Using EM for Text Classification on Arabic

Ghassan Kanaan
Arab Academy for Banking and Financial Sciences Amman Jordan
ghkanaan@aabfs.org

Mustafa Yaseen
Al- Ahlyya Amman University Amman Jordan
myaseen@ammanu.edu.jo

Riyad Al-Shalabi
Arab Academy for Banking and Financial Sciences Amman Jordan
rshalabi@aabfs.org

Bashar Al-Sarayreh
Arab Academy for Banking and Financial Sciences Amman Jordan
bsarayreh @aabfs.org

Ashraf Bany Mustafa
Arab Academy for Banking and Financial Sciences
Amman Jordan

Abstract

Text classification is getting more attention and there is an increased need for text classification technique that provides automatic, fast, and accurate semi-supervised classification with the least human interaction with such systems. In our work we incorporated a well experimented technique for classification that makes use of the famous EM algorithm in training the classifier to be more effective on Arabic language.

Introduction

An expectation-maximization (EM) algorithm is used in statistics for finding maximum likelihood estimates of parameters in probabilistic models, where the model depends on unobserved latent variables. EM alternates between performing an expectation (E) step, which computes an expectation of the likelihood by including the latent variables as if they were observed, and a maximization (M) step, which computes the maximum likelihood estimates of the parameters by maximizing the expected likelihood found on the E step. The parameters found on the M step are then used to begin another E step, and the process is repeated (Dempster et al. 1997). The EM algorithm is a popular class of iterative algorithms for maximum likelihood estimation for problems involving missing data. It is often used to fill the missing values in the data using existing values (Xiaoli Li et al. 2004).

In this method a set of labels used as representative for the documents are built for each class. It then uses these documents to label a set of documents for each class from a set of unlabeled documents to form the initial training set. The EM algorithm is then applied to build the classifier. The key issue of the approach is how to obtain a set of representative documents for each class to solve the problem; it uses the EM iterativealgorithm. Through this technique the classifiers learn by more of the unlabeled documents to classify the new ones (Xiaoli Li et al. 2004).
Text classification is the process of assigning predefined category labels to new documents based on the classifier learnt from training examples, in which document classifier is first trained using documents with reassigned labels or classes picked from a set of labels, which we call the taxonomy or catalog. Once the classifier is trained, it is offered test documents for which it must guess the best labels. Depending on the application, the label may correspond to a broad topic (e.g., a topic in the Yahoo! directory), a product category, or a user's personal taste in books, CDs, or Web sites (Sarawagi et al. 2003). Classifying textual data is considered as a very difficult task, and by the introduction of the web text classification became more difficult (Hirsh et al. 2000).

This problem of automatically classifying text documents is of great practical importance given the massive volume of online text available through the World Wide Web, Internet news feeds, electronic mail, corporate databases, medical patient records and digital libraries. Existing statistical text learning algorithms can be trained to approximately classify documents given a sufficient set of labeled training examples. These text classification algorithms have been used to automatically catalog news articles and web pages automatically learn the reading interests of users and automatically sort electronic mails (Nigam et al. 2000). Document classification may appear in many applications such as Email filtering, mail routing, news monitoring, Narrowcasting and content classification.

Applications of various machine-learning techniques attempted to solve this problem which includes categorization of Web pages into sub-categories for search engines, and classification of news articles by subject. These Machine learning classification programs, such as C4.5 and RIPPER, suffer from the limitation that the learning mechanism is based solely upon previously classified data. In traditional classification techniques, training examples are labeled with the same set of pre-defined category or class labels and labeling is often done manually. Many of these text classification techniques have been proposed and implemented e.g., the Rocchio algorithm, the naive Bayesian method (NB), support vector machines (SVM) and many others (Liui et al. 2003).
Since labeled data is difficult to obtain, and unlabeled data is readily available and plentiful. Castelli and Cover in 1996 showed in a theoretical framework that unlabeled data can

indeed be used to improve classification, although it is exponentially less valuable than labeled data. Fortunately, unlabeled data can often be obtained by completely automated methods (McCallum et al.).

Implementation overview

In our implementation of the EM algorithm classifier, we experimented this idea of making the classification learning process using unlabeled data proceeds as follows (Tsuruoka et al. 2003):

1. Train the classifier using only labelled data.
2. Classify unlabeled examples, assigning probabilistic labels to them.
3. Update the parameters of the model. Each probabilistically labeled example is counted as its probability instead of one.
4. Go back to (2) until convergence.

It's hard to realize how the unlabeled data would improve the classification of documents; it seems unrealistic to assert that these data can contribute to the classification. However new methodologies have proven that unlabeled documents contains some of the most important pieces of information that would provide more effective classification in the future. Recently, it has been shown in (Nigam et al. 1998) that unlabeled data is helpful in classifier building. This technique alleviates some labor-intensive effort (Lee et al. 2002). Although in some cases this approach did not do as expected but in contrast it made the classifier less effective. However, our work presents an experiment for this approach and tries to investigate its effectiveness using Arabic language data.

Existing text classification techniques can be grouped into three types, supervised learning, semi-supervised learning, and unsupervised learning (or clustering). The proposed technique in which we will be using is related to but significantly different from all these existing approaches, the effectiveness of various algorithms and approaches

have shown significant value in applications (Xiaoli Li et al. 2004), (Tsuruoka et al. 2003). Nigam (Nigam et al. 1998) used Expectation Maximization (EM) and a naive Bayes classifier. Nigam et al. present a number of experimental results that shown that error rates can be reduced significantly using unlabeled examples in this way (Hirsh et al. 2000). [8,7,2,9] developed a frame work for enhancing the classification using unlabeled data by enriching the classifier using an EM algorithm, in which they prove that unlabeled data can improve the text classification, by estimating the parameters of the classifier till they reach a convergence.

Arabic language provides a great challenge for such approaches, in our work we simulated this frame work for using unlabeled data to enhance the classifier, through a technique that keeps in mind all of the Arabic language features, characteristics and associated difficulties.

The data model is built, the algorithm based on EM algorithm is developed, and a representation approach for the documents is designed based on vectors, commonly used TF/IDF (Term Frequency/Inverted Document Frequency) weighting scheme are implemented and used. For classification algorithm, a probabilistic frame work is used to build the classifier. In our study we used the widely known naïve Bayesian to calculate the initial document labels. Then a final classifier is built using the EM algorithm.
The system is tested on a data set of 600 documents from 6 classes taken form Al-Jazeera satellite news, where these classes spans 6-differnet fields: Agriculture, Economy, Health-Medicine, Politics, Science and Sports. We will report details of the tests and the produced results in the final paper.

We made an initial run with 15-labeled documents for each class. Results of the run have shown that applying the EM algorithm presented a notable improvement for all classes with an average improvement of 20%. We also noticed that the effect of amount of unlabeled data for all classes having more unlabeled data helped more when there are little labeled ones.

Conclusion

When our assumptions of data generation are correct basic EM can effectively incorporate information from unlabeled data. However, the full complexity of real world text data cannot be completely captured by known statistical models (Nigam et al. 2000). Hence we have some unexpected distributions in our results.

We believe that our algorithm and others using unlabeled data require a closer match between the data and the model than those with only labeled data; if the intended target concept and model desire too much with the actual distribution of the data, then the use of unlabeled data will not help (Nigam et al. 1998).

However in our work the EM algorithm had provided us with an average improvement of 20% accuracy, still we have some classes that were out of range in the total improvement, due to its uncertainty of classification. Although we can say that EM has to be further investigated on Arabic documents, EM presented an advantage in different ways for Arabic documents.

Experimental results using EM algorithm with proposed constraint consistently reduced the average classification error rates when the amount of labeled data is small. The results also showed that use of unlabeled data is especially advantageous when the amount of labeled data is small (Tsuruoka et al. 2003).

References

Arthur Dempster, Nan Laird and Donald Rubin, "Likelihood from incomplete data via the EM algorithm". Journal of the Royal Statistical Society, Series B, 39(1): 1-38, 1977.

Andrew McCallum, Kamal Nigam, "Text Classification by Bootstrapping with Keywords, EM and Shrinkage", In ACL99, Workshop for Unsupervised Learning in NLP.

Bing Liu, Xiaoli Li, Wee Sun Lee and Philip S. Yu, “Text Classification by Labeling Words”, Proceedings of the national conference on AI, 2004.

Bing Liu ,Yang Dai , Xiaoli Li, Wee Sun Lee , Philip S. Yu , “ Building Text Classifiers Using Positive and Unlabeled Examples “, ICDM 2003, Third IEEE International Conference on Data Mining. 19-22 Nov. 2003, pg. 179-186.

Carlos Ordonez, Edward Omiecinski. “FREM: Fast and Robust EM Clustering for Large Data Sets”, Conference on Information and Knowledge Management, Proceedings of the 11th International Information and Knowledge Management, Virginia, USA 2002, pg. 590-599.

Jinxi Xu, Alexander Fraser, Ralph Weischedel,” Empirical Studies in Strategies for Arabic Retrieval”, Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, Tampere, Finland 2001, Pg. 269 - 274

Kamal Nigam, Andrew McCallum, Sebastian Thrun Tom Mitchell," Using EM to Classify Text from Labeled and Unlabeled Documents", CMU-CS-98-120, December 21, 1998.

Kamal Nigam, Andrew McCallum, Sebastian Thrun Tom Mitchell, "Text Classification from Labeled and Unlabeled Documents using EM ", Machine Learning, Springer Netherlands, Vol. 39, Numbers 2-3/May 2000, pg. 103-134.

Kamal Paul Nigam, “Using Unlabeled Data to Improve Text Classification“, Ph.D. Dissertation, Carnegie Mellon University, PA, USA, May 2001.

10- Mohammed Aljlayl, Ophir Frieder, & David Grossman, “On Arabic-English Cross-Language Information Retrieval: A Machine Translation Approach”, IEEE Computer Society, ITCC 2002.

Rayid Ghani, "Combining labeled and unlabeled data for text classification with a large number of categories", Proceedings of the IEEE International Conference on Data Mining 2001.

Sarah Zelikovitz, Haym Hirsh ,” Improving Short-Text Classification Using Unlabeled Background Knowledge to Assess Document Similarity”, Proceedings of the Seventeenth International Conference on Machine Learning, 2000.

Sunita Sarawagi, Soumen Chakrabarti, Shantanu Godbole, "Cross-training: Learning probabilistic mappings between topics", Proceedings of the ninth ACM SIGKDD international conference 2003

Wee Sun Lee, Bing Liu, Philip S. Yu, Xiaoli Li, “ Partially Supervised Classification of Text Documents”, Machine Learning-International Workshop Then Conference-, 2002

Yoshimasa Tsuruoka, and Jun’ichi Tsujii," Training a Naive Bayes Classifier via the EM Algorithm with a Class Distribution Constraint", Proceedings of the seventh conference on Natural language Processing, 2003