Asian Journal of Engineering and Technology Innovation 03 (06) 2015 (65-69)

Asian Journal of Engineering and Technology Innovation

ISSN : 2347-7385

REVIEW ARTICLE

Received on: 22-05-2015
Accepted on: 10-06-2015
Published on: 22-06-2015

Madhugandha Bhosale
madhugandha.bhosale@gmail.com

QR Code for Mobile users

Conflict of Interest: None Declared

Minimum Consistent Subset Cover Problem : A Minimization view Of Data Mining

Madhugandha Bhosale

Abstract: In this paper, we tend to introduce and study the minimum consistent set cowl (MCSC) drawback. Given a finite ground set and a constraint t, notice the minimum variety of consistent subsets that cowl X, wherever a set of X is consistent if it satisfies t. The MCSC drawback generalizes the normal set covering drawback and has minimum pack partition (MCP),a twin drawback of graph coloring, as Associate in Nursing instance. several common data processing tasks in rule learning, clustering, and pattern mining are often developed as MCSC instances. Especially, we tend to discuss the minimum rule set (MRS) drawback that minimizes model complexness of call rules, the converse k-clustering drawback that minimizes the quantity of clusters, and also the pattern report drawback that minimizes the number of patterns. For any of those MCSC instances, our planned generic formula CAG are often directly applicable. CAG starts by constructing a greatest optimum partial answer, then performs Associate in Nursing example-driven specific-to-general search on a dynamically maintained bipartite assignment graph to at the same time learn a collection of consistent subsets with little cardinality covering the bottom set.

Keywords: Minimum consistent set cowl, set covering, graph coloring, minimum pack partition, minimum star partition, minimum rule set, converse k-clustering, pattern report

Cite this article as:
Madhugandha Bhosale, Minimum Consistent Subset Cover Problem :A Minimization view Of Data Mining Asian Journal of Engineering and Technology Innovation 03 (06); 2015; 65-69.

INTRODUCTION

IN this paper, we tend to introduce and study the minimum consistent set cowl (MCSC) drawback that finds several applications in common data processing tasks, providing a minimization read of knowledge mining. Given a finite ground set X and a constraint t, the MCSC drawback finds the minimum number of consistent subsets that called X, wherever a set of X is consistent if it satisfies t.

The MCSC drawback provides a way of generalizing the traditional set covering drawback [15], wherever a set of X is consistent if it's a given set. completely different from set covering, in typical MCSC instances the consistent subsets are not expressly given and that they got to be generated. For example, minimum pack partition (MCP), a twin drawback of graph coloring, are often thought of as Associate in Nursing MCSC instance, where a set is consistent if it forms a pack and also the cliques don't seem to be given as input.

Scope& Importance

Data mining applications.- several common data processing tasks are often developed as MCSC instances. As an utilization of the MCSC drawback in rule learning, the minimum rule set (MRS) drawback finds a complete and consistent set of rules with the minimum cardinality for a given set of tagged examples. The completeness and consistency constraints need correct classifications of all the given examples. With the goal of minimizing model complexness, the MRS drawbacks are often motivated from each information classification and information description applications. The MRS drawback may be a typical MCSC instance, wherever a set is consistent if it forms an identical rule, i.e., the bounding box of the set contains no examples of alternative categories.

As a distinguished bunch model, k-clustering generates k clusters minimizing some objective, like most radius as within the k-center drawback or most diameter as within the pair wise bunch drawback [4], [12]. The radius of a cluster is that the most distance between the centre of mass and the points within the cluster. The diameter is that the most distance between any 2 points within the cluster. Since the number of clusters is commonly laborious to work out before hand, converse k-clustering are often a a lot of applicable bunch model, wherever a most radius or diameter threshold is given and also the variety of clusters k is to be decreased . The converse k-center and converse pair wise bunch issues area unit both MCSC instances, wherever a set is consistent if it forms a cluster satisfying a given distance constraint. Frequent pattern mining has been a trademark of knowledge mining. whereas mining potency has been greatly improved over the years, interpretability instead became a bottleneck to its winning application. As a noted drawback, the irresistibly large number of generated frequent patterns containing redundant info area unit if truth be told “inaccessible knowledge” that must be any mined and explored. Thus, report of enormous collections of patterns within the pursuit of usability has emerged as a vital analysis problem. The converse k-clustering models mentioned on top of as well as another MCSC formulations seem to be a reasonable and promising approach towards this drawback. The goal of knowledge mining is to extract attention-grabbing patterns [11]. Knowledge should be concise and ideally human-comprehensible, providing a generalization of knowledge. The deserves of minimalist (detailed in Section 3) of classification models are well mentioned and with success used [43], [33]. Many common data processing tasks are often viewed as a decrease process. The MCSC drawback we tend to study formalizes such a decrease views. within the drawback, the constraint t isused to check the “consistency” of partial information, i.e., subsets of the ground set X. every qualified consistent set corresponds to a motivating pattern, i.e., a rule or a cluster of certain size. The goal is to attenuate the model complexness in terms of variety of patterns.

Related work

A generic formula-several sensible MCSC instances feature antimonotonic constraints, i.e., constraints with the downward closure property, underneath that any set of a consistent set is additionally consistent. Antimonotonicity are often used to gain potency in finding MCSC instances, similar to the cases of frequent pattern mining, consecutive pattern mining, and topological space bunch [20]. We style a generic formula CAG which will be wont to solve Associate in Nursing MCSC instance that exhibits an antimonotonic constraint. CAG starts by constructing a greatest optimum partial answer, so performs Associate in Nursing example-driven specific-to-general search on a dynamically maintained bipartite assignment graph to at the same time learn a tiny low consistent set cowl.

We conjointly extend the applicable territory of CAG by introducing pivot antimonotonicity that generalizes antimonotonicity. We use the separate converse k-center and also the star partition issues as examples to indicate however CAG will be tailored to unravel such MCSC instances .Set covering. the standard set covering drawback [15] finds the minimum variety of subsets from a given collection of subsets that cowl a given ground set. It is one

of the foremost basic algorithmic issues that has many variants, settings, and applications. the matter is NP-hard and there's no constant issue approximation. It is approximate inside one þ log n by an easy greedy algorithm [15], that iteratively selects the set that covers the biggest variety of uncovered parts till the ground set is roofed. This greedy algorithmic rule basically adopts a separate-and-conquer approach as seen in most rule learners.

In the MCSC drawback we have a tendency to study, the subsets aren't explicitly given. Instead, a constraint is given and accustomed qualify the subsets which will be utilized in a canopy. Graph coloring. Graph coloring is twin drawback of the MCP drawback, therefore graph coloring heuristics are often applied to complementary graphs to unravel MCP instances. The decision drawback of graph coloring, the k-coloring number drawback, is NP-complete for absolute k [17]. The problem is soluble in polynomial time just for k ¼ a pair of, and for absolute k on some special graphs [10]. Graph coloring approximations square measure surveyed in [20]. Algorithms to unravel the graph coloring drawback represent three categories: precise strategies, metaheuristics, and construction methods. precise approaches embody number linear programming and branch-and-bound. Metaheuristics begin with some construction technique to quickly acquire Associate in Nursing initial solution, that is any improved with metaheuristic techniques like random native search, tabu search, simulated hardening [16], or genetic algorithms [8]. These techniques perform otherwise on differing kinds of graphs.

Construction strategies square measure additional sensible in real applications involving giant (more than one,000 vertices) and dense graphs thanks to their potency. they typically build possible colorings in Associate in Nursing progressive method, beginning with Associate in Nursing empty assignment and iteratively coloring the vertices till all vertices square measure coloured. DSATUR [7] is one in all the foremost fashionable construction heuristics. In DSATUR, we have a tendency to square measure given a listing of various colours indexed from one to end. The vertices square measure 1st sorted in decreasing order of degree and a vertex with the biggest degree is appointed the colour with very cheap index. Then at each construction step, following vertex to be coloured is chosen according to the saturation degree, that is, the amount of different colours appointed to adjacent vertices. The vertex with the largest saturation degree is chosen and appointed the colour with very cheap doable index such the partial coloring remains possible. Ties square measure broken pro the vertex with the largest variety of unassigned adjacent vertices. If ties still remain, they're broken indiscriminately. Theoretical analysis shows that DSATUR is precise for bipartite graphs.

Rule learning- within the past few decades, various rule learners are projected in the main from the machine learning community. the continual development of the AQ family could replicate this endless effort. While the first AQ member dates back to the late sixties of last century, the newest AQ21 [25] was free shortly past.

Most of the prevailing rule learners follow a separate-and-conquer approach, that originated from AQ and still enjoys quality. The approach searches for a rule that covers a vicinity of the given (positive) examples, removes them, and recursively conquers the remaining examples by learning additional rules till no examples stay. The approach is additionally referred to as sequent covering, learning one rule at a time till all the positive examples square measure covered. Besides the AQ family, alternative representative separate-and-conquer rule learners embody CN2 [6], RIPPER [7], etc. Separate-and-conquer rule learning is surveyed in [9]. In most of the prevailing rule learning approaches, the minimality of rule sets has not been a “seriously implemented bias” [13]. The RAMP [4] rule generation system, however, generates lowest classification rules from categorical information. The primary goal of RAMP is to attempt for a lowest rule set that is complete and in keeping with the coaching information, utilizing a logic reduction methodology known as R-MINI[13]. this system was 1st developed for programmable logic array circuit reduction, and is taken into account joined of the best famed 2-level logic reduction techniques. RAMP and a few alternative fashionable rule learners square measure surveyed in [3].

Minimum Rule Set-The minimum rule set downside finds a disjunctive set of ifthen rules with the minimum cardinality that cowl a given set of labeled examples utterly and systematically. A rule covers associate example if the attribute values of the instance satisfy the conditions per the antecedent (if-part) of the rule.

Another in style live for tree quality is that the total number of tests (internal nodes), that corresponds to the total variety of conditions during a rule set. the 2 measures tend to comply with one another. within the rule set case, fewer number of rules sometimes cause fewer variety of conditions, as incontestible in our experimental study.

The problem of inducting optimum trees is extremely troublesome. The NP-hardness results for various objectives area unit shown in [14], [19]. optimum call tree learning addresses a partitioning downside whereas optimum rule learning addresses a covering downside, that is mostly tougher than its corresponding partitioning downside within the sense that it has a a lot of larger search house.

Most rule learners, e.g., the AQ family, CN2 and liquidator, also implicitly cut back the quality of rule sets so as to achieve sensible generalization accuracy [9]. However, as pointed out by Hong [13], the minimality of rule sets has not been a “seriously implemented bias,” that is additionally evident in our experimental comparison study.

Rule learning. within the past few decades, varied rule learners are projected primarily from the machine learning community. the continual development of the AQ family would replicate this endless effort. While the first AQ member dates back to the late sixties of last century, the newest AQ21 [25] was discharged shortly one.

Most of the present rule learners follow a separate- and – conquer approach, that originated from AQ and still enjoys quality. The approach searches for a rule that covers a vicinity of the given (positive) examples, removes them, and recursively conquers the remaining examples by learning a lot of rules till no examples stay. The approach is additionally referred to as serial covering, learning one rule at a time till all the positive examples area unit covered. Besides the AQ family, alternative representative separate -and-conquer rule learners embrace CN2 [6], RIPPER [7], etc. Separate-and-conquer rule learning is surveyed in [9]. In most of the present rule learning approaches, the minimality of rule sets has not been a "seriously implemented bias" [13]. The RAMP [3] rule generation system, however, generates bottom classification rules from categorical information. The primary goal of RAMP is to try for a bottom rule set that is complete and according to the coaching information, utilizing a logic reduction methodology referred to as R-MINI [13]. this system was 1st developed for programmable logic array circuit reduction, and is taken into account united of the best renowned 2-level logic reduction techniques. RAMP and a few alternative in style rule learners area unit surveyed in [2].

Data mining frameworks. data processing analysis has been concentrating on developing algorithms for individual problems. one amongst the most open challenges in data processing is the development of a unifying theory [26]. For this purpose, discusses theoretical frameworks for data mining and proposes many doable approaches: probabilistic, information compression, political economy, and inductive databases. we predict that owing to the various nature, data mining will in all probability be higher viewed from completely different angles and represented by multiple complementing frameworks.

Though not formalizing a framework, this paper makes a shot to supply a reduction read by proposing associate improvement downside that generalizes several common data processing tasks.

Merits of minimality. The notion that the accuracy of associate explanation is related to its simplicity dates back as early because the fourteenth century, once William of Ockham posited the medieval rule of parsimony that came to be referred to as Occam's Razor. The principle states that one must not increase, on the far side what's necessary, the quantity of entities required to clarify something. It may be even as follows: first, nature exhibits regularity and natural phenomena area unit more usually easy than advanced. At least, the phenomena humans value more highly to study tend to own easy explanations. Second, there area unit so much fewer easy hypotheses than complex ones, so there's solely a tiny low likelihood that any easy hypothesis that's wildly incorrect are going to be consistent with all observations.

Machine learning researchers have followed the principle to favor hypotheses with easy representations. For example, by the minimum description length principle [21], an operational formalization of Occam's Razor, the best hypothesis for a given set of knowledge is that the one that results in the largest compression of the information. The deserves of minimality of classification models are well mentioned and with success utilized [22], [18].

Frequent pattern mining has been studied extensively for various styles of patterns together with item sets, sequences, and graphs [11]. whereas nice progress has been created in terms of potency improvement, interpretability of the results has become a bottleneck to self-made application of pattern mining attributable to the massive variety of patterns generally generated. A closely connected downside is that there's lots of redundancy among the generated patterns. Explainable and representative report of enormous collections of patterns has emerged as a vital analysis direction and numerous approaches are planned together with mining top frequent patterns, closed frequent patterns, and high k frequent closed patterns [11]. Some recent work aims at finding a hard and fast variety of patterns representing the whole set of frequent patterns further as doable[1]. Similar to the situation for k-clustering, the acceptable number k of patterns to summarize the frequent pattern set is hard to specify. However, the users might have the domain knowledge that however similar a gaggle of patterns ought to be so that they'll be drawn as an entire while not losing much info. In lightweight of this, converse k-clustering models, like converse k-center and converse pairwise clustering, seem to be terribly cheap and promising to provide elliptic and informative pattern report.

Such bunch models generate clusters, i.e., groups of patterns, with sure quality guarantee. In such a formulation, the objective is to reduce the amount of pattern groups necessary to hide the complete assortment of frequent patterns, that is natural for the aim of report since it maximizes interpretability of the result representing the collection and at constant time reduces redundancy. For the radius or diameter threshold, customary distance functions can be utilized, like the Jaccard's constant for item sets or edit distance for successive patterns.

BEYOND ANTIMONOTONICITY

Limitation by antimonotonicity- kind of like several knowledge mining algorithms [11], CAG needs antimonotonicity to work. what quantity limitation would the antimonotonicity requirement cause to the pertinency of CAG? really, not as much in concert would speculate. The MCSC downside minimizes the amount of consistent subsets. cheap constraints on subsets ought to have the tendency that the larger the subsets, the tougher for them to satisfy the constraints. Otherwise, the step-down method would become trivial with the bottom set X forming a single consistent set. antimonotonicity is a lot of restrictive but in line with this tendency.

CONCLUSION

This paper makes the subsequent main contributions.

1) We introduce the minimum consistent set cowl downside that finds applications in several common data processing tasks.

2) we tend to study the theoretical properties of the MCSC problem, supported that we tend to gift a generic formula CAG that solves MCSC instances with antimonotonic and pivot antimonotonic constraints.

3) we tend to perform comprehensive experiments on benchmark knowledge sets compared with start-of-the-art algorithms, demonstrating the effectiveness and potency of CAG

REFERENCES

1. F. Afrati, A. Gionis, and H. Mannila, "Approximating a Collection of Frequent Sets," Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), 2004.
2. GAO ET AL.: THE MINIMUM CONSISTENT SUBSET COVER PROBLEM: A MINIMIZATION VIEW OF DATA MINING 701 TABLE 1 MCP Results Fig. 3. Separability of clusters.
3. C. Apte´ and S. Weiss, "Data Mining with Decision Trees and Decision Rules," Future Generation Computer Systems, vol. 13, nos. 2/3, pp. 197-210, 1997.
4. S.H.S.P.B.R. Apte, "C. RAMP: Rules Abstraction for Modeling and Prediction," technical report, IBM Research Division, T.J. Watson Research Center, 1995.
5. M. Bern and D. Eppstein, "Approximation Algorithms for Geometric Problems," Approximation Algorithms for NP-Hard Problems, D.S. Hochbaum, ed., PWS Publishing Co., 1997.
6. D. Brelaz, "New Methods to Color the Vertices of a Graph," Comm. ACM, vol. 22, no. 4, pp. 251-256, 1979.
7. P. Clark and T. Niblett, "The CN2 Induction Algorithm," Machine Learning, vol. 3, no. 4, pp. 261-283, 1989.
8. W.W. Cohen, "Fast Effective Rule Induction," Proc. 12th Int'l Conf. Machine Learning (ICML), 1995.
9. A.E. Eiben, J.K. Van Der Hauw, and J.I. Van Hemert, "Graph Coloring with Adaptive Evolutionary Algorithms," J. Heuristics, vol. 4, no. 1, pp. 25-46, 1998.
10. J. Fu¨ rnkranz, "Separate-and-Conquer Rule Learning," Artificial Intelligence Rev., vol. 13, no. 1, pp. 3-54, 1999.
11. M.R. Garey and D.S. Johnson, Computers and Intractability: A Guide to the Theory of NP-Completeness. W.H. Freeman & Co., 1979.
12. J. Han, Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc., 2005.
13. D.S. Hochbaum, "Various Notions of Approximations: Good, Better, Best, and More," Approximation Algorithms for NP-Hard Problems, PWS Publishing Co., 1997.
14. S.J. Hong, "R-MINI: An Iterative Approach for Generating Minimal Rules from Examples," IEEE Trans. Knowledge and Data Eng., vol. 9, no. 5, pp. 709-717, Sept./Oct. 1997.
15. L. Hyafil and R. Rivest, "Constructing Optimal Binary Decision Trees is NP-Complete," Information Processing Letters, vol. 5, no. 1, pp. 15-17, 1976.
16. D.S. Johnson, "Approximation Algorithms for Combinatorial Problems," J. Computer and Systems Science, vol. 9, no. 3, pp. 256- 278, 1974.
17. D.S. Johnson, C.R. Aragon, L.A. McGeoch, and C. Schevon, "Optimization by Simulated Annealing: An Experimental Evaluation. Part i, Graph Partitioning," Operation Research, vol. 37, no. 6, pp. 865-892, 1989.
18. R. Karp, "Reducibility among Combinatorial Problems," Complexity of Computer Computations, R. Miller and J. Thatcher, eds., Plenum Press. 1972.
19. M. Mehta, J. Rissanen, and R. Agrawal, "MDL-Based Decision Tree Pruning," Proc. First Int'l Conf. Knowledge Discovery and Data Mining (KDD), 1995.
20. G. Naumov, "Np-Completeness of Problems of Construction of Optimal Decision Trees," Soviet Physics, vol. 36, no. 4, pp. 270-271, 1991.
21. V. Paschos, "Polynomial Approximation and Graph-Coloring," Computing, vol. 70, no. 1, pp. 41-86, 2003.
22. J. Rissanen, "Modelling by Shortest Data Description," Automatica, vol. 14, pp. 465-471, 1978.
23. J. Rissanen, Stochastic Complexity in Statistical Inquiry Theory. World Scientific Publishing Co., Inc., 1989.
24.] C. Toregas, R. SWain, C. Revelle, and L. Bergman, "The Location of Emergency Service Facilities," Operations Research, vol. 19, pp. 1363-1373, 1971.
25. J. Wojtusiak, R. Michalski, K. Kaufman, and J. Pietrzykowski, "Multitype Pattern Discovery via AQ21: A Brief Description of the Method and its Novel Features," Technical Report MLI 06-2, Machine Learning and Inference Laboratory, George Mason Univ., 2006.
26. Q. Yang and X. Wu, "10 Challenging Problems in Data Mining vol. 5, no. 4, pp. 597-604, 2006.
27. "Minimum Consistent Subset Cover Problem:A Minimization View Of Data Mining" Byron J. Gao, Martin Ester, Member, IEEE Computer Society, Hui Xiong, Senior Member, IEEE, Jin-Yi Cai, and Oliver Schulte