Association Rules Algorithms for Data Mining Process Based on Multi Agent System

Imane Belabed, Mohammed Talibi Alaoui, Jaara El Miloud, Abdelmajid Belabed

Association Rules Algorithms for Data Mining Process Based on Multi Agent System

Belabed Imane[1], Talibi Alaoui Mohammed[2], Jaara El Miloud[3] And Belabed Abdelmajid[4]
[1 3 4] University Mohammed The First,Oujda, Morocco
[2] Faculty of Science and Technology,University Mohammed Ben Abdellah,Fez ,Morocco
Belabedimane@gmail.com

Abstract. In this paper, we present a collaborative multi-agent based system for data mining. We have used two data mining model functions, clustering of variables in order to build homogeneous groups of attributes, association rules inside each of these groups and a multi-agent approach to integrate the both data mining techniques. For the association rules extraction, we use both apriori algorithm and genetic algorithm.
The main goal of this paper is the evaluation of the association rules obtained by running apriori and genetic algorithm using quantitative datasets in multi agent environment.
Keywords: Association Rules, Apriori, Clustering, Multi Agent System, Genetic Algorithm

1 Introduction

In recent years, more researchers have been involved in research on both agent technology and data mining. A clear disciplinary effort has been activated toward removing the boundary between them, which form the interaction and integration between agent technology and data mining.

DM (Data Mining) has evolved to become a well-established technology field with subfields such as classification, clustering, and rule mining.

In fact, the clustering encompasses a number of different algorithms and methods for grouping objects of similar kind into respective categories. Such algorithms or methods are concerned with organizing observed data into meaningful structures.

Otherwise, the association rules aims at finding strong relations between attributes.

In order to integrate the two data mining techniques, we used a multi agent system which is the combination of multiple agents. Indeed, an agent is a computer system that is capable of autonomous action on behalf of its user or owner. It is capable to figure out what it is required to be done, rather than just been told what to do.

In this work, we developed a multi agent data mining framework to extract useful rules from real data sets, relying on clustering of variables to build homogeneous groups of attributes and mine supervised association rules inside each cluster.

For the clustering step, we used the K-Means algorithm for variables and for the association rules step we compare the results obtained by using two association rules algorithm, the apriori algorithm and the genetic algorithm knowing that we deal only with quantitative datasets.

2 Related Work

In this section, we will briefly review of the previously proposed studies in autonomous intelligent agent systems or multi-agent systems for data mining and knowledge discovery in database. The techniques used in these studies include association rule mining, associative classification mining, computational intelligence and rule generation algorithms. The proposed approach is mainly related to two areas of research, knowledge extraction from large dataset and knowledge modeling using multi-intelligent agent system. Warkentin, Sugumaran, and Sainsbury produce a study in which they discuss the role of intelligent agents and data mining in electronic partnership management. The procedures of data mining used in this process can be enhanced by using intelligent agents [14]. Nahar, Imam, Tickle, and Chen discussed a paper in which they used association rule mining and a computational intelligence to identify the factors which contributes to heart diseases for males and females. This research presents rule extraction experiments on heart disease data using three rule generation algorithms apriori, predictive apriori and tertius [15]. Ait-Mlouk, Agouti and Gharnati propose an approach to discover a category of relevant association rules based on multi-criteria analysis to avoid redundant rules, they use multi agent system to manage and model the quality measurement according to six agents working in cooperation [16].

3 Data Mining Process

Currently, enormous volumes of data are being produced and stored in computer systems around the world. So, data mining techniques are adequate to address the problem of analyzing and understanding the massive datasets [12].

In this work, we use firstly, a combination of K-Means clustering for variables and supervised association rules i.e. the right part of the rule are always known (the variables to predict) Table 1. Secondly we automate the process by relying on a multi agent system. Through the research we were faced with several limitations such:

- Using K-Means for clustering variables.
- Using association rules algorithm for quantitative datasets.

To deal with the first limitation, we choose to deal only with quantitative datasets and transpose the data in order to cluster the variables instead of individuals.

For the second limitation, we choose to compare two approach of rule mining. The first one concerns the use of apriori algorithm, this after the discretization of all datasets. The second approach is using genetic algorithm for quantitative datasets.

We will apply the proposed system on a real dataset to illustrate how the proposed system can extract a set of rules from real dataset to construct knowledge base.

- Heart datasets: Heart disease.
- Pima datasets: Pima Indians Diabetes Database.
- Vehicle datasets: Use vehicle silhouette to predict the model of a vehicle.

The data used in this work comes from UCI archives, internet.

Table 1. Datasets description

Datasets	Number of variables	Number of lines	Variable to predict
Heart datasets	13	244	Presence of heart disease
Pima datasets	8	692	Presence of diabetes
Vehicle datasets	15	677	The model of vehicle

4 Overview of Basic Techniques Used in the Data Mining Process

4.1 K-Means Clustering

The K-Means algorithm takes two input parameters: the dataset of n objects, and k, the number of clusters to be created. The algorithm partitions the dataset of n objects into k clusters. Cluster similarity is measured by taking the Euclidean distance between objects. In this way, K-Means finds spherical or ball shaped clusters. The mean value of the objects in a cluster can be viewed as the cluster's center of gravity.

Formally, the K-Means clustering algorithm follows the following steps:
Step 1: Choose a number of desired clusters, k.
Step 2: Choose k starting points to be used as initial estimates of the cluster centroids. These are the initial starting values.
Step 3: Examine each point in the dataset and assign it to the cluster whose centroid is nearest to it.
Step 4: When each point is assigned to a cluster, recalculate the new k centroids.
Step 5: Repeat steps 3 and 4 until no point changes its cluster assignment, or until a maximum number of passes through the dataset is performed.

4.2 Discretization Pre-processing

Discretization is a data preprocessing technique which transforms continuous attributes into discrete ones by dividing the continuous values into intervals, or bins. In this work, we based our discretization process on Class-Attribute Interdependence Maximization (CAIM) which is a discretization algorithm of data where the classes

are known**.** In fact the CAIM algorithm works in a greedy top down manner. It starts with a single interval and divides it iteratively, using for the division the boundary that gave the highest values of the CAIM criterion. The algorithm assumes that every discretized attribute needs at least number of intervals equal to the number of classes.

Let us assume that we have a training data set consisting of M examples, and that each example belongs to only one of the S classes. F will indicates any of the continuous attributes. Then there exists a discretization scheme D on F, which discretizes the continuous domain of attribute F into n discrete intervals bounded by the pairs of numbers: [1].

D : $\{[d_0, d_1], (d_1, d_2], \dots (d_{n-1}, d_n]$

where d_0 is the minimal value d_n is the maximal value of attribute *F*, and the values are arranged in the ascending order. These values constitute the boundary set $\{d_1, d_2, d_3, \dots d_{n-1}, d_n\}$ for discretization *D*.

- Caim creterion

Given the quanta matrix defined in Fig.1, the Class-Attribute Interdependency Maximization (CAIM) criterion that measures the dependency between the class variable C and the discretization variable D for attribute F is defined as:

$$\text{CAIM}(C, D|F) = \sum_{i=1}^{n} \frac{\frac{max_i^2}{M_{ir}}}{n} \quad (1)$$

Where:

n is the number of interval
i iterates through all intervals, i.e. i=1,2,..n.
max_i is the maximum value among all q_{ir} values (maximum value within the i^{th} column of the quanta matrix), r=1,2,…S (see Fig.1).
M_{ir} is the total number of continuous values of attribute F that are within the interval $[d_r, d_{r-1}]$.

Class	Interval					Class Total
	$[d_0, d_1]$	…	$(d_{r-1}, d_r]$	…	$(d_{n-1}, d_n]$	
C_1	q_{11}	…	q_{1r}	…	q_{1n}	M_{1+}
:	:	…	:	…	:	:
C_i	q_{i1}	…	q_{ir}	…	q_{in}	M_{i+}
:	:	…	:	…	:	:
C_S	q_{S1}	…	q_{Sr}	…	q_{Sn}	M_{S+}
Interval Total	M_{+1}	…	M_{+r}	…	M_{+n}	M

Fig. 1. Quanta matrix

4.3 Association Rules

The most popular task of DM is to find trends in data that show associations between domain elements. This is generally focused on transactional data such as a database of purchases at a store. This task is known as Association Rule Mining (ARM). It was first introduced in Agrawal et al. [2]. Association rules identify collections of data attributes that are statistically related in the underlying data. An association rule is of the form $X \rightarrow Y$ where X and Y are disjoint conjunctions of attribute value pairs. The confidence of the rule is the conditional probability of Y given X, $Pr(Y|X)$, and the support of the rule is the prior probability of X and Y, $Pr(X \cap Y)$. Here probability is taken to be the observed frequency in the dataset.

The traditional ARM problem can be described as follows. Given a database of transactions, a minimal confidence threshold and a minimal support threshold, find all association rules whose confidence and support are above the corresponding thresholds.

Apriori Algorithm. The apriori algorithm iteratively identifies frequent intemsets FIs, in data by employing the "closure property" of itemsets in the generation of candidate itemsets, where a candidate (possibly frequent) itemset is confirmed as frequent only when all its subsets are identified as frequent in the previous pass. The closure property of itemsets can be described as follows: if an itemset is frequent then all its subsets will also be frequent; conversely if an itemset is infrequent then all its supersets will also be infrequent.

- Apriori

```
Input: (a) A transactional database Dt;
(b) A support threshold s;
Output: A set of frequent itemsets S;
1: begin:
2: k ←1;
3: S←  an empty set for holding the identified frequent
itemsets;
4: generate all candidate 1-itemsets from Dt;
5: while (candidate k-itemsets exist) do
6: determine support for candidate k-itemsets from Dt;
7: add frequent k-itemsets into S;
8: remove all candidate k-itemsets that are not suffi-
ciently supported to give frequent k-itemsets;
9: generate candidate (k + 1)-itemsets from frequent k-
itemsets using closure property;
10: k←k + 1;
11: end while
12: return (S);
13: end Algorithm
```

Note: A *k*-itemset represents a set of k items.

Genetic Algorithm. Genetic algorithms are stochastic search methods that mimic the metaphor of natural biological evolution [4].At each generation, a new set of approximations is created by the process of selecting individuals according to their level of fitness in the problem domain and breading them together using operators borrowed from natural genetics. Genetic algorithm take as an input the following elements : population , selection according to fitness, crossover to produce new offspring, and random mutation of new offspring.

- Initial population : The initial population of individuals is generated as follows: in the first individuals, the intervals $[l_i,u_i]$ represent the whole domain of the i^{th} numeric attribute, and the following individuals encode intervals with decreasing amplitudes (length of intervals) until they reach a minimum support in the dataset. Once the amplitudes are fixed for an individual, the bounds l_i and u_i are chosen at random.
- Mutation and crossover: The crossover operator consists in taking two individuals, called parents, at random and generating new individuals: Each attribute the interval is either inherited from one of the parents or formed by mixing the bounds of the two parents [5]. Mutation works on a single individual and increases or decreases the lower or upper bound of its intervals respectively. Moving interval bounds is done so as to discard/involve no more than 10% of tuples already covered by the interval [6].
- Fitness function: The fitness function used is based on the gain measure [7]. If the gain is positive (the confidence of the rule exceeds the minimum confidence threshold), we take into account the proportions of the intervals (defined as the ratios between the amplitudes and the domains). Moreover, rules with low supports are penalized by decreasing drastically their fitness values by the number of tuples in the database [8].

Algorithm. The algorithm starts with a set of rule templates and then looks dynamically for the "best" intervals for the numeric attributes present in these templates. An optimization criterion based on both support and confidence is used to keep only high quality and interesting rules [10]. The algorithm follows a prototypical genetic algorithm scheme. The inputs are the minimum support (MinSupp), the minimum confidence (MinConf), the population size (PopSize), the number of generations (GenNum), the fraction of population to be replaced by crossover (Cross) and the mutation rate (MutR).

```
Input: A dataset composed of NbTuples, PopSize, GenNb,
CR, MR, MinSupp, MinConf
Output: Quantitative association rules R
Select a set of attributes
```

```
Let Rt a set of rule templates defined on these attrib-
utes
foreach r ∈ Rt do
Generate a random population P of PopSize
while i ≤ GenNum do
Form the next generation of population by mutation and
crossover w.r.t. MutR and Cross.
Extract the itemsets that satisfy the best fitness to
constitute the association rule values
i++
Return R= max (fitness (r)); r belongs to P
```

4.4 Agent and Multi Agent System

Agents and multi-agent systems are an emergent technology that is expected to have a significant impact in realizing the vision of a global and informational rich services network to support dynamic decision making [3].

Agents. Agents are defined by Wooldridge [9] as computer systems that are situated in some environment and are capable of autonomous action in this environment in order to meet their design objectives.

Multi Agent Systems. By combining multiple agents in one system to solve a problem, the resultant system is a multi-agent system (MAS). These systems are comprised of agents that individually solve problems that are simpler than the overall system problem. They can communicate with each other and assist each other in achieving larger and more complex goals [13].

5 Multi Agent Framework for Data Mining Process

The proposed mining framework comprises four categories of agent Fig. 3:

- User agent: User agent is charged by the communication with the user interface.
- Coordinator agent: Coordinator Agent is focused on the correct message transmission among the agents. It takes the requirements (data, number of clusters...) and sends them to the corresponding agent.
- Data agent: Data Agent is in charge of a data source; it interacts and allows data access. There is one data agent per data source.
- Clustering agent: Clustering agent is concerned with a clustering K-Means algorithm.
- Association rules Agent: Association rules agent is in charge of extracting supervised rules though the genetic algorithm or apriori algorithm inside each cluster.

The sequence of operation between different agents constituted the system given in Fig.3. This diagram shows the sequence of operations during the execution of the proposed multi-agent system. However in this work, we will focus on the execution of association rules agent. Indeed there are two scenarios for the quantitative datasets. The first is to execute the apriori algorithm and this includes the use of the discretization preprocessing. The second is the use of the genetic algorithm Fig.2.

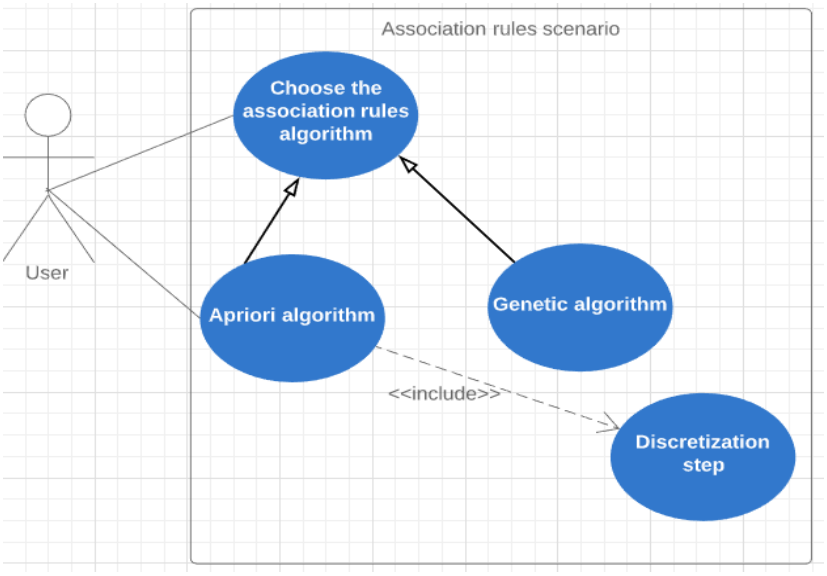

Fig. 2. Association rules use case.

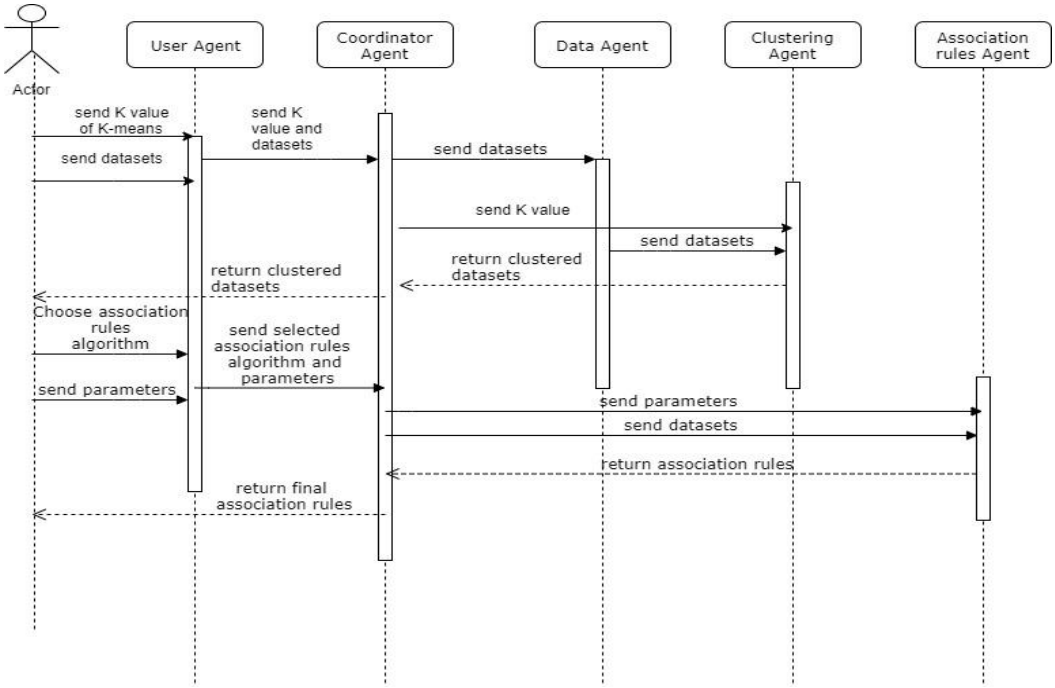

Fig.3. The sequence diagram of the proposed agent system.

5.1 Overview of the Implementation of the Data Mining Process

The implementation of the Multi-Agent System for centralized Data Mining framework was done by using java platform through Agent-Oriented Programming paradigm (AOP). In order to allow inter-agents communication, agents must share the same language, vocabulary and protocols so; we have followed the recommendations of the standard Foundation for Intelligent, Physical Agents (FIPA).

We have developed our proposed framework with Java Agent Development (JADE) [11] which is FIPA-compliant middleware that enables the development of applications based on the agent paradigm and is adequate to process large amounts of data with a data mining approach.

6 Results

In this section, we will illustrate the results obtained by running our data mining process on the three datasets; heart, pima and vehicle datasets.

As perquisite, for clustering part, we choose to fix the K value of K-Means into 3. Also, for association rules part, we fix the confidence threshold at 60%, the support threshold at 10%; particularly for genetic algorithm we fix the population size at 250, the crossover rate at 50%, generation number at 100 and the mutation rate at 40%.

1. Heart datasets :

Table 2. . Result of heart datasets.

Association rules agent				
Heart datasets	Number of rules with apriori	Confidence	Support	Execution time
Cluster 1	1	70%	35%	3.27E-4 seconds
Cluster 2	101	> 62%	> 10%	
Cluster 3	10	> 63%	> 10%	
Total of rules	112	-	-	
Heart datasets	Number of rules with genetic	Confidence	Support	Execution time
Cluster 1	1	72%	27%	2.46E-4 seconds
Cluster 2	45	> 63%	> 10%	
Cluster 3	6	> 65%	> 10%	
Total of rules	52	-	-	-

For the heart datasets, we find that, first; the association rules agent with apriori algorithm generates more rules than with genetic algorithm. Second we notice that there is a small difference in the execution time between the two algorithms.

However, in the case where the number of the rules generated is the same (cluster 1), we find that the confidence of the rule with genetic algorithm is higher than the one with apriori algorithm.

In this part and the next part, we specify that the execution time exclude the execution time of the dicretization in the case of apriori algorithm.

2. Pima datasets

Table 3. Result of Pima datasets.

Association rules Agent				
Pima datasets	Number of rules with apriori	Confidence	Support	Execution time
Cluster 1	6	> 65%	> 39%	2.79E-4 seconds
Cluster 2	2	>71%	> 16%	
Cluster 3	10	> 66%	> 45%	
Total of rules	18			
Pima datasets	Number of rules with genetic	Confidence	Support	Execution time
Cluster 1	10	> 60%	> 10%	2.62E-4 seconds
Cluster 2	2	> 81%	> 11%	
Cluster 3	14	> 74%	> 10%	
Total of rules	26			

The results of the pima datasets are the opposite of the results of heart datasets. In addition of the high confidence compared to apriori algorithm, the genetic algorithm extracts more rules than the apriori algorithm. Also, in the case where the rules extracted are the same (cluster 2), we notice that the confidence of rules with genetic algorithm is higher compared to apriori algorithm. Moreover, the execution time of genetic algorithm is lower than with apriori algorithm.

3. Vehicle datasets

Table 4. Result of vehicle datasets.

Association rules Agent				
Vehicle data	Number of rules with apriori	Confidence	Support	Execution time
Cluster 1	31	> 61%	> 10%	3.65E-4 seconds
Cluster 2	32	> 60%	> 10%	
Cluster 3	53	> 60%	> 11%	
Total of rules	116			
Vehicle data	Number of rules with genetic	Confidence	Support	Execution time
Cluster 1	25	> 79%	> 10%	0.003 seconds
Cluster 2	26	> 60%	> 10%	
Cluster 3	107	> 60%	> 10%	
Total of rules	158			

For the vehicle datasets, we conclude that genetic algorithm generates more rules than apriori algorithm, this with high confidence compared to apriori. However in this case the execution time of genetic algorithm is 10 times more than apriori algorithm.

From the results presented in this section, we can conclude that genetic algorithm is more performing than the apriori algorithm. In the majority of cases this is due to the quality of the discretization phase.

In fact, some rules with apriori algorithm are redundant, does not brings new information, this is due to the dicretization intervals.

Example:

Support = 42 (17%), confidence = 60 %: MaxHeartRate = [71.0-147.5] and SerumCholestoral = [126.0-272.0] --> class = presence of disease.

Support = 26 (10%), confidence = 86 %: MaxHeartRate = [71.0-147.5] and SerumCholestoral = [272.0-417.0] --> class = presence of disease

On the other hand, in the three datasets, we find that genetic algorithm brings new rules that involve more attributes.

As a conclusion, we find that genetic algorithm is more adequate taking into consideration to use of multi agent system. Firstly, it generates more significant rules;

secondly it avoids the discretization step which means that the gain of execution time in whole process including clustering and association rules will be considerable.

7 Discussion

The proposed approach illustrated in this work is more efficient compared to previous works such Ait-Mlouk (2016) [16]. In fact our approach deals with redundant rules by using genetic algorithm instead of multi criteria approach. That allows decreasing the number of agent in the association rules step. We have one association rules agent rather than six agents in that work.

Also our proposed approach surpasses that of Nahar [15], mainly in the execution time knowing that the both works processes heart datasets with the same number of variables.

The developed framework in this paper is presented with three real test cases from different domains such health and industry; however the framework is applicable to any other datasets. The use of multi agent system allows us to take advantage of four features: reactivity, autonomy, interaction and initiative. This makes our work extending to distributed and parallel paradigm.

8 Conclusion

We have proposed a new approach based on a Multi Agent framework for data mining process that includes genetic algorithm for extracting association rules, JADE frameworks and five different types of agents (user agent, clustering agent, association rule agent, data agent and coordinator agent).

In this work we focus on the association rules step, because we propose two scenarios for extracting rules. The experimental results proved that the extraction based on genetic algorithm is more adequate. In addition to the quality of rules extracted, the execution time of genetic algorithm is interesting because it avoids the discretization step.

Also, if we take into consideration the integration of the algorithm in data mining process using multi agent system and other algorithm that increase the execution time, such clustering, we conclude that the use of genetic algorithm is an optimization of the whole process.

As a perspective, we want to extend our approach to deal with a real time data sets from agriculture field in order to extract rules based on real time weather, ground composition. This, in order to improve the agricultural yields.

References

1. Ching J.Y., Wong A.K.C. & Chan K.C.C.:Class-Dependent Discretization for Inductive Learning from Continuous and Mixed Mode Data, IEEE Transactions on Pattern Analysis and Machine. (1995)
2. Author, R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 1993.
3. T. Marwala and E. Hurwitz. Multi-Agent Modeling using intelligent agents in a game of Lerpa. eprint arXiv:0706.0280, 2007.
4. Pei M., Goodman E.D., Punch F.(2000) Feature Extraction using genetic algorithm, Case Center for Computer-Aided Engineering and Manufacturing W. Department of Computer Science.
5. Guo, H., Zhou, Y.: An Algorithm for Mining Association Rules Based on Improved Genetic Algorithm and its Application. In: 3rd International Conference on Genetic and Evolutionary Computing, WGEC 2009, pp. 117–120 (2009).
6. Gonzales, E., Mabu, S., Taboada, K., Shimada, K., Hirasawa, K.: Mining Multi-class Datasets using Genetic Relation Algorithm for Rule Reduction. In: IEEE Congress on Evolutionary Computation, CEC 2009, pp. 3249–3255 (2009).
7. Tang, H., Lu, J.: Hybrid Algorithm Combined Genetic Algorithm with Information Entropy for Data Mining. In: 2nd IEEE Conference on Industrial Electronics and Applications, pp. 753–757 (2007).
8. Dou, W., Hu, J., Hirasawa, K., Wu, G.: Quick Response Data Mining Model using Genetic Algorithm. In: SICE Annual Conference, pp. 1214–1219 (2008).
9. M. Wooldridge. An Introduction to Multi-Agent Systems. John Wiley and Sons (Chichester, England), 2003.
10. Pei M., Goodman E.D., Punch F.(2000) Feature Extraction using genetic algorithm, Case Center for Computer-Aided Engineering and Manufacturing W. Department of Computer Science.
11. F. Bellifemine, A. Poggi, and G. Rimassi. JADE: A FIPA-Compliant agent framework. Proceedings Practical Applications of Intelligent Agents and Multi-Agents, 1999. http://www.jade.tilab.com.
12. J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufman Publi hers, San Francisco, CA, (Second Edition), 2006.
13. Popa, H., Pop, D., Negru, V., Zaharie, D.: AgentDiscover: A Multi-Agent System for Knowledge Discovery from Databases. In: Ninth International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, pp. 275–281. IEEE, Timisoara (2008).
14. Warkentin, M., Sugumaran, V., Sainsbury, R.: The role of intelligent agents and data mining in electronic partnership management. Expert Systems with Applications 39(18), 13277–13288 (2012).
15. Nahar, J., Imam, T., Tickle, K., Chen, Y.: Association rule mining to detect factors which contribute to heart disease in males and females. Expert Systems with Applications 40(4), 1086–1093 (2013).
16. Ait-Mlouk A., Agouti T., Gharnati F.: Multi-agent-based modeling for extracting relevant association rules using a multi-criteria analysis approach: Vietnam J Comput Sci (2016) 3:235–245 (2016).