Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 11, November 2014, pg.300 – 304

RESEARCH ARTICLE

IMPLEMENTING ONE TO MANY DATA LINKAGE USING ONE CLASS CLUSTERING TREE

M.REENA BANU[1], S.SUBASHINI[2]
PG STUDENT, ASSISTANT PROFESSOR
NPR COLLEGE OF ENGINEERING AND TECHNOLOGY, TAMIL NADU, INDIA
EMAIL: *mreena891@gmail.com* , *balasundar1975@gmail.com*

Abstract: ***The task of data linkage is performed among entities of the same type. The one to one data linkage links one record from one table and another one record in another table. It is extremely necessary to develop linkage techniques that link between matching entities of different types and also to improve one to one linkage to one to many data linkage as well. The proposed method emphasizes on one-class clustering tree (OCCT). This method characterizes the entities that should be linked together. This method enables easy understanding and transformation of the clusters into association rules. The association rules indicate that the inner nodes consist of only the features describing the first set of entities, while the leaves of the tree represent features of their matching entities from the second data set. The four splitting criteria, coarse grained jaccard coefficient , fine grained jaccard coefficient, least possible intersection and two different pruning methods which can be used for inducing the OCCT.***

Index Terms— Clustering, classification, data matching, decision tree induction

1. INTRODUCTION

Data mining refers to extracting or “mining” knowledge from large amounts of data stored either in data warehouses or other information repositories like database. Data mining has mostly get an important area for research. One-to-many information linkage is an important task in several domains. The projected

technique is predicated on a one-class clump tree (OCCT) that characterizes the entities that ought to be joined along. The tree is made such it's simple to know and remodel into association rules, i.e., the inner nodes consist solely of options describing the primary set of entities, whereas the leaves of the tree represent features of their matching entities from the second data set.

The proposed method is based on a one-class clustering tree (OCCT) that characterizes the entities that should be linked together. The OCCT was evaluated using data sets from three different domains: data leakage prevention, recommender systems, and fraud detection. In the data leakage prevention domain, the goal is to detect abnormal access to database records that might indicate a potential data leakage or data misuse. The goal is to match an action, performed by a user within a specific context, with records that can be legitimately retrieved within that context. In the recommender systems domain, the proposed method is used for matching new users of the system with the items that they are expected to like based on their demographic attributes. In the fraud detection domain, the goal is to identify online purchase transactions that are executed by a fraudulent user and not the legitimate user (i.e., identity theft). The results show that the OCCT performs well in different linkage scenarios.

2. LITERATURE SURVEY

2.1. A Decision Tree Based Recommender System

A new method for decision-tree-based recommender systems is proposed. The proposed method includes two new major innovations. First, the decision tree produces lists of recommended items at its leaf nodes, instead of single items. This leads to reduced amount of search, when using the tree to compile a recommendation list for a user and consequently enables a scaling of the recommendation system. The second major contribution of the paper is the

splitting method for constructing the decision tree. Splitting is based on a new criterion - the least probable intersection size.

2.2 Learning Decision Trees Using the Area under the ROC Curve

Here we show how a single decision tree can represent a set of classifiers by choosing different labellings of its leaves, or equivalently, an ordering on the leaves. In this setting, rather than estimating the accuracy of a single tree, it makes more sense to use the area under the ROC curve (AUC) as a quality metric. We also propose a novel splitting criterion which chooses the split with the highest local AUC.

2.3. Lifestyle Finder: Intelligent User Profiling Using Large-Scale Demographic Data

Number of approaches have been advanced for taking data about a user's likes and dislikes and generating a general profile of the user. These profiles can be used to retrieve documents matching user interests; recommend music, movies, or other similar products; or carry out other tasks in a specialized fashion. This article presents a fundamentally new method for generating user profiles that takes advantage of a large-scale database of demographic data. These data are used to generalize user-specified data along the patterns common across the population, including areas not represented in the user's original data. I describe the method in detail and present its implementation in the LIFESTYLE FINDER agent, an internet-based experiment testing our approach on more than 20,000 users worldwide.

2.4. Modeling User Search Behavior for Masquerade Detection

Masqueraders will likely not know the system and layout of another user's desktop, and would likely search more extensively and broadly in a manner

that is different than the victim user being impersonated. We identify actions linked to search and information access activities, and use them to build user models. The experimental results show that modeling search behavior reliably detects all masqueraders with a very low false positive rate of 1.1%, far better than prior published results. The limited set of features used for search behavior modeling also results in large performance gains over the same modeling techniques that use larger sets of features.

2.5. Semtree: Ontology- Based Decision Tree Algorithm for Recommender Systems

Recommender systems play an important role in supporting people when choosing items from an overwhelming huge number of choices. We are modeling user preferences with a machine learning approach to recommend people items by predicting the item ratings. Specifically, we pro- pose SemTree, an ontology-based decision tree learner, that uses a reasoner and an ontology to semantically generalize item features to improve the effectiveness of the decision tree built. We show that SemTree outperforms comparable approaches in recommending more accurate recommendations considering domain knowledge.

3. CONCLUSION

The main aim is based on a one-class clustering tree (OCCT) that characterizes the entities that should be linked together. Here we use four splitting criteria and two different pruning methods which can be used for inducing the OCCT. The method was evaluated using data sets from three different domains. The results affirm the effectiveness of the proposed method and show that the OCCT yields better performance in terms of precision data linkage method aimed at performing one-to-many linkage that can match entities of different types.

The modified entropy formula that considers the weight of the positive class in the given data set and assumes the number of negative examples in the unlabeled data according to the given distribution. Each leaf represents a cluster, while the characteristics of the cluster are represented by a set of rules.

REFERENCES

[1] A. Gershman et al., "A Decision Tree Based Recommender System," Proc. 10th Int'l Conf. Innovative Internet Community Services, pp. 170-179, 2010.

[2]C. Ferri, P. Flach, and J. Herna´ndez-Orallo, "Learning Decision Trees Using the Area under the ROC Curve," Proc. Ninth Int'l Conf. Machine Learning, pp. 139-146, 2002.

[3]B. Krulwich, "Lifestyle Finder: Intelligent User Profiling Using Large-Scale Demographic Data," Artificial Intelligence Magazine, vol. 18, no. 3, pp. 37-46, 1997.

[4] M.B. Salem and S.J. Stolfo, "Modeling User Search Behavior for Masquerade Detection," Proc. 14th Symp. Recent Advances in Intrusion Detection, 2011.

[5]A. Bouza, G. Reif, A. Bernstein, and H. Gall, " Sem-tree: Ontology- Based Decision Tree Algorithm for Recommender Systems," Proc. Int'l Semantic Web Conf., 2008.

