Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IMPACT FACTOR: 6.017

IJCSMC, Vol. 8, Issue. 2, February 2019, pg.7 – 12

Privacy-Preserving Knowledge Discovery in Distributed Databases

Sefer Kurnaz[1], Hanaa Khudhur Obaid[2]

[1]Computer Engineering& Altınbaş University, Turkey

[2]Information Technology & Altınbaş University, Turkey

[1] sefer.kurnaz@altinbas.edu.tr; [2] hanaaaliraqi0@gmail.com

Abstract— Data mining is worried with the mining of beneficial information since several kinds of information. In data that concern with human or other specific fields confidentiality anxieties are occupied additional extremely than further data mining responsibilities. In this study vertically, divider the information and mine these separated data at manifold sites. Then, the deep auto-encoders used to classify the yield data, which first auto-encoder learn important features from the input data and the output become input to the auto-encoder2. Furthermore, the auto-encoder2 also reduce the number of features and extracted only the affective features according. Then, the productivity of auto-encoder2 wired to the SoftMax and trained by using supervised technique to classify the features to the classes. Finally, the auto-encoders and SoftMax set and trained in supervised method using labelled data. Several datasets used to validate the proposed method and the obtained results compared with well-known researches in this filed.

Keywords— Auto-encoder, SoftMax, Privacy-Preserving Knowledge Discover, vertically separation, Cancer.

I. INTRODUCTION

The propagation of data knowledges and the internet in the last two decades has transported a prosperity of separate data into the indicators of profitable corporations and administration activities. As hardware prices go depressed, establishments discovery it calmer than forever to retain any portion of data learnt from the continuing actions of their customers. Information proprietors continually pursue to brand healthier usage of the information they own and apply information quarrying methods to mine valuable data and designs since the information. In outcome, there is a rising anxiety around the aptitude of information landlords, such as big companies and management activities, to misuse this information and cooperation the confidentiality of their customers anxiety which has been responded in the activities of lawmaking forms (for example, the discussion around and following removal of the Whole Data Consciousness task in the US [1]). This anxiety is worsened by real events that prove how problematic it is to usage and portion data while defensive persons' confidentiality. One instance is from 2006 [2], when AOL available on their webpage an information group of 20 million web hunts for investigation dedications. Though the feature set was supposed to be anonymized, New York Times journalists have exposed in what way the free data can be used to depiction the individualities of the rescuers and study fairly a ration about them. Additional instance tells to the Netflix prize contest1 that removed home among October 2006 and September 2009. Netflix has issued a features containing of extra than 100 million film scores since completed 480 thousand of its clienteles and asked the investigation municipal to struggle for developments to its endorsement process. To defend client confidentiality, Netflix detached wholly separate data recognizing separate clienteles and disconcerted some of the film assessments. Notwithstanding these defences, academics have exposed that with comparatively slight supplementary data anonymous clienteles can be re-identified [3].

In this paper, vertical and horizontal separation are combined with deep auto-encoder to classify cancer dataset that obtained from UCI.

II. Page Layout

In this section, the techniques that applied in this paper explained in detail phase such as deep auto-encoder and separation techniques.

A. Deep auto-encoder

An Auto-encoder (AE) is a kind of ANN applied to extract effective features coding in an unsupervised learning. The purpose of an AE is to extract a demonstration (encoding) for a group of features, naturally for dimensionality decrease. AE are able to explain the delinquent of “backpropagation without a teacher”. There are several of AE kinds like: sparse AE, denoising AE and in whole AE. AE consist of three layers, namely the input layer, hidden layer and the output layer. SAE is fundamentally a regular unsupervised NN, which is able to extract the sensitive features of an input features in an unsupervised way by decreasing the discrepancy between the input and output. Deep auto-encoders is a kind of DNN consist from SSAE and also SoftMax classifier is frequently applied as the last layer for cataloguing (which mean discrete output such as classify the images into two classes female and male) and regression problems (which mean the output of the model is continues such as estimate the temperature). In last years, the SAE have been usually used to several inclusive investigation spaces, like speech recognition, medical image classification and computer vision. The first stage is pre-processing stage that the SSAE1 is trained by retaining an unsupervised learning method. Fundamentally, the yield of the first SAE wired an input to the SSAE2, and the outcomes of SSAE2 wired as an input to the SoftMax classifier correspondingly. An instance AE is showed in Figure 1. The planning among input and output is assumed following equations:

$$\dot{x} = f(x) \tag{1}$$

$$n_1^{(1)} = M_f(w_{11}^{(1)} x_1 + \cdots w_{15}^{(1)} x_{5+} + b_1^{(1)}) \tag{2}$$

$$n_i^{(1)} = M_f(w_{i1}^{(1)} x_1 + \cdots w_{i5}^{(1)} x_{5+} + b_i^{(1)}) \tag{3}$$

Where M () is an instigation applying sigmoid logistic function.

The last appearance container be illustrated as surveys (4):

$$n_{w,b}(x) = M_f(w_{11}^{(2)} n_1^{(2)} + \cdots w_{15}^{2} n_5 + \cdots + b_1^{(2)}) \tag{4}$$

Where w represented the weights, x represented the input data (features), and b represented the bias. Then, n represented the new features that produce from each layer.

Moreover, M represented that activation function used in this problem for example sigmoid, rectified linear unit (ReLU) etc.

Generally, the any auto-encoder consist from input, hidden, and output layers. The input layer represented the data that we want to classify it. Furthermore, Hidden layer represented the layers that try to extracted the sensitive and effective features from input data where number of hidden layers depended on the problem and its properties and the number of nodes in hidden layer represented the number of features in this layer.

Finally, the output layer represented the number of classes that the data classified on it see Figure 1. For example in cancer dataset there are 13 features, then the input layer consist from 13 node each of them represented one feature. The hidden layers reduce the number of features using the Eq(1,2,3,4) from 13 to 12, 11 or 9 this determined by doing number of experiment and investigate the results. Finally, the output layer represented the number of classes there is cancer disease or not which mean there is to classes.

The inconsistency among the input and output is definite by output a fitness function. This functions' primary period mentions the MSE while the another unique is the regularization period. Dissimilar processes are favoured to resolve the best parameters of the net [4,5,8].

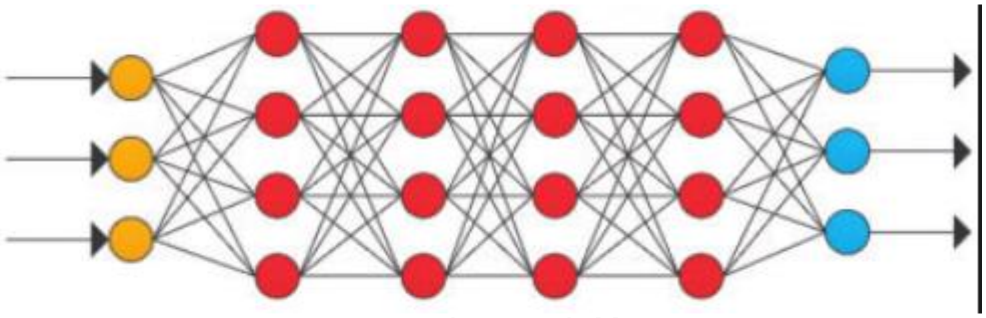

Fig. 1: Simple Auto-encoder [5]

B. Vertically Data Distribution

The information for a precise object is separated crossways numerous places, and each position has data for each particular object for an exact subsection of the features. We trust that realism of an object in a specific places record may be unprotected; it is the standards connected by an object that are subtle. The goal is to group the documented set of shared objects deprived of revealing slightly of the standards that the grouping [6].

C. Dataset

Examples land occasionally as Dr. Wolberg intelligences his medical suitcases. The database consequently reproduces this consecutive grouping of the data [9]. The first 10 attributes are the features that represented the cancer features which by investigating these features the physicians can decide there is disease or not and the last attribute 11 represented the target which mean there is cancer or not.

III.PROPOSED METHOD

A Novel approach for Privacy Preserving using vertically data distribution and deep learning techniques, Deep Learning is a machine learning technique that reproduces the mechanisms of the person brain in processing information and making designs for apply in choice creation. Deep learning is a subsection of machine learning in Artificial Intelligence (AI) that has nets accomplished of learning unsupervised after information that is shapeless or unlabelled. Then one of the most used deep learning which is AE used in this study. The proposed approach consists from two stages, in the first one the simple impression of perpendicular parting is that only the feature proprietor has the whole data set and individually 3rd gathering has only a ration of the features set. Therefore, feature privacy is endangered. Revenue the Wisconsin predictive breast cancer data set as an instance. Let's approximately the feature proprietor chooses to request numerous 3rd gatherings to study the feature. There are 9 features in this data set. We can eliminate one feature at a period and generate 9 sub-data sets. All deputise-dataset has lone eight features. Then, in the second stage the deep learning technique used to extract high level features (which mean minimum number of features which effected the accuracy result and produce maximum classification rate) and classify the data in to classes and repeat these operations in the number of row and compare the results see Figure 2. The aim of experiment is to classify the input data into there is cancer or not and try to find best accuracy of classification and investigate which features effected the accuracy when removed in Vertically Data Distribution operation to implement Privacy-Preserving Knowledge Discovery.

The proposed approach validated using number of datasets and obtained results compared with several studies proposed in this fields.

The steps of flowchart Figure 1 presented below in detail face.

1. Start
2. Read data (any data can be used in this thesis cancer data used)
3. Data Separation Techniques applied to the datasets that used to evaluate our study. In this cancer dataset that published in UCI website used. Furthermore, both Vertically Data Distribution and Horizontal Data Distribution are applied each of them alone to the cancer dataset.
4. Divide the cancer dataset into Train and Test groups.

5. The auto-encoder1 have five parameters identified by the user, hidden Size= 9, Max Epoch= 100, L2 Weight Regularization= 0.0011, Sparsity Regularization= 3, Sparsity Proportion= 0.002.
6. Train the auto-encoder1 in unsupervised fashion.
7. The auto-encoder2 have five parameters identified by the user, hidden Size= 8, Max Epoch= 50, L2 Weight Regularization= 0.0012, Sparsity Regularization= 2, Sparsity Proportion= 0.001.
8. Train the auto-encoder2 in unsupervised fashion by data that wired from auto-encoder1.
9. Train SoftMax in supervised fashion by data wired from auto-encoder2.

For instance, in cancer dataset there are 13 features, then the input layer consists from 13 nodes each of them represented one feature. The hidden layers reduce the number of features using the Eq (1,2,3,4) from 13 to 12, 11 or 9 this determined by doing number of experiment and investigate the results. Finally, the output layer represented the number of classes there is cancer disease or not which mean there is to classes.

10. Compare if training process complete or not? If complete go to 11 else go to 9.
11. Test the trained network by using testing data.
12. If the test process complete then go to 13 else go to 11.
13. Calculate the accuracy of the network.

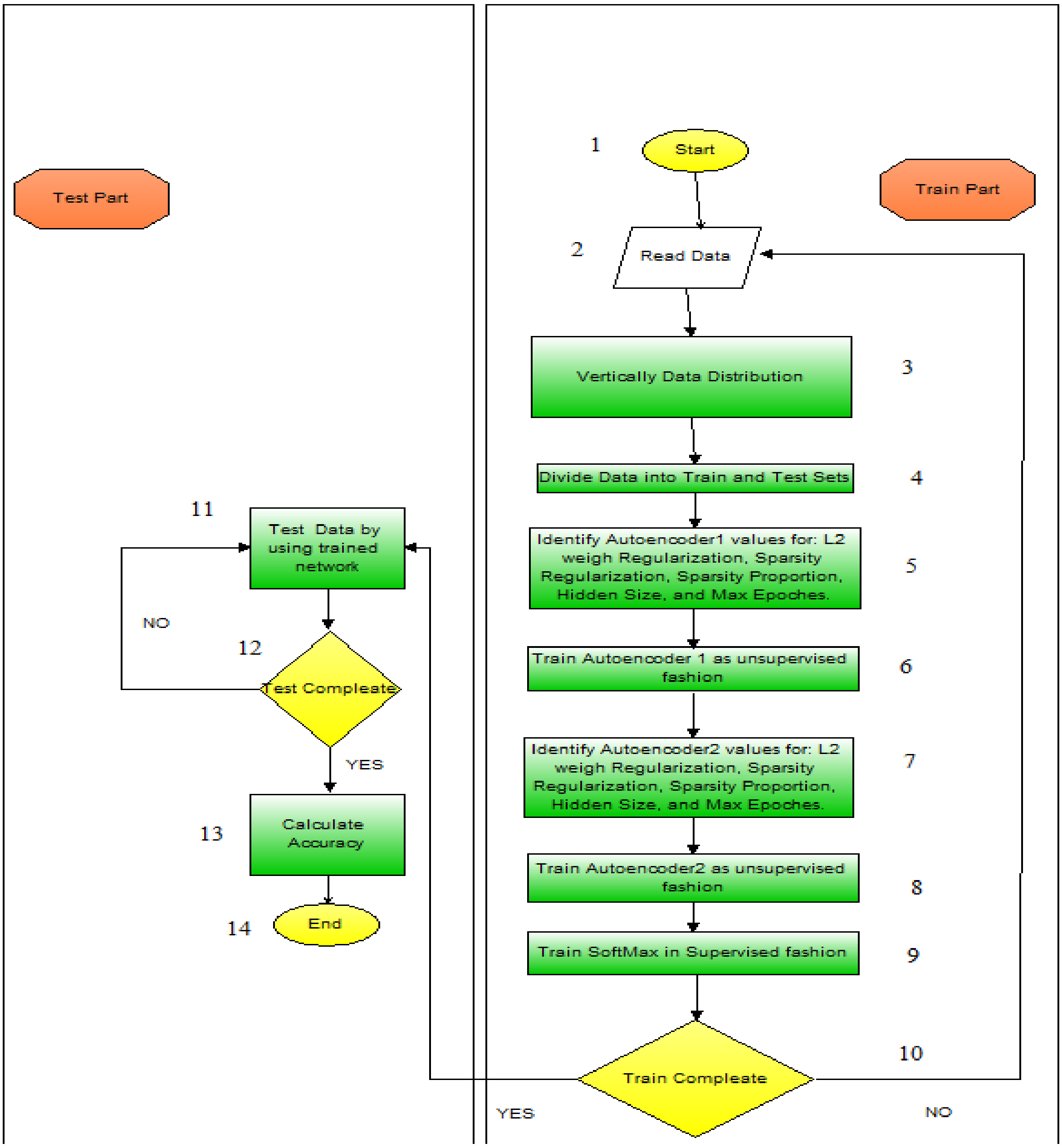
Test Part
Train Part
1
Start
2
Read Data
3
Vertically Data Distribution
4
Divide Data into Train and Test Sets
5
Identify Autoencoder1 values for: L2 weigh Regularization, Sparsity Regularization, Sparsity Proportion, Hidden Size, and Max Epoches.
6
Train Autoencoder 1 as unsupervised fashion
7
Identify Autoencoder2 values for: L2 weigh Regularization, Sparsity Regularization, Sparsity Proportion, Hidden Size, and Max Epoches.
8
Train Autoencoder2 as unsupervised fashion
9
Train SoftMax in Supervised fashion
10
Train Compleate
YES
NO
11
Test Data by using trained network
NO
12
Test Compleate
YES
13
Calculate Accuracy
14
End

Fig. 2: Proposed Method

IV. PROPOSED METHOD

A machine with Intel Core i5–6700 CPU 4-GB RAM is applied for achieving our method which applied to cancer dataset that obtained from UCI website. The cancer data that obtained from UCI website are become input to the auto-encoder1 which try extracted the sensitive features that presented high accuracy which only 8 features are extracted from input data. Then, autoencoder2 applied to extracted new features from output of auto-encoder1. Finally, SoftMax used to classify the obtained features.

The experimental results presented remarkable results when compared with well-known studies in this field see TABLE I.

TABLE I: RESULTS

Methods	Results
Gang Kou et al. [1]	96.68
Gang Kou et al. [1]	97.4
Our Method	99.5

V. CONCLUSIONS

Privacy-protection is a chief subject in medical information mining. This paper examines data parting methods in medical information recognition. The experimentations validate that data parting methods can not only defend information privacy, and sometime also growth classification accuracy.

In this thesis, deep auto-encoders used to classify the data sets by applying both vertical separation and horizontal separations. The extracted features by using auto-encoders are classified by using SoftMax. The SoftMax is multi-classes transfer function which used as output layer in this study. Then, the auto-encoders and SoftMax are set and trained in supervised technique to increase the classification accuracy which backpropagation applied to train the stacked model.

The proposed method presented 99.5% which is remarkable result when related with earlier results applied to the same dataset.

Furthermore, the proposed techniques can have applied to various critical fields personal information's that Sold by companies, hospital information's and data that sold by communication companies.

As future work the both separation techniques can be applied with new deep learning techniques like convolutional neural network, recurrent neural network and deep belief neural network. Moreover, new machine learning combinations can also have proposed by combining unsupervised techniques such as principal components analysis, self-organizing map and K-mean with supervised learning techniques such as: SVM, ANN and NB.

REFERENCES

[1] Charu C. Aggarwal. On k-anonymity and the curse of dimensionality. In VLDB, pages 901-909, 2005.

[2] Charu C. Aggarwal and Philip S. Yu. A condensation approach to privacy preserving data mining. In EDBT, pages 183-199, 2004.

[3] Charu C. Aggarwal and Philip S. Yu. Privacy-Preserving Data Mining: Models and Algorithms. Springer Publishing Company, Incorporated, July 2008.

[4] Karim, A. M., Güzel, M. S., Tolun, M. R., Kaya, H., and Çelebi, F. V., "A New Generalized Deep Learning Framework Combining Sparse Auto-encoder and Taguchi Method for Novel Data Classification and Processing," pp. 1–22.

[5] Karim, A. M., Güzel, M. S., Tolun, M. R., Kaya, H., & Çelebi, F. V. (2019). A new framework using deep auto-encoder and energy spectral density for medical waveform data classification and processing. Biocybernetics and Biomedical Engineering, 39(1), 148-159. doi:10.1016/j.bbe.2018.11.004.

[6] Kou, G., Peng, Y., Shi, Y., & Chen, Z. (2007). Privacy-Preserving Data Mining of Medical Data Using Data Separation-Based Techniques. Data Science Journal, 6, S429-S434. doi:10.2481/dsj.6.s429.

[7] M. Wegmuller, J. P. von der Weid, P. Oberson, and N. Gisin, "High resolution fiber distributed measurements with coherent OFDR," in *Proc. ECOC'00*, 2000, paper 11.3.4, p. 109.

[8] Karim, A. M., Çelebi, F. V., and Mohammed, A. S., "Software Development for Blood Disease Expert System," Lect. Notes Softw. Eng., vol. 4, no. 3, pp. 179–183, 2016.

[9] O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.