A Heuristic Approach for Minimum Set Cover Problem

Fatema Akhter
IEEE student member
Department of Computer Science and Engineering
Jatiya Kabi Kazi Nazrul Islam University Trishal,
Mymensingh-2220, Bangladesh

***Abstract*—The Minimum Set Cover Problem has many practical applications in various research areas. This problem belongs to the class of NP-hard theoretical problems. Several approximation algorithms have been proposed to find approximate solutions to this problem and research is still going on to optimize the solution. This paper studies the existing algorithms of minimum set cover problem and proposes a heuristic approach to solve the problem using modified hill climbing algorithm. The effectiveness of the approach is tested on set cover problem instances from OR-Library. The experimental results show the effectiveness of our proposed approach.**

***Keywords*—Set Cover; Greedy Algorithm; LP Rounding Algorithm; Hill Climbing Method**

I. Introduction

For a given set system on a universe of items and a collection of a set of items, Minimum Set Cover Problem (MSCP) [1] finds the minimum number of sets that covers the whole universe. This is a NP hard problem proven by Karp in [2]. The optimization has numerous applications in different areas of studies and industrial applications [3]. The applications include multiple sequence alignments for computational biochemistry, manufacturing, network security, service planning and location problems [4]–[7].

Several heuristics and approximation algorithms have been proposed in solving the MSCP [8]. Guanghui Lan et al. proposed a Meta-RaPS (Meta-heuristic for Randomized Priority Search) [9]. Fabrizio Grandoni et al. proposed an algorithm based on the interleaving of standard greedy algorithm that selects the min-cost set which covers at least one uncovered element [10]. Amol Deshpande et al. [11] proposed an Adaptive Dual Greedy which is a generalization of Hochbaums [12] primal-dual algorithm for the classical Set Cover Problem.

This paper studies some popular existing algorithms of MSCP and proposes a heuristic approach to solve MSCP using modified hill climbing method. Within our knowledge, the same approach for MSCP of this paper has not been yet reported. Although this work implements two popular algorithms, Greedy Minimum Set Cover [14] and Linear Polynomial Rounding (LP) algorithm [15] to find solutions to MSCP, this work does not focus on the strength and weakness of the algorithms. The proposed approach starts with an initial solution from Greedy approach and LP rounding and then the result is optimized using modified hill climbing technique. The computational results shows the effectiveness of the proposed approach.

The rest of the paper is organized as follows: Section II describes the preliminary studies for proposed approach. Section III describes the proposed algorithm for MSCP. Section IV presents the experimental results. Section V provides the conclusion and future work.

II. Background Theory and Study

This section briefly describes MSCP and presents some preliminary studies. This includes Greedy Algorithm, LP Rounding Algorithm, Hill Climbing Algorithm and OR Library of SCP instances.

A. Minimum Set Cover Problem

Given a set of n elements $U = [e_1, e_2, ..., e_n]$ and a collection $S = \{S_1, S_2, ..., S_m\}$ of m nonempty subsets of U where $\bigcup_{i=1}^{m} S_i = U$. Every S_i is associated with a positive cost $c(S_i) \geq 0$. The objective is to find a subset $X \subseteq S$ such that $\sum_{S_i \in X} c(S_i)$ is minimized with respect to $\bigcup_{S \in X} S = U$.

B. Minimum k-Set Cover Problem

An MSCP (U, S, c) is a k-set cover problem [13] if, for some constant k, it holds that $|S_i| \leq k$, $\forall S_i \in S$ represented as (U, S, c, k). For an optimization problem, x^{OPT} presents an optimal solution of the problem where $OPT = f(x^{OPT})$. For a feasible solution x, the ratio $\frac{f(x)}{OPT}$ is regarded as its approximation ratio. If the approximation ratio of a feasible solution is upper-bounded by some value k, that is $1 \leq \frac{f(x)}{OPT} \leq k$, the solution is called an k-approximate solution.

C. Greedy Minimum Set Cover Algorithm

Data: Set system $(U, S), c : S \rightarrow Z+$

Input: Element set $U = [e1, e2, ..., en]$, subset set $S = \{S1, S2, ..., Sm\}$ and cost function $c : S \rightarrow Z+$

Output: Set cover X with minimum cost

Algorithm 1 Greedy MSCP

```
procedure GREEDY(U, S, c)   ▷ Set system {U,S} and cost function, c (S)
    X ← φ
    while Σ_{i∈X} X_i ≠ U do   ▷ Continue until X = U
        Calculate cost effectiveness, α = c(S)/|S−X| for every unpicked set {S_1, S_2, ..., S_m}
        Pick a set S, with minimum cost effectiveness
        X ← X ∪ S
    end while
    return X   ▷ Output X, minimum number of subsets
end procedure
```

D. LP Rounding Algorithm

The LP formulation [15] of MSCP can be represented as

$$\begin{aligned} &\textit{Minimize:} \\ &\sum_{i=1}^{m} c_i \times X_i \\ &\textit{Subject To:} \\ &\sum_{i:e\in S_i} X_i \geq 1 \qquad \forall e \in U \\ &X_i \leq 1 \qquad \forall e \in \{1,2,...,m\} \\ &X_i \geq 0 \qquad \forall e \in \{1,2,...,m\} \end{aligned} \tag{1}$$

Algorithm 2 LP Rounding MSCP

```
procedure LPROUND(U, S, c)
    Get an optimal solution x* by solving the linear program for MSCP defined in Equation 1.
    X ← φ
    for each S_j do   ▷ Continue for all members of S
        if x*_j ≥ 1/f then
            X ← X ∪ S_j
        end if
    end for
    return X   ▷ The minimum number of sets
end procedure
```

E. Hill Climbing Algorithm

Hill climbing [16] is a mathematical optimization technique which belongs to the family of local search. It is an iterative algorithm that starts with an arbitrary solution to a problem, then attempts to find a better solution by incrementally changing a single element of the solution. If the change produces a better solution, an incremental change is made to the new solution, repeating until no further improvements can be found.

F. OR Library

OR-Library [17] is a collection of test data sets for a variety of Operations Research (OR) problems. OR-Library was originally described in [17]. There are currently 87 data files for SCP. The format is

Algorithm 3 Hill Climbing Algorithm

1: Pick a random point in the search space.
2: Consider all the neighbors of the current state.
3: Choose the neighbor with the best quality and move to that state.
4: Repeat 2 through 4 until all the neighboring states are of lower quality.
5: Return the current state as the solution state.

a) number of rows (m), number of columns (n)
b) the cost of each column $c(j), j = 1, 2, ..., n$

For each row $i(i = 1, ..., m)$: the number of columns which cover row i followed by a list of the columns which cover row i.

III. Proposed Algorithm

This work modified the conventional hill climbing algorithm for set cover problem. To avoid the local maxima problem, this work introduced random re-initialization. For comparisons, greedy algorithm and LP rounding algorithm are used to find the initial state for the modified hill climbing algorithm. The evaluation function for the modified hill climbing algorithm is described below.

A. Problem Formulation

- **Input:** $N = |U|$, $U = [e1, e2, ..., en]$, $M = |S|$, $S = \{S1, S2, ..., Sm\}$, $c = \{c_1, c_2, ..., c_m\}$
- **Output:**
  1) Minimum number of sets, $n(X) = |X|$.
  2) List of minimum number of Sets, $X = \{X_1, X_2, ..., X_{n(X)}\}$.
- **Constraint:** Universality of X must hold, that is $\sum_{i\in X} X_i = U$.
- **Objective:**
  1) Minimize the number of sets, X.
  2) Minimize the total cost, $c(X)$.

B. OR Library MSCP Formulation

The formulation of MSCP for OR Library is given below.

1) Let $M^{m\times n}$ be a $0/1$ matrix, $\forall a_{ij} \in M_{ij}, a_{ij} = 1$ if element i is covered by set j and $a_{ij} = 0$ otherwise.
2) Let $X = \{x_1, x_2, ..., x_n\}$ where $x_i = 1$ if set i with cost $c_i \geq 0$ is part of the solution and x_i = 0 otherwise.

$$\begin{aligned} &\textit{Minimize:} \\ &\sum_{i=1}^{n} x_i \times c(x_i) \\ &\textit{Subject To:} \\ &1 \leq \sum_{i=1}^{n} x_i \times a_{ij} \quad j \in \{1,2,...,m\} \\ &x_i \geq 0 \qquad \forall x_i \in \{0,1\} \end{aligned} \tag{2}$$

C. Proposed Algorithm

This section describes our proposed algorithm for MSCP. The algorithm finds an initial solution and then optimizes the result using modified hill climbing algorithm.

IV. Experimental Results and Discussions

This section presents the computational results of the proposed approach. The effectiveness of the proposed approach is tested on 20 SCP test instances obtained from Beasley's OR Library. These instances are divided into 11 sets as in Table I, in which Density is the percentage of nonzero entries in the SCP matrix. All of these test instances are publicly available via electronic mail from OR Library.

The approach presented in this paper is coded using C on an Intel laptop with speed of 2.13 GHz and 2GB of RAM under *Windows 7* using the codeblock,version-13.12 compiler. Note here that this study presented here did not apply any kind of preprocessing on the instance sets received from OR-Library. This paper did not report the CPU times or running time of the algorithm as they vary machine to machine and compiler to compiler.

TABLE I: Test instance details

Problem Set	Number of instances	Number of rows(m)	Number of columns(n)	Range of cost	Density %
4	10	200	1000	1-100	2%
5	10	200	2000	1-100	2%
6	5	200	1000	1-100	5%
A	5	300	3000	1-100	2%
B	5	300	3000	1-100	5%
C	5	400	4000	1-100	2%
D	5	400	4000	1-100	5%
NRE	5	500	5000	1-100	10%
NRF	5	500	5000	1-100	20%
NRG	5	1000	10000	1-100	2%
NRH	5	1000	10000	1-100	5%

A. Experimental Results of Weighted SCP

Table II presents the experimental results for the proposed approach for weighted SCP instances. The first column represents the name of each instance. The optimal or best-known solution of each instance is given in the 2nd column. The 3rd and 4th column represent the solution found using greedy and LP rounding approach. The 5th and 6th column represent the solutions found in [5] and [7]. The last two columns contain the result found using proposed approach, started from greedy approach and LP rounding approach respectively.

B. Experimental Results of Unweighted SCP

Table III presents the experimental results of the proposed approach for unweighted SCP instances. This paper used the same 20 instances of weighted SCP and made them unweighted by replacing the weights to 1 on these instances. The first column represents the name of each instance. The optimal or best-known solution of each instance is given in the 2nd column. The 3rd and 4th column represent the solution found using greedy and LP rounding approach. The 5th and 6th column represent the solutions found in [18] and [19]. The last two columns contain the result found using proposed approach, started from greedy approach and LP rounding algorithm respectively.

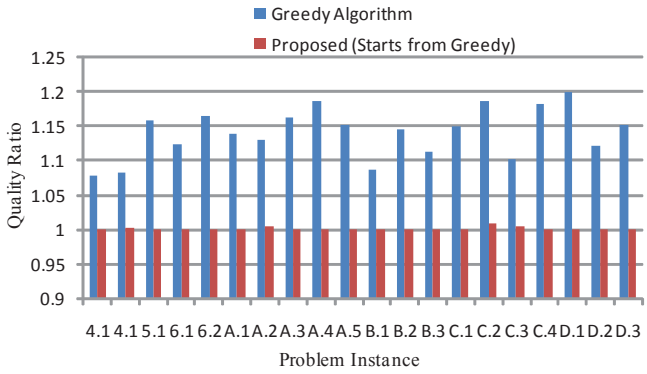

Fig. 1: Quality ratio of weighted problem instances for Greedy and Proposed Algorithm.

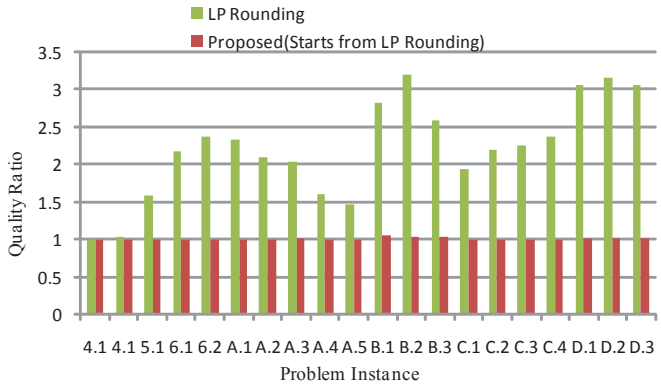

Fig. 2: Quality ratio of weighted problem instances for LP Rounding and Proposed Algorithm.

C. Result Summary

Summary: The optimal solution presented in Table II and Table III are taken from [7]. The quality of a solution derived by an algorithm is measured by *Quality Ratio* which is defined as a ratio of the *derived solution* to the *optimal solution*. The *quality ratio* for each instance for conventional greedy algorithm, LP rounding and Proposed algorithms, presented in this work are shown in Fig. 1, 2, 3 and 4. The figures show the ratio values, plotted as histogram for every instance, presented in this work.

$$Quality\ Measure\ Ratio = \frac{Derived\ Solution}{Optimal\ Solution} \qquad (3)$$

Another popular quality measurement reported in literature is called *GAP* which is defined as the percentage of the *deviation of a solution* from the *optimal solution* or *best known solution*. The summarized results, in terms of average quality and average GAP, for weighted set covering instances are presented in Table IV. For unweighted set covering instances it is represneted in Table V.

Algorithm 4 Proposed Algorithm

1: Preparation: In this step, elements of Universal set U, subsets of sets S and cost c of each set are taken as inputs.
2: Initial Solution: This step finds a solution X using *Greedy method* and *LP Rounding algorithm* of MSCP. X is considered the initial state for hill climbing optimization step. This study uses both the solutions and further optimizes for comparisons.
3: Hill Climbing Optimization: This Phase uses modified hill climbing algorithm and optimizes the cost of set cover problem.
4: Find the cost $c(X)$ from X. ▷ Initial best found cost, $c(X)$
5: Keep this (X) as the best found sets. ▷ Initial best found set, X
6: Calculate $R = n(S) \times n(U)$ ▷ number of elements, $n(U) = |U|$, number of subsets, $n(S) = |S|$
7: **for** M times of R **do** ▷ Here M is the *Set Minimization Repetition Factor*
8: Randomly select a set X^* from the selected sets. ▷ Random selection of a candidate redundant set
9: Mark this set X^* as *Unselected Set*.
10: **if** $X - X^* = U$ **then** ▷ Check whether the universality constraint holds
11: Stay with this state and find the cost, C_{new}.
12: Replace the best found cost C, with the current cost, C_{new}.
13: Remove set X^* from the selected sets, X.
14: Go back to step 8
15: **end if**
16: **for** K times **do** ▷ Here K is the Hill Climbing Repetition Factor
17: Randomly select a set Y from the unselected sets, $S - X$
18: Mark this set as *Selected*.
19: **if** $(X - X^*) \cup Y \neq U$ **then** ▷ Check whether the universality constraint holds
20: Go back to step 17
21: Find cost $C_{new}\ of\ c((X - X^*) \cup Y)$
22: **if** $C_{new} \leq C$ **then**
23: Replace the best found cost C, with the current cost, C_{new}.
24: Enlist Y in the *Selected Sets*.
25: Go back to step 17
26: **end if**
27: **end if**
28: **end for**
29: **end for**
30: Return best found list of sets X and minimum number of sets $n(X)$.

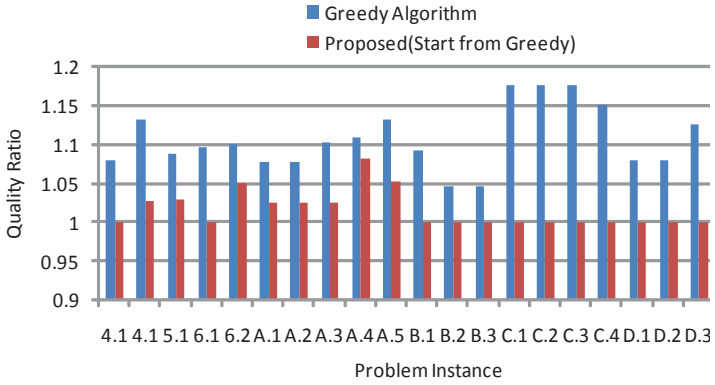

Fig. 3: Quality ratio of unweighted problem instances for Greedy and Proposed Algorithm.

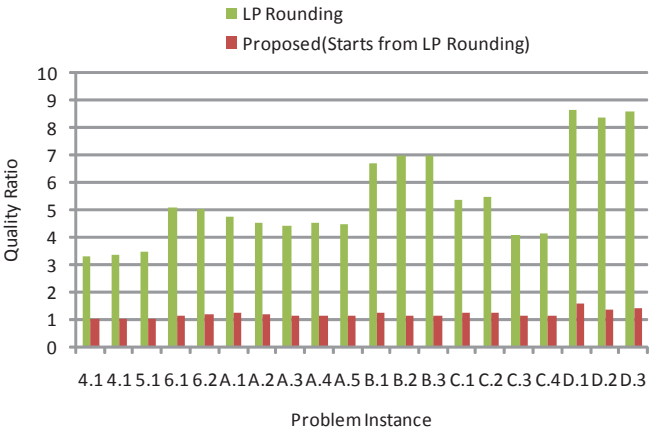

Fig. 4: Quality ratio of unweighted problem instances for LP Rounding and Proposed Algorithm.

$$GAP = \frac{Derived\ Solution - Optimal\ Solution}{Optimal\ Solution} \times 100\% \quad (4)$$

The proposed algorithm presented in this paper used conventional greedy algorithm and LP-Rounding Algorithm as initial solution. Then with the modified hill climbing method, these results are further optimized. Table IV and Table V compare the proposed heuristic approach to the original greedy approach and LP Rounding algorithm.

In Table IV, the average quality ratio and average GAP of original greedy are 1.14 and 14.10 respectively for weighted SCP while for proposed approach they are 1.00 and 0.09.

TABLE II: Experimental Results for Weighted SCP

Instance number	Optimal Solution	Greedy Algorithm	LP Rounding	[5] (Meta-RaPS)	[7] (Descent Heuristic)	Proposed Algorithm	
						Start from Greedy	Start from LP Rounding %
4.1	429	463	429	429	433	429	429
4.10	514	556	539	514	519	515	514
5.1	253	293	405	253	265	253	255
6.1	138	155	301	138	149	138	138
6.2	146	170	347	146	156	146	147
A.1	253	288	592	253	258	253	255
A.2	252	285	531	252	262	253	253
A.3	232	270	473	232	243	232	235
A.4	234	278	375	234	240	234	234
A.5	236	272	349	236	240	236	236
B.1	69	75	196	69	72	69	73
B.2	76	87	243	76	79	76	79
B.3	80	89	207	80	84	80	84
C.1	227	261	442	227	237	227	229
C.2	219	260	484	219	230	221	221
C.3	243	268	551	243	249	244	245
C.4	219	259	523	219	229	219	221
D.1	60	72	184	60	64	60	61
D.2	66	74	209	66	68	66	68
D.3	72	83	221	72	77	72	74

TABLE III: Experimental Results for Unweighted SCP

Instance number	Optimal Solution	Greedy Algorithm	LP Rounding	[18] (Tabu Search)	[19] Local Search for SCP	Proposed Algorithm	
						Start from Greedy	Start from LP Rounding %
4.1	38	41	125	38	38	38	38
4.10	38	43	127	38	38	39	39
5.1	34	37	117	35	34	35	34
6.1	21	23	107	21	21	21	23
6.2	20	22	101	21	20	21	23
A.1	39	42	186	39	39	40	47
A.2	39	42	176	39	39	40	46
A.3	39	43	172	39	39	40	44
A.4	37	41	167	38	37	40	41
A.5	38	43	170	38	38	40	43
B.1	22	24	147	22	22	22	27
B.2	22	23	154	22	22	22	25
B.3	22	23	154	22	22	22	25
C.1	40	47	214	43	43	40	49
C.2	40	47	220	44	43	40	49
C.3	40	47	163	43	43	40	45
C.4	40	46	165	43	43	40	45
D.1	25	27	216	25	25	25	39
D.2	25	27	209	25	25	25	34
D.3	24	27	206	25	24	24	33

TABLE IV: Average quality ratio and GAP for the Weighted Set Covering Problem

Algorithm	Average Quality Ratio	Average GAP
Greedy Algorithm	1.14	14.10
Proposed (greedy initial solution)	1.00	0.09
LP Rounding	2.22	122.57
Proposed (LP initial solution)	1.01	1.48

TABLE V: Average quality ratio and GAP for the Unweighted Set Covering Problem

Algorithm	Average Quality Ratio	Average GAP
Greedy Algorithm	1.11	10.66
Proposed (greedy initial solution)	1.02	1.58
LP Rounding	5.41	441.06
Proposed (LP initial solution)	1.18	17.6

The average quality ratio and average GAP of LP rounding are 2.22 and 122.57 respectively for weighted SCP while for proposed approach they are 1.01 and 1.48. It is clearly visible that original greedy and LP Rounding are deviated from the optimal solution by a high degree where proposed approach hardly deviates from the optimal solution.

In Table V, the average quality ratio and average GAP of original greedy are 1.11 and 10.66 respectively for unweighted SCP while for proposed approach they are 1.02 and 1.58. The average quality ratio and average GAP of LP rounding are 5.41 and 441.06 respectively for unweighted SCP while for proposed approach they are 1.18 and 17.6. It is clearly visible that original greedy and LP Rounding are highly deviated from the optimal solution where proposed approach hardly deviates from the optimal solution.

V. Conclusion and Future Work

This paper studies the existing approaches of MSCP and proposes a new heuristic approach for solving it. Appropriate theorems and algorithms are presented to clarify the proposed approach. The experimental results are compared with the existing results available in literature which shows the effectiveness of the proposed approach. This approach is tested only on OR-Libray in this work. In future this approach will be

tested on some other libraries of SCP like *Airline and bus scheduling problems* and *Railway scheduling problems*. The proposed algorithm can also be tested in another popular NP hard problem called *Vertex Cover Problem*.

ACKNOWLEDGEMENT

The author would like to express her greatest gratitude to the anonymous reviewers for their constructive feedback and critical suggestions that helped significantly to elicit the utmost technical attribute of this research work.

REFERENCES

[1] G. Gens and E. Levner, *"Complexity of approximation algorithms for combinatorial problems: a survey,"* ACM SIGACT News, vol. 12, no. 3, pp. 52-65, Fall 1980.

[2] R. M. Karp, *"Reducibility among combinatorial problems,"* Springer US, pp. 85-103, March 1972.

[3] T. M. Chan, E. Grant, J. Knemann and M. Sharpe, *"Weighted capacitated, priority, and geometric set cover via improved quasi-uniform sampling,"* Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms, pp. 1576-1585, SIAM, January 2012.

[4] F. C. Gomes, C. N. Meneses, P. M. Pardalos and G. V. R. Viana, *"Experimental analysis of approximation algorithms for the vertex cover and set covering problems,"* Computers and Operations Research, vol. 33, no. 12, pp. 3520-3534, December 2006.

[5] G. Lan, G. W. DePuy and G. E. Whitehouse, *"An effctive and simple heuristic for the set covering problem,"* European Journal of Operational Research, vol. 176, no. 3, pp. 1387-1403, February 2007.

[6] L. Ruan, H. Du, X. Jia, W. Wu, Y. Li and K.-I Ko *"A greedy approximation for minimum connected dominating sets,"* Theoretical Computer Science, vol. 329, no. 1, pp. 325-330, December 2004.

[7] N. Bilal, P. Galinier and F. Guibault, *"A New Formulation of the Set Covering Problem for Metaheuristic Approaches,"* International Scholarly Research Notices, pp.1-10, April 2013.

[8] Y. Emek and A. Rosen, *"Semi-streaming set cover,"* Automata, Languages, and Programming, pp. 453-464, Springer Berlin Heidelberg, July 2014.

[9] G. Lan, G. W. DePuy and G. E. Whitehouse, *"An effective and simple heuristic for the set covering problem,"* European journal of operational research, vol. 176, no. 3, pp. 1387-1403, February 2007.

[10] F. Grandoni, A.Gupta, , S. Leonardi, P. Miettinen, P. Sankowski and M. Singh, *"Set covering with our eyes closed,"* SIAM Journal on Computing, vol. 42, no. 3, pp. 808-830, May 2013.

[11] A. Deshpande, L. Hellerstein and D. Kletenik, *"Approximation algorithms for stochastic boolean function evaluation and stochastic submodular set cover,"* Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1453-1467, SIAM, January 2014.

[12] F. A. Chudak, M. X. Goemans, D. S. Hochbaum and D. P. Williamson, *"A primaldual interpretation of two 2-approximation algorithms for the feedback vertex set problem in undirected graphs,"* Operations Research Letters, vol. 22, no. 4, pp. 111-118, May 1998.

[13] F. Colombo, R. Cordone and G. Lulli, *"A variable neighborhood search algorithm for the multimode set covering problem,"* Journal of Global Optimization, pp. 1-20, August 2013.

[14] V. Chvatal, *"A greedy heuristic for the set-covering problem,"* Mathematics of operations research, vol. 4, no. 3, pp. 233-235, August 1979.

[15] B. Saha and S. Khuller, *"Set cover revisited: Hypergraph cover with hard capacities,"* Automata, Languages, and Programming, pp. 762-773, Springer Berlin Heidelberg, July 2012.

[16] K. J. Lang, *"Hill climbing beats genetic search on a boolean circuit synthesis problem of koza's,"* Proceedings of the Twelfth International Conference on Machine Learning, pp. 340-343, June 2014.

[17] J. E. Beasley, *"OR-library: distributing test problems by electronic mail,"* Journal of the Operational Research Society, vol. 41, no. 11, pp. 1069-1072, November 1990.

[18] G. Kinney, J. W. Barnes and B. Colleti, *"A group theoretic tabu search algorithm for set covering problems,"* Working paper, online available at http://www.me.utexas.edu/barnes/research/, 2004.

[19] N. Musliu, *"Local search algorithm for unicost set covering problem,"* Springer Berlin Heidelberg, pp. 302-311, June 2006.

[20] B. Yelbay, S. I. Birbil and K. Bulbul, *"The set covering problem revisited: an empirical study of the value of dual information,"* European Journal of Operational Research, September 2012.