Kernelised Parameter Generation for Solving Minimum Consistent Subset Cover Problem in Data Mining

MadhugandhaNivruttiBhosale

Aurangabad, Maharashtra

Abstract - In this paper, we tend to introduce and study the minimum consistent set cover (MCSC) drawback. The MCSC drawback generalizes the normal set covering drawback and has minimum clique partition (MCP), a twin drawback of graph coloring, as an 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 summarization 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 and at the same time learn a collection of consistent subsets with little cardinality covering the bottom set. In basic MCSC CAG algorithm is used to get the partial answer. It takes more computational time. So that over all execution time increases. Sothat there is kernelized parameter generation for purpose of optimize CAG. So that computational time reduced drastically. Optimize CAG algorithm linearly increases computational time.

Keywords: Antimonotonic constraints, Converse k-clustering, Graph coloring, Minimum clique partition, Minimum rule set, Minimum consistent set cover, Set covering

1.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 view of knowledge /data 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 is 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 clique partition (MCP), a dual drawback of graph coloring, are often thought of as associate in nursing MCSC instance, where a subset is consistent if it forms a clique and also the cliques don't seem to be given as input. In this we use optimize CAG algorithm. In which kernelised parameter generated for purpose of optimize CAG.

2.Scope& Importance

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 examples. The completeness and consistency constraints need to 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. Here, the data classification and data description are mentioned. Where data description means given data set is described by some attributes. And data classification means data classify by category, group, behavior or some different things.

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 two points within the cluster. Since the number of clusters is commonly laborious to work out before hand, converse k-clustering are often a lot of applicable models, wherever a most radius or diameter threshold is given and also the variety of clusters k is to be decreases. The converse k-center and converse pair wise clustering issues area 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 information 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 of classification models are well mentioned and with success used. 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 issued 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.

In MCSC problem, constraint t is used to test the consistency of subsets of ground set X. The constraint't' is worked as a function may call data selection function. It can be said as threshold. Each qualified consistent subset corresponds to an interesting pattern that is rule or cluster of certain size. The main goal of MCSC is to minimize model complexity in terms of number of patterns.

In basic MCSC problem we use two basic algorithms. CAG algorithm and RGB algorithm. The CAG algorithm takes more time for execution .It means CAG time exponentially increases. To overcome from this problem we use optimize CAG algorithm which computational time reduced drastically. And Optimize CAG time linearly increases.

3.Material & Methods

Set covering Problem

The traditional set covering problem finds [16] finds the minimum number of subset from a given collection of subsets that cover a given ground set. In this method greedy approach is used and essentially divide & conquer method is adopt. It has many variants, settings, and applications. The problem is NP-hard and there is no constant factor approximation. In MCSC method external subset is not given explicitly. Instead, a constraint is given and used to qualify the subsets that can be used in a cover. NP-ion hard means no constant factor approximation. NP-hard problem means a lot of times you can solve problem by reducing it to a different form of problem.

The set cover problem is a classic question in combinatoric , computer science and complexity theory. It is one of Karp's 21 NP =complete problems .The decision version of set covering is NP-Complete and optimization /search version of set cover is NP-hard.

If each set is assigned a cost, it becomes a weighted set cover problem.

The difference between MCSC instance and set covering problem is that in MCSC, subsets are not explicitly given. Instead, a constraint is given and used to qualify the subsets that can be used in a set cover problem.

The set covering problem can be considered as an MCSC instance where a subset is consistent if it is given. Consistent subsets are not explicitly given in typical MCSC instances. If we take a preprocessing step to generate all the consistent subsets then a MCSC instance becomes a set covering problem. Since the generated collection of subsets would be prohibitively large and this is not a feasible approach to solve MCSC instances.

Graph coloring-

Graph coloring is dual problem of Minimum Clique Partition (MCP) problem. The decision problem of graph coloring, the k-coloring number problem is NP-complete for arbitrary k.

In graph theory, graph coloring is a special case of graph labeling .It is an assignment of labels traditionally called "colors" to elements of a graph subject to certain constraints in its simplest form. It is a way of coloring the vertices of a graph such that no two adjacent vertices share same color; this is called a vertex coloring. Similarly, an edge coloring assigns a color to each edge so that no two adjacent edges share the same color, and a face coloring of planar

graph assigns a color to each face or region so that no two faces that share a boundary have the same color.
Algorithms to solve graph coloring fall into three categories.

1. Exact method
2. Metaheuristics
3. Construction Methods.

An exact approach includes integer linear programming. Metaheuristic method start with some construction method quickly obtain an initial soiclution which is further improved with metaheuristic techniques such as stochastic local search, tabu search, simulated annealing or genetic algorithms .And these techniques perform differently on different types of graphs. Construction methods are more and practical in real applications involving large vertices and dense graphs due to their efficiency. Generally build feasible coloring in an incremental way, starting with an empty assignment and iteratively coloring the vertices until all vertices are colored. DSATUR is most popular construction method.

Rule Learning

Mostly rule learners proposed from machine learning community. Most of rule learners follows divide-and –conquer method, which originated from AQ family. At a time single rule learn.
Drawback of data mining and machine learning programs is as they visualize simple representation languages, e.g., decision trees, Bayesian nets, neural nets, etc., which their capabilities limits in the range of patterns they can discover. As a result, such programs may not be able to discover simple patterns and also not easily represented by the program.

The methodology has been implemented in AQ21, a program that performs natural induction, by which we mean an inference process that strives to generate accurate inductive hypotheses that are represented in human-oriented forms resembling natural language descriptions, and are by that easy to interpret and understand.

The AQ21 program integrates several new features either non-existent or present only individually and in a more limited form in other programs. An important feature of AQ21 is that it can discover different types of regularities in data, depending on its parameter setting, such a conjunctive patterns, general rules with exceptions, consistent and complete data characterizations, and optimized alternative hypotheses.

In this sequential rule learning is also proposed as they learn single rule at a time. In this methodology part of example is taken and apply rule until given example is cover and get positive answer. Then apply this process on remaining example. This process is called sequential covering. Learn single rule at a time until all positive examples are covered.

4.Data mining framework

Mostly data mining research developing algorithms for individual problems. Main challenge in data mining is development of a unifying theory. To solve this drawback different possible proposed approaches are discussed [30]. Possible proposed approaches are like probabilistic, data compression, microeconomic and inductive databases**.**

5.Antimonotonic constraints

These constraintshaving no same purpose. Antimonotocity can be used to gain efficiency in solving MCSC instances. Antimonotonic constraints works exactly opposite to closure property or downward closure property. Downward closure property means operations are done on the set like multiplication, division, etc. The answer will not be from same set .It is frequent.

6.Beyond Antimonoticity

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.

7. Two Basic Algorithms Used

1] RGB Algorithm-

-These are rule governed behavioral algorithms.
-These are mostly focus on rectangle based & graph based rule.
-Due to Optimal CAG algorithm the runtime of RGB is less.
Focusing on rectangle based rule
1. First calls CAG algorithm and stores set of learned rectangle rules in R.
2. Then general to specific beam search is performed to remove redundant conditions.

Properties-

1. RGB does not follow a separate and conquer approach, instead all rules are learned simultaneously.
2. RGB starts with subset containing a maximal number of examples such that no pair of them can coexist in a single consistent rule, which constitutes a maximal optimal partial solution.
3. RGB naturally learns a set of rules for all classes simultaneously.
4. Unlike many existing methods that can apply learn either perfect modes such as early members of family, or approximate models such as CN2 and RIPPER,RGB has flexibility to learn both without resorting to post-processing.

2] Optimal Generic CAG Algorithm-

1. Each vertex in Independent set forms a singleton consistent subset, represented by so called "condensed vertex" in an assignment graph. The rest of vertices in graph are called "element vertex".
2. Element vertices are processed in sequential manner and assigned to condensed vertices.

-With growth of condensed vertices, some element vertices would get isolated and new condensed vertices have to be created for them.

-Upon completion, graph becomes edge free with all vertices assigned and the set of condensed vertices, each representing consistent subset, constitutes a solution for given set.

3. Most seen separate-and-conquer approach, for example the greedy algorithm for set covering and most existing rule learners, optimal CAG algorithm adopts an example-driven strategy to learn consistent subset simultaneously.

8.Evaluation Measures

1] Which element vertex is the next to be processed? Which condensed vertex should it be assigned to?

In principle, the element vertex with least degree should be considered first since it is the one most likely to get isolated.

Least degree first criterion can be used to choose the next element vertex to be processed.

2] How Optimal CAG works?

Basic CAG algorithm starts by constructing a maximal optimal partial solution and then performs an example-driven specific–to-general search on dynamically maintained bipartite assignment graph to simultaneously learn small consistent subset cover. In it kernelised parameter generate to less some steps in basic CAG to get fast partial solution. This is called optimal CAG algorithm. Kernelised parameter generation for purpose of optimize CAG. So computation time reduce drastically.CAG algorithm required time for execution exponentially increases. Optimal CAG algorithm required time for execution linearly increases. Kernelised stands for designing efficient algorithms that achieve efficiency.

9.Result

When we compare CAG and Optimal CAG, we get required time difference very large. As CAG execution time exponentially increases and Optimal CAG execution time linearly increases due to kernelised parameter generation. So overall time required for execution decreases drastically as RGB call Optimal CAG get partial solution for further process or execution.

10. Conclusion

This paper makes the subsequent main contributions.

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

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

3) We also tend to study properties of RGB algorithm and Optimal CAG algorithm.

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.GAO ET AL.: The Minimum Consistent Subset Cover Problem: A Minimization View of Data Mining 701 table 1 MCP Results

[2] 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.

[3] S.H.S.P.B.R. Apte, "C. RAMP: Rules Abstraction for Modeling and Prediction," technical report, IBM Research Division, T.J. WatsonResearch Center, 1995.

[4] M. Bern and D. Eppstein, "Approximation Algorithms for Geometric Problems," Approximation Algorithms for NP-Hard Problems, D.S. Hochbaum, ed., PWS Publishing Co., 1997.

[5] D. Brelaz, "New Methods to Color the Vertices of a Graph,"Comm. ACM, vol. 22, no. 4, pp. 251-

256, 1979.
[6] P. Clark and T. Niblett, "The CN2 Induction Algorithm," MachineLearning, vol. 3, no. 4, pp. 261-283, 1989.
[7] W.W. Cohen, "Fast Effective Rule Induction," Proc. 12th Int'l Conf.Machine Learning (ICML), 1995.
[8] 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.
[9] J. Fu¨ rnkranz, "Separate-and-Conquer Rule Learning," ArtificialIntelligence Rev., vol. 13, no. 1, pp. 3-54, 1999.
[10] M.R. Garey and D.S. Johnson, Computers and Intractability: A Guide to the Theory of NP-Completeness. W.H. Freeman & Co., 1979.
[11] J. Han, Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc., 2005.
[12] D.S. Hochbaum, "Various Notions of Approximations: Good, Better, Best, and More," Approximation Algorithms for NP-Hard Problems, PWS Publishing Co., 1997.
[13] S.J. Hong, "R-MINI: An Iterative Approach for Generating Minimal Rules from Examples," IEEE Trans. Knowledge and DataEng., vol. 9, no. 5, pp. 709-717, Sept./Oct. 1997.
[14] L. Hyafil and R. Rivest, "Constructing Optimal Binary Decision Trees is NP-Complete," Information Processing Letters, vol. 5, no. 1, pp. 15-17, 1976.
[15] D.S. Johnson, "Approximation Algorithms for Combinatorial Problems," J. Computer and Systems Science, vol. 9, no. 3, pp. 256- 278, 1974.
[16] 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.
[17] R. Karp, "Reducibility among Combinatorial Problems," Complexity of Computer Computations, R. Miller and J. Thatcher, eds., Plenum Press. 1972.
[18] M. Mehta, J. Rissanen, and R. Agrawal, "MDL-Based Decision Tree Pruning," Proc. First Int'l Conf. Knowledge Discovery and Data Mining (KDD), 1995.
[19] G. Naumov, "Np-Completeness of Problems of Construction of Optimal Decision Trees," Soviet Physics, vol. 36, no. 4, pp. 270-271, 1991.
[20] 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 MasonUniv., 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
[28] http;//en.m.wikipedia.org/wiki/set cover problem
[29]http://en.m.wikipedia.org/wiki/graph_coloring_problem
[30]] H. Mannila, "Theoretical Frameworks for Data Mining," SIGKDD Explorations, vol. 1, no. 2, pp. 30-32, 2000.