*pattern recognition, multispectral analysis, dimensionality reduction, feature selection*

Marcin MICHALAK[2], Adam ŚWITOŃSKI[1,2], Magdalena STAWARZ[2]

# SELECTION OF THE MOST IMPORTANT COMPONENTS FROM MULTISPECTRAL IMAGES FOR DETECTION OF TUMOR TISSUE

The problem raised in this article is the selection of the most important components from multispectral images for the purpose of skin tumor tissue detection. It occured that 21 channel spectrum makes it possible to separate healthy and tumor regions almost perfectly. The disadvantage of this method is the duration of single picture acquisition because this process requires to keep the device very stable. In the paper two approaches to the problem are presented: hill climbing strategy and some ranking methods.

## 1. INTRODUCTION

Reduction of the data dimensionality is a very important part of machine learning and data mining techniques. On the one hand every scientist would have as many analysed object's or phenomenon's features as possible but on the other hand to much information may introduce some disinformation. For example we can not get exact information about two numbers if we know only the result of their addition or multiplication. The ideal situation - when measurements do not introduce any error - is when we know results of the both operations. If calculation results are noised and contain not only the result of addition and multiplication but also for example result of raising into the power or rooting extracting the information about the original numbers will be also difficult. It will be caused by _ too much number of analysed objects features.

The other argument that speaks for dimensionality reduction is the cost of obtaining the whole set of attributes for the object. It may occur that there are some redundant variables which values are hard or expensive to be measured.

The last problem of multidimensional data is called "the curse of dimensionality" and means that increasing the number of attributes decreases the algorithm (classifier or regressor) ability to describe dependencies within the data.

Very similar approach of data analysis is called feature extraction. With this approach we try to evaluate new attributes on the basis of the original. It may occur that the smaller number of features calculated from the set of original features will give us better results of classification or regression. This approach is usually used when the cost of original attributes aquisition is negligible.

This article raises the problem of selection most significant color components for the task of skin tumor tissue detection. In our research the set of multispectral images was given where every pixel was described with 21 color components and the expert given value: healthy or tumor. Because using all components gave quite good results we tried to limit the number of components as aquisition of every component delays the total aqcuisition time. The whole process of data capturing is performed manually so the quality of multispectral images strongly depends on the aquisition duration.

This paper is organized as follows: the next section describes the research context and our previous works in this area. Then algorithms that were used for feature selection are described including some rankings and climbing strategy. Afterwards experiments and their results are presented. Finally certain summary conclusions are written and perspectives of further works are mentioned.

[1] Polish-Japanese Institute of Information Technology, ul. Koszykowa 86, 02-008 Warszawa, Poland.
[2] Silesian University of Technology, ul. Akademicka 16, 44-100 Gliwice, Poland,
email: {Marcin.Michalak, Adam.Switonski, Magdalena.Stawarz}@polsl.pl.

## 2. ANALYSIS BACKGROUND

Image analysis becomes more and more important in medicine. Our research deals with the analysis of multispectral skin pictures in the context of tumor tissue detection. Skin fragments were lighted up with white and blue light and the reflection was captured as the 21 components multispectral image. On the left side of the Fig. 1 the acquisition device is shown.

On the basis of the experts opinion about regions with tumor we tested most popular classifying algorithms [12] what gave quite satisfactory results: artificial neural network had the ability of infallibility when the skin was lighted up with the white light. Visualization of the results is shown on the right side of the Fig. 1.

To analyze the real quality of obtained pictures a 24 color specimen was used and the camera results were compared to the given spectral color definition [6]. Results of this experiment led us to the kernel postprocessing of the obtained multispectral images [5].

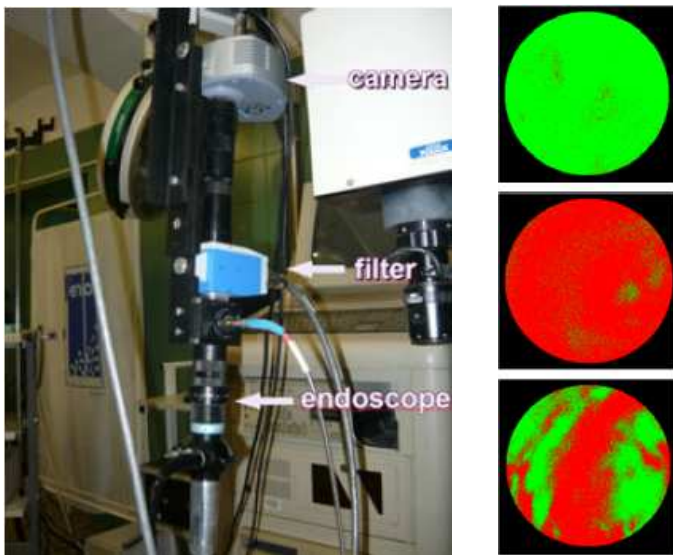


Fig. 1. Left: Acquisition device; Right: Sample results visualization (red – tumor, green – healthy).

## 3. FEATURE SELECTION

As it was mentioned in the introduction there are two main groups of algorithms dedicated for dimensionality reduction: feature extraction and feature selection. From our point of view in the situation when *ANN* can achieve almost a 100% accuracy there is no need to find some new attributes on the basis of the 21 spectral components because it can not improve the classification results. However, it is very interesting and desired to limit the number of components that would give the comparable level of tumor tissue detection.

The attributes selections can be performed by ranking based methods. In this approach we evaluate each of the attribute separately by some kind of the quality index and select the highest scored ones. The number of attributes to select could be given or could be determined based on the scores obtained. The crucial problem in the ranking approach is the way attributes are evaluated.

In order to compare the following methods were chosen: information gain (IG), gain ratio (GR), $\chi^2$ statistic, OneRule ranking, Gini Index (GI), Fisher ratio (FR) and t-test based method. First two ranks [14,7] are based on the entropy of the class with respect to given attributes. $\chi^2$ is the simple application of the $\chi^2$ statistics [10]. The One Rule classifier [4] determines a single rule based on the chosen single attribute to classify the instances. To find a such a simple rule, the OneRule algorithm tests every attribute. The Gini Index is a measure used in statistic to quantify inequality of random variable distribution [2,3,8]. Fisher ratio [13] uses means normalized by covariance to quantify the distance for the difference of feature means between two classes. The last ranking applied to ranking features is based on two-sample *t*-test with cumulative variance formula [15]. Test statistic value can be defined

as: $t(x) = (\mu_1 - \mu_2)^2 / \sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}$ , where $\mu_1, \mu_2$ and $\sigma_1^2, \sigma_2^2$ are accordingly means and variations of each feature observations for class 1 and 2. The alternative hypothesis $H_A$ is that values of compared classes are different, while the null hypothesis $H_0$ assumes that they are equal. Greater value of $t(x)$ indicates stronger evidence supporting the $H_A$, that means of two classes are different.

For the entropy and the chi square based rankers we have applied discretization based on the Minimal Descriprion Length Principle [1,9]. The idea of the discretization is to find the thresholds which minimize the average entropy of the class in the following buckets.

The climbing strategy is the heuristic that starts with empty set of output attributes and in each iteration the only one from remaining attributes is moved to the output set – the one that improves the classification accuracy the best. For example if the data contains *m* attributes we perform *m* classification experiments, for each attribute separately and observe the classification accuracy. As the first attibute the one that gives the best accuracy is moved to the output set. Then we have one attribute in the output set and *m-1* attributes in the input set. The next iteration contains *m-1* classification experiments and each of them is performed with the usage of one from the *m-1* attributes and the whole output attributes set. Completing of the output set stops when addition of any of the resting attribute does not improve the classification accuracy.

If we assume that the complexity of the single classification execution is constant (for the given number of data objects and the variable number of attributes) the complexity of this strategy is $O(a^2)$ where *a* is the number of attributes. The pseudocode of this algorithm is listed below. The function *classAccuracy(data, atts)* is the proper classifier that works on the subset of attributes *atts*.

LISTING 1. Climbing strategy.

```
function climbing(data, atts)
begin
 local_atts := atts;
 best_acc := 0;
 best_atts := ϕ;
 new_acc :=0;
 do
 begin
  best_acc := new_acc;
  new_acc :=0;
  temp_acc := 0;
  foreach latt ∈ local_atts
  begin
   t_acc := classAccuracy(data, best_atts U { latt });
   if (t_acc > new_acc)
   begin
    best_atts := best_atts U { latt };
    local_atts := local_atts \ { latt }
   end
  end
 end
 while new_acc > best_acc
 return best_atts;
end
```

## 4. EXPERIMENTS AND RESULTS

In our experiments 5000 randomly selected objects (multispectral pixels) from all pictures obtained were taken into the analysis. The images contain 21 channels with wavelengths ranging from 400 nm to 720 nm with 16 nm step. Each of classes was represented by almost the same number of objects (2632 pixels from tumor region). This data set was divided into three disjoint subsets: train (4000 objects with 2084 from the tumor region), tune (500 objects with 268 from the tumor region) and test (500 objects with 280 from the tumor region). On the basis of the result from the previous research [12] an artificial neural network was selected as the classificational tool. Classification was performed with the usage of the Statistica 9 automatically. On the basis of the train set and the error on the tune set the best network was determined. Afterwards, the result of test set classification was taken as the quality of the specific attributes subset.

In the Tab. 1 the results of seven ranking attributes are presented as the permutation of spectrum component numbers (400 nm component is numbered as 1, component 416 nm as 2 and so on). All rankings were calculated on the sum of train and tune set to separate these results from the data in the test set. The results of the climbing strategy are presented as the sequence of the spectrum components numbers subset. Last nonnegative number of the attribute means that adding the following attributes did not improve the classification accuracy.

Table. 1 Results of ranking and climbing strategy.

| ranking\rank | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| InfoGain | 4 | 3 | 5 | 6 | 12 | 11 | 10 | 13 | 7 | 9 | 14 | 15 | 8 | 21 | 2 | 20 | 1 | 16 | 17 | 19 | 18 |
| $\chi^2$ | 4 | 3 | 5 | 6 | 12 | 11 | 13 | 10 | 7 | 9 | 14 | 15 | 8 | 21 | 2 | 20 | 1 | 16 | 17 | 19 | 18 |
| GainRatio | 5 | 21 | 2 | 6 | 4 | 7 | 3 | 12 | 13 | 10 | 11 | 15 | 9 | 14 | 8 | 19 | 1 | 20 | 16 | 18 | 17 |
| OneR | 18 | 12 | 4 | 6 | 3 | 11 | 21 | 7 | 10 | 5 | 8 | 13 | 2 | 14 | 17 | 9 | 16 | 20 | 15 | 19 | 1 |
| ttest | 3 | 4 | 2 | 5 | 6 | 7 | 10 | 12 | 11 | 21 | 20 | 8 | 13 | 14 | 17 | 9 | 1 | 16 | 15 | 19 | 18 |
| FR | 3 | 4 | 2 | 5 | 6 | 7 | 10 | 12 | 11 | 21 | 20 | 8 | 13 | 14 | 17 | 9 | 1 | 16 | 15 | 19 | 18 |
| GI | 12 | 13 | 14 | 15 | 10 | 16 | 11 | 17 | 18 | 19 | 20 | 9 | 8 | 7 | 6 | 5 | 21 | 4 | 3 | 2 | 1 |
| climbing | 4 | 18 | 12 | 6 | 14 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

On the Fig. 2 we may observe that the climbing strategy achieves the maximal accuracy just in six steps. It means that over 70% of attributes contain redundant information. The OneR ranking seems to be the closest to the climbing strategy. This ranking should be also admitted as the best of other rankings together with the *GI*. They approach to their maximal values the most fast. Together they also have the greater accuracies than other rankings.

What is also interesting the two pairs of rankings occur to be equal and almost equal: InfoGain and $\chi^2$ differs on the 7$^{th}$ and 8$^{th}$ position; *t*-test and Gini Index generated exactly the same rankings.

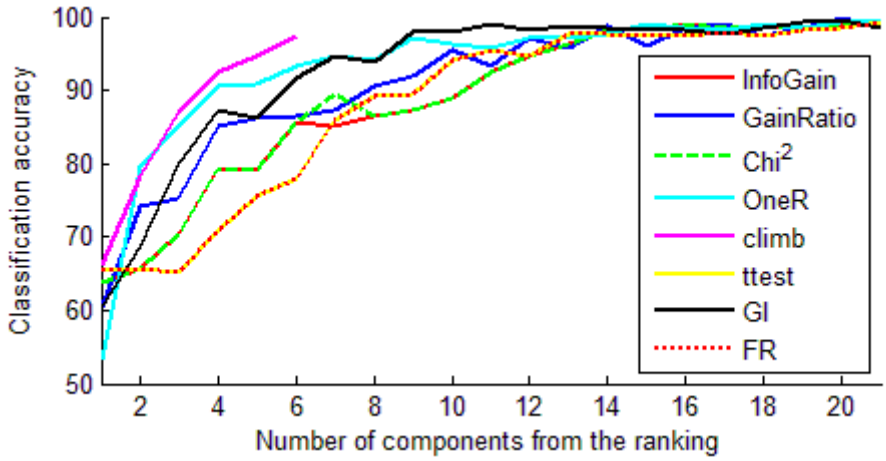


Fig. 2. Comparison of climbing hill strategy and rankings efficacy.

When we compare the first six components from rankings with the set of components given by the hill climbing strategy we may observe some interesting facts: the first component from the spectrum is the last one pointed by the hill climbing but in rankings it usually takes far places ($17^{th}$ or even the last one). Furthermore, only 2-3 from the first six components given by ranking was also pointed by the hill climbing strategy. This points, that evaluation of attribute relevance given by rankings and hill climbing strategy is very different.

## 5. CONCLUSIONS AND FURTHER WORKS

In this paper applications of several algorithms of feature selection were compared. This analysis is the direct continuation of the previous work around detection of skin tumor with the usage of multispectral pictures. Due to the fact that the full set of spectra components gives almost 100% accuracy of regions classification we tried to limit the number of necessary components what should shorten the multispectral pictures acquisition time and improve their quality (elimination of the human inaccuracy of keeping the acquiring device steady).

Two strategies of features selection was taken into consideration: climbing hill strategy and seven features rankings. The hill climbing strategy achieved the level of the reference accuracy (the accuracy of the full set of components classification) faster than every ranking strategy. Interestingly enough only 6 components give the same amount of information as the full 21 component spectrum. It means that the further data acquisition should be done four times faster.

Our further works will focus on developing the multispectral images database with only the most significant components given by the hill climbing strategy.

## 6. ACKNOWLEDGMENTS

This work was financed from the Polish Ministry of Science and Higher Education resources in 2009-2012 as a research project. Participation of the third Author was supported by the European Community from the European Social Fund.

## BIBLIOGRAPHY

[1] FAYAD U.M., IRANI K.B., Multi-interval discretization of continuousvalued attributes for classification learning, Thirteenth International Joint Conference on Articial Intelligence, 1993, pp. 1022-1027.

[2] GINI C., On the measurement of concentration and variability of characters, METRON - International Journal of Statistics, Vol. LXIII, No. 1, 2005, pp. 3-38.

[3] GINI C., Variabilit`a e mutabilit`a, 1912, Reprinted in Memorie di metodologica statistica (Ed. PIZETTI E., SALVEMINI T.), Rome: Libreria Eredi Virgilio Veschi, 1955.

[4] HOLTE R.C., Very simple classification rules perform well on most commonly used datasets, Machine Learning Vol. 11, No. 1, 1993, pp. 63-90.

[5] MICHALAK M., ŚWITOŃSKI A., Kernel postprocessing of multispectral images, Computer Recognition Systems 4, AISC 95, Springer, 2011, pp. 395-401.

[6] MICHALAK M., ŚWITOŃSKI A., Spectrum evaluation on multispectral images by machine learning techniques, BOLC L. et al. (Eds.): ICCVG 2010, Part II, LNCS 6375, Springer-Verlag Berlin Heidelberg, 2010, pp. 126-133.

[7] MITCHELL T.M., Machine Learning, The McGraw Hill, 1997.

[8] NISBET R., ELDER J., MINER G., Handbook of statistical analysis and data mining applications, Academic Press, 2009.

[9] PFAHRINGER B., Compression-Based Discretization of Continuous Attributes, Proc. of the 12th Int. Conf. on Machine Learning, 1995, pp. 456-463.

[10] PLACKETT, R.L., Karl Pearson and the Chi-Squared Test, International Statistical Review, Vol. 51, No. 1, pp. 59-72.

[11] STEIN C., A two-sample test for a linear hypothesis whose power is independent of the variance, The Annals of Mathematical Statistics, Vol. 16, No. 3, 1945, pp. 243-258.

[12] ŚWITOŃSKI A., MICHALAK M., JOSIŃSKI H., WOJCIECHOWSKI K., Detection of tumor tissue based on the multispectral imaging, BOLC L. et al. (Eds.), ICCVG 2010, Part II, LNCS 6375, Springer-Verlag Berlin Heidelberg, 2010, pp. 325-333.
[13] WEISS S.M., Indurkhya N., Predictive data mining: a practical guide, The Morgan Kaufmann Series in Data Management Systems, 1997.
[14] WITTEN I., Frank E., Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2005.
[15] ZHU W., Wang X., Ma Y., Rao M., Glimm J., Kovach, J.S., Detection of cancerspecific markers amid massive mass spectral data, Proceedings of the National Academy of Sciences of the United States of America, 100(25), 2003, pp. 14666-14671.