This paper proposes a series of algorithms using unlabeled text data to supplement scarce labeled text data to improve models for classification tasks of text documents, which is inspired by (1) high cost of hand-labeling, (2) availability of abundant unlabeled data, and (3) information about the joint distribution over features being contained in unlabeled data. According to assumptions about the generative process of text documents, a classic probabilistic classifier, naïve Bayes, is applied to estimate model parameters using labeled training data. Then, naïve Bayes is combined with Expectation-Maximization method (EM) for labeled data combined with unlabeled data: the classifier is used to assign class labels based on probabilistic weights for the unlabeled data, and new model parameters are estimated using all labeled and unlabeled data. This process is iterated until learned parameters converge. Furthermore, to address the negative influence of incorrect assumptions about real-world data, there are 2 augmentation to basic EM scheme: “weighting the unlabeled data” (by reducing the effect of violated assumptions on parameter estimation) and “multiple mixture components per class” (by removing/weakening a restrictive assumption).

As a basic framework, a probabilistic generative model takes two assumptions for the generative process: (1) the data are produced by a mixture model, and (2) there is a one-to-one correspondence between mixture components and classes. In this way, the likelihood of a document could be expressed as the sum of total probabilities (weight of a mixture component × the probability of the document generated by the mixture component) over all mixture components.

With additional “word independence assumption”: words of a document are generated independently of context (other words in the same documents given the class label) and their positions within the document, people develop naïve Bayes model for text classification tasks. In this viewpoint, the likelihood of a document is proportional to the product of likelihood of each word within the document. Thus, the parameters for each mixture component are a multinomial distribution over words, which constitute the parameter set of the generative model with prior probabilities over these multinomials for mixture components. For model training (estimation of parameters), basic ideas including (1) maximum a posteriori parameter estimation, (2) Bayes’ rule and (3) the method of Lagrange multipliers are applied, and “ratios of count” formulae for (1) word probability given class and (2) class prior probabilities are derived. Interestingly, naïve Bayes performs well although some assumptions could not be satisfied in real life.

Then, how can abundant unlabeled data be incorporated with Expectation-Maximization method to supplement scarce labeled data? As mentioned before, naïve Bayes is trained with labeled data as initialization of parameter estimation (“priming” M-step). Then, the following 2 steps are looped until convergence of parameter estimation: the learned model parameters are used to assign class labels for all unlabeled data (E-step), and new model parameters are estimated using all labeled and unlabeled data (M-step). It is great that this process is guaranteed to find model parameters having non-decreasing likelihood than that of previous iterations. Furthermore, notice that complete log likelihood replaces incomplete log likelihood approximately in practice for the sake of computational tractability, which gives a local optimum by a hill climbing procedure.

In contrast to naïve Bayes, the performance of EM is negatively affected when assumptions about the generative process are violated. Fortunately, the authors come out with 2 approaches to augment the EM method from different angels. The first, weighting the unlabeled data, adds a parameter to modulate the degree to which EM weights the unlabeled data. It modifies the formulae forms of parameter estimation (in M-step). The second, “multiple mixture components per class”, assumes a many-to-one correspondence between mixture components and classes instead of the original one-to-one relations, is beneficial for capturing co-occurrence patterns of words belonging to different sub-topics of same topics as well as relaxing an assumption of the generative model. In this way, restriction and renormalization of membership probability estimates of labeled data are needed (in E-step). It is worth mentioning that cross validations are conducted to determine the value of the modulating parameter and the number of mixture components per class.

Finally, the authors do some experiments in 3 real-world text classification tasks (20 Newsgroups, WebKB, Reuters) and verify that the proposed model gives better performance than state-of-the-art models. Also, the parameter decision process of augmented EM is included. Furthermore, the authors point out several directions of future work with unlabeled data such as active learning, incremental learning and technique application to similar domains.