Expected likelihood based query for active learning of Gaussian mixture models based classifiers

B H Iswanto*  
Department of Physics, Universitas Negeri Jakarta, Jl. Rawamangun Muka, Jakarta 13220, Indonesia  
*bhi@unj.ac.id

Abstract. Typically, a supervised machine learning requires sufficient labelled data to achieve satisfying performance. In many domains, however, the labelled data are often expensive to obtain and requires laborious human effort. In order to minimize the labelled data, we propose an active learning method by introducing the expected likelihood as a criterion in the query system, especially for the Gaussian mixture models based classifiers. In this scenario, the expected log-likelihood is calculated for all unlabelled examples in a pool data set, then an instant with the lowest value is chosen to query the object of the label. Experiments were performed to evaluate the effectiveness of the method. Empirically, using the two class artificial data set we observed that the proposed method can reduce significantly the number of labelled examples required to achieve a certain performance compared with passive learning.

1. Introduction  
In a passive machine learning framework, the training data are generally gathered by sampling at random from an underlying distribution, then a part of them is labelled according to class labels for training. The training algorithms in such learning system cannot control over information that it receives and accepts whatever training examples are given to them. Of course this learning process requires large labelled training data set to achieve a certain performance goal. On the other hand, in many machine learning problems, data may abundant but obtaining the class labels of data set can be expensive or time-consuming. For example, in domains such as medical image and remote sensing data we require expensive human expertise for labelling each object of the data set. In other domains, such as speech and sound recognition [1], web documents and text mining [2], the labelling of tweet corpus [3] may be time-consuming and need high patient. Therefore, training algorithms that can reduce the labelled data is needed in machine learning, both in supervised and semi-supervised learning.

In this paper, we present the active learning system [4], a machine learning system that attempt to achieve high accuracy using as few labelled instances as possible. The system asks interactively queries in the form of unlabelled instances to be labelled by an oracle, e.g. a human annotator that understands the nature of the problem or some other information source. Under this scenario, the machine learning can achieve greater accuracy with fewer training labels, thereby minimizing the cost of procuring labelled data.  

Generally, active learning methods can be categorized into two different approaches. The first attains optimal solution based on a statistically optimal solution [5-8], e.g. by choosing queries that optimize the learner variance when a new example is added to the training set. The second optimizes the other
criterion in order to maximize expected error reduction [9], e.g. by minimizing the size of the version space. Included in the method are Query-by-Committee [10] and active-SVM algorithm [11]. They tend to choose examples that close to the decision boundaries, so that more suitable for discriminative classifiers. The other methods take the prior data distribution into account to achieve better performance of generative classifiers. One of the others is the method proposed by McCallum and Nigam in McCallum and Nigam for Naive Bayes classifiers in semi-supervised learning frameworks, in which labeled and unlabeled examples are used to train the model [12]. They modified QBC algorithm by weighting the uncertainty measure with the density of the samples. However, the chosen examples are closest to the current classification boundary, which not represent the actual data distribution.

For active learning of generative classifiers based on Gaussian mixture models we propose the expected log-likelihood as a criterion in choosing queries. The algorithm takes the prior data distribution into account to achieve better performance. The method works under a pool-based sampling, in which learners have access to all possible queries.

2. Active learning in machine learning
Active learning is a special case of machine learning in which a learning algorithm is able to interactively query the learner to obtain the class labels of new training data points [4]. Typically, an active learning method consists of two parts: training a model of classifier and query system. First, a model is trained with a very small number of labelled examples selected at random or provided by an oracle. Thereafter, the query system poses queries to the oracle for labels of the specific examples according to their categories. Only a most informative example, i.e. resulting optimal response respect to the criterion, will be selected and requested to the oracle for labelling before being added to the training data set. This learning process is repeated until a certain number of query or satisfactory performance is met. Generally, the active learning algorithm is as follows

```
Input: a set of unlabelled examples, number of request Nreq
Initialize: a few numbers of training examples
Do for t < Nreq
    Training: update model using the current training data set
    Query: select an unlabelled example for querying
    Update the training data set
Return: trained classifier
```

3. Expected likelihood based queries
There are a number of active learning algorithms proposed by authors. Among others is pool-based sampling that assumes the availability of a base pool of instances from which to query objects of labels. The challenge in active learning algorithms is to design the exact strategies used for queries so which the selected examples are most informative, then will maximize the classification accuracy.

In this paper, we propose the expected likelihood as a criterion to choose a new example iteratively for labelling, especially for Gaussian mixture models based classifiers. The proposed method attempt to increase the performance of the classifiers by reducing the number of labelled examples. The method works under pool-based learning, in which learners have access to all possible queries. The method is applied in the query system, in which a new example is selected from the pool with the lowest expected-likelihood for labelling then added to the training data set. The expected-likelihood is computed using actual parameters of the model.

The Gaussian mixture model (GMM) is a probability density function formulated as a simple linear superposition of a number of Gaussian components [13]. We consider now a mixture models with \( M \) components, in which each component represents different class, as follows:
\[ p(x) = \sum_{k=1}^{M} \pi_j p(x | \ell_j) \]

where \( \pi_j \) and \( p(x | \ell_j) \) represents prior probability and class-conditional density of \( j \)-class. Let \( \Psi \) be the actual parameter of the model estimated using training data set \( L = \{ (x_1, z_1), (x_2, z_2), \ldots, (x_N, z_N) \} \) where \( z_i \) is a class label vector of \( x_i \) with \( z_i = (z_{i1}, z_{i2}, \ldots, z_{iL}) \) and \( z_j \in \{ 1, 0 \} \) according to whether \( x_i \) is generated or not from component \( j \)-th of the mixture of Gaussian. According to eq. (1), log-likelihood of the labelled training data set can be written in the form

\[ \mathcal{L}(L | \Psi) = \sum_{x_i \in L} \sum_j z_j l_i \pi_j p(x_i | \ell_j) \]

Suppose \( x^* \) is a new example sampled from a pool of unlabelled examples. When \( x^* \) is added to the training data set \( D \), the expected likelihood of the new training set \( D^* = \{ D \cup x^* \} \) is computed as expectation of log-likelihood as follows

\[ E_x[\mathcal{L}(L^* | \Psi)] = E_x[\mathcal{L}(L | \Psi)] + \sum_j \mathcal{H}(j | x^*) l_i \pi_j p(x_i | \ell_j) \]

Since the expected likelihood of the labelled examples is fixed, not depend on the considering example, therefore, it is sufficient to consider the second term as a selection criterion

\[ E_x[\mathcal{L}(L^* | \Psi)] = \sum_j \mathcal{H}(j | x^*) l_i \pi_j p(x_i | \ell_j) \]

After the training phase using the training data set, query selection can be done by assigning every example \( x \) in the data pool by a score of the expected likelihood. The learning system then selects a new example \( x^* \) with the lowest expected likelihood, namely

\[ x^* = \arg \min_{x \in \mathcal{L}} E_x[\mathcal{L}(x | \Psi)] \]

The model parameter \( \Psi \) is estimated using current training set \( D \) by employing the expectation-maximization (EM) algorithm [14]. The proposed active learning algorithm can be presented as follows

\[ \text{Input:} \text{ a pool of unlabelled examples } P, \text{ number of request } N_{\text{req}} \]
\[ \text{Initialize:} \text{ Labelled examples } D \text{ selected at random from } P \]
\[ \text{Do for } t = 1, \ldots, N_{\text{req}} \]
\[ \text{Training: Run EM to obtain parameter } \Psi^t \text{ given training set } D \]
\[ \text{Query:} \]
\[ \text{For each } x \in P \text{ assign to } x \text{ its expected likelihood } E_x[\mathcal{L}(L^* | \Psi^t)] \]
\[ \text{Select an example } x^* \text{ for which } E_x[\mathcal{L}(L^* | \Psi^t)] \text{ is the lowest.} \]
\[ \text{Update training set and the data pool: } D \leftarrow D \cup \{ x^* \}, \ P \leftarrow P \setminus \{ x^* \}, \text{ where } z^* \text{ given by an oracle.} \]
\[ \text{Return: Model parameter } \Psi^t \]
4. Experiment results
The expected likelihood based active learning algorithm has been evaluated on a two-dimensional artificial data set. The data are binary classification problems, each class was generated by one spherical normal distribution and stored in a pool data set. First, a small number of data were taken randomly to be labelled. For all experiments, the data includes a fixed 60%:40% training/test data set. The training data set was used to initialize the model of classifier and test data set to evaluate the accuracy of the classifiers.

Using the data set we compare our query system against randomly chosen data (passive learning). Each method was implemented on GMM based classifiers with spherical covariance parameters. Approximation of the model parameters was performed using the EM algorithm. The parameters of EM were initialize using k-Means algorithm for the centers of Gaussians and covariance matrices corresponding to the data points in their Voronoi region. The mixing parameters of the mixture were initialized as the proportion of the examples present in each Voronoi sets. Each experiment was conducted by 10-fold validation and performance of the classifiers was evaluated by using the confusion matrix. The results of the experiments are presented in Figure 1. The graph shows the generalization classification error of the proposed active learning and the passive learning method on the 2gauss data set and a pool of 400 examples. Horizontal- and vertical-axis represent the number of queries and classification error of the classifier respectively. The experiments were performed by using the GMM with a different number of component densities. First experiment used one component of the mixture model for each class. The passive learning reached an error rate of 4.50% after 81 queries, meanwhile, the same error was reached only after 36 queries by the proposed active learning. This means that active learning needs only about 44.40% labelled examples less than passive learning. Using five Gaussian components for each class, the proposed method also outperformed the passive, in which the number of examples required only about 39.50% less than the passive to achieve the same classification error. These results indicate that the addition of the data points that maximize the expected likelihood to the training data will allow to gradually build the density function, instead of minimizing the learner's error.

![Figure 1](image-url)

**Figure 1.** Performance of GMM based classifiers trained by using the proposed active learning and the passive learning method.

5. Conclusion
The expected likelihood can be used as a criterion in a query system of active learning frameworks, especially for the Gaussian mixture models based classifiers. Our experiment using two-class artificial
data sets show that the proposed method can reduce significantly until about 40% the number of labelled examples required to achieve a certain performance compared with passive learning.

References
[1] Jondya A G and Iswanto B H 2017 Indonesian’s traditional music clustering based on audio features, Procedia Computer Science 116 174-181
[2] Barfian E, Iswanto B H and Isa S M 2017 Twitter pornography multilingual content identification based on machine learning Procedia Computer Science 116 129-136
[3] Iswanto B H and Poerwoto V 2018 Sentiment analysis on Bahasa Indonesia tweets using unibigram models and machine learning techniques IOP Conference Series Materials Science and Engineering 434 (1)
[4] Aggarwal C, Kong X, Gu Q, Han J and Yu P 2014 Active Learning: A Survey, Data Classification: Algorithms and Applications (CRC Press)
[5] Cohn D A, Ghahramani Z and Jordan M I 1995 Active learning with statistical models Advances Neural Information Processing Systems 7 ed G Tesauro, D Touretzky and T Leen (Massachusset: MIT Press) pp 705-712
[6] Yangab Y and Loog M 2018 A variance maximization criterion for active learning Pattern Recognition 78 358-370
[7] Zhang T and Oles F 2000 A probability analysis on the value of unlabeled data for classification problems International Join Conference on Machine Learning 1191-1198
[8] Jie S, Xin F and Wen S 2009 Active learning for semi-supervised classification based on information entropy International Forum on Information Technology and Applications Chengdu 591-595
[9] Roy N and McCallum A 2001 Toward optimal active learning through sampling estimation of error reduction Proc. 18th International Conf. on Machine Learning (Morgan Kaufmann) pp 441-448
[10] Seung H S, Opper M and Sompolinsky H 9192 Query by committee Proc. 5th Workshop on Computational Learning Theory (Morgan Kaufman, San Mateo) pp 287–294
[11] Tong S and Koller D 2001 Support vector machine active learning with applications to text classification Machine Learning Research 2 45-66
[12] McCallum A and Nigam K 1998 Employing em and pool-based active learning for text classification Proc. of the Fifteenth Int. Conf. ICML 98 359-367
[13] Jiyong M and Wen G 1998 The supervised learning Gaussian mixture model Journal of Computer Science and Technology 13 (5) 471–474
[14] Dempster A P, Laird N M and Rubin D B 1977 Maximum likelihood from incomplete data via the EM algorithm Journal of the Royal Statistical Society Series 39 (1) 1–38