titel

***Abstract*—Text classification is significant task in the area of text mining, and it require now more than before due to the increasing of using the data on the Internet. In last year, researcher has been trying to find features that could help to create accurate model in order to recognize the data for all classes in classification. In a given a set of class, recognize one class from the other classes, is classification. Accurate classifier, which determines all known classes during training test, is called Close World. In constant, unknown classes during training test time is the Open World. Open-set World is a classifier that not identified unseen classes during training to other classes, which lead to lack of performance to state-of-art classifier. This survey presents study of open classification for text and image.**

I. Introduction

II. Open Set Classification

A. Open Set in Text:

DOC: Deep Open Classification of Text Documents paper:

(shu, xu,liu, 2017) proposed a novel deep learning approach for open classification world. It performs better than existing state-of-the-art techniques. They proposed method similar to CBS, but outperforms is signally better.[fei 2016]. (fei 2016) added capability of learning new class which is important because any system will not be able to learn by itself without the ability to learn and identify a new things.

Their proposed system is Deep Open Classification (DOC) which uses deep learning. To decrease open space risk, the use 1-vs-rest final layer of sigmoid with multi class classifier. DOC uses CNN because CNN perform well in text [4]. They used the same datasets that used in (fei,2016). However, they compared DOC with two state-of-the-arts. First one is cbsSVM for text classification (fei 2016). Second one is OpenMax that is for computer vision (boult,2016), but they adapt it for text classification. Their result show DOC is significantly better than cbsSVM and OpenMax. It performs better than in state-of-the-art for both text and image classification.

Social Media Text Classification under Negative Covariate Shift Paper:

Fei and Liu (2015) proposed a novel technique called Center Based Similarity (CBS) to solve the problem of covariate shift in classification. The key is the transformation of orignal document from D-space with n-gram space to CBS space because the CBS learning in similarity space. The new space in the proposed technique of covariate shift problem is qualify to built much better classifiers. The difference between D-space and CBS is document representation. However, That's is suitable for open classification (fei, 2016). The problem of covariate shift in Machine Learning is a type of sample selection bias. Also, they used Amazon review datasets. Fei and Liu (2015) conclude that the proposed technique performed better than baseline such as SVM, and it improve classification.

breaking the closed world Assumption in Text Classification paper geli fei and bing liu 2016:

(fei and liu, 2016) tried to solve open-set problem by proposing CBS space learning strategy in order to decrease open space risk, and balance the empirical risk for open classification.The researcher proposed CBS space method which calculate a center for every class and then transforms every document to vector to the center. After bulit the classifier by using transformed data, the surface is like a ball surround every class, but every outside the ball is become unknown.

They did extensive experiment, and they use two data sets. First data set is 20-newsgroup that contain 18828 documents. Second data set is Amazon reviews that contain 50 types of products, and every product has 1000 reviews [1] [2]. Their experiment show that cbsSVM on multi-class open set text classification makes an excellent classifier compared to state-of-the-art methods. The solution for CBS learning reduce the positive label area from infinite space to finite space which significantly decrease open space risk.

They explain that the proposal solution is hopeful, but they need to design robust solution that works with known class. The reason for classifier is not in-formed enough to refuse unknown classes is because of significant open space risk in their proposal solution.

Breaking a challenge for text classification in open world recognition (tri,kalata, 2017)

(doan and kalita, 2017) proposed Nearest Centroid Class classifier. Their goal of the classifier is to expose unknown classes incrementally. They test NCC on document classification on many domains, and they found promising results. They said that Rondom Forests is promising model more than SVM because of RF ability to remove and add components. Distance-based method such as NCM Nearest Class Mean present each class by mean, and sphere shape of centered class locate boundary of known class. Outside the sphere shape is outlier on unknown in the open space. However, their proposal model is inspired by the logic of NCM. They designed set of closest centroid class by using DBSCAN algorithm which is clusters that represent one data point.

They use the same data sets presented in " breaking the closed world Assumption in Text Classification paper geli fei and bing liu 2016:". Their model perform well, and gradual loss in performance as more as unknown class show in testing.

B. Open Set in Vision:

UNSEEN CLASS DISCOVERY IN OPEN-WORLD CLASSI- FICATION paper:

(Shu el at.,2018) trying to discovering unseen classes that not appear in training, and reject it. Also, classifier in open

world classification is to classify the test example data to seen class. Therefore, they proposed the first a joint open image classification model with sub-model for classifying to find out the pair example refers to same class or different class by using only the seen class training data. Their theory about that model because they have the data for seen classes, and they know similarity or difference for the test example from the same or different class. So, they assuming that may work on rejected example test. (SHu,...) aim to convey the class similarity recognition learned by clustering algorithm from seen class to unseen class, and it is from Supervised to Unsupervised learning[3]. They used combination of Open Classification Network (OCN) that is for class classification for seen and unseen classes, and Pairwise Open Classification (PCN), which classifies the two examples weather are from the same or different class. In addition, they used auto encoder from unlabeled example to learn Unsupervised representations, and they used clustering method that cluster the reject example. (shu,...) used two datasets MNIST and EMNIST. For evaluation, they performed two evolutions the Number and Quality clusters. (shu....) conclude that it is important to discover the hidden class during training from reject example. That's will lead to learn the system automatically. Their experiments explain the efficiency of the proposed work.

Towards Open Set Recognition paper 2013 *:

(Scheirer et al., 2013) integrate open space risk and empirical risk because of existence of space, and specified it as a relative measure. They devised an extension of existing binary SVMs and one class for problem of open classification. Open space risk realize that unknown classes are likely to bring errors to classification decisions. By trained classifier, empirical risk founded from test example that misclassified. Their proposed method result reduced risk by changing the half space of a binary linear classifier with a positive region limited by two parallel hyperplanes. Also, developed algorithm that modifies SVM liner, which incrementally moves the planes. While the positively labeled is decrease compared to the half space in SVM linear, their risk is still infinite.

Multi-class Open Set Recognition Using Probability of Inclusion 2014 paper:

[**?**].

(Scheirer et al., 2014) introduce the novel idea of fitting a robust single-class probability model over the positive class scores from a classifier. The using of binary classification model helps to distinguish the positive class from the known negative classes. The one class probability is detecting the decision boundary which makes unseen classes are not frequently misclassified as belonging to the positive class. Also, they introduced PI –SVM algorithm for modeling the unnormalized posterior probability of class inclusion. Their research extends the recent learning work, which is limited to closed set problems for Scheirer et al. [43,42], but their proposed work for inclusion for open set problems that extension directly models the probability of inclusion. (Scheirer et al., 2014) formulate Compact Abating Probability (CAP) model that show how to mange risk by thresholding the probabilistic output of one class RBF SVM. They use the probabilistic output of RBF One class SVM, and they combine RBF One class SVM with Weibull. Decision thresholds should be chosen depend on the previously knowledge of the ratio of unknown classes in testing that is a weakness of the methods.

Towards Open World Recognition paper 2015:

(Bendale and Boult, 2015) proposed theory to reduce the weighted sum of open space risk and empirical risk by using thresholding sums of monotonically decreasing recognition functions, and they extend an alternative approach Nearest Centroid Classifier (NCM) for open world recognition (Rocchio, 1971) [5]. This classifier performs classes by the mean feature vector, and unseen example test is allocating a class with close mean. For open classification, The Nearest Non-Outlier (NNO) algorithm (Bendale and Boult, 2015) adapts NCM that evolves model to manage open space risk. They obtained a protocol for evolution using Image-Net Large Scale Visual Recognition Competition 2010 dataset for open world recognition. That protocol performs significantly better on open world recognition on their NNO algorithm, and it is similar to NCM on close world. Also, they earn robust to open world.

III. Conclusion

do not forget to draw a figure