Available online at www.sciencedirect.com

ScienceDirect

Expert Systems with Applications 36 (2009) 1587–1592

Expert Systems with Applications

www.elsevier.com/locate/eswa

A novel hybrid intelligent method based on C4.5 decision tree classifier and one-against-all approach for multi-class classification problems

Kemal Polat *, Salih Güneş

Selcuk University, Electrical and Electronics Engineering, 42035 Konya, Turkey

Abstract

Generally, many classifier systems compel in the classification of multi-class problems. The aim of this study is to improve the classification accuracy in the case of multi-class classification problems. In this study, we have proposed a novel hybrid classification system based on C4.5 decision tree classifier and one-against-all approach to classify the multi-class problems including dermatology, image segmentation, and lymphography datasets taken from UCI (University of California Irvine) machine learning database. To test the proposed method, we have used the classification accuracy, sensitivity-specificity analysis, and 10-fold cross validation. In this work, firstly C4.5 decision tree has been run for all the classes of dataset used and achieved 84.48%, 88.79%, and 80.11% classification accuracies for dermatology, image segmentation, and lymphography datasets using 10-fold cross validation, respectively. The proposed method based on C4.5 decision tree classifier and one-against-all approach obtained 96.71%, 95.18%, and 87.95% for above datasets, respectively. These results show that the proposed method has produced very promising results in the classification of multi-class problems. This method can be used in many pattern recognition applications. In future, instead of C4.5 decision tree, other classification algorithms such as Bayesian learning, artificial immune system algorithms, artificial neural networks can be used.

Keywords: Hybrid systems; C4.5 Decision tree classifier; One-against-all approach; Multi-class dataset classification

1. Introduction

In this paper, C4.5 decision tree classifier and one-against-all method were combined to improve the classification accuracy for multi-class classification problems including dermatology, image segmentation, and lymphography datasets.

Classifying of multi-class classification problems such as document categorization, image recognition, signal classification etc. that have multi-class is important issue in pattern recognition applications. In this work, in order to solve this problem, the used datasets were classified as two groups using one-against-all method.

There have been several studies reported focusing on erythemato-squamous disease diagnosis using dermatology dataset. These studies applied different methods to the given problem and achieved high classification accuracies using the dataset taken from UCI machine learning repository. Among these studies, the first work on the differential diagnosis of erythemato-squamous diseases was conducted by Guvenir et al. In their study, they presented an expert system for differential diagnosis of erythemato-squamous diseases incorporating decisions made by three classification algorithms: nearest neighbor classifier, naive Bayesian classifier and voting feature intervals-5. Also, they obtained 99.2% classification accuracy on the differential diagnosis of erythemato-squamous diseases using voting feature intervals-5 and 10-fold cross validation (Demiroz, Govenir, & Ilter, 1998; Govenir & Emeksiz, 2000). Ubeyli and Guler (2005) obtained 95.5% classification accuracy by using

* Corresponding author. Tel.: +90 332 223 2056; fax: +90 332 241 0635.
E-mail addresses: kpolat@selcuk.edu.tr (K. Polat), sgunes@selcuk.edu.tr (S. Güneş).

0957-4174/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2007.11.051

ANFIS. Also, Nanni (2006) obtained 97.22%, 97.22%, 97.22%, 97.22%, 97.22%, 97.22%, 97.22%, and 97.22% using LSVM, RS, B1_5, B1_10, B1_15, B2_5, B2_10, and B2_15 algorithms, respectively. In this work, we obtained 99.00% classification accuracy. Polat et al. obtained 91.84% and 92.94% classification accuracies using combination of C4.5 decision tree with fuzzy weighted pre-processing and combination of C4.5 decision tree with *k*-NN based weighted pre-processing on the diagnosis of erythemato-squamous diseases, respectively (Polat & Güneş, 2006).

There have been several studies reported focusing on classification of image segmentation. Among these studies, Tin and Kwork obtained 83% classification accuracy using support vector machine (SVM) on the classification of image segmentation (Tin & Kwork, 1999). Tolson achieved 85.2% success rate with *k*-NN (*k*-nearest neighbor) classifier on the same dataset (Tolson (2001)). 87.6% and 86.7% classification accuracies were obtained using PNN (probabilistic neural network) and GRNN (generalized regression neural network) by Coşkun and Yıldırım (2003), respectively. Polat, Şahan, Kodaz, and Güneş (2005) achieved 88.2% and 90.00% classification accuracies using AIRS (artificial immune recognition system) and Fuzzy-AIRS on the classification of image segmentation dataset.

As far as we know, there are a few studies related with classification of lymphography dataset in literature. Newton Cheung obtained 79.72%, 80.88%, and 81.08% classification accuracies using Naive Bayes, BNNF (Bayesian Network with Naïve dependence & Feature selection), and BNND (Bayesian Network with Naïve Dependence), respectively (Cheung, 2001). (Polat & Güneş (2006)) achieved 83.138% and 90.00% classification accuracies using AIRS and Fuzzy-AIRS on the classification of lymphography dataset.

In this study, we have proposed a novel hybrid classification system based on C4.5 decision tree classifier and one-against-all approach to classify the multi-class problems including dermatology, image segmentation, and lymphography datasets taken from UCI (University of California Irvine) machine learning database. To test the proposed method, we have used the classification accuracy, sensitivity-specificity analysis, and 10-fold cross validation. In this work, firstly C4.5 decision tree has been run for all the classes of dataset used and achieved 84.48%, 88.79%, and 80.11% classification accuracies for dermatology, image segmentation, and lymphography datasets using 10-fold cross validation, respectively. The proposed method based on C4.5 decision tree classifier and one-against-all approach obtained 96.71%, 95.18%, and 87.95% for above datasets, respectively. These results show that the proposed method has produce very promising results in the classification of multi-class problems.

2. Used datasets

We have used three dataset including dermatology, image segmentation, and lymphography datasets taken from UCI (University of California Irvine) machine learning database. We have explained the above datasets in the following subsections.

2.1. Dermatology dataset

Melanin incontinence is a diagnostic feature for lichen planus, fibrosis of the papillary dermis is for chronic dermatitis, exocytosis may be seen in lichen planus, pityriasis rosea and seboreic dermatitis. Acanthosis and parakeratosis can be seen in all the diseases in different degrees. Clubbing of the rete ridges and thinning of the suprapapillary epidermis are diagnostic for psoriasis. Disappearance of the granular layer, vacuolization and damage of the basal layer, saw-tooth appearance of retes and a band like infiltrate are diagnostic for lichen planus. Follicular horn plug and perifollicular parakeratosis are hints for pityriasis rubra pilaris. The features of a patient are represented as a vector of features, which has 34 entries for each feature value. In the dataset, the family history feature has the value 1 if any of these diseases has been observed in the family and 0 otherwise. The age feature simply represents the age of the patient. Every other feature (clinical and histopathological) was given a degree in the range of 0–3. Here, 0 indicates that the feature was not present, a 3 indicates the largest amount possible and 1, 2 indicate the relative intermediate values. Each feature has either nominal (discrete) or linear (continuous) value having different weights showing the relevance to the diagnosis (Demiroz et al., 1998; Govenir & Emeksiz, 2000).

This erythemato-squamous diseases database comes from the Gazi University and Bilkent University and was supplied by Nilsel Ilter, M.D., Ph.D., and H. Altay Guvenir, Ph.D. This dataset contains 34 attributes, 33 of which are linear valued and one of them is nominal (Demiroz et al., 1998; Govenir & Emeksiz, 2000; Machine-learning-databases, 2007). This dataset originally contains 366 instances. Distribution according to class variable of this dataset is given in Table 1.

2.2. Image segmentation dataset

The problem to be solved here is classification of outdoor image dataset. This dataset was taken from Vision Group, University of Massachusetts in 1990 with the

Table 1
Distribution according to class variable of this dataset

Class code	Class	Number of instances
1	*Psoriasis*	112
2	*Seboreic dermatitis*	61
3	*Lichen planus*	72
4	*Pityriasis rosea*	49
5	*Cronic dermatitis*	52
6	*Pityriasis rubra pilaris*	20

contributions of Carla Brodley. In image segmentation data set, the instances were drawn randomly from a database of 7 outdoor images. The images were hand segmented to create a classification for every pixel. Each instance is a 3 × 3 region. In training data there are 210 instances and in test data there are 2100 instances with 19 continuous attributes (Coşkun & Yıldırım, 2003; Machine-learning-databases, 2007). In the data set, the third attribute is the same for all inputs therefore while the simulations are being done this attribute is not added to network. The existing seven classes are grass, path, window, cement, foliage, sky, and brickface (Coşkun & Yıldırım, 2003; Machine-learning-databases, 2007).

2.3. Lymphography dataset

This Lymphography database was obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. There are 148 instances in total and there are no missing attributes. There are 18 numeric valued attributes which are listed as follows (Machine-learning-databases, 2007):

- Lymphatic – A test for the overall lymphatic system; and value 1 for normal;
- Value 2 for arched, value 3 for deformed & value 4 for displaced;
- Block of afferent – value 1 for no & value 2 for yes;
- Block of lymph c – value 1 for no & value 2 for yes;
- Block of lymph s – value 1 for no & value 2 for yes;
- By pass – value 1 for no & value 2 for yes;
- Extravasates – expel from a vessel and is represented by 1 and 2;
- Regeneration – value 1 for no & value 2 for yes;
- Early uptake – value 1 for no & value 2 for yes;
- Lymph nodes dimension – ranges from 0 to 3;
- Lymph nodes enlarge – range from 1 to 4;
- Changes in lymph – value 1 for bean, value 2 for oval & value 3 for round;
- Defect in node – value 1 for no, value 2 for lacunars, value 3 for lacunars;
- Marginal & value 4 for lacunars central;
- Changes in node – value 1 for no, value 2 for lacunars, value 3 for lacunars;
- Marginal & value 4 for lacunars central;
- Changes in structure – the structure of the lymphatic system;
- Special forms – value 1 for no, value 2 for chalices & value 3 for vesicles;
- Dislocation of node – value 1 for no & value 2 for yes;
- Exclusion of node – value 1 for no & value 2 for yes;
- Number of nodes – ranges from 0 to 80.

There are four classes in the class variables: normal, metastases, malign lymph and fibrosis that are represented by integer 1, 2, 3 and 4, respectively.

3. The proposed method

3.1. Overview

In this paper, we have combined the C4.5 decision tree classifier and one-against-all method to classify Multi-Class datasets including dermatology, image segmentation, and lymphography datasets. The numbers of class are 6, 7, and 4 classes for dermatology, image segmentation, and lymphography datasets, respectively. Fig. 1 shows the working of proposed method for dermatology dataset. Fig. 2 presents the working of proposed method for image segmentation dataset. Fig. 3 displays the working of proposed method for lymphography dataset. The used sections in proposed method are explained in the following subsections.

3.2. C4.5 Decision tree classifier

C4.5 Decision tree learning is one of the most widely used and practical methods for inductive inference. It is a method for approximating discrete-valued functions that is robust to noisy data and capable of learning disjunctive expressions (Mitchell, 1997; Quinlan, 1986).

C4.5 Decision tree learning is a method for approximating discrete-valued functions, in which the learned function is represented by a decision tree. Learned trees can also be-represented as sets of if-then rules to improve human readability. These learning methods are among the most popular of inductive inference algorithms and have been successfully applied to a broad range of tasks from learning to diagnose medical cases to learning to assess credit risk of loan applicants. C4.5 Decision tree learning is a heuristic, one-step look ahead (hill climbing), non-backtracking search through the space of all possible decision trees (Mitchell, 1997; Quinlan, 1986; Yi & Zheng, 2005).

The aim of C4.5 Decision tree learning is recursively partition data into sub-groups. Working of C4.5 Decision tree learning is as follows:

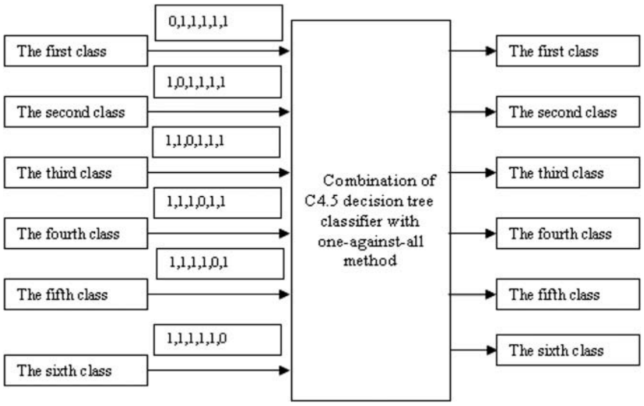

Fig. 1. The block scheme of proposed method for dermatology dataset (for six classes: these classes are psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis, pityriasis rubra pilaris).

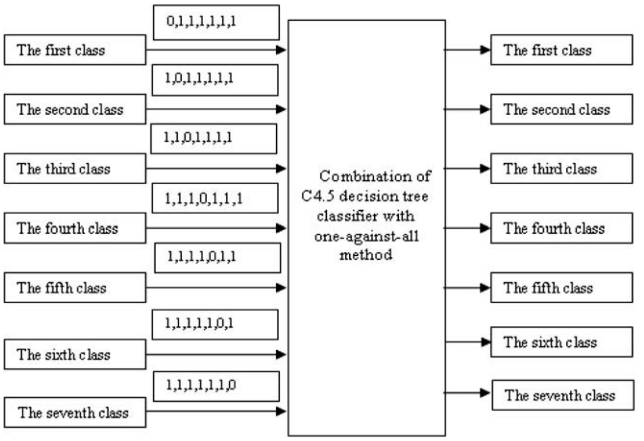

Fig. 2. The block scheme of proposed method for image segmentation dataset (for seven classes: these classes are grass, path, window, cement, foliage, sky, and brickface).

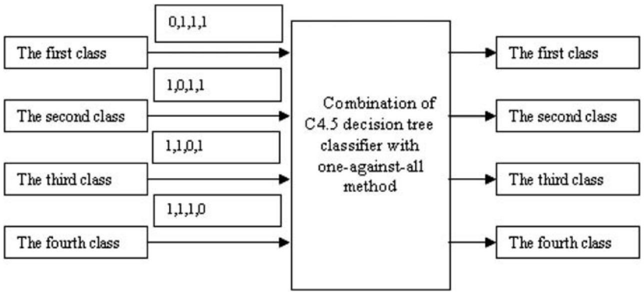

Fig. 3. The block scheme of proposed method for lymphography dataset (for four classes: these classes are normal, metastases, malign lymph and fibrosis).

- Select an attribute and formulate a logical test on attribute.
- Branch on each outcome of test, move subset of examples (training data) satisfying that outcome to the corresponding child node.
- Run recursively on each child node.
- Termination rule specifies when to declare a leaf node.

Training of C4.5 Decision tree learning is given in Fig. 4.

Definitions that used training of C4.5 Decision tree learning are explained as follows:

```
Decision Tree (examples)
Prune (Tree_Generation(examples)
Tree_Generation(examples)=
        IF termination_condition(examples)
                THEN leaf (majority_class (examples)
        ELSE
LET
Best_test=selection_function(examples)
IN
        FOR EACH value of Best_test
        Let subtree_v=Tree_Generation({e η example | e.Best_test=v})
                        IN Node (Best_test, subtree_v)
```

Fig. 4. Training algorithm of C4.5 decision tree classifier.

- Selection: used to partition training data.
- Termination condition: determines when to stop partitioning.
- Pruning algorithm: attempts to prevent overfitting.

3.3. One-against-all approach

Consider an M-class problem, where we have N training samples: $\{x_1, y_1\}, \ldots, \{x_N, y_N\}$. Here $x_i \in R^m$ is a m-dimensional feature vector and $y_i \in \{1, 2, \ldots, M\}$ is the corresponding class label (Yi & Zheng, 2005).

One-against all approach constructs M binary C4.5 decision tree classifier, each of which one class separates one class from all the rest. The ith C4.5 decision tree classifiers are trained with all the training examples of the ith class with positive labels and all the others with negative labels. Mathematically the ith C4.5 decision tree classifiers solve the following problem that yields the ith decision function $f_i(x) = w_i^T\phi(x) + b_i$:

$$\begin{aligned} &\text{minimize}: \quad L(w, \xi_j^i) = \frac{1}{2}\|w_i\|^2 + C\sum_{l=1}^{N} \xi_j^i \\ &\text{subject to}: \quad \tilde{y}_j(w_i^T\phi(x_j) + b_i) \geqslant 1 - \xi_j^i, \quad \xi_j^i \geqslant 0, \end{aligned} \tag{1}$$

where $\tilde{y}_j = 1$ if $y_j = i$ and $\tilde{y}_j = -1$ otherwise.

At the classification phase, a sample x is classified as in class i^* whose f_{i^*} produces the largest value.

$$i^* = \arg\ \max f_i(x) = \arg\ \max(w_i^T\phi(x) + b_i). \tag{2}$$

4. The empirical results and performance evaluation

In this section, we used the performance evaluation methods including classification accuracy, sensitivity and specificity analysis, and 10-fold cross validation to evaluate the proposed method.

4.1. Classification accuracy

In this study, the classification accuracies for the datasets are measured using Eq. (3):

$$\begin{aligned} &\text{accuracy}(T) = \frac{\sum_{i=1}^{|T|}\text{assess}(t_i)}{|T|}, \quad t_i \in T \\ &\text{assess}(t) = \begin{cases} 1, & \text{if classify}(t) = t.c \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{3}$$

where T is the set of data items to be classified (the test set), $t \in T$, $t.c$ is the class of item t, and classify(t) returns the classification of t by C4.5 decision tree classifier.

4.2. Sensitivity and specificity analysis

For sensitivity and specificity analysis, we use the following expressions.

Table 2
The obtained classification accuracies, sensitivity and specificity values for C4.5 decision tree classifier and proposed method with 10-fold cross validation

Used datasets	Method used	Classification accuracy (%)	Sensitivity (%)	Specificity (%)
Dermatology dataset	C4.5 Decision tree classifier	84.48	86.00	82.92
	1.Class (*Psoriasis*)	98.11	98.97	97.50
	2. Class (*Seboreic dermatitis*)	93.87	94.05	93.82
	3. Class (*Lichen planus*)	98.33	98.97	97.50
	4. Class (*Pityriasis rosea*)	93.87	94.05	93.82
	5. Class (*Cronic dermatitis*)	96.67	97.00	96.34
	6. Class (*Pityriasis rubra pilaris*)	99.43	99.00	98.79
	Combination of C4.5 decision tree and one against all approach	96.71	97.00	96.34
Image segmentation dataset	C4.5 decision tree classifier	88.79	89.36	88.32
	1.Class (grass)	91.87	90.90	92.42
	2. Class (path)	100	100	100
	3. Class (window)	92.80	92.78	92.53
	4. Class (cement)	92.07	87.72	96.69
	5. Class (foliage)	90.2	89.00	91.60
	6. Class (sky)	99.48	100	99.24
	7. Class (brickface)	99.87	100	100
	Combination of C4.5 decision tree and one against all approach	95.18	94.05	96.15
Lymphography dataset	C4.5 decision tree classifier	80.11	88.88	66.66
	1.Class (normal)	98.46	90.90	100
	2. Class (metastases)	74.21	80.00	60.00
	3. Class (malign lymph)	82.22	88.88	66.67
	4. Class (fibrosis)	96.92	100	83.33
	Combination of C4.5 decision tree and one against all approach	87.95	90.00	80.00

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}\ (\%) \tag{4}$$

$$\text{specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}}\ (\%) \tag{5}$$

where TP, TN, FP and FN denote true positives, true negatives, false positives, and false negatives, respectively.

4.3. k-Fold cross validation

k-Fold cross validation is one way to improve the holdout method. The dataset is divided into k subsets, and the holdout method is repeated k times. Each time, one of the k subsets is used as the test set and the other $k-1$ subsets are put together to form a training set. Then the average error across all k trials is computed. Every data point appears in a test set exactly once, and appears in a training set $k-1$ times. The variance of the resulting estimate is reduced as k is increased. A variant of this method is to randomly divide the data into a test and training set k different times (Jeff Schneider's Home Page, 2007).

4.4. Results and discussion

In this paper, we have proposed a novel hybrid classification system based on C4.5 decision tree classifier and one-against-all approach to classify the multi-class problems including dermatology, image segmentation, and lymphography datasets taken from UCI (University of California Irvine) machine learning database. To test the proposed method, we have used the classification accuracy, sensitivity-specificity analysis, and 10-fold cross validation. In this work, firstly C4.5 decision tree has been run for all the classes of dataset used and achieved 84.48%, 88.79%, and 80.11% classification accuracies for dermatology, image segmentation, and lymphography datasets using 10-fold cross validation, respectively. The proposed method based on C4.5 decision tree classifier and one-against-all approach obtained 96.71%, 95.18%, and 87.95% for above datasets, respectively. Table 2 shows the obtained results for dermatology, image segmentation, and lymphography datasets.

As can be seen from above results, the proposed method has produced very promising results on the classification of multi-class datasets including dermatology, image segmentation, and lymphography datasets. This method can be used in many pattern recognition applications. In future, instead of C4.5 decision tree, other classification algorithms such as Bayesian learning, artificial immune system algorithms, artificial neural networks can be used.

5. Conclusions and future work

Classifying of multi-class classification problems such as document categorization, image recognition, signal classification, etc. that have multi-class is important issue in pattern recognition applications. In this work, in order to solve this problem, the used datasets were classified as two groups using one-against-all method. In this work, firstly C4.5 decision tree has been run for all the classes of dataset used and achieved 84.48%, 88.79%, and 80.11% classification accuracies for dermatology, image segmentation, and lymphography datasets using 10-fold cross validation, respectively. The proposed method based on C4.5 decision tree classifier and one-against-all approach obtained 96.71%, 95.18%, and 87.95% for above datasets, respectively. These results show that the proposed method has produced very promising results in the classification of multi-class problems. This method can be used in many pattern recognition applications. In future, instead of C4.5 decision tree, other classification algorithms such as Bayesian learning, artificial immune system algorithms, artificial neural networks can be used.

Acknowledgement

This study has been supported by Scientific Research Project of Selcuk University.

References

Cheung, N. (2001). Machine learning techniques for medical analysis. *School of Information Technology and Electrical Engineering*, BSc thesis, University of Queenland.

Coşkun, N., Yıldırım, T. (2003). Image segmentation using statistical neural networks. In *Proceedings of international conference on artificial neural networks/international conference on neural information processing (ICANN/ICONIP)* (Istanbul, Turkey, June 26–29).

Demiroz, G., Govenir, H. A., & Ilter, N. (1998). Learning differential diagnosis of Eryhemato-Squamous diseases using voting feature intervals. *Aritificial Intelligence in Medicine, 13*, 147–165.

Govenir, H. A., & Emeksiz, N. (2000). An expert system for the differential diagnosis of erythemato-squamous diseases. *Expert Systems with Applications, 18*, 43–49.

Jeff Schneider's Home Page, <http://www.cs.cmu.edu/~schneide/tut5/node42.html> (last accessed: May, 2007).

ftp://ftp.ics.uci.edu/pub/machine-learning-databases (last accessed: May, 2007).

Mitchell, M. T. (1997). *Machine learning*. Singapore: McGraw-Hill.

Nanni, L. (2006). An ensemble of classifiers for the diagnosis of erythemato-squamous diseases. *Neurocomputing, 69*, 842–845.

Polat, K., & Güneş, S. (2006). Automated identification of diseases related to lymph system from lymphography data using artificial immune recognition system with fuzzy resource allocation mechanism (fuzzy-AIRS). *Biomedical Signal Processing and Control, 1*(4), 253–260.

Polat, K., & Güneş, S. (2006). The effect to diagnostic accuracy of decision tree classifier of fuzzy and k-NN based weighted pre-processing methods to diagnosis of erythemato-squamous diseases. *Digital Signal Processing, 16*(6), 922–930.

Polat, K., Şahan, S., Kodaz, H., & Güneş, S. (2005). Outdoor image classification using artificial immune recognition system (AIRS) with performance evaluation by fuzzy resource allocation mechanism. *LNCS, 3691*, 81–87, CAIP 2005.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*, 81–106.

Tin, J., & Kwork, Y. (1999). Moderating the outputs of support vector machine classifiers. *IEEE Transactions On Neural Networks, 10*(5), 1018–1031.

Tolson, E. (2001). Machine learning in the area of image analysis and pattern recognition. Advanced Undergraduate Project Spring.

Ubeyli, E. D., & Guler, I. (2005). Automatic detection of erythemato-squamous diseases using adaptive neuro-fuzzy inference systems. *Computers in Biology and Medicine, 35*, 421–433.

Yi Liu, Zheng, Y.F. (2005). One-against-all multi-class SVM classification using reliability measures. In *Proceedings of the IEEE international joint conference on neural networks IJCNN '05 (Vol. 2 pp. 849–854)*.