A Review on OCCT: A One Class Clustering Tree (OCCT) for Implementing One-to-Many Data Linkage

Vikram G. Lachake, Prakash P. Rokade

[1]ME-II, Computer Engineering, JES's SND COE & RC, Yeola,

[2]Head of Department, Information Technology, JES's SND COE & RC, Yeola

Abstract: ***There is increased consciousness in several nations of the potential of record linkage for recommender system, data leakage detection and fraud detection. Record linkage compares records in one data set with records in another data set to match them. Record linkage is traditionally performed among tables to cluster the data. The proposed method aims to perform One-to-many data linkage i.e. to associate one record in Table T_A with one or more matching records in Table T_B using OCCT tree. The OCCT tree provides One-to-One as well as One-to-many record linkage between objects of same or different types and these objects do not share common attribute. It is easy to build OCCT tree and convert into linkage rules. The inner nodes of OCCT tree contains attribute from table T_B and the leafs holds a compact representation of a subset of records from Table T_B which are more likely to be linked with matching record from Table T_A.The values of Table T_A's attribute are according to the path from the root of the tree to the leaf. The induced OCCT tree is small in size due to use of splitting and pruning methods. The OCCT tree contains lesser number of nodes to avoid over fitting. Old methods take long time for one-to-many record linkage. The OCCT based on One Class approach that is it considers only positive examples (matching examples). Hence the proposed method provides better performance in terms of precision and recall as compared to previous record linkage methods.***

Keywords: Data linkage, One Class Clustering Tree, Splitting, Pruning, Data linkage

1. Introduction

Data linkage is the task of identifying different entries (i.e., data items) 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 (i.e., a foreign key). Common data linkage scenarios include: linking data when combining two different databases [2]; data Deduplication (a data compression technique for eliminating redundant data), which is commonly done as a preprocessing step for data mining tasks; identifying individuals across different census data sets; linking similar DNA sequences; and matching astronomical objects from different catalogues.

The data linkage can be divided into two types: one to one and one to many. The goal is to associate an entity from one data set with a single matching entity in another data set is in one to one linkage. Most of the previous works focus on one-to-one data linkage. A new data linkage method aimed at performing one to many linkages we propose. The proposed data linkage technique can match entities of different types, while data linkage is usually performed among entities of the same type.

For example, in a student database we might want to link a student record with the courses she should take. The proposed method links between the entities using a one class clustering tree. Each of the leaves contains a cluster instead of a single classification. Each cluster is generalized by a set of rules that is stored in the appropriate leaf. The evaluation of the OCCT is done from three different domains: data leakage prevention, recommender systems, and fraud detection. The goal is to detect abnormal access to database record that might indicate a potential data linkage or data misuse in the data leakage prevention domain. With records that can be legitimately retrieved within that context, is the goal to match an action, performed by a user within a specific context. The proposed method is used for matching new user of the system with the items that they are expected to like based on their demographic attribute in the recommended system domain. The goal is to identify online purchase transactions that are executed by a fraudulent user and not the legitimate user in the fraud detection domain. The OCCT performs well in different linkage scenarios are shown in results. In addition, it performs at least as accurate as the well-known C4.5 decision tree data-linkage model, while incorporating the advantages of a one-class solution. As it can easily be translated to linkage rules, so the OCCT is preferable over the C4.5 decision tree. Twofold is the contribution of this work. A method that allows performing one to many linkage between objects of the same or of different types is in our proposed system. This is opposed to existing methods that are only able to link between objects of the same type. Second, we use a one-class approach. The advantage is that in certain domains obtaining meaningful non matching examples can be difficult. For example, in the fraud detection case, we can easily obtain genuine matching examples; these are actually legitimate transactions performed by users. Non matching examples (fraudulent transactions) are rare and more difficult to obtain. In such cases, nonmatching examples can be artificially created and added to the training set; however, we can receive examples that do not make sense. For example, a fraudulent customer purchases a product that is not being sold in the customer's country [1].

2. Literature Survey

The task of matching entities from two different data source that do not share a common identifier is a data linkage. To perform data linkage the entities should be of same type. Data linkage can be divided into two types one to one and one to many. The goal is to associate one record in table TA with a single matching record in table TB is done in one to one linkage. The goal is to associate one record in TA with one to many matching records in TB is one to many data linkage [1].

I.P. Fellegi and A.B. Sunter presented mathematical model within a context of which linkage rules developed to determine records in two different data sets link or non-link to provide guidance for handling of linkage problem. The linkage rules assigns the probabilities for taking each of the three actions i.e. link, non-link or possible error. They defined two types of error as error of decision. First unmatched linked records are actually unmatched and second non-linked records are actually matched [2].

F.De Comite, F.Denis, R. Gilleron and F.Letouzey introduced POSC4.5 algorithm for record linkage of positive and unlabeled examples. They considered binary classification and hence this method is not generalized. They require not only the data set but also the information of positive examples out of whole data set. The attraction of their work is that they presented modified entropy formula which that considers weight of positive examples in a given data set. They assumed that negative examples are in unlabeled data set as per given distribution [4].

M.D.Larsen and D.B.Rubin used maximum likelihood learning amongst candidate models. The maximum likelihood define some similarity measure between records in one data set and those in another data set. We can use maximum likelihood to classify potential record pairs as either match or nonmatch [5].

A.J.Strokey, C.K.I. Williams, E. Taylor and R.G. Mann presented One-to-Many record linkage based on Expectation Maximization algorithm. They use the Expectation Maximization algorithm to compute the probability of a given record pair being match and to learn the characteristic of matched records. The method is derived for specific astronomical problem of far-infrared observations to optical counterpart, but is generally applicable. They described theory of record linkage but does not discussed its application or its implementation [6].

H.Blockeel, L.D. Raedt and J.Ramon presented a top down induction of decision tree in which each leaves contains cluster instead of single classification. Each cluster is characterized logical expression representing records belonging to it [8].

3. The Proposed Method

The proposed method is divided into following steps:
1) Inducing a clustering tree linkage model.
2) Building probabilistic models to represent the leaves.
3) Linking items according to the induced model.

3.1 Inducing a clustering tree linkage model

The knowledge of which records are expected to match each other is encapsulated in linkage model. The induction process includes deriving the structure of the tree. Building the tree requires deciding which attribute should be selected at each level of the tree. In inner nodes of tree consist of attributes from table TA only. By using one of the possible splitting criteria presented, selecting the attribute is done. The splitting criteria ranks the attributes based on how good they are in clustering the matching examples. A Pre Pruning process is implemented. This means that the algorithm stops expanding a branch whenever the sub branch does not improve the accuracy of the model. The inducer is trained with matching examples only.

3.1.1 Splitting Criteria

During inducing the clustering tree is that it should contain smallest number of nodes. Reducing the size of the tree (in number of nodes) that performs well on training set. It is believed that small tree would better generalize, avoids the over fitting and forms simpler representation for human eye which is easy for human eye to understand. The proposed method will use four splitting criteria to evaluate the splitting of the tree based on attribute of Table A. Each splitting criteria is used to calculate the similarity between two record sets T_1 and T_2 and is indicated by sim(T_1 ,T2). The splitting criteria that is used determine the attribute that creates the best split of a table that is Table T divided into two Tables T_A and T_B, which differ from each other as much as possible. Each attribute in Table T_A is evaluated to determine the record that it achieves.

3.1.2 Coarse Grain Jaccord Coefficient(CGJ)

The Jaccord Coefficient is used to find similarity between two clusters. The attributes which are not selected as splitting attribute, the similarity between the attribute subsets is calculated which is denoted by sim(T_A,T_B)[11]. The similarity between two attribute subsets is computed using Jaccord Coefficient as the ratio number of records belonging to two subsets. The attribute with smallest similarity value is chosen as next splitting attribute which creates two subsets which are different as much as possible.

3.1.3 Fine Grain Jaccord Coefficient(FGJ)

Fine Grain Jaccord Coefficient is capable of identifying partial matches only on the other hand Coarse grain find exact matches only i.e. all attributes contain same values. The Fine Grain Jaccord Coefficient is calculated as ratio of number of attribute containing the same values in two attribute subsets and total number of attributes examined that do not contain null value in either of subsets [11].

3.1.4 Least Probable Intersection(LPI)

Amir Greshman, Amnon Miesel propose splitting criteria that computes the probability for getting the intersection for each potential split in a random split and select the split that generates least probable size of intersection [12]. This method is depend on the Cumulative Distribution Function (CDF) of poission distribution. This method select the optimal splitting attribute that result into minimum amount of items that are shared between two item sets.

3.1.5 Maximum Likelihood Estimation (MLE)

The maximum likelihood i.e. probability score is calculated for the attribute which is not selected as splitting attribute by giving the value of other attribute. The attribute having highest maximum likelihood score is selected as the next splitting attribute [10]. The complexity of this method depends on the size of input data set, maximum likelihood score calculated for number of attribute and method used to build or model the tree (e.g. decision tree).

3.1.6 Pruning

Pruning is the process to trim unnecessary branches to improve accuracy of model. Thus tree is induced using matching examples only. The pruning process is use to give compact representation of tree i.e. it contains small number of attributes. It also avoid over fitting and improve the time complexity. There are two types of pruning process 1) Prepruning and 2) Postpruning.

The prepruning work on top down approach i.e. the pruning process is done during the tree induction process when further split does not give complete knowledge of record matching. linkage model. The postpruning work on bottom up approach i.e. the pruning process done after completion of inducing linkage tree when further split does not give complete knowledge of record matching. In our proposed system we are using prepruning process to reduce the time complexity. The decision whether to prune the branch taken once best splitting attribute is chosen. We propose either MLE or LPI our system.

In Maximum Likelihood Estimation (MLE), a MLE score is calculated for each of splitting attribute. If none of the candidate attribute achieve MLE score greater than current node then branch is pruned and current node becomes leaf node.

3.2 Building Probabilistic Models to represent Leaves

Once the tree is built then each leaf contains the matching record from Table B. The probabilistic model is built for each attribute of Table B by giving the values of other attributes. There are two goals for this step, first to reduce the size of tree, there by produce the compact representation of tree and to avoid overfitting. It is not necessary to create probabilistic model for each attribute of Table A. The attributes having specific meaning in leaf, the models are created for those attributes only. These attributes are selected using feature selection process. The purpose of feature selection process to best represent of records in leaf.

3.3 Linking items according to the induced model

In this linkage step, Maximum Likelihood Estimation (MLE) is calculated foe each possible pair of records. The MLE score indicate the probability of record pairs are being match. The cardinality of record pairs are multiply by MLE score. Then MLE score is compared with given threshold to decide given record pairs are match. If MLE score of record pair is greater than threshold then record pairs are categorized as match otherwise it is categorized as nonmatch.

4. Conclusion

The One Class Clustering Tree (OCCT), a one-class decision tree approach for performing one-to-many and many-to-many data linkage. The presented method is based on a one-class decision tree model that encapsulates the knowledge of which records should be linked to each other. The presented method used four possible splitting criteria and two possible pruning methods that can be used for inducing the data models. The presented method is effective when applied in different domains. The objective of OCCT is to link a record from a table TA with records from another table TB. The One Class Clustering (OCCT) is in the form of a tree in which the inner nodes represent attributes from TA and the leafs hold a compact representation of a subset of records from TB which are more likely to be linked with a record from TA, whose values are according to the path from the root of the tree to the leaf [1].

5. Future Scope

For future work, we can compare the OCCT with other data linkage methods. In addition, we can extend the OCCT model to the many-to-many case and to handle continuous attributes [1].

References

[1] Ma'ayan Dror, Asaf Shabtai, Lior Rokach, and Yuval Elovici, "OCCT: A One-Class Clustering Tree for implementing One-to-Many Data Linkage", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 3, MARCH 2014.

[2] I.P. Fellegi and A.B. Sunter, "A Theory for Record Linkage," J. Am. Statistical Soc., vol. 64, no. 328, pp. 1183-1210, Dec. 1969.

[3] M. Yakout, A.K. Elmagarmid, H. Elmeleegy, M. Quzzani, and A. Qi, "Behavior Based Record Linkage", Proc. VLDB Endowment, vol. 3, nos. 1/2, pp. 439-448, 2010.

[4] F. De Comite´, F. Denis, R. Gilleron, and F. Letouzey, "Positive and Unlabeled Examples Help Learning", Proc. 10th Int'l Conf. Algorithmic Learning Theory, pp. 219-230, 1999.

[5] M.D. Larsen and D.B. Rubin, "Iterative Automated Record Linkage Using Mixture Models", J. Am. Statistical Assoc., vol. 96, no. 453, pp. 32-41, Mar. 2001.

[6] A.J. Storkey, C.K.I. Williams, E. Taylor, and R.G. Mann, "An Expectation Maximization Algorithm for One-to-Many Record Linkage", Univ. of Edinburgh Informatics Research Report, 2005.

[7] P. Christen and K. Goiser, "Quality and Complexity Measures for Data Linkage and Deduplication," Quality Measures in Data Mining, vol. 43, pp. 127-151, 2007.

[8] H.Blockeel, L.D.Raedt, and J.Ramon, "Top down Induction of Clustering Tree", ArXiv computer Science e-prints, pp 55-63,1998.

[9] D.J. Rohde, M.R. Gallaghar, M.J.Drinkwater, and K.A.Pimpplet," Matching of Catalogues by Probabilistic Pattern Classification", Monthly Notices

of the Royal Astronomical Soc., vol. 369, no. 1, pp. 2-14, May 2006.

[10] D.D.Dorfmann and E.Alf, “Maximum-Likelihood Estimation of Parameters of Signal-Detection Theory and Determination of Confidence Intervals-Rating-Method Data,” Journal of Math Psychology, vol. 6, no. 3, pp. 487-496, 1969.

[11] S.Guha, R.Rastogi and K.Shim” ROCK: A Robust clustering algorithm for Categorical attributes”, Information System, Vol.25, no.5, pp. 345-366, July 2000.

[12] A.Gershman et al., “A Decision Tree Based Recommender System,” in Proc. the 10th Int. Conf. on Innovative Internet Community Services, pp. 170-179, 2010.