Knowledge Assisted Visualization for Imbalanced Data Clustering

P. Alagambigai
Department of Computer Applications,
Easwari Engineering College,
Chennai, Tamilnadu, India - 600089.

K. Thangavel
Department of Computer Science
Periyar University
Salem, Tamilnadu, India - 636011.

ABSTRACT
The common challenge which is faced by much of the data clustering techniques is data complexity, which leads to many issues such as overlapping, lack of representative data and class imbalance. This may deteriorates the clustering process. The situation gets worse when the class imbalance is very high. To cluster such imbalanced data sets, better understandings of the dataset and efficient clustering algorithms are required. This could be achieved by integrating suitable domain intelligence into the clustering process. In this paper, Knowledge Assisted Visualization framework is proposed for imbalanced data clustering and validation. The proposed Knowledge Assisted Visualization framework integrates an efficient visual clustering framework with suitable domain intelligence acquired from domain experts and users into clustering process. An experimental analysis is carried out over a wide range of highly imbalanced data sets. Experiments demonstrate that the proposed method works well with imbalanced dataset and eases the cluster identification and validation in an effective way.

General Terms
Data Mining.

Keywords
Data Mining, Class Imbalance, Interactive Clustering, Knowledge Assisted Visualization, Visual Clustering.

1. INTRODUCTION
The continuous expansion of data availability in many large scale, complex and networked systems makes the fundamental understanding and knowledge discovery process tedious. This becomes worse when the data set is imbalance in nature. Although existing knowledge discovery and data engineering techniques have shown great success in many real-world applications, the problem of learning from imbalanced data (the imbalanced learning problem) is a relatively new challenge that has attracted growing attention from both academia and industry. The imbalanced learning problem is concerned with the performance of learning algorithms in the presence of underrepresented data and severe class distribution skews. Due to the inherent complex characteristics of imbalanced data sets, learning from such data requires new understandings, principles, algorithms, and tools to transform vast amounts of raw data efficiently into information and knowledge representation [13].

Cluster analysis is deemed one of the most difficult and challenging problems in machine learning, particularly due to its unsupervised nature. The unsupervised nature of the problem implies that its structural characteristics are not known, except if there is some sort of domain knowledge available in advance. Cluster analysis includes two major aspects: clustering and cluster validation. Clustering achieves to distinguish objects into groups according to certain criteria. The grouped objects are called clusters, where the similarity of objects is high within clusters and low between clusters. To achieve different application purposes, a large number of clustering algorithms have been developed [4, 14]. However, there are no general-purpose clustering algorithms that fit all kinds of applications; thus, the evaluation of the quality of clustering results plays the critical role of cluster analysis, i.e. cluster validation, which aims to assess the quality of clustering results and find a fit cluster scheme for a specific application [4]. Numerous studies have been adopted so far, in the literature [3, 14] to deal with clustering the large dataset automatically. However, clustering is still a challenging task, since many cluster algorithms fail to do well in scaling with the size of the dataset and the number of dimensions that describe the points, or in finding arbitrary shape of clusters, or dealing effectively with the presence of noise [5]. While dealing with cluster validation, the traditional statistical methods such as variance and intra cluster / inter cluster similarity are found to be ineffective to validate irregularly shaped clusters. To overcome the aforementioned issues, visualization techniques have been improved in the field of cluster analysis and validation.

Visualization techniques could enhance the current knowledge and data discovery methods by increasing the user involvement in the interactive process. Visualization has been proven as an efficient technique for cluster visualization and validation [1, 2, 6]. Most existing visualization techniques and systems were not designed to utilize the information and knowledge about the data or derived from the analysis and visualization process.

As visual data analysis is inherently an iterative and explorative process, it is highly desirable to enable more effective visualization by utilizing information about the visualization process itself (e.g., visualization parameters chosen by users) and knowledge about the data to be visualized (e.g., feature description from specialists). Collecting and leveraging such information and knowledge becomes important, especially when the cost of visualization is high or when the work requires collaborative efforts. The combination of such information from different visualization processes can also infer new knowledge that can aid data visualization in an intelligent manner [25].

The rest of paper is organized in the following manner. In Section 2, related work is described. In Section 3, the complexity in data is analyzed. The overview of the visualization is described in section 4. Section 5 deals with the proposed Knowledge Assisted Visualization system. The experimental analysis and results are discussed in section 6. Section 7 concludes the paper.

2. RELATED WORK
Interactive clustering differs from traditional automatic clustering in such a way that it incorporates user's domain knowledge into the clustering process. There are wide variety

of interactive clustering methods are proposed in recent years [1, 2, 6, 23, 24, 26, 27].

Star coordinate based visual cluster analysis system is proposed by Kandogan [17] to visualize and analyze the clusters interactively. In star coordinates, coordinate axes are arranged on a two dimensional surface, where each axes shares the same origin point. Each multidimensional data element is represented by a point, where each attribute of the data contributes to its location through uniform encoding. Interaction features of star coordinates provide users the ability to apply transformations dynamically, integrate and separate dimensions, analyze correlations of dimensions, view clusters, trends, outliers in the distribution of data, and query points based on data range.

Olga Sourina, et al. [24] proposed a novel interactive clustering method based on geometric model with implicit functions and visualization techniques integrated with Graphical User Interface. First, visual clustering with blobby model allows the user to see the result of clustering on the screen and set the appropriate parameters interactively. After that, the user can get data of cluster in two ways. First method implies usage of solid based subdivision algorithm. In the second method, the user needs to wrap the cluster he/she is interested in which geometric primitive solids that currently are cubes and/or spheres/ellipsoids. Geometric operations of union, intersection or subtraction can be performed over the geometric primitive solids to get the final wrapping shape. The user visually clusters the data and wraps the clusters with geometric shapes or even query clusters through graphics interface accessing dynamically 3D projections of multidimensional points from database or files.

Keke Chen and Liu, L. [6] proposed VISTA model, an intuitive way to visualize clusters. This model provides a similar mapping like star coordinates [17, 18], where a dense point cloud is considered a real cluster or several overlapped clusters. To overcome the arbitrary and random adjustments of star coordinates and its extensions, HOV^3 (Hypothesis Oriented Verification and Validation by Visualization) approach is proposed by Zhang, et al. [26, 27]. HOV^3 generalizes the visual adjustments by coefficient / measure vector. The performance of HOV^3 is better than star coordinates and its extensions on cluster detection.

Subspace clustering (or projective clustering) is very important for effective identification of patterns in a high dimensional data space which receives great attention in recent years. The recent survey [8] offers a comprehensive summary on the different applications and algorithms of subspace clustering. Domeniconi et.al [3] proposed an algorithm that discovers clusters in subspace spanned by different combinations of dimensions via local weighting of dimensions. The method associates to each cluster in weight vector, whose values capture the relevance of dimensions within the corresponding clusters.

Liping [16] proposed a new K-Means type algorithm for clustering high-dimensional objects in subspaces. They extend the K-Means by adding an additional step to the K-Means which automatically compute the weights of all dimensions in each cluster and use the weight values to identify the subsets of important dimensions that categorize different clusters.

3. DATA COMPLEXITY

Data complexity is a broad term that comprises issues such as overlapping, lack of representative data, small disjuncts, and others. It is the primary determining factor of clustering deterioration, which in turn, is amplified by the addition of class imbalance. In many practical domains, particularly in medical domains, the problem of class imbalance in the data is a common challenge facing machine learning methods. The situation gets worse when the size of minority class is very small [21].

To highlight the implications of the imbalanced learning problem in the real world, the following biomedical application is considered. The "Mammography Data Set," a collection of images acquired from a series of mammography exams performed on a set of distinct patients, which has been widely used in the analysis of algorithms addressing the imbalanced learning problem [10, 11,13].

Analyzing the images in a binary sense, the natural classes (labels) that arise are "Positive" or "Negative" for an image representative of a "cancerous" or "healthy" patient, respectively. From experience, one would expect the number of noncancerous patients to exceed greatly the number of cancerous patients; indeed, this data set contains 10,923 "Negative" (majority class) samples and 260 "Positive" (minority class) samples. Preferably, we require a classifier that provides a balanced degree of predictive accuracy (ideally 100 percent) for both the minority and majority classes on the data set [13].

The class imbalances can occur in different forms. Some forms of class imbalances are: between-class imbalance, within–class imbalance, relative imbalance and imbalance due to rare instances. In between-class imbalanced datasets, one of the classes is largely under-represented in comparison to others and usually this type of imbalances is innately binary (or two-class). It is noted that, there are multiclass data in which imbalances exist between various classes. The within-class imbalance concerns itself with the distribution of representative data for sub concepts within a class.The existence of within-class imbalances is closely intertwined with the problem of small disjuncts [13], which has been shown to greatly depreciate clustering performance. The following are the few datasets which are imbalanced in nature: Breast cancer data [10], Bupa Liver disorders [9], Pima Indian's diabetes database [9, 11], SPECTF Heart disease [15] and Spambase dataset [22] with binary class imbalance Dermatology [12] and Thyroid dataset [11] with multiclass imbalance. The binary class imbalance ratios of the datasets are tabulated in Table 1.

Table 1. Imbalance Ratio of Binary Class Datasets

Dataset	*Positive Samples*	*Negative Samples*	*Class Distribution*
Breast Cancer	241	458	0.34 : 0.66
Bupa	200	145	0.42 : 0.58
Pima Diabetes	268	500	0.35 : 0.65
SPECTF Heart	55	212	0.20 : 0.79
Spambase	1813	2788	0.39 : 0.60

4. VISUALIZATION

Visualization is defined by Ware as "a graphical representation of data or concepts" which is either an "internal

construct of the mind" or an "external artifact supporting decision making". Visualization is typically employed as an observational mechanism, to assist users with intuitive comparisons and better understanding of the studied data. Instead of precisely contrasting clustering results, the visualization techniques employed in cluster analysis; focus on providing the user with an easy and intuitive understanding of the cluster structure and to explore clusters randomly [6]. Due to this, visualization models have received great attention in the field data clustering, validation and result evaluation.

Visualization techniques could enhance the current knowledge and data discovery methods by increasing the user involvement in the interactive process. To incorporate visualization techniques, the existing clustering algorithms use the result of clustering algorithm as the input for visualization system. The drawback of such approach is that it can be costly and inefficient. The better solution is to combine two processes together, which means to use the same model in clustering and visualization. Interactive clustering allows the user to be involved into the clustering and visualizing process via interactive visualization [3, 4, 7].

4.1 Taxonomy of Visualization

Visualization has been categorized into major areas [19, 20]:

- Scientific visualization which involves scientific data with an inherent physical component.
- Information visualization – which involves abstract nonspatial data.

4.1.1 Scientific Visualization

Scientific visualization focuses primarily on physical data such as human body, the earth, molecules and so on. It also deals with multidimensional data, but most of the datasets used in this field use the spatial attributes of the data for visualization purpose; e.g. Computer Aided Tomography (CAT) and Computer Aided Design (CAD). Also, many of the Geographical Information Systems (GIS) use either the cartesian coordinate system or some modified geographical coordinates to achieve a reasonable visualization of the data.

4.1.2 Information Visualization

It focuses on abstract, nonphysical data such as text, hierarchies and statistical data. Data mining techniques are primarily oriented toward information visualization. The challenge for nonphysical data is in designing a visual representation of multidimensional samples. Multidimensional information visualization presents data that are not primarily plenary or spatial. Keim's [20] classification of the information visualization techniques are as follows;

Geometric Projection Techniques

Geometric Projection techniques aim at finding "interesting" projections of multidimensional datasets. The class of geometric projection techniques includes techniques for exploratory statistics such as Principal Component, Factor Analysis and multidimensional scaling, many of which are subsumed under the term "projection pursuit". Parallel coordinate visualization techniques and Radial Visualization (RadViz) also belong to this category of visualization.

Icon based Techniques

The idea of icon based technique or iconic display is to map each multidimensional data item into an icon (or glyph) whose visual features vary depending on the data values. Some of the most commonly used iconic displays are Chernoff, Stick Figure, Star Display, Shape coding, etc.

Pixel oriented Techniques

Pixel oriented techniques map each data values to a coloured pixel and present the data values belonging to one attribute in separate window. All pixel oriented techniques partition the screen into multiple windows. For data sets with m dimension, the screen is partitioned into m windows: one for each of the dimensions.

Hierarchical and Graph based Techniques

The hierarchical techniques subdivide the m-dimensional space and present the subspaces in a hierarchical fashion. Well known representatives of hierarchical techniques are n-Vision technique, the dimensional stacking, and treemaps [7]. The basic idea of the graph based techniques is to effectively present a large graph using specific layout algorithms, query languages and abstraction techniques.

4.2 Knowledge Assisted Visualization

Though Researchers have attempted to clarify the taxonomy of terms used in the visualization community, Min Chen et. al [7] attempts to offer a different taxonomy for visualization, that provides a new dimension on visualization processes. Min Chen et. al., [7] describes the visualization process as follows: Given a data set C_{data} and control information C_{ctrl}, user first makes decisions about which visualization tools to use for exploring the data set. The user then experiments with different controls, such as styles, layout, viewing position, color maps, and transfer functions, until the user obtains a satisfactory collection of visualization results, $C_{image.}$. The typical visualization process described by Min Chen et. al., is shown in Figure 1. For complex visualizations, interaction alone often can't reduce the search space rapidly. So, it is essential to advance the visualization technology, from today's interactive visualization to tomorrow's knowledge-assisted visualization. Based on these perspective, Min Chen et. al., [7] defined three types of visualization: Data, Information and Knowledge visualization.

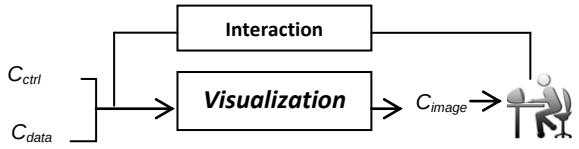


Fig 1: Typical Visualization Process

In information-assisted visualization, the system provides the user with a second visualization pipeline, which typically displays the information about the input data set. But it can also present attributes of the visualization process, the properties of the results, or characteristics of the user's perceptual behaviors. The user uses such information to reduce the search space for optimal control parameters, hence making the interaction much more cost effective.

Knowledge assisted visualization include sharing domain knowledge among different users and reducing the burden on users to acquire knowledge about complex visualization techniques [7]. In a visualization process, the user's knowledge is an indispensable part of visualization. For instance, the user might assign specific colors to different objects in visualization according to certain domain knowledge. The user might also choose different viewing

positions because the visualization results can reveal more meaningful information or a more problematic scenario that requires further investigation. Researchers and developers often incorporate some general or domain knowledge into visualization systems, either intentionally or unintentionally. It also enables the visualization community to learn and model the best practice, so that powerful visualization infrastructures can develop and evolve. Both information and knowledge assisted visualization plays a major role in the process of data mining [7]. They not only used for visualizing the results, but also complement and steer the mining process.

5. KNOWLEDGE ASSISTED VISUAL CLUSTERING

In this section, a Knowledge Assisted Visual (KAV) clustering system is presented and it integrates the VISTA model [6] and soft subspace clustering as proposed in Liping, et al. [16]. The subspace clustering aims at finding clusters from subspaces of data instead of entire data space. The major challenge of subspace clustering, which makes it distinctive from traditional clustering, is the simultaneous determination of both clustering memberships of objects and the subspace of each cluster.

In a subspace clustering each cluster is a set of objects identified by a set of dimensions and different clusters are represented in different subsets of dimensions. The cluster memberships are determined by the similarities of objects measured with respective subspaces. According to the ways of clustering, the subspace clustering methods are divided as hard subspace clustering, in which exact subspaces of different clusters are identified and soft subspace clustering, where clustering is performed in the entire data space and different weight values are assigned to different dimensions of clusters during clustering process [16].

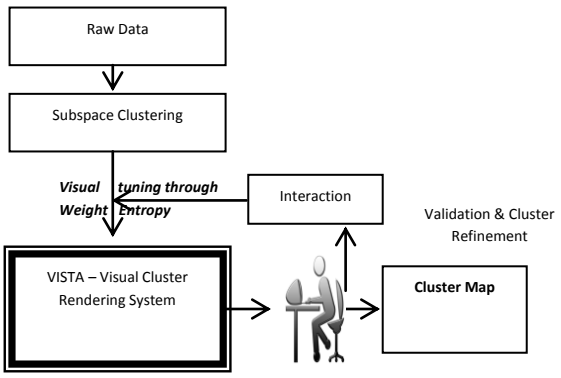


Fig 2: Knowledge Assisted Visual Clustering System

The aim of the proposed system is not just to show the utilization of the visualization, but to introduce knowledge assisted intelligent visual cluster rendering system, that identifies the contribution of individual dimension in forming the clusters as a measure of weight entropy, which is then used to refine the clustering process. Since every dimension contributes to the discovery of clusters, the dimensions with larger weights form the subsets of dimensions of the cluster.

The weight entropy of dimension in a cluster is inversely proportional to the dispersion of the values from the center in the dimension of the cluster. Since the dispersions are different in different dimensions of different cluster, the weight entropies for different clusters are different. The high weight indicates a small dispersion in a dimension of the cluster. Therefore, that dimension becomes more important in forming the cluster. The proposed Knowledge Assisted Visual clustering system applies this domain knowledge in finding the dominating dimension for visual tuning.

Initially the given data set is clustered by using centroid based hard clustering algorithm for instance K-Means [14], weighted K-Means [4] and so on. Based on the clustering results, the weight entropy is computed through soft subspace clustering. The clustered dataset is then explored in VISTA system, the validation, refinement and re-labelling is performed by the user based on the domain knowledge obtained from the weight entropies of individual dimension and his visual perception. The framework of the proposed system is depicted in Figure 2. The step by step procedure is described in Figure 3.

When the visual cluster rendering produces satisfactory visualization of well separated clusters, the boundaries of the clusters are drawn as a polygon. A unique identifier is assigned to each polygon and the data points which do not belong to any polygon is treated as outlier and merged into nearby clusters. For experimental purpose, the data objects are visualized in different coloured pixels. The colour is fixed based on their original cluster id.

Algorithm 1. Proposed Knowledge Assisted Visual Clustering

Input: p data sets with same underlying distribution
Output: partitions of m datasets
Procedure:
Step 1: Cluster each datasets by K-Means / Weighted K-Means algorithm and obtain class labels.
Step 2: Apply the following formula to find the weight entropy of dimension for each cluster.

$$\lambda_{li} = \frac{1}{\sum_{t=1}^{m}\left[\frac{\sum_{j=1}^{n}\omega_{lj}^{\eta}(z_{li}-x_{ji})^2}{\sum_{j=1}^{n}\omega_{lj}^{\eta}(z_{lt}-x_{jt})^2}\right]^{1/(\beta-1)}} \quad (1)$$

λ_{li} = weight for the i[th] dimension in the l[th] cluster.
ω_{lj} = degree of membership of the j[th] object belonging to the l[th] cluster.
x_{ji} = value of i[th] dimension in the j[th] object.
z_{li} = value of the i[th] component of the l[th] cluster.
β (>1) and $\eta(\geq)$ are two parameters greater than 1. subject to

$$\sum_{i=1}^{n}\omega_{lj}=1, \quad 1\leq j\leq n \quad (2)$$

since K-Means is hard clustering, $\omega_{lj}=1$ if j[th] object belongs to the l[th] cluster else =0

$$\sum_{i=1}^{m}\lambda_{li}=1, \quad 1\leq l\leq k, \quad 0\leq\lambda_{li}\leq 1 \quad (3)$$

Step 3: Explore the dataset with class labels in VISTA model.
Step 4: If the clusters are clearly distinguished from others perform free hand drawing and go to step 7 else
Step 5: Perform α- tuning using the following visual rendering rules. Omit the dimension whose weight entropy is less than γ, since it is the non-contributing dimension. Maximize the α parameters of dimensions with weight entropy greater than γ to either 1 or -1, to separate all clusters as possible and polish them.
Step 6: Re cluster the data points which are dense as new cluster.
Step 7: Label the outliers as members of the nearby clusters.

Fig 3: Step-by-step procedure for Knowledge Assisted Visual Clustering System

The weight entropy for a dimension in a cluster is inversely proportional to the dispersion of the values from the center in the dimension of the cluster. The high weight entropy indicates a small dispersion in a dimension of the cluster. Therefore, that dimension is more important in forming the cluster. Similarly, low weight entropy indicates large dispersion in the cluster structure. Hence, the proposed KAV system identifies the major, minor and non-contributing dimensions through the weight entropies and utilizes that domain knowledge for further refinement. This makes the visual clustering to be more meaningful and flexible and this kind of domain knowledge is very important to perform an actionable knowledge discovery.

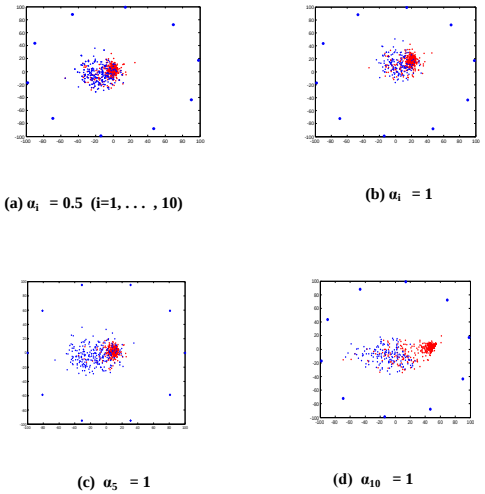
From the experiments, it is observed that dimensions which have least weight entropy i.e. entropy less than γ moves all data points together in a single direction, do nothing with the visualization and dimension with high weight makes a small dispersion either towards the center of the cluster or away from it. Hence, this dimension is considered as important in forming the cluster, whereas the dimension which makes a large dispersion in the cluster are non-contributing ones, which usually has skewed distribution. While performing the visual clustering with dimensions of higher weight entropies it is observed that, the clusters may separate as possible and increase the cluster cohesion. Similarly dimension with minimum weight entropy does nothing to visualization.

6. EXPERIMENTAL ANALYSIS AND DISCUSSION

In this section, the performance of proposed KAV system is evaluated with various benchmark datasets, in terms of two external validity measures: Rand index and Jaccard index, two internal validity measures DBIndex and H/S ratio. For experimental purpose the initial α values are set as 0.5, and the α variation is set as 0.001. The proposed KAV system introduces a knowledge assisted visual clustering framework that integrates efficient domain knowledge into the visual clustering process. Initially the data set is clustered by using K-Means or Weighted K-Means algorithm. From the resultant cluster distribution, the weight entropies of individual dimensions for each resultant clusters are computed. The cluster labels and weight entropies are further used as domain knowledge to identify the underlying pattern of the dataset, and enrich the clustering results.

The data objects belong to different clusters are explored in KAV with different colours and α tuning is performed based on the weight entropies of individual dimensions until satisfactory results are obtained. The original VISTA framework performs the visual clustering by sequential rendering each dimension, whereas the proposed KAV system identifies the visually dominating dimensions through domain knowledge based on weight entropy of individual dimension, thus increases the scalability of visual clustering process. The dimensions with high weight entropy identify the imbalanced nature of the dataset.

For instance, the following dimensions are identified as major contributing dimensions for Breast Cancer dataset: 10, 6, 3, 8, 5, 9, 4, 2, 7 and dimension 1 identified with least weight entropy, hence considered as non-contributing dimensions. During visual clustering process, when the α value of dimension 1 changes, it is observed that, the data points move in single direction, do nothing in visual rendering. The dimensions 10, 6, 3 are found to be with high entropy values and with the continuous change in α of dimension 10, it is observed that the data points of the two clusters move closer and increase the clusters cohesion. Similar results are obtained when the α values of dimensions 6, 3 change and so on. Thus, they agree with the major and minor contributing dimensions as proposed in VISTA model. While clustering the given dataset using K-Means (KM) and weighted K-Means(WKM) the following observations are recognized. Since the dispersions are different in different dimensions of different cluster, the weight entropies for different clusters are different. Each cluster is formed by different subsets of dimensions. For example, for given dataset the weight entropies of dimensions for cluster 1 need not be the same for cluster 1 and so on. The visual clustering on Breast Cancer dataset is shown in Figure. 4.



(a) $\alpha_i = 0.5$ (i=1, . . . , 10) (b) $\alpha_i = 1$

(c) $\alpha_5 = 1$ (d) $\alpha_{10} = 1$

Fig 4: Visual Clustering Process

The main objective of this experiment is to analyze the performance of the proposed KAV system against automatic clustering algorithms KM and WKM for ten different datasets. The best and average results of KAV in comparison with the results of KM and WKM for few imbalanced datasets, in terms of Rand Index, Jaccard coefficient, DBIndex and H/S ratio are shown in Table.2 and Table.3 respectively.

The scrutiny of the external validity measures reveals the fact that KAV wins in all datasets. The experimental results show that KAV works well in identifying the imbalanced nature of the dataset, irregularly shaped clusters and provide an efficient human-computer interaction. More specifically, the results of the proposed KAV system provide highly appreciable results for Breast Cancer, Dermatology, SPECTF Heart and Wine datasets. While considering internal validity measures, the proposed KAV system loses in all datasets.

From Table 2, it is observed that, the proposed KAV system yields better result than KM and WKM in terms of the best values of Rand Index, for all the datasets. It also provides consistent results for Breast Cancer, Bupa, Dermatology and Pima Diabetes datasets. It is also observed that the variation between best and average results of proposed KAV system for SPECTF Heart, Spambase and Thyroid datasets are high when compared with KM and WKM.

Since the α tuning is carried out in the sequential order in VISTA, variation between the averages and best is high for all datasets, whereas the KAV system starts the α tuning from dimensions with large weight entropy, the best and average result yields consistent values. Owing to the severity of imbalance and existence of highly overlapped clusters, the cluster quality becomes highly inconsistent for SPECTF Heart, Spambase and Thyroid dataset in KAV system. When the average result of Rand Index is compared, the following observations are identified: the proposed KAV system yields better result than KM, WKM for all datasets except Thyroid dataset. While considering Jaccard coefficient it is observed that the proposed KAV system yields better results than the existing methods. More importantly, it outperformed for Breast Cancer and SPECTF heart datasets.

The clustering results based on DBIndex are tabulated in Table.3. From the results, the following observations are recognized. The proposed KAV system performs worse than KM and WKM algorithm for almost all the datasets except Breast Cancer. It is also noted that, the consistent rate of KAV system is high for all datasets except Bupa, Dermatology and Wine datasets. From the H/S ratio, it is observed that the performance of the proposed KAV system is worse than KM and WKM for all datasets except for Breast Cancer Dataset. The variation between best and average results is also high in KAV for all datasets except Dermatology, Mammography dataset.

It is noted that the cluster quality obtained from all the methods are worse for the following datasets; Bupa Liver dataset, Pima and Thyroid datasets. As these datasets exhibit class imbalance and are hard to process, the automatic algorithms fail to properly represent the distributive characteristics of the data and resultantly provide unfavourable accuracies across the classes of the data.

In general, any cluster distribution with small value of DBIndex and H/S ratio suggests better quality. Similarly any method that increases the Rand and Jaccard index suggested being an efficient method. Although these statistical indices were proved effective in determining the optimal number of well separated spherical clusters, they do not work well for arbitrarily shaped clusters. Even if it is possible to find appropriate indices to deal with certain shapes, the statistical methods are not flexible enough to adapt any unanticipated shapes, especially sketch of clusters is not known earlier. However, the visualization of these datasets able to represent the imbalance nature, but the statistical validity measures greatly depreciates the cluster quality. Since K-Means and its variants work well in identifying convex shaped clusters, their external validity increases as internal validity decreases. But while considering KM, WKM and KAV system, it is observed that there exists a greater variation in internal and external cluster validity measures. In such cases, visual perception may play a vital role, and it is also believed that visual ability is better than the statistical methods especially in dealing with arbitrarily shapes [6].

Table.2: Comparison for KAV based on External Validity Measure

Dataset	Rand Index						Jaccard Coefficient					
	KM		WKM		KAV		KM		WKM		KAV	
	Avg.	Best	Avg.	Best	Avg.	Best	Avg.	Best	Avg.	Best	Avg.	Best
Breast Cancer	0.5209	0.5209	0.7548	0.7548	0.9294	0.9364	0.4141	0.4141	0.6644	0.6644	0.8776	0.8891
Bupa	0.5058	0.5058	0.5015	0.5015	0.5118	0.5236	0.4555	0.4555	0.4320	0.4320	0.4457	0.4457
Dermatology	0.7033	0.7189	0.7418	0.7418	0.8374	0.8507	0.1350	0.1460	0.2191	0.2191	0.5043	0.5588
Mammography	0.5574	0.5579	0.5706	0.5706	0.6060	0.6591	0.3928	0.3932	0.4182	0.4182	0.4736	0.4932
Pima Diabetes	0.5389	0.5513	0.5234	0.5234	0.5545	0.5672	0.4282	0.4584	0.3905	0.3905	0.4739	0.5459
SPECTF Heart	0.5938	0.5938	0.6133	0.6133	0.6919	0.8524	0.5801	0.5801	0.6016	0.6016	0.6774	0.8524
Spambase	0.5370	0.5370	0.5331	0.5331	0.6150	0.7062	0.5083	0.5083	0.5197	0.5197	0.5243	0.5606
Thyroid	0.5299	0.7069	0.6071	0.6071	0.5497	0.7103	0.4987	0.4987	0.5290	0.5290	0.5342	0.7048

Table.3: Comparison for KAV based on Internal Validity Measure

Dataset	DBIndex						H/S Ratio					
	KM		WKM		KAV		KM		WKM		KAV	
	Avg.	Best	Avg.	Best	Avg.	Best	Avg.	Best	Avg.	Best	Avg.	Best
Breast Cancer	0.4668	0.4668	1.2082	1.0282	0.0872	0.0722	0.1533	0.1533	0.5707	0.5707	0.4888	0.3547
Bupa	0.7679	0.7679	0.8570	0.8570	2.3422	1.5204	0.1747	0.1747	0.1970	0.1970	0.4373	0.3194
Dermatology	1.1444	1.0622	1.2231	1.2231	3.8017	2.9474	0.1798	0.1768	0.1998	0.1998	0.2941	0.3241
Mammography	0.6208	0.6199	0.6417	0.6417	1.9704	0.3549	0.2129	0.2126	0.2225	0.2225	0.4383	0.4198
Pima Diabetes	0.7368	0.7134	0.7599	0.7599	1.9863	1.0303	0.2154	0.2010	0.2299	0.2299	0.4773	0.4304
SPECTF Heart	1.4034	1.4034	1.3729	1.3729	1.9001	1.4140	0.2845	0.2845	0.2756	0.2756	0.4616	0.4119
Spambase	0.5856	0.5856	0.6113	0.6113	1.336	1.2874	0.0697	0.0697	0.1164	0.1164	0.3065	0.2190
Thyroid	1.0815	0.9720	1.2491	1.2491	2.8377	2.4955	0.2823	0.2604	0.3119	0.3119	0.3756	0.3464

7. CONCLUSION

In this paper, the framework of the proposed Knowledge Assisted Visual cluster rendering system is described. The comparative analysis of clustering results shows that the proposed KAV system works well in finding cluster distribution specifically for imbalanced dataset. Empirical evidence is provided for the proposed KAV system that exhibits better quality clusters than the K-Means, Weighted K-Means. Although the proposed Knowledge Assisted Visual cluster rendering system provides suitable visual impact on imbalanced datasets, more importantly, the knowledge assistance provided by the proposed system becomes more efficient in performing visual clustering by reducing computational complexity and making the human-computer interaction easy. It also provides visual clues regarding the cluster distribution, specifically while dealing with overlapped cluster distribution, and the visual tuning based on weight entropy outperformed in finding the cluster boundary. Depending upon the characteristics of individual dimension, the visual tuning process may result in best clusters. Method which identifies potential clusters automatically can be applied with the proposed framework, in future, to still reduce human errors and improve the efficiency of resultant cluster models.

8. ACKNOWLEDGMENTS

This research work is partially supported by special assistance program of University Grant Commission, New Delhi, India. [Grant No. F3-50/2011-SAP-II].

9. REFERENCES

[1] Alagambigai, P., Thangavel, K., "Visual Clustering through Weight Entropy," International Journal on Data Mining, Modelling and Management, Vol. 2(3), pp. 196-215, 2010.

[2] Alagambigai, P., Thangavel, K., Karthikeyani Vishalakshi, N, "Entropy Weighting Feature Selection for Interactive Visual Clustering," In: Proceedings of 4th International Conference on Artificial Intelligence, pp. 545-557, 2009.

[3] Ankerst M., Breunig M., Kriegel H. P., Sander J.,"OPTICS: Ordering Points To Identify the Clustering Structure," In: Proceedings of ACM SIGMOD '99, International Conference on Management of Data, Philadelphia, pp. 49-60, 1999.

[4] Ashok Kumar, "Intelligent Partitional Clustering," Ph. D Thesis, Gandhigram Rural University, Gandhigram, Tamil Nadu, India, 2007.

[5] Barbara. D., Chen. P., "Using the fractal dimension to cluster dataset", KDD'00 proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 260- 264.

[6] Chen K., Liu L., ''VISTA: Validating and Refining Clusters via Visualization," Information Visualization, Vol. 3(4), pp. 257-270, 2004.

[7] Chen M., Ebert D., Hagen H., Laramee R.S., Van Liere R., Ma K., Ribarsky W., Scheuermann G., Silver D., "Data Information, and Knowledge in Visualization," IEEE Computer Graphics and Applications, Vol. 29(1), pp. 12-19, 2009.

[8] Domeniconi, .C, Papadopoulos, P., Gunopulos, D., Ma, S.,"Subspace Clustering of High Dimensional Data. Proc. SIAM Int'l Conf. Data Mining, 2004.

[9] Doucette J., Heywood M. I., "GP Classification under Imbalanced Data Sets: Active Sub-Sampling AUC Approximation," LNCS, Vol. 4971, pp. 266-277, 2008.

[10] Estabrooks A., Jo T., Japkowicz N., "A Multiple Resampling Method for Learning from Imbalanced Datasets," Computational Intelligence, Vol. 20(1), pp. 18-36, 2004.

[11] Fernandez A., del Jesus M.J., Herrera F., "Multi-class Imbalanced Datasets with Linguistic Fuzzy Rule based Classification systems based on Pairwise Learning," Computational Intelligence for Knowledge-Based Systems Design, LNCS, Vol. 6178/2010, pp. 89-98, 2010.

[12] Fernandez A., del Jesus M.J., Herrera F., "On the Influence of an Adaptive Inference System in Fuzzy Rule Based Classification Systems for Imbalanced Datasets," Expert Systems with Applications, Vol. 36, pp. 9805 - 9812, 2009.

[13] He H., Garcia E. A., "Learning from Imbalanced Data," IEEE Transactions on Knowledge and Data Engineering, Vol. 21(9), pp. 1263-1284, September 2009.

[14] Jain, A.K., Murty, M.N., Flynn, P.J., "Data Clustering : A Review", ACM Computing Surveys, (1999).

[15] Jeatrakul P., Wong K. W., Fung C. C., Takama Y., "Misclassification Analysis for the Class Imbalance Problem," World Automation Congress (WAC) 2010, pp. 1-6, Sept 19-23, 2010.

[16] Jing L., Michael Ng K., Huang J. Z, "An Entropy Weighting K-Means Algorithm for Subspace Clustering of High-Dimensional Sparse Data," IEEE Transactions on Knowledge and Data Engineering, Vol. 19(8), pp. 1026-1041, 2007.

[17] Kandogan E., "Star Coordinates: A Multi-dimensional Visualization Technique with Uniform Treatment of Dimensions," IEEE Symposium on Information Visualization, Salt Lake City, Utah, pp. 4-8, 2000.

[18] Kandogan E., "Visualizing Multi-dimensional Clusters, Trends and outliers using star Co-ordinates," In: Proceedings of ACM KDD, 2001.

[19] Keim D. A, Hans-Peter, Kriegel, "Visualization Techniques for Mining Large Databases: A Comparison," IEEE Transactions on Knowledge and Data Engineering, Vol. 8(6), pp. 923-938, 1996.

[20] Keim, D. A., "Information Visualization and Visual Data Mining," IEEE Transactions on Visualization and Computer Graphics, Vol. 7(1), pp. 1-8, 2002.

[21] Klement W., Wilk S., Michalowski M., Matwin S., "Classifying Severely Imbalanced Data," Advances in Artificial Intelligence, LNCS, Vol. 6657, pp. 258-264, 2011.

[22] Liu Y., An A., Huang X., "Boosting Prediction Accuracy on Imbalanced Data Sets with SVM Ensembles," LNAI, Vol. 3918, pp. 107-118, 2006.

[23] Marie desJardins, James MacGlashan, Julia Ferraioli.: "Interactive visual clustering. Intelligent User Interfaces" , pp. 361-364, (2007).

[24] Sourina O., Liu D., "Visual Interactive 3-Dimensional Clustering With Implicit Functions," In: Proceedings of the IEEE Conference on Cybernetics and Intelligent Systems, Vol. 1, pp. 382-386, 1-3 December 2004.

[25] Wang. C., Ma. K., "Information and Knowledge assisted analysis and visualization of large-scale data", Proceedings of Ultrascale Visualization, 2008, UltraVis 2008.

[26] Zhang K. B., "Visual Cluster Analysis in Data Mining", Ph.D, Thesis, Department of Computing, Division of Information and Communication Sciences Macquarie University, NSW 2109, Australia, 2007.

[27] Zhang K. B., Orgun M. A., Zhang K., "HOV^3: An Approach for Visual Cluster Analysis," In: Proceedings of the 2nd International Conference on Advanced Data Mining and Applications (ADMA 2006), Xian, China, LNCS, Springer Press, Vol. 4093, pp.316-327, August 14-16, 2006.