Australian Journal of Basic and Applied Sciences, 9(11) May 2015, Pages: 790-794

ISSN:1991-8178

Australian Journal of Basic and Applied Sciences

Journal home page: www.ajbasweb.com

A New Method for Linguistic Modeling on Text Identification Using K-Nearest

Nasim Elyasi Soren

Department of Linguistics, Osmania University, Hyderabad, India

ARTICLE INFO

Article history:
Received 25 March 2015
Accepted 28 May 2015
Available online 14 June 2015

Keywords:
Bigram Language Model, Fuzzy K-Nearest Neighbor, Features Extraction, Tifinagh Character Recognition.

ABSTRACT

This hypothesis, develops a new method in Tifinagh personality, identity using both of Closer adjacent algorithm and the program language model simultaneously. The k-N algorithm propound candidates weighted of their membership grade, for each personality image, after the preprocessing of the text portrait and the word segmentation.And by utilizing the language of the Bigram model in order to choose the most appropriate sequence of personality. As a consequence, our method increment the recognition rate.

© 2015 AENSI Publisher All rights reserved.

To Cite This Article: Nasim Elyasi Soren, A New Method for Linguistic Modeling on Text Identification Using K-Nearest. ***Aust. J. Basic & Appl. Sci.,*** *9(11): 790-794, 2015*

INTRODUCTION

In pattern recognition, the k-Nearest Neighbors algorithm (or k-NN for short) is a non-parametric method used for classification andregression (Engin, 2009). In both cases, the input consists of the k closest training examples in the feature space. The output depends on whetherk-NN is used for classification or regression (Ellis, 1997):

1. In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.
2. In k-NN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbors.

K-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms.Both for classification and regression, it can be useful to weight the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of 1/d, where d is the distance to the neighbor (Engin, 2009). The neighbors are taken from a set of objects for which the class (for k-NN classification) or the object property value (for k-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required.A shortcoming of the k-NN algorithm is that it is sensitive to the local structure of the data. The algorithm has nothing to do with and is not to be confused with k-means, another popular machine learning technique.

As mentioned before, given a query point, KNN makes predictions based on the outcome of the K neighbors closest to that point. Therefore, to make predictions with KNN, we need to define a metric for measuring the distance between the query point and cases from the examples sample. One of the most popular choices to measure this distance is known as Euclidean. Other measures include Euclidean squared, City-block, and Chebyshev:

$$D(x,p) = \begin{cases} \sqrt{(x-p)^2} & \text{Euclidean} \\ (x-p)^2 & \text{Euclidean squared} \\ \text{Abs}(x-p) & \text{Cityblock} \\ \text{Max}(|x-p|) & \text{Chebyshev} \end{cases}$$

eve

Corresponding Author: Nasim Elyasi Soren, Department of Linguistics, Osmania University, Hyderabad, India

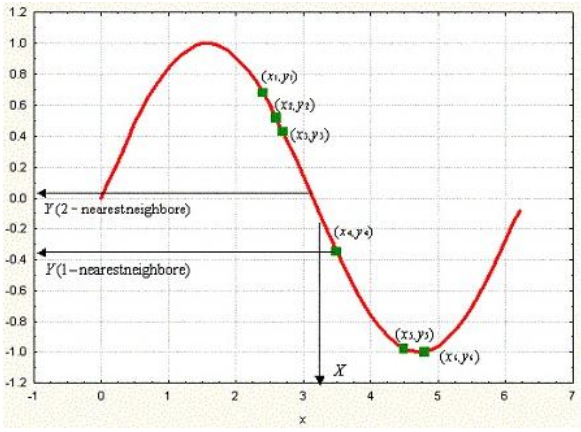

Fig. 1: City-block, and Chebyshev.

Tifinagh (Berber pronunciation: [tifinaɣ]; also written Tifinaɣ in the Berber Latin alphabet, □ □ □ □ □ □in□ Neo-Tifinagh) is a series of abjad and alphabetic scripts used to write Berber languages (Gahin, 2001). A modern derivate of the traditional script, known as Neo-Tifinagh, was introduced in the 20th century. A slightly modified version of the traditional script, called Tifinagh Ircam, is used in a number of Moroccan elementary schools in teaching the Berber language to children as well as a number of publications (Gahin, 2001; Gardner, 2006) The word tifinagh is thought to be a Berberized feminine plural cognate of Punic, through the Berber feminine prefix ti- and LatinPunicus; thus tifinagh could possibly mean "the Phoenician (letters)" (Engin, 2009) or "the Punic letters".

Fig. 2: Simulation of The Tifinagh Alphabet.

Methodology:

In advanced processing techniques similar word segmentation, thinning, slant modification and smoothening are practical.To extractive of the image features, we use a linear compound of the gravity center distance method and the pixel agglomeration method.These two methods need the spatial section of the personality image into W×H boxes so that the portions of the personality will be in some of these

boxes. Some boxes could be empty.The choice of the number of squares is obtained by probation.

Gravity Center Distance, the feature is then obtained by dividing the distance c among G and bottom left nook of the box by the length of the bevel of that box "Fig. 1".One can easily see that this particularity is invariant of the personality image dimensions.Hence an image of whatever size gets transformed into a vector of W×H predesignate dimensions.Our recognition system is made of three blocks, the pre-processing block, the isolated personality identification block and the word recognition block.First, the word illustration is in advance processed by thinning; slant modification and smoothening.The word image is then segmented into a sequence of personified images.For each personality image the fuzzy k-nearest adjacent proposes candidates, each one with a membership grade (the j_{th} candidate proposed for the i_{th} personality image).From these candidates we procreate all m conceivable combinations. The membership grade of each word is given by.

Table 1: Recognition rate.

Grid Size	3*3	3*5	5*5	5*7	7*5	9*11
FKNN	81.25	85.32	89.52	88.2	85.42	91.52
FKKN & Bigrams	83.2	85.21	89.63	90.12	86.45	88.2

Discussionofthe results:

In order to test our method, we developed a java desktop application.The data base used to train the fuzzy close adjacent has 1940 manuscript personality images.The culture used in the bigram model is a set of 1069 disparate Amazigh words written in Tifinagh.In addition to the personality images database and the dictionary size, Results obtained depends on many other parameters: the features elicitation method, the size of the grid used in the features elicitation, the number of adjacent k and the fuzziness parameter m.The gravity center interval method gives data about the circumstances of the darksome pixels in the box. Hence, two boxes with disparate amount of darksome pixels but with the same gravity center will have the same feature value. On the other hand, the pixel agglomeration method gives information about the amount of the darksome pixels in a box.The faint of this method is that it gives no data about the position of black pixels in that box.So, two boxes with the same quantity of darksome pixels will have the same feature value, even if the darksome pixel repartition is not the same.In order to take advantage of both procedures we used a linear combination of these two procedures.So it is the feature given by the pixel density procedure and the one given by the gravity center interval.

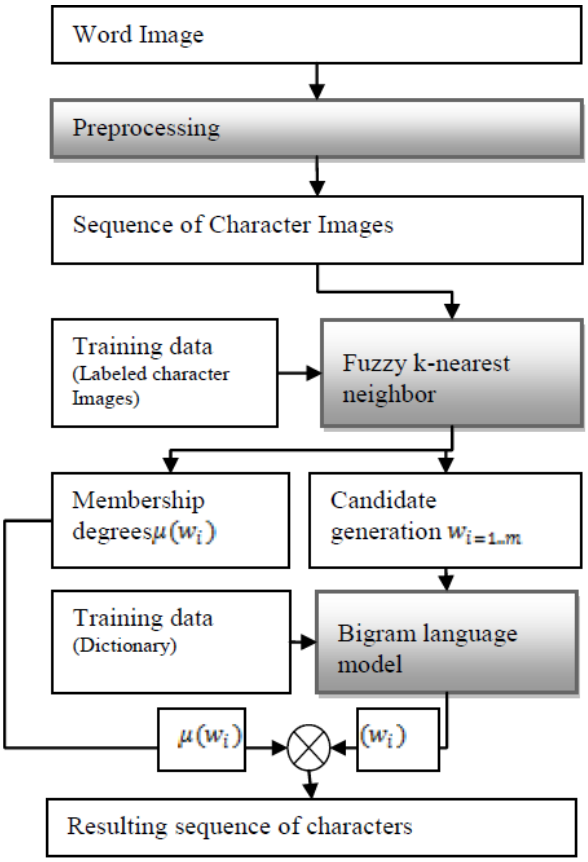

Fig. 3: Recognition System.

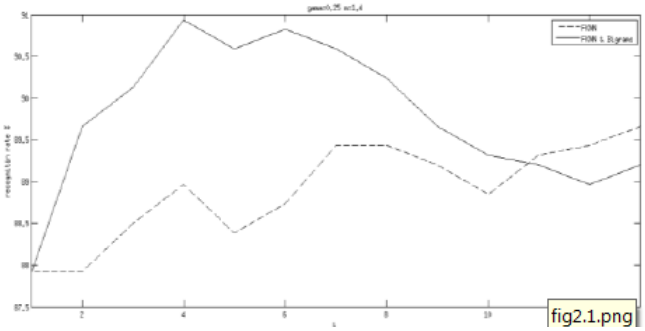

Fig. 4: Identification Rate Subordinate of K, Fuzzy KNN In Dashed Document, FKNN In Dashed Document.

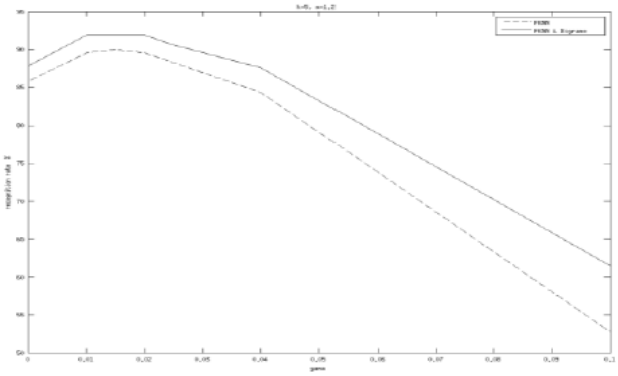

Fig. 5: Identification Measure Subordinate of Γ, Bigrams In Solid Document, FKNN In Dashed Document.

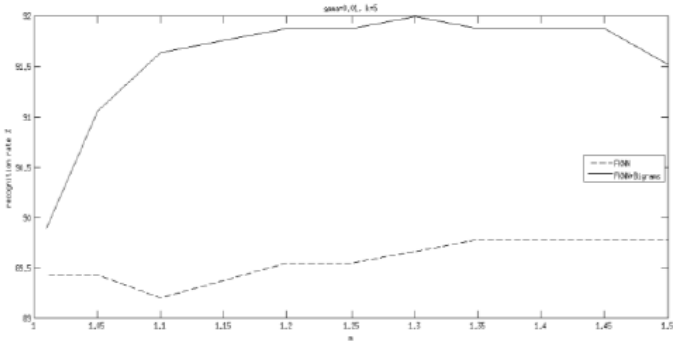

Fig. 6: Identification Rate Subordinate of M, Fuzzy KNN In Dashed Document &FKNN In Dashed Document.

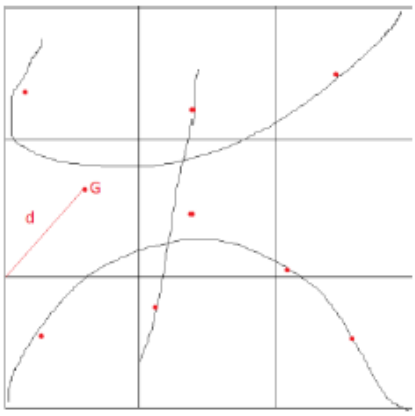

Fig. 7: Gravity Center Interval Procedure Applayed To 'Yaz' Personality

REFERENCES

Cross, G., 1971. The Effect of Belief Systems and the Amount of Debate Experience on the Acquisition of Critical Thinking. PhD Dissertation, University of Utah.

Ellis, R., 1997. Second Language Acquisition. Oxford: Oxford University Press.

Engin, A.O., 2009. Second language learning success and motivation. Social Behavior and Personality, 37: 1035-1041.

Fotos, S., R. Ellis, 1991. Communicating about grammar: A task-based approach. TESOL Quarterly, 25: 605-628.

Gahin, G., D. Myhill, 2001. The communicative approach in Egypt: Exploring the secrets of the pyramids. TEFL Web Journal, 1(2). Retrieved 24 March 2004.

Gardner, R.C., 1985. Social psychology and second language learning: The role of attitudes and motivation. London: Edward Arnold.

Gardner, R.C., 2006. The socio-educational model of Second Language Acquisition: A research paradigm. EUROSLA Yearbook, 6: 237-260.

Grabe,W., 1991. Current development in second language reading research. TESOL Quarterly, 25(3): 375-406.

Halvorsen, A., 2005. Incorporating critical thinking skills development into ESL/EFL courses. Internet TESL Journal, 11(3). Retrieved September 14, 2008.

Kasper, G., 1998. Analyzing verbal protocols. TESOL Quarterly, 32: 358-362.

Kormos, J., 1998. The use of verbal reports in L2 research. TESOL Quarterly, 32: 353-358.

Percival, B. Donald and T. Andrew Walden, 1993. Spectral analysis for physical applications:Multitaper and conventional univariate techniques. Cambridge University Press, Cambridge, UK.

Pressley, M., 2002. Metacognition and Self-Regulated Comprehension, Chapter 13, in Farstrup, A. E. & Samuels, S. Jay. What research has to say about reading instruction? Newark, Del: International Reading Association.

Sato, K., R. Kleinsasser, 1999. Communicative language teaching (CLT): Practical understandings. The Modern Language Journal, 83(4): 494-17.

Shumway, H. Robert and S. David Stoffer, 2006. Time series analysis and its applications. Springer Texts in Statistics. Springer, 2nd edition.

Slepian, David, 1964. Prolate spheroidal wave functions, Fourier analysis, and uncertainty—iv: Extensions to many dimensions; generalized prolate spheroidal functions. Bell System Technical Journal, 43: 3009–3058.

Zheng, X., B. Adamson, 2003. The pedagogy of a secondary school teacher of English in the People's Republic of China: Changing the stereotypes. RELC Journal, 34(3): 323-337.