International Journal of International Journal of Control and Automation
Vol. 12, No. 4, (2019), pp. 170-188

Survey on Classification and Clustering Schemes in Big Data using Image Processing

Navneet Kaur, Dr. SahilVerma*, Dr. Kavita, AnupLalYadav
Lovely Professional University, Jalandhar - Delhi G.T. Road, Phagwara, Punjab, India, navneetphul@gmail.com
Lovely Professional University, Jalandhar - Delhi G.T. Road, Phagwara, Punjab, India, sahilkv4010@yahoo.co.in
Lovely Professional University, Jalandhar - Delhi G.T. Road, Phagwara, Punjab, India, dhullkavita@yahoo.in
SRMIET, Bhurewala, Naraingarh, Ambala, Haryana, India, anupsaran@gmail.com
**corresponding author*
sahilkv4010@yahoo.co.in

Abstract

The various forms of voluminous data havelead to evolution of a variety of techniques forclassification and clustering in machine learning. This paper provides a brief description on big data, its background, evolution, its characteristics, an ampleoverview of different classification and clustering methods in machine and a literature review on different clustering and classification schemes in image processing that have been used so far in the recent years. Further, a taxonomy based on classification and clustering schemes has been presented along with its evolution in the recent years, and comparison of different papers. Since, images form a significant part of big data, so classification and clustering in images is the primary focus of this paper. The described research efforts will help in the identification of severaltechniques in classification and clustering. Also, few research directions have been presented in the field of classification and clustering in images.

Keywords: *Big data,Classification, Clustering, Image classification and clustering.*

1.1. Introduction

With the onset of cloud computing and internet era, the different applications in the field of video, audio, text are creating huge data. This massive volume of data refers to big data which means, simple words, extremely large datasetThis can be popularly associated with business or market data. According to International Data Corporation(IDC) in 2011[12], the creation and duplication of data will be 35 Zetabytes by 2020. Massive volumes of both structured and unstructured data are developing and evoluteseveryday.Big data is random and impetuous. How is it possible to analyze such data having the characteristics of randomness and spontaneity (as the data is generated and accumulated in case of stock markets which keep on changing from every single minute.Hence, such origination of data requires to be methodized clearly.The huge volume of data not only causes

problems in understanding, but also creates problems in generating the appropriate results based on the data set. Summarization of data is the key point here.

1.1.1 Motivation

After evaluating the recent research in classification and clustering techniques, there was a requirement for investigating the available literature for this topic. Motivation and novelty has been reflected in this section.

(1) The role of classification and clustering is to organize the data taking into consideration the kind of data and what kind technique needs to be applied.
(2) The existing classification and clustering algorithms have been categorized as support vector machines, decision trees classification and hierarchical, partitioning, density-based, grid based and ordered clustering respectively.
(3) Recent techniques in image processing
(4) Impending research directions for image clustering have been discussed.

Various quality journals have been referred for the purpose. This article has been organized in to different sections. Section 2 describes about the background of big data. Section 3 discusses about the review method used, planning, selection process, information sources, search principle, and selection and omission principle. Section 4 describes about the classification and clustering techniques in general and images in particular. Section 5 refers to conclusion and impending research directions.

1.2 Background of Big Data

The big data [13] is nothing but bulky repository of unstructured data. And this data is multiplying at a very faster rate as compared to decades before. The available data today is generated in a variety of forms as image, video, and text.As already mentioned, both structured and unstructured data [24] are produced everyday. We also have category known as semi-structured data. Structured data is something we generally expect in a database. The unstructured data is available in the wild. It comprises text, pictures, all formats of audio and video components The semi-structured data does not completely follow the rules of a database relational model but some of their features make it easier to organize them. Maximum of the content found on web can is the semi-structured data. Fig. 1.1 clearly depict the types of data.

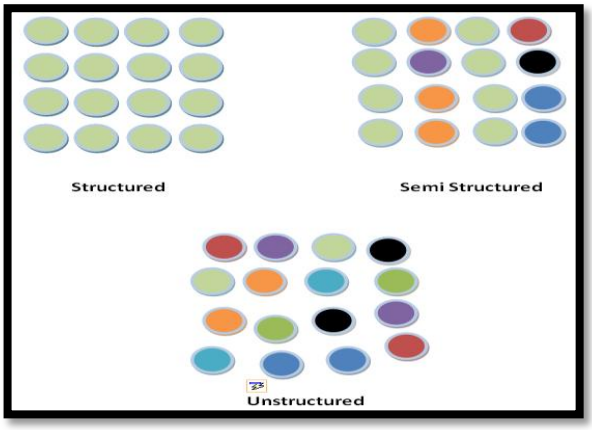

Figure 1: Types of data

1.2.1 Evolution of big data and its techniques

The invention of internet, developments in IOT and cloud computing has been the primary cause for the explosive expansion of data. The classification of big data can be done in two categories. The first category can be considered as the data that is generated from the physical world(the examples may include sensor data, experiments of science and its observations. The second category comprises the data which is obtained from the society itself(examples include social media, finance, healthcare, economics, transportation).

1.2.2 Big data 5Vs

A mention of features of big data is necessary here. The characterization of big data can be done as 5V- Volume, Velocity, Variety, Veracity, and Value [15].

- Volume refers to the massiveness of data which is multiplied or generated. These days distributed systems are being used for storing such massive amounts of data.
- Velocity refers to the rateat which the data is multiplied, accumulated, and evaluated. The analysis of big data can be done even when the data is being generated without storing the data into the database.
- Variety refers to the different kinds of data. The most common form of data now-a-days is unstructured. Structured data is now a thing of the past. Unlike unstructured data, structured data fits well into the table. Photos, video, audio and social media are common examples of unstructured data.
- Veracity refers to the quality of data. In other words, it is accuracy and that whether the data is trustworthy or not. This massive amount data generate everyday is of no use unless it is accurate, trustworthy, and reliable. A very good example of veracity is GPS. Whenever the GPS component loses signals, the location data from some other source needs to be merged with the actual data so as to provide an accurate location.

- Value refers to nothing but how much the data is useful. In other words, it refers to its worth. Turning such huge amount of data into a useful one is challenge here. So, it is very important to go through the costs and benefits of accumulating and evaluating the data, so that after extracting and analyzing whatever we achieve should be worthy and useful.

1.2.3 Big data characteristics

- Heterogeneity – The origination of data is from several sources that may comprise different formats. This is what gives rise to unstructured and semi-structured data. Examples comprise social media and instant messengers.
- Complexity – The massiveness or multiplicity of data refers to big data complexity. Complexity is
- Evolving – Big data evolves or changes at a very fast rate. The typical example is of the stock market which changes every second.

Due to the immense growth in the data over the decades, there was a need to group the data sensibly in an organized way. The variety of data makes this task cumbersome. The techniques for methodizing the data are classification and clustering. Classification is known as supervised learning(in which human intervention is not necessary). In other words, the items are labeled using which they can be easily classified. Clustering is known as unsupervised learning(in which human intervention is necessary) in which the items are unlabeled and hence are difficult to classify. An extremely popular technique is K means clustering. A middle approach is called as semi supervised or partially supervised learning[8]. Fig.1.2 depicts the simple classification and clustering concept.

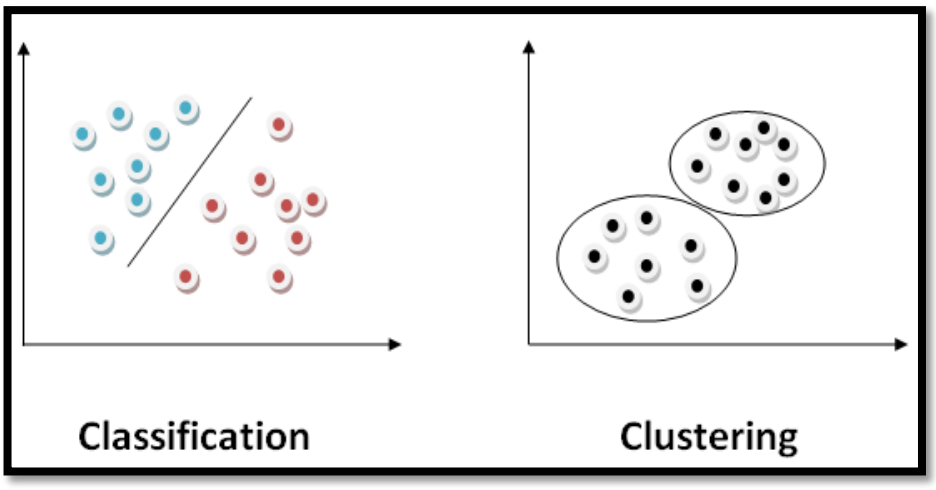

Figure 2: Classification and clustering

1.3 Review Method

The objective of this paper is to give its readers a go-through analysis of various techniques in clustering and classification. No simulation tools are referred here.

1.4 Review planning

The strategy comprises the examination of several databases, the analysis of current techniques. The study involves the sources from online databases with manual searches on various journals and conference proceedings. It involves some amount of omission principle so as to identify the important studies to elicitate the data.

1.5 Study and selection process

Keywords search has been used in the filteration of the search procedure. The search includes research papers on classification and clustering and its several challenges. The appropriate research papers were searched from various online databases and the question that are required to understand.

1.6 Information sources

The appropriate information was found out from a wide range of publications. The online databases used were:

- Springer
- IEEE transactions
- Science Direct
- Elsevier
- Google Scholar

1.7 Search principle

The basic search principle involved “Big data”, “Data mining”, “clustering”, “”classification”, “Big data in machine learning”, “Data mining in machine learning”, “Image clustering”. “Image classification”, “Image clustering and classification”.

1.8 Selection and omission principle

Some of the research papers were omitted as they didn’t provide enough knowledge regarding the concerned topic. Many of these required to be omitted in order to keep this article confined to big data clustering, classification, image classification and clustering.

1.9 Classification

There was a need for the classification and its techniques to come into survival taking into account the trending big data. Organized construction of data is required so that it can be accessed with minimal difficulty level. The two extensive categories of classification comprise – Supervised classification, and unsupervised classification. Classification is performed in two parts. First one is the training part. In this part, a training dataset is provided and the results are to be obtained from this training dataset. It is used to find the class. Here, the outcome of every element is known in advance. By analyzing the items, we can conclude the labels(class) for data based on their features or characteristics. The second part in classification comprises the test part. This part refers to the unlabeled data.Unlike the training part, outcome is unknown for items, hence we have the unlabeled data. The different techniques in supervised learning are as stated and explained below. The processing in supervised classification is faster as compared to unsupervised classification. Fig.1.3 mentions the types of classification.

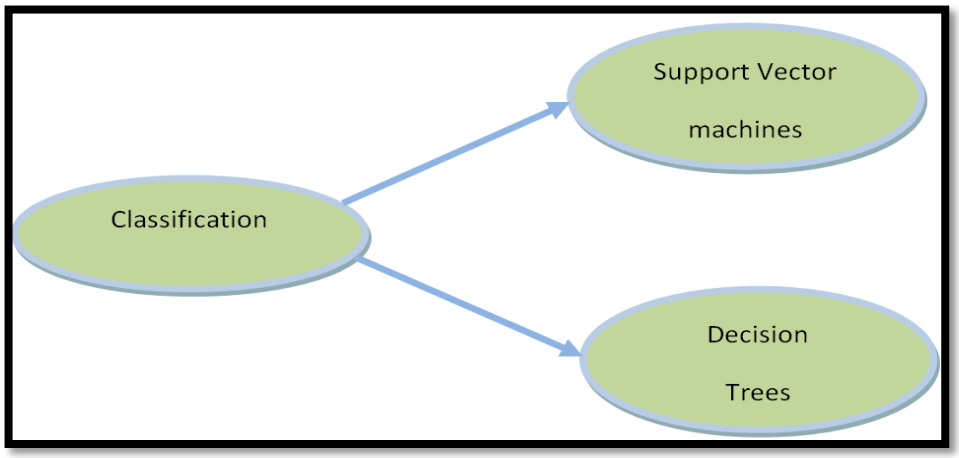

Figure 3. Types of classification Techniques

1.9.1 Support Vector Machine

Support Vector machines [2], also known as Support Vector Nerworksare a type of supervised learning approach. They are based on statistical learning theory. It was required to assess a function based on a collection of items which is known as the training dataset. In other words, to learn from training dataset. The technique was originally developed in 1992. The appropriate functions should be selected based on the training set. The complexity of these functions will determine the risk involved in the selection of the functions. The complexity is also determines the best function. If we consider, the items as points in space, the items can be made distinct with the help of a hyperplane. The items will belong to either side of the hyperplane depending on which category the items belong to. SVMs are widely used in categorizing text and image recognition.

1.9.2 Decision Trees

Decision trees [1] are used for the purpose of decision making or decision analysis. It is a representation in the form of a tree or a graph. Here, the nodes refer to the data where it divides and leaves refer to final decisions or answers. They can be classified as classification trees and regression trees. In classification kind of trees, the results are categorical while in the latter the results are continuous.

1.10 Clustering

Clustering is unsupervised learning. Dividing the similar data into distinct categories or groups refers to clustering.The similarity may be determined by their characteristics or behavior. Labels are allocated to these groups thereafter. Clustering finds a wide range of applications in digital image processing, data analytics, pattern recognition, biology etc.The massiveness of data generation in today's world has led to the development several clustering algorithms. To deal with big data we have single machine clustering techniques and multiple machine clustering techniquesTraditional or single machine clustering algorithms comprise hierarchical, partitioning, grid based, density based, and model based. Fig.1.4 depicts the types of clustering techniques.

- Hierarchical methods
  - Agglomerative algorithms
  - Divisive
- Partitioning relocation methods
  - Probabilistic clustering
  - *k*-medoids methods
  - *k*-means methods
- Density-based partitioning methods
  - Density-based connectivity clustering
  - Density functions clustering
- Grid-based methods

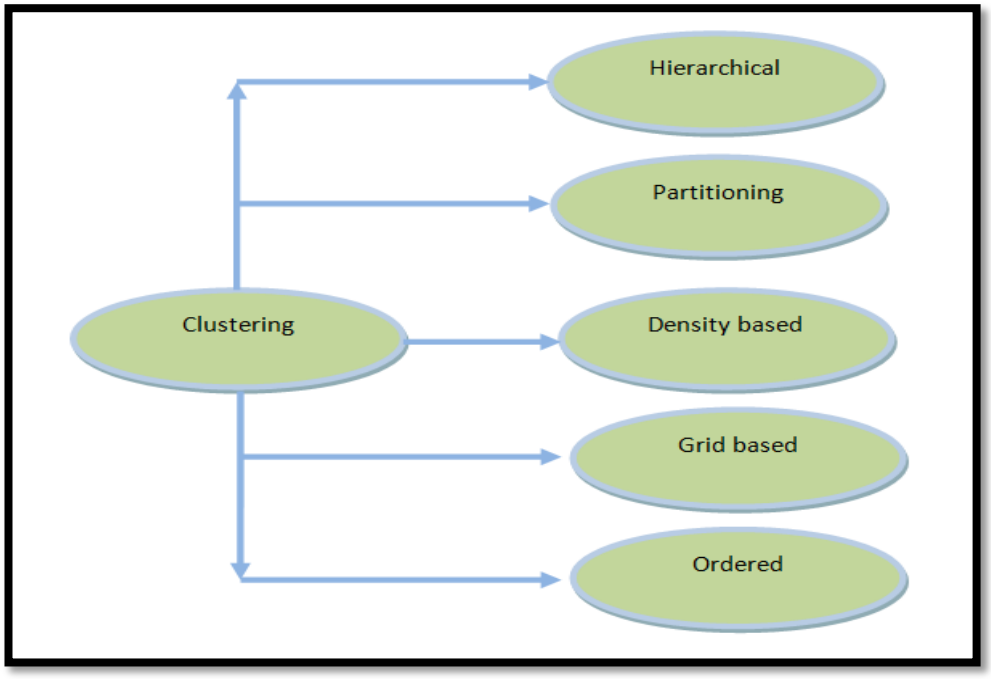

Figure 4. Types of clustering techniques

1.10.1 Hierarchical

This type of clustering is in the form of a tree referred to as the dendrogram. Here, the clusters are developed in a hierarchical manner. The two main approaches are bottom-up(agglomerative) and top-down(divisive), which represent the granularity level that the hierarchical clustering can handle.In agglomerative approach, the formation of clusters starts with one cluster and iteratively assimilating it into the similar clusters. The divisive approach too starts with a single cluster and iteratively dividing or splitting the cluster into relevant child clusters. An intrusion detection system was developed with a combination of hierarchical clustering and support vector machines[7]. The specific data are referred to as the leaf nodes. One of the disadvantages of hierarchical clustering is that the step performed cannot be reverted, whether it is merging or splitting. In other words, once the development of cluster takes place, we cannot go back in step. The other on is that sometimes it may be problematic to choose the basis for termination. The advantages comprise its adjustability and easy management.The examples of algorithms in this category are CHAMELEON and CURE.

1.10.2 Partitioning relocation method

The data is partitioned into multiple child clusters or small partitions. The classification of points is done on the basis of their similarities using a distance for classifying the points. There are some conditions that need to be taken into consideration for satisfying the criteria for partitioning. Every cluster must consist

of at least one item. K-means [3] and K-medoids are examples of algorithms used in this method.

1.10.3 Density based

In this type of clustering [5], density, boundaries, and connections are involved in the criteria for forming clusters. These algorithms find a significant role in identifying non-linear or arbitrary compositions based on density.In other words, clusters of arbitrary shapes can be identified using density based algorithms. Where ever the density leads to, the cluster formation matures in that direction. Two important terms form the foundation for the concept of density based algorithms.

-Density Reachability – A point 'a' will be called as density reachable from point 'b' if 'a' lies within a distance of$\in$ from 'b'. Also 'b' should have enough count of points in its neighbors which lie within a distance of '$\in$'.

-Density Connectivity – Points 'a' and 'b' are called as densely connected if there is an existence of 'c' having enough count of points in its neighbors. And also both these points should lie within a distance of $\in$.

The different algorithms under this category comprise DBSCAN, OPTICS, DBCLASD, and DENCLUE.

1.10.4 Grid based clustering

In this category of clustering [4], the division of data or space is performed into a number of grids or rectangular cells. These cells are equally sized. It basically comprises three steps - division of space into cells; removal of cells having low density; cluster formation by merging of adjacent cells with high density. Calculation is performed for the statistical cells. The fast processing duration is an advantage of grid based clustering.

1.10.5 Ordered Clustering using PROMTHEE method

The classical K means clustering does not takes into account the priority. In [20], the authors have adduced a technique for multi-criteria ordered clustering problems. The center of every cluster is calculated using the net outranking flow. Hence, the objective function refers to the sum of net outranking flow of all the alternatives.

1.11 Classification and clustering in images

Labeling of the image objects or components refer to classification while clustering refers to high level depiction of image composition or objects.

1.11.1 Object classification in images

Classification in images can be in a number of contexts such as from pixel classification, or prediction image or binary image. In [6], the authors have used BOV technique for object classification in HR(High Resolution) aerial images. The appearance of words is used as the training data set for BOV. Since the aerial images are characterized by their spatial and spectral composition, this technique has been proved to be very purposeful by using patch descriptors. Also, the outlier effect has been minimized with the help of a virtual word Comparisons show BOV technique outperformed low level features. Using the proposed technique, all types of simple, complex, and composite objects can be worked upon.

1.11.2 Semi-supervised classification

In [8], remote sensing images have been used as the input from which maps can be generated – both classification maps and confidence maps. These remote sensing images can be labeled using the proposed algorithm. Emphasis has been made to minimize the count of queries to acquire the classification results. Hierarchical segmentation has been used here. Dendrograms are obtained for minimal error classification. Both hyperspectral and multispectral images have been used as the dataset to perform the experiment. The proposed algorithm is very purposeful for various earth examination applications.

1.11.3Multiple Feature Learning

In order to classify the hyperspectral images, which usually comprises multiple features, a framework has been adduced in [16]. The adduced framework deals with both linear and non-linear features, and hence is flexible as compared to kernel learning. It doesn't require regularization parameters for weight control of features. Spectral and spatial features based on both original and kernel representations have been considered. The experimental results prove the framework, without provides better results without rise in the complexity. Also, it was found out that using kernel representation will boost computational complexity, so kernel features can be omitted for the proposed framework and can be considered only for specific applications. As a future scope of the adduced framework, inclusion of more relevant features can be considered. Development of the adduced framework can also be considered for GPUs and multi-GPU platforms.

1.11.4 Image clustering for site-specific management

In [21], wheat biomass estimation has been performed using image clustering and height of the crop. It is for observing the geographical modifications in the wheat biomass. For this purpose, digital images of fresh and dry biomass were captured using digital camera systems and instruments for measuring crop height were

used. NIR and NDVI images were used. Plant pixels and background pixels were disjointed from each other to estimate the plant coverage percentage in order to relate it with fresh and dry biomass. In experimental results, it was found out that NIR clustering was better than NDVI accurate estimates were found to be calculated by NIR clustering. But NIR was affected by scattering effects.

1.11.5 Fuzzy classifiers

Organizing the items into a fuzzy set refers to fuzzy classification. The function of the fuzzy set is denoted by the truth value of propositional function. In [22], a novel technique is adduced for classifying visual objects on the basis of fuzzy classifiers. These fuzzy classifiers perform classification by using use local image features to form a fuzzy rule base. A comparison has been made between the adduced technique and the bag-of-features image model. The adduced approach depicts a flexible system as there will be only a need to add new fuzzy rules in case new visual classes are added to the system. To select the most relevant features form the image, the authors of the technique have used AdaBoost algorithm.

1.11.6 Using convolutional Neural Networks

In [25], the problem of dense or pixel wise classification or labeling has been put forward which can be performed with the help of adduced technique. The technique uses CNN (Convolutional Neural Networks). A fully convolutional network architecture has been introduced. Satellite images are used as input for the framework. Imperfect training data issue has also been discussed. Due to the unavailability of accurate training data, the application of CNN may be limited. So, two step training approach has been utilized for this purpose. Firstly, using the imperfect data, CNNs are loaded and then using the accurate or processed data.

1.11.7 Hypergraph

The proposed technique based on hypergraph for the HIS(Hyperspectral Imagery) is a very powerful approach for clustering that is to reflect contextual correlation within HIS. The adduced technique is known as spatial-spectral locality elastic net hypergraph for HIS and in [28]. Exploring intimately correlated samples to build hyperedges is the core point to build a useful hypergraph. K relevant pixels are chosen that focuses on the most co-associated atoms. It also reduces the computational time as the technique is locality constrained. Pixels are represented as linear combination of atoms. Then, linking of the hyperedges is performed on pixels and their most associated pixels. And, hence high order relationship can be obtained. The experimental results are a proof of the adduced technique being an effective one.

1.11.8 Seed Point Selection

In [27] seed point selection technique has been adduced which is based on seed point, a significant feature of clustering. The selection of the earliest cluster center will decide how the algorithm will work in terms of performance parameters. Image RGB color characteristics of a color image was used to be applied upon by the proposed algorithm. Also, 2D data was considered along with Shannon's entropy and distance restriction. Both image data and discrete data were used for producing the results. The comparison was done with other seed selection techniques in terms of iterations and CPU time. As a future work, extension of the algorithm done to determine optimal number of clusters using unknown real datasets. The algorithm will become more efficient this way.

1.11.9 Multiview Learning

A multiview learning model has been adduced in [26]. This model is able to achieve semi-supervised classification, clustering and local structure learning at the same time. Division of the achieved graph into particular clusters can be done directly. Weights can be assigned to seach view without taking into consideration other parameters. This model outperformed other ones.

1.12 Taxonomy, evolution and comparison of image clustering and classification

Fig.1.5 depicts the taxonomy of image clustering and classification. Fig.1.6 refers to evolution of image clustering and classification techniques. Table 1.1 presents the comparison of techniques.

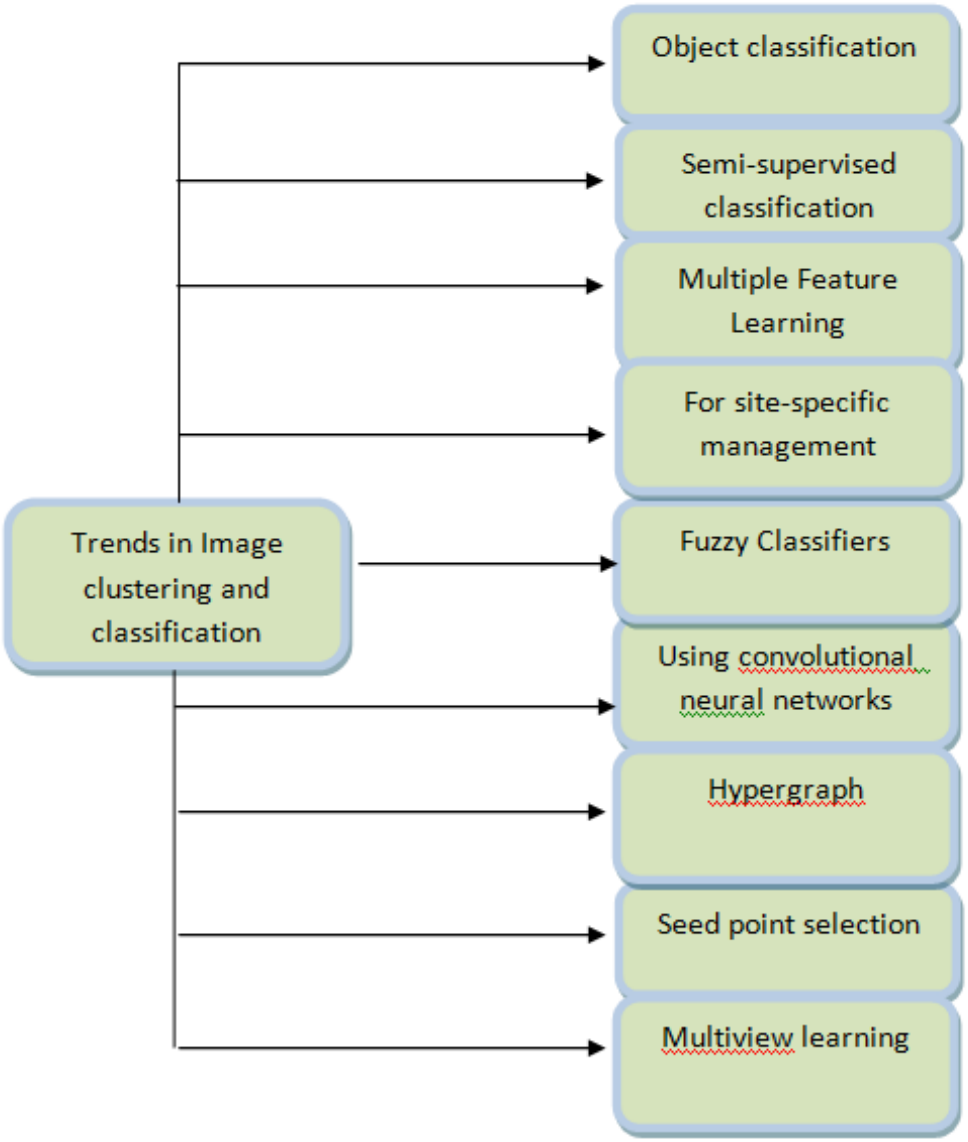

Fig.1.5. Taxonomy of clustering and classification in images

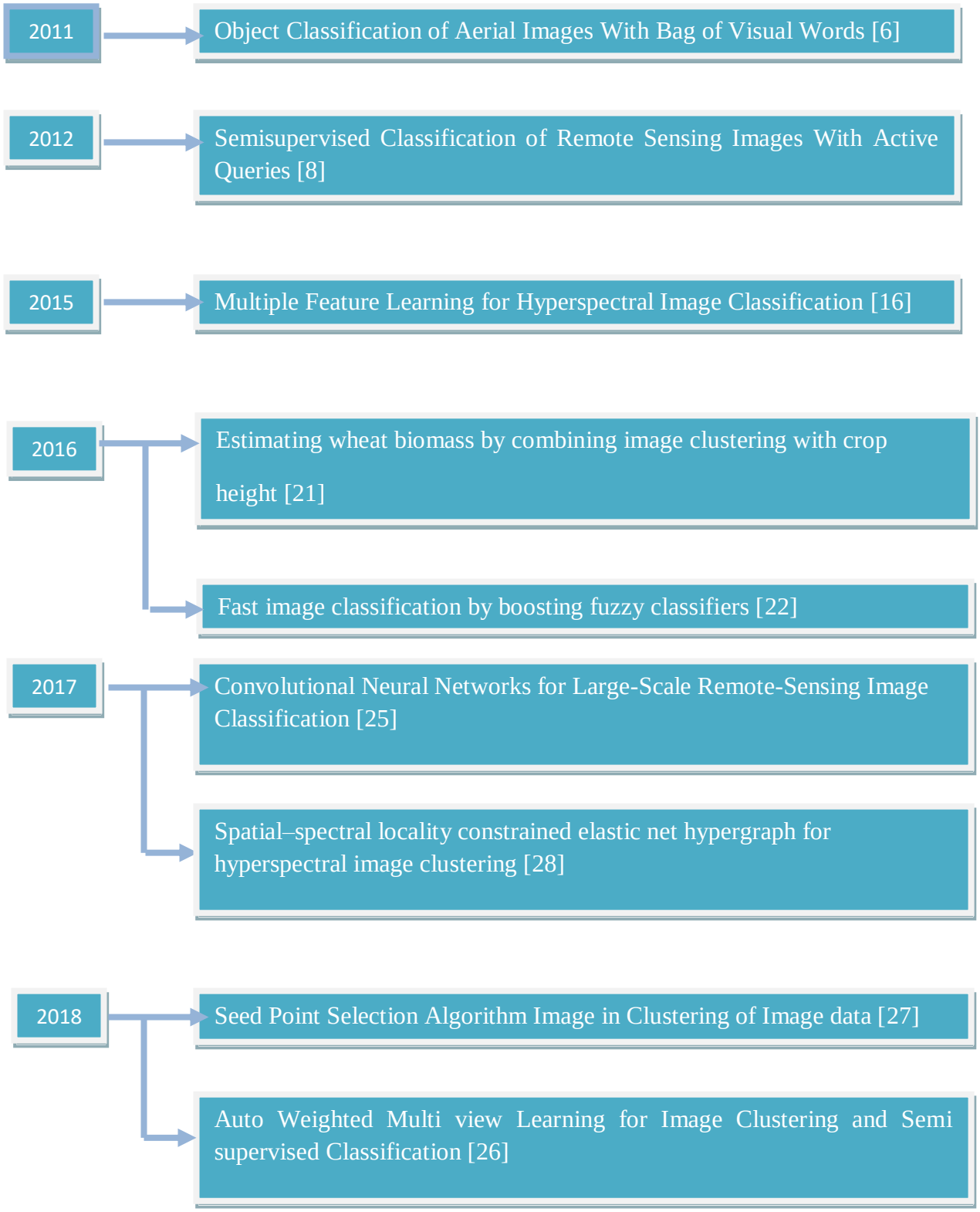

Fig.1.6. Evolution of image clustering and classification techniques

Table 1.1. Comparison table of various techniques in image clustering and classification

Authors	Technique	Description	Findings
Bong et al. [6]	Object Classification of Aerial Images With Bag of Visual Words	The appearance of words is used as the training data set for BOV. The aim is to express complex content of images with high resolution	Outperformed SIFT and good for VHR images
Munoz-Mari et al. [8]	Semisupervised Classification of Remote Sensing Images With Active Queries	Land cover map development using remote sensing images	Powerful classification and Purposeful for earth examination applications
Li et al. [16]	Multiple Feature Learning for Hyperspectral Image Classification	A framework for classification of hyperspectral images, which usually comprises multiple features	Reduces complexity
Schirrmann et al. [21]	Estimating wheat biomass by combining image clustering with crop height	Observation of geographical modifications in the wheat biomass	NIR image clustering outperformed NDVI
Korytkowski et al. [22]	Fast image classification by boosting fuzzy classifiers	Fuzzy classifiers perform classification by using use local image features to form a fuzzy rule base	Better accuracy and flexible

Maggiori et al. [25]	Convolutional Neural Networks for Large-Scale Remote-Sensing Image Classification	The problem of dense or pixel wise classification or labeling has been put forward which can be performed with the help of adduced technique	Outperformed other models in various aspects
Wang et al. [28]	Spatial–spectral locality constrained elastic net hypergraph for hyperspectral image clustering	Powerful approach for clustering that is to reflect contextual correlation within HIS	Better performance as compared to other methods
Chowdhury et al. [27]	Seed Point Selection Algorithm Image in Clustering of Image data	The selection of the earliest cluster center will decide how the algorithm will work in terms of performance parameters	The comparison was done with other seed selection techniques in terms of iterations and CPU time. As a future work, extension of the algorithm done to determine optimal number of clusters using unknown real datasets
Nie et al. [26]	Auto Weighted Multi view Learning for Image Clustering and Semi supervised Classification	This model is able to achieve semi-supervised classification, clustering and local structure learning at the same time	Better performance as compared to other techniques

1.13 Conclusion

This review depicts the various classical classification and clustering techniques and the recent trends in image clustering and classification in the past few years. A detailed description of the related work based on the recent trends in image classification and clustering has been presented in this paper using the taxonomies. A clear depiction of evolution of several techniques in image classifications and clustering has been presented. A comparison table has also been presented to justify the evolution. Some of the techniques were exceptionally outstanding as compared to the existing techniques while some were marginally better than the existing ones. Hence, taking into account the mentioned techniques in image classification and clustering, some future research directions have been recommended.

1.13.1 Impending research directions in image processing

- To describe complex content in images, gray level differences can also be included.
- For semi-supervised classification, both spatial and non-linear relationships can be considered.
- The adduced framework for multiple feature learning may include both linear and non-linear characteristics. Also, only important characteristics can be included for reducing complexity.
- Multiple object classes and shape priors can be included for image classification using convolutional neural networks.
- Several class data can be used for clustering using seed point selection algorithm. There is aneed to reduce the structural complexity to represent the image.

REFERENCES

[1]. C. Apte, and S. Weiss, “Data Mining with Decision Tress and Decision Rules”, Future Generation Computer Systems, Vol. 13, No. 2-3, pp. 197-210, 1997.
[2]. M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, “Support vector machines”, IEEE Intelligent Systems and their Applications, Vol. 13, No. 4, pp. 18-28, 1998.
[3]. A.K. Jain, “Data clustering: 50 years beyond K-means”, Pattern Recognition Letters, Vol. 31, No. 8, pp. 651-666, 2010.
[4]. I. Mr, V. Mohan, “A Survey of Grid Based Clustering Algorithms”, International Journal of Engineering Science and Technology, Vol. 2, pp. 3441-3446, 2010.
[5]. H. Kriegel, P. Kröger, J. Sander, and A. Zimek, “Density-based clustering”, Wiley Online Library, 2011.

[6]. S. Xu, T. Fang, D. Li, and S. Wang, "Object Classification of Aerial Images With Bag-of-Visual Words", IEEE Geoscience And Remote Sensing Letters, Vol. 7, No. 2, pp.366-370, 2010.

[7]. S. Horng, M. Su, Y. Chen, T. Kao, R. Chen, J. Lai, and C. D. Perkasa, "A novel intrusion detection system based on hierarchical clustering and support vector machines", Expert Systems with Applications, Vol. 38, No. 1, pp. 306-313, 2011.

[8]. J. Muñoz-Marí, D. Tuia, and G. Camps-Valls, "Semisupervised Classification of Remote Sensing Images With Active Queries", IEEE Transactions On Geoscience And Remote Sensing, Vol. 50, No. 10, pp. 3751-3763, 2012.

[9]. A.Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A. Y. Zomaya, S. Foufou, and A. Bouras, "A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis", IEEE Transactions on Emerging Topics In Computing, Vol. 2, No. 3, pp. 267-279, 2014.

[10]. A.S. Shirkhorshidi, S. Aghabozorgi, T. Y. Wah, and T. Herawan, "Big Data Clustering: A Review", International Conference on Computational Science and Its Applications, Vol. 8583, pp. 707-720, 2014.

[11]. F. Schwenker, and E. Trentin, "Pattern classification and clustering: A review of partially supervised learning approaches", Pattern Recognition Letters, Vol. 37, pp. 4-14, 2014.

[12]. M. Gu, X. Li, Y. Cao, "Optical storage arrays: A perspective for future big data storage", Light: Science & Applications, 2014.

[13]. X. Wu, X. Zhu, G. Wu, and W. Ding, "Data Mining with Big Data", IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 1, pp. 97-107, 2014.

[14]. B. Zerhari, A. A. Lahcen, and S. Mouline, "Big Data Clustering: Algorithms and Challenges", International Conference on Big Data, Cloud and Applications, 2015.

[15]. Ishwarappa, and Anuradha J, "A Brief introduction on big data 5Vs characteristics and Hadoop Technology", International Conference on Intelligent Computing, Communication and Convergence, 2015.

[16]. J. Li, X. Huang, P. Gamba, J. M. Bioucas-Dias, L. Zhang, J. A. Benediktsson, and A. Plaza, "Multiple Feature Learning for Hyperspectral Image Classification", IEEE Transactions On Geoscience And Remote Sensing, Vol. 53, No. 3, pp. 1592-1606, 2015.

[17]. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. U. Khan, "The riseof "big data" on cloud computing: Review and open research issues", Information Systems, Vol. 47, pp. 98-115, 2015.

[18]. X. Jin, B. W. Wah, X. Cheng, and Y. Wang, " Significance and Challenges of Big Data Research", Big Data Research, Vol. 2, No. 2, pp. 59-64, 2015.

[19]. Ajin V W, and L. D. Kumar, "Big Data and Clustering Algorithms", International Conference on Research Advances in Integrated Navigation Systems, 2016.
[20]. L. Chen, Z. Xu, and H. Wang, S. Liu, "An ordered clustering algorithm based on K-means and the PROMETHEE method", International Journal of Machine Learning and Cybernetics, Vol. 9, No. 6, pp. 917-926, 2016.
[21]. M. Schirrmann, A. Hamdorf, A. Garz, A. Ustyuzhanin, and K. Dammer. "Estimating wheat biomass by combining image clustering with crop height", Vol. 121, pp. 374-384, 2016.
[22]. M. Korytkowski, L. Rutkowski, and R. Scherer, "Fast image classification by boosting fuzzy classifiers", Information Sciences, Vol. 327, pp. 175-182, 2016.
[23]. P. Pandey, M. Kumar, and P. Srivastava, "Classification Techniques for Big Data: A Survey", International Conference on Computing for Sustainable Global Development, 2016.
[24]. S. Sato, A. Kayahara, and S. Imai, "Unstructured Data Treatment for Big Data Solutions", International Symposium on Semiconductor Manufacturing (ISSM), 2016.
[25]. E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Convolutional Neural Networks for Large-Scale Remote-Sensing Image Classification", IEEE Transactions on Geoscience and Remote Sensing, Vol. 55, No. 2, pp. 645-657, 2017.
[26]. F. Nie, G. Cai, J. Li, and X. Li, "Auto-Weighted Multi-view Learning for Image Clustering and Semi-supervised Classification", IEEE Transactions on Image Processing, Vol. 27, No. 3, pp. 1501-1511, 2017.
[27]. Kuntal Chowdhury, DebasisChaudhuri and Arup Kumar Pal, "Seed Point Selection Algorithm in Clustering of Image Data", Progress in Intelligent Computing Techniques: Theory, Practice, and Applications, pp.119-126, 2017.
[28]. S. Wang, Y. Sun, R. Hang, Q. Liu, X. Yuan, and G. Liu, "Spatial–spectral locality constrained elastic net hypergraph for hyperspectral image clustering", International Journal of Remote Sensing, pp. 7374-7388, 2017.
[29]. M. Gutoski, M. Ribeiro, N. Marcelo R. Aquino, L. T. Hattori, A. E. Lazzaretti, and H. S. Lopes, "Feature Selection Using Differential Evolution for Unsupervised Image Clustering", International Conference on Artificial Intelligence and Soft Computing, pp. 376-385, 2018.