INTELLIGENT RECOGNITION ALGORITHM FOR SOCIAL NETWORK SENSITIVE INFORMATION BASED ON CLASSIFICATION TECHNOLOGY

WEIPING LI*
School of Information Engineering, Wuhan University of Technology
Wuhan, China
Department of Police Technology, Railway Police College
Zhengzhou, China

HAIYAN WU
Department of Police Technology, Railway Police College
Zhengzhou, China

JIE YANG
School of Information Engineering, Wuhan University of Technology
Wuhan, China

ABSTRACT. In the social network, there is the problem of network sensitive information with low accuracy rate of information recognition. To effectively improve the accuracy of intelligent identification of sensitive information, an intelligent recognition algorithm for sensitive information based on improved fuzzy support vector machine is proposed in this paper. The information is collected. The trajectory of the best movement of the information node is found in the low energy cache. In the limited time, the performance of information acquisition is improved by using the mobility of information nodes. According to DFS criterion, the features are added into the feature subset or eliminate the sensitive information. The feature selection algorithm based on multi-label is applied to feature selection of the collected information, so that the information gain between information feature and label set can be used to measure the importance. The improved support vector machine classification algorithm is used to classify the information selected by feature selection, and select effective candidate support vector, reduce the number of training samples, and improve the training speed. The new membership function is defined to enhance the effect of support vector on the construction of fuzzy support vector machine. Finally, the nearest neighbor sample density is applied to the design of membership function to reduce the noise, and achieve intelligent recognition of the sensitive information in the social network. Experimental results show that the accuracy rate of sensitive information intelligent recognition can be effectively improved by using the proposed algorithm.

1. Introduction. Now mankind has entered the information age. The most important feature of this information age is data generalization, data storage and transmission intellectualization, data classification and efficient utilization. Massive
information is generated every day. In order to obtain and use this information conveniently and effectively, it is needed to be classified [7,12]. In this way, the artificial intelligence technology has been developed. The application of target recognition is widely used in the field of computer vision. It plays an important role in the industrial application, aerospace and military field [20]. For the perception of the world, most of the information comes from the vision. The description and recognition of the object feature is the key to cognition. The current computer visual perception is not able to meet human needs. Therefore, it is a major challenge in the field of recognition to simulate the learning of the human brain to accomplish the learning of the target, and it is also a hot research [4,21].

In the social network, the target recognition of sensitive information is also called as the pattern recognition of the vision. The purpose is to process the information by using the theory and method of information processing and pattern recognition, in order to determine whether it contains sensitive information, extract useful information, determine the location of the target, and realize the description, analysis, judgment and recognition of sensitive information. Pattern recognition is a subject of classification and description of the information or physical processes. However, research results show that the classification algorithms that use different theories on the same data will obtain different results. In practice, in order to achieve the best effect, researchers need continuous experiments [3,9].

In present, a target recognition algorithm for convolution neural networks based on unsupervised pre-training and multi-scale partitioning is proposed. A sparse automatic coder is trained by using a non-labeled image to obtain a filter set that meets the characteristics of the dataset with good initial value. The feature is finally used for information classification, and feature input classifier is used to achieve information target recognition. However, the classification performance of the method is poor.

2. Intelligent recognition algorithm for social network sensitive information based on classification technology.

2.1. Social network sensitive information collection. In the social network, when the data is collected in the data cache with the distance from network center \( l = \sqrt{2(R^2 - 2r^2)/2} \), the total energy consumption of the whole network is low [11].

Randomly select a small region \( dx\,dy \) on a circular ring with the center \( O \), which is given by \( \rho d\rho \,d\theta \) in the polar coordinate. The number of the node in the region is \( \frac{n_{pd}d\rho d\theta}{\pi R^2} \). When \( \rho < 1 - r \), at least \( \left\lfloor \frac{l-\rho}{r} \right\rfloor - 1 \) jump is required for each information node to reach the cache area. When \( l+r \leq \rho \leq R \), at least \( \left\lfloor \frac{l-\rho}{r} \right\rfloor - 1 \) jump is required.

Energy consumption for transmitting or receiving unit data is \( e \). The amount of data collected by each node is \( q \). The total energy consumption of the information node in the cache area is related to the data amount of the node [14,16]. Assume the data of all information nodes are transmitted to the sensitive information node, and the total energy consumption of the nodes to transmit and receive data is \( p_t \) and \( p_r \), respectively. Then the total energy consumption in the cache area is given by

\[
P_1 = p_t + p_r = nqe + (n - n_s)eq = (2n - n_s)eq \quad (1)
\]

In addition to the cache area, the total energy consumption of the information nodes in the other regions along the shortest path for data transmission is approximately expressed as
Let $f(l) = (l-r)^3 + (l-r)^3 - 3(l+r)R^2 + 2R^3$. When $f(l) = 6l^2 + 6r^3 - 3R^2 = 0$ and $l = \sqrt{2(R^2 - 2r^2)}/2$, $f(l)$ obtains the minimum value, that is, the total energy consumption of the network is low when the information node moves in the cache area with the distance $l$ from the network center.

Constrained by the mobility of information nodes, information nodes traverse all the nodes in the cache area sequentially, which takes longer time and longer distance to move, and the more frequent the information nodes move, the worse the network stability is. In order to further shorten the movement distance of information nodes, different access probabilities are set for nodes in the cache area, which is to further reduce the number of nodes directly visited by information nodes [5, 8]. When the information node accesses one of the nodes, all the neighbor nodes can communicate directly with the information node. So the information node does not need to move to the actual location of the neighbor nodes to collect data. In order to ensure when the information node moves in the cache area, all the nodes have the opportunity to communicate directly with the information node, the width of the cache area is set to $r$.

When the width of the cache area is $r$ and the mobile step of the social network information node is less than $\sqrt{3}r$, it can guarantee that all nodes in the network have the opportunity to communicate directly with the sensitive information node.

As shown in Fig. 1, the current location of the sensitive information node is $O$. The communication scope of the next node to be visited by the information node must cover the two points of $A$ and $B$ to ensure that the sensitive information node can communicate directly with all the nodes in the network. $A$ and $B$ is taken as the center, respectively, and $r$ is as the radius, then the curves intersect at the point $C$. Only when $OC$ is less than $\sqrt{3}r$, it can guarantee that the two points of $A$ and $B$ is in the communication scope, as shown in the shadow part in Fig. 1.

Let $SS = \{SS(1), SS(2), \ldots, SS(n_s)\}$ is the set of all nodes, $NS(i)$ is the number of all the neighbor nodes. Assume the first visited node of the sensitive information node is $SS(i)$. After collecting the data of this node, the node with the maximum probability is selected as the object of the next access [18, 23]. The transfer
The probability between nodes is expressed as

$$P(i, j) = \begin{cases} 
0, & d(i, j) > \sqrt{3}r \\
\alpha \frac{N_S(j)}{n_s} + \beta \frac{d(i, j)}{\sqrt{3}r}, & r < d(i, j) \leq \sqrt{3}r 
\end{cases}$$

(3)

where $\alpha$ and $\beta$ is a constant between 0 and 1, respectively, $d(i, j)$ is the distance between $SS(i)$ and $SS(j)$, $N_S(j)/n_s$ is the ratio of the number of neighbor nodes of $SS(j)$ to the total number of nodes in the cache area. The node with more neighbor nodes is more likely to be accessed by the sensitive information node. When a node is accessed by a sensitive information node, the access probability of all neighbor nodes will be 0. The value limit of $d(i, j)$ can ensure that all nodes have the opportunity to communicate directly with the sensitive information nodes to prevent the loss of data. The set of $d(i, j)/\sqrt{3}r$ makes the probability that the nodes in the shadow area far away from the current location of the sensitive information node become more likely to be the next access object. The moving step of the sensitive information node is increased, thus reducing the number of times the sensitive information node moves [22]. The probability matrix of the sensitive information node moving between nodes is expressed as

$$P = \begin{bmatrix} 
P(1, 1) & P(1, 2) & \cdots & P(1, n_s) \\
P(2, 1) & P(2, 2) & \cdots & P(2, n_s) \\
\vdots & \vdots & \ddots & \vdots \\
P(n_s, 1) & P(n_s, 2) & \cdots & P(n_s, n_s) 
\end{bmatrix}$$

(4)

The matrix $VS(i)$ denotes the optimal access point set of the sensitive information node as the start of $SS(i)$. $pID$ denotes the ID of the current node accessed by the sensitive information node, $nID$ denotes the ID of the next node to be accessed by the sensitive information node, $D(i)$ denotes the moving distance of the sensitive information node from the $SS(i)$ to $SS(i)$ in a sampling cycle. In order to obtain the minimum moving distance of the sensitive information node, a path for the sensitive information node with the smallest distance is selected from the $n_s$ candidate path as the optimal movement strategy [6, 13]. Then the optimal access node set of the sensitive information node is $VS(\arg \min D(i))$.

In a certain network delay time, according to the sum of the cache data of each node in the communication scope of sensitive information, the pause time can be changed to improve the data collection performance [17]. Assume the time limit is $T$ and the maximum moving speed of the sensitive information node is $v_m$, then the total moving time is $D(x)/v_m$. The total pause time and the pause time of each accessed node are given by

$$T_{pause} = T - \frac{D(x)}{v_m} = \frac{size(VS(x))}{\sum_{k=0}^{n_s}} T_k$$

(5)

$$T_k = \frac{T_{pause}}{n} \sum_{j=1}^{a_k} q_{kj}$$

(6)

where $a_k$ is the total number of the nodes communicated with the sensitive information node for the $k$th accessed point, $q_{kj}$ is the number of the member nodes of the $j$th node for the $k$th accessed point [1].

If multiple nodes transmit the data to the sensitive information node at the same time, it will cause collision conflict and the loss of information data. In order to avoid conflict, the TDMA mechanism is used in this paper. The sensitive
information node generates TDMA rules based on the pause time and the member number of each node. Then these rules are sent to the nodes in the communication scope. After receiving the rules or messages, the node is in the sleep state and only transmits data in the specific time. This not only reduces the loss rate of data and improves the efficiency of data acquisition, but also reduces the energy consumption of a single node and improves the utilization of energy.

Define the ratio of the collected data amount \( q_i \) of the \( i \)-th node to the buffered data amount \( q_{buffer} \) is the data collection rate of the \( i \)-th node. The computation is given by Eq. (7), where \( 0 \prec p_i \leq 1 \). \( p_i = 1 \) denotes all the buffered data of the node is transmitted to the sensitive information node. The greater the value of \( p_i \), the stronger the data collection ability.

\[
p_i = \frac{q_i}{q_{buffer}} \tag{7}
\]

The sensitive information collection on the social network is achieved as the above process.

2.2. Social network sensitive information feature selection. Assume two classification problems in \( m \)-dimensional real space \( \mathbb{R}^m \), The size of the training sample set of the sensitive information is \( n \), the number of samples of positive and negative classes are \( n_+ \) and \( n_- \). The training set is \( \{(x_k, y_k) | x_k \in \mathbb{R}^m, m > 0, y_k \in \{1, -1\}, k = 1, \ldots, n\} \parallel n_- \). The discernibility of feature subsets (DFS) of the sensitive information with \( i (i = 1, 2, \ldots, m) \) features is given by

\[
DFS_i = \frac{\sum_{j=1}^{i} (\bar{x}_j^+) - \bar{x}_j)^2 + (\bar{x}_j^-)^2}{\frac{1}{n_+} \sum_{k=1}^{n_+} (\bar{x}_{k,j}^+) - \bar{x}_j^+)^2}
\]

where \( \bar{x}_j, \bar{x}_j^+ \), and \( \bar{x}_j^- \) is the mean of the \( j \)-th feature in the whole dataset, the positive class dataset, and the negative class dataset, respectively, \( x_{k,j}^+ \) is the value of the \( j \)-th feature of the \( k \)-th positive class sample point, \( x_{k,j}^- \) is the value of the \( j \)-th feature of the \( k \)-th negative class sample point.

In Eq. (8), the numerator denotes the sum of the square of the distance between the mean vector of the feature subset with \( i \) features for the positive class and the negative class and the mean vector of the feature subset with the \( i \) features for the whole sample set. The denominator denotes the sum of the variance of the sensitive information feature subset with \( i \) features for the positive class and the negative class. The larger numerator represents the interclass of the feature subset is sparser and the smaller denominator represents the interclass is more clustered [2,15]. The larger value of DFS represents the discernibility ability is stronger.

For the \( l (l \geq 2) \) class classification problem, assume the size of the sensitive information training sample set is \( n \), the dimension of sample space is \( m \). The training sample set is \( \{(x_k, y_k) | x_k \in \mathbb{R}^m, m > 0, y_k \in \{1, \ldots, l\}, l \geq 2\} \), where the number of the \( j \)-th class samples is \( n_j \), \( y_k = j, k = 1, \ldots, l \). \( DFS_i \) of the \( i(i = 1, \ldots, m) \) features is defined by
Assume $i$-th class dataset, respectively, where $\bar{\text{discernibility}}$ ability.

When $i = 1$, $\text{DFS}_1$ becomes the criterion for interclass discernibility of a single feature, which is the improved F-score criterion, given by

$$
\text{DFS}_i = \frac{\sum_{j=1}^{l} \left\| \bar{x}^{(j)} - \bar{x} \right\|^2}{\sum_{j=1}^{l} \frac{1}{n_j-1} \sum_{k=1}^{n_j} \left\| \bar{x}^{(j)}_{k,i} - \bar{x}^{(j)} \right\|^2}
$$

where $\bar{x}$ and $\bar{x}^{(j)}$ is the mean vector of the feature subset in the whole dataset and the $j$-th class dataset, respectively, $x^{(j)}_{k,i}$ is the feature vector of the current $i$ features of the $k$th sample in the $j$th class.

As the nonnegativity of the information gain, if the label and the feature $x$ is independent, then $\text{IG}_S(x)$ is transformed. Assume the distribution of the sensitive information $x$ is the mean vector of the feature subset in the whole dataset and the label set $L$ have little correlation with the label set. Then $\text{IG}_S(x)$ is the value of the $i$-th feature of each class [19]. Therefore, $F_i$ represents the ratio of the distance to the variance of interclass of the $i$th feature. The larger value shows the stronger discernibility ability.

Given a sensitive information feature $x$ and label set $L = \{l_1, l_2, \ldots, l_m\}$. Assume $\text{IG}(l_i | x)$ is the information gain of the feature $x$ and the label $l_i$. Then the information gain of the feature $x$ and the label set $L$ is given by

$$
\text{IGS}(Y | x) = \sum_{i=1}^{m} \text{IG}(l_i | x)
$$

As $\text{IG}(l_i | x) \geq 0$, $\text{IGS}(L | x) \geq 0$. From the information theory, $\text{IGS}(L | x) \leq \sum_{i=1}^{m} H(l_i)$.

If the sensitive information feature $x$ and each label in the label set $L = \{l_1, l_2, \ldots, l_m\}$ are independent, the information gain has minimum value.

As the nonnegativity of the information gain, if the label and the feature $x$ are independent, then $\text{IG}(l_i | x) = 0$ and $\text{IGS}(L | x) = 0$. As $\text{GS}(L | x) \geq 0$, the information gain has minimum value.

If each label $l_i$ is decided by the feature $x$, the information gain has maximum value. Then $\text{IGS}(L | x) = H(l_i)$ and $\text{IGS}(L | x) = \leq \sum_{i=1}^{m} H(l_i)$. As $\text{IGS}(L | x) = \leq \sum_{i=1}^{m} H(l_i)$, the information gain has maximum value.

From $\text{IGS}(L | x)$, it can be found the feature which has no effect on the label set $L$. In this way, a reasonable threshold can be set to remove the features which have little correlation with the label set.

In order to compute the threshold, the information gain of different feature $x$ and the label set $L$ is transformed. Assume the distribution of the sensitive information gain obeys the normal distribution. It can be transformed to the standard normal distribution by using

$$
\text{IGS}(L | x) = \frac{\text{IGS}(L | x) - \mu}{\sigma}
$$
where $\mu$ is the mean of the sensitive information gain, $\sigma$ is the standard variance [10].

Given a threshold $\delta \succ 0$, if the absolute value of the information gain $|IGS(L|x)| \geq \delta$, the feature $x$ is related to the label set, otherwise, they are unrelated.

In order to make the information gain of the features and the label has the same measurement range, the information gain is normalized, that is, $IG(B|A) = IG(A|B)$, and $IG(B|A) \leq H(B), IG(A|B) \leq H(A)$.

$$SU(A, B) = 2 \left[ \frac{IG(B|A)}{H(A) + H(B)} \right]$$  \hspace{1cm} (13)

is used for normalization of the information gain $IG(B|A)$, and $SU(A, B) \in [0, 1]$. If $SU(A, B) = 0$, it represents $A$ and $B$ are independent. $SU(A, B) = 1$ represents one of $A$ and $B$ can determine the other.

From $IGS(L|x) = \sum_{i=1}^{m} IG(l_i|x)$, it can be known that the information gain is related to the number of labels. As the number of the datasets is different, one specific threshold is not suitable. To address this problem, it is necessary to design a method that can automatically calculate the threshold value of the sensitive information gain according to different applications [24].

Assume the number of the candidate labels $m = 20$. For two different features $x_1$ and $x_2$, $IGS(L|x_1) = 0.5$ and $IGS(L|x_2) = 1.6$. It can be seen that the importance of $x_2$ is greater than $x_1$, so the feature $x_1$ is reserved. Based on the standard normal transformation,

$$\delta = \frac{1}{n} \sum_{i=1}^{n} |IGZ(L|x_i)|$$ \hspace{1cm} (14)

is used for setting the threshold, where $|IGZ(L|x_i)|$ is the absolute value of the sensitive information gain $IGZ(L|x_i)$. In Eq. (14), the mathematical expectation of the absolute value of the transformed sensitive information gain is used as the current threshold, which can increase the adaptive ability of the algorithm. The selection of the sensitive information feature of the social network is achieved as above.

2.3. Classification and recognition of social network sensitive information. Combined with the feature selection, the two classification problems are used in this paper, which are the positive class and negative class training sample. Because the training speed of the support vector machine is related to the size of the training sample set, the support vector which is distributed on the class boundary plays a decisive role in the decision. Therefore, the number of sensitive information of training samples is reduced by preselecting effective candidate support vectors to improve the training speed. The sample selected from the center distance of the mutual center (the distance between the sample and the disparate class center) is less than the distance of two class sample centers as the effective candidate support vector.

(1) Linearly separable case. The known sample set is, then the average feature of the samples of this class sensitive information is called the center $m$, which is given by

$$m = \frac{1}{n} \sum_{i=1}^{n} x_i$$ \hspace{1cm} (15)

(2) Nonlinearly separable case. Known two vectors $x$ and $y$. They are mapped to the feature space $H$ by the nonlinear function, then the Euclidean distance of
the two vectors in the feature space is given by
\[ d^H(x, y) = \sqrt{K(x, x) - 2K(x, y) + K(y, y)} \] (16)
where \( K(\cdot) \) is the kernel function. The center vector \( m_\phi \) of the samples in feature space is given by
\[ m_\phi = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) \] (17)

According to Eq. (15) or Eq. (17), the class centers of the two classes are obtained, which are the positive class center \( m^+ \) and the negative class center \( m^- \). The distance of the two class centers is given by
\[ D = |m^+ - m^-| \] (18)

By using Eq. (19), the distance from all samples to the disparate class center \( m \) is calculated. The sample with the distance less that \( D \) is selected as effective candidate support vector.
\[ D' = |x_i - m| \] (19)

The sensitive information sample satisfying \( D' \prec D \) is retained as effective candidate support vector, shown in the arc part of Fig. 2. After preprocessing the sensitive information sample, the aim of reducing the number of training samples is achieved and the training speed is improved.

In order to reduce the influence of noise points on the construction of support vector machine, a fuzzy membership sensitivity information function based on class center is proposed. The membership function considers the location of noise point or outlier point far away from the class center. Therefore, the influence of noise point or outlier point can be reduced by giving smaller membership to the sample far from the class center. However, the design of the membership function ignores the support vector is also far from the class center. Thus, the reduction of the noise point or outlier point introduces the reduction of the support vector, as shown in Fig. 3. Therefore, a new membership function is designed to make the membership of the sample increases with the increase of the distance from the class center. In this way, the support vector will obtain larger membership.

The distance of each positive class sample and the positive class center is \( d_i^+ = |x_i^+ - m^+| \) and the distance of each negative class sample and the negative class center is \( d_i^- = |\bar{x}_i - m^-| \). Assume after preselecting support vector the positive
class sensitive information sample set is $X^+$, and the negative class sample set is $X^-$. Then the designed membership function is given by

$$s_{i1} = \begin{cases} 
\frac{d^+_i + \delta}{\max d^+_i}, y = +1, x^+_i \in X^+ \\
\frac{d^-_i + \delta}{\max d^-_i}, y = -1, x^-_i \in X^- 
\end{cases} \tag{20}$$

where $\delta$ is a small enough positive number to avoid $s(x_i) = 0$.

The membership function is based on the measurement of the farthest distance from the class center, which make the sample farthest from the class center has greater membership. In this way, the role of the support vector in the construction of the optimal classification surface is ensured. However, as the noise point is also located in the farthest region, it will enhance the noise point or outlier point.

The membership function is weighted by the nearest neighbor sample density to distinguish the noise point from the normal sample. The more samples in the nearest neighbor $e$ of the sample $x_1$ represent the larger nearest neighbor sample density and the less samples of the noise point $x_2$ represent the smaller nearest neighbor sample density. Therefore, the weighting of the membership function can suppress the noise point. The nearest neighbor sample density function of each sample is calculated to quantify the sample density of the nearest neighbor sensitive information. For each sample $x$, by calculating the sample $x_j$ with the distance satisfying Eq. (12) in the sample set, the nearest neighbor sample subset $X_i$ of the $i$th sample is formed.

$$d_{ij} \leq e_1; i, j = 1, 2, \ldots, \text{numX}; j \neq i \tag{21}$$

where $d_{ij}$ is the distance between the two samples of $x_j$ and $x_i$, $e_1$ is the nearest neighbor of the sample and $\min(d_{ij}) \leq e_1 < \max(d_{ij})$, $\text{numX}$ is the number of the samples in the sample set. Sample density is defined by the distance between the nearest neighbors. Assume there are $k$ nearest neighbor samples in the nearest neighbor sample subset $X_i$, the nearest neighbor sample density function is given by

$$z_i = \frac{1}{\sum_{i \neq j; j=1}^{k} d_{ij}} \sum_{i \neq j; j=1}^{k} \frac{1}{d_{ij} + a}, d_{ij} \leq e_1, i = 1, 2, \ldots, \text{numX} \tag{22}$$
where $a$ is a small penalty constant. The normalization of the nearest neighbor sample density is given by

$$w_i = \frac{z_i}{\sum_{i=1}^{numX} z_i}$$

(23)

where $z_i$ is the nearest neighbor sample density of the $i$th sample $x_i$. The samples in the nearest neighbor are more, $w_i$ is larger.

Because the influence of the classification of the nearest neighbor sample on the category of the sample is different, the nearest neighbor sample density can be adjusted. If the nearest neighbor sample subset is the similar class sample, the sample is not confused with the disparate class sample and the nearest neighbor sample density is kept constant. If the nearest neighbor sample subset includes the disparate class sample, the sample is confused with the disparate class sample and the membership is decreased. If the nearest neighbor sample subset is the disparate class sample, the membership is set to 0, in order to reduce the effect on construction of support vector machine.

The final membership function is given by

$$s_i = s_{i1} \cdot w_{i1}$$

(24)

The new membership function is combined with the membership function based on class center and nearest neighbor sample density, which can enhance the support vector and reduce the noise. The membership function obtained with Eq. (24) is used to train the fuzzy support vector machine for classification. The intelligent recognition of social network sensitive information is achieved with the classification result.

3. **Experimental results and analysis.** To verify the proposed algorithm, simulation experiment is carried out and analyzed with the current algorithm. In the experiment, 4 artificial data sets and 5 UCI data sets are used for classification. Experimental environment is with CPU Intel i52, 60GHz, 4GB RAM, 64 bit Windows8, and MATLAB2017.

![Figure 4. Classification result obtained with the current method](image-url)
In the experiment, the sensitive information training sample set is randomly generated two classes of two-dimensional samples with normal distribution, which are positive class sample and negative class sample, and the 2% random noise data is added. A random two-dimensional sample is taken as a test sample with addition of 1% noise data. The parameters of the two algorithms is selected the same \((C = 100)\). The selection of \(e_1\) is different with the sample and \(3\sim5\) times of \(\min(d_{ij})\). With the increase of the data set, the classification results of the two algorithms are shown in Fig. 4 and Fig. 5 for the case of the 200 positive samples and 200 negative samples. In Fig. 4 and Fig. 5, triangle represents positive class sample, circle represents negative class sample, and square represents the deleted sample after preselecting candidate support vector.

From Fig. 4 and Fig. 5, it can be known that, with the increasing number of sensitive information training samples, the proposed algorithm has improved in training time and classification accuracy compared with the current algorithm. A large number of kernel matrix operations and storage are carried out in support vector machine training to make the training slow, and the size of the matrix is related to the number of training samples. As the current method does not reduce the number of samples, the training time is longer, and the current algorithm adds class centrifugal degree to set membership, which results in the slower training speed. However, the proposed algorithm preselects candidate support vector and deletes some samples to reduce the number of training samples. Experiment proves that the time cost of the proposed algorithm in membership setting process is less than the time cost of training samples. In addition, the proposed algorithm gives support vector greater membership to enhance its role, and weights the membership with the consideration of smaller nearest neighbor sample density of the noise point and the outlier point, which can effectively reduce the noise point or the outlier point and improve the classification accuracy.

Assume \(S\) is the set of all sensitive information related to the information identification in the social network, \(R\) is the set of the recognized sensitive information, \(s\) is the number of the recognized related sensitive information for one detection, \(m\) is the number of the recognized unrelated sensitive information for one detection, and \(n\) is the number of unrecognized related sensitive information. The recall ratio
Re and the precision ratio Pr are expressed as

\[
Re = \frac{P(S \cap R)}{P(S)} = \frac{s}{s+n} \tag{25}
\]

\[
Pr = \frac{P(S \cap R)}{P(R)} = \frac{s}{s+m} \tag{26}
\]

The high recall ratio and precision ratio represents the good performance. But generally, they are contradictory. When the precision ratio is high, the recall ratio is low, and vice versa. Comparison of the recall ratio and the precision ratio between the proposed algorithm and the original algorithm is shown in Fig. 6. As the test sensitive information is increasing, the trend of the recall ratio is rising and the trend of precision ratio is down. As the support vector machine in the proposed algorithm eliminates some error samples which can affect the result, and the precision ratio and the recall ratio is improved compared with the original algorithm without adding location information. The results show that the information recognition rate of the proposed algorithm is high, and it can effectively improve the security of network operation.

4. Conclusions. For the poor accuracy of social network sensitive information recognition, the social network sensitive information recognition algorithm based on classification technology is researched in this paper. By using the improved support vector machine classification algorithm to identify sensitive information, the proposed algorithm improves the training speed and classification speed of fuzzy support vector machine without losing the classification accuracy, and effectively improves the accuracy rate of sensitive information recognition.

Acknowledgments.

1. The Key research projects of Ministry of public security: Railway security information system development based on cloud computing (No. 201202ZDYJ017);
2. The Science and technology project of Henan Province: Research on text sensitive information recognition and classification in big data environment (No. 2017172102210111);
3. The Key scientific research project plan of Henan province: The research on the application of clustering algorithm in the identification of OOV words in social network (No. 17A520053).

REFERENCES

[1] M. Abdelwahab and M. Abdelwahab, Human action recognition and analysis algorithm for fixed and moving cameras, in Electronics Letters, **23** (2015), 1869–1871.
[2] C.-C. Chung, W.-Y. Dzan, Y.-M. Cheng and S.-J. Lou, On the push-pull mobile learning of electric welding, Eurasia Journal of Mathematics Science & Technology Education, **13** (2017), 3235–3260.
[3] X. Du, Target recognition algorithm for fused hyperspectral image by using combined spectra, Spectroscopy Letters, **48** (2015), 251–258.
[4] X. Fan, K. Zheng, Y. Zhou and S. Wang, Pose locality constrained representation for 3d human pose reconstruction, Journal of Jilin University (Information Science Edition), 1–7.
[5] W. Gao, L. Zhu, Y. Guo and K. Wang, Ontology learning algorithm for similarity measuring and ontology mapping using linear programming, Journal of Intelligent & Fuzzy Systems, **33** (2017), 3153–3163.
[6] J. Hou, Z. C. Wen and J. F. Lai, A constrained optimization reformulation of the generalized nash equilibrium problem, Journal of Interdisciplinary Mathematics, **20** (2017), 27–34.
[7] Z. Huang, Improved adaboost detection algorithm and application in identity authentication, Bulletin of Science and Technology, 190–192.
[8] K. Kahl and J. S. Sirkis, Damage detection in beam structures using subspace rotation algorithm with strain data, Avaa Journal, **34** (2015), 2609–2614.
[9] K. Kim, S. H. Kong and S. Y. Jeon, Slip and slide detection and adaptive information sharing algorithms for high-speed train navigation systems, IEEE Transactions on Intelligent Transportation Systems, **51** (2015), 3193–3203.
[10] S. G. Konov, A. A. Khokholikov and V. V. Skvortsova, Algorithm for rapid recognition of measurement markers for non-contact measurement systems, Measurement Techniques, **58** (2015), 845–847.
[11] A. Leon M., L. David J., W. Steven C.R., J. Martha, M. Christoph, S. Johannes, J. Karl-Heinz, W. Christian and M. Andre F., Making use of longitudinal information in pattern recognition, Human Brain Mapping, **72** (2016), 4385–4404.
[12] B. Li, Y. Tang and T. Han, Research on human face recognition based on improved nmf algorithm, Computer Simulation, **3** (2016), 428–432.
[13] H. Lim, T. Park and N. S. Kim, Joint optimisation of computational accuracy and algorithm parameters for energy-efficient recognition algorithms, Electronics Letters, **51** (2015), 1238–1240.
[14] A. K. Malhi and S. Batra, An efficient certificateless aggregate signature scheme for vehicular ad-hoc networks, Discrete Mathematics and Theoretical Computer Science, **17** (2015), 317–338.
[15] G. Napolitano, A. Marshall, P. Hamilton and A. T. Gavin, Machine learning classification of surgical pathology reports and chunk recognition for information extraction noise reduction, Artificial Intelligence in Medicine, **70** (2016), 77–83.
[16] G. X. Ritter, J. A. Nieves-Vázquez and G. Urcid, A simple statistics-based nearest neighbor cluster detection algorithm, Pattern Recognition, **48** (2015), 918–932.
[17] D. Shashikumar and S. Srinivas, 3D human activity recognition by indexing and sequencing (risq), Nature, **367** (2015), 480–483.
[18] C. Szabó, L. C. Morgan, K. M. Karkar, L. D. Leary, O. V. Lie, M. Gironard and J. E. Cavazos, Electromyography-based seizure detector: Preliminary results comparing a generalized tonic-clonic seizure detection algorithm to video-eeg recordings, Epilepsia, **56** (2015), 1432–1437.
[19] D. Wang, H. Lu and M. H. Yang, Kernel Collaborative Face Recognition, **10**, Elsevier Science Inc., 2015.
[20] H. Wang, X. Su, X. Lu and M. Cao, Based on the improved grey relational algorithm platform for the airborne radar emitter recognition method, *Journal of China Academy of Electronics and Information Technology*, 523–526.

[21] Y. Wei, Assessment study on brain wave predictive ability to policemens safety law enforcement, *Journal of Discrete Mathematical Sciences & Cryptography*, 20 (2017), 193–204.

[22] Y. Xu, Z. Li, B. Zhang, J. Yang and J. You, Sample diversity, representation effectiveness and robust dictionary learning for face recognition, *Information Sciences*, 375 (2017), 171–182.

[23] G. Zengtai and W. Qian, On the connection of fuzzy hypergraph with fuzzy information system, *Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology*, 33, 1665–1676.

[24] X. Zhang, C. Mei, D. Chen and J. Li, Feature selection in mixed data: A method using a novel fuzzy rough set-based information entropy, *Pattern Recognition*, 56 (2016), 1–15.

Received June 2017; revised December 2017.

*E-mail address:* liweiping@rpc.edu.cn
*E-mail address:* wuhaiyan@rpc.edu.cn
*E-mail address:* jieyang@whut.edu.cn