International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 04, April 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

DATA CLASSIFICATION APPROACH USING MULTI VIEW GRAPH LEARNING TECHNIQUE

Omkar Chandole, Sanket Jangle, Satish Suryawanshi, Sagar Wagh

Computer Engineering,G.S Moze Collage Of Engineering,Savitribai Phule Pune University

ABSTRACT

Every day the mass of information available to us increases. This information would be irrelevant if our ability to efficiently access did not increase as well. For maximum benefit, we need tools that allow search, sort, index, store and analyze the available data. One of the promising area is the automatic text categorization. Imagine ourselves in the presence of considerable number of texts, which are more easily accessible if they are organized into categories according to their theme. Of course one could ask human to read the text and classify them manually. This tasks is hard if done on hundreds, even thousands of texts. So, it seems necessary to have an automated application, we present automated text categorization using machine learning approach. An increasing number of data mining applications involve the analysis of complex and structured types of data and require the use of expressive pattern languages. Many of these applications cannot be solved using traditional data mining algorithms. This observation forms the main motivation for the multi-disciplinary field of Multi-Relational Data Mining (MRDM).

Keywords:- Text categorization, Machine learning, Multi-Relational Data Mining

I. INTRODUCTION

Text classification is to map the text to one or more predefined categories using a kind of classification algorithm which is accomplished according to text content. A standard classification corpus has been established and a unified evaluation method is adopted to classify English text based on machine learning which has made a large progress now. In the last ten years content-based document management tasks (collectively known as information retrieval – IR) have gained a prominent status in the information systems field, due to the increased availability of documents in digital form and the ensuing need to access them in flexible ways. Text categorization (TC), the activity of labelling natural language texts with thematic categories from a predefined set, is one such **task.** Machine learning (ML) paradigm, according to which a general inductive process automatically builds an automatic text classifier by learning,

from a set of preclassified documents, the characteristics of the categories of interest. The advantages of this approach are an accuracy comparable to that achieved by human experts, and a considerable savings in terms of expert manpower, since no intervention from either knowledge engineers or domain experts is needed for the construction of the classifier

Current solution does not scale well and cannot realistically be applied when considering database containing huge amount of data. In order to learn classification models, we propose a multi-graph-view bag learning algorithm (MGVBL), which aims to explore sub graph features from multiple graph views for learning, which aims to learn a classifier from a set of labeled bags each containing a number of graphs inside the bag. A bag is labelled positive, if at least one graph in the bag is positive, and negative otherwise.

Such a multi-graph representation can be used for many real-world applications, such as webpage classification, where a webpage can be regarded as a bag with texts and images inside the webpage being represented as graphs. Another MGVBL application is scientific publication classification, where a paper and its references can be represented as a bag of graphs and each graph (i.e., a paper) is formed by using the correlations between keywords in the paper, as shown in Fig. 1.

International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 04, April 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

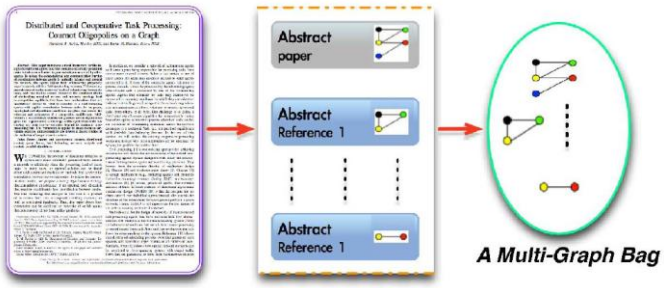

Fig. 1. Example of multi-graph representation for a scientific publication.

A bag is labelled positive, if the paper or any of its references is relevant to a specific topic. Similarly, for online review based product recommendation, each product receives many customer reviews. For each review composed of detailed text descriptions, we can use a graph to represent the review descriptions. Thus, a product can be represented as a bag of graphs. A product (i.e., a bag) can be labeled as positive if it receives atleast one positive review else negative. As a result, we can use MGVBL learning to help recommend products to customers.

Text categorization

Text categorization is the task of assigning a Boolean value to each pair <dj , ci> ∈ D×C, where D is a domain of documents and C = {c1, . . . , c|C|} is a set of predefined categories. A value of T assigned to <dj , ci> indicates a decision to file dj under ci, while a value of F indicates a decision not to file dj under ci. More formally, the task is to approximate the unknown target function ϕ^ : D ×C → {T, F} (that describes how documents ought to be classified) by means of a function ϕ: D × C → {T, F} called the classifier (aka rule, or hypothesis, or model) such that ϕ ^ and ϕ "coincide as much as possible".

Single-label vs. multi-label text categorization

Different constraints may be enforced on the TC task, depending on the application. For instance we might need that, for a given integer k, exactly k (or ≤ k, or ≥ k) elements of C be assigned to each dj ∈ D. The case in which exactly 1 category must be assigned to each dj ∈ D is often called the single-label (aka non-overlapping categories) case, while the case in which any number of categories from 0 to |C| may be assigned to the same dj ∈ D is dubbed the multi-label (aka overlapping categories) case.

II .IMPLEMENTATION DETAIL

Here we will see system architecture, algorithm and mathematical model.

A. SYSTEM ARCHITECTURE

In Fig. 2, the proposed multi-graph-view learning for graph bag classification (MGVBL). In each iteration, MGVBL selects an optimal sub graph g_ (step a). If the algorithm does not meet the stopping condition, g will be added to the sub graph set g or terminates otherwise (step c). During the loop, MGVBL update the weights for training graph-bags and graphs. The weights are continuously updated until obtaining the optimal classifier. Decision stump verifies the bag label if it is correct, if not it updates weight of each bag and the process is repeated unless we get a optimal result. View learner can specify no. of iterations, as no of iterations increases accuracy is more, but as number of iterations increases time complexity will also increase so we suppose to find best balance between number of iterations and time complexity for maximum accuracy. In propose system we are using Ant colony Optimization algorithm to check the maximum support (weight) for positive bag for particular class label.

International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 04, April 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

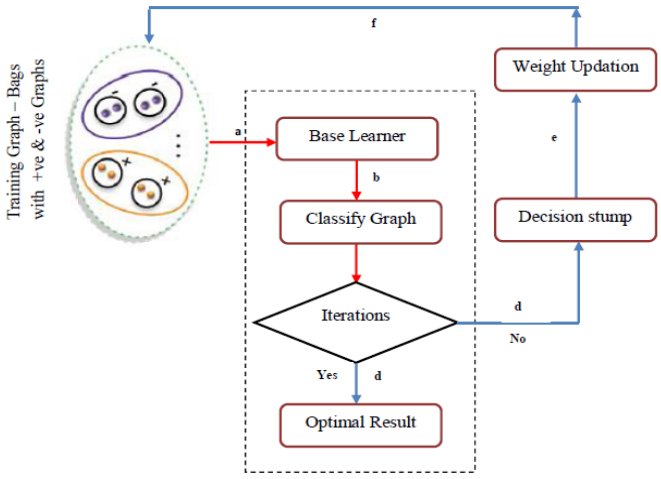

Fig. 2. Overview of the proposed bMGC framework.

B. Algorithm

Algorithm *(MGVBL)*: *multi-graph-view bag learning algorithm*

The document set provided as an input must be pre-processed as it contains texts that are irrelevant for classification. The pre-processing includes tokenization, stop word removal, stemming and term weighting.

1) Tokenization: It is the process of splitting stream of text into meaningful words or phrases. The words are split based on the special delaminating characters such as spaces, punctuation, and symbols etc.

2) Stop Word Removal: Frequently occurred words, like pronouns, prepositions and conjunctions in English e.g. 'it', 'in', 'and', etc. are known as stop words. These words from the text documents are having a very low discriminative value. It includes creating a list of stop words and then scanning the tokens to remove the stop words occurred.

3) Stemming: It is the process of finding the root word of the token. For example, the words "purification", "purity", "purify" and "purifying" having stemmed root as "pure". Stemming words helps to reduce the dimensionality of the feature space. The Porter stemming algorithm is used, which is a natural language processing (NLP) by removing the suffix, to narrow down the size of the feature space.

4) Term Weighting: TF-IDF is a term weighting approach which is one of the widely used methods to ealuate the importance of a term in the corpus or identifies how relevant a term is to the classification.

A. Feature Extraction

It is the process of converting the text feature into feature vector. For the representation of text we are going to use the vector space model in our proposed system.

B. Feature Selection

Feature selection is used for dimensionality reduction of original feature set to get the more relevant feature space for classification. In our proposed system we used the combination of ACO and GA proposed in.

C. Similarity Based techniques

1) Cosine Similarity: The similarity between two documents which are considered as nodes can be calculated using cosine similarity. The cosine similarity is calculated using formula (1).

2) Transition probability: The transition probability can be calculated using formula (3) as follows.

$$P_{ij} = \frac{\tau_j}{\sum_\tau \tau_i}$$

C. Evaluation Of Classification

To evaluate the performance of the classifier is evaluated according to the accuracy results. In order to compare the predicted categories assigned by classifier with the actual categories of the test documents, first of all the number of True Positives, False Negatives and False Positives are determined, then precision, recall and accuracy is computed using these values.

Algorithm1 ACO-GA

A. *Feature Extraction*

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

It is the process of converting the text feature into feature vector. For the representation of text we are going to use the vector space model in our proposed system.

B. *Feature Selection*

Feature selection is used for dimensionality reduction of original feature set to get the more relevant feature space for classification. In our proposed system we used the combination of ACO and GA.

C. *Similarity Based techniques*

1) Cosine Similarity: The similarity between two documents which are considered as nodes can be calculated using cosine similarity. The cosine similarity is calculated using formula (1).

2) Transition probability: The transition probability can be calculated using formula (3) as follows.

$$P_{ij} = \frac{\tau_j}{\sum_\tau \tau_i}$$

The next node for the classification will be selected by taking product of formula (1) and formula (3).

D. *ACO classification*

In the proposed system we adopt the algorithm of Ant Colony Optimization (ACO) suggested in [2]. Once we get the précised set of features, which is outcome of proposed ACO-GA approach for feature selection. On this feature set we will apply the ACO based approach for text categorization. There were several techniques available for text classification. The proposed approach will give the better results for classification of text documents into its correct category than other approaches. The proposed approach can improve the efficiency and performance of the classification.

E. *Evaluation Of Classification*

To evaluate the performance of the classifier is evaluated according to the accuracy results. In order to compare the predicted categories assigned by classifier with the actual categories of the test documents, first of all the number of True Positives, False Negatives and False Positives are determined, then precision, recall and accuracy is computed using these values.

Algorithm 2 TF-IDF

Step1: TF (t) = (Number of times term t appears in a individual node) / (Total number of terms in the bags).

Step 2: IDF(t) = log_e(Total number of bags / Number of no with nodes t in it).

Step 3: Term Weighting: TF-IDF is a term weighting approach which is one of the widely used methods to evaluate the importance of a term in the corpus or identifies how relevant a term is to the classification. It can be calculated as follows.

$$W(i,j) = tf_{ij} \cdot \frac{N}{df_{ij}}$$

Step 5: Finish Procedure

Algorithm 3 cosine similarity

Cosine Similarity: The similarity between two documentswhich are considered as nodes can be calculated usingcosine similarity. The cosine similarity is calculated usingformula (1).

2) Transition probability: The transition probability can becalculated using formula (3) as follows.

$$P_{ij} = \frac{\tau_j}{\sum_\tau \tau_i}$$

The next node for the classification will be selected by takingproduct of formula (1) and formula (3).

VI. ACKNOWLEDGMENT

I would like to express my sincere thanks of gratitude to my esteemed guide Prof. Sonia Mehta for her guidance and suggestions during this work. I am thankful to Prof. Priyadarshani Kalokhe, HOD of computer Dept for providing me the necessary lab facility and software. I am thankful to all who have directly or indirectly guided and helped me in delivery of this work.

I would like to thank the researchers as well and publishers for making their resources available and teachers for their guidance. I am thankful to the authorities of Savitribai Phule University of Pune and concern members of cPGCON2016 conference, organized by, for their constant guidelines and support. We are also thankful to the reviewer for their valuable suggestions.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Finally, we would like to extend a heartfelt gratitude to friends and family members.

V. CONCLUSION

In this work, we explored a novel Multi Graph Classification (MGC) issue, in which various charts frame a sack, with every pack being marked as either positive or negative. Multi-chart representation can be utilized to speak to some true applications, where name is accessible for a sack of items with reliance structures. To construct a learning model for MGC, we proposed a bMGC, which utilizes dynamic weight conformity, at both diagram and sack levels, to choose one subgraph in every cycle to frame an arrangement of powerless diagram classifiers. The MGC is accomplished by utilizing weighted blend of powerless chart classifiers. Probes two certifiable MGC assignments, including DBLP reference system and NCI concoction compound order, show that our technique is viable in finding useful subgraph, and its precision is fundamentally superior to anything gauge techniques.

REFERENCES

[1] Boosting for Multi-Graph Classification Jia Wu, Student Member, IEEE, Shirui Pan, Xingquan Zhu, IEEE TRANSACTIONS ON CYBERNETICS, VOL. 45, NO. 3, MARCH 2015

[2] R. Angelova and G. Weikum, "Graph-based text classification: Learn from your neighbors," in Proc. 29th Annu. Int. ACM SIGIR, Seattle, WA, USA, 2006, pp. 485–492.

[3] M. Kuramochi and G. Karypis, "Frequent subgraph discovery," in Proc. 1st ICDM, 2001, pp. 313–320.

[4] T. Dietterich, R. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," Artif. Intell., vol. 89, no. 1–2, pp. 31–71, 1997.

[5] X. Yan and J. Han, "gSpan: Graph-based substructure pattern mining," in Proc. 2nd ICDM, Washington, DC, USA, 2002, pp. 721–724.

[6] A Survey on Approaches of Multirelational Classification Based On Relational Database Shraddha Modi, Amit Thakkar, Amit Ganatra, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-1, Issue-3

[7] Lachetar, N. ; Comput. Sci. Dept., Univ. 20 Aout 1955, Skikda, Algeria ; Bahi, H. Application of an ant colony algorithm for text indexing, :Multimedia Computing and Systems (ICMCS), 2011 International Conference –IEEE 2011

[8] Ant Colony optimization L Jiao, L Feng - Information and Computing (ICIC), 2010 - ieeexplore.ieee.org

[9] Guo, H., Herna, L., Viktor.. Multirelational classification: a multiple view approach, Knowl. Inf. Systems, vol.17,pp.287–312,Springer-VerlagLondon.2008.

[10] W. Lian, D.-L. Cheung, N. Mamoulis, and S.-M. Yiu, "An efficient and scalable algorithm for clustering XMLdocuments by structure," *IEEE Trans. Knowl. Data Eng.* , vol. 16, no. 1, pp. 82–96, Jan. 2004.

[11] C. Chen *et al.*, "Mining graph patterns efficiently via randomized summaries," in *Proc. 35th Int. Conf. VLDB*,Lyon,France,2009,p.742–753.

[12] H. Wang, H. Huang, and C. Ding, "Image categorization using directed graphs," in *Proc. 11th ECCV*, Crete,Greece,2010,pp.762–775.

[13] R. Angelova and G. Weikum, "Graph-based text classification: Learn from your neighbors," in *Proc. 29th Annu.Int.ACM SIGIR*, Seattle, WA, USA, 2006, pp. 485–492.