ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

Data Classification Approach Using Multi View Graph Learning Technique

Deepak Singh, Prof. Ayonija Pathre

PG Students, Dept of Computer Science Engineering AISECT University, Bhopal, India

Asst Professor, Dept of Computer Science Engineering AISECT University, Bhopal, India

ABSTRACT: Data classification is necessity of today's world as large amount of information available to us in today's digital world and it increases continuously. This information would be useless and not relevant if we are not able 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 classification. Imagine we have considerable number of texts, which are more easily accessible if they are put into different categories according to their contents. 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 tasks includes the analysis of complex and structured types of data and make use of expressive pattern languages. Most from these applications can't be solved using traditional data mining algorithms. This is the cause of 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 aims to classify the text to some predefined categories using a kind of classification algorithm which is performed based on text content. The 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 past ten years management of document based contents (collectively known as information retrieval – IR) have become very popular in the information systems field, due to the greater availability of documents in digital form and resulting need to access them in flexible and efficient ways. Text categorization (TC), the activity of labeling natural language texts with one or more categories from a predefined set, is one such task. Machine learning (ML) methodology, according to which we can automatically build an automatic text classifier by learning, from a set of pre-classified text documents based on the characteristics of the categories of interest. The advantages of this approach is accuracy as compared to that obtained by human beings, and a considerable savings in terms of expert manpower, since no contribution from either domain experts or knowledge engineers is needed for the construction of the classifier.

Recent years have experienced an increasing number of applications where instances (or samples) are no longer represented by a flat table with instance-feature format, but share some complex structural dependency relationships. Typical examples include XML webpages (i.e. an instance) which point to (or are pointed to) several other XML webpages , a scientific publication with a number of references , posts and responses generated from social networking sites, and chemical compounds with molecules (i.e. nodes) linked together through some bounds (edge) . Given a collection of graph data {Gi, yi}, yi∈ {+1,-1}, each of which is labeled (+1 for +ve graphs, -1 for -ve graphs), we propose here graph classification model which aims to build prediction model with maximum accuracy in classifying previously unseen graphs. In order to learn classification models, we propose a multi-view-graph learning algorithm (MVGL), which aims to explore sub graph features from multiple graph views for learning, which aims from

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

a set of labeled bags to learn a classifier each having many number of graphs inside the bag. A bag is considered positive, if at least one graph in the bag is positive, and negative otherwise.

Such a multi-graph presentation can be used for large number of real-world applications, such as classification of webpage, where a webpage can be considered as a bag with texts and images inside it being represented as graphs. Another 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.

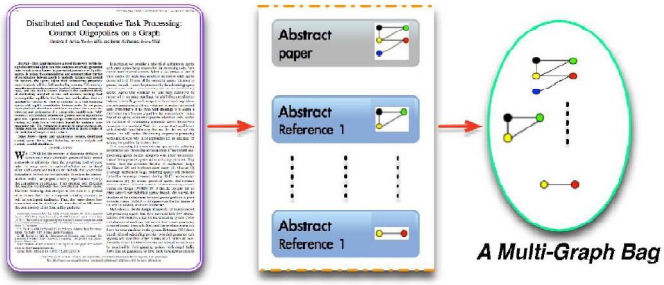

Fig. 1. Example of multi-graph representation for a scientific publication.

A bag is labelled +ve, if the paper 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 bag (product) can be labelled as positive if it receives at least one positive review else negative. As a result, we can use graph based classification learning to help recommend products to customers.

Text categorization

Text categorization is the duty of assigning a Boolean value to each pair <di , cj>∈ D×C, where D is a domain of documents & C = {c1, . . . , c|C|} is a set of predefined categories. A value of T assigned to <di ,cj> indicates a decision to file di under cj, while a value of F indicates a decision not to file di under cj. 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 such that ϕ ^ and ϕ "coincide as much as possible".

Single-label vs. multi-label text categorization

Various constraints may be applied on the Text Classification task, depending on the application. For instance we may need that, for a given integer n, exactly n (or ≤ n, or ≥ n) elements of C be assigned to each di∈ D. The case in which exactly 1 category must be assigned to each di ∈ D is often called the case of single-label, while the case in which any number of categories from 0 to |C| may be assigned to the same di ∈ D is dubbed the multi-labelcase.

II. LITERATURE SURVEY

The most relevant work to this paper is presented in [1]. The author presented a boosting based multi-graph classification framework (bMGC). Suppose a set of labeled multi-graph bags, bMGC employs dynamic weight adjustment at both bag- and graph-levels to select one subgraph in each iteration as a weak classifier. In each iteration, bag and graph weights are adjusted such that an miss classified bag will receive a higher weight because its predicted bag label conflicts to the genuine label, whereas an miss classified graph will receive a lower weight value if the graph is in a +ve bag (or a higher weight if the graph is in a -ve bag). Accordingly,

bMGC is able to properly differentiate graphs in +ve and -ve bags to derive effective classifiers to form a boosting model for MGC.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

MGC is a generalization of the MIL problem, which was first proposed by Dietterich et al. [4] for predicting of drug activity. The multiple instance problem occurs in tasks where the training examples are uncertain: a single object may have many different feature vectors (instances) that describe it, and only one of those feature vectors may be responsible for classification of the object.

Extracting important subgraph features, using some predefined criteria, to represent a graph in a vectorial space becomes a popular solution for graph classification. The most common subgraph selection criterion is frequency, which intends to select frequently appearing subgraphs by using frequent subgraph mining methods. For example, one of the most popular algorithms for frequent subgraph mining is gSpan [5]. Its uses depth first search (DFS) to search most frequent subgraph.

Application of an ant colony algorithm for text classification [7] [8]: Every day, the amount of information available to us increases. This information would be useless and not relevant if our ability to efficiently access didn't increase as well. For greater benefit, we need tools that allow us to search, sort, index, store, and analyze the available data. For greater benefit, we need tools that allow us to search, sort, index, store, and analyze the available data. We also need tools which helps us to find desired information in a reasonable time by performing certain tasks for us. One of the promising areas is the automated text classification. Just imagine we have considerable number of texts, which are more easily accessible if they are organized into categories according to their theme. Of course, we can ask a human to classify the texts by reading them manually. This task is very tough if we do it for hundreds, even thousands of texts. So, it seems necessary to have an automatic text classification application. In this paper author presents his experiments in automated text categorization, where author suggest the use of an ant colony algorithm.

Classification is an important task in data mining and machine learning, in which a model is formed based on training dataset and it is used to predict class label of unknown dataset. Now a days most real-world data are stored in relational databases. Therefore to classify objects in one relation, other relations provide crucial information. Relational databases are the popular format for structured data which consist of tables connected via relations (primary key/ foreign key). Thus relational databases are simply very much complex to analyze with a propositional algorithm of data mining [6].

To identify informative subgraphs with minimum redundancy, author proposes a subgraph assessment criterion to assess the in formativeness of individual features and the redundancy of the whole feature set at the same time. To capture instances represented by emerging concept drifting in graph streams, author employ a dynamic instance weighting mechanism, where graphs misclassified by the existing model receive a higher weight so the subgraph feature selection can emphasize on difficult graphs to find effective features to represent them[9].

III. PROBLEM STATEMENT

We propose automated text categorization application which gives classification of scientific publications (pdf) with maximum accuracy and minimum time complexity. In order to learn classification models, we propose a multi-view-graph learning algorithm (MVGL), which aims to explore sub graph features from multiple graph views for learning. By enforcing labeling constraints at both the bag and graph levels, MGVL is able to discover most effective sub graph features across all graph-views for learning.

IV. IMPLEMENTATION DETAIL

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

A. SYSTEM ARCHITECTURE

In Fig. 2, the proposed multi-view-graph learning for graph bag classification (MVGL). In each iteration, MVGL selects an optimal sub graph g_ (step a). If the algorithm fails to meet the stopping condition, g will be added to the sub graph set g or terminates otherwise (step d). During the loop, MVGL update the weights for training graph-bags and graphs. The weights are continuously updated until obtaining the optimal classifier.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

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 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 (ACO) algorithm for feature selection and Genetic algorithm (GA)for classification by checking the maximum fitness score (weight) for positive bag for particular class label.

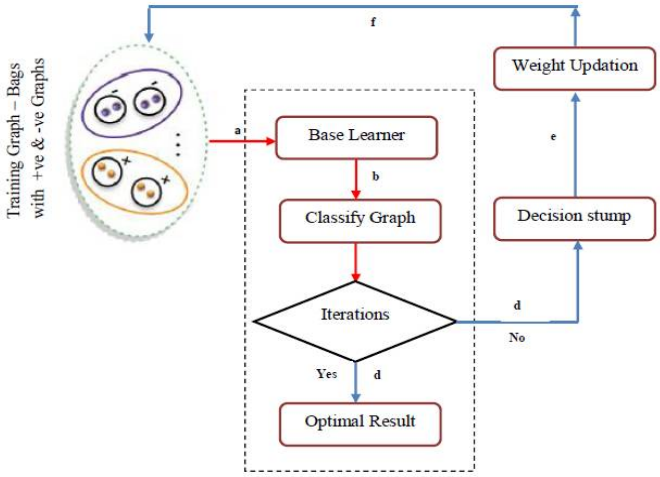

Fig. 2.Overview of the proposed MVGBL framework.

Creation of Training Data Set:
We had considered 100 scientific papers form different publication of different domains for creating training dataset. These papers are input for creating training dataset. This input is first preprocessed and most informative features are extracted using TF/IDF and cosine similarity. We have identified ten different domains from market and then extracted feature we have to put to corresponding domain where each domain is considered as one class that we will use for labeling in testing part and features are considered as nodes.

Feature Selection:
When .pdf (scientific papers) is inputted in testing, data preprocessing is done as in training for feature extraction and most informative features are selected using TF/IDF and cosine similarity. These features are considered as nodes of graph.

Graph Classification:
For classification these bags of graphs are compared against training master and weights are assigned to each of this bag as per their fitness score using GA algorithm. Each of this bag will be either positive or negative based on the assigned weight and thus each of bag (scientific paper) will be given single or multiple label based on for which domain it shows maximum weight.

B. Algorithm

Algorithm1 *(MVGL)*: *multi—view-graph 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.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

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 evaluate 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

Cosine Similarity and Transition probability techniques are used to calculate similarity between two documents.

D. 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.

Algorithm2 ACO- Ant Colony Optimization[8]

A. *Feature Extraction*

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

B. *Feature Selection*

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

C. *Similarity Based techniques*

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

$$\text{soft\_cosine}_1(a,b) = \frac{\sum_{i,j}^{N} s_{ij} a_i b_j}{\sqrt{\sum_{i,j}^{N} s_{ij} a_i a_j} \sqrt{\sum_{i,j}^{N} s_{ij} b_i b_j}}, \quad \ldots\ldots (1)$$

where s_{ij} = similarity(feature$_i$, feature$_j$).

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

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

$$P_{ij} = \frac{\tau_j}{\sum_\tau \tau_i} \quad \text{.....(2)}$$

The next node for the classification will be selected by takingproduct of formula (1) and formula (2).

$$\textbf{Products} = (s_{ij} * P_{ij})$$

D. *GA classification*

In this work finally system executes the classification phase[11]. Once the feature selection is done, we want to customized each instance for specific domain using Genetic evaluation approach. The algorithm can't be work base on variation basis it can execute once, The procedure for GA is given in below

- First we consider all the feature vectors as initial population.
- Set the generation count for GA
- Apply evolutionary algorithm.
- Calculate fitness of each chromosome using training dataset.
- Apply selection operator for select a best subset.
- Check generation is done or not.
- Terminate GA classify data domain wise.

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 3 TF-IDF

Step1: TF (t) = (Number of times term t appears in aindividual 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 approachwhich is one of the widely used methods to evaluate theimportance of a term in the corpus or identifies how relevant 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

C. Mathematical Model

Let S, be the proposed system which can be represented as
S = {{I}, {I_t, I_s, I_{st}, I_{tw}, I_{tr}, I_{ts}, I_{fs}, I_{cs}, $I_{tp,}$ I_r}, {R}}
Where,
I -> Input document collection (for Training and Testing)
I_t -> Abstract reading from input document
I_s ->Applying stop word removal on abstracts
I_{st}-> Applying Porter's Stemming on abstract
I_{tw}-> Feature Term Set with TF-IDF score
I_{tr}-> Training Feature Set
I_{ts}-> Test Feature Set
I_{fs} -> Test Feature subset depending upon the training feature set
I_{cs}-> Test feature subset with Cosine Similarity

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

I_{tp}-> Test feature subset with Transition Probability
R -> Test document labeled with the appropriate domain/ category

P is Learning Algorithm
Input = {Text Documents}
Output= {Categorized text with their labels}
Where, P represented as **Functions** like Tokenization, Stemming, Stop word Removal, Feature Selection and Feature Transformation.

P = {Fx | Input Output}
Fx is a function which takes input as text documents and give output as text indexing.
Let S, be the proposed system which can be represented as

$$S = \{\{I\}, \{P\}, \{O\}\}$$

Where,
I ->Input document collection (for Training and Testing)
P -> Functions used
O -> Test document labeled with the appropriate Domain
Where,

$$P = \{ f1, f2, f3, f4 \}$$

f1 ->Term Weighting (TF-IDF)
f2 -> Feature Selection Method (Evaluation Algorithm)
f3 -> Similarity Based Methods (Cosine Similarity and Transition Probability)
f4 -> Evaluation Parameters (Precision, Recall and Accuracy)

V. EXPECTED RESULT

The experimental results illustrate the benefits of multi-view learning methods compared to traditional single-view learning, Table 1 presents a list drawn from several published multi-view learning papers.(Varma and Babu, 2009 ,Zhu et al., 2012 and Rakotomamonjy et al., 2008 used the WebKB data as one of the evaluation datasets.

VI. 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 MVGBL, 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.System can able to classify data into view objects like one pdf can be classify into different domain base on features.

Future Work
Sometime system having a accuracy issues well false detection ratio, we can focus on such problems. The second part is system execution complexity when we work with high dimensional or big data. The system can be work with HDFS

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

framework for minimum time computation or parallel distribution. So For the enhancement system can be execute HDFS base architecture with parallel genetic algorithm.

ACKNOWLEDGMENT

I would like to express my sincere thanks of gratitude tomy esteemed guide Prof. Sonia Mehta for her guidanceand 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 ofSavitribaiPhule 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.

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

REFERENCES

[1] Boosting for Multi-Graph ClassificationJia 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 ShraddhaModi, AmitThakkar, AmitGanatra, 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] S. Pan, X. Zhu, C. Zhang, and P. Yu, "Graph stream classification using
labeled and unlabeled graphs," in *Proc. 29th IEEE ICDE*, Brisbane,
QLD, USA, 2013, pp. 398–409.

[10] M. Thoma et al., "Near-optimal supervised feature selection among
frequent subgraphs," in Proc. 9th SDM, 2009, pp. 1075–1086.

[11] Genetic algorithm Tutorial: Darrell Whitley, Computer Science Department, Colorada State University, Fort Collins CO 80523