Pictorial Data Analysis through Classification Techniques for Real World Sector: A survey

Swasthika Jain T J
Department of Computer Science and Engineering
GITAM School of Technology, Bengaluru campus, Karnataka, India

Dr. I Jeena Jacob
Department of Computer Science and Engineering
GITAM School of Technology, Bengaluru campus, Karnataka, India

Dr. M. Ajay Kumar
Department of Electrical, Electronics and Communication Engineering
GITAM School of Technology, Bengaluru campus, Karnataka, India

Dr. Brahmananda S H
Department of Computer Science and Engineering
GITAM School of Technology, Bengaluru campus, Karnataka, India

Abstract- The advancements in data analytics pave the path to utilize images as Big Data in real world applications, which is being used in many areas like education, government sector, healthcare centers, manufacturing units, finance filed, banks and centers of retail business. Many researches were done to process the huge volume of image data and to extract the necessary information from it. Categorization of images is very essential for utilizing it in real world environments. The classification can be done based on different features like sample data, parameter data and pixel information. This paper analyses the different categories of image classification techniques.

Keywords – minimum-distance-to-means classification, ANN, SVM, eCognition, ISODATA, Fuzzy c means, parallelepiped and maximum likelihood, K -means .

I. INTRODUCTION

The utilization of Internet is quickly expanding which leads to data detonation and accumulation of Big data. With the rise in storage capabilities and strategies of knowledge assortment, immense amounts of knowledge became simply accessible for processing [1]. The big data market is an industry that is expected to grow enormously into the future and offers the economy (business and government) great potential [2][3]. 2018). The processing and mining information from this huge available data is a tedious task. The accumulation in high magnitude of the data has led to a climb in need for carrying out cleansing, inspection, revolution as well as data modelling to pick up experiences from the data in order to evolve closure for better decision-making process. This practice is known as data analysis.

The data may be of 3 types, Structured, unstructured and semi structured. Data which contain explicit length and pattern, can easily storable and analyses with high degree of organization is defined as structured data. This implies the information is sorted out in recognizable structure to enable it reaction to inquiries to recover data for organization use [5]. 2009). Example for structured one is RDB (Relational Database) table like SQL (structured query language) or Access, which contains structured numbers, dates, gathering of numbers and word called content/string. Semi-structured is sporadic information that might be deficient and have a structure that changes quickly or eccentrically yet doesn't comply with a fixed or unequivocal pattern. This implies it isn't table arranged as in a RDB model or arranged chart as in OODB (Object Oriented Database). As per [6]. semi-structured information representation permits data from a few sources, with related to various properties, to be fit together in one entire, for instance, XML, email and doc documents. Unstructured information has no specific structure. Unstructured information ordinarily incorporates bitmap pictures/objects, email, content and other information variety that aren't part of a database [7]. In spite of the fact that messages are composed in a DB (database) design like in Microsoft Exchange and Lotus Notes, the texture of the message is in content organization without structure in any way. At the

other end, unstructured information involves reports such as PowerPoint used to depict organization system, spreadsheets of lead list, messages between collaborators, emails between co-workers and associations of clients on interpersonal organizations [8].

The measure of organized and unstructured information being created and put away has detonated as of late into exponential movement because of digitalization of information. The wellsprings of both organized and unstructured information incorporate every day exchanges, online networking, sensor produced, advanced pictures, recordings, sounds, and clickstreams which include commitments from associations and people [4]. In this way, there is a need to break down both organized and unstructured information in the everyday running of association to decide client responses, item inclination of customers, item personalization, and other organizational necessities. Unstructured information is continuously assumed control over the information arrangement of associations and Cisco (2015) discovered that information is developing at phenomenal rate as portrayed in Figure 1.

I.

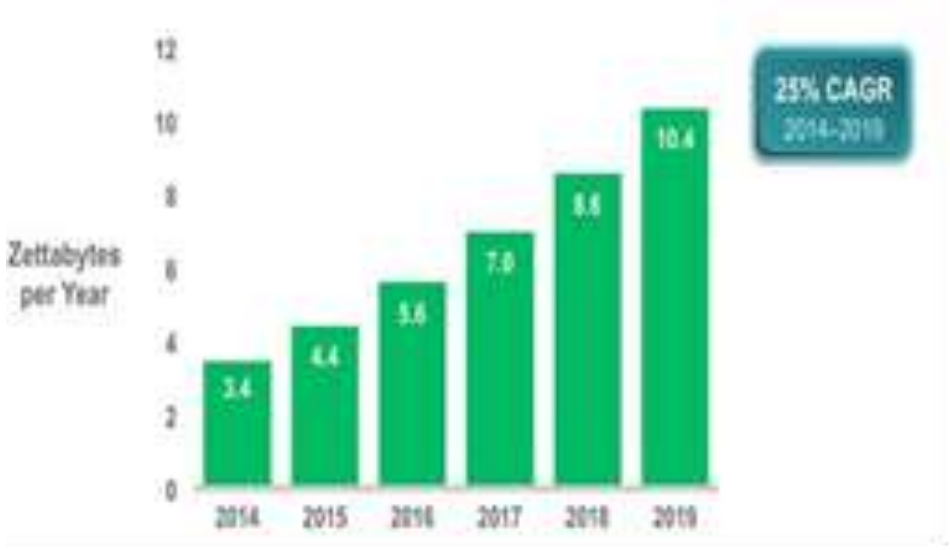

Fig 1. Yearly data development rate Source: Cisco Global Cloud Index

Image data is regularly used to speak to realistic or pictorial information. Picture data is unstructured data type. The term picture inalienably mirrors a realistic portrayal, and in the GIS world, varies fundamentally from raster information. Frequently, picture information is utilized to store remotely detected symbolism, for example satellite scenes or orthophotos, or auxiliary designs such as photos, filtered plan archives, and so on [9].

Image analysis or investigation is the extraction of important data from pictures; for the most part from computerized pictures by methods for advanced image handling or processing techniques [2]. Image analysis errands can be as plain as perusing bar code labels or as modern as distinguishing an individual from their face. It contains a lot of picture data. For the picture data analysis, image processing [10],[11] and information/data mining [4],[6] systems assume a significant role. Numerous product instruments [16],[21] such as R [24], WEKA [17],[18], RapidMiner [19], SciKit [20], KNIME [20], SparkMLlib [22] are accessible to deal with various sorts of data consequently. However, these tools are not adequate for detailed analysis of a specific data type, as the number of data mining system operations has been predetermined.[16]

Practitioners, analyst and Scientists have tried extraordinary endeavors in creating advanced classification approaches and method for remodelling classification efficiency [27],[29],[30],[31],[32],[33], [34].

The process of classifying pixels into limited arrangement of solitary classes dependent on their data value is known as image classification. Classification is one of the Data Mining method that is fundamentally utilized to indoctrinate a given dataset and takes each example of it and assigns this case to a specific class with the end goal that classification mistake will be least. If it satisfies the certain set of rules to fit in a particular class pixel is used to precise class. The classes can be known or unknown [33].

Classification is used to categorize each object into one of the predefined categories or classes in a set of data. This is used to elicitation models that accurately describes important data classes within the given dataset. The data analysis task classification is a technique or classifier is developed to anticipate categorical labels (class name traits). Classification is a data mining capacity that allocates things in an assortment to target categories or classes.[36]. It is two step procedure. In the time of initial level, the model is made by exploiting classification algorithm on training data set then in next level the extricate model is examined across a predefined test dataset to quantify the model trained presentation and efficiency. Along these lines, classification is the process to appoint class name from dataset whose class name is obscure. [38].

II. IMAGE CLASSIFICATION TYPE

The image classification approaches can be categorized into various types based on various features like training sample, parameter data, pixel information and count of each spatial element results [14].

2.1 Based on Training Sample

Based on training sample, there are two classifications, supervised and unsupervised. First classification is Supervised method is utilizing the samples of known informational classes (training datasets) to categized the pixels of unknown identity. Recognize known from the earlier by means of a blend of fieldwork, map investigation, and individual practice as preparing sites; the spectral characteristics of these sites are used to develop the classification algorithm for prospective land cover mapping of the carryover of the image. whole pixel with both inward and external, the training sites is then examined and empower to the class of which it has the most likelihood of being representative. Examples for supervised algorithms are minimum distance to means algorithm [44], parallelepiped algorithm [13] and maximum likelihood algorithm [45].

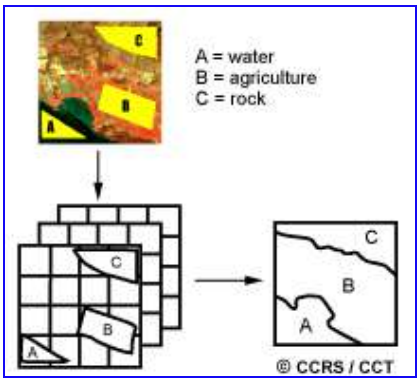

Figure 2. Supervised classification

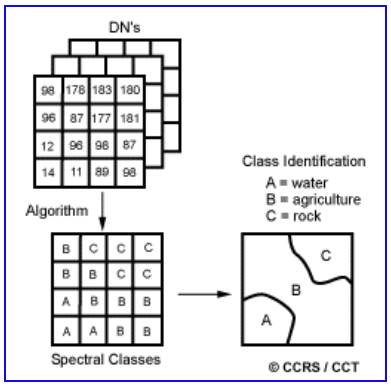

Figure 3. Unsupervised classification

Unsupervised Classification variety of classification is a process which audit a greater number of unknown pixels and breaks it into estimated classes based on essential groupings appears in the picture data values. Computer inclines spectrally separable class and then describes its data value. No comprehensive prior knowledge is prescribed. The algorithm naturally bunches of pixels with comparable spectral qualities (means, standard deviations, covariance, connection matrices, and so on.) into novel groups as indicated by a few statistically decided benchmark. The analyst then re-labels and associate the spectral clusters into information classes. Example for Unsupervised algorithm is K means clustering algorithm.[33],[43].

2.2. Assumption of Parameter on Data:

In view of parameter highlights are arranged into two techniques. In that first classifier is Parametric, uses parameters like covariance metrics and mean vector are used. There is a supposition of Gaussian dispersion. The parameters like mean vector and covariance grid are much of the time created from preparing tests. Example Models are Maximum likelihood and linear discriminant analysis. Second classifier is Non-Parametric has no hypothesis around the information. Non-parametric classifiers don't utilize measurable parameters to figure class partition. Example Models are SVM, ANS(Artificial Neural System), Decision tree classifier and Expert system analysis [35],[45].

2.3 Based on Pixel Information:

Based on pixel information categories are classified into four methods, perpixel, subpixel, prefield and object-oriented. Per- pixel classifier is ordinary technique in classification, which creates a mark by using the mix of the spectra of whole training set pixels from provided component. The commitments of whole materials commenced in the preparation set pixels are applicable in the subsequent mark. It tends to be parametric or nonparametric, but the exactness may not get together as a result of the effect of the mixed pixel issue. Examples are maximum likelihood, SVM, ANN and minimum distance. Sub-pixel classifiers is where value of spectral of every pixel is thought to be a straight or non-direct mix of characterized unadulterated materials called end individuals, giving relative membership of every pixel to each end part. Sub-pixel classifier has the capacity to maintain the mixed pixel problem, relevant for medium and coarse spatial resolution images. Examples are spectral mixture analysis, subpixel classifier and Fuzzy-set classifiers. The per-field classifier is proposed to deal with the issue of ecological heterogeneity, and furthermore improves the classification precision. For the most part utilized by GIS-based classification approaches. Object-oriented classifiers utilize pixels of the picture and joined them into articles and afterward classification is performed based on objects. It includes 2 phases like picture division and picture characterization. Picture division joins pixels into objects, and a classification is then executed based on objects. Model is e-Cognition [14].

2.4 In view of number of outputs for each spatial element:

In view of number of outputs for each spatial component, it has two techniques as hard characterization or classification and delicate or soft classification. Hard characterization is also called fresh order. In this, every pixel is required or compelled to demonstrate enrolment to a solitary class.eg maximum likelihood, minimum distance, artificial neural network, decision tree, and support vector machine. Soft classification is also called fuzzy grouping. In this, every pixel may show various and halfway class participation. It creates progressively precise outcome.

2.5 Spatial information

Based on spatial information, it has three methods as spectral classification, contextual classification and spectral-contextual classification. Spectral Classifier is picture characterization that utilizations unadulterated spectral data. Models are, minimum distance, Artificial neural network ANN and Maximum likelihood. Contextual Classifiers is picture characterization utilizes the spatially neighboring pixel data. Model is frequency-based contextual classifier. Spectral-contextual classifiers is characterization utilizes both phantom and spatial data starting grouping pictures are created utilizing parametric or non-parametric classifiers and afterward contextual classifiers are actualized in the ordered/classified pictures. Models are mix of parametric or non-parametric and contextual algorithm [13].

Table-1 Image classification methods

Algorithm	Categories based on features	Characteristics
minimum-distance-to-means classification [44]	Supervised classification, Parametric classifier, Per pixel classifier, Hard Classification, Spectral Classifiers, Spectral-contextual classifiers,	• Computational complexity is less • Efficient • Fast • No unclassified pixels • Does not incorporate variability of signatures • Digital parameter space is calculated for pixels of known classes.

Gaussian Maximum Likelihood Classifier [45]	Supervised classification, Parametric classifier, Per pixel classifier, Hard Classification, Spectral Classifiers, Spectral-contextual classifiers.	• Most accurate • Considers variability • Slow • Relies heavily on normally distributed signatures • Population Distribution does not follow the normal distribution. can't be applied the maximum likelihood method. • Quantitively evaluates both variance and covariance of the category used as classification in remote sensing. • Training sets uses the normal distribution. • Those training data sets is Gaussian in nature. • Which depict by the mean vector and covariance matrix
Parallelepiped [7]	supervised classification, Parametric classifier	• ***Level-Slice Classifier*** • Simplest of all classification method • To classify hyperspectral data basic decision rule can be used. • Computationally very efficient. • Two bands of image value are plotted in a scatter diagram.
K-means algorithm	Unsupervised classification, nonparametric classifier	• Iterative method. • Amount of results specified for spectral classes (let's consider n clusters) by the analyst. • In multidimensional data centroids of K clusters are located. • Nearest cluster assigned by pixel in the image. • Revised mean of every cluster reclassified by Image.
ISODATA algorithm (Iterative Self Organizing Data Analysis Technique) [35]	Unsupervised classification, nonparametric classifier	• Uses "spectral distance" between picture pixels in include space to characterize pixels into a predefined number of extraordinary spectral groups. • ISODATA algorithm allows for different number of clusters like k-means. • Clusters are combined if either the number of pixels in a cluster is less than a certain threshold or if the centers of two clusters are closer than a certain threshold. • Those groups or clusters split into two different clusters. • If the groups of standard deviation exceed a predefined value and the number of members (pixels) is double the limit for the minimum number of individuals.
Fuzzy c-means [43]	Unsupervised classification, Subpixel classifiers, Soft classification.	• In radiance of more than one feature pixel recorded are known as mixed pixel. • It is a member in more than one class. • To solve the unmixed problem, we need a fuzzy logic-based model. • The fuzziness of image pixels can be represented.
support vector machine [44]	Supervised classification, Parametric classifier, Per pixel classifier, Hard Classification, Spectral Classifiers, Spectral-contextual classifiers.	• This uses the pair of classification or regression challenges. • As a n- dimensional space data can be plotted (n is number of feature) with the value of each feature being the value of a particular coordinate. • SVMs can efficiently achieve a non-linear classification, essentially mapping their resource value into high-dimensional feature spaces.
ANN [45]	Supervised classification	• This can be applied where, what has happened in past is repeated almost exactly in same way. E.g. Black Jack against a computer game. • This is used rarely for predictive modelling. reason behind this is tries to over-fit the relationship. • many different coefficients, which it can optimize. Hence, it can handle much more variability as compared to traditional models.
eCognition [14]	Object-oriented classifiers	• eCognition to produce a supervised classified map based on object image analysis.

III. CONCLUSION

Categorization of images is very essential for utilizing it in real world environments. The classification can be done based on different features like sample data, parameter data, pixel information. This study analyses the different categories of image classification techniques. This paper endeavors to examine and gives a short information about the assorted image portrayal moves close and particular classification method. Most ordinary approaches for image characterization can be arranged as unsupervised and supervised, or parametric and nonparametric or spectral classifiers, contextual classifiers and spectral-contextual classifiers, subpixel, per-pixel, object-oriented and per-field, or hard and soft classification.

REFERENCES

[1] Nada Elgendy and Ahmed Elragal. Big Data Analytics: A Literature Review Paper. ICDM 2014 .Advances in Data Mining. Applications and Theoretical Aspects pp 214-227

[2] Solomon, C.J., Breckon, T.P. (2010). *Fundamentals of Digital Image Processing: A Practical Approach with Examples in Matlab*. Wiley-Blackwell. doi:10.1002/9780470689776. ISBN 978-0470844731.

[3] Analytics Comes of Age, McKinsey Analytics, January 2018 (PDF, 100 pp.)

[4] Chen, Y., Wang, W., Liu, Z., & Lin, X. (2009, June). Keyword search on structured and semi-structured data. In Proceedings of the 2009 ACM SIGMOD International Conference on Management of data (pp. 1005-1010). ACM

[5] Doan, A., Naughton, J., Baid, A., Chai, X., Chen, F., Chen, T., ... & Huang, J. (2009). The case for a structured approach to managing unstructured data.arXiv preprint arXiv:0909.1783

[6] Hänig, C., Schierle, M., & Trabold, D. (2010). Comparison of structured vs. unstructured data for industrial quality analysis. In Proceedings of The World Congress on Engineering and Computer Science

[7] Feldman, R., & Sanger, J. (2007). The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge University Press

[8] Maluf, D. A., & Tran, P. B. (2008, March). Managing Unstructured Data with Structured Legacy Systems. In Aerospace Conference, 2008 IEEE (pp. 1-5). IEEE.

[9] David J et al.1997.GIS introduction textbook.

[10] Amit Kumar and Fahimuddin Shaik. 2016. Importance of Image Processing. Image Processing in Diabetic Related Causes, Springer, Singapore, pp 5-7.

[11] Anil Jain. 1989. Fundamentals of digital image processing. (ACM) Prentice-Hall, Inc. Upper Saddle River, NJ, USA ©1989

[12] Press G, $16.1 billion big data market: 2014 predictions from IDC and IIA, Forbes, Tech. Rep. 2013

[13] D. Lu And Q. Weng, "A SURVEY OF IMAGE CLASSIFICATION METHODS AND TECHNIQUES FOR IMPROVING CLASSIFICATION PERFORMANCE" International Journal of Remote Sensing Vol. 28, No. 5, 10 March 2007

[14] Pooja Kamavisdar1, Sonam Saluja2, Sonu Agrawal. A Survey on Image Classification Approaches and Techniques. International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 1, January 2013.PP-1005-1009

[15] Ming-Syan Chen, Jiawei Han and Philip Yu. 1996. Data mining: an overview from a database perspective. In IEEE Transactions on Knowledge and Data Engineering, 8.6 (1996): 866-883.

[16] Ralf Mikut and Markus Reischl. 2011. Data mining tools. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1.5 (2011): 431-443.

[17] Geoffrey Holmes, Andrew Donkin and Ian H. Witten. 1994. Weka: A machine learning workbench. In proceedings of the 2ndIEEE Australian and New Zealand Conference on Intelligent Information Systems, pp. 357-361.

[18] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard P fahringer, Peter Reutemann, and Ian H. Witten, 2009. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter 11, no. 1 (2009): 10-1

[19] Radim Burget, Jan Karasek, Zdenek Smékal, Václav Uher, and Otto Dostal. 2010. Rapidminer image processing extension: A platform for collaborative research. In proceedings of the 33rd International Conference on Telecommunication and Signal Processing, pp. 114-118

[20] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel et al. 2011. Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research, 12 (2011): 2825-2830

[21] Xiangrui Meng, Joseph Bradley, Burak Yavuz, Evan Sparks, Shivaram Venkataraman, Davies Liu, Jeremy Freeman et al. 2015. Mllib: Machine learning in apache spark. The Journal of Machine Learning Research, 17.1 (2016): 1235-1241.

[22] Alexander Fölling, Joachim Lepping. Knowledge discovery for scheduling in computational grids. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Wiley, 2012, 2 (4), pp.287- 297. ff10.1002/widm.1060ff. ffhal-00758208f

[23] Luis Torgo. 2010. Data mining with R: learning with case studies. Chapman & Hall/CRC

[24] Michael Berthold, Nicolas Cebron, Fabian Dill, Thomas Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Christoph Sieb, Kilian Thiel,and Bernd Wiswedel. 2008. KNIME: The Konstanz information miner. Springer Berlin Heidelberg, pp. 319-326

[25] Anil Jain. 2010. Data clustering: 50 years beyond K-means. Pattern recognition letters, 31.8 (2010): 651-666.
[26] GONG, P. and HOWARTH, P.J., 1992, Frequency-based contextual classification and gray-level vector reduction for land-use identification. Photogrammetric Engineering and Remote Sensing, 58, pp. 423–437.
[27] FOODY, G.M., 1996, Approaches for the production and evaluation of fuzzy land cover classification from remotely-sensed data. International Journal of Remote Sensing, 17, pp. 1317–1340.
[28] APLIN, P., ATKINSON, P.M. and CURRAN, P.J., 1999a, Per-field classification of land use using the forthcoming very fine spatial resolution satellite sensors: problems and potential solutions. In P.M. Atkinson and N.J. Tate (Eds), Advances in Remote Sensing and GIS Analysis, pp. 219–239 (New York: John Wiley and Sons).
[29] KONTOES, C., WILKINSON, G.G., BURRILL, A., GOFFREDO, S. and MEGIER, J., 1993, An experimental system for the integration of GIS data in knowledge-based image analysis for remote sensing of agriculture. International Journal of Geographical Information Systems, 7, pp. 247–262.
[30] STUCKENS, J., COPPIN, P.R. and BAUER, M.E., 2000, Integrating contextual information with per-pixel classification for improved land cover classification. Remote Sensing of Environment, 71, pp. 282–296.
[31] SAN MIGUEL-AYANZ, J. and BIGING, G.S., 1997, Comparison of single-stage and multi-stage classification approaches for cover type mapping with TM and SPOT data. Remote Sensing of Environment, 59, pp. 92–104.
[32] PAL, M. and MATHER, P.M., 2003, An assessment of the effectiveness of decision tree methods for land cover classification. Remote Sensing of Environment, 86, pp. 554–565
[33] FRANKLIN, S.E., PEDDLE, D.R., DECHKA, J.A. and STENHOUSE, G.B., 2002, Evidential reasoning with Landsat TM, DEM and GIS data for land cover classification in support of grizzly bear habitat mapping. International Journal of Remote Sensing, 23, pp. 4633–4652
[34] F.J., 2004, Remote sensing and land cover area estimation. International Journal of Remote Sensing, 25, pp. 3019–3047
[35] Nupur Thakur and Deepa Maheshwari.2017. A Review Of Image Classification Techniques, IJERT. p-ISSN: 2395-0072
[36] G.Kesavaraj, Dr.S.Sukumaran.2013. A Study On Classification Techniques in Data Mining. 4th ICCCNT.
[37] Laurent Hyafil, RL Rivest, "Constructing Optimal Binary Decision Trees is NP-complete" , Information Processing Letters, Vol. 5, No. 1. (1976), pp. 15-1 R. Kohavi and J. R. Quinlan. Decision-tree discovery.
[38] Sagar S. Nika. 2015. A Comparative Study of Classification Techniques in Data Mining Algorithms. ISSN: 0974-6471 April 2015, Vol. 8, No. (1): Pgs. 13-19
[39] Badr HSSINA, Abdelkarim MERBOUHA, Hanane EZZIKOURI, Mohammed ERRITALI. A comparative study of decision tree ID3 and C4.5. 2013, (IJACSA).P-13-18
[40] http://rulequest.com/see5-comparison.html (article de R. Quinlan)
[41] Ankur Shrivastava and Vijay Choudhary ,Comparison between ID3 and C4.5 in Contrast to IDS Surbhi Hardikar, VSRD-IJCSIT, Vol. 2 (7), 2012, 659-667
[42] H. I shibuchi, K. Nozaki, N. Yamamoto, H. Tanaka (1995), "Selecting fuzzy if then rules for classification problems using genetic algorithm", IEEE Trans. Fuzzy System 3 (3) 260 – 270
[43] G. Fung, O. L. Mangasarian (Oct. 1999), "Semi supervised support vector machines for unlabeled data classification", Technical Report, Dept. of Computer science, University of Wisconsim.
[44] Rajni Bala, Dr. Dharmender Kumar(2017)," Classification Using ANN: A Review", International Journal of Computational Intelligence Research. pp. 1811-1820
[45] Lillesand,T.M and Kiefer.R.W.,Remote Sensing and Image Interpretation, Fourth Edition,John Wiley and Sons,2002,ISBN 9971-51-427-3