A Novel Graph Based Framework to build Multi Label Text Classifier

S.C.Dharmadhikari, Maya Ingle, Parag Kulkarni

Abstract: Text document is multifaceted object and associated with many properties such as multi labeledness. Under this a single text document can inherently belongs to more than one category simultaneously. Traditional single label and multi class text classification paradigms cannot efficiently classify such multifaceted text corpus. Through our paper we are proposing a graph based frame work for Multi Label Text Classification paradigm. Representing text documents in the form of graph vertices rather than the vector representation like Bag of Words allows pre-computing and storing of necessary information. It also models the relationship between text documents and class labels. We are using semi supervised learning technique in our proposed approach for effectively utilizing labeled and unlabeled data for classification .Our proposed approach promises better classification accuracy and handling of complexity. Our proposed framework is elaborated on the basis of standard dataset such as Enron, Slashdot, Bibtex and Reuters.

Keywords: Multi-label learning, text classification, semi-supervised learning, graph based learning,

1 INTRODUCTION:

The area of text classification forms one of the important step in the process of text mining. The major objective of text classification system is to organize the available text documents systematically into their respective categories[7]. This categorization of text documents facilititates 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.

S.C. Dharmadhikari is currently pursuing PhD degree program in computer science from Devi Ahilya Vishwa Vidyalaya , Indore, India.E-mail: d.shweta18@gmail.com.

Maya Ingle is currently working as Professor in Devi Ahilya Vishwa Vidyalaya,Indore,India. E-mail: maya_ingle@rediffmail.com.

Parag Kulkarni is Founder and Chief Scientist at EKlat - Research Pune,India. paragindia@gmail.com

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. 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. All the approaches needs initial step of text document representation[16]. The common approaches are vector space model using various term weighting schemes such as Boolean , word frequency count , term and document frequency , entropy encoding etc. All of these are popularly known as BOW (Bag Of Words) approaches[17]. Even though these are widely used but these ignores use of structural and semantic information in classification which may significantly improves accuracy. Other alternative to bag of words representation is graph based representation. The graph based representation offers much better document representation as it also considers relationship among documents in the form of edge of the graph[16].

Through our paper we are proposing a graph based frame work for Multi Label Text classifier with the setting of semi supervised learning as we want to use unlabeled data effectively for classification along with labeled data. 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 RCV1 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 overview of graph representation , sparcification . Section 4 describes our proposed graph based framework for building multi label text classifier followed by experiments and results in Section 5 , followed by a conclusion in the last section.

2 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].

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 framework in conjuction with semi supervised learning[16][17].

Table 1 summarizes few well-known representative methods for multi label text classifier ; few of them are based on simply 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**
Expectation Maximization (EM) based text classification[1999][7]	Uses the joint distribution over features other than the class labels.	WebKB,Reuters , 20 Newsgroups	successfully able to utilize unlabeled data alongwith labeled data	Applicable to single label text classifier
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.
Multi-label learning by using dependency among labels [2009][12]	Training the ordered list of classifiers.	Emotions, yeast and scene datasets.	Improved accuracy	More time complexity
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.

3 GRAPH CONSTRUCTION RELATED ISSUES

In this section we are introducing some notions related with graph construction in the setting of text classification. The process of graph construction deals with conversion of input text document corpus , X to graph G ie $\mathbf{X} \rightarrow G$, where $\mathbf{X}$ represents input text document corpus $x_1,x_2,..,x_n$ 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 = { $\mathbf{X}$, $\mathbf{Y}$, T , $\hat{y}$, h} where $\mathbf{X}$ represents entire input text document corpus = $\{x_1,x_2,..,x_n\}$. Out of these |L| numbers of documents are labeled and remaining are unlabeled. $\mathbf{Y}$ represents set of possible labels = $\{Y_1,Y_2,\ldots,Y_n\}$. T represents multilabel training set of classifier of the form $\{(x_1,Y_1), (x_2,Y_2),\ldots, (x_n,Y_n)\}$ where $x_i \in \mathbf{X}$ is a single document instance and $Y_i \subseteq \mathbf{Y}$ is the label set associated with x_i . $\hat{y}$ represents set of estimated labels = $\{\hat{Y}l , \hat{Y}u\}$. The goal of the system is to learn a function h ie
h : $\mathbf{X} \rightarrow \mathbf{2}^y$ from T which predicts set of labels for unlabeled documents ie $x_{l+1} .. x_n$

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.

4 PROPOSED FRAMEWORK

Many existing approaches dealing with multi label text classification treats all the class labels independently and unable to explore relationship between them. It may affect accuracy of classification because two document instances with no common classes can still related to each other if their assigned classes are related to each other.

Our framework is mainly based on smoothness assumption of semi supervised learning which 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. After that we performed graph sparcification by representing it in the form of diagonal matrix in order to reduce consideration of redundant data.

$$Wij = \text{arc cos} \quad \frac{X1 \,.\, X2}{|X1|.|X2|}$$

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.

While identifying relationships between class labels we computed correlation matrix C mxm where m is no. of class labels using cosine similarity measure again for ease of computation. 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.

$$E = \sum W_{ij} - D^{-1}C_{ij}$$

The labels are propagated based on minimum value of Energy function. It indicates that if two text documents are

ISSN 2229-5518

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.

The summary of our proposed approach is given as :

Input - T : The multi label training set $\{(x_1,Y_1), (x_2,Y_2),\ldots, (x_n,Y_n)\}$.

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

Output – The predicted label set for z .

Process:

- Compute the edge weight matrix W as $\boldsymbol{Wij} = \textbf{arc cos} \quad \frac{\boldsymbol{X1} \,.\, \boldsymbol{X2}}{|\boldsymbol{X1}|.|\boldsymbol{X2}|}$ and assign W_{ii}=0
- Sparcify the graph by computing diagonal degree matrix D as $D_{ii}=\sum_j W_{ij}$
- 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 $\hat{Y}^{(\infty)}{}_i$

5 EXPERIMENTS AND RESULTS

In this section, in order to evaluate our framework we conducted experiments on four text based datasets namely Enron , Slashdot , Bibtex and Reuters and measured accuracy of overall classification process. we used accuracy measure proposed by Godbole and Sarawagi in [13] . It symmetrically measures how close y_i is to Z_i 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 is as :

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

Datasets

Table I summarizes the statistics of datasets that we used in our experiments.

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]. Reuters 21578 is the most popular among all existing datasets used for text classification[21].

TABLE 2 : STATISTICS OF DATASETS

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

Experimental Results

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 are run on 32 bit machines with 1.3 GHz clock speed, allowing up to 2 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 ran experiment by slightly increasing no. of unlabeled documents and evaluated performance of our approach by measuring corresponding accuracy.

- Thus initially 5% of unlabeled and rest of labeled documents are used for classification (5% : 95%) and gradually these are increased upto 80% (80% : 20%).

Figure 1 to 4 represents the graph showing results in terms of accuracy . Table 3 represents the result after label propagation phase of semi supervised learning.

Figure 1 : ACCURACY MEASUREMENT ON ENRON DATASET

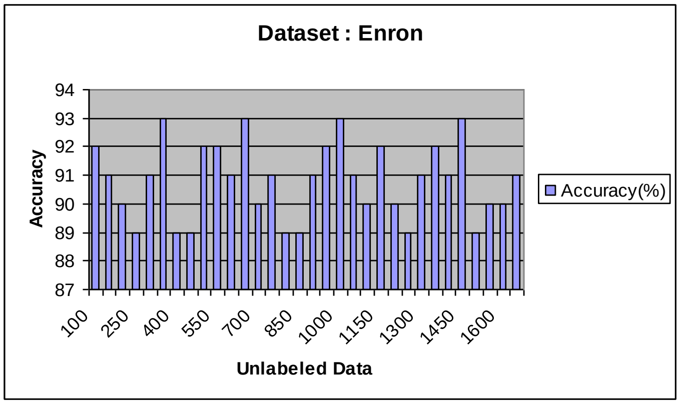

Figure 2 : ACCURACY MEASUREMENT ON SLASHDOT DATASET

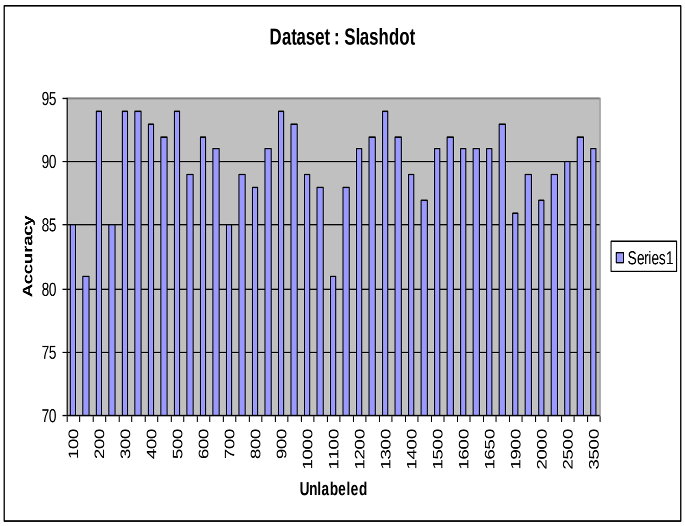

Figure 3 : ACCURACY MEASUREMENT ON BIBTEX DATASET

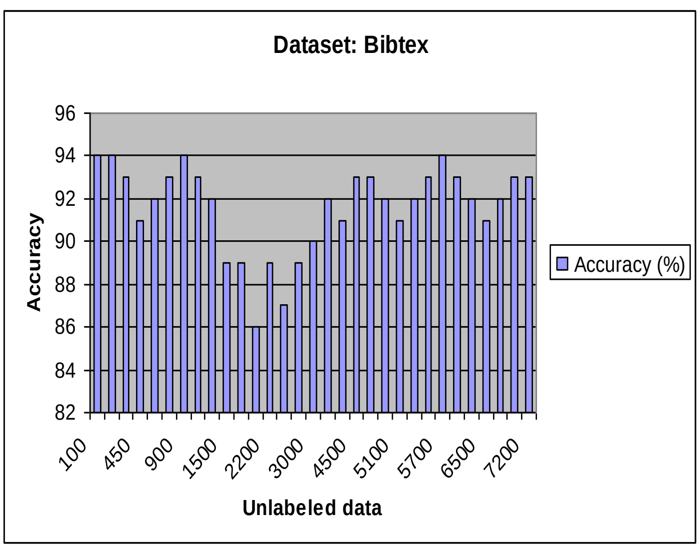

Figure 4 : ACCURACY MEASUREMENT ON REUTERS DATASET

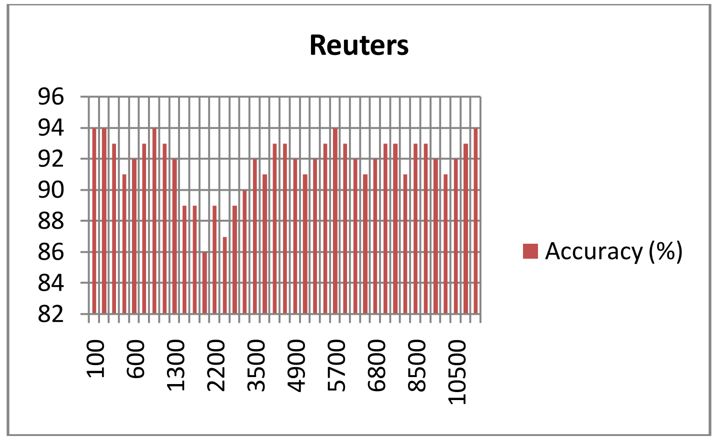

TABLE 3 : RESULTS AFTER LABEL PROPAGATION PHASE

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

CONCLUSION AND FUTURE WORK

We have proposed a novel graph based framework for multi label classifier. It works in conjunction with semi supervised learning setting by considering smoothness assumptions of data points and labels. The framework is evaluated using small scale datasets (Enron , Slashdot) as well as large scale dataset (Bibtex , Reuters). It is giving consistent results upto 85 : 15 split of unlabeled : labeled document ratio. We observed that the sparse representation of data in the matrix greatly affects the extraction of semantically associated features. But significant amount of computational time is observed to calculate similarity among documents as well as class labels with improvement in accuracy. In the future the use of feature extraction methods like NMF with Latent Semantic indexing may provide more stable results.

REFERENCES

1. [1] J. Zhu. Semi-supervised learning Literature Survey. Computer Science Technical Report TR 1530 , University of Wisconsin – Madison , 2005.
2. [2] Olivier Chapelle , Bernhard Schfolkopf , Alexander Zien. Semi-Supervised Learning2006 , 03-08 , MIT Press.
3. [3] G. Tsoumakas, I. Katakis. Multi-label classification: An overview. International Journal of Data Warehousing and Mining, 3(3):1-13, 2007.
4. [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. [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.

7. [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.

8.

9. [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.

10.

11. [8] Y. Liu, R. Jin, L. Yang. Semi-supervised Multi-label Learning by Constrained Non-Negative Matrix Factorization .In: AAAI, 2006.

12.

13. [9] Z. Zha, T. Mie, Z. Wang, X. Hua. Graph-Based Semi-Supervised Learning with Multi-label. In ICME. page 1321-

14. 1324, 2008.

15.

16. [10] G. Chen, Y. Song, C. Zhang. Semi-supervised Multi-label Learning by Solving a Sylvester Equation. In SDM, 2008.

17.

18. [11] Semi-supervised Nonnegative Matrix factorization. IEEE. January 2011.

19.

20. [12] Qu Wei , Yang, Junping, Wang. Semi-supervised Multi- label Learning Algorithm using dependency among labels. In IPCSIT vol. 3 2011.

21.

41.

22. [13] S. Godbole and S. Sarawagi , "Discriminative methods for multi-labeled classification", 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2004.

23.

24. [14] R. Angelova , G. Weikum . "Graph based text classification : Learn from your neighbours". In SIGIR'06 , ACM , 1-59593-369-7/06/0008".

25.

26. [15] T.Jebara , Wang and chang , "Graph construction and b-matching for semi supervised learning". In proceedings of ICML-2009.

27.

28. [16] Thomas, Ilias & Nello. " scalable corpus annotation by graph construction and label propogation". In proceedings of ICPRAM, 25-34, 2012.

29.

30. [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.

31. [18] X. Dai, B. Tian , J. Zhou , J. Chen. "Incorporating LSI into spectral graph transducer for text classification" . In the proceedings of AAAI. 2008.

32.

33. [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.

34.

35. [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.

36. [21] http://mulan.sourceforge.net/datasets.html

37.

38. [22] http://MEKA.sourceforge.net

39.

40. [23] www.cs.waikato.ac.nz/ml/weka/