A Hybrid Feature Selection Method based on IGSBFS and Naïve Bayes for the Diagnosis of Erythemato - Squamous Diseases

S. Aruna
Research Scholar
Dept of Computer Applications
Dr M.G.R Educational and
Research Institute University

L. V. Nandakishore
Asst Professor
Dept of Mathematics
Dr M.G.R Educational and
Research Institute University

S. P. Rajagopalan
Prof Emeritus
Dept of Computer Applications
Dr M.G.R Educational and
Research Institute University

ABSTRACT

In this paper a diagnostic model based on Naive Bayes developed to diagnose erytemato squamous diseases. The hybrid feature selection method, named IGSBFS (Information Gain and Sequential Backward Floating Search), combines the advantages of filters and wrappers to select the optimal feature subset from the original feature set. In IGSBFS, Information Gain acts as filters to remove redundant features and SBFS with Naïve Bayes acts as the wrappers to select the ideal feature subset from the remaining features We conducted experiments in WEKA with 10 fold cross validation. The algorithm selected an optimum feature subset of 10 features with 98.9% accuracy.

General Terms

Algorithms, Pattern Recognition.

Keywords

Erythemato squamous diseases, Feature selection, Information Gain, Naïve Bayes, Sequential backward floating search.

1. INTRODUCTION

Feature selection plays an important role in dimension reduction for classification [1]. The task of a feature selection algorithm (FSA) offers a computational solution to the problem motivated by a certain definition of relevance. Reduction of feature dimensionality is of much importance in data mining. Feature selection directly reduces the number of original features by selecting a subset of them that still retains enough information for classification. It reduces the dimension and lowers the computation consumption so that it can gain classification performance. The feature selection algorithms designed with different evaluation criteria are divided into two categories: the filter methods and wrappers methods [2]. Filter methods have high computational efficiency than wrapper methods. Wrapper approach achieves better results than filter methods. To combine the advantages of both methods we propose a hybrid algorithm to cut the feature space.

In dermatology the differential diagnosis of erythemato-squamous diseases is a difficult problem. The diseases in this group are psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, chronic dermatitis, and pityriasis rubra pilaris. They have little difference in their clinical and histopathological features. Because of this even with biopsy it is difficult to diagnose the disease. Moreover it may show the features of another disease at the beginning stage and their characteristic features at the following stages.

In this paper we propose a diagnostic model based on Naïve Bayes with hybrid feature selection method to diagnose the eythemato squamous diseases. Remaining of the paper organized as follows. Related work is given in Section 2, Section 3 gives a brief description about the FSA algorithms - Information Gain and SBFS, Naïve Bayes algorithm and dermatology dataset used for the experiment. Section 4 discusses the diagnostic model for erythemato squamous diseases, Section 5 gives the results obtained and the concluding remarks are given in Section 6 to address further research issues

2. RELATED WORK

There have been several studies reported focusing on the diagnosis of erythemato-squamous diseases using dermatology dataset. Among these studies Guvenir et al[3] conducted the first work on the differential diagnosis of erythemato-squamous diseases. They developed a new classification algorithm, called VFI5 (for Voting Feature Intervals) and applied it to problem of differential diagnosis of Erythemato-Squamous diseases. The VFI5 algorithm achieved 96.2% accuracy on the Dermatology dataset with 22 histopathological features. Guvenir and Emeksiz [4] presented an expert system for differential diagnosis of erythemato-squamous diseases incorporating decision made by three classification algorithms: nearest neighbor classifier, naïve Bayesian classifier and voting feature intervals-5. They obtained 99.2% classification accuracy on the differential diagnosis of erythemato-squamous diseases. Bojarczuk et al. [5] used a constrained-syntax GP algorithm. The results showed that the proposed GP obtained 96.64% classification accuracy rate considerably better than C4.5 (89.12%) a very well known advanced decision-tree algorithm often used in machine learning and data mining. Ubeyli and Guler proposed an approach based on adaptive neurofuzzy inference systems for detection of erythemato-squamous diseases [6], and got the classification accuracy of 95.5%.

Luukka and Leppalampi obtained 97.02% using fuzzy similarity classifier for diagnosis of erythemato-squamous diseases [7]. Polat and Gunes [8] proposed methods based on fuzzy weighted pre-processing, K-NN based weighted pre-processing, and decision tree classifier. Classification accuracy they reached are 88.18%, 97.57%, and 99.00% respectively. Nanni obtained 97.22%, 97.22%, 97.5%, 98.1%, 97.22%, 97.5%, 97.8%, and 98.3% using LSVM, RS, B1 5, B1 10, B1 15, B2 5, B2 10, and B2 15 algorithms [9]. Luukka presented similarity classifier using similarity measure derived from Yu's norm for classifying medical data sets [10], and the classification accuracy for the diagnosis of erythemato-squamous diseases was 97.8%. Ubeyli obtained 98.32% classification accuracy on the differential diagnosis of erythemato-squamous diseases [11], using multiclass support vector machines with the error correcting output codes. Polat and Gunes obtained 96.71% classification correct rate on diagnosis of erythemato-squamous diseases using a novel hybrid intelligence method based on C4.5 decision tree classifier and one against all approach for multi-class classification problem [12].

Ubeyli obtained about 97.77% classification accuracy using combined neural networks model to guide model selection for the diagnosis of erythemato-squamous diseases [13]. Liu et al. obtained 96.72%, 92.18%, 95.08%, and 92.20% using feature selection algorithm with dynamic mutual information, estimated using four typical classifiers named Naive Bayes, 1-Nearest neighbor, C4.5 and RIPPER [14]. Karabatak and Ince proposed a new feature selection method based on association rules and neural network for diagnosis of erythemato-squamous diseases, and their correct classification rate was 98.61%.[15]. Ubeyli and Dogdu [16] presented an approach based on the implementation of *k*-means clustering for automated detection of erythemato-squamous diseases. They obtained 94.22% prediction accuracy of the *k*-means clustering. Lekkas and Mikhailov [17] reviewed a method for evolving fuzzy classification which allows data to be processed in online mode by recursively modifying a fuzzy rule base on a per-sample basis from data streams. They obtained the accuracy of 97.55%. Xie and Wang [18] developed a diagnosis model based on support vector machine (SVM) and a novel hybrid feature selection method to diagnose erythemato-squamous diseases. They obtained a classification accuracy of 98.61%..

3. MATERIALS AND METHODS

3.1 Feature Selection Algorithms

3.1.1 Information Gain

Information Gain measures the entropy when the feature is given vs. absent to decide how important a given feature is. Let Attr be the set of all attributes and Ex the set of all training examples, value(x,a) with $x \in Ex$ defines the value of a specific example x for attribute $a \in Attr$, H specifies the entropy. The information gain for an attribute $a \in Attr$ is calculated as in equation 1.

$$\text{IG(Ex,a)=H(Ex)-} \sum_{v \in values(a)} \frac{|\{n \in Ex \mid value(x,a)=v\}|}{|Ex|} * H(\{x \in Ex \mid value(x,a)=v\}) \quad (1)$$

For each attribute values if a unique classification exists for the result attribute then the information gain is equal to the total entropy where the relative entropy is 0. Equation 2 defines the gain ratio for a test., where n is the number of examples left in the class after the test on the attribute. Equation 3 gives the Information Gain Ratio Calculation.

$$\text{IV(Ex,a)=-} \sum \frac{|\{n \in Ex\}|}{|Ex|} * \log_2\left(\frac{|\{n \in Ex\}|}{|Ex|}\right) \quad (2)$$

$$\text{IGR(Ex,a)} = \text{IG/IV} \quad (3)$$

3.1.2 Sequential backward floating search

Floating search methods [19-20], sequential forward floating selection (SFFS) and sequential backward floating selection (SBFS), are now considered as standard feature selection tools, providing good performance and close-to-optimum or optimum results in most tasks [21-22]. SFBS starts with the full feature set (of size m). It do the search until the reach of desired dimension d. The time complexity for SFBS is O (m-d). SBFS works best when the optimal feature subset has more number of features, since SBFS spends most of its time visiting large subsets.

3.1.3 Naïve Bayes

Naïve Bayes classifier based on Bayes theorem is a probabilistic classifier.It is called as the independent feature model because it considers that all attributes independently contribute to a certain decision. Hence for each class only the variances of the variables determined and not the entire covariance matrix.Taking into account the nature of the underlying probability model, the Naïve Bayes classifier can be trained very efficiently in a supervised learning setting, working better in many complex real-world situations, especially in the computer-aided diagnosis than one might expect [23-24].

3.1.4 Erythemato squamous data set

Nilsel Ilter M.D.,Ph.D is the owner of this data set obtained from UCI Machine learning depository (http://archive.ics.uci.edu/ml). This database has 34 attributes, 33 of which are linear valued and one of them is nominal. Patients were first evaluated clinically with 12 features. Then skin samples evaluated for 22 histopathological features by analyzing under a microscope. Number of Instances: 366, Number of Attributes: 34. Table 1, 2, 3 gives the information about the class, clinical and histopathological attributes. Most attributes take values 0,1,2,3. The 0 value indicates that the clinical or histopathological feature was not present, values 1 and 2 are intermediate for each feature while value 3 means that the feature was the strongest. In case of the family history feature value 0 means that symptoms were not observed in the family, value 1 otherwise.

Table 1. Class distribution

Class code	Class Name	# of Instances
1	psoriasis	112
2	seboreic dermatitis	61
3	lichen planus	72
4	pityriasis rosea	49
5	chronic dermatitis	52
6	pityriasis rubra pilaris	20

S. No	Attribute name
1	Erythema
2	scaling
3	definite borders
4	itching
5	koebner phenomenon
6	polygonal papules
7	follicular papules
8	oral mucosal involvement
9	knee and elbow involvement
10	scalp involvement
11	family history, (0 or 1)
34	Age (linear)

Table 2. Clinical attributes

Table 3. Histo pathological attributes

S. No	Attribute Name
12	melanin incontinence
13	eosinophils in the infiltrate
14	PNL infiltrate
15	fibrosis of the papillary dermis
16	exocytosis
17	acanthosis
18	hyperkeratosis
19	parakeratosis
20	clubbing of the rete ridges
21	elongation of the rete ridges
22	thinning of the suprapapillary epidermis
23	spongiform pustule
24	munro microabcess
25	focal hypergranulosis
26	disappearance of the granular layer
27	vacuolisation and damage of basal layer
28	spongiosis
29	saw-tooth appearance of retes
30	follicular horn plug
31	perifollicular parakeratosis
32	inflammatory monoluclear inflitrate
33	band-like infiltrate

4. DIAGNOSTIC MODEL

In the diagnostic model first the Information Gain method filters the redundant and irrelevant features which results in reduced feature subset. From this subset SBFS based on Naïve bayes searches for the most relevant features resulting in optimum feature set used for diagnosing the disease. Figure 1 explains the stages in the algorithm.

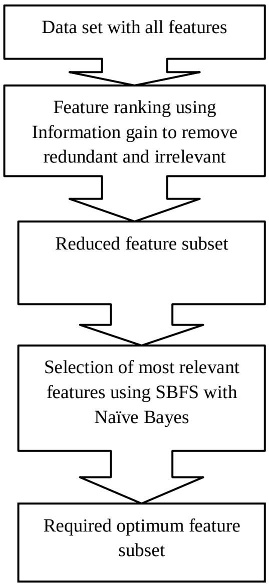
Data set with all features
Feature ranking using Information gain to remove redundant and irrelevant
Reduced feature subset
Selection of most relevant features using SBFS with Naïve Bayes
Required optimum feature subset

Fig 1. Stages in IGSBFS algorithm

4.1 IGSBFS algorithm

This is a greedy algorithm with best first search approach.We rank all the features with Information Gain. After ranking we remove $1/3^{rd}$ features with least gain ratio.With the resultant $2/3^{rd}$ features SBFS with Naïve bayes classifier searches for optimum feature subset. We store the features in a queue in such a way that the features with less gain ratio will be in the front of the queue. Naïve bayes classifier run to get the global optimum. Initially all the features will be in waiting state. When the queue has features in waiting state, a feature in front of the queue is sequentially deleted, kept in a temporary variable and the state of the feature is change to ready state. Naïve bayes classifier run with features in the queue and the accuracy obtained is local optimum. If the local optimum is more than global optimum set the global optimum as the local optimum. If the local optimum is less than the global optimum then backtracking is done i.e. the feature in the temp added to the end of the queue and its state changed to processed state. Otherwise the process continues until the queue has features in waiting state.

4.1.1.1 Algorithm

Input: Data set with all features

Output: Data set with optimum features

Step 1: Rank the features using Information Gain method

Step 2: Remove $1/3^{rd}$ of the features from the feature set which has least score of Gain ratio.

Step 3: Add the remaining $2/3^{rd}$ feature set in a queue in such a way that the front of the queue has features with least Gain ratio.

Step 4: Set the status of all the features in the queue to waiting state.

Step 5: Run the Naïve Bayes classifier and set the global optimum as accuracy obtained.

Step 6: If the queue has features in waiting state goto step 7 else goto step 12.

Step 7: Move the feature in the front of the queue to temp and change its state as ready

Step 8: Run the Naïve Bayes classifier with features in the queue and the accuracy obtained be local optimum.

Step 9: If the local optimum is more than the global optimum set the global optimum as the current local optimum.

Step 10: If the local optimum is greater than or equal to global optimum delete temp and goto step 6.

Step 11: If the local optimum is less than global optimum change the state of the feature in the temp as processed, add the feature at the end of the queue and goto step 6.

Step 12: The features in the queue is the required optimum feature subset necessary for the diagnostic model.

5. RESULTS

Experiments were conducted in WEKA with 10 fold cross validation. Ten fold cross validation has been statistically good enough in evaluating the performance of the classifier [25]. Figure 2 shows the feature ranking using Information gain.

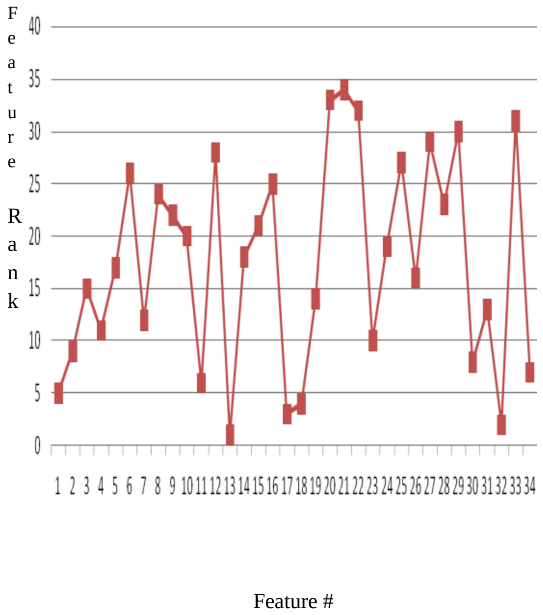

Fig 2. Feature ranking using Information Gain

Algorithm started with 34 features with accuracy of 97.2678%. After ranking with information gain we removed 11 features with least gain ratio namely 13, 32, 17, 18, 1, 11, 34, 30, 2, 23, 4. With the resultant 23 features namely 21, 20, 22, 33, 29, 27, 12, 25, 6, 16, 8, 28, 9, 15, 10, 24, 14, 5, 26, 3, 19, 31 and 7 the accuracy is 98.3607%, Then SBFS method used with Naïve bayes classifier to search for optimum feature subset. This algorithm ended with the accuracy of 98.9071% with 10 features namely 21, 33, 6, 8, 28, 15, 5, 26, 31 and 7 as optimum feature set. The steps carried out in each iteration is shown in Table 4.

Table 4. Steps carried out in each iteration

Iteration	Feature #		Accuracy	Comments
	Removed	Added		
1	7		98.0874	L.O<G.O-Backtrack
2	31	7	98.0874	L.O<G.O-Backtrack
3	19	31	98.3607	L.O=G.O
4	3		98.6339	L.O>G.O-Update G.O
5	26		97.8142	L.O<G.O-Backtrack
6	5	26	94.2623	L.O<G.O-Backtrack
7	14	5	98.9071	L.O>G.O-Update G.O
8	24		98.9071	L.O=G.O
9	10		98.9071	L.O=G.O
10	15		96.1749	L.O<G.O-Backtrack
11	9	15	98.9071	L.O=G.O
12	28		98.6339	L.O<G.O-Backtrack
13	8	28	98.6339	L.O<G.O-Backtrack
14	16	8	98.9071	L.O=G.O
15	6		98.6339	L.O<G.O-Backtrack
16	25	6	98.9071	L.O=G.O
17	12		98.9071	L.O=G.O
18	27		98.9071	L.O=G.O
19	29		98.9071	L.O=G.O
20	33		98.0874	L.O<G.O-Backtrack
21	22	33	98.9071	L.O=G.O
22	20		98.9071	L.O=G.O
23	21		97.2678	L.O<G.O-Backtrack

We used the data mining algorithms such as Decision tree (Simple Cart), RBF Network, SMO RBF Kernel and Poly Kernel to classify the erythemato squamous dataset with all features and with optimum features selected by our hybrid algorithm. Table 5 and 6 shows the results for the classifiers. The empirical comparison shows that the features selected by the hybrid algorithm also improved the accuracy of all the data mining algorithms. Table 7 shows the classification accuracies of our method and other classifiers from literature for the dermatology data set.

Table 5. Accuracy, mean absolute and root mean squared error for various datamining algorithms without feature selection

Algorithm	Accuracy	Mean absolute error	Root mean squared error
Simple Cart	93.9891	0.0267	0.1401
RBF Network	96.1749	0.0132	0.1112
RBF Kernel(c-10, G-0.1)	95.9016	0.2231	0.3117
SMO-Poly Kernel(E-1.0)	95.3552	0.2233	0.3119

Table 6. Accuracy, mean absolute and root mean squared error for various datamining algorithms with 10 features selected by our algorithm

Algorithm	Accuracy	Mean absolute error	Root mean squared error
Simple Cart	95.6284	0.0217	0.1189
RBF Network	97.2678	0.0117	0.0959
RBF Kernel(c-10, G-0.1)	98.0874	0.2228	0.3107
SMO-Poly Kernel(E-1.0)	98.3607	0.2226	0.3103

Table 7. Classification accuracies of our method and other classifiers from literature

Author[Ref]	Year	Method	Accuracy(%)
Guvenir et al. [3]	1998	VFI5	96.2
Guvenir and Emeksiz [4]	2000	Nearest neighbor classifier Naïve Bayesian classifier VFI5	99.2
Bojarczuk [5]	2001	A constrained-syntax genetic programming C4.5	96.64 89.12
Ubeyli and Guler [6]	2005	ANFIS	95.50
Luukka and Leppalampi [7]	2006	Fuzzy similarity-based classification	97.02
Polat and Gunes [8]	2006	Fuzzy weighted pre processing K-NN based weighted pre-processing Decision tree	88.18 97.57 99.00
Nani [9]	2006	LSVM RS B1_5 B1_10 B1_15 B2_5 B2_10 B2_15	97.22 97.22 97.50 98.10 97.22 97.50 97.80 98.30
Luukka [10]	2007	Similarity measure	97.80
Ubeyli[11]	2008	Multiclass SVM with the ECOC	98.32
Polat and Gunes [12]	2009	C4.5 and one-against-all	96.71
Ubeyli [13]	2009	CNN	96.71
Liu et al. [14]	2009	Naive Bayes 1-NN C4.5 PIPPER	96.72 92.18 95.08 92.20
Karabatak and Ince [15]	2009	AR and NN	98.61
Ubeyli and Dogdu [16]	2010	K-mean clustering	94.22
Lekkas and Mikhailov [17]	2010	Evolving fuzzy classification	97.55
Xie and Wang [18]	2011	IFSFS and SVM	98.61
This study (10 features)		IGSBFS with Naïve bayes	98.90

6. CONCLUSION

Feature selection, as a preprocessing step to machine learning especially in real world data, has been very effective in reducing dimensionality, removing irrelevant data, and noise from data to improving result comprehensibility. In this study a hybrid feature selection method based on Naïve bayes proposed for the diagnosis of erythemato squamous disease. It combines the advantages of filters and wrappers. Information gain ratio computed for the features act as evaluation criteria of filters to remove redundant and irrelevant features. SBFS with Naïve bayes acts as wrappers to select the optimum feature subset. Experiment results show that the new hybrid feature selection method obtained optimal classification accuracy with 10 features from a set of 34 features. The optimal feature subset thus obtained then trained with various data mining algorithms such as decision tree (CART), RBF Neural networks, SMO- Poly kernel and RBF kernel to diagnose erythemato squamous diseases. Experimental results show that classification accuracies improved for all data mining algorithms. We conducted the experiments in WEKA with 10 fold cross validation. In this study we used dermatology data set. We use this algorithm in future with various datasets and various combinations of feature selection algorithms will also be considered to select the optimal feature subset.

7. REFERENCES

[1] Y. liu and Y. F. Zheng, (2006) "FS_SFS: A novel feature selection method for support vector machines", *Pattern Recognition*, vol. 39, pp. 1333–1345.

[2] L. Talavera,(2005) "An evaluation of filter and wrapper methods for feature selection in categorical clustering" *Proceedings of 6th international symposium on intelligent data analysis*, Madrid, Spain, pp. 440-445.

[3] H. A. Govenir, G. Demiroz, and N. Ilter,(1998) "Learning differential diagnosis of Eryhemato-Squamous diseases using voting feature intervals", *Artificial Intelligence in Medicine*, vol. 13, pp. 147-165.

[4] H.A. Guvenir and N. Emeksiz,(2000) "An expert system for the differential diagnosis of erythemato-squamous diseases", *Expert Systems with Applications*, vol. 18, pp. 43–49.

[5] Bojarczuk , C. C., Lopes, H. S., Freitas, A. A., ,(2001) "Data Mining with Constrained-Syntax Genetic Programming: Applications in Medical Data Set", *Data Analysis in Medicine and Pharmacology (IDAMAP-2001), a Workshop at Medinfo-2001*, London, UK.

[6] E. D. Ubeyli and I. Guler,(2005) “Automatic detection of erythemato-squamous diseases usingadaptive neuro-fuzzy inference systems”, *Computers in Biology and Medicine*, 35:421-433.

[7] P. Luukka and T. Leppalampi.(2006) “Similarity classifier with generalized mean applied to medical data”, *Computers in Biology and Medicine*, 36:1026-1040.

[8] K. Polat and S. Gunes.(2006) “The effect to diagnostic accuracy of decision tree classifier of fuzzyand k-nn based weighted pre-processing methods to diagnosis of erythemato-squamousdiseases”, *Digital Signal Processing*, 16:922-930.

[9] L. Nanni.(2006) “An ensemble of classifiers for the diagnosis of erythemato-squamous diseases”, *Neurocomputing*, 69:842-845.

[10] P. Luukka,(2007) “Similarity classifier using similarity measure derived from yu's norms in classification of medical data sets”, *Computers in Biology and Medicine*, 37:1133-1140.

[11] E. D. Ubeyli,(2008) “Multiclass support vector machines for diagnosis of erythemato-squamous diseases”, *Expert Systems with Applications*, 35:1733-1740.

[12] K. Polat and S. Gunes,(2009) "A novel hybrid intelligent method based on C4.5 decision tree classifier and one-against-all approach for multi-class classification problems", *Expert Systems with Applications*, vol. 36, no. 2, pp. 1587-1592.

[13] E. D. Ubeyli,(2009) “Combined neural networks for diagnosis of erythemato-squamous diseases”, *Expert Systems with Applications*, 36:5107-5112.

[14] H. W. Liu, J.G. Sun, L. Liu, and H. J. Zhang.(2009) “Feature selection with dynamic mutual information”, *Pattern Recognition*, 42:1330-1339.

[15] M. Karabatak and M. C. Ince,(2009) “A new feature selection method based on association rulesfor diagnosis of erythemato-squamous diseases”, *Expert Systems with Applications*, 36: 12500-12505.

[16] E. D. Ubeyli and E. Dogdu,(2010) "Automatic Detection of Erythemato-Squamous Diseases Using k-Means Clustering", *Journal of Medical Systems*, vol. 34, pp. 179-184.

[17] Stavros Lekkas and Ludmil Mikhailov,(2010) "Evolving fuzzy medical diagnosis of Pima Indians diabetes and of dermatologica diseases," *Artificial Intelligence in Medicine*, vol. 50, pp. 117-126.

[18] J. Xie and Ch. Wang,(2011) "Using support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases", *Expert Systems with Applications*, vol. 38, no. 5, pp. 5809–5815.

[19] P. Pudil , J. Novovicova, J. Kittler, (1994) “Floating Search Methods in Feature Selection”, *Pattern Recognition Letters* 15 1119-1125.

[20] P. Pudil, J. Novovicova, P. Somol,(2003) “Recent Feature Selection Methods in Statistical Pattern Recognition”, *Pattern Recognition and String Matching*, Springer-Verlag, Berlin Heidelberg New York.

[21] A.K. Jain, D. Zongker,(1997) “Feature selection: evaluation, application and small sample performance”, *IEEE Trans*. PAMI 19 153-158.

[22] M. Kudo, J. Sklansky,(2000) “Comparison of algorithms that select features for pattern classifiers”, *Pattern Recognition* 33 25-41

[23] S. Belciug, (2008) “Bayesian classification vs. k-nearest neighbour classification for the non-invasive hepatic cancer detection”, *Proc. 8th International conference on Artificial Intelligence and Digital Communications*.

[24] F. Gorunescu, (2006) *Data Mining: Concepts, models and techniques*, Blue Publishing House, Cluj Napoca.

[25] P. Baldi, S. Brunak, Y. Chauvin, et al. (2000) “Assessing the accuracy of prediction algorithms for classification and overview”, *Bioinformatics*, 5(5):412–424.