Bioinformatics and Biology Insights

ORIGINAL RESEARCH

Automatic Identification of Algal Community from Microscopic Images

Natchimuthu Santhi[1], Chinnaraj Pradeepa[1], Parthasarathy Subashini[2] and Senthil Kalaiselvi[1]

[1]Department of Biochemistry, Biotechnology and Bioinformatics, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, Tamil Nadu, India. [2]Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, Tamil Nadu, India.
Corresponding author email: santhigowri@yahoo.com

Abstract: A good understanding of the population dynamics of algal communities is crucial in several ecological and pollution studies of freshwater and oceanic systems. This paper reviews the subsequent introduction to the automatic identification of the algal communities using image processing techniques from microscope images. The diverse techniques of image preprocessing, segmentation, feature extraction and recognition are considered one by one and their parameters are summarized. Automatic identification and classification of algal community are very difficult due to various factors such as change in size and shape with climatic changes, various growth periods, and the presence of other microbes. Therefore, the significance, uniqueness, and various approaches are discussed and the analyses image processing methods are evaluated. Algal identification and associated problems in water organisms have been projected as challenges in image processing application. Various image processing approaches based on textures, shapes, and an object boundary, as well as some segmentation methods like, edge detection and color segmentations, are highlighted. Finally, artificial neural networks and some machine learning algorithms were used to classify and identifying the algae. Further, some of the benefits and drawbacks of schemes are examined.

Keywords: Algae identification, segmentation, neural network, feature extraction, identification

Bioinformatics and Biology Insights 2013:7 327–334

doi: 10.4137/BBI.S12844

This article is available from http://www.la-press.com.

Introduction

Algae are a very huge and diverse collection of simple, normally autotrophic organisms, ranging from unicellular to multicellular forms. They affect water properties such as water color, odor, taste, and the chemical composition, which may cause potential hazards for human and animal health.[1] They are highly sensitive to the changes in their environment.[2] Shift in algal species and population can be used to identify the environmental changes and the status of nutrient content.[3] Algae are very good biological indications for water pollution assessment; therefore, they have long been used to assess the quality of waters in lakes, ponds, reservoirs, rivers, and so on. However, identification of algae at their taxonomy level and the application in environmental assessment is a difficult process. Several studies reported the conventional identification of algae by using microscopy images, which is a time consuming process. This has led many researchers to develop several systems to automate the analyzing and classifying algal images.[2,3] An automated computer-based recognition and classification system for the rapid identification of algae will definitely reduce the burden of routine identifications by taxonomists.[4–6] This identification and classification would allow many people to identify and know about the algae without any knowledge of algae.

Image processing is an effective technology to analyze the digital images for various applications in society. In that category, it is used in several places, such as in medical images, spatial images, underwater images, and other biological images. Several studies were carried out on the biodiversity of algae in India.[7–13] Very little research was identified on automatic algal identification using image processing techniques.

Most research applied image processing to detect, count, identify, and classify algal groups; some of this approach was efficient with 92% accuracy.[14] Some developed tools are used effectively for online monitoring, some for measurements of density of microorganism in water, and other tools were developed to assist in recognition process, such as enhancing images, noise elimination, and edge-extracted segmentation.[15–17] A combination of image processing techniques and Artificial Neural Network (ANN) algorithms are used to automate the process of detection and recognition.[18] Other techniques used included was image processing with genetic algorithms or ANN for recognition purpose.[15,19–22] MatLab based image processing tools were used for the complete enhancement and analytical operations. An automated object recognition segments the algal images and locates possible objects accurately by their boundary and texture without human interaction.[23] Automatic identification and classifications of diatoms with a circular shape were achieved by using contour and texture analysis.[24]

Image Processing Methodology

Identification of the algal community from images consists of various steps namely preprocessing, segmentation, morphological operations, feature extraction, classification, and identification. Figure 1 gives the architectural layout of the image processing method used in the identification and the classification of algae. In the following section, we will discuss the functionality of each processing technique.

Image Preprocessing

Correct object detection depends upon many factors, such as the type of illumination, the presence of shadows, the level of noise, the state of focus, the

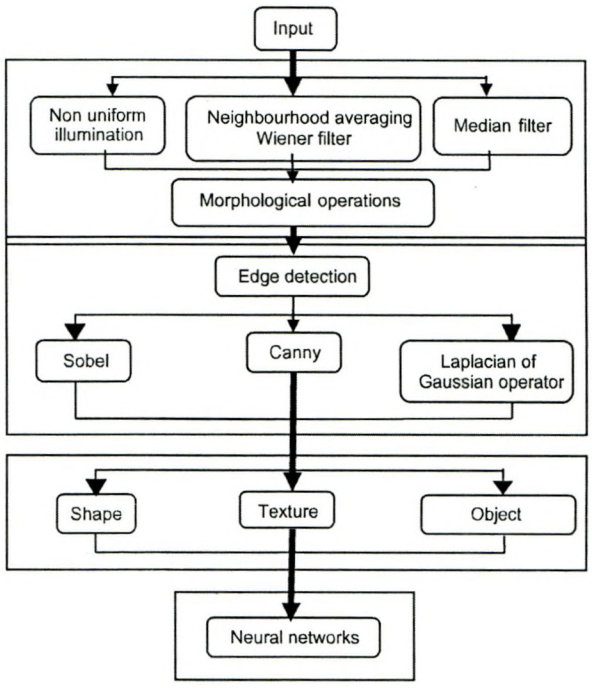

Figure 1. Proposed methodology of automatic algal identification.

overlapping of objects, as well as level of object similarity to the background.[25,26] The digital grayscale images captured from a microscope are preprocessed to reduce the effects of nonuniform illumination and other noise. A median filter (size 3×3 and 5×5) was used to reduce image noise.[15,27] In the present study, the neighborhood averaging technique was used to enhance the image and morphological features were processed for noise elimination, and to keep the cyanobacteria structure clear (Fig.2).

Nonuniform illumination was corrected using the top-hat filter. Neighborhood averaging technique using Wiener filter and median filter methods were used to reduce image noise and to preserve edges. The performance of the three methods were analyzed statistically and the results were shown in Table 1. Based on mean squared error and peak signal noise ratio values, the median method showed a better result than the other two methods.

Image Segmentation

Objects within each image are separated from the background via a process called segmentation. Segmentation is the key part in the image processing.[25,26] Algal images showed various shapes for the same species. The edges and contour of the objects are more meaningful. So far, much research on the automatic identification of algae has been done using edge detection; this is achieved by the Sobel edge detector.[28] Another algorithm called the Canny edge detector algorithm is a powerful edge detector for image segmentation.[15,24,29]

In this study, both the Canny and Sobel edge detection methods were adopted for image segmentation.[25] After the Sobel edge detector method is applied, the resulting images had many discontinuities. Laplacian of Gaussian operator was applied on the Sobel image to smooth the image.[28] The edges of the algae with minimum discontinuities were detected in the

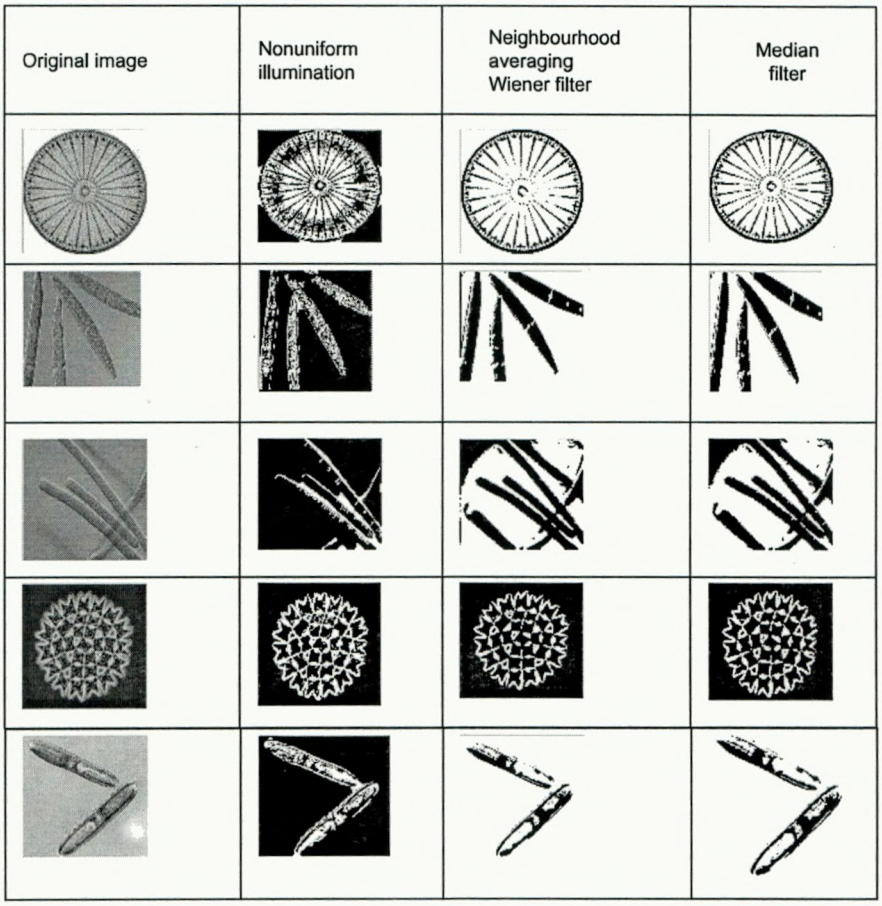

Figure 2. Pre processed images by various filters.
Note: The original images were collected from Algal Resource Database, Microbial cluture collection, National Institute for Environmental Studies. http://www.shigen.nig.ac.jp/algae/.

Table 1. Comparison of noise removal filters using MSE and PSNR metrics.

Image	Median filter		Wiener filter		Non uniform illumination using top-hat filter	
	MSE	PSNR	MSE	PSNR	MSE	PSNR
Diatom	0.0122	30.6193	0.0115	31.0761	*0.3481*	*23.3133*
Closterium acerosum	0.0152	30.8247	0.0120	35.4253	0.3542	23.1095
Oscillatoria	0.0076	33.4772	0.0078	43.4040	0.3090	23.9395
Pediastrum	0.0135	30.9478	0.0184	32.3668	0.4764	22.3336
Pinnularia	0.0058	35.6971	0.0069	36.3533	0.4965	24.4697

Canny edge detector method. To avoid the discontinuities, the same method was repeated for several times on the detected edges. A mean square error of the Canny edge detection method is slightly greater than the Sobel edge detection method. The peak signal noise ratio of the Canny method is slightly lesser than the Sobel method. Finally, the object result from the Sobel method was better than the Canny edge detection method; this is shown in Figure 3 and Table 2.

Feature Extraction

Feature extraction used to transform a binary and color image from the preprocessed stage into a set of param-

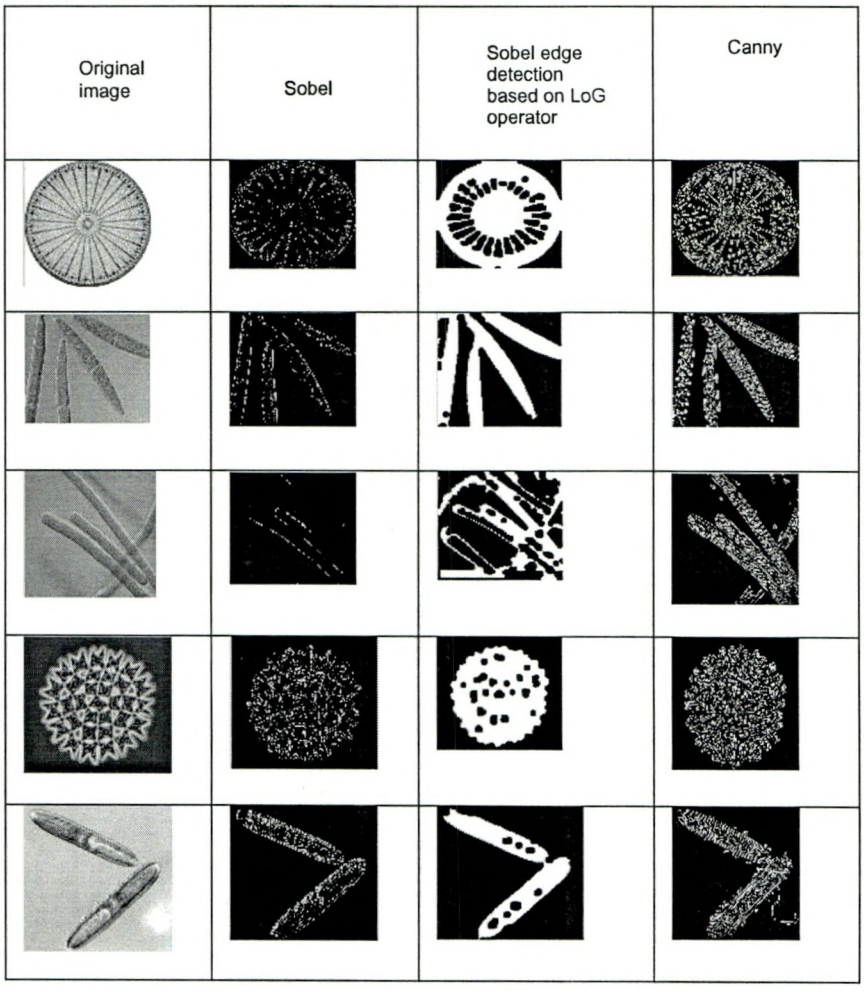

Figure 3. Edge detection methods.

Table 2. Comparison of the noise edge detection methods using MSE and PSNR metrics.

Image	Sobel		Canny	
	MSE	PSNR	MSE	PSNR
Diatom	0.4546	25.7925	0.4236	27.8187
Closterium acerosum	0.3674	24.4938	0.3630	27.0445
Gloeotrichia	0.3016	26.2720	0.3097	27.6404
Pediastrum	0.5193	24.7969	0.4967	27.1131
Pinnularia	0.5087	25.9998	0.4941	27.4304

eters that described the algae features.[15] Once an interesting feature has been detected, the illustration of this feature will be used to compare with all possible features known to the processor.

There are two main methods for object identification that use boundary information.[26] The first is the Fourier descriptor method, and the second is the moment invariant method. In the Fourier descriptor method, the boundary is divided into $N = 2^n$ parts to produce N equidistant boundary points. The coordinates of these points were now processed using fast Fourier transform. This will produce frequency classification of the boundary. The second method is finding moment invariants. In this technique, seven moment invariants can be derived, all of which are invariant to objects and changes made in magnification.[23]

Two-dimensional moment invariants of a digitally sampled $M \times M$ image.

$f(x, y)$, $(x, y = 0 \ldots M - 1)$ is given as,

$$m_{pq} = \sum_{x=0}^{x=M-1} \sum_{y=0}^{y=M-1} (x)^p \cdot (y)_p f(x,y) \qquad (1)$$

where $p, q = 0, 1, 2, 3$

The moments $f(x, y)$ translated by an amount (a, b), are defined as,

$$\mu_{pq} = \sum_x \sum_y (x+a)^p \cdot (y+b)^q f(x,y). \qquad (2)$$

Thus, the central moments m'_{pq} or μ_{pq} can be computed from (2) on substituting $a = -\bar{x}$ and $b = -\bar{y}$ as,

$$\bar{x} = \frac{m_{10}}{m_{10}} \text{ and } \bar{y} = \frac{m_{01}}{m_{00}},$$
$$\mu_{pq} = \sum_x \sum_y (x-\bar{x})^p . (y-\bar{y})^q f(x,y). \qquad (3)$$

When scaling normalization is applied the central moments change as,

$$\eta_{pq} = \mu_{pq}/\mu^{\gamma}{}_{00}, \gamma = [(p+q)/2] + 1. \qquad (4)$$

In particular, Hu defines seven values, which are computed by normalizing central moments through order three, which are invariant to object scale, position, and orientation.[30] In terms of the central moments, the seven moments are given as,

$$M1 = (\eta_{20} + \eta_{02}), \qquad (1)$$

$$M2 = (\eta_{20} - \eta_{02})^2 + 4\eta^2\ 11, \qquad (2)$$

$$M3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2, \qquad (3)$$

$$M4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2, \qquad (4)$$

$$M5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + 12)^2 - (\eta_{21} + \eta_{03})^2], \qquad (5)$$

$$M6 = (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] + 4\eta_{11}(\eta_{30} + {}_{\eta 12})(\eta_{21} + \eta_{03}), \qquad (6)$$

$$M7 = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] - (\eta_{30} + 3\eta_{12})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]. \qquad (7)$$

The moment invariant features are given in Table 3.

Walker et al[26] used new features to classify an object into one of the number of classes, (ie, *Microcystis*, *Anabaena*, and so on) it is essential to quantitatively measure characteristics of the object that may indicate its class membership. For example, the feature "area" is an excellent discriminator of class membership when classifying algae such as *Microcystis* and *Anabaena* cyanobacteria, as these two genera differ substantially in size. The features of each object, including morphometric properties (the area, circularity, and perimeter length), object boundary, shape features, frequency domain features, and spatial statistics containing Gray level co-occurrence matrix measures are used for identification.

The principal component analysis (PCA) method is widely used in most image processing applications

Table 3. Moment invariants for the algae.

Image	Moment invariant
Anabaena	0.0211, 0.0004, 0.0000, 0.0000, 0.0000, 0.0000,0
Closte	0.0189, 0.0004, 0.0000, 0.0000, 0.0000, 0.0000,0
Diatom	0.0191, 0.0004, 0.0000, 0.0000, 0.0000, 0.0000,0
Eremo	0.0183, 0.0003, 0.0000, 0.0000, 0.0000, 0.0000,0
Fibro	0.0184, 0.0003, 0.0000, 0.0000, 0.0000, 0.0000,0
Gloeo	0.0183, 0.0003, 0.0000, 0.0000, 0.0000, 0.0000,0
Microcystis	0.0225, 0.0005, 0.0000, 0.0000, 0.0000, 0.0000,0
Oscillatoria	0.0235, 0.0006, 0.0000, 0.0000, 0.0000, 0.0000,0
Penium	0.0189, 0.0004, 0.0000, 0.0000, 0.0000, 0.0000,0

to reduce the number of features by a normalization process.[1] PCA involves a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. The Fourier spectrum is ideally suitable for describing the directionality of periodic or almost periodic two-dimensional patterns in a round image.[24,30]

Identification

The classification method uses a set of features or parameters to differentiate each object, where these features should be related to the task at hand. A human expert has to determine into what classes an object may be categorized and also has provided a set of sample objects with known classes. This set of identified objects is called the training set. This is used to train the classification programs to learn how to classify objects.

Automated recognition of blue-green algae implemented a discriminant analysis for classification. It is a statistical method that provides a discriminator function for each different species. Discriminant analysis may be used for two

Table 4. Observation and analysis on existing system.

Author	Year	Objectives	Methods			Results
			Segmentation	Feature extraction	Classification	
Stefan et al	1995	Automated recognition of blue green algae	Sobel edge detection	Fourier descriptors and moment invariants	Discriminant analysis	98%
Gao et al	2011	Automatic identification of diatoms with circular shape using texture analysis	Canny edge detection	Fourier spectrum	Neural Networks	94.44%
Mansoor et al	2011	Automatic recognition system for some cyanobacteria using image processing techniques and ANN approach	Thresholding technique	Principal component analysis	Multilayer perceptron feed forward artificial neural networks	95%
Walker et al	2011	Fluroscence-assissted image analysis of freshwater microalgae	Binary segmentation	Co occurrence matrix measures	Bayes decision function	–
Fang et al	2011	Automatic identification of mycobacterium tuberculosis in acid-fast stain sputum smears with image processing neural networks	–	–	Perceptron and FFNN	100%
Anggraini et al	2011	Automated status identification of microscopic images obtained from malaria thin blood smears using bayes decesion	Edge detection, thresholding, segmentation and watershed algorithm	–	Bayes classifier	99.65%

objectives: to assess the adequacy of classification, given the group memberships of the objects under study; or to assign objects to one of a number of (known) groups of objects.

Gao et al[24] proposed a neural networks classification. Here, neural networks are designed with 15, 30, 40, 60, or 80 nodes in a single hidden layer and six nodes for each class in the output layer to test the performance.

Mansoor et al[1] presented multilayer perceptron feed forward ANN to perform an identification process for selected cyanobacteria. ANN architecture consists of six outputs, three outputs, and three neurons in a hidden layer—0.78 for learning rate, and 0.5 for momentum. The classifier is used to index the database content during the training mode for categorizing purposes.

Walker et al[26] implemented a general Bayes decision function for assumed Gaussian feature distributions with unequal variance–covariance matrices. The resulting decision surface is of hyperquadtric form. In this, the target is only the anabaena and microcystis genera. So, the microalgea in water samples were classified to the genus level.

Fang et al[19] used perceptron and the feed forward back propagation scheme of the neural network. The perceptron has six neurons and its accuracy is 100% sensitivity and 39.8% specificity. The result is 97.8% sensitivity and 72.4% specificity for this application.

Anggraini et al[27] implemented Bayes classifier in each node. The performance of this classification model was evaluated using 20 microphotographs obtained from different blood smears, which are identified as infested erythrocytes with sensitivity of 92.59%, specificity of 99.65%.

In this study, a back propagation neural network was used to classify the images that achieved 100% of classification accuracy on the trained images and 80% classification accuracy on tested images. The results are shown in Table 4.

Conclusion

This paper reviewed various techniques of preprocessing, segmentation, feature extraction, and classification in image processing. The achieved detection rate of combining all the features was more than 98%. Particularly, using the neural network, 86.5% of the identification rate was achieved. In total, 95% accuracy was achieved in the identification and classification of four genera of cyanobacteria using back propagation and shape boundary features. Then, 97% of the classification accuracy was achieved by object size, shape, and texture based on feature extraction techniques. For automatic algal identification, the identification accuracy was increased by several features such as shape, size, object boundary, and textures combined with morphological operators. The automatic identification rate is increased by using different segmentation methods and developing new features for microscopic algae images.

Author Contributions

Conceived and designed the experiments: NS, CP, PS, SK. Analyzed the data: Wrote the first draft of the manuscript: NS, CP, PS. Contributed to the writing of the manuscript: NS, CP, PS. Agree with manuscript results and conclusions: NS, CP, PS. Jointly developed the structure and arguments for the paper: NS, CP, PS, SK. Made critical revisions and approved final version: NS, CP, PS, SK. All authors reviewed and approved of the final manuscript.

Funding

Authors would like to thank University Grants Commission, Government of India, for funding to carry out this project.

Competing Interests

Author(s) disclose no potential conflicts of interest.

Disclosures and Ethics

As a requirement of publication the authors have provided signed confirmation of their compliance with ethical and legal obligations including but not limited to compliance with ICMJE authorship and competing interests guidelines, that the article is neither under consideration for publication nor published elsewhere, of their compliance with legal and ethical guidelines concerning human and animal research participants (if applicable), and that permission has been obtained for reproduction of any copyrighted material. This article was subject to blind, independent, expert peer review. The reviewers reported no competing interests.

References

1. Mansoor H, Sorayya M, Aishah S, Moshleh MAA. Automatic recognition system for some cyanobacteria using image processing techniques and ANN approach. *2011 International Conference on Environmental and Computer Science*. 2011;19:73–8.
2. Anton A. Algae in the conservation and management of freshwaters. Malayan Nature Society. *Intern Development and Research Centre of Canada*; 1991.
3. Culverhouse PF, Williams R, Benfield M, et al. Automatic image analysis of plankton: future perspectives. *Mar Ecol Progr Ser*. 2006;312:297–309.
4. Weeks PJD, Gauld ID, Gaston KJ, O'Neill MA. Automating the identification of insects: a new solution to an old problem. *Bull Entomol Res*. 1997;87(2):203–11.
5. Culverhouse PF, Williams R, Reguera B, Herry V, González-Gil S. Do experts make mistakes? A comparison of human and machine identification of dinoflagellates. *Mar Ecol Progr Ser*. 2003;247:17–25.
6. Patrick R. What are the requirements for an effective biomonitor? In: Loeb SL, Spacie A, editors. *Biological Monitoring of Aquatic Systems*. Boca Raton, FL: Lewis Publishers; 1994:23–9.
7. Anand N. *Indian Fresh Water Microalgae*. Dehradun, India: Bishen Singh Mahendrapal Singh; 1998.
8. Mishra PK, Srivastava AK, Prakash J, Asthana DK, Rai SK. Some fresh water algae of Eastern Uttar Pradesh, India. *Our Nature*. 2005;3:77–80.
9. Mahendraperumal G, Anand N. *Manual of Fresh Water Algae of Tamilnadu*. Dehradun, India: Bishen Singh Mahendrapal Singh; 2008:124.
10. Sankaran V. Fresh water algal biodiversity of the Anaimalai hills, Tamil Nadu-Chlorophyta-Chlorococcales. *Biology and Biodiversity of Microalagae*. 2009:84–93.
11. Arulmurugan PS, Nagaraj S, Anand N. Biodiversity of fresh water algae from temple tanks of Kerala. *Recent Research in Science and Technology*. 2010;2(6):58–72.
12. Makandar BM, Bhatnagar A. Biodiversity of microalgae and cyanobacteria from freshwater bodies of Jodhpur, Rajesthan (India). *Journal of Algal Biomass Utilization*. 2010;1(3):54–69.
13. Arulmurugan P, Nagaraj S, Anand N. Biodiversity of fresh water algae from Guindy campus of Chennai, India. *Journal of Ecobiotechnology*. 2011;3(10):19–29.
14. Jefferies HP, Berman MS, Poularikas AD, et al. Automated sizing, counting and identification of zooplankton by pattern recognition. *Marine Biology*. 1984;78:329–34.
15. Mosleh MA, Manssor H, Malek S, Milow P, Salleh A. A preliminary study on automated freshwater algae recognition and classification system. *BMC Bioinformatics*. 2012;13 Suppl 17:S25.
16. Katsinis C, Poularikas AD. Image processing and pattern recognition with applications to marine biological images. *Proceedings of the SPIE 7th Meeting on Applications of Digital Image Processing*. San Diego, CA; 1984:324–9.
17. Estep KW, MacIntyre F, Hjörleifsson E, Sieburth JM. MacImage: a user-friendly image-analysis system for the accurate mensuration of marine organisms. *Mar Ecol Prog Ser*. 1986;33:243–53.
18. Kamath SB, Chidambar S, Brinda BR, Kumar MA, Sarada R, Ravishankar GA. Digital image processing-an alternate tool for monitoring of pigment levels in cultured cells with special reference to green alga Haematococcus pluvialis. *Biosens Bioelectron*. 2005;21(5):768–73.
19. Cheng J, Ji G, Feng C, Zheng H. Application of connected morphological operators to image smoothing and edge detection of algae. *International Conference on Information Technology and Computer Science*. 2009:73–6.
20. Schultze-Lam S, Harauz G, Beveridge TJ. Participation of a cyanobacterial S layer in fine-grain mineral formation. *J Bacteriol*. 1992;174(24):7971–81.
21. Blackburn N, Hagstrom A, Wikner J, Cuadros-Hansson R, Bjornsen PK. Rapid determination of bacterial abundance, biovolume, morphology, and growth by neural network-based image analysis. *Appl Environ Microbiol*. 1998;64(9):3246–55.
22. Siena I, Adi K, Gernowo R, Miransari N. Development of algorithm tuberculosis bacteria identification using color segmentation and neural networks. *International Journal of Video and Image Processing and Network Security*. 2012;12(4):9–13.
23. Stefan UT, Ron JW, Davies LJ. Automated object recognition of blue-green algae for measuring water quality using digital image processing techniques. *Environ Int*. 1995;21(2):233–6.
24. Luo Q, Gao Y, Luo J, Chen C, Liang J, Yang C. Automatic identification of diatoms with circular shape using texture analysis. *Journal of Software*. 2011;6(3):428–35.
25. Gupta S, Purkayastha SS. Image enhancement and analysis of microscopic images using various image processing techniques. *Proceedings of the International Journal of Engineering Research and Applications*. 2012;2(3):44–8.
26. Walker RF, Ishikawa K, Kumagai M. Fluorescence-assisted image analysis of freshwater microalgae. *J Microbiol Methods*. 2002;51(2):149–62.
27. Anggraini D, Nugroho AS, Pratama C, Rozi IE, Pragesjvara V, Gunawan M. Automated status identification of microscopic images obtained from malaria thin blood smears using Bayes decision: a study case in Plasmodium falciparum. *Proceedings of the International Conference on Advanced Computer Science & Information Systems*. Jakarta, India; Dec 17–18, 2011:347–52.
28. Gonzalez RC, Woods RE. *Digital Image Processing*. 3rd ed. Readings, MA: Addison-Wesley Publishers; 1992.
29. Canny JA. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Intelligence*. 1986;8(6):679–98.
30. Jolliffe IT. *Principal Component Analysis*. 2nd ed. New York, NY: Springer; 2002.

IJAEGT
ISSN No: 2309-4893
International Journal of Advanced Engineering and Global Technology
Vol-2, Issue-7, July 2014

Transductive Support Vector Machine Web Log Classifier for Identifying Potential Users

S. Chitra
Assistant Professor
Department of Computer Science
Government Arts College (Autonomous)
Coimbatore - 641 018, Tamilnadu, India.
Email:chitra.sivakumar@ymail.com

B. Kalpana
Professor
Department of Computer Science
Avinashilingam Institute for Home Science and Higher Education for Women
Coimbatore - 641 043, Tamilnadu, India.

Abstract--- Web usage mining deals with analyzing the log data of a web server. Web log analysis provides information to predict user behaviour and the efficiency of web site design. Users' characteristics of a web site are analyzed by classifying the users into two categories: (1) Users with purchase interest, (2) Users without purchase interest. Prior to classification, the raw web logs are processed to identify the unique users and to construct sessions. This paper oversimplifies the method of classifying interesting users from a given set of web logs of an e-commerce web server. Modified Transductive Support Vector Machine (TSVM) algorithm is used to classify the web logs into these two categories.

Keywords---User identification, Session Construction, SVM (Support Vector Machine), Transductive SVM(TSVM), Web log Classification.

1. INTRODUCTION

The growth of World Wide Web is very rapid. Many research works are carried out to enhance the efficiency of services provided to the users over the internet. Web usage mining is an area where data mining techniques are applied to web server logs of a particular web site. It is also called web log mining. There are many types of web logs, but typically the log files share the same basic information such as client IP address, request time, requested URL, HTTP status code, referrer url, etc.

The interaction of any user with a website is recorded in the server's log file. That may be used for various analysis and all the data mining techniques like association, classification, clustering, etc., may be applied on those log data. The usefulness lies in improving the performance of the web site in terms of design and content-building.

This paper concentrates on classifying the users of a web site into 2 categories as i) users who are really interested in buying one or more products that are displayed in the pages of the web site. ii) users who browse the pages of the site just to get familiarization about the site, i.e., visitors without purchase interest. This work is divided into three phases. In first phase the data cleaning and path completion processes are done. Second phase consist of session construction and user identification. Third phase focuses on web log classification. This paper uses semi-supervised Trasductive SVM for web log classification. The remainder of the paper is organized as follows. In section 2, we discuss the related work. In section 3, Web Log Classification is discussed. In section 4 Transductive SVM Classification approach is discussed in detail. Results on the experiments conducted are discussed in section 5. Finally conclusion is given in section 6.

2. RELATED WORKS

Web log Classification using TSVM method is an approache that will improve the overall performance of the web server. Existing methods of weblog classification uses Decision Tree, Navie Bayesian Classification, etc,.

Jie Zhang, *et. al.*,[1] proposed that the role of Web usage mining is very important for personalization of Web services. Several approaches have been proposed for extracting the required user sessions from the Web server logs.

Alka Gangrade, *et. al.*,[2] discussed about the techniques for privacy preserving classification under multi-party environment. Further, the two approaches, the classification model and secure multi-party computation algorithms have also been reviewed. The performance analysis of the algorithms has been concentrated in connection with the classification.

Mahesh Thylore Ramakrishna, *et. al.*,[3] use the data-centric view to refine the definition of Web mining. Data-centric view defines web mining with respect to the web data used whereas the alternative method process-centric view defines web mining as a serial collection of operations.

Classification algorithms discussed by Hanady Abdulsalam, *et. al.*,[4], have a set of training samples with labels, a set of test records without labels and classifier to label the test records and rerun the classifier.

Hidenao Abe [5] proposed a classification framework by combining the temporal pattern extraction and rule mining. This framework has been developed for mining if-then rules consisting of temporal patterns in left hand side of the rules. The right hand side helps us to predict both of important events and temporal patterns of important index.

Smith Tsang, *et. al.*, [6], discussed the problem of constructing decision tree classifiers on data with uncertain numerical attributes devised for decision tree construction. Lots of applications for this algorithm are also described in this paper.

C4.5 algorithm for data classification is discussed in Veronica S. Moertini [7]. The algorithm was experimented in utilizing C4.5 for varied dataset.

Enhanced C4.5 algorithm was introduced by Salvatore Ruggieri [8]. It improves C4.5 by adopting the best among the strategies for computing the information gain of continuous attributes. All the strategies adopt a binary search of the threshold in the whole training set starting from the local threshold computed at a node.

The Decision Tree's can deal with one attribute per test node or with more than one. The former approach is called Univariate Decision Tree, and the latter is the Multivariate method. Thales Sehn Korting [9] explains the construction of Multivariate Decision Tree's and the C4.5 algorithm, used to build such trees.

Mahdi Khosravi *et. al.*,[10] proposed a dynamic mining approach for modeling and predicting users' navigation patterns. A.K. Santra, S. Jayasudha,[11] proposed a classification model using Naïve Bayesian classification method. They considered the page count, time spent on each page, number of pages visited, etc., as classifying attributes. The method discussed by Jeffrey Xu Yu , *et. al.*,[12] for classification motivated us to use TSVM classification model for web log data which takes less execution time and gives more accuracy.

3. WEB LOG CLASSIFICATION

Web usage mining will give us a way to analyze the user characteristics of a particular web site. The visitors of the site may be classified into two. 1. Users who are interested in buying the products posted in the web site. 2. Users who are not interested in the buying but browse the site by accident or they visit the site to familiarize the contents of the site. In addition to these two groups of customers, a special group may also be considered, "the network robots. Many search engines use network robots to scramble over the Web. The robots generate numerous access records in Web logs that seriously influence the discovery of customers' patterns. In our method robots are cleaned before classification process starts. We have developed a classification approach that provide web site administrators to know the different type of visitors of their site. Classification algorithm uses the logs and manually create an attribute set for training phase[12].

3.1. Web Site Visitor Characteristics

Extended log format contains ip-address, password, time stamp, url, status code, access methods, the transferred bytes, URLs of referrers' pages, and user agents. We use some of these fields as classification attributes. Our Classification algorithm uses the attribute discretization as Jeffery Xu Zu, et al.,[12].

Browsers of a web site who are really interested in buying the products have the following characteristics:

- Visitors spend time to read the matter present in the page. They spend large amount of time to read the contents of the page. And also the time taken to

navigate from one page to another page is also large.

- Visitors read all the topics, and they search some specific topics also.
- They often use the HTTP POST mode (which sends data to a Web server and retrieves a response), because they're interested in registering with Web sites and are willing to fill out forms with their own information.
- They often access images and graphic files.

On the other hand, visitors with no purchase interest exhibit these access patterns:

- They access many pages quickly to browse contents. The ratio between the time they need to read contents and the time they navigate from one page to another is almost 1.
- They don't navigate down to low-level pages but rather access a large number of high-level child pages, because they're not interested in any specific topics.
- They don't often use POST mode, because they're not interested in registering at Web sites.
- They don't access images and graphic files.

On these bases, we classify two types of accesses (Visitors with purchase interest, and Visitors without purchase interest), we select eight attributes to construct our classifier.

Table 1 shows the eight attributes grouped into three types: *temporal attributes* (A1–A3), *page attributes* (A4–A7), and *communication attributes* (A8).

Discretization of the attribute values are given in table 2. Then, with customers' browsing behaviour, we identify a small set of training data. The dataset's labels should reflect the customers' understanding of their own behaviors.

3.2 Classifier Design

Our algorithm classifies Web logs using Transductive Support Vector Machine Algorithm (TSVM). TSVM is used since our classification algorithms classify unlabelled samples also. TSVM uses a set of training data which have labels and a test data set that are to be labeled correctly using prediction. This is a semi-supervised classification algorithm where concrete training for labels cannot be given. Web logs don't show whether a particular visitor of a web site is really interested in buying the product(s) of the web site.

3.3 Evaluation

To test this algorithm, we conducted experiments on e-commerce web sites' data collected between Jan 2013 to July 2013 and discretize the data. Tables 1 and 2 show the classification attributes and discretized attribute values. The numbers of logs considered are 10,767. We selected 89 sessions for experimentation. We built the classifier using 16 records and the remaining 73 as test data. After data cleaning the count became 1,002. After user identification and session construction, the logs are classified. All the records have eight attributes. The number of the i^{th} element in a record indicates the i^{th} attribute value (table1). Table 3 shows a sample of 16 training records. For example, the first positive training record for a visitor with purchase interest (the first column of table 3) has eight values, each of which comes from table 2. The first value, zero, indicates that the value of attribute A1 is zero, which means that there is no night-accessing[12].

Table 1. Classification Attributes

Temporal Attributes	A1	Accessing between midnight and 7 a.m.
	A2	The total session time
	A3	Statistics such as the time a visitor accesses the site, the total time a visitor stays at the site, and the different amounts of time a visitor stays on various pages
Page Attributes	A4	The total number of accessed pages during the whole session
	A5	A5 The accessing width (the number of child pages accessed from a single page)
	A6	A6 The accessing depth (the depth of t he pages accessed from a single page)
	A7	A7 The percentage of graphic files requested compared to the total number of accessed pages
Communication Attributes	A8	Access methods (such as Get, POST, and Head) that visitors use to interact with the site

Table 2. Discretized Attribute Values

Attribute	Value 0	Value 1	Value 2	Value 3
A1	No	Yes	-	-
A2	≤ 2min	2-5min	5-15 min	15-30 min
A3	≤ 3 sec	3-30 sec	≥ 30 sec	-
A4	≤ 2 pages	2-5 pages	≥ 5 pages	-
A5	≤ 2 pages	2-5 pages	≥ 5 pages	-
A6	1 hierarch y	2-3 hierarch y	≥ 5 hierarchi es	-
A7	0%	0-20%	20-50%	50-100%
A8	Use Get	Use POST	Use Head	-

Table 3. SixteenTraining Records Containing Data on Nine Attributes

With purchase interest	Without purchase interest
0,0,1,2,2,1,3,0	0,0,0,0,0,0,0,0
0,0,0,2,2,2,3,0	0,2,1,0,0,1,3,0
0,0,2,1,1,1,0,1	0,0,0,0,0,0,3,0
0,3,2,2,2,0,1,0	0,0,1,1,1,1,0,0
1,0,0,0,0,0,0,1	0,0,0,1,1,1,0,0
0,2,1,2,2,1,3,0	0,3,1,0,0,0,0,0
1,2,2,1,1,0,2,0	
1,2,2,2,2,2,0,1	
1,1,2,2,2,1,1,0	
1,1,2,1,1,1,0,1	

4 TRANSDUCTIVE SVM CLASSIFICATION

4.1 Binary Classification

Given training data (x_i, y_i) for i = 1 . . . N ,*with* $xi \in Rd$ *and* $yi \in \{-1, 1\}$, learn a classifier f (x) such that

$$f(x_i)\{\geq 0\ yi = +1$$
$$\{< 0\ yi = -1$$

i.e.$y_i\ f(x_i) > 0$ for a correct classification.

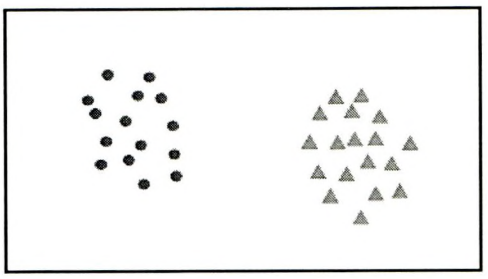

Figure 1 Binary Classification

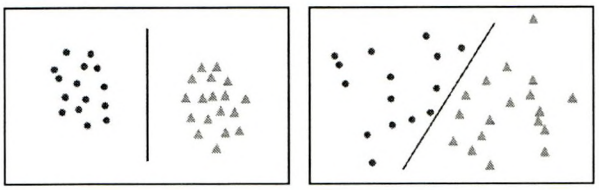

Figure 2Linearly Separable

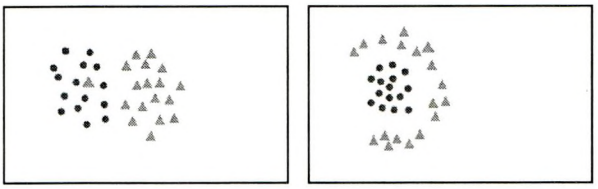

Figure 3 NonLinearly Separable

A linear classifier has the form $f\ (x) = w > x + b$

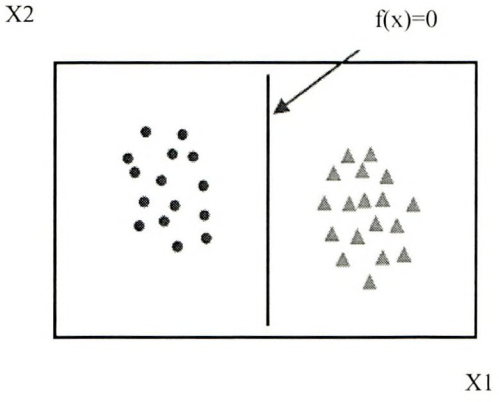

Figure 4 Linear Classifier

- in 2D the discriminant is a line
- w is the normal to the line, and b the bias
- w is known as the weight vector