Online Chinese Restaurant Process

Chien-Liang Liu
Industrial Technology Research Institute
Computational Intelligence Technology Center
Rm. 709, Bldg. 51, 195, Sec. 4. Chung Hsing Rd., Chutung.
Hsinchu, Taiwan 310, ROC
jackyliu@itri.org.tw

Tsung-Hsun Tsai
National Chiao Tung University
Department of Computer Science
1001 University Road,
Hsinchu, Taiwan 300, ROC
artsuu@gmail.com

Chia-Hoang Lee
National Chiao Tung University
Department of Computer Science
1001 University Road,
Hsinchu, Taiwan 300, ROC
chl@cs.nctu.edu.tw

ABSTRACT

Processing large volumes of streaming data in near-real-time is becoming increasingly important as the Internet, sensor networks and network traffic grow. Online machine learning is a typical means of dealing with streaming data, since it allows the classification model to learn one instance of data at a time. Although many online learning methods have been developed since the development of the Perceptron algorithm, existing online methods assume that the number of classes is available in advance of classification process. However, this assumption is unrealistic for large scale or streaming data sets. This work proposes an online Chinese restaurant process (CRP) algorithm, which is an online and nonparametric algorithm, to tackle this problem. This work proposes a relaxing function as part of the prior and updates the parameters with the likelihood function in terms of the consistency between the true label information and predicted result. This work presents two Gibbs sampling algorithms to perform posterior inference. In the experiments, the online CRP is applied to three massive data sets, and compared with several online learning and batch learning algorithms. One of the data sets is obtained from Wikipedia, which comprises approximately two million documents. The experimental results reveal that the proposed online CRP performs well and efficiently on massive data sets. Finally, this work proposes two methods to update the hyperparameter α of the online CRP. The first method is based on the posterior distribution of α, and the second exploits the property of online learning, namely adapting to change, to adjust α dynamically.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning—*Parameter Learning*; F.1.2 [**Theory of Computation**]: Modes of Computation—*Online computation*

KDD'14, August 24–27, 2014, New York, NY, USA.

http://dx.doi.org/10.1145/2623330.2623636.

General Terms

Algorithms, Experimentation, Theory

Keywords

Adaptive Learning; Chinese Restaurant Process; Nonparametric; Online Learning

1. INTRODUCTION

Online learning has attracted a significant amount of interest in the field of machine learning, and an important family of efficient and scalable machine learning algorithms exist for large-scale applications. Like massive data sets, streaming data sources are commonly important, particularly when data sets are generated in real time, as are many log files, sensor data, and network data. The distributions that underlie the streaming data typically change over time, so the predictive model should be adapted accordingly. One important property of online learning is that the true label of the instance is available after the prediction is made. The true label information can then be used to refine the predictive model, so that subsequent predictions will be closer to the true labels.

Although many online learning algorithms have been developed over the past few decades, most online learning algorithms require the number of classes of the clasification model to be determined in advance of information processing. However, this requirement is unreasonable in real world applications, since the number of classes is usually unknown for massive data sets. Additionally, many application settings require models to be able to deal with a *nonexhaustive* training data set, from which some classes may be missing and therefore absent from the training data set. Detecting new classes from streaming or massive data sets is an important issue in the era of big data, motivating the development herein of a new algorithm to solve this problem. For example, the biosensing problem requires rapid identification of new, emerging classes of microorganisms, which are not represented in the initial training library. Detecting the sudden presence of a new class is an important element of an automated outbreak-identification strategy [11]. E-mail classification is another typical example of online learning. An e-mail system that can identify new classes that are not present in the existing class set and suggest possible labels could constitute an extension of an existing e-mail classification system.

Bayesian nonparametric (BNP) models provide an alternative means of solving this problem; their complexity increases with the number of observed data instances. BNP models can incorporate prior knowledge into the model by placing a prior on an unbounded number of parameters. The Dirichlet process (DP), introduced by Ferguson [14], is a stochastic process that is frequently used as a prior in Bayesian nonparametric statistics. The idea of using a Dirichlet process as the prior of the mixing proportions of a simple distribution was first introduced by Antoniak [2]. The DP is an elegant alternative to parametric model selection using the Dirichlet process mixture model, which comprises a countably infinite number of components [2, 31, 28]. The Chinese restaurant process (CRP) [1] mixture model is one of the representations of the DP mixture model. The CRP induces an exchangeable distribution over partitions, so that the joint distribution is invariant to the order in which observations are assigned to clusters, and each observation can be treated as the last in a sequence of observations. This property yields a very useful representation in the inferences of DP mixture models [36, 28].

This work develops a nonparametric online learning algorithm that is called the online Chinese restaurant process. The proposed algorithm retains the ability of nonparametric models to automatically infer an adequate model complexity from data without determining the number of classes in advance. The online CRP is also an online learning algorithm, indicating that the online CRP is adapted as more data instances are observed. This work proposes a relaxing function to represent part of the prior and updates the parameters according to the likelihood function in terms of the consistency between the true label information and the predicted result. The online CRP is influenced by the true label information, which is used to adjust the parameters of the model, so the online CRP relaxes the assumption of exchangeability used by CRP. Unlike existing online learning algorithms, the online CRP is a nonparametric method, so its model complexity is determined by data. Additionally, the proposed algorithm allows the creation of new classes, and so is more practical and flexible model when applied to massive data sets. In the experiments, three massive data sets are used to assess the online CRP and compare it with several online learning and batch learning algorithms. One of the data sets is obtained from Wikipedia, and includes approximately two million articles. The experimental results demonstrate that the online CRP works well and efficiently on the massive data sets.

The main contribution of this paper is to develop a nonparametric online learning algorithm. To the best of our knowledge, this is the first work that combines Bayesian nonparametric learning with online learning, allowing the classification model to grow with the data rather than requiring that the number of classes be determined in advance of the classification process. This work presents two Gibbs sampling algorithms for making posterior inferences. One of these algorithms is a collapsed Gibbs sampling algorithm, which is more efficient than the other and is used in the experiments. This work proposes two methods for updating hyperparameter α of the online CRP; the first one is based on the posterior distribution of α, while the second exploits the property of online learning, namely adapting to change, to adjust α dynamically.

The rest of this paper is organized as follows. Section 2 then reviews the Chinese restaurant process, and introduces online Chinese restaurant process algorithms. Next, Section 3 summarizes the results of several experiments. Section 4 discusses the experimental results and analyzes the online Chinese restaurant process. Section 5 draws conclusions.

2. ONLINE CHINESE RESTAURANT PROCESS

The goal of clustering is to identify latent information in data, such that objects in the same cluster are more similar to each other than objects from different clusters. The finite mixture model (FMM) is commonly used in clustering applications. The FMM assumes the existence of finite models, each of which is associated with a parameter. Each data point is generated by one of the mixture models. Inferring the parameters associated with the models and identifying which model generated each data point yields the clustering of the data points. However, the FMM must determine the number of models in advance of the clustering. The usual trade off in model order selection problems arises: with too many components, the mixed models may overfit the data, while a mixture with too few components may not be sufficiently flexible to approximate the true underlying model [15]. The Bayesian nonparametric mixture model, which is called a Chinese restaurant process mixture or a Dirichlet process mixture, infers the number of clusters from the data and allows the number of clusters to grow as new data instances are observed [17].

2.1 Notation

This section presents the notation that will be used throughout this study. The notation is based on the metaphor of the CRP, which involves customers, tables and dishes. Customer $\mathbf{x}_i$ corresponds to a data point, while the table is the cluster in the CRP or a class in the online CRP. Accordingly, the customers who sit at the same table correspond to data points in the same class. Meanwhile, the dish at table j is the parameter of the class j, denoted as θ_j, and the number of customers who sit at table j is m_j. The preference of customer $\mathbf{x}_i$ for the dish served at table j is $H(\mathbf{x}_i, \theta_j)$, which is the likelihood that $\mathbf{x}_i$ belongs to class j. Unlike the CRP, the online CRP allows each customer to move to another table after he has first been assigned to a table. The assigned table and the final table at which customer $\mathbf{x}_i$ sits are z_i and y_i, respectively.

The online CRP is a nonparametric method: the number of possible tables is infinite. In this work, the number of occupied tables is denoted as k. The base distribution and the concentration parameter are G_o and α, respectively. Finally, this work proposes a relaxing function $g(\gamma_1, \gamma_2, e_j, f_j)$ to adjust the prior in the online CRP, in which γ_1 and γ_2 are regret rates, and f_j and e_j are used to track the misassignment of the previous $i-1$ customers on table j. Finally, $\mathbf{x}_{-i}$ is introduced to represent all of the customers except for customer i, and $\mathbf{x}_{-i,c}$ represents all of the customers in class c except for customer i. The terms y_{-i} and z_{-i} have similar meanings.

2.2 Chinese Restaurant Process

The Chinese restaurant process (CRP) [1], a discrete-time stochastic process, defines a distribution over partitions that

embodies the assumed prior distribution over cluster structures [30]. The CRP receives the name from the following metaphor. Imagine a Chinese restaurant with an infinite number of tables each with infinite capacity, and a sequence of n customers who enter the restaurant and sit down. The first customer enters the restaurant and sits at the first table. The ith subsequent customer sits at an occupied table, or at the next unoccupied table as follows.

$$P(z_i = j|z_{-i}, \alpha) \propto \begin{cases} \frac{m_j}{i-1+\alpha} & \text{if } j \leq k \\ \frac{\alpha}{i-1+\alpha} & \text{if } j = k+1 \end{cases} \quad (1)$$

The CRP induces an exchangeable distribution over partitions, so that the joint distribution is invariant in the order in which observations are assigned to clusters. Hence, each customer's table assignment z_i can be made by pretending that $\mathbf{x}_i$ is the last person to sit down. As given by Equation (1), the ith customer sits at an occupied table j with a probability that is proportional to the number of already seated customers m_j, or sits at a new table with a probability that is proportional to α. The tables are analogous to clusters, and customers to observations or data points. A larger m_j is more likely to grow. Therefore, the CRP exhibits a clustering property due to a rich-gets-richer phenomenon. Additionally, as a prior on the number of tables, the CRP is a nonparametric model, meaning that the number of occupied tables grows as more customers enter the restaurant.

The exact computation of posterior expectations for a DP mixture model is infeasible when the data include just a few observations. Markov chain methods enable the systematic computation of the posterior distribution of the parameters [12, 36, 26, 13, 28, 20]. Several Markov chain methods for sampling from distribution of a DP mixture model have been presented [28]. Markov chain Monte Carlo (MCMC) sampling methods can be slow to converge and their convergence can be difficult to diagnose. Variational inference is an alternative, giving deterministic approach to approximate likelihoods and posteriors [6, 35].

The CRP offers great flexibility for clustering applications, and has motivated many extensions. For example, the nested Chinese restaurant process (nCrp) [5, 4] extends the CRP to a hierarchy of partitions, allowing arbitrarily large branching factors. While exchangeability is commonly considered to be an advantageous property, much of the data in text, image and audio domains are not exchangeable. The distance-dependent Chinese restaurant process (dd-CRP) [3] relaxes the assumption of exchangeability, and provides a better fit to sequential data and network data. Socher et al. [33] further uses spectral clustering [32, 29] to reduce dimensionality and cluster data using the dd-CRP on spectral space.

2.3 Relaxing Function

The metaphor of the online CRP is an extension of that of the CRP. The first customer sits at the first table, as in the CRP. When a subsequent customer enters the restaurant, the table is assigned differently from that in the CRP. In the online CRP, a waiter tracks the dishes at all the tables, and assigns the customer to a appropriate table based on the waiter's prior knowledge of the dishes and the customer's preference for dishes. However, the customer may be dissatisfied with the table assignment, and move to another table. The restaurant will be in chaos, if the customers move too frequently. To reduce the probability of misassignment, the waiter tracks the movements of customers to adjust the probabilities associated with the assignment of the next new customer to the various tables, and adjusts table dish information to avoid the misassignment. It is noted that table assignments for ith customer comprise two results, one is assigned by waiter, denoted as z_i, and the other one is the final sitting table, represented as y_i. For table j and customer i, f_j and e_j track the misassignment of the previous $i-1$ customers as shown in Equation (2), where $\mathbb{1}$ is an indicator function.

$$\begin{aligned} f_j &= \sum_{a=1}^{i-1} \mathbb{1}\{y_a = j \wedge z_a \neq j\} \\ e_j &= \sum_{a=1}^{i-1} \mathbb{1}\{z_a = j \wedge y_a \neq j\} \end{aligned} \quad (2)$$

$$g(\gamma_1, \gamma_2, e_j, f_j) = (1+\gamma_1)^{f_j}(1-\gamma_2)^{e_j} \quad (3)$$

For table j, $f_j > 0$ indicates that the assignment to table j should be increased, since customers tend to move to table j. Conversely, the assignment to table j should be reduced if $e_j > 0$, since the assignments to table j may cause the customers to move to other tables. Inspired by regret theory [25], this work proposes a relaxing function, Equation (3), where γ_1 and γ_2 are regret rates, and the information about misassignment that is carried by f_j and e_j are viewed as prior knowledge in determining the distribution of table assignment probabilities.

$$\begin{aligned} & P(z_i = j|z_{-i}, \mathbf{x}_i, y_i, \theta, G_0, \alpha) \\ \propto & \begin{cases} g(\gamma_1, \gamma_2, e_j, f_j)\frac{m_j}{i-1+\alpha}H(\mathbf{x}_i, \theta_j) & \text{if } j \leq k \\ \frac{\alpha}{i-1+\alpha}\int H(\mathbf{x}_i, \theta_j)\mathrm{d}G_0(\theta_j) & \text{if } j = k+1 \end{cases} \end{aligned} \quad (4)$$

Equation (4) presents the posterior distribution estimation of online CRP, in which the prior comprises the relaxing function and the prior of the CRP, and $H(\mathbf{x}_i, \theta_j)$ denotes the likelihood that datum $\mathbf{x}_i$ is a member of class j. Notably, the online CRP reduces to the CRP when $f_j = 0$ and $e_j = 0$ for all j. Although the metaphor of the online CRP is similar to that of the CRP, several differences exist between the two processes. First, the CRP is an unsupervised learning method, and it is usually used in clustering applications. Conversely, the online CRP is an online algorithm, in which the label information y_i is available after the prediction of $\mathbf{x}_i$ is made. Second, the prior of the online CRP differs from that of the CRP. Third, the movement of the datum $\mathbf{x}_i$ between classes simultaneously alters the parameters of the classes that are indexed as z_i and y_i in the online CRP.

2.4 Inference and Learning

Figure 1 shows the graphical model of the online CRP, in which the observed random variables include the datum $\mathbf{x}_i$ and its corresponding label y_i. The sampling process is as follows:

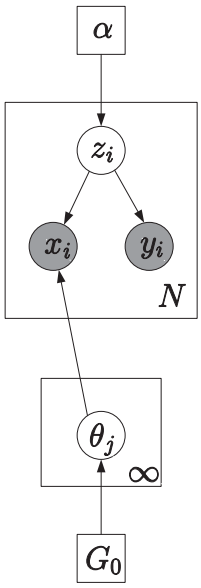

Figure 1: Graphical Model of Online Chinese Restaurant Process

$$\begin{aligned} z_i|\alpha &\sim \text{Mult}(\text{Online CRP}(\alpha)) \\ \theta_{z_i}|G_0 &\sim G_0 \\ \mathbf{x}_i|\theta, z_i &\sim F(\theta_{z_i}) \end{aligned} \tag{5}$$

Equation (5) specifies that z_i is a class indicator variable, referring to the predicted class of datum $\mathbf{x}_i$. For each datum $\mathbf{x}_i$, the predicted label z_i can be sampled using the online CRP. The class parameter θ_{z_i} is drawn from the base distribution, G_0, and each data point $\mathbf{x_i}$ is generated by a distribution F associated with parameter θ_{z_i}. The proposed algorithm is an online algorithm, so the correct label information y_i becomes available after the prediction is made. The model uses z_i and y_i to update class parameters.

Unlike most online learning algorithms, such as Perceptron and Winnow [24], the online CRP is not a mistake-driven model, but updates parameters whenever correct or incorrect predictions have been made. If the prediction is incorrect, $z_i \neq y_i$, then the likelihood that datum $\mathbf{x}_i$ belongs to class z_i is overestimated. To adjust the estimate, a pseudo data point $\mathbf{x}'_i = -b \times \mathbf{x}_i$ is assumed to join class z_i, where $b > 0$ is a constant that specifies the weighting of the pseudo data point. Then, the model samples a new class parameter θ_{z_i} with the posterior probability of θ_{z_i} that is conditional on $\mathbf{x}'_i$, $\mathbf{x}_{-i,z_i}$, α, G_0 and all of the pseudo data points that are associate with table z_i. Conversely, the likelihood that datum $\mathbf{x}_i$ belongs to class y_i is underestimated, so a pseudo data point $\mathbf{x}''_i = b \times \mathbf{x}_i$ rather than $\mathbf{x}_i$ is assumed to join class y_i. Similarly, the model samples a new class parameter θ_{y_i} with the posterior probability of θ_{y_i} that is conditional on $\mathbf{x}''_i$, $\mathbf{x}_{-i,y_i}$, α ,G_0 and all of the pseudo data points that are associate with table y_i. Conversely, if the prediction is correct ($z_i = y_i$), then the model samples a new class parameter θ_{z_i} using the posterior probability of θ_{z_i} that is conditional on $\mathbf{x}_i$, $\mathbf{x}_{-i,z_i}$, α, G_0 and all of the pseudo data points that are associated with table z_i.

The graphical model that is presented in Figure 1 reveals that random variables z and θ are latent variables. In this work, Gibbs sampling is conducted to carry out posterior inference. The Gibbs sampling involves iterations that alternatively draw samples from one of the variables while the others are kept fixed. Accordingly, the conditional posterior distributions of these variables must be derived. The conditional posterior distribution of θ_j is given above. The sampling of z_i is based on the following equation, in which $P(z_i = j|z_{-i}, y_i, \alpha)$ is the prior, and $P(\mathbf{x}_i|\theta_j)$ is the likelihood that θ_j generates datum $\mathbf{x}_i$:

$$P(z_i = j|z_{-i}, y_{-i}, \mathbf{x}_i, \theta_j, \alpha, G_0) \propto P(z_i = j|z_{-i}, y_{-i}, \alpha)P(\mathbf{x}_i|\theta_j) \tag{6}$$

This study uses the prior of online CRP to represent $P(z_i = j|z_{-i}, y_{-i}, \alpha)$, and uses $H(\mathbf{x}_i, \theta_j)$ to represent likelihood, $P(\mathbf{x}_i|\theta_j)$. Hence, Equation (6) becomes Equation (4) in the online CRP. Algorithm 1 is the online CRP algorithm. Notably, the true label information is used to adjust model parameters, so the online CRP is not exchangeable.

Algorithm 1: Online Chinese Restaurant Process

Input: The dispersion parameter α and base distribution G_0, and regret rates γ_1 and γ_2

Initialize m_s, f_s, and e_s as 0 for all $s \in \mathbb{N}$
$k \leftarrow 0$
$b \leftarrow 1$
for $i \leftarrow 1$ **to** ∞ **do**
  Get a data $\mathbf{x}_i$
  if $k = 0$ **then**
    $z_i \leftarrow 1$
  else
    Sample z_i from $P(z_i = j|z_{-i}, \mathbf{x}_i, \theta, G_0, \alpha) \propto$
    $\begin{cases} (1+\gamma_1)^{f_j}(1-\gamma_2)^{e_j}\frac{m_j}{i-1+\alpha}H(x_i, \theta_j) & \text{if } j \leq k \\ \frac{\alpha}{i-1+\alpha}\int H(\mathbf{x}_i, \theta_j)\mathrm{d}G_0(\theta_j) & \text{if } j = k+1 \end{cases}$
    Get label information y_i
    if $z_i = y_i$ **then**
      Sample a new θ_{z_i} from the posterior of θ_{z_i} conditional on G_0 and all the real and pseudo data points associated with the table z_i
      $m_{z_i} \leftarrow m_{z_i} + 1$
    else
      Sample a new θ_{z_i} from the posterior of θ_{z_i} conditional on G_0 and all the real and pseudo data points associated with the table z_i
      Sample a new θ_{y_i} from the posterior of θ_{y_i} conditional on G_0 and all the real and pseudo data points associated with the table y_i
      $m_{y_i} \leftarrow m_{y_i} + 1$
      $e_{z_i} \leftarrow e_{z_i} + 1$
      $f_{y_i} \leftarrow f_{y_i} + 1$
    end
  end
end

2.5 Collapsed Sampling

The marginalization of some variables from a joint distribution always reduces the variance, consistent with the Rao-Blackwell Theorem [21]. In a conjugate context, we can

integrate analytically over θ_j, eliminating θ_j from the algorithm to simplify the sampling process. Then, for each data point $\mathbf{x}_i$, only z_i has to be sampled using Equation (7), in which $F(\theta_j)_{-i}$ is the posterior probability of θ_j conditional on $\mathbf{x}_{-i,j}$, the pseudo data points in table j and G_0.

$$P(z_i = j | z_{-i}, \mathbf{x}_i, y_{-i}, \theta, G_0, \alpha) \propto \begin{cases} g(\gamma_1, \gamma_2, e_j, f_j) \frac{m_j}{i-1+\alpha} \int H(\mathbf{x}_i, \theta_j) \mathrm{d}F(\theta_j)_{-i} & \text{if } j \leq k \\ \frac{\alpha}{i-1+\alpha} \int H(\mathbf{x}_i, \theta_j) \mathrm{d}G_0(\theta_j) & \text{if } j = k+1 \end{cases} \quad (7)$$

Algorithm 2: Collapsed Online Chinese Restaurant Process

Input: The dispersion parameter α and base distribution G_0, and regret rates γ_1 and γ_2

```
1  Initialize m_s, f_s, and e_s as 0 for all s ∈ ℕ
2  k ← 0
3  b ← 1
4  for i ← 1 to ∞ do
5      Get a data x_i
6      if k = 0 then
7          z_i ← 1
8      else
9          Sample z_i from the distribution as listed in
           Equation (7)
10         Get label information y_i
11         if z_i = y_i then
12             Update the sufficient statistics of table z_i on
               joining x_i to table z_i
13             m_{z_i} ← m_{z_i} + 1
14         else
15             Update the sufficient statistics of table z_i on
               joining −b × x_i to table z_i
16             Update the sufficient statistics of table y_i on
               joining b × x_i to table y_i
17             m_{y_i} ← m_{y_i} + 1
18             e_{z_i} ← e_{z_i} + 1
19             f_{y_i} ← f_{y_i} + 1
20         end
21     end
22 end
```

Algorithm 2 shows the collapsed sampling for the online CRP, and Algorithm 2 is used in the experiments herein. The proposed algorithm is an online algorithm, so the algorithm processes datum $\mathbf{x}_i$ sequentially, as in Line 5. The first datum is assigned to the first class, as in Line 7. For subsequent data, z_i are sampled using Equation (7), as in Line 9. Since the number of classes is k, the sampling process must be conducted k times to estimate the posterior probabilities for all the k classes. The model should consider the probability of assigning $\mathbf{x}_i$ to a new class, labeled as $k+1$ in the algorithm. The model selects the most likely class from the $k+1$ classes, which is given by z_i in the algorithm. When the predicted class is available, the model receives the true label information y_i. The model compares z_i with y_i and updates parameters based on the results of the comparison. As in Line 12, the model updates the sufficient statistics of class z_i with the added $\mathbf{x}_i$ and increases the number of data within class z_i by one when $z_i = y_i$. However, classes z_i and y_i update the sufficient statistics of classes z_i and y_i with the addition of a pseudo data point when $z_i \neq y_i$. Besides sufficient statistics, the algorithm updates the information about the number of data and misassignment. The above processes are found in Lines 15-19. Additionally, this work develops two methods for updating hyperparameter α of the online CRP. The first is based on the posterior distribution of α, while the second exploits the property of online learning, namely adapting to change, to adjust α dynamically. They are described in the Appendix. In the experiments herein, the second method is used to adjust α dynamically with an initial value of one.

3. EXPERIMENTS

The implementation of the proposed online CRP assumes that the text of a document follows a multinomial distribution, and that the parameters of the multinomial follow a Dirichlet distribution, such that the conjugate prior can be used to perform collapsed sampling as in Algorithm 2. In this work, three data sets are used to assess system performance and several methods are compared with the proposed algorithm.

3.1 Data Corpora

The 20 Newsgroups, RCV1, and Wikipedia are all popular data sets that are commonly used in text analysis experiments.

- The 20 Newsgroups data set is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. The 20 Newsgroups data set has become popular for experiments involving text applications of machine learning techniques, such as text classification and text clustering.

- Reuters Corpus Volume I (RCV1) [22] is an archive of over 800,000 manually categorized newswire stories that has recently been made available by Reuters, Ltd. for research purposes. RCV1 documents are categorized with respect to three controlled vocabularies - topics, industries, and regions. We follow the category tree and use the second layers as the labels of documents. In the experiments, the documents that appear only in the first layers and those documents that belong to more than one category are removed. The remaining number of categories is 53, and the total number of documents is 534,135.

- The Wikipedia data set is obtained from the Wikimedia database dump, which comprises approximately three million documents. Terms with frequency ranking lower than 50,000 are used as features and the documents without any of those features are removed. The number of remaining documents is approximately two million. The class label information is the category information that is presented in a Wikipedia document. The Wikipedia data set comprises 24 categories.

In the preprocessing stage, the stop words are removed from the data sets, since they fail to provide sufficient information to be useful in the clustering task. Punctuation marks are removed and all English letters are converted to lower case. Finally, stemming is applied to the words.

3.2 Evaluation Measurements

Classification F_1 metric is commonly used to assess the performance of text classification applications. However, the use of classification F_1 assumes that the number of classes is fixed. The proposed algorithm is a nonparametric method in which the number of classes is unbounded. Therefore, the classification F_1 metric could not be used in the experiments herein. The generated classes are compared using the clustering F_1 metric [27]. The clustering F_1 metric considers both precision and recall, which are computed over pairs of documents whose label assignments either agree or disagree. Similarly, one can obtain true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) in terms of the clustering conditions of the pairs. Then, precision, recall and clustering F_1 metric can be obtained from the confusion matrix. Not only clustering F_1 metric, but also error rates are used as evaluation measures. The execution time is also recorded.

3.3 Online Learning Experiments

The experiments herein involve massive data sets, so classification takes a relatively long time. Different approaches are used in the experiments on the 20 Newsgroups data set and the other two data sets. The 20 Newsgroups data set comprises approximately 20,000 documents, and five-fold cross-validation is used on this data set. Experimental results are presented as average and standard deviation, and the mean plus or minus two standard deviations is presented in the tables. The RCV1 and Wikipedia data sets contain numerous documents, so uniform random sampling is used to select training data and testing data from these. The experiments on the two data sets are only conducted once, since each experiment takes a long time. All the methods are implemented using Matlab. The online CRP involves a parameter b that specifies the weighting of the pseudo data point, and its value is one in the experiments.

Hoi et al. [18] developed a library for online learning algorithms that is called LIBOL, which comprises a large family of recently developed state-of-the-art online learning algorithms for large-scale online classification tasks. The LIBOL provides binary and multi-class classification algorithms. The experiments involve multi-class data sets, so multi-class classification algorithms are considered in the comparison of methods. Among these multi-class algorithms, the first-order algorithms include the online gradient descent (OGD) [38] algorithm, passive aggressive (PA) [8] algorithms, and the relaxed online maximum margin algorithm (ROMMA) [23]. The second-order algorithms include adaptive regularization of weight vectors (AROW) [10] algorithm, multi-class confidence weighted (CW) algorithms [9], and soft confidence-weighted learning (SCW) [19] algorithms. The LIBOL further includes variants of some of these algorithms, all of which are used in the experiments. Crammer et al. [8] developed three variants of the passive-aggressive algorithms - PA, PA-I and PA-II. The LIBOL includes the multi-class aggressive ROMMA algorithm called aROMMA. Finally, the LIBOL includes two variants of SCW called SCW-I and SCW-II, respectively.

Besides LIBOL, this work uses additional methods in the experiments, including several variants of Perceptron and online logistic regression [37]. Perceptron is a typical online learning algorithm, and Freund and Schapire [16] proposed several variants of online Perceptron algorithms, including voted Perceptron, average Perceptron, last Perceptron and random Perceptron. This work uses the first three variants in the experiments. The Perceptron algorithm is well known to be able to be employed in a nonlinear way by means of the kernel trick, which depends on only the dot products between the vectors in feature space, and selects the mapping such that these high-dimensional dot products can be computed in the original space by means of a kernel function. The kernelized algorithms of the variants of Perceptron are applied to the data sets. Although these variants of Perceptron belong to batch learning algorithms, they can process data sequentially in the training phase, explaining why they are used in the experiments.

The algorithms that are implemented in LIBOL belong to online learning, while the variants of Perceptron belong to batch learning. Therefore, the experiments herein are conducted using online learning and batch learning settings. The online learning setting concerns the algorithms in the LIBOL, since these do not have a training phase, and they evaluate system performances from when they receive the first data point until they receive the last data point. They are online learning algorithms, so the models that are used in the algorithms can be continually adapted. In contrast, the algorithms of the variants of Perceptron comprise a training phase and a prediction phase, so the batch learning setting uses supervised learning approach to evaluate system performances, in which the algorithms adapt their models during the training phase, and use the trained model to predict testing data without any further change to the model. The proposed online CRP can use online learning and batch learning settings, and we compare online CRP with the comparison algorithms with the two settings.

The first data set to be used in the experiments is the 20 Newsgroups data set, which includes 20 categories and approximately 20,000 documents. Table 1 presents the experimental results obtained using the online learning setting, and Table 2 lists the experimental results obtained using the batch learning setting. All of the algorithms complete classification of the 20 Newsgroups data set. The second data set to be used is RCV1, which is massive. In a practical application setting, classification performance and execution time must both be considered. The methods in the LIBOL fail to complete the task owing to insufficient memory problem, so they are absent from the experimental results, and the batch learning setting is used in the experiments to compare the proposed online CRP with variants of Perceptron algorithms. As indicated in Table 2, the online CRP, the variants of Perceptron and their kernelized extensions can complete the classification in reasonable time. However, the kernelized extensions require the computation of the kernel matrix, which encounters the memory problem with RCV1 and Wikipedia data sets. Additionally, the execution time of the variants of Perceptron are all much longer than the proposed algorithm on the RCV1 and Wikipedia data sets. Therefore, the remaining experiments were performed using the online CRP and last Perceptron. Table 3 and Table 4 present the experimental results obtained using the RCV1 and Wikipedia data sets, respectively. Table 4 presents only the error rate, since clustering F_1 metric depends on pairwise comparisons between all of the classification results. The number of documents in Wikipedia is approximately two millions, yielding a large number of pairwise combinations.

Table 1: Experimental Results on 20 Newsgroups with Online Learning Setting

	Error Rate	Cluster F_1	Execution Time (sec)
Online CRP	0.2471 ± 0.0057	0.6056 ± 0.0045	277 ± 7
ROMMA	0.9451 ± 0.0123	0.0950 ± 0.0003	92 ± 0
aROMMA	0.9449 ± 0.0128	0.0950 ± 0.0003	84 ± 1
OGD	0.3931 ± 0.0396	0.4056 ± 0.0432	85 ± 1
PA	0.9432 ± 0.0170	0.0950 ± 0.0003	84 ± 0
PA-I	0.2762 ± 0.0042	0.5507 ± 0.0058	82 ± 2
PA-II	0.2693 ± 0.0036	0.5605 ± 0.0046	82 ± 2
CW	0.2362 ± 0.0027	0.6076 ± 0.0040	$48,382 \pm 1,167$
AROW	0.4423 ± 0.0197	0.3322 ± 0.0232	$46,865 \pm 675$
SCW-I	0.2358 ± 0.0036	0.6083 ± 0.0047	$46,276 \pm 432$
SCW-II	0.2345 ± 0.0034	0.6102 ± 0.0054	$46,360 \pm 455$

Table 2: Experimental Results on 20 Newsgroups with Batch Learning Setting

	Error Rate	Cluster F_1	Execution Time (sec)
Online CRP	0.2010 ± 0.0100	0.6790 ± 0.0224	313 ± 9
Last Perceptron	0.3904 ± 0.0313	0.4005 ± 0.0302	256 ± 27
Voted Perceptron	0.3206 ± 0.0127	0.4938 ± 0.0187	267 ± 11
Average Perceptron	0.3928 ± 0.0391	0.4100 ± 0.0460	269 ± 9
Kernel Last Perceptron	0.3903 ± 0.0304	0.4060 ± 0.0395	63 ± 3
Kernel Voted Perceptron	0.3100 ± 0.0287	0.5085 ± 0.0399	407 ± 19
Kernel Average Perceptron	0.3936 ± 0.0209	0.3972 ± 0.0360	404 ± 5
Online Logistic Regression	0.3177 ± 0.0162	0.5010 ± 0.0196	$113,527 \pm 2,755$

Table 3: Experimental Results on RCV1 Data Set

	Error Rate	Cluster F_1	Execution Time
Online CRP	0.1477	0.8517	54,958
Last Perceptron	0.1176	0.8905	259,512

Table 4: Experimental Results on Wikipedia Data Set

	Error Rate	Execution Time
Online CRP	0.3602	112,030
Last Perceptron	0.3888	2,249,000

4. DISCUSSION AND ANALYSIS

Table 1 reveals that the online CRP performs similar to some of the second-order algorithms in the LIBOL. In the LIBOL, the second-order online learning methods such as CW, SCW-I and SCW-II perform well on the 20 Newsgroups data set, but their execution times greatly exceed those of first-order methods. For the high-dimensional data sets, the calculation of covariance matrix is highly complex and therefore difficult. Table 2 and Table 4 show that the online CRP outperforms the variants of Perceptron on 20 Newsgroups and Wikipedia data sets.

Last Perceptron outperforms the proposed method on the RCV1 data set as shown in Table 3, but its execution time is much longer. In the experiments on the compared methods that involve binary classifiers, such as Perceptron variants, the one-against-all technique is used to extend these methods to multi-class scenarios. Therefore, the number of classes should be available when a k-class problem is decomposed into a series of two-class problems. In contrast, the proposed online CRP grows to accommodate the complexity of the data without requiring the number of classes in advance, providing flexibility in processing massive or streaming data sets. To evaluate the proposed algorithm when used on a streaming data set, the proposed algorithm is applied to a sorted 20 Newsgroups data set, in which the documents are sorted by timestamps. The experimental results are almost the same as those presented in Table 2.

The experiments involve balanced and imbalanced data sets, and the experimental results demonstrate that the proposed method can perform well on the two kinds of data sets. The experimental results also indicate that the proposed method can complete classification within reasonable time, even though the Wikipedia data set comprises two million documents. The proposed method is analyzed, compared with supervised learning methods, and the performance is assessed with different numbers of training documents. Finally, the settings of the parameters are discussed.

4.1 Analysis of Online CRP Algorithm

The proposed online CRP is an online learning algorithm. The prediction depends on the posterior probabilities of all the existing classes and a new class, so the model need re-

tain sufficient statistics only for these classes. When the prediction of a data point is completed, the model updates the required sufficient statistics rather than undergoing retraining. Thus, the online CRP can process the streaming data in almost real-time. Additionally, the model is memory-efficient, since only the model and a single data point have to be retained in memory.

The sampling process that is listed in Line 9 of Algorithm 2 shows that the computational cost of each sampling update is proportional to the number of classes, k. The value of k approaches $\alpha \log(n)$ asymptotically as n approaches infinity [2, 34]. Meanwhile, the worst case for the model to update parameters is $z_i \neq y_i$, since the model should update the parameters of class z_i and y_i. The time complexity for updating model parameters is a constant, since the update involves only the sufficient statistics. Therefore, the total time complexity for processing n data points is $O(k \times n) = O(\alpha n \log(n))$. In most cases, the time complexity approaches $O(n)$ since $n \gg k$. The experimental results are consistent with the analysis.

4.2 Comparison with Supervised Learning

The proposed method is compared with supervised learning methods on the three data sets using SVM, logistic regression, and Naive Bayes methods. Table 5 presents the experimental results on the 20 Newsgroups. Experiments on SVM are conducted using libsvm [7] with linear kernel. The above supervised learning methods require that model parameters to be determined in classifying documents, so a subset of the data is used as a validation set to determine appropriate values of parameters. Table 6 and Table 7 present the experimental results concerning RCV1 and Wikipedia data sets, respectively.

The experimental results reveal that SVM generally outperforms other methods, but it fails to classify the Wikipedia data set, owing to the computation of the kernel matrix. The proposed online CRP outperforms Naive Bayes, and is comparable to logistic regression. The above three supervised learning methods require that the number of classes is known in advance of classification process. Although many differences exist between the proposed online CRP and the CRP, the proposed online CRP and the CRP share an important property: both are nonparametric methods. The number of classes in the online CRP is determined by the data, and provides a practical and flexible model for dealing with massive data sets.

4.3 Assessment of Performance with Different Numbers of Training Documents

Experiments were conducted herein to determine whether the online CRP benefits from more training examples. Although the proposed method is an online learning algorithm, the online CRP can be very easily converted into a supervised learning algorithm. We can use the online CRP to process available labeled examples sequentially, yielding a classification model. The classification model can then be used to classify the testing examples sequentially, and the classification performance can be evaluated.

In the experiments, 15,000 documents were randomly selected from the RCV1 data set as testing documents. To analyze further the impact of the number of training documents on classification performance, various numbers of labeled examples are used to train the classification model. Figure 2 presents the experimental results, in which the testing error rates for various numbers of training documents are recorded. The experimental results indicate that the testing error generally declines as the number of training documents increases.

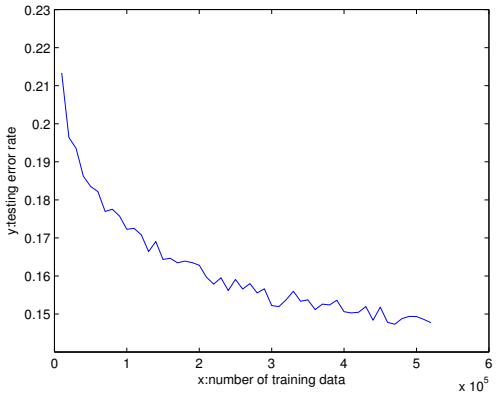

Figure 2: Testing Error on RCV1 Data Set

4.4 Parameter Setting

This section conducts experiments on 20 Newsgroups and evaluates performance with various settings of the parameters γ_1, γ_2, b, and α. First experiments with various values of γ_1 and γ_2 are conducted and the performance is evaluated. The settings are simplified by setting γ_1 equal to γ_2. When their values are less than 0.5, clustering F_1 metric is around 0.66. In contrast, increasing γ_1 and γ_2 significantly degrades performance. A small value of regret rate was preferred in the experiments. Second, experiments were performed using various values of b, which specifies the weighting of the pseudo data point. The experimental results indicate that when b is ten, the best F_1 value, 0.6949, is obtained. If b is larger than 100 or less than 0.1, the performance will be poor. The above evaluation suggests the setting of parameter b. Finally, the technique that is mentioned in the Appendix is used to enable the model to automatically determine α from the data. Experiments with two initial values, 1 and 1000, are conducted. The above experimental settings can evaluate whether the proposed scheme can automatically adjust α to maintain performance. The experimental results indicate that their performance is almost the same as that in Table 2. Therefore, even if the initial value of α is unreasonable, the proposed dynamic setting scheme can adjust α according to the data.

5. CONCLUSION

This work develops a nonparametric online learning algorithm that adjusts the parameters and complexity of the model as more data instances are observed. The online CRP is more flexible in dealing with massive and streaming data sets than are traditional online learning algorithms. This work proposes a new prior and model parameter updating mechanism that are based on the property of online learning. The experimental results indicate that the online CRP performs well and efficiently on massive data sets. A simple mechanism for converting the online CRP into a supervised learning algorithm is also proposed, and experiments are

Table 5: Compared with Supervised Learning on 20 Newsgroups

	Error Rate	Cluster F_1	Execution Time
Online CRP	0.2010 ± 0.0100	0.6790 ± 0.0224	312.8182 ± 8.9205
SVM	0.2030 ± 0.0092	0.6516 ± 0.0104	$5,931.9255 \pm 56.2376$
Logistic Regression	0.1793 ± 0.0195	0.6929 ± 0.0285	$8,576.3864 \pm 315.3372$
Naive Bayes	0.2276 ± 0.0120	0.6139 ± 0.0217	$2,909.1719 \pm 30.3903$

Table 6: Compared with Supervised Learning on RCV1 Data Set

	Error Rate	Cluster F_1	Execution Time
Online CRP	0.1477	0.8517	54,958
SVM	0.0748	0.9318	346,112
Logistic Regression	0.1409	0.8394	12,367
Naive Bayes	0.1800	0.7850	143,831

Table 7: Compared with Supervised Learning on Wikipedia Data Set

	Error Rate	Execution Time
Online CRP	0.3602	112,030
Logistic Regression	0.3488	40,952
Naive Bayes	0.3887	389,209

performed with various numbers of training examples. The experimental results indicate that the online CRP benefits from more training examples, but the performance gain is small when the number of training examples exceeds a specific threshold. This finding gives a direction for the future work: the online CRP with the stochastic gradient descent (SGD) trick can be used to deal with massive data sets when training time is an important issue. Central to the proposed method is nonparametric online learning approach, which we argue provides a more flexible and realistic means of dealing with streaming data sets. The nonparametric models make weaker assumptions and let the data "speak for themselves", and online learning allows the model to adapt to observed data.

6. ACKNOWLEDGMENTS

This work was sponsored by Ministry of Economic Affairs, Taiwan, R.O.C. through project No. D352B23100 conducted by Industrial Technology Research Institute (ITRI).

7. REFERENCES

[1] D. Aldous. Exchangeability and related topics. In *École d'Été St Flour 1983*, pages 1–198. Springer-Verlag, 1985. Lecture Notes in Math. 1117.

[2] C. E. Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *Annals of Statistics*, 2(6), November 1974.

[3] D. M. Blei and P. I. Frazier. Distance dependent chinese restaurant processes. *J. Mach. Learn. Res.*, 12:2461–2488, Nov. 2011.

[4] D. M. Blei, T. L. Griffiths, and M. I. Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *J. ACM*, 57(2):7:1–7:30, Feb. 2010.

[5] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *NIPS*, 2003.

[6] D. M. Blei and M. I. Jordan. Variational inference for dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–144, 2006.

[7] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[8] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer. Online passive-aggressive algorithms. *J. Mach. Learn. Res.*, 7:551–585, Dec. 2006.

[9] K. Crammer, M. Dredze, and A. Kulesza. Multi-class confidence weighted algorithms. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 496–504, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[10] K. Crammer, A. Kulesza, and M. Dredze. Adaptive regularization of weight vectors. *Machine Learning*, 91(2):155–187, 2013.

[11] M. Dundar, F. Akova, A. Qi, and B. Rajwa. Bayesian nonexhaustive learning for online discovery and modeling of emerging classes. In *ICML*, 2012.

[12] M. D. Escobar. Estimating normal means with a dirichlet process prior. *Journal of the American Statistical Association*, 89(425):268âĂŞ277, 1994.

[13] M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577âĂŞ588, 1995.

[14] T. S. Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, page 209âĂŞ230, 1973.

[15] M. A. T. Figueiredo and A. K. Jain. Unsupervised learning of finite mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(3):381–396, Mar. 2002.

[16] Y. Freund and R. E. Schapire. Large margin classification using the perceptron algorithm. *Mach. Learn.*, 37(3):277–296, Dec. 1999.

[17] S. J. Gershman and D. M. Blei. A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1):1–12, Feb. 2012.

[18] S. C. Hoi, J. Wang, and P. Zhao. *LIBOL: A Library for Online Learning Algorithms*. Nanyang Technological University, 2012.

[19] S. C. H. Hoi, J. Wang, and P. Zhao. Exact soft confidence-weighted learning. In *ICML*. icml.cc / Omnipress, 2012.

[20] H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

[21] S. M. Kay. *Fundamentals of statistical signal processing: estimation theory*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.

[22] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, 5:361–397, Dec. 2004.

[23] Y. Li and P. M. Long. The Relaxed Online Maximum Margin Algorithm. *Mach. Learn.*, 46(1-3):361–387, Jan. 2002.

[24] N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Mach. Learn.*, 2(4):285–318, Apr. 1988.

[25] G. Loomes and R. Sugden. Regret theory: An alternative theory of rational choice under uncertainty. *Economic Journal*, 92(368):805–24, 1982.

[26] S. N. MacEachern. Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics - Simulation and Computation*, 23(3):727–741, 1994.

[27] C. D. Manning, P. Raghavan, and H. Schtze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[28] R. M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, pages 249–265, 2000.

[29] A. Y. Ng, M. I. Jordan, and Y. Weiss. On Spectral Clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14*, pages 849–856. MIT Press, 2001.

[30] J. Pitman. Combinatorial stochastic processes. Technical Report 621, Dept. Statistics, U.C. Berkeley, 2002. Lecture notes for St. Flour course, July 2002.

[31] C. E. Rasmussen. The infinite gaussian mixture model. In *NIPS*, pages 554–560, 1999.

[32] J. Shi and J. Malik. Normalized Cuts and Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000.

[33] R. Socher, A. L. Maas, and C. D. Manning. Spectral chinese restaurant processes: Nonparametric clustering based on similarities. *Journal of Machine Learning Research - Proceedings Track*, 15:698–706, 2011.

[34] E. B. Sudderth. *Graphical models for visual object recognition and tracking*. PhD thesis, Cambridge, MA, USA, 2006. AAI0809973.

[35] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1-2):1–305, Jan. 2008.

[36] M. West, P. Müller, and M. D. Escobar. Hierarchical priors and mixture models, with applications in regression and density estimation. *Aspects of Uncertainty: A Tribute to D. V. Lindley*, pages 363–386, 1994.

[37] F. Zhdanov and V. Vovk. Competitive online generalized linear regression under square loss. In *Proceedings of the 2010 European conference on Machine learning and knowledge discovery in databases: Part III*, ECML PKDD'10, pages 531–546, Berlin, Heidelberg, 2010. Springer-Verlag.

[38] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *ICML*, pages 928–936, 2003.

APPENDIX

A. SAMPLE α WITH POSTERIOR PROBABILITY

As in the CRP mixture, the concentration parameter α can be sampled from the posterior probability conditional on $\mathbf{z}$ as indicated in Equation (8). Equation (9) gives the likelihood that α generates $\mathbf{z}$, where n is the number of data points, $g(\gamma_1, \gamma_2, e_{z_i}, f_{z_i})$ denotes the relaxing function for $\mathbf{x}_i$ joining to table z_i, m_{z_i} represents the number of people who sit at the table z_i when $\mathbf{x}_i$ joins it. If the prior distribution of α is available, we can sample α using Equation (8) and the observed data points.

$$P(\alpha|\mathbf{z}) \propto P(\mathbf{z}|\alpha)P(\alpha) \tag{8}$$

$$P(\mathbf{z}|\alpha) = \prod_{i=1}^{n} \frac{\mathbb{1}\{m_{z_i} = 0\}\alpha + \mathbb{1}\{m_{z_i} \neq 0\} g(\gamma_1, \gamma_2, e_{z_i}, f_{z_i}) m_{z_i}}{\alpha + \sum_{j=1}^{k} g(\gamma_1, \gamma_2, e_j, f_j) m_j} \tag{9}$$

B. DYNAMIC α

This work proposes to use a dynamic approach that is adapting to change to determine the value of α. The value of α specifies the probability of creating a new class. If the class prediction of $\mathbf{x}_i$ is a new class, but the true class of $\mathbf{x}_i$ is in one of the existing ones, the value of α should be reduced. In contrast, the value of α should be increased if the predicted class is in one of the existing classes, but the true class label is new. This approach is used in the experiments to adjust α dynamically.