

1. はじめに

d 次元ユークリッド空間 R^d 内の点集合 $S = A \cup B$ (A : 正例の集合, B : 負例の集合; $A \cap B = \emptyset$) において, 正例の集合の幾何学的分布と負例の集合の分離を行うことは, 機械学習やデータ解析の基盤技術である. 具体的には R^d を幾何学的な領域 X とその補集合 $R^d \setminus X$ に分割し, X を正例の集合の近似として示すものであり, $X \cap A$ は大きく, $X \cap B$ は小さい集合になることが望ましい.

さまざまな幾何学的集合 $X \subset R^d$ について研究がなされている. 例えば, サポートベクトルマシンでは X が半空間であり, R^d を X と $R^d \setminus X$ に分離する超平面を求める問題として定式化される. その他の幾何学的集合についても多く研究がなされており, データマイニングなどでの重要な要素技術となっている. [1]

本論文では正例と負例が混在するデータのセットカバー問題を取り上げ, 自然な最適化問題に対して焦点を当てる.

具体的には正例と負例の差を最大化することを考え, 多項式時間アルゴリズムが設計できる条件と定式化を考察する.

2. 関連研究

① GEOMETRIC SET COVER

SET COVER 問題とは, 全体集合 U とその部分集合族 $U_1, \dots, U_m$ が与えられた時, U の要素をすべて被覆するように部分集合族から最小個数の部分集合を選ぶ問題である. この幾何バージョンもまた古典的な組み合わせ最適化問題であり, データベースや VLSI 回路など, 多くの応用がある. この問題は実に多くの設定で研究がなされている.

一般の設定の集合被覆問題は NP 完全であることが知られている [2]. たいていの幾何的版本での設定でも, NP 困難であるか, 近似解を求めるのでさえ困難であることが知られている [3, 4, 5]. しかし, 幾何的な性質を利用し, 近似率を保証した効率的な解法も得られる. たとえば, Mustafa ら [6] は 3 次元空間において, 円板領域集合被覆問題に対する PTAS の存在を示している.

② RED-BLUE SET COVER (RBSC)

要素数の和が N となる赤色の集合 R と青色の集合 B , および要素数が n となる $R \cup B$ の部分集合族 S が与えられた時, RED-BLUE SET COVER 問題は, B の要素をすべてカバーし R の要素のカバー数を最小にするような S の部分集合を求める問題である. SET COVER 問題の自然な一般化

であり, A が空集合のとき, SET COVER 問題と一致する. Carr ら [7] によって提唱され, また, $P=NP$ でない限りは任意の定数 $c < \frac{1}{2}$ に対して, $2^{\log^{1-\delta} n}$ 倍に近似できないことが証明されている ($\delta = 1/\log^c \log n$).

幾何バージョンでは, Chan ら [8] が二次元平面において S を単一正方形集合に限定した場合でも NP 困難であること, また, PTAS が存在することを示した.

3. 定義

本論文ではセットカバー問題の自然な拡張として新たな定式化を導入し, 一次元の場合にはあらゆるインスタンスに対して効率的なアルゴリズムが存在することを示す. また, 問題が二次元の場合は NP 困難であることを示し, インスタンスを特別な場合に限定すれば多項式時間アルゴリズムが存在すること, およびその条件について詳説する.

問題 1 最大差セットカバー問題

インスタンス: R^d 上の正例点集合 A , 負例点集合 B および領域集合 X

タスク: 部分集合 $X' \subseteq X$ で, $|X' \cap A| - |X' \cap B|$ を最大にするものを見つける

4. 提案手法

$d=1$ のとき, 任意のインスタンスが DAG 上の最長経路問題に変換できることを示す.

アルゴリズム 1

入力: 一次元 ($d=1$) の最大差セットカバー問題のインスタンス

出力: インスタンスの最適解の要素のインデックス, および最適値

方法: x_i を右端座標の昇順に並び替える. インスタンスから重み付き有向グラフ $G(V, E)$ を構成する. ダミー頂点 v_0 と, $x_i \in X$ に対応する頂点 v_i を V に追加する. $i < j$ に対して $x_i \not\subseteq x_j$ なら, 辺 (v_i, v_j) を E に追加する. 辺の重み $w((v_i, v_j)) := |(x_i \cup x_j)/x_i|$ とする. 生成されたグラフ G 上で v_0 からの最長経路 P を求める. P 上の各頂点のインデックスと経路長をそれぞれもとの問題の最適解と最適値として出力する.

生成されたグラフ G は DAG になっており, DAG 上の最長経路問題は最短経路問題に帰着することで解くことができる. このアルゴリズムは, 辺の重み $w((v_i, v_j))$ が領域 $x_1, \dots, x_{i-1}$ に依存せず x_i と x_j のみ

¹ A VARIATION OF GEOMETRIC RED-BLUE SET COVER PROBLEM² Fujihara Naoki, Takeshi Tokuyama, Graduate School of Science and Technology, Kwansei Gakuin University

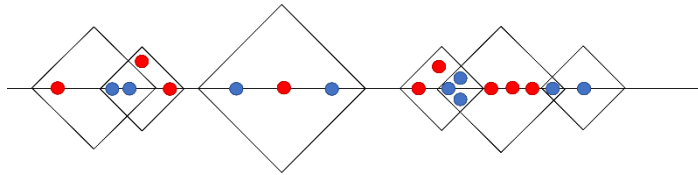
より一意に決定されることと、辺の重みの定義から任意の経路長が元の問題の対応するカバーリングと一致することから、正当性が保証される。またこのアルゴリズムの計算時間は $O(N|X| + |X|^2)$ である。

$d=2$ のとき、インスタンスで与えられる領域集合を単一正方形に限った場合でも NP 完全であることが、Planar 3-SAT からの帰着によって証明できるが、ここでは詳細は省略する。P=NP でない限りは多項式時間アルゴリズムが存在しないが、インスタンスを特別な場合に限定すれば問題を効率的に解くことができる。

5. 限定されたインスタンス

ただし、領域集合が特別な場合であれば次元と同様に重み付き有向グラフ上の最長経路問題に帰着することで問題を解くことができる。以下に、その一例となるインスタンスと、インスタンスが満たすべき条件を詳説する。

問題2 最大差セットカバー問題 (限定版 $d=2$)
インスタンス: R^2 上の正例点集合 $A = \{a_i \in R^2\}$, 負例点集合 $B = \{b_i \in R^2\}$, 水平直線 L および中心が L 上に乗る (L_1 ノルムでの) 円集合 $D = \{d_i \in R^2\}$
タスク: 部分集合 $D' \subseteq D$ で, $ D' \cap A - D' \cap B $ を最大にするものを見つける

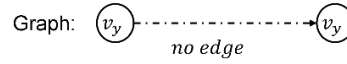
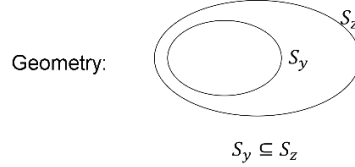
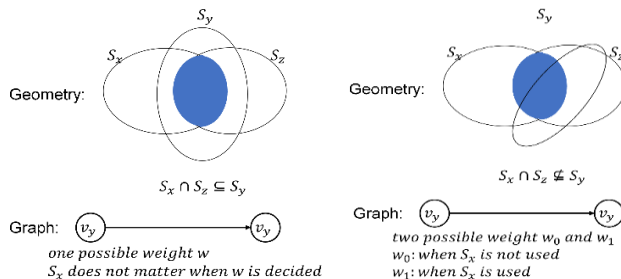


問題2の問題例。任意の問題例を DAG 上の最長経路問題に変換できる。

一般に、インスタンスからグラフを構成できる条件として、辺の重みが一意に決定されることがあげられる。これを達成するにはインスタンスの入力で与えられる領域集合をある順に並べたときに、次を満たす必要がある。

$$\forall i < j < k \ (x_i \cap x_k \subseteq x_j) \vee (x_j \subseteq x_k)$$

また、これが満たされているとき、任意のインスタンスがグラフに変換可能である。以下に参考の図を示す。



青色に示される領域 $S_x \cap S_z$ で, S_y に含まれない部分に正例点または負例点が存在するとき、グラフの辺 $w((v_y, v_z))$ の重みが一意に決定されない。すなわち, S_x がカバーに使用されるかどうか、グラフ上で言えば、経路に v_x を訪れたかどうかで $w((v_y, v_z))$ が変化してしまう。以上から、次の定理が言える。

定理1

最大差セットカバー問題のインスタンスで与えられる領域集合 X が $\forall i < j < k \ (x_i \cap x_k \subseteq x_j) \vee (x_j \subseteq x_k)$ を満たすとき, $O(N|X| + |X|^2)$ 時間で最大経路問題に帰着可能。

6. おわりに

正例と負例を持つセットカバー問題を考察し、その問題が NP 完全に属すること、一次元の場合には効率的なアルゴリズムが存在することを示し、二次元以上の場合についてはインスタンスを限定すれば効率的なアルゴリズムが存在することを示した。また、問題が多項式時間アルゴリズムを持つようなインスタンスの条件を示した。

7. 参考文献

- [1] 福田 剛志・森本 康彦・徳山 豪 著 (2001) 『データマイニング』 共立出版
- [2] Juris Hartmanis. Computers and intractability: a guide to the theory of npcompleteness (michael r. Garey and david s. Johnson). Siam Review, Vol. 24, No. 1, p. 90, 1982.
- [3] Piotr Berman and Bhaskar DasGupta. Complexities of efficient solutions of rectilinear polygon cover problems. Algorithmica, Vol. 17, No. 4, pp. 331-356, 1997
- [4] Nimrod Megiddo and Kenneth J Supowit. On the complexity of some common geometric location problems. SIAM journal on computing, Vol. 13, No. 1, pp. 182-196, 1984.
- [5] Nimrod Megiddo and Arie Tamir. On the complexity of locating linear facilities in the plane. Operations research letters, Vol. 1, No. 5, pp. 194-197, 1982
- [6] Nabil H Mustafa and Saurabh Ray. Improved results on geometric hitting set problems. Discrete & Computational Geometry, Vol. 44, No. 4, pp. 883-895, 2010.
- [7] R.D. Carr, S. Doddi, G. Konjevod, M. Marathe. On the red-blue set cover problem. Proc. ACM-SIAM Symposium on Discrete Algorithms (SODA) (2000), pp. 345-353
- [8] Timothy M. Chan, Nan Hu, Geometric red-blue set cover for unit squares and related problems, Computational Geometry, Volume 48, Issue 5, 2015, Pages 380-385