OCCT: A One –Class Clustering Tree for Implementing One – to- Many and Many – to- Many Data Linkage

Manali Pare Guha
Barktulla University of Instittute & Technology Bhopal

Anju Singh, PhD
Barktulla University of Instittute & Technology Bhopal

Divaker Singh, PhD
Barktulla University of Instittute & Technology Bhopal

ABSTRACT
One to many & many to many data linkage are necessary in data mining. OCCT Implementation for one to many & Many to many Data Linkage is to identify different entities across different Data sources. Data Linkage is linking data between two different database. One to many data linkage is associated an entity from first data set with a group matching from the other data set. In many to Many Data Linkage method the entities of same type and different nature should be arrange with Map Reduce method. In the OCCT was evaluated after using data sets from three different domains: , recommender system, data leakage prevention and fraud detection. data leakage prevention domain, the goal is to detect abnormal access. Recommender system, the method is used for matching new users of the system with the items. In fraud detection legitimate transactions performed by users.

Keywords
Clustering, classification, data matching, decision tree induction keywords, Map Reduce, Data Linkage Matching

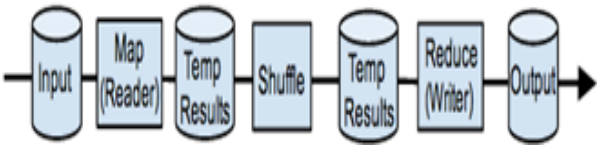
1. INTRODUCTION
With the quick development in information technology, the scale of data is increasing growing up. In that huge data poses a great challenge for data processing and arrange proper classification. For classifying the data we use several algorithm for efficient cluster data. For that we use Random Forest algorithm. This Random Forest classifier uses a many more decision tree, in order to improve the classification rate.

2. MANUSCRIPT WORK
A. Linkage of Data Set
In one to one data linkage, it is associate one record in table TA with a single matching record in table TB. In terms of one to many data linkage e, there is represent one record in TA with one or more matching records in TB. One-to-one data linkage was implemented as an SVM classifier that is trained to distinguish between matching and nonmatching record pairs . maximum - likelihood estimation (MLE) is the process to find the Probability of a record matching of two-dimensional clustering process [1] in which entities & attributes are clustered at the same time. The OCCT model also results in clusters of instances means entity, each may be described with a different set of attributes.

B. Decision Tree
Trees are generally used for classification and regression tasks. The training set use d for inducing the tree must be labelled [2]. An acquiring a labelled data set is a costly task for an adaptation of the C4.5 algorithm for learning from true Positive values that is unlabelled. In the given data sets, it necessary to sought out the knowledge of the ratio of positive values from whole data set. From this approach over the C4.5 algorithm is modified entropy formula which considers the weight of the positive class in the given data set and then assumes the number of negative data value in the unaddressed data depending on to the given distribution. In addition, or union only binary classification problems can be considered [9]. Create a forest of trees by iterating over different possible ratios of the positive class[10]. Accurate model which is the most required model for use.

C. Record Linkage
In Record linkage data is linked for the observation or micro level, for that it recognize the pair of record data files it is identical for representing. It is used for gathering of the whole data that wil be obtain from different origins for testing, research, updating or extending data files, post enumeration, DE duplication of sampling frames etc. Its advantage are that when the data are not able collect parallel that time record linkage composition an substitute, also its very affordable if it measure the direct collection of data.

D. Record Linkage using Map Reduce
The Hadoop Map Reduce framework is used in many to many data linkage to sort the output of the maps, which are then input to the reduce tasks. Both input and the output of the job are stored in a file system [3]. For the execution of data linkage in a distributed and parallel surrounding using Map reduce. Map reduce is a function which finds the output with a key and list of values related with that key [4], the library of Map Reduce function have collection of output classes .



D. Shuffle
The use of shuffle is firstly as a cloud storage bucket in Google it may be either as a default or specified. If shuffle is being started in that for that they all are grouped together with same keys, and every key has single list value output. If it may find the same pair of key value used by more than once then in shuffle value is associated appear multiple times output for that particular key .

E. Map Function
The Map Reduce library include a mapper class that perform the map stage in many to many data linkage[5]. The map stages uses an input reader that deliver data one record at a time using map function().

3. PROCESS FOR LINKAGE OF DATA SET

In this process of linkage we derive the structure of tree for one to many data linkage, in that inner node of tree Table TB to TA. Using pre pruning method it stop expanding a branch. For easy understanding of the tree structure four splitting criteria are used is: CGJ, FGJ, MLE, LPI.

A.Coarse Grained Jaccard Coefficient (CGJ)

The algorithm used in clustering it measure the similarity between the cluster. It generate the subset that are different from each other. Records of the Table may be concatenated as a string and use to find intersection only if record are identical.

B. Fine Grained Jaccard Coefficient (FGC)

It identify partial record matches, and better than CGJ which identify exact matches only. It not only consider identical match record but also extent each possible pair of record which is similar.

C. Maximum Likelihood Estimation (MLE)

This splitting criteria use for selecting the appropriate the next splitting attribute for the forthcoming attribute that yet to split.

D. *Least Probable Intersection (LPI)*

Captions IN this splitting criteria used cumulative distribution function (CDF). In this splitting attribute it must be least amount of identifier that are shared. Least probable to generate the subset randomly which has achieved the higher score. And this method is affordable than other.

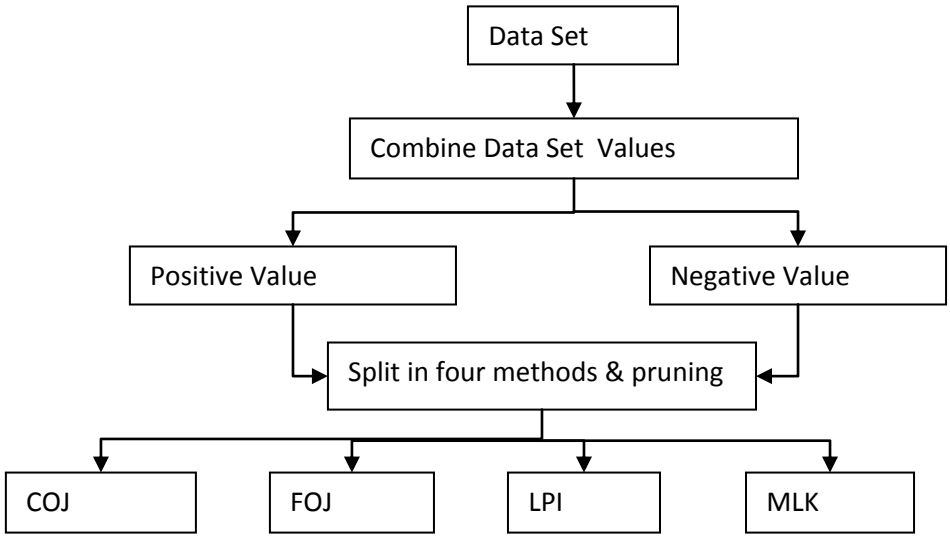


Fig 1: Use Case Diagram

4. SHARDING AND PRUNING METHODS

In one to many data Linkage for accuracy of tree we use pruning, it is pre pruning and post pruning. Pre pruning methods are MLE & LPI. In Linkage process there is a testing phase each pair of record in the testing set is cross valid against linkage model. In output true match score is calculated using MLE. Pair match in threshold value is FPR False positive rate & TPR True positive rate. OCCT evaluated the domains are: leakage prevention, recommender system, and fraud detection.

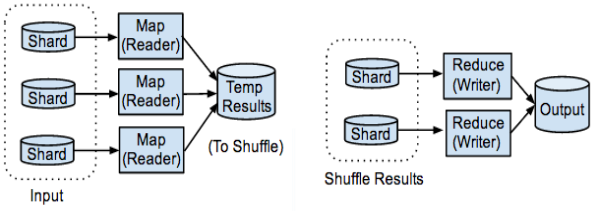
A. Subsections

The heavy data produces a great challenge classification and for data processing. In order to arrange the data, there were several algorithm suggested to accurate cluster of the data. One of that is the random forest algorithm, which is used for the feature subset selection. The feature selection involves identifying a subset of the most useful features that construct the performing results as the original complete set of features. It is catalogue after arranging the given data. The efficiency is calculated based on the time required to find a subset of features, the successful in producing a desired result is related to the quality of the subset of features. The system deals with fast in less time clustering based feature selection algorithm, which is demonstrate to be strong, but when the size of the dataset growing rapidly, the current algorithm is found to be less productive quality as the clustering of datasets takes quiet more number of time.

B. Sharding : Parallel Processing

Sharing is parallel processing in which stage of inputs are further spread in multiple sets. The data sets are called as shards are processing parallels. In the sharding concept table will be splitter in row and column form horizontally and vertically, each splitter data set called shard have been store or get memory in different server. Partition of row and column is done with help of primary key that also called as a index or also by other. Master- Slave replication done by sharding to deliver better high availability without address scalability. Here normalization concept again dominated because of sharding process as they are not improve efficiency of the frequent output of the huge table. Same type of data are placed in the particular data sets or the related data are making the group queries are minimal Form in nodes are less-expensive commodity machines. The capacity of shards function in the databases, is ideal for applications. But in database data are in various forms and shapes, object

oriented, relational, partitioning, replicated, hierarchical, in-memory, they occup the same machine. In the internet form Actions range from deleting creating of accounts, searching a phrase, rating commenting, to just clicking an advance, scrolling through a page. No, navigating a link in database can record the volume of data provoked. MySQL is also the form of Cluster that is a real-time open source transactional distributed database designed for quick, always –on access to data under high throughput conditions. MySQL, Cluster is a form of Shards all are the datasets. So sharding is build for high read write ratio & quick scalability and My SQL is designed for fast response, reliability & handling high read write ratio.



C. Pruning

Pruning is a method in which accurate tree is produce. Decision tree has prepruning & postpruning approach.

In prepruning current node of the tree is much beneficial so have not begin splitter more & in postpruning is a bottom up process to find the non beneficial branch, and its grown completely. We use prepruning in the MLE & LPI method. Leaf representation is shown a s a set of probability of attribute in models[6]. It is not required to save model for all possible attributes of particular Table. If the attribute have more significant effect then of particular leaf then only models are created for that feature selection process is used that represent by Weka's J48 decision trees as probability estimation measurement.

5. EVALUATION

1. Examine where the 4 methods of splitting criteria after pruning which one is suitable.
2. Looking for the binary tree and OCCT for one to many data linkage [7].
3. Verifying the suggested method that is universal, it is used for data linkage under various framework and implement on different domain.

A. The Database Misuse Domain

The purpose of using OCCT in this domain is to link a set of records, that shows the idea of request a set of data illustrated that can be actionable within the specific idea. Misuse of the domain will be explaining in example first malicious be like if employee searching for costumer record at opening hours and another malicious be searching record at closing hours.

B. The Movie Recommender Domain

Now days recommender system is very useful to comparing rating in different session and season. Similarly we talk about a system which work for the rating in different attributes like programmer if choose the rating in drama, comedy, crime, romance then again by artist and student may be male- female there thought are different so they are going to rate differently[11]. So that rating can be calculated by CGJ, FGJ, LPI & MLE method [12]. Further we have to compare this methods which will be best fit.

D. The Fraud Detection Domain

Fraud detection is to identify the factor that can lead to fraud, the persons who intentionally act secretly for their own benefit [8]. In this development age the new technologies has also provided various alternative in which criminals may commit fraud done in online business, reengineering, reorganization, credit cards, debit cards etc.

Advantage:

We suggested a new one-to-many data linkage procedure that links between entities of many more type of natures. The suggested procedure is based on a one-class clustering tree (OCCT) which distinguish the attributes that should be connect together.

We suggest four subdivided criteria and two different pruning procedure, which can be used for persuade the OCCT. This procedure was assessed to use this datasets from three dissimilar domains.

The output for the requested agreement is potent powerful of the suggested method and represent that the OCCT yields better achievement in terms of exactness accuracy and recall (in most cases it is statistically noteworthy) when differentiate, to a C4.5 decision tree-based linkage method.

Disadvantage:

We contemplate appraise the entity resolution (ER) problem (also known as reduplication, or merge--purge), in which records possessing to show the same real-world entity are accomplish located and merged. We structured the generic ER problem, it behave towards the functions for comparing and merging documents as black-boxes, which officially allow to effective conveying and substantial ER solutions.

We recognize four important properties that, if pleased by the merge and match functions, authorize much more optimal ER algorithms. We develop three optimal ER algorithms: G-Swoosh is the example where the four methods do not hold, and R-Swoosh and F-Swoosh that exploit the four possessions collectively. F-Swoosh the sound produce by a sudden rush of air or liquid, in summation assumes data collection of the advantageous example recorded entities may be used by the function match () .

We practically calculate the result by algorithms using distinguish different shopping data from Yahoo! Shopping and hotel information data from Yahoo! Travel. We also show that R-Swoosh and F-Swoosh that also be used even when the four merge and match features do not hold, if an "al mostly " result is sustainable.

Algorithm:

Lets have a class to be show that {C1, C2,…, Ck}. Three possibilities for the data set for training samples In the decision tree T is the given tree node:

1. In the decision tree T is the training sample which hold one or more samples, which is associated to a single class Cj. The decision tree for T is a leaf recognize class Cj.

2. T accommodate no samples. The decision tree is once more a leaf, but the class to be belonging with the leaf must be governed from learning sequential data other than T, such as the overall larger part of class in T. C4.5 algorithm uses as a principle or standard the most customarily class at the parent of the given node.

3. T accommodate samples that associate to a blend classes. In this set of circumstances, the project proposal is to filter cleanse T that is subdivided into subsets. This sample which is subdivided that are set as a title concern single-class grouping of samples. An suitable proper test is selected, based on single allotted element, that has one or more mutually dependent completely productive results {O1, O2, …,On}: T is segregation into subsets T1, T2, …, Tn where Ti hold all the specimen in T that have resulted Oi of the selected test. The decision tree for T composed of a decision node recognize the test and one branch for each possible productive result.

6. CONCLUSION

OCCT, a one-class decision tree process towards accomplish one -to- many and many -to- many data linkage using different methods like CGJ, FGJ, LPI & MLE for one to many data linkage & MapReduce function is for many to many data linkage presented in the paper. The initiated method is based on a one-class decision tree model that encapsulates the knowledge with records should be linked to each other. Implementation using all four method and after comparing them for suitable output & Map Reduce will diminish the prosecution time of many-to-many data linkage. It extent parallelism as linkage is implement in a distributed environment. The proposed method will be very efficient for large datasets.

We are showing here OCCT, a one-class decision tree closer to execute one-to-many and many-to-many data linkage. The suggested procedure is based on a one class decision tree model that enclose the apprehension of which documentation should be connected to each other. In summation, we proposed four possible splitting criteria and two feasible pruning methods that can be used for persuade the data models. Our judgment results show that the suggested algorithm is implied powerful when it is use in many other domains . Our goal is to link a documentation from a table TA with documentation from another table TB. The produce model is in the form of a tree in which the inner nodes demonstrate attributes from TA and the leafs hold a dense demonstration of a subset of documentation from TB which are more likely to be linked with a record from TA, whose values are dependent on whether to the path from the root of the tree to the leaf.

7. REFERENCES

[1] S Mallela, Dhillon I.S and D.S Modha, " Co – Clustering and information – Theoretic" Prof. SIGKDD Ninth ACM Int'l Conf. Data Mining and knowledge Discovery , pp. 89-98, 2003.

[2] Dr. Anju Singh, Dr. Divakar Singh, Gopal Patidar " Document Clustering approach using Hebbian-type Neural Network and Agglomerative Clustering " vol. 75, issue 9, 2013.

[3] , A.K. Elmagarmid, M. Yakout, H. Elmeleegy, M. Quzzani, and A.Qi, " Record Linkage Behaviours," Proc. Endowment VLDB, vol. 64, no 328.

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

[5] Baxter Rf. And M Guha. Gurprit L., " Linkage record which based on Decision Models ," Data Mining, , pp. 146-169. vol. 3755

[6] Goshair R. k. and Christaen p. , "Complexity for Reduplication in Data Mining and Quality Measures of Linkage Data in Data Mining ," pp. 127-151, 2007 vol. 43,.

[7] Christean P. , " Indexing Techniques survey for Scalable Reduplication and Linkage Record ," IEEE Transmission . and Data Eng and Knowledge., doi:10.1109/TKDE. 2011.127. vol. 24, no. 9, pp. 1537-1555, Sept. 2012,

[8] Dr. Divakar Singh" Intrusion Detection based System on Probabilistic Neural Network and Fuzzy C Means Clustering , D Singh – 2013 vol. 74, issue 2, pp. 30-33

[9] Grahmens A. " A Data Mining Decision Tree Recommender smart System ," Proc. 10th Int'l Conf Community Services . Innovative Internet, pp. 170-179, 2010.

[10] Flach P. , Ferri C., M. Guha and Herna´ndez-Orallo J., "DecisionTrees Model showing the Use of Area under curve ROC . Machine Learning, pp. 139-146, 2002.

[11] Gopandi N., Korean Y. , and Lempeal R., Adaptive internet based business or other enterprise Systems Using Decision Trees updated," Proc. Fourth ACM Int'L Conf. for Data Mining and Web Search , pp. 595-604, 2011.

[12] Adomavicius G. and Tuzhilin A. Its for the Next Generation of Data Mining Recommender smart Systems: A Survey of the Possible Extensions and State-of-the-Art Data Engineering and IEEE T Knowledge Transmission ., vol. 17, no. 6, pp. 739-749,June2005.