IT in Industry, Vol. 9, No.2, 2021 Published Online 09-April-2021

Categorization of Text Documents Based on Fuzzy Attitude and Neural Networks

Amir Rajaei[1*], Khadijeh Seimari[2]

[1]Department of Computer Engineering, VelayatUniversity, Iranshahr, IRAN Email:a.rajaei@velayat.ac.ir
[2]Department of Computer Engineering, Hatef University, Zahedan IRAN Email: seimari.kh@gmail.com

Abstract: With the exciting development of Internet and the increasing use of it for providing or acquiring information, we are witnessing an enormous volume of text documents and online images. This is considered as information redundancy, which is one of the prominent features of modern day life. In this regard, fast and accurate access to important and favorite resources is one of the concerns of users of these enormous resources of information. Today, what is of great importance is the lack of methods to find and optimally exploit the information available, rather than the shortage or lack of information. The problem with big image data, the effort to eliminate noise and visual disturbances such as parameters from inappropriate light sources, the inadequacy of color combinations, and many other factors in received images, are very important issues in working on images and processing them. In this regard, the method of classification of the texts from the images using a fuzzy system and neural network based algorithm is suggested. In this method, the location of the fuzzy system is introduced at the begin and end of the neural network synchronized with fuzzification operation and fuzzy inversion. In fact, the main idea in this article is to eliminate or minimize noise in classifying the documents with high inaccuracy.

Keywords: Fuzzy Logic, Neural Network, Natural language processing, Optical Character Recognition

1. Introduction

Today, researchers acknowledge that despite continuous research in their field of work, they cannot keep up-to-date their science together with knowledge progression. For example, the Medline database currently contains 20 million abstracts, and between seven and eight thousand abstracts are added to it each week. While not all articles may be related to a particular knowledge, the number of specialized articles in a particular research area is so large that one cannot claim to have studied all of them. Additionally, the role of in-depth and extend studies and extracting new ideas and knowledge from the material studied is obvious to anyone. The Internet, as the largest source of public information, is made up of hundreds millions of pages of information that, for the sake of its universality and lack of foresight at the time of its formation and growth, bear the same information as authors, researchers, scholars etc. wrote it. The lack of a comprehensive and rigorous standard in the arrangement of texts and the placement of this huge collection of unstructured or partially structured information has caused some confusion for informational society in access to the information needed. So that they incur many time cost for finding the material they need. Researchers have suggested ways to structure

information and have largely prevented this informational disintegration by presenting standard markup languages such as XML. However, what remains is many unstructured texts. To this end, the provision of tools capable to analyze texts by examining them, has led to the formation of a new field in artificial intelligence and information technology known as text learning [1].

Text mining, as a method of extracting knowledge from text, is one of the important issues in a wide range of information management practices. However, providing solutions to deal with this enormous amount of information and to make the best use of it in the creation of mass wisdom is of the utmost importance. In recent years, the importance of texts as high-potential information resources has been widely acknowledged. As, the knowledge discovery from texts is considered one of the most important activities of artificial intelligence and information technology researchers [2].

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

This area covers all activities that somehow seek to acquire knowledge from the text. Text data analysis by machine learning techniques, intelligent information recovery, natural language processing, or other related techniques are all in the context of text learning. One of the methods mentioned is the use of machine learning techniques in the field of text processing. The significant issue is that these techniques were first applied for structured data. So they created a science called data mining. Structured data refers to data that are completely independent of one another but structurally are collected in an identical place. Types of databases are examples of this kind of information. Therefore, the data-mining problem is the acquisition of information and knowledge from this structured collection. However, largely unstructured or semi-structured texts must first be structured by some methods and then the suggested approaches may be used to extract information and knowledge. However, the use of data mining for texts has created another branch in the field of artificial intelligence called text mining. One of the most important activities in this area is the classification of texts. Briefly, text mining tasks include texts classification, clustering of texts, extracting the meaning, generating their categorization, analyzing the emotions, summarizing documents, and modeling entity relationships. Text analysis engages with information recovery, lexical analysis to study vocabulary frequency distribution, pattern recognition, labeling, annotation, information extraction, data mining techniques including connection and link analysis, visualization, and predictive analysis. The ultimate goal is basically to convert text to data for analysis through the use of natural language processing and analytical methods [2].

Natural language processing is one of the most important sub-fields in the broad field of computer science, artificial intelligence, and computational linguistics science that deals with the interaction between computers and natural languages. Thus, processing natural languages focus on communication between human and computer. Therefore, the main challenge in this regard is to understand the natural language and mechanizing the process of understanding and deducing the concepts expressed in a natural language. More precisely, natural language processing involves the use of computer for spoken and written language processing. That is, to enable computers to analyze and infer the speech or text produced in the form and structure of a natural language, or to generate it. Natural language processing applications can be divided into two general categories, including writing applications and speech applications. Writing applications refer to extract specific information from a text, translate a text into another language, or to find some specific documentation in a textual database, such as finding related books in a library. Examples of speech applications in processing languages include human-computer question-and-answer systems, automated telephone-to-customer communication services, learner training systems, or voice control systems [3].

One of the first steps in text mining is text processing. There are a few important things to do in text processing, such as cleaning text, removing ads from web pages, normalizing converted text from binary formats, working with tables, shapes and formulas. The next is text markup, i.e. to divide a string of characters into a set of marks that overcome problems such as apostrophes like "he's", polymorph words like "data-base", "database" or "data base", words like C ++, A / C, or to find signs like "..." or answers for questions like "Whether white space matters or not". The next topic is Parts Of Speech tagging, or the word marking process of a text with its corresponding parts of speech. This is related to grammatical rules based on the sequential probabilities of different words, and requires a corpus (a set of texts or statements) to be tagged manually for machine learning[4].

The next issue is word sense disambiguation, that is to determine which meaning of a word with multiple meanings, is desired in the sentence. In semantic structures, we have two methods: one is full parsing which creates a parse tree for the sentence and the other is partial parsing which creates syntactic structures such as noun expressions and verb groups for the sentence. Which is better? Creating a full parse tree often fails because of grammatical inaccuracies, bad punctuation, new vocabulary, incorrect separation of sentences, incorrect POS tags. Therefore, partial parsing is more commonly used[2].

Texts in the document must be recognized in some way to allow the use of document image written information. This is performed by Optical Character Recognition (OCR) software. The word OCR was originally used only to identify digits and printed letters. The optical suffix in this phrase is in contrast to the magnetic compound term, to distinguish it from the older method of character recognition based on magnetic ink (i.e. Magnetic Ink Character Recognition). Over time and remarkable progress in this field, some techniques have been introduced for handwriting and printed texts recognition which extended the scope of the work on words and phrases. OCR is now widely used to recognize printed documents such as book pages, magazines and print letters [5].

The character recognition system, like a typewriter or typist, reads the text of the document and converts it into a suitable format for storing on a computer. Usually a scanner provides the document image for the OCR. The character recognition system recognizes objects in a document image, which are digits, letters, symbols, and words and stores the corresponding string in the appropriate format. Character recognition systems, like many other intelligent systems, are highly complex. Image processing and pattern recognition are two main bases of these systems. The complexity of these systems varies for different languages. For example, it is easier to write OCRs for Latin languages, because their letters are spelled separately than languages such as Farsi and Arabic that combine letters of a word. This, in addition to the small population of Farsi users, has led to fewer Farsi-language

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

recognition systems. Text classification, namely assigning text documents based on content to one or more predefined categories, is one of the most important issues in text mining. Real-time sorting e-mails or files into hierarchies of folders, identifying the subject of the text, structured search or finding documents related to user's interest, are among the applications of text classification. It is expensive and time-consuming to classify, therefore restricts its application, and thus there is an increasing interest in the development of technologies for automatic text sorting.

1.1. Importance and Necessity of the Research

However, in today's informational societies, what is increasingly important is information and its exchange. The development of related technologies in this regard, is considered. However, a completely new and more prominent step in trans-industrial societies is the creation of new knowledge from previous information. These societies see it as a key to their future success and are working hard in this area. We must address the more serious issues that fall under the scope of Information High Technology, while upgrading the country's information technology and creating the necessary infrastructure as soon as possible. Our purpose in this regard is to use fuzzy logic to classify outputs of the classification, to reduce inaccuracy in the output. In fact, our main idea in this article is to provide management strategies to eliminate or minimize noise in the classification of documents with high inaccuracy.

1.2. Classification

The purpose of data classification is to organize and assign data to separate classes. In this process, an initial model is created based on the training dataset, and then it is used to classify new data. Thus, by applying the obtained model, new data belongings to a given class may be predicted. In other words, classification involves examining the properties of a new object and assigning them to one of the predefined sets [6].

1.3. Noise

Noisy text is a stored electronic connection that cannot be properly classified by a text mining software. In an electronic document, a text is characterized by the difference between the letters and symbols in the HTML code and the desired meaning of the author. Noisy text does not comply with the rules that the program uses to define and classify words, idioms, phrases, and classes in a particular language. Idioms, acronyms or business-exclusive language may all create text. These texts are quite common in chat, blog posts, short messages etc. Other reasons of creating noisy text include poor spelling and punctuation, typographical errors, and poor translations of speech and Optical Character Recognition (OCR) programs. The following are some methods the text mining uses to extract information from the text [7].

1.4. Text Mining Techniques

In general, the methods used in text mining are as follows:

1.4.1. Extract Information and Features

In extracting information, key terms and their relevance are identified in the text. This is done by pattern matching processing. In addition, the extracted phrases and expressions should be standardized. For example, learning and acquisition are recognized as an identical word. The first step in text mining is to extract features in a set of documents so that one can perform calculations and use statistical methods. Text mining uses the words "corpus" and "lexicon". Corpus means a set of documents, and many of the extraction methods depend on corpus. Lexicon (glossary) means a collection of unique words in the corpus. Text documents are represented by the vocabularies (features) they have and the relationships among them. Two main approaches to represent documents are: "bag of words" and "vector space". Text mining usually does not go into deep linguistic analysis but relies on simple representation of the text with the " Bag-of-Words " technique based on "Vector Space". There are different approaches for determining patterns, such as dimensionality reduction, clustering and automatic classification [8].

A) Bag-of-Words

If w is a dictionary, and the set of all words that occurred at least once in a set of documents is D, the Bag-of-Words representation of the document dn is a vector of weights (W1n, ..., Wjn). In other words, the weights are either 0 or 1, indicating whether a word exists in that document or not. It can be said that Win represents the frequency of the ith word in the nth document, which represents the repetition of words [9].

B) Feature Selection

This is to select a subset of features to display a document. The greater the number of features, the less the semantic burden there are. For example, if there are too many features, there may be stop words. Some features are misleading, some are overlapping and lead to double counting, and some algorithms work better with lesser features, because, the more features, the more complex classifiers are created. Therefore, the resulting classifiers space becomes too large. Two ways to select a feature are:

I. Selecting the feature before using it in the classifier. This requires a feature ranking method and the number of choices becomes too many.

II. Selecting the feature based on how well they act in a classifier. This is often an iterative procedure and the

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

classifier is also a part of the feature selection procedure.

The first part of feature extraction is preprocessing the lexicon. This part usually consist of three sections [1]. The details of these sections are as follows.

I. To remove stop words: These are words, which do not change the conceptual content of the sentence, for example "and" and "or" which can be predetermined.

II. Rooting: The process of removing prefixes and suffixes and obtaining the root word. Rooting is often used to retrieve information when the goal is to improve system performance and reduce the number of unique words. Rooting and deleting the stop words reduce the size of the lexicon, thus saving computational resources. Porter rooting algorithm is very common for rooting.

III. Term weighting: One way of coding a text is to count the number of times a phrase appears in the text, called the term-frequency method. Higher frequency phrases, of course, are not necessarily more important, so we weigh the word according to the text, document, or corpus. One of the most popular term weightings seems to be the reverse frequency of a document, where the word frequency is weighted by the total number of times that phrase appears in the corpus.

C) Dimensionality Reduction

The space in which the document is located, usually has thousands dimensions. Given the number of documents along with the corresponding internal distance matrix, it is often prefer to find a lower dimensional space for further analysis. This makes visualization, clustering and sorting easier. Applying dimensionality reduction can then eliminate the noise of data and make better use of statistical data mining methods to find the relationships between documents. The method of dimensionality reduction is called latent semantic indexing analysis in text mining and natural language processing. In addition, there are other methods for dimensionality reduction obtained from matrices extracted from the Term-Document matrix. Another new method of multidimensional scaling is nonlinear dimensionality reduction via isometric mapping. The idea of this method is that sometimes Euclidean distance is not better for representing the distance between two objects, but the least distance between the surfaces that describe the dataset in terms of parameters, is better [2].

1.4.2. Clustering

Clustering is a technique for categorizing documents, which today plays a vital role in information recovery methods. Its purpose is to place similar documents in one cluster such that they are different from those in the other cluster. Unlike classification, groups in clustering are not known beforehand, and it is unclear the clustering is performed on which feature. Clustering algorithms calculate clusters based on data characteristics and measuring similarities or non-similarities. In document clustering, a high volume of documents is given and we have no idea what the documents are, so we use clustering of documents [10].

1.4.3. Classification

Classification is the recognition of the main subject of a document. The purpose of classification is to make it possible to use a model to predict a class of objects labeled as unknown. Classification is used in cases such as validating, identifying groups of clients that share common characteristics and interests, identifying the effectiveness of medications, and the effectiveness of treatment. The purpose of classifying texts is to assign predefined classes to the text documents. In classifying, there is a training set of documents with certain classes. Using this set, the classification model is determined and the new document class is specified. To measure the efficiency of the classification model, a test set is considered independent of the training set. Estimated lables are compared with actual lables of documents. The ratio of properly classified documents to the total documents is called accuracy [10].

1.4.4. Summarizing

Summarizing is an operation to reduce the amount of text in a document while retaining its original meaning. In summarizing, the user specifies that the summarized text to be what percentage of the original text. To summarize is the process of constructing a set of basic concepts of text only in a few lines, this type of text mining seems to yield no new information from the text. Because the author himself probably knew what he was trying to say, and the summary of his writings did not add new information. However, this can make it easier for users to review the contents of the documentation and speed them up on the path to get what they need [10].

1.4.5. Tracking the Subject

The subject tracking system predicts other documents that may be of interest to the user, by keeping user profiles and based on the documents the user has seen so far [2].

1.4.6. Connector of Concepts

This connects existing documentation through identifying common concepts they have, so users will be able to find information they cannot find through traditional search methods. Among the facts that can be gleaned from a set of

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

texts is the relevance and dependence of some concepts to other concepts[1].

1.4.7. Information Representation

This puts a lot of text resources into the visual hierarchy or map and makes it possible to search for them. Informatik V'S docminer is a tool capable of displaying large amounts of information on a map and thereby providing visual analysis [1].

1.4.8. Questions and Answers

In answering natural language questions, the best way to find the answers related to these questions is considered. The START system is capable of answering natural language questions [1].

1.4.9. Text-based Exploration

Allows the user to scroll through a set of documents based on relevant topics and specific terms and to identify key concepts [2].

1.4.10. Analysis of Attitudes

Used to identify attitudes in documentation collected over a specified period of time. For example, it is used to determine if a company has changed its interests from one topic to another [1].

1.4.11. Finding and Analyzing Trends

To describe this application, suppose you are the manager of a business company. Specifically, you should always monitor the activities of your competitors. This could be any information you have obtained from news, stock exchanges, or documents produced by the same company. Today, information is increasingly on the rise, so managing all of these data resources is certainly not possible only with the help of eyes. Text mining allows to automatically find new trends and changes. In fact, what should be expected from text mining is to tell what news is related to what you want in a range of news and what is new among them, what your business developments are, what your interests are, and how the current trends and interests are and how they change. Using this information, the manager need only to use the information discovered to check the competitor's status [10].

2. Literature Review

This section provides an overview of existing algorithms for text document segmentation. In this regard, Bahrestaghi [11] developed an algorithm for structural recognition of the Persian handwritten letters. The purpose of this article is to recognize the Persian handwritten letters separately written by different people. The proposed algorithm is divided into two basic parts. In the first part, attributes are extracted, and in the second part, the letters are recognized by a decision tree as a classifier. At the feature extraction phase, the first step is a pre-processing step to remove the cuts, cavities and noise. Then, by applying the component labeling algorithm, the dots and serifs are separated from the main body of the letter and its type is detected. Finally, the letters are extracted and identified by grouping the sections of the different rows of the following letter image matrix. For a set consisting of 61 samples per letter, the recognition rate was 91.5%, and for another set consisting of 40 samples per letter, 92.12% letters were correctly recognized [11].

Identification of Farsi typed letters with different fonts was suggested by Namazi[12] using fuzzy neural network. Alphabet recognition or optical character recognition is the automation of the process of entering information into a computer. That is, instead of entering data through the keyboard, the scanned information is detected by the computer. In this research, using the results of fuzzy logic and combining it with neural network, it is tried to distinguish Persian printed letters with different fonts. In this method, pseudo-Zernike moments are used as the main feature of the input image. After fuzzification of these moments by the corresponding membership functions, several neural networks with different topologies perform the alphabetical classification and finally, the final classifier identifies the input patterns using these networks [12].

ShahHosseini[13] suggested Persian handwritten letters identification using a neural network. The used neural network is a three-layer perceptron classifier with the post-diffusion learning law. Three types of geometrical moments have been used as features. The major phase of the recognition algorithm is to group the letters in 17 groups. Seventeen groups are defined on the basis of the results of the classification of letters into 33 classes by a neural network. Each group contains one, two, three, or four letters, which are confused with each other, regarding the features used in this classification. To categorize letters, a network with 22 input nodes with the number of moment features, 35 hidden layer nodes and 17 external nodes was used. In the next step of letter recognition algorithm, each group is identified by smaller size networks. The set of sample letters in this article consists of 1650 training samples and 1565 test samples. These samples were imaged with a handheld scanner with a resolution of 200 dpi. 97.82% of training samples and 90.35% of test samples were correctly classified. The mean rate of letter recognition in seventeen groups is 89.48% [13].

Sadri [14] performed the identification of letters and marks of Persian texts using fuzzy logic. This paper presents a new method for recognizing Persian distinct letters, digits and characters. In this research, the principles of pattern recognition and its steps are discussed in general and two methods of pattern recognition are provided and compared.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Then, they describe the characteristics of the Latin and Persian script and compare them from the OCR perspective [14].

Mozafari and Safabakhsh [15] performed identification of discrete letters and digits of Persian handwriting based on neural network, radial basis functions and fractal properties. This paper presents a method based on fractal codes for identifying discrete letters and numbers of Persian manuscripts. Each fractal code, consisting of six parameters as length and width coordinates of the blocks of the domain and the corresponding range, was applied with equal brightness and conversion number. The set of these fractal codes as features of each letter or number is given to the neural network of radial basis functions and the multilayer perceptron neural network. In this way, it is attempted by applying the preprocessing operations, the rotation of the resize and displacement of numbers and letters to be ineffective in identifying them [15].

Kamali*et al.*,[16] studied the correct extraction of the letter features. In this paper, a new method is suggested for feature extraction. Two features are extracted from each box by this method: A) a fixed point for each box whose distance to the origin of the box (bottom left corner) is the average distance of each of the black pixels to the origin. B) An angle equal to the average of the angles of each of the black pixels relative to the horizon. These two features are extracted for each box, and the total of these features are fed into a fuzzy system for training. Due to the fact that in this method, two features of distance and angle are appropriately utilized, the letters recognition accuracy is very high and reaches by 99%. This method is not dependent on the font type and the input letter size and can efficiently work with a slight change in the preprocessing stage for each language [16].

Ahmadi *et al.*,[17] proposed the use of geometrical and textual criteria for detecting objects in plaque images. Plaque location is considered as the first step in identifying a vehicle plaque. This paper presents a method for locating motorcycle plaque in images with diverse and complex properties and backgrounds. Their proposed algorithm is implemented in two steps, once on the original image and once again on the central part of the image that is sized to the original image. Each step exploits simple concepts such as thresholding, morphology and edge detection using gradient calculation for image segmentation. Then, by examining some geometrical criteria of plaque such as area, ratio of length to width, density of edges, and ratio of mean to variance of gradient in each connected area of the image, plaque candidate areas are identified. Finally, a confidence criterion is used to determine the similarity between the candidate area and the plaque. Then, the plaque location is determined by comparing the two-step output. This method yields good results for some of actual images with different imaging conditions. It shows the method independence on environmental conditions, such as lighting conditions, camera angle, the object distance to camera, and plaque orientation [17].

Frey [18] developed an algorithm for identifying English letters using a fuzzy neural network. This paper presents a serious method for identifying English letters based on the Gray Level Co-occurrence Matrix (*GLCM*). After preprocessing on the input image, the GLCM matrix is calculated in four directions (θ = 0, θ = 45, θ = 90 and θ = 135). To account for the tilt when writing a letter, some deviation is also included in the calculation of each direction. This consideration of the little deviation in the calculation of each direction has a large effect on the overall percentage of correct recognition [18].

Asadi[19] presented an algorithm for off-line detection of Persian discrete handwritten letters. Their main purpose is to provide a practical algorithm for reading and recognizing Persian handwritten letters by computer. In this paper, first, different methods and their accuracy are investigated. Then, to increase the accuracy, the combination of these methods is discussed. The main part of this article examines different feature extraction methods for the letter recognition problem. In each method, different parameters and preprocessors as well as three neural network clustering protocols, nearest neighbor and support vector machines are utilized and compared. At the end, the best mode of each method is selected. After this step, existing systems are combined using voting methods. In order to speed up the desired system in each method, the feature extraction section is implemented in parallel. Testing results show that the highest percentage of recognition before combining was for Kirsch function with 97.44% accuracy, and it was 97.73% after combining these methods using voting [19].

Razavi and Kabir[20] recognized the distinctive Persian online letters. In this paper, a method is suggested for online recognition of Persian handwritten letters. In their proposed method for recognizing Persian distinctive handwritten letters, corresponding knowledge of the main body and micro-motions is used simultaneously. In this paper, the distinctive handwritten letters of Persian language are categorized into 18 groups according to the similarity of the main body, and into 11 groups based on the similarity of the micro-motions. In this paper, four point feature sets and one global feature set are extracted from pre-processed samples. In order to obtain the best feature set, several experiments have been performed using point features set as well as using point features along with global features. In order to reduce the computational cost and increase the resolutions of the features, the feature vector dimension is reduced from 102 features to 17 features by using feature dimensionality reduction techniques such as Linear Discriminant Analysis (*LDA*) and Principal Component Analysis (*PCA*). The Support Vector Machine (*SVM*) classification is used to classify the main body of the letters as well as micro-motions. The results show that through using the proposed method, about 98% of Persian online distinctive handwritten letters are correctly recognized [20].

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Soleimani[21] studied the identifying Persian letters using artificial neural networks and vocabulary scope. Due to the different fonts and handwritings, it is a difficult task to recognize letters. In this process, obtaining accurate features of the letters and sub-letters will be a correct process for correctly identifying the letters. This article uses a new approach to integrate fuzzy and neural systems. So, as information differs across lines and varies from a font or personal to other font or person, information processing is also better to be performed in fuzzy. One of the milestones of this system is to detect the network error tolerance against noise. Then it recognizes the letters using the fuzzy system. Using the proposed method, the detection efficiency will

reach up to 99.1% for different letters with different fonts. Finally, using the vocabulary scope, accuracy of the identified letters was checked. This will eventually improve the efficiency of the network by up to 99.3% in word (connected letters) recognition [21].

With the review of related works, it was revealed that a lot of studies have been conducted on the identification and classification of text documents. Therefore, in current paper we try to use fuzzy logic and neural networks for document recognition.

2. Proposed Method

The purpose of this article is to classify text documents based on fuzzy attitude. Also, the most important factor in the confirmation and recognition of documents is fuzzy logic [22]. Given the large number of images and text documents in the dataset, the challenge of memory and time on one hand, and the achievement of acceptable accuracy on the other hand, get importance. Innovation of the proposed method in this paper is to reduce the time and eliminate or minimize noise in classifying documents with a high degree of inaccuracy. With maintained or increased accuracy, we can achieve both goals simultaneously. In this paper, text image segmentation was implemented using fuzzy inference system. Designed using adaptive neural learning techniques, this system is applied on a sample image as input and indicates the probability of existing a particular color for each pixel of the image. The brightness of each pixel will indicate this probability in the gray level of output image. After selecting the threshold value, a binary image is obtained, which can serve as a mask for color segmentation of the input text image. To show the efficiency of the proposed method, we used the method to segment text documents and the results are provided.

In fact, the main idea in this paper is to eliminate or minimize noise in classifying the documents with high degrees of inaccuracy. The algorithms based on fuzzy logic[22] and neural network [23] are used to classify. In addition, attempts are made to eliminate noise and visual differences such as parameters from inappropriate light sources, color mismatches, and other factors at the pre-processing stage. The parameters generating the fuzzy rules and used for optimization are the neural network inputs after the first layer as well as the neural network outputs after the last layer. This algorithm is implemented in MATLAB software. Because of comparing the proposed method with the existing ones, the data set is selected from the UCI site.

The steps of the proposed algorithm are as follows: pre-process is first done on the input image, thereby eliminating the existing noise. Then, the characters in the text are separated from each other and recognized by an artificial neural network. After reconnecting the letters and forming the words, a fuzzy system is started to identify the content, and classify the existing image, identifying what the image concept is and to which class it belongs. The proposed algorithm is as follows:

1. Read the input image
2. Remove the noises
3. Convert to gray level
4. Convert to binary image
5. Remove dots less than 30 pixels in the image
6. Separate found characters
7. Extract the features of each character
8. Fuzzify inputs
9. Train neural network
    I. Build a network using 70% of the data
    II. Perform the training step *i.e.* giving each data with all its features to the first layer of the network
    III. Update network weights
    IV. Calculate output tags
    V. Calculate the return weights for updating weights i.e. learning operations
    VI. Continue the training steps of the network until a threshold error is reached
10. Test 30% of the data
    I. Assign each data to the network input layer.
    II. Examine each test data and specify on their output label
    III. Calculate the least squares error
11. Return neural network results from fuzzy to normal
12. Combine the characters found together
13. Form words and sentences of text as an input to a fuzzy neural system
14. Train neural fuzzy network based on 70% of word inputs from input images
15. Test neural fuzzy network based on 30% of word inputs from input images
16. Identify the class related to the input image

A binary image is an image whose pixels have only one of two possible values 0, 1 or 0 and 255. In MATLAB, binary images can be stored and introduced as intensify images or indexed images. Intensity image or gray level image is defined as an image which only has brightness values and lacks color

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

properties such as shade and purity. These images are defined by two-dimensional matrices in MATLAB, so the value of each element of this matrix represents its corresponding pixel brightness in the related image. The range of elements of this matrix may vary from 0 to 1 or from 0 to 255. In the first case, the matrix data will be of double precision type and in the second case, it will be of type uint8.

Image analysis and refinement consist of three operations: to obtain the value of the image dots and to apply statistical operations on them (image analysis to extract information about its overall structure), image refinement (in order to more clarify image details) and noise elimination (for subsequent processing operations). Additionally, an operation is performed on binary images with the aim of modifying or correcting the components within a binary image. This operation is performed usually in a step before the final processing operation. The final processing operation is the operation in which information is extracted from the image.

3.1. Artificial Neural Network

This step of proposed method is performed for character recognition as well as the classification of the input image based on the content.

Input: Pre-processed data as images of characters.

Output: Input character type.

This is the structure of a network of nodes. These components are in fact simple processing units. For our purpose, here a neuron is a set of nodes in their simplest form, that is, a set containing four nodes: two nodes for input, one bias node and one network node. The neural network is built on a computational unit, so that it captures a vector of inputs with real values and computes a linear combination from these inputs. If the resulted value exceeds a threshold value, the output will be equal to 1 and otherwise equal to -1 [23].

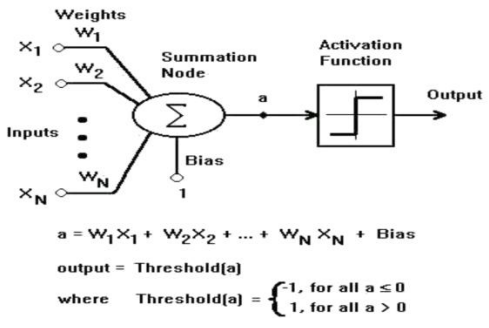

Figure 1. Architecture of Neural Network

Learning after adding the bias:

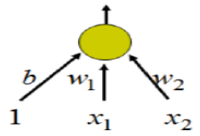

Figure 2. After Learning

$$\hat{y} = w_0 + \sum_{i=1} x_i w_i \qquad (1)$$

Learning can be considered as a level of decision-making in the n-dimensional space of the samples. For the samples on one side of the page it generates a value of 1 and for the values of the other side it generates a value of -1 [6].

A weight is the amount given to the bond or link, and is useful in the learning process. This value is immediately updated by the learning function and naturally is based on a specific rule. Keep in mind that the ultimate goal of the network is to learn to provide the right answers based on the training data provided [6][23].

$$w_i = w_i + \Delta w_i \qquad (2)$$

$$\Delta w_i = \eta(t - o)x_i \qquad (3)$$

T: output

O: Network-generated output

η: the learning rate

It seems that a perfectly appropriate rule for updating the weights may be to randomly assign a value to them until the answer is reached [24]. Theoretically, this can make the network's working time longer than when explicit rules are provided. When the samples are not linearly separable, the network will not converge. Delta law is used to overcome this problem. The main idea of the law is to use the descending gradient to search the space of the hypothesis of possible weights. The basic rule is the back propagation method, which is used to train the network with multiple neurons connected. It is also a basis for a variety of learning algorithms that must search for a hypothetical space containing various continuous hypotheses [24].

3.2. Fuzzy System

It first appeared for new computation following the fuzzy set theory proposed by Professor Lotfizadeh, 1995. The word Fuzzy means inaccurate, imprecise and vague. The application of this area in software science can be easily defined as follows: Fuzzy logic goes beyond the logic of the values of "zero and one" of classical software and opens a new window to the world of software and computing science. Because it also exploits and challenges the infinite and floating space between zero and one digits in its logic and reasoning. The

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

fuzzy set theory provides a framework used for the realization of fuzzy rules based systems, or user interfaces, for many disciplines such as control, decision-making and pattern recognition systems. The rule-based fuzzy system consists of a fuzzy user interface, a rule-base, a database, a decision making unit, and finally a non-fuzzy user interface [22]. These five functional parts are described as below:

- A base contains a number of IF-THEN fuzzy rules.
- A database that defines membership functions (MFs) for fuzzy sets.
- A fuzzy interface that converts the value of each input to a fuzzy value.
- A decision-making unit that performs the inference operations on the rules and generates the fuzzy results.
- A non-fuzzy interface that converts fuzzy results to non-fuzzy results.

The fuzzy system design and implementation steps are as follows[22]:

A. Creating fuzzy sets and mapping input numbers to fuzzy membership values.
B. Formulation of fuzzy rules and calculation of fuzzy inputs and outputs.
C. If an output depends on more than one rule, then this step becomes a single unit.
D. Converting fuzzy values to normal input equivalent values.

In order to fuzzify the inputs, first the fuzzy membership functions are defined, then the properties of each input image after extraction are shown as composed of some number as:

$$X= [x_1, x_2, \dots , x_D] \quad (4)$$

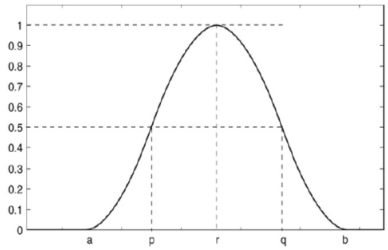

Figure 3. View of Membership Function

The value of r or membership function center [25] for each feature is equal to the mean of the training samples for the same feature, (*i.e.* y).

$$r = mean\ (y) \quad (5)$$

$$p = mean\ (y) - \frac{[\max(y) - \min\ (y)]}{2} \quad (6)$$

$$q = mean\ (y) + \frac{[max(y) - min\ (y)]}{2} \quad (7)$$

$$f_i = 2^{m-1}[\frac{(x-a)}{(r-a)}]^m \quad (8)$$

However, for values less than "a" and more than "b", the output fi equals to zero.For the values $p < x <= r$, the output is:

$$f_i = 1 - 2^{m-1}[\frac{(x-a)}{(r-a)}]^m \quad (9)$$

For the values $r< x<=q$, the output is:

$$f_i = 2^{m-1}[\frac{(x-a)}{(b-a)}]^m \quad (10)$$

For the values $q< x<=b$, the output is:

$$f_i = 1 - 2^{m-1}[\frac{(b-x)}{(b-r)}]^m \quad (11)$$

For an input image, after the fuzzification operation, the membership matrix is as follows:

$$F(x) = \begin{bmatrix} f_{1.1}(x_1) & f_{1.2}(x_1) & \dots & f_{1.C}(x_1) \\ f_{2.1}(x_2) & f_{2.2}(x_2) & \dots & f_{2.C}(x_1) \\ \dots & \dots & \dots & \dots \\ f_{D.1}(x_D) & f_{D.2}(x_D) & \dots & f_{D.C}(x_D) \end{bmatrix} \quad (12)$$

$F_{d,c}(x_i)$*:* Membership value for the D^{th} feature of the C^{th}class. For example, $F_{4,5}(x_4)$ corresponds to the 4th feature of 5^{th} class of input image. For the fuzzy recovery operation in the output layer of the neural network, the output of the network belongs to the class which membership function has the most value among all the character classes [22][25].

3.3. Algorithm Settings

It is necessary to initialize the variables before the simulation. The values of these variables are:

1. The three-layer neural network consists of an input layer, a hidden layer and an output layer.
2. The number of input layer neurons is equal to the number of features extracted from the input image.
3. The number of hidden layer neurons is equal to the average number of neurons in the input and output layers.
4. The number of output layer neurons is equal to the number of classes or is the same number of characters in the first network as well as the number of content categories for the second network.
5. In order to extract features, 50x50 image sizes are considered.
6. Neural network learning rate is 0.01
7. The number of neural network courses is 0.2

However, in order to check different values for the best results in different implementations, we change these values and compare the results with each other and with other articles.

In order to simulate the proposed method, MATLAB software is used in this research and statistical analysis of input and output data

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

is performed in Excel software. A number of scanned images with different content are used for initial examination of the proposed system, and the proposed method is tested according to them.

Number of training samples is 3823. Number of testing samples is 1797. Number of features is 1 + 64 output class labels. All features values is a number between 0 and 16. Values of output classes labels is a number between 0 and 9.

Table 1. Distribution of Classes in Training Samples

Class No.	0	1	2	3	4	5	6	7	8	9
Number of Samples	376	389	380	389	387	376	377	387	380	382

Distribution of classes in testing samples:

Table 2. Distribution of Classes in Testing Samples

Class No.	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Number of Samples	178	182	177	183	181	182	181	179	174	180

After simulation and testing the proposed method, it is necessary to analyze the results, so the following measures are suggested.

A. Calculating overall system error

B. Calculating the execution time of different parts of the system including preprocessing, neural network and fuzzy system, as well as the total time of content recognition or classification of an input image document.

C. Accuracy of neural network in training and testing stages

D. Precision of the Fuzzy Neural System

Training and test samples are used to calculate the accuracy of the system. Determining the error rate in the system and the degree of overall accuracy of the system depend on the correct or incorrect recognition of the samples at each stage of training or testing. The parameters evaluated are the accuracy of the fuzzy neural system, the overall error and each algorithm separately and the time of implementation of the proposed method. Fuzzy rules are applied to optimize the neural network inputs, after the first layer, as well as the neural network outputs, after the last layer.

In the first step of the proposed method, the texts of images must be converted into text documents. The output of this phase is a text file in which the letters are displayed as a text file. Firstly, input images are received. Then the necessary conversions are made on them, so the modifications can be easily done. The purpose of these collections is to construct a template from the input character. It means that all the characters are read first and placed together.

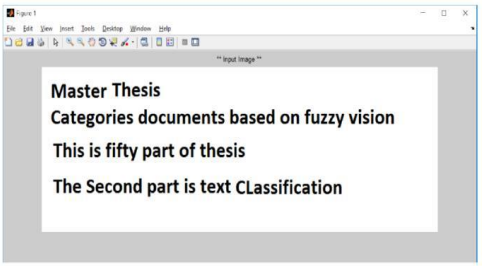

Fig.4. Input Image

As shown in the figure, the input image is a photo containing the characters of the image. The output of this step will be as follows.

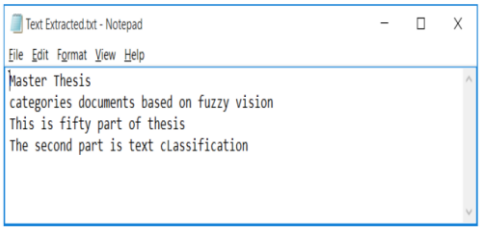

Fig.5. Output Image

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

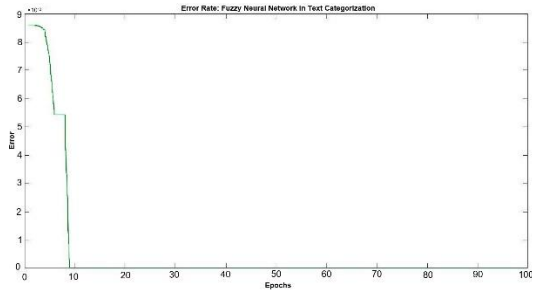

Figure 5. Calculation of Error Rate

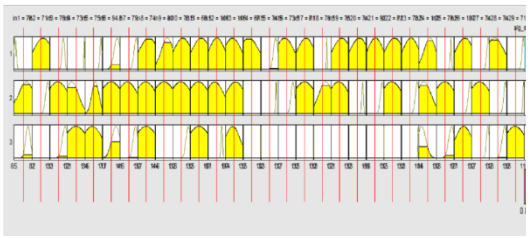

Figure 6. The First Fuzzy Rues

The following figure displays fuzzy rules made by the algorithm using features and output labels. It shows the input fuzzy membership functions, and the output is in the right bottom.

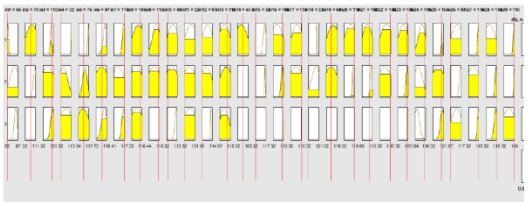

Figure 7. The Second Fuzzy Rues

In the input part of each data, one can enter the ASCII codes of each data. The upper part of the figure shows the fuzzy rules for each character. Also, the last column shows the classification result which is between 0.8 to 3.2.

For example, the sentences entered into the proposed algorithm are:

"Mathematicians seek out patterns and use them to formulate new conjectures. Mathematicians resolve the truth or falsity of conjectures by mathematical proof. When mathematical structures are good models of real phenomena, then mathematical reasoning can provide insight or predictions about nature. Through the use of abstraction and logic, mathematics developed from counting, calculation, measurement, and the systematic study of the shapes and motions of physical objects. Practical mathematics has been a human activity from as far back as written records exist. The research required to solve mathematical problems can take years or even centuries of sustained inquiry., Math"

A text file is displayed in the output of the algorithm, where all the characters are displayed by uppercase and lowercase letters. Then the text extracted from the images is classified with the purpose of feature extraction. Initially, we extract the features of some of the texts about computer, math and physics and save them in a file, where each of extracted features and the following labels are assigned to each group. These features are obtained based on ASCII code of each character.

$$(Computer = 1 \quad - \quad Math = 2 \quad - \quad Physics = 3)$$

In the second step of the proposed method, the characters stored in the file, are removed, leaving only the codes served as input of the algorithm. At this step, a neural fuzzy algorithm is developed and the total sum of data is given to this algorithm for classification. The results of this phase show the reduction rate of measurement error of the proposed method as shown in the following figure.

The calculation of error rate of suggested method for different steps of neural fuzzy algorithm:

It can be deduced from the above text that it is mathematical; our solution to this problem is approximately a number equal to 2. The results of the other data tests are shown in the table below:

Today, computers help make jobs that used to be complicated much simpler. For example, you can write a letter in a word processor, edit it anytime, spell check, print copies, and send it to someone across the world in a matter of seconds. All of these activities would have taken someone days, if not months, to do before computers. Also, all of the above is just a small fraction of what computers can do.,Computer	\
The basic components of a modern digital computer are: Input Device, Output Device, Central Processor Unit (CPU), mass storage device and memory. A Typical modern computer uses LSI Chips. ,Computer	\

The results show that 98% of the recognitions of each text category are correctly selected.

IV. Conclusion

Today, humans have concluded that information plays the most important role, both in life and in business. The huge amounts of data collected in different ways need to be processed. One way to analyze the problem is to use different patterns and to find the common within the data. Due to the increasing growth of access to electronic documents, the classification has become particularly important. In general, classification means assigning documents to predefined classes, which has attracted attention in recent years. Classification can be used in high volume information management. Automated classification is not separated from this topic and is widely used in various fields such as, archiving documents, categorizing new articles, categorizing web pages, finding a person's home page, automatically learning users' research interests, automatic filtering of e-mails on the basis of content, and so on. Segmentation is one of the most important pre-processing steps for pattern recognition and image understanding. In this paper, a text image segmentation method is suggested and

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

implemented using fuzzy inference system. The data used in this article are in the form of images with different properties. These documents have been prepared in many different ways. It also seems that using the hybrid approach can improve the accuracy of the model. Designed using adaptive neural learning techniques, the system is applied to a sample image as input and shows the probability that a particular color exsists for each pixel of the image. The brightness of each pixel reflects this probability in the gray level output image. After selecting the threshold value, a binary image is obtained, which can serve as a mask to segment the color in the input text image.

References

[1] Balamurugan, E., Sangeetha, K., Sengottuvelan, P. (2011). Document Image Analysis - A Review. *International journal of Computer application*, 1(1).

[2] Doermann, D., Tombre, K.(2014). Handbook of Document Image Processing and Recognition. Springer.

[3] Huang, X., Wang, R., Shen, Gao, T.C., (2015). Text Detection and Recognition in Natural Scene Images. *International Conference on Estimation, Detection and Information Fusion (ICEDIF)*.

[4] Chaithanya, C.P., Manohar, N., Issac, A.B.,(2019). Automatic Text Detection and Classification in Natural Images. *International Journal of Recent Technology and Engineering*, 7(5).

[5] Cheung, A., Bennamoun, M.,(2001). An Arabic Optical Character Recognition System using Recognition-based Segmentation. *Pattern Recognition*, Vol.34, pp 215 -233.

[6] R.O. Duda, P.E. Hart, D.G. Stork, "Pattern Classification", 2nd Edition, Wiley publisher, 2000.

[7] Baird, H.S. , (2014). A Brief History of Documents and Writing Systems. *Handbook of Document Image Processing and Recognition*, pp.3-10.

[8] Baird, H.S. , Tombre, K., (2019). The Evolution of Document Image Analysis. *Handbook of Document Image Processing and Recognition*, pp. 63-71.

[9] Kavithaa, A.S., Shivakumara, P., Kumar, G.H., Lu, T., (2016). Text Segmentation in Degraded Historical Document Images. *Egyptian Informatics Journal*, 17(2), pp. 189-197.

[10] Gonzales, R.C., Woods, R.E., (1992). Digital Image Processing. *Prentice Hall Professional Technical Reference*, 2nd Edition.

[11] Bahresraghie, G. A., (1993). Structural Identification of Persian Handwritten Letters. *Msc Dissertation*, Ferdowsi University.

[12] Namazi, M., (1994). Identification of Persian Typefaces with Different Fonts using Neural Network Fuzzy. *MSc Dissertation*, Islamic Azad University.

[13] ShahHosseini, A., (1995). Identification of Persian Handwritten Letters using Neural Network, *MSc Dissertation*, TarbiatModares University.

[14] Sadri, J., (1997). Identification of Letters and Signs of Persian Texts with the Help of Fuzzy Logic. *MSc Dissertation*, Ferdowsi University.

[15] Mozafari, S., Safabakhsh, R., (2004). Identification of Letters and Numbers of Grid-based Persian Handwriting Neural Radial basis Functions and Fractal Properties, *3rd Conference on Vision and Image Processing Machine*.

[16] Kamali, M., Izadyan, J., Qoachani, S.R., Kheyrkhah, A.R., (2007). Recognition of Handwritten Letters by Fuzzy System using Framing Method in Feature Extraction, *1st Joint Congress on Fuzzy Systems and Intelligent Systems*.

[17] Ahmadi, R. , Hosseinzadeh, G., Aarabi, B.N., (2007). Using Geometric Criteria and Texture for Detecting Objects in Different and Complex Images Useful in Automatically Locating the Plaque, *The 1st Joint Congress on Fuzzy Systems and Intelligent Systems*.

[18] Feri, A., (2011). Identification of English Letters using Fuzzy Neural Network. *MSc Dissertation*, Islamic Azad University (In Persian).

[19] Asadi, M., (2011). Offline Detection of Persian Handwritten Cross Section Letters. *MSc Dissertation*, University of Kashan (In Persian).

[20] Razavi, S.M., Kabir, E.A., (2004). Recognizing Persian Fonts. *6th National Conference on Intelligent Systems* (In Persian).

[21] Soleimani, S.M., (2013). Identification of Persian Letters using Neural Networks Synthesis and Scope of Vocabulary. *MSc Dissertation*, Islamic Azad University (In Persian).

[22] Zadeh, L.A., (1965). Fuzzy Sets, *Information and Control*, 8(3), pp.338-353.

[23] Patel, C., Patel, R., Patel, P.,(2011). Handwritten Character Recognition using Neural Network, *International Journal of Scientific & Engineering Research*, 2(5).

[24] Gupta, M., Jin, L., Homma, N., (2003). Static and Dynamic Neural Networks: From Fundamentals to Advanced Theory, *Wiley-IEEE Press*.

[25] Ghosh, A., Shankar, B.U., Meher, S.K., (2009). A Novel Approach to Neuro-Fuzzy Classification, *Neural Networks*, 22(1), pp.100-109.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731