Link Prediction-based Multi-label Classification on Networked Data Using Apriori Algorithm

Ashutosh Patel[1], Jitendra Agrawal[2], Sanjeev Sharma[3]
[1]student, SOIT RGPV, Bhopal, India;
[2,3]Assistant Professor, SOIT RGPV, Bhopal, India;

Abstract
During this work, we tend to have an interest to tackle the problem of link prediction in complicated net-works. Especially, we tend to explore topological two approaches for link prediction. Different topological proximity measures have been studied in the scientific literature for finding the probability of appearance of new links in a complex network. Link Prediction is a part of great interest in social network analysis. Previous works within the area of link prediction have only targeted on networks wherever the links once created can't be removed. The rapid growth of social networks shows the increasing quality of those services among the users. The expansion of social networks happens as results of adding new users and new links between users. Link prediction has many applications and, it offers many benefits to the users of social networking services such as providing fast and accurate recommendations or suggestions to the users. There are various tries to handle the problem of link prediction through various approaches. Commonest means is to measure the closeness /similarity of nodes to every different in terms of various social aspects

***Keywords**:* networked data; multi-label classification; link prediction

1. Introduction

The past decade has witnessed a speedy development and change of the net and internet. The advancement in computing and communication technologies is drawing individuals along in innovative ways in which. Prodigious numbers of on-line volunteers collaboratively write cyclopedia articles of previously not possible scopes and scales; on-line marketplaces suggest product by work user looking behavior and interactions; Political movements are making new sorts of engagement and collective action. Varied participatory internet and social networking sites are cropping up, empowering new sorts of collaboration, communication and emerging intelligence.

Classification may be a well-known task in data processing that aims to predict the category of an unseen instance as accurately as possible. Whereas single label classification, that assigns every rule in the classifier the most obvious label, has been wide studied very little work has been done on multi-label classification. Most of the work up to now on multi-label classification is said to text categorization. There are several approaches for building single category classifiers from information, like divide-and-conquer and separate-and-conquer. Most traditional learning techniques derived from these approaches, such as decision trees, and statistical and covering algorithms, are unable to treat problems with multiple labels.

The most common multi-label classification approach is one-versus-the rest (OvR), which constructs a set of binary classifiers obtained by training on each possible class versus all the rest. OvR approach performs the winner-take-all strategy that assigns a real value for each class to indicate the class membership.

1.1. Social networks

A social network may be a structure consists of entities which might be people, teams or organizations, and also the relations or associations among them. With the emergence of the net, the net social networks are gained increasing quality. On-line social networks has become one in all the most influential and key supply of service providing, information/knowledge sharing and lots of other web primarily based activities. Social networks are composed of users (nodes) and associations (edges) among them. The users are often people, groups, organizations, etc. Users join a social network, publish their own content, profile and create links to other users in the network by making “friendships”. The meaning of a “friendship” depends on the network. It are often a standard relationship, scientific collaboration, account, etc. the expansion of social networks happens as a results of adding new users and adding new links. Social networks serve a spread of advantages to its users:

Support for organizing & sharing contents to form friendships: Most social networking services give platforms for users to make share and organize their own profiles. These services has become very well-liked because of availability of user oriented, increased strategies to act with different users. Social networking sites such as Facebook (over 1 billion users), Twitters (over 200 million users), are examples of wildly popular networks used to share and organize the contents, finding friends. Social networks such as Flickr, YouTube {, are examples for social networks for sharing multimedia content such as photos, videos. Support for sharing knowledge, learning & collaboration: Social networks enhance informal learning and support social connections within users or organizations for sharing their profiles for educational and business functions. The users of such service will notice the appropriate candidates WHO match the non-public or structure interests. LinkedIn, a social network created up over 200 million professionals is an example for educational similarly as business oriented social networks.

Support for communication: E-mail networks are an example of communication social networks. The new e-mail system has incorporated state-of-art communication technology like dialing, chatting, video conferencing so as to empower the users. Its allowable advanced heterogeneous social connections between users. Modern multi-relational, heterogeneous social networks are analyzed exploitation completely different approaches like graph theory, graph mining. Social Network Analysis (SNA) is that the study of relations between people additionally because the analysis of social structures, social position, role analysis, and much of others. Normally, the link between people, e.g., kinship, friends, neighbors, etc. are given as a network. Traditional science involves the circulation of questionnaires, asking respondents to detail their interaction with others. Then a network is also created supported the response, with nodes representing people and edges the interaction between them. This type of knowledge assortment range traditional SNA to a restricted scale, usually at the foremost several actors in one study.

2. Link prediction

Link prediction is the most fundamental problem that attempts to infer which new links are likely to occur in the near future based on the topological, node and edge properties during a given network. That is, if we tend to are given with a snap of a network at this time, the goal is predicting links that may occur within the next time step. As a part of the recent surge of analysis on large, complicated social networks and their properties, a substantial quantity of attention has been dedicated to the process analysis of social network evolution. In social networks nodes represent people or alternative entities embedded during a social context, and whose edges represent interaction, collaboration, or influence between entities. Link prediction drawback has taken and outlined in some ways. We tend to discuss few of them shortly. In data processing perspective, link prediction drawback as a link mining task as a result of several real-world networks composed of kind of entity sorts joined via multiple kinds of relations. A rising challenge for link mining is that the drawback of mining richly joined datasets to explore the data behind the links or relationships. This information provides extra advantage which will be useful for several data processing tasks. However multi-relational information violates the standard assumption of independent, identically distributed information instances that gives the idea for several applied math machine learning algorithms. Therefore, new approaches are required which will exploit the dependencies across the attribute and link structure [27]. Link prediction can be divided in to two cases:

(1) Predicting entirely new links which means those links are never appeared in the network. New links emerge in between existing nodes as well as by adding new nodes. Predicting links added by latter case is extremely hard problem. Thus, most of the research has been attempting to find methods to predict links among the existing nodes.

(2) Predicting repeating links, that is, some links are not visible in the network during the observed period of time but they appeared either before or after the observed prod of time. However, if time is a part of the predictive model, then repeating link prediction refer to the same task which is to predict the evolution of a network in terms of new edges that will be added in the future. According to the probabilistic perspective, Link prediction is an estimate of the likelihood or probability of the future occurrence of a link in a network or estimating the probability of whole network taking a particular form by adding set of new links. In both cases the complex dependencies among the links are required to address using probabilistic and statistical models.

3. Proposed Methodology

Our projected algorithmic rule consists of 3 phases: rules generation, algorithmic learning and classification. Within the initial part, it scans the training information to get and generate an entire automotive. within the second part, MMAC issue to get a lot of rules that pass the MinSupp and MinConf thresholds from the remaining unclassified instances, till no more frequent things are often found. Within the third part, the principles sets derived at every iteration are incorporated to make a worldwide multi-class label classifier which will then test against test information. Figure 4.1 represents a general description of our planned technique that we'll justify in additional detail below. Training attributes may be categorical, i.e. attributes with restricted distinct values, or continuous, i.e., real and number attributes. For categorical attributes, we tend to assume that every one possible value is mapped to a collection of positive integers. At the current time, our technique doesn't treat continuous attributes.

To increase the efficiency of frequent things discovery and rules generation, MMAC employs a new technique supported an intersection technique that has been conferred. We've got extended their technique to accomplish classification. Our technique scans the training information once to count the occurrences of single things, from that it determines people who pass MinSupp and MinConf thresholds, and stores them alongside their occurrences (rowIds) within fast access information structures. Then, by across the rowIds of the frequent single things discovered up to now, we will simply get the possible remaining frequent things that involve over one attribute. The rowIds for frequent single things are helpful data, and may be wont to find things simply within the training information so as to get support and confidence values for rules involving over one item. To clarify the picture, consider for instance frequent single items A and B, if we intersect the rowIds sets of A and B, then the resulting set should represent the tuples where A and B happen to be together in the training data, and therefore the classes associated with A^B can be easily located, in which the support and confidence is accessed and calculated, that they'll be accustomed decide whether or not or not A^B could be a frequent item and a candidate rule out the classifier. Since the training data are scanned once to find and generate the principles, this approach is very effective in runtime and storage as a result of it doesn't rely on the traditional approach of discovering frequent items, which requires multiple scans.

Once an item has been known as a frequent item, MMAC checks whether or not or not it passes the MinConf threshold. If the item confidence is larger than MinConf, then it'll be generated as a candidate rule out the classifier. Otherwise, the items are discarded. Thus, all things that survive MinConf are generated as candidate rules within the classifier.

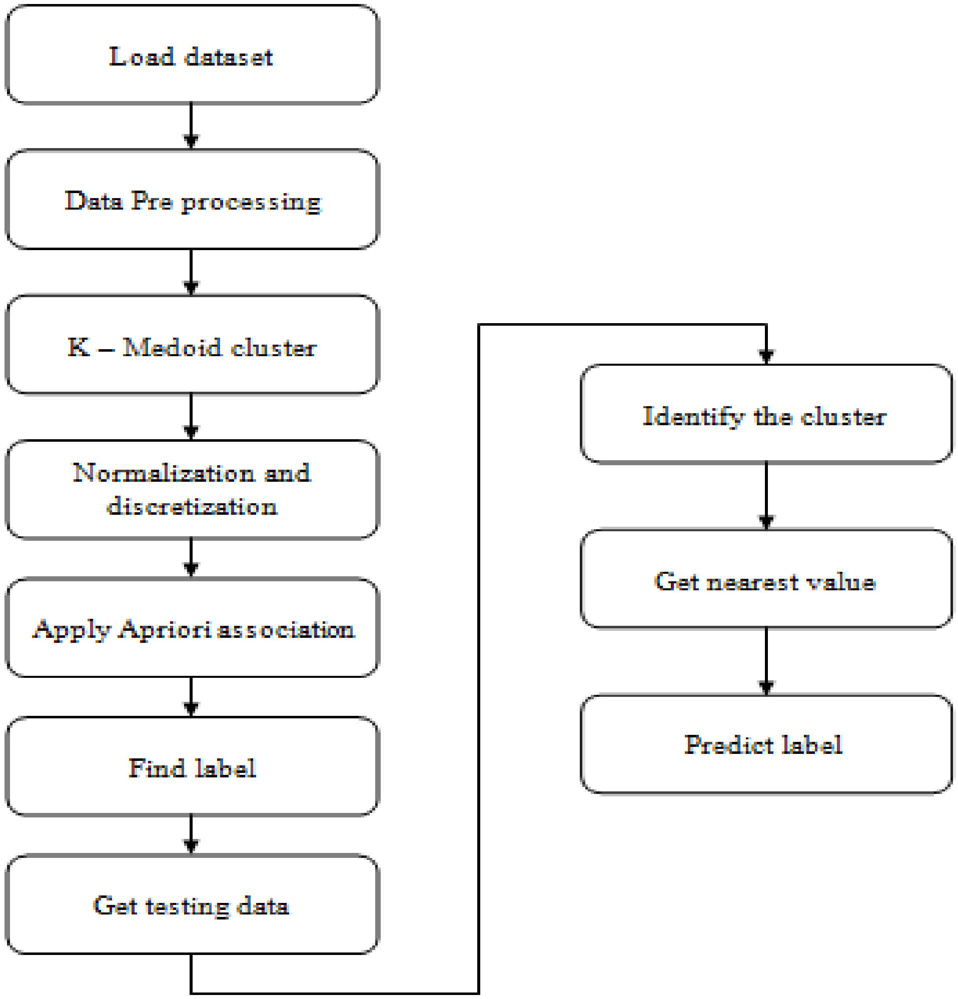

Fig.3.1 flow diagram of proposed system

3.1 Advantages

- Our approach against 19 different datasets from as well as a different datasets for forecasting the behavior of an optimization heuristic within a hyper heuristic framework
- The choice of such learning methods is based on the different strategies they use to generate the rules.

3.2 Algorithm

```
Step 1: Input: Adjacency matrix
Step 2: MaxIndex=70;    /* number of the regions in connectivity matrix*/
Step 3: TopRank=-1;
Step 4: for i=1 to MaxIndex-1 do
Step 5:        for j=i+1 to MaxIndex do
Step 6:          if Matrix (i,j)==0 then
Step 7:              TempRank=Score Function Matrix (i,j);
Step 8:                  if TempRank>TopRank then
Step 9:                          x=i;
Step 10:                          y=j;
Step 11:                         TopRank=TempRank;
Step 12:              end if
Step 13:          end if
Step 14:      end for
Step 15: end for
```

4. Result:

Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in massive databases. it's meant to identify robust rules discovered in databases exploitation some measures of interestingness.[1] based on the conception of robust rules, Rakesh Agrawal, Tomasz Imielinski and Arun swami [2] introduced association rules for locating regularities between product in large-scale transaction data recorded by point-of-sale (POS) systems in supermarkets.

Table 1: Performance of link prediction

	Dataset	**SRW**	**RWR**
Base Paper	**IMDB**	0.9988	0.9940
	DBLP	0.9607	0.9826
Proposed	**Dataset**	**k-medoids**	**Apriori**
	MYD	0.8988	0.8740
	MYI	0.8918	0.8856

5. Conclusion and Future Work:

A new approach for multi-class, and multi-label classification has been proposed that has many distinguishing features over traditional and associative classification methods in that it (1) produces classifiers that contain rules with multiple labels, (2) presents three evaluation measures for evaluating accuracy rate, (3) employs a new method of discovering the rules that require only one scan over the training data, (4) introduces a ranking technique which prunes redundant rules, and ensures only high effective ones are used for classification, and (5) integrates frequent items set discovery and rules generation in one phase to conserve less storage and runtime. Performance studies on 19 datasets from Weka data collection and 9 hyperheuristic scheduling runs indicated that our proposed approach is effective, consistent and has a higher classification rate than the-state-of-the-art decision tree rule (PART), CBA and RIPPER algorithms.

References:

[1] Zhao, Yinfeng, Lei Li, and Xindong Wu. "Link Prediction-Based Multi-label Classification on Networked Data." Data Science in Cyberspace (DSC), IEEE International Conference on. IEEE, 2016.

[2] Bilgic, Mustafa, Galileo Mark Namata, and Lise Getoor. "Combining collective classification and link prediction." Data Mining Workshops, 2007. ICDM Workshops 2007. Seventh IEEE International Conference on. IEEE, 2007.

[3] Chen, Zheng, et al. "Marginalized Denoising for Link Prediction and Multi-Label Learning." AAAI. 2015.

[4] Al Hasan, Mohammad, and Mohammed J. Zaki. "A survey of link prediction in social networks." Social network data analytics. Springer US, 2011. 243-275.

[5] Kong, Xiangnan, Bokai Cao, and Philip S. Yu. "Multi-label classification by mining label and instance correlations from heterogeneous information networks." Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2013.

[6] Vijayan, Priyesh, Shivashankar Subramanian, and Balaraman Ravindran. "Multi-label collective classification in multi-attribute multi-relational network data." Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on. IEEE, 2014.

[7] Shi, Chuan, et al. "Multi-label classification based on multi-objective optimization." ACM Transactions on Intelligent Systems and Technology (TIST) 5.2 (2014): 35.

[8] Wang, Xi, and Gita Sukthankar. "Link prediction in heterogeneous collaboration networks." Social network analysis-community detection and evolution. Springer International Publishing, 2014. 165-192.

[9] Afrati, Foto, Aristides Gionis, and Heikki Mannila. "Approximating a collection of frequent sets." Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004.

[10]Rexeena, X., B. Suganya Devi, and S. Saranya. "Risk Assessment for Diabetes Mellitus using Association Rule Mining." Age 30.30to (2014): 50.

[11]Wang, Chao, and Srinivasan Parthasarathy. "Summarizing itemset patterns using probabilistic models." Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006.

[12]Xin, Dong, et al. "Extracting redundancy-aware top-k patterns." Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006.

[13] R. Agrawal, T. Amielinski and A. Swami. Mining association rule between sets of items in large databases. In Proceeding of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, DC, May 26-28 1993, pp. 207-216.

[14] R. Agrawal and R. Srikant. Fast algorithms for mining association rule. In Proceeding of the 20th International Conference on Very Large Data Bases, 1994, pp. 487 – 499.

[15] M. Boutell, X. Shen, J. Luo and C. Brown. Multi-label semantic scene classification. Technical report 813, Department of Computer Science,