Generalized Class Cover Problem with Axis-Parallel Strips

Apurva Mudgal* and Supantha Pandit

Department of Computer Science and Engineering
Indian Institute of Technology Ropar, Nangal Road, Rupnagar, Punjab-140001
{apurva,supanthap}@iitrpr.ac.in

Abstract. We initiate the study of a generalization of the class cover problem [1,2], the generalized class cover problem, where we are allowed to misclassify some points provided we pay an associated positive penalty for every misclassified point. We study five different variants of generalized class cover problem with axis-parallel strips and half-strips in the plane, thus extending similar work by Bereg et al. [2] on the class cover problem. For each of these variants, we either show that they are in P, or prove that they are NP-complete and give constant factor approximation algorithms.

Keywords: class cover problem, axis-parallel strips, approximation algorithms, geometric set cover.

1 Introduction

The class cover problem [1,2] is the following: given a set \( R \) of red points, a set \( B \) of blue points, and a set \( O \) of geometric objects, find a minimum cardinality subset \( O' \subseteq O \) which covers every blue point, but does not cover any red point. If we identify the blue points as positive examples and the red points as negative examples, the set \( O' \) gives us a classifier of complexity \( |O'| \), since every point \( p \in R \cup B \) can be correctly classified as blue or red using the disjunction of \( |O'| \) queries of the form “Does \( p \) lie inside object \( o \in O' \)?”.

In this paper, we study the generalized class cover problem where the classifier \( O' \subseteq O \) is allowed to misclassify some blue points as red and some red points as blue. We measure the amount of misclassification by a penalty function \( \mathcal{P} : R \cup B \rightarrow \mathbb{R}^+ \) assigning positive penalties to every point, where the penalty of a point is understood to be the cost we pay for misclassifying it using our classifier i.e., for reporting a blue point as red and vice versa. The objective now is to minimize the complexity of the classifier (i.e., the cardinality of the set \( O' \subseteq O \)) plus the sum of penalties of every point incorrectly classified by \( O' \).

Allowing for misclassification in the class cover problem can be useful in several ways. First, a small fraction of the red and blue points in the data may be

* Partially supported by grant No. SB/FTP/ETA-434/2012 under DST-SERB Fast Track Scheme for Young Scientist.
“outliers” and allowing for misclassification can identify these points and lead to a classifier with much smaller complexity. Secondly, there may occur scenarios where no subset $O' \subseteq O$ can completely separate blue points from red points (one such case occurs when every object $o \in O$ contains a red point). In such cases, all classifiers are “approximate” and the generalized class cover problem gives us an approximate classifier which minimizes the sum of classifier complexity and penalty due to misclassification. Finally, not all data points may be equally important and the practitioner can fine-tune the point penalties to reflect the relative importance of a data point as a positive or negative example for the classifier.

In the following, we define $P = R \cup B$, $n = |B|$, $m = |R|$, and for any function $f : U \to \mathbb{R}$ and a subset $S \subseteq U$, we use $f(S)$ to denote the sum $\sum_{x \in S} f(x)$. We assume that no two points have the same $x$- or $y$-coordinates. We consider two versions of generalized class cover problem, single coverage and multiple coverage, which differ in the way the penalty of misclassified red points is counted:

1. **Generalized Class Cover (Single Coverage).** The penalty of each red point $r \in R$ covered by $O'$ is counted exactly once. The cost $c_1(O')$ is defined as $|O'| + \mathcal{P}(B') + \mathcal{P}(R')$, where $B' \subseteq B$ is the set of blue points that are not covered by objects in $O'$ and $R' \subseteq R$ is the set of red points that are covered by objects in $O'$.

2. **Generalized Class Cover (Multiple Coverage).** The penalty of a red point $r \in R$ is counted once for every object containing $r$ in $O'$. The cost $c_2(O')$ of $O'$ is defined as $|O'| + \mathcal{P}(B') + \sum_{r \in R} \mathcal{P}(r) \cdot m(r, O')$, where $B' \subseteq B$ is the set of blue points not covered by any object in $O'$, and $m(r, O')$ is the number of objects in $O'$ which contain point $r$.

Note that if the penalty of each red and blue point is infinity, both the single and multiple coverage versions of generalized class cover problem reduce to the class cover problem.

In general, the set $O$ of geometric objects may be triangles, circles, axis-parallel squares and rectangles, etc. [1][2], but in this paper we assume $O$ to consist of only axis-parallel strips and half-strips in the plane. A horizontal (resp. vertical) strip $(a, b)$ is the set of points $(x, y) \in \mathbb{R}^2$ satisfying the equation $a \leq y \leq b$ (resp. $a \leq x \leq b$). A half-strip $(a, b, c)$ extending to infinity in the southern direction is the set of points satisfying $a \leq x \leq b$ and $y \leq c$. Similarly, we can define half-strips extending to infinity in the northern, eastern, and western directions. We now define the set $O$ of geometric objects for six different generalized class cover problems with strips and half-strips:

1. **STRIP.** All vertical and horizontal strips.
2. **HS-1D.** All half-strips extending to infinity southwards.
3. **HS-2D-SAME.** All half-strips extending to infinity in two opposite directions.
4. **HS-2D-DIFF.** All half-strips extending to infinity in two mutually orthogonal directions.
5. **HS-3D.** All half-strips extending to infinity in three different directions.
6. **HS-4D.** All half-strips extending to infinity in four different directions.