International Journal of Engineering Research and Development
ISSN: 2278-067X, Volume 1, Issue 9 (June 2012), PP.09-14
www.ijerd.com

Multi Label Text Classification through Label Propagation

Shweta C. Dharmadhikari[1], Maya Ingle[2], Parag Kulkarni[3]
[1,3]*Pune Institute of Computer Technology - EkLat Solutions, Pune , Maharashtra, India*
[2]*Devi Ahilya Vishwa Vidyalaya, Indore, Madhya Pradesh , India*

Abstract: *Classifying text data has been an active area of research for a long time. Text document is multifaceted object and often inherently ambiguous by nature. Multi-label learning deals with such ambiguous object. Classification of such ambiguous text objects often makes task of classifier difficult while assigning relevant classes to input document. Traditional single label and multi class text classification paradigms cannot efficiently classify such multifaceted text corpus. Through our paper we are proposing a novel label propagation approach based on semi supervised learning for Multi Label Text Classification. Our proposed approach models the relationship between class labels and also effectively represents input text documents. We are using semi supervised learning technique for effective utilization of labeled and unlabeled data for classification .Our proposed approach promises better classification accuracy and handling of complexity and elaborated on the basis of standard datasets such as Enron, Slashdot and Bibtex.*

Keywords: *Label propagation , semi-supervised learning , multi-label text classification.*

I. INTRODUCTION

The amount of textual data being produced through internet is growing faster than the ability of information consumers to search, digest and use it. Textual data is difficult to effectively understand and categorize because the relationship between its sequence of words and its content is less clear as compared to numerical. Such data includes technical article, memos, manuals, electronic mail, books, online news paper, journal articles and many other forms of texts. Thus text classification has become an active research topic now a day. It classifies document under a predefined category. Categories may be represented numerically or using single word or phrase or words with senses, etc. In traditional approach, classification of text was carried out manually using domain experts. The human expert was required to read and sort the input text document to predefined category or set of categories. Thus this approach requires extensive human efforts and error prone also. This leads to the scheme of automated text classification scenario. This automated text document classification facilitates ease of storage, searching, retrieval of relevant text documents or its contents for the needy applications. Three different paradigm exists under text classification and they are single label(Binary) , multiclass and multi label. Under single label a new text document belongs to exactly one of two given classes, in multi-class case a new text document belongs to just one class of a set of m classes and under multi label text classification scheme each document may belong to several classes simultaneously [3]. In real practice many approaches are exists and proposed for binary case and multi class case even though in many applications text documents are inherently multi label in nature. Eg. In medical diagnosis a document report containing set of symptoms can belong to many probable disease categories. Multilabel 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 classification. Eg. In the process of classification of online news article the news stories about the scams in the commonwealth games in india can belong to classes like sports, politics , country-india etc. 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.

Multilabel 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 classification.. 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. Generally supervised methods from machine learning are mainly used for realization of multi label text classification. But as it needs labeled data for classification all the time, semi supervised methods are used now a day in multi label text classifier. Many approaches are preferred to implement multi label text classifier. Through our paper we are proposing label propagation approach for multi label text classifier , it uses existing label information for identifying labels of unlabeled documents.We are representing input text document corpus in the form of graph to exploit the ambiguity among different text documents. The ambiguity is represented in the form of similarity measures as a weighted edge between text documents . With the setting of semi supervised learning we have focused on not only graph construction but also sparsification and weighting of graph to improve classifiers accuracy. We apply the proposed framework on standard dataset such as Enron, Bibtex and slashdot.

The rest of the paper is organized as below. Section 2 describes literature related to semi supervised learning methods for multi label text classification system ; Section 3 highlights mathematical modeling of our approach . Section 4 describes our proposed label propagation approach for building multi label text classifier followed by experiments and results in Section 5 , followed by a conclusion in the last section.

II. RELATED WORK

Multilabel 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. Unsupervised methods relies heavily on only unlabeled text documents; whereas semi supervised methods can effectively use unlabeled data in addition to the labeled data[1][2]. The traditional approach towards multi-label learning either decomposes the classification task into multiple independent binary classification tasks or identifies rank to find relevant set of classes. But these methods do not exploit relationship among class labels. Few popular existing methods are binary relevance method, label power set method, pruned sets method, C4.5, Adaboost.MH & Adaboost.MR, ML-kNN , Classifier chains method etc[20]. But all these are lacking the capability of handling unlabeled data ie these are based on principle of supervised learning.

While designing a multi label text classifier the major objective is not only to identify the set of classes belonging to given new text documents but also to identify most relevant out of them to improve accuracy of overall classification process. Graph based approaches are known for their effective exploration of document representation and semi supervised methods explores both labeled and unlabeled data for classification thatswhy accuracy of multi label text classifier can be improved by using graph based representation of input documents in conjunction with label propagation approach of semi supervised learning[16][17].

Table 1 summarizes few existing well-known representative methods for multi label text classifier based on semi supervised learning , few uses only graph based framework and few uses both.

Table 1: *Statistics of popular algorithms for MLTC based on semi supervised learning and graph based representation.*

Algorithm and Year of proposal	Working Theme	Datasets used for experimentation	Merits	Demerits
Multi-label classification by Constrained Non-Negative Matrix Factorization [2006] [8]	Optimization of class labels assignment by using similarity measures and non negative matrix factorization.	ESTA	Powerful representation of input documents using NMF and also works for large scale datasets	Parameter selection is crucial.
Graph-based SSL with multi-label [2008][9]	Exploits correlation among labels along with labels consistency over graph.	Video files : TECVID 2006.	Effective utilization of unlabeled data.	Can not applicable to text data , more effective on video data.
Semi supervised multi-label learning by solving a Sylvester Eq [2010][10]	Graph construction for input documents and class labels.	Reuters	Improved accuracy	May get slower on convergence.
Semi-Supervised Non negative Matrix Factorization [2009][11].	Performs joint factorization of data and labels and uses multiplicative updates performs classification.	20-news, CSTR, k1a,k1b,WebKB4, Reuters	Able to extract more discriminative features	High computational complexity.

In preprocessing stage graph based approaches can effectively represents relationship between labeled and unlabeled documents by identifying structural and semantical relationship between them for more relevant classification ; and during training phase semi supervised methods can propagate labels of labeled documents to unlabeled documents based on some energy function or regularizer. Our proposed work is based on the same strategy.

III. MATHEMATICAL MODEL OF PROPOSED SYSTEM

In this section we are introducing some notions related with text classification. We are firstly representing the input document corpus in the form of graph.The process of graph construction deals with conversion of input text document corpus , X to graph G ie X → G , where X represents input text document corpus x1,x2,..,xn wherein each text document instance x_i in turn represented as m-dimensional feature vector. And G represents overall graph structure as G=(V,E) where V = set of vertices corresponding to document instance x_i ; E represents set of weighted edges between pair of vertices where associated edge weight corresponds to similarity between two documents. Generally weight matrix W is computed to identify the similarity between pair of text documents. Various similarity measures such as cosine, Jacobi or kernel functions K(.) like RBF kernel , Gaussian kernel can be used for this purpose.

Now we are defining our graph based multi label text classifier system S as follows :
S = { X , Y , T , $\hat{y}$, h}; where X represents entire input text document corpus = {x1,x2,..,xn}. Out of these |L| numbers of documents are labeled and remaining are unlabeled.Y represents set of possible labels = {Y1,Y2,…,Yn}. T represents multilabel training set of classifier of the form {(x1,Y1), (x2,Y2),….., (xn,Yn)} where xi ∈ *X* is a single document instance

and Yi $\subseteq$ Y is the label set associated with xi . $\hat{y}$ represents set of estimated labels = $\{\hat{Y}l , \hat{Y}u\}$. The goal of the system is to learn a function h ie

h : X → 2y from T which predicts set of labels for unlabeled documents ie xl+1 ..xn

With this graph based setting, we are using semi supervised learning to propagate labels on the graph from labeled nodes to unlabeled nodes and compare the estimated labels $\hat{y}$ with the true labels.

IV. PROPOSED APPROACH

We are mainly using theme of smoothness assumption of semi supervised learning to propagate the labels of labeled documents to unlabeled documents. Smoothness assumption of semi supervised learning states that "if two input points x1,x2 are in a high-density region are close to each other then so should be the corresponding outputs y1,y2" . Thus based on this we mainly emphasized on exploiting relationships between input text documents in the form of graph and relationship between the class labels in the form of correlation matrix. The purpose behind this is to reduce classification errors and assignment of more relevant class labels to new test document instance.

During classifiers training phase we are computing similarity between input documents to identify whether they are in high density or low density region. We evaluated relationships between documents by using cosine similarity measure and represented it in the form of weighted matrix, W as :

$$W_{i,j} = \exp(-\lambda (1- \cos(d_i,d_j)))$$

Where X1 and X2 are two text documents represented in the feature space. Large cosine value indicates similarity and small value indicates that documents are dissimilar.

After that we performed graph sparcification by representing it in the form of diagonal matrix in order to reduce consideration of redundant data. So we normalize the term $1/\| d_i\| \|d_j\|$, we calculated the diagonal matrix as $Dcos(d_i, d_j) = 1/\sqrt{f(i)f(i)^T}$ where F(i) is the ith row vector of F. While identifying relationships between class labels we computed correlation matrix C mxm where m is no. of class labels using RBF kernel. Each class is represented in the form of vector space whose elements are said to be 1 when corresponding text document belongs to the class under consideration.

Then in testing phase, in order to provide relevant label set to unlabeled document we computed energy function E to measure smoothness of label propagation. This energy function measures difference between weight matrix W and dot product of sparcified diagonal matrix with correlation matrix.

$$\mathbf{E = \sum Wij - DCij}$$

The labels are propagated based on minimum value of Energy function. It indicates that if two text documents are similar to each other then the assigned class labels to them are also likely to be closer to each other. In other words two documents sharing highly similar input pattern are likely to be in high density region and thereby the classes assigned to them are likely to be related and propogated to those documents which in turn resides in same high density region.

After this label propagation phase, we obtained labels of all unlabeled document instances. We computed accuracy to verify correct assignment of label sets. The corresponding results are given in table [III]. We once again ensured the working by applying all this document and label set to existing classifier chains method which is supervised in nature. We used decision tree(J48 in WEKA),SVM (SMO & libSVM) separately as base classifiers and computed the results. The corresponding results are given in table [IV].

The summary of our proposed label propagation approach is given as :

Input - T : The multi label training set {(x1,Y1), (x2,Y2),….., (xn,Yn)}.

z : The test document instance such that z $\in$ X

Output – The predicted label set for z .

Process:

- Compute the edge weight matrix W as $\boldsymbol{Wij = \arccos \frac{X1 . X2}{|X1|.|X2|}}$ and assign W_{ii}=0
- Sparcify the graph by computing diagonal degree matrix D as $D_{ii}=\sum_j W_{ij}$
- Compute the label correlation matrix C_{mxm} using RBF kernel method
- Initialize $\hat{Y}^{(0)}$ to the set of ($Y_1,Y_2,\ldots,Y_l,0,0,\ldots\ldots,0$)
- Iterate till convergence to $\hat{Y}^{(\infty)}$

1. $E = \sum W_{ij} - D^{-1}C_{ij}$
2. $\hat{Y}^{(t+1)} = E$
3. $\hat{Y}^{(t+1)}{}_l = Y_l$

- Label point z by the sign of $\hat{Y}^{(\infty)}_{i}$

V. EXPERIMENTS AND RESULTS

In this section, in order to evaluate our approach we conducted experiments on three text based datasets namely Enron , Slashdot , Bibtex and measured accuracy of overall classification process. Table II summarizes the statistics of datasets that we used in our experiments.

Table II: *Statistics Of Datasets*

Dataset	No. of document instances	No. of Labels	Attributes
Slashdot	3782	22	500
Enron	1702	53	1001
Bibtex	7395	159	1836

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 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:

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

We also computed precision , recall and F-measure values , the formula used to compute them is as follows:

$$\textbf{F-Measure} = \frac{2.0 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

$$\textbf{F-Measure} = \frac{1}{N}\sum_{i=1}^{N}\frac{2\left[Y_i \cap Z_i\right]}{\left[Z_i\right] - \left\lfloor Y_i \right\rfloor}$$

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 and after label propagation phase is over. Table III enlists accuracy measured for each dataset.

Table III: *Results after Label Propagation Phase*

Evaluation Criterion	Enron	Slashdot	Bibtex
Accuracy	90	89	92
Precision	50	49	48
Recall	49	47	46
F-measure	50	47	47

After label propagation phase , we obtained labels of all unlabeled documents. Thus we get entire labeled dataset as a result now. We applied this labeled set to Ensemble of classifier chains method which is supervised in nature[24] and measured accuracy ,precision, recall on three different base classifiers of decision tree(J48 in WEKA) , and two variations of support vector machine (SMO in WEKA , libSVM).We also measured overall testing and building time required for this process. The Ensemble of classifier chains method (ECC) is proven and one of the efficient supervised multi label text classification technique , we verified our entire final labeled dataset by giving input to it. The results are enlisted in table IV

Table IV: *Result after using supervised multi label classifier*

DATASET: SLASHDOT

Parameters	BS : SMO	BS : libSVM	BS : J48
LCard(training)	2.42	2.27	2.23
#(training Samples)	1135.0	1135.0	1135.0
LCard(testing)	2.32	2.20	2.1
#(testing Samples)	756.0	756.0	756.0
Test time	69.5	28.9	29
Build time	173.9	2609.078	2546.17
Recall	0.40	0.56	0.53
Threshold	0.01	0.2	0.2
F1_micro	0.56	0.52	0.48
Precision	0.93	0.49	0.44
Accuracy	0.41	0.41	0.32

DATASET: ENRON .

Parameters	BS : SMO	BS : libSVM	BS : J48
LCard(training)	2.42	2.26	2.23
#(training Samples)	1135.0	1135.0	1135.0
LCard(testing)	2.32	2.20	2.13
#(testing Samples)	756.0	756.0	756.0
Test time	69.516	28.89	28.97
Build time	173.891	2609.078	2546.17
Recall	0.40	0.56	0.53
Threshold	0.0010	0.2	0.2
F1_micro	0.56	0.52	0.48
Precision	0.94	0.49	0.44
Accuracy	0.41	0.41	0.32

DATASET : BIBTEX

Parameters	BS : SMO	BS : libSVM	BS : J48
LCard(training)	2.45	2.27	2.23
#(training Samples)	1135.0	1135.0	1135.0
LCard(testing)	2.32	2.20	2.13
#(testing Samples)	756.0	756.0	756.0
Test time	69.516	28.89	28.9
Build time	173.8	2609.078	2546.172
Recall	0.40	0.56	0.53
Threshold	0.0010	0.2	0.2
F1_micro	0.56	0.52	0.48
Precision	0.93	0.49	0.44
Accuracy	**0.41**	**0.41**	**0.32**

VI. CONCLUSION AND FUTURE WORK

We have proposed a novel label propagation based approach for multi label classifier. It works in conjunction with semi supervised learning setting by considering smoothness assumptions of data points and labels. The approach is evaluated using small scale datasets (Enron , Slashdot) as well as large scale dataset (Bibtex). It is also verified against traditional supervised method. Our approach shows significant improvement in accuracy by incorporating unlabeled data along with labeled training data. But significant amount of computational time is required to calculate similarity among documents as well as class labels. The input text corpus is well exploited as a graph however, in the future the use of feature extraction methods like NMF with Latent Semantic indexing may provide more stable results.

REFERENCES

[1]. J. Zhu. Semi-supervised learning Literature Survey. Computer Science Technical Report TR 1530 , University of Wisconsin – Madison , 2005.

[2]. Olivier Chapelle , Bernhard Schfolkopf , Alexander Zien. Semi-Supervised Learning2006 , 03-08 , MIT Press.

[3]. G. Tsoumakas, I. Katakis. Multi-label classification: An overview. International Journal of Data Warehousing and Mining, 3(3):1-13, 2007.

[4]. A. Santos , A. Canuto, and A. Neto, "A comparative analysis of classification methods to multi-label tasks in different application domains", International journal of computer Information systems and Industrial Management Applications". ISSN: 2150-7988 volume 3(2011), pp. 218-227.

[5]. R.Cerri, R.R. Silva , and A.C. Carvalho , "Comparing Methods for multilabel classification of proteins using machine learning techniques",BSB 2009, LNCS 5676,109-120,2009.

[6]. G. Tsoumakas , G. Kalliris , and I. Vlahavas, " Effective and efficient multilabel classification in domains with large number of labels", Proc. Of the ECML/PKDD 2008 workshop on Mining Multidimensional Data (MMD' 08)(2008) 30-44.

[7]. Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. M. (2000). Text classification from labeled and unlabeled documents using EM. Machine Learning,39, 103–134.

[8]. Y. Liu, R. Jin, L. Yang. Semi-supervised Multi-label Learning by Constrained Non-Negative Matrix Factorization .In: AAAI, 2006.

[9]. Z. Zha, T. Mie, Z. Wang, X. Hua. Graph-Based Semi-Supervised Learning with Multi-label. In ICME. page 1321-1324, 2008.

[10]. G. Chen, Y. Song, C. Zhang. Semi-supervised Multi-label Learning by Solving a Sylvester Equation. In SDM, 2008.

[11]. Semi-supervised Nonnegative Matrix factorization. IEEE. January 2011.

[12]. Qu Wei , Yang, Junping, Wang. Semi-supervised Multi- label Learning Algorithm using dependency among labels. In IPCSIT vol. 3 2011.

[13]. S. Godbole and S. Sarawagi , "Discriminative methods for multi-labeled classification", 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2004.

[14]. R. Angelova , G. Weikum . "Graph based text classification : Learn from your neighbours". In SIGIR'06 , ACM , 1-59593-369-7/06/0008".

[15]. T.Jebara , Wang and chang , "Graph construction and b-matching for semi supervised learning". In proceedings of ICML-2009.

[16]. Thomas, Ilias & Nello. " scalable corpus annotation by graph construction and label propogation". In proceedings of ICPRAM, 25-34, 2012.

[17]. P. Talukdar , F. Pereira. " Experimentation in graph based semi supervised learning methods for class instance acquisition". In the proceedings of 48th Annual meet of ACL. 1473-1481.2010.

[18]. X. Dai, B. Tian , J. Zhou , J. Chen. "Incorporating LSI into spectral graph transducer for text classification" . In the proceedings of AAAI. 2008.

[19]. S.C. .Dharmadhikari , Maya Ingle , parag Kulkarni .Analysis of semi supervised methods towards multi-label text classification. IJCA , Vol. 42, pp. 15-20 ISBN :973-93-80866-84-5.

[20]. S.C. .Dharmadhikari , Maya Ingle , parag Kulkarni .A comparative analysis of supervised multi-label text classification methods. IJERA , Vol. 1, Issue 4 , pp. 1952-1961 ISSN : 2248-9622.

[21]. http://mulan.sourceforge.net/datasets.html

[22]. http://MEKA.sourceforge.net

[23]. www.cs.waikato.ac.nz/ml/weka/

[24]. J. Read, B. Pfahringer, G. Homes , and E.Frank , "Classifier chains for multi-label classification"., Proc. Of European Conference on Machine Learning and knowledge discovery in Databases, LNAI 5782(254-269),2009.

[25]. Griffiths and Ghahramani. Infinite latent feature models and the Indian buffet process. In proc. of NIPS, 2005.

[26]. Rousu , saunders. On maximum margin hierarchical multi-label classification. In Proc. Of NIPS workshop on Learning with structured outputs ,2004.

[27]. S. Zhu, Ji, Xu and Y. Gong. Multi-labelled classification using maximum entropy method. In Proc. Of SIGIR , 2005.