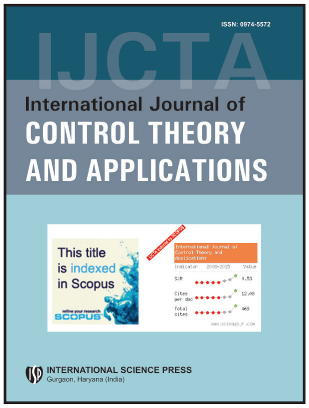

International Journal of Control Theory and Applications

ISSN : 0974-5572

© International Science Press

Volume 10 • Number 19 • 2017

Efficient Subgraph Selection using Principal Component Analysis with Pruning Methods in Multi Task Graph Classification

S. Thanikaivelan[a] and K. Rajiv Gandhi[b]

[a]Research Scholar, PRIST University, Thanjavur, Tamilnadu.
[b]Assistant Professor, Department of Computer Science, Alagappa University College of Arts and Science, Paramakudi, Ramanathapuram, Tamilnadu.

Abstract: Recently, the organization of graph information has emerged to be a significant and lively investigation interest in the past epoch, having numerous real world applications. The present research on graph classification is focused over single label settings. Nonetheless, in several applications, every graph information can be allotted with a set consisting of numerous labels at the same time. The multi task graph application is a huge issue and therefore in the already available system, joint structure feature exploration and regularization technique is proposed. It is employed for finding discriminative sub graphs that is shared by all tasks for the purpose of learning. But the available research has problems with the extraction of good features making use of multiple labels of the graphs prior to graph classification. In the work proposed, multi-task graph classification is carried out by exploiting Norm Lasso regularization. The feature sub graphs are chosen by applying Principal Component Analysis (PCA), utilized for selecting the most frequent sub graph features on the basis of the learning model. Thereafter, the pre-pruning and post-pruning steps are applied to minimize the sub graph search space with more efficacy. It is aimed at extracting the vital sub graph features from the graph sample dataset. Then these chosen sub graph features are merged and provided to the graph classification. In this step, the discriminative score is computed and the best sub graphs get classified. The number of iterations is decreased through the selection of the most discriminative score in addition to best sub graph feature selection. The results indicated on real-world responsibilities show that this feature selection method can efficiently improve the multi-label graph classification recitals and reveals added effectiveness over the clipping of the sub graph search space employing numerous labels. The accuracy of graph organization is enhanced by making use of the PCA and pruning technique proposed. The performance metrics include accuracy, Area Under Cure (AUC), recall, *f*-measure and running time. The newly introduced PCA with pruning technique yields better performance with respect to optimal sub graph classification results.

1. INTRODUCTION

Designing computational methods that are based on classification for determining the identity of data indicated as graphs has gained the focus in the form of a theme in data mining research. Graph classification offers enormous strong advantages in various applications. Graph-based classification can find its application in the drug discovery problems, in which it is helpful in learning the structural characteristics of a chemical compound and the impact which they have over treating a specific disease. In the past few years, different advanced classification procedures for graph-based information have been designed [1] [2]. Many of the algorithms are dependent on the underlying supposition that the inherent characteristics of a graph are provided by its underlying substructures (nodes, edges, paths, strongly connected components, trees, sub graphs, etc). These substructure-based algorithms detect the significant elements available in every graph and then consequently utilize them for discriminating graphs from diverse classes [3].

Over the years, however, recurrent sub graph grounded method was presented where frequent sub graph mining procedures are used for generating all the sub graphs which happen in an adequately huge numeral of times in the graph databank and for constructing feature vectors on the basis on the haul out sub graphs [4]. Through the transformation of every graph into its respective feature vector, any one among the available classification algorithms like Support Vector Machine (SVM), boosting, decision trees, or rule-based classifiers is subsequently applied to construct a classification replica for the graph databank.

Graph mining techniques ascertains the sub graph patterns that appear frequently that could be exploited in the form of features for the successive organization or regression. But the frequent patterns do not unavoidably provide information for the learning problem given. A mathematical programming increasing technique (gBoost) is proposed which progressively gathers the informative patterns. In this technical work, g Boost can construct the forecast rule with less number of iteration. In order to use the improving technique to graph data, a branch-and-bound pattern search procedure is designed on the basis of the DFS code tree. The search space constructed is reclaimed in the repetitions coming later for the purpose of minimizing the computational time [5].

Multitask learning refers to the task of forecasting the efficiency of different amalgamations of drugs that are associated. During the forecast of the development of disease, the outcome forecast at every time point can be treated as a task and these responsibilities are associated temporally. In the case of multitask learning, these associated tasks are learnt at the same time through the extraction and utilization of suitable communal data across errands. Simultaneously learning multiple connected tasks is helpful in increasing the example size for every job efficiently, and enhances the forecast recital. This way, multi-task learning is particularly advantageous if the training sample size is smaller for every job [6].

Multi-task learning (MTL) targets at boosting the simplification recital of administered regression or classification through the simultaneous learning of numerous related tasks. Recently, MTL has gained remarkable research focus in the data mining and machine learning community [7]. It has been noticed that simultaneous learning of multiple related tasks frequently enhances the demonstrating precision and makes it improved. Consider a challenge of predicting the cancer condition on the basis of Microarray data sets, in which there exist numerous data sets for various kinds of cancers [8]. Every data set consists of numerous Microarray information from patients who do or do not have the particular cancer. Few cancers have “similarity” with one another (e.g. breast cancer vs ovary cancer) where as few are some what diverse (e.g. breast cancer vs prostate cancer). For comparable kinds of cancer (tasks), learning replicas constructed for those comparable cancer kinds (tasks) are anticipated to be sharing similar features; for the case of dissimilarity, learning replicas are supposed to choose diverse features. Nonetheless, present feature selection techniques for MTL [9] choose a subset of features common crosswise all the responsibilities.

Graph classification has attracted admirable interest in the recent time sowing to the rising numeral of applications that involve objects having complicated structure associations. Till date, all of the available graph classification procedures make an assumption, either openly orindirectly, that misclassification of occurrences in various classes consumes an equivalent sum of cost (or risk) that is frequently not in real-life requests (in which misclassification of a particular class of samples, like unhealthy patients, is prone to added expenditure compared to others). Even though cost-effective learning has been studied extensively, all the techniques are dependent on information having instance-feature demonstration. However, graphs, does not possess features that are obtainable for learning and then the feature space of graph information are possibly immeasurable and requires to be cautiously investigated so as to favor the classes having a huge expense [10].

2. RELATED WORK

Vogelstein et. al., [11] introduced a new graph/class model for the purpose of statistical investigation. It explores two approaches for the estimation of the signal-sub graph: the first making use of just the vertex label information, the second also making use of graph structure. The signal-sub graph estimators are employed for improving the classification performance. But this work poses problems with misclassification error in some cases.

Pan et. al., [12] proposed an imbalanced graph boosting algorithm, ig Boost for dealing with the Imbalanced Class Distributions and Noise. It chooses informative sub graph designs progressively from excessive graph information to aid learning. In this research, a boosting framework takes the class distributions into consideration and the distance consequence of every hart to its class center to weigh separate graphs. Depending on the weight values it combines the sub graph assortment and margin optimization between optimistic along with the negative graphs to create an increasing framework, therefore the sub graph designated can assist in finding better margins and the optimized margins more over aid in selecting improved sub graph features. Nonetheless, under-sampling actually makes changes in the sample distributions and might result in loss of valued data present in the tested graph information, resulting in low-quality results.

Thoma et. al., [13] introduced Correspondence-based Quality Criterion for effective feature selection among frequent sub graphs. It integrates two major benefits. First, it does the optimization of a sub modular quality criterion that indicates that a near-optimal solution can be provided making use of greedy feature selection. Secondly, this sub modular quality function standard can be combined into gSpan, the benchmarked device for frequent sub graph mining, and helps in pruning the search space for discriminative recurrent sub graphs even while performing recurrent sub graph mining. But it has problem with process being slower.

X. Kong et. al., [14] explored the issue of semi-supervised feature assortment for graph organization and introduces a new answer, referred to as gSSC, to effectively search for optimal sub graph features along with categorized and unlabeled graphs. This work proposed a feature assessment standard, known as *g* Semi, for the evaluation of sub graph features having both characterized and unlabeled graphs, and obtains an upper-bound for *g* Semi for the purpose of pruning the sub graph search space. Thereafter, a branch-and-bound algorithm is used for proficiently finding a regular finest sub graph feature, helpful for the graph organization. Experiential results have revealed that feature assortment method for graph organization performs better than supervised and unsupervised methods. It is actual effective as it prunes the sub graph search space exploiting together labeled and unlabeled graphs.

Godbole et. al., [15] proposed techniques of improving the available discriminative classifiers for multi-labeled forecasts. Discriminative techniques such as support vector machines achieve better for classification tasks involving uni-labeled text. This work yields a novel method for integrating text features and features specifying associations amongst classes that could be utilized with any kind of discriminative procedure. It also presents two improvements to the margin of SVMs for constructing improved replicas with the overlying classes extant. The outcomes of trials over real world text standard datasets are presented. These new techniques exhausted the precision of the available techniques with considerable improvements statistically.

Liu et. al., [16] investigated the issue of joint feature selection transversely a cluster consisting of associated jobs having applications in several areas inclusive of biomedical informatics and computer apparition. It takes the $l_{2,1}$ - norm regularized regression replica into consideration for joint feature assortment from different tasks that could be obtained in the probabilistic structure. In order to speed up the computation, first it is proposed to have the reformulation of it into two equal smooth convex optimization issues, and thereafter resolve the reformulations by means of the Nesterov's technique. Nonetheless this research imposes challenges with computational complexity.

Fei et. al., [17] introduced a new feature selection technique for graph classification. Through the ranking of features on the basis of their spatial distribution and thereafter the influences to organization, a feature selection technique (and many differences) known as structure grounded feature assortment technique is designed. This work is focused over the selection of a lesser subset of features to construct an effective graph organization prototype. But it does not have the capability for dealing with semi supervised feature selection techniques.

Shi et. al., [18] suggested a Bayesian-based transfer learning model. This work is focused over the evaluation of the resemblance amongst the objective and the basis datasets through the estimation of the degree that is shared on their important sub graphs. Then, this dataset comparison is utilized for judicious selection of significant sub graphs from similar kind of (associated) datasets for the objective dataset. A problem of optimization is derived in order to increase the probability that the sub graphs designated are important in the objective dataset.

3. PROPOSED METHODOLOGY

A. Multi-task Graph Classification

The newly introduced multi task graph classification of PCA with pruning technique yields optimal sub graph feature selection to be used for the graph dataset. The entire block illustration of the proposed scheme is revealed in the Figure 1.

With the aim of achieving multi-task graph organization, the subject is about using multi-task for guiding an iterative sub graph exploration process so as to attain the lowest regularized empirical risks for all the tasks. This can be expressed in the objective function below [12]:

$$= \min_{\mathrm{W}, b} \sum_{t=1}^{\mathrm{T}} \frac{1}{n_t} \sum_{i=1}^{n_t} \mathcal{L}(y_t, i, f_t(x_t, i)) + \gamma \mathrm{R}(\mathrm{W}) \tag{1}$$

Where, $\mathrm{W} = [w_1, ..., w_{\mathrm{T}}]$ refers to a weight matrix that indicates the weights of every sub graph corresponding to various tasks, $b = [b_1, ..., b_{\mathrm{T}}]$ refer to the bias parameters for every function f_t and n_t stands for the number of training graphs present in task t. It gives the measure of the loss over the training graphs for every task where $\mathcal{L}(y_t, i, f_t(x_t, i))$refers toa loss function, giving the measure of the misclassification penalty of a graph. $\gamma\mathrm{R}(\mathrm{W})$ stands for a regularization term for enforcing sparse solutions. The Logistic loss function

$$\mathcal{L}(y_t, i, f_t(x_t, i)) = \log(1 + \exp\{-y_t, if_t(x_t, i)\}) \tag{2}$$

To derive a sparse answer on W, which means, a determinate set of sub graph features that is shared by entire tasks, deliberate the regularizes below:

$$l_1\text{-Norm Lasso regularization } \mathrm{R}(\mathrm{W}) = \sum\nolimits_{k,t} \left| \mathrm{W}_{k,t} \right|$$

The rational idea is that the l_1-norm regularizer can generate solutions by means of several coefficients being 0, referred to as Lasso and has been extensively used for different assortments. A interpretation of Lasso in MTG is about using a parameter to have a switch over the regularization of all the responsibilities, supposing that various tasks have the similar sparsity factor.

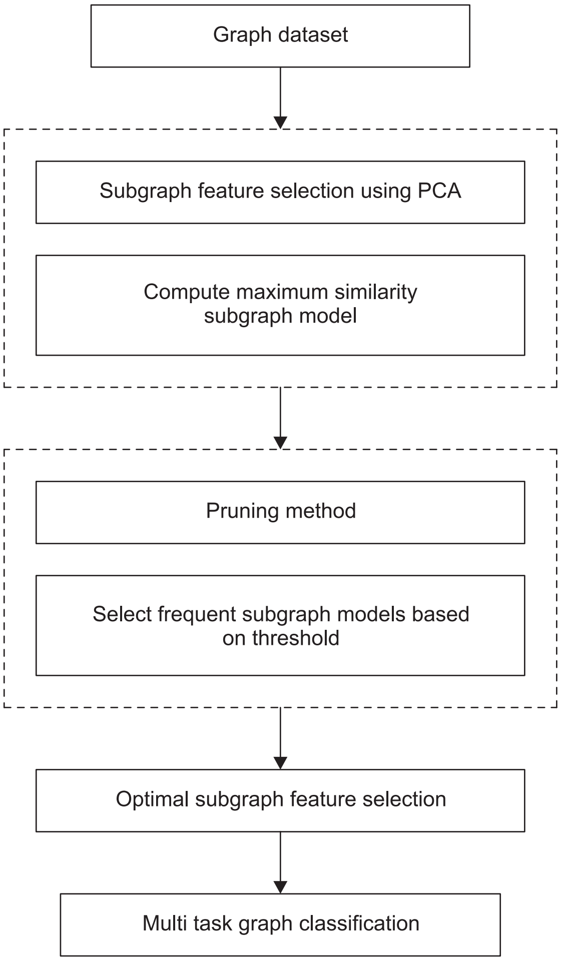
Graph dataset
Subgraph feature selection using PCA
Compute maximum similarity subgraph model
Pruning method
Select frequent subgraph models based on threshold
Optimal subgraph feature selection
Multi task graph classification

Figure 1: Overall Architecture Diagram for Proposed Methodology

$l_{2,1}$-norm normalization: Since the total sub graph space is endlessly large and it has to choose only a subset of most significant sub graphs among all the likely sub graphs, a mixed-norm regularizer is proposed to be used.

$$\| \mathrm{W} \|_{2,1} = \sum_{k=1}^{m} \sqrt{\sum_{t=1}^{\mathrm{T}} | \mathrm{W}_{k,t} |} = \sum_{k=1}^{m} | \mathrm{W}_k |_2 \qquad (3)$$

Where W_k refers to the k-th row of W. The $l_{2,1}$ regularizer at first computes the l_2-norm of every row in W and there after computes the l_1-norm of the vector $d(\mathrm{W}) = (| \mathrm{W}_1 |_2, ..., \| \mathrm{W}_m \|_2)$.

When the entire feature set $\mathrm{F} = \{g_1, ..., g_m\}$ is not large and is available for learning, the objective function in Eq. (3) can be efficiently resolved by making use of an available toolbox [19] for either ‘1 or ‘2;1 norm regularization. For the graph data, but, there are two problems: (4) the entire feature set F is implied and is unobtainable, and enumeration of the sub graph features is NP-complete; and the numeral of sub graphs is very large and probably infinite ($m \rightarrow +\infty$).

In order to resolve the trials mentioned above, this work introduces to iteratively have the features/sub graphs included into objective function. Otherwise said, multi-task sub graph selection and model learning are combined into one objective function for mutual benefits. To be more specific, it carries out sub graph selection on the basis of the sub gradient of the objective function $\mathcal{J}$, hence the empirical loss can be minimized always while choosing and having the most discriminative sub graph added to the existing sub graph feature set. Once a new sub graph is integrated, the new restricted master problem is resolved in Eq. (5).

$$f_t(x_{t,i}) = x_{t,i} \cdot w_t + b_t = \sum_{g_k \in \mathrm{F}} w_{t,k} h_{g_k}(\mathrm{G}_{t,i}) + b_t \tag{4}$$

$$\mathcal{J}_1 = \min_{(w^{(s)}) \cdot b^{(s)}} \sum_{t=1}^{\mathrm{T}} \frac{1}{n_t} \sum_{i=1}^{n_t} \mathcal{L}(y_t, i, f_t(x_{t,i}^{(s)}) + \mathrm{YR}(\mathrm{W}^{(s)})) \tag{5}$$

Where $\mathrm{W}^{(s)}$ and $b^{(s)}$ refer to the solutions on the basis of the selected features in the s-th iteration, and $x_{t,i}^{(s)}$ stands for the feature representation of $x_{t,i}$ corresponding to the selected features.

The feature selection and model learning process mentioned above goes on till there is algorithm convergence. In order to deal with the big sub graph space, this work introduced Principal Component Analysis (PCA) for the sub graph features reduction. It is helpful in reducing the search space.

Principal Component Analysis (PCA) derives a new coordinate system. The axes of the new coordinate system are ortho normal to one another and are known as principal components (PC). Every TAG present in a set of TAGs is indicated as a matrix, G. Element G_{ij} provides the total number of bytes that are transferred from node i to node j within the interval of observation. The vectorized form of G is treated to be a vector in one dimension.

B. Feature Selection using PCA

Input: number of sub graph features

Output: Selection of sub graph results from PCA σ^2_{sel}

```
$\sigma^2_{\mathrm{ub}} \leftarrow$ get upper bound $(miy)$
$\sigma^2_{\mathrm{est}} \leftarrow \sigma^2_{\mathrm{ub}}$
For $i = 1$ to $i_{\max}$
    $\sigma^2_{\mathrm{next}} \leftarrow$ get next variance estimate $(miy, \sigma^2_{\mathrm{est}}, \sigma^2_{\mathrm{ub}})$
    If $\sigma^2_{\mathrm{est}} = \sigma^2_{\mathrm{next}}$ then
        Return $\sigma^2_{\mathrm{est}}$
    End if
    $\sigma^2_{\mathrm{est}} \leftarrow \sigma^2_{\mathrm{next}}$
End for
Return $\sigma^2_{\mathrm{est}}$
```

Maximum possible p is equivalent to 1, associated with the entire dataset sample features having edges. Next, it ignores the sub graph features having the maximum variance by reducing p to $1 - \Delta p$, $1 - 2\Delta p$, and continues, till p is lesser than $p_{\min}$. Upper bound σ^2_{ub} is used as an additional test of the accuracy of the estimate computed. Function Apply PCA computes $\tilde{\lambda}_{\mathrm{MIY},\,i}$, $i = 1, \ldots, \mathrm{M}$.

C. Get next Subgraph Feature Selection

Input: Medical image with noise *miy*, Noise variance estimation outcome from PCA σ^2_{est}, upper bound σ^2_{ub}

Output: Next sub graph feature estimate σ^2_{next}

```
1.  $p \leftarrow 1$
2.  $\sigma^2_{\text{next}} \leftarrow 0$
3.  While $p \geq p_{\text{min}}$ do
4.  $\tilde{\lambda}_{\text{MIY, 1}}, \ldots, \tilde{\lambda}_{\text{MIY, M}} \leftarrow$ Apply PCA(B($p$))
5.  $\sigma^2_{\text{next}} \leftarrow \tilde{\lambda}_{\text{MIY, M}}$
6.  If $\tilde{\lambda}_{\text{MIY, M}} - m + 1 - \tilde{\lambda}_{\text{MIY, M}} < \text{T}\sigma^2_{\text{est}} / \sqrt{|B(p)|}$ and
7.  $\sigma^2_{\text{next}} \leq \sigma^2_{\text{ub}}$ then
8.  Return $\sigma^2_{\text{next}}$
9.  End if
10. $p \leftarrow p - \Delta p$
11. End while
12. Return $\sigma^2_{\text{next}}$
```

When considering the execution time of the program,, the function Apply PCA has to be focused on, as it is referred to inside two loops: the initial loop starts from lines 3-9 of function sub graph feature estimation [20]; and the subsequent loop begins from lines 3–10 of function Get Next sub graph feature estimate. Function Apply PCA consists of two elements:

1. Computation of the sample covariance matrix

$$\frac{1}{|\text{B}(p)|-1}\left(\sum_{miy_i\,\in\,\text{B}(p)} \text{MIY}_i\text{MIY}_i^{\text{T}} - \frac{1}{|\text{B}(p)|}\sum_{miy_i\,\in\,\text{B}(p)} \text{MIY}_i \sum_{miy_i\,\in\,\text{B}(p)} \text{MIY}_i^{\text{T}}\right) \tag{6}$$

The amount of the operations is proportional with $|\text{B}(p)|\text{M}^2$.

2. Calculation of the Eigen values of the samples covariance matrix.

In lieu of the fact that, $|\text{B}(p)| \gg \text{M}$, the calculation of the trial covariance matrix is the most important component of the function Apply PCA. Suppose $\text{C}_{\text{MIX}} = \Sigma_{\text{MIX}_i\,\in\,\text{MIX}}\text{MIY}_i\text{MIY}_i^{\text{T}}$ and $\text{C}_{\text{MIX}} = \Sigma_{\text{MIX}_i\,\in\,\text{MIX}}\text{MIY}_i$. It is to be noticied that for disjoint samples MIX_1 and MIX_2, $\text{C}_{\text{MIX}_1\,\cup\,\text{MIX}_2}$ and $\text{C}_{\text{MIX}_1\,\cup\,\text{MIX}_2} = \text{C}_{\text{MIX}_1\,+\,\text{MIX}_2}$. At that instant (2) can be re-expressed as,

$$\frac{1}{|\text{B}(p)|-1}\left(\text{C}_{\text{B}(p)} - \frac{1}{|\text{B}(p)|}\text{C}_{\text{B}(p)}c^{\text{T}}_{\text{B}(p)}\right) \tag{7}$$

Function Apply PCA is invoked only with arguments

$$\text{B}(1) \supset \text{B}(1-\Delta p) \supset \ldots \supset \text{B}(1-n\Delta p) \tag{8}$$

Where $n = [1 - p_{\text{min}}]/\Delta p$ and = [mix] indicates the highest integer not above mix. For $j = 0, \ldots, n-1$, consider,

$$\mathrm{MIY}_j = \{\mathrm{miy}_i | \mathrm{Q}(1 - (j + 1)\Delta p) < s^2(\mathrm{miy}_i) \leq \mathrm{Q}(1 - j\Delta p)\} \tag{9}$$

Subsequently,

$$\mathrm{B}(1 - j\Delta p) = \mathrm{B}(1 - (j + 1)\Delta p \cup \mathrm{MIY}_j) \tag{10}$$

During the beginning of the PCA methods for noise variance estimation, the covariance matrices $\mathrm{C}_{\mathrm{B}(1 - j\Delta p)}$ and vectors $\mathrm{C}_{\mathrm{B}(1 - j\Delta p)}$, $j = 0, ..., n$ are precomputed. Matrices $\mathrm{C}_{\mathrm{B}(1 - n\Delta p)}$, $\mathrm{C}_{\mathrm{MIY}_{0, ..., }\mathrm{MIY}_{n-1}}$ and vectors $\mathrm{C}_{\mathrm{B}(1 - n\Delta p)}$, $\mathrm{C}_{\mathrm{MIY}_{0, ..., }\mathrm{MIY}_{n-1}}$,

$$\mathrm{C}_{\mathrm{B}(1 - j\Delta p)} = \mathrm{C}_{\mathrm{B}(1 - (j + 1)\Delta p)} + \mathrm{C}_{\mathrm{MIY}_j} \tag{11}$$

$$c_{\mathrm{B}(1 - j\Delta p)} = c_{\mathrm{B}(1 - (j + 1)\Delta p)} + c_{\mathrm{MIY}_j} \tag{12}$$

D. Subgraph Feature Pruning Method

Pruning criteria allows then arrowing down of the search space while searching for discriminative sub graphs [21]. It is actually clipping the sub graph search space making use of together labeled and unlabeled graphs.

Input: Multi label graphs $X \in G^n$, set of frequent sub graphs S, frequent sub graph S, threshold min Sup

```
1.  Function sub graph mining (X, S, S, min Sup)
2.  If S is not minimal then
3.  Return
4.  End if
5.  $S \leftarrow S \cup \{S\}$
6.  Compute all uninterrupted children C of S
7.  For every C:C is a (potentially negligible) child of S do
8.  If C.D. size $\geq$ min Sup then
9.  Subgraph mining (X, S, C, min Sup)
10. End if
11. End for
12. End function
```

Any graph S can be extended repetitively into a child C so that C is a super graph of S. Every time, it reaches a graph S that is not frequent in X, it backtracks to the nurturing frequent sub graph and tries its subsequent child.

However, to prevent duplicate graphs, interest is only shown in least BFS code allowances. Non-minimal BFS codes can only be partly discovered by means of a series of rapidly obtainable pre-pruning characteristics. They can be used in line 7 of process 3. In order to guarantee the unfussy of a BFS code, a more complex test process has to be carried out. Therefore, it will only be carried out on BFS codes that have been already seen to be often. It limits the unnecessary features through the comparison of the minimum threshold value.

E. Multi-Task Graph Classification Algorithm

For a sub graph pattern g_k, its definition of discriminative notch over all T tasks is as below:

$$\Theta(g_k) = |\Delta C_k \,.\, 1| = \left| \sum_{t=1}^{T} \nabla C_{k,\, w_t} \right| \tag{13}$$

In which ΔC_k and $\Delta C_{k,\, w_t}$ are defined $\Delta C_{k,\, w_t}$ refers to the gradient of loss term C on the sub graph feature g_k corresponding to *t*-th task as $\Delta C_{k,\, w_t}$. ΔC_k refers to the gradient vector of feature g_k completed all the T tasks.

At first, the weights for all the training graphs in every job are fixed correspondingly to be $1/n_t$ (n_t refers to the amount of labeled sub graphs in task *t*), and the lively set F1 is set to be unfilled.

The algorithm does the mining of a set of sub graphs P from F3 having the biggest MTG discriminative scores as distinct by Eq. (9). This step consists of a multi-task driven sub graph mining process that will be dealt with in the subsequent subsection. In order to minimize the number of iterations for sub graph mining, top K sub graphs are employed in every repetition(rather than the finest sub graph).

$\{(G_t, y_{t,\,1}), ..., (G_{t,\,n}\, y_{t,\,1})\}$, $t \in \{1, 2, ..., T\}$ graph dataset from tasks

γ: Predetermined regularization parameter

α_{ti}: Weight for each graph example

K: Number of optimal sub graph patterns

F_1: Already selected subgraph set

Output:

$P = \{g_k\}_{k = 1, ..., k}$: The top – K subgraphs

```
$\eta = 0, P \leftarrow \phi$
while recursively visit the Breath First Search (BFS) tree
$g_p \leftarrow$ Current visited subgraph in BFS code tree
if $g_p$ has been examined then
Continue
Compute score $\Theta(g_p)$ and $Y(g_k)$ for subgraph $g_p$
Based on (9) and (10)
if $g_p \in F_3^{r'} \,\&\, \Phi(g_p) > \eta$ then
$P \leftarrow P \cup g_p$
if $|P| > K$ then
$g^* \leftarrow \text{argmin}_{g_k \in P\Theta(g_k)}$
$P \leftarrow P/g^*$
$\eta \leftarrow \min_{g_k \in P\Theta(g_k)}$
if $\widehat{\Theta}(g_p) > \eta \,\&\, \widehat{Y}(g_p) > Y$ then
Breath first hunt the sub tree rooted from node $g_p$
Return $P = \{g_k\}_{k = 1, ..., K}$;
```

Multitask driven sub graph mining procedure is provided in Algorithm 2. The least value h in optimal set P is set on step 1. Duplicated sub graph types are pruned on the steps 4-5, and the discriminative score $\Theta(g_p)$ and conditional score $Y(g_p)$ for g_p are computed on step 6. If g_p is included in the current candidate set $F_3^r\, i = \{g_k | g_k \in F_2, Y(g_k) > Y + \varepsilon$ and $\Theta(g_p)$ is bigger than η, it is added g_p to the feature set P (steps 7-8). If the size of P does exceed beyond the predetermined size K, the sub graph with the least discriminative notch is eliminated (steps 9-11). Then, the algorithm updates the minimum optimal value h on step 12, and makes use of two branch-and-bound pruning rules, Theorems II and III, for clipping the search space on step 13. The two rules will minimize the unpromising candidates by employing discriminative scores and conditional scores, respectively. Finally, the optimal set P is derived on step 15.

4. EXPERIMENTAL RESULT

This section uses Predictive Toxicology Challenge Dataset (PTC). The PTC challenge consists a amount of carcinogenicity jobs for forecast of toxicology of chemical compounds. The dataset comprises of 417 compounds along with four sorts of test animal: MM (male mouse), FM (female mouse), MR (male rat), and FR (female rat). every compound possesses one label chosen from {CE, SE, P, E, EE, IS, NE, N} that refers to Clear Evidence of Carcinogenic Activity (CE), Some Evidence of Carcinogenic Activity (SE), Positive (P), Equivocal (E), Equivocal Evidence of Carcinogenic Activity (EE), Inadequate Study of Carcinogenic Activity (IS), No Evidence of Carcinogenic Activity (NE), and Negative (N). Similar to [22], we set {CE, SE, P} as positive labels, and {NE, N} as negative labels. In order to make a formulation an MTG dataset, 417 compounds are haphazardly divided into four equal and disjoint subsets. For every subset, one kind of carcinogenicity experiment is considered to be its learning task. The performance metrics are assessed by making use of the available MTL, MTG technique and the PCA with pruning technique proposed.

A. Accuracy

Accuracy is defined to be the whole exactness of the replica and is computed as the sum of definite classification factors ($T_p + T_n$) isolated by the total amount of classification factors ($T_p + T_n + F_p + F_n$)

$$\text{Accuracy} = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \tag{14}$$

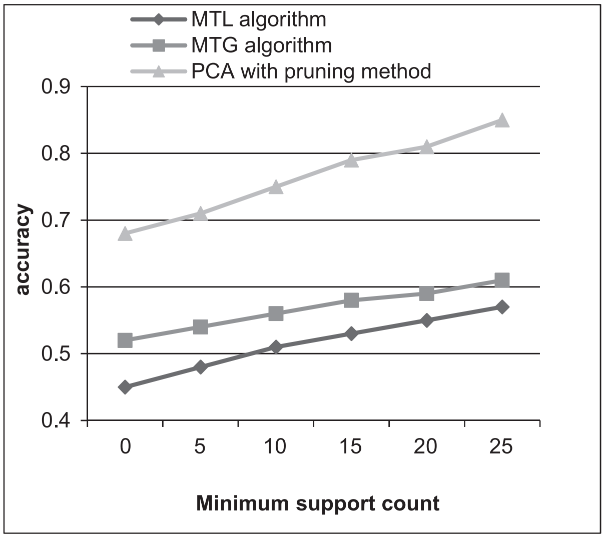

Figure 2: Accuracy metric

The accuracy comparison for graph dataset is illustrated from the graph in Figure 2. The minimum support count is plotted along x-axis and the accuracy is plotted along the y-axis. The experimental results show that the available MTL and MTG technique have provided only lower accuracy results. The PCA with pruning technique proposed show the higher classification accuracy results.

B. Recall

Recall value: Recall value is computed based on the repossession of data at true positive prediction, false negative. Typically it can be defined as

$$\text{RECALL} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}} \tag{15}$$

Recall is also known as sensitivity. Recall in the retrieval of information indicates the proportion of the pages having relevance with the query, successfully retrieved.

$$\text{RECALL} = \frac{|\{\text{relevant pages}\} \cap \{\text{retrieved pages}\}|}{|\{\text{relevant pages}\}|} \tag{16}$$

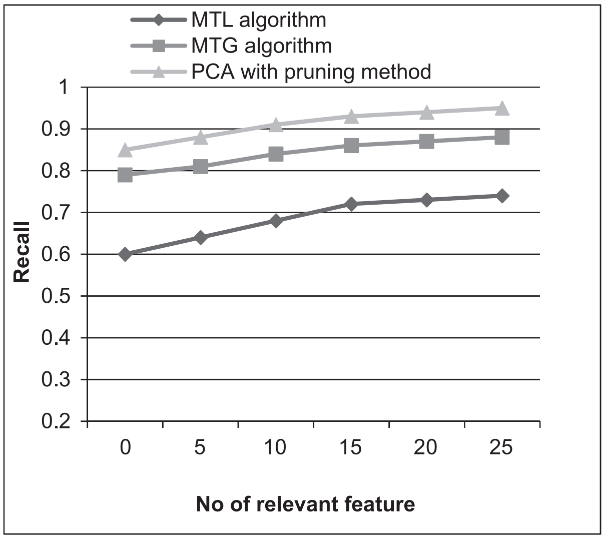

Figure 3: Recall Metric

Figure 3 shows the results of No of relevant feature Vs recall for the available MTL and MTG technique and PCA with pruning technique. It is revealed from the outcomes that the recommended PCA with pruning technique has higher recall compared to the available MTL and MTG technique.

C. Precision

Precision is defined to be the fraction of the true positives against both true positives and false positives results for intrusion and real features. It is defined as below

$$\text{Precision} = \frac{\text{T}_p}{\text{T}_p + \text{F}_p} \tag{17}$$

Figure 4 illustrates the results of No of relevant feature Vs precision for existing MTL and MTG technique and proposed PCA with pruning technique. It is revealed from the results that the proposed PCA with pruning technique has higher precision compared to the available MTL and MTG technique.

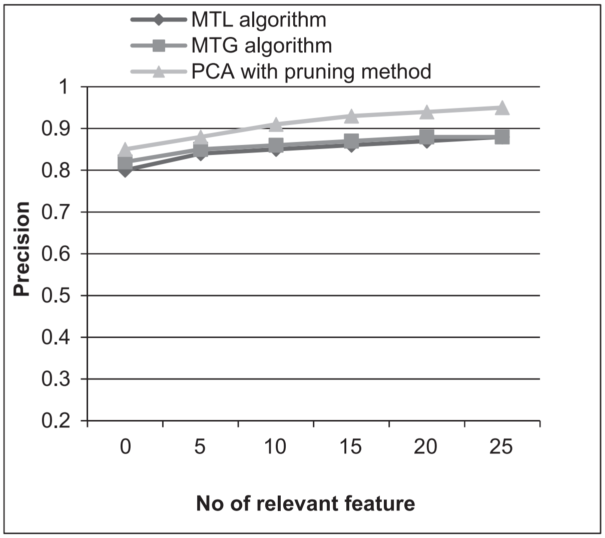

Figure 4: Precision Metric

D. F-measure

It measures the accuracy of the test. It takes both the accuracy *p* and the recalll *r* of the test into consideration for computing the score: *p* refers to the number of correct positive results divided by the numeral of all the positive outcomes, and *r* views for the numeral of correct positive outcomes divided by the numeral of positive outcomes, which ought to be returned.

$$F_1 = \frac{\text{precision . recall}}{\text{precision + recall}} \tag{18}$$

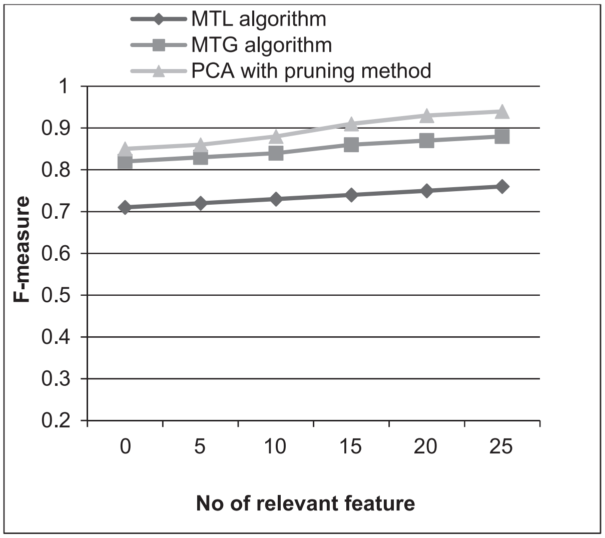

Figure 5: F-measure Metric

The graph from the above Figure 5 illustrates the accuracy comparison made for graph dataset. The number of relevant feature is plotted along the x-axis and the accuracy is plotted along the y-axis. The experimental results shows that the available MTL and MTG technique provides the lower accuracy results. The PCA proposed with pruning technique yields the higher classification *f*-measure results.

E. AUC

The area under the curve (AUC) specifies the area underneath the curve (mathematically called as definite integral) in a plot of concentration.

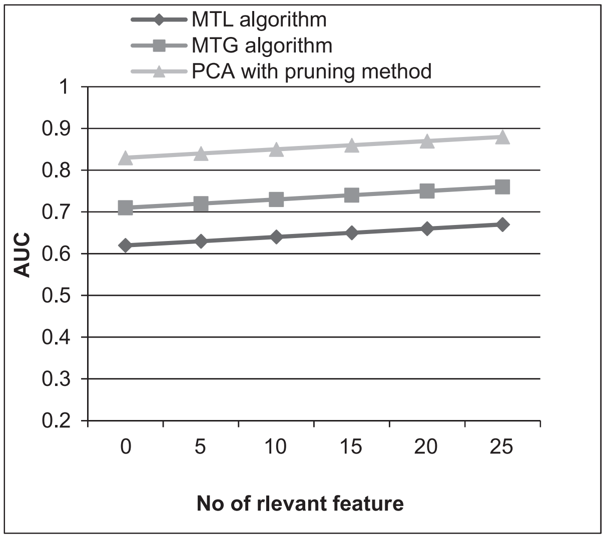

Figure 6: AUC Metric

Figure 6 illustrates the results of the number of relevant feature Vs recall for the available MTL and MTG technique and PCA with pruning technique proposed. It is experimental from the results that the PCA proposed with pruning technique has a higher AUC metric compared to the available MTL and MTG technique.

F. Running Time

This section shows that the efficiency of pruning is measured through the reduction of the search space for sub graph feature exploration, since the entire sub graph searches space is exponentially huge (or infinite). A threshold value min sup, denoting the minimum frequency of every qualified sub graph feature in the training graph datasets, is for enclosing the number of sub graphs in the search space. In this manner, it is aware of the

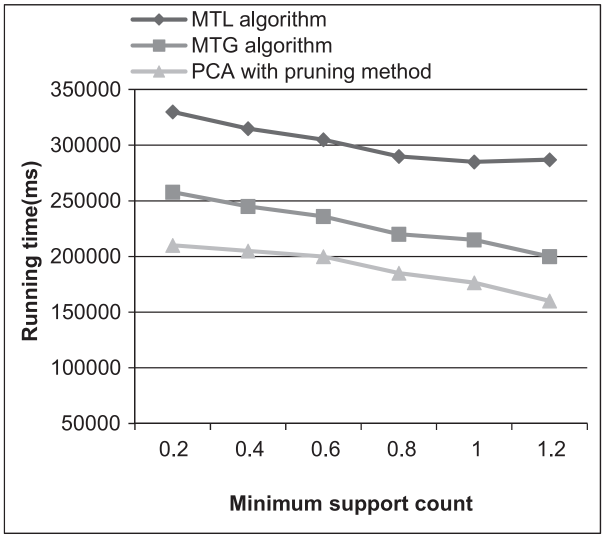

Figure 7: Running Time

total number of sub graph candidates, due to which the mechanism can analyze the pruning efficiency by having a check over the percentage of candidates that are pruned by the pruning procedure.

The graph from the Figure 7 above illustrates the comparison of running time for graph dataset. The minimum support count is plotted along the *x*-axis and the execution time is taken along the *y*-axis. The experimental results show that the available MTL and MTG technique have illustrated the higher running time complexity. The PCA with pruning technique show that it provides lower running time complexity results.

5. CONCLUSION

In this research work, PCA with pruning technique is introduced to improve the graph classification accuracy. The system proposed is different from the available MTG technique. The already available system does not achieve the best sub graph classification. In order to improve the sub graph classification performance, the novel system anticipated PCA method and pruning techniques. This novel system comprises of three phases like feature selection making use of PCA, sub graph feature selection making use of pruning technique and multi task graph classification procedure. The PCA algorithm proposed is focused towards selecting the most relevant and optimal sub graph features by the covariance matrix computation. It is utilized for increasing the maximum classification accuracy among the number of training samples. Thereafter, the pruning technique is employed for reducing the unwanted sub graph features from the multi task graph dataset. It is also aimed at selecting the more informative sub graph features that satisfies the minimum support threshold value and it is most frequent features in the given graph dataset. The prepruning and post pruning approach ignores the duplicate graphs and it also reduces the complexity of running time through the computation of the FS search. By integrating multiple tasks to direct the sub graph feature exploration and the successive learning process, the process of multi-task graph classification has obvious benefits in exploring good sub graph features and preventing over fitting, rather than the models learned from single tasks only. This approach guarantees that including the sub graph features can lead to reduced loss inregularization, which again results in optimal learning models. The PCA proposed and pruning technique yields a higher performance with respect to greater accuracy, recall, *f*-measure, AUC and lower time complexity values compared to the earlier research. Nonetheless this approach failed to manage large multi dimensional dataset with more efficiency.

REFERENCES

[1] M. Deshpande, M. Kuramochi, N. Wale and G. Karypis, Frequent substructure-based approaches for classifying chemical compounds, *IEEE Transactions on Knowledge and Data Engineering, 17(8),* 2005, 1036-1050.

[2] N. Wale and G. Karypis, Acyclic Subgraph-based Descriptor Spaces for Chemical Compound Retrieval and Classification, In *Proc of IEEE International Conference on Data Mining* (ICDM), 2006.

[3] H.D.K. Moonesinghe, H. Valizadegan, S. Fodeh and P.N. Tan, A probabilistic substructure-based approach for graph classification, In *19th IEEE International Conference on Tools with Artificial Intelligence, 1*, 2007, 346-349.

[4] M. Deshpande, M. Kuramochi, N. Wale and G. Karypis, Frequent substructure-based approaches for classifying chemical compounds, *IEEE Transactions on Knowledge and Data Engineering*, *17(8),* 2005, 1036-1050.

[5] H. Saigo, S. Nowozin, T. Kadowaki, T. Kudo and K. Tsuda, gBoost: a mathematical programming approach to graph classification and regression, *Machine Learning, 75(1),* 2009, 69-89.

[6] J. Zhou, J. Chen and J. Ye, *Malsar: Multi-task learning via structural regularization* (Arizona State University, 2011).

[7] X. Chen, W. Pan, J.T. Kwok and J.G. Carbonell, Accelerated gradient method for multi-task sparse learning problem. In *2009 Ninth IEEE International Conference on Data Mining*, 2009, 746-751.

[8] H. Fei and J. Huan, Structured feature selection and task relationship inference for multi-task learning. *Knowledge and information systems*, *35(2),* 2013, 345-364.

[9] Y. Zhang, D.Y. Yeung and Q. Xu, Probabilistic multi-task feature selection. In *Advances in neural information processing systems*, *2010, 2559-2567.*

[10] S. Pan, J. Wu and X. Zhu, Cogboost: boosting for fast cost-sensitive graph classification, *IEEE Transactions on Knowledge and Data Engineering*, *27(11),* 2015, 2933-2946.

[11] J.T. Vogelstein, W.G. Roncal, R.J. Vogelstein and C.E. Priebe, Graph classification using signal-subgraphs: Applications in statistical connectomics, *IEEE transactions on pattern analysis and machine intelligence*, *35(7)*, 2013, 1539-1551.

[12] Pan, S., & Zhu, X. (2013, August). Graph classification with imbalanced class distributions and noise, In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence* (pp. 1586-1592). AAAI Press.

[13] A. Pnueli, *Near-optimal supervised feature selection among frequent subgraphs*, 2009.

[14] X. Kong and P.S. Yu, Semi-supervised feature selection for graph classification, In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010, 793-802.

[15] S. Godbole and S. Sarawagi, Discriminative methods for multi-labeled classification, In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2004, 22-30.

[16] "Challenges and Surveys in Key Management and Authentication Scheme for Wireless Sensor Networks" in Abstract of Emerging Trends in Scientific Research 2014-2015.

[17] http://econpapers.repec.org/article/pkpabetsr/Impact Factor: 0.119

[18] "Biologically Inspired Intelligent Robots Using Artificial Muscles" ,International Journal of pharma and bio sciences, Impact Factor = 5.121(scopus indexed)

[19] X. Shi, X. Kong and S.Y. Philip, *Transfer Significant*, 2012.

[20] J. Zhou, J. Chen and J. Ye, *Malsar: Multi-task learning via structural regularization.* Arizona State University, 2011.

[21] S. Pyatykh, J. Hesser and L. Zheng, Image noise level estimation by principal component analysis. *IEEE transactions on image processing*, *22(2),* 2013, 687-699.

[22] A. Pnueli, *Near-optimal supervised feature selection among frequent subgraphs*, 2009.

[23] T. Kudo, E. Maeda and Y. Matsumoto, An application of boosting to graph classification. In *Advances in neural information processing systems*, 2004, 729-736.