International Journal of Scientific Engineering and Technology
Volume No.5 Issue No.5, pp: 319-322

ISSN:2277-1581
01 May 2016

Feature Selection in Data Mining using Chemical Reaction Optimization Algorithm

E. Karam Kiani[1], M. Sadeghzadeh[2]
[1]Departmant of computer, Ardabil Science and Research Branch, Islamic Azad University, Ardabil, Iran
Departmant of computer, Ardabil Branch, Islamic Azad University, Ardabil, Iran
[2]Scientific mission of Islamic Azad University, Mahshahr branch,Mahshahr,Iran
elhamkiani2012@yahoo.com, Sadegh_1999@yahoo.com

Abstract: One of the characteristics of recent problems can be referred to the great number of features that have led to slowing down the classification systems, decreased efficiency and rising the costs of such systems. In recent years, feature selection problem has been investigated in data mining field when encountering data sets with many features. This study aimed to present new and optimized application in order to use metaheuristic algorithms in the feature selection problem in data sets in which through large number of useless features can decrease these features and their required time to implement various algorithms. In this study, a metaheuristic algorithm called Chemical Reaction Optimization Algorithm was used that is among the modern and most powerful evolutionary optimization techniques introduced in 2010. At the end, the proposed method was analyzed along with available standard data sets of UCI. The results indicated high efficiency based on two criteria of classification accuracy and small subset selection of features as the salient features at the same time.

Keywords: **data mining, feature selection, chemical reactions**

Introduction

Data mining is the process of pattern discovering in data that should be automated or semi-automated. Nowadays, with the expansion of databases and huge amount of data stored in these systems, there is a need for a tool that can process these stored data and convert them into a series of useful information [1].
Progresses in data collection and storage capabilities in recent years have provided huge amounts of information in various sciences. Data sets that have many dimensions, despite creating many opportunities, present computational challenges. One of the problems with high dimensionality data is that most often, all data features are not crucial to find the knowledge that lies in data. Therefore, data dimension reduction has remained as one of the significant discussions in many fields [2]. Data dimension reduction methods are divided into feature extraction-based methods and feature selection-based methods. This study has concentrated on feature selection-based methods. These methods try to decrease data dimensions through selecting a subset of initial features. Sometimes, data analyses such as classification on the reduced space perform better compared to the main space. Various methods of feature selection try to choose the best subset among the two candidate subsets. In all of these methods, based on application and definition, the subset that is capable to evaluate the value of a function will be chosen as the solution. Despite each method tries to choose the best feature, regarding the extent of possible answers and that the answer sets increase by N, finding the optimal solution is difficult and is costly in medium and large Ns [3]. To solve this problem, it has been referred to one of the intelligent methods called Chemical Reaction Optimization (CRO). This intelligent method has been discussed in this paper [4].

Introducing Chemical Reaction Optimization Algorithm

Chemical Reaction Optimization (CRO) is a technique that solves optimization problems using the concepts of chemical processes [3]. In this section, it is not possible to explain all the available hypotheses in chemical process; therefore, the general concepts of these processes are addressed.
In CRO technique, four initial reactions have been defined regarding molecular behavior.

- **On-Wall Ineffective Collision**

On-Wall Ineffective Collision indicates a condition in which the molecule collides with the wall of the area where the molecule is located and returns, but still stays mono-molecular and does not decompose.

- **Decomposition**

Decomposition describes a situation in which the molecule collides with the wall and divides into several parts. Suppose that ω is divided into ω'_1 and ω'_2. Every mechanism that is able to produce ω'_1 and ω'_2 is allowed to indicate this regarding having more chance.

- **Synthesis**

Synthesis is against decomposition. This happens when the molecules collide with each other in opposite direction and combine. If only one molecule is created, we have $\omega 1 + \omega 2 \rightarrow \omega'$.

- **Inter-Molecular Ineffective Collision**

This collision happens when several molecules collide with each other (two molecules are supposed) and return to each side. The molecules stay stable before and after collision.

$\omega 1 + \omega 2 \rightarrow \omega' 1 + \omega' 2$

The first two cases belong to the mono-molecular collisions and the last two cases belong to the inter-molecular collisions. Decomposition and synthesis cause more changes in molecular structure. These classifications are shown in Figure (1).

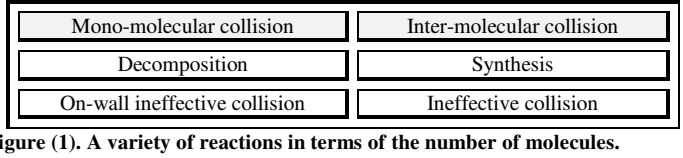

Figure (1). A variety of reactions in terms of the number of molecules.

International Journal of Scientific Engineering and Technology
Volume No.5 Issue No.5, pp: 319-322
ISSN:2277-1581
01 May 2016

Optimization means to study problems in order to minimize or maximize a particular parameter. This parameter is calculated through a function called "the objective function" that works on dummy variables of selection set. It is possible to consider various applications on molecules in CRO equal to optimization. The step by step similarity of this technique to optimization problems as well as its purpose to bringing the matter to the minimum energy have made this possible to use CRO in order to solve optimization problems.

CRO is a population based metaheuristic method. Regarding the hypotheses related to energy consumption and its conversion through initial reactions, it is possible to manipulate optimization solutions in order to obtain optimal solutions.

In this section, different parts of CRO are explained and the way to convert them into a real algorithm is described.

In the following, we will explain these parameters.

1. Molecular structure: is used to indicate solution in optimization problems and does not have a clear structure. It can be a number, curve or matrix regarding the problem.
2. Potential energy: is considered as the objective function in optimization problem.
3. Kinetic energy: is a non-negative value that indicates the worse solution in the case of withdrawal.
4. Number of collisions: when the molecules collide with each other, one of the four initial reactions happens and the change of molecule structure is possible. This parameter indicates the number of molecular collisions.
5. Minimum structure: is a molecular structure that has the minimum potential energy.
6. Minimum potential energy: potential energy is the minimum structure of molecule.
7. Minimum collisions: the number of molecule collisions to reach the minimum structure.

The Conversion of Chemical Collision Algorithm to use in Feature Selection

As mentioned, feature selection problems include a number of features that are selected in order to decrease subset calculations. This subset should be the best possible subset in terms of efficiency compared to the initial subset. Therefore, the following components from feature selection problem should be written based on CRO.

In order to describe algorithm, a general problem is supposed. This example consists of a, b, c, and d. This example will be expanded to various components.

- **Molecular Structure**

In previous section, the components of CRO were described. The most important conversion regarding problem solving was feature selection by CRO. The molecular structure expresses problem solution. In the present study, binary structure was used for implementation. Binary structure means that each equals the length of features and this length is 4 for this example. Selecting or deselecting features are simulated with 0 and 1. In the following, the converted structure of a solution to CRO can be observed. One of the available solutions is selecting features (a) and (b) as the result. The output structure can be observed in Figure (2).

W	a	B	c	D
1001	1	0	0	1

Figure (2). Sample molecule structure

1001 structure presented in Figure (2) means that (a) and (d) are selected and (b) and (c) are not.

In order to show the solution, it is possible to use other structures such as decimal structure, but this method has disadvantages such as complexity, lack of resolution for features and length of variable.

- **Calculating Neighborhood Molecules after Collisions**

After each collision, based on the behavior of algorithm, it is possible to have one or two new molecules. The new molecules should have a new structure that is created by neighborhood generation methods. Neighborhood methods are calculated for each kind of collisions, separately.

On-Wall Ineffective Collision

In this method, two neighborhoods have been used that half uses the first type neighborhood and the other half uses the second type neighborhood.

First type neighborhood: moves two features that have been selected randomly. As an example, suppose four features for a hypothetical problem. One of the solutions is the W that can be seen in the following figure. In order to obtain the neighborhood, first, 2 features are selected. Suppose (a) and (c) are selected. These two are moved.

W				W'			
0	1	1	0	0	0	1	1
d	c	B	a	D	c	B	A

Figure (3). First type neighborhood

Second type neighborhood: one of the features is selected randomly and its value will be determined. Suppose that in initial W structure, the variable (c) is selected that follows the following figure.

W				W'			
0	1	1	0	0	0	1	0
d	C	b	A	D	c	B	A

Figure (4). The structure of second type neighborhood

Decomposition

Neighborhood: in this method, two molecules are decomposed. In this manner, two new molecules are created from odd and even features of initial molecules and the left features are completed randomly.

International Journal of Scientific Engineering and Technology
Volume No.5 Issue No.5, pp: 319-322
ISSN:2277-1581
01 May 2016

W				W'			
d	c	B	a	D	R	b	R
				0	0	1	1
0	1	1	0				
				1	1	1	0
				R	c	R	a

Figure (5). Neighborhood structure of decomposition

Synthesis

In this manner, two molecules are taken and a new molecule is created. One random number is considered regarding neighborhood and synthesizes from the beginning of molecular structure until its number from the first molecule and synthesizes the rest from the second molecule. According to the initial example, suppose the random number as 2 (R=2).

W1							
d1	c1	b1	a1				
0	1	1	0	W'			
				d2	c2	b1	a1
W2				1	1	1	0
d2	c2	b2	a2				
1	1	0	1				

Figure (6). Neighborhood structure of synthesis

Ineffective Collision of Molecules

In this method, two molecules collide with each other and create two new molecules. Regarding neighborhood, each of the molecules were sent to the neighborhood function separately and two new molecules are calculated through selecting a feature and random value change. Suppose that two initial molecules are according to Figure (7) and for the first molecule, feature (a) and for the second molecule, feature (b) have been selected. New molecules will be as follows.

W1				W1			
d1	c1	b1	a1	d1	c1	b1	a1
0	1	1	0	0	1	1	1
W2				W2			
d2	c2	b2	a2	d2	c2	b2	a2
1	1	0	1	0	1	0	1

Figure (7). Neighborhood structure of collision

- **Potential Energy**

Potential energy aims to find the best molecule or solution and this can be defined by optimizer function or objective function. Objective function gives a value to each of the molecular structures that expresses the level of satisfaction regarding the proposed solution. The objective function is generally between 0 and 1. According to the initial definition of CRO algorithm, the smaller the objective function is, the more optimized the solution will be. The objective function used in this study will be described in the following.

In the present study, Pearson correlation coefficient has been used to evaluate the molecular structure that includes the potential energy. Pearson correlation coefficient evaluates the linear correlation of two random variables. The value of this coefficient varies between -1 and 1, so that "1" means complete positive correlation, "0" means no correlation, and "-1" means negative correlation. This correlation that is highly applicable in statistics, was presented by Pearson according to the idea of Francis Galton.
Pearson Correlation coefficient between two random variable equals to their covariance divided by their standard deviation.
For a sample population, the correlation coefficient of a community is defined as follows:

$$\rho X,Y = \frac{cov(X,Y)}{\sigma X \sigma Y} = \frac{E[(X-\mu X)(Y-\mu Y)]}{\sigma X \sigma Y}$$

Where *cov* is covariance, σX is the standard deviation of X, μ_X is the mean of E and X.
For a sample population with odd n, Pearson correlation coefficient is defined as follows:

$$r = \frac{\sum_{i=1}^{n}(Xi-\bar{X})(Yi-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(Xi-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Yi-\bar{Y})^2}}$$

The following definition is the same as above definition:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{(Xi-\bar{X})}{s_X}\right)\left(\frac{(Yi-\bar{Y})}{s_Y}\right)$$

In which the quantities are defined as follows:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} Xi, \text{ and} s_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Xi-\bar{X})^2}$$

Finally, for its usage in the objective function, the following is used:

$$F_k = \left(\frac{1}{N}\right)\sum_{i=1}^{N}\left|C_{iR}\right| - \alpha\left(\frac{2}{N(N-1)}\right)\sum_{i=1}^{N}\sum_{j=i+1}^{N}\left(\left|C_{ij}\right| - \beta\right)$$

In the above equation, Fk is the objective function that is implemented for each of the molecules. N is the number of selected features in the proposed solution (molecule) that must have values above 2. If N equals 1, then the second part of the equation leaves the circuit and only the first part is considered.
CiR indicates the relationship between I and R, so that Cij indicates the relationship between I and j. α is the correlation

ISSN:2277-1581
01 May 2016

penalty factor and β is the final penalty factor. Based on decomposition, if Cij is larger than 0.95, it takes 1; otherwise, it takes 0. In this problem, regarding the obtained results, α has been considered as 0.5 and the effects of both are the same.

Results

In Table (1), the data sets have been summarized [5].

Table (1). The summary of data sets

Classes No	Features No	Samples No	Data Set
3	4	150	Iris
7	10	214	Glass identification
2	13	270	Heart
7	17	101	Zoo
2	60	208	Sonar
3	4	24	Lenses
3	4	625	Balance scale

In the following, the proposed method for the data sets has been implemented and for each data set, a number of features were selected. In order to understand whether the selected set has an acceptable level of efficiency, the obtained features were calculated.

In order to evaluate the efficiency of feature selection algorithm, three machine learning algorithms were used in the experiments including NB, j48, and KNN.

The presented classification methods in this study, were implemented in two phases: one for all features and one for the selected features. In order to calculate the results, we considered classification accuracy percentage.

Table (2). Summarizing the obtained results

KNN	J48	NB	Selected Feature	KNN	J48	NB	Feature No	Data Set
97/22	97/13	96/25	2	96/00	96/00	96/00	4	Iris
63/20	76/32	74/35	4	62/62	65/89	65/98	10	Glass identification
78/12	76/82	77/32	1	75/19	76/67	76/67	13	Heart
94/35	92/15	93/13	5	94/20	92/08	92/08	17	Zoo
80/34	75/65	75/65	23	85/10	71/15	71/15	60	Sonar
80/20	84/00	85/13	1	78/25	83/33	83/33	4	Lenses
79/48	76/64	76/64	4	79/48	76/64	76/64	4	Balance scale

Table (3) has been presented based on the following factors regarding compare the available methods:

The ratio of feature reduction to total features is obtained; the less it is, the better will be, since it indicates that more features have been factored. However, this does not mean that reduction always improves the methods, since some of the features are probably removed.

The increase ratio that is obtained through dividing the selected features to the total features, will be better if is in higher rates. Therefore, the more the better.

Table (3). Feature reduction-accuracy increase

Classification accuracy increase ratio			Features reduction ration	Data sets
KNN	J48	NB		
1/01	1/01	1/00	0/50	Iris
1/01	1/16	1/13	0/40	Glass identification
1/04	1/00	1/01	0/08	Heart
1/00	1/00	1/01	0/29	Zoo
0/94	1/06	1/06	0/38	Sonar
1/02	1/01	1/02	0/25	Lenses
1/00	1/00	1/00	1	Balance scale
1/00	1/05	1/03	0/41	Mean of data
0/03	0/06	0/05	0/29	Standard deviation

Regarding small standard deviation, it can be concluded that the mean of data indicates good distribution and through comparing mean and the results of other methods, it is possible to examine the performance of algorithm.

Conclusion and Future Works

Generally, there is no algorithm that performs better than others regarding problem solving. Studies aimed to present an understanding for the capabilities and limitations of various algorithms. Correlation-based feature selection algorithms can improve the efficiency of machine learning algorithms in many cases and at the same time, decrease the features used in learning. These algorithms may encounter failure in selecting the related features when there are features with strong interactions with estimated values in small area of the sample.

The main correlation-based algorithm limitations is failure in selecting features that have estimated values and in general, they will be worthless. While such a unique feature can be considered for a small section of data, a number of such features can cover an important part of data sets. Although redundancy on methods such as j48 and KNN has less effects, it causes detrimental effect on NB. Features with these properties can have some of the unrelated values and the existence of these characters can decrease the efficiency of sample-based learner efficiency. Of course, it is possible to select feature sets for each sample using package methods.

References

i. I. H. Witten,D Mining Practical Machine Learning Tools and Techniques, Published: 2000.

ii. Mingqiang, Y; Kidiyo, K. and Joseph, R. "A survey of shape feature extraction techniques", Author manuscript, published in "Pattern Recognition, Peng-Yeng Yin (Ed.) 43-90".(2008).

iii. Lam, A and Li, V. "Chemical Reaction Inspired Metaheuristic For Optimization" , IEEE transaction on evolutionary computation, Vol.14, NO.3, pp.381-399. (2010).

iv. Lam, A and Li, V. "Chemical Reaction Optimizations Tutorial", springer, DOI 10.1007/s12293-0120075-1, Memetic comp, pp.3-17. (2010).

v. archive.ics.uci.edu/ml/datasets.html. (2013).