Artificial Immune Systems and Missing Features Classification Approach

Štefánia GALLOVÁ, *Member, IAENG,* Sakena Saied Alsadig ALTAHR, Tomáš PENZEŠ

Abstract**—Artificial Immune Systems are one part of the computational intelligence area. They are inspired by the biological immune systems. Learning, inference and prediction in the presence of missing data are important problems in machine learning and data analysis. This paper provides a contribution to some classification problem solving with missing attributes values presence within artificial immune system environment.**

Index Terms**—artificial immune system, missing value, classification, imputation method, antigen, lymphocyte, clonal selection, stimulation threshold**

I. Introduction

This paper provides the contribution to theme how to deal with missing values presence within artificial immune systems environment.

There are several reasons for the data missing existence. Data may be missing because equipment malfunctioned, observations become incomplete due to people becoming ill or observations which are not entered correctly. The data can be missing completely at random - MCAR, but obviously they are not completely missing at random. Then, they may be classifiable as missing at random - MAR
We have the following definitions.

1. MCAR: (*Missing completely at random).* We have a random variable X. The missing data for a random variable X are *missing completely at random* if the probability of having a missing value for X is unrelated to the values of X itself or to any other variables in the set of data.
2. MAR: (*Missing at random).* We have a random variable X. The missing data for a random variable X are *missing at random* if the probability of missing data on X is unrelated to the value of X, after controlling for other random variables in the analysis.
3. NMAR: (*Not missing at random).* When the instance probability with a missing value for an attribute could depend on the value of the attribute.

Generally, we have two approaches to handle with missing data. [10], [15].

- Handling by analogy with deliberately missing data in survey samples, then model the probability of being missing and use probability weighting to estimate complete data summaries.
- Model the missing data distribution, use explicit imputation or maximum likelihood that does implicit imputation.

There are some problems existing with missing data:

- Precision loss due to less data.
- Bias due to data distribution distortion.
- Computational difficulties due to dataset holes.

II. PROBLEM FORMULATION

The selected approach to handle missing data in process modeling is based on class of the expectation maximization techniques. This method was improved with some specific procedures and can be illustrated by following steps:

- An initial set of parameters selection for a model and linearly interpolated values for the missing data substitution.
- Missing data expected values determination.
- Squared errors sum minimization and new model parameters induction

There are some measuring performance dimensions of supervised learning algorithms. We use, first of all, the following three dimensions:

1) *Accuracy.* This dimension contains classification accuracy description concept.

2) *Learning Rate.* This concept contains the speed at which classification accuracy increases during process of training. It is a very useful indicator of the learning algorithm performance, than is accuracy for finite-sized training sets.

3) *Generality.* We have concepts class that is describable by the representation and learnable by the algorithm.

Generally, we know the following approaches for description of missing attributes values.

Štefánia GALLOVÁ, Associate Professor, Pavol Jozef Šafárik University in Košice, Faculty of Science, Institute of Computer Science, phone: + 421 948 158 082, e-mail: stefania.gallova@upjs.sk).
Tomáš PENZEŠ, Faculty of Science, Institute of Computer Science, Pavol Jozef Šafárik University in Košice, 040 00 Slovak Republic, (e-mail: tomas.penzes@upjs.sk)

- *Most Common Attribute Value.* The attribute value that occurs most often is selected to be the value for all the unknown attribute values. This is a simple method. For example, the CN2 algorithm uses this approach.
- *Concept Most Common Attribute Value Method (Maximum relative frequency method, or Maximum conditional probability method-given concept).* This is a restriction approach and does not pay any attention to the relationship between attributes and decision. The attribute value which occurs the most common within the concept is selected to be the value for all the unknown values of the attribute [8], [10]. *C4.5 method.* We consider entropy and the example splitting with missing attribute values to all concepts.
- *All Possible Attribute Values Assigning Method.* We replace an example with missing attribute by a set of new examples, in which the missing attribute value is replaced by all possible values of the attribute. If example contains more that one unknown attribute value, at first – we will do our substitutions for one attribute, then – we will do the substitution for the next attribute, and so on, until all unknown attribute values are replaced by new known attribute values.[9].
- *All Possible Values of the Attribute Restricted to the Given Concept Assigning Method.* This approach is not related to the concept. We have a restriction of the method of assigning all possible values of the attributes to the concept, indicated by an example with a missing attribute value.
- *Ignoring Examples with Unknown Attributes Values Method..* This approach ignores the examples that have at least one unknown attribute value. The next step is using the rest of the table as input to the successive learning process realization.
- *Event Covering Method.* This probabilistic method makes covering or selecting a subset of statistically interdependent events in the outcome space of variable-pairs, disregarding whether or not the variables are statistically interdependent.
- *Missing Attribute Values as Special Values Treating Method.* We don´t try to find some known attribute value as its value. We treat "unknown" itself as a new value for the attributes that contain missing values and treat it in the same way as other values [12], [13] .
- *A Method based on Special LEM2 Algorithm.* This method omits the examples with unknown attribute values when building the block of attribute. We induce a set of rules by using the original LEM2 method.

We create the reasonable sized disjoint subsets by the originally large ratings data dividing. By this way, we evaluate the proposed system on the real world datasets from machine tool diagnostic domain. We apply "the mean absolute error" (MAE) that computes the average of the absolute difference between the true ratings and the realized predictions. Now, we have original rating data and the imputed rating data. Disjoint subsets contain missing rates ranged from 71 % to 99 %.

Implementation of machine learning classifiers creates "pseudo rating matrices" for each judged item. We train a machine learner on content data and set the class label.

We have the machine learner algorithm that is applied within artificial immune system environment as a special procedure framework. There was devised an ensemble classifier by using imputed data from ten classifiers that have the top performance and use the threshold of seven for the majority voting. This strategy has the best results in comparison process with other threshold values in many previous realized experiments. The sense of above mentioned approach is the following. If we have no class value receives at least seven votes from classifiers, it will be left as missing.

The idea of imputation method for missing attributes value in solved three databases is the following. Our goal is to replace missing values with attribute mean. Suppose that used dataset contains some missing attribute values. Suppose "Obrab". as a dataset of R records in which each record contains A attributes. So, we have A x R attribute values in the dataset "Obrab". Missing attribute values (inside that database) may be represented by a non numeric string. Within replacing missing values process we remove all the data rows that are having missing values. This procedure will provide a missing value removed database "obr" with total records "r". Use an attribute mean is one of the different approaches for missing values handling in solved database. This methodology realizes an attribute missing values replacing with the mean value for that attribute in the database. If we have discrete attributes of missing values we realize median value for that attribute in the database.

If we train on content data, the performance of above mentioned approach using machine learning classifiers has usually the following ranking. (for example):

- *significantly better than…*(with 1-tailed t-test, prediction<0.04 or 96 % confidence interval)
- *better than …*(with 0.04 <= prediction < 0.15)
- *shlightly better than (or equivalent with)…*(with 0.15 <= 0.4).

As imputation techniques are often used to handle with missing data, we consider first using some imputation method to fill in solved tables, and then making predictions based on this imputed data, anticipating this may yield more accurate predictions within artificial immune system environment.

III. ARTIFICIAL IMMUNE SYSTEM

A biological immune systems effort is the body protection from foreign molecules named as antigens. The relevant idea was the following. The immune system maintains an idiotypic network of interconnected B cells for antigen recognition. These cells interconnect with each other in some ways that lead to the network stabilization. Two B cells are connected if the affinities they share exceed a certain threshold. The connection strength is directly proportional to the affinity they share. The artificial immune network theory has been introduced by Jerne (1974). The clonal selection methodologies are the whole process of antigen recognition, cell proliferation and differentiation into memory cell (Burnet, 1959). There was proposed a CLONALG algorithm, which is a clonal selection algorithm for learning and optimization (Castro, Zuben, 2002). This algorithm generates a population of N antibodies, each specifying a random solution for the optimization process. During each iteration some of the best existing antibodies are selected, cloned and mutated in order to construct a new candidate population. Later, new antibodies are evaluated and some percentage of the best antibodies is added to the original population.

The model for data clustering and classification, for example, was created as a Tree Structured Artificial immune Network (TSAIN), (Zhang, Yi, 2007). In this model, a topological link is setup between two antibodies immediately after one has reproduced by another with no need to set a threshold for this connection. The foreign antigen creates by means of surface contact. Antibodies on the surface of the B cell react to that complex system as receptors. Competent B cells bind themselves to the antigen, complex on the macrophage´s surface. Stimulated B cells begin to grow in size and divide via mitosis. Then they produce numerous identical clones. Some B cells transform into plasmatic cells producing antibodies. Others B cells become immunological memory calls that enable the organism to respond immediately in case of a future infection by the same pathogen. The basic element conditioning the action of the artificial immune system is the stimulation of the lymphocyte by the antigen. The object classification plays the key role in the understanding process. A classification hierarchy implementation for solved domain database (or knowledge-base) is given below:

The foreign antigen creates by means of surface contact. Antibodies on the surface of the B cell react to that complex system as receptors. Competent B cells bind themselves to the antigen, complex on the macrophage´s surface. Stimulated B cells begin to grow in size and divide via mitosis. Then they produce numerous identical clones. Some B cells transform into plasmatic cells producing antibodies. Others B cells become immunological memory calls, that enable the organism to respond immediately in case of a future infection by the same pathogen. The basic element conditioning the action of the artificial immune system is the stimulation of the lymphocyte by the antigen.

The object classification plays the key role in the understanding process. A classification hierarchy implementation for solved domain database (or knowledge-base) is given below:

```
Object([entity,part, episode, relation])
 Entity([machinery,non/machinery])
  Machinery([machine_tool, cutting_machine])
   Machine_tool([drill,lathe])
    Drill([pneumatic_drill, hydraulic_drill])
     Pneumatic_drill([horizontal, vertical])
………
  Part([component, non_component])
    Component([spindle,…..]
………..
  Episode([state, event, goal])
    State([standard, non_standard,….])
  Relation[(loading, vibration, parts_ connection,…])
```

At first, judged objects are divided into three categories, such as entities, parts and relations. Relations involve causal relationships, temporal relationships, configurations, connections and so on. Episodes are recognized (for example) of four types - standard, non_standard, events, goals. Entities are presented as various types of machinery equipment, divided until the basic concepts are reached.

We have a general rule-group entity:

```
(Rule group number
Number of rules
Initial conditions
Terminating conditions)
()
:gettable-instance-variables
:settable- instance-variables
:inittable- instance-variables)
```

Above-mentioned outlined algorithm is the first step within hierarchical classification process realization, that is implemented as a logical knowledge representation scheme, coded in terms of rules and facts. The started knowledge structure provides information also for running inferences directed by rules. The solved problem continues with concept hierarchical classification in more detail within developed artificial immune systems methodologies. Next, we apply learning methodology approach by artificial immune system principles applying. We have the classification system implementation. In the starting case of random generation, we encounter the problem of high dissimilarity level between the lymphocytes and the antigens. It seems to be necessary to generate the initial population based on the teaching series division into two subseries. The first determines the initial population, the second remains the actual teaching series. This structure allows to avoid the situation when the lymphocytes are alternately stimulated by different-class antigens, that leads to the system constantly getting out of order instead of learning.[1], [17].

IV. PROBLEM SOLVING

Lymphocyte is represented by a chain of 55 numbers in the <0,22>bracket. The numerical equivalent of the antigen class is a judged value. We have stimulation of the domain knowledge of how much time has passed since the last stimulation of the antigen. We have a numerical value that describes the antigens affiliation with a certain class. Solved learning process consists in presenting the system with full sets of antigens *n* times. The similarity level of the lymphocytes system is checked for each presented antigen. The activity value of lymphocytes whose similarity level is below the stimulation threshold is lowered. If we have a state, that a lymphocyte crosses the stimulation threshold, clonal selection takes place. A stimulated lymphocyte activity level is set to maximum. Its clones replace cells with activity level *0* and after this setting mutate. Subsequent lymphocytes are then checked. Finally, at the end of the cycle, the stimulation threshold lowers [1].

We solve the following learning algorithm [2] [3]

```
Read antigen population from file
Generate initial lymphocyte population
Teaching cycle number > 0
    For each antigen
        Get next B cell
        Calculate similarity level
        If similarity level crosses similarity threshold delete a lymphocyte with activity level equal 0,
then replace it with B cell
clone mutate B cell
        If similarity level does not cross similarity threshold,
then decrease lymphocyte activity
decrease stimulation treshold
decrease teaching cycle number by 1
```

Relations for mutation procedure are following [4], [5]:

$$m^{t+1} = m' + (N/M) \,.\, e^{f}_{n}(m) \qquad (1)$$

where m_{t+1} is a new mutate cell of origin parent m_t,
0<N<1 is a random number with normal distribution of probability, M is constant value, which adapts the size of mutation for used application, $f_n(m)$ is normed fitness function with the following equation:

$$fn(m) = \frac{f(m) - \min f(m)}{\max f(m) - \min f(m)} \qquad (2)$$

where $f(m)$ is fitness in maximize case.
Solved artificial immune system algorithm combines local and global searching. This algorithm requires to realize fitness evaluation several times in each generation. To compare with others types of evolutionary techniques, artificial immune system is computationally more complex

V. AN ARTIFICIAL IMMUNE SYSTEM ALGORITHM (AIM)

1) Creation of initial random population of individuals within *Initialisation* stage.
2) The antigens are presented to the antibodies [1]. Antibodies suffer hypermutation in order to better fit antigens or antigens-antibody interactions within *affinity maturation* stage.
3) Those antibodies that are more stimulated are selected to be cloned, and the solved network grows within *clonal expansion* stage.
4) The interaction between antibodies is quantified within *network suppression* stage. If one antibody recognizes another antibody, one of them is removed from the pool of cells - i.e.antibody - antibody interactions. Parameter p_t controls this stage. By this way we determine the size of the suppression radii of the antibodies.

If we realize previous procedures and achieve the final network, we continue with graph theory tool application. The Minimum Spanning Tree tool is built on the resultant antibody network [1]. Inconsistent edges in this network are then identified and removed, thus performing the network data separation into clusters. Within mentioned procedure we apply a *discriminating criterion* that takes into account the relative density of graphically represented objects (points in the space) to determine the edges of the tree to be pruned. The edge of the Minimum Spanning Tree connecting the two

TABLE 1
ACCURACY VERSUS TRESHOLD VALUE INCREASING

Tolerance	Sensitivity	Accuracy	Specificity
0.5	0.901	28.056	0.008
0.6	0.889	28.631	0.009
0.7	0.823	30.562	0.39
0.8	0.752	33.989	0.099
0.9	0.714	40.009	0.135
0.95	0.692	56.550	0.225
1.00	0.606	62.411	0.409

TABLE 2
SOME RELEVANT EXPERIMENT FACTS

Cycle number	Population size	Simulation Threshold	Clonal selection
1	2200	290	12
3	2200	290	12
2	2200	290	12
2	2200	290	11
1	2600	300	11
3	2600	300	12
3	2700	290	12
4	2800	290	12
2	2600	300	12
3	2600	290	11
3	2600	300	11
2	2200	300	14
1	2200	360	13
2	2700	360	13
1	2400	300	12
1	2200	320	11
1	2200	320	11
1	2200	380	11
1	2200	350	11
2	2200	300	14
2	2300	300	17
2	2700	300	17
2	2400	320	16
2	2200	320	15
2	2400	400	15
2	2600	400	24
2	2600	400	24
2	2200	380	33
2	2200	380	33
2	2200	340	35
2	2300	390	35
2	2300	390	41
2	2200	400	41
2	2200	400	43

clusters must be removed from being considered too long when compared to neighbor edges from both sides. Above mentioned algorithm AIM was coded in Matlab.

We have the following *Hierarchical clustering algorithm*

1) Defining an initial value for the parameter p_t, i.e. suppression threshold.
2) Setting up a decaying rate for this parameter, which is: 0<q<1
3) Running the learning algorithm with the set of given parameters.
4) Resultant network generates an offspring network of the tree in the next level. By this way, we gain a new branch of solved tree. Each cluster, which is detected by the artificial immune system algorithm

VI. COMPUTATIONAL EXPERIMENTS - AIM ALGORITHM TESTING

We have achieved some experimental results.

The first experiment was focused on AIM algorithm testing. Data set data was normalized. A percentage of self examples was chosen at random and removed from the data set. After this procedure they became the example that the detectors were trained. This experiment was focused on test tolerance value change effect from sensitivity, accuracy and specificity point of view. The number of training examples was in the range from 100 to 700. The detectors number for the training is from 300 to 1200, and the final detectors number is from 200 to 600. The value of the test tolerance was from 0.5 to 1.0. The results are summarized in table 1. The experiment provides the following result. If threshold value has increased, the results of accuracy were improved. This result is confirmed also in reverse direction. [1], [16], [17].

We experimentally tested the classification effectiveness in relation to set functioning parameters. A higher threshold of stimulation is represented by a lower numerical value, see table 1.

VII. COMPUTATIONAL EXPERIMENTS – CLASSIFICATION EFFICIENCY EXPERIMENT

The second experiment was focused on classification efficiency testing in relation to set functioning parameters. Table 2, 3, 4 illustrate some relevant experiment facts. One interesting fact is that a higher stimulation threshold is represented by a lower numerical value. Generally, for evolutionary algorithms seems to be relevant parameter the *size of the antibody pool.* There is also a clonal selection algorithm performance.The number of clones created in each generation is equal to one hundred, which is the same for all population size. If w1e compare the obtained results to other evolutionary tech1niques and methodologies, we can say the following. Obtained results show a good used algorithm performance.

If we have a clones number within the tested range from 60 to 270, a small size of population of 15 to 25 seems to be an appropriate choice for most graph instances.

TABLE 3
SOME RELEVANT EXPERIMENT FACTS

Cycle number	Lymphocyte activity	Locus number	Treshold decrease
1	44	15	0.018
3	44	15	0.018
2	44	14	0.018
2	44	14	0.017
1	45	15	0.017
3	45	15	0.017
3	44	15	0.017
4	45	16	0.017
2	44	16	0.017
3	45	16	0.017
3	45	16	0.015
2	44	14	0.015
1	44	15	0.017
2	49	15	0.017
1	49	15	0.017
1	49	15	0.017
1	49	14	0.016
1	50	14	0.016
1	50	14	0.017
2	50	15	0.017
2	50	14	0.017
2	55	14	0.019
2	59	12	0.019
2	59	12	0.019
2	59	12	0.019
2	59	12	0.019
2	60	12	0.019
2	60	12	0.019
2	60	12	0.019
2	60	12	0.019
2	81	11	0.019
2	81	11	0.019
2	83	11	0.019
2	83	11	0.019

A size of population, which is large, will slow down the search process. Too small size of population will affect the diversity of selections.

Another relevant judged parameter is the clones number generated each turn. This parameter influences the search speed. Experiments confirm the following result. Regardless of the size of an instance the clones number is best kept at a low level of about 45 to 90 clones, if we have a pool size of 28. Experiences also show that for the sake of simplicity, it is possible to create the same number of clones for all antibodies, regardless of their quality.

Fig. 1 shows missing values percentages versus accuracy results. In table 4 we investigate (for each cycle) system efficiency without missing values SE, system efficiency with missing values (train) SEtrain and system efficiency with missing value (class) SEclass.

TABLE 4
SOME RELEVANT EXPERIMENT FACTS

Cycle number	SE	SEttrain	SEclass
1	88.665	86.989	76..030
3	85.552	91.987	74.999
2	92.956	92.935	75.655
2	92.002	92.001	76.344
1	93.623	91.997	76.922
3	83.998	90.879	76.354
3	91.884	90.698	74.869
4	92.996	90.322	75.699
2	92.898	90.021	77.674
3	93.961	89.902	77.003
3	93.878	88.599	78.006
2	92.001	88.002	77.641
1	92.132	87.566	76.060
2	92.823	86.208	77.499
1	91.996	86.002	77.060
1	91.899	85.690	74.969
1	92.456	85.633	76.098
1	92.934	84.900	77.006
1	93.032	84.697	75.523
2	92.997	84.444	78.909
2	93.010	83.030	79.663
2	93.445	83.612	79.902
2	94.002	82.698	80.699
2	94.023	82.633	81.900
2	93.888	82.002	80.001
2	94.068	80.978	82.412
2	92.036	88.999	82.006
2	91.699	86.698	83.001
2	91.681	89.989	82.523
2	92.597	87.899	81.999
2	93.020	90.004	84.623
2	92.898	89.654	86.004
2	94.987	89.663	85.899
2	94.996	90.098	86.002

VIII. CONCLUSION

This paper contains a contribution to classification problem solving based on the artificial immune systems. The aim of biological immune systems is the body protection from foreign molecules named as antigens. To compare with other types of evolutionary techniques, artificial immune systems are computationally more complex. Solved artificial immune system algorithm combines local and global searching. This algorithm requires realization of the fitness evaluation several times in each generation. Future experiments will be focused first of all on processes with introduction of weights to stimulate threshold.

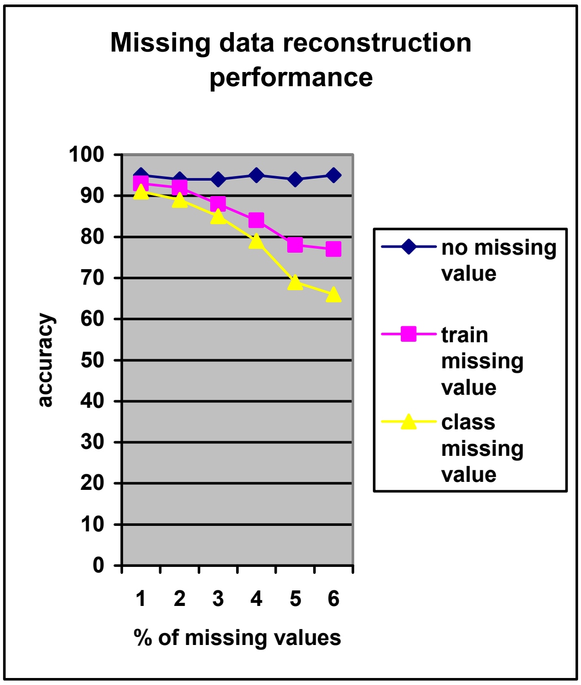

Fig.1. Missing values percentages versus Accuracy.

This paper also illustrates how a missing attribute value presence and handling in used datasets could be effectively realized in combination with an artificial immune system implementation. The choice of the method for missing values handling depends upon the motivation for solving the missing data problem. The imputation method application was a good idea how to solve the problems of this type. The performance of the missing value imputation algorithm was measured with respect to different percentage of missing values in the used data sets. Used imputation method reconstructed the missing values within artificial immune system environment in a better manner than other constant substitution methods that produce poor results. This fact is more evident in these cases when the percentage of missing values presence is increasing.

Nowadays, there are exist other methods of missing data imputation that contain more sophisticated reconstruction approaches for multiple from minimal data, method for handling inapplicable and unknown missing data., robust Bayesian estimators and nonparametric kernel classification rules derived from incomplete or missing data [2], [3],[5], [6].

REFERENCES

[1] S. Gallova.,"Contribution to Incomplete and Noisy information Problem Solving by artificial Intelligence Principles Applying"., Lecture Notice in Engineering and Computer Science, Newswood Limited , London, UK, 2010

[2] G.Gessert, "Handling Missing Data by Using Stored Truth Values.", SIGMOD record, vol.20(3), pp.30-42, 2001

[3] F.Kahl, "Minimal Projective Reconstruction Includin Missing Data.", IEEE Trans. Pattern Anal. Mach. Intell., vol.23(4), pp.418-424, 2001

[4] T. Lumley, "Missing data", A lecture Notes BIOST 570, 2005-11-9, 2005.

[5] E.Pesonen,"Treatment of missing data values in a neural network based decision support system for acute abdominal pain.", Artificial Intelligence in Medicine, vol.13(3), pp.139-146, 1998M.Pawlak,"Kernel classification rules from missing data." IEEE Transactions on Information Theory, pp.979-988, 1993

[6] M. Ramoni, P.Sebastiani.”Robust Learning with Missing Data.” Machine Learning 2001, vol. 45(2), 2001

[7] B. Caruanna, A.Niculescu-Mizil, “Data mining on Metric Space:: An Empirical Analysis of Supervised Learning Performance Criteria” KDD, 319–352, 2004.

[8] P. Clark, T.Niblett,”The CN2 Inductiion Algorithm..” Machine learning vol 3., pp.261-283, 1989.

[9] J.W.Grzymala-Busse,”On the unknown attribute values in learning from examples.”, Proc… of the ISMIS-91, 6th International Symposium on Methodologies for Intelligent systems, Charlotte, North Carolina, October 1991, Lecture Notes in Artificial intelligence, vol.542, Springer-Verlag, Berlin, Heidelberg, New York, pp.368-377, 1991.

[10] I.Knonenko, E.Roskar,”Experiments in automatic learning of medical diagnostic rules.”, Technical Report, J.S.Institute, Ljubljana, 1984.

[11] B. Marlin, “Modeling user rating profiles for collaborative filtering.”, Proceedings of the seventeenth Annual conference on Neural Information Processing systems (NIPS-2003), 2003.

[12] B. Marlin, “Modeling user rating profiles for collaborative filtering.”, Proceedings of the seventeenth Annual conference on Neural Information Processing systems (NIPS-2003), 2003.

[13] J.R.Quinlan,,”C4,5:Programs for Machine Learning.”, Morgan Kaufman Publishers, San Matteo, CA, 1993.

[14] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[15] S.C.Zhang, “Information Enhancement for Data Mining”, IEEE Intelligent Systems, Vol. 19 (2), pp.12-13,2004

[16] G.B.Bezera, T.V.Barra, L.N.Castro, F.J.von Zuben,,”Adaptive Radius immune Algorithm for Data clustering”, Lecture notes in Computer science, vol. 3627, pp.290-303, Springer, august 2005

[17] M. Morkowski, R.Nowicki,”Information theory inpired weighted Immune Classification Algorithm”, ICAISIC, International conference, Zakopane, 2008