Genetic Algorithms for Thyroid Gland Ultrasound Image Feature Reduction

Ludvík Tesař[1], Daniel Smutek[2], and Jan Jiskra[2]

[1] Institute of Information Theory and Automation, Czech Academy of Sciences, Prague, Czech Republic
[2] 3rd Department of Medicine, 1st Medical Faculty, Charles University, Prague, Czech Republic

Abstract. The problem of automatic classification of ultrasound images is addressed. For texture analysis of ultrasound images quantifiable indexes, called features, are used. Classification was performed using Gaussian mixture model based on Bayes classifier. The common problem of texture analysis is a feature selection for classification tasks. In this work we use genetic algorithms for a feature subset selection. Total number of 387 features was used, consisting of spatial an co-occurance statistical texture features (proposed by Muzzolini and Haralick). The classification infers between healthy thyroid gland and thyroid gland with chronic inflammation.

1 Introduction

Ultrasound imaging is a very important cost-effective method for diagnostics of thyroid gland diseases. For most thyroid disorders it surpasses the more expensive magnetic resonance. Image analysis can give more objective way of diagnosing patient than a physician who relies on his experience only.

We use Bayes classifier, where diagnose was verified by other methods. The method we use very successfully (See [1, 2]) employs Gaussian Mixture model in feature space. Reduction of number of features can help to reduce number of computations substantially, and to better understand, which pattern feature characterize thyroid gland inflammation in ultrasound image. In this work we are using genetic algorithm for feature selection.

2 Classification Method Description

Texture features are computed from a set of fixed-size rectangular regions referred to as texture samples. The non-overlapping samples are obtained from a manually segmented sonographic B-mode image of thyroid gland. Haralick texture features [4] were computed from the co-occurrence matrix. Muzzolini's spatial features, originally suggested by Muzzolini [5], are based on the original pixel gray levels. For closer feature description refer to [3]. Finally 387 different texture feature values were known for each patient.

The vector Y of features, is modeled using Gaussian mixture model:

$$p(Y) = \sum_{i=1}^{n} \frac{\alpha_i \exp\left[-\frac{1}{2}(Y - Y_i)^{\mathrm{T}} C_i^{-1}(Y - Y_i)\right]}{(2\pi)^{\frac{d}{2}} |C_i|^{\frac{1}{2}}} \tag{1}$$

Where n is order of the mixture, d dimension of vector Y, $|\cdot|$ denotes determinant of matrix, symbol $()^{\mathrm{T}}$ denotes transposition. Important condition is, that $\sum_{i=1}^{n} \alpha_i = 1$.

During learning process, parameters of the model (1) are estimated. Two sets of parameters are calculated, one for healthy and one for inflamed tissue. The diagnosis is obtained by applying features of given patient to probability density functions from equation (1) fitted to healthy and unhealthy patients. The method is more more in detail described in [2]. The result of classification method is used in genetic algorithm to evaluate the fitness of an individual, by selecting only features that are attached to this individual and by doing classification only on such subset of features.

3 Genetic Algorithm Description

In explanation of the genetic algorithm, we will use terms *population*, *generation* and *individual* as follows: Each individual in our population was representing the set of features (its chromosomes). The population is set of individuals. In every generation, chromosomes (features) of individuals in the population are crossed in order to create the next generation. Fitness of individual is evaluated according to the success of its features in classification.

Every individual in our population have given number (in our experiments 5 or 10) of chromosomes. Every chromosome represents one feature. Features are numbered by numbers between 1 and 387, so chromosome is simply one number between 1 and 387. If two parents are to have an offspring, chromosomes of an offspring are randomly selected from its parents. Let $D_1 = [c_{1,1}, c_{1,2}, \ldots, c_{1,n}]$ is an ordered n-tuple representing chromosomes of the first parent and $D_2 = [c_{2,1}, c_{2,2}, \ldots, c_{2,n}]$ of the second parent, then $D_3 = [c_{2,1}, c_{2,2}, \ldots, c_{2,n}]$ defined as $D_3 = [c_{i(1),1}, c_{i(2),2}, \ldots, c_{i(n),n}]$ is genetic information of the offspring, where $i(1), i(2), \ldots, i(n)$ is vector of binary random numbers ($i(k) \in \{1, 2\}$), and n is number of chromosomes in individuals of our population. In every new offspring, mutation was made randomly with mutation rate given in per-individual basis, i.e. if mutation rate was 0.5, it means that there was one mutation per two new offsprings in average.

Algorithm is started by randomly chosen generation $\mathcal{G} = \{D_1, D_2, \ldots, D_h\}$ of h individuals with n chromosomes. Following steps are repeated for given number of generations:

1. **Selection.** Fitness of every individual $D_k \in \mathcal{G}$ is evaluated and only first ℓ are selected (Fitness of the individual is evaluated based on classification method described in Section 2). I.e. worst $h - \ell$ individuals are removed from

population. The new set will be $\mathcal{G} := \{D_1, D_2, \ldots, D_\ell\}$, assuming that D_k was already sorted by fitness. Note that in algorithm, we are using the same letter to represent different thing.

2. **Making offsprings.** h new individuals are created as offsprings of individuals in set $\mathcal{G}$. I.e. two individuals from the set $D_1, D_2 \in \mathcal{G}$ are repeatedly randomly selected as parents and new individual E_k is created (as explained above) as their offspring for $k \in \{1, 2, \ldots, h\}$. Now the set $\mathcal{G}$ is changed to hold the new generation: $\mathcal{G} := \{E_1, E_2, \ldots, E_h\}$.
3. **Mutation.** Every individual $E_k \in \mathcal{G}$ is mutated with given mutation rate.

Chromosomes of individuals from the last generation represent an optimal selection of features.

4 Testing of the Genetic Algorithm

We tested the proposed algorithm using the data with 100 subjects, of which 62 were patients with lymphocytic thyroiditis. The diagnosis was confirmed by fine-

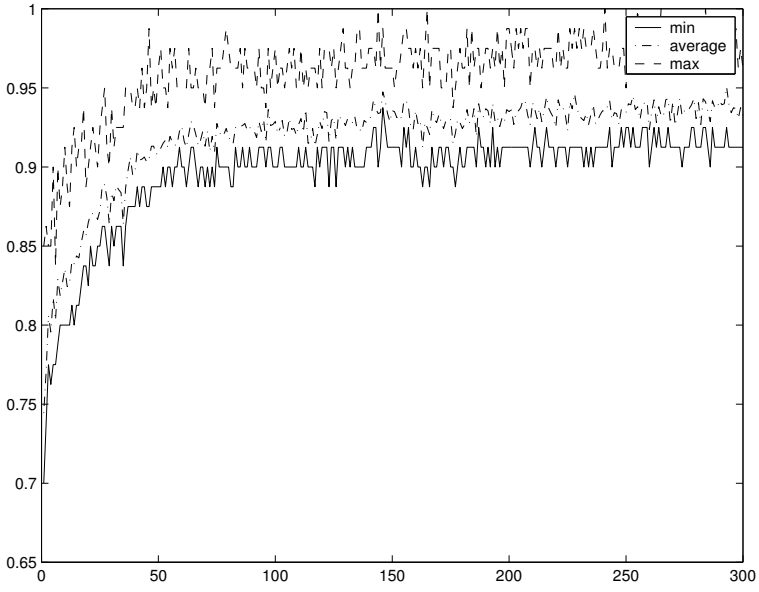

Fig. 1. Results of run with 300 generations. Graphs show minimum, average and maximum of the success rate of individuals from the population.

needle aspiration biopsy, an increased level of antibodies (anti-thyroperoxidase and anti-thyroglobulin) and by clinical examination. Another 38 subjects were

healthy test persons (volunteers) with mean age 28 ± 14 years with no known thyroid disease.

The principal parameters of the sonograph were fixed in the study. Details concerning data acquisition and processing are the same as in [3].

We made several runs with number of generations between 10 and 300 and with number of chromosomes between 5 and 10. Number of mutations was 0.5 to 3 per generation. Population size h was between 40 and 80 and parameter ℓ was selected to be 10 to 20.

Graph in Figure 1 shows minimum, average and maximum of the success rate of individuals from the population in the typical run of algorithm. We can see that genetic algorithm converged after 70 generations already.

5 Conclusions

Compared to our earlier paper [6], results of classification are better, because of much better classifier, which was developed in [2]. We have found the most suitable quantitative indicators of an ultrasound examination of thyroid gland, assuming they include the highest amount of information for texture recognition of chronic inflammation in thyroid tissue. Such indicators enable reproducibility of the examination, facilitate an assessment of changes of the disease in time and make the comparison of different physicians' ultrasound findings possible.

6 Acknowledgment

Paper was supported by Grant agency of Academy of Sciences of the Czech Republic by project 1ET101050403.

References

1. Tesař, L., Smutek, D.: Bayesian classification of sonograms of thyroid gland based on Gaussian mixtures. In: Proceedings of Norwegian Conference on Image Processing and Pattern Recognition, NOBIM 2004, Stavanger, Norway (2004) 36–40
2. Smutek, D., Šára, R., Jiskra, J., Tesař, L.: Ultrasound of thyroid gland - what is hidden inside and physician does not see. In Cikes, Nada, eds.: Lijecnicki Vjesnik; Abstracts from European Congress on Ultrasound in Medicine and Biology 126 (Suppl. 2), Zagreb, Croatia, Kratis - Zagreb (2004) 57
3. Smutek, D., Šára, R., Sucharda, P., Tjahjadi, T., Švec, M.: Image texture analysis of sonograms in chronic inflammations of thyroid gland. Ultrasound in Medicine and Biology **29** (2003) 1531–1543
4. Haralick, R.M., Shapiro, L.G. In: Computer and Robot Vision. Volume 1. Addison-Wesley Publishing Company, Reading MA (1993) 453–508
5. Muzzolini, R., Yang, Y.H., Pierson, R.: Texture characterization using robust statistics. Pattern Recognition **27** (1994) 119–134
6. Smutek, D., Semecký, J.: Feature selection by genetic algorithms in image texture analysis of thyroid gland ultrasound. In: IFBBE Proceedings: 2nd European Medical and Biological Engineering Conference EMBEC'02. Volume 3., Verlag der Technischen Universität Graz (2002) 878–879