Volume: 03 Issue: 01 June 2014, Page No. 38-41
ISSN: 2278-2419

A Novel Approach to Mathematical Concepts in Data Mining

I.Benjamin Franklin[1], V.Julian Arockiaraj[2]
St. Joseph's College of Arts & Science (Autonomous),Cuddalore,
Email:[1]franklinbenj@gmail.com , [2]sunil29101988@gmail.com

Abstract-This paper describes three different fundamental mathematical programming approaches that are relevant to data mining. They are: Feature Selection, Clustering and Robust Representation. This paper comprises of two clustering algorithms such as K-mean algorithm and K-median algorithms. Clustering is illustrated by the unsupervised learning of patterns and clusters that may exist in a given databases and useful tool for Knowledge Discovery in Database (KDD). The results of k-median algorithm are used to collecting the blood cancer patient from a medical database. K-mean clustering is a data mining/machine learning algorithm used to cluster observations into groups of related observations without any prior knowledge of those relationships. The k-mean algorithm is one of the simplest clustering techniques and it is commonly used in medical imaging, biometrics and related fields.

Keywords: Data mining, K-means algorithm, K-median algorithm, clustering.

I. INTRODUCTION

Mathematical program programming, that is optimization subject to constraints, in a broad discipline that has been applied to a great variety of theoretical and applied problems such as operations research [6], network problems [5], games theory and economics [9], engineering mechanics [7] and more recently to machine learning [6]. In this paper we describe three recent mathematical-programming-based developments that are relevant to data mining; feature selection [10], clustering [11] and robust representation [8]. We note at the outset that we do not plan to survey either the fields of data mining of mathematical programming, but rather highlight some recent and highly effective applications of the latter to the former. We will, however, point out other approaches that are mostly not based on mathematical programming.

Basic Description

The fundamental nonlinear programming problem consists of minimizing an objective function subject to inequality and equality constraints and is typically written as follows min f(x) subject to g(x) _< 0, h(x)=0, where x is an n-dimensional vector of real variables, f is a real- valued function of x, g and h are finite dimensional vector functions of x. if all the functions f, g and h are linear then the problem simplifies to a linear program [4,] which is the classical problem of mathematical programming. If x is two-dimensional, a linear program can be thought of as the problem of finding a lowest point (not necessarily unique) on a tilted plane surrounded by a piecewise-linear fence. Extremely efficient algorithms exist for the solution of linear programs. Thus reducing a problem to a single or finite sequence of linear programs is tantamount to solving the problem.Another reason for emphasizing mathematical programming in this work is the very broad applicability of the optimization- under-constraints paradigm; a great variety of problems from many fields can be formulated and effectively solved as mathematical programs. According to the great eighteenth century mathematician Leonhard Euler: "Nothing happens in the universe that does not have a sense of either certain maximum or minimum" [8,p 1]. From the point of view of applicability to large-scale data mining problems, the proposed algorithms employ either linear programming which is polynomial-time-solvable [9], or convex quadratic programming (section 4) which is also polynomial-time-solvable.

Extremely fast linear and quadratic programming codes [4] that are capable of solving linear program with millions of variables[8, 4] and very large quadratic programs, make the proposed algorithms easily scalable and effective for solving a wide range of problems. One limitation however is that the problem features must be real numbers or easily mapped into real numbers. If some of the features are discrete and can be represented as integers, then the techniques of integer programming [5, 4] can be employed. Integer programming approaches have been applied for example to clustering problems [6, 1], but will not be described here, principally because the combinatorial approach is fundamentally different that the analytical approach of optimization with real variables. Stochastic optimization methods based on simulated annealing have also been used in problems of inductive concept learning.The problems considered in this paper are:

II. FEATURE SELECTION

A. Introduction

A feature selection algorithm can be seen as the combination of a search technique for proposing new feature subsets, along with an evaluation measure which scores the different feature subsets. The simplest algorithm is to test each possible subset of features finding the one which minimises the error rate. This is an exhaustive search of the space, and is computationally intractable for all but the smallest of feature sets.

B. Subset selection

Subset selection evaluates a subset of features as a group for suitability. Subset selection algorithms can be broken up into Wrappers, Filters and Embedded. Wrappers use a search algorithm to search through the space of possible features and evaluate each subset by running a model on the subset. Wrappers can be computationally expensive and have a risk of

over fitting to the model. Filters are similar to Wrappers in the search approach, but instead of evaluating against a model, a simpler filter is evaluated. Embedded techniques are embedded in and specific to a model.The choice of evaluation metric heavily influences the algorithm, and it is these evaluation metrics which distinguish between the three main categories of feature selection algorithms: wrappers, filters and embedded methods. The feature selection problem treated is that of discriminating between two finite point sets in n-dimensional feature space by a separating plane that utilizes as few of the features as possible. The problem is formulated as a mathematical program with a parametric objective function and linear constraints [10]. A step function that appears in the objective function is approximated by a concave exponential on the nonnegative real line instead of the conventional sigmoid function of neural networks. This leads to a very fast iterative linear-programming-based algorithm for solving the problem that terminates in a finite number of steps. On the Wisconsin prognosis blood cancer (WPBC) [5] database by 35.4% while reducing problem feature from 32 to 4.

III. CLUSTERING

A. Definition of clustering

The process of grouping a set of physical of abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. A data can objects can be treated collectively as one group and so may be considered as a form of data compression. The clustering problem considered in this paper in that of assigning m points in the n-dimensional real space R^n to d clusters. The problem is formulated as that of determining k centers in R^n such that the sum of distances of each point to the nearest center is minimized. Once the cluster centers are determined by a training set, a new point is assigned the cluster with the nearest cluster center; if a polyhedral distance (such as the 1-norm distance) is used, the problem can be formulated as that of minimizing a piecewise-linear concave function on a polyhedral set which is shown to be equivalent to a bilinear program: minimizing the product of two linear functions on a set determined by satisfying a system of linear inequalities [11]. Although a bilinear program is a no convex optimization problem (I.e. minimizing a function that is not valley-like), a fast finite k-median algorithm consisting of solving few linear programs in closed form leads to a stationary point. Computational testing of this algorithm as a KDD tool [11] has been quite encouraging; on the Wisconsin prognosis blood cancer database (WPBC), distinct and clinically important survival curves were discovered from the data base by the k-Median Algorithm, whereas the traditional k-Mean Algorithm [6], which uses the square of the 2-norm distance, thus emphasizing outliers, failed to obtain such distinct survival curves for the same database. On four other publicly available databases each of the k-mean algorithms did best on two of the databases.

B. k-means clustering

K-Means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. *K-Means* clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into voronoi cells.The problem is computationally difficult however, there are efficient heuristic algorithms that are commonly employed and converge quickly to a local optimum. These are usually similar to the expectation-maximization algorithm for mixtures of Gaussian via an iterative refinement approach employed by both algorithms. Additionally, they both use cluster centers to model the data; however, *k*-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes.

Description

Given a set of observations $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$, where each observation is a *d*-dimensional real vector, *k*-means clustering aims to partition the *n* observations into *k* sets $(k \leq n)$ $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares (WCSS):

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

Where $\boldsymbol{\mu}_i$ is the mean of points in S_i.

C. Standard algorithm

The most common algorithm uses an iterative refinement technique. Due to its ubiquity it is often called the ***k*-means algorithm**; it is also referred to as **Lloyds'** algorithm, particularly in the computer science community.Given an initial set of *k* means $m_1^{(1)},\ldots,m_k^{(1)}$ (see below), the algorithm proceeds by alternating between two steps: [7]

Assignment step:

Assign each observation to the cluster whose mean yields the least within-cluster sum of squares (WCSS). Since the sum of squares is the squared Euclidean distance, this is intuitively the "nearest" mean. (Mathematically, this means partitioning the observations according to the voronoi diagram generated by the means).

$$S_i^{(t)} = \left\{ x_p : \left\| x_p - m_i^{(t)} \right\|^2 \leq \left\| x_p - m_j^{(t)} \right\|^2 \forall j, 1 \leq j \leq k \right\},$$

Where each x_p is assigned to exactly one $S^{(t)}$, even if it could be is assigned to two or more of them.

Mean algorithm step:

Calculate the new means to be the centroids of the observations in the new clusters.

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

Since the arithmetic mean is a least-squares estimator, this also minimizes the within-cluster sum of squares (WCSS) objective.

- Lloyd's k-means algorithm has polynomial smoothed running time. It is shown that for arbitrary set of n points

in $[0, 1]^d$, if each point is independently perturbed by a normal distribution with mean 0 and variance σ^2, then the expected running time of k-means algorithm is bounded by $O(n^{34}k^{34}d^8log^4(n)/\sigma^6)$, which is a polynomial in n, k, d and $1/\sigma$.

- Better bounds are proved for simple cases. For example, showed that the running time of k-means algorithm is bounded by $O(dn^4M^2)$ for n points in an integer lattice $\{1, \ldots, M\}^d$.

Variations

- K-Median clustering uses the median in each dimension instead of the mean, and this way minimizes L_1 norm (Taxicab geometry).
- K-Medoids (also: Partitioning around Medoids, PAM) uses the medoid instead of the mean, and this way minimizes the sum of distances for *arbitrary* distance functions.
- Fuzzy C-Means Clustering is a soft version of K-means, where each data point has a fuzzy degree of belonging to each cluster.
- Gaussian mixture models trained with Expectation-maximization algorithm (EM algorithm) maintains probabilistic assignments to clusters, instead of deterministic assignments, and multivariate Gaussian distributions instead of means.
- Several methods have been proposed to choose better starting clusters. One recent proposal is k-means++
- The filtering algorithm uses k d-trees to speed up each k-means step.
- Some methods attempt to speed up each k-means step using corsets or the triangle inequality.
- Escape local optima by swapping points between clusters.
- The Spherical k-means clustering algorithm is suitable for directional data.
- The Murkowski deals with irrelevant features by assigning cluster specific weights to each feature

IV. K-MEDIAN ALGORITHM

K-medians **clustering** are a cluster analysis algorithm. It is a variation of k-means clustering where instead of calculating the mean for each cluster to determine its centroid, one instead calculates the median. This has the effect of minimizing error over all clusters with respect to the1-norm distance metric, as opposed to the square of the 2-norm distance metric. This relates directly to the *k*-median proble**m** which is the problem of finding *k* centers such that the clusters formed by them are the most compact. Formally, given a set of data points *x*, the *k* centers c_i are

A. Medians and medoids

As the median is computed in each single dimension, the individual attributes will come from the data set, making this algorithm more reliable for discrete or even binary data sets. The means will however not necessarily be instances from the data set, as the attributes may come from different instances. This algorithm is often confused with the k-medoids Algorithm. However, a medoid has to be an actual instance from the dataset, while for the (multivariate) median this only holds for single attribute values. The actual median can thus be a combination of multiple instances. Given the vectors $(0,1)$, $(1,0)$ and $(2,2)$, the median obviously is $(1,1)$ and does not exist in the original data, and thus cannot be a medoid.

B. Initialization methods

Commonly used initialization methods are Forgy and Random Partition. The Forgy method randomly chooses *k* observations from the data set and uses these as the initial means. The Random Partition method first randomly assigns a cluster to each observation and then proceeds to the update step, thus computing the initial mean to be the centroid of the cluster's randomly assigned points. The Forgy method tends to spread the initial means out, while Random Partition places all of them close to the center of the data set. According to Hamerly et al., the Random Partition method is generally preferable for algorithms such as the *k*-harmonic means and fuzzy *k*-means.

C. Robust Representation

This problem deals with modeling a system of relations within a database in a manner that preserves, to the extent possible, the validity of the representation when the data on which the model is based changes. This problem is closely related to the generalization problem of machine learning of how to train a system on a given training set so as to improve generalization on a new unseen testing set [6]. We use here a simple linear model [7] and will show that if a sufficiently small error is purposely tolerated in construction the model, then for a broad class of perturbations the model will be a more accurate representation than one obtained by a conventional zero error tolerance. A simple example demonstrates this result.

V. DATA MINING AND KDD PROCESS

Data mining is a detailed process of analyzing large amounts of data and picking out the relevant information. It refers to extracting or mining knowledge from large amounts of data. Data Mining is the fundamental stage inside the process of extraction of useful and comprehensible knowledge, previously unknown, from large quantities of data stored in different formats, with the objective of improving the decision of companies, organizations where the data can be collected. However data mining and overall process known as Knowledge Discovery from Databases (KDD) is usually an expensive process, especially in the stages of business objectives elicitation, data mining objectives elicitation, and data preparation. This is especially the case each time data mining is applied to a blood bank.

VI. CONCLUSION

A number of ideas based on mathematical programming have been proposed for the solution of the fundamental problems of

Volume: 03 Issue: 01 June 2014, Page No. 38-41
ISSN: 2278-2419

feature selection, clustering and robust representation. Examples of applications of these ideas have been given to show their effectiveness. We discuss now some issues associated with these approaches.All the methods here used variables. Even though the class of problems falling in this category is quite broad, this requirement imposes a restriction on the type of problems that can be handled. Nevertheless the proposed methods can be applied to problems with discrete variables of one is willing to use the techniques of integer and mixed integer programming [5] which are more difficult. In fact one proposed algorithms, the k-Median algorithm, whose finite termination is established for problems with real variables, is directly applicable with no change to problems with ordered discrete variables such as integers. How well it performs on such problems would be an interesting problem to examine. We conclude with the hope that the problems solved demonstrate the theoretical and computational potential of mathematical programming as a versatile and effective tool for solving important problems in data mining and knowledge discovery in database

REFERENCES

[1] k. AI-Sultan. A Tabu search approach to the clustering problem. Pattern recognition, 28(9):1443-1451, 1995.
[2] K. P. Bennett and O. L. mangasarian. Robust linear programming discrimination of two linearly inseparable sets. Optimization Methods Software, 1:23-34, 1992
[3] D. P. Bertsekas. Nonlinear Programming. Athena Scientific, Belmont, MA, 1995.
[4] K.G. Murty. Linear Programming. John Wiley & Sons, New York, 1983.
[5] K.G. Murty. Network Programming. Prentice Hall, Englewood Cliffs, New Jersey, 1992.
[6] K.G. Murty. Operations Research. Prentice Hall, Englewood Cliffs, New Jersey, 1995.
[7] R. T. Rockafellar. Convex Analysis. Princeton University Press, Princeton, New Jersey, 1970.
[8] Y. Z, Tsypkin. Foundations of the Theory of Learning Systems. Academic Press, New York, 1973.
[9] V. N. Vapnik. The Nature of Statistical Learning Theory. Springer, New York, 1995.
[10] D.H. Wolpert, editor. The Mathematics of Generalization, Reading, MA, 1995. Addison Wesley.
[11] U. Fayyad, G. Piatetsky - Shapiro, and P. Smyth. The KDD process for extraction useful knowledge from volumes of data.Communications of the ACM,39:27-34,1996.