Semi Supervised Learning Based Text Classification Model for Multi Label Paradigm

Shweta C. Dharmadhikari[1], Maya Ingle[2], and Parag Kulkarni[3]

[1] Pune Institute of Computer Technology,
Pune, Maharashtra, India
[2] Devi Ahilya Vishwa Vidyalaya, Indore, Madhya Pradesh, India
[3] EkLat Solutions
Pune, Maharashtra, India
{d.shweta18,paragindia}@gmail.com, maya_ingle@rediffmail.com

Abstract. Automatic text categorization (ATC) is a prominent research area within Information retrieval. Through this paper a classification model for ATC in multi-label domain is discussed. We are proposing a new multi label text classification model for assigning more relevant set of categories to every input text document. Our model is greatly influenced by graph based framework and Semi supervised learning. We demonstrate the effectiveness of our model using Enron, Slashdot, Bibtex and RCV1 datasets. We also compare performance of our model with few popular existing supervised techniques. Our experimental results indicate that the use of Semi Supervised Learning in multi label text classification greatly improves the decision making capability of classifier.

Keywords: Automatic text categorization, Multi-label text classification, graph based framework, semi supervised learning.

1 Introduction

Automatic text classification (ATC) is a prominent research area within Information retrieval. Multi label text classification problem refers to the scenario in which a text document can be assigned to more than one classes simultaneously during the process of text classification. The inherent ambiguity present in the content of textual data often makes the text document to be the member of more than one class simultaneously[3]. It has attracted significant attention from lot of researchers for playing crucial role in many applications such as web page classification, classification of news articles, information retrieval etc. Multi label text classifier can be realized by using supervised, unsupervised and semi supervised methods of machine learning. In supervised methods only labeled text data is needed for training. But availability of labeled data all the time is rare and processing of is expensive. Unsupervised methods relies only on unlabeled text documents; but it does not shows remarkable improvement in the performance. Semi supervised methods effectively uses unlabeled data in addition to the labeled data. Majority of existing approaches are supervised in nature[16]. Most of these lacking in considering relationship

V.V. Das and P.M. El-Kafrawy (Eds.): SPIT 2012, LNICST 117, pp. 178–184, 2014.

between class labels, input documents and also relying on labeled data all the time for classification. And also not capable of utilizing information conveyed by unlabeled data[17].

Hence through our paper we are proposing a multi label classification model using semi supervised learning so that classifier can handle labeled and unlabeled data. We are also aiming at handling input documents similarity along with correlation existing between class labels to improve decision making capability of our proposed classifier. We apply the proposed model on standard dataset such as Enron, Bibtex and RCV1 and Slashdot to test the performance. We also compare performance of our model with few popular existing supervised techniques.

The rest of the paper is organized as below. Section 2 describes relevant literature related to our proposed system; Section 3 describes our proposed classification model. Section 4 describes experiments and results, followed by a conclusion in the last section.

2 Related Work/Literature

Multi label learning problem is generally realized by problem transformation and algorithm adaptation methods. Few popular algorithms under these categories are binary relevance method, label power set method, pruned sets method, C4.5, Adaboost.MH & Adaboost.MR, ML-kNN, Classifier chains method etc[20]. These methods either decomposes classification task into multiple independent binary classification tasks[6], one for each category or the ranking function of category labels from the labeled instances and apply it to classify each unknown test instance by choosing all the categories with the scores above the given threshold[20]. Almost all of these methods are supervised in nature. These methods cannot utilize information conveyed by unlabeled data. The other common drawbacks include inability to handle relationship among class labels and can not scale to large data set.

Recently some new approaches for multi-label learning that consider the correlations among categories have been developed. Few eg. are generative model proposed by Ueda[26], Bayesian model proposed by Griffiths [27], Hierarchical structure considered by Rousu [28], Maximum entropy method proposed by Zhu[29], Latent variable based approach proposed by McCallum. But all these methods are also supervised in nature.

Few recent approaches effectively used semi supervised learning for multi label text classification. In 2006 Liu, Jin and Yan proposed Multi-label classification approach based on constrained non negative matrix factorization [8]. In this approach parameter selection affects the overall performance of the system. Zha and Mie proposed Graph-based SSL for multi-label classification in the year 2008[9]. But this approach was purely intended for classification of video files and not for documents. Chen,Song and Zhang proposed Semi supervised multi-label learning by solving a Sylvester Eq in the year 2010 [10]. In this approach they constructed graph for input representation and class representation as well but this approach is getting slower on convergence when applied in the situation where large number of classes and input

data exists. In 2009 Lee, Yoo and Choi proposed Semi-Supervised Non negative Matrix Factorization based approach [11]. But this approach was not specifically meant for multi-label text classification.

Thus by identifying limitations of all these methods we feel that there is need to build intelligent text classifier for multi label scenario which can efficiently handle all these said issues.

3 Proposed Classifier Model

The objective behind designing the proposed classifier model is to improve accuracy of multi label text classification process by assigning more relevant set of classes to unclassified documents. Following Fig.1 shows architecture proposed by us to achieve the said objective. We are using both labeled and unlabeled documents for training as our classifier is based on semi supervised learning.

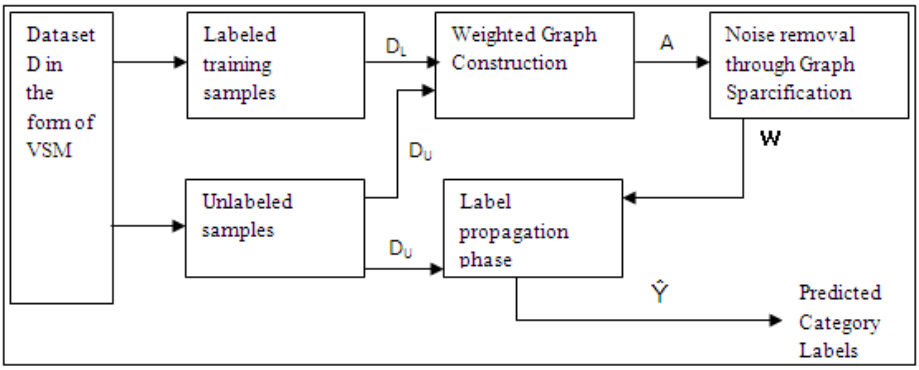

Fig. 1. Architecture of Classifier Model

We considered dataset D, which is in the VSM representation format. Out of which$|D_L|$ documents are already labeled and $|D_U|$ are unlabeled. We constructed Graph G(V,E) out of it. This graph is represented in the form of adjacency matrix A. Graph G consists of "n" no. of vertices such that $n=|D_L| + |D_U|$ .The objective is to predict set of labels for D_U. Each vertex Vi represents document instance di . Relationship between pair of vertices is represented by edge E. The adjacency matrix $A \in Rnxn$ is computed to represent the edge weight using cosine similarity measure. We have captured the correlation among different classes by computing matrix [B]kxk for representing relationship between classes.

In the next phase we attempted to remove noise by eliminating irrelevant documents prior to classification. We constructed graph W from Graph A through graph sparcification process; $A \Rightarrow W \in Rnxn$. In this, Matrix A is specified and reweighted using Knn approach and produce matrix W. This graph specification can lead to improve efficiency in the label inference stage. This stage is followed by classifier training phase which estimates a continuous classification function F on W

i.e. F∈R|v|x|c| where |v| is number of vertices and |c| is number of class labels.F: W→Ŷ …Where Ŷ is estimated label set.It estimates soft labels of unlabeled doc. By optimizing the energy function by generating confidence matrix [P]nxn. To this phase specified graph W acts as an input. Given this graph W and label information. This phase infers labels of unlabeled documents.

In the last Prediction phase we employed label propagation approach . It works on the smoothness assumption of SSL which states that "If two input points x1, x2 are in high density region are closer to each other then so should be the corresponding outputs y1, y2". Closeness between the two document instance can be identified by W. Relation between corresponding class labels can be computed by weighted dot product piBpj . If assignment of class labels pi and pj are relevant to doc. di and dj then we would expect Wi,j ≈ piBpj and uses following smoothness function to predict the labels of unlabeled doc.

$$\emptyset = \sum_{i,j=1}^{n} \left(W_{i,j} - \sum_{i,k=1}^{m} p_i B p_j \right)$$

4 Experimentations and Result Discussion

We evaluated our approach under a WEKA-based [23] framework running under Java JDK 1.6 with the libraries of MEKA and Mulan [21][22]. Jblas library for performing matrix operations while computing weights on graph edges. Experiments ran on 64 bit machines with 2.6 GHz of clock speed, allowing up to 4 GB RAM per iteration. Ensemble iterations are set to 10 for EPS. Evaluation is done in the form of 5 × 2 fold cross validation on each dataset. We first measured the accuracy, precision, Recall after label propagation phase is over. We conducted experiments on four text based datasets namely Enron, Slashdot, Bibtex and Reuters. Table 1 summarizes the statistics of datasets that we used in our experiments.

Table 1. Statistics of Datasets

Dataset	No. of document instances	No. of Labels	Attributes
Slashdot	3782	22	500
Enron	1702	53	1001
Bibtex	7395	159	1836
RCV1	12,000	135	5000

Enron dataset contains email messages. It is a subset of about 1700 labeled email messages [21]. BibTeX data set contains metadata for the bibtex items like the title of the paper, the authors, etc. Slashdot dataset contains article titles and partial blurbs mined from Slashdot.org [22]. We measured accuracy, precision, recall and F-measure of overall classification process. Fig. 2 shows the result comparison for these different datasets.We used accuracy measure proposed by Godbole and Sarawagi in [13]. It symmetrically measures how close y_i is to Zi ie estimated labels and true labels. It is the ratio of the size of the union and intersection of the predicted and

actual label sets, taken for each example and averaged over the number of examples. The formula used by them to compute accuracy is as follows:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{Y_i \cap Z_i}{Y_i \cup Z_i}\right]$$

In order to evaluate the performance of our classifier model using SSL approach, we compared the results of few popular supervised algorithm such as C4.5, Adaboost, ML-kNN, BP-MLL, SVM-HF (Algorithm adaptation method) and BR,RAkEL, MetaLabeler, CC,PS and EPS (problem transformation method).

Fig. 2 shows comparison of accuracy measured for each dataset; whereas Fig. 3 and Fig. 4 represents comparison of accuracy measured during experimentation between our classifier (referred as GB-MLTC) and supervised approaches on the same set of datasets.

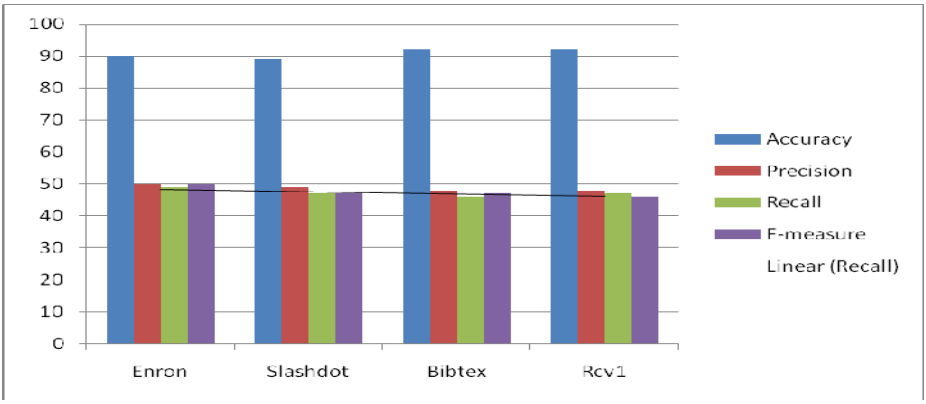

Fig. 2. Comparison of Results Measured Using GB-MLTC on Different Datasets

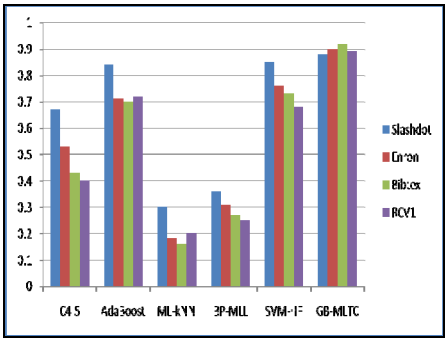

Fig. 3. GB-MLTC Vs Supervised Algorithm Adaptation Methods

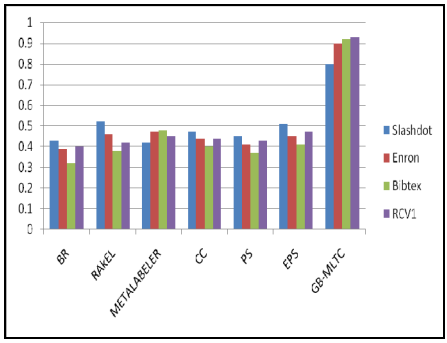

Fig. 4. GB-MLTC Vs Supervised P.T. methods

5 Conclusion and Future Work

In our classification model we incorporated document similarity along with class label correlation in order to improve accuracy of multi label text classifier. We have used semi-supervised learning to utilize the unlabeled data for text classification. Experimental results show that our model offers reasonably good accuracy. Use of cosine similarity measure may ignore some aspects of semantic relationship between text documents which can affect accuracy. However In future, along with vector space model of text representation use of more robust feature extraction technique like LSI or NMF may be incorporated in order to reduce rate of misclassification.

References

1. Zhu, J.: Semi-supervised learning Literature Survey. Computer Science Technical Report TR 1530, University of Wisconsin – Madison (2005)
2. Chapelle, O., Schfolkopf, B., Zien, A.: Semi-Supervised Learning, 03-08. MIT Press (2006)
3. Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. International Journal of Data Warehousing and Mining 3(3), 1–13 (2007)
4. Santos, A., Canuto, A., Neto, A.: A comparative analysis of classification methods to multi-label tasks in different application domains. International Journal of Computer Information Systems and Industrial Management Applications 3, 218–227 (2011) ISSN: 2150-7988
5. Cerri, R., da Silva, R.R.O., de Carvalho, A.C.P.L.F.: Comparing methods for multilabel classification of proteins using machine learning techniques. In: Guimarães, K.S., Panchenko, A., Przytycka, T.M. (eds.) BSB 2009. LNCS, vol. 5676, pp. 109–120. Springer, Heidelberg (2009)
6. Tsoumakas, G., Kalliris, G., Vlahavas, I.: Effective and efficient multilabel classification in domains with large number of labels. In: Proc. of the ECML/PKDD 2008 Workshop on Mining Multidimensional Data (MMD 2008), pp. 30–44 (2008)
7. Nigam, K., McCallum, A.K., Thrun, S., Mitchell, T.M.: Text classification from labeled and unlabeled documents using EM. Machine Learning 39, 103–134 (2000)
8. Liu, Y., Jin, R., Yang, L.: Semi-supervised Multi-label Learning by Constrained Non-Negative Matrix Factorization. In: AAAI (2006)
9. Zha, Z., Mie, T., Wang, Z., Hua, X.: Graph-Based Semi-Supervised Learning with Multi-label. In: ICME, pp. 1321–1324 (2008)
10. Chen, G., Song, Y., Zhang, C.: Semi-supervised Multi-label Learning by Solving a Sylvester Equation. In: SDM (2008)
11. Semi-supervised Nonnegative Matrix factorization. IEEE (January 2011)
12. Wei, Q., Yang, Z., Junping, Z., Wang, Y.: Semi-supervised Multi- label Learning Algorithm using dependency among labels. In: IPCSIT, vol. 3 (2011)
13. Godbole, S., Sarawagi, S.: Discriminative methods for multi-labeled classification. In: 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining (2004)
14. Angelova, R., Weikum, G.: Graph based text classification: Learn from your neighbours. In: SIGIR 2006. ACM (2006) 1-59593-369-7/06/0008
15. Jebara, T., Wang, Chang: Graph construction and b-matching for semi supervised learning. In: Proceedings of ICML- 2009(2009)

16. Thomas, Ilias, Nello: Scalable corpus annotation by graph construction and label propogation. In: Proceedings of ICPRAM, pp. 25–34 (2012)
17. Talukdar, P., Pereira, F.: Experimentation in graph based semi supervised learning methods for class instance acquisition. In: The Proceedings of 48th Annual Meet of ACL, pp. 1473–1481 (2010)
18. Dai, X., Tian, B., Zhou, J., Chen, J.: Incorporating LSI into spectral graph transducer for text classification. In: The Proceedings of AAAI (2008)
19. Dharmadhikari, S.C., Ingle, M., Kulkarni, P.: Analysis of semi supervised methods towards multi-label text classification. IJCA 42, 15–20, ISBN: 973-93-80866-84-5
20. Dharmadhikari, S.C., Ingle, M., Kulkarni, P.: A comparative analysis of supervised multi-label text classification methods. IJERA 1(4), 1952–1961, ISSN: 2248-9622
21. http://mulan.sourceforge.net/datasets.html
22. http://MEKA.sourceforge.net
23. http://www.cs.waikato.ac.nz/ml/weka/
24. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. In: Buntine, W., Grobelnik, M., Mladenić, D., Shawe-Taylor, J. (eds.) ECML PKDD 2009, Part II. LNCS, vol. 5782, pp. 254–269. Springer, Heidelberg (2009)
25. Schapire, R.E., Singer, Y.: Boostexter: A boosting based system for text categorization. Machine learning 39(2-3) (2000)
26. Ueda, Saito, K.: Parametric mixture models for multi-labelled text. In: Proc. of NIPS (2002)
27. Griffiths, Ghahramani: Infinite latent feature models and the Indian buffet process. In: Proc. of NIPS (2005)
28. Rousu, Saunders: On maximum margin hierarchical multi-label classification. In: Proc. of NIPS Workshop on Learning with Structured Outputs (2004)
29. Zhu, S., Ji, X., Gong, Y.: Multi-labelled classification using maximum entropy method. In: Proc. of SIGIR (2005)
30. Ding, C., Jin, R., li, T., Simon, H.: A learning framework using Green's Function and Kernel Regularization with application to Recommender System. ACM, San Jose (2007) 978-1-59593-609-7/07/0008