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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 7, July 2016

New Approach for Data classification using Multi view graph learning Technique

Miss. Smital B. Deore Prof. Archana Jadhav

Department of Computer Engineering, Alard College of Engineering and Management, Pune, India

ABSTRACT: Text classification approach gaining more importance because of the accessibility of large number of electronic documents from a variety of resource. Text categorization (Also called Text Categorization) is the task of assigning predefined categories to documents. It is the method of finding interesting regularities in large textual, where interesting means non trivial, hidden, previously unknown and potentially useful. The goal of text mining is to enable users to extract information from textual resource and deals with operation such as retrieval, classification, clustering, data mining, natural language preprocessing and machine learning techniques together to classify different pattern. A major characteristic or difficulty of text categorization is high dimensionality of feature space. The reduction of dimensionality by selecting new attributes which is subset of old attributes is known as feature selection. Feature selection methods are discussed in this paper for reducing the dimensionality of the dataset by removing features that are considered irrelevant for the classification. This paper surveys of text classification, several approaches of text classification, feature selection methods and applications of text classification. 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.

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. MGC is a generalization

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 7, July 2016

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. PROPOSED SYSTEM

Text classification is a fundamental task in document processing. The goal of text classification is to classify a set of documents into a fixed number of predefined categories/classes. A document may belong to more than one class. When classifying a document, a document is represented as a “bag of words”. It does not attempt to process the actual information as information extraction does. Rather, in simple text classification task, it only counts words (term frequency) that appear and, from the count, identifies the main topics that the document covers e.g. if in the document, cricket word comes frequently then “cricket” is assigned as its topic (or class) [1] [2]. There are two phases in the Classification. They are Training Phase and Testing phase.

A. Training Phase: It is also called as Model Construction or Learning Phase; the set of documents used for model construction is called training set. It describes a set of predetermined classes. Each document/sample in the training set is assumed to belong to a predefined class (labelled documents). The model is represented as classification rules, decision trees, or mathematical formulae [2] [3].

B. Testing Phase: This is the 2nd step in classification and also called Mode Usage or Classification Phase. It is used for classifying future or unlabelled documents. The known label of test document/sample is compared with the classified result to estimate the accuracy of the classifier. For e.g. the labelled documents of the training set, is used further to classify unlabelled documents. Test set is independent of training set [1] [2] [3].

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 7, July 2016

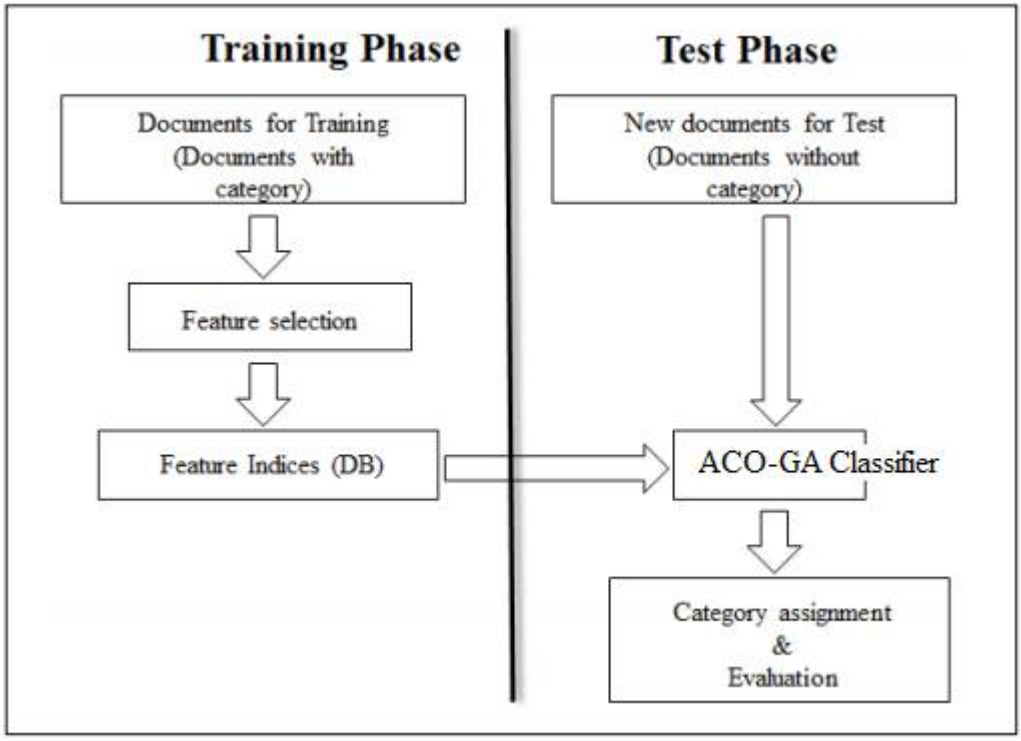

Fig 1 .Proposed System Architecture

In Fig. 1, 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. 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.

IV. RESULTS AND DISCUSSION

The experimental results illustrate the benefits of multi-view learning methods compared to traditional single-view learning, Below graph shows the system accuracy of proposed system. For the given execution we given 70-30 ratio for IEEE input papers. We got the classification accuracy on satisfactory level.

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 7, July 2016

Fig.2 System Accuracy

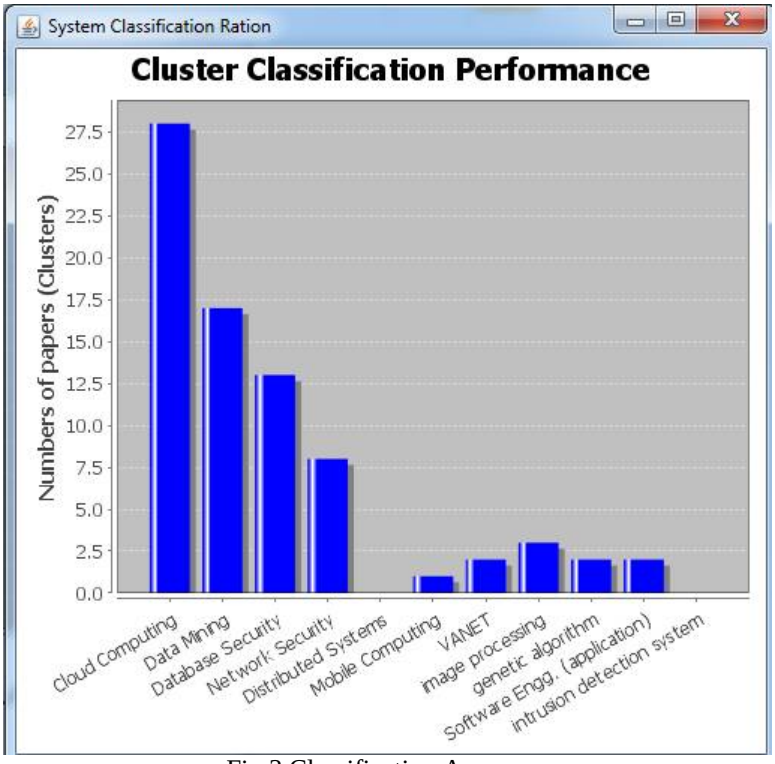

Fig.3 Classification Accuracy

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 7, July 2016

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 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.

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

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 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