Information Sciences 523 (2020) 63–76

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Three-way decisions based blocking reduction models in hierarchical classification

Wen Shen, Zhihua Wei*, Qianwen Li, Hongyun Zhang, Duoqian Miao

Department of Computer Science and Technology, Tongji University, Shanghai, China

ARTICLE INFO

Article history:
Received 29 March 2019
Revised 6 February 2020
Accepted 8 February 2020
Available online 10 February 2020

Keywords:
Hierarchical classification
Blocking reduction
Three-way decisions
Category relation mining
Topic model

ABSTRACT

Hierarchical classification (HC) is effective when categories are organized hierarchically. However, the blocking problem makes the effect of hierarchical classification greatly reduced. Blocking means that samples are easily getting misclassified in high-level classifiers so that the samples are blocked at the high-level of the hierarchy. This issue is caused by the inconsistency between the artificially defined hierarchy and the actual hierarchy of the raw data. Another issue is that it is flippant to strictly process data following the hierarchy. Therefore, special treatment is required for some uncertain data. To address the first issue, we learn category relationships and modify the hierarchy. To address the second issue, we introduce three-way decisions (3WD) to targetedly deal with the ambiguous data. We extend original studies and propose two HC models based on 3WD, collectively referred to as TriHC, for carefully modifying the hierarchy to alleviate the blocking problem. The proposed TriHC model learns new category hierarchies by the following three steps: (1) mining category relations; (2) modifying category hierarchies according to the latent category relations; and (3) using 3WD to divide observed objects into three regions: positive region, boundary region, and negative region, and making decisions based on different strategies. Specifically, based on different category relation mining methods, there are two versions of TriHC, cross-level blocking priori knowledge based TriHC (CLPK-TriHC) and expert classifier based TriHC (EC-TriHC). The CLPK-TriHC model defines a cross-level blocking distribution matrix to mine the category relations between the higher and lower levels. To better exploit category hierarchical relations, the EC-TriHC model builds expert classifiers using topic model to learn latent category topics. Experimental results validate that the proposed methods can simultaneously reduce the blocking and improve the classification accuracy.

© 2020 Elsevier Inc. All rights reserved.

1. Introduction

Hierarchical classification (HC) is an effective method to solve multiclass classification problems, especially when categories are organized hierarchically. Many important real-world classification problems naturally can be treated as HC problems, such as text categorization [4,14], protein function prediction [27,33], image classification [2,6] etc. Using HC methods, a large-scale classification task can be divided into several small-scale tasks, so as to reduce the difficulty of classification. Usually, HC method uses a top-down level-based strategy, in which the class hierarchy typically stored as a tree. Such a

* Corresponding author.
E-mail address: zhihua_wei@tongji.edu.cn (Z. Wei).

https://doi.org/10.1016/j.ins.2020.02.020
0020-0255/© 2020 Elsevier Inc. All rights reserved.

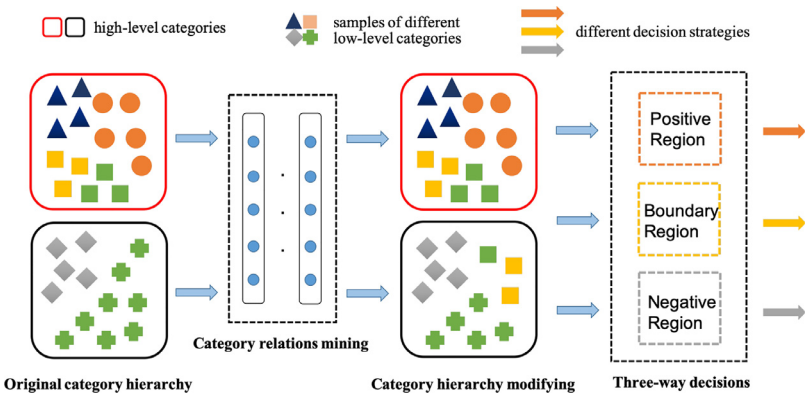

Fig. 1. The basic idea of the TriHC model. Each high-level category includes several low-level categories represented by different shapes. Each low-level category has a different color that represents the difference from other categories. There may be similarities between low-level categories that belong to different high-level categories (different shapes, but the same color). That is why blocking occurs. The TriHC model contains 3 steps: (1) mining category relations; (2) modifying category hierarchy according to the latent category relations; and (3) using 3WD to divide the observed objects into three regions: positive region, boundary region and negative region, and making decisions based on different strategies.

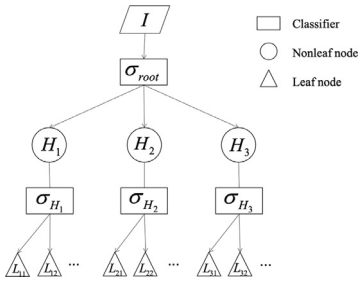

Fig. 2. Illustration of hierarchical classification. Take a two-level hierarchy as an example.

strategy is simple, intuitive, and interpretable. However, due to the complexity of category relationships, samples are easily misclassified in higher-level classifiers, *i.e.* blocking. It is clear that blocking is one of reasons that HC performs worse than the traditional flat classification.

Most existing blocking reduction strategies in HC can be divided into two types depending on whether changing the category hierarchy: methods that will change the hierarchy and methods that will not change the hierarchy. For the first type, there are Restricted Voting method (RVM) [37], Priori Knowledge Based Hierarchical Classification method (PKHC) [40], data-driven hierarchical structure modification approach (Global-INF) [25], and etc. These strategies give the blocked samples another chance to return to the correct categories by changing the original hierarchy. For the second type, there are Multiplicative method (MM) [7], Extended Multiplicative method (EMM) [37], Threshold Reduction Method (TRM) [37], and etc. These strategies work on probabilities of base classifiers, taking into account the horizontal or vertical base classifiers' results and helping blocked samples return to the correct category. In fact, methods of the second type have little improvement in the blocking problem. Since blocking is mainly caused by the complexity of category relationships, methods which change the hierarchy are more suitable for the blocking reduction problems. The two blocking reduction strategies proposed in this paper also belong to the first type.

Although there are many methods of the first type have been presented in the past, less of them considers the inconsistency between the artificially defined hierarchy and the actual hierarchy of the data. Some of existing methods take into account the relationship between base classifiers, but overrefine the hierarchical topological structure, which bring new blocking. Take PKHC method as an example, let us consider the most extreme case, high-level category H_1 has two subcategories L_{11} and L_{12}, and all samples of L_{11} are misclassified as H_2 at the high-level of the hierarchy, while all samples of L_{12} are classified correcetly as H_1, see Fig. 2. PKHC thinks that samples of H_1 are easily to be misclassified as H_2, thus it will add both paths from H_2 to L_{11} and L_{12}, it increases training costs and increases the uncertainty of classifier σ_{H_2}. Actually, only the path from H_2 to L_{11} need to be added. See Section 3 for more details.

To this end, in this study, we propose a new method for blocking reduction problems, the illustration of this method see Fig. 1. In a nutshell, the core of our method is to reconstruct the topology of the hierarchical classification model by learning category relations from the original category hierarchy, then using 3WD method to divide observed objects into three regions: positive region, boundary region and negative region, specially paying attention to the boundary region. We utilize two different methods to mine category relations. The first method takes into account the cross-level blocking priori knowledge based on the method proposed in paper [40], which only considers the relationships between high-level categories

but neglects the relationships between high- and low-level categories. However, this method cares only the one-to-one relationships between categories, the second method applies the topic model based label grouping algorithm proposed in paper [39] to learn the many-to-many relationships amongst categories.

The proposed TriHC method can effectively reduce the blocking error and has less impact on categories that are not prone to blocking. Contributions of this study could be summarized as follows.

- We propose two hierarchical models to modify the category hierarchy to deal with blocking problems caused by the inconsistency between the artificially defined hierarchy and the actual hierarchy of the data.
- We explore three-way decisions to alleviate the classification errors caused by data uncertainty.

Our experiments on the challenging DeepFashion dataset [23] and Stanford Dogs dataset [13] demonstrate the effectiveness of the proposed models. We validate that our TriHC models can significantly reduce the blocking problem when compared with several previous HC models. The classification accuracy can even be higher than that of the well trained deep convolutional neural network model.

2. Related work

2.1. Hierarchical classification

In recent years, the development of deep learning has promoted many state-of-the-art models for the classification task [9,15,35,38]. But this kind of flat classification (FC) model wastes the hierarchical relationship between categories. Also when the number of categories is huge, the training of FC models is difficult. The hierarchical classification methods deal with multi-classification problems by dividing a large-scale classification task into several small-scale tasks [8,12,26]. A hierarchical classifier consists of two parts: category hierarchy and classifier. Depending on the storage structure used, the category hierarchy can be divided into two types: tree-structured category hierarchies and DAG-structured (Directed Acyclic Graph-structured) class hierarchies. The storage structure will influence the degree of difficulty of the underlying hierarchical classification problems, and most HC models use tree structure because it is easier than DAG structure [34]. Error-correcting output codes (ECOC) [5] is another extensively used framework to decompose a multi-class problem into several sub-problems, which takes binary tree structure. ECOC and its variants have achieved superior performance in various multi-classification tasks [1,21,22,28,50]. This work also focuses on HC with tree-structured hierarchy. The top-down level-based strategy is easy to cause blocking problem, because the classification error of high-level nodes can not be corrected by the low-level nodes. Many blocking reduction methods have been proposed, for example, RVM [37], MM [7], PKHC [40], Global-INF [25] etc.

Though these methods have achieved good performance, they have not considered the unreliability of the original category relationships, which is inconsistent with the relationships learned from the raw data. PKHC [40] considers the blocking between the high-level categories and rebuilds the hierarchy to reduce the blocking error. However, it overrefines the hierarchical structure and brings new blocking. Liu et al. [22] propose a joint binary classifier learning (JCL) method, which takes into account the inherent relationships among the base classifiers in ECOC-based methods. The JCL method attempts to find out the most valuable classifiers so that alleviate the negative impacts of such relationships on the learning performance of base classifiers. In this work, we also consider the relationships among the classifiers. We extend the original works and present a new HC method based on category relationships mining.

2.2. Three-way decisions

The three-way decision (3WD) notion was first proposed to explain the three regions of probabilistic rough sets [41], and has been constructed to a general theory by embracing ideas from interval sets, rough sets, decision-theoretic rough sets, fuzzy sets, and shadowed sets [44]. The 3WD theory is formulated based on the notions of acceptance, rejection and non-commitment. The 3WD introduces a third option and then collects more information to make a more confident decision, which alleviates the uncertainty of the two-way decision making [20,30,42,43]. There are many applications use 3WD to handle the situation of uncertion decision making. For instance, three-way spam filtering systems [19,49], medical decision making [24,29], classification and clustering [16,46,48], 3WD itself is constantly evolving. Traditional 3WD only considers a static and one-step decision strategy [18]. To consider the cost of obtaining required evidence or information, 3WD is further extended into sequential three-way decision-making [31,45,47]. Sequential 3WD has good performance in the field of face recognition. Li et al. [17] proposed a cost-sensitive sequential 3WD method for face recognition, which seeks a decision which minimizes the misclassification cost rather than misclassification error in each decision step. On this basis, they introduced a deep neural network to extract sequential granular features [18]. On the one hand, DNN improves the capacity for representation of the original sequential 3WD method. On the other hand, the sequential 3WD method reduces the time cost of DNN extracting features. Determining the scope of the three decision regions is an important part of the 3WD theory. Many approaches divide the three regions based on prediction probability. Bayesian Factor [36] is a widely-used way to define the three decision regions, which does not require information about the prior and posterior probabilities. The reject option [11] permits a special treatment of the instances close to the decision boundary. Recently proposed distance factor [31] is a more robust way to define the three decision regions.

Numerous studies have shown that 3WD is a rigorous and comprehensive theory and have made outstanding contributions in the fields of uncertain reasoning and decision making. In this paper, we use three-way decisions (3WD) to targetedly deal with the ambiguous data and we define the three decision regions based on category relationships.

3. Three-way decisions based blocking reduction models

In this section, we first introduce the notation used in this paper. Then, we present two blocking reduction models based on 3WD.

3.1. Notation

Let c_i be a nonleaf category and σ_{c_i} be the classifier of it. Let $w(\sigma_{c_i})$ be the work domain of classifier σ_{c_i}. Suppose there are N_i sub categories of c_i, then $w(\sigma_{c_i}) = \{c_{i1}, c_{i2}, \ldots, c_{iN_i}\}$. Let c_j be another nonleaf category that at the same level as c_i. Assume that the input I has ground truth label c_{i1} but is classified as c_j, then we say I is blocked.

3.2. Cross-level blocking priori knowledge based TriHC model

3.2.1. Blocking priori knowledge based hierarchical classification model

Unlike most previous HC methods which blindly accept the original hierarchical structures, PKHC boldly breaks the original hierarchy and rebuilds the hierarchy by analyzing the distribution of blocked samples. PKHC thinks the reason for the blocking is essentially the uncoordination of the prior knowledge and actual data, that is the prior hierarchical structure of experts can not correctly reflect the hierarchy of actual data. One way to solve this problem is to open the amended paths for unreliable classification results at the high-level. As Fig. 3 shows, the samples of category c_m are easily to be misclassified to category c_n, these blocked samples are absolutely impossible to be classified as the right category in the low-level. Thus, PKHC adds paths from high-level category c_n to low-level sub categories of category c_m. This means that even if the samples of category c_m are misclassified to category c_n by the high-level classifier σ_{root}, they still have the opportunity to be classified correctly by the low-level classifier σ_{c_n}.

However, PKHC only considers the blocking between the same level categories and overrefines the hierarchical topological structure, which brings new blocking, as aforementioned in Section 1. To this end, we propose an improved version of PKHC, cross-level blocking priori knowledge based TriHC model (CLPK-TriHC). CLPK-TriHC analyzes the cross-level blocking distribution of the samples. There are two steps, the first step is cross-level blocking priori knowledge learning and the second step is 3WD based category hierarchy rebuilding.

3.2.2. Cross-level blocking priori knowledge learning

We first define the cross-level blocking distribution matrix (CLBDM), which is an $l \times h$ two-dimensional matrix, where l is the category number of low-level and h is the category number of high-level (see Fig. 3). Suppose the sizes of work domains $w(\sigma_{c_1})$, ..., $w(\sigma_{c_m})$, $w(\sigma_{c_n})$, ..., and $w(\sigma_{c_h})$ are respectively N_1, ..., N_m, N_n, ..., and N_h. Then $l = N_1 + \ldots + N_m + N_n + \ldots + N_h$. CLBDM indicates the blocking distribution across high- and low-level categories, where $b^{mi}_{(p)(n)}$ $(p = 1, \ldots, l,\ n = 1, \ldots, h)$ represents the number of misclassifed samples, whose ground truth label are c_{mi} $(i = 1, \ldots, N_m)$ but being misclassified as c_n at the high level of the hierarchy. It is worth noting that $b^{mi}_{(p)(n)} = 0$ when $m = n$. Each row of the matrix reflects the cases where samples of the low-level category being blocked to high-level categories, while each column of the matrix reflects

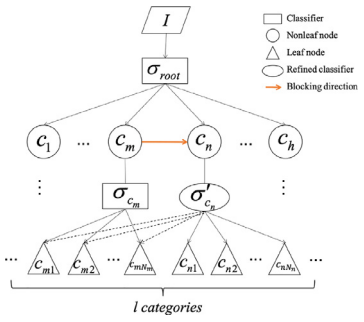

Fig. 3. Illustration of the PKHC algorithm. The solid line indicates the prior hierarchical relationship, and the dotted line indicates the amended hierarchical relationship.

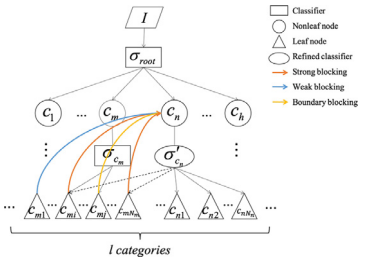

Fig. 4. Illustration of the CLPK-TriHC model. The solid line indicates the prior hierarchical relationship, and the dotted line indicates the amended hierarchical relationship.

the source of the blocked samples received by the high-level category.

$$CLBDM = \begin{pmatrix} 0 & \cdots & b^{11}_{(1)(m)} & \cdots & b^{11}_{(1)(n)} & \cdots & b^{11}_{(1)(h)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \\ b^{mi}_{(p)(1)} & \cdots & 0 & \cdots & b^{mi}_{(p)(n)} & \cdots & b^{mi}_{(p)(h)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \\ b^{hN_h}_{(l)(1)} & \cdots & b^{hN_h}_{(l)(m)} & \cdots & b^{hN_h}_{(l)(n)} & \cdots & 0 \end{pmatrix} \tag{1}$$

We use blocking value (Bv) to measure the degree of blocking, as in (2), where $f(c_{mi}, c_n)$ measures the proportion of samples blocked to category c_n in all samples of category c_{mi}, as in (3), and $t(c_{mi}, c_n)$ measures the proportion of samples from category c_{mi} in all samples classified as category c_n, as in (4).

$$Bv(c_{mi}, c_n) = \begin{cases} \sqrt{\log\left(\dfrac{2f(c_{mi}, c_n)t(c_{mi}, c_n)}{f(c_{mi}, c_n) + t(c_{mi}, c_n)} + 1\right)}, & f(c_{mi}, c_n) > 0 \wedge t(c_{mi}, c_n) > 0 \\ 0, & f(c_{mi}, c_n) = 0 \vee t(c_{mi}, c_n) = 0 \end{cases} \tag{2}$$

$$f(c_{mi}, c_n) = \frac{b^{mi}_{(p)(n)}}{\sum CLBDM_{p,:}}, \tag{3}$$

$$t(c_{mi}, c_n) = \frac{b^{mi}_{(p)(n)}}{\sum CLBDM_{:,n}}, \tag{4}$$

where $\Sigma CLBDM_{p,:}$ and $\Sigma CLBDM_{:,n}$ respectively represent the sum of the elements of the pth row and nth column of the CLBDM. (c_{mi}, c_n) indicates a blocking pair, and the collection of (c_{mi}, c_n) is blocking set (*BS*). According to the blocking degree of the blocking pairs, the *BS* can be divided into three subsets: strong blocking set (*SBS*), weak blocking set (*WBS*) and boundary blocking set (*BBS*). The three subsets are pair-wise disjoint.

3.2.3. Three-way decision based category hierarchy modification

After learning the cross-level blocking priori knowledge, we get the blocking directions from low-level categories to high-level categories. According to the blocking directions, we modify the original hierarchical structure. As in Fig. 4, (c_{mi}, c_n) and (c_{mN_m}, c_n) are strong blocking pairs, which should be added to *SBS*. (c_{m1}, c_n) is a weak blocking pair, which should be added to *WBS*. (c_{mj}, c_n) is a boundary blocking pair, which should be added to *BBS*. We define two parameters α and β as thresholds for dividing strong or weak blocking pairs, see (5).

$$\begin{aligned} SBS_{(\alpha,\beta)}(Bv) &= \{(c_{mi}, c_n) \in BS | Bv(c_{mi}, c_n) \geq \alpha\}, \\ WBS_{(\alpha,\beta)}(Bv) &= \{(c_{mi}, c_n) \in BS | Bv(c_{mi}, c_n) \leq \beta\}, \\ BBS_{(\alpha,\beta)}(Bv) &= \{(c_{mi}, c_n) \in BS | \beta < Bv(c_{mi}, c_n) < \alpha\}. \end{aligned} \tag{5}$$

For each blocking pair (c_{mi}, c_n) in *SBS*, we add the path from high-level category c_n to low-level category c_{mi}, in other words, add c_{mi} to $w(\sigma_{c_n})$, the work domain of classifier σ_{c_n}. During the test procedure, even the observed objects of category c_{mi} is blocked to c_n, they still have a chance to be correctly classified as c_{mi}. For the blocking pair (c_{mj}, c_n) in *BBS*, c_{mj} and c_n are not strongly related, thus more information should be provided to process the observed objects. Our intuition is that, for a boundary blocking pair (c_{mj}, c_n) in *BBS*, if the observed object I is classified as c_n at the high-level of the hierarchy but I has the smaller distance from the c_{mj} than c_n, we think that blocking may have happened. In light of this, we define a distance function J to measure the distance between sample I and category c.

$$\mathrm{J}(I, c) = \|I - \mu(c)\|. \tag{6}$$

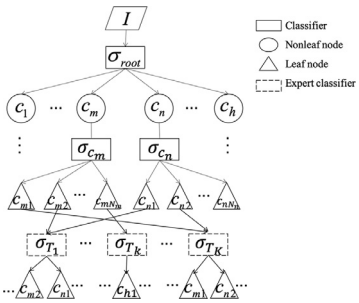

Fig. 5. Illustration of the EC-TriHC model.

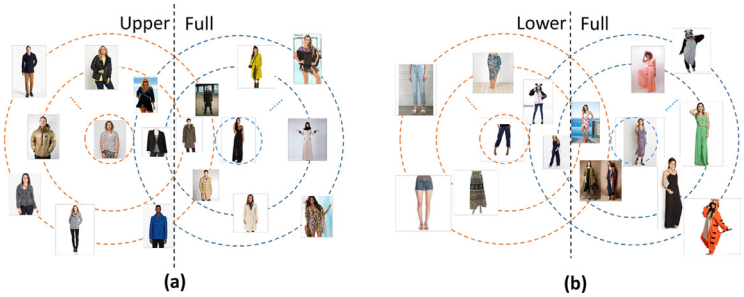

Fig. 6. Illustration of cross-class sample similarity in DeepFashion dataset. The "*spread out*" phenomenon between "*upper cloth*" and "*full cloth*" is shown in (a) and the "*spread out*" phenomenon between "*lower cloth*" and "*full cloth*" is shown in (b). Samples farther away from the center, the easier to be misclassified as other category.

When the input is an image, $\mu(c)$ returns a matrix, where each value is the mean pixel value of all the samples of category c. During the hierarchical classification process, if the observed object I is classified as c_n but $J(I, c_{mj}) > J(I, c_n)$, then we run the parent classifier of c_{mj}, that is σ_{c_m}, to reclassify it.

CLPK-TriHC can effectively reduce the blocking problem. First, compared with SHC, CLPK-TriHC reduces the inconsistency between the original hierarchical structure and the hierarchy of the actual data and helps the blocked samples return to correct categories. Second, compared with PKHC, CLPK-TriCH considers the cross-level blocking instead of the same-level blocking and only adds path between those strong blocking pairs, which reduces the complexity of the high-level base classifiers. Last, CLPK-TriHC can alleviate the uncertainty of the boundary blocking pairs.

3.3. Expert classifier based TriHC model

CLPK-TriHC only considers the hierarchical relationships across high- and low-level categories, which are one-to-one relationships. In addition, when CLPK-TriHC adds low-level category c_{mi} to high-level category c_n, it assumes that all the low-level categories of c_n are similar to c_{mi}. In fact, it may be that only one or several low-level categories of c_n are similar to c_{mi}, which causes c_{mi} easily to be blocked to c_n. Therefore, we pursue a category regrouping method that can mine the hidden relationships between multiple low-level categories at the same time.

In light of this, we propose another variant of TriHC method named expert classifier based TriHC model (EC-TriHC). EC-TriHC makes an attempt to give samples a chance to make the right prediction when the samples are misclassified in high-level classifiers by using expert classifiers to make a secondary classification. Inspired by papers [32,39], we use topic model to learn the latent relationships between multiple low-level categories and build the expert classifiers for those samples suspected to be blocked. Fig. 5 shows the EC-TriHC process.

The EC-TriHC model mainly contains three parts: topic learning based category relation mining, expert classifier building and voting.

3.3.1. Topic learning based category relation mining

A good dataset requires that the samples with the same labels be as close as possible, while the samples with different labels be as far apart as possible. But this is actually difficult to achieve, because the data in reality is diverse. Samples that are divided into the same category according to certain characteristics may be very different in other characteristics. In other words, it is difficult to cover all characteristics with a single label. This phenomenon is also called "*spread out*" in other studies [10]. The "*spread out*" phenomenon may cause the intra-class distance to be greater than the inter-class distance, see Fig. 6. Our intuition is that first group together the categories which "*spread out*" mutually to build expert classifiers, then vote on several classifiers' results to get the final result.

A topic model based label grouping (TMLG) algorithm in paper [39] was proposed to learn the relationship between different categories and mine the latent topics. A topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents. The "topics" here are clusters of similar words. In light of this, the key elements of a topic model are: documents, words and topics. Suppose $D = \{d_1, d_2, \ldots, d_n\}$ denotes the input of a topic model, where d_i indicates the term frequency statistics of the ith document. In this work, the available data is the predection scores S returned at softmax layer of a convolutional neural network (CNN), $S = \{s_1, s_2, \ldots, s_n\}$, where $s_i = \{s_{i1}, s_{i2}, \ldots, s_{iC}\}$ indicates the predicted classification scores of the ith example. By taking the score times a constant A as the term frequency ("tf"), see (7), TMLG successfully introduced topic models to mine label subsets.

$$tf_{ic} = round(A * s_{ic}) + B * \mathbb{1}[c == c_{gt}], \tag{7}$$

where $\mathbb{1}[\cdot]$ is the indicator function whose value is 1 when the statement is true and 0 otherwise, and c_{gt} indicates the ground truth label of the example. B is to prevent when the probability of the real category is really small, which leads to the tf_{ic} being 0.

After mapping the prediction results to documents, we use Latent Dirichlet Allocation (LDA) [3] topic model to learn relationships between low-level categories from different high-level categories and get several topics to build expert classifiers. The output of the LDA topic model contains a topic-category relationship matrix TC. Each row of TC can be viewed as the category distribution of a certain topic. The procedure of TMLG see Algorithm 1.

Algorithm 1 TMLG algorithm.

Require: Prediction scores of the original SHC model, $S = \{s_1, s_2, \ldots, s_n\}$, where $s_i = \{s_{i1}, s_{i2}, \ldots, s_{iC}\}$;
Ensure: K topic subsets, $T = \{T_1, \ldots, T_k, \ldots, T_K\}$;
1: Take each $s_i \in S$ as a "document" and compute the tf value for each score in s_i. Then obtain all the "documents" frequency term statistics, TF;
2: Training an LDA topic model with input TF and obtain the topic-category relationship matrix TC;
3: For all the low-level categories, assign category c with $p_c^k > th$ to each topic T_k;
4: **return** T;

$$TC = \begin{pmatrix} \cdots & p^1_{c_{mi}} & \cdots & p^1_{c_{nj}} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \cdots & p^k_{c_{mi}} & \cdots & p^k_{c_{nj}} & \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \cdots & p^K_{c_{mi}} & \cdots & p^K_{c_{nj}} & \cdots \end{pmatrix} \tag{8}$$

We only keep topics that contain low-level categories from different high-level categories because this work only focus on blocking problem. By topic learning, we obtain the topic subsets $T = \{T_1, \ldots, T_k, \ldots, T_K\}$.

It can be known from the Algorithm 1 that samples belonging to the same topic are easily to be confused by the SHC. Therefore, for these examples, directly accepting the decision of SHC is unreliable and more information is required to make the final decision. Our intuition is that for an observed object x in dataset X and the predicted category is $\hat{c}$, we define two regions: acceptance region (ACC) and boundary region (BND). It should be noted that there is no rejection region here, in other words the rejection region is an empty set.

$$BND = \{BND_1, \ldots, BND_k, \ldots, BND_K\}, \tag{9}$$

where

$$BND_k = \{x | \hat{c} \in T_k\}. \tag{10}$$

Then,

$$ACC = X/(BND_1 \cup \cdots \cup BND_k \cup \cdots \cup BND_K). \tag{11}$$

For the observed object $x \in ACC$, directly accept the prediction results of SHC. While for the observed object $x \in BND_k$, more information should be provided for making the final decision. In light of this, we train an expert classifier for each topic subset and make a secondary classification for x.

3.3.2. Expert classifier building

In this paper, an expert classifier is trained for distinguishing between certain similar categories, which are easily to be blocked. After topic learning, we obtain the category subsets that the expert classifiers are trained on. The expert classifiers can be trained entirely independently. At test time, if the predicted category from the HC model belongs to a certain topic subset T_k, it indicates that blocking may occur and the relevant expert classifier σ_{T_k} need to be run.

3.3.3. Voting

As aforementioned, for the observed object $x \in BND_k$, in addition to the base classifier, there are expert classifiers that classify it, so it will have two or more classification results (x may be too uncertain so that more than one expert is required). To get a unified result, EC-TriHC applies the idea of voting, which means that combine decisions from several independent classifiers to classify a sample.

In this paper, we use information gain (IG) as the basis for voting. Our intuition is that the larger the IG, the more information the classifier learns and the more reliable the classification result is. Suppose the predicted result of a classifier is $p = \{p_1, \ldots, p_i, \ldots, p_C\}$, where p_i is a score returned at softmax layer of a CNN, it indicates the probability of being the ith category. C is the number of categories. The IG is computed as:

$$IG = H_r - H_p, \tag{12}$$

where H_r is the entropy of the randomly classified result and H_p is the entropy of the prediction result P. The H_r is calculated as.

$$H_r = -\sum_{i=1}^{C} \frac{1}{C} \log \frac{1}{C} = \log C, \tag{13}$$

where $\frac{1}{C}$ indicates the randomly predicted probability of all the categories. The H_p is calculated as:

$$H_p = -\sum_{i=1}^{C} p_i \log p_i, \tag{14}$$

3.4. Evaluation metric

In this paper, we take 3 metrics to evaluate the models, including accuracy, $F1$ measure, and blocking correction rate.

3.4.1. Accuracy

Accuracy is defined as the ratio of the number of samples correctly classified by the classifier to the total number of samples for a given set of test data. In this paper, we take top-1 accuracy and top-3 accuracy to evaluate the models, where the top-3 accuracy is the fraction of test images for which the correct label is among the 3 labels considered most probable by the model [15].

3.4.2. F1-measure

As in (15), the $F1$-Measure is

$$F1 = \frac{2PR}{P + R} \tag{15}$$

where P is precision rate and R is recall rate. In this paper, we also take the Macro-F1 score to evaluate the models. The Macro-F1 score calculates metrics for each label, and find their unweighted mean.

3.4.3. Blocking correction rate

To evaluate the blocking reduction capability of the HC models, we define the blocking correction rate (BC):

$$BC = \frac{N_{bc}}{N_b}, \tag{16}$$

where N_b is the number of blocking samples, that is, the samples misclassified by the high-level classifiers. N_{bc} is the number of corrected blocking samples, that is, the samples misclassified by the high-level classifiers but corrected by the low-level classifiers.

4. Experiments

In this section, we evaluate the CLPK-TriHC and EC-TriHC on fashion image classification task. Our experiments on the DeepFashion dataset [23] and the Stanford Dogs dataset [13] demonstrate that the proposed methods perform better than several previous HC methods, and even surpass the well trained convolutional neural network (CNN) in some cases.

4.1. Experimental datasets

In this study, we did experiments on two benchmark datasets, *i.e.* the DeepFashion dataset [23] and the Stanford Dogs dataset [13]. In the DeepFashion dataset, each image is labeled with one of 50 categories. The reason we choose this dataset is that, in addition to the 50 labels, it also provides two sets of labels with different granularities, a coarse-grained one and a fine-grained one. The coarse-grained one contains 3 labels, "*upper cloth*", "*lower cloth*", and "*full cloth*". The fine-grained one contains 5620 labels. There is a hierarchical relationship between these three label sets: high-level labels (the 3 labels),

middle-level labels (the 50 labels), and the low-level labels (the 5620 labels). Since there are too many labels of fine-grained set, which will take a lot of time and space costs, thus, in this work, we only take the high-level labels and the middle-level labels, the illustration of the hierarchy see Fig. 7. Where σ_{root} is the classifier of "root" node, its work domain $w(\sigma_{root}) = \{U, L, F\}$, where "*U*" represents "*upper cloth*", "*L*" represents "*lower cloth*", and "*F*" represents "*full cloth*". In the training and testing process, 4 of the 50 labels are merged into one label. Therefore, there are totally 46 labels, among which, 20 labels belong to "*upper cloth*", 16 labels belong to "*lower cloth*", and 10 labels belong to "*full cloth*". Besides, we used a subset of the Stanford Dogs dataset as the second experimental dataset in this study. This subset also provides two sets of labels with different granularities, a coarse-grained one that contains 4 labels, and the fine-grained one that contains 15 labels, *I.e.* {*English setter, Gordon setter, Irish setter*,}, {*Siberian husky, Malamute*}, {*Great dane, Boxer, Bull mastiff, Eskimo dog, Saint bernard*}, and {*Golden retriever, Chesapeake bay retriever, Curly coated retriever, Flat coated retriever, Labrador retriever*}.

As mentioned above, the essential reason for blocking is the "*spread out*" phenomenon. The "*spread out*" phenomenon may cause the intra-class distance to be greater than the inter-class distance, as shown in Fig. 6. Representative images are in the center of the category, which can be easily distinguished with other categories. And images distant to the center are less representative. Take the "*upper cloth*" and the "*full cloth*" in the DeepFashion dataset as examples, the typical "*full cloth*" should be long and can cover most of the lower body, however the non-representative "*full cloth*" may be short and can not contain the lower body, which will confuse the classifier so that lead to the blocking.

4.2. Experimental results on the DeepFashion dataset

4.2.1. Comparison methods

In order to verify the superiority of the proposed methods, we do experiments with several other methods: flat classification method (FC), standard hierarchical classification method (SHC), multiplicative method based HC (MMHC), and PKHC. Note that although the JCL method [22] also takes into account the relationships among the base classifiers, we did not compare the proposed methods with JCL, as well as other ECOC-based methods. The reason is ECOC-based methods take binary tree structured hierarchies and all the base classifiers are binary classifiers, while this work only focuses on HC models with multi-class base classifiers.

FC: We train the FC model by fine-tuning a pre-trained CNN (GoogLeNet on ImageNet LSVRC). The top-1 accuracy is 61.02%, the top-3 accuracy is 80.31%, the F1 score is 60.23%, and the Macro-F1 score is 33.56%. Note that all the base classifiers used in HC, including SHC, MMHC, PKHC, and the proposed TriHC models, are trained by fine-tuning the pre-trained CNN mentioned above.

SHC: We take the structure in Fig. 7 to build the SHC model. The top-1 accuracy is 60.04%, the top-3 accuracy is 75.77%, the F1 score is 59.62%, and the Macro-F1 score is 32.38%. For the convenience of description, we think the BC rate of SHC as 0%.

MMHC: We take the same structure in Fig. 7 to build the MMHC. The top-1 accuracy is 60.21%, the top-3 accuracy is 78.92%, the F1 score is 59.76%, and the Macro-F1 score is 32.54%. The BC rate is 1.45%, compared with SHC, MMHC reduces the blocking slightly.

PKHC: To demonstrate the roles of "cross-level" in CLPK-TriHC, we implement the PKHC algorithm in paper [40] as another baseline. As defined in [40], the threshold θ directly determines the final structure, here we shows all the 7 cases, see Fig. 8. We use 5-fold cross-validation on an independent validation set to select the case which performs the best. Case (e) is selected after the cross-validation procedure. Then we test the performance of PKHC on an independent test set. It achieves the top-1 accuracy of 60.72%, top-3 accuracy of 78.19%, F1-score of 60.52%, and Macro-F1 score of 31.60%. The BC rate is 3.87%, 2.42% higher than that of MMHC. PKHC does perform better than MMHC in blocking reducing.

4.2.2. Results of CLPK-TriHC

We compute the *CLBDM* based on an independent validation set and then calculate the blocking values. To select the suitable parameters α and β, we use 5-fold cross-validation on an independent validation set. We test the model performance under $(\alpha, \beta) = \{(0.2,0.1), (0.3,0.2), (0.4,0.3), (0.5,0.4), (0.6,0.5), (0.7,0.6)\}$. CLPK-TriHC with $(\alpha, \beta) = (0.7, 0.6)$ performs the best all the time, thus $(\alpha, \beta) = (0.7, 0.6)$ is selected after the cross-validation procedure.

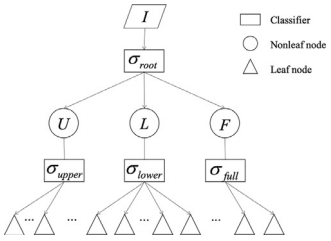

Fig. 7. Illustration of DeepFashion dataset hierarchy. Only the first- and second-level of DeepFashion are utilized in this work.

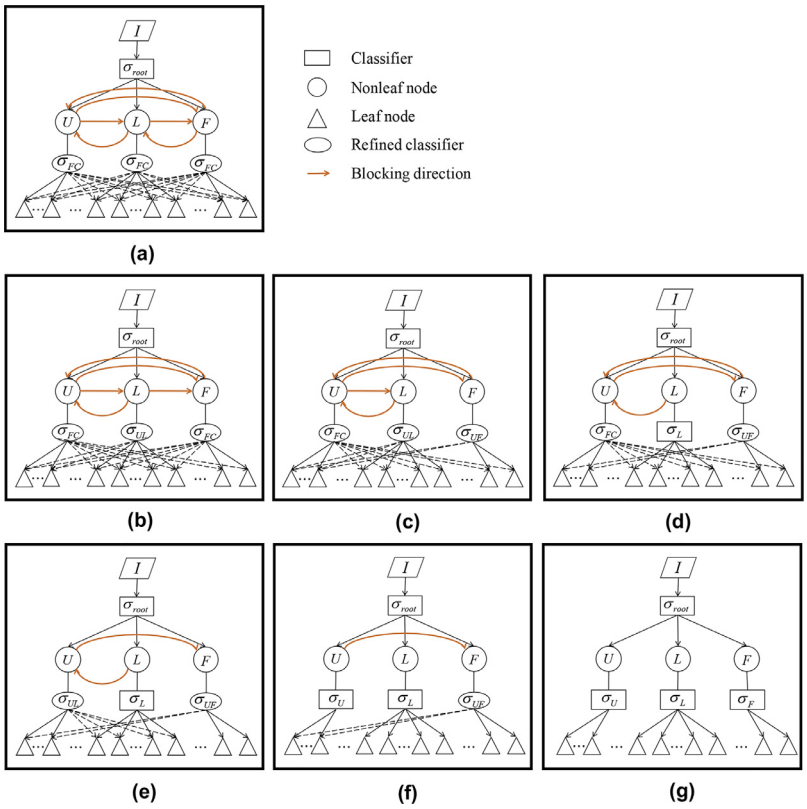

Fig. 8. Illustration of the modified hierarchies under different parameters: (a): $\theta \in [0.00, 0.44]$, (b): $\theta \in (0.44, 0.46]$, (c): $\theta \in (0.46, 0.67]$, (d): $\theta \in (0.67, 0.68]$, (e): $\theta \in (0.68, 0.69]$, (f): $\theta \in (0.69, 0.70]$, and (g): $\theta \in (0.70, 1.00]$, where (a) is equivalent to a flat classifier and (g) is equivalent to the SHC. The solid line indicates the prior hierarchical relationship, and the dotted line indicates the amended hierarchical relationship. σ_{FC} stands for the flat classifier of all the low-level categories, σ_{UL} stands for the merged classifier of upper and lower cloth categories, and σ_{UF} stands for the merged classifier of upper and full cloth categories.

We then test the performance of CLPK-TriHC on the same test set as the PKHC uses. CLPK-TriHC achieves the top-1 accuracy of 61.42%, top-3 accuracy of 80.67%, F1-score of 60.90%, and Macro-F1 score of 32.98%. The BC rate is 12.32%, 10.87% higher than that of MMHC, and 8.45% higher than that of PKHC. Also worth noting is that CLPK-TriHC performs better in Macro-F1 than PKHC. The reason for this may be that PKHC crudely merged all the lower-level sub-categories of similar high-level categories, causing those categories that are not easily blocked to be involuntarily involved in a new "war", which will result in a reduction in the accuracy of these categories. Therefore, although the overall accuracy rate increases, the macro-F1 score decreases. In a nutshell, PKHC sacrifices the accuracies of many other categories to increase the accuracies of several certain categories, while CLPK-TriHC is more fair-minded.

In order to verify the need for 3WD, we do experiment on a binary-decision variant of CLPK-TriHC, where $\alpha = \beta$. When $\alpha = \beta = 0.7$, the top-1 accuracy is 61.23%, and the BC rate is 11.97%. When $\alpha = \beta = 0.6$, the top-1 accuracy is 61.01%, and the BC rate is 12.39%. These results indicate that introducing the boundary blocking relations can help imporve the trade-offs between accuracy and blocking. When $\alpha = \beta = 0.7$, fewer paths of blocking pairs are added, lead to more blocking errors. When $\alpha = \beta = 0.6$, more paths of blocking pairs are added thus more complex the base classifiers are, lead to less blocking errors but more base classifier errors.

4.2.3. Results of EC-TriHC

We compute term frequencies based on an independent validation set and then use TMLG algorithm to mine the latent topics. Following [39], we use perplexity to select the appropriate topic number K. Since we only keep topics that contain low-level categories from different high-level categories, some of the learned topics are abandoned. We get three topic sebsets: T_1 = {*Tee, Blouse, Top, Tank*}, T_2 = {*Dress, Jacket, Skirt, Blazer, Blouse*}, and T_3 = {*Cardigan, Sweater, Shorts, Dress*}, where the categories in T_1 are belonging to the same high-level category, thus we keep T_2 and T_3 to construct expert classifiers.

Table 1 shows the accuracy (%), F1 score (%), Macro-F1 score (%), and the BC rate (%) of the EC-TriHC model. In order to show the effectiveness of different voting strategies, we conduct experiments on 3 different voting strategies. The three voting strategies are: (1) **+Base+P**: comparing the probabilities of both base and expert classifiers, (2) **+P**: comparing the probabilities of expert classifiers, and (3) **+IG**: comparing the information gain of expert classifiers. Note that, first, the

voting mechanism was triggered only if the sample is sent to the expert classifier for secondary classification. Second, in this experiment, we did not evaluate the top-3 accuracy metric because the category numbers of expert classifiers are small, thus the top-3 accuracy metric is meaningless here.

From Table 1, first, we can observe that the EC-TriHC model with voting strategy (3) achieves the highest Top-1 accuracy (62.03%) and the highest F1-score (61.34%). The Macro-F1 score is 33.21%, which is the highest except that of FC. The BC rate is 11.77%, slightly smaller than that of CLPK-TriHC. EC-TriHC does not change the original base classifiers, so it will not produce more base classifier errors, that is why the overall classification performance of EC-TriHC is better than CLPK-TriHC. Second, when comparing voting strategies (1) and (2), we could see the effectiveness of only considering expert classifiers. Third, when comparing voting strategies (2) and (3), we could see the effectiveness of information gain. Note that, the EC-TriHC method equals to the SHC when removing the *BND* region, naturally verify the need for 3WD.

4.2.4. Comparison and analysis

Table 2 shows the Top-1 accuracy (%), Top-3 accuracy (%), F1 score (%), Macro-F1 score (%), and the BC rate (%) of the comparison methods and the proposed methods. From these results, we can observe that the proposed methods achieve the best results in multiple evaluation metrics. First, in terms of accuracy, EC-TriHC achieves the highest top-1 accuracy, which is 1.99% higher than that of SHC, and CLPK-TriHC achieves the highest top-3 accuracy, which is 4.9% higher than that of SHC. Second, in terms of F1-score, EC-TriHC achieves the highest F1-score, which is 1.72% higher than that of SHC. However, none of the HC models outperforms FC on the Macro-F1 metric. This is because, in order to reduce blocking errors, more or less the model will sacrifice the accuracy of certain categories. Nonetheless, it is worth noting that the Macro-F1 scores of the proposed methods in this paper are relatively high, this is due to our methods focus more on the categories that are confused with each other, thereby reducing the impact on other categories. Last, the proposed TriHC models achieve the highest and second highest BC rate, indicate the superiority of the proposed TriHC models in blocking reduction.

4.3. Experimental results on the Stanford Dogs dataset

To further verify the effectiveness of our method, we did experiments on the subset of the Stanford Dogs dataset as introduced in Section 4.1.

4.3.1. Comparison methods

As shown in Table 3, the FC method achieved the accuracy of 73.00%, the SHC method achieved the accuracy of 72.67%, the MMHC method achieved the accuracy of 73.33%, and the PKHC method achieved the accuracy of 74.33%. For the evalua-

Table 1
The evaluation results on the DeepFashion dataset with different voting strategies.

Strategy	Top-1 acc.	F1	Macro-F1	BC rate
+Base+P	60.23	60.01	32.42	9.01
+P	61.77	61.10	33.19	10.21
+IG	***62.03***	***61.34***	***33.21***	***11.77***

Table 2
The evaluation results on the DeepFashion dataset with different methods.

Method	Top-1 acc.	Top-3 acc.	F1	Macro-F1	BC rate
FC	61.02	80.31	60.23	***33.56***	–
SHC	60.04	75.77	59.62	32.38	0
MMHC	60.21	78.92	59.76	32.54	1.45
PKHC	60.72	78.19	60.52	31.60	3.87
CLPK-TriHC	61.42	***80.67***	60.90	32.98	***12.32***
EC-TriHC	***62.03***	–	***61.34***	33.21	11.77

Table 3
The evaluation results on the subset of Stanford Dogs dataset with different methods.

Method	Top-1 acc.	F1	Macro-F1	BC rate
FC	73.00	73.00	73.32	–
SHC	72.67	72.67	72.87	0
MMHC	73.33	73.33	73.46	2.17
PKHC	74.33	74.33	74.45	***8.70***
CLPK-TriHC	74.50	74.50	74.61	6.52
EC-TriHC	***74.67***	***74.67***	***74.76***	5.43

tion of BC rate, the MMHC method achieved the BC rate of 2.17%, and the PKHC method achieved the BC rate of 8.70%. This indicates that the PKHC method can correct more blocked samples.

4.3.2. Results of CLPK-TriHC

In this experiment, we adopted the same hyper parameters and thresholds as introduced in Section 4.2.2. The reconstructed topology of the hierarchical model is {{*English setter, Gordon setter, Irish setter*,}, {*Siberian husky, Malamute, Eskimo dog*}, {*Great dane, Boxer, Bull mastiff, Eskimo dog, Saint bernard, Siberian husky*}, and {*Golden retriever, Chesapeake bay retriever,Curly coated retriever,Flat coated retriever,Labrador retriever*}}. The CLPK-TriHC method achieved the accuracy of 74.50%, and the BC rate of 6.52%.

4.3.3. Results of EC-TriHC

Jist like experiments on the DeepFashion dataset, we computed term frequencies based on an independent validation set and then used TMLG algorithm to mine the latent topics. We get two topic sebsets: T_1 = {*Siberian husky, Malamute, Eskimo dog*}, and T_2 = {*Labrador retriever, English setter, Golden retriever, Irish setter*}. In this experiment, we adopted the **+IG** voting strategy. The EC-TriHC method achieved the accuracy of 74.67%, and the BC rate of 5.43%.

4.3.4. Comparison and analysis

In comparison, the EC-TriHC method achieved the highest accuracy, F1 score, and Macro-F1 score. For the evaluation of the BC rate, the PKHC method achieved the highest score of 8.70%. This indicates that, in some cases, the PKHC method sacrifice the base classifier' accuracy, because the PKHC method overrefines the hierarchical topological structure and weakens the performance of the base classifiers.

4.4. Discussion

In this section, we first discuss the superiority of the proposed methods by analyzing the effectiveness of the two category relations mining methods. Then we discuss the application scope of the proposed methods.

4.4.1. Superiority

First, due to the uncoordination of the prior knowledge and actual data, most blocking errors can not be corrected in the original hierarchy. Thus, the methods that will not change the hierarchy have little improvement in the blocking problem. The classification performance of MMHC proves this. Therefore, this paper studies the problem of constructing a better hierarchy. Second, some existing reconstructing hierarchy methods, such as PKHC, crudely combine all the sub-categories of similar high-level categories, which although improves the classification accuracy, but also sacrifices some of the low-level categories that are not easily to be blocked. The low Macro-F1 scores of PKHC model prove this.

The proposed category regrouping methods can alleviate the above problems. First, by mining the relationship between categories and using 3WD to reduce the uncertainty of label relation dividing, the uncoordination of the original hierarchical structure and the actual data can be eliminated partially. Then, modify the original hierarchy to help reduce blocking errors. Second, the CLPK-TriHC model considers the cross-level blocking and cuts unnecessary path between high- and low-level categories, which reduces the complexity of the high-level classifiers and avoids sacrificing some of the low-level categories that are not prone to blocking. This can be reflected from the higher macro-F1 scores of the proposed methods. Third, the EC-TriHC model introduces expert classifiers by mining the latent topics from low-level categories, which only focus on the similar categories, while in CLPK-TriHC, all low-level categories of the high-level category at the end of the blocking direction will be involved in the modification process. Thus both CLPK-TriHC and EC-TriHC can reduce the impact on other categories that are not easily to be blocked.

4.4.2. Application scope

TriHC methods are suitable for alleviating the blocking problem caused by the inconsistency between the artificially defined hierarchy and the actual hierarchy of the data. But TriHC methods are not quite suitable for binary tree-structured hierarchies, because when the number of categories is large, the numbers of hierarchy levels and base classifiers increase and the relationship mining becomes complicated. Besides, the number of layers and categories of a dataset will affect the selection of hyper parameters and thresholds.

5. Conclusion and future works

In this paper, we proposed a three-way decision based hierarchical classification model (TriHC) to alleviate the blocking problem. The TriHC model learns category relations to rebuild the category hierarchy and uses 3WD to targetedly deal with the uncertain data. Adopting different category relation mining methods, we proposed two variants of TriHC, CLPK-TriHC model and EC-TriHC model. Specifically, in CLPK-TriHC model, we considered the cross-level category relationship between the blocked low-level categories and the high-level categories in the blocking direction. In EC-TriHC model, we construct experts which focus only on the relationship between the low-level categories so that narrowing down the blocking area. Experimental results on fashion classification demonstrated that the proposed methods can achieve better performance than

well trained deep convolutional neural networks and previous HC methods. Results also show the superiority of the proposed TriHC models in blocking reduction.

In future, we plan to take into account visual attributes in category relation mining step to better eliminate the inconsistency between the artificially defined hierarchy and the actual hierarchy of the data. Also, we will pay more attention to the generalization ability of the model in the later research.

Declaration of Competing Interest

There is no declaration of interest statement of this paper.

CRediT authorship contribution statement

Wen Shen: Conceptualization, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft, Writing - review & editing. **Zhihua Wei:** Supervision. **Qianwen Li:** Investigation. **Hongyun Zhang:** Supervision. **Duoqian Miao:** Supervision.

Acknowledgments

The work is partially supported by the National Key Research and Development Project (No. 213), the National Nature Science Foundation of China (No. 61573259, 61976160, 61573255), the Special Project of the Ministry of Public Security (No. 20170004), and the Key Lab of Information Network Security, Ministry of Public Security (No. C18608).

References

[1] X. Bai, S.I. Niwas, W. Lin, B.-F. Ju, C.K. Kwoh, L. Wang, C.C. Sng, M.C. Aquino, P.T. Chew, Learning ECOC code matrix for multiclass classification with application to glaucoma diagnosis, J. Med. Syst. 40 (4) (2016) 78.
[2] A. Binder, M. Kawanabe, U. Brefeld, Efficient classification of images with taxonomies, in: Asian Conference on Computer Vision, 2009, pp. 351–362.
[3] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[4] S. Chakrabarti, B. Dom, R. Agrawal, P. Raghavan, Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies, Vldb J. 7 (3) (1998) 163–178.
[5] T.G. Dietterich, G. Bakiri, Solving multiclass learning problems via error-correcting output codes, J. Artif. Intell. Res. 2 (1994) 263–286.
[6] I. Dimitrovski, D. Kocev, S. Loskovska, S. Deroski, Hierarchical annotation of medical images, Pattern Recognit. 44 (10) (2011) 2436–2449.
[7] S. Dumais, H. Chen, Hierarchical classification of web content, in: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2000, pp. 256–263.
[8] T. Gao, D. Koller, Discriminative learning of relaxed hierarchy for large-scale visual recognition, in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 2072–2079.
[9] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, IEEE, 2016, pp. 770–778.
[10] S. Jiang, M. Shao, C. Jia, Y. Fu, Learning consensus representation for weak style classification, IEEE Trans. Pattern Anal. Mach. Intell. PP (99) (2017). 1–1
[11] F. Kamiran, S. Mansha, A. Karim, X. Zhang, Exploiting reject option in classification for social discrimination control, Inf. Sci. 425 (2018) 18–33.
[12] S.W. Ke, W.C. Lin, C.F. Tsai, Y.H. Hu, Soft estimation by hierarchical classification and regression, Neurocomputing 234 (C) (2017) 27–37.
[13] A. Khosla, N. Jayadevaprakash, B. Yao, F.-F. Li, Novel dataset for fine-grained image categorization: Stanford Dogs, in: Proc. CVPR Workshop on Fine–Grained Visual Categorization (FGVC), 2, 2011.
[14] D. Koller, M. Sahami, Hierarchically classifying documents using very few words, in: International Conference on Machine Learning, 1997.
[15] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: International Conference on Neural Information Processing Systems, 2012, pp. 1097–1105.
[16] F. Li, D. Miao, W. Zhang, Three-way decisions based multi-label learning algorithm with label dependency, in: International Joint Conference on Rough Sets, Springer, 2016, pp. 240–249.
[17] H. Li, L. Zhang, B. Huang, X. Zhou, Sequential three-way decision and granulation for cost-sensitive face recognition, Knowl. Based Syst. 91 (2016) 241–251.
[18] H. Li, L. Zhang, X. Zhou, B. Huang, Cost-sensitive sequential three-way decision modeling using a deep neural network, Int. J. Approx. Reason. 85 (2017) 68–78.
[19] J. Li, X. Deng, Y. Yao, Multistage email spam filtering based on three-way decisions, in: International Conference on Rough Sets and Knowledge Technology, Springer, 2013, pp. 313–324.
[20] J. Liu, H. Li, X. Zhou, B. Huang, T. Wang, An optimization-based formulation for three-way decisions, Inf. Sci. 495 (2019) 185–214.
[21] K.-H. Liu, Z.-H. Zeng, V.T.Y. Ng, A hierarchical ensemble of ECOC for cancer classification based on multi-class microarray data, Inf Sci 349 (2016) 102–118.
[22] M. Liu, D. Zhang, S. Chen, H. Xue, Joint binary classifier learning for ECOC-based multi-class classification, IEEE Trans. Pattern Anal. Mach. Intell. 38 (11) (2015) 2335–2341.
[23] Z. Liu, P. Luo, S. Qiu, X. Wang, X. Tang, Deepfashion: powering robust clothes recognition and retrieval with rich annotations, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1096–1104.
[24] J.D. Lurie, H.C. Sox, Principles of medical decision making, Spine 24 (5) (1999) 493–498.
[25] A. Naik, H. Rangwala, Inconsistent node flattening for improving top-down hierarchical classification, in: IEEE International Conference on Data Science & Advanced Analytics, 2017.
[26] A. Naik, H. Rangwala, Improving large-scale hierarchical classification by rewiring: a data-driven filter based approach, J. Intell. Inf. Syst. 52 (1) (2019) 141–164.
[27] F.E. Otero, A.A. Freitas, C.G. Johnson, A hierarchical classification ant colony algorithm for predicting gene ontology terms, in: European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, Springer, 2009, pp. 68–79.
[28] A. Passerini, M. Pontil, P. Frasconi, New results on error correcting output codes of kernel machines, IEEE Trans. Neural Networks 15 (1) (2004) 45–54.
[29] S.G. Pauker, J.P. Kassirer, The threshold approach to clinical decision making, New Engl. J. Med. 302 (20) (1980) 1109–1117.
[30] J. Qiao, Hesitant relations: novel properties and applications in three-way decisions, Inf. Sci. 497 (2019) 165–188.
[31] A. Savchenko, Sequential three-way decisions in multi-category image recognition with deep features based on distance factor, Inf. Sci. 489 (2019) 18–36.

[32] W. Shen, Z. Wei, C. Zhao, D. Miao, A self-adaptive cascade ConvNets model based on three-way decision theory, in: CCF Chinese Conference on Computer Vision, 2017, pp. 433–444.
[33] G. Sherlock, Gene ontology: tool for the unification of biology, Can. Inst. Food Sci. Technol. J. 22 (4) (2009) 415.
[34] C.N. Silla, A.A. Freitas, A survey of hierarchical classification across different application domains, Data Min. Knowl. Discov. 22 (1–2) (2011) 31–72.
[35] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, Proceedings of the International Conference on Learning Representations (2015) 1–14.
[36] D. Ślezak, Rough sets and Bayes factor, in: Transactions on Rough Sets III, Springer, 2005, pp. 202–229.
[37] A. Sun, E.-P. Lim, W.-K. Ng, J. Srivastava, Blocking reduction strategies in hierarchical text classification, IEEE Trans. Knowl. Data Eng. 16 (10) (2004) 1305–1308.
[38] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[39] Z. Wei, W. Shen, C. Zhao, D. Miao, A self-adaptive cascade ConvNets model based on label relation mining, Neurocomputing 328 (2019) 29–38.
[40] L. Wen, M. Duo-Qian, W. Zhi-Hua, W. Wei-Li, Hierarchical text classification model based on blocking priori knowledge, Pattern Recognit. Artif. Intell. 23 (4) (2010) 456–463.
[41] Y. Yao, Three-way decision: an interpretation of rules in rough set theory, in: International Conference on Rough Sets and Knowledge Technology, 2009, pp. 642–649.
[42] Y. Yao, The superiority of three-way decisions in probabilistic rough set models, Inf. Sci. 181 (6) (2011) 1080–1096.
[43] Y. Yao, Three-way decisions with probabilistic rough sets, Inf. Sci. 180 (3) (2011) 341–353.
[44] Y. Yao, An outline of a theory of three-way decisions, in: International Conference on Rough Sets and Current Trends in Computing, Springer, 2012, pp. 1–17.
[45] Y. Yao, X. Deng, Sequential three-way decisions with probabilistic rough sets, in: IEEE International Conference on Cognitive Informatics & Cognitive Computing, 2011, pp. 120–125.
[46] H. Yu, P. Jiao, Y. Yao, G. Wang, Detecting and refining overlapping regions in complex networks with three-way decisions, Inf. Sci. 373 (2016) 21–41.
[47] L. Zhang, H. Li, X. Zhou, B. Huang, Sequential three-way decision based on multi-granular autoencoder features, Inf. Sci. 507 (2020) 630–643.
[48] Y. Zhang, D. Miao, J. Wang, Z. Zhang, A cost-sensitive three-way combination technique for ensemble learning in sentiment classification, Int. J. Approx. Reason. 105 (2019) 85–97.
[49] B. Zhou, Y. Yao, J. Luo, Cost-sensitive three-way email spam filtering, J. Intell. Inf. Syst. 42 (1) (2014) 19–45.
[50] J.D. Zhou, X.D. Wang, H.J. Zhou, J.M. Zhang, N. Jia, Decoding design based on posterior probabilities in ternary error-correcting output codes, Pattern Recognit. 45 (4) (2012) 1802–1818.