Final Report (type I)

Name: Guanxiong Wang
ID: gwang43@hawk.iit.edu

1. Introduction

In traditional machine learning, the observed and unobserved samples are assumed to be drawn independently from an identical distribution. Classification problems are solved using only instances' content and labels. Relations between instances are not taken into consideration. However, in addition to content, connectivity information is often available and can be an important factor in determining the node labels. Link-based classification takes into consideration the links between the instances in order to improve the estimation performance.

Iterative classification algorithm (ICA) is a successful approximate inference algorithm used for collective classification. Local approaches (iterative classification) learn a model locally without considering unlabeled data and apply the mode iteratively to classify unlabeled data. Global approached (such as pairwise Markov Random Fields), on the other hand, exploit unlabeled data and the links occurring between labeled and unlabeled data for learning.

2. Approach

To determine the label of the test data, Iterative Classification Algorithm (ICA) assumes that all of the neighbors' attributes and labels of the instance are already known. Then, it calculates the most likely label with a local classifier that uses instance content and neighbors' labels. Since that all instances do not have equal number of neighbors, it is hard to implement the local classifier which should take constant number of inputs. To solve the problem, we introduce an aggregation operator such as Count, Proportion, Exist, and Mode. For example, count aggregation operator returns the number of occurrences of each label among the neighbors. Finally, ICA repeats the process iteratively until all of the label assignments become stabilized.

3. Datasets

3.1 Cora

The Cora dataset consists of 2708 Machine Learning papers classified into one of seven classes such as Case_based, Genetic_Algorithms, Neural_Networks and Probabilistic_Methods. As features, the words that occur at least 10 times are used. For each paper, whether or not it contains a specific word, which class it belongs to, which papers it cites and which papers it is cited by are known. The .content file contains descriptions of the papers and the .cites file contains the citation graph of the corpus.

3.2 Citeseer

The Citeseer dataset consists of 3312 scientific publications classified into one of six classes such as AI, DB, IR, and HCI. As features, the words that occur at least 10 times are used. For each paper, whether or not it contains a specific word, which class it belongs to, which papers it cites and which papers it is cited by are known. The .content file contains descriptions of the papers and the .cites file contains the citation graph of the corpus.

3.3 PubMed Diabetes

The PubMed Diabetes dataset consists of 19717 scientific publications classified into one of three classes. Compared with the Cora and Citeseer dataset, each publication in the dataset is described by a TF/IDF weighted word vector from a dictionary, which consists of 500 unique words. The .paper.tab file contains descriptions of the papers and the .cites.tab file contains the citation graph of the corpus.

3.4 Wikipedia

I choose the smaller version (nips2010data) for research. Each document of the dataset belongs to one of 19 distinct categories. Each document in the dataset is described by a TF/IDF weighted word vector from a dictionary, which consists of 4973 unique words.

4. Experimental methodology

Baseline:

1. Load a networked dataset -------------------100%
2. 5-fold cross-validation ----------------------100%
3. Content-only classification ------------------100%
4. ICA --100%

Evaluation metric:

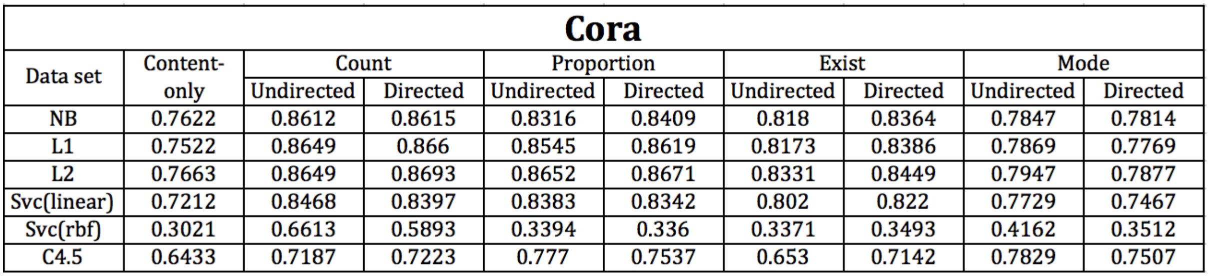
Repeat the ICA process ten times and use the result to calculate the accuracy.

Evaluation methodology:

Use the 5-fold cross-validation to split the data into train and test splits. Each time do the Content-only classification and ICA and we can get 5 results and the average.

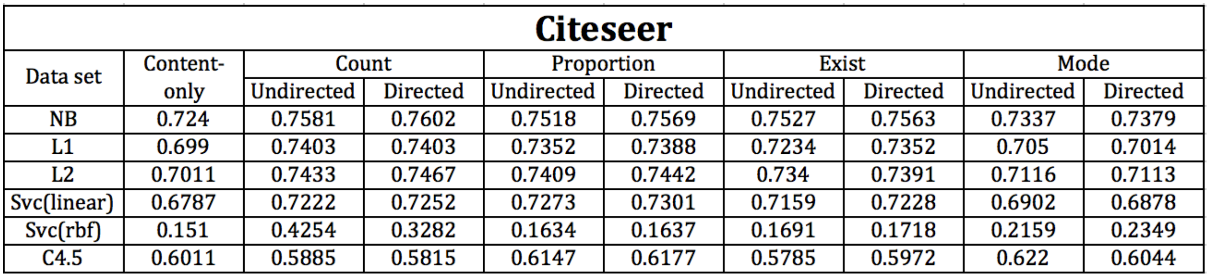
Use the Content-only classification to train a model and test it using only the node attribute.

5. Results



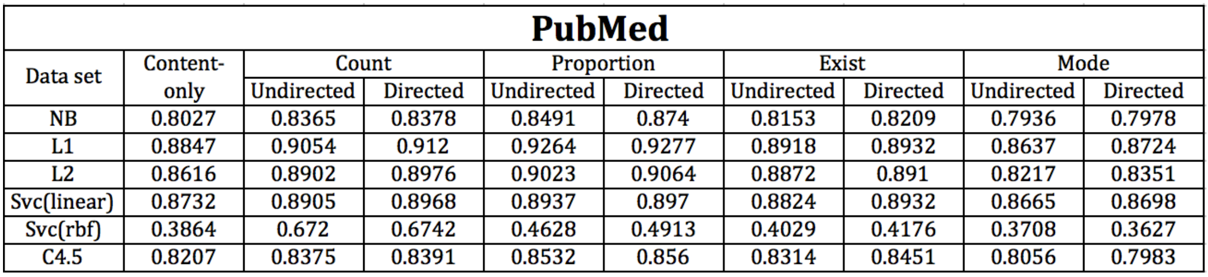
Cora

Data set	Content-	Count		Proportion		Exist		Mode	
	only	Undirected	Directed	Undirected	Directed	Undirected	Directed	Undirected	Directed
NB	0.7622	0.8612	0.8615	0.8316	0.8409	0.818	0.8364	0.7847	0.7814
L1	0.7522	0.8649	0.866	0.8545	0.8619	0.8173	0.8386	0.7869	0.7769
L2	0.7663	0.8649	0.8693	0.8652	0.8671	0.8331	0.8449	0.7947	0.7877
Svc(linear)	0.7212	0.8468	0.8397	0.8383	0.8342	0.802	0.822	0.7729	0.7467
Svc(rbf)	0.3021	0.6613	0.5893	0.3394	0.336	0.3371	0.3493	0.4162	0.3512
C4.5	0.6433	0.7187	0.7223	0.777	0.7537	0.653	0.7142	0.7829	0.7507



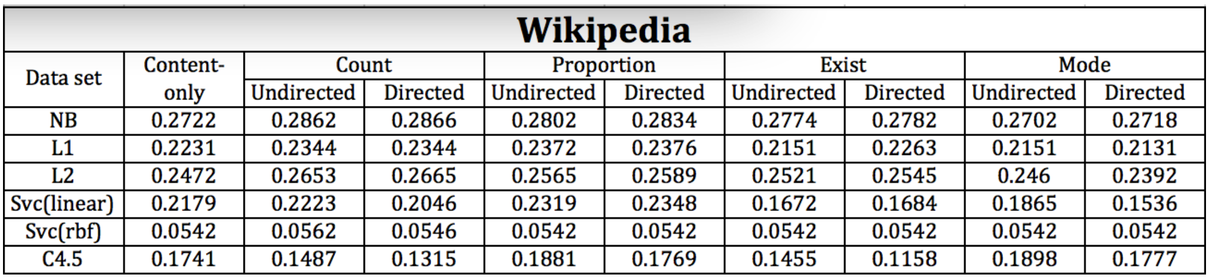
Citeseer

Data set	Content-	Count		Proportion		Exist		Mode	
	only	Undirected	Directed	Undirected	Directed	Undirected	Directed	Undirected	Directed
NB	0.724	0.7581	0.7602	0.7518	0.7569	0.7527	0.7563	0.7337	0.7379
L1	0.699	0.7403	0.7403	0.7352	0.7388	0.7234	0.7352	0.705	0.7014
L2	0.7011	0.7433	0.7467	0.7409	0.7442	0.734	0.7391	0.7116	0.7113
Svc(linear)	0.6787	0.7222	0.7252	0.7273	0.7301	0.7159	0.7228	0.6902	0.6878
Svc(rbf)	0.151	0.4254	0.3282	0.1634	0.1637	0.1691	0.1718	0.2159	0.2349
C4.5	0.6011	0.5885	0.5815	0.6147	0.6177	0.5785	0.5972	0.622	0.6044



PubMed

Data set	Content-	Count		Proportion		Exist		Mode	
	only	Undirected	Directed	Undirected	Directed	Undirected	Directed	Undirected	Directed
NB	0.8027	0.8365	0.8378	0.8491	0.874	0.8153	0.8209	0.7936	0.7978
L1	0.8847	0.9054	0.912	0.9264	0.9277	0.8918	0.8932	0.8637	0.8724
L2	0.8616	0.8902	0.8976	0.9023	0.9064	0.8872	0.891	0.8217	0.8351
Svc(linear)	0.8732	0.8905	0.8968	0.8937	0.897	0.8824	0.8932	0.8665	0.8698
Svc(rbf)	0.3864	0.672	0.6742	0.4628	0.4913	0.4029	0.4176	0.3708	0.3627
C4.5	0.8207	0.8375	0.8391	0.8532	0.856	0.8314	0.8451	0.8056	0.7983



Wikipedia

Data set	Content-	Count		Proportion		Exist		Mode	
	only	Undirected	Directed	Undirected	Directed	Undirected	Directed	Undirected	Directed
NB	0.2722	0.2862	0.2866	0.2802	0.2834	0.2774	0.2782	0.2702	0.2718
L1	0.2231	0.2344	0.2344	0.2372	0.2376	0.2151	0.2263	0.2151	0.2131
L2	0.2472	0.2653	0.2665	0.2565	0.2589	0.2521	0.2545	0.246	0.2392
Svc(linear)	0.2179	0.2223	0.2046	0.2319	0.2348	0.1672	0.1684	0.1865	0.1536
Svc(rbf)	0.0542	0.0562	0.0546	0.0542	0.0542	0.0542	0.0542	0.0542	0.0542
C4.5	0.1741	0.1487	0.1315	0.1881	0.1769	0.1455	0.1158	0.1898	0.1777

6. Related Work

Multi-label Classification:

Most classification problems associate a single class to each example or instance. However, there are many classification tasks where each instance can be associated with one or more classes.

7. Future Work

One of the future work directions is the analysis of different results with different feature selection. For example, compare with the undirected aggregation, the directed mode aggregation reduce the accuracy.
Second, I will try to shorten the amount of the code and strengthen the code reuse.

8. Conclusion

We have perform collective classification algorithm ICA with different local classification methods. Most of the results have shown that collective classification achieves slightly better accuracy then content only classification. These have also shown that feature selection dramatically improves the performance of collective classification.

References

[1] Prithviraj Sen, Galileo Mark Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad, “Collective classi- fication in network data,” AI Magazine, vol. 29, no. 3, 2008.

[2] Qing Lu and Lise Getoor, “Link-based Classification,” 20th International Conference on Machine Learning, Washington, DC, August 2003.

[3] Galileo Mark Namata, Prithviraj Sen, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad, “Collective Classification for Text Classification”, Book Chapter in Text Mining: Classification, Clustering, and Applications (Mehran Sahami and Ashok Srivastava, eds), Taylor and Francis Group, 2009.