A SURVEY ON COLLECTIVE CLASSIFICATION IN INFORMATION NETWORKS

Krishnapriya V
Computer Science and Engineering
N.S.S. College of Engineering
Palakkad

Anuraj Mohan
Assistant Professor
Computer Science and Engineering
N.S.S. College of Engineering
Palakkad

***Abstract*—Classification is a problem of machine learning and statistics which identifies the set of sub categories to which a new observation belongs. As the quantity and quality of data to be dealt with have achieved tremendous changes, classification has seized new levels of perception. Here comes the significance of different techniques for the classification of network data or more precisely, the graph data. Graph classification is a more general framework where focus of this survey lies in the classification of nodes of the network data. Since the data elements represented using graphs have relationships among them, during the classification the test labels have to be determined jointly. This is called collective classification. It acquires different forms based on the type of networks viz Homogeneous and Heterogeneous Networks. Various state-of-the-art approaches existing for classification on these networks, where the data chosen may be pre existing graphs for experiments or graphs constructed from raw data ,is studied and the compared through this survey. .**

keywords: Node classification,Semi supervised Learning,Homogeneous Networks,Heterogeneous Networks.

I. INTRODUCTION

Network data or graph data is a form of data representation in which the data objects along with their relationship to other data objects are mathematically represented and processed. The data objects are represented using nodes/vertices and their relationships using edges/links. Mining data, exploiting these structures is a hot area of research these days. The graph classification problem is important among them. It is the process of learning the graphs in the graph database to classify them into separate, individual categories with similar characteristics. A number of algorithms have been introduced for the graph classification problem[27]. It finds wide applications in social network analysis, text data network analysis, web mining, bioinformatics etc. The major approaches in graph classification are vertex classification and sub graph classification [28].

Sub-graph classification is the process of assigning with labels for the sub-graphs in a network whose labels are unknown. A more revolutionary method to classify networks is node classification. Node classification gives us insight to the classes to which a particular node in a graph belongs to. The applications are co-authorship prediction, citation analysis, protein-protein network analysis etc. Most traditional classification approaches assume that the training samples are drawn independently and identically from the unknown data, generating distribution. But to handle realistic situations where batches or sub-groups have internal correlations, dependencies are introduced among the data objects which is a major challenge in node classification. Most traditional method to achieve this is collective classification [29].

Collective classification can be defined as a classification method in which class labels for a group of linked instances are correlated and need to be predicted simultaneously[16]. Collective classification has a wide variety of real world applications, e.g. hyperlinked document classification, social networks analysis and collaboration networks analysis.

Node classification can be performed using the three modes of classification namely Supervised, Unsupervised and Semi Supervised methods[28]. Supervised learning is the data mining task of inferring a function from labeled training data. A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. In Data mining, the problem of unsupervised learning is that of trying to find hidden structure in unlabeled data. Semi Supervised learning(SSL) is a subfield of Supervised Learning [30] . Here, the major challenge faced by Supervised Learning, i.e., scarcity of data, is solved. Classification can be performed using fewer amount of labeled data. And hence it is very easy to solve real world problems using SSL. As a result SSL techniques are used widely in node classification[31],[32],[33],[34],[35].

Nodes of these graphs have labels. Large graphs are encountered in real world situations (e.g., dealing with social networks). A subset of the nodes of these networks may be labeled. These labels can indicate demographic values, interest, beliefs or other characteristics of the nodes (users). Among these nodes most of them may be unlabeled. Hence our aim is to use this information to extend the labeling so that all the nodes are assigned with a label (or labels). According to the number of labels available, classification can be of two types. They are: Single label collective classification It is the task of assigning a class label for each node present in an information network where the class labels belong to a set of disjoint labels L. The main characteristic of this method is that each of these nodes belongs to a single class and hence is called single label classification.

The complexity of different approaches varies based on the types of graph to be dealt with and the number of labels possible for each node. The types of graph being discussed in this survey are:(1)Homogeneous Graphs composed of similar kinds of nodes and links. Skype is an example for homogeneous networks where most of the value is derived from a single class of users all interested in placing a phone call.(2)Heterogeneous Graphs dissimilar kinds of objects and links, and this links may be found between any two vivid objects. We can find that most real world networks are heterogeneous in nature.

The remainder of this article is structured as follows. Section 2 defines the problem. Section 3 explains various techniques of node classification on Homogeneous networks. This section gives an insight to various local and global collective classification approaches. Section 4 explains various collective classification techniques on Heterogeneous graphs. Finally Section 5 describes challenges in graph classification.

II. PROBLEM DEFINITION

Given a network and a node V in the network there are three distinct types of correlations that can be utilized to determine the class of V. They are: a) Correlations between label of V and observed attributes of V. b) Correlations between label of V and observed attributes (including observed labels) of nodes in neighborhood of V. c) The correlations between the label of V and the unobserved labels of objects in neighborhood of V. Collective classification refers to the combined classification of a set of interlinked objects using all three types of information just described .

There are two kinds of collective classification: (a) Single Label collective classification and (b) Multi label collective classification. Problem of collective classification on the homogeneous and heterogeneous networks are defined as follows:

A. Single label collective classification on homogeneous networks

An information network, which is a directed graph G=(V,E) with object type mapping function ϕ:$V \rightarrow O$,$v \in V$ and link type mapping function ϕ:$E \rightarrow R$ where O is the object type set such that $\phi(v) = O, v \in V$ and R is the relation type set such that $\phi(e) = R, e \in E$ is said to be homogeneous if all the relation type, starting object type and ending object type are the same and if there is a set of disjoint labels L, such that $|L| > 1$,then for each of the vertices $v \in V$ of the homogeneous graph, inferring a single label $l \in L$ such that class $C(v) = l$ is termed as single label collective classification on Homogeneous Networks.

B. Multi label collective classification on homogeneous networks

An information network, which is a directed graph G=(V,E) with object type mapping function $\phi : V \rightarrow O, v \in V$ and link type mapping function $\phi : E \rightarrow R$ where O is the object type set such that $\phi(v) = O, v \in V$ and R is the relation type set such that $\phi(e) = R, e \in E$ is said to be homogeneous if all the relation type, starting object type and ending object type are the same. If there exist a set of labels L for this homogeneous network, such that each node has a vector of label variables $Y_i = (Y_{i1} .... Y_{iq}) \in (0, 1)^q$ and if the training set being $Y = Y_{ii=1}^n$ have a subset of nodes Y^L with known set of labels $Y^L = y_i|v_i \in V^L$ where $Y_i = (Y_{i1} .... Y_{iq}) \in (0, 1)^q$ then the task of inferring values $Y_i \in Y^U$ for remaining nodes in test set $V^U = V - V^L$ is termed as multi label collective classification on homogeneous networks.

C. Single Label Classification on heterogeneous Networks

An information network, which is a directed graph G=(V,E) with object type mapping function ϕ:$V \rightarrow O$,$v \in V$ and link type mapping function ϕ:$E \rightarrow R$ where O is the object type set such that $\phi(v) = O, v \in V$ and R is the relation type set such that $\phi(e) = R, e \in E$ is said to be heterogeneous information network if the types of objects $|A| > 1$ or the types of relations $|R| > 1$ and while defining a relation the relation type, starting object type and ending object type are dissimilar. If there is a set of disjoint labels L, such that $|L| > 1$, then for each of the vertices $v \in V$ of the homogeneous graph, inferring a single label $l \in L$ such that class $C(v) = l$ is termed as single label collective classification on Homogeneous Networks.

D. Multi Label Collective classification on Heterogeneous Networks

An information network, which is a directed graph G=(V,E) with object type mapping function ϕ:$V \rightarrow O$,$v \in V$ and link type mapping function ϕ:$E \rightarrow R$ where O is the object type set such that $\phi(v) = O, v \in V$ and R is the relation type set such that $\phi(e) = R, e \in E$ is said to be heterogeneous information network if the types of objects $|A| > 1$ or the types of relations $|R| > 1$ and while defining a relation the relation type, starting object type and ending object type are dissimilar. If there exist a set of labels L for this homogeneous network, such that each node has a vector of label variables $Y_i = (Y_{i1} .... Y_{iq}) \in (0, 1)^q$ and If the training set being $Y = Y_{ii=1}^n$ have a subset of nodes Y^L with known set of labels $Y^L = y_i|v_i \in V^L$ where $Y_i = (Y_{i1} .... Y_{iq}) \in (0, 1)^q$ then the task of inferring values $Y_i \in Y^U$ for remaining nodes in test set $V^U = V - V^L$ is termed as multi label collective classification on homogeneous networks.

III. NODE CLASSIFICATION ON HOMOGENEOUS NETWORKS NETWORKS

Node classification which is also called Graph Labeling is the method of assigning class labels to unlabeled nodes in a network. Node classification differs from the traditional classification tasks as it has to deal with the data objects which are not independent and identically distributed. So the classification of nodes in real world networks can only be performed jointly avoiding the concept of i.i.d. Collectively exploiting the different kinds of correlations in the nodes to perform node classification is called collective classification.

Collective approach considers both the node characteristics and network topology while assigning class labels for a node. The features of the test node as well as attributes and labels of the neighboring nodes are considered for this. There are two broad categories of approaches for the classification of nodes in the network (i) Within-network (ii)Across-network inference Within-network classification, for which training nodes are connected directly to other nodes, whose labels are to be classified, differs from across-network classification where models learnt from one network are applied to another similar network.

The collective classification on Homogeneous networks are classified as:

A. Single Label Approaches

Iterative algorithms[1] [2]are used when very poor information about the data object is available. Since the feature set of such data objects would be weak, the classification algorithm is first applied to get a richer set of putative information about the neighborhood, giving more insight to the label of the test node. Classification takes place iteratively until it converges or a specific number of iterations are reached. Label information is spread from the neighborhood nodes and hence acquires the form of semi supervised classification. After addition of features to the node using the neighborhood information, apply a local classifier like Nave bayes or decision tree to generate label of node by training on the newly labeled examples. Traditional classification techniques on the other hand, have a richer set of feature set right from the beginning of the classification which will not be altered. But these methods are quiet inefficient in sparsely labeled, large scale and multi dimensional networks[36].

A novel iterative approach is developed with the application of label-dependent features, is introduced in [3] to provide more accurate generalization for sparse datasets. Additional steps are needed to extract new input features based on graph structures, limited to the nodes of the given label. A separate set of structural features is provided for each label. Iterative approaches can be local if they use a local classifier and global if it uses a global optimization function.

Another approach which has gained importance in the area of collective classification is *Ensemble collective classifiers*. Ensemble methods are learning algorithms that construct a set of classifiers and then classify new data points by taking a weighted-vote of their predictions [4]. Ensemble methods perform accurately for i.i.d. data, but the research on these methods is comparatively less for relational data. The extension of i.i.d. ensembles are used to improve classification accuracy for relational domains[5]. A special case of these methods are *Hybrid Classifier methods*. These mainly help to deal with the sparsely labeled networks which was a challenge faced by the iterative algorithms. Novel combinations of classifiers are formed to exploit the different characteristics of the relational features vs. the non-relational features of the graph[6][7].

Apart from these there are other techniques that have been suggested which uses the graph structure directly to perform the labeling task. The labels are propagated from the nodes whose labels are known to those nodes with labels unknown, and observe that many of these methods perform random walks over the network to determine a global function for labeling. These methods are hence referred to as *Random Walk Based methods*. All these methods can be described using iterative matrix formulations but instead of using local classifiers, a global optimization function is used. Therefore these techniques are closely related to iterative approaches with simple classifiers[8],[9].

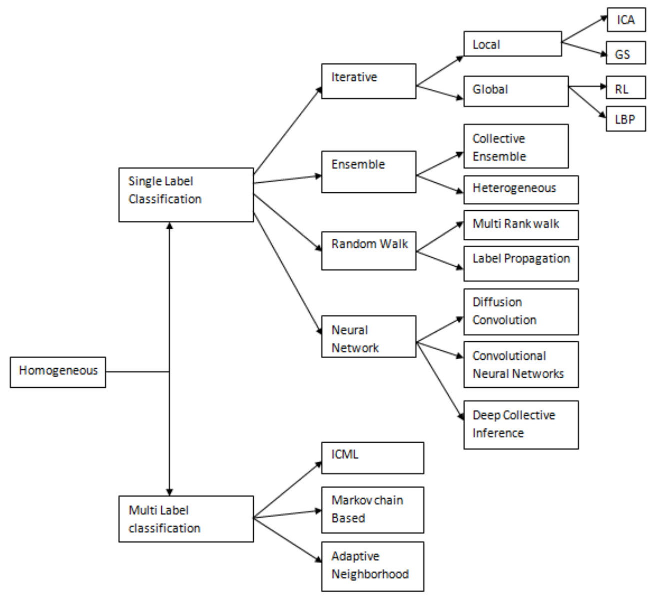

Fig. 1. Collective Classification Approaches in Homogeneous networks

The idea underlying the random walk methods is as follows: the probability of labeling a node $v_i \in V$ with label $c \in Y$ is the total probability that a random walk starting at v_i will end at a node labeled c. The various methods proposed in the literature differ in their definition of the random walk used for labeling. For this to provide a complete labeling, the graph G is often assumed to be label connected. That is, it is possible to reach a labeled node from any unlabeled node in finite number of steps.Label propagation is the most common random walk approach.

Graph based neural network models are yet another category of node classification technique. Recurrent Neural networks were first used in this framework which requires the repeated application of contraction maps as propagation functions until node representations stabilizes to a fixed point. Later Convolution like propagation rules were used to perform graph classification[11][12]. Graphs are converted locally into sequences and fed into a conventional 1D convolutional neural network, where the input nodes were ordered in the preprocessing step[10]. But the major drawback with this technique is the increasing memory requirement with the scaling size of the dataset. Also the edge features are not supported here and can be efficient only with undirected graphs.

B. Multi Label Approaches

All the approaches mentioned before were aimed at finding the class of nodes with single labels. But the real world data might require approaches which can classify the nodes which have multiple labels. Here various kinds of dependencies like intra node cross label dependencies, inter node single label dependencies, inter node cross label dependencies etc have to be dealt with. This in turn improves the complexity of the algorithm and is referred to as *Multi label Collective classification* models.

Collective classification is particularly challenging in multi-label settings. The reason is that, in the single-label settings, conventional collective classification methods can classify a group of related instances simultaneously by considering the dependencies among related instances for one label concept. But in multilabel settings, each instance can have multiple label concepts within its label set, and the dependencies among related instances with multiple labels are more complex *ICML* model can exploit three types of dependencies (1) Intra-instance cross-label dependencies,(2)Inter-instance single-label dependencies, (3)Inter-instance cross-label dependencies [12].

The rationale behind collective classification using *Adaptive neighborhood* stems from the fact that an entity in a network (or relational data) is most likely influenced by the neighboring entities, and can be classified accordingly, based on the class assignment of the neighbors[13].It is based on a neighborhood ranking method for multi-label classification.

In Markov chain based approach, the idea is to model the problem as a Markov Chain with restart on transition probability graphs, and to propagate the ranking score of labeled instances to unlabeled instances based on the affinity among instances. The affinity among instances is set up by explicitly using the attribute features derived from the content of instances as well as the correlation features constructed from the links of instances. Intuitively, an instance which contains linked neighbors that are highly similar to the other instances with a high rank of a particular class label, has a high chance of this class label [14].

IV. NODE CLASSIFICATION ON HETEROGENEOUS NETWORKS

Recently, more and more researchers begin to consider the interconnected, multi-typed data as heterogeneous information networks, and develop structural analysis approaches by leveraging the rich semantic meaning of structural types of objects and links in the network. The heterogeneous information network contains richer structure and semantic information, which provides plenty of opportunities as well as a lot of challenges for data mining. The meta path concept in [15], Heterogeneous Information Networks(HIN) have become a hot topic in HIN data mining.

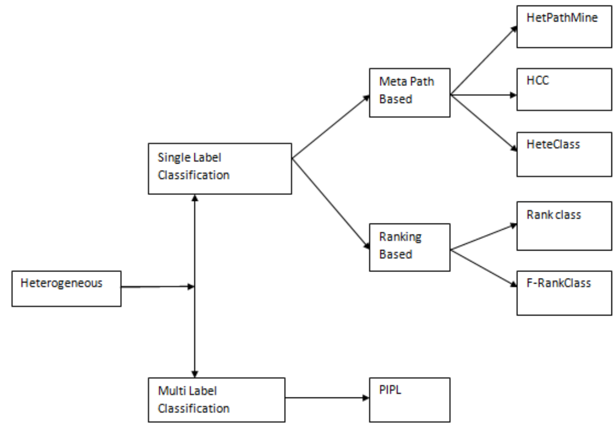

Fig. 2. Examples of meta paths in bibliographic data

A. Single Label Approaches

The collective classification on HIN can be classified into two categories: (1)Metapath based (2)Ranking Based.

A *Meta path* P is a path defined on a schema S = (A, R), and is denoted in the form of :

$$A_1 \xrightarrow{\text{R1}} A_2 \xrightarrow{\text{R2}} ... \xrightarrow{\text{Rl}} A_{l+1}$$

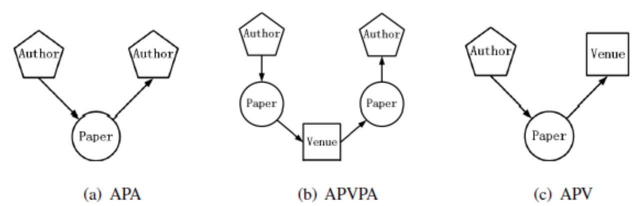

Fig. 3. Examples of meta paths in bibliographic data

which defines a composite relation R = R1 o R2 o...o Rl between objects A1, A2, ..., Al+1, where o denotes the composition operator on relations. As examples shown in Fig. authors can be connected via meta paths "Author-Paper-Author" (APA) path, "Author-Paper-Venue-Paper-Author" (APVPA) path, and so on.It is obvious that semantics underneath these paths are different which is an important characteristic of HIN. Based on different meta paths, objects have different connection relations with diverse path semantics, which may have an effect on many data mining tasks.

A novel solution, called *Hcc* (Meta-path based Heterogeneous Collective Classification)[16], is developed to effectively assign labels to a group of instances that are interconnected through different Meta-paths. The proposed Hcc model can capture different types of dependencies among objects with respect to different meta paths.Hcc can utilize the Meta-path based dependencies to exploit the heterogeneous network structure more effectively.

HetPathMine[17], a transductive algorithm introduced based on Meta-path principle is an accurate method where a novel Meta-path selection model is proposed to calculate the different weight of each relation paths. Finally, by using the different weight of each meta path, the transductive classification framework, is proposed. The weight obtained by HetPathMine for each Meta-path is consistent with human intuition or real-world situations.

Another Meta-path based framework, *HeteClass*, for transductive classification of target type objects explores the network schema of the given network and can also incorporate the knowledge of the domain expert to generate a set of Meta-paths[18]. The regularization based weight learning method proposed in HeteClass is effective to compute the weights of symmetric as well as asymmetric Meta-paths in the network, and the weights generated are consistent with the real-world understanding. Using the learned weights, a homogeneous information network is formed on target type objects by the weighted combination, and transductive classification is performed. The proposed framework HeteClass is flexible to utilize any suitable classification algorithm for transductive classification and can be applied on heterogeneous information networks with arbitrary network schema.

Apart from metapath based approaches single label classification can be performed on the network using a rank based collective classification method ,ie,*Rankclass*. It is beneficial to integrate classification and ranking in a simultaneous, mutually enhancing process to iteratively compute the ranking distribution of the objects within each class. At each iteration, according to the current ranking results, the graph structure used in the ranking algorithm is adjusted so that the subnetwork corresponding to the specific class is emphasized, while the rest of the network is weakened[19]. Integrating ranking with classification not only generates more accurate classes than the state-of-art classification methods on networked data, but also provides meaningful ranking of objects within each class, serving as a more informative view of the data than traditional classification.

Another heterogeneous information network mining algorithm, feature-enhanced RankClass (*F-RankClass*)is proposed which extends RankClass to a unified classification framework that can be applied to binary or multiclass classification of unimodal or multimodal data[20].The merits of the F-RankClass framework are: 1) F-RankClass does not require features extracted from multimodal data to have a fixed number of dimension. Only extra links between nodes representing features and data objects are added. 2) Inheriting benefits from RankClass, F-RankClass identifies irrelevant objects by ranking, and thus improves classification accuracy. 3) All information, including features and multimodal structures, are encoded in the heterogeneous information network without further information loss. 4) F-RankClass provides a unified framework for both binary and multiclass classification of multimodal and unimodal classification without further modification.

B. Multi Label Approaches

The key challenge of multi-label classification comes from the large space of all possible label sets, which is exponential to the number of candidate labels. Most previous work focuses on exploiting correlations among different labels to facilitate the learning process.

A method called *PIPL* is introduced which facilitate the multilabel classification process by mining the correlations among instances and labels from heterogeneous information networks[21].The algorithm assign a set of candidate labels to a group of related instances in heterogeneous information networks. Different from previous work, the proposed PIPL, can exploit various types of dependencies among both of instances and labels based upon different meta-paths in heterogeneous information networks. By explicitly exploiting these Meta-path based dependencies,PIPL method can effectively capture the diverse and complex relationships among instances and labels.

Very few works have been proposed in Multi Label collective classification on HIN .It is still a hot area of research.

V. CHALLENGES IN COLLECTIVE CLASSIFICATION

The major challenges in dealing with the collective classification is when the graphs are scaled to *Big networks*[22][24]. To deal with big graphs approaches of MapReduce are introduced [23][25].Distributed graph processing is still complex to be implemented.

Another problem with collective classification is Dynamic graphs[26]. The data in this era is updated with every nanoseconds. But when this is conceived in a graphic framework it is not easy to perform knowledge mining tasks like collective classification on such graphs.

VI. CONCLUSION

Collective Classification is a hot area of research in the field of Statistical Relational Learning. This survey explores various approaches for performing collective classification on two types of graphs namely homogeneous graphs and heterogeneous graphs. The collective classification is done exploiting the dependencies among the different data objects jointly. The challenge in collective classification is dealing with big graphs and dynamic graphs.

REFERENCES

[1] Kajdanowicz, T.,Kazienko,P. & Jancza,M., *Collective classification techniques: an experimental study.* In New Trends in Databases and Information Systems, pp. 99-108. Springer, Berlin, Heidelberg, 2013.

[2] McDowell, L. K.,Gupta,K. M.,& Aha, D. W., *Cautious collective classification.* Journal of Machine Learning Research 10, pp.2777-2836 ,2009

[3] Kazienko, P . & Kajdanowicz, T.,*Label-dependent node classification in the network.* Neurocomputing 75, pp. 199-209 ,2012.

[4] Eldardiry, H., & Neville, J. *An analysis of how ensembles of collective classifiers improve predictions in graphs.*In Proceedings of the 21st ACM international conference on Information and knowledge management, pp. 225-234, 2012.

[5] Eldardiry, H. & Neville, J. *Across-model collective ensemble classification.* In AAAI, San Franciscol, CA, 2011.

[6] Cataltepe, Z.,Sonmez, A.,Baglioglu, K., & Erzan, A., *Collective Classification Using Heterogeneous Classifiers.* In MLDM, vol. 6871, pp. 155-169, 2011.

[7] McDowell, L. K., & David W. A.,*Semi-supervised collective classification via hybrid label regularization.*In Proceedings of the 29th International Conference on International Conference on Machine Learning, pp. 1243-1250 , 2012.

[8] Lin, F., & William W. C. *Semi-supervised classification of network data using very few labels.* In Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on, pp. 192-199., 2010.

[9] Nandanwar, S. , & M. N. Murty. *Structural neighborhood based classification of nodes in a network.* In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1085-1094 , 2016.

[10] Atwood, J., & Don T. *Diffusion-convolutional neural networks.* In Advances in Neural Information Processing Systems, pp. 1993-2001, 2016.

[11] Moore, J. & Neville, J., *Deep Collective Inference*In Proc. of AAAI 0.5em minus 0.4empp. 2364-2372. ,2017.

[12] Kong, X., Xiaoxiao, S., & Philip S. Y., *Multi-label collective classification*, in Proc. of SIAM International Conference on Data Mining pp. 618-629 ,2011.

[13] Saha T.,Huzefa R. & Carlotta D. *Multi label collective classification using adaptive neighborhoods*,In 11th International conference on ICMLA pp. 427-432 , 2012

[14] Wu, Q., Michael K. N.,Yunming Y.,Xutao L.,Ruichao S. & Yan L., emphMulti Label collective classification via markov chain based learning method,Knowledge based Systems 63, pp.1-14,2014.

[15] Meng C. ,Reynold C., Silviu M.,Pierre S. & Wangda Z. *Discovering metapaths in large heterogeneous networks* In proc. of 24th international conference on world wide web, pp.754-764. ,2015.

[16] Kong, X. , Philip S. Y., Ying D., & David J. W. *Meta path-based collective classification in heterogeneous information networks.*In Proceedings of the 21st ACM international conference on Information and knowledge management, 1998. pp. 1567-1571, 2012.

[17] Luo, C., Renchu G., Zhe W., and Chenghua L. *HetPathMine: A Novel Transductive Classification Algorithm on Heterogeneous Information Networks.* In ECIR, pp. 210-221,2014.

[18] Gupta, M., Pradeep K., & Bharat B. *HeteClass: A Meta-path based framework for transductive classification of objects in heterogeneous information networks.* Expert Systems with Applications 68, pp. 106-122 ,2017

[19] Ji, M., Jiawei H., & Marina D. *Ranking-based classification of heterogeneous information networks.* In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, , pp. 1298-1306, 2011.

[20] Chen, S. D., Ying-Yu C., Jiawei H., & Pierre M. *A feature-enhanced ranking-based classifier for multimodal data and heterogeneous information networks.* In Data Mining (ICDM), 2013 IEEE 13th International Conference on, pp. 997-1002, 2013.

[21] Kong, X., Bokai C., & Philip S. Y. *Multi-label classification by mining label and instance correlations from heterogeneous information networks.* In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, , pp. 614-622, 2013

[22] Talukdar, P., & William C. *Scaling graph-based semi supervised learning to large number of labels using count-min sketch.* In Artificial Intelligence and Statistics, , pp. 940-947, 2014.

[23] Indyk, W., Tomasz K., Przemysaw K., & Sawomir P. *Mapreduce approach to collective classification for networks.* In International Conference on Artificial Intelligence and Soft Computing, , pp. 656-663, 2012.

[24] Ravi, S., & Qiming D. *Large scale distributed semi-supervised learning using streaming approximation.*In Artificial Intelligence and Statistics,, ,pp. 519-528.,2016.

[25] Kajdanowicz, T., Wojciech I., & Przemyslaw K. *MapReduce approach to relational influence propagation in complex networks.*Pattern Analysis and Applications 17, no. 4 pp. 739-746,2014

[26] Guo, T., Xingquan Z., Jian P., & Chengqi Z. *Snoc: streaming network node classification.* In Data Mining (ICDM), 2014 IEEE International Conference on, pp. 150-159, 2014.

[27] Ketkar, N. S., Lawrence B. H., & Diane J. C. *Empirical comparison of graph classification algorithms.*In Computational Intelligence and Data Mining,2009. CIDM'09. IEEE Symposium on, pp. pp. 259-266, 2009.

[28] Zuber, M. *A survey of data mining techniques for social network analysis.* International Journal of Research in Computer Engineering & Electronics 3, no. 6 pp.203-219 ,2014.

[29] Sen, P., Galileo N., Mustafa B., & Lise G. *Collective classification.* In Encyclopedia of Machine Learning, pp. 189-193. Springer US, 2011.

[30] Goldberg, A. B., and Xiaojin Z. *New directions in semi-supervised learning.* PhD diss., University of Wisconsin-Madison, ch. 1, pp. 2-4, 2010

[31] Kingma, D. P., Shakir M., Danilo J. R., & Max W. *Semi-supervised learning with deep generative models.* In Advances in Neural Information Processing Systems, pp.3581-3589, 2014.

[32] Reitmaier, T., Adrian C., & Bernhard S. *Transductive active learninga new semi-supervised learning approach based on iteratively refined generative models to capture structure in data.* Information Sciences 293 pp. pp. 275-298,2015

[33] He, Y., & Deyu Z. *Self-training from labeled features for sentiment analysis.*Information Processing & Management 47, no. 4 pp. 606-616,2011

[34] Gibson, B. R., Timothy T. R., & Xiaojin Z. *Human semisupervised learning.* Topics in cognitive science 5, no. 1 pp. 132-172,2013.

[35] Zhou, Zhi-Hua, and Ming Li. "Semi-supervised learning by disagreement." Knowledge and Information Systems 24, no. 3 , pp. 415-439.,2010

[36] Berlingerio, M., Michele C., Fosca G., Anna M., & Dino P.*Foundations of multidimensional network analysis.* In Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on, pp. 485-489, 2011.