ISSN: 2277-9655
Impact Factor: 1.852

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Value Decomposition and Dimension Selection in Multi-Dimensional Datasets using Map-Reduce Operation

Preethi V[*1], Suriya M[2]

[*1]PG Scholor, Assistant Professor, Department of Information Technology, SNS College of Engineering, Coimbatore, India

prethiinfo@gmail.com

Abstract

The datasets which are in the form of object-attribute-time format is referred to as three-dimensional (3D) data sets. Clustering these three-dimensional (3D) data sets is a difficult task. So the subspace clustering method is applied to cluster the three-dimensional (3D) data sets. But finding the subspaces in the these three-dimensional (3D) dataset which is changing over time is really a difficult task. Sometimes this subspace clustering on three-dimensional (3D) data sets may produce the large number of arbitrary and spurious clusters. So to cluster these three-dimensional (3D) data sets a new centroid based concept is introduced called CATS. This CATS allows the users to select the preferred objects as centroids. This algorithm is not the parallel one. So it increases the time and space requirements which are needed to cluster the three-dimensional (3D) data sets. And in CATS no optimal centroids have been chosen to cluster the three-dimensional (3D) datasets. Since the CATS clusters the data based on the fixed centroids, the CATS cannot produce the good quality clusters. So for the first time in the proposed method the CPSO technique is introduced on the three-dimensional (3D) data sets to overcome all these drawbacks which clusters the three-dimensional (3D) datasets based on the optimal centroids and also it acts as the parallelization technique to tackle the space and time complexities.

Keywords: 3D subspace clustering, singular value decomposition, numerical optimization technique, protein structural data analysis, financial and stock data analysis.

Introduction

Clustering is the task used to group the similar objects and because of its applications, clustering is popular with a large diversity of domains, such as geology, marketing, etc. Across the years, the tremendous amount of growth in the data has created a lot of high-dimensional data sets in these domains. As a consequence, the deviation between any of the two objects becomes same or similar in the high dimensional data, which reduces the meaning of the cluster. A technique introduced to handle this high dimensional dataset is by clustering the dataset in its subspaces, and so the objects lie in a group is enough to be similar to a subset of attributes called subspace, rather than of living similar over the entire set of attributes called full space. Through SVD technique the subspace clustering will be made. But most the high-dimensional datasets in the domain like stock market can possibly change over time. So to handle this dataset which is changing over time is the difficult task. The data which is changing over time is referred to as three-dimensional (3D) dataset.

These three-dimensional (3D) data sets can be generally stated in the form of object-attribute-time, for instance the stock-ratio-year data in the finance area, and the residues-position-time protein structural data in the biological science, among others .In such data sets, discovering subspace clusters per timestamp may produce many spurious and arbitrary clusters, thus it is worthy to detect clusters that remain same in the database across the specified amount of time period. These three-dimensional (3D) dataset clustering is a difficult task. So a centroid based clustering concept is introduced to group the data. The algorithm called CATSeeker used to cluster the 3D data based on the centroid value. But this CATSeeker chooses the fixed object as centroid which reduces the cluster quality. So there is a need to find out the optimal centroid to group the 3D dataset which can efficiently handle the dataset that is changing over time.

Space and time complexity problem also rises while clustering very large amount of data sets that contain large numbers of records with high dimensions is considered. This is a very important issue nowadays. Examples are the clustering of profile pages in social networks, Bioinformatics applications, and article clustering of big libraries. Most sequential clustering algorithms suffers from the problem that they do not scale with larger sizes of data sets, and most of them are computationally expensive in memory space and time complexities. For these reasons, the parallelization of the data clustering algorithms is paramount in order to deal with large scale data. To develop a good parallel clustering algorithm that takes big data into consideration, the algorithm should be efficient, scalable and obtain high quality clusters.

So definitely an algorithm needed to handle the dataset which is changing over time (3D dataset) and to reduce the time and space complexity while clustering the data have more timestamp and high dimensions. So the MapReduce method is introduced with CPSO algorithm which can handle the data that is changing over time based on the optimal centroid value and the parallelization will be employed by this MapReduce CPSO to reduce the time and space complexity. The general concept behind the MapReduce method is that the problem is made as the functional abstraction using two important operations: first is the Map operation and the next is Reduce operation. The Map operation employed on a large amount of records and extracts interesting information from each record and all the values have the same key are sent to the same Reduce operation. Furthermore, the Reduce operation aggregates intermediate results with the same key that is generated from the Map operation and then generates the final results.

So the clustering task simply expressed as an optimization problem to obtain the best solution based on the minimum distances between the data points and the cluster centroids. For this task, we used Particle Swarm Optimization (PSO) as it performs a globalized search to find the best solution for the clustering task problem (this solves the K-means sensitivity of the selection of the initial cluster centroids and avoids the local optima convergence problem). PSO is the common basic and important optimization technique or method that iteratively moves to discover the most beneficial solution based on a particular value. PSO has been applied to solve a clustering task, where the problem discussed was document clustering. The results compared with K-Means, whereby the PSO algorithm tested to get more compact clustering outcomes. This PSO approach can be applied to more generalized and much larger datasets.

In addition, the MapReduce framework has been chosen as the parallelization technique in order to tackle the computational and space complexities that large datasets incur causing an efficiency degradation of the clustering. To the best of our knowledge, this is the first work that implements PSO clustering with MapReduce. Our goal is to show that PSO clustering benefits of the MapReduce framework and works with large datasets achieving high clustering quality, scalability, efficiency and a very good speedup.

Problem Definition

Nowadays the data size is growing rapidly in all the fields. These data referred to as the High Dimensional Data. Clustering in the high dimensional data is the difficult task. Thus the subspace clustering can be employed on this dataset. But subspace clustering is the difficult task when the dataset has so many timestamp. The dataset which is changing over time referred to as the three-dimensional (3D) data. The data like stock details and protein structure can be said 3D data which is changing over time.

While grouping the stock details based on its profitability the user has to know about the utility of the each stock in the organizations. Finding the particular object's usefulness and utility factor is a big problem in the subspace clustering .The usefulness or utility factor of subspace clusters, is used to find the concrete actions. Such defined patterns are called as actionable patterns, and these actionable patterns generally linked with the total amount of profits or benefits. The subspace clusters usability and utility factor can be increased by permitting users to integrate their domain knowledge in the clusters. To attain usability, we permit users to choose their preferable objects as centroids, and the objects are clustered which are similar to the centroids. Thus this subspace clustering with incorporated domain knowledge used to make important decisions in financial domain.

In Financial model the Value investors verify details and past histories of the organizations which are the critical indicators of their future subsequent stock price and cost movements. For instance, the investors who are going to invest money in the particular organization have to know about the organization histories to make the business successful. Experts in financial areas have suggested certain reliable financial ratios and their respective values. For instance, the experts in the finance domain will select the objects which produce higher profits than the others to find the additional object which has the same earning and profit value similar to the selected object called the centroid. Even though there is no practical evidence to examine their accuracy, and the choice of the suitable financial ratios and their values has stayed subjective.

On the other hand, the experts and value investors know a small amount of stocks or objects which is profitable so only these objects used as the centroids to find the other profitable stocks. Since the investor simply do not know about all the profitable stocks the cluster quality simply reduced. So to increase the cluster quality the user or investor must know about all stock and object details which is not possible in very large amount of data sets.

Dynamics and flexibility are the basic and important properties of biological molecules, example, and proteins. The B-factor value simply indicates the flexibility of the protein structure whereas the positional dynamics indicate the dynamics of the protein structural data which is changing over time. The catalytic residues can also be used as centroids. Based on the catalytic residue value the regulating residues will be found. The selected regulating residue should be similar to the mentioned catalytic residue value called centroids. For instance, a biologist has chosen 61 as the catalytic residue value and which is used as centroids to choose the other regulating residues in the protein. These two instances highlight the needs to discover the actionable clusters of objects that create more profits and benefits in the stock data and find regulating residues based on the catalytic residues in biological data. So the clusters must be homogeneous, actionable, consistent, and correlated with which the datasets change over a time.

Limitations of Existing Approaches

Most of the existing algorithms for subspace clustering only deals with the two-dimensional (2D) data which is in the form of object-attribute. The algorithms like STATPC and Maxn-Cluster can only handle the two-dimensional (2D) dataset and these algorithms cannot handle the three-dimensional (3D) data which change over time. Surviving 3D subspace clustering algorithms are not efficient in mining actionable 3D subspace clusters. Because these algorithms do not incorporate the domain knowledge which increases the cluster results. For example in protein structural data, and stock data clustering domain knowledge is important to choose the best object as centroid. But the existing algorithms are lagging in this area.

The protein structural data and stock data will change over time and these datasets will not remain same. So the existing algorithms are not sufficient to handle the data which change to every time stamp. So we cannot produce the homogenous cluster in changing time period. So the 3D subspace cluster should focus on both subsets of attributes and a subset of time stamps to increase the cluster quality. The existing algorithms like GS-search and MASC do not create 3D subspace clusters which occur for each time stamp.

The existing algorithm simply depends on the users to set the tuning parameters. But the clustering results should be insensitive to the tuning up parameters. The existing algorithms like GS-search and Tricluster require users or investors to set up the parameters which strongly influence the results. So this should be avoided while grouping the large amount of datasets. Even though the algorithm called CATSeeker effectively handles the 3D dataset this CATSeeker cluster the objects based on the centroid value. So the centroid value concept applied to high dimensional data to cluster dataset which is changing over time. But in CATSeeker only the fixed centroid has been chosen to cluster the data which reduces the cluster quality. And the time needed to cluster the data is high when fixed centroids are used to cluster the dataset.

Proposed Solution

We propose mining using Mapreduce particle swarm optimization which uses the concept of centroids, to solve the above issues.

SVD pruning

SVD is the process of calculating singular value for each attribute. In Singular Value Decomposition first the matrix for input dataset is constructed. Then based on the matrix the factorization value is constructed. In this the entire matrix is reduced to bidiagonal matrix then the decomposed values for each attribute will be found.

Introduction to Particle Swarm Optimization

To solve the above mentioned issues the MapReduce particle swarm optimization algorithm is introduced to cluster the large set of data. This algorithm produces the efficient clusters based on the set of centroids.

PSO is the swarm intelligence technique. The of particle swarm optimization is explained by the groups of birds that are looking for the optimal food origins. To get the optimal food sources the birds should move in one direction. This movement of birds referred to as current movement. If any of the birds in the group get the optimal food sources then the other birds in the group go in the same direction to get the optimal food sources. In PSO the particles search for the optimal position by moving through the search space. While moving the particle resides in two locations. They are personal finest location, and the global finest location. A particle resides in the swarm and the swarm comprises of many particles. The particles reside in the swarm owns a fitness value. This fitness value is described by the objective function which is based on the particle's placement.

And the particle may have additional information, fitness value and velocity (position) which is useful in the motion of the particle.

In PSO the particle resides in the temporary personal location with the fitness value. And also the particle bears the finest global location with the best fitness value. Based on the temporary personal location the global position of the particle will be found. In the proposed system the Global Best (or) optimal Particle Swarm Optimization technique is used. The following equations are used to relocate or move the particles within the problem search space.

$Y_j (k + 1) = Y_j (k) + Velo_j (k + 1)$ (1)

Where Y_j is the position of particle j, k is the iteration number and $Velo_j$ is the velocity of particle j. PSO uses the following equation to update the particle velocities,

$Velo_j (k + 1) = Z \cdot Velo_j (k) + (ran1 \cdot constant1) \cdot [YPar_j - Y_j (k)] + (ran2 \cdot constant2) \cdot [YG - Yj (k)]$ (2)

Where Z is inertia weight, ran1 and ran2 are randomly generated numbers, constant1, constant2 are constant coefficients, $YPar_j$ is the current best position of particle j and YG is the current best global position for the whole swarm.

Proposed MapReduce PSO Clustering Algorithm (MRCPSO)

The MapReduce-CPSO algorithm the clustering task is considered as an optimization technique to obtain the best and optimal clustering result based on the optimal centroid value. The optimal solution is obtained by calculating the distance between the data points and the centroid. The MapReduce-CPSO is similar to the K-means clustering algorithm. In k-means algorithm the centroid value depends on the weighted average value of all the points within the cluster. But in MapReduce-CPSO particle's velocity used to update the centroid value. In MR-CPSO the particles contain the information which is used to accelerate the clustering task

The MapReduce-CPSO algorithm deals with the two main operations called fitness evaluation and particle centroid updating. The equations 1 and 2 used to calculate the updated centroids in each iteration. The particle centroid updating mainly depend on the PSO movement. Sometimes the centroid value update takes a long time when the particle swarm size is large. The proposed MR-CSPO is an algorithm in which the optimal centroids have been chosen to cluster the data during clustering process, rather than choosing fixed centroids. Choosing the best optimal centroid values improves the clustering results in 3D attributes. The MapReduce function contains the following important terms.

- Centroids Vector (CV): Current cluster centroid vector.
- Velocities Vector (VV): Current velocity vector.
- Fitness Value (FV): Current fitness value for the particle at iteration t.
- Best Personal Centroids (BPC): Best personal centroid seen so far for Pi.
- Best Personal Fitness Value (BPCFV): Best personal fitness value seen so far for Pi.
- Best Global Centroid (BGC): Best global centroid seen so far for whole swarm.
- Best Global Fitness Value (BGCFV): Best global fitness value seen so far for whole swarm.

The fitness function in the MapReduce-CPSO plays the important role. The fitness function simply measures the distance between all the points and particle centroids. The average distance is calculated from all the measures. For instance nj is the number of records that belong to cluster j; Ri indicate the ith record; k is the total number clusters; Distance (Ri, Cj) is the distance between record Ri and the cluster centroid Cj. The fitness value calculation takes a long time when working with large datasets.

Updating centroids

The first task in the MR-CPSO is to update the particle centroids. There are two functions in MapReduce technique. They are map function and reduce function. First the map function is used to get the particles which have the identification numbers. The particle ID called as the Map key whereas the particle indicates the value. All the particle information like CV, VV, FV, BPC and BGC is associated with the map value. Using PSO the centroids are updated in the map function. Entropy values also used while updating the particle centroids. The entropy values are inactivity weight (W), PSO coefficients named constant 1 and constant 2. These entropy values are applied in the equation to get the updated centroid value. Finally the map function finds the updated centroid value. And this value will be given to the Reduce function for further processing.

The reduce function also called as identity reduce function. This reduce function is used to sort out the results which are made by the Map function. And also the reduce function used to aggregate the results in a single file which is made by the Map function. Thus the centroid value created in the Map function will be preserved for future operation.

Updating fitness value

The fitness value is calculated during the Reduce function. The map function is used to get the records which have the identification numbers (record ID). The record ID called as the Map key whereas the data record indicates the value. The distributed cache is used to store the particles. The map function collects the particles from this cache. The MapReduce technique uses the cache concept to increase the clustering speed. For every particle the centroid vector is extracted from the Map function and now the distance between the centroid and the record is calculated. This minimum distance is represented by the centroidID. Thus using the ParticleID and centroidID which has the small and minimum distance the composite key is formulated.

Now the new value is found by the minimum distance. And for each iteration the Map function produces the new key and the new value. Finally these new values and new keys handed over to the reduce function. The reduce function aggregates all the values and keys to obtain the average distance. This average distance is referred as fitness value. Now this fitness value acts as the centroids until the reduce function emits the key with new average distances. And after the reduce function emitting the key with new average distance, this new average distance considered as centroids. And this process repeats until get the good quality clusters.

Merging

The third task is to merge the results from the first two tasks. The final fitness value is calculated by taking the summation of all the centroids' which is generated in the updating fitness value (second) task. Then BPCFV is calculated for each particle. The calculated BPCFV compared with the fitness value. If the new particle fitness value is less than the BPCFV, the centroid and BPCFV are updated. And also the BGCFV is calculated for each particle. The calculated BGCFV compared with the fitness value. If the new particle fitness value is less than the BGCFV, the centroid and BGCFV are updated. At last in the distributed file system the new swarm with new information is saved which is the input to the next iteration.

The another algorithm called MapReduce K-means clustering algorithm also used to deal with the outlier detection problem. This algorithm is also the efficient algorithm which handles the large data sets. So clustering with MapReduce framework can work well with large amounts of data with the parallelization concept. But this MapReduce K-means clustering does not scale well for the data that is changing over time. And also this algorithm cannot handle with the increasing data sizes. And the other new algorithm called fast clustering algorithm also proposed to handle the high dimensional data. This algorithm uses the concept of constant factor approximation. In this algorithm only the samples are taken to cluster the datasets. But this algorithm cannot produce the good quality clusters since the sampling is made based on the sampling datasets not with the original datasets. And also this algorithm is time consuming one because it takes more time to search the sampling dataset from the original dataset.

Another technique called BOW (Best Of both Worlds) is introduced. This is a kind the subspace clustering technique which can handle large amount of datasets in efficient time. In this algorithm only the small amount of disk and network delay experienced. But this algorithm has some small amount of performance loss and experiences more cost like I/O cost and network cost. SO as to avoid the drawbacks of the above three techniques the MapReduce CPSO algorithm is produced. Moreover, the algorithm is faster than other parallel algorithms for very large data sets. And the performance of this algorithm is higher than other algorithms. And also this MapReduce CPSO algorithm experiences less cost than other algorithms.

Probability Estimation

BCLM(Bound-Constrained Lagrangian Method) technique is applied to calculate the probability value. Probability value will be calculated for dataset. Based on this probability value the data will be clustered to particular group. Then the objective function is formulated as follows.

$$f(P) = \sum_{o \in O} \sum_{a \in A} \sum_{t \in T} p_{oat} h(v_{oat}) util(u_{ot})$$

Where , o : object ,a-attribute ,t-time
v_{oat} : value of object
u_{ot} : utility of object
p_{oat} : probability of object

Cluster Extraction

In cluster extraction phase the probability value will be converted to "0" and "1". The object which holds 1 is added to cluster. The object which hold 0 is not added to the particular cluster and the process is repeated to find out the best location for all the data values.

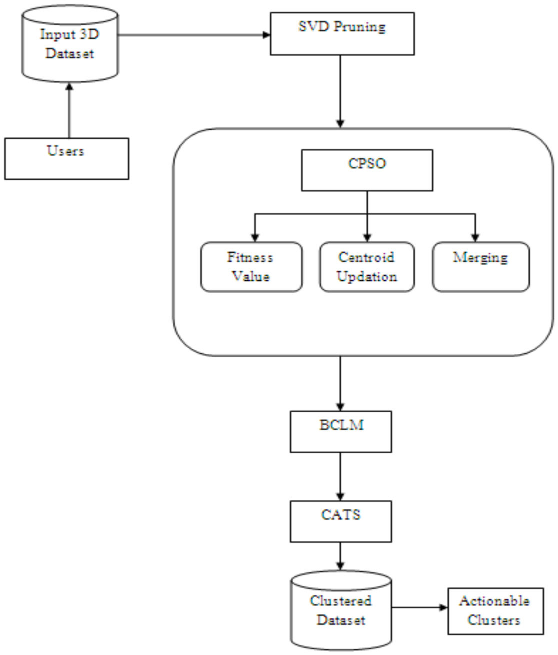

Figure 3.1 System Architecture

Applications

The CPSO algorithm can be used on the real world data set and synthetic data set, to evaluate the accuracy and actionability of the clustering results. So in our proposed work the algorithm MapReduce CPSO applied to the stock details data and protein structure data to produce the centroid based actionable subspace clusters. The MapReduce CPSO clusters these datasets based on the optimal centoids which is appropriate for changing data over a time period and while clustering these datasets the MapReduce CPSO acts as a parallelization technique to reduce the time and space complexity.

Clusters of Profitable Stocks Data

Clustering profitable stocks tasks include two phases. They are buy phase and sell phase. Based upon the details in these two phases the clustering made on the stock details data. Using this results the users and investors can make the efficient decision which will increase the profit. To cluster these stock data the utility value u_{min} is used. The u_{min} value for all the objects in the stock dataset will be found. And based upon these values the optimal value has been chosen and this optimal value finally acts as the optimal centroid to cluster the most profitable stocks. Finally the cluster results contain the profitable stocks or objects using which the users and value investors can make good decisions and attain more profits. Thus this clustering result used to make important decisions this clustering task referred as actionable task and the algorithm referred as an actionable clustering algorithm.

Clusters of Protein Structural Data

Protein structural data also change over time. The protein structure consists of amino acids called residues. The residues can be classified into catalytic residue and regulating residue. The catalytic residue value has been chosen as centroids. Based on the catalytic residue value the regulating residues will be clustered. Since drug molecules should bind to the catalytic site of the target (disease) protein, the catalytic residue chosen as centroids. In clustering the protein structural data also the threshold u_{min} value used. This threshold u_{min} is necessary for the regular functioning of the proteins therefore reduces the unwanted side effects

Instead of choosing the conserved catalytic site, it is preferable to looking for an alternative site, called allosteric site. The allosteric site is made by the regulating residues in which the drug molecules can bind selectively only with the targeted (diseased) protein but not with the other proteins in the family. Identifying the allosteric site is a difficult task. Because it is less effective than catalytic site or residue. So in most of the protein structural data clustering regulating residues has been chosen as centroids rather than allosteric site. The B-factor value also used to find the motion of residues which change over time. B-factor value can also be used to find the residues' flexibility. The residues dynamics' which is chosen from the molecular dynamics can be used to find the protein structure which changes over the time period.

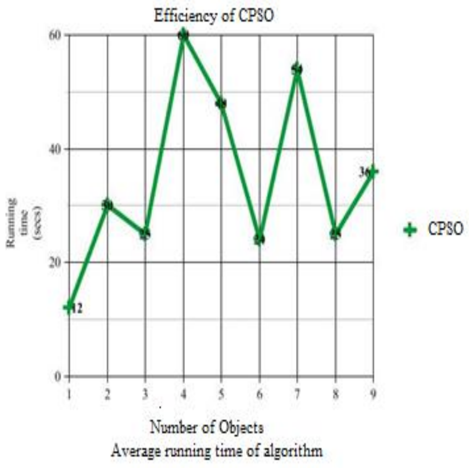

Figure 4.1 Efficiency Analysis of CPSO

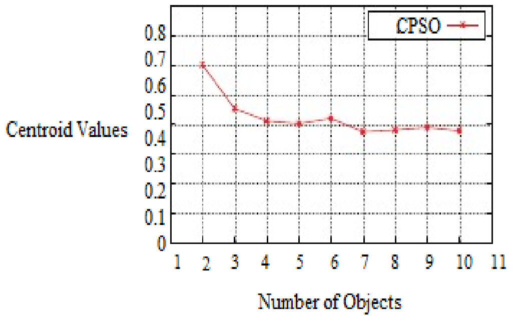

Figure 4.2 CPSO centroid scaleup

Conclusion

The proposed method called MapReduce CPSO is applied to the large amount datasets. The MapReduce CPSO is the optimization and parallel methodology technique which is used to obtain the best clustering results. In MR-CPSO the clustering is made based on the centroid value. Since MR-CPSO is the optimization technique it is used to find the optimal centroids based on the velocity of the particle. The centroid value for each iteration is updated using particle's velocity. Since MR-CPSO is the parallel methodology it is used to reduce the time and space complexity. This MR-CPSO can be applied to both real-world and synthetic datasets. This MapReduce CPSO can work well with the increasing data sizes which is used to increase the cluster quality with minimal time and space requirement.

References

[1] *K.S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When Is 'Nearest Neighbor' Meaningful?" Proc. Seventh Int'l Conf. Database Theory (ICDT), pp. 217-235, 1999.*

[2] *K. Sim, G. Liu, V. Gopalkrishnan, and J. Li, "A Case Study on Financial Ratios via Cross-Graph Quasi-Bicliques," Information Sciences, vol. 181, no. 1, pp. 201-216, 2011*

[3] *J. Kleinberg, C. Papadimitriou, and P. Raghavan, "A Microeconomic View of Data Mining," Data Mining Knowledge Discovery, vol. 2, no. 4, pp. 311-324, 1998.*

[4] *K. Wang, S. Zhou, and J. Han, "Profit Mining: From Patterns to Actions," Proc. Eighth Int'l Conf. Extending Database Technology: Advances in Database Technology (EDBT), pp. 70-87, 2002.*

[5] *K. Wang, S. Zhou, Q. Yang, and J.M.S. Yeung, "Mining Customer Value: From Association Rules to Direct Marketing," Data Mining Knowledge Discovery, vol. 11, no. 1, pp. 57-79, 2005.*

[6] *H.-P. Kriegel et al., "Future Trends in Data Mining," Data Mining Knowledge Discovery, vol. 15, no. 1, pp. 87-97, 2007.*

[7] *B. Graham, The Intelligent Investor: A Book of Practical Counsel. Harper Collins Publishers, 1986.*

[8] *X. Xu, Y. Lu, K.-L. Tan, and A.K.H. Tung, "Finding Time-Lagged 3D Clusters," Proc. IEEE Int'l Conf. Data Eng. (ICDE), pp. 445-456, 2009.*

[9] *K. Kailing, H.P. Kriegel, P. Kroger, and S. Wanka, "Ranking Interesting Subspaces for Clustering High Dimensional Data," Proc. Practice of Knowledge Discovery in Databases (PKDD), pp. 241-252, 2003.*

[10] *C.H. Cheng, A.W. Fu, and Y. Zhang, "Entropy-Based Subspace Clustering for Mining Numerical Data," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 84-93, 1999.*

[11] *S. Schubbert, K. Shannon, and G. Bollag, "Hyperactive Ras in Developmental Disorders and Cancer," Nature Rev. Cancer, vol. 7, no. 4, pp. 295-0308, 2007.*

[12] *B.J. Grant et al., "Novel Allosteric Sites on Ras for Lead Generation," PLoS One, vol. 6, no. 10, p. e25711, 2011.*

[13] *S. Bochkanov and V. Bystritsky, "ALGLIB 2.0.1 L-BFGS Algorithm for Multivariate Optimization," http://www.alglib.net/optimization/lbfgs.php, 2009.*

[14] *H. Cheng, K.A. Hua, and K. Vu, "Constrained Locally Weighted Clustering," Proc. VLDB Endowment, vol. 1, no. 1, pp. 90-101, 2008.*

[15] *J.Y. Campbell and R.J. Shiller, "Valuation Ratios and the Long Run Stock Market Outlook: An Update," Advances in Behavioral Finance II, Princeton Univ. Press, 2005.*