Survey of active learning for networked data

Xinlei Wang[1, a], Bo Yang[2] and Jing Huang[3]

[1]The College of Computer Science and Technology, Jilin University, Changchun Jilin, 130012, China

[2]The College of Computer Science and Technology, Key laboratory of Symbolic Computation and knowledge Engineering, Jilin University, Changchun Jilin, 130012, China

[3]The College of Computer Science and Technology, Jilin University, Changchun Jilin, 130012, China

[a]wxl162531@163.com

Keywords: Active Learning, Networked Data, Classification, Machine Learning

Abstract. Classification in networked data is a popular research of complex network. Because of the large scale of networked data and the shortage of training data, active learning, which is an effective classification method for sparse data in machine learning, is often applied to networked data classification problems. Introduced in this paper are classification methods based on active learning for networked data classification problems with basic concepts and algorithms. Finally, according to the existing research, some problems for the future developing and research of networked data classification issues is presented.

Introduction

Supervised learning is a process to study the learning model based on labeled data set as training set, which is widely use in classification problem. But unlabeled data in the real world tends to be huge and labeling is more expensive and time-consuming. Sometimes labeling can be very difficult, while the unlabeled data set is relatively easier to get. For instance, for large-scale text categorization, the labeling for all text one by one is difficult. On the contrary, it is easier to select some unlabeled text. In order to solve this problem, active learning is proposed. When initial labeled data is sparse with quite rich unlabeled data, and the manual labeling of the data is very expensive, we can offer some instances which are filtered by active learning model to be labeled by specialists. Then these instances would be put into the training set to get new learning model iteratively. In general, the purpose of active learning is to use the least amount of labeled data and obtain relatively good classification effect in limited time and resources.

Because the supervise learning of networked data has obtained some success, some researches extend to the field of active learning. Compared with traditional passive learning, active learning can more effectively reduce sample complexity and has become a research hot topic in the field of complex network, machine learning and data mining. With the increase of attention, it is considered that effective active learning algorithm should not only focus on instance itself, network structure should also be taken into account. Due to networked data are often large and disorderly with sparse labeling, it is difficult to label on this data. Such as social networks are based on specific attributes to be classified. So active learning just can solve this problem. But the networked data is abstracted into network structure composed of nodes and the relationships between nodes. The instances of data set are the nodes in the network and the network edges are the relationships between instance nodes. In the past active learning research, the relationships between instance nodes of data set have not been considered. So the application of active learning in networked data classification problems has the increased difficulty. Besides, it becomes another important research of active learning for application.

For such problems, we usually give definitions as follows: networked data classification problems with active learning can be represented and modeled by triple form *(U,L,G)*.We define data as a network $G=(V,E)$, $v_i \in V$ $(1 \le i \le l+n)$ is one instance node of networked data, which

has $l+n$ instance nodes ($l \ll n$) including n unlabeled data instance nodes $U=\{x_1,x_2,\ldots,x_n\}$ and l labeled data instance nodes $L=\{(x_{n+1},y_{n+1}),\ldots,(x_{l+n},y_{l+n})\}$. $x_i(1\le i\le l+n)$ represents attribute vector of i-node and its correlative labeling class is y_i. As the edge of network, $e_{ij}\in E$（$1\le i$，$j\le l+n$）shows relationship between node v_i and node v_j. According to the existing query method Q, instances in unlabeled data set U are selected to be labeled and join to the L. A classification model would be constructed by cyclic iterative learning with new training set L.

Researches of networked data classification problem with active learning can be considered from multiple perspectives as follow: (1) According to the query processing mode, it can be divided into single mode and batch mode. (2) But in the earlier researches, whether it is a single mode or a batch mode, the number and type of class are known in advance, which is the static setting of data classes. So the static setting or dynamic change of classes can also be a new point of networked data classification.

Active Learning for Networked Data: Algorithms and Applications

For active learning, how to select the best sample (e.g. instance node of networked data) for query has been a popular topic. At present, there are parts of researches about this problem. First of all, we introduce some single mode query processing modes as follow.

Single mode of active learning for query in static setting: According to pool-based active learning method, Bilgic et al. [3] combine edge information and node information together in networked data. For each node to calculated utility scores. Three main tasks in active learning can be completed: utility score calculation, the choice of sample pool and the choice of the sample. Then complete the classification of networked data. But this algorithm is only based on the original pool-based active learning with edge information for a simple extension. So the algorithm improvement is relatively small.

Zhu et al. [4] put forward a classification method based on harmonic functions and Gaussian random fields. They also combine with supervised learning and active learning to construct learning model. This algorithm use the spread process of labeling behavior in the network. Through empirical risk minimization, the process of binary classification of networked data would be done. But the calculation is large.

Then, based on the idea of [4], Macskassy et al. [5] think that the links or edges in the network can make prediction accuracy maximization. Besides, the problem of accelerating the active learning algorithm in networked data classification is also thought up. They use graph-based centrality measures from social network analysis to replace empirical risk minimization method, which greatly improve the speed and performance of the algorithm. It can be found that the improved sample selection method will greatly speed up the running time of algorithm. So, the acceleration of active learning still has a lot of development space to grow.

Yan et al. [6] complete the sample selection through two methods: one is maximizing mutual information between two node attributes, which is often used to the active learning. The other one is based on the conditional Gibbs distribution to maximize the average agreement between two independent instance nodes. And they both eventually got a better classification effect. But for networked assortativity or disassortativity, they did not give a clear assuming. Based on the leftover problem of [6], Yan's research team discuss it in [7]. They put forward information-theoretic technique, as a new method of active learning, information-theoretic technique is proposed, which can be well to predict and identify assortativity or disassortativity network community. But the above two algorithms are not really applied to large-scale real networks for validation, which are not suitable for sparse networks and heterogeneous degree distributions. So the range of application is narrow.

Bilgic et al. [8] mainly focus on graph criterion to define data information and chose some instance nodes with high information. Train the classifier by combining edge information and instance node attributes in networked data. Select instance nodes by multiple inquire strategy and

put forward active learning for networked data(ALNET) to construct initial labeled set with clustering technology. For each iteration, there are three classifiers: uncertainty sampling, committee-based sampling, and clustering. Then we can calculate a disagree score for each instance node to choose the highest score instance node for labeling. ALNET, however, is based on a restricted assumption: the initial instance nodes of training set are from all possible classes. So, this algorithm is lack of ability to build and find a new class.

Cesa-Bianchi et al. [9] simplified the network data classification problem for tree learning problems. The two target of active learning are as follow: one is to minimize the number of mistakes on the non-queried vertices under a certain query budget, the other one is to minimize the sum of queries and mistakes under no restriction on the number of queries. But whether such an algorithm can be applied to all graph problems has not yet received verification.

In addition to the above some theoretical research, the application of active learning for networked data classification is gradually applied in many specific areas in recent years.

Seliya et al. [10] propose neural-network-based active learning algorithm, which is applied to computer network intrusion detection. Instance nodes can be divided into two classes with normal or attack. Because the networked data is very large and sample selection of active learning algorithm can reduce the size of the training set, which will not weaken the accuracy of the classification. Besides, compared with C4.5 decision tree algorithm, it can obtain better accuracy. You can see that active learning in computer network intrusion detection still can be improved from time and space.

Since Microblogging data is huge with a large amount of unlabeled data, Hu et al. [11] propose Active learning framework for the classification of Networked Texts (ActNeT) for Microblogging data. This algorithm select representative and informative instance nodes and use a regularization to ridge regression to represent the link relations. However, we can still improve the criterion of the instance nodes selection. And more hidden information can be considered (e.g. geographical information, temporal information).

Batch mode of active learning for query in static setting: But some methods mentioned above [3 -11] are computationally intensive algorithms. Each query will only pick out one instance node to label and add it to training set to learn a new classifier. The whole process is an iterative learning process with high time complexity. Gradually people find that if only one instance node is selected for the query, it is not only a waste of time and resources, but also can produce redundancy. According to previous researches, it can be found that batch mode active learning can save resources comparing to single mode active learning. So people start from the original single mode research to batch mode research. SHI et al. [1] propose Batch Mode Active Learning (BMAL) for Networked Data. Because it is networked Data, inevitably, it needs to consider instance node information and edge information simultaneously. They use a random walk model to give a similarity matrix between node and node and represent the similarity between instance nodes by using radial basis function (RBF). Criteria are given to represent instance node information, especially on the definition of the redundancy, which make the selected labeled instance nodes representative, uniform distribution, and not too concentrated. Selected instance nodes should be close to the classified boundary, which are of large uncertainty. According to the definition of instance node information, the larger information value, the more representative instance node. If select a series of such instance nodes, it is easy to find the boundary of the classification to classify. According to the above situation, give three criteria called maximum uncertainty, maximum impact, and minimum redundancy to define the objective function. Then optimize the objective function. But the similar researches of batch mode active learning are still scarce. Future research can be applied to more different network for experimental verification and improve the application range of algorithm. Then the definition of redundancy, the definition of node information, instance node selection criteria and other issues still have very large development space.

Dynamic changes of active learning for networked data: it's not hard to find that, in the above researches, whether it is a single mode or a batch mode, it belongs to static setting of class. So people are not satisfied with only considering the problem of instance node selection and begin to

consider the dynamic change of class. The typical researches of dynamic change methods mainly include [2], and research from the perspective of network structure dynamic change [12]. For the problems existing in [8], Fang et al. [2] propose a new method, ADLNET (active class discovery and learning for networked data), which defined the entire class distribution by Dirichlet process to find new classes actively. Then explicitly model active learning general function. In previous researches, they often give all possible classes. While in real world, in the beginning, we may not find all classes, and gradually find the new class in the learning process. Based on Dirichlet process, this algorithm define the class distribution and find the new class in the whole process, which does not belong to the existing set of classes. ADLNET still has a problem in improving the effectiveness of the algorithm in large-scale networked data. Besides, networked dynamics can also be adding to the algorithm to improve ADLNET.

Then for streaming networked data, Yang et al. [12] propose a novel active learning algorithm. Taking into consideration the dynamic changes of data distribution and data structure, with the help of Markov random field, they put forward new instance node selection method, which obtain good classification accuracy on the four data set. Besides, the social network and incorporate social factors can be taken into account to improve the definition of the problem.

Conclusion

Active learning aim at a large number of unlabeled data, query and label the instance nodes with highest information, which helps classification or clustering of unlabeled data set. So it has wider application range, such as in document information classification, etc. At the beginning of the research, active learning select instance nodes by using a single model, which may lead to selected instance nodes with approximate function, more concentrated and large redundancy. Besides, single model may waste resources by querying to label selected instance node each time. So recently researches on the batch mode have been started to reduce these phenomena. Previous researches do not consider edge information mostly, that is to say, they only consider the instance node information to describe the characteristics of the network. Then later, people begin to attach importance to network edge. Combine instance node information with edge information together for simulation network structure. The results of active learning under this condition, comparing the results only considering instance node information condition, is of improved accuracy. So active learning can develop a lot of directions, which includes the classification of networked data. In addition, active learning research problem of networked data has been extended to the dynamic changes of network structure, data sets, and the set of classes.

In most cases, active learning will transform research problems and frameworks into formulas with the same meaning. Then problem optimization will naturally become formula optimization problem. So that the problem definition will be formalized, which more facilitates analysis. Therefore, the definition of the problem becomes crucial.

Open questions and future outlook

Now the research of active learning of networked data has very large development space. It can be found that previous researches need to be improved as follow: (1) how to define redundancy and what method is used to reduce the redundancy of candidate sets. Then (2) for networked data, instance node information and edge information should be considered simultaneously. Previous researches simply represent edge information with degree of network, whether we can define the edge information from other levels. Besides (3) if we can independently label instance nodes in training set without querying experts.

According to the current development situation and combined with the existing researches, we can consider how to improve the effectiveness of algorithm in large-scale networked data. And, different with simple node distribution of data, there is a certain relationship between instance nodes in networked data. Although a lot of literature made their definition on network structure, it is not perfect enough. So for how to define network structure to capture more labeled correlation is an

important direction for future development. And applying the existing active learning method to the edge prediction of networked data is also an interesting topic. In addition, there are researches mostly pool-based active learning method, and few are stream-based active learning. So we need to consider if the nodes and edges in a network will make dynamic change over time, and what kind of effect would be haven in the application of active learning. Besides, for the question of networked data classification, combining of batch mode and dynamic change problem together is also a large challenge for the further research.

References

[1] Shi L, Zhao Y, Tang J. Batch mode active learning for networked data [J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2012, 3(2): 33.

[2] Bilgic M, Mihalkova L, Getoor L. Active learning for networked data[C]//Proceedings of the 27th international conference on machine learning (ICML-10). 2010: 79-86.

[3] Bilgic M, Getoor L. Link-based active learning[C]//NIPS Workshop on Analyzing Networks and Learning with Graphs. 2009.

[4] Zhu X, Lafferty J, Ghahramani Z. Combining active learning and semi-supervised learning using gaussian fields and harmonic functions[C]//ICML 2003 workshop on the continuum from labeled to unlabeled data in machine learning and data mining. 2003: 58-65.

[5] Macskassy S A. Using graph-based metrics with empirical risk minimization to speed up active learning on networked data[C]//Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009: 597-606.

[6] Yan X, Zhu Y, Rouquier J B, et al. Active Learning for Hidden Attributes in Networks[J]. arXiv preprint arXiv:1005.0794, 2010.

[7] Moore C, Yan X, Zhu Y, et al. Active learning for node classification in assortative and disassortative networks[C]//Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2011: 841-849.

[8] Bilgic M, Mihalkova L, Getoor L. Active learning for networked data[C]//Proceedings of the 27th international conference on machine learning (ICML-10). 2010: 79-86.

[9] Cesa-Bianchi N, Gentile C, Vitale F, et al. Active learning on trees and graphs[J]. arXiv preprint arXiv:1301.5112, 2013.

[10] Seliya N, Khoshgoftaar T M. Active learning with neural networks for intrusion detection[C]//Information Reuse and Integration (IRI), 2010 IEEE International Conference on. IEEE, 2010: 49-54.

[11] Hu X, Tang J, Gao H, et al. ActNeT: Active Learning for Networked Texts in Microblogging[C]//SDM. 2013: 306-314.

[12] Yang Z, Tang J, Zhang Y. Active Learning for Streaming Networked Data[C]//Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. ACM, 2014: 1129-1138.