MATRUSRI ENGINEERING COLLEGE

(Sponsored by Matrusri Education Society, Estd.1980)
(Approved by AICTE & Affiliated to Osmania University)
#16-1-486, Saidabad, Hyderabad – 500 059, Ph: 040-24072764
Email: matrusri.principal@gmail.com, www.matrusri.edu.in

INDEX

3.2.2 Number of books and chapters in edited volumes/books published and papers published in national/ international confer proceedings per teacher during last five years

A.Y: 2015-16

Sl. No.	Name of the teacher	Title of the book/chapters published	Title of the paper	Title of the proceedings of the conference	Year of publication	ISBN/ISSN number of the proceeding	Whether at the time of publication Affiliating Institution Was same Yes/NO	Name of the publisher	Page No
1	Dr. Pallavi Khare	-	Image Segmentation: Fuzzy C-Mans Clustering with Kernel metric and Local Information	National conference on Advances in Information Technology NCAIT 2016	2015-16	ISBN : 978-93-5230-099-0	Yes	MVSR Engineering College,Hyderabad	1-6
2	Dr. P. Vijaya Pal Reddy	-	Emprical evaluatons using character and word n-grams on authorship attribution for telugu text	Springer transactions on International Intelligent computing and applications	2015-16	ISBN: 978-81-322-2268-2	Yes	Advances in Intelligent Systems and Computing book series (AISC, volume 343)	7-2(

Principal

Principal
Matrusri Engineering College
Saidabad, Hyderabad.

16.

Proceedings

National Conference on Advances in Information Technology

22-23 January, 2016

Organised By

Department of Information Technology

M. V. S. R. Engineering College

(Affiliated to Osmania University, Hyd. & Recognised By AICTE, New Delhi)

NADERGUL (P.O), HYDERABAD-501 510.

Technical Sponsors

IEEE

ISTE

COMPUTER SOCIETY OF INDIA

Proceedings of the NCAIT 2016
National Conference on Advances in Information Technology, 22nd – 23rd January, 2016
Department of Information Technology MVSR Engineering College, Nadergul, Hyderabad.

Image Segmentation: Fuzzy C- Means Clustering with Kernel Metric and Local Information

Dr. Pallavi Khare[1] Dr. Akhil Khare[2]

Abstract – **Image segmentation has been an intriguing area for research and developing efficient algorithms, playing a paramount role in high caliber image interpretation and image analysis. Segmentation of images plays an imperative role in medical diagnosis. Such segmentation demands a robust segmentation algorithm against noise. The legendary orthodox fuzzy c-means algorithm is proficiently exploited for clustering in medical image segmentation. FCM is highly sensitive to noise due to the practice of only intensity values for clustering. Thus this paper aims to apply the 'kernel method', instituted on the conventional fuzzy clustering algorithm (FCM) to swap the Euclidean metric norm to a novel kernel induced metric in the data space. Images can be segmented by pixel classification through clustering of all features of interest. In unsupervised methods of clustering algorithms utilizing kernel method, a nonlinear mapping is operated initially in order to map the data into a much higher space feature, and then clustering is executed. The integer of clusters in the multidimensional feature space thus represents the number of classes in the image. As the image is sorted into cluster classes, segmented regions are obtained by examination of the neighborhood pixels for the same class label. Since clustering produces disjointed regions with holes or regions with a single pixel, a post processing algorithm such as region growing, pixel connectivity, or a rule-based algorithm is applied to obtain the final segmented regions.**

Index Terms: **Segmentation, clustering, kernel, nonlinear, FCM**

I. INTRODUCTION

Imaging science has long-drawn-out primarily along three distinct but related lines of research: segmentation, registration and visualization. Registration involves finding the transformation that brings different images of the same object into strict spatial (and/or temporal) congruence. And visualization involves the display, manipulation, and measurement of image data. Finally, segmentation is defined as the process of partitioning an image into a set non-overlapping regions whose union is the entire image where these regions should ideally correspond to objects and their meaningful parts, and background. Most image segmentation algorithms are based on two basic properties that can be extracted from pixel values-discontinuity and similarity-or a combination of them. Segmentation of nontrivial images is a grim problem-made even rigid by non- uniform lighting, shadows, overlapping objects, poor contrast between objects and background, and so on with some degree of success to this date. Image segmentation can be approached as three philosophical perspectives - region, boundary and edge. Image processing techniques for quantitative analysis are primarily used in computational medical analysis. Computer analysis, if performed with the appropriate care and logic, can potentially add objective strength to the interpretation of the expert. Thus, it becomes possible to improve the diagnostic accuracy. It is the important yet elusive capability to accurately recognize and delineate all the individual objects in an image scene.

[1]Assistant Professor, ECE Dept., Matrusri Engineering College, Saidabad, Hyderabad
[2]Professor, CSE, MVSR Engineering College, Hyderabad India

II. EXISTING METHODS

Rudimentally we can cerebrate of several rudimental concepts for segmentation. Pixel predicated methods only utilize the gray values of the individual pixels. Region-predicated methods analyze the gray values in more immensely colossal areas. Conclusively, edge-predicated methods detect edges and then endeavor to follow them. The prevalent constraint of all these approaches is that they are predicated only on local information. Even then they utilize this information only partly. Pixel predicated techniques do not even consider the local neighborhood. Edge-predicated techniques look only for discontinuities, while region predicated techniques analyze homogeneous regions. In situations where we ken the geometric shape of an object, model-predicated segmentation can be applied.

A.PIXEL-PREDICATED DIRECT CLASSIFICATION

The pixel-predicated direct relegation methods use histogram statistics to define single or multiple thresholds to relegate an image pixel by pixel. The threshold for relegating pixels is achieved from the investigation of the histogram of the image. A humble line of attack is to examine the histogram for bimodal dispersal. If the histogram is bimodal, the threshold can be set to the gray assessment analogous to the inmost point in the histogram valley. If not, the image can be fenced off into two or more constituencies utilizing some heuristics about the assets of the image. The histogram of every partition can then be utilized for decisive thresholds. The image f(x, y) can be segmented into two classes using a gray value threshold T such that

$G(x, y) = 1$ if $f(x, y) > T$;else is 0 if $f(x, y) \leq T$

Where G(x, y) is the segmented image with two classes of binary gray values, "1"and "0", and T is the threshold selected at the valley point from the histogram.

B.THRESHOLDING

Thresholding is one of the simplest methods to attain a crusty segmentation a uni-spectral image. Thresholding engenders a binary image in which the pixels belonging to objects have the value 1 whereas the pixels belonging to the background have the value 0. Images are normally acquired as gray-scale images. Idyllically, objects in the image should appear steadily brighter (or darker) than

the background. Using Xi for location (i, j), the thresholded image is given by

IT (Xi) = {1; I (Xi) ≥ τ
0; I (Xi) < τ}

Where τ is the threshold value

Detriments

- Little tolerance to intensity rescaling.
- Difficult to set threshold.
- Slight use of spatial information

C. REGION –BASED SEGME NTAT I O N

Region-growing based segmentation algorithms inspect pixels in the vicinity grounded on a predefined resemblance criterion. The neighborhood pixels with similar properties are merged to form closed regions for segmentation. The region-growing approach can be extended to merging regions instead of merging pixels to form larger meaningful regions with similar properties. Such a region-merging approach is quite effective when the original image is segmented into a large number of regions in the preprocessing phase. Large meaningful regions may provide better correspondence and matching to the object models for recognition and interpretation.

Difficulties

- Low restraint to intensity rescaling.
- Challenging to set mounting criteria and preventing criteria.
- Needs human intervention for defining seed point.

D. EDGE-BASED IMAGE SEGMENTATION

Edge-based methodologies use a spatial filtering method like the Laplacian mask to work out the first-order or second-order gradient statistics of the image. The segmentation of an image into separate objects can be achieved by finding the edges of those objects. This method involves computation of an edge image, containing all (plausible) edges of an original image, then processing the edge image so that only closed object boundaries remain, and finally transforming the result to an ordinary segmented image by filling in the object boundaries.

Disadvantages

- Transforming an edge image to closed boundaries often requires the removal of edges that are caused by noise or other artifacts,
- Intelligent decisions should be made to connect the edge parts that make up a single object where detection of edges remains ambiguous.

E. LEVEL SET METHOD

The level set method was devised by Osher and Sethian to embrace the topology ups and downs of curves. The level set method has been very prosperous in computer graphics and vision, widely used in medical imaging for segmentation and shape recovery. Based on geometric deformable model, the level set scheme translates the tricky evolution 2-D (3-D) close curve (surface) into the evolution of level set function in the space with sophisticated dimension to obtain the benefit in handling the topology changing of the shape.

Drawbacks

•It is difficult to obtain a perfect result when there is a fuzzy or discrete boundary in the region, and the leaking problem is inevitable.

- Solving the partial differential equation of the level set function requires numerical processing at each point of the image domain which is a time consuming process.
- The iteration time increases greatly for too large or too small contour causing the convergence of evolution curve to the contour of object incorrectly.

F. CLUSTERING

Clustering is a process which partitions a given data set or data points into homogeneous groups predicated on given features such that kindred objects are kept in a group whereas dissimilar objects are in different groups. A common approach to image clustering involves addressing the following issues: Image representation, Organizing data, classification of image to a group. The similarity of feature vectors can be represented by an appropriate distance measure such as Euclidean or Mahalanobis distance. Each cluster is represented by its mean (centroid) and variance (spread) associated with the distribution of the corresponding feature vectors of the data points in the cluster. The materialization of clusters is optimized with reverence to an objective function involving pre-specified distance and similarity measures; along with additional constraints such as smoothness. It is the most paramount unsupervised learning quandary. It deals with finding structure in an accumulation of unlabeled data.

1. CALCULATING DISTANCE BETWEEN CLUSTERS

The distance between the centroids of two clusters, i.e., dis (Ki, Kj) = dis(Ci , Cj)

Centroid: the "middle" of a cluster

$$C\,m = \frac{\sum_{i=1}^{N} (tip)}{N}$$

Medoid is defined as the distance between the medoids of two clusters, i.e., dis (K i, K j) = dis (M i , Mj) where medoid is defined as one chosen, centrally located object in the cluster.

Distances are normally used to measure the similarity or dissimilarity between two data objects

a.Minkowski distance

The Minkowski metric favors the prime scaled feature, which dominates others. The problem can be addressed by proper normalization or other weighting schemes applied in the feature space where d is the dimensionality of the data.

d (i ,j)=q√(|xi1-xj1|q+|xi2-xj2|q+…)

Where i = (xi1, xi2, …, xip) and j = (xj1, xj2, …, xjp) are two p-dimensional data objects, and q is a positive integer.

b. If q = 1, d is Manhattan distance

d(i ,j)=|xi1-xj1|+|xi2-xj2|+…+|xip-xjp|

c. If q = 2, d is Euclidean distance which is defined as the distance between two points as the length of the line segment connecting them. The Euclidean distance has an intuitive appeal as it is commonly used to evaluate the proximity of objects in 2- or 3D space. It works well when a data set has "compact" or "isolated" clusters. The advantage of Euclidean distance is that it is intuitively obvious. The disadvantages are costly calculation due to the square root, and its Hon-integral value.

a.Mahalanobis distance
Linear correlation among features can also distort distance measures. This distortion can be alleviated by applying a whitening transformation to the data by using the Mahalanobis distance measure

d(x, y) as
d(x, y) = ((x - y)A-1 (x – y)T)1 /2
where A is a covariance matrix

A Kernel based similarity measure Mercer Kernel functions map data from input space to high, possibly infinite, dimensional feature space. For a finite sample of data X, the kernel function yields a symmetric and positive definite matrix K, where the K (i, j) entry corresponds to the dot product between f (x i) and f (x j) as measured by the kernel function. In feature space, the distance measure between any two points is given by:
f (x i) - f (x j) 2 =<f (x i) ,f (x j)> + < f (x i) ,f (x j) > - 2<f (x i), f (x j)> - f (x i) - f (x j) 2 =k (i ,i) k(i ,j) 2k (i ,j)

2.CLUSTERING CLASSIFICATION
Clustering relegation techniques can be grouped into two main types: supervised and unsupervised. Supervised relegation relies on having example pattern or feature vectors which have already been assigned to a defined class. By contrast, unsupervised relegation does not rely on possession of subsisting examples from a kenned pattern class. The examples are not labeled and we seek to identify groups directly within the overall body of data and features which enables us to distinguish one group from another. Clustering techniques are an example of unsupervised relegation.

a.K-means clustering
K-means (MacQueen, 1967) is one of the simplest unsupervised learning algorithms. It outlines a conceptually simple way to partition a data set into a specified number of clusters k. The algorithm aims to iteratively minimize a simple squared error objective function of the form

$$J = \sum_{j=1}^{k} \sum_{\text{all } i \text{ in class } j} |(xi)j - cj|2 ,$$

Where cj denotes the coordinate vector of the jth cluster and {xij} are the points assigned to the jth cluster. Minimizing J equivalently means reaching that configuration at which switching any point to a cluster other than its currently assigned one will only increase the objective function.

Algorithm for K-means clustering
1 Select the number of desired clusters k. Place the k cluster centers at different initial locations in the image.
2 Assign each data point to the cluster whose center is nearby.
3 Re-compute the cluster centers; the cluster center should be at the average coordinates (center of gravity) of the data points that make up the cluster.
4 Go to step 2 until no more changes befall or a determined number of iterations is reached.

Advantages
1 Fast, robust and more facile to understand.
2 Gives best result when data set are distinct or well disunited from each other.
Shortcomings
1 The learning algorithm requires apriori designation of the number of cluster centers.
2.If there are two highly overlapping data then k- denotes will not be able to resolve that there are two clusters.
3.The learning algorithm is not invariant to non-linear transformations i.e. with different representation of data we get different results (data represented in form of Cartesian co-ordinates and polar coordinates will give different results).
4.Euclidean distance measures can unequally weight underlying factors.
5.The learning algorithm provides the local optima of the squared error function.
6.Randomly culling of the cluster center cannot lead us to the fruitful result.
7.Applicable only when mean is defined i.e. fails for categorical data.
8.Unable to handle strepitous data and outliers.

b.FUZZY C-MEANS CLUSTERING
Fuzzy c-means (FCM) is a scheme of clustering which allows one section of data to belong to dual or supplementary clusters. This method was developed by Dunn in 1973 and enriched by Bezdek in 1981 and it is habitually used in pattern recognition. Main objective of fuzzy c-means algorithm is to minimize:

$$J(U, V) = \sum_{i=1}^{c} \sum_{j=1}^{n} (uij)m\|xi-vj\|2$$

Where, '||xi – vj||' is the Euclidean distance between ith data and jth cluster center.

Algorithm for Fuzzy c-means clustering
1.Randomly select cluster centers.
2.Calculate the fuzzy membership.
3.Compute the fuzzy centers.
4.Repeat step 2) and 3) until the minimum value of the objective function is achieved.
Advantages
1.FCM gives best result for overlapped data set and is comparatively better then k-means algorithm.
2.Data point is assigned membership to each cluster center as a result of which data point may belong to more than one cluster center.
Weaknesses
1.Apriori measurement of the number of clusters.
2.With subordinate value of β we get the better result but at the overhead of extra number of iteration.
3.Euclidean distance measures can inequitably weight underlying factors.

III. PROPOSED METHOD

A. KERNEL FUZZY C MEANS CLUSTERING

The kernel metric Fuzzy C-Means minimizes the following objective function.

$$J\Phi = \sum_{i=1}^{k} \sum_{j=1}^{n} uijm\|\Phi(xj)-\Phi(vi)\|2$$

where, uij denotes the membership of xj in cluster i, Φ(vi) is the center of cluster i in the feature space, and Φ is the mapping from the input space X to the feature space F. Minimization of the function has been proposed only in the case of a Gaussian kernel.

B. KFCM Algorithm:
1. Select initial class prototype {Vi} c
2. Update all memberships Uij

3.Obtain the prototype of clusters in the forms of weighted average Repeat step 2-3 till termination. The termination criterion is □Vnew – Vold ⩽ε
Where □□is the Euclidean norm. V is the vector of cluster centers ε is a small number that can be set by user (here ε =0.01).

C.Kernel-based clustering algorithms have the following main advantages.
1.We can obtain a linearly separable hyper-plane in the high-dimensional, or even in an infinite feature space.
2.They can identify clusters with arbitrary shapes.
3.Kernel-based clustering algorithms, have the capability of dealing with noise and outliers.
4.There is no requirement for prior knowledge to determine the system topological structure.
5.The kernel matrix can provide the means to estimate the number of clusters.

D. Downsides of KFCM
A precarious issue related to KFCM clustering is the selection of an "optimal" kernel for the problem at hand and on the setting of the involved parameters. The kernel function in use must conform to the learning objectives in order to obtain meaningful results for un-labeled data.

IV. BLOCK DIAGRAM OF PROPOSED METHOD

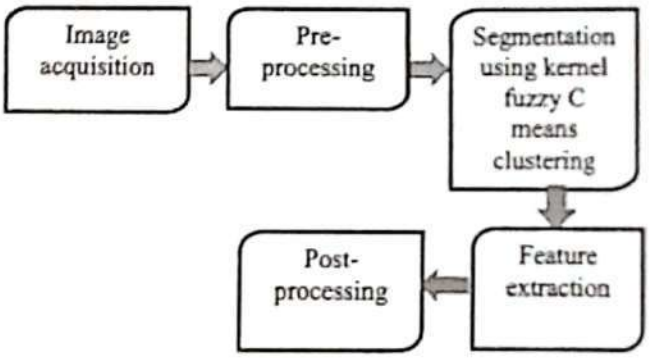

Figure1: Block diagram of the proposed system

The system consists of the following blocks. The raw data is passed through the system as numerical data or in the form of waves. Applicable techniques are applied to get the preprocessed data. Further, clustering returns the cluster centers. Feature extraction is then performed to obtain the attributes that can downright exemplify a given instance. Next a post processing is used to enhance the quality of the final segmented image.

A.Pre-processing
The pre-processing stage is performed to convert all attributes of the data into a numeric form that can be used by the clustering process. This is extremely useful for reduction in dimension of the dataset using normalization. If the values of some attributes vary in different ranges then to reduce the effect of such attributes, all values of the attributes are normalized to lie in some common range, like [0, 1].Pre-processing enhances the visual appearance of images and manipulates datasets.

B.Clustering
The clustering is an important step, as it is an essential precursor to the feature extraction. The input for feature extraction is the pre-processed data, where the labels are stripped off. Clustering is a form of unsupervised learning that helps to find the inherent structure in the data.

C.Feature extraction
It is the process by which certain features of interest within an image are detected and represented for further processing. It marks the transition from pictorial to non-pictorial (alphanumerical, usually quantitative) data representation which can be subsequently used as an input to a number of pattern recognition and classification techniques, which will then label, classify, or recognize the semantic contents of the image or its objects.

D.Post-processing
Image post processing enhance the quality of the finished image, by filtration and other treatments. Here algorithm such as region growing, pixel connectivity or a rule-based algorithm is applied to obtain the final segmented regions.

V. APPLICATIONS

1.Quantitative or semi-quantitative diagnostic image analysis.
2.Surgical planning.
3.Computer assisted surgery

VI. CONCLUSION

In this paper we have discussed the performance of the three algorithms FCM, KFCM and K means. A comparative study suggests the effectiveness of the KFCM algorithm over FCM and K means clustering algorithm.

REFERENCES

[1]. Fast and Robust Fuzzy C-Means Clustering Algorithms Incorporating Local Information for Image Segmentation by Weiling Cai, Songcan Chen and Daoqiang Zhang.
[2]. S. Krinidis and V. Chatzis, "A robust fuzzy local information C-means clustering algorithm," IEEE Trans. Image Process., vol. 19, no. 5, pp. 1328–1337, May 2010.
[3]. P.Sivasangareswari, K.Sathish Kumar, "Fuzzy C-Means Clustering With Local Information and Kernel Metric For Image Segmentation", International Journal of Advanced Research in Computer Science & Technology, ISSN - 2347 - 8446, 2014.
[4]. Ajala Funmilola A, Oke O.A, Adedeji T.O, Alade O.M, Adewusi E.A, "Fuzzy k-c means Clustering Algorithm for Medical Image Segmentation", Journal of Information Engineering and Applications, Vol. 2, ISSN - 2225-0506, 2012.
[5]. J.C.Bezdek ,Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press ,New York,1981
[6]. S. Chen and D. Zhang, "Robust image segmentation using FCM with spatial constraints based on new kernel-induced distance measure," IEEE Trans. Syst., Man, Cybern., B, Cybern., vol. 34, no. 4, pp. 1907–1916, Aug. 2004.
[7]. Y. Tolias and S. Panas, "Image segmentation by a fuzzy clustering algorithm using adaptive spatially constrained membership functions," IEEE Transactions on Systems, Man and Cybernetics, vol. 28, no. 3, pp. 359– 369, March 1998.
[8]. L.Szilagyi ,Z. Benyo, S. Szilagyii, and H. Adam, "MR brain image segmentation using an enhanced fuzzy C-means algorithm," in proce.25th Annu.I nt. Conf. IEEE EMBS, Nov 2003, pp. 17-21.
[9]. F.Masulli and A.Schenone, "A fuzzy clustering based segmentation system as support to diagnosis in medical imaging," med, vol. 16, no 2, pp. 129-147, 1999.Fu, S.K.—Mui, J.K., A Survey on Image Segmentation. Pattern Recognition, Vol. 13, 1981, pp. 3–16
[10]. Huang M, D. ang D, 2001 "An efficient fuzzy entropy clustering algorithm". In Proc. the 2001 IEEE International Conference on Data Mining
[11]. "Medical image analysis" , Atam P. Dhawan, Second edition, Published by John Wiley & Sons, Inc., Hoboken, New Jersey

ABOUT MVSR ENGINEERING COLLEGE

MVSR Engineering College, sponsored by Matrusri Education Society, was founded in 1981 with three disciplines. Over the years it has grown in stature and is presently acknowledged as a premier technical institute in Telangana. The Institution has over 225 well qualified and experienced faculty members and ably supported by technical assistants. The college offers BE 4-Year Degree Programs in Civil, ECE, EEE, CSE, IT, Mechanical and Automobile Engineering disciplines. The Programs are affiliated to Osmania University and are approved by AICTE. B.E. Programs are accredited by the NBA of AICTE. The college also has post-graduate program in M.E.Mech. (CAD / CAM), ECE (Embedded Systems & VLSI Design), M.Tech. (CSE) and M.B.A.

ABOUT DEPARTMENT OF INFORMATION TECHNOLOGY

The Department of Information Technology was started in the year 2000. The current annual intake is 90. The curriculum has been designed after carefully considering the specific needs of IT industry, comprising of a judicious mix of electronics, computer science, and information technology courses with equal emphasis on practical and theoretical aspects, broadly covering computers, communications, controls, and information processing. A sizeable population of students are placed on campus with leading information technology companies such as Infosys, CTS, Wipro, Capgemini, Delloite etc. A significant number of talented students have successfully pursued higher studies at leading institutions in India and abroad. Institutions such as IITs, IIITs, and NITs, Stanford, USC and other top American Universities.

BSP BS Publications
A Unit of BSP Books Pvt. Ltd.
4-4-309/316, Giriraj Lane, Sultan Bazar, Hyderabad - 500 095
Ph: 040-23445688 / 605. Fax: +91-40-23445611.
e-mail: info@bspbooks.net, marketing@bspbooks.net
Website: www.bspbooks.net

ISBN: 978-93-5230-099-0
9 789352 300990

Empirical Evaluations Using Character and Word N-Grams on Authorship Attribution for Telugu Text

Intelligent Computing and Applications pp 613-623 | Cite as

- S. Nagaprasad (1) Email author (nagkanna80@gmail.com)
- T. Raghunadha Reddy (2)
- P. Vijayapal Reddy (3)
- A. Vinaya Babu (4)
- B. VishnuVardhan (5)

1. Department of CSE, Aacharya Nagarjuna University, , Guntur, India
2. Department of CSE, Swarnandhra Institute of Engineering and Technology, , Narsapur, India
3. Department of CSE, Gokaraju Rangaraju Institute of Engineering and Technology, , Hyderabad, India
4. Department of CSE, J.N.T.U. College of Engineering, , Hyderabad, India
5. Department of IT, J.N.T.U. College of Engineering, , Nachupally, Karimnagar, India

Conference paper
First Online: 24 February 2015

Part of the Advances in Intelligent Systems and Computing book series (AISC, volume 343)

Abstract

Authorship attribution (AA) is the task of identifying authors of anonymous texts. It is represented as multi-class text classification task. It is concerned with writing style rather than topic matter. The scalability issue in traditional AA studies concerns with the effect of data size, the amount of data per candidate author. Most stylometry researches tend to focus on long texts per author, but it is not probed in much depth in short texts. This paper investigates the task of AA on Telugu texts written by 12 different authors. Several experiments were conducted on these texts by extracting various lexical and character features of the writing style of each author, using word n-grams and character n-grams as a text representation. The support vector machine (SVM) classifier is employed in order to classify the texts to their authors. AA performance in terms of F_1 measure and accuracy deteriorates as the number of candidate author's increases and size of training data decreases.

Keywords

Authorship attribution Telugu language Support vector machine
Evaluation measures Word n-grams Character n-grams Text classification

References

1. Zhao, Y., Zobel, J.: Searching with style: authorship attribution in classic literature
Google Scholar (https://scholar.google.com/scholar?q=Zhao%2C%20Y.%2C%20Zobel%2C%20J.%3A%20Searching%20with%20style%3A%20authorship%20attribution%20in%20classic%20literature)

2. Stamatatos, E., Fakotakis, N., Kokkinakis, G.: Automatic text categorization in terms of genre and author. Comput. Linguist. **26**, 471–495 (2000)
CrossRef (https://doi.org/10.1162/089120100750105920)
Google Scholar (http://scholar.google.com/scholar_lookup?title=Automatic%20text%20categorization%20in%20terms%20of%20genre%20and%20author&author=E.%20Stamatatos&author=N.%20Fakotakis&author=G.%20Kokkinakis&journal=Comput.%20Linguist.&volume=26&pages=471-495&publication_year=2000)

3. Holmes, D.I.: Authorship attribution. Comput. Humanit. **28**(2), 87–106 (1994)
CrossRef (https://doi.org/10.1007/BF01830689)
Google Scholar (http://scholar.google.com/scholar_lookup?title=Authorship%20attribution&author=DI.%20Holmes&journal=Comput.%20Humanit.&volume=28&issue=2&pages=87-106&publication_year=1994)

4. Zhai, C.X., Lafferty, J.: Model-based feedback in the KL-divergence retrieval model. In: Proceedings of the 10th ACM CIKM International Conference on Information Knowledge Management, ACM Press, Atlanta, Georgia, USA, pp. 403–410 (2001)
Google Scholar (https://scholar.google.com/scholar?q=Zhai%2C%20C.X.%2C%20Lafferty%2C%20J.%3A%20Model-based%20feedback%20in%20the%20KL-divergence%20retrieval%20model.%20In%3A%20Proceedings%20of%20the%2010th%20ACM%20CIKM%20International%20Conference%20on%20Information%20Knowledge%20Management%2C%20ACM%20Press%2C%20Atlanta%2C%20Georgia%2C%20USA%2C%20pp.%20403%E2%80%93410%20%282001%29)

5. Bozkurt, D., Baglıoglu, O., Uyar, E: Authorship attribution: performance of various features and classification methods. Computer and information sciences (2007)
Google Scholar (https://scholar.google.com/scholar?q=Bozkurt%2C%20D.%2C%20Bagl%C4%B1oglu%2C%20O.%2C%20Uyar%2C%20E%3A%20Authorship%20attribution%3A%20performance%20of%20various%20features%20and%20classification%20methods.%20Computer%20and%20information%20sciences%20%282007%29)

6. Zhao,Y., Zobel, J., Vines, P.: Using relative entropy for authorship attribution. In: Proceedings of the 3rd AIRS Asian Information Retrieval Symposium, Springer, Singapore, pp. 92–105 (2006)
Google Scholar (https://scholar.google.com/scholar?q=Zhao%2CY.%2C%20Zobel%2C%20J.%2C%20Vines%2C%20P.%3A%20Using%20relative%20entropy%20for%20authorship%20attribution.%20In%3A%20Proceedings%20of%20the%203rd%20AIRS%20Asian%20Information%20Re

trieval%20Symposium%2C%20Springer%2C%20Singapore%2C%20pp.%2092%E2%80%93105%20%282006%29)

7. Vishnu Vardhan, B., Padmaja Rani, B., Kanaka Durga, A., Pratap Reddy, L., Vinay Babu, A.: Analysis of N-gram model on telugu document classification. In: Proceedings of 2008 IEEE Congress on Evolutionary Computation (CEC 2008), Hong Kong, pp. 3198–3202(1–6 June 2008)
Google Scholar (https://scholar.google.com/scholar?q=Vishnu%20Vardhan%2C%20B.%2C%20Padmaja%20Rani%2C%20B.%2C%20Kanaka%20Durga%2C%20A.%2C%20Pratap%20Reddy%2C%20L.%2C%20Vinay%20Babu%2C%20A.%3A%20Analysis%20of%20N-gram%20model%20on%20telugu%20document%20classification.%20In%3A%20Proceedings%20of%202008%20IEEE%20Congress%20on%20Evolutionary%20Computation%20%28CEC%202008%29%2C%20Hong%20Kong%2C%20pp.%203198%E2%80%933202%281%E2%80%936%20June%202008%29)

8. Klinkenberg, R., Joachims, T.: Detecting concept drift with support vector machines. In: Proceedings of the 17th ICML International Conference on Machine Learning, Morgan Kaufmann Publishers, Stanford, California, USA, pp. 487–494 (2000)
Google Scholar (https://scholar.google.com/scholar?q=Klinkenberg%2C%20R.%2C%20Joachims%2C%20T.%3A%20Detecting%20concept%20drift%20with%20support%20vector%20machines.%20In%3A%20Proceedings%20of%20the%2017th%20ICML%20International%20Conference%20on%20Machine%20Learning%2C%20Morgan%20Kaufmann%20Publishers%2C%20Stanford%2C%20California%2C%20USA%2C%20pp.%20487%E2%80%93494%20%282000%29)

9. Mosteller, F., Wallace, D.: Inference and Disputed Authorship: The Federalist. Addison-Wesley Publishing Company, USA (1964)
zbMATH (http://www.emis.de/MATH-item?0122.14106)
Google Scholar (http://scholar.google.com/scholar_lookup?title=Inference%20and%20Disputed%20Authorship%3A%20The%20Federalist&author=F.%20Mosteller&author=D.%20Wallace&publication_year=1964)

10. Yang, Y.M., Zhang, J., Kisiel, B.: A scalability analysis of classifiers in text categorization. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, Toronto, Canada, pp. 96–103 (2003)
Google Scholar (https://scholar.google.com/scholar?q=Yang%2C%20Y.M.%2C%20Zhang%2C%20J.%2C%20Kisiel%2C%20B.%3A%20A%20scalability%20analysis%20of%20classifiers%20in%20text%20categorization.%20In%3A%20Proceedings%20of%20the%2026th%20Annual%20International%20ACM%20SIGIR%20Conference%20on%20Research%20and%20Development%20in%20Information%20Retrieval%2C%20ACM%20Press%2C%20Toronto%2C%20Canada%2C%20pp.%2096%E2%80%93103%20%282003%29)

11. Yule, G.U.: On sentence-length as a statistical characteristic of style in prose, with applications to two cases of disputed authorship. Biometrika **30**, 363–390 (1938)
CrossRef (https://doi.org/10.1093/biomet/30.1-2.1)
Google Scholar (http://scholar.google.com/scholar_lookup?title=On%20sentence-length%20as%20a%20statistical%20characteristic%20of%20style%20in%20prose%2C%20with%20applications%20to%20two%20cases%20of%20disputed%

20authorship&author=GU.%20Yule&journal=Biometrika&volume=30&pages=363-390&publication_year=1938)

12. Holmes, D.I.: The analysis of literary style: a review. Roy. Stat. Soc. A **148**(4), 328–341 (1985)
CrossRef (https://doi.org/10.2307/2981893)
Google Scholar (http://scholar.google.com/scholar_lookup?title=The%20analysis%20of%20literary%20style%3A%20a%20review&author=DI.%20Holmes&journal=Roy.%20Stat.%20Soc.%20A&volume=148&issue=4&pages=328-341&publication_year=1985)

13. Baayen, H., Halteren, H.V., Neijt, A., Tweedie, F.: An experiment in authorship attribution. In: Proceedings 6th International Conference on the Statistical Analysis of Textual Data, pp. 29–37 (2002)
Google Scholar (https://scholar.google.com/scholar?q=Baayen%2C%20H.%2C%20Halteren%2C%20H.V.%2C%20Neijt%2C%20A.%2C%20Tweedie%2C%20F.%3A%20An%20experiment%20in%20authorship%20attribution.%20In%3A%20Proceedings%206th%20International%20Conference%20on%20the%20Statistical%20Analysis%20of%20Textual%20Data%2C%20pp.%2029%E2%80%9337%20%282002%29)

14. Diederich, J., Kindermann, J., Leopold, E., Paass, G.: Authorship attribution with support vector machines. Appl. Intell. **19**(1–2), 109–123 (2003)
CrossRef (https://doi.org/10.1023/A%3A1023824908771)
zbMATH (http://www.emis.de/MATH-item?1040.68120)
Google Scholar (http://scholar.google.com/scholar_lookup?title=Authorship%20attribution%20with%20support%20vector%20machines&author=J.%20Diederich&author=J.%20Kindermann&author=E.%20Leopold&author=G.%20Paass&journal=Appl.%20Intell.&volume=19&issue=1%E2%80%932&pages=109-123&publication_year=2003)

15. Holmes, D.I., Robertson, M., Paez, R.: Stephen Crane and the New York tribune: a case study in traditional and non-traditional authorship attribution. Comput. Humanit. **35**(3), 315–331 (2001)
CrossRef (https://doi.org/10.1023/A%3A1017549100097)
Google Scholar (http://scholar.google.com/scholar_lookup?title=Stephen%20Crane%20and%20the%20New%20York%20tribune%3A%20a%20case%20study%20in%20traditional%20and%20non-traditional%20authorship%20attribution&author=DI.%20Holmes&author=M.%20Robertson&author=R.%20Paez&journal=Comput.%20Humanit.&volume=35&issue=3&pages=315-331&publication_year=2001)

16. Juola, P., Baayen, H.: A controlled-corpus experiment in authorship identification by cross-entropy. Literary Linguist. Comput. **20**, 59–67 (2003)
CrossRef (https://doi.org/10.1093/llc/fqi024)
Google Scholar (http://scholar.google.com/scholar_lookup?title=A%20controlled-corpus%20experiment%20in%20authorship%20identification%20by%20cross-entropy&author=P.%20Juola&author=H.%20Baayen&journal=Literary%20Linguist.%20Comput.&volume=20&pages=59-67&publication_year=2003)

17. Burrows, J.: Delta: a measure of stylistic difference and a guide to likely authorship. Literary Linguist. Comput. **17**, 267–287 (2002)
CrossRef (https://doi.org/10.1093/llc/17.3.267)
Google Scholar (http://scholar.google.com/scholar_lookup?title=Delta%3A%20a%20measure%20of%20stylistic%20difference%20and%2

0a%20guide%20to%20likely%20authorship&author=J.%20Burrows&journal=Literary%20Linguist.%20Comput.&volume=17&pages=267-287&publication_year=2002)

18. Stamatatos, E., Fakotakis, N., Kokkinakis, G.: Automatic authorship attribution. In: Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Bergen, Norway, pp. 158–164 (1999)
Google Scholar (https://scholar.google.com/scholar?q=Stamatatos%2C%20E.%2C%20Fakotakis%2C%20N.%2C%20Kokkinakis%2C%20G.%3A%20Automatic%20authorship%20attribution.%20In%3A%20Proceedings%20of%20the%209th%20Conference%20of%20the%20European%20Chapter%20of%20the%20Association%20for%20Computational%20Linguistics%2C%20Association%20for%20Computational%20Linguistics%2C%20Bergen%2C%20Norway%2C%20pp.%20158%E2%80%93164%20%281999%29)

19. Stamatatos, E., Fakotakis, N., Kokkinakis, G.: Computer-based authorship attribution without lexical measures. Comput. Humanit. **35**(2), 193–214 (2001)
CrossRef (https://doi.org/10.1023/A%3A1002681919510)
Google Scholar (http://scholar.google.com/scholar_lookup?title=Computer-based%20authorship%20attribution%20without%20lexical%20measures&author=E.%20Stamatatos&author=N.%20Fakotakis&author=G.%20Kokkinakis&journal=Comput.%20Humanit.&volume=35&issue=2&pages=193-214&publication_year=2001)

20. Burrows, J.: Word patterns and story shapes: the statistical analysis of narrative style. Literary Linguist. Comput. **2**, 61–70 (1987)
CrossRef (https://doi.org/10.1093/llc/2.2.61)
Google Scholar (http://scholar.google.com/scholar_lookup?title=Word%20patterns%20and%20story%20shapes%3A%20the%20statistical%20analysis%20of%20narrative%20style&author=J.%20Burrows&journal=Literary%20Linguist.%20Comput.&volume=2&pages=61-70&publication_year=1987)

21. Binongo, J.N.G.: Who wrote the 15th book of Oz? An application of multivariate statistics to authorship attribution. Comput. Linguist. **16**(2), 9–17 (2003)
MathSciNet (http://www.ams.org/mathscinet-getitem?mr=1982502)
Google Scholar (http://scholar.google.com/scholar_lookup?title=Who%20wrote%20the%2015th%20book%20of%20Oz%3F%20An%20application%20of%20multivariate%20statistics%20to%20authorship%20attribution&author=JNG.%20Binongo&journal=Comput.%20Linguist.&volume=16&issue=2&pages=9-17&publication_year=2003)

22. Pol, M.S.: A stylometry-based method to measure intra and inter-authorial faithfulness for forensic applications. In: SIGIR Workshop on Stylistic Analysis of Text for Information Access, ACM Press, Salvador, Bahia, Brazil (2005)
Google Scholar (https://scholar.google.com/scholar?q=Pol%2C%20M.S.%3A%20A%20stylometry-based%20method%20to%20measure%20intra%20and%20inter-authorial%20faithfulness%20for%20forensic%20applications.%20In%3A%20SIGIR%20Workshop%20on%20Stylistic%20Analysis%20of%20Text%20for%20Information%20Access%2C%20ACM%20Press%2C%20Salvador%2C%20Bahia%2C%20Brazil%20%282005%29)

23. Kukushkina, O.V., Polikarpov, A.A., Khmelev, D.V.: Using literal and grammatical statistics for authorship attribution. Probl. Inf. Transm. **37**(2), 172–184 (2001)

CrossRef (https://doi.org/10.1023/A%3A1010478226705)
zbMATH (http://www.emis.de/MATH-item?1008.62118)
MathSciNet (http://www.ams.org/mathscinet-getitem?mr=2099901)
Google Scholar (http://scholar.google.com/scholar_lookup?title=Using%20literal%20and%20grammatical%20statistics%20for%20authorship%20attribution&author=OV.%20Kukushkina&author=AA.%20Polikarpov&author=DV.%20Khmelev&journal=Probl.%20Inf.%20Transm.&volume=37&issue=2&pages=172-184&publication_year=2001)

24. Yang, Y.M., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Proceedings of the 14th ICML International Conference on Machine Learning, Morgan Kaufmann Publishers, Nashville, Tennessee, USA, pp. 412–420 (1997)
Google Scholar (https://scholar.google.com/scholar?q=Yang%2C%20Y.M.%2C%20Pedersen%2C%20J.O.%3A%20A%20comparative%20study%20on%20feature%20selection%20in%20text%20categorization.%20In%3A%20Proceedings%20of%20the%2014th%20ICML%20International%20Conference%20on%20Machine%20Learning%2C%20Morgan%20Kaufmann%20Publishers%2C%20Nashville%2C%20Tennessee%2C%20USA%2C%20pp.%20412%E2%80%93420%20%281997%29)

25. Farringdon, J.M.: Analysing for Authorship: A Guide to the Cusum Technique. University of Wales Press, UK (1996)
Google Scholar (http://scholar.google.com/scholar_lookup?title=Analysing%20for%20Authorship%3A%20A%20Guide%20to%20the%20Cusum%20Technique&author=JM.%20Farringdon&publication_year=1996)

26. Benedetto, D., Caglioti, E., Loreto, V.: Language trees and zipping. Am. Phys. Soc. **88**(4), 048702 (2002)
Google Scholar (http://scholar.google.com/scholar_lookup?title=Language%20trees%20and%20zipping&author=D.%20Benedetto&author=E.%20Caglioti&author=V.%20Loreto&journal=Am.%20Phys.%20Soc.&volume=88&issue=4&pages=048702&publication_year=2002)

27. Jurafsky, D., Martin, J.H.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Prentice Hall, USA (2000)
Google Scholar (http://scholar.google.com/scholar_lookup?title=Speech%20and%20Language%20Processing%3A%20An%20Introduction%20to%20Natural%20Language%20Processing%2C%20Computational%20Linguistics%20and%20Speech%20Recognition&author=D.%20Jurafsky&author=JH.%20Martin&publication_year=2000)

28. Juola, P.: What can we do with small corpora? Document categorization via cross-entropy. In: Proceedings of the Interdisciplinary Workshop on Similarity and Categorization, Edinburgh, UK (1997)
Google Scholar (https://scholar.google.com/scholar?q=Juola%2C%20P.%3A%20What%20can%20we%20do%20with%20small%20corpora%3F%20Document%20categorization%20via%20cross-entropy.%20In%3A%20Proceedings%20of%20the%20Interdisciplinary%20Workshop%20on%20Similarity%20and%20Categorization%2C%20Edinburgh%2C%20UK%20%281997%29)

29. Kjell, B.: Authorship attribution of text samples using neural networks and bayesian classifiers. In: Proceedings of IEEE International Conference on Systems, Man and Cybernetics, IEEE Press, San Antonio, Texas, pp. 1660–1664 (1994a)

Google Scholar (https://scholar.google.com/scholar?q=Kjell%2C%20B.%3A%20Authorship%20attribution%20of%20text%20samples%20using%20neural%20networks%20and%20bayesian%20classifiers.%20In%3A%20Proceedings%20of%20IEEE%20International%20Conference%20on%20Systems%2C%20Man%20and%20Cybernetics%2C%20IEEE%20Press%2C%20San%20Antonio%2C%20Texas%2C%20pp.%201660%E2%80%931664%20%281994a%29)

30. Domingos, P., Pazzani, M.J.: On the optimality of the simple bayesian classifier under zerone loss. Mach. Learn. **29**(2/3), 103–130 (1997)
CrossRef (https://doi.org/10.1023/A%3A1007413511361)
zbMATH (http://www.emis.de/MATH-item?0892.68076)
Google Scholar (http://scholar.google.com/scholar_lookup?title=On%20the%20optimality%20of%20the%20simple%20bayesian%20classifier%20under%20zerone%20loss&author=P.%20Domingos&author=MJ.%20Pazzani&journal=Mach.%20Learn.&volume=29&issue=2%2F3&pages=103-130&publication_year=1997)

31. Gabrilovich, E., Markovitch, S.: Text categorization with many redundant features: using aggressive feature selection to make SVMs competitive with C4.5. In: Proceedings of the 21st ICML International Conference on Machine Learning, ACM Press, Banff, Alberta, Canada, pp. 321–328 (2004)
Google Scholar (https://scholar.google.com/scholar?q=Gabrilovich%2C%20E.%2C%20Markovitch%2C%20S.%3A%20Text%20categorization%20with%20many%20redundant%20features%3A%20using%20aggressive%20feature%20selection%20to%20make%20SVMs%20competitive%20with%20C4.5.%20In%3A%20Proceedings%20of%20the%2021st%20ICML%20International%20Conference%20on%20Machine%20Learning%2C%20ACM%20Press%2C%20Banff%2C%20Alberta%2C%20Canada%2C%20pp.%20321%E2%80%93328%20%282004%29)

32. Zheng, R., Li, J., Chen, H., Huang, Z.: A framework for authorship identification of online messages: writing-style features and classification techniques. J. Am. Soc. Inf. Sci. Technol. **57**(3), 378–393 (2006)
CrossRef (https://doi.org/10.1002/asi.20316)
Google Scholar (http://scholar.google.com/scholar_lookup?title=A%20framework%20for%20authorship%20identification%20of%20online%20messages%3A%20writing-style%20features%20and%20classification%20techniques&author=R.%20Zheng&author=J.%20Li&author=H.%20Chen&author=Z.%20Huang&journal=J.%20Am.%20Soc.%20Inf.%20Sci.%20Technol.&volume=57&issue=3&pages=378-393&publication_year=2006)

33. Vishnu Vardhan, B.,Vijaypal Reddy, P., Govardhan, A.: Corpus based extractive summarization for Indic script. In: International Conference on Asian Language Processing (IALP) IEEE computer society (IALP 2011) pp. 154–157
Google Scholar (https://scholar.google.com/scholar?q=Vishnu%20Vardhan%2C%20B.%2CVijaypal%20Reddy%2C%20P.%2C%20Govardhan%2C%20A.%3A%20Corpus%20based%20extractive%20summarization%20for%20Indic%20script.%20In%3A%20International%20Conference%20on%20Asian%20Language%20Processing%20%28IALP%29%20IEEE%20computer%20society%20%28IALP%202011%29%20pp.%20154%E2%80%93157)

34. Pal Reddy, P.V., Vishnu Murthy, G., Vishnu Vardhan, B., Sarangam, K.: A comparative study on term weighting methods for automated telugu text categorization with effective classifiers. Int. J. Data Min. Knowl. Manage. Process (IJDKP) 3(6) (Nov. 2013)

Google Scholar (https://scholar.google.com/scholar?q=Pal%20Reddy%2C%20P.V.%2C%20Vishnu%20Murthy%2C%20G.%2C%20Vishnu%20Vardhan%2C%20B.%2C%20Sarangam%2C%20K.%3A%20A%20comparative%20study%20on%20term%20weighting%20methods%20for%20automated%20telugu%20text%20categorization%20with%20effective%20classifiers.%20Int.%20J.%20Data%20Min.%20Knowl.%20Manage.%20Process%20%28IJDKP%29%203%286%29%20%28Nov.%202013%29)

35. Vishnu Vardhan, B., Pal Reddy, P.V., Govardhan, A.: Analysis of BMW model for title word selection on Indic scripts. Int. J. Comp. Appl. (IJCA) **18**(8), 21–25 (2011)
Google Scholar (http://scholar.google.com/scholar_lookup?title=Analysis%20of%20BMW%20model%20for%20title%20word%20selection%20on%20Indic%20scripts&author=B.%20Vishnu%20Vardhan&author=PV.%20Pal%20Reddy&author=A.%20Govardhan&journal=Int.%20J.%20Comp.%20Appl.%20%28IJCA%29&volume=18&issue=8&pages=21-25&publication_year=2011)

36. Luyckx, K: Scalability issues in authorship attribution. Ph.D thesis, Faculty of Arts and Philosophy, Dutch UPA University (2010)
Google Scholar (https://scholar.google.com/scholar?q=Luyckx%2C%20K%3A%20Scalability%20issues%20in%20authorship%20attribution.%20Ph.D%20thesis%2C%20Faculty%20of%20Arts%20and%20Philosophy%2C%20Dutch%20UPA%20University%20%282010%29)

37. Houvardas, J., Stamatatos, E.: N-gram feature selection for authorship identification. In: Euzennat, J., Domingue, J. (eds.) Proceeding of Artificial Intelligence: Methodology, Systems, and Applications (AIMSA), pp. 77–86. Springer, Berlin (2006)
CrossRef (https://doi.org/10.1007/11861461_10)
Google Scholar (http://scholar.google.com/scholar_lookup?title=N-gram%20feature%20selection%20for%20authorship%20identification&author=J.%20Houvardas&author=E.%20Stamatatos&pages=77-86&publication_year=2006)

38. Stamatatos, E.: Author identification: using text sampling to handle the class imbalance problem. Inf. Process. Manage. **44**(2), 790–799 (2008)
CrossRef (https://doi.org/10.1016/j.ipm.2007.05.012)
Google Scholar (http://scholar.google.com/scholar_lookup?title=Author%20identification%3A%20using%20text%20sampling%20to%20handle%20the%20class%20imbalance%20problem&author=E.%20Stamatatos&journal=Inf.%20Process.%20Manage.&volume=44&issue=2&pages=790-799&publication_year=2008)

Copyright information

About this paper

Cite this paper as:

Nagaprasad S., Raghunadha Reddy T., Vijayapal Reddy P., Vinaya Babu A., VishnuVardhan B. (2015) Empirical Evaluations Using Character and Word N-Grams on Authorship Attribution for Telugu Text. In: Mandal D., Kar R., Das S., Panigrahi B. (eds) Intelligent Computing and Applications. Advances in

Intelligent Systems and Computing, vol 343. Springer, New Delhi. https://doi.org/10.1007/978-81-322-2268-2_62

- First Online 24 February 2015
- DOI https://doi.org/10.1007/978-81-322-2268-2_62
- Publisher Name Springer, New Delhi
- Print ISBN 978-81-322-2267-5
- Online ISBN 978-81-322-2268-2
- eBook Packages Engineering Engineering (R0)

Personalised recommendations

Empirical Evaluations Using Character and Word N-Grams on Authorship Attribution for Telugu Text

S. Nagaprasad, T. Raghunadha Reddy, P. Vijayapal Reddy, A. Vinaya Babu and B. VishnuVardhan

Abstract Authorship attribution (AA) is the task of identifying authors of anonymous texts. It is represented as multi-class text classification task. It is concerned with writing style rather than topic matter. The scalability issue in traditional AA studies concerns with the effect of data size, the amount of data per candidate author. Most stylometry researches tend to focus on long texts per author, but it is not probed in much depth in short texts. This paper investigates the task of AA on Telugu texts written by 12 different authors. Several experiments were conducted on these texts by extracting various lexical and character features of the writing style of each author, using word n-grams and character n-grams as a text representation. The support vector machine (SVM) classifier is employed in order to classify the texts to their authors. AA performance in terms of F_1 measure and accuracy deteriorates as the number of candidate author's increases and size of training data decreases.

Keywords Authorship attribution · Telugu language · Support vector machine · Evaluation measures · Word n-grams · Character n-grams · Text classification

S. Nagaprasad (✉)
Department of CSE, Aacharya Nagarjuna University, Guntur, India
e-mail: nagkanna80@gmail.com

T. Raghunadha Reddy
Department of CSE, Swarnandhra Institute of Engineering and Technology, Narsapur, India

P. Vijayapal Reddy
Department of CSE, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India

A. Vinaya Babu
Department of CSE, J.N.T.U. College of Engineering, Hyderabad, India

B. VishnuVardhan
Department of IT, J.N.T.U. College of Engineering, Nachupally, Karimnagar, India

D. Mandal et al. (eds.), *Intelligent Computing and Applications*,
Advances in Intelligent Systems and Computing 343,
DOI 10.1007/978-81-322-2268-2_62

1 Introduction

Authorship attribution (AA) is the technique of determining the author of a text when it is ambiguous to identify the real author [1]. Every author has his own writing style. Invariably, AA applications are plagiarism detection, resolving disputed authorship, criminal law, civil law and data security [2]. AA can be viewed as problem of text classification (TC), but it is different from classification in terms of considering embedded author style in addition to text content. Hence, AA is more challenging compared with text classification. In TC, the problem is viewed as identification of related topic or the given test document, whereas in AA, it is viewed as assigning the test document to one the many predefined authors [1]. AA is a research field that is in the last decade on various data sets of various languages. AA is identified as a stylometry problem [3] till computational techniques were enough matured. Usage of computational techniques in AA gives pathway for considering many other aspects other than linguistic features. Different data sets with various sizes were experimented in combination with different features and with different machine learning algorithms. Based on the data set size and the based on the number of authors, features and machine learning approaches behave differently [4]. In order to evaluate the proposed AA method thoroughly, its performance is measured under various conditions [5] such as training corpus size and number of candidate authors. Unless these issues are addressed, it is impossible to claim superiority of any type of features for AA. In this paper, a systematic study of the features of AA, such as effect of author set size; data size on performance; and influence of lexical and character features in a categorization approach using support vector machine (SVM) was presented. The behaviour of SVM and the predictive strength of different types of features using various author sets sizes and varying data sizes on Telugu data set were compared. To our knowledge, this is the first study of these aspects of AA on Telugu data set.

2 Related Work

AA can be viewed as one of the oldest problems and one of the newest research problems in the field of information retrieval. Stylometry is the statistical analysis of literary style. The main assumption behind stylometry is that the authors make certain subconscious and conscious choices in their writing. Some of the features that were used in stylometry include average sentence length, average syllables per word, average word length, distribution of parts of speech, function word usage, the type–token ratio, Simpson's Index, Yule's Characteristic K, entropy, word frequencies and vocabulary distributions [6]. Some models that were used in stylometry include n-grams [7], feature counts, inductive rule learning, Bayesian networks, radial basis function networks, decision trees, nearest neighbour classification and SVM [8]. Mosteller and Wallace [9] propose to select semiautomatically the most

common terms composed mainly by various function words for AA. The earliest studies of AA were reported by Yang et al. [10] and Yule [11], in which statistical methods were used limit data, not only the size of the experimental corpus but also the size of feature set. Yang et al. [10] graphically represented the word length as characteristic curves, and he also in [11] used sentence length to differentiate between authors text.

Many types of lexical features were proposed [12–16] including token-level style markers, frequency of word usage, richness of the vocabulary, including the distribution of vocabulary, the number of hapax legomena. Burrows [17] extracted commonly used words from the collection as the features. Stamatatos in [18] was experimented with vocabulary richness. Stamatatos also pointed out in [19] that merely using features at the token level are not be sufficient for reliable AA. Burrows in [20] first proposed the use of function words as style markers for AA. Baayen in [13] experimented with 42 common function words and eight punctuation symbols. A set of 50 common function words was selected as style markers by Holmes in [15] in order to discriminate between two authors on disputed journal articles. Binongo in [21] also used 50 common function words to examine the authorship. More function words were used by Juola and Baayen [16]. Pol [22] has carried out experiments to discriminate the power of different lexical features. Grammatical-based or syntax-based features in AA were applied by several researchers [13, 23]. Chi-square (χ^2) measure is often used to determine relevant features in authorship attribution [19, 24]. The cumulative sum technique [25] looks at the frequencies of a range of possible habits in use of language. Principal component analysis (PCA) [15–17], Markov chains [13] and compression-based techniques [26] are typical of computational approaches that were proposed for AA. N-grams are widely used in authorship attribution [18, 27]. Juola in [28] proposed a similar approach that was applied to AA, in which the unigram model on the character level was used. Benedetto in [26] used compression approach to different applications including AA. Machine learning approaches were applied to AA in recent years, including neural networks [29], Bayesian classifiers [30], SVMs [31] and decision trees [32].

In this paper, Sect. 3 discusses about the different steps in proposed model such as data preprocessing, feature extraction, feature selection and machine learning approach. The AA accuracy and F_1 measure in terms of data set size and author set size were evaluated in the Sect. 4. The description about the data set collection and the language characteristics was presented in Sect. 4. Section 5 summarizes the work done, and the conclusions were drawn from the results and possible extensions.

3 Author Attribution Model

Authorship attribution is viewed as an automatic text classification task that assigns documents according to a set of predefined author set. AA model consists of various steps as shown in Fig. 1. They are data preprocessing, feature extraction and feature selection. These three steps are performed on both training set and testing set.

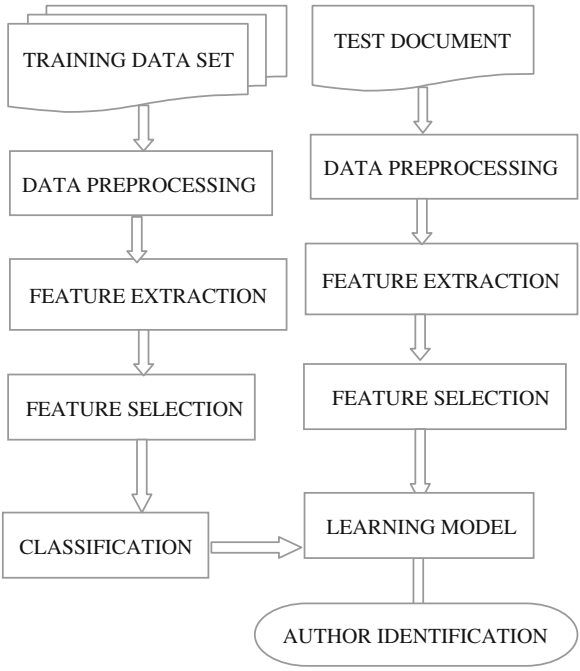

Fig. 1 Text authorship attribution model

A learning model was generated for each author using machine learning technique SVM, and finally, test document is assigned to one of the known authors using the learning model.

3.1 Data Preprocessing

The raw text documents are not suitable for processing by the machine learning algorithms. In this scenario, there is a need to convert these raw documents into a suitable format such as attribute-value representation. This step contains the tokenization, stemming and stop word removal as in [33].

Tokenization is a process of dividing the raw text into meaningful elements. The elements are in the form of paragraphs, sentences, phrases, words and also characters. Based on the characteristics of the language, various types of elements may have various meanings. Tokens such as punctuation symbols, whitespace and numbers are not included in the input token list as they are not deriving any meaning to be extracted from the text. This token acts as an input data for the remaining steps in the proposed model.

To reduce the feature space of the token set, all the words are reduced to their stem form. Stemming is the process of deriving root or base form of the original word which is from the token set. The stemmed word may not be the root of the original word, but

all the related words are mapped to a single word known as stemmed word. For the Telugu data set used in the paper, the stemmed words are derived using the Telugu stemmer called Telugu morphological analyser (TMA) as in [34].

3.2 Feature Extraction

The style of a particular author is generally identified by extracting various features from the text. The various features of a text are broadly categorized into three types as in [32], namely lexical, syntactic and structural features. In this paper, lexical n-grams such as syllable n-grams and word n-grams as features were considered. Vishnu in [35] claimed that lexical n-grams are best features for Telugu text classification. As in [36, 37], lexical features are good for small data sets and also able to capture nuances in different linguistic levels: it is considered that the syllable unigram, syllable bigram, syllable trigram and syllable tetragram at syllable level and also word unigram, word bigram, word trigram and word tetragram as features at word level.

3.3 Feature Selection

The extracted features from the previous step may increase the dimensionality space of the input set. The machine learning classifiers suffer with the problem of curse of dimensionality as the dimensionality space increases. Hence, it is required to remove irrelevant or not most relevant features from the features set. It is carried out by various measures such as document frequency, DIA association factor, chi-square, information gain, mutual information, odds ratio, relevancy score and GSS coefficient. In this paper, chi-square (χ^2) metric [25] is used as a measure for feature selection, which is the most effective feature selection metric in the literature [26]. Chi-square measures the correlation between feature and author set. The relevance of feature t with the author set c is calculated as follows:

$$\chi^2(t,c) = \frac{N * (\mathrm{AD} - \mathrm{BC})^2}{(A+C) * (B+D) * (A+B) * (C+D)}$$

where A is the number of times both feature t and author set c exist; B is the number of times feature t exists, but author set c does not exist; C is the number of times feature t does not exist, but author set c exists; D is the number of times both feature t and author set c does not exist; N be the total number of the training samples. As the value is more, the feature t is more relevant to the set c. Some of the features whose chi-square value is less than the threshold value are considered as non-relevant to the class c.

3.4 Classification

SVM is proved to be an effective machine learning algorithms for text categorization. In [38] for AA, SVM is used to generate learning model by using lexical features such as character n-grams and word n-grams to represent the text. SVM classifier is used to learn the boundaries between author sets where author sets are treated as classes. The learned model generated from the SVM is used for author identification of unknown text as shown in Fig. 1.

3.5 Author Identification

In this step, author is assigned for a given unknown text document. Unknown author text is processed through the various steps as shown in Fig. 1. The vector representation of the text document after dimensionality reduction is given as input to learn the model which is generated from the classifier. The learned model assigns one of the authors from author set to the test document.

4 Results and Discussions

The following Sect. 4.1 briefly describes the characteristics of the language; Sect. 4.2 describes the data set. The different evaluation measures used in authorship attribution are explained in Sect. 4.3; Sect. 4.4 discusses about the influence of the number of authors on AA. Sect. 4.5 presents the influence of the size of the data for each author on AA.

4.1 About the Language

There are more than 150 different languages spoken in India today. Indian languages are characterized by a rich system of inflectional morphology and a productive system of derivation. This means that the number of surface words found to be very large and so the raw feature space, leading to data scarcity. Dravidian languages such as Telugu and Kannada are morphologically more complex and comparable to the languages such as Finnish and Turkish as in [34]. The main reason for richness in morphology of Telugu (and other Dravidian languages) is the significant part of grammar that is to be handled by syntax in English (and other similar languages) to be handled within morphology. Phrases including several words in English are mapped on to a single word in Telugu. Hence, there is a necessity to study the influence of features and different AA approaches on Indian context.

4.2 Data set Description

To address the problem of authorship attribution on Indian context for Telugu language, the data set is collected from Telugu newspapers. The collected data cover various topics. The data set contains 300 news articles written by 12 authors. The average numbers of words are 547 per document. In our experiments, the data set is separated into two groups such as training and testing data. The training set contains 20 text articles for each author. On the other hand, for the test set consists of 5 text articles for each author. The training data set is used to generate learning model using SVM algorithm. Each test document is assigned to one of the authors from the author set using the learning model.

4.3 Evaluation Measures

The performance of the SVM in combination with various lexical features for various data set sizes and author set sizes is evaluated using accuracy and F_1 measure. The accuracy and F_1 measure are defined as follows:

Accuracy is the number of text articles from test set for which the author is correctly assigned over the total number of articles in the test set as in Eq. 1

$$\text{Accuracy} = \frac{\text{Number of documents that are correctly assigned}}{\text{Total number of test documents}} \tag{1}$$

F_1 is calculated as in Eq. 2

$$F_1 = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}} \tag{2}$$

where

$$\text{precision} = \frac{\text{Number of documents correctly author assigned}}{\text{Number of documents author assigned}} \tag{3}$$

and

$$\text{recall} = \frac{\text{Number of documents correctly assigned}}{\text{Total number of test documents}} \tag{4}$$

4.4 Influence of Number of Authors in the Training Set on Authorship Attribution

In this phase, 120 text documents were considered in training set and five documents for each author in the test set. Initially, we have considered 6 authors each with 20 documents in the training set. With the help of training and testing set documents, the precision, recall, F_1 measure and accuracy were evaluated using SVM classifie with different feature vectors. The obtained results were presented in Table 1. Similarly, process is continued by considering 8, 10 and 12 authors in the training set with 15, 12 and 10 text documents for each author subsequently. From the obtained results, it is clear that as the number of authors in the training set increases, there is a significant decrease in the performance both in terms of F_1 and accuracy measures. Word unigram is outperforming compared with all other features. After the word unigram, character trigram is performing well compared with the remaining features. Character-level features performance is good when compared with word-level features.

4.5 Influence of Amount of Data per Author in the Training Set on Authorship Attribution

Articles are collected from various topics written by 12 authors. In total, we have collected 300 news articles from the websites. From the collected text documents, 240 documents are treated as training set and the remaining 60 documents are considered as test set. For the experimental evaluation, each time we have considered 5, 10, 15 and 20 documents per author. The performance of the each feature using SVM in terms of F_1 and accuracy is calculated for each data set as shown in Table 2. From the obtained results, it is clear that as the number of documents in the

Table 1 F_1 and accuracy values for number of authors with number of features using SVM for fixed data size

Feature	F_1 value				Accuracy			
	Number of authors							
	6	8	10	12	6	8	10	12
Character unigram	0.68	0.64	0.61	0.58	0.74	0.71	0.69	0.65
Character bigram	0.75	0.71	0.68	0.63	0.78	0.75	0.70	0.64
Character trigram	**0.82**	0.78	0.75	0.69	**0.85**	0.81	0.74	0.71
Character tetragram	0.79	0.75	0.76	0.67	0.82	0.79	0.77	0.74
Word unigram	**0.84**	0.81	0.77	0.71	**0.87**	0.83	0.79	0.76
Word bigram	0.76	0.73	0.67	0.64	0.79	0.75	0.72	0.67
Word trigram	0.68	0.66	0.64	0.61	0.73	0.71	0.68	0.64
Word tetragram	0.64	0.62	0.60	0.56	0.69	0.70	0.63	0.60

Bold indicates outperforming value compared with other values

Table 2 F_1 and accuracy values for number of documents with number of features using SVM for fixed number of authors

Feature	F_1 value				Accuracy			
	Number of documents per author							
	5	10	15	20	5	10	15	20
Character unigram	0.51	0.58	0.62	0.65	0.58	0.65	0.68	0.71
Character bigram	0.55	0.63	0.65	0.71	0.61	0.64	0.70	0.75
Character trigram	0.59	0.69	0.73	**0.75**	0.65	0.71	0.76	**0.81**
Character tetragram	0.60	0.67	0.70	0.73	0.63	0.74	0.73	0.78
Word unigram	0.66	0.71	0.76	**0.85**	0.68	0.76	0.83	**0.89**
Word bigram	0.58	0.64	0.69	0.76	0.62	0.67	0.69	0.77
Word trigram	0.52	0.61	0.63	0.69	0.59	0.64	0.66	0.69
Word tetragram	0.49	0.56	0.59	0.66		0.60		

Bold indicates outperforming value compared with other values

training set increases there is significant increase in the performance both in terms of F_1 and accuracy measures. Word unigram is outperforming compared with all other features. After the word unigram, character trigram is performing well compared with the remaining features. On an average, the character-level features are exhibiting good performance compared with word-level features. The reason for good performance is, in general, character-level features will gather clues from lexical, syntactic and structural levels, and also character n-grams reduce the sparseness of the data.

5 Conclusion

Addressing the problem of authorship attribution on Telugu text is not yet attempted by any other researcher still now. The work is carried out in this paper is a real motivation towards the language. We have viewed the AA problem as a text classification problem. In this paper, it is evaluated the influence of various lexical features at character and word level with varying lengths and also empirically evaluated the impact of number of authors in the training set by keeping the total number of documents in the training set constant. Similarly, we also studied the influence of data set by keeping the number of authors in the training set as constant. In both cases, we obtained the expected results. In most cases, word unigrams and character trigrams are performing well in terms of F_1 metric and accuracy compared with other features. In this paper, for each feature vector, learning model is generated using SVM.

As a future work, we can investigate the influence of various machine learning algorithms that were investigated for generating best suitable learning models. And, it also considered various other features and its combinations to increase the performance of author identification. We can also extend the scope of study on AA by increasing the data set size.

References

1. Zhao, Y., Zobel, J.: Searching with style: authorship attribution in classic literature
2. Stamatatos, E., Fakotakis, N., Kokkinakis, G.: Automatic text categorization in terms of genre and author. Comput. Linguist. **26**, 471–495 (2000)
3. Holmes, D.I.: Authorship attribution. Comput. Humanit. **28**(2), 87–106 (1994)
4. Zhai, C.X., Lafferty, J.: Model-based feedback in the KL-divergence retrieval model. In: Proceedings of the 10th ACM CIKM International Conference on Information Knowledge Management, ACM Press, Atlanta, Georgia, USA, pp. 403–410 (2001)
5. Bozkurt, D., Baglıoglu, O., Uyar, E: Authorship attribution: performance of various features and classification methods. Computer and information sciences (2007)
6. Zhao,Y., Zobel, J., Vines, P.: Using relative entropy for authorship attribution. In: Proceedings of the 3rd AIRS Asian Information Retrieval Symposium, Springer, Singapore, pp. 92–105 (2006)
7. Vishnu Vardhan, B., Padmaja Rani, B., Kanaka Durga, A., Pratap Reddy, L., Vinay Babu, A.: Analysis of N-gram model on telugu document classification. In: Proceedings of 2008 IEEE Congress on Evolutionary Computation (CEC 2008), Hong Kong, pp. 3198–3202(1–6 June 2008)
8. Klinkenberg, R., Joachims, T.: Detecting concept drift with support vector machines. In: Proceedings of the 17th ICML International Conference on Machine Learning, Morgan Kaufmann Publishers, Stanford, California, USA, pp. 487–494 (2000)
9. Mosteller, F., Wallace, D.: Inference and Disputed Authorship: The Federalist. Addison-Wesley Publishing Company, USA (1964)
10. Yang, Y.M., Zhang, J., Kisiel, B.: A scalability analysis of classifiers in text categorization. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, Toronto, Canada, pp. 96–103 (2003)
11. Yule, G.U.: On sentence-length as a statistical characteristic of style in prose, with applications to two cases of disputed authorship. Biometrika **30**, 363–390 (1938)
12. Holmes, D.I.: The analysis of literary style: a review. Roy. Stat. Soc. A **148**(4), 328–341 (1985)
13. Baayen, H., Halteren, H.V., Neijt, A., Tweedie, F.: An experiment in authorship attribution. In: Proceedings 6th International Conference on the Statistical Analysis of Textual Data, pp. 29–37 (2002)
14. Diederich, J., Kindermann, J., Leopold, E., Paass, G.: Authorship attribution with support vector machines. Appl. Intell. **19**(1–2), 109–123 (2003)
15. Holmes, D.I., Robertson, M., Paez, R.: Stephen Crane and the New York tribune: a case study in traditional and non-traditional authorship attribution. Comput. Humanit. **35**(3), 315–331 (2001)
16. Juola, P., Baayen, H.: A controlled-corpus experiment in authorship identification by cross-entropy. Literary Linguist. Comput. **20**, 59–67 (2003)
17. Burrows, J.: Delta: a measure of stylistic difference and a guide to likely authorship. Literary Linguist. Comput. **17**, 267–287 (2002)
18. Stamatatos, E., Fakotakis, N., Kokkinakis, G.: Automatic authorship attribution. In: Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Bergen, Norway, pp. 158–164 (1999)
19. Stamatatos, E., Fakotakis, N., Kokkinakis, G.: Computer-based authorship attribution without lexical measures. Comput. Humanit. **35**(2), 193–214 (2001)
20. Burrows, J.: Word patterns and story shapes: the statistical analysis of narrative style. Literary Linguist. Comput. **2**, 61–70 (1987)
21. Binongo, J.N.G.: Who wrote the 15th book of Oz? An application of multivariate statistics to authorship attribution. Comput. Linguist. **16**(2), 9–17 (2003)

22. Pol, M.S.: A stylometry-based method to measure intra and inter-authorial faithfulness for forensic applications. In: SIGIR Workshop on Stylistic Analysis of Text for Information Access, ACM Press, Salvador, Bahia, Brazil (2005)
23. Kukushkina, O.V., Polikarpov, A.A., Khmelev, D.V.: Using literal and grammatical statistics for authorship attribution. Probl. Inf. Transm. **37**(2), 172–184 (2001)
24. Yang, Y.M., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Proceedings of the 14th ICML International Conference on Machine Learning, Morgan Kaufmann Publishers, Nashville, Tennessee, USA, pp. 412–420 (1997)
25. Farringdon, J.M.: Analysing for Authorship: A Guide to the Cusum Technique. University of Wales Press, UK (1996)
26. Benedetto, D., Caglioti, E., Loreto, V.: Language trees and zipping. Am. Phys. Soc. **88**(4), 048702 (2002)
27. Jurafsky, D., Martin, J.H.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Prentice Hall, USA (2000)
28. Juola, P.: What can we do with small corpora? Document categorization via cross-entropy. In: Proceedings of the Interdisciplinary Workshop on Similarity and Categorization, Edinburgh, UK (1997)
29. Kjell, B.: Authorship attribution of text samples using neural networks and bayesian classifiers. In: Proceedings of IEEE International Conference on Systems, Man and Cybernetics, IEEE Press, San Antonio, Texas, pp. 1660–1664 (1994a)
30. Domingos, P., Pazzani, M.J.: On the optimality of the simple bayesian classifier under zerone loss. Mach. Learn. **29**(2/3), 103–130 (1997)
31. Gabrilovich, E., Markovitch, S.: Text categorization with many redundant features: using aggressive feature selection to make SVMs competitive with C4.5. In: Proceedings of the 21st ICML International Conference on Machine Learning, ACM Press, Banff, Alberta, Canada, pp. 321–328 (2004)
32. Zheng, R., Li, J., Chen, H., Huang, Z.: A framework for authorship identification of online messages: writing-style features and classification techniques. J. Am. Soc. Inf. Sci. Technol. **57**(3), 378–393 (2006)
33. Vishnu Vardhan, B.,Vijaypal Reddy, P., Govardhan, A.: Corpus based extractive summarization for Indic script. In: International Conference on Asian Language Processing (IALP) IEEE computer society (IALP 2011) pp. 154–157
34. Pal Reddy, P.V., Vishnu Murthy, G., Vishnu Vardhan, B., Sarangam, K.: A comparative study on term weighting methods for automated telugu text categorization with effective classifiers. Int. J. Data Min. Knowl. Manage. Process (IJDKP) 3(6) (Nov. 2013)
35. Vishnu Vardhan, B., Pal Reddy, P.V., Govardhan, A.: Analysis of BMW model for title word selection on Indic scripts. Int. J. Comp. Appl. (IJCA) **18**(8), 21–25 (2011)
36. Luyckx, K: Scalability issues in authorship attribution. Ph.D thesis, Faculty of Arts and Philosophy, Dutch UPA University (2010)
37. Houvardas, J., Stamatatos, E.: N-gram feature selection for authorship identification. In: Euzennat, J., Domingue, J. (eds.) Proceeding of Artificial Intelligence: Methodology, Systems, and Applications (AIMSA), pp. 77–86. Springer, Berlin (2006)
38. Stamatatos, E.: Author identification: using text sampling to handle the class imbalance problem. Inf. Process. Manage. **44**(2), 790–799 (2008)