A Novel Approach for Determination of Clusters from Unlabeled Data Sets

Jammi Ashok¹, Srinivas Bobba²
¹ Professor, Dept. of CSE, Gurunanak Institute of Technology, Hyderabad
² Assistant Professor, Dept. of IT, Geethanjali college of Engg. & Tech, Hyderabad

Abstract - We introduce the Novel approach for determination of clusters from unlabeled data sets. We investigate a new method called Extended Support vector Machine (ESVM) along with existing Dark Block Extraction (DBE) which is based on an existing algorithm for Visual Assessment of Cluster Tendency (VAT) of a data set, using several common image and signal processing techniques. Its basic steps include 1) Generating a VAT image of an input dissimilarity matrix, 2) Performing image segmentation on the VAT image to obtain a binary image, followed by directional morphological filtering, 3) Applying a distance transform to the filtered binary image and projecting the pixel values onto the main diagonal axis of the image to form a projection signal, 4) Smoothing the projection signal, computing its First-order derivative and then detecting major peaks and valleys in the resulting signal to decide the number of clusters, and 5) The K-Means algorithm is applied to the major peaks. We also implement the Cluster Count Extraction (CCE), which uses VAT and the combination of several image processing techniques. In both the methods we use Reordered Dissimilarity Image (RDI), which highlights potential clusters as a set of “Dark blocks” along the diagonal of the image, corresponding to sets of objects with low dissimilarity, which is implemented using VAT algorithm.

Keywords - Clustering, Clustering tendency, Extended support vector machine (ESVM), DBE, VAT

1. INTRODUCTION

One of the major problems in cluster analysis is the determination of the number of clusters in unlabeled data, which is a basic input for most clustering algorithms. In this project, we investigate a new method called Dark Block Extraction (DBE) for automatically estimating the number of clusters in unlabeled data sets, which is based on an existing algorithm for Visual Assessment of Cluster Tendency (VAT) of a data set, using several common image and signal processing techniques. A general question in the data mining community is how to organize observed data into meaningful structures (or taxonomies). As an exploratory data analysis tool, cluster analysis aims at grouping objects of a similar kind into their respective categories. There have been a large number of clustering algorithms reported. In general, clustering of unlabeled data poses three major problems: assessing cluster tendency, i.e., how many clusters to seek or partitioning the data into meaningful groups, and validating the clusters discovered. This project addresses the first problem, i.e., determining the number of clusters c prior to clustering. Many clustering algorithms require the number of clusters c as an input parameter, so the quality of the resulting clusters is largely dependent on the estimation of c. For some applications, users can determine the number of data themselves.

1.1 PREVIOUS WORK

A general question in the data mining community is how to organize observed data into meaningful structures (or taxonomies). As an exploratory data analysis tool, cluster analysis aims at grouping objects of a similar kind into their respective categories. There have been a large number of clustering algorithms reported. In general, clustering of unlabeled data poses three major problems: assessing cluster tendency, i.e., how many clusters to seek or partitioning the data into meaningful groups, and validating the clusters discovered. This project addresses the first problem, i.e., determining the number of clusters c prior to clustering. Many clustering algorithms require the number of clusters c as an input parameter, so the quality of the resulting clusters is largely dependent on the estimation of c. For some applications, users can determine the number of data themselves.

2. LITERATURE SURVEY

The three clustering algorithms considered in this article are the well-known K-Means and single linkage algorithms and a recently developed simulated annealing (SA) based clustering technique that uses probabilistic redistribution of points. The K-Means algorithm is an iterative scheme that evolves K crisp, compact, and hyperspherical clusters in the data such that a measure is minimized. Here, the K cluster centers are initialized to K randomly chosen points from the data, which is then partitioned based on the minimum squared distance criterion. The cluster centers are subsequently updated to the mean of the points belonging to them. This process of partitioning followed by updating is repeated until either the cluster centers do not change or there is no significant change in the J values of two consecutive iterations. [2]

The single linkage clustering scheme is a no iterative
method based on a local connectivity criterion, and is usually regarded as a graph theoretical model. Instead of an object data set X, single linkages process sets of n2 numerical relationships, between pairs of objects represented by the data. The number represents the extent to which object j and k are related in the sense of some binary relation. It starts by considering each point in a cluster of its own. The single linkage algorithm computes the distance between two clusters S and T: Based on these distances, it merges the two closest clusters, replacing them by the merged cluster. The distance of the remaining clusters from the merged one is recomputed as above. The process continues until the single cluster, comprising all the points, is formed. The third clustering algorithm considered in this article for the purpose of comparison is a simulated annealing (SA) based scheme with probabilistic redistribution of the data points. The SA algorithm starts from a random initial configuration at high temperature Tmax. [2]

In Intelligent statistical data analysis or unsupervised classification, cluster analysis is to determine the cluster number or cluster membership of a set of given samples, by its mean vector. In most cases, the first step of the clustering is to determine the cluster number. The second step is to design a proper clustering algorithm. In recent years, several clustering analysis algorithms have been developed to partition samples into several clusters, in which the number of clusters is predetermined. The most notable approaches are, for example, the mean square error (MSE) clustering and finite mixture model algorithms. The MSE clustering algorithm typically is implemented by the well-known -mean algorithm. This method requires specifying the number of clusters, , in advance. If is correctly selected, then it can produce a good clustering result; otherwise, data sets cannot be grouped into appropriate clusters. However, in most cases the number of clusters is unknown in advance. Because it is difficult to select appropriate number of clusters, some heuristic approaches have been used to tackle this problem. The rival penalized competitive learning (RPCL) algorithm has demonstrated a very good result in finding the cluster number. However, there is still no appropriate theory being developed. [4]

In the mixture model cluster analysis, the sample data are viewed as two or more mixtures of normal (Gaussian) distribution in varying proportion. The cluster is analyzed by means of mixture distribution. The likelihood approach to the fitting of mixture models has been utilized extensively. However, the determination of the appropriate cluster number still remains one of the most difficult problems in cluster analysis. The Bayesian–Kull back Ying–Yang (BYY) learning theory has been proposed. The BYY learning is a unified algorithm for both unsupervised and supervised learning which provides us a reference for solving the problem of selecting cluster number. The experimental results worked very well for a large set of samples when the smoothing parameter. However, for a relatively small set of samples, the maximum likelihood (ML) method with the expectation-maximization (EM) algorithm for estimating mixture model parameters will not adequately reflect the characteristics of the cluster structure. In this way, the selected cluster number is incorrect. To solve the problem for the small set of samples, the BYY theory for data smoothing is developed is approach considers the nonparametric density estimation and the smoothing factor in the Parzen window. [4]

We are living in a world full of data. Every day, people encounter a large amount of information and store or represent it as data, for further analysis and management. One of the vital means in dealing with these data is to classify or group them into a set of categories or clusters. Actually, as one of the most primitive activities of human beings [14], classification plays an important and indispensable role in the long history of human development. In order to learn a new object or understand a new phenomenon, people always try to seek the features that can describe it, and further compare it with other known objects or phenomena, based on the similarity or dissimilarity, generalized as apical distortion while using as few dimensions as possible. Also note that, in practice, many (predictive) vector quantizes are also used for (non predictive) clustering analysis.

We consider a type of preliminary data analysis related to the pattern recognition problem of clustering. Clustering or cluster analysis is the problem of partitioning a set of objects into c self-similar subsets based on available data and some well-defined measure of (cluster) similarity. In some cases, a geometric description of the clusters (e.g. by “cluster centers” in data space) is also desired and some clustering methods are able to produce such geometric descriptors. The type of clusters found is strongly related to the properties of the mathematical model that underlies the clustering method. All clustering algorithms will find an arbitrary number of clusters, even if no “actual” clusters exist. Therefore, a fundamentally important question to ask before applying any particular (and potentially biasing) clustering algorithm is: Are clusters present at all? The problem of determining whether clusters are present as step prior to actual clustering is called the assessing of clustering tendency. Various formal (statistically based) and informal techniques for tendency assessment are discussed in Jain and Dubes and
Everitt. None of the existing approaches is completely satisfactory (nor will they ever be). [17]

The purpose of this note is to add a simple and intuitive visual approach to the existing repertoire of tendency assessment tools. Visual approaches for various data analysis problems have been widely studied in the last 25 years; Tukey and Cleveland are standard sources for many visual techniques. The visual approach for assessing cluster tendency introduced here can be used in all cases involving numerical data. It is both convenient and expected that new methods in clustering have a catchy acronym. Consequently, we call this new tool VAT (visual assessment of tendency). The VAT approach presents pair wise dissimilarity information about the set of objects O = {o1, on} as a square digital image with n2 pixels, after the objects are suitably reordered so that the image is better able to highlight potential cluster structure. To go further into the VAT approach requires some additional background on the types of data typically available to describe the set. There are two common data representations upon which clustering can be based. When each object is represented by a (column) vector, the set is called an object data representation of O. [7]

If the original data has missing components (is incomplete), then any existing data imputation scheme can be used to “fill in” the missing part of the data prior to processing. The ultimate purpose of imputing data here is simply to get a very rough picture of the cluster tendency in O. Consequently, sophisticated imputation schemes, such as those based on the expectation-maximization (EM) algorithm in Dempster, Laird and Rubin, are unnecessarily expensive in both complexity and computation time. For incomplete object data, we would suggest the Dixon scheme, which generates a pair wise Euclidean (or other norm) dissimilarity Rij from incomplete xi and xj simply by using all features common to both object data, and then properly scaling the result, based on how many of the s possible features are actually used. For missing dissimilarity (Rij), one of the triangle inequality schemes in Hathaway and Bezdek [9] should be sufficiently accurate. We refer the reader interested in learning more about missing data and imputation to little and Rubin and Schafer. [7]

So, we can assume without loss that dissimilarity data of the type needed for a VAT display can be easily obtained, whether the original data description of O is object or relational, and whether the data are complete or incomplete.

Cluster analysis is a method for clustering a data set into groups of similar characteristics. It is an approach to unsupervised learning and also one of the major techniques in pattern recognition. The conventional (hard) clustering methods restrict each point of the data set to exactly one cluster. Since Zadeh (1965) proposed fuzzy sets that produced the idea of allowing to have membership functions to all clusters, fuzzy clustering has been widely studied and applied in a variety of substantive areas (Bezdek, 1981; Ho’ppner et al., 1999; Yang, 1993; Baraldi and Blonda, 1999a,b). In the fuzzy clustering literature, the fuzzy means (FCM) clustering algorithm and its variation are the most well-known and used methods (Bezdek, 1981; Ho’ppner et al., 1999; Yang, 1993). However, it is necessary to preassume the number c of clusters for these fuzzy clustering algorithms.[6] In general, the number c is unknown. The problem for finding an optimal c is usually called cluster validity (Bezdek, 1974a, b). The issue of cluster validity methods is a broad one. In this paper we focus on the cluster validity of fuzzy partitions. The objective is to find optimal c clusters that can validate the best description of the data structure. Each of these optimal c clusters should be compact and separated from other clusters. The first proposed fuzzy cluster validity functions associated with FCM are the partition coefficient (PC) and partition entropy (PE) (Bezdek, 1974a, b). Dave (1996) proposed a modified partition coefficient (MPC) index by changing its range to the interval [0, 1]. The MPC index behaves like the fuzzy set decomposition of Backer and Jain (1981). Above indexes have the disadvantage of lack for the connection to the geometrical structure of data. The separation coefficient proposed by Gunderson (1978) was the first validity index that explicitly takes into account the data geometrical properties. Indexes in this class include the XB index proposed by Xie and Beni (1991), FS index proposed by Fukuyama and Sugeno (1989), SC index proposed by Zahid et al. (1999), the fuzzy hypervolume (FHV) and partition density (PD) indexes proposed by Gath and Geva (1989). [11]

Once the partition is obtained by a clustering method, the validity function can help us to validate whether it accurately presents the data structure or not. We know that there are broadly extended types of FCM in the literature, such as Gustafson and Kessel (1979), Krishnapuram and Kim (1999), Bezdek et al. (1981a,b), Dave (1992), Gath and Geva (1989), Wu and Yang (2002), Yu and Yang (accepted for publication), etc.[15] However, validity indexes are considered to be independent of clustering algorithms. Most clustering algorithms, such as mentioned above, can generate fuzzy partitions and cluster centers for a given data set. [8]Thus, we only consider the standard FCM clustering algorithm for all validity indexes. In this paper, we present a new validity index for fuzzy clustering called a partition coefficient and exponential separation (PCAES) index. It uses the factors from a
normalized partition coefficient and an exponential separation measure for each cluster and then pools these two factors to create the PCAES validity index. We also discuss the case that validity indexes face in a noisy environment. Most validity indexes measure the degree of compactness and separation for the data structure in all of c clusters and then find an optimal c that each one of these optimal c clusters is compact and separated from other clusters. If the data set contains some noisy points that may be far away from other clustered points, it can be visualized that validity indexes will take the noisy point into a compact and separated cluster. However, the proposed new validity measure can give another point of view in this noisy environment. For each identified cluster, we can measure the potential for the cluster to be a well identified cluster. Under this criterion, a noisy point will not have enough potential to be a cluster. This consideration has a different merit from most cluster validity indexes. The merit of the proposed validity index is not only to find an optimal cluster number estimate and also to provide more information about the data structure in a noisy environment. [11]

In medical informatics area, an abundance of digital data has promised a profound impact in both the quality and rate of discovery and innovation. Modern experimental and observational instruments generate and collect large sets of data of varying types at increasing speeds. Worldwide health scientists are producing, accessing, analyzing, integrating, and storing massive amounts of digital medical data daily, through observation, experimentation, and simulation. If we were able to effectively transfer and integrate data from all possible resources, then a deeper understanding of all these data sets and better exposed knowledge, along with appropriate insights and actions, would be granted. Unfortunately, in many cases, the data users are not the data producers, and they thus face challenges in harnessing data in unforeseen and unplanned ways. In order to obtain the ability to integrate heterogeneous data, and thereby efficiently revolutionize the traditional medical and biological research, new methodologies built upon the increasingly pervasive cyber infrastructure are required to conceptualize data, and acquire the “deep” knowledge out of original data thereafter. In particular, ontologies are formal, declarative knowledge representation models. They form a semantic foundation for many domains, and thus render great help to medical informatics researchers in better capturing hidden knowledge from large amounts of original data. The most renowned example of applying ontological techniques into medical and biological research is the Gene Ontology (GO) project. [9]

Note that a lot of time and efforts have been spent in every search for available information in each small medical informatics sub area. This situation is further aggravated by great complexity and imprecise terminologies, which characterize typical medical and biological fields. A great deal of variety has been identified in the adoption of different biological terms, along with the relationships among all these terms. Such variety has inhibited effective information acquisition by both human and computers. Therefore, there is a need to explore innovative, cyber-enabled computing frameworks that are based on ontological techniques. Such frameworks will facilitate knowledge acquisition from existing resources, assist biologists in better understanding important biological functions at different levels, and ultimately, help clinicians in making sound decisions when treating their patients. [13]

Ontology is a computational model of some portion or domain of the world. The model describes the semantics of the terms used in the domain. Ontology is often captured in some form of a semantic network, i.e., a graph whose nodes are concepts or individual objects and whose arcs represent relationships or associations among the concepts. The semantic network is augmented by properties and attributes, constraints, functions, and rules, which govern the behavior of the concepts. In brief, ontology consists of a finite set of concepts, along with these concepts’ properties and relationships. In addition, most real-world ontologies have very few or no instances, i.e., they only have the aforementioned graphical structure (also known as “schema”). [12]

Ontology Heterogeneity is an inherent characteristic of ontologies developed by different parties for the same (or similar) domains. The heterogeneous semantics may occur in two ways. (1) Different ontologies could use different terminologies to describe the same conceptual model. That is, different terms could be used for the same concept, or an identical term could be adopted for different concepts. (2) Even if two ontologies use the same name for a concept, the associated properties and the relationships with other concepts are most likely to be different. Ontology Matching is short for “Ontology Schema Matching”, also known as “Ontology Alignment,” or “Ontology Mapping.” It is the process of determining correspondences between concepts from heterogeneous ontologies (often designed by distributed parties). Such correspondences include many relationships, for example, equivalentWith, subClassOf, superClassOf, and siblings. Machine Learning is a scientific discipline that is concerned with the design and development of algorithms that allow computers to change behavior based on available data (also known as “training data”). A major focus of
machine learning research is to automatically learn to recognize complex patterns and make intelligent decisions based on data. Ontological techniques have been widely applied to medical and biological research. The most successful example is the GO project, which is a major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases. GO provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data, as well as tools to access and process such data. There are many research directions in ontologies, e.g., automatic ontology generation, ontology engineering, and ontology matching, etc., with ontology matching the most related one to this paper’s theme. [13].

According to the classification, most ontologymatching techniques can be divided into two categories: rule-based approaches and learning-based approaches. 1) Rule-Based Ontology Matching: N.F. Noy and M.A. Musen describe PROMPT, a semiautomatic approach to ontology alignment. By performing some tasks automatically and guiding the user in performing other tasks for which intervention is required, PROMPT helps in understanding ontologies covering overlapping domains. S. Castano et al. present H-MATCH. The authors divide the semantics of a concept into its linguistic and contextual parts. The former captures the meaning of terms used as concept names, while the latter evaluates the semantic affinity between two concepts by taking into account the affinity between their contexts, which are concept properties and relationships. D. Dou et al. view ontology translation as ontology merging and automated reasoning, which are in turn implemented through a set of axioms. [18]

### 3. EXTENDED SUPPORT VECTOR MACHINE ALGORITHM

- Dissimilarity Transformation and Image segmentation.
- Directional Morphological filtering of binary image.
- Distance transform and diagonal projection of filtered image.
- Detection of major peaks and valleys in the projected signal.

### 3.1 DBE ALGORITHM

- **Step 1**: Input data set is retrieved rows and corresponding columns are transformed.
- **Step 2**: We retrieve value from each cell and store it in dark blocks.
- **Step 3**: We compute the validation between diff columns of data.
- **Step 4**: We apply some operations to compute the distance between combination of values.
- **Step 5**: The data is filtered in this stage.
- **Step 6**: The above all the steps are performed until null variation & centurion calculated.
- **Step 7**: Values are stored as groups and displayed through VAT technique.

### 3.2 VAT ALGORITHM

- **Step 1**: The regions of interest given as input from DBE algorithm.
- **Step 2**: We retrieve e MAX and MIN values.
- **Step 3**: The classes are compared until no variance.
- **Step 4**: These values are ordered and displayed as an image.

### 4. RESULTS

The concept of this paper is implemented and different results are shown below.

In the above screen select the X-axis attribute and Y-axis attribute and then press on the compute clusters button, will get number of clusters, each cluster is having a group of data items. select any cluster from the existed and display the cluster.
Extended Support Vector Machine (ESVM) works for classification. It is used as a supervised technique for analyzing and recognizing patterns where we give Data Sets as input. This technique can be very efficiently used in case of high dimensional data such as unlabeled data.

Graphical display of the clusters.

5. CONCLUSION

In this paper investigates a nearly parameter-free method for automatically estimating the number of clusters in unlabeled data sets. The only user-defined parameter that must be chosen to control the filter size. It is relatively easy to make a realistic and useful choice, since it essentially specifies the smallest cardinality of a cluster relative to the number of objects in the data. Our experiments confirm what many users of clustering believe: that most methods prefer “larger” rather than “smaller” clusters. The cluster number corresponding to the detection of major peaks extracted by DBE appears to be increasingly reliable. As long as the filter sizes are set to be less than the minimum meaningful cluster size, the larger they are, the more reliable the estimation of the cluster number should be.

6. REFERENCES


AUTHORS BIOGRAPHY

Prof. J. Ashok
Currently working as Professor and Head at Gurunanak Institute of Technology, Hyderabad, A.P., INDIA. He has received his B.E. Degree from Electronics and Communication Engineering from Osmania University and M.E. with specialization in Computer Technology from SRTMU, Nanded, INDIA.

His main research interest includes neural networks, Bioinformatics and Artificial Intelligence. He has been involved in the organization of a number of conferences and workshops. He has been published more than 35 papers in International journals and conferences. He is currently doing his Ph.D from Anna University and is at the end of submission.

Srinivas Bobba
Received his B.E (CSE) from Gulberg University, Karnataka, India in 1997. He worked as senior consultant with ESPN, USA. He has 2 years of experience in teaching and 5 years in Industry. He is pursuing M.Tech degree in software engineering from JNTU Hyderabad. His areas of interest are Databases and Data Mining.