Analysis of semi supervised learning methods towards multi label text classification

S.C.Dharmadhikari
PICT
Katraj,Pune,India

Maya Ingle
DAVV
Indore,India

Parag Kulkarni
EklaT Solutions
Pune , India

ABSTRACT

The area of multi label text classification is getting more attention of researchers because of its role in the field of information retrieval , text mining , web mining etc. Supervised methods from machine learning are mainly used for its realization. But as it needs labeled data for classification all the time , semi supervised methods are now a day getting popular in the MLTC domain. The goal of Semi supervised learning is to reduce the classification errors using readily available unlabeled data in conjunction with available labeled data.

This paper mainly provides survey and analysis of various semi supervised methods used in multi label text classification task ; This overview concludes that consideration of semantic aspects of input document datasets , their representation in conjunction with smoothness and manifold assumptions in semi supervised learning may give more relevant classification results.

General Terms

Machine learning , Information retrieval , text classification.

Keywords

Semi supervised learning, Multi label text classification, Smoothness and manifold assumptions.

1. INTRODUCTION

Recently the area of Multi label text classification has attracted significant attention from lot of researchers, as playing a crucial role in many applications such as web page classification , classification of news articles, information retrieval etc[6].Generally Supervised methods are used in working principle of multi label classification. But in real practice availability of labeled data is rare and that of unlabeled data is plenty [9]. Major limitation of existing supervised algorithms for multi label text classifiers is that they need labeled training data to learn accurately[9][10]. But acquisition of labeled training data is not as easy as that of getting unlabeled data. We need human intervention to label the given text document which is not only time consuming but error prone also [14]. This demands other sources of information that can reduce the need for labeled data. So now a day many researchers are looking towards semi supervised learning as promising solution to the give problem.

Semi supervised learning is closely related with the inference of information from data and thus plays crucial role in the classification applications [1][2]. It is generally most useful whenever there are far more unlabeled data than labeled. Its goal is to reduce the classification errors using unlabeled data in conjunction with labeled data [2][9]. So through our paper, we are providing comparative study on various multi label text classification approaches based on semi supervised learning and their analysis. We are also providing the overview of evaluation measures that can be used to evaluate multi label text classifier in the setting of semi-supervised learning. This entire work we have organized as follows in this paper.

Section 2 briefly describes the multi label text classification scenario and its general types

Section 3 highlights working methodologies of some popular semi supervised multi label text classification techniques and shows their analysis in terms of merits and demerits.

In section 4 we discussed various evaluation measures used in this problem setting Section 5 concludes this analysis and provides our insight on this upcoming scenario.

2. OVERVIEW OF MULTI LABEL TEXT CLASSIFICATION

The goal of text classification system is to determine the correct class of a new text document based on some training examples. Thus consideration of semi supervised machine learning method for building text classifier is an interesting area for research. Some of the research in the area of text classification focuses on some specific properties of text data. One such a property is its multi-labelity [3]. Multi-label text classification system is one key domain in this research area. Multi-label classification studies the problem in which a data instance can have multiple labels [4]. Semi supervised methods for text classification is also present in the literature. But very few techniques are available for solving multi-label text classification problem.

Multi-label classification is the task of assigning an object simultaneously to one or multiple classes [4]. In the literature, different methods have been proposed to be applied to multi-label text classification problems. These methods are falling under two broader categories of problem transformation methods and algorithm adaptation methods. Under each of this stated method many algorithms are proposed in the literature. Almost all these existing algorithms are supervised in nature , that means set of labels associated with each instance are already provided in the training data.

A) Problem transformation Methods (PT):
Problem transformation is the process whereby a multi-label problem is transformed into one or more single-label problems. In this scheme, single-label classifiers are employed, and their single-label predictions are transformed into a multi-label prediction[3][7]. The prime advantage of problem transformation is flexibility. In this single-label classifier can be used to suit requirements of the application. The performance of the classifier mainly depends on the application context.
Another advantage of problem transformation is that it can abstract away from classifier specifics and be more generally applicable by focusing on issues relevant to all multi-label domains such as modeling label correlations[4].
Some of the well known supervised problem transformation methods are Binary relevance method, Pairwise classification method, Label powerset method , Prunned sets method, Ensembles of pruned sets method, Random k-labelsets method, Ranking by pairwise comparison method, Caliberated label ranking method, Collective multi-label classifier, Metalabeler and classifier chains method.

B) Algorithm Adaptation Methods(AA):
In this approach, internal working mechanisms of single-label classifier is modified in such a way that they can solve multi label problems. The new algorithms can be developed specifically for multi label problems. This approach is often highly suited to specific domains or contexts but not as flexible as a problem transformation approach[4][5].

Algorithm adaptation methods are usually designed with a specific domain in mind. For example, decision trees are typically used on biological datasets while Bayes-based mixture models are commonly used specifically on text data
Some of the well known supervised algorithm adaptation methods are C4.5 , AdaBoost.MH, AdaBoost.MR, Multilabel k-Nearest Neighbours, Back-Propogation Multi-label Learning(BPMLL).
Table 1 presents a comparative analysis of few popular approaches from problem transformation and algorithm adaptation methods with respect to their merits and demerits[8].

Table 1: Comparative analysis of supervised multi label text classification algorithms namely PT and AA.

Algorithms	**Merits**	**Demerits**
BR	conceptually simple and relatively fast	It does not explicitly model label cor-relations. And affected by class-imbalance , needs all labeled data for classification
PW	conceptually simple	Time complexity is an issue for PW ,This method is criticized for not dealing well with overlapping labels , needs all labeled data for classification.
Label Power set	can take into account label correlations	*Can also be computationally complex ,Leads to overfitting of the training data.
Pruned Sets Method	*Run much faster , handles irregular labeling ,Can take into account label correlations.	It relies on prediction confidence distributions of the base classifier. It can not handle unlabeled data.
Ensembles of Pruned Sets(EPS)	*Provides increased predictive performance & parallelism	Cannot utilize available unlabeled data for classification
C4.5	*It allows attributes selection for splitting the data for easy learnability	*It does not take into account the correlation among the classes & It cannot able to utilize the unlabeled data for classification.
AdaBoost .MH& AdaBoost .MR	*Improved accuracy and minimization of Hamming loss error.	*Attempts for generalization results into decrease in performance & Cannot utilize unlabeled data for classification.
ML-kNN	*Improved performance & Can work well on image as well as textual data.	* Cannot utilize unlabeled data for classification.
Back-propogation algorithm for multilabel learning(BP-MLL)	*Gives better generalization capability to learning system.	*Computational complexity in training phase is high because of use of neural networks & Cannot able to utilize unlabeled data for classification.
SVM-HF	*Take into account correlation among classes & multilabel data.	Accuracy reduces with consideration of unlabeled data

If we observe the demerits of these supervised methods then we found that as per their very basic nature they are not capable of utilizing the available unlabeled data effectively for classification. But in practice large number of unlabeled data

is readily available . Also, the task of manually assigning labels to the unlabeled documents is error prone as well as time consuming. That's why one needs the approach which can utilize unlabeled data effectively along with the available small amount of labeled data for classification. Semi supervised learning can play important role in this regard. Semi supervised learning addresses this issue by using large amount of unlabeled data along with labeled data. So major objective behind using semi supervised techniques for multi-label text classification is to maintain or improve accuracy of these classifiers using unlabeled data. Following section describes few approaches which have successfully applied semi supervised techniques to to train multi label text classifier.

3. SEMI SUPERVISED MULTI LABEL TEXT CLASSIFICATION APPROACHES

This section presents the sort of analysis of recently proposed semi supervised multi label text classification approaches with respect to their working strategy , datasets used for experimentation , merits and shortfalls.

3.1 Expectation Maximization (EM) based text classification.

Nigam and Mccallum [9] developed this algorithm in 1999. It was very popular attempt to introduce semi supervised learning for text document classification. In this technique the authors have proposed updation in the basic EM technique by considering unlabeled data as incomplete data as it is coming without labels. EM is a class of iterative algorithms for max. Likelihood or max. a posteriori estimation in problems with incomplete data.

They carried out experiments on WebKB , Reuters , 20 Newsgroups datasets and computed accuracy.

The mode of working of this method is as follows :

1. The algo. First trains the classifier using available labeled documents.
2. This classifier is then used to assign class labels to unlabelled docs probabilistically by calculating the expectation of missing class labels.
3. It then trains a new classifier using all the documents through iterations.
4. EM performs hill-climbing approach in data likelihood space, finding the classifier parameters that locally maximize the likelihood of all the data.

Merits :

This method successfully able to utilize unlabeled data alongwith labeled data by using the joint distribution over features other than the class labels to classify the text documents.

Demerits :

It works for single label text classification only . It does not works well for multiclass and multi label text classification.

3.2 Multi-label classification by Constrained Non-Negative Matrix Factorization.

Y. Liu, R. Jin, L. Yang [10] proposed this algorithm in 2006. In this proposal the classification of multi label text documents is performed by considering the relationship between class labels and similarity between input documents.

They carried out experiments on the textual data of ESTA and computed precision, Recall , F-measure.

The mode of working of this method is as follows :

1. It first represent all the documents by feature matrix
2. Computes similarity matrix by calculating similarity between documents using cosine similarity measure.
3. Computes similarity between class labels using RBF kernel.
4. Performs label assignment to unlabeled data based on minimum difference between the two sets of similarity calculated in step 2 and step 3. This difference between two sets of similarities is represented as constrained non negative matrix factorization.

Merits :

- It able to effectively utilize the semi-supervised learning scenario in conjunction with low rank representation of input documents using Constrained Non-Negative Matrix Factorization.
- It even works well in the scenario of small number of training data and large no. of class labels.

Demerits :

- There is strong influence of two parameters on the performance of this method , these two parameters are namely latent variables and tuning parameter. Wrong value chosen greatly degrades the performance.

3.3 Graph-based SSL with multi-label

Z. Zha, T. Mie, Z. Wang, X. Hua [11] proposed this algorithm in 2008 . In this work graph based learning framework is proposed. This framework is exploiting correction among labels along with label consistency over graph in the setting of semi-supervised learning with multi label text .
They carried out experiments on video files annotation dataset TRECVID 2006 and computed average precision.

The mode of working of this method is as follows :

1. The vector valued function over the Graph is defined as :

$$E(F) = E_l(F) + Es(F) + Ec(F)$$

The function exhibits 3 properties as

-E_l (F) corresponds to the property that function should be close to the given labels.
-Es (F) corresponds to the property that function should be smooth over the graph.
-Ec(F)Consistent with label correction.

2. Capture the label correction using symmetric matrix to represent the correlation between labels.
3. Uses equation of Ec(F) to predict the consistency between the multiple labels.
4. Uses equation of E(F) to predict the labels of unlabeled instances.

Merits:

It offers effective utilization of large amount of unlabeled data and also able to exploit relationship between labels.

Demerits :

Applicable to video files not works well for text.

3.4 Multi-label learning by using dependency among labels:

Wei, Yang, Zhu and Wang proposed [14] this method in 2011.

They carried out experiments on emotions, yeast, and scene datasets and computed accuracy.

The mode of working of this method is as follows:

1. Explored relationship among labels by feature selection algo.

2. Adopted cross-validation method to calculate the classification accuracy on each label.

3. Based on the accuracy, classification order for labels is determined.

4. Trained a list of classifiers according to the order each classifier in list with additional features which are provided by the outputs of the previous classifiers in the list.

Merits :

Improved accuracy using SSL setting.

Demerits:

Increased time complexity for large datasets.

3.5 Semi supervised multi-label learning by solving a Sylvester Eg [SIAM]

Chen , Song , Wang , Zhang proposes this algorithm[12] in 2008 .
They carried out experiments on Reuters datasets and computed accuracy.

The mode of working of this method is as follows:

1. Two graphs are constructed for input documents and class labels resply.

2. For input documents the nodes in the graph represents both labeled & unlabeled instances. Each edge between nodes represents similarity between pair wise instances.

3. In the similar fashion the graph is constructed for class labels.

4. Defined a quadratic energy function on each graph.

5. The labels for unlabeled data can be inferred by minimizing the combination of the two energy terms.(By solving Sylvester eg)

Merits:

It offers effective utilization of large amount of unlabeled data and also able to exploit relationship between labels. Significant improvement in the accuracy.

Demerits :

May get slower on convergence for large data set.

3.6 Semi-Supervised Non negative Matrix Factorization.

Lee , Yoo & Choi proposed this algorithm in 2009. They developed semi-supervised non negative matrix factorization by using multiplicative updates[13].
They carried out experiments on 20-news , CSTR , k1a,k1b, Reuters, WebKB4 and computed accuracy.

The mode of working of this method is as follows:

1) Represent documents as data matrix.
2) Represent relationship between labels as label matrix.
3) Performs joint factorization of data and label matrix by using equation involving data matrix, common factors, basic feature matrix.
4) Computed multiplicative updates for common factor matrix, basic feature matrix and data matrix.

Merits:

Use of NMF in conjunction with SSL setting ables to extract more discriminative features than the standard NMF.

Demerits:

High computational complexity

4. EVALUATION MEASURES USED

This section addresses the issue of evaluating performance of multi label text classifier irrespective of whether algorithm is based on supervised , unsupervised or semi supervised method.

In a multi-label text classification problem an example may be associated with set of labels therefore classification of an example may be partially correct or partially incorrect[3]. This can happen when a classifier correctly assigns an example to at least one of the labels it belongs to, but does not assign to all labels it belongs to. Also, a classifier could also assign to an example to one or more labels it does not belong to [7].

The commonly used performance evaluation measures for multi-label classifiers are broadly categorized in two groups namely bipartition-based and ranking-based [3]. Bipartition-based measures are again having two types called examples-based measures and label-based measure. Example-based measures evaluate bipartition over all the examples of the evaluation dataset. Label-based measures decomposes the evaluation process into the separate evaluations for each label. Whereas the ranking-based measures evaluate ranking with respect to the ground truth of multi-label dataset.

However, for the definitions of these measures, let an evaluation dataset of multi-label examples be denoted as $(x_{i,} y_i), i = 1$ to N, $y_i \subseteq L$, is the set of true labels and L={x $_i$ $j =, m\}$ is the set of all labels . Given an examples x , the set of labels that are predicted by an multi-label method is denoted as z. while the rank predicted for a label is denoted as Z_i ,the most relevant label receives the highest rank(1),while the least relevant one receives the lowest rank(M)[3][4].

Example based measures includes Exact match (accuracy), Hamming loss, Precision ,Recall, F-measure. Label based measure includes macro-averaging and micro-averaging. Whereas ranking based measures includes one-error , coverage and average precision, log-loss.

We evaluated the said algorithms by measuring values of some of the representatives from above measuring techniques. This includes Accuracy , Example based accuracy that is Exact-match, F-measure and log loss.

Exact Match : It is the accuracy measure in the example based scheme. It is computed as :

$$\textbf{EXACT-MATCH(D)} = \frac{1}{N}\sum_{i=1}^{N} 1yi = yi$$

F-Measure: F-measure is a combination of precision and recall. It is the harmonic average of the two metrics and it is used as an aggregated performance score.

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

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

Accuracy: This measure is proposed by Godbole and Sarawagi in [15] which is independent of example – based and label – based accuracy measures. It is now a day most popular multi-label accuracy measure. It symmetrically measures how close y_i is to Z_i . 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.

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

Log-Loss: This measure is introduced by Jeese Read [4] to overcome some of the limitations of the ranking loss measures. Under this each label error is graded by the confidence at which it was predicted. It also takes into account label relevances at the time of predictions.

LOG-LOSS(D)=

$$\frac{1}{NL}\sum_{i=1}^{N}\sum_{j=1}^{L}\min\left(-LOG-LOSS(wj, yj), \ln(N)\right)$$

5. CONCLUSION

In this paper, we have presented a analysis of some popular semi supervised learning based multi label text classification methods that have proposed since last few years. This analysis provides a useful insight to researchers who are willing to work on semi-supervised techniques for multi label classification.

Out of these methods the first one works only for single label text classification task the rest are applicable to multi label domain. But the said method can also be extended to work for multi label domain using algorithm adaptation strategy and may be well suited for applications which are iterative in nature.

The methods described at subsection 3.2 and 3.6 are focusing on documents representation and its optimization also in the form of non negative matrix factorization approach. However to reduce computational complexity and to provide scalable operation , one may use Alternating least square method or gradient decent method in the feature this representation step.

The methods described at subsection 3.3 and 3.5 are based on graph based framework in conjunction with similarity measures such as RBF kernel, cosine and Sylvester equation. The use of Gaussian kernel measures can also lead to better performance.

Based on this analysis we may like to provide method which can take into account manifold and smoothness assumption of semi supervised learning to improve classification accuracy of multi label text classifier. Use of more powerful input document representation such as latent semantic indexing may help in improving accuracy as impurities, noise, existence of redundant features may reduce accuracy of text classifier.

6. 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, "Multi-label text classification for automated tag suggestion", Proc. Of the ECML/PKDD 2008 Discovery Challenge, Antwerp , Belgium(2008)

[7] 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] 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.

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

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

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

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

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

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