SNS Privacy Protection based on the ELM Integration and Semi-supervised Clustering

Kunlun Li
College of Electronic an Information Engineering, Hebei Universit, Baoding, China
Email:likunlun@hbu.edu.cn

Zhe Wang, Chao Ma, Song Song
College of Electronic an Information Engineering, Hebei Universit, Baoding, China
Email: zhegejiayou@163.com

***Abstract*—Social network service (SNS) is a new emerging Web application. With the growth of SNS in application, the web security is facing more serious threats and the leak of individual privacy is a major one. Because of vast number of unlabeled data and small amount of labeled data in SNS web, the availability of data is relatively poor for the study of SNS privacy preservation. In order to solve the above problem, this paper proposed an ELM ensemble algorithm based on Bagging combined with semi-supervised Seeds set clustering for privacy preserving. The main process is as follows: first, the ensemble ELM is used to label the unlabeled data to enlarge the scale of Seeds set; second, the Seeds set is used to initialize the center of clustering; and finally, the algorithm adopts semi-supervised clustering to achieve K-anonymity. Experimental results show that the method can improve the usability of the released data while preserving privacy.**

***Index Terms*—SNS, Privacy preservation, Semi-supervised clustering, ELM, K-anonymity**

I. Introduction

Social network service (SNS) is a new emerging Web application form. As a new internet business model, SNS has become a hot topic and aroused interest from the public. In the United States, SNS Facebook beat Google and became the largest website in America. Statistics show that 7.07% of American Internet users visit SNS. Until 2011, the number of SNS web site users in China is up to 235 million. At the same time, several problems in SNS site have gradually emerged, especially which involve user's privacy issues. In December 2011, CSDN, Renren nets and many others SNS web sits in China encountered hacker attacks, leading to 43 million users' privacy leak out. If the privacy information is used by criminals, it will cause a lot of trouble for SNS users, such as spam harassment, fake identity fraud, human flesh search, etc.

The K-anonymity model can achieve the purpose of preventing privacy [1]. It demands that the release of data cannot be distinguished in the presence of at least K records in quasi identifier, so that the attacker cannot distinguish the privacy information of specific individuals, thus protecting personal privacy. It needs data privacy in quasi identifier of the attribute values for generalized data processing, in order to eliminate link attack. Currently, the K-anonymity model research focuses on the protection of personal privacy information, meanwhile, improves the availability of data.

G.Aggarwal's research shows that, the optimal data anonymity problem (i.e., in the realization of sensitive attribute anonymous protection at the same time, the information loss minimization) is a NP hard [2,3]. Focusing on how to reduce the information loss when anonymous protection, researchers have put forward a variety of heuristic data anonymous algorithms. The advantage of the heuristic algorithms is that they can be used universally in many anonymous rules. V.Iyengar proposed an algorithm based on incompletely random search method, to solve the K-anonymity of combination flooding problem [4]. K.LeFevre provided a practical framework for implementing one model of K-anonymity, called full-domain generalization [5]. R.Bayardo proposed a new approach to exploring the space of possible anonymity that tames the combinatorial analysis of the problem, and developed data-management strategies to reduce reliance on expensive operations such as sorting [6]. T.Iwuchukwu observed that since building an index over a data set leads to a natural partitioning of the data set, K-anonymity can be introduced by enforcing a minimum occupancy threshold on partitions [7].

The above works achieved K-anonymity mainly through generalizations and suppressions. However, the technologies have been proved low efficiency and the data availability is relatively poor. In recent years, clustering algorithms are applied to K-anonymity, which make up for the deficiency of the generalization and suppressions technology. S.L.Hansen presented an efficient polynomial time algorithm to solve the univariate microaggregation problem. Optimal partitions are shown to correspond to shortest paths in a network [8]. A.Sonlanas proposed a new blocking method based on 2d-trees for intelligently partitioning very large data sets for microaggregation [9]. J.Domingo-ferrer extended the use of microaggregation for K-anonymity to implement the recent property of p-sensitive K-anonymity in a more unified and less disruptive way [10].

The traditional clustering for the initial value is sensitive, i.e., different initial value can lead to different

doi:10.4304/jsw.8.1.160-167

clustering results, and sometimes the clustering is in local optimum, not the global optimal. Aiming at this problem, this paper proposed an ELM ensemble algorithm based on Bagging combined with semi-supervised Seeds set clustering for privacy preserving. Experimental results show that the proposed method can improve the usability of the released data and preserve privacy at the same time.

II. Concept Of K-Anonymity

K-anonymity is a privacy protection technology. The objective of K-anonymous is to make every tuple of privacy-related attributes in a published table identical to at least (k-1) other tuples. As a result, no privacy-related information can be easily inferred. Generalization and suppressions are traditional technology of K-anonymous, but they have low efficiency and the data availability is relatively poor. Therefore, clustering algorithm is applied to the K–anonymity. In order to comprehend K-anonymity better, we introduce the following definitions [11].

DEFINITION 1 (K-anonymity) A dataset is said to satisfy K-anonymity for k>1 if, for each combination of values of quasi-identifiers (e.g. name, address, age, gender, etc.), at least k records exist in the dataset sharing that combination.

DEFINITION 2 (Median) Given an ordinal scale C= $\{c_1<c_2<\ldots<c_o\}$, the median of the set S= $\{a_1, a_2, \ldots, a_n\}$(with $a_i\in C$)is the category that occupies the central position in S once S is ordered. In terms of frequencies, the median is a category such that its predecessors and successors in the ordered S have equal frequency.

DEFINITION 3 (Quasi-identifier Attribute Set) A quasi-identifier attribute set is a set of attributes in a table that potentially reveal private information, possibly by joining with other tables.

DEFINITION 4 (equivalence class) An equivalence class of a table with respect to an attribute set is the set of all tuples in the table containing identical values for the attribute set.

DEFINITION 5 (K-anonymity Property) A table is K-anonymous with respect to a quasi identifier if the size of every equivalence class with respect to the attribute set is K or more.

The main objective of the k-anonymity model is thus to transform a table so that no one can make high-probability associations between records in the table and the corresponding entities. In order to achieve this goal, the k-anonymity model requires that any record in a table be indistinguishable from at least (k−1) other records with respect to the pre-determined quasi-identifier. A group of records that are indistinguishable to each other is often referred to as an equivalence class. By enforcing the k-anonymity requirement, it is guaranteed that even though an adversary knows that a k-anonymous table contains the record of a particular individual and also knows some of the quasi-identifier attribute values of the individual, he/she cannot determine which record in the table corresponds to the individual with a probability greater than 1/k.

The key idea underlying our approach is that the K-anonymization problem can be viewed as a clustering problem. Clustering is the problem of partitioning a set of objects into groups such that objects in the same group are more similar to each other than objects in other groups with respect to some defined similarity criteria. Intuitively, an optimal solution of the k-anonymization problem is indeed a set of equivalence classes such that records in the same equivalence class are very similar to each other, thus requiring a minimum generalization.

Realizing the anonymous by clustering algorithm is divided into two steps. (1) By clustering, the table T is divided into K parts and new data table is composed by the G equivalence classes. (2) The algorithm substitutes the clustering center for classes, and achieves k-anonymity. For example: the original data table T (see table 1) contains quasi-identifier QI = {Sex, Zip, Disease}, sensitive attributes= {Disease}. The process of K-anonymity is as follows: (1) Deleting identifiers, data were initialized, such as a table one. (2) Data table T in QI attribute value is standardized. And T is clustered based on the QI. (3) The normalized value recovers original values (using average value as the center) and we get two each equivalence class. Anonymity results see table 2.

TABLE 1
SNS PERSONAL PRIVACY DATA

Birth Date	Sex	Zip	Disease
1980/05/01	Male	100110	AIDS
1980/05/31	Famale	100200	FLU
1975/08/10	Famale	100138	HEPATITIS
1975/02/12	Famale	100140	FEVER

TABLE 2
ANONYMOUS RESULTS

Birth Date	Sex	Zip	Disease
1980/05	Male	100150	AIDS
1980/05	Male	100150	FLU
1975	Famale	100139	HEPATITIS
1975	Famale	100139	FEVER

Data anonymous can be regarded as specific constrained clustering problems which satisfied the requirements of anonymous model. K-anonymity clustering is obvious difference from the general clustering problem. The traditional clustering methods are not suitable for solving data anonymity. It requires the number of class be fixed. However, K-anonymity clustering problem does not limit the number of classes, and contains at least K records compulsorily. In the situation that Anonymous model satisfy the requirements, groups object is similar as much as possible and different groups object is not similar as much as possible [12].

At the heart of every clustering problem are the distance functions that measure the dissimilarities among data points and the cost function which the clustering problem tries to minimize. The distance functions are usually determined by the type of data being clustered, while the cost function is defined by the specific objective of the clustering problem. In this section, we describe our distance and cost functions which have been specifically tailored for the K-anonymization problem.

The data can be divided into two categories: continuous data (such as zip code, income, etc) and categorical data (such as color, position titles, etc). Different types of data require different similarity measurement and center definition.

(1) Definition of continuous data distance

Continuous data usually use the Euclidean distance definitions. It is defined as

$$d(X,Y)=\sum_{i=1}^{p}(X_i-Y_i)^2 \quad (1)$$

Information loss in the single equivalent class is defined as

$$LL=\sum_{j=1}^{n_i}d(X_{ij},\bar{X}_i),\ \text{where}\ \bar{X}=\frac{1}{n_i}\sum_{j=1}^{n_i}X_{ij} \quad (2)$$

Information loss in all equivalence classes is defined as

$$CL=\sum_{j=1}^{n_i}LL(G_i)=\sum_{i=1}^{g}\sum_{j=1}^{n_i}d(X_{ij},\bar{X}_i) \quad (3)$$

where g is the number of categories.

To meet the K-anonymity model at the same time, the smaller LL is, the stronger internal homogeneity is.

(2) Categorical data distance definition

The data from the relatively simple type definition

$$\text{IF}\ \ X_i=Y_i, d(X_i,Y_i)=0$$
$$\text{ELSE}\ \ d(X_i,Y_i)=1 \quad (4)$$

III. Semi-Supervised Clustering

Because of vast number of unlabeled data and small amount of labeled data in SNS web, limited training data does not provide enough data distribution information, so the cluster of data can not be satisfied with the anonymous results. Therefore, this paper proposes to use semi-supervised clustering method to solve the problem of K-anonymity.

Clustering algorithms try to find the structure information in unlabeled data to construct a classifier, which need no prior knowledge and be seen as one of unsupervised learning usually. Because of no labeled information on data distribution, when the objective function is unsuitable to data set, clustering methods may give useless partition results in practical problems. Semi-supervised clustering using little prior knowledge to improve the performance of clustering algorithms became a novel hotpot in machine learning recent years

Semi-supervised clustering algorithm can be divided into three categories:

(1) Constrained-based. The method uses supervised information to constrain clustering search process. By using the given labeled data set or other constraints to carry out clustering, the method can get more heuristic information and reduce the searching blindness. Its direct target is to make a better clustering effect. The typical algorithm is Seeded-K-means and Constrained-K-means [13]. Wagstaff et al proposed semi-supervised algorithm COP-K-means [14].

(2) Similarity measurement-based. The method makes use of labeled data to meet the marker or constraint distance measure function, and its main purpose is to meet certain conditions given distance function for clustering. The typical algorithm is that Klein gave a method for inducing spatial effects of pairwise constraints and had demonstrated that it substantially outperforms previous approaches, exhibiting behavior which is both quantitatively superior and qualitatively more natural [15].

(3) Combination of the former two methods [16, 17]. The typical algorithm is MPC-K-means [18]. The algorithm is integration method based on the constraints and the similarity measurement. The algorithm introduces pairwise constraints as supervisory information. Reflecting in the objective function is to increase C which does not satisfy the penalty. In order to join algorithm based on the similarity thought, it introduces a symmetric positive definite matrix, using new parameters of Euclidean distance.

Jia LV proposed a novel semi-supervised ,which can not only handle semi-supervised binary classification problem but also deal with semi-supervised multi-class classification problem [19]. Tao Guo proposed a algorithm, which combines graph-based semi-supervised learning with collaboration-training [20].

As classical clustering algorithms, K-means and fuzzy c-means have been widely used in many fields. To the integrality of our paper, firstly, we give a short review of K-means. Let X is the dataset of N samples and D dimensions. $\mathrm{X}=\{\mathrm{x}_1,\ldots,\mathrm{x}_\mathrm{N}\}$. Our goal is to assign the data points into K partitions. Assume that the K centers are $m_1,m_2,\ldots,m_k$ and in cluster K there is N$_k$ instances. So we get $m_k=\frac{1}{N}\sum_{i=1}^{N_k}x_i$, for k=1,…,k. Based on squared Euclidean distances and the criteria of within-groups sum of squares error, the objective function of K-means clustering can be presented as: $J=\sum_{k=1}^{K}\sum_{i=1}^{N_k}\|x_i-m_k\|^2$.

Based on classical K-means, Basu et al proposed a semi-supervised K-means method, using a small amount of labeled data. A seeds set is introduced which contain a little labeled data and assume that all K clusters are covered. In each cluster there is typically at least one seed point. Assign the instances in seeds set to K clusters to give an initial partition, optimize the objective function using EM algorithm and gain much better result than classical K-means.

Using seed set, Basu gave two semi-supervised K-means algorithms: Constrained-K-means and Seeded-K-means. In Constrained-K-means, the cluster memberships of the data points in the seeds set are not re-computed and thus the cluster labels of the seed data remain unchanged, and only the labels of the non-seed data are re-estimated.

In Seeded-K-means, the user-specified labeling of the seed data may be changed.

```
Input: Set of data points $X = \{x_1,...,x_N\}$, number
of cluster K, set $S = \bigcap_{l=1}^{K} S_l$ of initial seeds
Output: Disjoint K partitioning $\{X_l\}_{l=1}^{K}$ of X such
that K-means objective function is optimized
Method:
1   initialize: $u_h^{(0)} \leftarrow \frac{1}{S_h}\sum_{x\in S_h} x$ for h=1,…,K;
2   Repeat until convergence
        Assign-cluster: Assign each data point x to the
3       cluster h*(i.e. set $X_{h^*}^{(t+1)}$ ), for
        h* = argmin $\left\| x - u_h^{(t)} \right\|^2$
        stimate-means:
4       $u_h^{(t+1)} \leftarrow \frac{1}{x_h^{(t+1)}}\sum x \in x_h^{(t+1)} x$
5       t ← (t+1)
```

Ⅳ. The Ensemble Elm Algorithm Based On Bagging

Semi-supervised clustering algorithm based on seed set is affected by Seeds set size and quality. In this paper, we use ELM to train the learner and label unlabeled data with a small amount of labeled data, and utilize Bagging algorithm to integrate the ELM learner, so as to improve the accuracy of the mark and generalization ability.

A. Elm (Extrem Learning Machine)

ELM algorithm was proposed by G.B Huang in 2004 [21].It is a new learning scheme of feed forward neural network. Before training, ELM set weights and bias values from hidden layer to the input layer. The output layer weight generates a unique solution. The essence of algorithm does not need to adjust the hidden layer but try to reach the minimum training error and minimum output weight. This algorithm has better generalization performance. The learning speed is very fast.

For N arbitrary distinct samples (x_i, t_i) , standard SLFNS with N hidden neurons and activation function g(x) are mathematically modeled as

$$\sum_{i=1}^{\tilde{N}} \beta_i g(w_i \bullet x_j + b_j) = o_j, j = 1,...,N \qquad (5)$$

That standard SLFNS with N hidden neurons with activation function g(x) can approximate these N samples with zero error means that $\sum_{j=1}^{N} \left\| o_j - t_j \right\| = 0$, i.e. there exist β_i w_i b_i such that

$$\sum_{i=1}^{\tilde{N}} \beta_i g(w_i \bullet x_j + b_j) = t_j, j = 1,...,N \qquad (6)$$

The above N equations can be written compactly as:

$$H\beta = T,$$

$$\beta_{l\times m} = [\beta_1, \beta_2, ..., \beta_l]\ T_{n\times m} = [T_1, T_2, ..., T_n] \qquad (7)$$

H is called the hidden layer output matrix of the neural network: the ith column of H is ith hidden neuron's output vector with respect to inputs $x_1, x_2, ..., x_n$.

One solution is to $\beta = H^{'}T$

ELM algorithm steps can be summed up in the following three steps:

(1)Randomly generated hidden neuron parameter

(2)Calculation of the output matrix of hidden layer

(3)Calculation of output weights $\beta = H^{'}T$

The learning speed of ELM is extremely fast. It can train SLFNs much faster than classical learning algorithms. Previously, it seems that there exists a virtual speed barrier which most (if not all) classic learning algorithms cannot break through and it is not unusual to take very long time to train a feed forward network using classic learning algorithms even for simple applications. Unlike the traditional classic gradient-based learning algorithms which intend to reach minimum training error but do not consider the magnitude of weights, the ELM tends to reach not only the smallest training error but also the smallest norm of weights. Thus, the proposed ELM tends to have the better generalization performance for feed forward neural networks. Unlike the traditional classic gradient-based learning algorithms which only work for differentiable activation functions, the ELM learning algorithm can be used to train SLFNs with non-differentiable activation functions

B. The Design Of Ensemble Learner

Considering accuracy rate of a single weak learner is not high, we use the learner for integration, so the accurate rate is improved. Ensemble learning is a combination learner method, making the learner exhibit better performance than a single learner. To enhance the learner generalization ability, on one hand we should improve the generalization ability of single ELM, on the other hand, we should increase the difference between training set.

An ensemble is itself a supervised learning algorithm, because it can be trained and then used to make predictions. The trained ensemble, therefore, represents a single hypothesis. This hypothesis, however, is not necessarily contained within the hypothesis space of the models from which it is built. Thus, ensembles can be shown to have more flexibility in the functions they can represent. This flexibility can, in theory, enable them to over-fit the training data more than a single model would, but in practice, some ensemble techniques (especially bagging) tend to reduce problems related to over-fitting of the training data.

Empirically, ensembles tend to yield better results when there is a significant diversity among the models [22]. Many ensemble methods, therefore, seek to promote diversity among the models they combine [23,24]. Although perhaps non-intuitive, more random algorithms (like random decision trees) can be used to produce a stronger ensemble than very deliberate algorithms (like entropy-reducing decision trees)[25]. Using a variety of strong learning algorithms, however, has been shown to be more effective than using techniques that attempt to dumb-down the models in order to promote diversity.

Bootstrap aggregating, often abbreviated as bagging, involves having each model in the ensemble vote with equal weight. In order to promote model variance, bagging trains each model in the ensemble using a randomly drawn subset of the training set. As an example, the random forest algorithm combines random decision trees with bagging to achieve very high classification accuracy [26]. Boosting involves incrementally building an ensemble by training each new model instance to emphasize the training instances that previous models mis-classified. In some cases, boosting has been shown to yield better accuracy than bagging, but it also tends to be more likely to over-fit the training data. By far, the most common implementation of Boosting is Adaboost, although some newer algorithms are reported to achieve better results.

The existing ensemble methods get bigger difference degree of individual networks through the disturbance training data. For example, by boosting the network training set is determined by the prior network performance. The sample which is misjudged by existing network will probably appear in the new network training set [27]. The Bagging algorithm is the basis for repeated sampling. The algorithm trains the network using examples which are randomly sampled from the original training. The algorithm increases discrimination of integration, by repeating the selection of training set, in order to enhance the generalization ability [28]. Gang Zhang proposed to ensemble with base learners trained by both labeled and unlabeled data, by adopting data dependant kernel mapping [29].

This paper uses ensemble ELM based on the Bagging to mark unlabeled data. As follows:

(1) We randomly sample 1/2 samples from the training data set (have back samples) and train the ELM learner by the samples. The classifier is marked as ht h_t .

(2) We can use the same method to form multiple ELM classifiers and get a prediction function sequence $h_{1,}h_2,...,h_t$

(3) For unknown sample classification, each classifier can give a classification results. Then T classifiers cast a vote. The classification result with most votes is classification results for unknown sample.

C. Semi-Supervised Clustering Based On Elm-Bagging

As mentioned above, this paper proposes an ELM ensemble algorithm based on Bagging combined with semi-supervised Seeds set clustering for privacy preserving. The main process is as follows: firstly the ensemble ELM is used to label the unlabeled data to enlarge the scale of Seeds set, secondly seeds set is used to initialize the center of clustering. Finally, the algorithm adopts semi-supervised clustering to achieve K-anonymity. Specific algorithm is described as follows:

```
Input: data set $X=\{X_1,X_2,...,X_n\}$ , Anonymous requirements K, set S=$\bigcup_{l=1}^{K} S_l$ of initial seeds, the number of classifier T, ELM hidden neurons $P=\{P_1,P_2,...,P_n\}$
Output: K-anonymity data set, $Y=\{Y_1,Y_2,...,Y_n\}$, all equivalence classes of data information loss CL.
for i ∈ {1,2,…,T} do //Initialize the T learner
    Si ← BooststrapSample (S)
    Hi ← Learn (S$_i$)
end
Unlabeled data U=X-S
for i ∈ {1,2,…,T} do
    New label data Vi ← Hi (U)
end
V ← Bagging (Vi ) //ensemble
int M ← X/K //the number of cluster M
L ← V+S
if M= =label number of L //If there is a lack of class label, we will add to complete
    C ← center (L)
else
    C ← center (L+ random (X))
end
bool =0,pass=0
while(bool= =0 or pass<10000)
    D ← Distribute(X) //distribute data to nearest class
    C ← center(D) //Calculate the cluster centers
    If C’= =C
        bool=1;
    else
        C ← C’
    pass++
    end
end
```

28	$CL = \sum_{j=1}^{n_i} LL(G_i) = \sum_{i=1}^{g}\sum_{j=1}^{n_i} d(X_{ij}, \bar{X}_i$

V. Experiment

All experiments are run on machine with Pentium(R) Dual-Core CPU. The operating system on the machine is Microsoft Windows XP Professional Edition, and the implementation is built and run in Matlab7.1.

We do our experiment on the Adult Database from the UCI Machine Learning Repository. The data set comprises part of the United States census data. Adult data set is privacy protection referenc9;e test set. We remove tuples with missing values and there are 45,222 tuples in data set. In order to verify the effectiveness of the proposed algorithm, we randomly select two groups of data to make the same test, respectively 500 and 1000. In the test, age, fnlwgt and education-num should be quasi identifier and salary should be sensitive attributes. Considering the different effect of different identifiers for anonymity, we make all the data normalized, such as formula:

$$Y_i = (X_i - \min(X)/\max(X) - \min(X)) * 2 - 1$$

In the experiment, let the initial label rate be 10%, then by the ELM-Bagging algorithm Seeds will be increased to 20%, 30%,40%, 50%. Algorithm makes semi-supervised clustering based on Seeds sets and computes information loss.

An ensemble learner of the generalization ability is up to output difference degree of each learner. In the circumstance that the outputs are accurate, ELM guarantees the difference of output by adjusting the parameters. From the statistical machine learning perspective, when the base learner number is large, ensemble learner output results become closer to the expected value, however, when the base learner number increases, learning calculation time and complexity will increase. In conclusion, experiment trains seven base learners. By adjusting the hidden neurons, algorithm make sure each learner of accuracy higher than 80%, otherwise you will add too much noise, influencing clustering effect.

Ensemble learning is a powerful machine learning paradigm which has exhibited apparent advantages in many applications. By using multiple learners, the generalization ability of an ensemble can be much better than that of a single learner. A serious deficiency of current ensemble methods is the lack of comprehensibility, i.e., the knowledge learned by ensembles is not understandable to the user. Improving the comprehensibility of ensemble is an important yet largely understudied direction. Another important issue is that currently no diversity measure is satisfying although it is known that diversity plays an important role in ensembles. If those issues can be addressed well, ensemble learning will be able to contribute more to more applications. Ensemble learning has already been used in diverse applications such as optical character recognition, text categorization, face recognition, computer-aided medical diagnosis, gene expression analysis, etc. Actually, ensemble learning can be used wherever machine learning techniques can be used.

First, the algorithm expands seeds through ELM-Bagging. Because ELM is an extremely fast learner, so the time cost is negligible in this part. The time cost is mainly concentrated on the redistribution of data points and calculating the number of iterative clustering center. We assume that the data set size is n, dimension is D, and iterative number is m, therefore, the algorithm complexity is O (mknd) in the worst condition.

500 data (table 1)

K rate	10	15	20
10%	69.125	76.120	82.252
20%	69.106	75.836	81.125
30%	68.256	75.165	80.655
40%	68.201	73.256	78.663
50%	67.168	72.066	76.125

1000 data (table 2)

K rate	10	15	20
10%	103.889	116.865	128.102
20%	102.562	115.251	127.369
30%	101.056	115.021	126.562
40%	101.568	113.712	125.210
50%	98.658	112.258	123.267

We can see from the results, when the label rate is same, along with the anonymity number increasing, the quantity loss of information increases gradually. This is because the K-anonymity itself is an NP hard. And in order to get better effectiveness of privacy protection, it will increase the quantity loss of information. But when the anonymity number is same, along with the Seeds set size increasing, the quantity loss of information reduces. This is because the ELM ensemble learner can make Seeds set size increase and improve quality, in the mean time supervising the clustering process effectively. Ultimately it can improve the performance of clustering effect.

VI. Conclusion

Social network service (SNS) is a new emerging Web application form. As a new internet business model, SNS has become a hot topic and aroused interest from the public. For this kind of network privacy security problem, this paper adopt K-anonymous model for privacy protection and propose an ELM ensemble algorithm based on Bagging combined with semi-supervised Seeds set clustering. In SNS, there is vast unlabeled data with a few labeled data. This method can train the unlabeled data by the ELM-Bagging algorithm, increasing the labeled data size. It can supervise and guide the completed anonymous task. Experimental results show that, the method can effectively protect the personal privacy, and along with the marker information continues to expand, the quantity of information loss reduces gradually. At the same time, because the ELM ensemble learner is very fast, the method did not increase the time complexity of the algorithm than the general method which is semi-supervised clustering.

In the future, we will consider K-degrees anonymity to protect privacy in SNS. K-degrees anonymity and traditional K-anonymity is very similar, and the difference is that it focuses the "degree" nodes in topological structure. To meet the K-degrees anonymity, for each node, there are at least k-1 nodes which have the same degree in the network.

ACKNOWLEDGMENTS:

This work is supported by the National Natural Science Foundation of China under Grant (No.61073121), Natural Science Foundation of Hebei Province of China (No. F2009000215) and Biomedical Multidisciplinary Research Center of Hebei University (No.BM201102).

REFERENCES

[1] P.Samarati, L.Sweeney. *Generalizing data to provide anonymity when disclosing information*. Proc of the 17th ACM SIGMOD SIGACT SIGART Symposium. Seattle, WA, USA: IEEE press.(1998)188.

[2] G.Aggarwal, T.Feder, K.Kenthapadi, et al. *Achieving anonymity via clustering*. //Pro of the 25th ACM SIGMOD- SIGACT-SIGART Symp. New York: ACM, (2006) 153-162.

[3] A.Meyerson, R.Williams. *On the complexity of optimal k-anonymity*.//Proc of the 23rd ACM SIGACT- SIGMOD-SIGART Symp. New York: ACM, (2004) 223-228.

[4] V.Iyengar. *Transforming data to satisfy privacy constraints*. Proc of the 8th ACM SIGKDD Int'l Conf. New York: ACM, (2002)279- 288.

[5] K.LeFevre, DeWitt D, Ramakrishnan R. *Incognito: Efficient full- domain k-anonymity*. //Proc of the 24th ACM SIGMOD Int'l Conf. New York: ACM, (2005) 49-60.

[6] R.Bayardo, R.Agrawal. *Data privacy through optimal k-anonymization*.//Proc of the 21st Int'l Conf .Los Alamitos: IEEE Computer Society, 2005. 217-228.

[7] T.Iwuchukwu, J.F.Nauchton. *K-anonymization as spatial indexing: toward scalable and incremental anonymization*. VLDB. 07:Proceedings of the 33rd international conference on Very large data bases. VLDB Endowment,(2007)746-757.

[8] S.L.Hansen, S.Mukherjee. *A polnomial algorithm for optimal univarate microaggregantion*. IEEE Transaction Knowledge and Data Engineering, 15(4) (2003)1043-1044.

[9] Solanas A, Martinez Balleste A, Domingo Ferrer J, etal. *A2d-tree-based blocking method for microaggreagting very large data sets*. //Proc of the First International Conference on Availability, Reliability and Security. Vienna, Austrias: IEEE press, 2006. 922-928

[10] Domingo Ferrer J. *Microaggregation for database and location privacy*. Proc of Next Generation Information Technologies and Systems. Kibbutz Shefayim, Israel:Springer Verlag, 2006. 106-116.

[11] Doming-Ferrer J, TorraV. Torra. *Continuous and heteroge-neous k-anonymitythrough microaggreagtion*. Journal of Data Mining and Knowledge Discovery,2005,11(2):195-202

[12] Han J, Kamber M. Data Mining: *Concepts and Techniques*. 2nd ed. San Francisco: Morgan Kaufmann Publishers, 2006. 383-386

[13] Basu S,Banerjee A,Mooney R.*Semi-supervised clustering by seeding*.//Proc of the 19 Internal Conference on Machine Learning(ICML-2002), Sydney,Australia,July 2002:19-26

[14] Wagstaff K, CardieC,RogersS, et al *Constrained K-means Clustering with Background Knowledge*. //Proc of 18th International Conference on Machine Learning San Francisco USA,2001:577-586

[15] Klein D, Kamvar S D, Manning C. *From Instance-Level Constraints to Space-Level Constraints Making the Most of Prior Know ledge in Data Clustering*. //Proc of International Conference on Machine Learning, Sydey ,Australia, 2002

[16] Kun-lun LI, Chao ZHANG, et al. *Semi-supervised kernel clustering algorithm based on seed set*. Computer Engineering and Applications, 45(20) (2009) 154-157.

[17] Kun-Lun LI, Zheng CAO, et al. *Some developments on semi-supervised clustering*. Pattern Recognition and Artificial Intelligence, 22(5) (2009)735-742.

[18] Bilenko M, Basu S, Mooney R J. *Integrating Constraints and Metric Leaning in Semi-Supervised Clustering*. //Proc of the 21st International Conference on Machine Learning Banff, Canada, 2004:81-88

[19] Jia LV. *Semi-supervised Learning Using Local Regularizer and Unit Circle Class Label Representation.* JOURNAL OF SOFTWARE, VOL.7,NO.5,MAY 2012.

[20] Tao Guo. *Confidence Estimation for Graph-based Semi-supervised Learning*. JOURNAL OF SOFTWARE, VOL.7 NO.6, JUNE 2012

[21] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. *Extreme Learning Machine: A New Learning Scheme of Feedforward Neural Networks*. //Pro of Internation Joint Conference on Neural Network (IJCNN2004), Budapest, Hungary, July, (2004).

[22] L.Kuncheva, and C.Whitaker, *Measures of diversity in classifier ensembles*, Machine Learning, 51, pp. 181-207, 2003

[23] P.Sollich, and A.Krogh, *Learning with ensembles: How overfitting can be useful*, Advances in Neural Information Processing Systems, volume 8, pp. 190-196, 1996.

[24] G.Brown, and J.Wyatt, and R.Harris, and X.Yao, *Diversity creation methods: a survey and categorisation.*, Information Fusion, 6(1), pp.5-20, 2005.

[25] J. J. García Adeva, Ulises Cerviño, and R. Calvo, *Accuracy and Diversity in Ensembles of Text Categorisers*. CLEI Journal, Vol. 8, No. 2, pp. 1 - 12, December 2005.

[26] T.Ho, Random Decision Forests, *Proceedings of the Third International Conference on Document Analysis and Recognition*, pp. 278-282, 1995.

[27] R.E.Schapire. *The strength of weak learnability*. Machine Learning, 1990, 5(2): 197-227

[28] L.Breiman. *Bagging predictors*.Machine Learning, 1996, 24(2): 123-140

[29] Gang Zhang. *Data Dependant Learners Ensemble pruning*. JOURNA OF SOFTWARE ,VOL.7, NO.4, APRIL 2012.

KunLun Li, born in 1962, received the PhD from Beijing Jiaotong University, China, in 2004. He is working as a Professor in College of Electronic and Information Engineering, Hebei University. His main research interests include information security, pattern recognition, and biology information technology. In these areas, he has published over 40 technical papers in refereed international journals or conference proceedings.

Zhe Wang, born in 1986, received the bachelor's degree from the Tianjin University, China, in 2010. He is currently working toward the master degree in College of Electronic and Information Engineering, Hebei University. His research interests include information security and data mining.

Chao Ma, born in 1987, received the bachelor's degree from the Hebei University, China, in 2010. He is currently working toward the master degree in College of Electronic and Information Engineering, Hebei University. His research interests include pattern recognition and biology information technology.

Song Song, born in 1983, received the bachelor's degree from the Hebei University, China, in 2011. He is currently working toward the master degree in College of Electronic and Information Engineering, Hebei University. His research interests include pattern recognition and artificial Intelligence.