Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 3 (2011) 1412–1417

Procedia Computer Science

www.elsevier.com/locate/procedia

WCIT 2010

Classification of confidential documents by using adaptive neuro-fuzzy inference systems

Erdem Alparslan [a,b] , Adem Karahoca [a] , Hayretdin Bahşi [b]

[a]Bahcesehir University, Ciragan 34353 Besiktas, Istanbul, Turkey
[b]TUBITAK UEKAE, 41470 Gebze,Kocaeli, Turkey

Abstract

Detecting the security level of a confidential document is a vital task for organizations to protect the confidential information encapsulated in. Diverse classification rules and techniques are being applied by human experts. Increasing number of confidential information in organizations are making difficult to classify all the documents carefully with human effort. A hybrid approach involving support vector classifier and adaptive neuro-fuzzy classifier is proposed in this study. Also states preprocessing tasks required for document classification with natural language processing. To represent term-document relations a recommended metric TF-IDF was chosen to construct a weight matrix. Agglutinative nature of Turkish documents is handled by Turkish stemming algorithms. At the end of the article some experimental results and success metrics are projected with accuracy rates.

Selection and/or peer-review under responsibility of the Guest Editor.

Keywords: Document Classification, Expert Systems, ANFIS, SVM, Turkish NLP

1. Introduction

In recent years, protecting secure information became a challenge for military and governmental organizations. As a result, well defined security level contents and rules are more preferable than in the past [11]. Each piece of information has its own security level. Correct detection of this security level may lead to apply correct protection rules on information [9].

Determining the security level of a document by using expert systems is a document classification problem. The aim of classification of confidential documents is to assign predefined class labels to a new document that is not classified 1. An associated classification framework provides training documents with existing class labels. Therefore supervised, semi-supervised or unsupervised classification algorithms are fitting as a solution to the classification problem.

Classification accuracy of textual data is highly related to preprocessing tasks of training and test data 2. These tasks become more difficult in processing unstructured textual data than in structured data. Unstructured nature of

* Erdem Alparslan Tel.: +90-262-648-1576; fax: +90-262-648-1100
E-mail address: ealparslan@uekae.tubitak.gov.tr

doi:10.1016/j.procs.2011.01.023

data needs to be formatted in a relational and analytical form. TF-IDF (term frequency-inverse document frequency) is preferred to represent text based contents of documents.

Another important task of formatting textual data is stemming. Stemming the Turkish documents is more difficult than the other studies based on English or in other Latin-based languages. Turkic languages which are agglomerative languages, involve diverse exceptional derivation rules. Therefore stemming of Turkish terms provides some unstable rules varying from structure to structure [10].

In this study, internal documents of TUBITAK UEKAE (National Research Institute of Electronics and Cryptology) are classified into three classification levels: “secret, restricted and unclassified” by a hybrid approach which uses SVM outputs as ANFIS inputs. With this hybrid approach classification accuracy of 97% has became reached as well for Turkish confidential document set.

2. Adaptive Neuro-Fuzzy Inference System Algorithm

The "fuzzy logic" was firstly proposed by Zadeh 4 to describe complicated systems. It has became very popular and been used successfully in various problems especially on control processes such as chemical reactors, electronic motors, automatic trains and nuclear reactors. More recently, fuzzy logic has been highly recommended for modeling data mining and knowledge engineering problems. A neural network has the ability to learn from the environment (input–output pairs), self-organize its structure, and adapt to it in an interactive manner. Because of this ability of neural networks, we prefer using adaptive neuro-fuzzy inference system for predicting the security level class labels of text documents. Also the fuzzy nature of our classification problem is making it more convenient to use fuzzy inference system. Because the discrimination between class labels in a security level classification problem may be sometimes indiscernible. Fuzzy membership of documents to more than one class is fitting better in this study.

Fuzzy inference system can be considered by two inputs, x and y, and one fuzzy output z. A first-order Sugeno model proposes these two rules below as an example:

Rule 1: If x is and y is then = * x + * y +

Rule 2: If x is and y is then = * x + * y +

Where and are the fuzzy sets, , and are the design parameters which are defined by training phase 5.

ANFIS architecture covers 5 layers which are presented in figure 1.

Brief introduction of the model is as follows 67:

Layer 1: The adaptive nodes of this layer generate the membership grades for their appropriate fuzzy sets. The output functions of this layer are presented as:

$O_{1,i} = \mu_{A_i}$ for i = 1 , 2 (1)

$O_{1,i} = \mu_{B_{i-2}}$ for i = 3 , 4 (2)

, (small, big, etc.) are the linguistic labels characterized by the membership functions , respectively and *x*, *y* are the crisp inputs to node *i*.

Layer 2: In this layer the outputs are the result of a simple multiplication process:

$O_{2,i} = w_i = \mu_{A_i}(x)\,\mu_{B_i}$ for i = 1 , 2 (3)

Layer 3: This layer calculates the ratio of each rule’s firing strength to overall firing strength.

$O_{3,i} = \overline{w_i} = \frac{w}{w_1 +}$ for i = 1 , 2 (4)

Layer 4: The output function of this layer is calculated as:

$O_{4,i} = \overline{w_i}\, f_i = \overline{w_i}\,(p_i x + q_i y +$ For i = 1, 2 (5)

p_i, q_i are the consequent parameters. Regulating these parameters yields adaptive learning.

Layer 5: This layer contains only one node which computes the summation of all output nodes.

$$O_{5,1} = overall\ output = \sum_i \overline{w_i}\, f_i = \frac{\sum_i w}{\sum_i} \tag{6}$$

Fig. 1 ANFIS Structure

3. Experimental Settings

This study aims to develop a framework which has an ability to classify 222 internal documents of TUBITAK UEKAE (National Research Institute of Electronics and Cryptology) according to their security levels by using a hybrid approach containing adaptive neuro-fuzzy inference systems and support vector machines.

First, all of 222 internal documents are classified into correct security levels (secret, restricted, unclassified) according to the general policies of TUBITAK UEKAE with the help of a subject matter expert. (The numbers of secret, restricted and unclassified documents are 30, 165 and 27 respectively.) Then these classified documents are converted into UTF-8 encoded txt based file format. Training and test documents have totally about 2.5 millions of words except stopping words. All the documents are grouped and arranged in a relational database structure by using the open source database PostgreSQL. These 2.5 million words are unstemmed Turkish words. A comprehensive stemmer library Zemberek is used to find out roots of unstemmed words. Our stemming system selects the structure that has the biggest probability of semantic and morphologic patterns of the Turkish language.

Performing the stemming process we obtained approximately 9000 distinct terms. These 9000 terms may cause a high dimensionality and a time consuming classification process. Therefore a chi-square statistics with a threshold (100) was performed to select important features of classification. By the feature selection, the size of selected features is reduced from ~9000 to ~2000. This is known as the corpus of classification process.

The final task performed for the preprocessing phase was constructing a TF-IDF value matrix of all the features for all documents. The application was calculating TF-IDF values for each feature-document pair in the corpus.

4. Results

Adaptive Neuro-Fuzzy Inference Systems (ANFIS) can be widely used in classification of structured data. The problem of security level classification which is subject to our study, implies fuzzy class labels. By the nature, security level classification does not yield distinct and discrete security labels. For example a document labeled as “restricted” can be quite (30%) “unclassified” and more obviously (70%) “restricted”. So we label this document as “restricted”.

Thus, using ANFIS to classify security levels of documents is more suitable because of the continuous outputs of ANFIS. Having these continuous outputs a supervised discretization algorithm (CACC) 8 can be used to detect discrete class labels of documents.

In the other hand the nature of ANFIS algorithm is not suitable to handle thousands of input parameters. So a taxonomy based pre-classification has to be performed to provide input to the ANFIS algorithm.

Regarding Table 1 a sub-level classification according to the document types (tech test report, tech guide, meeting report, quality procedure, ITSM audit etc...) is performed by using Support Vector Machines. There exist 9 different document types. SVM-multiclass scores all train and test documents according to their document types. Example SVM-multiclass scoring outputs for the first 20 documents are represented in Table 1 below:

Table 1. SVM results for document area classification

tech test report	**tech guide**	**spec doc**	**quality procedure**	**meeting report**	**itsm audit**	**travel report**	**traning**	**test procedure**
9,362026	18,85194	-0,99012	-5,47238	-8,29389	-17,0853	-11,7173	48,87856	-17,614
24,28767	-17,9092	-33,8235	-13,8527	23,23329	-24,2238	38,07242	63,16465	-24,3409
-33,8924	-3,2924	13,78726	-17,0198	-14,2971	-19,9289	-8,8031	133,9987	-5,26039
8,438817	10,85226	-21,8353	-11,4376	2,460903	-30,7703	-5,59678	48,23171	20,76673
-7,92581	12,82986	-5,2941	-6,54043	-9,65095	-9,89681	-5,28716	48,52536	6,264983

These document type scores are the inputs of Adaptive Neuro-Fuzzy Inference System (ANFIS). ANFIS can handle less than 10 or 15 inputs because of the time complexity of the algorithm. The input variables of ANFIS are the document type scores represented in Table 10 and the outputs are continuous variables indicating security levels.

In this fuzzy inference structure 2 membership functions with trimf input membership type and constant output membership type provide the most appropriate, fitting FIS structure. The resulting output of ANFIS is shown in figure 2 below:

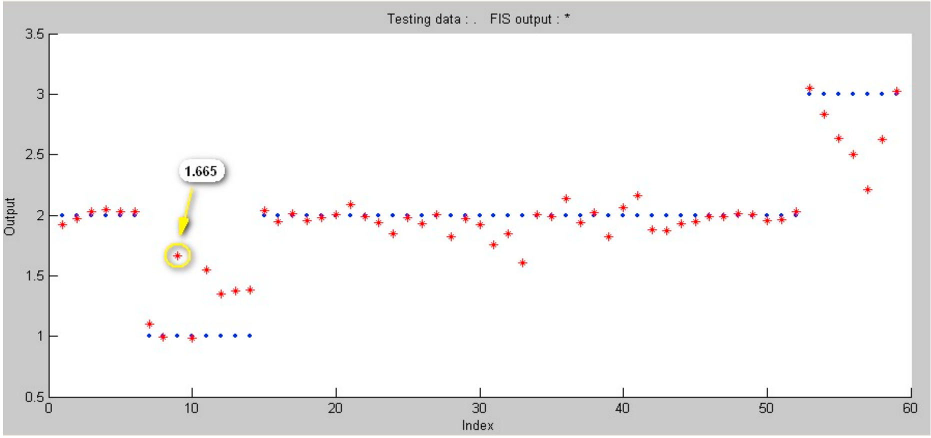

Fig. 2 Continuous FIS outputs of security level classification

In this figure each dot represents a test document. Blue ones and red ones represent the real security label of documents and predicted security level of documents respectively. It is noticeable that many of the predicted values are scored between two class labels. For example the 9th sample introduces a score of **1.665** which is between 1 (secret) and 2 (restricted).

To find out the corresponding security label of this level score we have to discretize the continuous outputs to the discrete labels. After using the CACC algorithm introduced in section 4 the discrete security labels are detected. CACC algorithm suggests these thresholds for discretization:

If the ANFIS score of document is smaller than **1.68** the security label must be **1 (secret)**.

If the ANFIS score of document is between **1.68** and **2.36** the security label must be **2 (restricted)**.

If the ANFIS score of document is bigger than **2.36** the security label must be **3 (unclassified)**.

The resulting scheme is summarized in table 2:

Table 2. Accuracy results of hybrid approach

	Secret	**Restricted**	**Unclassified**	**ACCR**
Secret	8	0	0	100%
Restricted	1	43	0	97.6%
Unclassified	0	1	6	85.7&
Overall Accuracy Rate :	**96.67% (57/59)**			

5. Discussion and Conclusion

This study classifies internal Turkish documents of TUBITAK UEKAE (a military-governmental organization) according to their security levels by using a hybrid approach containing support vector classifiers and adaptive neuro-fuzzy inference systems.

Documents are classified according to their types (tech test report, meeting report etc…) by using SVM. The outputs of SVM are also the inputs of ANFIS which is giving the security level rates of documents as output. These rates are rendered discretize by using a discretization algorithm, CACC. The accuracy rate of this constructed framework is very satisfied with 96%.

As we mentioned before, security level classification can constitute a basis for the extended detection capabilities of data loss / leakage prevention solutions. For each security problem, according to the nature of the business processes of documents for each organization, support vector phase of our hybrid approach may be reorganized. In TUBITAK UEKAE document types are the essential indicators to detect the security level. Therefore this study uses document types as support vector outputs and ANFIS inputs.

References

1. T. Joachims, Text Categorization with Support Vector Machines: Learning with Many Relevant Features. European Conference on Machine Learning, 1998.
2. J.W. Han and M. Kamber, Data Mining Concept and Techniques. Second Edition, 2007.
3. R. Cooley, Classification of News Stories Using Support Vector Machines. IJCAI Workshop on Text Mining, 1999.
4. L. Zadeh and J. Kacprztk, Computing with words in information intelligent systems. New York, Springer, 1999.
5. C. Cortes and V. Vapnik, Support-vector Networks. Machine Learning, 20:273-297, November, 1995.
6. J. Shing and R. Janj, ANFIS: Adaptive Network-Based Fuzzy Inference System. 1993.
7. I. Guler and E. Ubeyli, Adaptive neuro-fuzzy inference system for classification of EEG signals using wavelet coefficients. Journal of Neuroscience Methods vol:148 pp: 113–121, 2005.
8. C.J. Tsai, C.I. Lee, A discretization algorithm based on Class-Attribute Contingency Coefficient. Information Sciences vol:178 pp: 714–731, 2008.
9. Sakalli, M. (2009). The Frequent Use of Teaching Strategies/Methods Among Teachers According to the Teacher Candidates Observation. Cypriot Journal Of Educational Sciences, 2(1). Retrieved November 15, 2010, from http://www.world-education-center.org/index.php/cjes/article/view/13.
10. Meerah, T., Halim, L., Rahman, S., Abdullah, R., Harun, H., Hassan, A., & Ismail, A. (2010). Teaching marginalized children primary science teachers professional development through collaborative action research. Cypriot Journal Of Educational Sciences, 5(1). Retrieved November 15, 2010, from http://www.world-education-center.org/index.php/cjes/article/view/146.
11. Peck, B., Deans, C., & Stockhausen, L. (2010). The Tin-Man and the TAM: A Journey Into M-Learning in the Land of Aus.. World Journal On Educational Technology, 2(1). Retrieved November 15, 2010, from http://www.world-education-center.org/index.php/wjet/article/view/62