Clustering Medical Data Using Subspace and Parallel Approximation Algorithm

B. Thenmozhi[1], P. Shanthi[2]

[1]Research Scholar, M.Sc., Department of Computer Science,
Sri Jayendra Saraswathy Maha Vidyalaya College of Arts and Science, Coimbatore-5, India

[2]Assistant Professor, Research Supervisor, Department of Information Technology, Sri Jayendra Saraswathy Maha Vidyalaya College of Arts and Science, Coimbatore- 05, India

Abstract: ***In high-dimensional feature spaces traditional clustering algorithms tend to break down in terms of efficiency and quality. Nevertheless, the data sets often contain clusters which are hidden in various subspaces of the original feature space. In high dimensional data, however, many of the dimensions are often irrelevant. These irrelevant dimensions confuse clustering algorithms by hiding clusters in noisy data .In this paper we propose parallel approximation algorithm localize the search for relevant dimensions allowing them to find clusters that exist in multiple, possibly overlapping subspaces. A broad evaluation based on real-world medical data sets demonstrates that is suitable to find all relevant subspaces in high dimensional, sparse data sets and produces better results than existing methods.***

Keywords: Subspace clustering, Dimensionality Reduction, Redundancy Awareness, Detecting Relevant Attributes, Greedy optimization.

1. Introduction

Several algorithms for discovering clusters of points in subsets of attributes have been proposed in the literature. They can be classified into two categories: subspace clustering algorithms, and projected clustering algorithms [1]. Subspace clustering [3] algorithms search for all clusters of points in all subspaces of a data set according to their respective cluster definition. A large number of overlapping clusters is typically reported. To avoid an exhaustive search through all possible subspaces, the cluster definition is typically based on a global density threshold that ensures anti-monotonic properties necessary for an Apriori style search [4]. However, the cluster definition ignores that density decreases with dimensionality. Large values for the global density threshold will result in only low-dimensional clusters, whereas small values for the global density threshold will result in a large number of low-dimensional clusters (many of which are meaningless), in addition to the higher-dimensional clusters.

Projected clustering algorithms define a projected cluster as a pair ($X; Y$), where X is a subset of data points, and Y is a subset of data attributes, so that the points in X are \close" when projected on the attributes in Y , but they are \not close" when projected on the remaining attributes. Projected clustering algorithms have an explicit or implicit measure of \closeness" on relevant attributes (e.g., small range/variance), and a \non-closeness" measure on irrelevant attributes (e.g., uniform distribution/large variance). A search method will report all projected clusters [2] in the particular search space that an algorithm considers. If only k projected clusters are desired, the algorithms typically use an objective function to define what the optimal set of k projected clusters is.

2. Problem Statement

Based on our analysis, we argue that a first problem for both projected and subspace clustering is that their objectives are stated in a way that it is not independent of the particular algorithm that is proposed to detect such clusters in the data - often leaving the practical relevance of the detected clusters unclear, particularly since their performance also depends critically on different set parameter values. A second problem for the most previous approaches is that they assume, explicitly or implicitly, that clusters have some point density controlled by user-defined parameters, and they will (in most cases) report some clusters. However, we have to judge if these clusters \stand out" in the data in some way, or, if, in fact, there are many structures alike in the data. Therefore, a density criterion for selecting clusters should be based on statistical principles.

3. Related Works

CLIQUE, ENCLUS, MAFIA, Cluster are grid based subspace clustering algorithms that use global density thresholds in a bottom-up, Apriori style discovery of clusters. Grid-based subspace clustering algorithms are sensitive to the resolution of the grid, and they may miss clusters inadequately oriented or shaped due to the positioning of the grid. SCHISM uses a variable, statistically aware, density threshold in order to detect dense regions in a grid-based discretization of the data. However, for the largest part of the search space, the variable threshold equals a global density threshold. SUBCLU [6] is a grid free approach that can detect subspace clusters with more general orientation and shape than grid-based approaches, but it is also based on a global density threshold [7, 8].

Large amounts [9] of data are ubiquitous today. Data mining methods like clustering were introduced to gain knowledge from these data. Recently, detection of multiple clustering's has become an active research area, where several alternative clustering solutions are generated for a single dataset. Each of the obtained clustering solutions is valid, of importance, and provides a different interpretation of the data.

In DUSC [5], a point is initially considered a core point if its density measure is *F* times larger than the expected value of the density measure under uniform distribution, which does not have anti-monotonic properties, and thus cannot be used for pruning the search space. As a solution, DUSC modifies the definition of a core point so that it is anti-monotonic, which, however, introduces a global density threshold. Several subspace clustering algorithms attempt to compute a succinct representation of the numerous subspace clusters that they produce, by reporting only the highest dimensional subspace clusters, merge similar subspace clusters, or organize them hierarchically. In this paper we propose HSM [10], which defines a new pattern model for heterogeneous high dimensional data. It allows data mining in arbitrary subsets of the attributes that are relevant for the respective patterns. Based on this model we propose an efficient algorithm, which is aware of the heterogeneity of the attributes.

4. Proposed System

Motivated by these observations, we propose a novel problem formulation that aims at extracting from the data axis-parallel regions that “stand out" in a statistical sense. Intuitively, a *statistically significant* region is a region that contains significantly more points than expected. In this paper, we consider the expectation under uniform distribution. The set of statistically significant regions *R* that exist in a data set is typically highly redundant in the sense that regions that overlap with, contain, or are contained in other statistically significant regions may themselves be statistically significant. Therefore, we propose to represent the set *R* through a reduced, non-redundant set of axis-parallel statistically significant regions that in a statistically meaningful sense “explain" the existence of all the regions in *R*. We will formalize these notions and formulate the task of finding a minimal set of statistically significant \explaining" regions as an optimization problem. Exhaustive search is not a viable solution because of computational infeasibility. Propose parallel approximation algorithm for 1) selecting a suitable set $R^{reduced}$ in which we can efficiently search for 2) a smallest set *P** that explains (at least) all elements in $R^{reduced}$. Our comprehensive experimental evaluation shows that parallel approximation significantly outperforms previously proposed projected and subspace clustering algorithms in the accuracy of both cluster points and relevant attributes found.

4.1 Statistical Quality

Let *H* be a hyper-rectangle in a subspace *S*. We use the methodology of statistical hypothesis testing to determine the probability that *H* contains *AS* (*H*) data points under the null hypothesis that the *n* data points are uniformly distributed in subspace *S*. The distribution of the test statistic, *AS* (*H*), under the null hypothesis is the Binomial distribution with parameters *n* and *vol* (*H*).

$$AS(H) \sim Binormal(n, vol(H)) \quad (1)$$

The quality level α of a statistical hypothesis test is a fixed probability of wrongly rejecting the null hypothesis, when in fact it is true. α is also called the rate of false positives or the probability of type I error. The critical value of a statistical hypothesis test is a threshold to which the value of the test statistic is compared to determine whether or not the null hypothesis is rejected. For a one-sided test, the critical value θ_α is computed based on

$$\alpha = Probability(AS(H) \geq \theta_\alpha \quad (2)$$

for a two-sided test, the right critical value μR is computed by (2), and the left critical value θ_α^L is computed based on

$$\alpha = Probability(AS(H) \leq \theta_\alpha^L \quad (3)$$

Where the probability is computed in each case using the distribution of the test statistic under the null hypothesis.

A statistically significant hyper-rectangle *H* contains significantly more points than what is expected under uniform distribution, i.e., the probability of observing *AS*(*H*) many points in *H*, when the *n* data points are uniformly distributed in subspace *S* is less than α0.

4.2 Relevant vs. Irrelevant Attributes

Let *H* be a hyper-rectangle in a subspace *S*. As the dimensionality of *S* increases, *vol*(*H*) decreases towards 0, and, consequently, the critical value θ_α decreases towards 1. Thus, in high dimensional subspaces, hyper-rectangles *H* with very few points may be statistically significant.

Also, assume *H* is a statistically significant hyper-rectangle in a subspace *S*, and assume that there is another attribute $a \notin S$ where the coordinates of the points in *Supp Set*(*H*) are uniformly distributed in *dom*(*a*). We could then add the smallest interval *I0* = [*l; u*] to *H* that satisfies *attr*(*I0*) = *a* and *Supp Set*(*I0*) = *H*, i.e., *l* = *minfx:ajx 2 SuppSet*(*H*)*g*, and *u* = *maxfx:ajx 2 SuppSet*(*H*)*g*. The resulting hyper rectangle *H0* will then be statistically significant in subspace *S0* = *S [fag*. This happens roughly speaking because the support stays the same, but the volume does not increase (*AS*(*H*) = *AS*(*H0*), *vol*(*H0*) · *vol*(*H*)); for a formal proof. Clearly, reporting statistically significant hyper rectangles such as *H0* does not add any information, since their existence is “caused" by the existence of other statistically significant hyper rectangles to which intervals have been added in which the points are uniformly distributed along the whole range of the corresponding attributes.

To deal with these problems, we introduce the concept of \relevant" attributes versus \irrelevant" attributes. To test whether points in *Supp Set*(*H*) are uniformly distributed in the whole range of an attribute *a* we use the Kolmogorov Smirnov goodness of first test for the uniform distribution with a significance level of the test of α*K*.

4.3. Redundancy-Oblivious

Given a data set *D* of *n d*-dimensional points, we would like to find in each subspace all hyper rectangles that satisfy the number of hyper rectangles in a certain subspace can be infinite. However, we consider, for each subspace, all unique Minimum Bounding Rectangles (MBRs) formed with data points instead of all possible hyper-rectangles. The reason is that adding empty space to an MBR keeps its support

constant, but it increases its volume; thus, it only decreases its statistical significance.

Redundancy-oblivious problem: Find all unique subspace clusters in a set of *n d*-dimensional points.

For any non-trivial values of *n* and *d*, the *size of the search space* for the redundancy-oblivious problem is obviously very large. There are 2*d* ¡ 1 subspaces, and the number of unique MBRs in each subspace *S*, that contain at least 2 points, assuming all coordinates of *n* points to be distinct in *S*, is at least *choose*(*n;* 2) and upper bounded by *choose*(*n;* 2) + *choose*(*n;* 3) + …….. + *choose* (*n;* 2 *X dim*(*S*)).

$$R = T \cup \epsilon \cup F \quad (4)$$

Conceptually, the solution *R* to the redundancy-oblivious problem contains three types of elements: 1) a set of subspace clusters *T* representing the "true" subspace clusters, 2) a set *representing* the false positives reported by the statistical tests, and 3) a set of subspace clusters *F* representing subspace clusters that exist only because of the subspace clusters in *T*. We argue that reporting the entire set *R* is not only computationally expensive, but it will also overwhelm the user with a highly *redundant* amount of information, because of the large number of elements in *F*.

4.4. Subspace Relationship

Our goal is to represent the set *R* of all subspace clusters in a given data set by a reduced set P^{opt} of subspace clusters such that the existence of each subspace cluster $H \in R$ can be *explained* by the existence of the subspace clusters P^{opt}, and P^{opt} should be a smallest set of subspace clusters with that property. Ideally, $P^{opt} = T \cup \epsilon$.

To achieve this goal, we have to define an appropriate *Explain* relationship, which is based on the following intuition. We can think of the overall data distribution as being generated by the \true" subspace clusters, which we hope to capture in the set P^{opt}, plus background noise. We can say that the actual support *AS*(*H*) of a subspace cluster *H* can be *explained* by a set of subspace clusters *P*, if *AS*(*H*) is consistent with the assumption that the data was generated by only the subspace clusters in *P* and background noise.

More formally, if we have a set *P* = {*P*1…. P_K }of sub space clusters that should explain all subspace clusters in *R*, we assume that the overall distribution is a distribution mixture of *K* + 1 components, *K* components corresponding to (derived from) the *K* elements in *P* and the *K* + 1 component corresponding to background noise, i.e.,

$$f(x;\Theta) = \sum_{k=1}^{k+1} \mu_k f_k(x;\theta_k) \quad (5)$$

Where μk are the parameters of each component density, and μk are the proportions of the mixture.

In the following, we show how to define the *Explain* relationship assuming that all component densities are Uniform distributions. Let the *K* + 1 component is the uniform background noise in the whole space, i.e.

$$f_k(x) \sim Uniform([0,1]X \ldots\ldots X[0,1]) \quad (6)$$

For the other components, corresponding to *Pk 2 P*, we assume that data is generated such that in *sub space*(*Pk*), 1 · *k* · *K*, the points are uniformly distributed in the corresponding intervals of *Pk* (and uniformly distributed in the whole domain in the remaining attributes, since these are the irrelevant attributes for *Pk*).

$$f_k(x) \sim Uniform(I_1^k \ldots\ldots I_{mk}^k X [0,1] \ldots\ldots [0,1]) \quad (7)$$

To determine whether the existence of a subspace cluster *H* = *IH* …… *IH mH* is consistent with such a model, we have to estimate the possible contribution of each component density to *H*. For a component density *fk*, that contribution is proportional to the volume of the intersection between *fk* and *H* in the subspace of *H*, i.e., we have to determine the part of *fk* that lies in *H*.

$$\pi_H(p_k) = I_1^{\pi H} \ldots\ldots I_{mH}^{\pi H} \quad (8)$$

Because *fk* is a uniform distribution, the number of points in *H*(*Pk*) generated by *fk* follows a Binomial distribution.

$$n_i = AS(p_i) - \sum_{1 \le j \le k+1} \frac{vol(\pi p_i(P_j))}{vol(P_j)} XN_j \quad (9)$$

Say that a set of subspace clusters *P*, plus background noise, explains a subspace cluster *H* if the observed number of points in *H* is *consistent* with this assumption and not *significantly* larger or smaller than expected. From the Binomial distributions, we can derive a lower and an upper bound on the number of points in *H* that could be generated by component density *fk*, without this number being statistically significant.

Subspace clusters in *P*, plus background noise, i.e.

$$ES_H^L = \sum_k^{k+1} \theta_{\alpha 0}^L(k) \quad (10)$$
$$ES_H^U = \sum_k^{k+1} \theta_{\alpha 0}^L(k) \quad (11)$$

If *AS*(*H*) falls into this range, we can say that *AS*(*H*) is consistent with *P*, or that *P* is in fact sufficient to explain the observed number of points in *H*.

4.6. Redundancy Awareness

The problem of representing *R* via a smallest (in this sense non-redundant) set of subspace clusters Note that the optimization problem has always a solution.

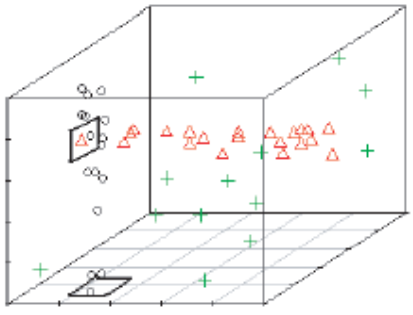

Figure 1: Example data

We emphasize the fact that the redundancy-aware problem definition avoids shortcomings of existing problem definitions in the literature. First, our objective is formulated through an optimization problem, which is independent of a particular algorithm used to solve it. Second, our definition of subspace cluster is based on statistical principles; thus, we can trust that P^{opt} stands out in the data in a statistical way, and is not simply an artifact of the method. Enumerating all elements in R in an exhaustive way is computationally infeasible for larger values of n and d. finding a smallest set of explaining subspace clusters by testing all possible subsets of R has complexity $2^{|R|}$, which is in turn computationally infeasible for typical sizes of R. We ran an exhaustive search on several small data sets where some low dimensional subspace clusters were embedded into higher dimensional spaces, similar to and including the data set depicted in Figure 1. The result set P^{opt} found for these data sets was always containing only the embedded subspace clusters (i.e., we did not even have any false positives in these cases); In Figure 1, the two depicted 2-dimensional rectangles indicating the embedded subspace clusters represent in fact the subspace clusters found by the exhaustive search.

5. Parallel Approximation Algorithm

In order to find the solution P^{opt} to the redundancy aware problem definition, we need heuristics to 1) Find a good set $R^{reduced} \subseteq R$ in which we can efficiently search for 2) a smallest set P^{sol} that explains (at least) all elements in $R^{reduced}$. Ideally, P^{opt} μ $R^{reduced}$. Propose parallel approximation algorithm that follows this schema. To construct a good set R^{educed}, it constructs subspace clusters by analyzing subspaces and local neighborhoods around individual data points Q. For this step we first suggest a method for identifying and refining candidate subspaces based on fixed neighborhoods around Q, and second a method for finding a locally optimal subspace cluster in the neighborhood of Q, given the constructed candidate subspaces for Q. To solve the optimization problem on $R^{reduced}$ heuristically, we propose a greedy strategy.

5.1. Detecting Relevant Attributes

For a given data point Q, we want to determine if there is a subspace cluster around Q. The neighborhoods we consider in this stage are *all* 2-dimensional rectangles with Q in the center and side length 2.

We propose to rank the 2-dimensional rectangles according to their actual support and select, based on an analysis of this ranking, a set of attributes, called *signaled* attributes, which are, with high probability, relevant for one of the true subspace clusters around Q.

When one or more true subspace clusters exist around Q, the actual support of the 2-dimensional projections that involve attributes of the true subspace clusters may not be statistically significant, nor higher than the support of some 2-dimensional rectangles formed by uniformly distributed attributes. However, the actual support is likely to be at least in the higher range of possible support values under uniform distribution. This does not mean that the top M pairs consist mostly of relevant attributes, but it means that the frequency with which individual relevant attributes are involved in the top M pairs is likely to be significantly higher than the frequency of a randomly chosen attribute.

5.2. Refining candidate subspaces

Let $S0$ be a set of signaled attributes. We observe that if $S0$ is only a subset of the relevant attributes for a true subspace cluster around Q, then, by considering the points in a hyper-rectangle W of width 2 around Q in subspace $S0$, we capture a fraction of the true subspace cluster's points, which is often large enough to allow us to determine more of the relevant attributes; these are attributes where the points in $SuppSet(W)$ are *not* uniformly distributed. Based on this observation, we can obtain a candidate subspace around Q through an iterative refinement of $S0$, as follows. Let $S1$ be the set of relevant attributes for W in $subspace(S0)$. If $S0 * S1$, return the empty set. If $S0 = S1$, return $S0$. Otherwise, we repeat with $S1$, selecting the relevant attributes of W in $subspace(S1)$, and so on, until no more attributes can be added.

5.3. Detecting a locally optimal subspace cluster

Let S be a candidate subspace. To determine if a subspace cluster around Q exists in S, we build a series of MBRs in S, starting from Q, and adding in each step to the current MBR the point that is closest to the current MBR in subspace S. For efficiency reasons, and because a cluster contains typically only a fraction of the total number of points, we only build $0{:}3\ n$ MBRs around Q in subspace S. Regarding the value for ±, there is no “best" value and to improve our chances of detecting a true subspace cluster, we suggest to use several different values. We simply try the 3 values $0{:}05$, $0{:}1$, $0{:}15$ for ±, resulting in *up to* three candidate subspaces for each point Q that we consider.

To construct a set *Rreduced*, tries to find subspace clusters around data points as described. The first point to consider is selected randomly from the set of all points. Subsequent points are selected randomly from the points that do not belong to detected subspace clusters in previous steps. Building *Rreduced* terminates when no data point can be selected for further subspace cluster search.

5.4. Greedy optimization

Although $|R^{reduced}| < |R|$, solving the optimization problem on *Rreduced* by testing all possible subsets is still computationally too expensive in general. Thus, we construct *greedily* a set *Psol* that explains all subspace clusters in $R^{reduced}$, but may not be the smallest set with this property. We build *Psol* by adding one subspace cluster at a time from *Rreduced*. At each step, let *Cand* be the set of subspace clusters in $R^{reduced}$ that are not explained by the current P^{sol}. Thus, subspace clusters in *Cand* can be used to extend *Psol* further, until *Psol* explains all subspace clusters in $R^{reduced}$.

6. Experimental Evaluation

Real Data: We test the performance of the compared algorithms on the following data sets from the UCI machine learning repository 4: Pima Indians Diabetes (768 points, 8

attributes, 2 classes); Liver Disorders (345 points, 6 attributes, 2 classes); and Wisconsin Breast Cancer Prognostic (WPBC)(198 points, 34 attributes, 2 classes).

Performance Measures: We use an *F* value to measure the clustering accuracy. We refer to implanted clusters as *input* clusters, and to found clusters as *output* clusters. For each output cluster *i*, we determine the input cluster *ji* with which it shares the largest number of points. The *precision* of output cluster *i* is defined as the number of points common to *i* and *ji* divided by the total number of points in *i*. The *recall* of output cluster *i* is defined as the number of points common to *i* and *ji* divided by the total number of points in *ji*. The *F* value of output cluster *i* is the harmonic mean of its precision and recall. The *F* value of a clustering solution is obtained by averaging the *F* values of all its output clusters. Similarly, we use an *F* value to measure the accuracy of found relevant attributes based on the matching between output and input clusters.

Table 1: Number Relevant Attributes with F Value

Algorithm/Performance	*F value - Cluster Points*	*log (time in*
ORCLUS	0.23	1
MAFIA	0.28	4
P3C	0.5	5
PRIM	0.4	7
MINECLUS	0.58	6
PROCLUS	0.61	4
PARALLEL	0.8	8

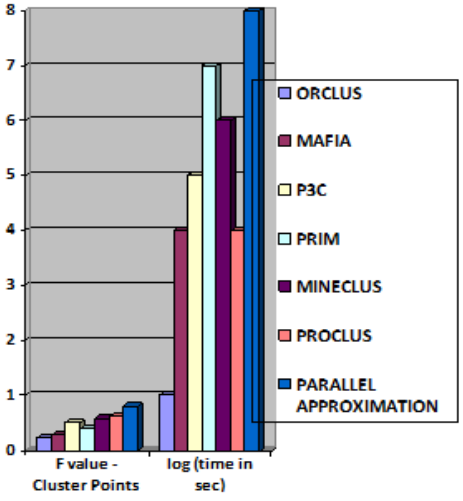

Figure 2: Liver Disorders

Parallel approximation algorithm computes subspace clusters that are statistically significant. The other algorithms sometimes compute statistically significant subspace clusters, other times they do not, depending on parameter values and on the density of the implanted clusters (denser clusters are easier to detect). The classes in the real data sets form statistically significant clusters, and these clusters stay statistically significant when adding uniform attributes

7. Conclusion

In this paper, we proposed dimensionality reduction based clustering methods for high dimensional data and we analyzed that by reducing the dimension and clustering, produced the best clustering. We proposed a parallel approximation algorithm for clustering high dimensional data with greedy optimization.

8. Future Work

In the future we will study cell-based subspace clustering and density-based subspace clustering.

References

[1] C. Aggarwal, J. Wolf, P. Yu, C. Procopiuc, and J. Park. Fast algorithms for projected clustering. In SIGMOD, pages 61-72, 1999.
[2] C. Aggarwal and P. Yu. Finding generalized projected clusters in high dimensional spaces. In SIGMOD, pages 70-81, 2000.
[3] R. Agrawal, J. Gehrke, D. Gunopulos, and
[4] P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In SIGMOD, pages 94-105, 1998.
[5] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In VLDB, pages 487{499, 1994.
[6] Assent, R. Krieger, E. Miuller, and T. Seidl. DUSC: Dimensionality unbiased subspace clustering. In ICDM, pages 409-414, 2007.
[7] Assent, R. Krieger, E. Miuller, and T. Seidl. INSCY: Indexing subspace clusters with in-process-removal of redundancy. In ICDM, pages 719-724, 2008.
[8] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases. In KDD, pages 226{231, 1996.
[9] E. Miuller, I. Assent, R. Krieger, S. Giunnemann, and T. Seidl. DensEst: Density estimation for data mining in high dimensional spaces. In SDM, pages 173-184, 2009.
[10] E. Miuller, I. Assent, R. Krieger, T. Jansen, and T. Seidl. Morpheus: Interactive exploration of subspace clustering. In KDD, pages 1089-1092, 2008.
[11] E. Miuller, I. Assent, and T. Seidl. HSM: Heterogeneous subspace mining in high dimensional data. In SSDBM, pages 497-516, 2009.

Author Profile

Mrs. B. Thenmozhi has done M. Sc (SS). She is pursuing M. Phil (CS) in Sri Jayendra Saraswathy Maha Vidyalaya CAS, Coimbatore-5. Her area of interest: includes Data mining, Networking, Software Engineering

Mrs. P. Shanthi has done M.C.A. and M.Phil. Presently she is wokring as Assistant Professor in Department of Information Technology- Sri Jayendra Saraswathy Maha Vidyalaya CAS, Coimbatore-5