Privacy preserving data mining using hiding maximum utility item first algorithm by means of grey wolf optimisation algorithm

M.T. Ketthari*

St. Peter's University,
Avadi, Chennai, India
Email: kettharithandapani12@gmail.com
*Corresponding author

Rajendran Sugumar

Department of Computer Science and Engineering,
Velammal Institute of Technology,
Chennai, India
Email: sugu16@gmail.com

Abstract: In the privacy preserving data mining, the utility mining casts a very vital part. The objective of the suggested technique is performed by concealing the high sensitive item sets with the help of the hiding maximum utility item first (HMUIF) algorithm, which effectively evaluates the sensitive item sets by effectively exploiting the user defined utility threshold value. It successfully attempts to estimate the sensitive item sets by utilising optimal threshold value, by means of the grey wolf optimisation (GWO) algorithm. The optimised threshold value is then checked for its performance analysis by employing several constraints like the HF, MC and DIS. The novel technique is performed and the optimal threshold resultant item sets are assessed and contrasted with those of diverse optimisation approaches. The novel HMUIF considerably cuts down the calculation complication, thereby paving the way for the enhancement in hiding performance of the item sets.

Keywords: data mining; privacy preserving utility mining; sensitive item sets; optimal threshold; grey wolf optimisation; GWO.

Reference to this paper should be made as follows: Ketthari, M.T. and Rajendran, S. (xxxx) 'Privacy preserving data mining using hiding maximum utility item first algorithm by means of grey wolf optimisation algorithm', *Int. J. Business Intelligence and Data Mining*, Vol. X, No. Y, pp.xxx–xxx.

Biographical notes: M.T. Ketthari received his BE degree from the Anna University, Chennai, India in 2010, MTech degree from the College of Engineering Guindy, Anna University, Chennai, India, in 2012, and is pursuing his PhD degree in the St. Peter's University, Avadi, Chennai, India, since 2014. His research interests include data mining and image processing. He has published research articles in various international journals and conference proceedings.

Sugumar Rajendran received his BE degree from the University of Madras, Chennai, India in 2003, MTech degree from the Dr. M.G.R. Educational and Research Institute, Chennai, India, in 2007, and PhD degree from the Bharath University, Chennai, India, in 2011. From 2003 to 2014, he worked at different levels in various reputed engineering colleges across India. He is currently working as an Associate Professor in the Department of Computer Science and Engineering at the Velammal Institute of Technology, Chennai, India. His research interests include data mining, cloud computing and networks. He has published more than 25 research articles in various international journals and conference proceedings. He is acting as a reviewer in various national and international journals. He has chaired various international and national conferences. He is a life time member of ISTE and CSI.

1 Introduction

In a changing world, the data management does not merely mean to store and retrieve data efficiently but also to derive meaningful information out of it. The recent advances in the networking technologies have enabled the collection and sharing of large amounts of data, which rendered the distributed data mining an essential part of the data management (Emekci et al., 2007). With the explosive growth of the hardware and software along with immense computing and communication power of the system and devices, it is unbelievably easy to store, retrieve and process large amounts of information. A good amount of privacy issues also arises with the proliferation of the digital technologies (Ukil, 2010). Moreover, privacy is an important issue in many data mining applications that deal with the healthcare, security, financial and other types of the sensitive data (Venkatesan et al., 2016). The actual anxiety of people is that their private information should not be misused behind the scenes without their knowledge. The real threat is that once the information is unrestricted, it will be impractical to stop the misuse (Kamakshi and Babu, 2010). Parallel grouping is the most minor order issue in which the information test has a place with one of the two target class names (Pratama et al., 2015a). Restorative conclusion, biometric security, and other comparable applications are cases of paired arrangement. At the point when the aggregate number of target class marks is more prominent than two, it is called multi-class arrangement (Pratama et al., 2016a). Privacy can, for instance, be threatened when the data mining techniques use the identifiers which themselves are not very sensitive, but are used to connect personal identifiers such as the addresses, names etc., with other more sensitive personal information (Pratama et al., 2015b). The simplest solution to this problem is to completely hide the sensitive data or not to include such sensitive data in the database. But this solution is not ideal and accurate in many applications, like the medicine research, DNA research, etc., (Phake et al., 2015). On the off chance that utilising the possess characterisation yields, the classifier would prepare its own particular blunders over and over into its structure, which after some time would prompt to a diminished prescient performance (Lughofer and Pratama, 2017). As foreshadowed, in single-name arrangement issues, each of the example information is related with a one of a kind target class name from a pool of target class names (Pratama et al., 2016a, 2016b, 2016c, 2016d, 2016e, 2016f). Single label classifiers can be further grouped into paired classifiers and multi-class classifiers.

There are generally two types of definitions for privacy. One type of definitions is micro databased. K-anonymity (Sweeney, 2002) and l-diversity (Machanavajjhala et al., 2007) are the two typical examples. K-anonymity requires that a published data set should have at least k rows (called a group) sharing the same QID value. Moreover, privacy preserving data mining (PPDM) finds numerous applications in surveillance which is naturally supposed to be the 'privacy-violating' applications. The key is to design methods which continue to be effective, without compromising security. One of the sources of privacy violation is called the data magnets (Rezgur et al., 2003). The data magnets are techniques and tools used to collect personal data (Pratama et al., 2016a, 2016b, 2016c, 2016d, 2016e, 2016f). In many cases, the users may or may not be aware that information is being collected or do not know how the relative information is collected (Laudon, 1996). Worse is the privacy invasion occasioned by the secondary usage of data when individuals are unaware of the 'behind the scenes' uses of data mining techniques (John, 1999). In particular, personal data can be used for secondary usage largely beyond the users' control and privacy laws (Lughofer et al., 2015). Then again, unsupervised strategies rank sentences by remarkable quality scores which are evaluated in light of factual and etymological components and remove the main ones to constitute the synopsis (Zhang et al., 2013). This scenario has led to an uncontrollable privacy violation, not because of the data mining itself, but fundamentally because of the misuse of data. Instabilities in the information streams can not be taken care of by the sort one concealed hubs since information streams can not express the issue being illuminated precisely, along these lines bringing about estimated parameter recognisable proof (Pratama et al., 2016d). However, many of the research have indicated that these privacy models are vulnerable to various privacy attacks (Wong, et al., 2007; Zhang et al., 2007; Ganta et al., 2008), and provide insufficient privacy protection. The adequacy of our developing web news mining strategies is numerically approved and looked at against best in class calculations (Za et al., 2017; Joo et al., 2016).

Moreover, the differential privacy (Dwork et al., 2006) has recently received considerable attention as a substitute for the partition-based privacy models for privacy preserving. However, so far most of the research on the differential privacy concentrates on the interactive setting with the goal of reducing the magnitude of the added noise (Dinur and Nissim, 2003; Roth and Roughgarden, 2010), releasing certain data mining results (Bhaskar et al., 2010; Chaudhuri et al., 2012a, 2012b), or determining the feasibility and infeasibility results of the differentially private mechanisms (Blum et al., 2008; McGregor et al., 2010). In recent years, different protocols have been proposed for different data mining tasks including the association rule mining (Vaidya and Clifton, 2002), clustering (Vaidya and Clifton, 2003), and the classification (Lindell and Pinkas, 2002). However, none of these methods provide any privacy guarantee on the computed output (i.e., classifier, association rules). Dominant part of developing classifiers work in the completely managed preparing situation, which expect all information streams to be completely marked (Pratama et al., 2016a, 2016b, 2016c, 2016d, 2016e, 2016f). Moreover, the randomisation technique is an inexpensive and efficient approach for the PPDM. In order to assure the performance (Zhu and Liu, 2004) of data mining and to preserve individual privacy, these randomisation schemes need to be implemented (Pratama et al., 2016a, 2016b, 2016c, 2016d, 2016e, 2016f). Moreover, the privacy preservation of multiple data sets is an efficiently challenging task for data mining.

Nowadays, to overcome the leakage most of the optimisation algorithms are introduced which maintain the reduced cost and securely share the data (Pratama et al., 2014).

2 Literature review

The literary arena is flooded with various techniques which have been launched for the privacy preserving of data sharing. Recounted below are some of most modern published works in this regard.

Kumari et al. (2013) proposed the modified maximum sensitive item-sets conflict first algorithm (MSICF) for hiding the sensitive item-sets. The method found the sensitive item sets and modified the frequency of high valued utility items. But, the performance of this method was deficient if the utility value of the items was same and the modified MSICF algorithm computed the sensitive item sets by utilising the user defined utility threshold value.

Moreover, Yang et al. (2013) have explained the privacy preserving data obfuscation scheme used in data statistics and data mining. Here they allocated different keys to users, and different users were given different permissions to access to data. To achieve this, their scheme contained four steps. Firstly, an improved cloud model was explained to generate an accurate 'noise'. Next, an obfuscation algorithm was introduced to add noise to the original data. Then, an initial scheme for the dataset obfuscation was explained, including the grouping and key allocating processes. In the final step, a fine-grained grouping scheme based on similarity was explained. The experiments showed that their scheme obfuscated the date correctly, efficiently, and securely.

Moreover, Yang et al. (2014) have explained the PPDM algorithm based on modified particle swarm optimisation. The algorithm was based on the centralised database, and it was used in the distributed database. The algorithm was divided into two steps in the distributed database. In the first step, the modified particle swarm optimisation algorithm was used to get the local Bayesian network structure. The purpose of the second step was getting the global Bayesian network structure by using the local ones. In order to protect the data privacy, the secure sum was used in the algorithm. The algorithm was proved to be convergent on theory. Some experiments were done on the algorithm, and the results proved that the algorithm was feasible.

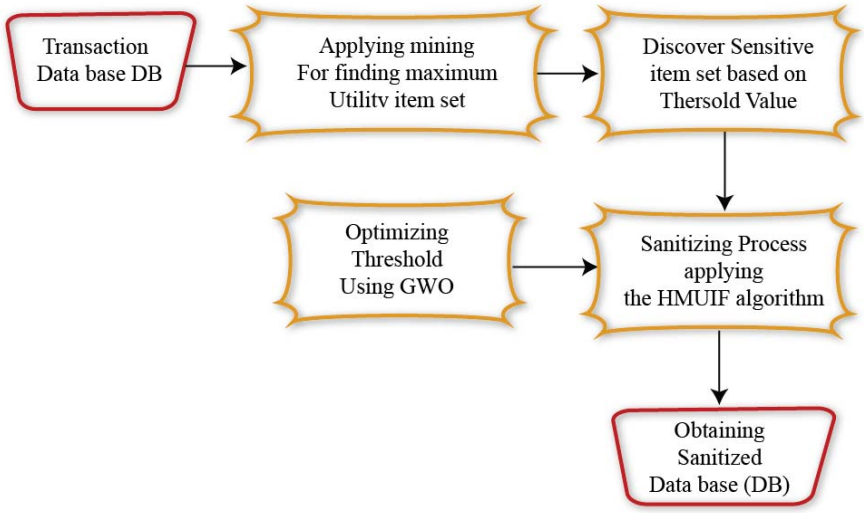
The modified MSICF algorithm was presented by Selvaraj and Kuthadi (2013). The result showed that the performance was improved and examined by using the hiding factor, miss cost (MC) and dissimilarity (DIS) and adapted the database transactions containing the sensitive item sets to reduce the utility value below the given threshold while averting the reconstruction of the original database from the sanitised one.

3 Proposed methodology

The underlying motive behind the novel technique is dedicated to build up the PPDM employing the hiding maximum utility item first (HMUIF) together with the optimisation technique. The HMUIF approach is entrusted with the task of considerably cutting down utility of sensitive data hidden in the item set. The steps considered by the utility mining illustrate the optimisation of the threshold value and sanitised database. It is effectively used to locate the sensitive data from the utility data set with the help of transactions and

external utility values. With an eye on slashing down the utility of each sensitive item sets, the new-fangled approach modernises the item quantity with the maximum utility value in the sensitive item set *S*. The variation between the threshold and high sensitive utility value estimated in the original data is revised for the sanitising procedure. With the intent to conceal the sensitive item sets, the frequency value of the items is suitably varied. In the HMUIF process the threshold is treated as the most significant constraint and in the data hiding approach, the relative threshold is decided by employing the grey wolf optimisation (GWO) approach. The threshold value is estimated in the optimisation method by utilising the product of the item set frequency value for the transaction table and the external utility value with the constant values of the variables. The efficiency of the threshold value is appraised with the help of constraints like hiding failure (HF), MC and the DIS. The efficiency in performance of the innovative technique is assessed and contrasted with those of various optimisation methods including the genetic (GA) and adaptive genetic algorithms (AGAs) by means of the captioned constraints. The record-breaking PPDM technique is proficiently performed in the working platform of the MATLAB software; it is shown in Figure 1.

Figure 1 Structure for proposed method (see online version for colours)



3.1 Utility mining process

The Utility item set mining, also generally called the utility pattern mining, was first introduced in every item in the item sets is connected with an additional value, called the internal utility which is the quantity (i.e., count) of the item. This process is used to find the utility values in the item sets.

Utility mining is used to find all the item set's utility values.

The utility value of item set I_v in transaction T_o is defined as,

$$U(I_v, T_o) = k(I_v) \times m(I_v, T_o) \tag{1}$$

Whereas final item set and m as final transaction and also $1 \leq k \leq m$

The utility value of an item set S in transaction T_o is defined as,

$$U\left(S, T_o\right) = \sum_{i_m \in S} U\left(I_v, T_o\right) \tag{2}$$

Thereafter, such item sets are ascertained whose utility value exceeds the user-defined threshold value τ, where τ. Represents the minimum utility threshold. The item set S characterises a high utility item set, if $U(S) \geq \tau$. These high utility item sets are amassed in $H = \{S_1, S_2, S_i\}$ and these item sets are called the sensitive item sets. Further, it is essential to head the sensitive item sets based on certain safety stratagems.

Pseudo code of the GWO

Initialise the grey wolf population $T_i = (1, 2, \ldots\ldots n)$
Initialise *a, A, and* *C*
Find the fitness of each search agent population by using equation (3)
T_a = the best search agent
T_β = the second best search agent
T_δ = the third best search agent
While (t < Max number of iteration)
  For each search agent population
    Update the position of the current search agent by using equation (3)
  End for
    Update *a, A, and* *C*
    Calculate the fitness for all search agents
    Update T_a, T_β and T_δ
    t = t + 1
End while
Return T_a

3.2 Threshold optimisation using GWO

In the privacy preserving process, the privacy threshold τ which is the proportion of sensitive patterns is still discovered from the sanitised database. In the process, all sensitive patterns can be discovered. The advantage of the threshold mechanism is that, the users can balance the privacy and disclosure of information. The high utility value item sets are hidden by modifying the frequency values of items contained in the sensitive item sets based on the minimum utility threshold value. The threshold value τ ranges from 100 to 2000, and the optimal threshold is found out using the GWO algorithm (Yusof and Mustaffa, 2015), this procedure of GWO shown in below pseudo code.

3.2.1 Initial solutions

In the initial solution, the random values of the threshold are chosen and the search agent parameters such as-a, A and C, the coefficient vectors are initialised and the population size is indicated by using the number of transactions and number of items.

3.2.2 Fitness calculation

The fitness function is used to find the minimum value of the difference between the threshold value and the multiplication product of the external utility value and the data item set value. The fitness is denoted by means of the following equation:

$$F = T - \left[\left(\sum_{i=1}^{N} \left(\sum_{j=1}^{m} U_j D(i, j) \right) * \omega_i \varepsilon \right) \right] \tag{3}$$

where T represents the best threshold value, the external utility data value, the utility data value and ω, ε characterise the variables, the number of transaction and, the number of item sets.

The condition is checked immediately after the fitness calculation. If the condition is satisfied, the result gives optimal value of the best solution. Otherwise, it is determined by the new first best, second best and the third best solution in the population agent. Thereafter, the new fitness value is compared with the previous one. If the minimum fitness value is achieved, then that value gives the best threshold value. Otherwise, it will update the new solution, by using the updating equations.

3.2.3 Based on the fitness separate the solution

Now, we find the fitness separate solution (threshold) based on the fitness value. Let the first best fitness solutions be α, the second best fitness solutions β and the third best fitness solutions δ.

3.2.4 Update the position

We assume that the alpha (the best candidate solution) beta and delta have the improved knowledge about the potential location of prey in order to mathematically reproduce the hunting behaviour of the grey wolves. As a result, we hoard the first three best solutions attained so far and require the other search agents (including the omegas) to revise their positions according to the position of the best search agent. For revision of the novel solution, $T(t + 1)$ the below-mentioned formulas are employed.

$$P^a = \left| C_1.T_a - T \right|, P^\beta = \left| C_1.T_\beta - T \right|, P^\delta = \left| C_1.T_\delta - T \right| \tag{4}$$

$$T_1 = T_\alpha - A_1.(P_\alpha), T_2 = T_\beta - A_2.(P_\beta), T_3 = T_\delta - A3(P_\delta) \tag{5}$$

To have the hyperspheres with different random radii the arbitrary parameters A and C help the candidate solutions. The investigation and utilisation are guaranteed by the adaptive values of A and a. With decreasing A, half of the iterations are dedicated to the investigation ($|A| < 1$) and the other half are devoted to the utilisation. Encircling the

behaviour, the subsequent equations are employed in order to carry out the mathematical modeling.

$$P = | C.T_P(t) - T(t) \tag{6}$$

For find, the coefficient vectors use equation (27):

$$A = 2a.r_1 - a, \quad C = 2.r_2 \tag{7}$$

where t indicates the current iteration, A and C are coefficient vectors, T_p is the position vector of the prey T and indicates the position vector of a grey wolf. The components of a are linearly decreased from 2 to 0 over the course of the iterations and r_1, r_2 are random vectors in [0, 1].

The GWO has only two main parameters to be adjusted (a and C). However, we have kept the GWO algorithm as simple as possible with the fewest operators to be adjusted. The maximum utility obtained in the process will be continued.

3.3 Hiding maximum utility item set (HMUIF) technique

The main objective of the HMUIF algorithm is to diminish the utility value of each sensitive item set by modifying the quantity values of items contained in the sensitive item sets. The pseudo-code of the HMUIF algorithm is given below:

Pseudo code of the HMUIF

Input: Original database D, minimum utility threshold τ and sensitive item set $H = \{s_1, s_2, \ s_i\}$

Output: Sanitized data base DB

For each sensitive item set $s_i \in H$

$diff = U(S_i) - \tau$ // the utility value needs to be reduced

While ($diff > 0$)

{

$U(I_v, T_0) = \arg\max_{(i \in s_i, s_i \subseteq T)}(U(i, T))$

Optimise T using GWO algorithm

Modify $m(I_v, T_0)$

$$m(I_v, T_0) = \begin{cases} 0 & ,if\ U(I_v, T_0) < dif \\ m(I_v, T_0) - \left\lceil \dfrac{diff}{s(i_m)} \right\rceil & ,if\ U(I_v, T_0) < dif \end{cases}$$

$$diff = \begin{cases} diff - U(I_v, T_0) & ,if\ U(I_v, T_0) < diff \\ 0, & ,if\ U(I_v, T_0) > diff \end{cases}$$

}

The above pseudo code shows the maximum sensitivity data hiding process, this process continuously becomes lower than τ. The HMUIF process calculates the difference between the utility of item set and minimum utility threshold. If the difference value is greater than zero it means that the utility value is calculated in the sensitive data items

and also takes the maximum sensitive data in transaction process that is $\max_{(i \in s_i, s_i \subseteq T)}$. In the next step find the modified quantity of an item I_v in transaction T_o. If the value is zero the maximum utility is lesser than the difference value, otherwise, it is the maximum value. The process continues until the utility value of each sensitive item set is below τ and this process is based on the original database to the sanitised database.

Example: Let us consider a transaction database with four numbers of transactions and three different items with their external utility values are shown in Table 1.

Table 1 Item set with external utility values

Transaction	*A*	*B*	*C*
T_1	2	0	0
T_2	1	2	0
T_3	2	0	3
T_4	3	0	1
External utility value	3	1	2

By using transaction Table 1 used to find the high utility item sets are,

Table 2 Item set with utility values

Item set	*Utility value*
A	**24**
B	2
C	8
AB	11
BC	8
AC	**15**

Table 2 the high utility items are A and AC, the utility value is 24 and 15.

Table 3 High utility value

High utility item set	*Utility value*
A	**24**
AC	15

In this example, we set the threshold value $\tau = 20$. After that, compared the utility values of both items sets A and C in the transactions T3 and T4

Table 4 Transaction values

TID *item*	*A*	*C*	
T_3	2	**3**	◄——Highest utility 4*3
T_4	3	1	

Table 5 New utility Item set

High utility item set	*Utility value*
A	**12**
AC	3

In the above-mentioned example, the initial high utility value is 24 and the threshold value is greater than the A item set utility value. So using HMUIF algorithm to hide the high utility sensitive data is converted to other, which are 24 converted to 12. The HMUIF algorithm generates no artificial item sets from the sanitised database.

4 Result and discussion

The novel HMUIF technique is performed on the working platform of MATLAB version 2014. The innovative approach employs two datasets I and II for their performance appraisal and the relative datasets encompass a diverse number of transactions and effectively a large number of item sets. The ensuing performance appraisal constrains privacy preserving are effectively achieved in the innovative HMUIF approach together with the GWO method and they are assessed and contrasted with the relative outcomes of the modern method.

4.1 Dataset description

In the proposed work, two datasets are utilised for the performance analysis of our proposed HMUIF with threshold optimisation algorithm. The datasets I and II respectively contain 100 and 200 transactions with ten different items. These two datasets are described in Table 6.

Table 6 Database details

Dataset	*Number of transactions*	*Distinct items*
Dataset I	200	10
Dataset II	100	10

Table 7 Sample dataset I

Attribute transaction	*A1*	*A2*	*A3*	*A4*	*A5*	*A6*	*A7*	*A8*	*A9*	*A10*
T1	0	2	0	2	1	2	0	2	0	2
T2	0	2	0	2	0	2	0	2	0	2
T3	0	2	0	2	0	2	0	2	0	2
T4	0	0	0	0	1	0	0	0	0	0
T5	0	5	0	5	0	5	0	5	0	5
T6	0	0	0	0	5	0	5	0	0	0
T7	0	0	0	0	5	0	0	0	0	0
T8	0	0	0	0	0	0	0	0	0	0
T9	0	0	4	0	0	0	0	0	0	0

Table 8 Sample dataset II

Attribute / *Transaction*	*A1*	*A2*	*A3*	*A4*	*A5*	*A6*	*A7*	*A8*	*A9*	*A10*
T1	4	0	1	0	8	7	9	10	3	0
T2	0	0	0	0	1	0	0	4	0	0
T3	8	0	0	4	10	0	6	2	7	8
T4	0	0	0	10	3	0	2	10	2	1
T5	9	8	8	4	9	0	0	9	6	2
T6	0	1	5	8	6	0	4	10	3	8
T7	0	4	0	7	6	2	8	10	2	10
T8	4	8	0	0	9	4	4	7	4	7
T9	0	0	0	0	0	0	7	2	4	4

4.2 Performance analysis

The efficiency of the novel technique is evaluated by means of calculating certain efficiency metrics. In the innovative technique, various parameters such as the HF, MC and DIS factors are effectively evaluated for ascertaining the efficiency of the novel technique. The captioned parameters are elucidated as follows.

4.2.1 Hiding failure

The HF measures the percentage of the sensitive item sets discovered from *DB′*. It is measured by the sensitive item sets of both the original database and the sanitised database, which is stated as follows,

$$HF = \frac{|S(DB')|}{|S(DB)|} \tag{8}$$

where $S(DB')$ and $S(DB)$ represents sensitive item set from the original database *DB* and sensitive item set from sanitised database *DB′*.

4.2.2 Miss cost

The MC measures the difference ratio of the valid item sets presented in the original database and the sanitised database. Its value is computed as:

$$MC = \frac{|N(DB) - N(DB')|}{|N(DB)|} \tag{9}$$

where $N(DB)$ and $N(DB')$ denotes the non-sensitive item sets discovered from the original database *DB* and the sanitised database *DB′*, respectively.

4.2.3 Dissimilarity

The DIS between the original database *DB* and the sanitised database *DB′* is calculated as furnished below.

$$DS = \frac{1}{\sum_{i=1}^{m} f_{DB}(i)} \left(\sum_{i=1}^{m} \left[f_{DB}(i) - f_{DB}'(i) \right] \right) \tag{10}$$

where $f_{DB}(i)$ and $f_{DB'}(i)$represents the frequency of the i^{th} item in the database *DB* and the frequency of the i^{th} item in the database *DB*.

Table 7 Performance analysis parameter with optimal threshold in dataset I

Threshold value (τ)	*Optimal threshold (τ_{opt})*	*Performance parameter*		
		HF	*MC*	*DIS*
100–500	276	0.750575758	0.751497	0
500–1000	823	0.506585366	2.567073	1.65E-279
1000–1500	1,422	0.395833333	6.697368	5.29E-121
1500–2000	1,836	0.171428571	15.71429	1.16E-27

Table 8 Performance analysis parameter with optimal threshold in dataset II

Threshold value (τ)	*Optimal threshold (τ_{opt})*	*Performance parameter*		
		HF	*MC*	*DIS*
100–500	310	0.139359699	0.527307	0
500–1000	615	0.033783784	1.739865	5.94E-101
1000–1500	1,331	0.016853933	4.100629	1.88E-25
1500–2000	1,786	0.119266055	6.440367	1.82E-61

Figure 2 Convergence graph (see online version for colours)

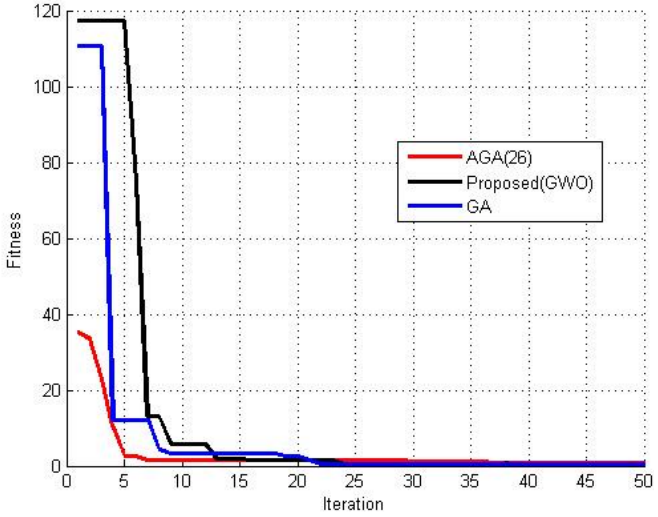
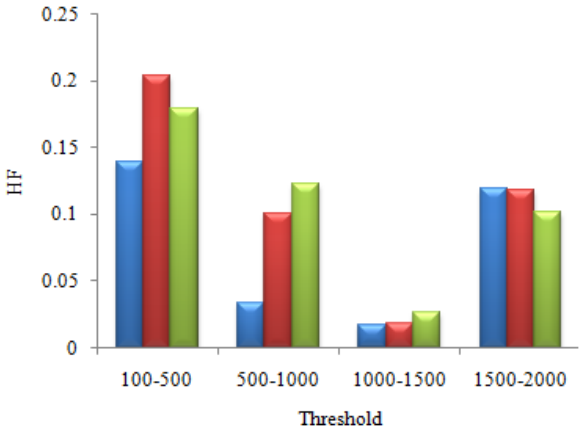
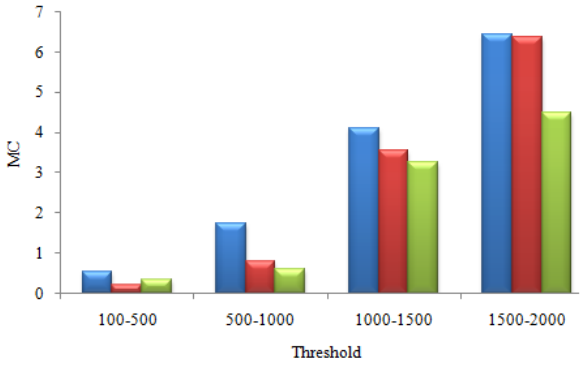


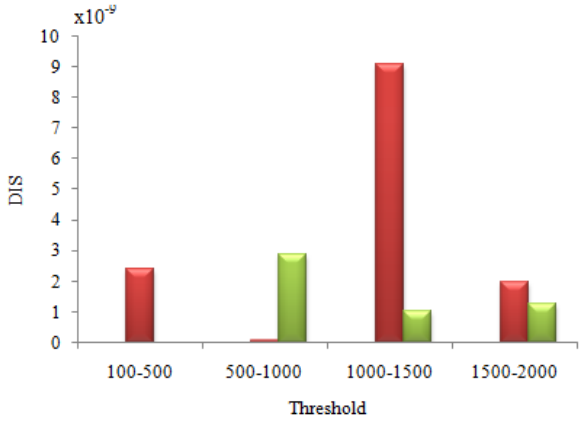
Figure 3 Comparison graph for dataset I (see online version for colours)



(a)

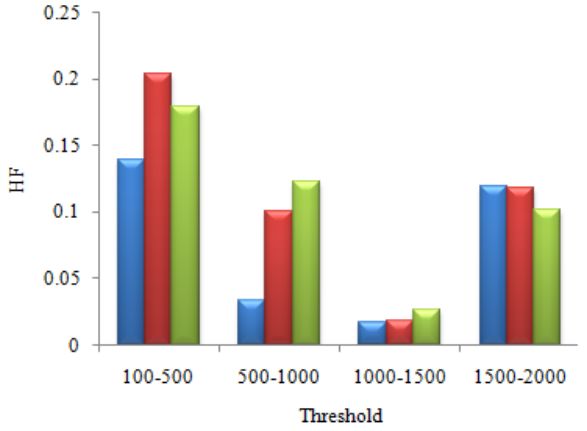


(b)

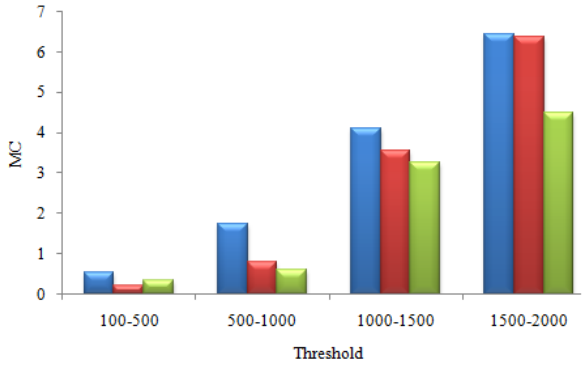


(c)

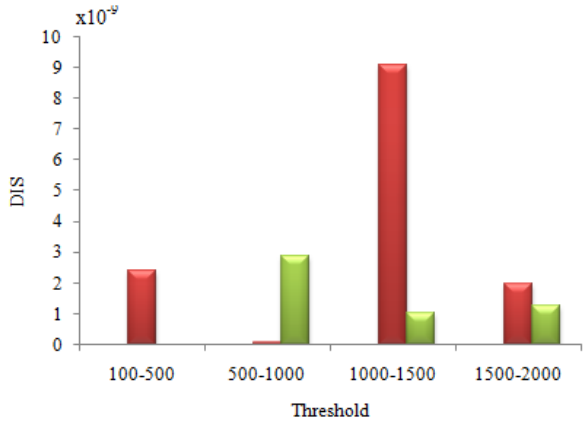
Figure 4 Comparison graph for dataset II (see online version for colours)



(a)



(b)



(c)

Tables 7 and 8 respectively show the threshold value range and the optimal threshold in the GWO technique with performance evaluation parameters in two different databases. The performance parameter HF yields minimum value and the remaining parameters which are MC and DIS yield a higher value.

The Figure 2 depicts the convergence graph for diverse techniques like the genetic algorithm (GA), the AGA and the GWO. It is crystal clear from the captioned figure that the proposed GWO algorithm beats the peer algorithms hands down, by amazingly achieving the minimum error in relation to the rival algorithms. The GA algorithm attains higher value at the beginning, but it remains constant after 60th iterations whereas the AGA takes minimum iteration for convergence and remains constant after just ten iterations. But it achieves the maximum value than AGA. On the other hand, the proposed grey wolf algorithm attains the superior value at the beginning and achieves the minimum value between 35 to 40 iterations than GA and is able to attain the convergence with extremely minimum value when compared with the other two traditional algorithms.

Figures 3 and 4 smartly show the novel optimal threshold algorithm assessed by utilising the constraints such as the HF, MC and the DIS. In the dataset I, Figure 3(a) illustrates the HF of the optimal threshold values for various conceptual threshold values. Out of these values representing the minimum threshold value of the entire HF values, only the GWO technique ushers in the optimal threshold and its relative hiding values are contrasted with those of several approaches such as the GA, AGA and GWO. The graph unambiguously illustrates the fact the cuckoo search algorithm is competent to furnish the optimal value. Figure 3(b) elegantly exhibits the comparison graph for the MC factor for various threshold values in respect of dataset I. From these values, the maximum values are selected for the best output outcomes, which show that the grey wolf algorithm tops the list by achieving the maximum outcome vis-à-vis the other parallel techniques. Figure 3(c) fascinatingly demonstrates the DIS graph for diverse threshold values achieved by the diverse techniques employed. The maximum DIS value offers the best threshold value from the diverse threshold values. In dataset I, the optimal threshold values are in diverse ranges from 200 to 2000. Figures 4(a), 4(b) and 4(c) effectively exhibit the performance constraints comparison graph for dataset II. In the first and second data set the minimum HF value is achieved and in the case of others, values are high. In respect of dataset I, the novel technique when compared to the parallel technique, the differences are found to be 85.42%, 86.23% and 81.28% respectively. So is the case with dataset II.

5 Conclusions

The novel HMUIF privacy preserving utility mining algorithm is effectively employed for concealing the high utility sensitive item sets and it successfully hides the sensitive item sets from the rivals irrespective of whether the item utility value is identical or different. In our new-fangled technique, the identical utility values of the high sensitive item set along with the diverse utility values of the high sensitive item set are well-elucidated and have been able to superb yields. For the choice of threshold values in this paper, we have employed the cuckoo search algorithm which is found to yield the optimal threshold value. The efficiency of the optimal threshold is estimated by employing the HF, MC and DIS factors and it is found to yield the finest threshold value.

The epoch-making technique fantastically fine-tunes the database transactions by possessing the sensitive item sets to restrict the utility value well below the specified threshold value. The cheering test outcomes have established without any bit of doubt that the performance of our masterpiece HMUIF algorithm with the grey wolf algorithm has scored a clear and unbeatable edge over the traditional techniques, by scaling newer and newer heights, thereby forcing the peer competitors to be mere silent spectators.

References

Bhaskar, R., Laxman, S., Smith, A. and Thakurta, A. (2010) 'Discovering frequent patterns in sensitive data', *Proc. ACM Int'l. Conf. Knowledge Discovery and Data Mining (SIGKDD'10)*, pp.503–512.

Blum, A., Ligett, K. and Roth, A. (2008) 'A learning theory approach to non-interactive database privacy', *Proc. ACM Symp., Theory of Computing*, Vol. 60, No. 2, p.12.

Chaudhuri, K., Monteleoni, C. and Sarwate, A. (2012a) 'Differentially private empirical risk minimization', *J. Machine Learning Research*, Vol. 12, pp.1069–1109.

Chaudhuri, K., Sarwate, A.D. and Sinha, K. (2012b) 'Near-optimal differentially private principal components', *Proc. Conf. Neural Information Processing Systems*, pp.989–997.

Dinur, I. and Nissim, K. (2003) 'Revealing information while preserving privacy', *Proc. ACM Symp. Principles of Database Systems*, pp.202–210.

Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006) 'Calibrating noise to sensitivity in private data analysis', *Proc. Theory of Cryptography Conf.*, pp.265–284.

Emekci, F., Sahin, O.D., Agrawal, D. and Abbadi, A.E. (2007) 'Privacy preserving decision tree learning over multiple parties', *Data and Knowledge Engineering*, Vol. 63, No. 2, pp.348–361.

Ganta, S.R., Kasiviswanathan, S. and Smith, A. (2008) 'Composition attacks and auxiliary information in data privacy', *Proc. ACM Int'l. Conf. Knowledge Discovery and Data Mining*, pp.265–273.

John, G.H. (1999) 'Behind-the-scenes data mining', *New Letter of ACM SIG on KDDM*, Vol. 1, No. 1, pp.9–11.

Joo, M., Zhang, Y., Wang, N. and Pratama, M. (2016) 'Attention pooling-based convolutional neural networks for sentence modeling', *Journal of Information Sciences*, Vol. 373, pp.388–403.

Kamakshi, P. and Babu, A.V. (2010) 'Preserving privacy and sharing the data in distributed environment using cryptographic technique on perturbed data', *Journal of Computing*, Vol. 2, No. 4.

Kumari, J., Murthy, N. and Suresh, S.S. (2013) 'An optimization based modified maximum sensitive item-sets conflict first algorithm (MSICF) for hiding sensitive item-sets', *Journal of Computer Applications*, Vol. 2, No. 4, pp.1–7.

Laudon, K.C. (1996) 'Markets and privacy', *Communication of the ACM*, Vol. 39, No. 9, pp.92–104.

Lindell, Y. and Pinkas, B. (2002) 'Privacy preserving data mining', *J. Cryptology*, Vol. 15, No. 3, pp.177–206.

Lughofer, E. and Pratama, M. (2017) 'On-line active learning in data stream regression employing evolving generalized fuzzy models with certainty sampling', *Journal IEEE Transactions on Fuzzy Systems*, pp.1–5.

Lughofer, E., Cernuda, C., Kindermann, S. and Pratama, M. (2015) 'Generalized smart evolving fuzzy systems', *Journal of Evolving Systems*, Vol. 6, No. 4, pp.269–292.

Machanavajjhala, A., Kifer, D., Gehrke, J. and Venkitasubramaniam, M. (2007) 'l-diversity: privacy beyond k-anonymity', *ACM Transaction on Knowledge Discovery from Data*, Vol. 1, No. 1.

McGregor, A., Mironov, I., Pitassi, T., Reingold, O., Talwar, K. and Vadhan, S. (2010) 'The limits of two-party differential privacy', *Proc. IEEE Symp., Foundations of Computer Science*, pp.81–90.

Phake, S.S., Moon, V.S., Waghmare, A.A. and Ijare, G.B. (2015) 'Preserving privacy using secret sharing in distributed environment on perturbed data', *International Journal of Innovative Research in Computer and Communication Engineering*, Vol. 3, No 3.

Pratama, M., Anavatti, S. and Lu, J. (2016a) 'Recurrent classifier based on an incremental metacognitive-based scaffolding algorithm', *Journal of IEEE Transactions on Fuzzy Systems*, Vol. 23, No. 6, pp.2048–2066.

Pratama, M., Lu, J., Lughofer, E., Zhang, G. and Joo, M. (2016b) 'Incremental learning of concept drift using evolving type-2 recurrent fuzzy neural network', *Journal of IEEE Computational Intelligence Society*, p.1.

Pratama, M., Lughofer, E., Lim, C.P., Rahayu, W., Dillon, T. and Budiyono, A. (2016c) 'pClass+: a novel evolving semi-supervised classifier', *Journal of Fuzzy Systems*, pp.1–18.

Pratama, M., Zhang, G. and Joo, M. (2016d) 'An incremental type-2 meta-cognitive extreme learning machine', *Journal IEEE Transactions on Fuzzy Systems*, Vol. 47, No. 2, pp.339–353.

Pratama, M., Lua, J., Anavattib, S., Lughoferc, E. and Lim, C.P. (2016e) 'An incremental meta-cognitive-based scaffolding fuzzy neural network', *Journal of Neurocomputing*, Vol. 171, pp.89–105.

Pratama, M., Lub, J., Lughofer, E., Zhang, G. and Anavattid, S. (2016f) 'Scaffolding type-2 classifier for incremental learning under concept drifts', *Journal of Neurocomputing*, Vol. 191, pp.304–329.

Pratama, M., Anavatti, S.G. and Joo, M. (2015a) 'pClass: an effective classifier for streaming examples', *Journal of IEEE Computational Intelligence Society*, Vol. 23, No. 2, pp.369–386.

Pratama, M., Lu, J. and Zhang, G. (2015b) 'Evolving type-2 fuzzy classifier', *Journal of IEEE Computational Intelligence Society*, pp.574–589.

Pratama, M., Joo, M., Anavatti, S., Lughofer, E., Wang, N. and Arifin, I. (2014) 'A novel meta-cognitive-based scaffolding classifier to sequential non-stationary classification problems', in *Proceedings of IEEE International Conference on, IEEE*, pp.6–14.

Rezgur, A., Bouguettaya, A. and Eltoweissy, M.Y. (2003) 'Privacy on the web: facts, challenges, and solutions', *IEEE Security and Privacy*, Vol. 1, No. 6, pp.40–49.

Roth, A. and Roughgarden, T. (2010) 'Interactive privacy via the median mechanism', *Proc. ACM Symp. Theory of Computing*, pp.765–774.

Selvaraj, R. and Kuthadi, V.M. (2013) 'A modified hiding high utility item first algorithm (HHUIF) with item selector (MHIS) for hiding sensitive itemsets', *Journal of Innovative Computing, Information and Control*, Vol. 9, No. 12, pp.4851–4862.

Sweeney, L. (2002) 'k-anonymity: a model for protecting privacy', *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, Vol. 10, No. 5, pp.1–14.

Ukil, A. (2010) 'Privacy preserving data aggregation in wireless sensor networks', *IEEE Wireless and Mobile Communications*, pp.435–440.

Vaidya, J. and Clifton, C. (2002) 'Privacy preserving association rule mining in vertically partitioned data', *Proc. ACM Int'l. Conf. Knowledge Discovery and Data Mining*, pp.639–644.

Vaidya, J. and Clifton, C. (2003) 'Privacy-preserving k-means clustering over vertically partitioned data', *Proc. ACM Int'l. Conf. Knowledge Discovery and Data Mining*, pp.206–215.

Venkatesan, R., Joo, M., Dave, M., Pratama, M. and Wu, S. (2016) 'A novel online multi-label classifier for high-speed streaming data applications', *Journal of Evolving Systems*, pp.1–13.

Wong, R.C.W., Fu, A.W.C., Wang, K. and Pei, J. (2007) 'Minimality attack in privacy preserving data publishing', *Proc. Int'l. Conf. Very Large Data Bases*, pp.543–554.

Yang, L., Wu, j., Peng, L. and Liu, F. (2014) 'Privacy-preserving data mining algorithm based on modified particle swarm optimization', *Intelligent Computing Methodologies Lecture Notes in Computer Science*, Vol. 8589, pp.529–541.

Yang, P., Gui, X., Tian, F., Yao, J. and Lin, J. (2013) 'A privacy-preserving data obfuscation scheme used in data statistics and data mining', *IEEE International Conference on Embedded and Ubiquitous Computing*, pp.881–887.

Yusof, Y. and Mustaffa, Z. (2015) 'Time series forecasting of energy commodity using grey wolf optimizer', in *Proceedings of the International Multi Conference of Engineers and Computer Scientists IMECS 2015*, pp.25–30.

Za, C., Pratama, M., Lughofer, E. and Anavattic, S. (2017) 'Evolving type-2 web news mining', *Journal of Applied Soft. Computing*, Vol. 54, pp.200–220.

Zhang, L., Jajodia, S. and Brodsky, A. (2007) 'Information disclosure under realistic assumptions: privacy versus optimality', *Proc. ACM Conf. Computer and Comm. Security (CCS'07)*, pp.573–583.

Zhang, Y., Joo, M. and Zhaom, R. (2013) 'Multi-view convolutional neural networks for multi document extractive summarization', *Journal of IEEE Transactions on Cybernetics*, pp.1–13.

Zhu, Y. and Liu, L. (2004) 'Optimal randomization for privacy preserving data mining', *ACM*, pp.761–766.