

# Text Classification Using the $\sigma$ -FLNMAP Neural Network

Vassilios Petridis<sup>(1)</sup>, Vassilis G. Kaburlasos<sup>(1)</sup>, Pavlina Fragkou<sup>(1)</sup>, and Athanasios Kehagias<sup>(2)</sup>

Aristotle University of Thessaloniki, GR-54006 Greece  
Faculty of Engineering

<sup>(1)</sup> Department of Electrical and Computer Engineering

<sup>(2)</sup> Dept. of Mathematical, Physical and Computational Sciences

## Abstract

A novel neural network, namely *sigma Fuzzy Lattice Neural network with MAPping* or  $\sigma$ -FLNMAP for short, is presented and applied to classification of text (documents) from the Brown Corpus benchmark collection of documents. The  $\sigma$ -FLNMAP is presented here as an enhanced extension of the fuzzy-ARTMAP neural network in the framework of fuzzy lattices. An individual  $\sigma$ -FLNMAP's classification accuracy is improved by training an ensemble of  $\sigma$ -FLNMAP modules on different permutations of the training data. Several different vector representations of a document are employed. The results, in a series of experiments, compare favorably with the results by other classification algorithms including K-Nearest Neighbor and Naïve Bayes Classifiers.

## 1 Introduction

In line with the ongoing proliferation- and the increasing interconnectivity of computers there emerges the need for automated classification of text. For instance, the problem of text classification appears in such applications as information retrieval (IR) [10], data mining [8], and Web searching [6]. A number of algorithms for text classification has been reported in the literature [10, 11].

For text classification purposes a document is, typically, represented by a high dimensionality vector of the *words* which appear in it [3, 16, 9]. This work employs, in addition, vectors of *senses* or, equivalently, vectors of *meanings of words* in order to test whether *senses* give better text classification results than *words*. Nevertheless, the emphasis of this work is on text /document classification using a novel neural network, namely *sigma Fuzzy Lattice Neural network with MAPping* or  $\sigma$ -FLNMAP for short. Comparative results by other classification algorithms such as K- Nearest Neighbor (KNN) and Naïve Bayes Classifiers (NBC) are also

presented as well as the effect of alternative vector representations of a document.

Section 2 reviews the  $\sigma$ -FLN neural network for clustering in the framework of fuzzy lattices. Section 3 describes both the  $\sigma$ -FLNMAP neural network for classification and the “ $\sigma$ -FLNMAP with Voting” neural model. Section 4 describes the benchmark collection of documents used in this work and it discusses alternative vector representations of a document. Section 5 presents comparative experimental results and, finally, section 6 provides the concluding remarks.

## 2 The $\sigma$ -FLN Neural Network for Clustering

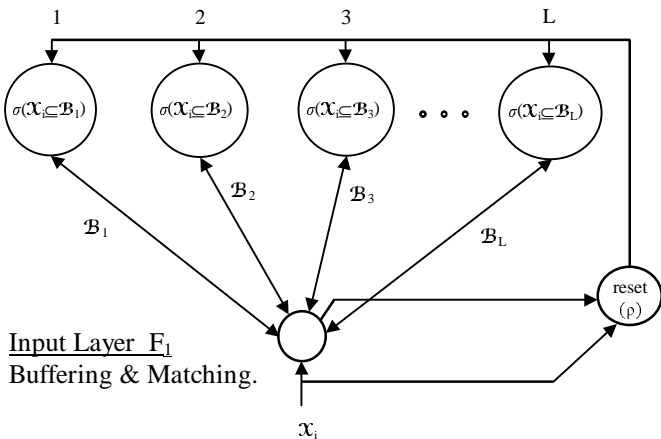
Apart from the  $N$ -dimensional Euclidean space “learning”, by a neural network, can also be effected in other domains as it has been demonstrated by the  $\sigma$ - Fuzzy Lattice Neural network or for short  $\sigma$ -FLN, [7]. The  $\sigma$ -FLN is applicable in the *framework of fuzzy lattices*, where a *fuzzy lattice* is a conventional (mathematical) lattice  $\mathbb{L}$  such that the ordering relation has been extended to all elements in  $\mathbb{L}$  in a fuzzy degree of truth sense. In particular, it has been shown in [7] that the inclusion relation in a conventional lattice  $\mathbb{L}$  can be extended to all elements of the Cartesian product  $\mathbb{L} \times \mathbb{L}$  using an axiomatically defined *inclusion measure* function  $\sigma: \mathbb{L} \times \mathbb{L} \rightarrow [0,1]$ . “Learning” is achieved in  $\sigma$ -FLN by computing intervals of lattice elements, and an interval is regarded as a cluster. Note that in the  $N$ -dimensional Euclidean space a lattice interval is a *hyperbox*, or *box* for short. The  $\sigma$ -FLN neural network architecture is shown in Fig.1.

“Learning” in the  $\sigma$ -FLN neural network is effected by a clustering algorithm similar to fuzzy-ART's algorithm as explained in [7]. In particular, it is shown in [7] that the  $\sigma$ -FLN is an enhanced extension of fuzzy-ART in the framework of fuzzy lattices such that: 1) the  $\sigma$ -FLN can deal with inputs both trivial and non-trivial intervals,

whereas fuzzy-ART deals solely with trivial intervals (points) inputs, 2) the learning behavior of  $\sigma$ -FLN can be *fine tuned* by selecting properly the underlying positive valuation function  $v(x)$ , whereas fuzzy-ART employs implicitly always the same positive valuation function  $v(x)=x$ , and 3) the  $\sigma$ -FLN is applicable in a (mathematical) lattice domain including fuzzy-ART's domain, that is the unit  $N$ -dimensional unit hypercube. Note that the degree of inclusion of a vector (in an upper layer neuron) in an input datum, as well as the degree of inclusion of an input datum in a vector of an upper layer neuron, are calculated using an inclusion measure function  $\sigma$  which is defined based on a real function  $v(x)$ , namely *positive valuation function*, as detailed in [7].

Category Layer  $F_2$

Competition : winner takes all.



**Fig. 1:** The two layer  $\sigma$ -FLN architecture. A *Category Layer* neuron employs a lattice inclusion measure  $\sigma$  as its activation function in order to specify the fuzzy degree of inclusion of input  $\mathcal{X}_i$  to weight box  $\mathcal{B}_k, k=1, \dots, L$ . The *Input Layer* buffers an input. A “reset” node is characterized by the system’s vigilance parameter ( $\rho$ ) and it is used for resetting the activity of a node in the *Category Layer*.

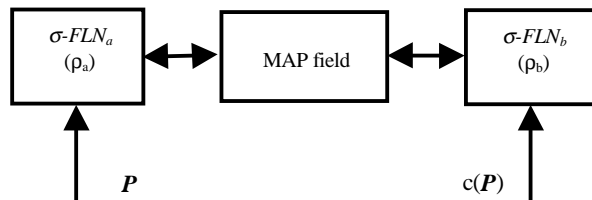
**3 Neural Models for Classification**

**3.1 The  $\sigma$ -FLNMAP Neural Network for Classification**

The  $\sigma$ -FLNMAP neural network for classification emerges by the synergetic combination of two  $\sigma$ -FLN modules, namely module  $\sigma$ -FLN<sub>a</sub> and module  $\sigma$ -FLN<sub>b</sub>, interconnected via a MAP field. Module  $\sigma$ -FLN<sub>a</sub> is trained using the data for training, whereas module  $\sigma$ -FLN<sub>b</sub> is trained using the corresponding category labels of the

training data. The MAP field maps a box/cluster from module  $\sigma$ -FLN<sub>a</sub> to a category/box in module  $\sigma$ -FLN<sub>b</sub> (Fig.2). The interconnection of the modules shown in Fig.2 is similar to the interconnection between the corresponding modules in fuzzy-ARTMAP neural network [2]. Recall that at the end of the previous section there have been enumerated the reasons for regarding the  $\sigma$ -FLN as an *enhanced* extension of fuzzy-ART; likewise it can be argued that the  $\sigma$ -FLNMAP neural network can be regarded as an *enhanced* extension of fuzzy-ARTMAP in the framework of fuzzy lattices.

The  $\sigma$ -FLNMAP is set to compute “uniform” boxes in the training data, that is all the data which give rise to a particular box/cluster belong to the same category. Moreover note that as the training data enter the system, the  $\sigma$ -FLNMAP computes the largest “uniform” boxes in the training data. It is known that the boxes /clusters calculated by  $\sigma$ -FLNMAP, and ultimately the classification decisions made by  $\sigma$ -FLNMAP, depend on the order of data presentation [7]. It has been confirmed experimentally that the classification performance of  $\sigma$ -FLNMAP can be stabilized and improved using an *ensemble* of  $\sigma$ -FLNMAP modules as explained in the following.



**Fig. 2:** The  $\sigma$ -FLNMAP neural network is a synergy of two  $\sigma$ -FLN modules, namely  $\sigma$ -FLN<sub>a</sub> and  $\sigma$ -FLN<sub>b</sub>, which are interconnected via the MAP field.  $P$  denotes a training datum whereas  $c(P)$  denotes its corresponding category.

**3.2 “ $\sigma$ -FLNMAP with Voting”: An Ensemble of  $\sigma$ -FLNMAP’s for Classification**

The inspiration for the “ $\sigma$ -FLNMAP with Voting” neural model derives from *statistical learning theories* [1, 15]. The idea is to train an ensemble of  $\sigma$ -FLNMAP’s using different permutations of the training data. In conclusion, a testing datum is classified to the category which receives the majority vote from the individual  $\sigma$ -FLNMAP voters. An “ $\sigma$ -FLNMAP with Voting” neural model is characterized by two parameters: 1) the vigilance parameter  $\rho_a$  of module  $\sigma$ -FLN<sub>a</sub>, and 2) the number  $n_V$  of  $\sigma$ -FLNMAP voters in the ensemble.

## 4 The Brown Corpus and Document Representations

In the experiments, the Brown Corpus benchmark collection of documents has been used which is distributed along with the *Wordnet* lexical database as described in the following.

### 4.1 The Brown Corpus Semantic Concordance

Wordnet is an on-line lexical database which was developed under the direction of G.A. Miller [12]. Wordnet is similar to an electronic thesaurus and is organized around the distinction between words and senses. It contains a large number of nouns, verbs, adjectives and adverbs of the English language, reaching a total of nearly 130,000 words as well as a total of nearly 100,000 senses. The Brown Corpus collection of documents is distributed along with Wordnet and it includes 500 documents which are classified into fifteen categories: 1) Press: Reportage, 2) Press: Editorial, 3) Press: Reviews, 4) Religion, 5) Skills and Hobbies, 6) Popular Lore, 7) Belles Lettres, Biography, Memoirs, 8) Miscellaneous, 9) Learned, 10) General Fiction, 11) Mystery and Detective Fiction, 12) Science Fiction, 13) Adventure and Western Fiction, 14) Romance and Love Story, and 15) Humor; for an extended description see in [5]. The Brown Corpus collection is a *semantic concordance*, that is a combination of documents and a thesaurus; the documents are combined in manner such that every substantive word in each document is linked to its appropriate sense in the thesaurus. The Brown Corpus semantic concordance makes use of 352 out of the 500 Brown Corpus documents. Linguists involved in the Wordnet project manually performed *semantic tagging*, i.e. annotation of the 352 texts with WordNet senses. In conclusion a number of alternative representations of a document have been possible as described below.

### 4.2 Document Representations

Each document from the Brown Corpus semantic concordance is represented by a vector of either words or senses. The *vocabulary* in a classification problem, involving a set of documents, is defined to be the set of  $N_w$  words  $w_1, \dots, w_{N_w}$  (or, the set of  $N_s$  sense  $s_1, \dots, s_{N_s}$ ) which appear in at least one document. A few different document representations are presented in the following for “words”. The same representations are used for “senses” as well.

A document  $d$  may be represented by a *frequency vector*  $\mathbf{d}=[d_1, \dots, d_n, \dots, d_{N_w}]$  where  $d_n$  is the number of times the  $n$ -th word  $w_n$  appears in document  $d$ . Moreover, a document  $d$  may be represented by a *Boolean vector*  $\mathbf{d}=[d_1, \dots, d_n, \dots, d_{N_w}]$  where  $d_n$  is either 1 or 0 when, respectively, the  $n$ -th word  $w_n$  appears or does-not appear

in document  $d$ . A *relative frequency* representation  $\mathbf{d}=[d_1, \dots, d_n, \dots, d_{N_w}]$  defines vector component  $d_n$  as

$$d_n = \frac{\text{no. of times the } n\text{-th word } w_n \text{ appears in document } d}{\text{total no. of words in document } d}$$

Finally, the *normalized frequency* representation  $\mathbf{d}=[d_1, \dots, d_n, \dots, d_{N_w}]$  defines vector component  $d_n$  as

$$d_n = \frac{\text{no. of times word } w_n \text{ appears in document } d}{\text{maximum no. of times word } w_n \text{ appears in any document}}$$

## 5 Experimental Results

Experimental results of text classification using the “ $\sigma$ -FLNMAP with Voting” neural model are presented comparatively with results obtained by other classification algorithms including the K- Nearest Neighbor (KNN) and the Naïve Bayes Classifier (NBC). The operation of KNN and NBC is described briefly in the following. On the one hand for the KNN, a “testing datum” is classified to the category which receives the majority vote among K “voter” in the training data which are the nearest to the testing datum in question in an L1-distance sense. For further details the reader may refer to [4]. The KNN has been employed here with both a *relative frequency* and a *Boolean* representation. On the other hand, the NBC assigns a “testing datum” to the category which maximizes the conditional probability of occurrence of the document in question *given a category*. A detailed description of the NBC algorithm appears in [13]. The NBC has been employed here with the *frequency* representation of a document.

Several classification problems have been dealt with as explained in the following. In particular either *all the 15 categories* of the Brown Corpus with 352 documents, or *only 3 categories* (i.e. categories 1, 2 and 10) of the Brown Corpus with 100 documents have been considered. For each one of the latter two problems either 1) the “nouns”, verbs”, “adjectives” and “adverbs”, or 2) only the “nouns” and verbs” have been considered in a document. For all previous  $2 \times 2 = 4$  combinations of problems, vectors of either *words* or *senses* have been produced from the documents and the corresponding vector lengths are shown in Table 1. Moreover, for all combinations of 1) parts of speech, and 2) no. of categories, a large number of experiments was carried out involving either words or senses such that three different classification algorithms have been employed. Each time, an algorithm has been applied, for several values of its parameter vector, on ten different random partitions of  $f$  data set such that  $2/3$  of the data were used for training and the remaining  $1/3$  for testing. The overall results are summarized in Table 2.

**Table 1** The lengths of feature vectors used in the experiments for various combinations of *parts of speech* and *no. of categories*. The code-words “n”, “v”, “adj” and “adv” stand, respectively, for “noun”, “verb”, “adjective” and “adverb”.

Parts of speech	no. of categories	Length of feature vectors	
		words	senses
n, v, adj, adv	3	10,890	9,348
	15	25,683	22,101
n, v	3	8,448	7,068
	15	18,806	15,728

**Table 2** The average % classification accuracy in 10 experiments for various combinations of 1) *parts of speech* and *no. of categories* as specified in the corresponding row, and 2) a *classification algorithm* (and a *document representation*) as specified in the corresponding column. The code words “n”, “v”, “adj” and “adv” stand, respectively, for “noun”, “verb”, “adjective” and “adverb”.

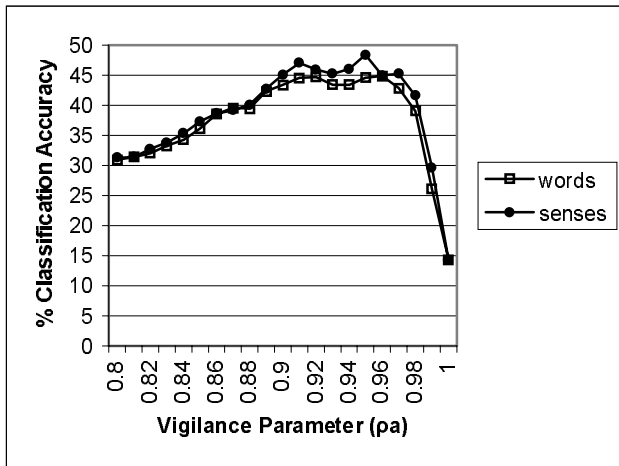
Parts of speech	no. of categories	KNN (Boolean)		KNN (Rel. Freq.)		Naïve Bayes (Frequency)		$\sigma$ -FLNMAP with Voting (Boolean)		$\sigma$ -FLNMAP with Voting (Norml. Freq.)	
		words	senses	words	senses	words	senses	words	senses	words	senses
n, v, adj, adv	3	80.00	76.00	76.00	80.00	53.12	56.87	79.68	81.25	80.00	82.18
	15	45.00	46.00	42.00	46.00	39.60	38.21	43.83	46.60	44.91	48.39
n, v	3	80.31	75.63	75.94	80.31	55.31	61.24	80.62	81.25	80.62	82.50
	15	45.71	47.50	41.16	46.16	41.00	38.57	41.60	47.23	44.01	47.85

By looking at Table 2 it follows that “senses” performed better than “words” in 36 out of 40 senses. Only in 3 cases the performance of senses was 5 percentage points or higher than the performance of words. Therefore it is concluded that, in the context of this work, the use of senses only marginally improves the classification accuracy. Among the three classification algorithms the “ $\sigma$ -FLNMAP with Voting” performed best whereas NBC performed worst. The KNN has given fairly good classification results especially for the *Boolean* representation. For the “ $\sigma$ -FLNMAP with Voting” the senses have always given better results than words; moreover the *normalized frequency* representation has implied better results than the *Boolean* representation, as expected, since the former representation includes all information in the latter representation.

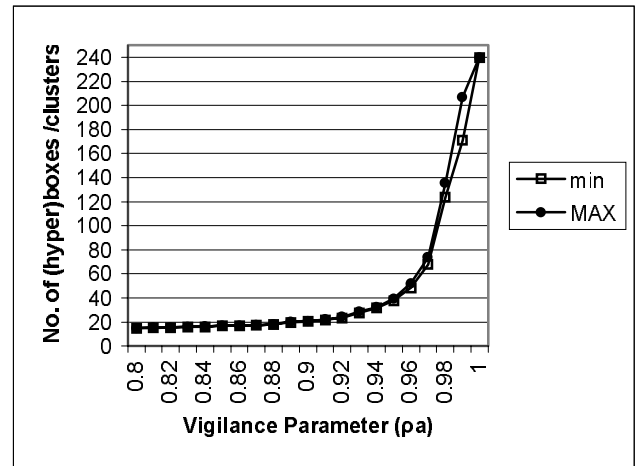
The “ $\sigma$ -FLNMAP with Voting” neural model outperforms all other text classification algorithms used in this work. Its good performance is attributed to both the effectiveness of inclusion measure  $\sigma$  and the model’s capacity for generalization which is based on the calculation of the largest “uniform” boxes in the training data. The performance of “ $\sigma$ -FLNMAP with Voting” remains quite stable for a fairly wide range of values of the vigilance parameter ( $\rho_a$ ) and it drops sharply as  $\rho_a$  approaches 1 as

shown in Fig.3 and Fig.5 for the 15- and the 3- categories problem, respectively.

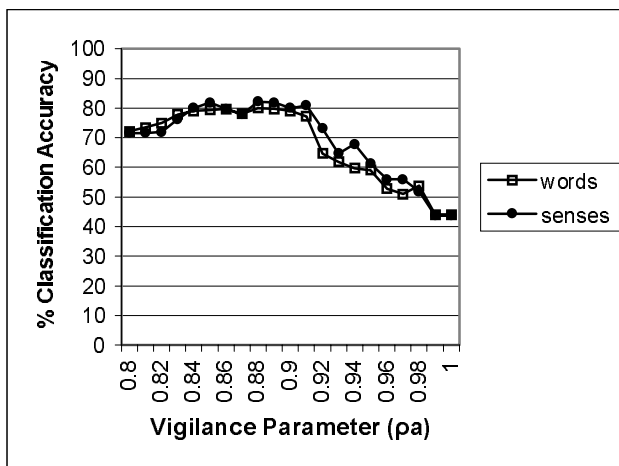
Fig.4 shows that the number of boxes/clusters computed by “ $\sigma$ -FLNMAP with Voting” increases exponentially as  $\rho_a$  approaches 1; in particular for  $\rho_a=1$  different training data give rise to different (trivial) boxes/clusters. Fig.6 illustrates the stability effects in classification accuracy of using an ensemble of  $\sigma$ -FLNMAP voters which are trained on different random permutations of the training data; note that an individual  $\sigma$ -FLNMAP’s classification accuracy in the ensemble may fluctuate in a wide range as shown in Fig.6. Fig.6 also demonstrates that for selected values of the vigilance parameter ( $\rho_a$ ) an ensemble of  $\sigma$ -FLNMAP voters can perform better than the individual  $\sigma$ -FLNMAP voters in the ensemble. The latter improvement is attributed to the noise-cancellation effects of the random permutations of the training data used to train the  $\sigma$ -FLNMAP’s in the ensemble. Note that Fig.3 through Fig.6 refer to experiments where all parts of speech have been employed. Finally note that both the the vigilance parameter ( $\rho_a$ ) and the number ( $n_v$ ) of voters can be estimated from the training data using several random partitions (of the training data) into a subset for training and another subset for validation.



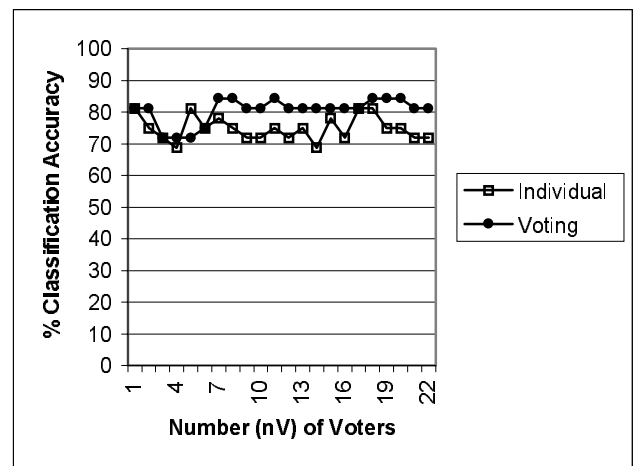
**Figure 3** Average classification accuracy of *words* and *senses* versus the vigilance parameter ( $\rho_a$ ) for 10 random training/testing data sets and 15 categories of the Brown Corpus documents. Senses have resulted in a marginally better classification accuracy.



**Figure 4** Average *minimum* and *maximum* number of (hyper)boxes /clusters computed by “ $\sigma$ -FLNMAP with Voting” versus the vigilance parameter ( $\rho_a$ ) for 10 random training/testing data sets and 15 categories of the Brown Corpus documents. The number of boxes increases exponentially as  $\rho_a$  approaches 1; for  $\rho_a=1$  the number of (hyper)boxes computed is equal to the number of training data.



**Figure 5** Average classification accuracy of *words* and *senses* versus the vigilance parameter ( $\rho_a$ ) for 10 random training/testing data sets and 3 categories of the Brown Corpus documents. Senses have resulted in a marginally better classification accuracy.



**Figure 6** Classification accuracy of individual  $\sigma$ -FLNMAP's and “ $\sigma$ -FLNMAP with Voting” with an increasing number  $n_V$  of voters versus the number  $n_V$  of voters for  $\rho_a=0.89$ . The “ $\sigma$ -FLNMAP with Voting” has resulted in both stability and improvement in classification performance.

## 6 Discussion and Conclusion

The “ $\sigma$ -FLNMAP with Voting” neural model has been introduced and applied in this work for classification of documents from the Brown Corpus benchmark collection of documents. In a series of classification experiments involving several alternative document representations the “ $\sigma$ -FLNMAP with Voting” outperformed conventional classification algorithms including K- Nearest Neighbor (KNN) and Naïve Bayes Classifiers (NBC). The same experiments have also shown that the use of “senses”, as features for representing a document, only marginally improves the classification accuracy over the use of “words”.

The  $\sigma$ -FLNMAP neural network could be regarded as an enhanced KNN classifier, as it will be elaborated elsewhere, which computes “uniform boxes” in the training data set. Perhaps this is the reason why  $\sigma$ -FLNMAP exhibits better generalization than KNN. Moreover, due to its applicability in (mathematical) lattices the  $\sigma$ -FLNMAP has the potential to represent a document using features other than vectors as well, e.g. graphs as detailed in [14].

## References

- [1] L. Breiman, “Bagging predictors”, Technical Report 421, Dept. of Statistics, Univ. of California at Berkeley, 1994.
- [2] G.A. Carpenter, S. Grossberg, N. Markuzon, J.H. Reynolds, and D.B. Rosen, “Fuzzy ARTMAP: A Neural Network Architecture for Incremental Supervised Learning of Analog Multidimensional Maps”, *IEEE Transactions on Neural Networks*, 3(5), 698-713, 1992.
- [3] H. Drucker, D. Wu, and V.N. Vapnik, “Support Vector Machines for Spam Categorization”, *IEEE Trans. on Neural Networks*, vol. 10, no. 5, pp. 1048-1054, 1999.
- [4] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern classification*, Wiley, 2001.
- [5] W.N. Francis, and H. Kucera, *Frequency Analysis of English Usage: Lexicon and Grammar*, Houghton Mifflin Company: Boston, 1982.
- [6] A. Fujii, and Ishikawa, “Utilizing the World Wide Web as an Encyclopedia: Extracting Term Descriptions from Semi-structured Texts”, *Proc. 38<sup>th</sup> Annual Meeting Association for Computational Linguistics (ACL-2000)*, pp. 488-495, 2000.
- [7] V.G. Kaburlasos, and V. Petridis, “Fuzzy Lattice Neurocomputing (FLN) Models”, *Neural Networks*, vol. 13, no. 10, pp. 1145-1170, 2000.
- [8] R. Kosala, and H. Blockeel, “Web Mining Research: A Survey”, *ACM SIGKDD Explorations*, vol. 2, pp. 1-15, 2000.
- [9] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-Based Learning Applied to Document Recognition”, *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324, 1998.
- [10] D.D. Lewis, and K.S. Jones, “Natural Language Processing for Information Retrieval”, *Communications of the ACM*, vol. 39, pp. 92-101, 1996.
- [11] C.D. Manning, and H. Schuetze, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999.
- [12] G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K.J. Miller, “Introduction to WordNet: An on-line Lexical Database”, *International Journal of Lexicography*, vol. 3, pp. 235-244, 1990.
- [13] T.M. Mitchell, *Machine Learning*, The McGraw-Hill Companies, Inc., 1997.
- [14] V. Petridis, and V.G. Kaburlasos, “Clustering and Classification in Structured Data Domains Using Fuzzy Lattice Neurocomputing”, *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 2, 2001.
- [15] R.E. Schapire, “The Strength of Weak Learnability”, *Machine Learning*, vol. 5, no. 2, pp. 197-227, 1990.
- [16] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer: New York, 1995.