Cognitive Computation (2019) 11:599–612
https://doi.org/10.1007/s12559-019-09664-w

Research Progress on Semi-Supervised Clustering

Yue Qin [1,2] · Shifei Ding [1,2] · Lijuan Wang [1,2,3] · Yanru Wang [1,2]

Received: 8 November 2018 / Accepted: 3 July 2019 / Published online: 17 July 2019

Abstract
Semi-supervised clustering is a new learning method which combines semi-supervised learning (SSL) and cluster analysis. It is widely valued and applied to machine learning. Traditional unsupervised clustering algorithm based on data partition does not need any property; however, there are a small amount of independent class labels or pair constraint information data samples in practice; in order to obtain better clustering results, scholars have proposed a semi-supervised clustering. Compared with traditional clustering methods, it can effectively improve clustering performance through a small number of supervised information, and it has been used widely in machine learning. Firstly, this paper introduces the research status and classification of semi-supervised learning and compares the four classification methods as follows: decentralized model, support vector machine, graph, and collaborative training. Secondly, the semi-supervised clustering is described in detail, the current status of semi-supervised clustering is analyzed, and the Cop-kmeans algorithm, Lcop-kmeans algorithm, Seeded-kmeans algorithm, SC-kmeans algorithm, and other algorithms are introduced. The introduction of several semi-supervised clustering methods in this paper can show the advantages of semi-supervised clustering over traditional clustering, and the related literature in recent years is summarized. This paper summarized the latest development of semi-supervised learning and semi-supervised clustering and discussed the application of semi-supervised clustering and the future research direction.

Keywords Semi-supervised learning · Clustering · Semi-supervised clustering · Pairwise constraints · Labeled

Introduction

With the advent of the big data era, the amount of data is increasing and becoming more and more complex. In order to extract the internal relations between the massive data and relevant useful knowledge information, data mining comes into being. Traditional cluster analysis is an unsupervised and exploratory analysis. In the process of data classification, it does not rely on any background knowledge and corresponding assumptions. It is naturally divided according to the similarity, so the data in the same class should be as similar as possible, and the data belong to different classes should be as different as possible.

✉ Shifei Ding
dingsf@cumt.edu.cn

[1] School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China

[2] Mine Digitization Engineering Research Center of Ministry of Education of the People's Republic of China, Xuzhou 221116, China

[3] School of Information and Electrical Engineering, Xu Zhou College of Industrial Technology, Xuzhou 221400, China

However, not all data are independent in real life. They are not only interrelated, but also have common attributes and own characteristics. The labeled data and unlabeled data often coexist with each other. If only use traditional unsupervised learning clustering analysis, the labeled data will be wasted. So in order to use the background knowledge of data objects reasonably, semi-supervised learning is proposed.

The two basic learning methods of machine learning are supervised learning and unsupervised learning. Supervised learning, also known as teacher learning or directed learning, the output value of the sample, is predicted by the corresponding relationship between input and output. The second method of learning is unsupervised learning, also known as undirected learning, which is learning with independent labels. It is a kind of potential structure that can be learned from massive data samples without labels and human intervention. As a learning method that combines supervised learning and unsupervised learning, semi-supervised learning can utilize both labeled and unlabeled data and can be divided into semi-supervised classification and semi-supervised clustering.

Different from traditional clustering algorithms such as k-means algorithm [1] and EM algorithm [2], semi-supervised clustering is a new research algorithm, which combines clustering with semi-supervised learning, and the clustering

performance can be improved through a small amount of labeled data and prior knowledge. In general, semi-supervised clustering methods can be divided into the following three categories: (1) semi-supervised clustering algorithm based on distance; (2) semi-supervised clustering algorithm based on constraint; (3) semi-supervised clustering algorithm based on constraint and distance.

Semi-supervised clustering has a good application prospect in the field of data mining. In recent years, there are more and more research achievements on semi-supervised clustering. Compared with clustering and semi-supervised learning, semi-supervised clustering has fewer review papers and published for a long time, as a result, the achievements in recent years have not been discussed. For example, Li Kunlun [3] has elaborated the progress of semi-supervised clustering, mainly discussed the classification of semi-supervised clustering, and proposed the semi-supervised fuzzy c-means clustering algorithm based on constraints. However, the research status of semi-supervised clustering and the description of relevant type algorithms are not particularly comprehensive. Xiong [4] has summarized semi-supervised learning and analyzed the advantages and limitations of several typical semi-supervised clustering algorithms; however, they only mentioned a few typical algorithms and did not describe algorithms in detail.

This paper summarizes the semi-supervised clustering, aiming to introduce the new algorithm to readers in detail, so that relevant professionals can understand the relevant knowledge, typical algorithms and research status of semi-supervised clustering clearly. The structure of this paper is as follows: in part 2, we introduce the concept and classification of semi-supervised learning; in part 3, we expound the concept and classification of semi-supervised clustering and introduce the typical algorithms of semi-supervised clustering in detail. The last part summarizes the whole paper and discusses the future research direction.

Semi-Supervised Learning

As early as the 1980s and 1990s, semi-supervised learning (SSL) became a research hotspot in machine learning. In the face of massive data, how to mine useful information in data and optimization algorithm is a crucial issue. Semi-supervised learning can be trained together with both labeled and unlabeled data. Literature [5] describes the basic concepts, development history, classification, and related algorithms of semi-supervised learning.

Research Status of Semi-Supervised Learning

Semi-supervised learning is developed from self-learning. In the subsequent development, transductive learning, generative model, and other learning methods have gradually emerged. The self-training approach proposed by Scudder [6], Fralick [7], and Agrawala [8] is the first method to use a sample of unlabeled for supervised learning. It mainly uses the idea of iteration, keeps repeating the supervised learning, applies the optimized labeling results to the next round, and adds them into the sample set together with the class label to continue to supervise the result. The advantage of this method is that it is easy to understand and operate, but the error of iteration can lead to a vicious circle.

Semi-supervised was first proposed by Merz [9]. Shahshahani [10] has proposed that the use of unlabeled samples can help alleviate the "Hughes" phenomenon in small samples; this idea has led to widespread attention to unlabeled samples and semi-supervised learning. Wang [11] has proposed the bivariate graph-based semi-supervised learning method, which used the binary class labels information and the continuous classification functions to optimize the learning problem. Klein [12] first proposed a semi-supervised distance metric learning method of clustering. Cheng [13] has proposed a particle swarm optimization algorithm based on the semi-supervised classifier to solve the Chinese text classification problem. Wang [14] has proposed the semi-supervised hashing method of large-scale image retrieval problems. Semi-supervised learning can also be combined with deep learning. Kingma [15] has proposed the semi-supervised learning of the deep generation model. Zhang [16] has presented an unsupervised deep learning framework for dimension reduction.

Classification of Semi-Supervised Learning

According to the different learning scenarios, semi-supervised learning is divided into four categories as follows: semi-supervised dimensionality reduction [17], semi-supervised regression [18], semi-supervised clustering [19], and semi-supervised classification [20]. Generally, semi-supervised learning is mainly divided into the following: semi-supervised learning based on the generative model, semi-supervised learning based on a support vector machine, semi-supervised learning based on the graph, and collaborative training [21].

Semi-Supervised Learning Based on Generative Model

The generative model (GM) [22] models the joint distribution $p(x|y)$ of x and y, obtains $p(x|y)$ and $p(y)$, then the posterior distribution $\mathrm{p}(y|x) = \frac{\mathrm{p}(x|y)p(y)}{p(x)}$ is obtained by Bayes criterion, and finally makes predictions through the maximum posterior criterion. The main idea of the generative learning algorithm (GLA) [23] is to assume that the sample and class labels are some kind of probability distribution. According to the

generative model, $p(y)$ is a prior distribution and $p(x|y)$ is a class conditional distribution, and the iteratively sample x and y, where x belongs to $p(x|y)$ and y belongs to $p(y)$, then the posterior probability distribution $p(y|x)$ is obtained, and the label with the highest posterior probability distribution $\text{argmax}p(y|x)$ is found. In general, the generative method is to put the test cases into each model and finally compare the results and select the best as the class label.

Commonly generated models include Gaussian mixture model (GMM), naive Bayesian model (NBM), hidden Markov model (HMM), and hidden Markov random field (HMRF).

(1) GMM

GMM is mainly subject to Gaussian discriminant analysis (GDA). Baudat and Anouar [24]applied fisher linear discrimination to other problems and obtained GDA. GDA assumes an input characteristic $x \in Rn$, and $p(x|y)$ obeys a multidimensional normal distribution. The GMM obeys the Gaussian distribution and the sample set input feature x is a continuous random value, as shown in Eq. (1):

$$p(x|y) = \frac{1}{(2\pi)^{n/2}|\Sigma x|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \sum^{-1}(x-\mu_0)\right) \quad (1)$$

(2) HMM

Markov process is the basis of the hidden Markov model (HMM). If the present state x_t is known, the probability of the future state value $x_u(u > t)$ (or some state) is independent of the past state value $x_s(s > t)$, that is, the past, present, and future state conditions are independent and irrelevant of each other; this processed is called the Markov process. In the real world, many random processes are Markov random processes, such as the Brownian motion of particles in liquids, the number of people infected by infectious diseases, and the number of people waiting at the station.

HMM is the most commonly used mathematical model for solving time series. According to the Markov process, the transfer probability under a certain state should be transferred from one state to another state. The current state only depends on the previous state, and the samples are connected to each other through the hidden state to form a sequence. The basic model of HMM [25] is shown in Fig. 1:where y_{t-2}, y_{t-1}, y_t is the observation variable and x_{t-2}, x_{t-1}, x_t is the hidden variable. Because there must be noise in the process of observation, the state you observe is equivalent to an estimate of its true state. It is expressed in mathematical language as $p(y_t|x_t)$, which is the measurement model or the measurement probability. On the other hand, the real current state should be related to the previous observation state, there exists such a distribution $p(x_t|x_{t-1})$, which is called the transfer model or transfer probability in the model. Namely, the expression of two important probabilities in the HMM model is obtained, as shown in Eqs. (2) and (3):

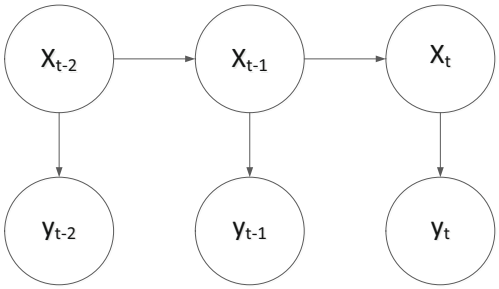

Fig. 1 The hidden Markov model

$$P(x_t|x_{t-1}, x_{t-2}, ..., x_1, y_1, ..., y_{t-1}) = P(x_t|x_{t-1}) \quad (2)$$

$$P(y_t|x_t, x_{t-1}, ..., x_1, y_1, ..., y_{t-1}) = P(y_t|x_t) \quad (3)$$

where Eq. (2) is the discrete transition probability; Eq. (3) is the continuous (or discrete) measurement probability.

Semi-Supervised Learning Based on Support Vector Machine

Transductive support vector machine (TSVM) proposed by Vapnik and Sterin [26] is an extension algorithm of the support vector machine [27] (SVM) for semi-supervised problems. In SVM, we need to find an optimal dividing line to make it the largest to both edges. The boundary is equidistant from the nearest point of the two types of points and the gap is the largest. Advances in semi-supervised support vector machines can be found in the literature [28]. In this case, a few data points with bold edges are called support vectors. Such as, Zhang H [29] has proposed a locality correlation preserving support vector machine based on traditional SVM. In order to ensure the consistency between the optimization goal of weak learners and the enhancement scheme, Tao [30] used to regard cost-sensitive SVM as basic weak leaner. For the tradeoff between speed and accuracy when classifying large data sets, Tang [31] combined of k-means clustering and multi-kernel SVM.

However, in many cases, a test set needs to be given and classified. In this case, this method is more effective than traditional SVM. Combining SVM and transduction reasoning, TSVM can make use of unmarked data, and TSVM [32] performs better in performance than traditional SVM. TSVM can use the core idea of transduction learning to solve the problem iteratively by using local search strategy and use labeled and unlabeled samples to determine the optimal classification boundary, find the largest edge and maximize the hyperplane. It cannot be only separate the existing data with minimum error, but also satisfy the classification of the

original unlabeled sample data. The TSVM consumes a lot of time because it has to traverse all unmarked samples. K2 TSVM [33] uses k-means to cluster unlabeled samples and deletes some samples according to the distribution of positive and negative data; the test data are input into TSVM. The K2 TSVM method does not need to traverse all data, which reduce the time compared with traditional TSVM.

TSVM feature vector set X and tag set Y have a geometric relationship. Suppose the hyperplane in space can be expressed as shown in Eq. (4):

$$w*x + b = 0 \tag{4}$$

Here, the edge value of the hyperplane of X is the distance between the hyperplane and the nearest sample as shown in Eq. (5) and is explained in more detail in the literature [34]:

$$\min_{i\in[1,2,\ldots,n]} \left[\frac{y_i}{\|w\|}(w*x + b) \right] \tag{5}$$

The objective function of TSVM is shown in Eq. (6):

$$\min_{w,b,Y_u} \frac{1}{2}\|w\|^2 + C_1 \sum_{i=1}^{l} V(y_i, f(x_i)) + C_2 \sum_{i=l+1}^{l+u} V\left(\hat{y}_i, f(x_i)\right) \tag{6}$$

In Eq. (6), $V(y_i, f(x_i))$ is the loss function and C_1 and C_2 are the regularization parameters. As the loss weight value on the label-free sample, the sample complexity, and the actual error are weighed.

Semi-Supervised Learning Based on Graph

Graph-based semi-supervised learning has been widely developed in recent years. The main idea is to construct a sample graph which the vertices represent samples and define the relationship of edges. The weight of the edges represents the degree of similarity between the two points. An effective energy function is used as the evaluation standard. In graph-based semi-supervised learning, the Min-cuts algorithm, and manifold regularization algorithm are typical algorithms.

The Min-cuts algorithm was first proposed by Blum and Chawla [35]. The source node is the positive marking instance, and the target node is the negative marking instance. After finding a set of deleted edges, the source node and the target node can be separated. There are other algorithms gradually appeared, for example, Richard Szeliski [36] conducted a study on energy function minimization and confirmed that the graph cut algorithm can guarantee high efficiency. Meanwhile, Zhu [37] has proposed the proportional cut method, and Zhou [38] has proposed the normalized cut method. The manifold semi-supervised learning method is mainly summarized by Belkin [39] and the manifold regularization method is proposed. After that, Goldberg [40] has proposed the regularization of online manifolds, which improved the applicability of regularization of manifolds in large-scale data. Balcan [41] used strong domain knowledge to construct the graph and then conducted character recognition of the semi-supervised learning based on the graph. The edges in the figure are combined by time edge, color edge, and face edge, a graph like this can reflect strong domain knowledge, a deep understanding of the problem structure, and how to use unannotated data.

Collaborative Training

Blum and Mitchell [42] have proposed the original method of collaborative training. Sample data set is set to two fully redundant views, which are first used for the labeled samples to obtain the classifier. When the sample is sufficiently large, the unlabeled sample with higher credibility in any classifier is selected. Mark it out and add it to the training data of another classifier, update it, and repeat it until you get a strong learner on any of the attribute sets. Goldman [43] proposed an improved algorithm that does not require sufficient redundant views; this algorithm needs to estimate the credibility of unlabeled data on two classifiers and classify them according to credibility.

The comparison of the four methods is shown in Table 1.

Semi-Supervised Clustering

In real life, how to use unlabeled data for clustering is always the research direction. The comprehensive development of semi-supervised learning lays a foundation for semi-supervised clustering. Semi-supervised clustering can use a small amount of prior information to process unlabeled data. On the basis of clustering and semi-supervised learning, semi-supervised clustering can improve the results of traditional clustering and get better results.

The semi-supervised clustering algorithm mainly contains two kinds of prior knowledge supervision information: independent class label and the pairwise constraint.

(1) Independent class label

In the unsupervised clustering algorithm, the class labels of all data samples are unknown in advance. When we know the real class label of a small number of samples, the original data set X can be expressed as follows:

$$X^l = \left\{ \boldsymbol{x}_1^l, \boldsymbol{x}_2^l, \boldsymbol{x}_3^l, \ldots, \boldsymbol{x}_{n_l}^l, \boldsymbol{x}_1^u, \boldsymbol{x}_2^u, \boldsymbol{x}_3^u, \ldots, \boldsymbol{x}_{n_u}^u \right\}$$

Table 1 Comparison of four classification algorithms for semi-supervised learning

Based on	Advantages	Disadvantages
GM	(1) Simple and intuitive (2) Better performance can be achieved when there are few labels.	(1) Sufficient prior knowledge is required. (2) When the model hypothesis is inconsistent with the data distribution, it is easy to reduce the model generalization ability.
SVM	(1) TSVM is better than conventional SVM in the high-dimensional data.	(1) If the loss function is not convex, it will fall into the local minimum point and affect the generalization ability.
Graph	(1) This method is more intuitive and has the good explanatory ability and learning performance.	(1) A good foundation in mathematics is needed. (2) It is difficult to carry out semi-supervised learning on large-scale unlabeled data.
Collaborative training	(1) Compared with the generative model and EM algorithm, when the feature set is large enough, the best effect can be obtained by the collaborative training.	(1) If the initial classifier is weak, unlabeled data may be mislabeled, causing noise to be introduced.

In the above formula, $X^l = \left\{\boldsymbol{x}_1^l, \boldsymbol{x}_2^l, \boldsymbol{x}_3^l, ..., \boldsymbol{x}_{n_l}^l\right\}$ represents a sample set of n_l known class labels, and $X^u = \left\{\boldsymbol{x}_1^u, \boldsymbol{x}_2^u, \boldsymbol{x}_3^u, ..., \boldsymbol{x}_{n_u}^u\right\}$ represents a sample set of n^u unknown class labels. In general, $n^l << n^u$, the number of known class label samples is much smaller than the number of unknown class label samples.

(2) Pairwise constraint relation

Constraint relation is also called the constraint condition. In real life, it is difficult to obtain samples of independent class labels. If you insist on using independent data, it will take a lot of effort to collect, but it is much easier to find the relationship between two pairs of data. In order to apply constraint relation to clustering, Wagstaff [44] proposed two paired constraints, namely Must-Link (ML) and Cannot-Link (CL), in which ML is positive association constraint relation, indicating that two data instances must be in the same class, while CL is the opposite, indicating that two data instances must be in different classes. That is, suppose two data instances x_i and x_j belong to class C_i and C_j respectively, if $(x_i, x_j) \in \text{Must} - \text{Link}$, $i = j$ is required, if $(x_i, x_j) \in \text{Cannot} - \text{Link}$, then $i \neq j$.

The pairwise constraint can only represent the obvious constraint relationship of the sample data. It can be constrained by a small amount of easily available information because of the little knowledge required. And, ML and CL constraints have symmetry and transitivity:

Symmetry:

$$\begin{aligned}&(x_i, x_j) \in ML \Rightarrow (x_j, x_i) \in ML\\ &(x_i, x_j) \in CL \Rightarrow (x_j, x_i) \in CL\end{aligned}$$

Transitivity:

$$\begin{aligned}&(x_i, x_j) \in ML \,\&\, (x_j, x_k) \in ML \Rightarrow (x_i, x_k) \in ML\\ &(x_i, x_j) \in ML \,\&\, (x_j, x_k) \in CL \Rightarrow (x_i, x_k) \in CL\end{aligned}$$

Classification of Semi-Supervised Clustering

In general, semi-supervised clustering methods are divided into the following three categories:

The first category is the constraint-based semi-supervised clustering method (CBSSC). This algorithm is to add constraint restriction information to optimize the clustering effect based on traditional clustering. Common methods are the following: (1) let the sample forcibly satisfy the constraints, and ensure that the final clustering results meet all given constraints. (2) Adding a penalty factor to the objective function, penalizing the data samples that violate the constraint condition, so that the final clustering result satisfies the constraint condition to the maximum extent. This algorithm is not mandatory, so the final result may be unsatisfied. (3) A given constraint is an independent class label, also known as a seed set (seeds); it is used to initialize the cluster center and require the cluster center to satisfy the given constraints.

The second category is the distance-based semi-supervised clustering method (DBSSC). The characteristic of this kind of algorithm is that in the process of data preprocessing, the similarity measure between samples is transformed to obtain a new measurement function, which is closer to the associated positive constraint samples, and the negative samples are more opposite. Common methods are (1) using the shortest path algorithm to adjust the sample distance, so that the adjusted distance satisfies the given constraint; (2) the horse

distance based on the convex optimization problem training; (3) in the spectral clustering algorithm, the constraint information is used to adjust the distance between samples, and the clustering problem is transformed into the spectral decomposition of the image Laplace matrix.

The third category is constraint and distance-based semi-supervised clustering method (CDBSSC). This is a new algorithm that combines the above methods.

Semi-Supervised Clustering Algorithm

Semi-supervised clustering is improved on traditional clustering. The semi-supervised clustering mainly uses the following distance equation (Eq. (7)–(10)) in distance measurement:

1. Euclidean distance

$$Dist(x_i, x_j) = \sqrt{(x_{i1}-x_{j1})^2 + (x_{i2}-x_{j2})^2 + \cdots + (x_{id}-x_{jd})^2} \quad (7)$$

2. Manhattan distance

$$Dist(x_i, x_j) = |x_{i1}-x_{j1}| + |x_{i2}-x_{j2}| + \cdots + |x_{id}-x_{jd}| \quad (8)$$

3. Murkowski distance

$$Dist(x_i, x_j) = \left(|x_{i1}-x_{j1}|^h + |x_{i2}-x_{j2}|^h + \cdots + |x_{id}-x_{jd}|^h\right)^{\frac{1}{h}} \quad (h > 0 \wedge h \in Z) \quad (9)$$

4. Cosine distance

$$Dist(x_i, x_j) = 1-sim(x_i, x_j) = 1-\frac{\sum_{k=1}^{d}(x_{ik} \cdot x_{jk})}{\sqrt{\sum_{k=1}^{d} x_{ik}^2 \cdot \sum_{k=1}^{d} x_{jk}^2}} \quad (10)$$

K-means algorithm is the representative of the traditional algorithm, and its central idea is to assign each data object to the class closest to it. The objective function definition of the k-means algorithm is shown in Eq. (11):

$$J_{K-means} = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left\|x_i - u_j\right\|^2 \quad (11)$$

where u_j represents the center point of the class C_j to which the sample x_i belongs. $J_{K-\text{means}}$ is the sum of the squares of the distance from the data sample to the corresponding class center. The K-means algorithm process is shown in Table 2:

Many new research algorithms have been obtained in the continuous development. Here, the most typical algorithm of semi-supervised clustering is introduced in detail.

Cop-kmeans Algorithm

The Cop-kmeans algorithm is a semi-supervised clustering algorithm, which is often used to introduce pairwise constraint information into the K-means algorithm. The basic clustering idea is the same as K-means, except that in the data distribution process, the data object must satisfy the ML constraints and CL constraints. Suppose that the number of clusters in a data set is $K = 2$; C_1 and C_2 represent two parts of the data set; u_1 and u_2 represent the center points of each part; and x_i and x_j are two data instances. As shown in Fig. 2, the solid lines represent ML constraint between two data instances. And, in Fig. 3, the dashed line represents the CL constraint between two data instances.

As shown in Fig. 2, when x_i has been allocated to the nearest C_1 class after the initial distance calculation, x_j is the sample that is not currently assigned, since x_i and x_j have a positively associated constraint relationship $(x_i, x_j) \in \text{ML}$. The Cop-kmeans algorithm does not calculate the distance between the sample x_j and the center points of the two classes, but will directly assign x_j to the C_1 class. Even though x_j may be closer to the center point of the C_2 class, this phenomenon is because of the ML constraint, two data instances are required to be assigned to the same class.

As shown in Fig. 3, when x_i is already allocated to the C_2 class, x_j is the sample that is not currently assigned. When the number of clusters $K = 2$, x_j is assigned directly to class C_2. If the number of clusters is greater than 2, x_j will be assigned to the nearest class except for C_1.

Table 2 k-means algorithm process

Input: number of clustering K, the data set $X = \{x_0, x_1, ..., x_{N-1}\}$
Output: initial cluster division $C = \{C_1, ..., C_k\}$:
Step 1 Randomly selects k center points from N data objects.
Step 2 Traverses all data and divides each data into the nearest cluster.
Step 3 After each iteration, calculate the average value of clustering.
Step 4 Continues to repeat steps 2 and 3 until the k midpoint no longer changes (converges) or stops when enough iterations are performed.

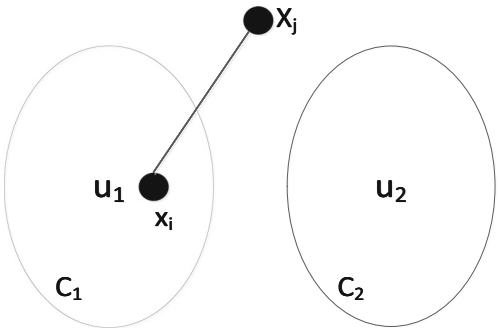

Fig. 2 ML constraint

The Cop-kmeans algorithm process is shown in Table 3:

The result of the Cop-kmeans algorithm is that the constraint must be satisfied. Due to the symmetry and transitivity of ML and CL, if the cluster meeting the constraint is not found, the return value is empty. The noise of the data lead to the instability of the convergence of the final result, and finally, the algorithm fails. There are three specific reasons for failure:

Situation 1: The ML constraint causes constraint violations.

As can be seen from Fig. 4a, it is assumed that the number of clustering $k=2$, x_j has been allocated to C_1, x_i has been allocated to C_2, and$(x_j, x_k) \in \mathrm{ML}$, $(x_i, x_k) \in \mathrm{ML}$, this indicates that the real class labels for x_i and x_j should be the same. When the data instance x_k is allocated, because x_k and x_j have positive association constraints, it should be assigned to class C_1. However, x_k and x_i also have positive association constraints. As a result, x_k cannot find the proper cluster allocation. The collision caused the algorithm to fail.

However, the premise is that they have been divided into different classes, and the ML constraint violation will not occur until x_k is allocated. However, according to the transitivity of constraints, if $(x_i, x_j) \in \mathrm{ML}$ & $(x_j, x_k) \in \mathrm{ML} \Rightarrow (x_i, x_k) \in \mathrm{ML}$, in Fig. 4b, x_i and x_j must be divided into the same class before allocating x_k, so this does not result in a constraint violation.

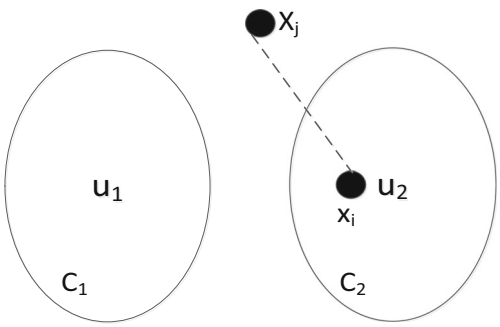

Fig. 3 CL constraint

Table 3 Cop-kmeans algorithm process

Input the initial number of clustering K, data set $X=\{x_0, x_1, ..., x_{N-1}\}$
Output the set of initial cluster division cluster $C=\{C_1, ..., C_k\}$
Step 1 Like the k-means algorithm, randomly selected K data instances as the initial center point.
Step 2 Select some data randomly from the data set and generate ML constraint sets and CL constraint sets respectively.
Step 3 For any sample x_j to be allocated:
If $(x_i, x_j) \in \mathrm{ML}$, and x_i has been allocated to class C_i, then x_j is also allocated to class C_i;
If $(x_i, x_j) \in \mathrm{CL}$, and x_i has been assigned to class C_i, then x_j is allocated to the nearest class C_j except for class C_i;
If x_i and x_j have no constraint, then x_j is allocated to the nearest class;
Step 4 After each iteration, update the center point of each class.
Step 5 Repeat steps 3–4 until the change of C_i value is less than the given threshold value or the number of iterations reaches the maximum value.

Situation 2: The relationship between ML and CL creates constraint violations.

It can be seen from Fig. 5a that both x_i and x_j have been assigned to the C_1 class, and there is no constraint relation between x_i and x_j, but $(x_j, x_k) \in \mathrm{CL}$, $(x_i, x_k) \in \mathrm{ML}$. Because x_k and x_i have ML constraints, when the data instance x_k is allocated, it should be assigned to class C_1. However, x_k and x_j have CL constraints and cannot be assigned to C_1 class. Due to both positive and negative association constraints, x_k cannot be properly allocated and the algorithm fails.

In Fig. 5b, it can be seen from $(x_i, x_k) \in \mathrm{ML}$ & $(x_j, x_k) \in \mathrm{CL} \Rightarrow (x_i, x_j) \in \mathrm{CL}$ that x_i and x_j must be divided into different classes before allocating x_k. At this point, $i \neq j$, which is inconsistent with the condition that x_i and x_j are in the same class. Therefore, this also does not result in a constraint violation.

Situation 3: The ML constraint causes constraint violations.

As can be seen from Fig. 6, it is assumed that the number of clustering $k=2$, x_i has been allocated to C_1, x_j has been allocated to C_2, and $(x_j, x_k) \in \mathrm{CL}$, $(x_i, x_k) \in \mathrm{CL}$, this indicates that the real class labels for x_i and x_j should be the same. But when the data instance x_k is allocated, it cannot be assigned to the C_1 class. At the same time, x_k and x_j also have negative association constraints and cannot be assigned to the C_2 class. This also indirectly results in x_k being unable to find the appropriate class to allocate. Therefore, it is because CL does not have the transitivity that the algorithm fails. And only the Cannot-Link constraint will result in a constraint violation.

Lcop-kmeans Algorithm

Lcop-kmeans algorithm improved the problem of constraint violation. It is an improved Cop-kmeans algorithm based on

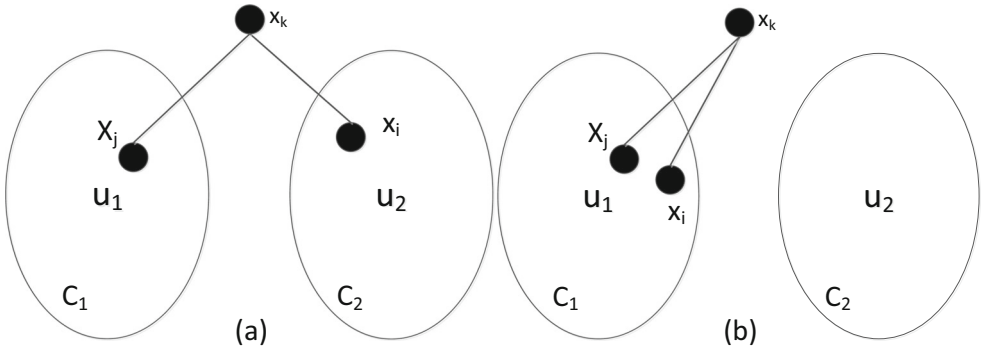

Fig. 4 ML constraint causes constraint violations

breadth-first searches (BFS), also be named linked to the Cop-kmeans algorithm [45] (Lcop-kmeans). The main idea is in the process of data iteration; the data with constraints are different from the data without constraints. Unconstrained data can be directly assigned to the nearest cluster when it is classified, then the constraint relations between other ML and CL constraint can be determined. When the ML constraint exists, it is distributed according to the transitivity to ensure that the constraint is not violated and the distribution results are obtained. When the CL constraint exists, the BFS search algorithm is adapted to insert the data of all CL constraints associated with a vertex into a first-in-first-out list structure in order, after which all the vertices adjacent to that vertex are accessed. This method can prevent data from being accessed repeatedly and ensure that the CL constraint is not violated.

The process of Lcop-kmeans algorithm is shown in Table 4:

Seeded-kmeans Algorithm

Different from the paired constraint algorithm previously introduced, the Seeded-kmeans algorithm aims to introduce the mark sample into Kmeans. The marked sample can be very small. As seeds set, it can be divided into K clusters. See Table 5 for the process of the Seeded-kmeans algorithm:

SC-kmeans Algorithm

Chen [46] has proposed Kmeans based on seed set and pairwise constraints (SC-kmeans algorithm), which is an active semi-supervised clustering algorithm based on seed and pairwise constraints and improved the Seeded-kmeans algorithm. In this algorithm, the pairwise constraints need to be expanded to obtain New-ML and CL constraint sets, namely New-ML and New-CL constraints. According to the transfer and symmetry of ML, a new equivalence class TML-Set is obtained, where$\mathrm{TML{-}Set} = \bigcup_{i=1}^{s} T_i$.

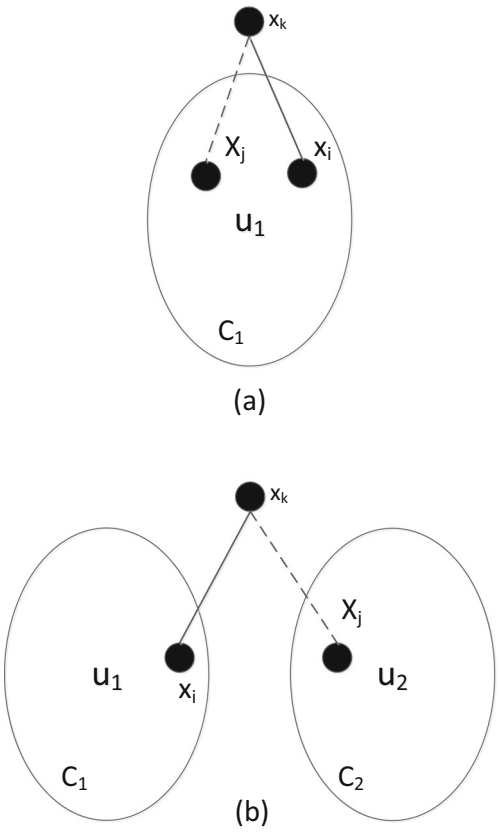

Fig. 5 ML and CL constraint causes constraint violations

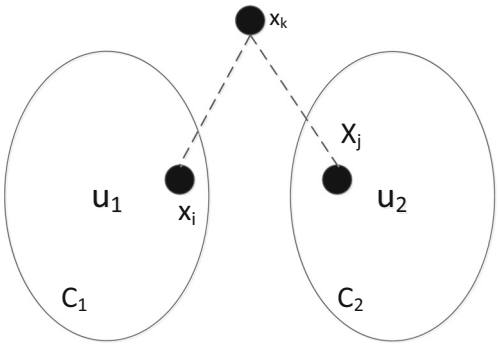

Fig. 6 ML constraint causes constraint violations

Table 4 Lcop-kmeans algorithm process

Input cluster number K, the data set $X = \{x_0, x_1, ..., x_{N-1}\}$

Output initial clustering division $C = \{C_1, ..., C_k\}$:

Step 1 Randomly select K data instances as the initial center point.

Step 2 Selects part of data to generate ML and CL and makes a preliminary extension of them with transitivity.

Step 3 For the unconstrained data x_i, allocate it to the nearest C_i class.

Step 4 If the sample x_i has a ML, the data instances with ML relation to x_i are saved in the list in turn, assign all instances to the Ci class, and mark them as assigned.

Step 5 If x_i has CL, save the data instances that have a direct CL relationship with x_i in the list. According to the CL constraint rule, they are assigned to the appropriate classes.

Step 6. Each iteration is completed, and the center point of each class is updated.

Step 7 Repeat steps 3–6 until the change of C_i value is less than the given threshold value or the number of iterations reaches the maximum value.

The process of SC-kmeans algorithm is shown in Table 6:

Other Algorithms

Davidson has proposed the CVQE [47] algorithm. Dan [48] has improved on the basis of CVQE and proposed the LCVE algorithm. The CVQE algorithm is a soft constrained clustering algorithm; it does not enforce constraints, and to some extent allows constraint violations, even if the data concentration is noisy or incomplete, it does not cause the algorithm to be affected. The CVQE algorithm attempts to minimize the following objective functions 12:

$$J_{CVQE} = \sum_{i=1}^{n} \left\|x_i - \mu_{l_i}\right\|^2 + \sum_{(xi,xj\in m, l_i \neq l_j)} \left\|\mu_{l_i} - \mu_{l_j}\right\|^2 + \sum_{(xi,xj\in c, l_i \neq l_j)} \left\|\mu_{l_i} - \mu_{hl_j}\right\|^2 \quad (12)$$

Table 5 Seeded-kmeans algorithm process

Input data set D, seeds set S;

Output K clustering division;

Step 1 Calculate k initial clustering centers $\mu_h = \frac{1}{|S_h|} \sum_{x \in S_h} x, h = 1, 2, 3, ..., k$, from seeds set S.

Step 2 For each sample x_i in D, calculate the distance between x_i and k clustering centers.

Step 3 Divides x_i into the cluster x_h, so that the cluster center of x_h is closest to x_i.

Step 4 Update new clustering center.

Step 5 If all k clustering centers are not changed, end the algorithm; Otherwise, go back to step 1 and continue the iteration.

Table 6 SC-kmeans algorithm process

Input data set D, seeds set S, ML constraint set New-ML, and CL constraint set New-CL

Output clustering divisions:

Step 1 Calculate K initial cluster centers $\mu_h = \frac{1}{|S_h|} \sum_{x \in S_h} x, h = 1, 2, ..., k$ from the seeds set S.

Step 2 According to New-ML, calculate the transfer closure TML Set = $\{T_1, T_2, ..., T_S\}$.

Step 3 Initializing the clustering center for each transitive closure T_i in the TML-Set, denoted as centers.

Step 4 For each sample, x_i in T_i calculate the distance from x_i to K initial cluster centers; add the cluster center closest to distance x_i to centers; perform cluster center Sort according to the frequency of each cluster center. Generating T_i cluster center sequence queue centers sort, when centers sort is not empty, take out the team head element as μ_h; if x_i and cluster center μ_h represent cluster x_h do not violate the known CL constraint, add all samples in T_i to cluster x_h; otherwise remove μ_h from centers sort.

Step 5 For other samples, x_j in D that is not TML-Set, calculate the distance from x_j to K cluster centers and find the nearest center point μ_h; if x_j and μ_h represent the cluster, x_h does not violate the known CL constraint, assign x_j to cluster x_h; otherwise, do not do any processing on x_j.

Step 6 Recalculate the new cluster center: $\mu_h = \frac{1}{|S_h|} \sum_{x \in S_h} x, h = 1, 2, ..., k$.

Step 7 If the K cluster centers have not changed, end the algorithm; otherwise, repeat iteration step 1 until convergence.

Research Status of Semi-Supervised Clustering

According to the classification of semi-supervised clustering, the status of such research can be introduced from the following two aspects:

(1) Research status of semi-supervised clustering based on constraint

The most typical and basic algorithms in the semi-supervised clustering algorithm based on constraints are the pairwise constraint thought proposed by early Wagstaff [49] and the idea through seed improved by Basu [50]. Wagstaff applied the pairwise constraint idea to the traditional Kmeans algorithm and proposed the Cop-kmeans algorithm to improve the clustering performance by using the constraint rules between data. This algorithm is a mandatory constraint algorithm. Basu also improved the traditional Kmeans algorithm, using the maximum expectation algorithm and seeds set to divide a small amount of marker data, optimized the initialization of traditional clustering algorithm, and proposed Seeded-kmeans algorithm and Constrained-kmeans algorithm which can improve clustering performance. These two algorithms are crucial in the development of semi-supervised clustering.

Later generations have improved these two algorithms. For example, Li [51] has proposed the semi-supervised hierarchical clustering algorithm, which can also use pairwise

constraints in hierarchical clustering. Literature [52] has proposed that the semi-supervised clustering algorithm can randomly walk in the lower part of the vertex and the upper part of the component. The low-level random walk obtains the constraint ability range of the constrained vertex to other vertex as a component and propagates the constraint on the high-level random walk, which ultimately comes down to a cluster. The literature [53] has proposed an improved density sensitive semi-supervised clustering algorithm. Yang [45] has proposed a semi-supervised clustering integration framework based on constrained self-organization mapping and improved Cop-kmeans. Shi [54] has proposed an efficient semi-supervised cooperative spectral clustering algorithm. In the literature [55], a semi-supervised clustering method based on the AP algorithm is proposed to optimize the clustering effect indirectly. Yang [56] has improved the cop-kmeans algorithm with Mapreduce, so that large-scale data sets can be operated in parallel. Sun et al. [57] have proposed an improved k-means clustering algorithm based on clustering for balancing constraints. Yin [58] has proposed a discriminant semi-supervised cluster analysis method based on paired constraints and added penalty factors to reduce complexity. Wei et al. [59] have proposed a semi-supervised clustering set based on pairwise constraints and metrics. In the case of pairwise constraints and labeled data, use constraint-based semi-supervised clustering and metric-based semi-supervised clustering to generate different partitions. And then get the target cluster by integration. This method can improve the clustering accuracy. Li [60] has proposed the method which uses the cross-entropy of samples to express the pair constraint information and can achieve higher results with better clustering accuracy when the pair constraint information is less. Ding [61] has proposed a local adaptive multi-core K-means based on a shared nearest neighbor. Chai [62] has proposed a semi-supervised k-means clustering algorithm with active learning. In this method, active learning is added into the constrained semi-supervised clustering, and the most useful unlabeled data samples in each cluster can be selected actively. Compared with the traditional semi-supervised clustering algorithm, this method has higher efficiency and better results.

In the field of Markov random airport, semi-supervised clustering can also be involved. For example, Basu [63] proposed a semi-supervised clustering based on the paired constraints on the hidden Markov random airport. Jia [64] proposed a semi-supervised approximate spectral clustering algorithm based on the HMRF model.

(2) Research status of semi-supervised clustering based on distance

In the distance-based semi-supervised clustering, Saha [65] studied the distance measurement of genetic semi-supervised clustering based on point symmetry. According to the label information, the other literature [66] proposes a semi-supervised clustering algorithm based on spectral clustering; the distance matrix was determined for spectral clustering. Bilenko [67] combined with constraints and distance to come up with a new approach. Alok [68] combined feature selection with semi-supervised clustering, solved the feature selection problem under the unsupervised classification framework, and compared multi-objective-based automatic clustering techniques, single-objective clustering techniques, and traditional K-means clustering. It is concluded that the semi-supervised feature selection technique can detect appropriate feature combinations and appropriate partitions from data sets with point-symmetric clustering.

Application

Semi-supervised clustering has gradually become a research hotspot in the field of machine learning and can solve practical problems, including road detection, image, classification, information retrieval, speech recognition, bioinformatics, and other fields. The following introduces the application of semi-supervised clustering in recent years through different fields.

First, the most typical application is road detection in GPS data. A variety of applications are currently supported on many maps, such as the ability to generate detailed custom driving directions. However, we do not know whether the road information is very rough, whether the road is smooth, and how much traffic the vehicle has a certain period of time. In the article [44], the K-means algorithm with background knowledge is improved, and the above questions are also answered. The main idea to solve this problem is according to the situation of the vehicle driving in the lane boundaries, the lane should be denser over time, and the lane boundaries should be sparse. Car position data are then aggregated and the position of each lane is automatically determined using two specific heuristics that produce constraints: tracking proximity and maximum separation. The tracking approach is equivalent to Must-Link, which means that all data points generated by the same car in one direction are in the same lane without changing the lane. The maximum distance is equivalent to Cannot-Link. When the distance between the two points exceeds 4 m, then we have a constraint that prevents the two points of being placed in the same cluster. Finally, by comparing several different constraints, we conclude that the constrained cop-kmeans algorithm has achieved very good results in processing road detection.

Second, in the area of bioinformatics, semi-supervised clustering has also been applied. Gui [69] has combined the semi-supervised learning with tumor classification and proposed a semi-supervised tumor classification method based on graphs; a new multi-step dimensionality reduction method

was used to extract tumor features and to evaluate performance using semi-supervised learning. This algorithm is better than SVM, and the method is effective and feasible. Similarly, in the area of cancer, Saha [70] applied semi-supervised clustering to the division of gene expression data sets to address cancer tissue classification problems, especially for three open source benchmark cancer data sets. In the end, the classification scheme can be obtained within a reasonable time frame. Yu [71] proposed a semi-supervised clustering integration framework for knowledge applied to biomolecules. Shiga [72] proposed a semi-supervised clustering algorithm based on information multi-graphs and used in bimolecular data sets. The literature [73] has applied semi-supervised clustering integration in bimolecular model mining. Due to the huge amount of cancer data, clustering can easily and clearly classify huge amounts of data, but in cancer data sets, it may also appear some problem like known cancer diagnosis information, including expert opinion guidance, and related cancer conditions. To make cancer diagnosis more accurate, this paper proposes two kinds of semi-supervised clustering integration frameworks based on this background knowledge. It can effectively remove noise and irrelevant data in the data, get accurate potential relationships, and help cancer and biological fields. Another literature [74] was the application of semi-supervised clustering in the electrocardiogram; the results were simple and direct; training time is also shorter.

Thirdly, in terms of safety, such as collecting wrong information and constraint violation, they will lead to reducing clustering performance. Gan [75] has developed a partial homogeneous and consistent semi-supervised clustering with the class label as prior knowledge. The basic idea is that when the labeled sample is at risk, the predictions of the labeled sample and its closest homogeneous unlabeled sample should be similar. And create a partial graph that models the relationship between a labeled sample and its nearest homogeneous unlabeled sample by unsupervised clustering results. A graph-based regularization term is then constructed to approximate the prediction of the labeled sample to the prediction of the local isomorphic neighbor. It is expected to reduce the risk of label samples. Other scholars have made relevant contributions to safe semi-supervised clustering. Such as Syed F H et al.[76] are proposed the safe semi-supervised multi-target regression for new targets learning. Wang Y et al. [77] have proposed the safe semi-supervised collaborative filtering the recommendation algorithm.

In the field of image segmentation, semi-supervised clustering has been widely used. The quality of image segmentation results will affect the subsequent work and analysis. The essence of image segmentation is to divide the similarity between pixels. Lu [78] used the semi-supervised clustering of graphs in the interactive image classification. Portela [79] also used semi-supervised clustering in MR brain image segmentation. Hasnat M A [80] has put forward the joint color-spatial-directional clustering and region merging for unsupervised RGB-D image segmentation. An [81] studied the problem of unclear image segmentation by using a semi-supervised k-means algorithm, labeled several independent pixels, and clustered them using seed set. In the face of noisy images, a robust semi-supervised clustering segmentation algorithm is proposed to the literature [82], in the stage of data preprocessing, guiding filter preprocessing is carried out first, and then the semi-supervised constraint is carried out, which can make the image segmentation effect on the condition of noise unaffected. The inconsistency between text features and visual content results in poor image search results. To solve this problem, a ranking model [83] that can utilize visual features and click features is proposed. In the same field, a sparse COD method based on multi-mode hypergraph learning [84] is also proposed predicting the clicks of images and obtained click data are applied to image reordering.

In other fields of computer, such as Cheng [85] proposed a semi-supervised clustering method-based test cases selection techniques, for software regression testing of a large number of revised set of use cases, regression test case selection technology is combined with a semi-supervised clustering, the discriminative semi-supervised K-means clustering method (DSKM) is obtained. It is still using the pairwise constraint information and label data in the background knowledge and uses semi-supervised clustering to obtain more optimized tests result. Improve the efficiency of regression testing.

In the field of deep learning, semi-supervised clustering is also applied. In the distance metric learning [86], a structured ranking model is adapted to utilize both visual and click features. MFH [87] algorithm can use their correlation to achieve more effective fusion of multi-modal features, taking into account both the content sensitivity of the image shared and the trust that the user is given to view the image, the fine-grained privacy settings for social image sharing [88] are developed. Literature [89] has proposed an end-to-end location recognition model based on the new deep neural network, which can use the spatial pyramid structure of images to enhance VLAD features. In the gesture recovery, a new retrieving framework [90] can be obtained by learning the method of joint locating and retrieving. Multi-view location sensitive sparse retrieval [91] can solve the high and low efficiency of dimensions. And non-linear mapping and multi-layer deep neural network [92] can also play a good role in human pose recovery.

The same idea was applied to intrusion detection, for example, Mukkamala [93] used SVM to extract the features of the intrusion system. In terms of unsupervised learning, Zhang [94] has proposed creating ensembles of classifiers via fuzzy clustering and deflection. Depren [95] has proposed the intelligent intrusion detection system for anomaly and misuse detection in a computer network. However, due to a lot of noise and incomplete information on big data, especially in the case

of some prior knowledge, the effect of supervised learning and unsupervised learning is not very ideal. At this time, the idea of combining with semi-supervised learning is putting forward. For example, Fiore U [96] has proposed using the limited Boltzmann machine for network anomaly detection, and Liang [97] has proposed new semi-supervised intrusion detection methods, which are achievements in this field. In other related fields, the literature studies a new semi-supervised intrusion detection method. The cop-kmeans algorithm can increase the number of labels, reduce the complexity of the algorithm, and be practical. Peng [98] also combined semi-supervised clustering in micro video direction; Zhong S [99] also involved in document clustering.

Summary and Outlook

This review starts from the two perspectives of semi-supervised learning and semi-supervised clustering and introduces the comparison of the advantages and disadvantages of semi-supervised learning and its classification. On this basis, semi-supervised clustering can be better understood. At the same time, it is suggested that the design of the semi-supervised clustering algorithm can be carried out from multiple perspectives. The future research direction is as follows:

(1) According to different clustering ideas, different semi-supervised clustering algorithms can be formed, such as semi-supervised hierarchical clustering, semi-supervised density clustering, and semi-supervised spectral clustering.
(2) Focus on the shortcomings of semi-supervised clustering, such as how to reduce data noise, how to improve accuracy, and how to make full use of prior knowledge.
(3) It can integrate multiple fields at the same time, apply the idea of semi-supervised clustering algorithm in different fields, add knowledge in different fields, and get more optimized results.
(4) The current data volume is very large. In the case of large-scale data and high-dimensional data, how to solve this problem is also what semi-supervised clustering algorithm needs to face.
(5) Whether it is semi-supervised clustering, traditional clustering, or even other algorithms, data preprocess is an important step to ensure the optimal final output.
(6) The evaluation criteria for semi-supervised clustering are also very important. Reasonable formulas and similarity evaluation criteria are used to judge whether the algorithm is good or not.

Funding information This work is supported by the National Natural Science Foundation of China under Grant Nos .61672522 and No.61379101.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Informed Consent All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1975, as revised in 2008 (5). Additional informed consent was obtained from all patients for which identifying information is included in this article.

Human and Animal Rights This article does not contain any studies with human or animal subjects performed by any of the authors.

References

1. Hartigan JA, Wong MA. Algorithm AS 136: a k-means clustering algorithm. J R Stat Soc Ser C Appl Stat. 1979;28(1):100–8.
2. Maddah M, Crimson WEL, Warfield SK. Statistical modeling and EM clustering of white matter fiber tracts. IEEE International Symposium on Biomedical Imaging: Nano To Macro. IEEE; 2006. p. 53–56.
3. Li KL, Cao Z, Cao LP, et al. Some developments on semi-supervised clustering. Int J Pattern Recognit Artif Intell. 2009;22(5):735–42.
4. Chen WJ. Semi-supervised learning study summary. Comput Knowl Technol. 2011;07(16):3887–9.
5. Liu JW, Liu Y, Luo XL. Semi-supervised learning methods. Chin J Comput. 2015;38(08):1592–617.
6. Scudder HI. Probability of error of some adaptive pattern-recognition machines. IEEE Trans Inf Theory. 1965;11(3):363–71.
7. Fralick S. Learning to recognize patterns without a teacher. IEEE Trans Inf Theory. 2003;13(1):57–64.
8. Agrawala A. Learning with a probabilistic teacher. IEEE Trans Inf Theory. 1970;16(4):373–9.
9. Merz CJ, St. Clair DC, Bond WE. Semi-supervised adaptive resonance theory (SMART2). Int Jt Conf Neural Netw IEEE. 1992;3:851–6.
10. Shahshahani BM, Landgrebe D. The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon. IEEE Trans Geosci Remote Sens. 1994;32(5):1087–95.
11. Wang J, Jebara T, Chang SF. Semi-supervised learning using greedy max-cut. J Mach Learn Res. 2013;14(1):771–800.
12. Klein D, Kamvar SD, Manning CD. From instance-level constraints to space-level constraints: ,making the most of prior knowledge in data clustering. The Nineteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc. 2002. p. 307–314.
13. Cheng S, Shi Y, Qin Q. Particle swarm optimization based semi-supervised learning on Chinese text categorization. IEEE Congress on Evolutionary Computation Cec; 2012. p. 1–8.
14. Wang J, Kumar S, Chang SF. Semi-supervised hashing for scalable image retrieval. IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, Ca, Usa, 13–18 June. DBLP, 2010:3424–3431.
15. Kingma DP, Rezende DJ, Mohamed S. Semi-supervised learning with deep generative models. Adv Neural Inf Proces Syst. 2014;4:3581–9.
16. Zhang J, Yu J, Tao D. Local deep-feature alignment for unsupervised dimension reduction. IEEE Trans Image Process. 2018:1–10.
17. Zhang D, Zhou ZH, Chen S. Semi-supervised dimensionality reduction. Siam International Conference on Data Mining, April 26-28, 2007, Minneapolis, Minnesota, USA. DBLP; 2007. p. 11–393.

18. Zhou ZH, Li M. Semi-supervised regression with co-training. International Joint Conference on Artificial Intelligence. Morgan Kaufmann Publishers Inc.; 2005. p. 908–913.
19. Mehrkanoon S, Alzate C, Mall R, et al. Multi-class semi-supervised learning based upon kernel spectral clustering. IEEE Trans Neural Netw Learn Syst. 2015;26(4):720–33.
20. Callut J, Francoisse K, Saerens M, et al. Semi-supervised classification from discriminative random walk. Lect Notes Comput Sci. 2008;5211:162–77.
21. Zhou ZH. Machine learning. Tsinghua University Press; 2016.
22. Castelli V, Cover TM. On the exponential value of labeled samples. Elsevier Science Inc.; 1995.
23. Cozman FG, Cohen I. Unlabeled data can degrade classification performance of generative classifiers. Fifteenth International Florida Artificial Intelligence Society Conference. 2009. p. 327–331.
24. Baudat G, Anouar F. Generalized discriminant analysis using a kernel approach. Neural Comput. 2000;12(10):2385–404.
25. Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. Read Speech Recognit. 1990;77(2):267–96.
26. Vapnik V, Sterin A. On structural risk minimization or overall risk in a problem of pattern recognition. Autom Remote Control. 1977;10(10):1495–503.
27. Zhang M, Pang L. Review of domestic application research of big data mining technology-SVM in credit risk evaluation. 3rd International Seminar on Education Innovation and Economic Management, Penang, Malaysia, 2018. p. 286.
28. Ding SF, Zhu ZB, Zhang XK. An overview on semi-supervised support vector machine. Neural Comput Applic. 2017;28(5):969–78.
29. Zhang H, Cao L, Gao S. A locality correlation preserving support vector machine. Pattern Recogn. 2014;47(9):3168–78.
30. Tao XM, Li Q, Guo WJ. Self-adaptive cost weights-based support vector machine cost-sensitive ensemble for imbalanced data classification. Inf Sci. 2019:487.
31. Tang T, Chen S, Zhao M. Very large-scale data classification based on K-means clustering and multi-kernel SVM. Soft Comput. 2018;1:3793–801.
32. Bruzzone L, Chi M, Marconcini M. A novel transductive SVM for semi-supervised classification of remote-sensing images. IEEE Trans Geosci Remote Sens. 2006;44(11):3363–73.
33. Yu LI, Feng A, Zou SR. TSVM learning algorithm based on improved K-nearest neighbor. Comput Modern. 2018:22–5.
34. Chapelle O, Vapnik V, Bousquet O, et al. Choosing multiple parameters for support vector machines. Mach Learn. 2002;46(1–3):131–59.
35. Blum A, Chawla S. Learning from labeled and unlabeled data using Graph Mincuts. Eighteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc.; 2001. p. 19–26.
36. Szeliski R, Zabih R, Ssharstein D, et al. A comparative study of energy minimization methods for Markov random fields. European Conference on Computer Vision. Berlin: Springer; 2006. p. 16–29.
37. Zhu X, Lafferty J. Harmonic mixtures: combining mixture models and graph-based methods for inductive and scalable semi-supervised learning. Int Conf DBLP. 2005:1052–9.
38. Zhou D, Scholkopf B. Learning from labeled and unlabeled data using random walks. Berlin Heidelberg: Springer; 2004.
39. Belkin M, Niyoge P, Sindhwani V. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. J Mach Learn Res. 2006;7(1):2399–434.
40. Goldberg AB, Li M, Zhu X. Online manifold regularization: a new learning setting and empirical study. European Conference on Machine Learning and Knowledge Discovery in Databases. Verlag: Springer; 2008. p. 393–407.
41. Balcan MF, Blum A, Choi PP, et al. Person identification in webcam images: an application of semi-supervised learning. International Conference on Machine Learning; 2005.
42. Blum A. Combining labeled and unlabeled data with co-training. Conf Comput Learn Theor 1998;92–100.
43. Coldman SA, Zhou Y. Enhancing supervised learning with unlabeled data. 2000. p. 327–334.
44. Wagstaff K, Cardie C, Rogers S, et al. Constrained K-means clustering with background knowledge. Proceedings of 18th International Conference on Machine Learning. Morgan Kaufmann Publishers Inc;2001. p. 577–584.
45. Yang Y, Tan W, Li T, et al. Consensus clustering based on constrained self-organizing map and improved Cop-Kmeans ensemble in intelligent decision support systems. Knowl-Based Syst. 2012;32(32):101–15.
46. Chen ZY, Wang MJ, Hu M, et al. An active semi-supervised clustering algorithm based on seed set and pairwise constraints. J Jilin Univ (Sci Ed). 2017;55(3):664–72.
47. Davidson I, Ravi S. Clustering with constraints: feasibility issues and the k-means algorithm. SDM. 2005;16(95):1147–57.
48. Dan P, Baras D. K-means with large and noisy constraint sets. Mach Learn ECML. 2007;2008:674–82.
49. Wagstaff K, Cardie C. Clustering with instance-level constraints. 17th International Conference on Machine Learning; 2000. p. 1097–1103.
50. Basu S, Banerjee A, Mooney R. Semi-Supervised Clustering by Seeding. 19th International Conference on Machine Learning; 2002. p. 19–26.
51. Zheng L, Li T. Semi-supervised hierarchical clustering. 11th International Conference on Data Mining; 2011. p. 982–991.
52. He P, Xu X, Lu L. Semi-supervised clustering via two-level random walk. J Softw. 2014;25(5):997–1013.
53. Wang L, Bo LF, Jiao LC. Density-sensitive semi-supervised spectral clustering. J Softw. 2007;18(10):2412–22.
54. Shi X, Fan W, Yu P. Efficient semi-supervised spectral co-clustering with constraints. International Conference on Data Mining, 2010.
55. Tang Q, Liao ZG. A semi-supervised clustering method based on affinity propagation algorithm. Electron Inf Warfare Technol. 2017;32(1):8–12.
56. Yang Y, Rutayisire T, Lin C, et al. An improved cop-Kmeans clustering for solving constraint violation based on map reduce framework. Fundam Inf. 2013;126(4):301–18.
57. Sun Y, Xin L, Cheng W. A modified k-means algorithm for clustering problem with balancing constraint. Third International Conference on Measuring Technology and Mechatronics Automation. IEEE; 2011. p. 127–130.
58. Yin SS, Hu SL, Chen SC. Discriminative semi-supervised clustering analysis with pairwise constraint. J Softw. 2008;19(11):2791–802.
59. Wei S, Li Z, Zhang C. Combined constraint-based with metric-based in semi-supervised clustering ensemble. Int J Mach Learn Cybern. 2018;9(7):1085–100.
60. Li CM, Xu SB, Hao ZF. Cross-entropy semi-supervised clustering based on pairwise constraints. Pattern Recogn Artif Intell. 2017;30(7):598–608.
61. Ding S, Xu X, Fan SY, Xue Y. Locally adaptive multiple kernel k-means based on shared nearest neighbors. Soft Comput. 2018;22(14):4573–83.
62. Chai BF, Lu F, Li WB. Semi-supervised Kmeans clustering algorithm based on active learning priors. Comput Appl. 2018;38(11):93–7.
63. Basu S, Bilenko M, Mooney RJ. A probabilistic framework for semi-supervised clustering. 2004;59–68.
64. Ding S, Jia H, Du M, et al. A semi-supervised approximate spectral clustering algorithm based on HMRF model. Inf Sci. 2018;429:215–28.

65. Saha S, Bandyopadhyay S. Semi-GAPS: a semi-supervised clustering method using point symmetry. IOS Press; 2009.
66. Si WW, Qian YT. Semi-supervised clustering based on spectral cluster. Comput Appl. 2005;25(6):1347–9.
67. Bilenko M, Basu S, Mooney R J. Integrating constraints and metric learning in semi-supervised clustering. International Conference. DBLP, Banff, Alberta, Canada, 2004;11.
68. Alok AK, Saha S, Ekbal A. Feature selection and semi-supervised clustering using multi-objective optimization. Springer Plus. 2014;3(1):1–12.
69. Gui J, Wang SL, Lei YK. Multi-step dimensionality reduction and semi-supervised graph-based tumor classification using gene expression data. Artif Intell Med. 2010;50(3):181–91.
70. Saha S, Kaushik K, Alok AK, et al. Multi-objective semi-supervised clustering of tissue samples for cancer diagnosis. Soft Comput. 2016;20(9):3381–92.
71. Yu J, Tao D, Li J, et al. Semantic preserving distance metric learning and applications. Inf Sci. 2014;281:674–86.
72. Shiga M, Mamitsuka H. Efficient semi-supervised learning on locally informative multiple graphs. Pattern Recogn. 2012;45(3): 1035–49.
73. Chen HS. Semi-supervised clustering ensemble for bio-molecular pattern mining. South China University of Technology; 2016.
74. Orozco-Duque A, Bustamante J, Castellanos-Dominguez G. Semi-supervised clustering of fractionated electrograms for electroanatomical atrial mapping. Biomed Eng Online. 2016;15(1):44.
75. Gan H, Fan Y, Luo Z. Local homogeneous consistent safe semi-supervised clustering. Expert Syst Appl. 2017;97:384–93.
76. Syed FH, Tahir MA. Safe semi supervised multi-target regression (MTR-SAFER) for new targets learning. Multimed Tools Appl. 2018;77:29971–87.
77. Wang Y, Chen J. Safe semi-supervised collaborative filtering recommendation algorithm. Comput Eng Appl. 2018;54(8):107–11.
78. Lu Z, Ip HHS. Combining context, consistency, and diversity cues for interactive image categorization. IEEE Trans Multimed. 2010;12(3):194–203.
79. Portela NM, Cavalcanti GDC, Ren TI. Semi-supervised clustering for MR brain image segmentation. Expert Syst Appl. 2014;41(4): 1492–7.
80. Hasnat MA, Alata O, Tremeau A. Joint color-spatial-directional clustering and region merging (JCSD-RM) for unsupervised RGB-D image segmentation. IEEE Trans Pattern Anal Mach Intell. 2016;1–1.
81. An QQ, Zhang F, Li ZX. Research on image segmentation based on machine learning. Automation & Instrumentation. 2018;6:29–31.
82. Li YW. Research on robust segmentation algorithm based on semi-supervised fuzzy clustering. Xi'an: Xi'an University of Posts & Telecommunications; 2018.
83. Yu J, Tao D, Wang M, et al. Learning to rank using user clicks and visual features for image retrieval. IEEE Trans Cybern. 2015;45(4): 767–79.
84. Yu J, Rui Y, Tao D. Click prediction for web image reranking using multimodal sparse coding. IEEE Trans Image Process. 2014;23(5): 2019–32.
85. Cheng XM, Yang QH, Zhai YP, et al. Test case selection technique base on semi-supervised clustering method. Comput Sci. 2018;45(1):249–54.
86. Yu J, Yang X, Gao F. Deep multimodal distance metric learning using click constraints for image ranking. IEEE Trans Cybern. 2016;1–11.
87. Yu Z, Yu J, Xiang C, et al. Beyond bilinear: generalized multimodal factorized high-order pooling for visual question answering. IEEE Trans Neural Netw Learn Syst. 2018;(99):1–13.
88. Yu J, Kuang Z, Zhang B, et al. Leveraging content sensitiveness and user trustworthiness to recommend fine-grained privacy settings for social image sharing. IEEE Trans Inf Forensics Secur. 2018;13(5): 1317–32.
89. Yu J, Zhu C, Zhang J, et al. Spatial pyramid-enhanced NetVLAD with weighted triplet loss for place recognition. IEEE Trans Neural Netw Learn Syst. 2019;(99):1–14.
90. Yu J, Hong C, Rui Y, et al. Multi-task autoencoder model for recovering human poses. IEEE Trans Indust Electron. 2018;(99):1–1.
91. Hong C, Yu J, Tao D, et al. Image-based three-dimensional human pose recovery by multiview locality-sensitive sparse retrieval. IEEE Trans Ind Electron. 2015;62(6):3742–51.
92. Hong C, Yu J, Wan J, et al. Multimodal deep autoencoder for human pose recovery. IEEE Trans Image Process. 2015;24(12):5659–70.
93. Mukkamala S, Sung AH. Feature ranking and selection for intrusion detection systems using support vector machines. Proceed the Second Digital Forensic Research Workshop. 2002;4(3):72.
94. Zhang H, Lu J. Creating ensembles of classifiers via fuzzy clustering and deflection. Fuzzy Set Sys. 2010;161(13):1790–802.
95. Depren O, Topallar M, Anarim E, et al. An intelligent intrusion detection system (IDS) for anomaly and misuse detection in computer networks. Expert Syst Appl. 2005;29(4):713–22.
96. Fiore U, Palmieri F, Castiglione A, et al. Network anomaly detection with the restricted Boltzmann machine. Neuro Comput. 2013;122:13–23.
97. Liang C, Li CH. Novel intrusion detection method based on semi-supervised clustering. Comput Sci. 2016;43(5):87–90.
98. Peng TL, Zhang WJ, Lan JL, et al. Micro video annotation method based on semi-supervised clustering. Appl Res Comput. 2016;33(3):948–52.
99. Zhong S. Semi-supervised model-based document clustering: a comparative study. Mach Learn. 2006;65(1):3–29.