Active learning by increasing model likelihood for Gaussian mixture models based classifiers

B H Iswanto*
Complex Systems Research Group, Department of Physics, Universitas Negeri Jakarta, Jalan Rawamangun Muka, Jakarta 13220, Indonesia

*bhi@unj.ac.id

Abstract. A problem in many classification tasks is to acquire labelled data, while large amounts of unlabelled data are available. One way to overcome these problems is to apply active learning. This technique aims to select the most informative examples and to build optimally classifiers. In this paper, we propose an active learning algorithm for Gaussian mixture model classifiers by maximizing the current model likelihoods. The method assumes a large pool of unlabelled examples is available, the examples are i.i.d according to some underlying distribution, and the labels are distributed according to the class-conditional distribution. Experiments using artificial and real datasets show that, the proposed method outperform and show efficiency in the context of the query number compared with random and expected likelihood queries.

1. Introduction
Nowadays machine learning plays an important role in many tasks of data mining and pattern recognition. Among the tasks is the classification, which is the task of assigning objects to one of several predefined categories. In machine learning, the assigning objects are employed by classification models (or classifiers) built by using supervised learning algorithms. For any supervised learning system to perform well, a large number of labelled instances are needed for training the models. However, for many domains more sophisticated supervised learning tasks, labelled instances are very expensive, difficult, and time-consuming to obtain [1].

Active learning systems attempt to overcome the labeling bottleneck by asking queries in the form of unlabelled instances to be labeled by an oracle (e.g., a human annotator). In this way, the active learner aims to achieve high accuracy using as few labeled instances as possible, thereby minimizing the cost of obtaining labeled data [2]. There are several scenarios in which active learners may pose queries, and different query strategies to decide which instances are most informative. In most active learning experiments, typically assumed that a large pool of unlabeled data is available, the most informative instance is selected to be labeled by the oracle based on an uncertainty sampling query strategy. Some strategies have been proposed in the literature. For example, select query instances that have the least label certainty under the currently trained classifier model [3-6]. Similarly, query-by-committee algorithms [7] and active Support Vector Machines [8] minimize the version space of the models. The other strategies take the prior densities into account to achieve better performance of generative models, such as Transductive Active Learning [9] and active learning with the probabilistic RBF classifier [10]. However, the proposed methods are not taken into account singularities problems that maybe occur during query selection and learning process.
In this paper, we propose the Increasing Model Likelihood, a query method for active learning of Gaussian mixture models based classification. In this method, an example is selected from a pool of unlabeled data that increases the model likelihood, instead of the expected likelihood as proposed in a previous paper [2]. We are primarily interested in how best to use the increasing model likelihood comparing to the expected likelihood in query selection of active learning.

2. Methods

2.1. Gaussian mixture models based classifiers

Classification is a task to find a rule based on observations for assigning an object to one class of a set of classes. In a statistical framework, the classification system can be built by using the optimal Bayes’ decision rule. Suppose that a dataset contains examples from \( c \) classes \( C_1, ..., C_c \). Under this framework, a classifier can be expressed in the form:

\[
\phi(x) = \arg \max_{c_k} P(C_k \mid x)
\]

(1)

According to Bayes’ theorem, the posterior probabilities of class membership are defined as:

\[
P(C_k \mid x) = \frac{P(C_k)p(x \mid C_k)}{\sum_{k=1}^{c} P(C_k)p(x \mid C_k)}
\]

(2)

The prior probability \( P(C_k) \) is given by the fraction of examples from class \( C_k \) in the training dataset.

Here, Gaussian mixture models (GMM) can be applied to estimate the class-conditional densities \( p(x \mid C_k) \) by utilizing training data labeled according to their class. Therefore, the problem of classification can be regarded as class-conditional densities estimation problem, since the class-conditional densities determine decision boundaries and classification accuracy. The GMM is a statistical model to estimate the data distribution by using several Gaussian components [11]. Each component corresponds to a data density or cluster and its parameters describe the corresponding density in terms of its center and spread. More formally, assume that there are \( M \) components of Gaussians and \( N \) objects, \( X = \{x_1, \ldots, x_N\} \). Let the \( j \)-component have a parameter \( \theta_j \) and prior probability \( \pi_j \), \( \Psi \) be set of model parameters, i.e. \( \Psi = \{\pi, \theta\} \). Then, the \( d \)-dimensional data distribution is defined as a linear combination of Gaussians components in the form:

\[
p(x \mid \Psi) = \sum_{j=1}^{M} \pi_j p(x \mid \theta_j)
\]

(3)

where \( p(x \mid \theta_j) \) is Gaussian distribution with parameter \( \theta_j = \{\mu_j, \Sigma_j\} \).

2.2. EM algorithm

Various procedures have been developed for determining the parameters of the Gaussian mixture models, among them based is Expectation and Maximization (EM) algorithm. The EM algorithm is an iterative method by starting from some initial parameter values \( \Psi^{(0)} \). Each iteration consists of two steps: expectation step (E-step) and maximization step (M-step). In E-Step, the EM algorithm calculates the expected likelihood using parameters from the previous iteration. In the M-step, use these parameters to compute a new estimate for the parameters that maximize the likelihood. This iteration continues until the estimates converge to a specific threshold. Since the EM is sensitive to initialization, we employed the LBGU-EM algorithm [12] to determine the model parameters.

2.3. Active learning by the Increasing Model Likelihood

In most supervised learning algorithms, classifiers are trained on a set of labeled examples by sampling at random from an underlying distribution. Such learning methods are called passive learning. The other methods are known as active learning, in which a learner attempt to improve performance by carefully
choosing examples for labeling. Active learning is called also as query learning, since a learner can interactively present one or some unlabeled examples as queries to an oracle, a human annotator, or some other information sources, to label the new examples with the desired classes. To select the informative examples a query algorithm is applied. The goal of active learners is to achieve high classification accuracy using as few labeled examples as possible. The challenge is thus to determine which examples would be most informative, improve the classifier the most if the examples were labeled and used as training examples [13].

In this paper, we present a query strategy, the Increasing Model Likelihood (IML) method. The method assumed a pool of unlabeled examples is available, the examples are i.i.d according to some underlying distribution, and the labels are distributed according to the class-conditional distribution. It would be expected by making a small number of queries the models close to the data distributions and perform effective decision boundaries. Briefly, the learning process may begin with a weak classifier initialized with only a small number of labeled examples. The query algorithm is used to select a new additional example among a pool of examples for labeling. The new labeled example is added to the training dataset, and the learner proceeds in supervised learning. These processes are repeated iteratively until some queries are met.

In the initialization stage, a small number of examples $D$ is taken from the pool data $P$ to be labelled. Then, the query algorithm is applied to select an example $x^*$ from the pool that increases the likelihood. The algorithm consists of two part. First, compute the expected likelihood $Q_x$ of each example in the pool for all possible class memberships $z_k$ as follows:

$$Q_x(x^*) = \mathbb{E}_x [\mathcal{L}(D|\Psi_{D^*})] = \frac{1}{|\mathcal{C}|} \sum_{z \in \mathcal{C}} \mathcal{L}(D|\Psi_{D^*})$$

(4)

where $\mathcal{L}(D|\Psi_{D^*})$ is the model likelihood given example set $D$ and the model parameters $\Psi_{D^*}$. Update the parameters using EM algorithm given new training data $D^* = D \cup x^*$. Next, an informative example $x^*$ is selected that increases the expected likelihood as follows:

$$x^* = \arg \max_{x \in \mathcal{U}} Q_x(x)$$

(5)

The proposed active learning method is presented in Table 1.

| Table 1. Algorithm 1: Increasing model Likelihood active learning. |
|---------------------------------------------------------------|
| **input** a pool of unlabelled data $P$, several queries $N_q$ |
| **begin initialization** labelled dataset $D \leftarrow P \backslash D$ |
| update $P \leftarrow P \backslash D$ |
| initialize model $\Psi_0$ given $D \cup P$ |
| repeat |
| compute $Q_x$ for each $x \in P$ by Eq. (4) |
| select $x^*$ by Eq. (5), label $x^*$ by an oracle |
| update $D \leftarrow D \cup x^*$, $P \leftarrow P \backslash x^*$ |
| update model by EM algorithm given $D$, $\Psi^{t+1} \leftarrow \Psi^t$ |
| until size of $D \leq N_q$ |
| return model parameters $\Psi^{t+1}$ for GMM classifier $\phi(x)$ |
| end |

3. Results and discussion

For the experimental study of the proposed learning method, we used two datasets of two-category: 2Gauss and Heart dataset, each consisting of 100 examples with balance categories. The 2Gauss dataset is an artificial dataset drawn from a multivariate Gaussian distribution of two components with a common covariance matrix. The Heart is a 13-dimensional dataset extracted from the Heart disease data
set of UCI Machine Learning Repository. The performance of the proposed method was evaluated in comparison with passive learning and the Expected Likelihood Query method [2]. Our experiments used classifiers based on two components Gaussian with spherical covariance matrices, each represents one category. We initialized the learning process by using small labeled examples of sampling at random from the pool dataset. The model parameters were estimated using the LBGU-EM algorithm [12]. To cut computation cost in the query process, the iteration of EM was limited to up 50 iterations. The performance of the classifiers was evaluated using the test dataset.

The experiment results are presented in the following figures, in which the generalization errors plotted against the number of queries. It can be seen that on both datasets, 2Gauss and Heart dataset, the Increasing Model Likelihood method performed better than random queries (passive learning) and the Expected Likelihood Query method. Figure 1 show experiment results using the 2Gauss dataset with 100 observations. We compared the Increasing Model Likelihood method with random queries and the Expected Likelihood Query method using correct models. The experiment results show that the Increasing Model Likelihood method performed better than random selection and ELQ in terms of achieving errors with the smallest number of queries. The Increasing Model Likelihood needs only a few queries to achieve certain performance comparing with the others. Similar results also were shown in experiments using the high-dimensional Heart dataset as presented in Figure 2. It is seen, the Increasing Model Likelihood method only requires ten queries, half of needed by Expected Likelihood Query to achieve the same generalization error, while the random method requires four times.

![Figure 1. Error reduction on 2Gauss dataset for random sampling and active learning using EMQ and IML query method. Each point represents a mean of 50 trials for all.](image)

4. Conclusion

We proposed the Increasing Model Likelihood, a query strategy for Gaussian mixture models based classifiers in active learning frameworks. The method selects carefully a new example from a pool of unlabeled data that has the largest contribution to increase the model likelihood. We tested the algorithm using two datasets of binary classification problems. The performance of the method has been compared with other query methods, random selection and the Expected Likelihood Query method, using repeated validation tests. Our experimental results show that the proposed method works well in all of the two-categorical datasets and show efficiency in the context of the query number compared with random selection and the Expected Likelihood Query method.
Figure 2. Error reduction on Heart dataset for random sampling, EMQ and IML.

References
[1] Settles B 2012 Active learning Synthesis Lectures on Artificial Intelligence and Machine Learning vol 6 no 1 (Morgan & Claypool) pp 1-114
[2] Iswanto B H 2019 Expected likelihood based query for active learning of Gaussian mixture models-based classifiers J. Phys. Conf. Ser. 1402 066026
[3] Liu D and Liu Y 2019 An active learning algorithm for multi-class classification Pattern Analysis and Applications 22(3) 1051-1063
[4] Bull L A, Rogers T J, Wickramarachchi C, Cross E J, Worden K and Dervilis N 2019 Probabilistic active learning: An online framework for structural health monitoring Mechanical Systems and Signal Processing 134 106294
[5] Tuer G, Hakkani-Tuer D and Schapire R 2005 Combining active and semi-supervised learning for spoken language understanding Speech Communication 45(2) 171-186
[6] Min F, Liu F-L, Wen L-Y and Zhang Z-H 2019 Tri-partition cost-sensitive active learning through kNN Soft Computing 23(5) 1557-1572
[7] Freund Y, Seung H, Shamir E and Tishby N 1997 Selective sampling using the query by committee algorithm Machine Learning 28 133-168
[8] Yin L, Wang H and Fan W 2018 Active learning based support vector data description method for robust novelty detection Knowledge-Based Systems 153 40-52
[9] Reitmaier T, Calma A and Sick B 2015 Transductive active learning – a new semi-supervised learning approach based on iteratively refined generative models to capture structure in data Information Sciences 293(293) 275–298
[10] Likas A 2008 Semi-supervised and active learning with the probabilistic RBF classifier Neurocomputing 71 13–15
[11] Iswanto B H 2009 Regularization of unlabeled data for learning of classifiers based on mixture models Int. Conf. on Instrumentation, Communication, Information Technology, and Biomedical Engineering (Bandung) (IEEE) pp 1-5
[12] Iswanto B H and Fritzke B 2003 LBGU-EM algorithm for mixture density estimation Between Data Science and Applied Data Analysis (Berlin: Springer-Verlag) pp 252-260
[13] Duda R O, Hart P E and Stork D G 2000 Pattern Classification 2nd ed. (Canada: John Willey & Sons)