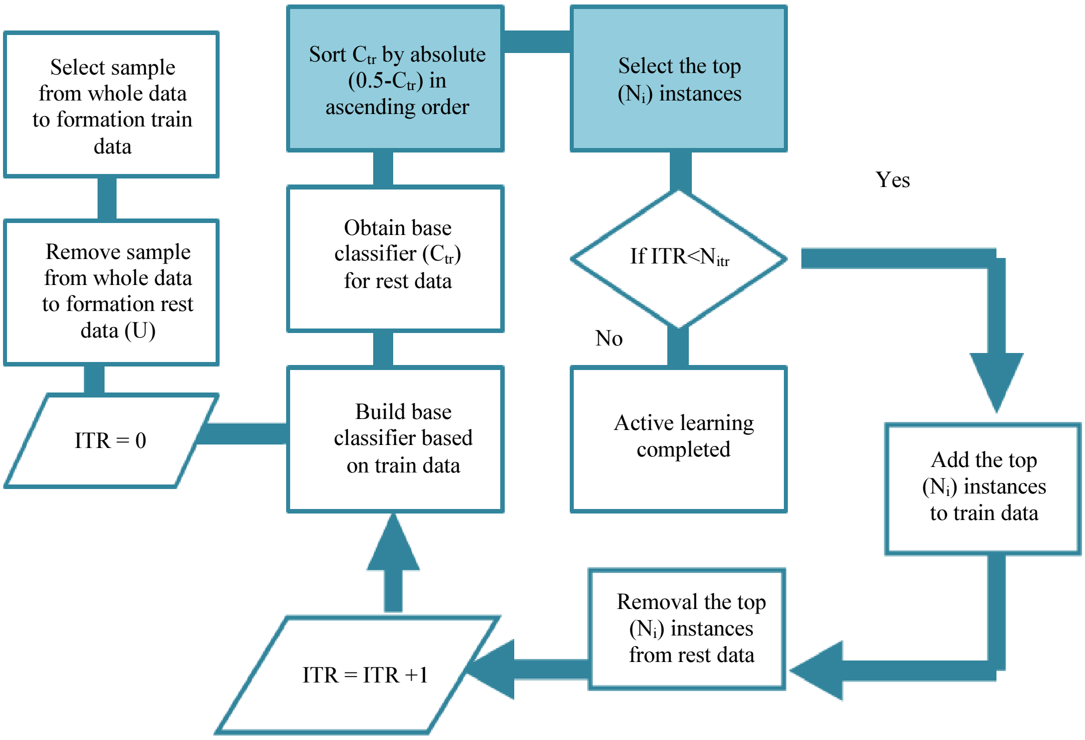
Select sample from whole data to formation train data
Sort C_{tr} by absolute (0.5-C_{tr}) in ascending order
Select the top (N_i) instances
Remove sample from whole data to formation rest data (U)
Obtain base classifier (C_{tr}) for rest data
If ITR<N_{itr}
Yes
No
ITR = 0
Build base classifier based on train data
Active learning completed
Add the top (N_i) instances to train data
Removal the top (N_i) instances from rest data
ITR = ITR +1

Fig. 2: Uncertainty sampling in active learning

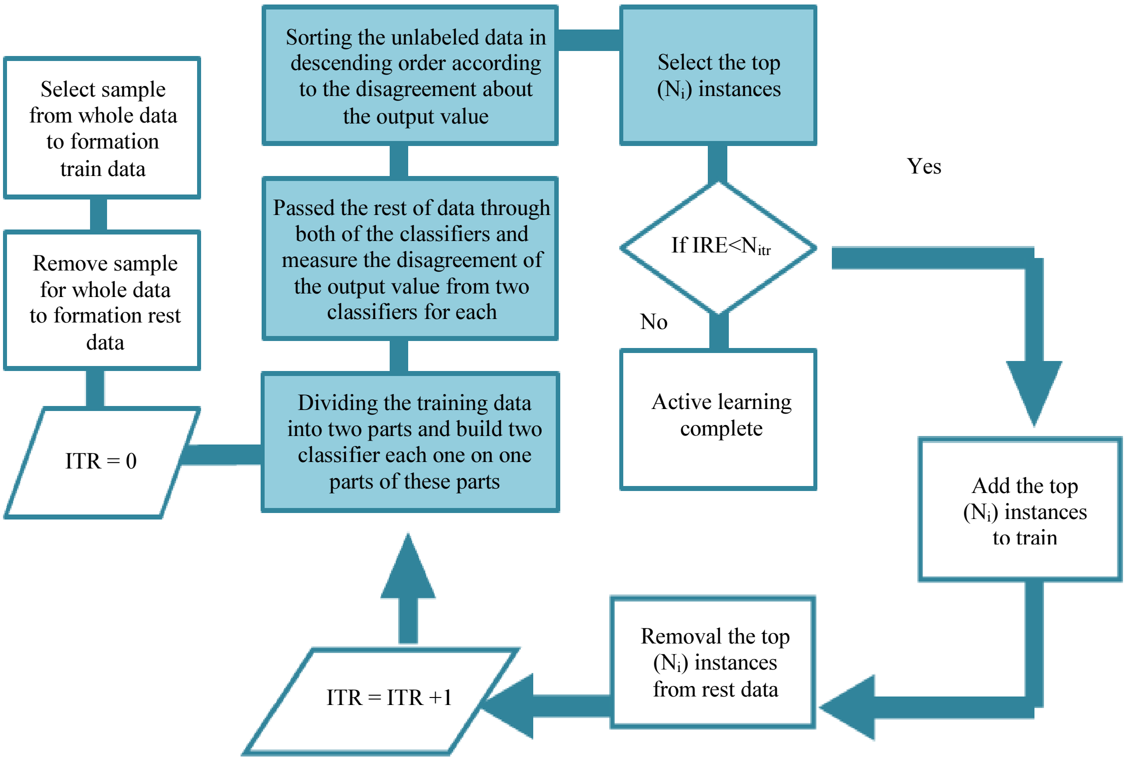
Select sample from whole data to formation train data
Remove sample for whole data to formation rest data
ITR = 0
Dividing the training data into two parts and build two classifier each one on one parts of these parts
Passed the rest of data through both of the classifiers and measure the disagreement of the output value from two classifiers for each
Sorting the unlabeled data in descending order according to the disagreement about the output value
Select the top (N_i) instances
If IRE<N_{itr}
Yes
No
Active learning complete
Add the top (N_i) instances to train
Removal the top (N_i) instances from rest data
ITR = ITR +1

Fig. 3: QBC in Active Learning

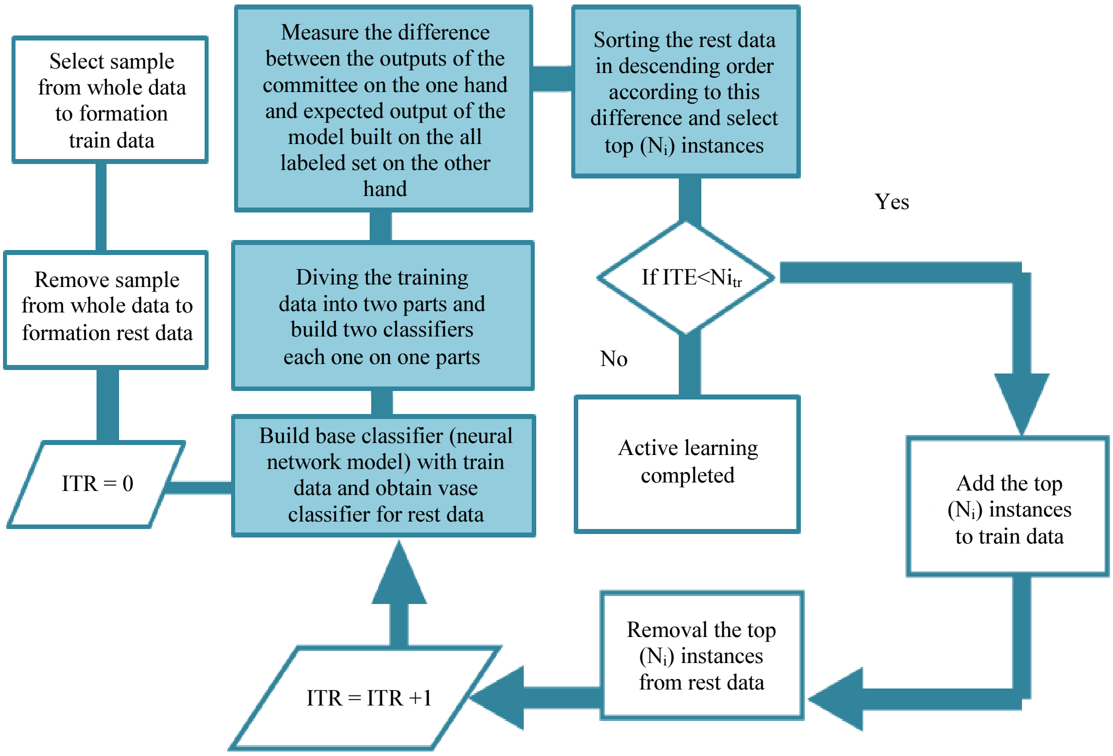
Select sample from whole data to formation train data
Remove sample from whole data to formation rest data
ITR = 0
Build base classifier (neural network model) with train data and obtain vase classifier for rest data
Diving the training data into two parts and build two classifiers each one on one parts
Measure the difference between the outputs of the committee on the one hand and expected output of the model built on the all labeled set on the other hand
Sorting the rest data in descending order according to this difference and select top (N_i) instances
If ITE<Ni_{tr}
Yes
No
Active learning completed
Add the top (N_i) instances to train data
Removal the top (N_i) instances from rest data
ITR = ITR +1

Fig. 4: Expected model change in active learning

Will be build the committee of two classifiers (NN) trained on the different parts on labeled set, the first classifier trained on the first 50 instances and the second classifier trained on the second 50 instances:

1. The base classifier trained on all the labeled set (the first 100 instances)
2. Pass the rest of data onto the two classifiers and take the difference between the outputs from the two classifiers for each instance
3. Sort the instances in descending order based on the difference and take the top N_i instances
4. Build the new classifier from this committee based on these instances
5. Pass the rest of the data onto both, the new classifier and the base classifier that has trained on all data
6. Take the difference between the output from the new classifier and the base classifier for each instance, sort the instances in descending order and take the top Ni instances
7. Add these instances on the labeled set and remove it from the rest
8. Build the classifier based on these instances

This process repeated until Nitr times.

Experimental Results

Parameters Setup

Active Learning Parameters

Firstly, we selected three numbers of iterations in our study 3, 5 and 7. These numbers of iterations were chosen based on the initial experiments that show that these numbers fit with the size of the whole data and forms the articulated points after trying many of the iteration numbers. For the number of instances that selected from the whole data in each iteration, we used two different numbers of instances (N_i) 50 and 30. These numbers were chosen also based on the experiments that show that these numbers provide preference accuracy from others and appropriate to make the comparison. Number of classifiers in the committee for the QBC method and expected model change method was two classifiers. NN was used as classifier. The parameters that used in our study are given in Table 1.

Table 1: Active learning Parameters

Number of iterations	3, 5, 7
Number of instances in each iteration	30, 50
Size of committee	Two classifiers
Classifier type	Neural network

Neural Network Parameters

We used the Neural Network (NN) classifier as classification algorithm. Many network architectures were used of the beginning of our study. NN with three hidden layers and one output with different number of neurons in the hidden layer such as 20-20-20-1, 40-40-40-1 and 30-30-30-1. In addition, we also used the architecture of one hidden layer and one output such as 30-1 architecture which denotes one output unit and one hidden layer and 20-1, 40-1 architectures. The 30-1 architecture was adopted in our study; it provided a good efficiency in learning. The log-sigmoid transfer function was used for hidden layer and output layer. The neural network was trained 10000 epochs at each active learning iteration. The parameters that used in our study are given in Table 2.

Results on Testing Data

The performance of the ID models based on the sampling methods that use in active learning was compared. This was done to evaluate the best of sampling approach that provides a good accuracy.

As shown previous the size of dataset is 5,000 instances; we selected randomly 100 instances initial training data and 800 instances the testing data. We made testing for three of sampling methods on six different initial samples with size of 100. The ratios for presence the normal instances and attack instances approximately are equal for these samples; 50% attack instances and 50% normal instances. These ratios reflect the ratios of the whole dataset. In our experimentation we took in consideration the following points to make comparison between three of sampling methods:

- We tested three of sampling methods on three different numbers of iterations 3, 5 and 7 to know the influence of change the number of iterations on the performance for these methods
- We tested these methods in two cases; the first case we considered the number of instances that will be select in each iteration 50 and in the second case the number of instances was 30
- We are tested the methods on six different initial samples with size of 100 to know the influence degree for selecting the initial samples on performance to these methods

Table 2: NN network parameters

Number of epochs in network	10000
Learning rate	0.06
Number of hidden layers	1
Number of neurons in hidden	30
Initial weights	Random between (0-1)
Performance function	Mean Square Errors

Results

Results Based Selecting 50 Instances

Results Based Uncertainty Sampling Method

The performance of uncertainty sampling method for six different initial samples with size of 100 instances and 50 instances selecting in each iteration is shown in Fig. 5. We note that the iterations 5 and 7 approximately have the same accuracy with slice preference to iteration 7 on 5, then the iteration 3. In iteration 7, the accuracy of 3 samples was 92% and above, one sample with accuracy 96% in iteration 3 and one sample with accuracy 95% in iteration 5. In addition, we note that the highest accuracy was in two samples; the sixth sample in first place and in the second place the first sample as shown in Fig. 5.

Results Based Query by Committee Method

The performance of the QBC method for six different initial samples with size of 100 instances and 50 instances selecting in each iteration is shown in Fig. 6. We note that the accuracy at iteration 3 was better than iterations 5 and 7, in the second place comes the iteration 7 and then comes iteration 5. In iteration 3, the accuracy of four of samples was 92-97%, three samples with accuracy 91-97% in iteration 7. We note that the highest accuracy was in two samples; the sixth sample in first place and in the second place the first sample.

Results Based Expected Model Change Method

The performance for expected model change for six different initial samples with size of 100 instances and 50 instances selecting in each iteration is shown in Fig. 7. We note that the accuracy at iteration number 3 was the best, iterations 5 and 7 approximately have the same accuracy with slice preference to iteration 7 on 5. In iteration 3, the accuracy of three of samples was 91-93%, two samples with accuracy 90% in iteration 5 and two samples with accuracy 94-95% in iteration 7. We note that the highest accuracy was in two samples; the third sample in first place and in the second place the sixth sample.

Comparison Results on Three Sampling Methods

The performance of the active learning procedure based on three sampling approaches for six different initial samples with size of 100 is show in Fig 8 to 13, the points represent the three sampling methods QBC, uncertainty sampling and expected model change. Figure 8 shows that the QBC method has the highest accuracy in two iterations 3 and 7 and the expected model change method has the highest accuracy in iteration 5. In general, the QBC method has the highest total accuracy in the initial sample number 1 and the expected model change method in the second place, then the uncertainty sampling method.

Figure 9 shows that the uncertainty sampling method has the highest accuracy in two iterations, 5 and 7. The expected model change method has the highest accuracy at iteration number 3. In general, the uncertainty sampling method has the highest total accuracy in the initial sample number 2, the expected model change method comes in the second place and then comes the QBC method.

Figure 10 shows that the expected model change method has the highest accuracy in all iterations 3, 5 and 7. In general, the expected model change method has the highest total accuracy in the initial sample number 3, the uncertainty sampling method in the second place and then the QBC method.

Figure 11 and 12 show that in sample number 4, the QBC method has the highest accuracy in all iterations 3, 5 and 7. In sample number 5, the QBC method has the highest accuracy at iteration number 3, the expected model change method has the highest accuracy in iteration 5 and the uncertainty sampling method in iteration 7. In general, the QBC has the highest total accuracy in two samples 4 and 5, the expected model change method in second place and then comes the uncertainty sampling method.

Figure 13 shows that the uncertainty sampling method has the highest accuracy in two iterations 3 and 7 and the QBC method has the highest accuracy at iteration 5. In general, the uncertainty sampling method has the highest total accuracy in the initial sample number 6, the QBC method comes in the second place, then the expected model change method. We note that the results of accuracy were the highest in sample number 6 from the other samples.

Results Based Selecting 30 Instances

Results Based Uncertainty Sampling Method

The performance of the uncertainty sampling method for six different initial samples with size of 100 and 30 instances in each iteration is shown in Fig. 14.

This Figure illustrates that the iterations 5 and 3 approximately have the same accuracy with slice preference to iteration 3 on 5 and then comes iteration 7. In iteration 3, the accuracy of four samples was 90-97%. In iteration 5, the accuracy of three samples was 93-96%. In iteration 7, two samples with accuracy 95-96%. We note that the highest accuracy was in two samples; the sixth sample in first place and in the second place the first sample. In general, we conclude that the accuracy of the uncertainty sampling method in the case of 30 instances in each iteration is better than the case of 50 instances in each iteration.

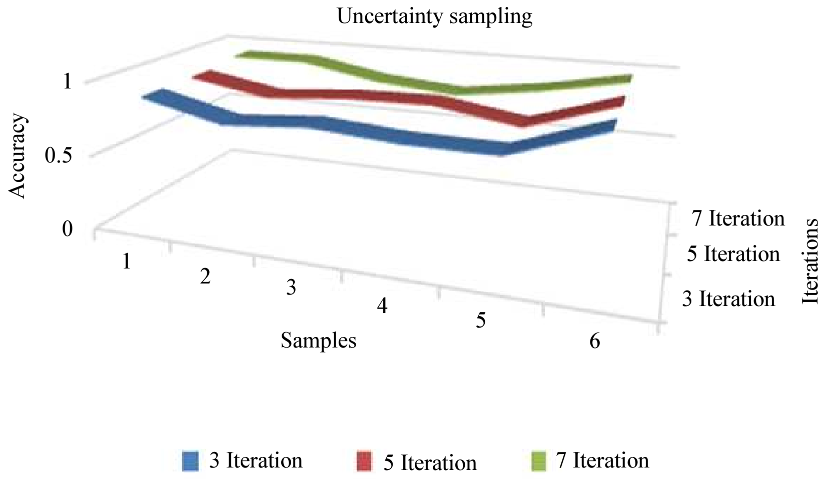

Fig. 5: Accuracy for uncertainty method based selecting 50 instances

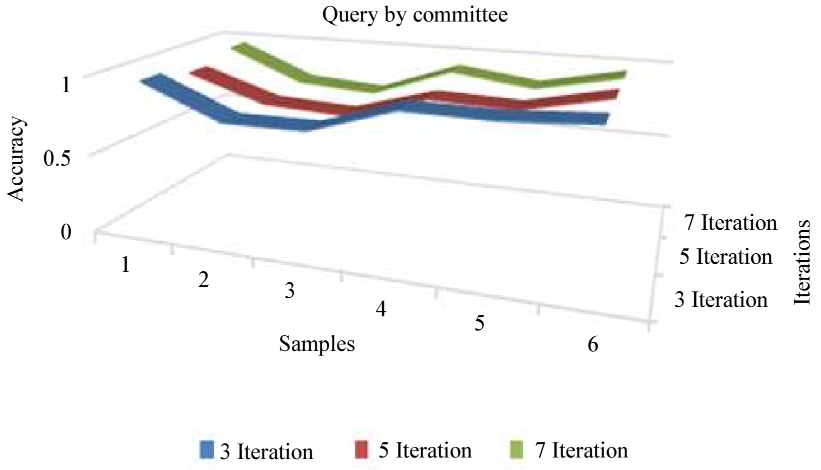

Fig. 6: Accuracy for QBC method based selecting 50 instances

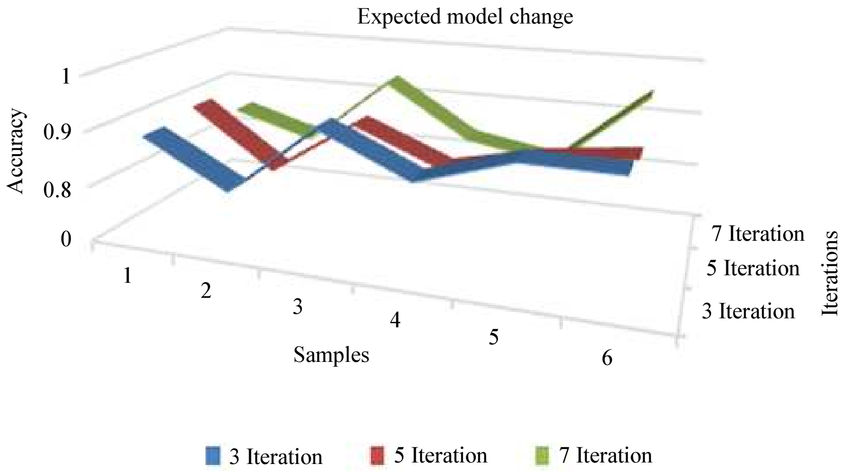

Fig. 7: Accuracy for expected model change method based selecting 50 instances

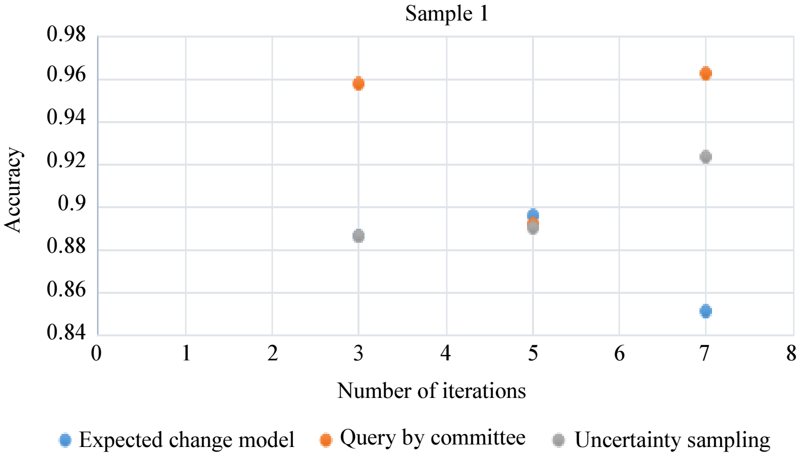

Fig. 8: Accuracy for all methods based selecting 50 instances

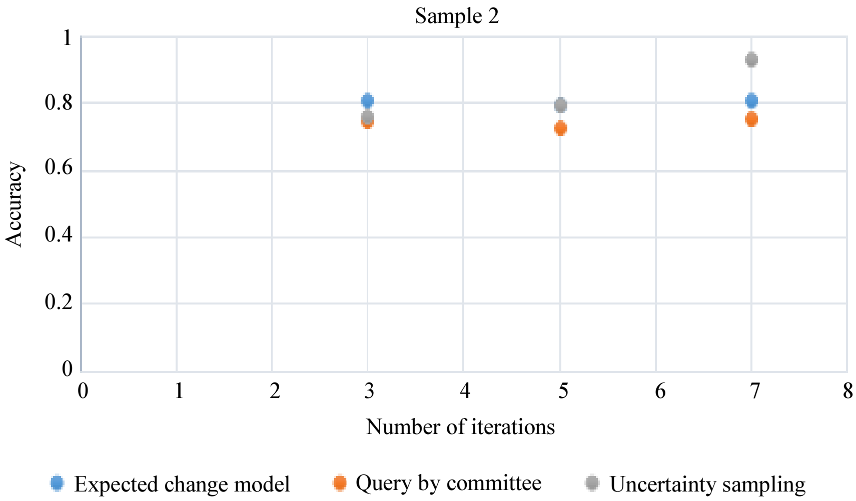

Fig. 9: Accuracy for all methods based selecting 50 instances

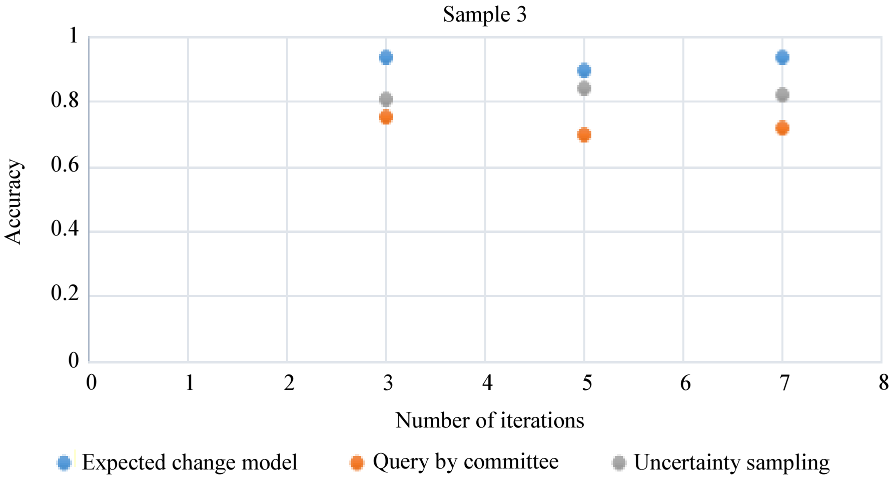

Fig. 10: Accuracy for all methods based selecting 50 instances

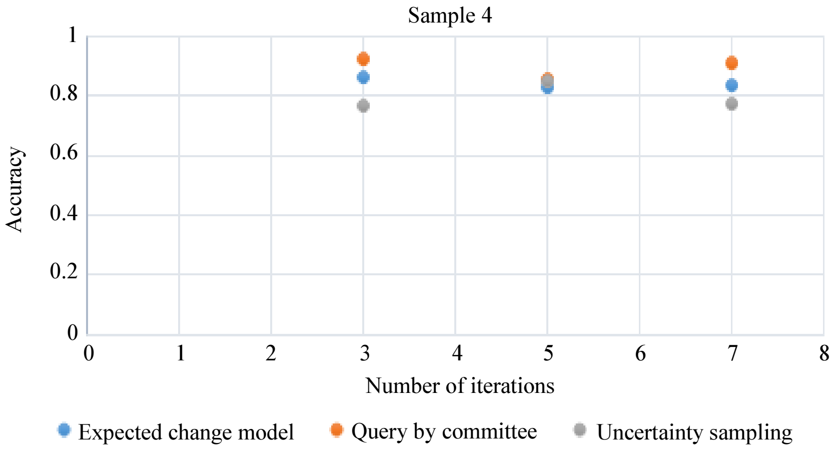

Fig. 11: Accuracy for all methods based selecting 50 instances

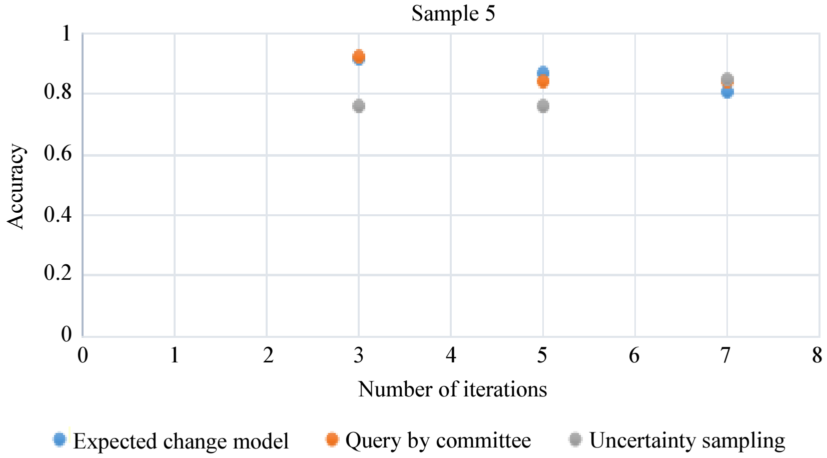

Fig. 12: Accuracy for all methods based selecting 50 instances

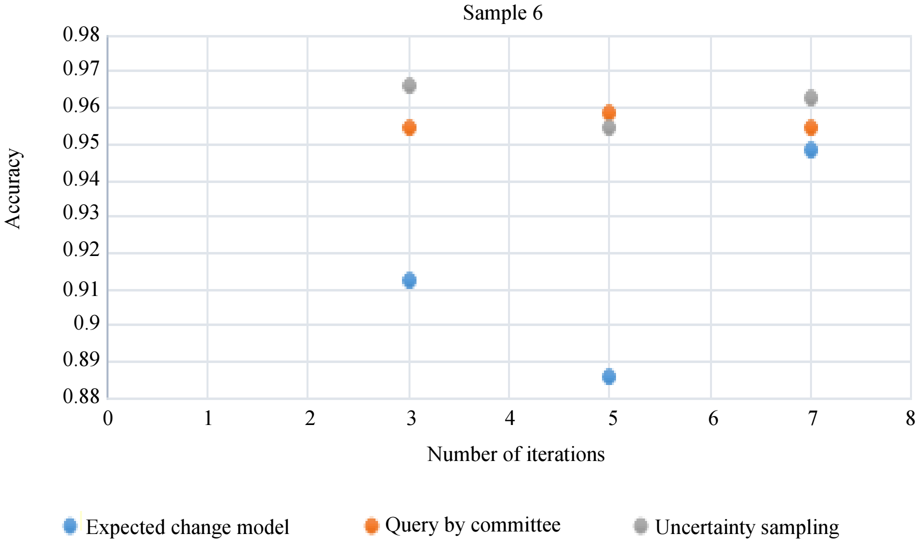

Fig. 13: Accuracy for all methods based selecting 50 instances

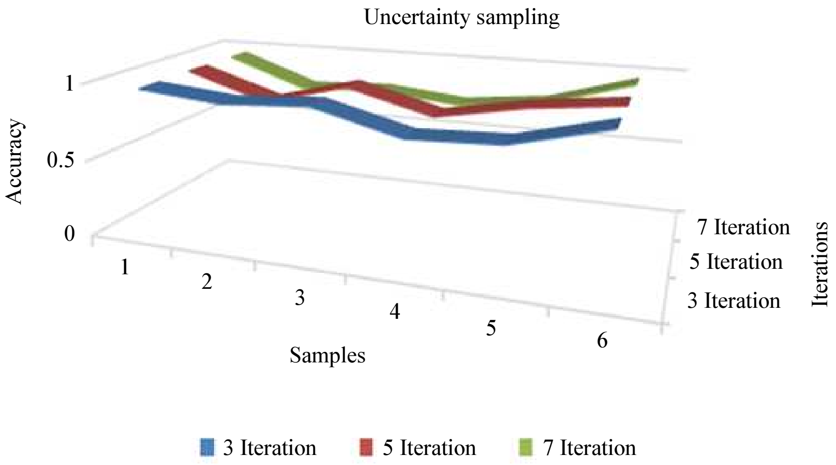

Fig. 14: Accuracy for uncertainty sampling based selecting 30 instances

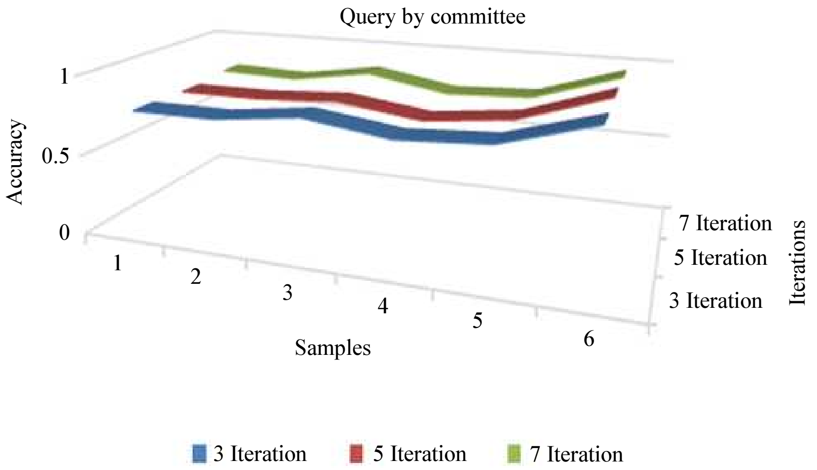

Fig. 15: Accuracy for QBC method based selecting 30 instances

Results Based Query by Committee Method

The performance of QBC method for six different initial samples with size of 100 and 30 instances selecting in each iteration is shown in Fig. 15.

This Figure shows that the accuracy in almost all samples is equal, with slice preference to iteration 7 on other iterations. Also, shows that the highest accuracy was in two samples, the sixth sample in first place and in the second place the third sample. In general, we conclude that the accuracy for the query by committee method in the case of selecting 30 instances in each iteration is better than the case of 50 instances in each iteration.

Results Based Expected Model Change Method

The performance of expected model change for six different initial samples with size of 100 and 30 instances in each iteration is shown in Fig. 16.

This Figure illustrates that the iterations 7 and 3 approximately have the same accuracy with slice preference to iteration 7 on 3, after that, comes iteration 5. In iteration 3, the accuracy of three samples was 90-97%. In iteration 7, the accuracy of three samples was 95-96%. In iteration 5, two samples with accuracy 90%. We note that the highest accuracy was in two samples, the sixth sample in first place and in the second place the first sample. In general, we conclude that the accuracy of the expected model change method in the case of 50 instances in each iteration was better than 30 instances in each iteration.

Comparison results on three sampling methods

Figure 17 to 22 show the results in the case of 30 instances in each iteration for the three sampling methods QBC, uncertainty sampling and expected model change. Figure 17 illustrates that the uncertainty sampling method has the highest accuracy in two iterations, 3 and 5. The uncertainty sampling method and the expected model change method having approximately the same and highest accuracy in iteration 7. In general, the uncertainty sampling method has the highest total accuracy in the initial sample number 1, expected model change method comes in second place, then the QBC method.

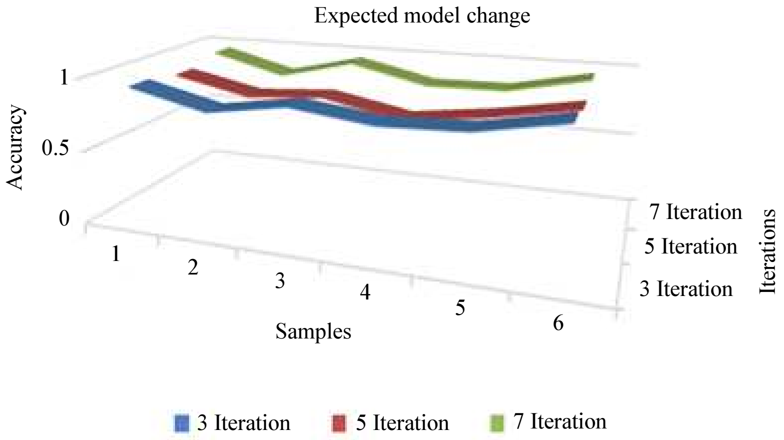

Fig. 16: Accuracy for expected model change based selecting 30 instances

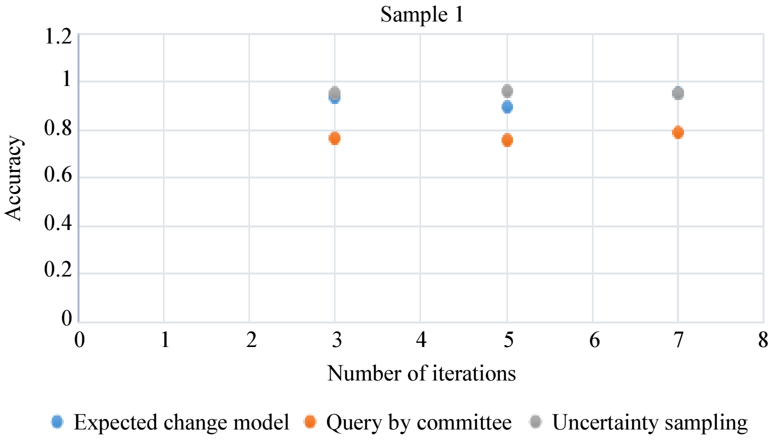

Fig. 17: Accuracy for all methods based selecting 30 instances

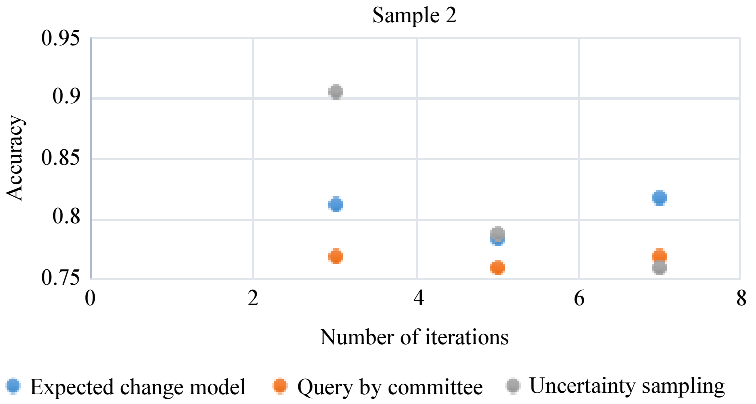

Fig. 18: Accuracy for all methods based selecting 30 instances

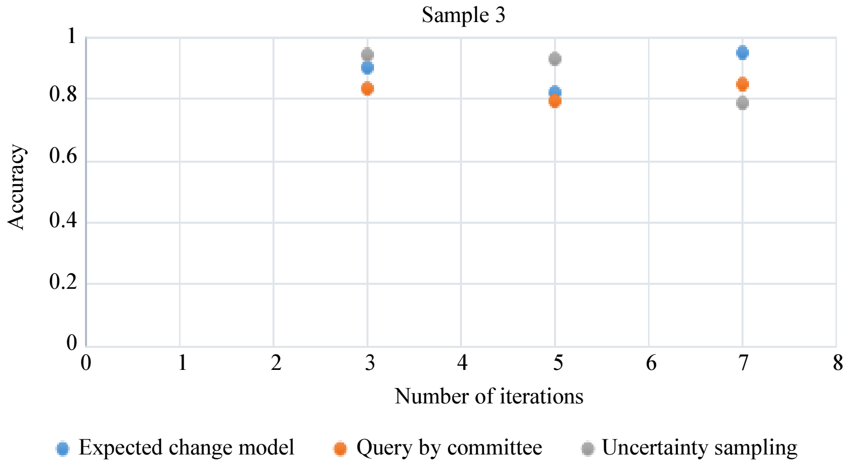

Fig. 19: Accuracy for all methods based selecting 30 instances

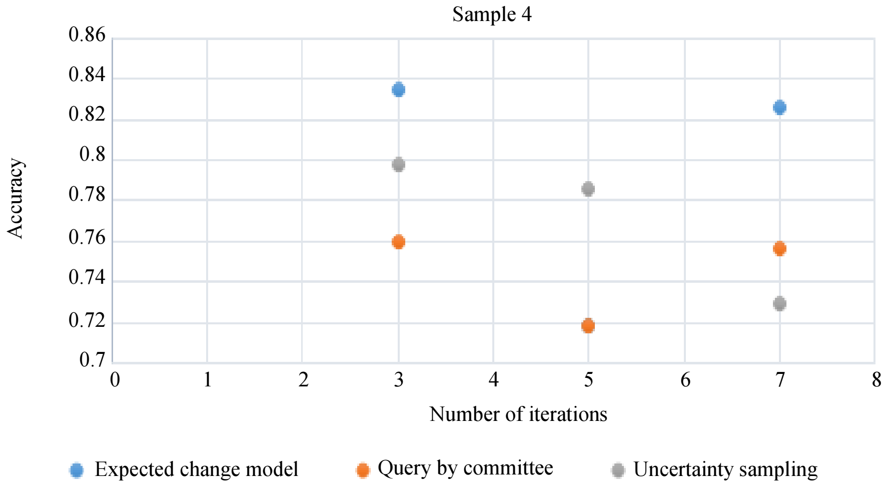

Fig. 20: Accuracy for all methods based selecting 30 instances

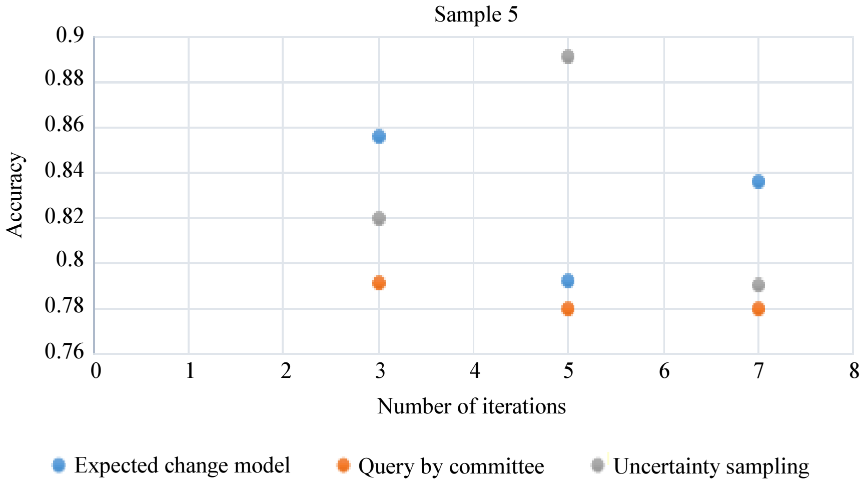

Fig. 21: Accuracy for all methods based selecting 30 instances

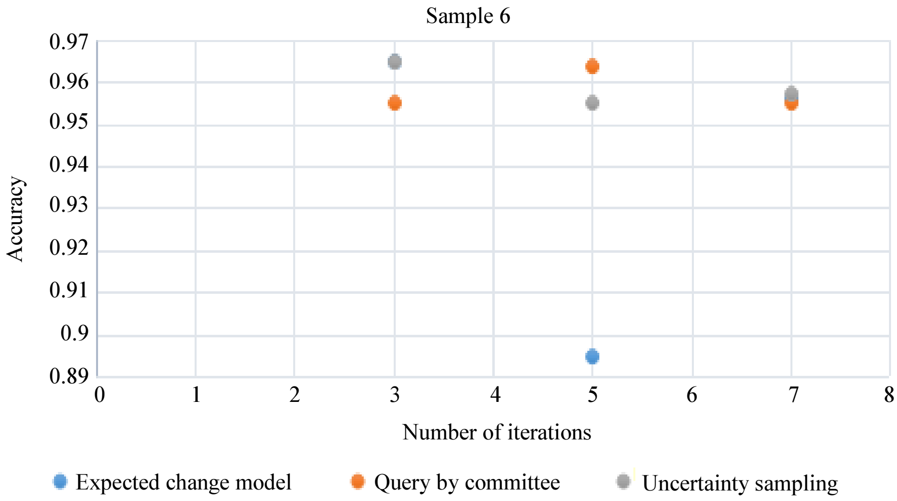

Fig. 22: Accuracy for all methods based selecting 30 instances

Figure 18 shows that the uncertainty sampling method has the highest accuracy in two iterations number 3 and 5. The expected model change method has the highest accuracy at iteration number 7. In general, the uncertainty sampling method and the expected model change method having approximately the same total accuracy in the initial sample number 2 with slice preference to the expected model change method on the uncertainty sampling method. After that, comes the QBC.

In Fig. 19, the uncertainty sampling method has the highest accuracy in two iterations number 3 and 5. The expected model change method has the highest accuracy at iteration number 7. In general, the uncertainty method has the highest total accuracy in the initial sample number 3 and QBC method has the least accuracy in this sample.

Figure 20 and 21 shows that in samples 4 and 5, the expected model change method has the highest accuracy in two iterations, 3 and 7. The uncertainty sampling method has the highest accuracy at iteration number 5. In general, the expected model change method has the highest total accuracy in sample number 4, the uncertainty sampling method achieves the highest accuracy in sample number 5, then the QBC.

In Fig. 22, the uncertainty sampling method has the highest accuracy in two iterations 3 and 7, the QBC method has the highest accuracy in iteration number 5. The uncertainty sampling method and the QBC method having approximately the highest total accuracy in the initial sample number 6, then the expected model change method. We note that the results of accuracy were the highest in sample number 6.

Discussion

Based Selecting 50 Instances

Looking at results in the previous section. Firstly, we will discuss the influence of the number of iterations for six different samples in the case of selecting 50 instances in each iteration. From the Fig. 5 to 7, show that in iteration number 3, the query by committee has the highest accuracy in three samples 1, 4 and 5. The expected model change has the highest accuracy in two samples 2 and 3. According to sample number 6, the uncertainty sampling method has the highest value of accuracy.

In iteration number 5, the expected model change method has the highest accuracy in three samples 1, 3 and 5. The QBC method achieves the highest accuracy in two samples 4 and 6. According to sample number 2, the uncertainty method is achieving the highest value of accuracy.

In iteration number 7, the uncertainty sampling method has the highest accuracy in three samples 2, 5 and 6. The QBC method has the highest accuracy in two samples 1 and 4. According to sample number 3, the expected model change method has the highest value of accuracy.

As a result of foregoing, the QBC method was the best method in iteration number 3. The expected model change method was the best in iteration number 5. The uncertainty sampling method was the best in iteration number 7.

The values of the accuracy depending on three sampling approaches for six different initial samples with size 100 appear in Table 3 to 8.

In sample 1, the optimal Performance was for the QBC method and iteration 7.

Table 3: Comparative performance for all methods in first sample

Active learning iterations	Expected change model	QBC	Uncertainty sampling
3	0.8863	0.9575	0.8863
5	0.8963	0.8925	0.8900
7	0.8513	0.9625	0.9238

Table 4: Comparative performance for all methods in second sample

Active learning iterations	Expected change model	QBC	Uncertainty sampling
3	0.8050	0.7475	0.7600
5	0.7913	0.7263	0.7938
7	0.8063	0.7513	0.9275

Table 5: Comparative performance for all methods in third sample

Active learning iterations	Expected change model	QBC	Uncertainty sampling
3	0.9338	0.7513	0.8050
5	0.8950	0.7025	0.8388
7	0.9363	0.7175	0.8225

Table 6: Comparative performance for all methods in fourth sample

Active learning iterations	Expected change model	QBC	Uncertainty sampling
3	0.8625	0.9263	0.7638
5	0.8300	0.8563	0.8513
7	0.8375	0.9100	0.7725

Table 7: Comparative performance for all methods in fifth sample

Active learning iterations	Expected change model	QBC	Uncertainty sampling
3	0.9138	0.920	0.7600
5	0.8663	0.840	0.7600
7	0.8100	0.841	0.8513

Table 8: Comparative performance for all methods in sixth sample

Active learning iterations	Expected change model	QBC	Uncertainty sampling
3	0.9125	0.9550	0.9663
5	0.8863	0.9588	0.9550
7	0.9488	0.9550	0.9625

In sample 2, the optimal Performance was for the uncertainty sampling method and iteration 7.

In sample 3, the optimal Performance was for the expected change model method and iteration 3.

In sample 4, the optimal Performance was for the QBC method and iteration 3.

In sample 5, the optimal Performance was for the QBC method and iteration 3.

In sample 6, the optimal Performance was for the uncertainty sampling method and iteration 7.

From previous tables we note that the highest accuracy was in the first place, in sample 6. In the second place, sample 1. In the third place, sample 5. In the fourth place, sample 4. In the fifth place, sample 3. In the sixth place, sample 2. We conclude that the samples 6 and 1 involve of informative instances more than other initial samples. Therefore, choosing a good initial sample containing informative instances will be lead to increases in accuracy.

Based Selecting 30 Instances

Now we will discuss the influence of the number of iterations for six different samples in the case of selecting 30 instances in each iteration. From the Fig. 14 to 16 show that in iteration number 3, the uncertainty sampling method has the highest accuracy in three samples 1, 2 and 3. The expected model change method has the highest accuracy in two iterations 4 and 5. In sample number 6, the uncertainty sampling method and the expected model change method have the same accuracy.

In iteration number 5, the uncertainty sampling achieves the highest accuracy in the first five samples, in the second place comes the QBC method achieves the highest accuracy in sample 6, then comes the expected model change method achieves the least accuracy.

In iteration number 7, the expected model change method has the highest accuracy in four samples 2, 3, 4 and 5. The uncertainty sampling method has the highest accuracy in sample numbers 6. The expected model change method and the uncertainty sampling method have the same accuracy in sample 1.

As a result of foregoing, the uncertainty sampling method was the best method in iterations number 3 and 5, the expected model change in iteration number 7.

The values of the accuracy depending on three sampling approaches for six different initial samples with size 100 appear in Table 9 to 14.

In sample 1, the optimal Performance was for the uncertainty sampling method and iteration 7.

In sample 2, the optimal Performance was for the uncertainty sampling method and iteration 3.

In sample 3, the optimal Performance was for the expected change model method and iteration 3.

In sample 4, the optimal Performance was for the expected change model method and iteration 3.

In sample 5, the optimal Performance was for the uncertainty sampling method and iteration 3.

In sample 6, the optimal Performance was for the uncertainty sampling method and iteration 3.

From previous tables we note that the highest accuracy was in the first place, in sample 6. In the second place, sample 1. In the third place, sample 3. In the fourth place, sample 5. In the fifth place, sample 2. In the sixth place, sample 4. Therefore, we conclude that the samples 6 and 1 involve of informative instances more than other initial samples.

In general, the method that achieves the highest accuracy in the case of selecting 50 instances in each iteration, the expected model change. In the second place, comes the QBC method. In the third place, comes the uncertainty sampling method. According to the case of selecting 30 instances in each iteration, the uncertainty method achieves the highest accuracy. In the second place, the expected model change method. After that, comes the QBC method.

Table 9: Comparative performance for all methods in first sample

Active learning iterations	Expected change model	QBC	Uncertainty sampling
3	0.9338	0.7688	0.9550
5	0.8938	0.7600	0.9588
7	0.9550	0.7863	0.9550

Table 10: Comparative Performance for all Methods in Second Sample

Active learning iterations	Expected change model	QBC	Uncertainty sampling
3	0.8125	0.7688	0.9050
5	0.7838	0.7600	0.7875
7	0.8175	0.7688	0.7600

Table 11: Comparative performance for all methods in third sample

Active learning iterations	Expected change model	QBC	Uncertainty sampling
3	0.9000	0.8325	0.9413
5	0.8238	0.7913	0.9325
7	0.9500	0.8513	0.7850

Table 12: Comparative performance for all methods in fourth sample

Active learning iterations	Expected change model	QBC	Uncertainty sampling
3	0.8350	0.7600	0.7975
5	0.7188	0.7188	0.7863
7	0.8263	0.7563	0.7288

Table 13: Comparative performance for all methods in fifth sample

Active learning iterations	Expected change model	QBC	Uncertainty sampling
3	0.8563	0.7913	0.8200
5	0.7925	0.7800	0.8913
7	0.8363	0.7800	0.7900

Table 14: Comparative Performance for all Methods in Sixth Sample

Active learning iterations	Expected change model	QBC	Uncertainty sampling
3	0.9650	0.9550	0.9650
5	0.8950	0.9638	0.9550
7	0.9563	0.9550	0.9575

Finally, we conclude from all previous results that the total accuracy in all iterations is the highest in the case of selecting 50 instances in each iteration. The expected model change method achieves the highest total accuracy than other methods. Sample 6 was the best initial sample for starting active learning procedure for all methods. The iterations 3 and 7 have the highest accuracy.

Summary

In this study, we made a comparison study from more than one side for three of sampling methods- the QBC method, the uncertainty sampling method and the expected model change method. We used three different numbers of iterations, two different numbers of instances that are selecting in each iteration and six different initial samples to study the effect of changing these factors on the effectiveness of the methods used in the sampling phase of active learning in general, we found the total accuracy over all iterations was the highest in the case of selecting 50 instances in each iteration. The sample 6 was the best initial sample for starting active learning procedure for all methods. The iterations 3 and 7 have the highest accuracy. The expected model change method achieves the highest total accuracy than other methods.

Conclusion and Future Work

The main goal of the active learning is selecting of the most informative instances from unlabeled data set to obtain an accurate model, this process falls under the sampling stage which forms the main issue in the active learning, so the focus of our study was on this phase. The most important characteristic of this paper from previous studies is that we implemented more than one strategy to select most informative instances in NIDS to determine the best strategy through making a comparison study in detail from more than one side for these sampling methods. We used three different numbers of iterations, two different numbers of instances that are selecting in each iteration and six different initial samples to know the effect of changing these factors on the effectiveness of the methods used in the sampling phase of active learning. Thus identifying the most appropriate method for NIDS in all cases. We selected the intrusion detection project to apply these sampling methods to active learning based neural network and the KDDCUP 1999 from UCI repository. A small number of labeled data used as initial training set, this set used to build the base classifier. In each iteration, in an active way, the new instances were selected based on the method that used. After that, the new instances added to the initial training data.

Our experiments showed that when comparing the three sampling methods QBC, expected model change and uncertainty sampling. The expected model change method achieves the highest accuracy, the uncertainty sampling method in the second place and then comes the QBC method.

The reason for this superiority of the expected model change method is that it is able to assess the classification score change for whole instances and then choose the instance with the highest effect.

In future research we will study the comparisons in more details, to know influence the selecting each parameter in active learning and apply these methods to other datasets.

Acknowledgment

The authors would like to express their gratitude to the anonymous referees for their valuable comments and suggestions for improving the paper.

Author's Contributions

Ghofran Mohammad Alqaralleh: Implement the code.

Mohammad Aref Alshraideh: Writing the paper.

Ali Alrodan: Review the code and the paper.

Ethics

Authors confirm that there are no ethical issues are involved.

References

Bloodgood, M., 2018. Support vector machine active learning algorithms with query-by-committee versus closest-to-hyperplane selection. Proceedings of the IEEE 12th International Conference on Semantic Computing, Jan. 31-Feb. 2, IEEE Xplore Press, Laguna Hills, CA, USA, pp: 148-155. DOI: 10.1109/ICSC.2018.00029

Cohn, D., L. Atlas and R. Ladner, 1994. Improving generalization with active learning. Machine Learn., 15: 201-221. DOI: 10.1007/BF00993277

Farhan, S., M. Alshraideh and T. Mahafza, 2015. A medical decision support system for ent disease diagnosis using artificial neural networks. Int. J. Artificial Intell. Mechatron., 4: 45-54.

Fu, Y., X. Zhu and B. Li, 2013. A survey on instance selection for active learning. Knowl. Inform. Syst., 35: 249-283. DOI: 10.1007/s10115-012-0507-8

Gilad-Bachrach, R., A. Navot and N. Tishby, 2006. Query by committee made real. Proceedings of the Advances in Neural Information Processing Systems, (IPS' 06), pp: 443-450.

Iglesias, J.E., E. Konukoglu, A. Montillo, Z. Tu and A. Criminisi, 2011. Combining generative and discriminative models for semantic segmentation of CT scans via active learning. Proceedings of the Biennial International Conference on Information Processing in Medical Imaging, (PMI' 11), Springer, Berlin, Heidelberg, pp: 25-36. DOI: 10.1007/978-3-642-22092-0_3

Joshi, A.J., F. Porikli and N. Papanikolopoulos, 2009. Multi-class active learning for image classification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Jun. 20-25, IEEE Xplore Press, Miami, FL, USA, pp: 2372-2379. DOI: 10.1109/CVPR.2009.5206627

Konyushkova, K., R. Sznitman and P. Fua, 2015. Introducing geometry in active learning for image segmentation. Proceedings of the IEEE International Conference on Computer Vision, Dec. 7-13, IEEE Xplore Press, Santiago, Chile, pp: 2974-2982. DOI: 10.1109/ICCV.2015.340

Lewis, D.D. and W.A. Gale, 1994. A sequential algorithm for training text classifiers. Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Jul. 03-06, Springer-Verlag, Dublin, Ireland, pp: 3-12.

Long, J., E. Shelhamer and T. Darrell, 2015. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Jun. 7-12, IEEE Xplore Press, Boston, MA, USA, pp: 3431-3440. DOI: 10.1109/CVPR.2015.7298965

Luo, T., K. Kramer, D.B. Goldgof, L.O. Hall and S. Samson *et al*., 2005. Active learning to recognize multiple types of plankton. J. Machine Learn. Res., 6: 589-613.

Mamitsuka, N.A.H., 1998. Query learning strategies using boosting and bagging. Proceedings of the 15th International Conference on Machine Learning, Jul. 24-27, Morgan Kaufmann Pub., USA, pp: 1-9.

Maystre, L. and M. Grossglauser, 2015. Just sort it! A simple and effective approach to active preference learning. arXiv preprint arXiv:1502.05556.

Olsson, F., 2009. A literature survey of active machine learning in the context of natural language processing. SICS Technical Report.

O'Neill, J., S.J. Delany and B. MacNamee, 2016. Model-based and model-free active learning for regression. Proceedings of the 16th Annual UK Workshop on Computational Intelligence, Sept. 7-9, Lancaster.

Qatawneh, Z., M. Alshraideh, N. Almasri, L. Tahat and A. Awidi, 2017. Clinical decision support system for venous thromboembolism risk classification. Applied Comput. Informat. DOI: 10.1016/j.aci.2017.09.003

Roy, N. and A. McCallum, 2001. Toward optimal active learning through sampling estimation of error reduction. Proceedings of the 18th International Conference on Machine Learning, Jun. 28-Jul. 01, Morgan Kaufmann Publishers Inc., Williamstown, pp: 441-448.

Salah, B., M. Alshraideh, R. Beidas and R. Hayajneh, 2011. Skin cancer recognition by using a neuro-fuzzy system. Cancer Informat., 10: 1-11. DOI: 10.4137/CIN.S5950

Settles, B., 2010. Active learning literature survey. Computer Sciences Technical Report.

Settles, B., M. Craven and S. Ray, 2007. Multiple-instance active learning. Proceedings of the 20th International Conference on Neural Information Processing Systems, Dec. 03-06, Curran Associates Inc., Vancouver, British Columbia, Canada, pp: 1289-1296.

Seung, H.S., M. Opper and H. Sompolinsky, 1992. Query by committee. Proceedings of the 5th Annual Workshop on Computational Learning Theory, Jul. 27-29, ACM, Pittsburgh, Pennsylvania, USA, pp: 287-294. DOI: 10.1145/130385.130417

Singla, A., S. Tschiatschek and A. Krause, 2016. Actively learning hemimetrics with applications to eliciting user preferences. Proceedings of the 33rd International Conference on International Conference on Machine Learning, Jun. 19-24, JMLR.org, New York, pp: 412-420.

Sznitman, R. and B. Jedynak, 2010. Active testing for face detection and localization. IEEE Trans. Patt. Anal. Mach. Intell., 32: 1914-1920. DOI: 10.1109/TPAMI.2010.106

Tong, S. and D. Koller, 2001. Support vector machine active learning with applications to text classification. J. Machine Learn. Res., 2: 45-66.

Tong, S. and D. Koller, 2001. Support vector machine active learning with applications to text classification. J. Mach. Learn. Res., 2: 45-66.

Vezhnevets, A., V. Ferrari and J.M. Buhmann, 2012. Weakly supervised structured output learning for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Jun. 16-21, IEEE Xplore Press, Providence, RI, USA, pp: 845-852. DOI: 10.1109/CVPR.2012.6247757

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

Zhao, W., J. Long, J. Yin, Z. Cai and G. Xia, 2012. Sampling attack against active learning in adversarial environment. Proceedings of the 9th International Conference on Modeling Decisions for Artificial Intelligence, Nov. 21-23, Springer, Girona, Catalonia, Spain, pp: 222-233. DOI: 10.1007/978-3-642-34620-0_21

Zhao, Y., C. Xu and Y. Cao, 2006. Research on query-by-committee method of active learning and application. Proceedings of the International Conference on Advanced Data Mining and Applications, (DMA' 06), Springer, Berlin Heidelberg, pp: 985-991. DOI: 10.1007/11811305_107