Selection of Parameters in Building Fuzzy Decision Trees

Xizhao Wang[1], Minghua Zhao[1], and Dianhui Wang[2]

[1] Machine Learning Center, School of Mathematics and Computer Science
HeBei University, Baoding, China
Phone/Fax: +86-312-5079638
wangxz@mail.hbu.edu.cn
[2] Department of Computer Science and Computer Engineering
La Trobe University, Melbourne, VIC 3083, Australia
Phone: +61-3-94793034, Fax: +61-3-94793060
csdhwang@ieee.org

Abstract. Compared with conventional decision tree techniques, fuzzy decision tree (FDT) inductive learning algorithms are more powerful and practical to handle with ambiguities in classification problems. The resultant rules from FDTs can be used in decision-making with similar nature of our human beings by inference mechanism. A parameter, namely significant level (SL) α, plays an important role in the entire process of building FDTs. It greatly affects the computation of fuzzy entropy and classification results of FDTs, however this important parameter value is usually estimated based on users by domain knowledge, personal experience and requirements. As a result of this, it will be hard to build a high performance FDT without an optimal SL option in practice. This paper aims at developing a method to determine an optimal SL value through analyzing the relationship between the fuzzy entropy and α. The main contribution of this work is to provide some guidelines for selecting the SL parameter α in order to build FDTs with better classification performance. Six data sets from the UCI Machine Learning database are employed in the study. Experimental results and discussions are given.

1 Introduction

Inductive learning [1], an important branch in machine learning field, extracts generic rules from a collection of data. Decision tree based inductive learning algorithm ID3 proposed by Quinlan in 1986 is well known and representative. Decision tree, as an effective method for knowledge representation, has some advantages in descriptions, effectiveness and efficiency. Many algorithms in building DTs are based on assumptions that attribute-values and classification-values are exactly known. This constrains application scopes and is not a suitable manner to meet the request for automated acquisition of uncertain knowledge. In order to overcome this shortcoming,

T.D. Gedeon and L.C.C. Fung (Eds.): AI 2003, LNAI 2903, pp. 282-292, 2003.

M.Umanol and C.Z.Janikow proposed fuzzy-ID3 algorithm [3,4] as a fuzzy version of traditional ID3 algorithm [2]. The significance of use of fuzzy logic in decision tree techniques is that it helps to handle ambiguity in classification problems with a parallel manner to our human's thought and sense.

In building FDTs, each expanded attribute will not classify data in crisp way as the original decision trees perform, attribute values will be allowed to have some overlaps. The entire process of building FDT is based on a significance level α [5]. This parameter plays an important role and controls the degree of such overlaps to some extent, which, however, is usually assigned by domain experts based on their working experience and knowledge base. Therefore, it is difficult to reach the best classification performance. Through analyzing the relationship between α and fuzzy entropy, this paper discusses analytical properties of the fuzzy entropy function associated with the parameter α, and shows that the change trend of fuzzy entropy function is problem dependent. The main contribution of this paper is to develop an experimental method for obtaining the optimal value of α so that it will result in the best FDT in terms of classification performance.

2 FDT Inductive Learning and Definition of the Parameter α

Optimization in DT inductive learning was proven to be a NP problem [6]. Recently, researchers working in this field focus on developing some good heuristic algorithms, where the key is to properly select the attributes in building the DT [7-9]. The concept of entropy is derived from information theory proposed by Shannon [10]. Quinlan's ID3 algorithm was based on entropy to select expanded attribute. The fuzzy-ID3 algorithm was constructed based on fuzzy entropy, a fuzzy extension of the concept of entropy in FDT induction learning. The following sections give some details on algorithm descriptions and relevant parameters.

2.1 FDT Heuristic Algorithm ID3 Based on Fuzzy Entropy

We view the nodes in FDT as fuzzy subset defined on universe *E*, which are usually obtained by ∩ operator with different linguistic values. Let *D* be a considered node.

Definition 1. We use i attributes of A_i to partition node *D*, let $Range(A_i) = \{T_{i1}, T_{i2}, \cdots, T_{ik_i}\}(1 \le i \le m)$, for each $j(1 \le j \le k_i)$, T_{ij} is a fuzzy vector defined on *E*. A partition is shown as below:

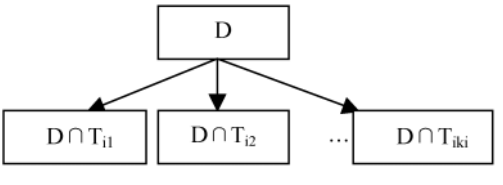

Fig. 1. A fuzzy partition

The fuzzy set $\{T_{i1}, T_{i2}, \cdots, T_{ik_i}\}$ defines a partition of D. Then, partition fuzzy entropy involved by A_i can be calculated as below. For the node D, a crisp partition about class {P, N} forms a fuzzy partition of D , that is, $\{P \cap D, N \cap D\}$. The fuzzy entropy of this partition is defined as:

$$E(D \cap T_{ij}) = -\frac{a}{a+b}\log_2 \frac{a}{a+b} - \frac{b}{a+b}\log_2 \frac{b}{a+b} \quad j = 1,2,\cdots,k_i, \tag{1}$$

where $a = M(P \cap D \cap T_{ij}), b = M(N \cap D \cap T_{ij})$.

Let $m^* = \Sigma_j M(D \cap T_{ij})$. Then, partition fuzzy entropy associated with A_i in node D is measured by:

$$FE(D, A_i) = \sum_{i=1}^{k_i} \frac{M(D \cap T_{ij})}{m^*} E(D \cap T_{ij}) \tag{2}$$

,

where M is the summation operator on all membership of the cases in considered fuzzy set, and

$$f_P(D) = M(D \cap P) / M(D),\ f_N(D) = M(D \cap N) / M(D) \tag{3}$$

denote the relative frequency of cases in node D belonged to P class and N class, respectively. Using these notations, we can describe the fuzzy ID3 algorithm.

Given a leaf criterion$\delta(0<\delta<1)$, consider node D:

***Step 1*:** calculate $f_P(D)$ and $f_N(D)$, if they are larger than the δ, then sign the node as leaf, turn *step 4*.

Step 2: calculate each fuzzy entropy $FE(D, A_j)$ of the attribute, which are not used in father node, select the attribute possessing the minimum as expanded attribute on this node, signed as A_i.

Step 3: generate child node based on the branch of attribute *Ai* in node D, sign D as expanded node.

***Step 4*:** judge if there are other non-leaf nodes that are not expanded:
If yes, consider one of them, turn *step 1*,
Else, stop, export the result, end.

The process of FDT inductive learning can be summarized below:

Step 1: fuzzify the training data using fuzzy cluster techniques based on iterative self-organization algorithm.

Step 2: build up FDT using fuzzy ID3 algorithm based fuzzy entropy mentioned above.

Step 3: translate FDT to a set of rules. Each rule is a path from root to leaf.

Step 4: apply the rules to perform classification task.

2.2 The Introduction of Parameter α

Definition 2. Given a fuzzy set E with membership $E(x)$, select a significance level α, $0<\alpha<1$, the fuzzy subset E_α is defined as

$$E_\alpha = \begin{cases} E(x) & if\ E(x) \geq \alpha \\ 0 & if\ E(x) < \alpha \end{cases} \tag{4}$$

The fuzzy subset E_α is a filter of E, from the view of evidence theory, E is evidence, and while E_α is strong evidence. The significance level α provides a filter to reduce the ambiguity (overlapping) in partitioning. The higher the α takes, the lower the ambiguity will be. However, a high α may lead to the omission of some objects that have no evidence exceed α therefore they do not belong to any subset of the partition. In the process of building FDT, we use fuzzy-ID3 algorithm based on fuzzy entropy described by Definition 1. Before building the FDT, the original data is intercepted by α to reduce the fuzziness of partition.

3 The Effect of α in FDT Inductive Learning

3.1 The Importance of Parameter α

In the process of expanding decision tree by ID3 algorithm, intercepted data will join the calculation of fuzzy entropy and further influence the selection of expanded attribute and the size of generated FDT, and finally the result in terms of rule complexity and classification accuracy. The following flow chart of building the FDT will show the importance of parameter α in FDT:

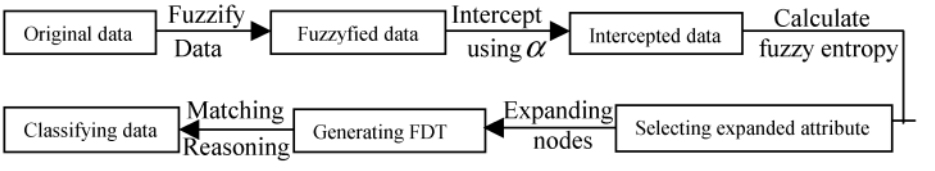

Fig. 2. The process of the influence of α to FDT

The α plays a key role for selection of expanded attribute, generation of decision tree and final classification result. So the relationship between α and fuzzy entropy is needed to explore, which will establish the base for further study on the parameter α.

The dataset *iris* in UCI-machine-learning database is employed as an example. The same cases will belong to different classes with different α. By attribute A4 of *iris*, Table 1 shows the change of class that case 16 belongs to □class of case 16□and the influence of α to the classification result of FDT with fuzzy entropy, rule number and classification accuracy for *iris* database. Here the true level β =0.9.

As α increases, the fuzzy entropy of A4 decreases gradually in the root node, so the information gain of A4 increases, while rule number and the overlapping that each class cover cases decrease, and classification accuracy becomes better. Obviously, different α will lead to different classification performance. Because the significant influence of α to FDT performance, it is essential to optimize the parameter α for obtaining the better results.

3.2 The Analytical Relationship between α and Fuzzy Entropy

From the above discussion, the higher the α takes, the less the classification fuzziness will be. Also, the crisper the classification performs, the fewer the fuzzy entropy of extended attribute is. This leads to stable classification accuracy at the higher level and the stable leaf number. Now, we extend Definition 1 on fuzzy entropy to more than two classes case.

Definition 3. Let U be a set of examples considered, $U=\{u_i \mid i=1,2,\cdots,N\}$, there are k classes $C_1, C_2, \cdots, C_k$. Each attribute of each example is defined as a linguistic value, and each linguistic value, which is the attribute-value of examples, is a fuzzy set defined on U. For simplicity, let A be an attribute-value, $A=(a_1, a_2 \cdots, a_N)$, $a_i (i=1, 2, \cdots, N)$ denotes the degree which the attribute of i example belongs to its attribute-value A. Then the fuzzy entropy of nodes is defined as

$$E(D)=-\sum_{j=1}^{K}\frac{\sum_{i=1}^{N}(C_{ji}\Lambda a_i)}{\sum_{j=1}^{K}\sum_{i=1}^{N}(C_{ji}\Lambda a_i)}\log\frac{\sum_{i=1}^{N}(C_{ji}\Lambda a_i)}{\sum_{j=1}^{K}\sum_{i=1}^{N}(C_{ji}\Lambda a_i)} \tag{5}$$

Table 1. The influence of α to the classification result of FDT

α	Class of case 16	Fuzzy entropy	Rule number	Train accuracy	Test accuracy
0.1	C1 or C2	0.391275	11	0.920026	0.972589
0.2	C1 or C2	0.340673	10	0.933118	0.976523
0.3	C1 or C2	0.226829	10	0.945236	0.977778
0.4	C1	0.157356	9	0.962377	1
0.5	C1	0.121731	3	0.962377	1

Observe the changing trend of fuzzy entropy function with the increase of α from (5) for a simple case, a crisp partition of class{P,N}, (5) becomes

$$E(D) = -\frac{\sum_{i=1}^{N}(P_i \Lambda a_i)}{\sum_{i=1}^{N} a_i} \log \frac{\sum_{i=1}^{N}(P_i \Lambda a_i)}{\sum_{i=1}^{N} a_i} - \frac{\sum_{i=1}^{N}(N_i \Lambda a_i)}{\sum_{i=1}^{N} a_i} \log \frac{\sum_{i=1}^{N}(N_i \Lambda a_i)}{\sum_{i=1}^{N} a_i} \tag{6}$$

The left hand side of (6) is a function with $3N$ dimension, which becomes $3N+1$ dimension while adding the α.

Let attribute I be 3 attribute-value such as $attribute\,1 = \{A, B, C\}$ in which $A = \{a_1, a_2 \cdots, a_N\}$, $a_i \in [0,1]\,(i = 1,2,\cdots, N)$ denotes the degree which the attribute of i example belongs to its attribute-value A, consider 2-class problem such as (0,1), then (6) becomes (7) below:

$$E = f(x) = -x \log x - (1-x)\log(1-x),\ x \in [0,1]. \tag{7}$$

Variable x is also a function of α in (7), the form is

$$x = g(\alpha) = \frac{a_1 \wedge 1 + a_2 \wedge 0 + \ldots + a_N \wedge 1}{a_1 + a_2 + \ldots + a_N}, \tag{8}$$

where a_i is the value of attribute-value A. $E = f(x)$ will increase as x increases in the interval $x \in [0,0.5]$, and will decrease as x increases in the interval $x \in [0.5,1]$. Function $x = g(\alpha)$ is a step-function; the piece number of the ladder is N, which is the number of attribute-value. The monotone character of function (8) is discussed using two instances as follows:

With the increase of α from 0 to 1, if it causes some item of numerator and denominator of (8) decrease at the same time, then x decrease;

If the increase of α cannot influence numerator of (6), for example, α increase at a_2 in (8), and any nonzero item of numerator does not become 0, while the corresponding item of denominator of (8) become 0, then x increase.

Here α does not appear in (8) explicitly, while α can intercept a_i using the method like (4).

4 Parameter Optimization

Firstly, we analyze the change trend of fuzzy entropy with the increase of α. From (8), only if α takes N value as $a_i \in [0,1]\,(i = 1,2,\cdots, N)$, then the fuzzy entropy E in (7) will be influenced. So, the relationship between α and fuzzy entropy involves two functions: $E = f(x)\ x \in [0,1]$, $x = g(\alpha)\ \alpha \in [0,1]$. The fuzzy entropy E becomes E' while taking the value of α as $a_1, a_2, \cdots, a_N$ below:

Step1: based on some idiographic attribute-value $A=\{a_1,a_2,\cdots,a_N\}$ and (8), calculate the value of x and E;

Step2: for (8), if $x\in[0,0.5]$, with the increase of α, there are two probabilities:

A□Some item of numerator and denominator of (8) both become 0, then x decrease to x', then $x=g(\alpha)$ is a step-down function which leads to $x'\in[0,0.5]$, because $E=f(x)$ is a step-up function as $x\in[0,0.5]$, $E=f(g(\alpha))$ is also a step-down function, so we have $E'<E$.

B□The numerator of (8) keeps no change, one item of denominator become 0, then x increases to x', in this case $x=g(\alpha)$ is a step-up function. There are two in stances:

(a) If $x'<0.5$, and that $E=f(x)$ is a step-up function at [0,0.5], then $E=f(g(\alpha))$ is a step-up function, $E'>E$;

(b) If $x'>0.5$ and $x'-0.5>x-0.5$, based on the character of $E=f(x)$,

$E=f(g(\alpha))$ is a step-down function, $E'<E$;

if $x'-0.5<x-0.5$, based on the character of $E=f(x)$,

$E=f(g(\alpha))$ is a step-up function, $E'>E$;

Step3: for (8), if $x\in[0.5,1]$, with the increase of α, there have two probability:

A□The numerator of (8) don't change, one item of denominator become 0, then x increase to x', then $x=g(\alpha)$ is a step-up function which leads to $x'\in[0.5,1]$, because $E=f(x)$ is step-down function when $x'\in[0.5,1]$, $E=f(g(\alpha))$ is step-down function, $E'<E$;

B□Some item of numerator and denominator of (8) both become 0, then x decrease to x', then $x=g(\alpha)$ is a step-down function, here are 2 instances as below:

(a) If $x'>0.5$, and that $E=f(x)$ is a step-down function at [0.5,1], then

$E=f(g(\alpha))$ is a step-up function, $E'>E$;

(b) If $x'<0.5$, and $x'-0.5>x-0.5$, based on the character of $E=f(x)$, $E=f(g(\alpha))$ is step-down function, $E'<E$;

if $x'-0.5<x-0.5$, based on the character of $E=f(x)$, $E=f(g(\alpha))$ is a step-up function, $E'>E$;

In this way, we can know when using some attribute-value as partition node, the change trend of fuzzy entropy of that node will increase along with the increase of α.

Secondly, we give a guideline for obtaining the optimal parameter. A small database car_type is employed to study the change trend of fuzzy entropy.

Table 2. Car_type Dataset

No	Hight(H)	Weight(W)	Length(L)	class
1	2.2	2.8	4	P
2	3.2	4.4	16	N
3	3.8	10	6	P
4	3.0	18	7	N
5	3.0	25	6	N
6	3.8	5.1	6	P
7	3.0	17	14	N
8	3.4	19	13	N

In the algorithm described above, each attribute-value obtain a minimum fuzzy entropy (min_entropy) and the corresponding α (optimal α), at the same time, each attribute obtain a minimum fuzzy entropy (min_entropy_sum) and the corresponding α (optimal_sum α). The obtained results are showed in Table 3.

Table 3. The minimum fuzzy entropy of each attribute-value and its corresponding α

Database name	Attr	Attr-Value	Min_ entropy	Optimal α	Min_entropy _sum	Optimal _sum α
Car_type	A_1	A_{11}	0.000000	0.23	0.539762	0.23
		A_{12}	0.625712	0.49		
	A_2	A_{21}	0.595286	0.40	0.287884	0.40
		A_{22}	0.000000	0.40		
		A_{23}	0.000000	0.40		
	A_3	A_{31}	0.666574	0.12	0.419186	0.12
		A_{32}	0.000000	0.49		

From Table 3, we can see that fuzzy entropy of attribute A_2 is the smallest for car_type dataset, i.e., 0.287884. For minimum fuzzy entropy of the 3 attribute-value of A_2, the fuzzy entropy of A_{22} takes the smallest value, the corresponding parameter α is α=0.40, which can chose as the optimal value of α.

This experiment demonstrates that in the process of building fuzzy decision tree the selected extended attribute at root-node is also A_2 at the moment. Based on this rule, the method of obtaining optimal α is proposed as follows: as a new data comes, the optimal attribute of root-node is firstly tested, then, the attribute with minimum fuzzy entropy is detected. As α takes N values as $a_i \in [0,1]\,(i = 1,2,\cdots,N)$, for each attribute-value of this optimal attribute, we can obtain the value $E^{'}$ of N fuzzy entropy using the above algorithm, each attribute-value select minimum $E^{'}$ and its corresponding parameter α, in which we take the biggest α as the optimal α of database for classification practice. While applying the above algorithm for different databases, it is observed that the change trend of fuzzy entropy varies in different ways. As a result of this, the optimal value of α will be problem dependent.

5 Experimental Results and Discussions

In this study, several datasets downloaded from UCI Machine Learning database are used. Table 4 below gives some statistics of these data sets.

Firstly, let α=0.5. It was found that the attribute take the minimum fuzzy entropy at root node for each database, details are given in Table 5-1 below.

Secondly, let α take N values as $a_i \in [0,1]\,(i = 1,2,\cdots,N)$. Each attribute of these databases gains a set of fuzzy entropy, from which we chose the attributes corresponding to the minimum fuzzy entropy. The smallest one in these minimum fuzzy entropies is denoted by min_attr_entropy, its corresponding attribute and parameter are denoted by min_attr and optimal_cut_sum, respectively. Furthermore, for the value of this attribute (attr_value), we denote its minimum fuzzy entropy as min_attr_value_entropy, and its corresponding parameter α as optimal_cut.

From Table 5-1 and Table 5-2, the attribute having minimum fuzzy entropy is consistent with the selected attribute at root node. Attribute A4 of *iris* is the selected expanded attribute at root node, the optimal α which the 3 attribute-value of A4 corresponding to is 0.27, 0.4, 0.46 respectively, so the biggest value 0.46 is selected which is consistent with the optimal α that attribute A4 corresponds to. From Table 5-2, the optimal parameters α for these databases are: pima: 0.48, liver: 0.46, ecoli: 0.49, wine: 0.49, rice: 0.44, housing: 0.49 respectively.

Now, we study the effect of α on classification results. Constraining $0 < \alpha < 0.5$, we observe the change trend of classification rate for training set, test set and rule number for the databases. The results are depicted below.

It is observed in Figure 3 that the optimal parameter α obtained by the proposed algorithm makes the classification performance best in terms of recognition rates for training set and test set, and rule number as well. For example, as α=0.46 for liver, and α=0.44 for rice, the classification rates reach the highest points, and the rule numbers become the lowest.

Table 4. Statistics of 6 databases

Database name	Attribute number	The number of attribute-value
Pima	8	2
liver	6	3
ecoli	7	2
iris	4	3
wine	13	2
Rice	5	3

Table 5-1. The optimal attributes at root node

Database name	Pima	Liver	Ecoli	Iris	Wine	Rice	Housing
The optimal attribute	A2	A5	A6	A4	A7	A3	A13

Table 5-2. The attribute of minimum fuzzy entropy and the corresponding α

Database name	min_ attr	min_attr _entropy	optimal_ cut_sum	Attr _value	min_attr_ value_entropy	Opti- mal_cut
Pima	A2	0.541514	0.48	A21	0.457574	0.47
				A22	0.666760	0.48
Liver	A5	0.660609	0.46	A51	0.692567	0.00
				A52	0.589134	0.46
				A53	0.543974	0.46
Ecoli	A6	0.903120	0.49	A61	0.897752	0.49
				A62	0.912927	0.49
Iris	A4	0.121731	0.46	A41	0.000000	0.27
				A42	0.248067	0.46
				A43	0.130226	0.40
Wine	A7	0.629253	0.49	A71	0.561929	0.47
				A72	0.672104	0.49
Rice	A4	0.390087	0.44	A41	0.192071	0.43
				A42	0.688028	0.23
				A43	0.199210	0.44
Housing	A13	0.729959	0.49	A131	0.809863	0.46
				A132	0.744533	0.38
				A133	0.340317	0.49

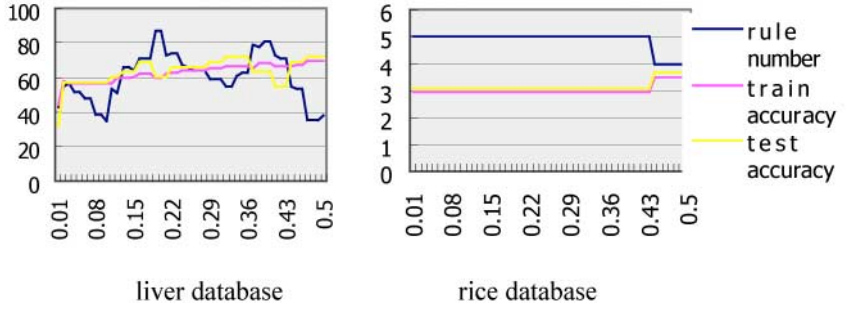

Fig. 3. The value of α vs. classification rates and rule number

6 Conclusion

The significance level (SL) greatly affects the computation of fuzzy entropy and FDT's performance in terms of rule number and classification accuracy. This paper develops a method to determine an optimal SL value through analyzing the relationship between the fuzzy entropy and α. The main contribution of this work is to provide some guidelines for selecting the SL parameter α in order to build FDTs with better classification performance.

Instead of taking one optimal value of α in the whole tree, this paper proposes to use different values of α in different nodes. In this way, a fuzzy decision tree may

correspond to many local optimal α. This idea will further improve the power and quality of FDT, but the computational burden will be increased.

References

[1] R.S.Michalski, J.G.Carbonell and T.M.Mitchell(Eds.). Machine Learning: An Artificial Intelligence Approach. Vol. I,Tioga, Palo Alto, CA, (1983)98-139
[2] Quinlan J.R. Induction of Decision Trees. Machine Learning, 1(1986) 81-106
[3] M.Umano, H.Okamolo, I.Hatono, H.Tamura, F.Kawachi, S.Umezu and J.Kinoshita. "Fuzzy Decision Trees by Fuzzy ID3 Algorithm and Its Application to Diagnosis System", Proceedings of Third IEEE International Conference on Fuzzy Systems, 3(1994) 2113-2118
[4] C.Z.Janikow, "Fuzzy Processing in Decision Trees," Proceeding of the International Symposium on Artificial Intelligence, (1993)360-370
[5] Y.Yuan and M.J.Shaw, Induction of fuzzy decision trees, Fuzzy Sets Syst., 69(1995)125-139
[6] J. R. Hong, A new learning algorithm for decision tree induction (in Chinese), Journal of Computer, 18(1995) 470-474
[7] Breiman L., Friedman J. H., Olshen R.A. and Stone C.J. Classification and Regression Trees. Wadworth International Group (1984)
[8] Smyth P. and Goodman R.M.. Rule Induction Using Information Theory. Knowledge Discovery in Database, MIT Press, (1990)
[9] Press W.H., Teukolsky S.A. et al. Numerical Recipes in C: The Art of Scientific Computing. Cambridge University Press, (1988)
[10] C. E. Shannon, A Mathematical Theory of Communication, Bell. Syst. Teh. Journal, 27(1948) 1-65