Tel Aviv University
The Raymond and Beverly Sackler Faculty of Exact Sciences
School of Mathematics
Department of Statistics and Operations Research

An Analytic Approach to Clustering Models with Outliers, and a Capacitated Vertex Cover Problem

A dissertation submitted for the degree of
Doctor of Philosophy

By
Einat Or

Submitted to the senate of Tel-Aviv University
August 2005

ACKNOLEDGMENTS

I wish to thank my supervisor, Prof. Refael Hassin. Rafi, I am grateful for your endless patience, for the professional instruction, and for being always, as you are, pleasant and kind.
I wish to thank my mother and my sister for pushing, and my father and my brother for not.
I wish to thank Daniel and Attay.

0.1 Abstract

This dissertation contains approximation algorithms for clustering problems with outliers, and the capacitated vertex cover problem.

Clustering problems are widely studied in operations research and computer science. Traditionally these problems are investigated under the assumption that all objects must be clustered. A significant shortcoming of this formulation is that a few very distant objects, called *outliers*, may exert a disproportionately strong influence over the solution. In this work we investigate clustering problem, while addressing outliers in a meaningful way. In statistics, *outliers* are defined as data objects which originated from a different probabilistic mechanism [3]. When clustering of data is considered, the intuitive definition of outliers becomes "points which do not belong to any of the clusters".

In the first chapter we present a generalization of the following clustering models on a graph, the k-CENTER, k-MIN-SUM, and k-MIN-MAX-DIAMETER, allowing nodess that are not clustered. A node either belongs to a cluster and influences the target according to the clustering model, or is not clustered and a penalty is paid for it. Formally, given a complete graph $G = (V, E)$, and a weight function $w : E \rightarrow I\!N_0$, on its edges, let $w(S) = \sum_{e=\{i,j\}\subset S} w_e$. The k-MIN-SUM problem is the problem of finding a partition of V to k sets, $\{S_1, \ldots, S_k\}$, minimizing $\sum_{i=1}^{k} w(S_i)$. Let $pV \rightarrow I\!N_0$ denote the penalty function on the nodes of G. For $S \subseteq V$ let $p(S) = \sum_{i\in S} p_i$. The PENALIZED k-MIN-SUM PROBLEM (we denote this problem by kPMS, and use PMS to denote 1PMS.) is the problem of finding a partition of V to $k+1$ sets, $\{S_1, \ldots, S_{k+1}\}$ minimizing $\cdot \sum_{i=1}^{k} w(S_i) + p(S_{k+1})$. The PENALIZED K-CENTER PROBLEM is the problem of choosing k nodes $\{s_1, \ldots, s_k\}$, and a set C of nodes, as to minimize $\max_{v\in C}\{\min_{i=1,\ldots,k}\{w(v, s_i)+p(V\setminus C)\}\}$. The PENALIZED k-MIN-MAX CLUSTERING PROBLEM (kPMD and 1PMD is denoted by PMD) is the problem of partitioning V into $k+1$ sets $\{S_1, \ldots, S_{k+1}\}$, as to minimize $\max_{1\le i\le k} \max_{u,v\in S_i}\{w(u, v)\} + p(S_{k+1})$.

For kPMS we present the following results: an efficient 2-approximation to PMS using a primal-dual algorithm. We prove that PMS is NP-hard even if w is a metric and present a randomized approximation scheme for the metric PMS where the ratio between the minimal and the maximal penalty is bounded, based on methods used to approximate MIN-BISECTION and 2-MIN-SUM [15, 16]. While the approach in [16] is a PTAS for metric PMS when

the cluster includes most of the nodes, it gives poor results if the cluster is smaller. The approach in [15] is the basis for a PTAS for metric PMS where the cluster and the set of non-clustered nodes are both large, but it gives poor approximation if one of the parts is small. Therefore we present a combination of the two approaches. For the metric kPMS we offer a 2-approximation by generalizing [24]. For the PENALIZED K-CENTER PROBLEM we present a 3-approximation, and for kPMD we show it is NP-hard even if $k = 1$ and w is a metric, we present an α-approximation algorithms for PMD, where $\alpha \leq 2$ is the approximation rate of the VERTEX COVER problem and prove it is best possible, and we present a 2-approximation for kPMD when k is fixed.

In the second chapter we model, and analyze the CLUSTER IDENTIFICATION OF MOLECULES (CIM), which is a clustering problem in a finite metric space. CIM has the following characteristics which separate it from other clustering models:
1. In most models outliers are a small portion of the data set, whereas in CIM they may be the vast majority of the objects. (see Figure 2.1)
2. The clusters identified by CIM are compact and their diameter is bounded.
3. There is a lower bound on the number of objects in a cluster.
4. Clusters may be very close to one another, as a result of the bound on the diameter. What may be considered as one cluster in other clustering models is considered as several clusters in CIM. (see Figure 2.2).
5. The number of clusters is not known a-priori to the clustering procedure.

We present CIM and model it as a MAXIMUM PROFIT COVERAGE PROBLEM (MPCP). The model is a measure to be optimized, rather then a heuristic.

Consider a finite set S in a metric space M with a distance function d. A ball with center t and radius r is the subset $B(t,r) = \{x \in M | d(t,x) \leq r\}$. We say that the ball *covers* the points of S that it contains. Given a set of balls $\mathcal{B}$ of radius r, a *coverage* $P = \{S'_1, \ldots, S'_l\}$ is a set of clusters such that each of them consists of points covered by a single ball of $\mathcal{B}$. Let $S'_P = \cup_{i=1}^{l} S'_i$, and define the *profit* of P as $\sum_{q \in S'_P} w_q - c|P|$, where c is the cost of a ball used by P, and w_q is a revenue obtained by covering $q \in S$. MPCP is the problem of finding a coverage with maximum profit.

We present two models for CIM, one as a PARAMETER ESTIMATION VIA LIKELIHOOD MAXIMIZATION and the other as MPCP. We introduce a polynomial time approximation scheme (PTAS) to MPCP in Euclidean space

using the shifting strategy [34, 21]. We present two practical heuristics for MPCP, one greedy and the other random, which introduce good results in numerical studies of CIM. We also study the problem of MPCP when the number of clusters is bounded using dynamic programming and the shifting strategy [34, 21].

In the third chapter we present the CAPACITATED VERTEX COVER PROBLEM. Let $G = (V, E)$ be an undirected graph with vertex set $V = \{1, \ldots, n\}$ and edge set E. Suppose that w_v denotes the weight of vertex v and k_v denotes the capacity of vertex v (we assume that k_v is an integer). A *capacitated vertex cover* is a function $x : V \to \mathbb{N}_0$ such that there exists an orientation of the edges of G in which the number of edges directed into vertex $v \in V$ is at most $k_v x_v$. (These edges are said to be *covered* by or *assigned* to v.) The *weight* of the cover is $\sum_{v \in V} x_v w_v$. The MINIMUM CAPACITATED VERTEX COVER problem is that of computing a minimum weight capacitated cover. The problem generalizes the MINIMUM WEIGHT VERTEX COVER problem which can be obtained by setting $k_v = |V| - 1$ for every $v \in V$. The main difference is that in vertex cover, by picking a node v in the cover we can cover all edges incident to v, in this problem we can only cover a subset of at most k_v edges incident to node v.

The main results that we show are as follows. We give a primal-dual algorithm that yields a factor 2 approximation for the basic problem. We also consider a generalization where each edge has a "demand" of d_e which has to be assigned to an adjacent vertex. For this generalization we show a factor 3 approximation. We also show how to relate this problem to work done in [31] on orientations of graphs. When the graph is a tree we show that the problem can be solved in polynomial time, but for the more general version with edge demands the problem is NP-hard.

אוניברסיטת תל-אביב
הפקולטה למדעים מדויקים על שם ריימונד ובברלי סאקלר
בית הספר למתמטיקה
המחלקה לסטטיסטיקה וחקר ביצועים

גישה אנליטית לבעיות של חלוקה לאשכולות עם נתוני שוליים ובעיית כיסוי צמתים עם קיבול

חיבור לשם קבלת התואר "דוקטור לפילוסופיה"

מאת

עינת אור

חיבור זה הוגש לסנאט של אוניברסיטת תל אביב

אוגוסט 2005

תקציר

מחקר זה מכיל אלגוריתמים לבעיות חלוקה לאשכולות עם נקודות מרוחקות ולבעיית כיסוי הצמתים עם קיבול.

בעיות של חלוקה לאשכולות נחקרו רבות בחקר ביצועים ובמדעי המחשב. באופן מסורתי בעיות אלה נחקרות תחת ההנחה שכל הנקודות צריכות להשתייך לאשכולות. חיסרון בולט של גישה זו הוא שמספר קטן של נקודות רחוקות, הקרויות רעש או שוליים, עלולות לגרום להשפעה לא פרופורציונאלית על הפיתרון. בעבודה זו אנו חוקרים בעיות של חלוקה לאשכולות תוך התייחסות בעלת משמעות לשוליים. בסטטיסטיקה שוליים מוגדרים כנתונים שיוצרו על ידי מכניזם הסתברותי השונה מהמכניזם ממנו נוצרו יתר הנתונים [3]. כאשר מדובר בחלוקה של נתונים לאשכולות ההגדרה הופכת להיות "נקודות שאינן שייכות לאף אחד מהאשכולות".

בפרק הראשון אנו מציגים הכללה של בעיות החלוקה לאשכולות על גרף, בעיית ה ***k-מרכז***, בעיית ה ***k-מינ-מקס*** ובעיית ה ***k-מינ-סכום*** המאפשרת צמתים שאינם שייכים לאף אחד מהאשכולות. צומת יכול או להשתייך לאשכול ולהשפיע על פונקציית המטרה בהתאם למודל, או לא להשתייך לאף אחד מהאשכולות וקנס ישולם עליו. באופן פורמאלי, בהינתן גרף $G=(V,E)$ ופונקציית משקל $w: E \to N_0$ על קשתות הגרף, יהי $w(S)=\sum_{e=(i,j)\subset S} w_e$. בעיית ה ***k-מינ-סכום*** היא בעיית המציאה של חלוקה של V ל-k קבוצות $\{S_1,...,S_k\}$ המביאה למינימום את $\sum_{i=1}^{k} w(S_i)$. תהי $p: V \to N_0$ פונקציית הקנס על צמתי הגרף G. עבור $S \subseteq V$ תהי $p(S)=\sum_{v\in S} p_v$. בעיית ה ***k-מינ-סכום עם קנסות*** (המסומנת ב kPMS כאשר 1PMS מסומנת ב PMS) היא הבעיה של חלוקת V ל-k+1 קבוצות $\{S_1,...,S_{k+1}\}$ המביאה למינימום את $\sum_{i=1}^{k} w(S_i)+p(S_{k+1})$. בעיית ה ***k-מרכז עם קנסות*** היא הבעיה של בחירת k צמתים $\{s_1,...,s_k\}$ וקבוצה C של צמתים על מנת להביא למינימום את $max_{v\in C} min_{i=1,...,k}\{w(v,s_i)+p(V\cap \bar{C})\}$. בעיית ה ***k-מינ-מקס עם קנסות*** (המסומנת ב kPMD כאשר 1PMD מסומנת ב PMD) היא הבעיה של חלוקת V ל-k+1 קבוצות $\{S_1,...,S_{k+1}\}$ המביאה למינימום את $max_{i=1,...,k} \max_{u,v\in S_i}\{w(v,u)\}+p(S_{k+1})$.

עבור kPMS אנו מציגים את התוצאות הבאות: קרוב שתיים יעיל ל PMS על ידי שימוש באלגוריתם פרימלי-דואלי. אנו מוכיחים ש PMS היא NP-קשה גם אם $w: E \to N_0$ הוא מטריקה ומציגים סכמת קרוב פולינומית רנדומאלית ל PMS מטרי כאשר היחס בין הקנס המקסימאלי לקנס המינימאלי חסום. הקרוב מתבסס על שיטות שהוצגו לקרוב בעיית מינימום חציה (BISECTION) ומינימום ***2-מינ-מקס*** [15],[16]. הגישה המוצגת ב [16] היא סכמת קרוב פולינומית ל PMS מטרי כאשר האשכול כולל את רוב הצמתים בגרף, הוא נותן תוצאות לא מספקות כאשר האשכול קטן. הגישה ב [15] היא הבסיס לאלגוריתם המאפשר סכמת קרוב פולינומית ל PMS מטרי במקרה בו האשכול וקבוצת הצמתים שאינם משתייכים לאשכול הן שתיהן גדולות, אך גישה זו אינה טובה אם אחת הקבוצות קטנה בהרבה מהאחרת. לכן אנו מציגים שילוב של שתי השיטות. ל kPMS המטרי אנו מציגים קרוב שתיים המכליל את [24]. עבור בעיית ***k-מרכז עם קנסות*** אנו מציגים קרוב שלוש ועבור kPMD אנו מראים שהיא NP-קשה אפילו אם k=1 ו $w: E \to N_0$ הוא מטרי. אנו מציגים קרוב α ל PMD כאשר α הוא יחס הקרוב הטוב ביותר לבעיית מינימום כיסוי צמתים, ומראים שקרוב זה הוא הטוב ביותר האפשרי. אנו מציגים קרוב שתיים ל kPMD מטרי כאשר k קבוע.

בפרק השני אנו ממדלים ומנתחים את בעיית ***זיהוי האשכולות של מולקולות קטנות*** (CIM)., שהיא בעיית חלוקה לאשכולות במרחב מטרי. ל CIM יש את המאפיינים הבאים המבדילים אותה ממודלים אחרים של חלוקה לאשכולות:

1. במודלים רבים השוליים הם חלק קטן מהנתונים בשעה שבמקרה זה הם יכולים להוות את הרוב המוחלט של הנתונים.
2. האשכולות הקיימים ב CIM הם קומפקטיים והקוטר שלהם חסום.
3. קיים חסם תחתון על מספר המולקולות באשכול.
4. אשכולות יכולים להיות קרובים מאוד זה לזה כתוצאה מהחסם על הקוטר. מה שעשוי להיחשב לאשכול יחיד בבעיות אחרות של חלוקה לאשכולות, יהיה במקרה זה מספר אשכולות.
5. מספר האשכולות אינו ידוע מראש.

אנו מציגים את CIM וממדלים אותה כבעיית ***כיסוי חלקי להשגת רווח מקסימאלי*** (MPCP). המודל הוא פונקציה שיש להביא לאופטימום ולא יוריסטיקה.

נתונה קבוצה סופית S במרחב מטרי M עם פונקצית מרחק d. כדור בעל מרכז t ורדיוס r הוא תת הקבוצה $B(t,r)=\{x\in M \mid d(t,x)\le r\}$. אנו אומרים שכדור מכסה את הנקודות מ S אותן הוא מכיל. בהינתן קבוצת כדורים $\boldsymbol{B}$ ברדיוס r, כיסוי חלקי $P=\{S'_1,...,S'_l\}$ הוא קבוצה של אשכולות כך שכל אשכול הוא קבוצת נקודות מ S המכוסה על ידי כדור יחיד מ $\boldsymbol{B}$. יהי $S'_P=\bigcup_{i=1}^{l} S'_i$ ונגדיר את הרווח מ P על ידי $\sum_{q\in S'_P} w_q + c\,|P|$ כאשר c היא עלות כדור בכיסוי החלקי ו w_q הוא ההכנסה מכיסוי הנקודה $q\in S$. MPCP היא הבעיה של מציאת כיסוי חלקי בעל רווח מקסימאלי.

אנו מציגים שני מודלים של CIM, האחד כ***הערכת פרמטרים של ערבוב גאוסיניים*** והשני ב MPCP. אנו מציגים סכמת קרוב פולינומית ל MPCP במרחב אוקלידי תוך שימוש באסטרטגית הזזה [34,21]. אנו מציגים שתי יוריסטיקות ישימות ל MPCP, אחת חמדנית והשנייה רנדומית, המציגות תוצאות טובות הניסויים מספריים של CIM. אנו מנתחים בנוסף את בעיית MPCP כאשר מספר האשכולות חסום ומציגים סכמת קרוב פולינומית עבור הבעיה במרחב אוקלידי תוך שימוש בתכנות דינאמי ובאסטרטגית הזזה [34,21].

בפרק השלישי אנו מציגים את בעיית ***מינימום כיסוי הצמתים עם קיבול***. יהי $G=(V,E)$ גרף לא מכוון עם קבוצת צמתים $V=\{1,...,n\}$ וקבוצת קשתות E. נניח ש w_v הוא המשקל של צומת v ו- k_v מסמן את הקיבול שלו (אנו מניחים ש k_v הוא שלם). ***כיסוי הצמתים עם קיבול*** הוא פונקציה $x:V\to N_0$ כך שקיימת הכוונה של הקשתות ב G שבה מספר הקשתות המכוונות לצומת $v\in V$ הוא לכל היותר $k_v x_v$. (קשתות אלה נקראות קשתות מושמות או מכוונות ל v). משקל הכיסוי הוא $\sum_{v\in V} w_v x_v$.

בעיית ***מינימום כיסוי הצמתים עם קיבול*** היא הבעיה של מציאת כיסוי הצמתים עם קיבול מינימאלי. ההבדל מבעית מינימום כיסוי הצמתים הקלאסית היא שבכיסוי צמתים בחירת צומת משמעות כיסוי כל הקשתות השכנות לצומת, בשעה שבמקרה של הבעיה עם קיבול חסום ניתן לכסות רק חלק של לכל היותר k_v מהקשתות השכנות.

התוצאות העיקריות שאנו מציגים הן התוצאות הבאות. אנו נותנים אלגוריתם פרימאלי-דואלי המבטיח קרוב שתיים לבעיה הבסיסית. אנו מציגים גם הכללה שבה לכל צומת יש "ביקוש" d_e שיש לכוון אל צומת שכן. להכללה זו אנו מראים אלגוריתם קרוב שלוש. אנו מראים כיצד לקשר בעיה זו לבעיה שהוצגה ב [31] על כיוון קשתות בגרף. כאשר הגרף הוא עץ אנו מראים שהבעיה הבסיסית פתירה בזמן פולינומי, אך למקרה הכללי יותר של "ביקוש" על הקשתות הבעיה היא NP-קשה.