A. Analysis of Generalization Error

We represent the generalization error of our ensemble model in Eq. 10, where e_i indicates the generalization error of the ensemble model on each data chunk D_i, N indicates the number of data chunks, and R^m indicates the feature space with m selected features.

$$\bar{g}e = \Sigma_{i=1}^{N} \Delta ge_i \tag{10}$$

$s.t., \Delta ge_i = \Sigma_{X_j \in D_i}(P(F_{R^m(D_i)}(X_j) \in Y_j) - max_{k \neq j}$ $P(F_{R^m(D_i)}(X_j) \in Y_k)) \simeq \Sigma_{X_j \in D_i} P(F_{R^m(D_i)}(X_j) \notin Y_j)$.

To make sure of the prediction accuracy, it is possible to guarantee under a realistic assumption that our ensemble model generated incrementally using N data chunks are asymptotically arbitrarily close to the one produced by the batch learner using all instances. In order to make this statement precise, we define the difference of the generalization error between our models generated using data chunks and using all instances, denoted as Eq. 11.

$$\Delta G = \Sigma_{i=1}^{N} \Delta ge_i - \Delta ge \tag{11}$$

$s.t., \Delta ge = \Sigma_{X_j \in D}(P(F_{R^m(D)}(X_j) \in Y_j) - max_{j \neq k}$ $P(F_{R^m(D)}(X_j) \in Y_k)) \simeq \Sigma_{X_j \in D} P(F_{R^m(D)}(X_j) \notin Y_j)$.

In this paper, we introduce the statistical method known as the Hoeffding Bound (or additive Chernoff bound) inequation popularly used in the data stream classification, that is, consider a real-valued random variable e whose range is R, in this paper, the variable e indicates the probability of classification error, its range is one. Suppose we have made n independent observations of this variable in each data chunk, and computed their mean $\bar{e}$. The Hoeffding bound states that, with probability 1-δ, the true mean of the variable is at least $\bar{e} - \epsilon$, where $\epsilon = \sqrt{\frac{R^2 ln(1/\delta)}{2n}}$. According to Hoeffding Bound inequation, we can get the following inference in Eq. 12, where $N \cdot n$ indicates the number of all instances, N is the total number of data chunks and n indicates the average number of instances in a data chunk.

$$P(\frac{\Delta G}{N \cdot n} < \epsilon) \leq 1 - \delta \tag{12}$$

Proof: $P(\frac{\Delta G}{Nn} < \epsilon) = P(\frac{1}{Nn}(\Sigma_{i=1}^{N} \Delta ge_i - \Delta ge) < \epsilon)$

$$= P((\frac{1}{Nn}\Sigma_{i=1}^{N}\Sigma_{X_j \in D_i} P(F_{R^m(D_i)}(X_j) \notin Y_j) - \frac{1}{Nn}\Sigma_{X_j \in D} P(F_{R^m(D)}(X_j) \notin Y_j)) < \epsilon)$$

$$= P((\frac{1}{n}\Sigma_{X_j \in D_i} P(F_{R^m(D_i)}(X_j) \notin Y_j) - \frac{1}{n}\Sigma_{X_j \in D_i} P(F_{R^m(D)}(X_j) \notin Y_j)) < \epsilon)$$

Let $Z_j = P(F_{R^m(D_i)}(X_j) \notin Y_j)$, $Z = P(F_{R^m(D)}(X_j) \notin Y_j)$, because the former is evaluated by the model learned from some instances instead of all instances used in the latter, we get $E(Z_j) \geq E(Z)$, where $E(\cdot)$ indicates the expectation function. In this case, $P(\frac{\Delta G}{Nn} < \epsilon) = P(\frac{1}{n}\Sigma_j^n Z_j - \frac{1}{n}\Sigma_j^n E(Z) < \epsilon) \leq P(\frac{1}{n}\Sigma_j^n Z_j - \frac{1}{n}\Sigma_j^n E(Z_j) < \epsilon)$. According to Hoeffding Bound inequation, $P(\frac{1}{n}\Sigma_j^n Z_j - \frac{1}{n}\Sigma_j^n E(Z_j) < \epsilon) = 1 - \delta$, it is hence proved.

In the above analysis, the critical issue is to decide how many instances are necessary in a data chunk D_i (namely the value of n) for a lower value of ΔG. In terms of the statistical result in the Hoeffding Bound, let us specify the significant level δ = 0.05, if we want to maintain the value of ΔG lower than $\epsilon = 0.1$, we only require 150 instances at most in each data chunk. In the following experiments, we specify n =200.