RESEARCH REVIEW JOURNALS
Volume-04
Issue-04
April -2019

ISSN: 2455-3085 (Online)
RESEARCH REVIEW International Journal of Multidisciplinary
www.rrjournals.com [UGC Listed Journal]

Unsupervised Approach for the Detection of Attacks in Recommender System

Ms. Nirmal Kaur

Research Scholar, Department of Computer Science and Application, Sant Baba Bhag Singh University

Introduction

Data mining is the process of finding confusing, patterns and relationships between large data sets to predict results. Using a variety of strategies, you can use this information to increase revenue, reduce costs, improve customer relationships, reduce risks and more. Privacy preserving data mining (PPDM) refers to a data mining environment that protects sensitive information from disclosing unsolicited or unauthorized information. Many common methods of data mining analyze and generate statistical data with privacy concerns that protect data from exposure.

Privacy measures include some form of data conversion by reducing representation granularity. This results in loss of efficiency of mining algorithms. The most widely used data modification techniques are the random method, anonymity model, L-diversity, distributed privacy protection. The PPDM research center has developed solutions based on a variety of approaches. A major challenge that needs to be addressed in any of these privacy practices is the ability of the data mining algorithm to extract the required patterns from the site even after anonymity. The performance of the classifier is analyzed by the parameters such as classification accuracy, average Positive Predictive value and average sensitivity with classifiers designed by K anonymization features, PSO features and PSO suppression features.

- Memetic improvement with the ACO algorithm is suggested to improve silence. Carrying out location searches for solutions generated by a memetic algorithm used by ACO. Qualification function is set to increase the accuracy of the sections. Tests are performed on an adult data set. Using selected features the facilitator is trained. Separator performance is analyzed by parameters such as phase accuracy, average Positive Predictive value and intermediate sensitivity with dividers designed for GA features, proposed ACO features .

The requirement for data mining confidentiality has become increasingly important in recent times due to the increasing need to maintain personal information. In this chapter, the attribute compression method uses the Particle swarm optimization algorithm and the standard optimization method is used to obscure it. The relevant features found by the PSO and the anonymous methods are verified using a split algorithm. The results of K-Anonymization irritate the selection of the PSO feature improved in terms of Program accuracy, Positive Guidance and Sensitivity.

Several strategies such as anonymity of k, l variability, Randomization and cryptographic methods have been suggested to make PPDM. To further enhance privacy, this chapter uses the PSO algorithm to compress and standardize the anonymity of K. The importance of the issue of privacy has been debated in a number of domains such as the Internet, Statistical Disclosure Control (SDC) and cryptography. Data entry methods have been developed to generate information to support a number of domains such as marketing, weather forecasts, medical diagnoses and national security. However, it is challenging to extract certain types of data without violating the privacy of the data owner. For example, the issue of mine patient data is a major factor in the use of health care.

Concerns about privacy are increasingly urgent as data mining spreads in a more complete way (Kavitha & Vanathi 2014). Privacy refers to the right of individuals to keep information confidential so as not to disclose it to others. Although privacy and confidentiality are often used as synonymous, there are significant differences between the two terms. Confidentiality is associated with an individual while confidentiality is associated with non-disclosure of information other than those authorized.

"Drawbacks of randomization technique include:

1) it is not adequate when database involves many features
2) it is an extremely slow method as data collectors collect data from data providers
3) data providers append some noise in data, so reordering data takes a longer time (Dhandeet al., 2013).

The drawbacks of Generalization-based k-anonymization is that it experiences failure in higher dimensional data because of the curse of dimensionality. It causes data loss because of uniform distribution assumption and the database does not maintain confidentiality.

Optimization refers to the procedure of discovering the best way of using available resources, however simultaneously, no violation of any of the needed conditions. It tries to maximize favorable characteristics and decreases on-favorable characteristics (Stanarevic 2012). Users typically demand that a practical minimization method ought to satisfy particular requisites:

- Capability for handling various kinds of issues.
- Easy to utilize with fewer control parameters.
- Excellent convergence method to global minimum in consecutive independent trials.

The above are basis of numeric analysis of data. The methods are primarily combinatorial and may be expressed as discrete optimization issues. The objective of almost all data mining jobs lends themselves to discrete NP-hard optimization issues. Apart from the problem of complexity, huge scale of real-life data mining issues is a problem that arises from optimization-based datamining research (Shi et al., 2011).

SI is a comparatively newer inter-disciplinary domain of research and this has obtained a lot of popularity recently. Domains which are apart of the field are inspired by collective intelligence which emerges from social behavior of insects such as bees or wasps. When functioning as a community, the insects even with restricted intelligence can cooperate and perform several tasks required for survival. Issues such as the discovery as well as storage of food, choosing as well as picking up material for future use requires detailed planning, and are resolved by insect colonies with no supervisor or controller present. PSO is a popular SI protocol for global optimization over continuous search space. PSO has obtained the interest of other researchers across the world leading to several variations of the fundamental protocol and other variable automation schemes.

Method to derive

PSOA is a strong stochastic optimization technique expressed based on the progress and intelligence of swarms. PSO is associated with the notion of social behaviour for resolution of problems. Kennedy and Eberhart introduced this concept in 1995. It uses a set of particles that denote swarms flying around in search space which provides optimum solutions. Every particle is regarded as a point in N-dimensional spaces which modify their flying habit according to their own experiences as well as the experience of the swarm entirely (Cao et al. 2007 and Latiffet al. 2007).

Every particle keeps track of the coordinates in the solution space which is associated with the best position arrived at so far by the particle. The value is understood as the personal best, *pbest*. The best value that is followed by PSOA is the best position reached by the neighbouring particles and this is called the *gbest*.

The basic concept of PSOA is in the acceleration of every particle toward pbest and gbest positions and the random weighted acceleration at each time step is given in Figure 3.1.

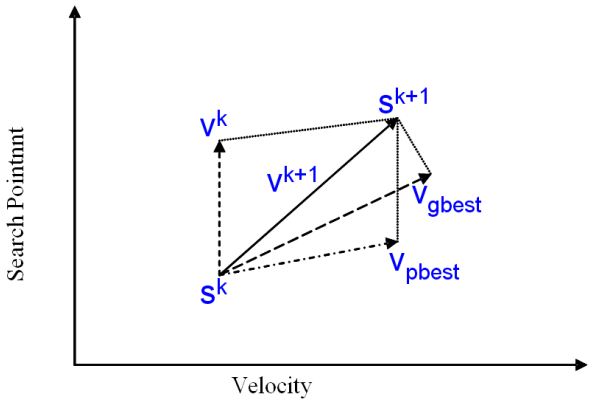

Figure 4.1 Concept of modification of a searching point by PSO

wherein, *Sk* Current searching point

S^{k+1} Altered searching point

V k Current velocity

V^{k+1} Altered velocity

pbest V Velocity on the basis of pbest

gbest V Velocity on the basis of gbest

Every particle attempts to modify the position via use of data provided:

• current position,

• current velocity,

• distance between current position and pbest,

• Distance between current position and gbest.

PSOA is a population based heuristic search technique which is generally used resolution of NP-hard optimization problems, inspired by the social activity of birds, bees or fish. All individuals in the swarm are denoted by vectors in multi-dimensional search space. The vector has an assigned vector that determines subsequent movement of the particle and is known as the velocity vector. PSOA decides the method of updating velocity of particles. All particles update their velocities on the basis of current velocity as well as best position visited as of then, as well as on the basis of the global best position visited by the swarm (Khan et al. 2010).

PSO Algorithm

Step 1 : Randomly initialize the position and velocity of each particle

Step 2 :particle fitness evaluation

if fitness of xi>pbesti

pbesti = xi

if fitness of pbesti>gbesti

gbesti = pbesti

Step 3 : Updated the velocity of particle *i*

Step 4 :if stopping criterion is not met, continue Steps 2 and 3.

Step 5 : Return gbest and its fitness values.

PSOA procedure is then iterated a certain number of times or till a minimal error based on favoured performance index is reached. It has been proven that the simplistic model is capable of dealing with hard optimization issues in an efficient manner. PSOA was built for real valued spaces initially, but several issues are defined however for discrete valued spaces wherein the domain of parameters is finite. Classical instances of such issues include: integer programming, scheduling as well as routing. In the simulation process, By initializing the population number equals to 30, the maximum iteration as500, the inertia value assigned to be 0.5, the cognitive and social parameterc1 & c2 as 2, fitness as RMSE the experiment is conducted . The algorithms were run ten times separately and the average values are used to do the comparison in the experiment.

Results of PCO

In this section, feature selection using PSO optimization algorithm and Generalization technique is employed. K-anonymity is accomplished by generalization and suppression of the original dataset. To validate the results of feature selection of Particle Swarm Optimization (PSO) and anonymization, classification accuracy is measured. K-Anonymization outrages PSO feature selection which is evaluated in terms of Classification accuracy, Positive Predictive value and Sensitivity. For different levels of k anonymity experiments were performed and the results achieved are evaluated. The adult dataset from UCI machine learning repository is used for evaluating the results. The values for classification accuracy, average Positive Predictive value and average sensitivity are compared and are shown in Table.

K Anonymization size	K Anonymization	PSO Optimization	PSO Suppression Only
No Anonymization	87. 99	85. 7351	83. 3281
K= 10	85. 72	83. 4634	82. 0608
K= 15	84. 72	83. 0583	80. 2289
K= 20	83. 4	81. 9866	80. 2289
K= 25	82. 66	81. 1503	79. 6517
K= 30	82. 14	81. 747	79. 6517
K= 35	82.23	80.535	79. 1341
K= 40	81. 84	81. 4897	78. 6891
K= 45	81. 41	80.5738	78. 359
K= 50	81. 7	80. 7474	77.8937

Table 4.1 Classification Accuracy

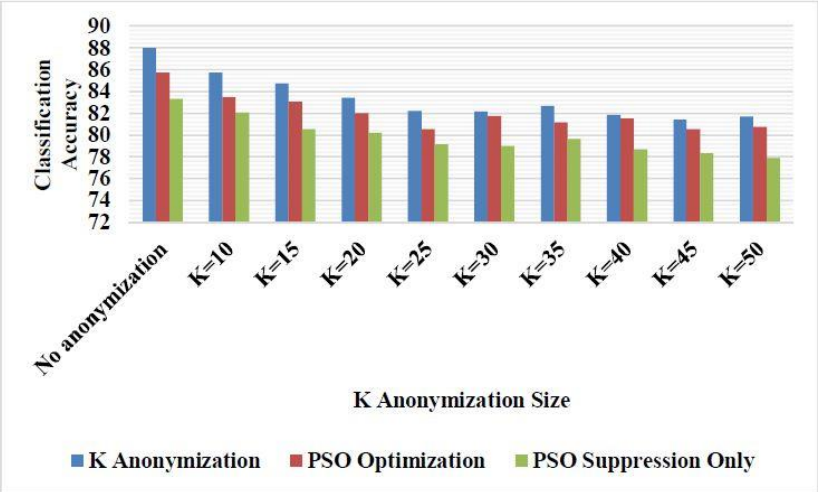

Figure 4.2 Classification Accuracy

From the Figure 4.2, it can be observed that the K anonymization has higher classification accuracy by 4.36% for K=10, by 3.87% for K=30and by 4.77% for K=50 when compared with PSO Optimization and by1.69% for K=10, by 3.4% for K=30 and by 3.59% for K=50 when compared with PSO suppression only.

The attribute suppression technique using Particleswarm optimization algorithm and a generalization technique is employed for anonymization. K-Anonymization is achieved through Suppression and generalization. To validate the results, classification accuracy of Particle Swarm Optimization (PSO) and k anonymization technique is measured.

The performance of the classification algorithm is analyzed by the parameters such as classification accuracy, average Positive Predictive value and average sensitivity and compared with parameters of K anonymization features, PSO features and PSO suppression features. Results show that the K anonymization has higher classification accuracy by 4.36% for K=10, by3.87% for K=30 and by 4.77% for K=50 when compared with PSO Optimization and by 1.69% for K=10, by 3.4% for K=30 and by 3.59% for K=50 when compared with PSO suppression only.

Feature Selection Through Ant Colony Optimization

Benefits of the fundamental PSOA are its basis intelligence in terms of solving a problem and its application in both scientific researches as well as for engineering purposes. PSOA has no overlapping or mutation computation. But the Drawbacks of the fundamental PSO protocol include:

1) The technique is vulnerable to partial optimism that causes regulation of speed as well as direction to be fuzzy,
2) The technique is not able to figure out the problem of non-coordinate system, like solution to the energy field as well as the moving rules of the particles in energy field, and
3) The technique is not able to figure out the issues of scattering as well as optimization (Selvi&Umarani 2010).

ACO is a popular optimization algorithm which reflects the food searching behavior of ants.

It is popularly known for its advantages that include:

1. Inbuilt parallelism,
2. Fast detection of new solutions,
3. Used to solve the Dynamic Problems.

By taking the advantages of ACO, this chapter deals with the selection of features through ACO optimization algorithm. Experiments are conducted using adult data set available in online repository. The Result implies that the ACO algorithm improved the performance with considerable amount in terms of classification accuracy, Average Positive Predictive value and Average Sensitivity than PSO optimization.

PPDM should protect against the disclosure of sensitive data at the time of publication of individual data. For maintaining privacy, several methods are suggested for modification or transformation of the data. For preventing the misuse of data, data is anonymized. The methods for PPDM

have their basis in cryptography, data mining, as well as information hiding(Agrawal&Srikant 2000). Generally, statistics-based as well as crypto-based methods are utilized for handling PPDM. In statistics-based method, data owners sanitize the data via perturbation or generalization prior to publishing.

The benefit of statistics-based method is efficient because it can handle huge volumes of data sets (Malinet al. 2011). In crypto-based PPDM method, data owners cooperate to execute specifically formulated data mining protocols(Singh et al. 2010). Although the protocols attain verifiable privacy protection, as well as improved data mining performance, it is vulnerable to scalability problems (Sharkey et al. 2008).

SI is inspired by swarms of social insects like ants or termites that display intelligent social behavior. Ants, for instance, communicate in an indirect way with their environments through deposition of a substance known as pheromone. Path with greater pheromone content has more probability to be selected and thereby, reinforced, and those paths which are not selected lose pheromone because of evaporation. This indirect communication is called stigmergy, and hits helps ants in finding shortestpaths to food (Martens et al. 2007).

Methodology

ACO utilizes artificial ants which cooperate for finding goodsolutions for discrete optimization issues. The software agents imitate theforaging activity of the biological ants in discovering the shortest route to thefood source. The first protocol follows how ants iteratively build solutionsand add pheromone to the routes associated with these solutions. Routeselection is a stochastic process that has its basis in two variables, pheromoneas well as heuristic values. Pheromone value indicates the quantity of antswhich have selected the trail in recent times, whereas the heuristic value is a problem-dependent quality metric. When ants reach decision points, it is more probable that they select the trail with more pheromone as well as heuristic value. When the ants reach the destination, solutions associated with theroutes followed by the ants are assessed and pheromone values are updated respectively. Moreover, evaporation makes the pheromone levels of other trails decrease in a gradual manner. Therefore, trails which are not reinforced from time to time, will lose pheromone and thereby decrease the likelihood of being selected by ants.

ACO is based on the following idealized rules (Parpinelliet al.2002):

- All paths taken by ants are related to potential solutions for a particular issue.
- When ants follow a particular route, the quantity of pheromone deposited on the path is proportional to the quality of the related potential solution for the target issue.
- When ants have to select between two or more routes, the one with greater quantity of pheromone is more likely to be selected.

Because of this, ants converge to the shortest path, and hopefully optimal or near-optimal solution for the issue is reached.

ACO protocol's design requires the following specifications:

- An adequate representation of the issue that permits the ants to incrementally build or alter solutions by using the probabilistic transition rule, on the basis of the quantity of pheromone in the trail as well as on a local, problem dependent heuristic.
- A technique for enforcing the building of valid solutions, i.e. solutions which are legal in the real world as per the problem definition.
- A problem-dependent heuristic function (□) which assesses the quality of the items which may be appended to the current partial solution.
- A rule for pheromone updating, that specifies the modification of pheromone trail (□).
- A probabilistic transition rule on the basis of the value of heuristic function (□) as well as on the contents of the pheromone trail (□) which is utilized to iteratively build a solution.

Artificial ants possess various features like real ants, which are:

- Artificial ants have a probabilistic preference for routes with greater pheromone.
- Shorter routes have bigger growth rates in the quantity of pheromone.
- Ants utilize indirect communication system on the basis of the quantity of pheromone in every route.

The objective of Ant-Miner is the extraction of classification rulesfrom data. A High-Level Description of Ant-Miner Protocol:

Training Set = {all training cases};
Discovered Rule List = [ ];
WHILE (Training Set >Max_uncovered_cases
t=1;
j=1;
Initialize all trails with the same amount of pheromone;
REPEAT
Antt starts with an empty rule and incrementally
Constructs a classification rule
Rt by adding one term at a time to the current rule;
Prune rule Rt;

Update the pheromone of all trails by increasing pheromone

In the trail followed by Ant (proportional to the quality of Rten decreasing pheromone in the other trails (simulating pheromone evaporation);

Results and Discussion

In this section, K Anonymization is improved through the selection of features via ACO optimization algorithm. Experiments were conducted using adult data set available in online repository. The classification accuracy, average Positive Predictive value and average sensitivity are shown in table 4.2. Results show that the ACO algorithm improved the performance with considerable amount in terms of classification accuracy, Average Positive Predictive value and Average Sensitivity than PSO optimization.

K Anonymization size	PSO Optimization	With ACO Optimization	ACO Suppression Only
No Anonymization	87. 7351	87. 99	85. 9851
K= 10	84. 4634	86. 93	83. 0654
K= 15	84. 0583	87. 01	83. 2639
K= 20	83. 9866	87. 16	83. 2849
K= 25	82. 1503	85. 83	82. 6668
K= 30	82. 747	85. 76	82. 7897
K= 35	79.535	85. 7	82. 1341
K= 40	79. 4897	85. 65	82. 6891
K= 45	79.5738	83. 16	82. 359
K= 50	79. 7474	83. 01	79.6757

Table 4.2 Classification Accuracy

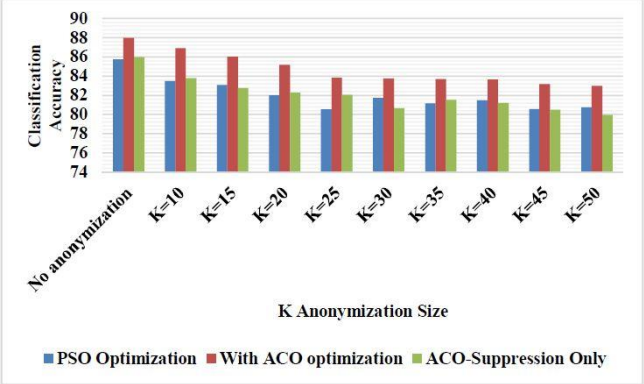

Figure 4.3 Classification Accuracy

From the Figure 4.3, it can be observed that the ACO optimization has higher classification accuracy by 4.06% for K=10, by 2.43% for K=30and by 2.76% for K=50 when compared with PSO optimization and by 3.68%for K=10, by 3.75% for K=30 and by 3.72% for K=50 when compared withACO suppression only.

Conclusion

PPDM strategies, however, are the modification of data to remove sensitive or anonymous information at certain levels. PPDM not only protects sensitive data but also ensures valid data mining results. In this function, the ACO algorithm is used for feature selection and the Adult database available on the online database is used for testing. The results show that ACO optimization has a maximum accuracy of 4.06% for K = 10, 2.43% for K = 30 and 2.76% for K = 50 compared to PSO preparation and 3.68% for -K = 10, by 3.75% K = 30 and by 3.72% at K = 50 compared to ACO pressure only."

