Efficient Graph Classification In Shifted Datasets using Weighted Correlated Feature Selection

Md. Samiullah[1], Chowdhury Farhan Ahmed[2], Anna Fariha[1], Akiz Uddin Ahmed[1]

[1] Department of Computer Science and Engineering, University of Dhaka, Bangladesh

[2] ICube Laboratory, University of Strasbourg, France

Presenter:

Chowdhury Farhan Ahmed

OUTLINE

- Introduction
- Background Study
- Related works & Motivation
- Contributions
- Proposed Approach
- Conclusions

INTRODUCTION

- Recently Machine Learning and Data Mining fields are experiencing a trend of dataset shift
- Inter-domain Knowledge is extracted and utilized to improve learning system performance
- Graph, a sophisticated data structure, is capable of capturing effective correlation among objects
- Classification of graphs have great impact on real life applications e.g. human behavior prediction in social networks
- Applications involved knowledge transferring requires specialized classifier model
- Designing a classifier, capable of inter-domain knowledge capturing, is challenging and demanding

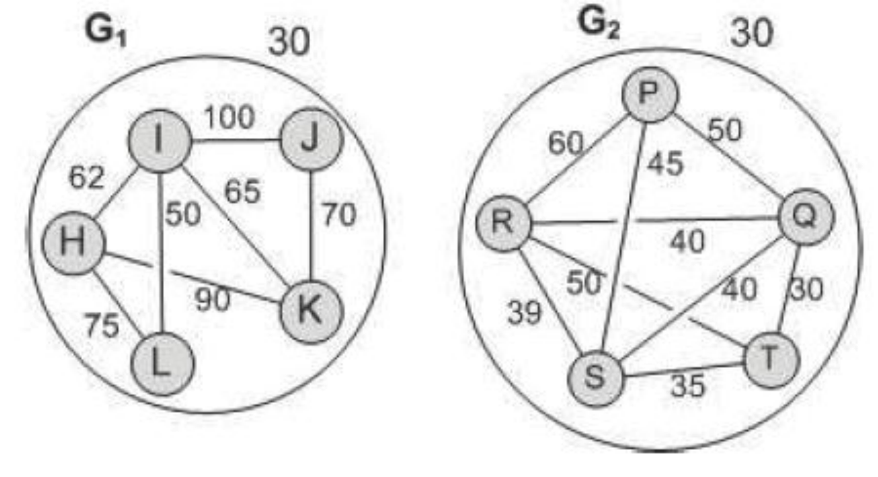
BACKGROUND

- Graph classification is important for predicting behavior or type or class of unknown graphs
- Graph classification is done by:
  - Feature selection
  - Classification
- *Correlation of a graph* G_s, $gConfidence(G_s) = \frac{|G_S = subgrapg(G \epsilon GD)|}{\max_{\forall e_i \in E(G_S \in GD)} \{freq(e_i)\}}$

Here for G1, Numerator = 30, Denominator = 100
Therefore, gConfidence(G1) = 0.3 and for G2, Numerator = 30, Denominator = 60. Therefore, gConfidence(G2) = 0.5

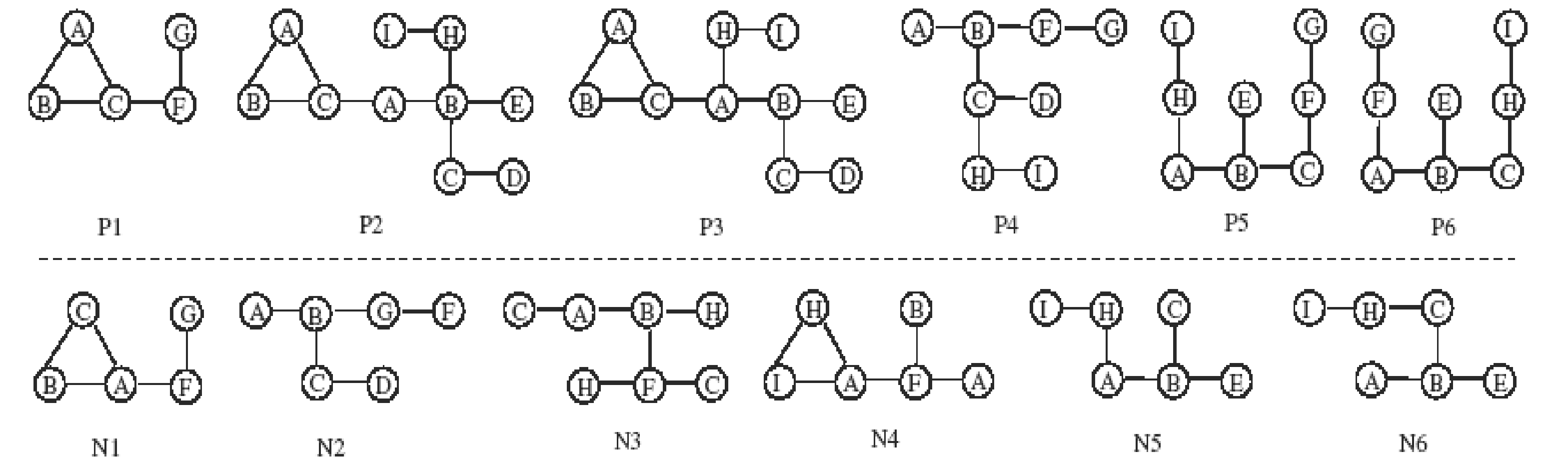
gConfidence(G1) < gConfidence(G2)

Hence, G2 is more inherently correlated than G1

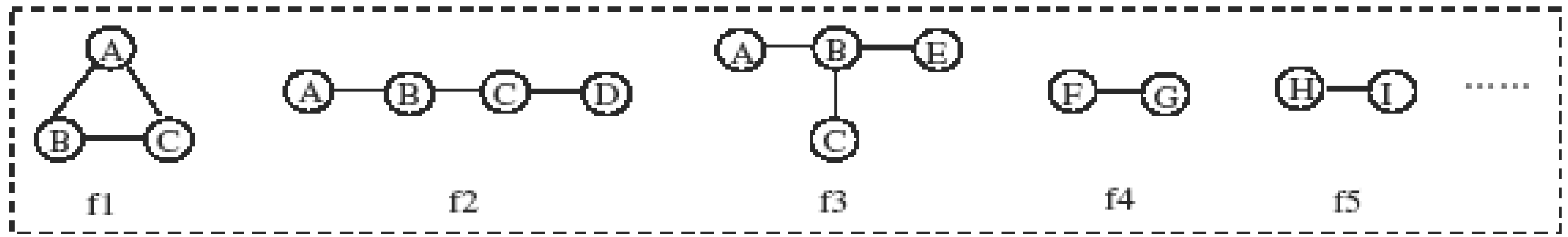


BACKGROUND …

- Few existing classifiers use discriminative fragments, extracted from a set of frequent features, for classifier model construction
- f1 has a discriminative score = log(3/1)= 0.477. Similarly, discriminative score of f4 = log(4/2)= 0.301.
- It is clear that f1 is more discriminative wrt f4 and can be a better candidate in constructing classifier



Sample DB

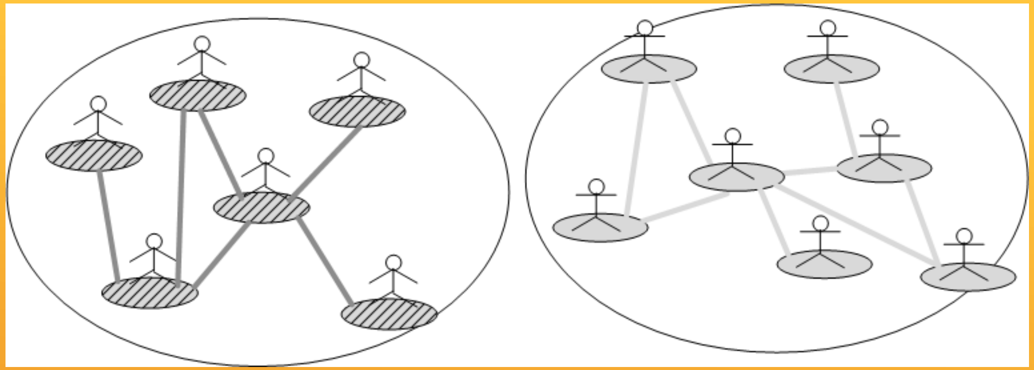


Frequent Fragments

Discriminative score

$$score(f) = \log\frac{r^+(f)}{r^-(f)}, \text{ where } r^+(f) = \frac{|supp(f, D^+)|}{|D^+|} \text{ and } r^-(f) = \frac{|supp(f, D^-)|}{|D^-|}$$

RELATED WORKS & MOTIVATIONS



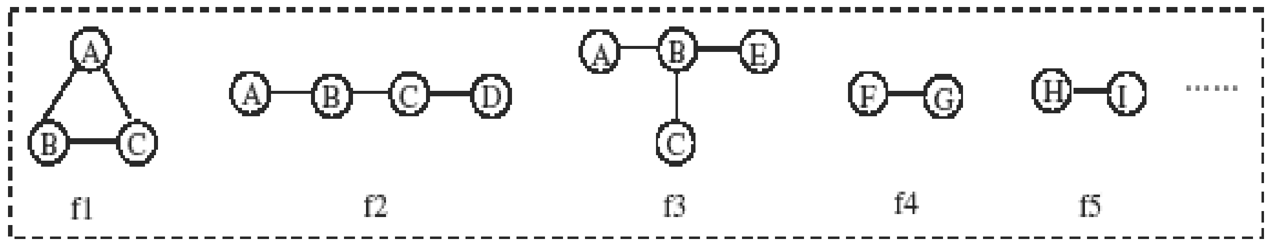
- For classifying graphs and model construction, some approaches [*Yan et al, SIGMOD'08*] mined features from frequent graphs and some [*Jin et al, CIKM'09*] considered co-occurring subgraphs. There exist some approaches [*Zhu et al, CIKM'12*] which additionally considered diversity among features for better classification.
- Existing graph based transfer learning approaches mined significant (frequent) subgraphs using semi-supervised learning method [*Shi et al, SDM'12*] and some alleviated knowledge from domain to domain using graphs [*Jingrui et al, CIKM'09*].
- No other earlier model considered correlation among entities during classification
- Correlation is a key indicator for classification of graphs and transferring knowledge between domains
- Scenario:
  - A social network (facebook, Twitter, LinkedIn) with several groups can be divided into classes based on their behavior.
  - The behavioral classification can be more effective by considering correlation among the persons. A group of researcher, students, businessmen and other professionals in a network like FaceBook can be better classified and the knowledge can be used in Twitter.

CONTRIBUTIONS

- Correlation based feature selection
- New diversity capturing score is proposed
- Effective classifier construction for classifying graphs
- Neural Network based leaning method to adjust the classifier in transferred environment

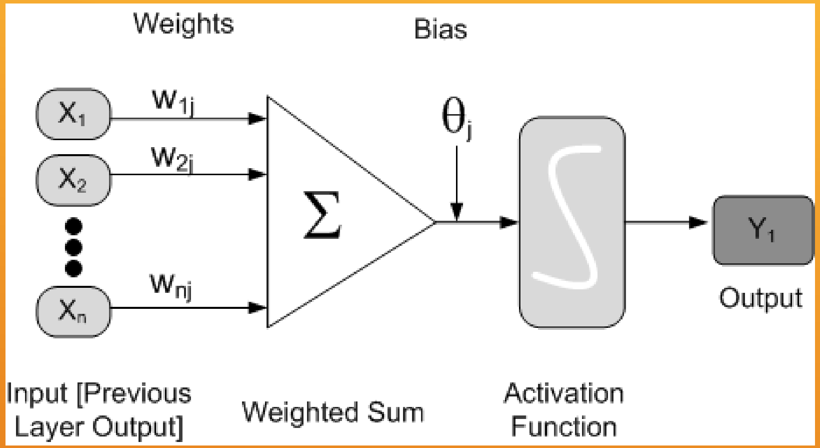
PROPOSED APPROACHES: MEASURES

- $gDistance(G_1, G_2\ ) = \sum_{i=1}^{5} a_i\ \times\ cf_i$
  - a_i represents the number of operation to convert G_2 into G_1 for
    - Addition of **(*i*=1)** vertices, **(*i*=2)** edges,
    - Deletions of **(*i*=3)** vertices, **(*i*=4)** edges and
    - Reversing **(*i*=5)** edge directions
  - cf_i is the cost factor (weight of operations)



- $\mathrm{gDistance(f1, f2)} = 3$. Since, the edge $(\mathrm{A, C})$ need to be deleted and the vertex "D" along with edge $(\mathrm{C, D})$ should be added to convert $\mathrm{f1}$ into $\mathrm{f2}$.
- $Diversity(G) = \min_{\forall sG,\, sG = subset(G)} gDistance(G, sG)$
- $CDDF(G) = w\ \times gConfidence(G) + (1 - w) \times Diversity(G)$
- $Score_{\ CDDF} = log\frac{CDDF^{+}(G)}{CDDF^{-}(G)}$
- The score will be positive for positive labeled graphs and similarly negative for negative graphs.

PROPOSED APPROACHES: MODEL



- $X_1, X_2 \dots X_n$ are treated as input layer units for n Independent Variables and initialized as usual
- Classification output is Y, the dependent variable with the domain $\{0,1\}$ for binary classification
- The model with m number of hidden layer and h number of units per layer have following units : $\mathrm{H}_{1,1}, \mathrm{H}_{1,2}, \dots, \mathrm{H}_{1,h}, \mathrm{H}_{2,1}, \mathrm{H}_{2,2}, \dots, \mathrm{H}_{2,h}, \dots \dots \dots \mathrm{H}_{\mathrm{m},1}, \mathrm{H}_{\mathrm{m},2}, \dots, \mathrm{H}_{m,h}$
- Connection weights between q^{th} and $(q+1)^{th}$ layer units r and s, respectively, are denoted as $\mathrm{W}_{(q,r),(q+1,s)}$.

PROCEDURES

- *Two Sets of best-k features F^+ and F^- are extracted from two source datasets D^+ and D^- which constitutes a set of n-features* {f1, f2, …, fn} with highest correlated and discriminative feature score is selected and a weight of 1 is assigned *[where $n \leq 2*k$]*
- Suppose in target domain, there are a few graphs G_L for learning and several unlabeled graph G_T for testing.
- Then NN based backpropagation algorithm is applied to adjust the weights of the features wrt the class labels of the graphs G_L.
- The feature weight adjustment is performed for every graph of the training dataset.
- Then for any newly coming graph of G_T with unknown label is tested for coverage by the features in both dataset.
- Now, the sum of positive feature weights and sum of negative feature weights are considered for log ratio calculation to estimate the label of the testing graph in target domain.

CONCLUSIONS

- Graph classification, especially in the transfer learning domain, is one of the most crucial topics in state-of-the-art machine learning research
- For the first time we have proposed a correlation based graph classification approach as well as a new diversity capturing measure, which are used for developing classification model
- Neural network based learning model is proposed that can be used for compensating the changes in transferred learning environments
- Currently we are working on experimental analysis to evaluate the performance of our approach and its supremacy.

Thanks …