An Implementation of Integer Programming Techniques in Clustering Algorithm

S. Shenbaga Ezhil[1] and Dr. C. Vijayalakshmi[2]
Department of Mathematics
Sathyabama University, Chennai – 119
e-mail: [1]shenbaga_ezhil@rediff.com, [2]vijusesha2002@yahoo.co.in

Abstract

This paper mainly deals with the analysis of IPP in clustering. Clustering is exemplified by the un-supervised learning of patterns and clusters that may exist in a given database and is a useful tool for Knowledge Discovery in Database (KDD). A mathematical programming formulation of this problem is proposed that is theoretically justifiable and computationally implementable in a finite number of steps. The clustering algorithm applies the hierarchical clustering methodology [1] where points or clusters with the shortest distance are merged into a cluster until the desired number of clusters is achieved. Numerical examples are given for the above algorithmic approach.

1.0 Introduction

The Mathematical programming model subject to some constraints is a broad discipline that has been applied for various theoretical and applied problems. In this paper the described various integer programming model for clustering is implemented.

The fundamental Non Linear programming problem, consists of minimizing the objective function subject to inequality and equality constraints and it is written as

$$\min_x g(x) \text{ subject to } f(x) \le 0,\ h(x) = 0$$

where x is an n-dimensional vector of real variables, f is a real-valued functions of x, f and h are finite dimensional vector functions of x. If all the functions f, g and h are linear then the problem is simplified to a linear program [8] which is the classical problem of mathematical programming.

The clustering problem considered in this paper is that of assigning m points in the n-dimensional real space R^n to k clusters. If a polyhedral distance (such as the 1-norm distance) is used, the model can be formulated as that of minimizing a piece-wise linear concave function. Although a bilinear program is a non convex optimization problem (i.e., minimizing a function that is not valley-like) a fast finite K-median

algorithm consisting of solving few linear programming in closed from leads to a stationary point. On four other publically available database of the K median and K mean algorithm did best on two of the database.

2.0 The Mixed Integer Optimization Model

The training data consists of n observations (x_i, y_i), i = 1, 2, …, n with $x_i \in R^d$ and $y_i \in (0, 1)$. Let $M_0 = \{1, 2, \ldots, m_0\}$, $M_1 = \{1, 2, \ldots, m_1\}$, $\overline{K} = \{1, 2, \ldots, K\}$. Let m_0 and m_1 be the number of class 0 and class 1 points respectively. Note that class 0 and class 1 points by x_i^0, i = 1, 2, …, m_0 and x_i^1, i = 1, 2, …, m_1 respectively.

Let G_K be the set indices of class 1 points that are in Group K, where $\bigcup_{K\in\overline{K}} G_K = M_1$ and $\bigcap_{K\in\overline{K}} G_K = \phi$. Thus the following system is infeasible, for all $i \in M_0$ and $K \in \overline{K}$

$$\left.\begin{aligned} \sum_{j\in G_K} \lambda_j x_j^1 = x_i^0, \\ \sum_{j\in G_K} \lambda_j = 1, \\ \lambda_j \geq 0, \quad j \in G_K. \end{aligned}\right\} \tag{1}$$

From Farka's lemma, system (1) is unfeasible if and only if the following problem is feasible.

$$\left.\begin{aligned} p^1 x_i^0 + q < 0 \\ p^1 x_j^1 + q \geq 0, \quad j \in G_K \end{aligned}\right\} \tag{2}$$

We consider the optimization problem

$$\left.\begin{aligned} &z = \text{maximize } \delta \\ &\text{subject to} \\ &P_{k,i}^1 x_i^0 + q_{k,i} \leq -\delta, \quad i \in M_0; K \in \overline{K} \\ &P_{k,i}^1 x_i^1 + q_{k,i} \geq 0, \quad i \in M_0; K \in \overline{K}, j \in G_K. \\ &\delta \leq 1 \end{aligned}\right\} \tag{3}$$

In order to determine if we can assign class 1 points into K groups such that z > 0, we define decision variable for $K \in \overline{K}$ and $j \in M_1$.

$$a_{k,j} = \begin{cases} 1, & y_j^1 \text{ is assigned to Group K} \\ 0 & \text{otherwise.} \end{cases}$$

We include the constraints $P_{k,i}^1 x_{j-1}^1 + q_{k,i} \geq 0$ in problems (3) if and only if $a_{k,j} = 1$.

i.e., $P_{k,i}^1 x_{j-1}^1 + q_{k,i} \geq M(a_{k,j-1})$

where M is the large positive constant. Now we can assume M = 1

i.e., $P_{k,i}^1 x_{j-1}^1 + q_{k,i} \geq a_{k,j-1}$.

Thus we can check whether we can partition class 1 points into K disjoint groups such that no class 0 points is assigned by solving by the mixed-integer programming.

3.0 Mathematical Formulation

z^* = maximize δ

subject to

$$\left.\begin{aligned} & P_{k,i}^1 x_i^0 + q_{k,i} \le -\delta, \quad i \in M_0; K \in \overline{K} \\ & P_{k,i}^1 x_j^1 + q_{k,i} \ge a_{k,j} - 1, \quad i \in M_0; K \in \overline{K}, j \in M_1 \\ & \sum_{k=1}^{K} a_{k,j} = 1, \quad j \in M, \\ & \delta \le 1 \\ & a_{k,j} \in \{0,1\} \end{aligned}\right\} \tag{4}$$

If $z^* > 0$, the partition into K groups is feasible, while if $z^* = 0$, it is not requiring us to increase the value of 0.

4.0 The Clustering Algorithm

A hierarchical clustering algorithm is developed based on preprocesses the data to create clusters of class 0 and class 1 points. Collection of class 0 (class 1) points are considered a cluster if there are no class 1 (class 0) points in their convex hull.

Equation (4) becomes expensive to solve. Alternatively, we can drastically decrease the dimension of problem (4) by solving hyperplane for clusters of points at a time instead of point by point.

The following linear optimization problem is solved to check whether class 1 clusters r and s can be merged.

$$\delta^* = \text{maximize } \delta \tag{5}$$

subject to

$$P_i^1 x_i^0 + q_i \le -\delta, \quad i \in M_0$$

$$P_i^1 x_j^1 + q_i \ge \delta, \quad j \in C_s \cup C_r.$$

where C_r and C_s are set of indices of class 1 points.

5.0 Clustering via Mathematical Programming

The unsupervised assignment of elements of a given set into groups or clusters of like points, is the objective of cluster analysis. A principal motivation behind the mathematical programming approach is a precise and concise statement of the clustering problems as a concave minimizing problems.

For a given set of m points in R^n represented by the matrix $A \in R^{m\times n}$ and a number K of desired clusters, we formulate the clustering problems as follows. Find cluster centers C_ℓ, $\ell = 1, \ldots, k$ such that the minima over $\ell \in \{1, 2, \ldots, k\}$ of the 1-norm distance between each point, A_i, i = 1, 2, …, m, and the cluster centers C_ℓ, $\ell = 1, \ldots, k$ is minimized so that the following integer programming problem can be solved.

$$\underset{C,D}{\text{minimize}} \sum_{i=1}^{m} \min_{\ell=1,\ldots k} \{e^T D_{i,\ell}\}$$

subject to

$$-D_{i,\ell} \le A_i^T - C_\ell \le D_{i,\ell}, \quad i = 1,2,\ldots,m, \ \ell = 1,2,\ldots,k \tag{6}$$

This is not the case of 2-norm or p-norm, $p \neq 1$. We state the bilinear programming formulation and K median algorithm for solving the clustering problem.

6.0 Clustering as a Bilinear Program

The clustering problem (6) is equivalent to the following bilinear program

$$\text{Minimize} \sum_{i=1}^{m}\sum_{\ell=1}^{K} e^T D_{i,\ell} T_{i,\ell}, \quad C_\ell \in R^n, D_{i,\ell} \in R^n, T_{i,\ell} \in R^n$$

subject to the constraints

$$-D_{i,\ell} \le A_i^T - C_\ell \le D_{i,\ell}, \quad i = 1,2,\ldots,m, \ \ell = 1,2,\ldots,k$$

$$\sum_{\ell=1}^{K} T_{i\ell} = 1, \quad T_{i\ell} \ge 0, \quad i = 1,2,\ldots,m, \ \ell = 1,2,\ldots,k \tag{7}$$

Here because of the simple structure of the bilinear program (7), the two linear programs can be solved explicitly in closed form. This leads to the following algorithmic implementation.

6.0 Algorithm of K Median

Given the clusters $A_1^j, A_2^j, \ldots, A_k^j$ at iteration j, compute $A_1^{j+1}, C_2^{j+1}, \ldots, A_k^{j+1}$ by the following two steps.

(a) Cluster Assignment: For each C_i^T, i = 1, 2, …, m determine ℓ(i) such that $A_\ell^j(i)$ is closed to C_i^T in the one norm.

(b) Cluster Center Update: For ℓ = 1, 2, …, k choose A_ℓ^{j+1} as a median of all C_i^T assigned to A_ℓ^j.

Stop when $A_\ell^{j+1} = A_\ell^j$. Assign each point to a cluster where center is closed in the 1-norm to the point.

7.0 Mathematical Formulation of K Means Clustering using Integer Programming Problem

The objective function is to maximize the total number of correctly classified data points as in equation (1). There are two sets of constraints used to ensure that the training samples are classified based on the noting

nearest neighbor rules as in equation (2) and (3). The equation (4) is a logical constraint used to ensure that at least one feature is used in the voting nearest neighbor rule.

The mixed integer programming is given by

The objective function is to maximize the total correct classification in equation (6). There are two sets of constraints used to ensure that the training sample are classified based on the distance averaging nearest neighbor rules as in equation (6) and (8). There is a set of logical constraint in (9) used to ensure that at least one feature is used in the distance averaging nearest neighbor rule. The integer programming problem for average SFM is given by

$$\max \sum_{i=1}^{n} y_i \tag{6}$$

subject to the constraints

$$\sum_{j=1}^{m} \overline{d_{ij}} x_j - \sum_{j=1}^{m} d_{ij} x_j \le M_{1i} y_i \quad \forall i = 1,2,...,n \tag{7}$$

$$\sum_{j=1}^{m} d_{ij} x_j - \sum_{j=1}^{m} \overline{d_{ij}} x_j \le M_{2i} (1 - y_i) \quad \forall i = 1,2,..., n \tag{8}$$

$$\sum_{j=1}^{m} x_j \ge 1 \tag{9}$$

$x \in \{0, 1\}^m$, $y \in \{0, 1\}^n$

where d_{ij} is the average statistical distance between sample i and all other samples from the same class at feature j (intra-class distance) $\overline{d_{ij}}$ is the average statistical distance between sample i and all other samples from different class at feature j (inter class distance), $M_{1i} = \sum_{j=1}^{m} \overline{d_{ij}}$ and $M_{2i} = \sum_{j=1}^{m} d_{ij}$.

Data Analysis using K means Clustering.

Table 1.

	Statistics for continuous variable: south west Number of clusters: 2 Total number of training cases: 12		
	Cluster 1	Cluster 2	Overall
Minimum	230.9000	667.8000	230.90
Maximum	894.6000	667.8000	894.60
Mean	594.509	667.8000	600.62
Standard deviation	220.4758	0.0000	44638.17

Table 2.

	Statistics for continuous variable: north east Number of clusters: 2 Total number of training cases: 12		
	Cluster 1	Cluster 2	Overall
Minimum	337.4000	1069.300	337.40
Maximum	769.1000	1069.300	1069.30
Mean	534.6727	1069.300	579.23
Standard deviation	124.4493	0.000	37898.53

Table 3.

	Statistics for continuous variable: winter se Number of clusters: 2 Total number of training cases: 12		
	Cluster 1	Cluster 2	Overall
Minimum	2.8000	52.7000	2.800
Maximum	204.1000	52.7000	204.100
Mean	44.7345	52.7000	45.398
Standard deviation	58.3340	0.0000	3098.791

Table 4.

	Statistics for continuous variable: hot seaso Number of clusters: 2 Total number of training cases: 12		
	Cluster 1	Cluster 2	Overall
Minimum	138.8000	458.6000	138.80
Maximum	461.9000	458.6000	461.90
Mean	321.2545	458.6000	332.70
Standard deviation	103.3986	0.0000	11291.32

Conclusion

According to the great eighteenth century Mathematician Leonhard Euler, “Nothing happens in the universe that does not have a sense of either contain maximum or minimum from his point of view in application of large scale data mining problems can be solved easily using quadratic programming with linear programs contains million of variables. But if some of the features are discrete and can be represented using integers, then the techniques of integer programming can be adapted. Hence integer programming approaches

have been applied for examples of annual rainfall data for Kanyakumari district from 1997 to 2010 obtained from meteorogical department.

By using Algorithmic approach, the binary program for discrete data is implemented by using clustering technique.

References

[1] Johnson, R.A., Wichern, D.W., (1998), Applied multivariate statistical analysis, 4th edition, Prenctice Hall, N.J.
[2] Robert, J. Vendertei, Linear programming: Foundations and extensions, Kluwer Academic Publisheres, Hingham, M.A., 1997.
[3] W.K.G. Icoontz, P.M. Narendra and K. Fukunga, A branch and bound clustering algorithm, IEEE Transactions on Computers C-24 (1975), 908–915.
[4] J.F. Macrotorchino and P. Michaud, Optimization en analysis odinale des donnees (Masson, Paris 1979).
[5] S. Regnier, Sur quelques aspects mathematiques at Sciences humanities 82 (1983), 85–111.
[6] G. Diehr, Evaluation of a branch and bound for clustering, SIAM Journal of Scientific and Statistical Computing, 6 (1985), 268–284.
[7] L. Kaufman and P.J. Rousseeuw, Finding groups in data: An introduction to clustering analysis (New York; Wiley, 1990).
[8] S. Chopra and M.R. Rao, On the multiway cut polyhedron, Networks, 21 (1991), 51–89.
[9] G. Klein and J.E. Aronson, Optimal clustering: A model and method, Novel Research Logistics, 38 (1991), 447–461.
[10] U. Dorndorf and E. Pesch, Fast Clustering Algorithms, ORSA Journal on Computing, 6 (1994), 141–153.
[11] S. Chopra and J.H. Owen, Extended formulations for the A-cut problem, Mathematical Programming, 73 (1996) 17–30.