A semisupervised classification algorithm combining noise learning theory and a disagreement cotraining framework

Zaoli Yang[a], Weijian Zhang[b], Chunjia Han[c], Yuchen Li[a*], Mu Yang[c], Petros Ieromonachou[d]

[a]College of Economics and Management, Beijing University of Technology, Beijing, China [b]Department of Statistics, Tianjin University of Finance and Economics, Tianjin, China

[c]School of Business, Economics & Informatics, Birkbeck, University of London, London, UK

[d]Department of Systems Management and Strategy, University of Greenwich, London, UK *Corresponding author: liyuchen@bjut.edu.cn

Abstract: In the era of big data, the data in many business scenarios are characterized by a small number of labelled samples and a large number of unlabelled samples. It is quite difficult to classify and identify such data and provide effective decision support for a business. A commonly employed processing method in this kind of data scenario is the disagreement-based semisupervised learning method, i.e., exchanging high-confidence samples among multiple models as pseudolabel samples to improve each model's classification performance. As such pseudolabel samples inevitably contain label noise, they may interfere with the subsequent model learning and damage the robustness of the ensemble model. To solve this problem, a semisupervised classification algorithm based on noise learning theory and a disagreement cotraining framework is proposed. In this model, first, the probably approximately correct (PAC) estimation theory under label noise conditions is applied, the relationship between the label noise level and model robust estimation in the process of multiround cotraining is discussed, and a disagreement elimination algorithm framework based on multiple-model (feature argument and select (FANS) algorithm and L1 penalized logistics regression (PLR) algorithm) cotraining is constructed based on this theoretical relationship. The experimental results show that the algorithm proposed in this paper gives not only a high-confidence sample set that meets the upper bound constraint of the label noise level but also a robust ensemble model capable of resisting sampling bias. The work performed in this paper provides a new research perspective for semisupervised learning theory based on disagreement.

Keywords: semisupervised classification, noise learning theory, disagreement cotraining, feature argument and select algorithm, L1 penalized logistics regression algorithm

1. Introduction

The rapid development of the digital economy provides great convenience for the collection and recording of various business data. For example, in addition to various identity information actively uploaded by users, customer information collection systems also include all types of electronic information automatically collected by intelligent devices, e.g., geographical location and path, application (app) installation list, social records, and hundreds or even thousands of other features [1-2]. However, in many business scenarios, it is extremely difficult to automatically identify data samples and assign category labels. Although manual labelling may be introduced, it has the disadvantages of high cost and low efficiency. Such scenarios include financial risk prediction [3], text classification [4-5], medical image recognition [6] and potential customer classification [7]. The data collected in such scenarios show that the number of labelled samples is quite small, while the number of unlabelled samples is quite large, showing typical characteristics

of semisupervised learning problems [8]. The way to maximize the use of data with this structure and to build a robust machine learning system is an urgent problem to be solved.

The semisupervised learning method is applied to solve the problem of combining a large amount of unlabelled data with a small amount of labelled data for learning [9-11]. In this method, the learner automatically modifies the statistical model constructed based on a small amount of labelled data with unlabelled data. Currently, there are four mainstream paradigms of semisupervised learning methods, i.e., generative model-based methods [12-14], graph-based methods [15-17], semisupervised support vector machine methods [18-20], and disagreement-based methods [21-23]. The disagreement-based method is the main form discussed in this paper. Such a form is less affected by model assumptions and loss function forms and has the characteristics of wide applicability [23]. Disagreement-based semisupervised learning begins with the cotraining model proposed by Blum and Mitchell [24]; such a model naturally requires data to have a variety of full views. For example, image data include two views: graphics and their accompanying text descriptions. Based on these two views, two different models are constructed. After completing the model training, the models select a part of specific samples (high-confidence samples, that is, the probability prediction given by the model is close to 0 or 1), attach the category labels predicted by the model (pseudolabel samples), and add them to each other's training set to mutually improve the model performance. This process of "mutual learning and common progress" will continue to iterate until the two learners do not change or reach the predetermined number of learning rounds. The method of building multiple models based on multiple views for alternate learning also derives the collaborative regularization method [25] and collaborative expectation-maximization method [26]. The precondition of this method is that sufficient multiple views are provided. However, in most scenarios, the data do not have the feature of multiple views. Even if there are multiple views, it is difficult to ensure their sufficiency for estimating the probably approximately correct (PAC) learner.

The single view scenario is relative to the multiple view scenario. The typical case of disagreement-based, semisupervised learning in the single view scenario is the ‘tri-training method’ proposed by Zhou and Li [27]. The core idea of this method is to build three different learners on the same dataset and label the sample points that can be reached by any two learners with category labels to train the third learner. Motivated by this idea, the "co-forest method" has been derived [2730]. In this method, multiple classifiers are utilized for simultaneous learning. For a certain unlabelled sample point, class label labelling by the majority model group and learning by the minority model group are performed. Various mechanisms are introduced to ensure that there is significant disagreement between two classifiers. Based on this approach, many scholars have proposed multicotraining (MCT) to improve the performance of document classification [31]; deep multiplanar cotraining [32]; cotraining-based, semisupervised, attribute reduction algorithms [33]; a deep cotraining method by ensemble of deep segmentation models [34]; a self-paced cotraining framework [35]; and cotraining-based noise correction [36].

> According to the comparison of the abovementioned two different styles of the single view method and multiple view method, the premise for the performance of the disagreement-based method is that there are enough disagreements among multiple learners and that it is required to continuously maintaining the disagreements among learners during the learning process, regardless of whether the disagreements originate from differences in views or differences in model principles.

A disagreement between two learners is vitally important to the performance of the constructed ensemble model. Under the cotraining framework, if the disagreement between two learners is eliminated too early, it is difficult to improve the performance of the ensemble model. In addition, in this process, the category labels marked by the learners will always make mistakes. If the mistakes are not correctly handled, the impact of such label noise will continue to spread and amplify, eventually leading to a significant decline in the performance of the constructed ensemble model. Therefore, there are two core topics in the

disagreement-based, semisupervised learning method, namely, disagreements among learners and pseudolabel sample noise.

In terms of a disagreement between two learners, it is relatively simple to create a disagreement between two initial models. The feasible methods include view-difference-based methods, featuregrouping-based methods, data-grouping-based methods, and model-principle-based methods [2123]. However, it is extremely difficult to maintain this disagreement for a long time during the process of exchanging samples and updating models. In view of the abovementioned problems, the collaborative forest method deliberately adopts a variety of difference introduction mechanisms to slow the premature fitting of the learning process and thus to maintain the disagreement between two models [29-30]. Moreover, Malach and Shalev-Shwartz also proposed the decoupling method [37]. In this method, only those samples with inconsistent predictions are used to learn the model to strengthen the difference between the two models and to ensure the performance of the ensemble model. In terms of the noise of pseudolabel samples, the way to deal with the label noise introduced by "pseudolabel samples" is more complex. Angluin [38] was an early scholar who discussed how to effectively learn when the samples were mixed with noise. Angluin [38] proposed a compromise formula between the number of noise samples and the noise level under the condition that the label noise and the observation characteristics are independent. Wang and Zhou [39] pointed out that in the insufficient feature space, the learning process of the disagreement-based method would be restricted by "label noise" and "sampling bias". It is difficult to learn the PAC model only by providing pseudolabel samples to each other in the way of cotraining. However, if the base learner can provide additional confidence estimation results about the relatively accurate probability by providing prediction categories, the interference of labelling noise and sampling bias can be alleviated to a certain extent by “adaptively” adjusting the number of pseudolabel samples in different rounds, and then the performance of the ensemble model can be improved by effectively utilizing unlabelled samples. The essence of this method is to combat pseudolabel noise by fixing the confidence level and adjusting the number of pseudolabel samples. In addition, MentorNet [4041], Coteaching [42-43], and other methods proposed by other scholars update the model based on the fixed number of samples with the highest confidence level. These two methods reduce the noise level in the pseudolabel samples from two perspectives with the confidence estimation of the model. The two core topics on semisupervised learning discussed above triggered our assessment. Regarding the handling of disagreement, previous work focused on the creation and maintenance of disagreement, disregarding the information gathered in the process of exchanging pseudolabel samples among models to eliminate their disagreement. Why can one set of pseudolabel sample models be exchanged to eliminate disagreement, while another set of pseudolabel sample models cannot be exchanged to eliminate disagreement? What are the properties of the pseudolabel sample set that may eliminate disagreements among models? For the treatment of pseudolabel noise, a previous work used an empirical method; however, there is no systematic solution associated with the specific data status. If it is conservative and the number of samples exchanged in each round is small, the higher the sample confidence level is, the stronger the consistency of different models on these samples, leading to slow model improvement and fast disagreement elimination among models. However, if it is more radical and the number of samples exchanged in each round is large, the label noise caused by the mislabelling will be quite large due to the low sample confidence level, thus destroying the robustness of the model estimation and causing the performance of the subsequent ensemble model to be inferior to that of the initial base learner. Considering the differences in the characteristics of different datasets, it is clear that there is no unified sample exchange quantity. How can we correctly set the exchange sample quantity according to the dynamic characteristics in the learning process?

Based on the consideration of these two problems, this paper focuses on the process from disagreement to agreement between the two models in cotraining. From the perspective of the nature of the high-

confidence sample set exchanged in this process, a disagreement elimination algorithm framework based on the cotraining of feature argument and select (FANS) algorithm and L1 penalized logistics regression (PLR) algorithm is constructed. Then, the noise level of the accumulated high-confidence sample set in the exchange process is detected and used as a guide to correctly adjust the number of samples exchanged according to the characteristics of the dataset. The core contributions made by this research are as presented as follows: first, the static PAC estimation theory under noise conditions is deduced to the dynamic PAC estimation theory under multiple round cotraining scenarios, thereby providing a solid theoretical foundation for the disagreement-based cotraining method; second, by accurately limiting the proportion of the expanded number of pseudolabel samples to the number of current labelled samples (initial labelled samples and previous pseudolabel samples) in each round of update, the upper bound of the noise level of the high-confidence sample set accumulated in previous exchanges at the time of final convergence (multimodel agreement) is calculated; third, by forcing multiple models to achieve convergence, the "safety problem" of semisupervised learning can be solved, that is, the resulting ensemble model can resist "label noise" and "sampling bias" and give robust classification results; fourth, combining the noise of the biased statistical learning theory with the FANS and PLR algorithms, we give the upper bound of error for the results of the classification criteria, which is not involved in the existing related methods and has significant application value in some fields sensitive to discrimination accuracy, such as the risk control or investment fields. We also propose a method to control the upper bound of the error of the final classification results by adjusting the amplification coefficient. This adjustment method is often reflected as the penalty coefficient for the estimated parameters in existing algorithms, but existing works only qualitatively describe how the adjustment penalty coefficient can increase or decrease the reliability of the classification results, while our method provides a quantitative description of this degree of reliability of specific values.

The remainder of the content is arranged as follows: in Section 2, the dynamic noise PAC estimation in the cotraining scenario is introduced; in Section 3, a disagreement elimination algorithm architecture based on noise learning is proposed; in Section 4, empirical results and discussions are conducted, this method is tested on two different datasets, and relevant conclusions are verified; and in Section 5, conclusions and prospects are presented.

2. Dynamic noise PAC estimation in the cotraining scenario

The essence of machine learning is to define a hypothesis space via the model structure and then obtain a hypothesis that best matches the training data in this hypothesis space by utilizing the training data. Notably, the larger the amount of training data is, the closer the matched assumptions are to the real system law. If the assumptions that we have estimated are not much different from the real system law, it can be concluded that the current estimation result is the PAC identification [44]. If the training data are mixed with label noise, the sample size required to achieve PAC estimation will be greatly increased. PAC estimation theory under noise conditions describes the mathematical law of this issue. In the process of cotraining, the pseudolabel samples exchanged between two models inevitably contain label noise, and the relationship between noise and model estimation is alternately evolving and dynamically changing. Noise will affect model estimation, and the affected model estimation will cause more noise. This situation is referred to as "dynamic noise PAC estimation in the cotraining scenario". In this chapter, the three level-by-level PAC estimation will be discussed in turn, providing a theoretical basis for the subsequent construction of a disagreement elimination algorithm based on noise learning theory.

2.1 PAC Identification

In the field of machine learning, the hypothesis refers to a discriminant rule. If a sample meets this discriminant rule, it will be labelled class 1. If the sample does not meet this rule, it will be labelled class 0 [45]. The essence of machine learning is to select hypothesis H that best matches the data in hypothesis space $\mathcal{H}$ by using a dataset $\{x_i, y_i\}$. Data $\{x_i, y_i\}$ are derived from a potential hypothesis H^*. A good machine learning method should be able to minimize a certain distance between H and H^* based on data $\{x_i, y_i\}$.

For the mathematical description, several definitions should be made. First, assuming the data population $U = \{x_i\}$ and that different assumptions will cause different discrimination results for the same sample point x_i, the set constructed by all sample points in the data population that meet hypothesis H is referred to as the hypothesis response set, recorded as S_H:

$$S_H = \{x \in U \mid H(x) = 1\} \quad (1)$$

Second, based on the hypothetical response set, the distance between two assumptions is defined as the difference between their response sets:

$$d(H_1, H_2) = \sum_{x \in S_{H1} \triangle S_{H2}} Pr_U(x) \quad (2)$$

where $S_{H_1} \triangle S_{H_2}$ is the symmetric difference operation between two sets and $Pr_U(x)$ represents the probability of extracting sample x from data population U.

Third, if the probability of the distance between hypothesis H and the potential real hypothesis H^* greater than ε is less than δ,

$$Pr(d(H, H^*) > \varepsilon) < \delta \quad (3)$$

then assume that H is the PAC identification of the potential real hypothesis H^*. It can be considered with great confidence (the probability greater than $1 - \delta$) that the difference between hypothesis H and the potential real hypothesis H^* is very small (less than).

A certain number of samples are required to realize PAC estimation. If the hypothesis space is limited and there are at least N assumptions in total,

$$m \geq \frac{1}{\varepsilon}\left(\ln N + \ln \frac{1}{\delta}\right) \quad (4)$$

samples are required for the realization of PAC estimation [44].

2.2 PAC identification under noise conditions

In some special cases, for a group of samples $\{x_i, y_i\}$ labelled by the potential real hypothesis H^*, its label y_i will be disturbed by noise, and the probability of $\eta(< 0.5)$ will be reversed, that is, from class 1 to class 0 or from class 0 to class 1. Here, we record the data disturbed by label noise as $\{x_i, y_i^+\}$. The research results by Angluin [38] showed that if the number of samples m containing noise of level η reached the following formula,

$$2 \qquad 2N$$

$$m \geq \frac{1}{\varepsilon(1-2\eta)^2}\ln\left(\frac{2N}{\delta}\right) \quad (5)$$

then any hypothesis H' that best matches the sample data is the PAC identification of the potential real hypothesis H^*:

$$Pr\left(d\left(H, H^*\right) < \varepsilon\right) > 1-\delta \quad (6)$$

where N is the assumed quantity in the hypothesis space [38].

2.3 Dynamic noise PAC estimation theory in the cotraining scenario

In the process of disagreement-based learning, the pseudolabel samples exchanged by the two algorithms are always bound to have label errors, i.e., label noise, making the subsequent rounds of learning to be performed in a noisy learning environment. Starting from the PAC recognition theorem under the condition of noise, we know that if we want to combat the noise introduced by the exchange of pseudolabel samples, the number of pseudolabel samples exchanged this time should reach the required scale m. However, the calculation of this specific quantity m depends on the noise level η and hypothesis space scale N, and these two parameters can hardly be measured. If the problem is stated in the opposite way, i.e., if the training samples are expanded to β times the original samples in this round of update, what is the noise level that the sample expansion can resist? Therefore, this paper further analyses the PAC recognition process under noise conditions as follows:

Step 1: the formula of PAC recognition under noise conditions is transformed to obtain

$$\varepsilon^2 = \frac{2}{m(1-2\eta)^2}\ln(2N) \quad (7)$$

Step 2: As ε represents the distance between the poor estimation result and the potential real hypothesis in the case of a small probability, some pseudolabel samples are expanded so that the training set of the same learning model reaches β times the original set, and the distance becomes:

$$\varepsilon'^2 = \frac{2}{\beta m(1-2\eta)^2}\ln(2N) \quad (8)$$

where β is referred to as the expansion ratio.

Step 3: As the expected distance between the estimated hypothesis and the potential real hypothesis after sample expansion is smaller than the original distance, that is, $\varepsilon'^2 < \gamma\varepsilon^2, \gamma < 1$, γ is referred to as the convergence rate. Thus, it is required that:

$$\frac{2}{m(1-2\eta)^2}\ln\left(\frac{2N}{\delta}\right) > \frac{2}{\beta m(1-2\eta)^2}\ln\left(\frac{2N}{\gamma\delta}\right) \quad (9)$$

The common terms are deleted to perform the conversion accordingly:

$$\frac{1-\eta}{\sqrt{\beta}} > 0.5(1-2\eta) \quad (10)$$

Step 4: Assume that there is no noise in the initial training set, that is, $\eta = 0$. If the training set is expanded β times in some way and the distance between the estimated hypothesis and the potential real hypothesis is shortened γ times, the noise level of the training set used in this round is:

$$1 \quad 0.5 \qquad (11)$$

As it is assumed that there is no noise in the initial training set, the new noise sample points are completely introduced by the pseudolabel samples; thus, the noise level $\hat{\eta}$ of these pseudolabel sample points is:

$$(0.5 \quad) \Big/ \qquad (1 \quad) \qquad (12)$$

Step 5: It is explained in step 4 that if the convergence ratio is γ after sample update, the noise level $\hat{\eta}$ of the newly added pseudolabel samples will have an upper bound. However, the potential real hypothesis is unknown; thus, there is no way to test the distance between the estimated hypothesis and the potential real hypothesis. However, the sample expansion and hypothesis learning can be repeated from the two hypotheses with large differences in the hypothesis space, forming two trajectories in the hypothesis spaces. If these two trajectories meet the hypothesis space and the distance is reduced to almost 0, it can be held that these two trajectories effectively approach the potential real hypothesis. If the distance between the two hypotheses is less than 0.001 and this approximation is completed in n update rounds, as long as $(0.001)^{1/n} \leq \gamma$, there is ample reason to believe that the convergence rate of each update round does not exceed γ. If the two models reach an agreement during the disagreement cotraining, the pseudolabel samples exchanged during this period have a noise level upper bound of

$$(0.5 \quad) \Big/ \qquad (1 \quad) \qquad (13)$$

For example, the convergence ratio γ is required to be 0.7 and the expansion ratio is required to be 2. If convergence is achieved under this condition, the upper bound of the noise level of the exchanged pseudolabel group $\bar{\eta}$=15.5%.

Based on the abovementioned analysis, the following two conclusions are drawn: First, if the disagreement cotraining converges in a specific round according to a certain expansion ratio, the lower bound of the prediction accuracy of a part of the unlabelled samples (pseudolabel group) can be given without using the verification set. Second, in general, we have identified an effective method to test semisupervised learning. If the test is passed, it is proven to be a robust model closer to the real hypothesis. Such a method basically solves the problem of "safe" semisupervised learning proposed by Li and Zhou [46]: how to ensure that the ensemble model constructed by semisupervised learning is robust in the absence of a verification set rather than overfitting on a limited number of labelled samples.

3. Framework of the disagreement elimination algorithm based on noise learning theory

Next, an algorithm framework to realize dynamic noise PAC estimation theory in a cotraining scenario is constructed.

First, the selection of two base learners is introduced. In most cases, only a small number of labelled samples are provided for semisupervised learning; thus, any nonlinear model will usually fall into the

problem of overfitting on such small sample data, and consequently, the high-confidence sample points are not reliable. Therefore, two linear models are selected here as the base learners, namely, FANS [47] and L1 penalized logistics regression (PLR) [48].

Second, the FANS and PLR algorithms classify data from the perspectives of kernel space and feature space, respectively. The FANS algorithm classifies data from the perspective of kernel space as the FANS algorithm will first construct Gaussian kernel density estimation on a single feature and then convert the feature information to density ratio information. The subsequent classification model is set based on the density ratio information. This approach is similar to the principle of the SVM algorithm; thus, we refer to it as classification in the kernel space. The kernel space and feature space carry the neighbourhood information and pattern information, respectively, of the sample group, which are relatively independent. Therefore, strong disagreement exists between the two linear classifiers. In this paper, the default Silver rule is adopted as the calculation method of the bandwidth parameter, and the penalty parameter is determined by the leave-one-out cross-validation method.

The pseudocode of the disagreement elimination algorithm framework under noise learning theory is described as follows:

1. Initialization operation:
   a) Labelled sample set L and unlabelled sample set U are set
   b) Expansion ratio β and upper iteration limit N are set
   c) The training set of the FANS algorithm and PLR algorithm are initialized as an empty set:

$T_{F,0} = \{\}$; $T_{L,0} = \{\}$

Then, the labelled sample set L is merged with the training sets of the two algorithms.

$$T_{F,0} \leftarrow T_{F,0} \cup L$$

$$T_{L,0} \leftarrow T_{L,0} \cup L$$

   d) An empty set of high-confidence samples is initialized: $\phi = \{\}$
   e) The training status label $\Delta = 1$ and iteration count $k = 0$ are initialized

2. Main process: The following operations are performed in the iteration until the iteration is exited or the upper iteration limit is reached, as shown in Fig. 1:

Step 1: The FANS algorithm is trained based on the FANS algorithm training set $T_{F,k}$.

Step 2:

1. Probabilistic prediction of unlabelled sample set U is performed based on the trained FANS model.
2. The samples are sorted according to the confidence level.

3. The number of samples $N(T_{L,k})$ in the PLR algorithm training set at this time is calculated, the $(\beta - 1) * N(T_{L,k})$ unlabelled samples with the highest confidence ranking are removed, the category labels given by the FANS algorithm are attached to form the pseudolabel sample set $\rho_{F,k}$, and this group of samples is merged with the high-confidence sample set and the PLR algorithm training set, namely:

$$T_{L,k+1} \leftarrow T_{L,k} \cup \rho_{F,k}$$

$\phi \leftarrow \phi \cup \rho_{F,k}$

4. The number of samples $N(T_{L,k+1})$ in the training set of the PLR algorithm after expansion are determined. If $\frac{N(T_{L,k+1})}{N(T_{L,k})} < 1.001$, the training status label $\Delta = 0$, and the iteration will stop normally.

Step 3: The PLR algorithm is trained based on the PLR algorithm training set $T_{L,k}$.

Step 4:

1. Probabilistic prediction of the unlabelled sample set based on the trained PLR model is performed.
2. The samples are sorted according to the confidence level.
3. The number of samples $N(T_{F,k})$ in the FANS algorithm training set at this time are calculated, the $(\beta - 1) * N(T_{F,k})$ unlabelled samples with the highest confidence ranking are removed, the category labels given by the PLR algorithm are attached to form the pseudolabel sample set $\rho_{F,k}$, and this group of samples is merged with the high-confidence sample set and FANS algorithm training set, namely:

$$T_{F,k+1} \leftarrow T_{F,k} \cup \rho_{L,k}$$

$\phi \leftarrow \phi \cup \rho_{L,k}$

4. The number of samples $N(T_{F,k+1})$ in the training set of the FANS algorithm after expansion are determined. If $\frac{N(T_{F,k+1})}{N(T_{F,k})} < 1.001$, the training status label $\Delta = 0$, and the iteration will stop normally.
5. The sample quantity $N(\phi)$ in the high-confidence sample set ϕ is checked. If $\frac{N(\phi)}{N(U)} > 0.9,\ \Delta = 1$, and the iteration will stop abnormally.
4. $k \leftarrow k + 1$

3. If the training status label $\Delta = 0$, agreement is reached between the two models under the specified conditions; then, the following conclusion is obtained:

a) The upper bound of the noise level of the high-confidence sample set is:

$$\bar{\square}\square\ (0.5\square\ \frac{1}{\sqrt{\square}})\Big/\ (1\square\ \frac{1}{\square})$$
$$4\square\square \qquad \square$$

where the assumed convergence ratio is $\gamma = \sqrt[k]{0.001}$.

b) The current ensemble model composed of two models is a robust estimation model. If the training status label is $\Delta = 1$, no agreement is reached between the two models under the specified conditions, and the parameter setting needs to be adjusted for retraining. Generally, the corresponding operation is to reduce the expansion ratio β.

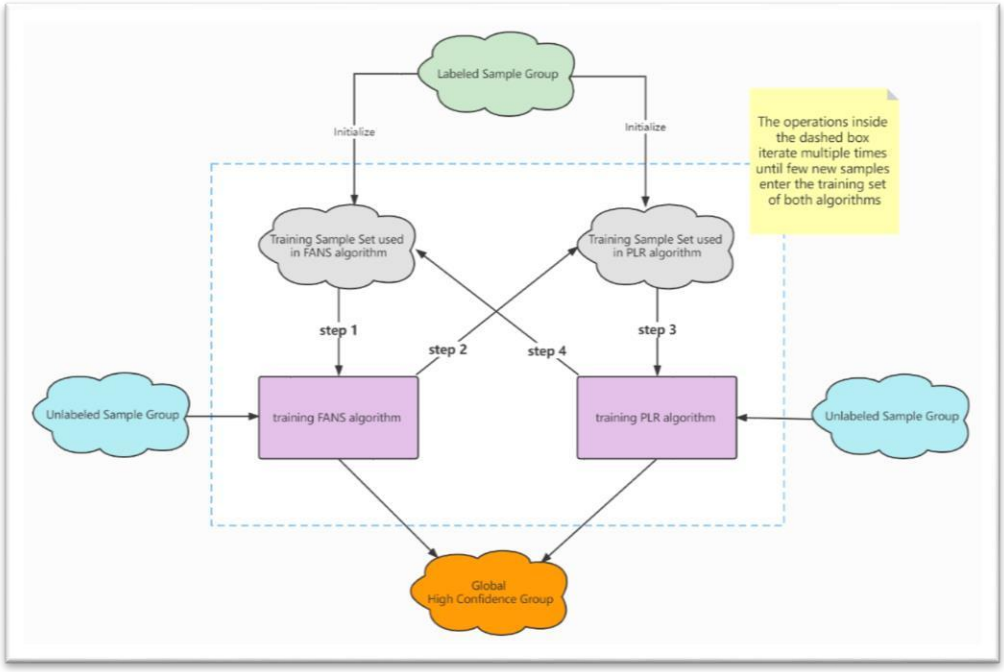


Fig. 1 Logic block diagram of the proposed algorithm framework

There are five points to explain in the algorithm framework:

First, the expansion ratio β should not exceed 2; otherwise, the number of new pseudolabel samples will always be greater than the number of existing training samples, inevitably leading to the failure to trigger the normal stop condition of iteration.

Second, if the expansion ratio β is set to be small and the algorithm stops iterating normally early, the size of the high-confidence sample set will be small, even though its error level is relatively low, which is not conducive to improving the performance of the ensemble model to the greatest extent. Therefore, the expansion ratio β should be further increased to make the algorithm converge to a larger high-confidence sample set and to simultaneously improve the performance of the ensemble model.

Third, if the two models fail to reach an agreement and exit the iteration, this indicates the noise level of the samples exchanged is higher than the upper bound of the theoretical calculation result. At this time, reducing the expansion ratio β and retraining may usually promote the two models to reach an agreement For the following reasons: first, the noise level in the sample set is proportional to its confidence; thus, reducing the number of samples exchanged in each round will significantly reduce the noise level; second, the consistency of different models is stronger in the higher confidence region. Therefore, the test should start with a larger expansion ratio β (generally 1.99) and then gradually reduce to obtain the maximum expansion ratio β, making the two models reach an agreement.

Fourth, the criteria for the agreement between the two models are presented as follows: whether the number of samples added to the training set exceeds a very small threshold after one base learner updates the training set for the other base learner. If this threshold is not exceeded, the important reference sample points considered by learner A are already the reference sample points for building learner B, and learners A and B have almost reached an agreement. As there are no new training sample points and learner B will not change, the reference sample points learner B gives to learner A will also be the same as those in the previous round, showing no need for iterative processing. In addition, if the algorithm still fails to reach an agreement between the two models when the unlabelled sample set U is almost exhausted or if the number of iterations reaches the upper bound, the two hypothesis trajectories will not intersect. It is indicated that the noise level of the highconfidence sample set generated in the exchange process exceeds the upper bound

specified in the formula, and the ensemble model may also have the problems of overfitting the labelled sample group L and random label noise.

Fifth, the FANS algorithm is more inclined to robust estimation in principle. Therefore, before the first sample exchange, the number of samples is extremely small. The FANS algorithm should be trained and then should provide pseudolabel samples for the PLR algorithm. In the comparative test, it is determined that the training from the FANS algorithm is significantly better than that from the PLR algorithm.

4. Empirical results and discussion

To verify whether the disagreement elimination algorithm based on noise learning theory can effectively achieve the two preset goals, i.e., to provide a high-confidence sample group with a distinct upper bound of noise level and to ensure a robust ensemble model that can combat sampling bias, we have performed an empirical analysis of the algorithm on two different training datasets. The two datasets have a large gap in all aspects; thus, the effectiveness of the method proposed in this paper will be comprehensively reflected.

4.1 Dataset

Repurchase user classification of the internet credit platform: The machine learning model should determine whether the user will apply for a loan on the platform again based on a group of user personality characteristics, which is a typical binary classification. There are 285 features in this dataset, including both continuous features and category features. There are 200 labelled samples and 25,000 unlabelled samples (this part actually has labels to facilitate the verification of model performance). When only based on 200 labelled samples, the highest accuracy and the area under the curve (AUC) value of various learning algorithms for this classification are 63% and 0.72, respectively. Later, we will refer to this test as test A.

Spam identification: This step requires a machine learning model to determine whether the email is spam based on the lexical features of the email content, which is a binary classification. The dataset has 57 features, including both continuous features and category features. There are 20 labelled samples and 3,400 unlabelled samples (this part actually has labels to facilitate the verification of the model performance). When only based on 20 labelled samples, the highest accuracy and AUC value of various learning algorithms for this classification are 85% and 0.93, respectively. Later, we will refer to this test as test B[1].

4.2 Analysis of experimental results

In semisupervised learning, the small number of labelled samples will inevitably lead to the problem of "sampling bias". This problem is one of the main factors that perplex the generalization ability of the machine learning model in the semisupervised learning scenario. In the empirical test, several different labelled sample sets are randomly selected for the same classification problem to test whether the disagreement elimination algorithm based on noise learning theory can overcome the sampling bias and achieve the two preset goals. The corresponding test results are recorded in Tables 1 and 2.

Table 1 Test A - Summary statistics of multiple group tests

Test ID*	Expansion ratio β	Rounds of convergence n (convergence rate	Upper bound of theoretical noise $\bar{\eta}$	Size of high-confidence sample set	Measured noise level	AUC value of base learner	AUC value of ensemble learner

[1] The datasets and code for this paper are included in https://github.com/xiaojianyang820/DiEliRecog

A-500	21(0.720)	0.160	5105	0.067	0.679/0.713	0.716	1.95
A-501**	1.50	-(-)	-	-	-	0.663/0.661	-
A-502	1.99	17(0.666)	0.132	4920	0.086	0.698/0.715	0.715
A-503	1.95	25(0.759)	0.182	4073	0.173	0.676/0.664	0.677
A-503+	1.90	15(0.631)	0.091	3042	0.085	0.676/0.664	0.684
A-503++	1.85	15(0.631)	0.081	2916	0.081	0.676/0.664	0.677
A-504	1.97	21(0.720)	0.163	4313	0.083	0.686/0.708	0.705

Note: *The test ID consists of two parts. The capital letters in the front represent the test dataset, and the numbers in the back represent the random seeds used in this test. The random seeds will affect the random sampling results.
**In the A-501 test, the quality of the labelled samples is too poor, leading to the failure of convergence, although the expansion ratio is already set to the lower bound.
+/++A specific experiment was added to further analyse the influence of the expansion ratio on the whole algorithm.

Table 2 Test B - Summary statistics of multiple group tests

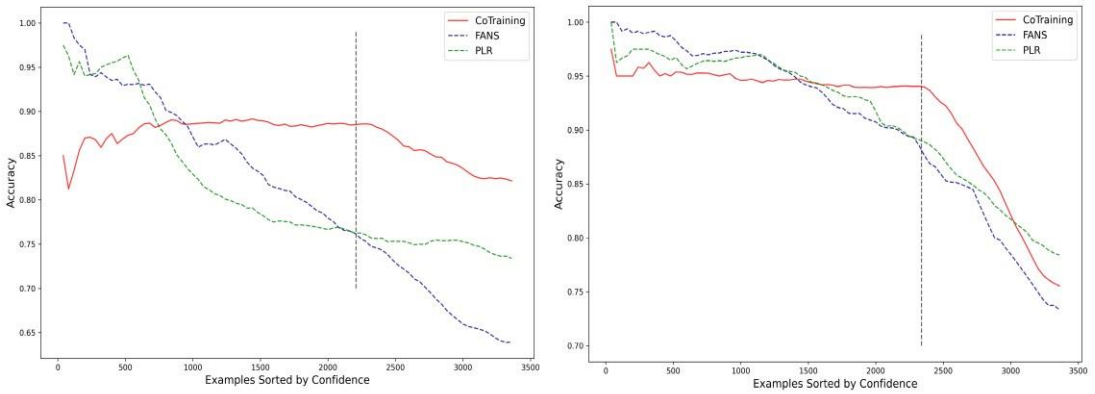
Test ID*	Expansion ratio β	Rounds of convergence$n$ (convergence rate γ)	Upper bound of theoretical noise $\bar{\eta}$	Size of high-confidence sample set	Measured noise level	AUC value of base learner	AUC value of ensemble learner
B-500	26(0.767)	0.191	1342	0.112	0.878/0.889	0.887	1.99
B-501	1.99	38(0.834)	0.225	2339	0.058	0.929/0.916	0.937
B-502	1.95	26(0.767)	0.187	2121	0.148	0.860/0.800	0.842
B-503	1.90	40(0.841)	0.228	2511	0.100	0.936/0.881	0.910
B-504	1.98	36(0.825)	0.220	2208	0.114	0.799/0.770	0.897
B-504+	1.92	20(0.708)	0.148	1012	0.060	0.799/0.770	0.914
B-504++	1.86	14(0.681)	0.067	742	0.056	0.799/0.770	0.901

Note: *The test ID consists of two parts. The capital letters in the front represent the test dataset, and the numbers in the back represent the random seeds used in this test. The random seeds will affect the random sampling results.
+/++A specific experiment was added to further analyse the influence of the expansion ratio on the whole algorithm.

As shown by the two groups of experiments, the proposed algorithm has basically achieved two preset goals. First, the upper bound of noise given by the algorithm is consistent with the measured noise level. Although the measured noise levels of the A-503^{++} tests approach the upper bound, the measured noise level is calculated based on one sample, and there is an estimation deviation. Therefore, this situation is acceptable. However, this closeness also shows that the upper bound estimation is quite compact. Second, this algorithm effectively ensures the robustness of the ensemble model. When the quality of labelled samples is poor, the recognition performance of the ensemble model remains at a high level.

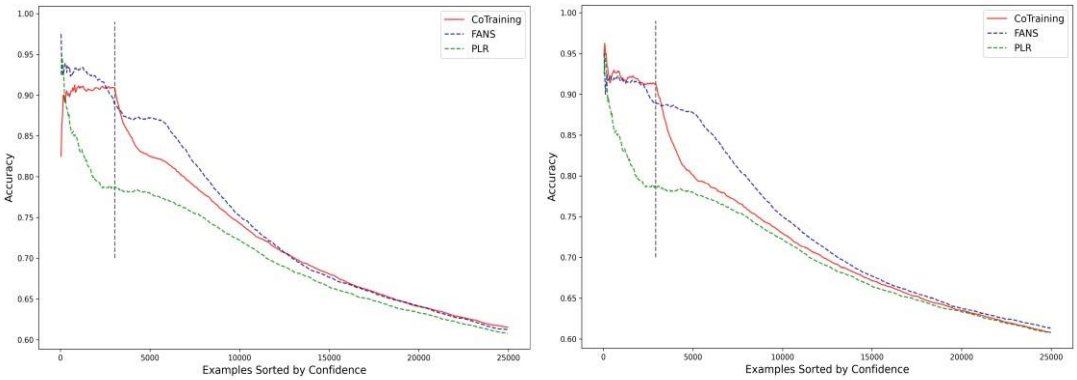
Furthermore, three different tests around the random seed B-504 are designed. The randomly labelled sample set corresponding to this random seed may probably have a large sampling bias; thus, the generalization performance of the base learner based on these data is poor. In the first test B-504, the largest expansion ratio β that makes the two models reach agreement is selected, and in the subsequent two tests B-504^{+} and B-504^{++}, the expansion ratio β is reduced. From the theoretical analysis of the algorithm framework proposed in this paper, reducing the expansion ratio β will reduce the noise level of the pseudolabel samples exchanged and improve the consistency of multiple models. Therefore, the model proposed in this paper reflects the decrease in the measured noise level, convergence rounds and the size of

the high-confidence sample set in the measured data. The comparison of the results of these three tests basically verifies this conclusion. Similarly, this phenomenon was also observed in the control group composed of three tests: A-503, A-503^{+} and A-503^{++}.



Note: After arranging the samples according to the classification confidence from large to small, the classification accuracy of the first n sample points is calculated, wherein n is the scale on the x-coordinate. The position of the grey vertical line is the size of the high-confidence sample set.

Fig. 2. Comparison curve of classification accuracy of B-501 (left) and B-504 (right)



Note: After arranging the samples according to the classification confidence from large to small, the classification accuracy of the first n sample points is calculated, wherein n is the scale on the x-coordinate. The position of the grey vertical line is the size of the high-confidence sample set.

Fig. 3 Comparison curve of classification accuracy of A-503^{+} (left) and A-503^{++} (right)

In test A-501, the expansion ratio β is reduced to the greatest extent, and there is still no way to reach an agreement between the two models. After inspection, it is determined that the noise level of the two base learners fitted by this randomly labelled sample set reaches 0.221 at the 100 sample points with the highest confidence. This finding indicates that the sample set with an unreasonably large sampling bias cannot support disagreement-based semisupervised learning, and the first sample exchange will destroy the model estimation due to the excessive noise of pseudolabel samples, causing a vicious circle.

In addition to the overall statistical indicator comparison, the comparison curve between the ensemble model and the base learner in the classification accuracy is drawn, as shown in Fig. 2. The two figures compare the classification accuracy comparison curves of B-501 (best sample representativeness) and B-504 (worst sample representativeness). The confidence of the sample points in the high-confidence sample set is set to the maximum; thus, the corresponding accuracy at the grey vertical line is the overall accuracy of the high-confidence sample set. The figure shows that the accuracy of the ensemble model is higher than that of the base learner at this specific point. Viewed from the overall data, the accuracy of the ensemble model is also slightly better than that of each base learner. Figure 3 also shows the classification accuracy in test A, which is similar to that in test B.

4.3 Effect of the expansion ratio

To introduce in more detail the role of the expansion ratio β in the algorithm and the dynamic characteristics of the change curve between the observable size and the unobservable accuracy of the high-confidence sample set, we set the expansion ratio β as 1.99, 1.95, 1.90 and 1.85 based on a group of randomly labelled sample sets and draw the change curve between the size and the accuracy of the high-confidence sample set in multiple rounds, as shown in Figs. 4 and 5.

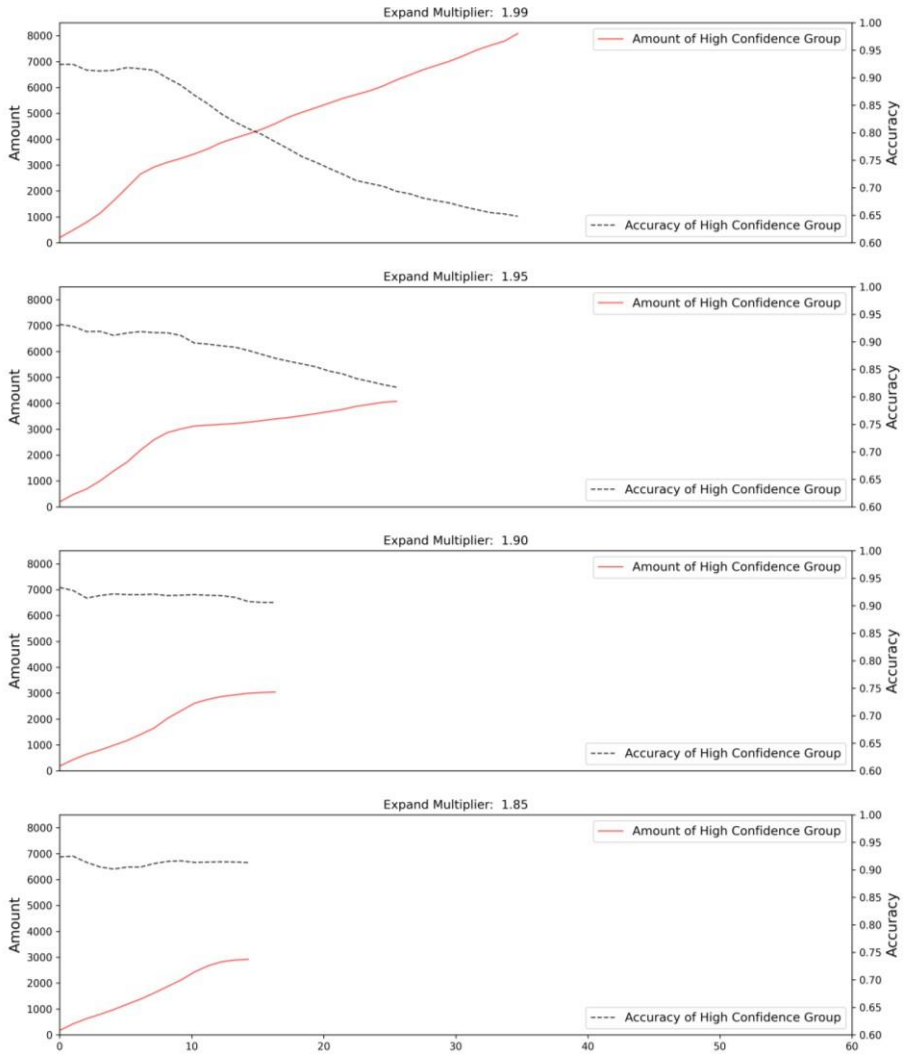
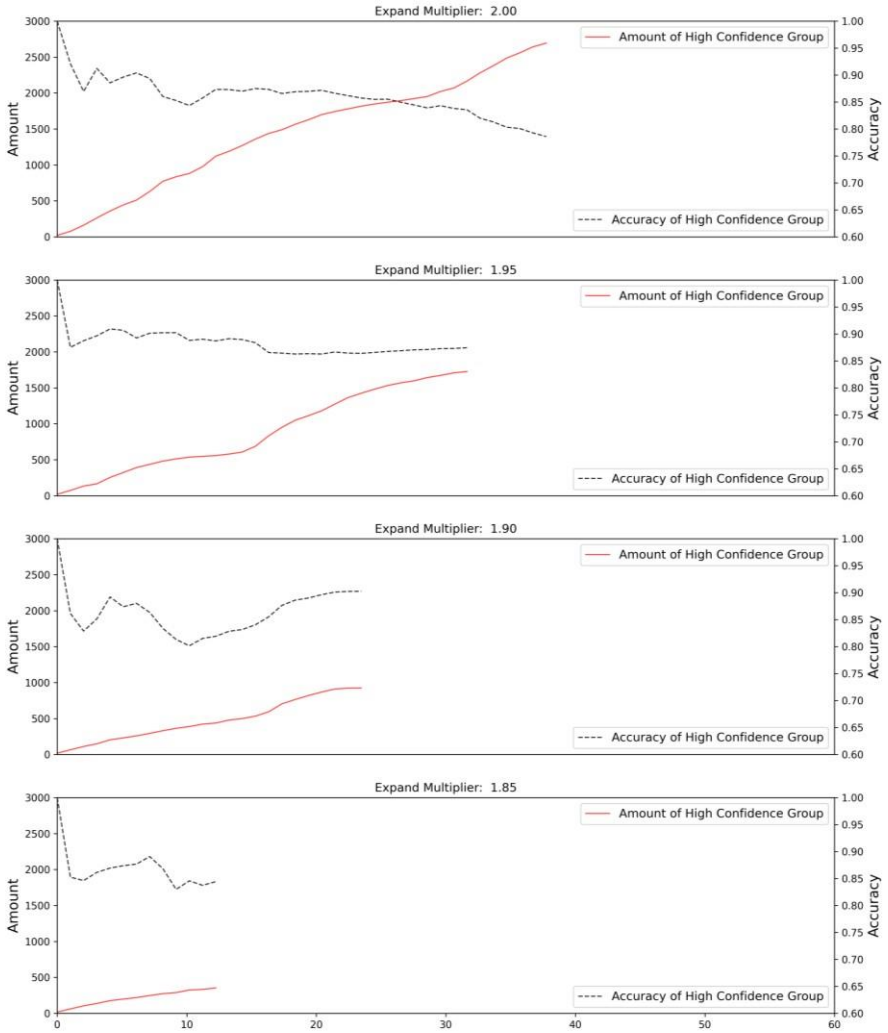


Fig. 4 Dynamic curve of the number and accuracy of high-confidence sample sets (test A)



It is observed that as the expansion ratio β decreases, the algorithm converges faster, the accuracy of the high-confidence sample set increases, and the size of the high-confidence sample set decreases. Simultaneously, the scale change curve and accuracy change curve of the highconfidence sample set are almost symmetrical. In both tests, $\beta = 1.95$ is the critical value of convergence and divergence of the algorithm. When $\beta > 1.95$, the size of the high-confidence sample set will continue to linearly increase without any tendency to converge; thus, the corresponding accuracy will also linearly decrease. When $\beta < 1.95$, the growth curve of the highconfidence sample set size will slow early and become a horizontal line, and the accuracy will also stay at a high level. The accuracy curve cannot be observed in the practical application of the algorithm; thus, the expansion ratio β can be adjusted by observing the curve of the highconfidence sample set size.

5. Conclusion

In this study, the research perspective of cotraining is changed from maintaining the disagreements among models to eliminating the disagreements among models for the first time. The role of noise learning theory in this process is discussed in detail, and the dynamic noise PAC estimation theory under the cotraining scenario is given. Based on this theory, a new disagreementbased, semisupervised classification algorithm is proposed: a disagreement elimination algorithm under noise conditions. Under the framework of the

algorithm, it is confirmed that if the two models with great disagreement eliminate their disagreement and reach an agreement by exchanging highconfidence samples, the set of high-confidence samples exchanged will have a certain upper bound of noise and that the algorithm can effectively combat the problem of sampling bias of small set samples. Otherwise, the two models can reach an agreement at another level by reducing the expansion ratio β and adjusting the high-confidence sample exchange process. Different from the premise of other disagreement-based semisupervised learning methods, the core skill of this algorithm framework is to eliminate the disagreements among multiple models in a planned way rather than to maintain these disagreements, thereby determining the quality level of the samples exchanged. The algorithm gives a high-confidence sample set with an upper bound of the noise level and a robust ensemble model.

This algorithm is an effective semisupervised learning framework that provides a reliable set of high-confidence samples. In addition, this algorithm is applicable to semisupervised problems in various fields, in which the number of labelled samples is small and the number of unlabelled samples is large. When a large number of unlabelled samples are employed to assist labelled samples in constructing classification models and in evaluating the reliability of classification results, the method proposed in this paper can be employed.

In future studies, these pseudolabel samples can be put into other models with higher complexity to further study the data. In the algorithm framework proposed in this paper, the base learner is a linear algorithm (FANS model and PLR model), and disagreement is created through different algorithm principles. In future studies, attempts may be made to replace the base learner in this paper with a neural network and then extend it from a double model to multiple models. In addition, if subsequent researchers discovered that the data distribution of the dataset was a Gaussian distribution and subject to other distributions, then the FANS algorithm of the Gaussian kernel can also be replaced with other kernel functions, such as the polynomial kernel function or triangle kernel function, to obtain meaningful results.

Acknowledgement

This research was supported by the Natural Science Foundation of China (Grant No. 70901006) and the Beijing Social Science Foundation of China (Grant No. 19GLC053)

Conflict of Interest Statement

We declare that there are no conflicts of interest regarding the publication of this paper.

References

[1]. Castillo, A., Benitez, J., Llorens, J. and Luo, X.R., 2021. Social media-driven customer engagement and movie performance: Theory and empirical evidence. *Decision Support Systems*, *145*, p.113516.

[2]. Rahim, M.A., Mushafiq, M., Khan, S. and Arain, Z.A., 2021. RFM-based repurchase behavior for customer classification and segmentation. *Journal of Retailing and Consumer Services*, *61*, p.102566.

[3]. Wang, D.N., Li, L. and Zhao, D., 2022. Corporate finance risk prediction based on LightGBM. *Information Sciences*, *602*, pp.259-268.

[4]. Pintas, J.T., Fernandes, L.A. and Garcia, A.C.B., 2021. Feature selection methods for text classification: a systematic literature review. *Artificial Intelligence Review*, *54*(8), pp.6149-6200.

[5]. Liang, Y., Li, H., Guo, B., Yu, Z., Zheng, X., Samtani, S. and Zeng, D.D., 2021. Fusion of heterogeneous attention mechanisms in multi-view convolutional neural network for text classification. *Information Sciences*, *548*, pp.295-312.

[6]. Mishra, S., Zhang, Y., Chen, D.Z. and Hu, X.S., 2022. Data-Driven Deep Supervision for Medical Image Segmentation. *IEEE Transactions on Medical Imaging*. DOI: 10.1109/TMI.2022.3143371

[7]. Wu, Z., Jing, L., Wu, B. and Jin, L., 2022. A PCA-AdaBoost model for E-commerce customer churn prediction. *Annals of Operations Research*, pp.1-18.

[8]. Bhattacharya, S., Nurmi, P., Hammerla, N., & Plötz, T. (2014). Using unlabeled data in a sparse-coding framework for human activity recognition. *Pervasive and Mobile Computing*, 15, 242-262.

[9]. Zhou, Z. H., & Li, M. 2010. Semi-supervised learning by disagreement. *Knowledge and Information Systems*, 24(3), 415-439.

[10]. Zhu, X. and Goldberg, A.B., 2009. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, *3*(1), pp.1-130.

[11]. Van Engelen, J.E. and Hoos, H.H., 2020. A survey on semi-supervised learning. *Machine Learning*, *109*(2), pp.373-440.

[12]. Jain, A. K. 2010. Data clustering: 50 years beyond K-means. *Pattern recognition letters*, *31*(8), 651-666.

[13]. Castillo-Navarro, J., Le Saux, B., Boulch, A. and Lefèvre, S., 2021. Energy-based models in earth observation: From generation to semi-supervised learning. *IEEE Transactions on Geoscience and Remote Sensing*.

[14]. Bai, R., Huang, R., Qin, Y., Chen, Y. and Lin, C., 2022. HVAE: A Deep Generative Model via Hierarchical Variational Auto-Encoder for Multi-view Document Modeling. Information Sciences. https://doi.org/10.1016/j.ins.2022.10.052

[15]. Song, Z., Yang, X., Xu, Z. and King, I., 2022. Graph-based semi-supervised learning: A comprehensive review. *IEEE Transactions on Neural Networks and Learning Systems*.

[16]. Wang, Z., Zhang, L., Wang, R., Nie, F. and Li, X., 2022. Semi-supervised Learning via Bipartite Graph Construction with Adaptive Neighbors. *IEEE Transactions on Knowledge and Data Engineering*.

[17]. Xu, L., Chen, C.P. and Han, R., 2022. Graph-based sparse bayesian broad learning system for semi-supervised learning. *Information Sciences*, *597*, pp.193-210.

[18]. Zheng, X., Zhang, L. and Xu, Z., 2021. L1-norm Laplacian support vector machine for data reduction in semisupervised learning. *Neural Computing and Applications*, pp.1-18.

[19]. Xue, Y. and Zhang, L., 2021. Laplacian pair-weight vector projection for semi-supervised learning. *Information Sciences*, *573*, pp.1-19.

[20]. Sun, Y., Ding, S., Guo, L. and Zhang, Z., 2022. Hypergraph Regularized Semi-supervised Support Vector Machine. *Information Sciences*.

[21]. Meng, Y., Li, W. and Kwok, L.F., 2014, June. Enhancing email classification using data reduction and disagreement-based semi-supervised learning. In 2014 IEEE International Conference on Communications (ICC) (pp. 622-627). IEEE.

[22]. Li, W., Meng, W. and Au, M.H., 2020. Enhancing collaborative intrusion detection via disagreement-based semi-supervised learning in IoT environments. *Journal of Network and Computer Applications*, *161*, p.102631.

[23]. He, H., Han, D. and Dezert, J., 2020. Disagreement based semi-supervised learning approaches with belief functions. *Knowledge-Based Systems*, *193*, p.105426.

[24]. Blum, A. and Mitchell, T., 1998, July. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory* (pp. 92-100).

[25]. Sindhwani, V., Niyogi, P. and Belkin, M., 2005, August. A co-regularization approach to semi-supervised learning with multiple views. In *Proceedings of ICML workshop on learning with multiple views* (Vol. 2005, pp. 74-79). Citeseer.

[26]. Nigam, K. and Ghani, R., 2000, November. Analyzing the effectiveness and applicability of co-training. In *Proceedings of the ninth international conference on Information and knowledge management* (pp. 86-93).

[27]. Zhou, Z.H. and Li, M., 2005. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on knowledge and Data Engineering*, *17*(11), pp.1529-1541.

[28]. Li, M. and Zhou, Z.H., 2007. Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *37*(6), pp.1088-1098.

[29]. Zhou, Q. and Duan, L., 2021. Semi-supervised recommendation attack detection based on CoForest. *Computers & Security*, *109*, p.102390.

[30]. Wang, Y. and Li, T., 2018. Improving semi-supervised co-forest algorithm in evolving data streams. *Applied Intelligence*, *48*(10), pp.3248-3262.

[31]. Kim, D., Seo, D., Cho, S., & Kang, P. 2019. Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec. *Information Sciences*, *477*, 15-29.

[32]. Zhou, Y., Wang, Y., Tang, P., Bai, S., Shen, W., Fishman, E., & Yuille, A. 2019. Semi-supervised 3D abdominal multi-organ segmentation via deep multi-planar co-training. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 121-140). IEEE.

[33]. Gao, C., Zhou, J., Miao, D., Wen, J., & Yue, X. 2021. Three-way decision with co-training for partially labeled data. *Information Sciences*, *544*, 500-518.

[34]. Peng, J., Estrada, G., Pedersoli, M., & Desrosiers, C. 2020. Deep co-training for semi-supervised image segmentation. *Pattern Recognition*, *107*, 107269.

[35]. Gong, M., Zhou, H., Qin, A.K., Liu, W. and Zhao, Z., 2022. Self-Paced Co-Training of Graph Neural Networks for Semi-Supervised Node Classification. *IEEE Transactions on Neural Networks and Learning Systems*.

[36]. Dong, Y., Jiang, L. and Li, C., 2022. Improving data and model quality in crowdsourcing using co-trainingbased noise correction. *Information Sciences*, *583*, pp.174-188.
[37]. Malach, E. and Shalev-Shwartz, S., 2017. Decoupling" when to update" from" how to update". *Advances in Neural Information Processing Systems*, *30*.
[38]. Angluin, D. and Laird, P., 1988. Learning from noisy examples. *Machine Learning*, *2*(4), pp.343-370.
[39]. Wang, W. and Zhou, Z.H., 2013, October. Co-training with insufficient views. In *Asian conference on machine learning* (pp. 467-482). PMLR.
[40]. Jiang, L., Zhou, Z., Leung, T., Li, L.J. and Fei-Fei, L., 2018, July. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning* (pp. 2304-2313). PMLR.
[41]. Cordeiro, F.R. and Carneiro, G., 2020, November. A Survey on Deep Learning with Noisy Labels: How to train your model when you cannot trust on the annotations?. In *2020 33rd SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)* (pp. 9-16). IEEE.
[42]. Zhang, Q., Lee, F., Wang, Y.G., Ding, D., Yang, S., Lin, C. and Chen, Q., 2021. CJC-net: A cyclical training method with joint loss and co-teaching strategy net for deep learning under noisy labels. *Information Sciences*, *579*, pp.186-198.
[43]. Wang, B., Shen, H., Lu, G. and Liu, Y., 2022. Graph Learning with Co-Teaching for EEG-Based Motor Imagery Recognition. *IEEE Transactions on Cognitive and Developmental Systems*.
[44]. Haussler, D., 1990. *Probably approximately correct learning*. Santa Cruz, CA, USA: University of California, Santa Cruz, Computer Research Laboratory.
[45]. Wang, H., Z. Lei, X. Zhang, B. Zhou, and J. Peng. 2016. "Machine learning basics." *Deep learning*, 98-164.
[46]. Li, M. and Zhou, Z.H., 2005, May. SETRED: Self-training with editing. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 611-621). Springer, Berlin, Heidelberg.
[47]. Fan, J., Feng, Y., Jiang, J. and Tong, X., 2016. Feature augmentation via nonparametrics and selection (FANS) in high-dimensional classification. *Journal of the American Statistical Association*, *111*(513), pp.275-287.
[48]. Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), pp.267-288.

Zaoli Yang: Conceptualization, Funding acquisition, Supervision, Software, Resources, Data curation, Supervision, Writing - review & editing

Weijian Zhang: Data curation, Software, Resources, Writing - review & editing

Chunjia Han: Formal analysis, Investigation, Writing - review & editing

Yuchen Li: Conceptualization, Data curation, Resources, Formal analysis, Investigation, Supervision.

Mu Yang: Formal analysis, Resources, Investigation, Writing - review & editing.

Petros Ieromonachou: Formal analysis, Investigation, Software,Writing - review & editing.