First of all, the purpose of this paper is to define the concept of EM. The traditional data model is Spuervised machine learning, but the disadvantage is that it requires a large number of data sets, which is not desirable in real life. So the idea of using labeling function LF to produce a lot of matching or non-matching labels containing noise. In this paper, the author focuses on using simple and powerful weakly supervised label task, in order to complete entity matching, using EM-specific transfer characteristics. The conclusion of themethod presented in this paper is superior to the best existing 10 usual weakly supervised data sets. In an entity EM dataset comparing two single tables and nine double tables, our labeling model results are 9% higher than the average of the best existing methods, and our method can improve labeling efficiency.

1. introduction

What is Entity matching EM? Dealing with whether a pair of data from two data sources is the same entity in the real world. EM has many uses such as matching product lists for competing products and building knowledge maps.The need to build weak supervision in EM,Users write LF functions to label data rather than writing them manually.To answer the question of whether it is possible to think outside the box, use more intelligent classification methods such as random forests to label models and avoid using existing complex methods. A yes answer will be presented here, and in the third chapter a simple method will be shown to perform better on 10 data sets than the existing method.

The main contributions of this paper are as follows: 1. A general labeling model is proposed; 2. This model is a simple and powerful labeling model; 3. Perfect solution for transitivity. 4, ML foundation to solve the transitivity.

2. problem definitionThe whole data label definition, elicits the use of label functions, and needs to meet the transitivity of the entity.

3. put forward the label model

The general form of the weakly supervised task labeling model proposed by us is introduced.

3.1 General view of label model

A formula expression is introduced where X represents the labeling matrix and y represents the predicted labeling vector, and an algorithm is required as a function parameter G satisfying y = (, Θ). It is assumed that the data points are independent of each other, so the predicted label values are also independent of each other. g is used as a label value function to predict each independent data point. For the i th data point, it refers to the i th row in matrix X. For example, if g(,), where is the parameter value of g. If (,) ≥ 0.5 when hit i=1, otherwise hit i=0. Here g is a classifier that takes a feature variable xi and predicts its label value. In principle, can be any classifier, how to learn parameters without labels, here leads to the expected maximum algorithm. This algorithm is mainly used to learn the model parameter , and the function is roughly the negative logarithmic likelihood function.

The process of the algorithm is roughly as follows:

(1) The initial estimate of hidden truth tag is obtained by the majority of votes, which is proved to be very effective in practice.

(2) M step: In the case that in the label is the current fixed value, the value of the model parameter is estimated by minimizing the previous equation 1. This step is to learn the model parameter by estimating the current label.

(3) Step E: Update the estimated label value and use the sliding model parameter value in step M.
(4) Repeat steps 2 and 3 until the results converge.

It is easy to find from formula 1 that by derivation and setting the derivation result to 0, the minimum value can be obtained from formula 1 in the M step, which can measure the difference between the predicted model value of the objective function and the current actual value. Therefore, the objective function can be replaced by any loss function in the M step to obtain similar results. And some classes have their own specific loss function form, so it is possible to use any loss function instead of the objective function while achieving similar results in the M step, that is, g can be instantiated to any classifier.

Instantiate different label models

The expected maximum algorithm can be instantiated into different existing model methods, where g can be used as a different hypothetical method. A common construction is to use confusion matrix for modeling and develop a series of truth inference methods on this basis. The other hypothesis is the Markov hypothesis, which takes the probability graph as the core hypothesis, so the truth inference problem can be transformed into a PGM with hidden variables.

3.2 Simple Algorithm

In principle, g as a classifier can be any classifier, but the goal we choose is that g should not be too large, otherwise the content of g learned in the first iteration of the M step is limited, on the other hand, g should not be too small, because we need to meet the interaction and dependency of different LF features.In fact, different existing ways and through different assumptions can limit the ability of g. This raises the question, do we need to design our own responsible models to limit capacity, or can we directly use existing generic classifiers as g to achieve similar results? The answer is yes, we can use generic classifiers such as random forest as g to get better performance. In this way, there is no need to spend a lot of effort to create some complex models, and the result may not be good, and directly using the existing model can get a good performance, of course, is a better choice. Here we learn to first learn the model from the principle, and then demonstrate that the model can not be too big or too small, and finally propose that since it is a classifier, Can the existing excellent classifiers be used directly to achieve the effect? If so, there is no need to spend more time thinking of more complex classifiers; if not, then find a way to optimize from the existing classifiers.

Here, the author thinks that in order to use a general classifier, the more direct choice is to use a linear classifier such as logistic regression, which simplifies the problem to use a weighted majority voting method. In essence, linear regression is to assign different weights to each LF, and finally synthesize all LF to obtain the final label result. However, logistic regression is limited in that it cannot capture more complex dependencies or interactions before different LFS.

The goal of selecting a suitable classifier is that the first is to be able to express the previous interactions of different LF, and the second is to be able to limit the capacity of g. The tree-based method can satisfy the interactivity between LFS, such as making decisions based on different levels and characteristics of the tree in the decision tree, which actually shows that the tree-based method performs well in structured data. Second, random forests can be used to limit the capacity of g. The capacity of the restricted classifier is also called regularization, and cross-validation is done in a display - and data-driven manner.

When g is a random forest classifier, step M of the expectation maximization algorithm is used to train the classifier with the current estimated label and step E is used to predict the training set, thus obtaining an updated version of .

Class unbalance

In real life, there are many category imbalances in data sets, among which category imbalances refer to the large difference in the number of samples of different categories in the data set. If the difference is too large, the classifier's ability to identify certain categories will become poor in the training process, which will affect the training effect. In particular, there are unbalanced data samples in EM's data set. In order to solve this problem, the state-of-art technique is adopted, which is the same as the typical class imbalance problem in ML. Solve the class imbalance problem at step M when training the classifier by adding a few class points to match the most class points [17] SMOTE is a simple technique to deal with the class imbalance problem in practice. For example, if you have two positive classes,(x1, 1),(x2,1), use SMOTE to create new data points ((x1+x2)/2, 1) based on these two positive classes, and then train the model by adding a few classes to the model in M steps. After training, the original data point set is predicted in step E, and the updated value is obtained. Because we do not know if there is a class imbalance problem in a given data set, SMOTE is performed at all M steps.

Here's the pseudocode for SMOTE:

Algorithm 1: SIMPLE

Input: Labeling matrix X
Output: Estimated soft labels $\boldsymbol{\gamma}$

$\boldsymbol{\gamma} \leftarrow$ majority vote on X
while *Not Converged* **do**
  M Step
    Obtain hard labels $\hat{\boldsymbol{y}}$ by binarilize the soft labels $\boldsymbol{\gamma}$.
    Make the classes balanced: $X', \hat{\boldsymbol{y}}' = \text{SMOTE}(X, \hat{\boldsymbol{y}})$
    Select random forest parameters $d_{\max}$ and ccp_alpha with cross validation on data $(X', \hat{\boldsymbol{y}}')$
    RandomForestClassifier.fit$(X', \hat{\boldsymbol{y}}')$
  E Step
    $\boldsymbol{\gamma} \leftarrow$ RandomForestClassifier.predict_proba(X)
end
return $\boldsymbol{\gamma}$

Computational complexity

The complexity of each iteration depends on training a random forest classifier whose time complexity is (log( )), where N is the number of tuple pairs of the candidate set. M1 is defined as the number of iterations and the total time complexity is (log( )). The experiment results show that 10 iterations can meet the requirements, and the space complexity is ( ).

Discuss

Here we propose a pseudo-labeling method similar to ours, which is based on semi-supervised learning. In the pseudo-labeling method, the model is first trained in the labeled training set, and the unlabeled data is

predicted, and then the data set of reliability prediction is added to the training set, and then the process is iterated continuously in the new training set. Our approach is similar to the pseudo-annotation approach in that both use the prediction data as the annotation data for the next iteration. The difference is that pseudo annotation methods are semi-supervised methods, and ours are unsupervised methods. The advantage of our method of using weak supervision is that the LF provided by each user is definitely better than a random guess. This allows for a good initial value estimate, and secondly, the purpose of the pseudo-labeling method is to train the model, while our focus is to obtain labels for all data points.

4. incorporating transitivity

Several cases are divided to explain the usage of transitivity and the actual algorithm process, and the algorithm derivation is used to simplify the calculation results, and other algorithm conclusions are used to reduce the calculation amount. Finally, the algorithm complexity is calculated.

5.experiment

Different data sets and different perspectives are used to illustrate that the algorithm proposed by the author is better, and there are corresponding comparative data and conclusions for each perspective. Emphasize what the different data say, emphasize what?

6. Related work

Entity matchingAlthough the supervised algorithm performs well in entity matching, it needs a lot of manpower to label data. Transfer learning can also be applied to existing source data or trained language models, and its robustness will deteriorate when the target data set changes, as can be seen from the experimental data. Different from the previous existing methods, for entity matching, weak supervision is adopted here, and the data label is generated programmatically, which saves a lot of manpower, and the matching effect is better. Especially when there is no annotated data, a new entity matching method is provided. Truth Inference. Truth inference

The general truths that exist are designed for general tasks. The Simplex -EM method proposed by the author is specially designed for EM. It combines the unique transitiveness of EM, and has better performance than the existing general truth value inference methods, lower time complexity, easier to understand, lower model complexity, and no need to manually label the data set based on weak supervision. Most of the existing methods need manual annotation, and the models are complex with various assumptions. In contrast, our approach requires no assumptions and is generic for all datasets, yielding better results on both general and EM tasks.

7. Summary

In this paper, a new EM matching method is proposed innovatively. By combining different LF functions to apply to weakly supervised data sets, it does not need data labeling and saves a lot of manpower. This matching method is proposed for the first time, it is simple and efficient, and this method is applied to EM, adding the unique transitivity of EM. The experimental data show that our method performs better than the existing method in 10 weakly supervised data sets from different angles by comparing the existing methods from various angles.

To solve the EM problem, the author first compared the existing methods and found the defects of the existing methods, such as the complexity of the model and the need for a lot of manpower to label the data. Based on this, a new innovative method is proposed, and the unique transitivity of EM is combined with the credibility and dexterity of the method from a theoretical point of view, and then several groups of experimental data are compared with different existing methods and from different angles, and it is found that the existing method performs better. From this paper, we learn that in view of the shortcomings of existing methods, we constantly ask ourselves whether we can do better in what aspects and from what angles. Then I searched a lot of literature to find out the feasibility of this problem. Finally, I thought from the experimental point of view, which aspects can better demonstrate the superiority of this method, and everything was based on data and feedback conclusions from data results. This kind of thinking is what we need to learn, whether it is any topic or any difficulty, it is to think about the defects in the existing methods, and make good use of the existing theoretical basis, and constantly improve and progress little by little. The last is to learn the idea of data demonstration, everything must have the data to speak, the author a large length of experimental data, different angles to demonstrate the method of data superiority, from the perspective of data is very convincing, but also have a global thinking, the Angle should be complete, can not only from a favorable perspective to explain, so it is not credible.

Literature reference