One-to-Many Data Linkage By One-Class Clustering Tree

P. Aruna kumari, Dr. Srinivas Prasad
GMR Institute of Technology
Arunakumari1264@gmail.com, Srinivas_prasad@hotmail.com

ABSTRACT

Data linkage is the task of identifying different entries that refer to the same entity across different data sources. The goal of the data linkage task is joining data sets that do not share a common identifier. Data linkage is divided into two types: one-to-one and one-to-many. In one-to-one data linkage, the goal is to associate an entity from one data set with a single matching entity in another data set. In one-to-many data linkage, the goal is to associate an entity from the first data set with a group of matching entities from the other data set. We implement a new one-to-many data linkage method that links between entities of different natures. In addition, while data linkage is usually performed among entities of the same type, the data linkage technique can match entities of different types. The data linkage method is based on a one-class clustering tree (OCCT) that characterizes the entities that should be linked together.

Key words: Clustering, classification, data matching, decision tree induction.

1. INTRODUCTION

Data linkage is named as Record linkage. Data linkage refers to the task of finding different entries that refer to the same entity across different data sources [1]. The main aim of the data linkage task is to join the data sets that do not share a common identifier. The various data linkage scenarios are linking data when combining two different databases [15]. Data linkage is divided into two types namely one-to-one and one-to-many. In one-to-one data linkage, the goal is to associate an entity from one data set with a single matching entity in another data set [12]. In one-to-many data linkage, the goal is to associate an entity from the first data set with a group of matching entities from the other data set. Most of the previous works focus on one-to-one data linkage.

In this paper, a new record linkage method which performs one-to-many linkage is proposed. This method links the entities using a One-Class Clustering Tree (OCCT) [1]. A clustering tree is a tree in which each of the leaves contains a cluster instead of a single classification [3]. Each cluster is generalized by a set of rules that is stored in the appropriate leaf.

The OCCT can be used in different domains like fraud detection, recommender systems and data leakage prevention. In fraud detection domain, the main aim is to find the fraudulent users and identify actions by the wrong people. In recommender systems domain, is to satisfy the customer’s product needs by providing them recommendation based on products. The recommender system mainly evaluates the customer’s profile by seeing what choice is it making or using while rating items. In data leakage prevention domain, the main aim is to detect the abnormal access to the database records that indicates data leakage or data misuse.

2. RELATED WORK

Record linkage is a process of matching entities from two different data sources that may or may not share a common identifier. One-to-one data linkage was developed using algorithms like SVM classifier—which is trained to differentiate between matching and the non matching record pairs [11]. Maximum Likelihood Estimation (MLE) helps in determining the probability of record pair being matched [10].

Most of the previous works have addressed one-to-many data Linkage only. Storkey et al. [4] used the Expectation Maximization algorithmic rule. During this algorithmic rule, they’re scheming the likelihood of a given record try that is matched and to learn the characteristics of the matched records. A mathematician mixture model was accustomed model the conditional magnitude distribution.

Ivie et al. [5] performed data linkage using five attributes
namely, individual’s name, gender, date of birth, location, and the relationships between the individuals. A decision tree was induced using those five attributes. But, however, the main disadvantage of this method is that it performs matches using the specific attributes and therefore becomes very hard to generalize.

Christen and Goiser [6] used a decision tree to determine which records must be matched to one another. In their work, they compare different decision trees which are built based on different string comparison methods. However, in their method, the attributes according to which the matching is performed are predefined and only one or two attributes are usually used.

In this paper, we propose a new data linkage method aimed at performing one-to-many linkage that can match entities of different types. The inner nodes of the tree consist of attributes referring to both of the tables being matched ($T_A$ and $T_B$). The leaves of the tree will determine whether a pair of records described by the path in the tree ending with the current leaf is a match or a non match.

Decision trees are predominantly used for regression tasks and for classification. However, the training set which is used for the induction of tree should not be unlabeled. Besides, getting a labeled dataset is bit expensive task. Therefore, instead of using training set with labeled dataset, we can highly prefer a decision model which requires examples of one class only. If we compare, traditional decision trees with clustering trees, they are different based on their structure [7]. A clustering tree is a tree in which each of the leaves contains a cluster whereas a traditional tree consists of a single classification. Each cluster in the clustering tree is generalized by a set of rules. Each leaf of the tree is characterized by a logical expression representing the instances belonging to it. According to the main advantage of using clustering trees is that they provide a description for each of the clusters using a logical expression.

3 INDUCING THE LINKAGE MODEL

In the planned technique, the primary step is linkage model induction. The linkage model receives the information concerning records that square measure expected to match every alternative. This method consists of explanation and structure of the tree. The tree building needs the decision of which attributes should be selected at each level of the tree. The inner nodes of the tree consist of attributes from table $T_B$. The leaf contains the cluster that is matching attributes from table $T_B$ to $T_A$. We can use any one of the splitting criteria for the selection of attributes. The splitting criteria ranks the attributes based on their clustering of matching examples.

In this planned technique, a pre-pruning approach is implemented. In this approach, the algorithm stops expanding a branch whenever the sub-branch does not improve the accuracy of the given model. The inducer is actually trained with matching examples only.

Any one of the splitting criteria can be used to derive the OCCT. The splitting criterion is used to determine which attribute must be used in each step of constructing the tree. Our main task is to achieve a tree that has less number of nodes, as smaller trees easily generalize the data by avoiding over fitting. At the same time, it will also be simpler for the human eyes to understand the tree structure [8]. The four types of splitting criteria used in this system are: Coarse Grained Jaccard Coefficient (CGJ), Fine Grained Jaccard Coefficient (FGJ), Maximum Likelihood Estimation (MLE) and Least Probable Intersections (LPI).

4 SPLITTING CRITERIA

The goal is to achieve a tree which contains a minimum number of nodes. Smaller trees generalize the data in a better way, avoid over fitting, and will be simpler for the human eye to understand. Therefore, it is crucial to use an effective splitting criterion in order to build the tree.

4.1 Coarse-Grained Jaccard Coefficient

The Jaccard similarity coefficient, a measure that is commonly used in clustering, measures the similarity between clusters [8]. The goal is to choose the splitting attribute which leads to the smallest possible similarity between the subsets. Records are considered in the intersection only if they are completely identical. To minimize the computational complexity of building the model using the CGJ criterion, the values of the fields from $T_B$ can be expressed as a single string. Then, a string matching algorithm can be used to find the intersection between the two subsets of records.

4.2 Fine-Grained Jaccard Coefficient (FGJ)

The fine-grained Jaccard coefficient is capable of identifying partial record matches [8], as opposed to the coarse-grained method, which identifies exact matches only. It not only considers records that are exactly identical, but also checks to what extent each possible pair of records is similar.
4.3 Least Probable Intersections (LPI)

We refer to a distinct combination of attributes as a unique identifier of an entity. Therefore, our goal is to find a splitting attribute for which there is the least amount of identifiers that are shared, in comparison to a random split of the same size [10]. The goal is to find the splitting attribute, which is the least probable to generate the two subsets randomly. Therefore, the candidate splitting attribute with the highest score is chosen as the next attribute for split.

4.3 Maximum-Likelihood Estimation

Splitting criterion uses the MLE to choose the attribute that is most appropriate to serve as the next splitting attribute [9]. Each candidate attribute from the set of attribute splits the node data set into subsets according to its possible values. We aim to choose the split that achieves the maximum likelihood and hence we choose the attribute that has the highest likelihood score as the next splitting criterion in the tree.

The computational complexity of building a decision model using the MLE method is dependent on the complexity of building a statistical model and the time it takes to calculate the likelihood.

4.4 Dealing with Multi-valued Splits

The three measures which are described above (CGJ, FGJ, and LPI) were presented for binary attributes. When multi-values attributes exist in the data set, the proposed splitting criteria are adapted as follows. The score of a candidate splitting attribute is calculated as a weighted average of a series of possible binary splits. Multi-values split is because these methods are capable of measuring the similarity of only two record sets at a time, and therefore, an adjustment is necessary for multi-valued splits. The fourth measure (MLE) does not measure the similarity between two given record sets, and it is computed individually for each possible subset.

5. PRUNING

Pruning is an important task in the tree induction process the main purpose of victimization pruning is to make a tree with accuracy and conjointly to avoid over fitting. Pruning will be tired to completely different ways: pre-pruning and post-pruning [7]. In pre-pruning, the branches are pruned at the time of induction process if there are no possible splits found. In post-pruning, the tree is built completely followed by a bottom-up approach in order to determine which branches are not beneficial.

In our system we have adopted a pre-pruning approach. Pre-pruning approach was selected because it reduces the time complexity of the algorithm. The decision made to prune the branch or not is taken once the next attribute for split is chosen. In this proposed system, two pre-pruning methods are used. They are maximum likelihood estimation (MLE) and least probable intersections (LPI).

In the maximum-likelihood method, an MLE score is computed for each of the possible splits. If none of the candidate attributes achieve a MLE score which is higher than the current node’s MLE score, the branch is pruned and the current node becomes a leaf.

In the least probable intersections method is calculated for each possible split. If all possible splitting attributes is smaller than a predefined threshold, then all of the possible splits are likely to be formed by a random split.

6. USE OF OCCT FOR DATA LINKAGE

Linkage is a process. During the linkage (i.e., testing) phase, each pair of records in the testing set is cross validated against the linkage model. The output is a score representing the probability of the record pair being a true match. The score is calculated using MLE [11]. The likelihood score for a match between the records is calculated by using the probability of each value, given all other values and appropriate model.

It is determined whether the given records are a match or not by comparing the likelihood scores that was calculated to the given threshold. If the pairs score is greater than threshold, it is classified as a match; otherwise, it is classified as a non match. The threshold is defined by taking into consideration the tradeoff between the false positive rate (FPR) and the true positive rate (TPR).

7. RESULT ANALYSIS

The input to the algorithm is an instance from table A i.e., TA and an instance from table B i.e., TB. The output of this algorithm is a Boolean value determining whether the instances should be matched or not. The likelihood score for a match between the records is calculated by using the probability of each value, given all other values and appropriate model. Eventually, the determination of the given records is found match or not by comparing the likelihood score which was calculated earlier with the threshold value. The pair is found to be matched if the pair’s score is greater than the threshold value. It is considered as a non-match if the pair’s score is less than the threshold value.
8. CONCLUSIONS AND FUTURE WORK

The main aim is based on a one-class clustering tree (OCCT) that characterizes the entities that should be linked together. To summarize, this method allows performing one-to-many linkage while the traditional methods followed one-to-one linkage. Here we use four splitting criteria and two different pruning methods which can be used for inducing the OCCT. These methods are made in decision tree that reduce the time complexity of algorithm by reducing the size of tree. The method was evaluated using data sets from three different domains. Previous we have used a one-class approach which results in matching pairs are only required in the training set, as more number of non-matching pairs will confuse the model and it will lead to a less accurate model. Another advantage of using OCCT model is that the solution can be easily transformed to rules. Effectiveness and performance is high in terms of precision data linkage method aimed at performing one-to-many linkage that can match entities of different types.

The future work may include comparing the OCCT with the other data linkage methods. Also it can be extended to perform many-to-many linkage.

8. REFERENCES