International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Computing the Probability on Socio Economic Factors to Predict the Crime Locations by Means of Joint Probability Based AMABC-FCIL

Sujatha Radhakrishnan[1]*, Ezhilmaran Devarasan[2]

[1]*Vellore Institute of Technology, Vellore, India*
[2]*School of Advanced Sciences, India*
*Corresponding author's Email: sujatha0180@gmail.com

Abstract: The Frequent Itemsets Mining (FIM) is a demanding task common to several important data mining applications that look for interesting patterns within the databases. Several techniques have been proposed to mine the frequent closed itemsets. In this paper, we have proposed a frequent closed itemset mining technique based on probability. The socio economic factors are clustered with the help of the Adaptive Mutation based Artificial Bee Colony (AMABC) Algorithm after fetching them from the database. After clustering the attributes, the rules are generated and the Joint Probability Function (JPF) is computed. The rules satisfy the joint probability cutoff which is selected to construct the Frequent Closed Itemset Lattice (FCIL). The rules which satisfy the support threshold after constructing the FCIL are selected as the frequent closed itemsets. Finally, a testing process is included in which the known test data are provided. To analyze the performance of the proposed technique, certain performance metrics like time, memory, accuracy, lift and confidence rates are utilized and the performance of the proposed technique is improved with the existing Sliding with Itemsets Factor (SIF) based FCIL.

Keywords: Frequent Closed itemsets lattice (FCIL); Crime locations prediction; Joint Probability based Adaptive Mutation based Artificial Bee Colony (JP-AMABC); and Probability based mining.

1. Introduction

Now-a-days, security is considered to be one of the major concerns and the issue is continuing to grow in intensity and complexity. Security is an aspect that is given top priority by all the political and government worldwide and is aimed at reducing the crime incidence [1]. As both the frequency and the complexity of crime are dramatically increasing these days, the occurrence of human errors and the huge amount of time required for the data analytics follow the same pattern [2]. Data mining is the process of discovering patterns in the data [3]. Data-mining techniques are often more powerful, flexible, and efficient for exploratory analysis than are statistical techniques [4]. Using data mining in crime mapping can help authorities identify frequent crime patterns and hotspots that can help in crime intelligence [5]. The goal of crime data mining is to understand patterns in criminal behavior in order to predict crime anticipate criminal activity and prevent it [3]. The crime data mining techniques can reduce the time and effort required for analyzing a large amount of data set, as well as extracting the useful information. Human analysts may typically take weeks to months to discover useful information from a large data set. As a result, investigators and other practitioners would spend less time on non-value-adding task; rather they can focus on more value-adding tasks based on the identified 'hidden' information in a large amount of data [2].

The Frequent Itemset Mining (FIM) has been an essential part of the data analysis and data mining [6]. It tries to extract information from the databases based on frequently occurring events. In this regard, an event, or a set of events, is deemed to be interesting if it occurs frequently in the data, according to a user-specified minimum frequency threshold [7]. Many techniques have been invented to mine databases for frequent events [7]. Marghny *et al* [8] proficiently presented the probability of frequent itemset (PFI) algorithm. The PFI algorithm mainly concentrated on reducing dataset searching area vertically and horizontally. Their proposed technique was used for scanning the database only

one time and processing different stages of large itemsets at the same time. The PFI offered reasonable performance but the accuracy of the system depended on a factor and while increasing the number of system nodes, the performance of the HoriVertical technique was better than the performance of the PFI.

Sujatha *et al* [9] proficiently presented a frequent itemset mining approach to forecast the offense locality by assessing the offense records by means of an Adaptive Mutation based Artificial Bee Colony (AMABC) technique, which made use of socio-monetary traits and grouping outcomes in the offense locality estimation procedure. Prominent among the forecast offense localities by the AMABC method was a sky-scraping offense locality which was estimated by extracting the models by means of the Association Rule Mining (ARM) algorithm. Although the collection of all frequent itemsets was typically very large, the subset that was really interesting for the user usually contained only a smaller number of itemsets. As a consequence, the mining task became rapidly intractable by the traditional mining algorithms which tried to extract all the frequent itemsets. In this context, the closed itemsets mining is a solution to this problem [10].

Vo *et al* [11] elegantly launched an algorithm to construct the frequent-closed-itemset pattern and also derived a concept for trimming the nodules present in the pattern in order to generate the rule. Thus, they evolved a technique for extracting the Most Generalization Association Rules (MGARs) from the pattern vastly. In the event that affiliation controls under different least Confidence qualities were to be mined, the lattice needed to be fabricated just once thus the re-mining time was altogether diminished. Despite the fact that it performed better, constructing a FCIL from FCIs was very drawn out, particularly when the quantity of FCIs was huge. The Association Rule mining is an important task of data mining that finds the probability of co-occurrence of items in a collection [12]. Interesting patterns often occur at varied levels of support. The classic association mining based on a uniform minimum support, such as Apriori either misses the interesting patterns of low support or suffers from the bottleneck of the itemset generation [13].

Karegar *et al* [14] effectively suggested a technique using the probability-trees to have a process on different type of data stored in the databases and give familiar information to system users. These patterns were designed using probability rules in decision trees and were cared to be valid, novel, useful and understandable. The use of the probability in the pattern made it so sensible, understandable and easy to be performed. By using the suggested patterns in data-mining, the system got efficient information about the data stored in its data-bases and used them in the best planning for special objectives. Discovering relations that connect variables in a database is the subject of data mining. Raghu *et al* [15] developed a Decision Support in Heart Disease Prediction System (DSHDPS) employing the data mining modeling technique, namely, probability. Using the medical profiles such as age, sex, blood pressure and blood sugar it was possible to predict the likelihood of patients getting a heart disease. They implemented the technique as web based questionnaire application.

We deliver a new Computing Probability on the Socio Economic Factors to detect the Crime Localities considering the Joint Probability depended on the AMABC-FCIL in this article. Frequent closed item set mining scheme depends only on the suggested algorithm which is the prime probability implemented in the following that database that is endangered to clustering procedure by considering the AMABC to set the like characteristics. The principles are produced with the help of the Association Rule Mining (ARM) and also the Joint Probability Function (JPF) is calculated to build the Frequent Closed Item set. Lattice rules that fulfill the support threshold after building the FCIL are picked as the recurrent closed item sets. Based on the rules fashioned, the test datasets are appraised and also inspected at last. Its function is better in the name of the calculation time, accurateness, memory assurance, lift and the number of rules.

The remnants of this work are systematized as below. The issue of redundancy and also accurateness are deliberated preceding Segment II. In Segment III, the anticipated Frequent Closed Item set Mining Procedure is elucidated in detail. In Sections IV, the imitation consequences are specified. Conclusions along with the future work are accessible in V and VI correspondingly.

2. Problem Definition

The Mining of frequent itemsets is a fundamental and essential problem in many data mining applications such as the discovery of association rules, strong rules, pattern predictions and many other important discovery tasks. In the AMABC-ARM technique [9], the high volume crime locations are predicted. While mining the locations, the technique requires more computation

time and scanning time, since the number of itemsets is large. So to avoid these negative aspects, AMABC is combined with modified FCIL to mine the location in less time [16]. Here we have mined only closed itemsets rather than the frequent itemsets. Mining closed itemsets is beneficial than mining frequent itemsets but still it suffers from the problem of redundancy and accuracy. To overcome those, here, probability based frequent closed itemsets mining technique is presented to get the richest information to improve accuracy.

3. Proposed Frequent Closed Itemset Mining Technique

In the proposed probability based frequent closed itemset mining technique, initially the attributes such as Household Income, Peoples under Poverty, Percentage of Population, Education, Crime Rate are grouped by means of AMABC after gathering them from the database. Then probability is computed for each rule generated using ARM and the rules which satisfy the predefined threshold are selected to construct the FCIL. Finally, FCIL is constructed with the consideration of sliding window to extract the frequent closed itemset and extracted rules are utilized to predict the crime location.

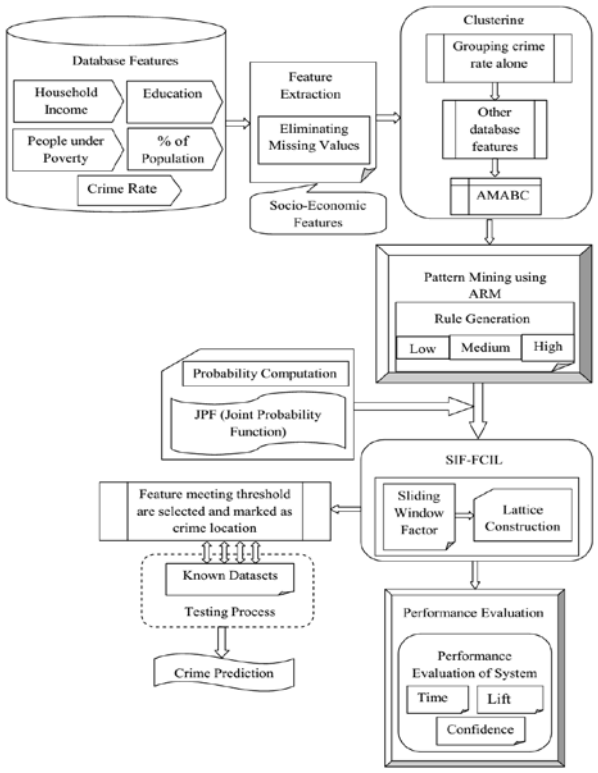


Figure.1 Overall Architecture Diagram for Crime Prediction

Let us consider the database $D=\{A_{ij}\}$; $i=1,2,\cdots W$, $j=1,2,\cdots Q$, which denotes the database attributes values. As the prediction accuracy may be decreased by the presence of the missing values, they are removed from the database and the resultant database is denoted by $D^{'}(M\times N^{'})$. The modified database is subjected to clustering process to group the similar attributes which will increase the performance of locality prediction.

Before entering in to the clustering phase, the crime values in the modified database are grouped by means of calculating the distance (dt). Since the crime values are located at the last column of the database ($j==Q^{'}$), the minimum (m) and maximum value (mx) of the last column ($j==Q^{'}$) are found out and they are clustered based on the difference between the maxima (mx) and minima (m) which is given in eqn. (1).

$$dt=(A_{ij})_{Q^{'}}^{mx}-(A_{ij})_{Q^{'}}^{m} \tag{1}$$

3.1 Clustering: Adaptive Mutation based ABC

After clustering the crime values, the remaining attributes are grouped with the help of AMABC [12] which utilizes the clustered crime values also. AMABC has three phases namely,

- Employed bee
- Onlooker Bee
- Scout bee

The generated fs and S are mapped $D^{'}$, for creating a matrix M is computed as follow,

$$M=(fs)^{T}S \tag{2}$$

The obtained M has the size of $R\times Q^{'}-1$ which is computed by multiplying the input vectors. fs is a one dimensional vector with R number of values with variable S that also is a one dimensional vector with $Q^{'}-1$ values. After that, each row vector value of M is multiplied by $D^{'}$ values by,

$$I_i=M\begin{pmatrix}11\ 12\ \ldots\ Q'-1\\ \ldots\ \ldots\ \ldots\ \ldots\\ R\ \ldots\ \ldots\ Q'-1\end{pmatrix}\times D^{'}\begin{pmatrix}11\ 12\ \ldots\ Q'-1\\ \ldots\ \ldots\ \ldots\ \ldots\\ W\ \ldots\ \ldots\ Q'-1\end{pmatrix}^{T} \tag{3}$$

After that, we find the fitness function for fs by exploiting the function as stated below.

$$FF=\min(F_k) \tag{4}$$

Where,

$$F_k=\left(\frac{1}{R^2-R}\left(\sum_{i=1}^{W}\sum_{\substack{i^{'}=1\\ i^{'}\neq i}}^{W}dt_{ii^{'}}\right)\right)^{1/2} \tag{5}$$

$$dt_{ii'} = \begin{cases} (dt_i - dt_{i'})^2; & if\ dt_i\ \&\ dt_{i'} \in c_1 \\ \dfrac{1}{(dt_i - dt_{i'})^2}; & otherwise \end{cases} \quad (6)$$

FF is computed fs using Eqn. (4) and we select the best fs which has the minimum FF value among the k number of generated fs. Next, we generate the new fs in employed bee phase by using the selected best fs. In employed bee phase the new fs are generated as given in eqn. (7)

$$fs_k^n = fs_{k\min} + rand(-1,1)(fs_{k\min} - fs_{k\max}) \quad (7)$$

The adaptive mutation function in the fs generation process will increase the ABC performance. The new adaptive mutation function in the AMABC is described in Eqn. (8):

$$AM = \chi\left(1 + rand\frac{(mx'-m')^{\omega} - ag^{\omega}}{\delta(mx'-m')^{\omega} - ag^{\omega}}\right) \quad (8)$$

$$\delta = \left(\frac{mx'-m'}{ag}\right)^{\omega} \quad (9)$$

Where, χ - represents Mutation probability, ω, φ - symbolize the mutation coefficient factors, mx', m', ag - denotes the maximum, minimum and average fitness of the food sources. Probability is computed to select the fs which are to be entered in to the onlooker bee phase.

$$p_j = \frac{FF_j}{\sum_{j=1}^{d} FF_j} \quad (10)$$

Where, p_j represents the probability of the j^{th} parameter. After computing p_j of all fs, average of eqn. (10) is calculated and fs which are greater than the average probability are selected for updation by means of the onlooker bee phase.

if $p_j > avg(p_j)$ *then* update *fs* using onlooker bee

The selected fs are also updated in the onlooker bee phase by using the neighborhood search and adaptive mutation Equation. The process is continued until the maximum iteration is reached. Once the termination criterion is met, the best fs are selected and it is replaced in eqn. (2) to get the final clustered result.

3.2 Rules Generation

The cluster centroid values (cc) are compared with the predefined factor (α) and the divided clusters named by setting priority as (i) Low (L) (ii) Medium (M) and (iii) High (H). The crime values clustered using eqn. (1) are also divided as L, M and H. Next to that, the rules (rl) are generated by combining both clustered values. In one combination, the support of each generated rl is compared with the minimum support value and whose support value is greater than α is selected. Using the selected rl , two combinations rl are generated and the same process is repeated up to five combinations.

3.3 Probability Computation

Based on the priority (pr) of the crime value, rl are listed out and the joint probability (jp) of each rl under each pr is computed. If the probability of rl satisfy the cutoff factor (cf) then rl is selected for the further process. Here, we have made the cutoff factor (cf) as 0. Most techniques use only the support and confidence to mine rl . To address the problem of redundancy, many techniques have been proposed. Most of them are based on user-defined templates or item constraints, interestingness measures to select only the interesting rl . Here, we have utilized the joint probability criterion for each rl which not only reduce the redundant rl but also help to mine more interesting rl . Because, low support and low confidence factors lead to the generation of more number of rl . The joint probability of each rule is calculated by eqn. 11 and eqn.12

$$P(rl_j) = \sum_{i=1}^{O} \beta_i \Big/ O \quad (11)$$

$$\beta_i = \begin{cases} 1, & rl_j = rl_i \mid j \neq i \\ 0, & else \end{cases} \quad (12)$$

O- Total number of rules under corresponding priority level (*pr*)

3.4 Mining frequent closed itemsets by FCIL construction

The Sliding window is moved on the selected rl and subsequently the FCIL is constructed to extract the rl . At first, the selected rules after calculating jp and its corresponding support of *s* are obtained and the root node of the FCIL is set as empty node (i.e $r = e_n$). The Number of attributes in each rule is counted and the priority levels are

named low (l_{h-2}), high (l_h) and medium (l_{h-1}). For each rule, priority (pr) of the last attribute is checked. Since the last attribute has three levels of pr, three FCILs are constructed. After checking pr of the last attribute, l_h among the remaining attributes is checked. The attribute which has l_h is added as the child node (c_n) of the root node (r_n) directly. If no attribute has l_h, e_n is added to r_n. This process is continued until the number of the attributes of the rule is reached. If two attributes have same pr, then both attributes are added as c_n. Once the count of the number of attributes in rl is reached, next rl is taken into consideration and the same process is continued. When the number of rl is greater than one, then pr of the last attribute in the same rl is checked. If pr of the last attribute already exists, then the already existing tree is modified, else a new tree is constructed. Finally, a support factor threshold (τ) has been set and the features in the lattice tree (T) which meet the threshold factor (τ) are selected. Subsequently, the corresponding attributes for the selected features are marked as frequent closed itemsets and are used for crime location prediction.

3.5 Sliding Window Factor

While employing the lattice tree technique, several novel collections linked linear forecast hassles crop up which are marginally divergent from the latest problems. The daunting challenges may be mitigated by means of bringing in a family of "sliding" window approaches. The amazing advantage of applying the sliding window aspect is the incredible cutback in the rules in relation to the parallel techniques. In the sliding window method, the support counts for the entire itemsets are appropriately estimated and brought under the umbrella of three itemsets categorized as the low, high and medium rows. The database traits such as the household income, people under poverty, percentage of population, education and crime rate are duly taken into account. The created rules subjected to the sliding window elegantly bring in the abridged rules, thereby facilitating the effortless and timely forecast of the crime rate. Though there is a host of techniques doing the elegant rounds dedicated for the forecast of the crime rate, the integration of the novel sliding windows technique takes the approach to the zenith of consistency.

3.6 Testing Process

The suggested technique involves the clustering process with the able assistance of the Joint Probability based Adaptive Mutation based Artificial Bee Colony (JP-AMABC) technique, in which the frequent intimately linked itemsets are grouped. In the new-fangled approach, the dataset traits gathered from the databases including the Household Income, Peoples under Poverty, Percentage of Population, Education and Crime Rate have to be subjected to the clustering function.

The data traits are grouped by way of an intimately linked approach. The rules are effectively produced for the related datasets with the help of the Association Rule Mining (ARM). Subsequently, the captioned traits are categorized into three distinct types such as the High, Low and Medium priority values, subsequent to the probability evaluation by means of the joint probability function, which is carried out prior to initiating the task of the FCIL creation with sliding window to significantly scale down the number of rule generations. During the testing phase, several pre-specified datasets are gathered from the database and thereafter the data traits are classified in accordance with the priorities such as the High, Low and Medium category. The test datasets are assessed and analyzed with the rules created, which leads to the achievement of superior outcomes by the novel approach by means of the performance of the investigation function. For instance, if 100 datasets are shortlisted for the testing procedure, each and every data is analyzed with the produced rules independently. Thereafter, the outcomes are appraised and contrasted in respect of the authentic technique and the test data outcomes for arriving at the precise outcome.

3.7 Performance Evaluation using Benchmark Functions

The standard functions are effectively employed to estimate the efficiency in execution. In the document, a cluster of four standard functions are extensively utilized to appraise the performance of the innovative Joint Probability based Adaptive Mutation Based Artificial Bee Colony algorithm (AMABC) approach with the modern Artificial Bee Colony (ABC) technique. Certain vital standard functions made use of in the novel are furnished as follows.

a) ***Axis parallel hyper-ellipsoid function***

It is identical to the function of De Jong and is otherwise known by the name 'weighted sphere model'. It glistens with the appealing traits of being

nonstop, convex and unimodal. Broadly, it may be characterized as detailed below.

$$g(y) = \sum_{k=1}^{n} (k \,.\, y_k{}^2) \qquad (13)$$

The test area is typically controlled within the hypercube $-50 \le y_k \le 50, k = 1,2,...,n$. Global minimum $g(y) = 0$ is obtainable for $y_k = 0, k = 1,2,...,n.$

b) *De Jong's function*

It represents one of the easiest, convenient and straightforward test criteria. It is endowed with the qualities of being incessant, convex and unimodal. It may be generally defined as follows.

$$g(y) = \sum_{k=1}^{n} x_k{}^2 \qquad (14)$$

As a rule, the test area is limited to the hypercube $-50 \le y_k \le 50, k = 1,2,...,n$. The global minimum $g(y) = 0$ is realizable in respect of $y_k = 0, k = 1,2,...,n.$

c) *Rotated hyper-ellipsoid function*

An enlargement of the axis parallel hyper-ellipsoid represents the Schwefel's function. In relation to the coordinate axes, the captioned function effectively generates the rotated hyper-ellipsoids. It has the attributes of being uninterrupted, convex and unimodal. It can be broadly represented by the following expression.

$$g(y) = \sum_{k=1}^{n} \sum_{l=1}^{m} y_l{}^2 \qquad (15)$$

The test area is normally limited to the hypercube $-50 \le y_k \le 50, k = 1,2,...,n$. The global minimum $g(y) = 0$ is obtainable for $y_k = 0, k = 1,2,...,n.$

d) Griewangk's function

This function is identical to that of the Rastrigin. It boasts of a host of extensive local minima evenly disseminated. It can be defined by means of the following expression.

$$g(y) = \frac{1}{4000} \sum_{k=1}^{n} y_k{}^2 - \prod_{1}^{n} \cos\left(\frac{y_k}{\sqrt{k}}\right) + 1 \qquad (16)$$

The test area is habitually limited within the hypercube $-50 \le y_k \le 50, k = 1,2,...,n$. The global minimum $g(y) = 0$ is achievable in respect of $y_k = 0, k = 1,2,...,n.$

Table 1. The performance of the Joint Probability based AMABC method technique

SI. No	Benchmark Functions	Support Counts	Basic ABC method	Performance of ARM-AMABC method	Performance of SIF-FCIL technique	Performance of Joint Probability based AMABC method
1	Axis parallel hyper-ellipsoid function	2	198	157	174	98
		3	328	334	48	28
		4	99	97	47	10
		5	167	115	63	2
2	De Jong's function	2	45	37	29	15
		3	65	62	18	7
		4	41	31	18	4
		5	67	36	25	1
3	Rotated hyper-ellipsoid function	2	150	147	81	40
		3	216	249	54	16
		4	183	123	53	8
		5	151	146	0.003	1
4	Griewangk's function	2	7	0.006	0.002	-0.002
		3	4	0.003	0.003	-0.001
		4	3	0.006	75	-0.0007
		5	5	0.008	0.005	-0.001

It is heartening to note that the Joint Probability based AMABC charismatically ushers in an amazing pace of convergence. In this regard, a host of standard functions are in vogue devoted for the appraisal of the convergence rates. A further quantitative appraisal is carried out regarding the number of generations essential by both the techniques to ascertain the global minima. It is illustrated without an iota of ambiguity that the epoch-making JP-AMABC technique exhibits the requisite skills essential for the achievement of the minimum values vis-à-vis peer approaches, while appraising the efficiency in performance by means of the captioned standard functions mentioned elsewhere.

4. Experimental Results and Discussion

The proposed crime location prediction technique based on probability is implemented in the working platform of JAVA (JDK 1.6) with Intel dual core Processor, Windows 7 Operating System, 2GB RAM and 3.06 GHz speed of CPU.

4.1 Database Description

Utilize the UCI Machine Learning Repository-Communities and Crime Data Set furnished in [17] to perform the crime location estimation. The data integrates the socio-economic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR. The UCI dataset dimension is characterized as 1994x127, in other words, it is home to 1994 rows and 127 columns of data. At the outset, we carry out the feature extraction procedure on the captioned UCI dataset. In the feature extraction the columns having the no value are summarily removed from the dataset. Thus, the dataset dimension is dwindled to 1994x102, in other words, the column dimension is decreased to 102 from 127. The resultant abridged 1994x102 dataset is employed in the supplementary clustering and FCIL frequent rule mining procedure.

4.2 Performance Analysis

The performance of the proposed technique is evaluated in terms of the computation time, confidence, lift and number of rules and also the system performance is evaluated with accuracy and memory to prove the efficiency of our method.

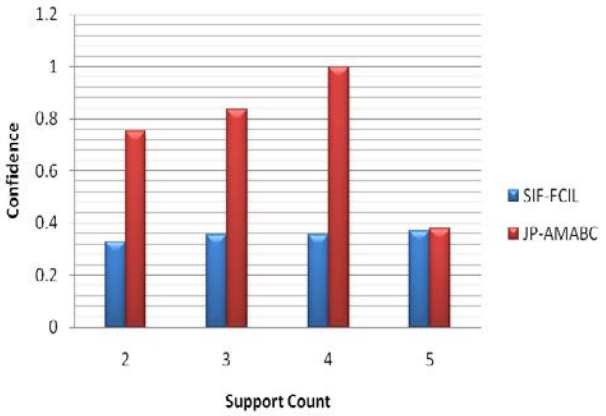
4.2.1 Time

To analyze our proposed method performance, the threshold is varied from 2 to 5 and the corresponding proposed and comparison mining methods for the database features like Percentage of Population, Household Income, Peoples under Poverty, Education and Crime Rate results are obtained in terms of operational time. Here, the time is calculated in terms of milliseconds and the support count is taken as 2, 3, 4 and 5. While observing the metrics, it shows that the operation time decreases as the support count increases. The operation time for our proposed technique and the existing SIF-FCIL by varying the support count shows increasing amount of time requirement for our proposed technique than the existing one. Here, the operation time for the proposed technique is more than that of SIF-FCIL due to more processing performed.

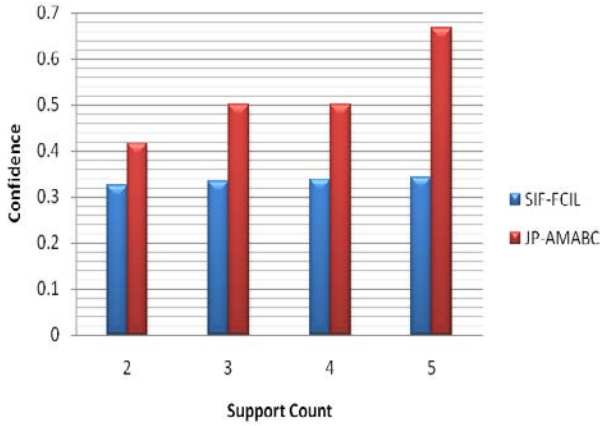
4.2.2 Confidence

It is a measure of the intensity of the consistency of the related rule. Fig. 2 shows the confidence of the rules obtained using both the proposed and the SIF-FCIL techniques by varying the support count from two to five. Confidence value increases as the support count increases. The proposed technique yields higher confidence than that of SIF-FCIL. The Confidence attained for the proposed technique is varying while increasing the support counts. It shows that proposed technique mines the most interesting rules.

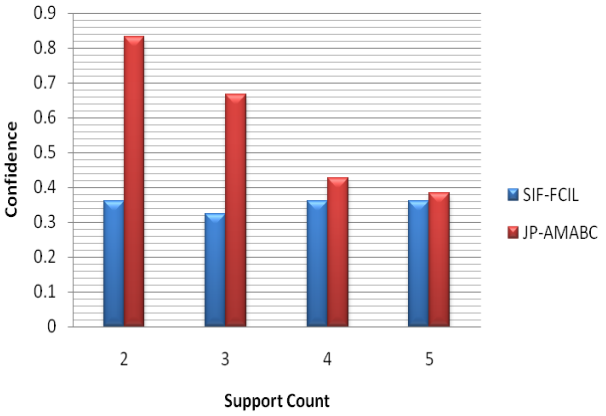
Discussion: Fig. 2 shows the compared results of the proposed technique and the existing SIF-FCIL in terms of confidence is obtained for all the database features like Percentage of Population, Household Income, Peoples under Poverty, Education and Crime Rate. The proposed shows better results than the existing SIF-FCIL. The proposed method provides greater confidence value even for higher support counts than the existing ones. For the database features like Peoples under Poverty and education, the confidence level is decreasing while on increasing the support counts but not less than the existing technique. The technique maintains the maximum confidence level for the features like Percentage of Population, Household Income and Crime Rate up to the support counts 2 to 4 and is reduced at support count 5 only for the Percentage of Population and Crime Rate features.



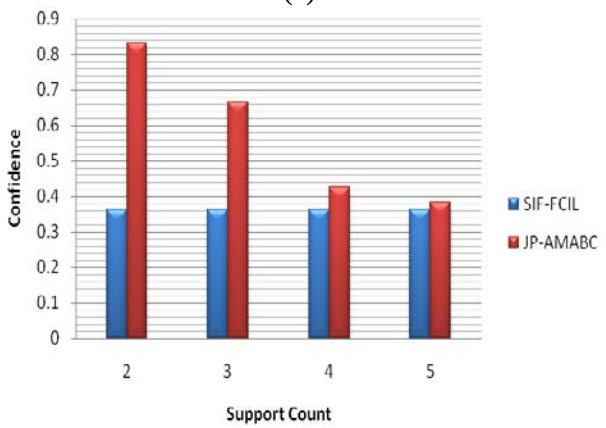
(a)



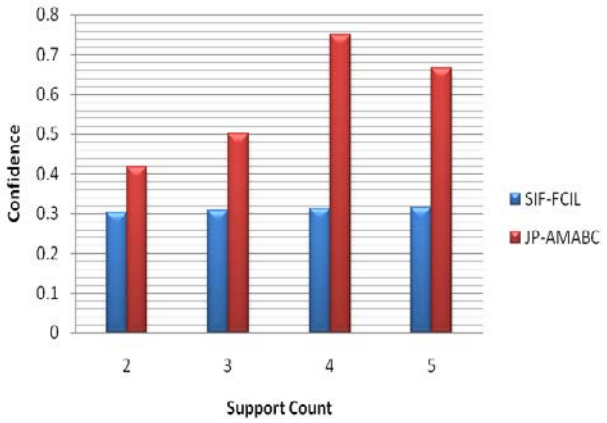
(b)



(c)



(d)



(e)

Figure.2 Comparison results of our proposed JP-AMABC technique and existing SIF-FCIL in terms of Confidence with varying support counts obtained for database features like (a) Percentage of Population, (b) Household Income, (c) Peoples under Poverty, (d) Education and (e) Crime Rate

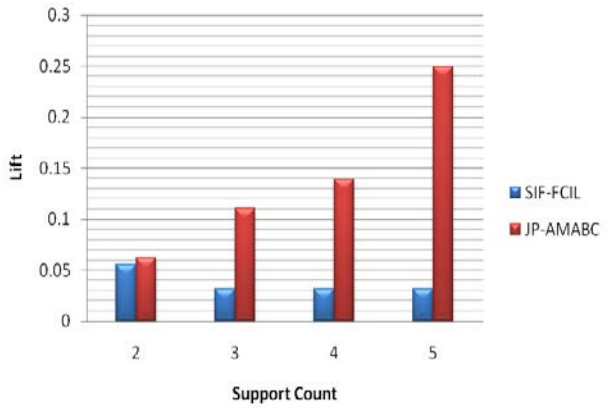
4.2.3 Lift

Lift of an association rule is defined as the ratio of confidence to the share of the entire isolates which come under the ambit of RHS. Fig 3 shows the lift values for our proposed technique and the existing SIF-FCIL technique by varying the support count. On looking at the line indicating the proposed technique the curve increases as the support count increases but this does not happen in the existing technique. Also the lift values of the rules obtained using the proposed technique is higher than that of SIF FCIL based technique.

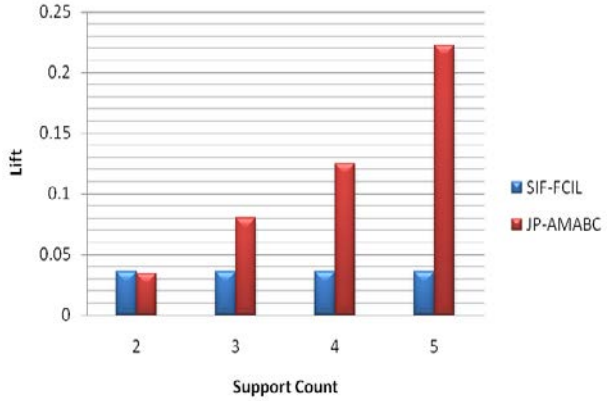
Discussion: Fig. 3 shows the compared results of the proposed technique and the existing SIF-FCIL in terms of the lift. The proposed technique shows better results than the existing SIF-FCIL. Also, it is seen that with the increase in the support count the lift ratio value also gets increased for the proposed technique while the existing technique maintains a specific lift value without any improvement even after for varying the support values.

4.2.4 Number of Rules

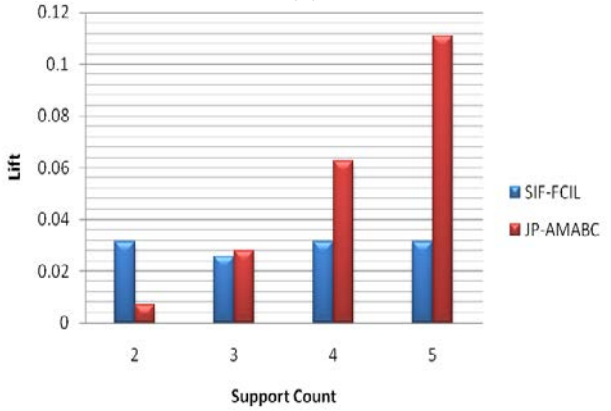
Fig.4 shows the number of rules extracted for the proposed technique and the existing technique SIF-FCIL by varying the support count. We can observe that the number of rules moderately equal when compared with SIF-FCIL. The number of rules generation is same for all the database features as because, the same number of rules were only generated during the rule generation process.



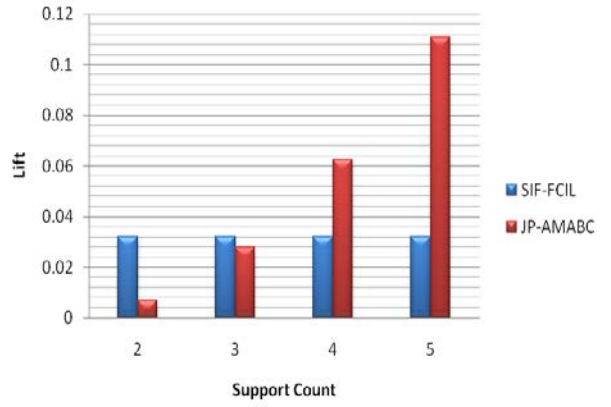
(a)



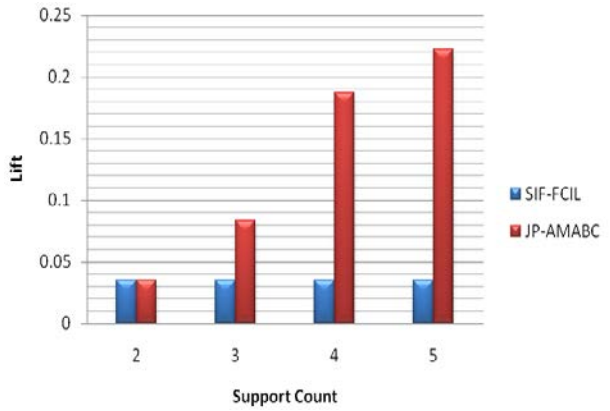
(b)



(c)



(d)



(e)

Figure.3 Comparison results of our proposed JP-AMABC technique and existing SIF-FCIL in terms of Lift with varying support counts obtained for database features like (a) Percentage of Population, (b) Household Income, (c) Peoples under Poverty, (d) Education and (e) Crime Rate

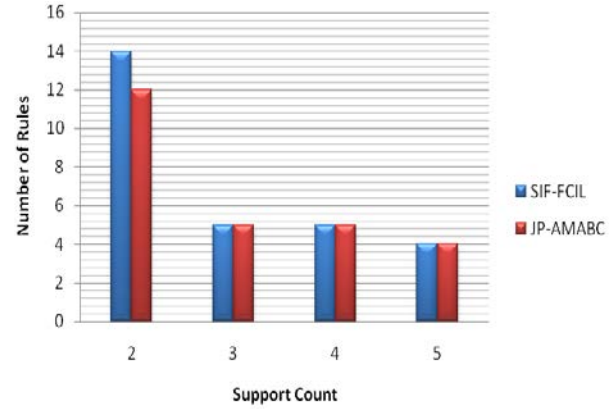


Figure.4 Comparison results of our proposed JP-AMABC technique and existing SIF-FCIL in terms of Number of Rules with varying support counts

Discussion: Fig. 4 shows the compared results of the proposed technique and the existing SIF-FCIL in terms of the number of rules. The proposed technique and existing system yields same number of rules.

4.2.6 Accuracy

To ensure the performance of our system the prediction accuracy along with the memory space requirement is compared with the existing techniques.

Fig. 5 shows the accuracy for the proposed technique and the existing technique SIF-FCIL by varying the support count. When the support count is increased, number of rules extracted is decreased. As the number of rules is decreased, the prediction accuracy is also increased. Thus the prediction accuracy of the proposed technique is higher than that of existing SIF-FCIL.

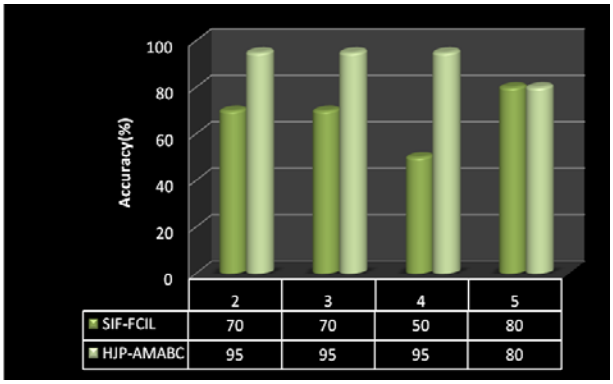


Figure.5 Comparison results of our proposed technique and the existing SIF-FCIL technique in terms of accuracy

Discussion: Fig. 5 shows the compared results of the proposed technique and the existing SIF-FCIL in terms of the accuracy. The proposed technique shows better results than the existing SIF-FCIL. The maximum value of accuracy is obtained for our proposed method on the support counts below 5, (i.e.,) 95% accuracy can be achieved for low support counts projecting our proposed method to be better than the existing SIF-FCIL. It is clearly seen that the Joint Probability based AMABC method produces 95% system accuracy for the thresholds 2, 3 and 4 showing the effectiveness of our technique.

4.2.7 Memory

Fig. 6 shows the memory space required for the proposed technique and the existing technique SIF-FCIL by varying the support count. When the support count is increased, the space requirement is decreased. Since the proposed technique yields number of rules without redundancy, it requires lower memory space than the existing SIF-FCIL.

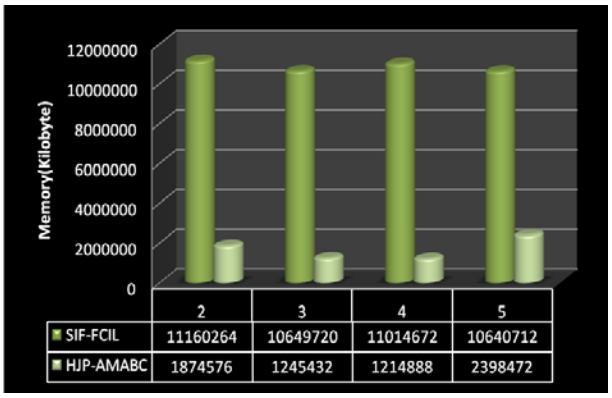


Figure.6 Comparison results of our proposed technique and the existing SIF-FCIL technique in terms of memory space

Discussion: Fig. 6 shows the compared results of the proposed and the existing SIF-FCIL in terms of the memory requirements. The proposed methods show better results than the existing SIF-FCIL which can be verified by means of the above graph. As the existing SIF-FCIL utilizes nearly the same memory space for all the support counts but our proposed method takes only lesser memory space requirement which could also be decreased on the increase of the support counts.

5. Conclusion

Here, a novel crime location forecast method by means of calculating probability and FCIL algorithm is launched to locate the elevated crime locations from the UCI data. Initially, the socio economic features are gathered and clustered by means of the AMABC algorithm. The optimal clustered outcomes from the AMABC algorithm are furnished to the FCIL to build high frequent crime locations. Our ambitious crime location forecast approach has shown its mettle by profitably forecasting the crime locations in accordance with the extracted outcomes accompanied by the diminished calculation time-frame in relation to the modern crime location forecast methods. What is more, our crime location forecast scheme with SIF-FCIL performance is assessed and contrasted with the current SIF-FCIL technique. The charismatic outcomes from the diverse mining algorithms have unequivocally established the fact that our milestone crime location forecast technique has ushered in excellence in performance vis-à-vis its peers in respect of the memory and accuracy.

6. Future Work

In our proposed method, a Joint Probability based Adaptive Mutation based Artificial Bee Colony (AMABC) algorithm is used for clustering the closely connected features. The proposed technique is usually incorporated with several benchmark functions and it is verified mathematically for its enhanced performance evaluation. Further, the actual performance evaluation may be enhanced by means of incorporating additional number of benchmark functions so that we could obtain more accurate evaluated values for the performance measure of any technique. In addition, future works may be carried out for reducing the time consumption by incorporating certain other Optimum clustering methods which could be attained by employing the Optimum Based Fuzzy C Means (OBFCM) clustering.

References

[1] Malathi and S. S. Baboo, “An Enhanced Algorithm to Predict a Future Crime using Data Mining”, *International Journal of Computer Applications*, Vol. 21, No. 1, pp. 1-6, 2011

[2] Rad, A. Ashoury, Y. Ham and Y. Song, “Evaluation, Prediction, and Visualization of Spatio-Temporal Crime Patterns in Washington DC Area”, 2012.

[3] Wang, Rudin, Wagner and Sevieri, "Learning to detect patterns of crime", *Machine Learning and Knowledge Discovery in Databases, Springer Berlin Heidelberg*, pp. 515-530, 2013.

[4] Hochachka, Caruana, Fink, Munson, Riedewald, Sorokina and Kelling, “Data-Mining Discovery of Pattern and Process in Ecological Systems”, *The Journal of Wildlife Management*, Vol. 71, No. 7, pp. 2427-2437, 2007.

[5] Sandig, Somoba, Concepcion and Gerardo, “Mining Online GIS for Crime Rate and Models based on Frequent Pattern Analysis”, *In Proceedings of the World Congress on Engineering and Computer Science*, Vol. 2, pp. 23-27, 2013

[6] C. Bhadane, K. Shah and P. Vispute, “An Efficient Parallel Approach for Frequent Itemset Mining of Incremental Data”, *International Journal of Scientific & Engineering Research*, Vol. 3, No. 2, 2012

[7] Moens, Aksehirli and Goethals, “Frequent Itemset Mining for Big Data”, *In IEEE Proceedings of International Conference on Big Data*, pp. 111-118, 2013

[8] Marghny and Hosam, “A fast Parallel Association Rule Mining Algorithm Based on the Probability of Frequent Itemsets”, *International Journal of Computer Science and Network Security*, Vol. 1, pp. 1-11, 2013

[9] Sujatha and Ezhilmaran, "An Adaptive Method for Analyzing and Predicting the Crime Locations by means of AMABC and ARM", *Journal of Theoretical and Applied Information Technology*, Vol. 59, No. 1, pp. 45-56, 2014

[10] Bonchi and C. Lucchese, "On closed constrained frequent pattern mining", *In IEEE Proceedings of Fourth International Conference on Data Mining*, pp. 35-42, 2004.

[11] B. Vo, T. P. Hong and B. Le, "A lattice-based approach for mining most generalization association rules", *Knowledge-Based Systems*, Vol. 45, pp. 20-30, 2013.

[12] S. Sharma and V. Chopra, “Association Rule Mining: A Multi-Objective Genetic Algorithm Approach Using Pittsburgh Technique”, *International Journal of Recent Technology and Engineering*, Vol. 2, No. 4, pp. 121-124, 2013.

[13] K. Wang, Y. He and J. Han, “Mining Frequent Itemsets Using Support Constraints”, *In Very Large Databases*, pp. 43-52, 2000.

[14] Karegar, Isazadeh, Fartash, Saderi and Navin, “Data-mining by the Probability-based Patterns”, *In Proceeding of the 30th International Conference on Information Technology Information Technology Interfaces*, pp. 353-360, 2008.

[15] D. Raghu, Srikanth and R. Jacub, “Probability based Heart Disease Prediction using Data Mining Techniques”, *International Journal of Computer Science & Technology*, Vol. 2, No. 4, pp. 66-68, 2011

[16] R. Sujatha and D. Ezhilmaran, “A new efficient SIF-based FCIL (SIF– FCIL) mining algorithm in predicting the crime locations”, *Journal of Experimental & Theoretical Artificial Intelligence*, Vol. 28, No. 3, pp. 561-579, 2015.

[17]http://archive.ics.uci.edu/ml/machine-learning-databases/communities/