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. 4, Issue. 3, March 2015, pg.56 – 60

RESEARCH ARTICLE

IDT: MINING DATA STREAMS USING INDETERMINATE DECISION TREE ALGORITHM

Vinodhini.K, Manju Bala

Dept. of Computer Science and Engineering, IFET College of Engineering, Villupuram, Tamilnadu, India

Dept. of Computer Science and Engineering, IFET College of Engineering, Villupuram, Tamilnadu, India

vinoguberan@gmail.com

pkmanju26@gmail.com

Abstract— One-to-many data linkage is an important task in many domains, yet only a handful of prior publications have addressed this issue. The technical challenge in implementing one to many data linkage using one class clustering tree is it do not support uncertain data in the sense of real time incoming data and it support only for data mining but not for data stream. The system investigates the problem of implementing uncertain data conditions and supports the data linkage for the data stream. The tree is built using Indeterminant decision tree algorithm such that it is easy to understand and transform. There are two type of pruning method used for inducing the indeterminate decision tree algorithm. They are Basic pruning and end point pruning. The system gives more efficient result in terms of number of iteration when compared to one class clustering tree based data linkage.

Index terms- Data stream, Indeterminate decision tree

I. INTRODUCTION

Data linkage [11], is the important process in the data mining because we have large number of data which are collected from the huge information. The data will be presented in the different data collections and stored in the different tables without the same entity. In that scenario there is a need of the enhanced data linkage process based on the decision making process in terms of the data clustering tree. In this paper we propose the new method for data linkage based on the one class clustering tree. And the enhanced pruning method based on the novel splitting criteria technique.

Nowadays, there is a need to found the techniques to link **the datasets that does not share the common identity**; it comes under the data linkage process. The data linkage's major task is to identify the different objects which were used to refer the same entity across the different source of data. It is much need in the combining different databases or like the preprocessing step process in the dataset oriented process. Data linkage is split up into two types they are one-to-one and one-to-many linkage. One-to-one type is used to combine the single object with the one data set with the single matching. Other type is one-to-many type which is used to combine one data set with the group of objects to the other data set.

II. RELATED WORK

The presentation of arithmetical revelation switch approaches for micro data is dignified in relations of the usefulness and the revelation risk related to the endangered micro dataset by Josep Domingo, Ferrer, and Vicenc Torra [1]. The interloper is expected to know an outside data set, whose annals are to be related to those in the endangered data set; the out of a hundred of properly related best pairs is an amount of revelation risk. This paper reviews conservative best connection, which shoulders communal variables between the external and the

© 2015, IJCSMC All Rights Reserved

endangered data sets, and then shows that record linkage. The drawback of this paper is it cannot takes place the experiments in the systematic way, they did not implement the process to suit for large number of records.

A one-class clustering tree (OCCT) [11] characterizes the entities that should be linked together. The tree is built such that it is easy to understand and transform into association rules, i.e., the inner nodes consist only of features describing the first set of entities, while the leaves of the tree represent features of their matching entities from the second data set. We propose four splitting criteria and two different pruning methods which can be used for inducing the OCCT.

S. Dhillon, Subramanyam mallela, Dharmendra S. Modha [2], present a groundbreaking co-clustering procedure that monotonically upsurges the conserved mutual info by interweaving both the row and column clustering sat all stages. Using the applied instance of concurrent word-document clustering, they prove that our procedure works well in practice, particularly in the attendance of thinly and high-dimensionality. The drawback here is it cannot consider the co-clustering for the joint distributions of the two random variables.

The demonstration how a single decision tree can signify a set of classifiers by selecting dissimilar classification of its verdures, or consistently, a collation on the leaves is done by [3] Cesar Ferri, Peter Flach, and Jose Hernandez-orallo. In this setting, rather than approximating the correctness of a single tree, it makes more sense to use the area under the roc curve (AUC) as a quality metric. They also propose a novel excruciating standard which chooses the split with the uppermost local AUC. The only disadvantage here is they did not consider the other learning methods that may use for the partition to get the instance space.

The paper given by Gediminas Adomavicius and Alexander Tuzhilin [4] labels various limits of present reference methods and deliberates likely postponements that can recover reference competences and make recommender systems appropriate to an even wider range of requests. These postponements include, among others, development of sympathetic of users and items, combination of the background info into the reference process, support for multi-criteria ratings, and delivery of more supple and less intrusive types of recommendations. The drawback of this paper is here they did not support for multi-criteria ratings, and they did not provide the provision based on the more flexible. Hendrik Blockeel, Luc De Raedt, Jan Ramon [5], introduced a Top-down induction of clustering tree it employments the values of example founded knowledge. The subsequent practice is applied in the tic (top down induction of clustering trees) scheme for first order gathering. The tic system employments the first order logical decision tree picture of the inductive logic programming sys-tem tilde. Various experiments with tic are presented, in both propositional and relational domains. But the disadvantage here is they did not consider the first order distance measures.

S.Mathew, M. Petropoulos, H.Ngo, And S.Upadhyaya [6] produced a paper in which they propose the new direction of obtaining the maliciously harvest data. By model the users access patterns in terms of profiling the data points based on the user access which is contritely analyze the query expressions. Their proposed data-centric approach is based on the key observation that has the query syntax alone is a poor discriminator of user intent, which is much better, rendered by what is accessed. They present a feature-extraction method to perfect users' access patterns. Statistical learning algorithms are trained and tested using data from a real graduate admission database. They did not refer directly to the elements of the base database schema in terms of the syntax. So this approach is does not impose the significant additional burden to the database server.

N.Golbandi, Y.Koren and R. Lembel [7] presents a technique for provoking info from new users in a way foremost to high excellence references, while reducing user effort in the procedure. The technique includes a meeting process, where users are asked for their sentiments on certain purposely selected crops. In order to achieve better correctness and user knowledge, the meeting procedure familiarizes to user replies, such that the response to a query influences the system's choice of the following question. But they did not consider the cost functions to improve the whole performance of the process. They did not attain the computation efficiency. A.Kamra, E.Terzi and E. Bertino [8] suggested a mechanism. Their method is founded on mining SQL queries stowed in database audit log files. The consequence of the mining procedure is used to form profiles that can perfect normal database admission behavior and classify intruders. They consider two different scenarios while addressing the problem. In the first case, they assume that the database has a role based access control (RBAC) model in place. Under a RBAC system permissions are associated with roles, grouping several users, rather than with single users. Drawback here is they did not consider the query semantics in terms of the data values which are used in the predications and the amount of the data accessed by the queries. They did not address the dynamic nature of the underlying data in the database.

The modeling of user search conduct to detect nonconformities representative a masquerade attack is introduced by [9] M.B.Salem and S.J.Stolfo. They imagine that each separate user knows their own le scheme well

© 2015, IJCSMC All Rights Reserved

sufficient to search in an incomplete, beleaguered and sole fashion in order to and info relevant to their present task. Masqueraders, on the other hand, will likely not know the le scheme and plan of another user's desktop, and would likely search more extensively and broadly in a manner that is different than the victim user being mimicked. They classify actions related to search and info access activities, and use them to build user models. The drawback here is they did not resolve the efficient detection issue when the attacker has the knowledge about the victim's behavior. They did not implement the monitoring strategy to counter the evasive tactic. G. Adomavicius and A. Tuzhilin [10] produced a paper which presents an overview of the field of recommender systems and describes the current generation of recommendation methods that are usually classified into the following three main categories: content-based, collaborative, and hybrid recommendation approaches. This paper also describes various limitations of current recommendation methods and discusses possible extensions that can improve recommendation capabilities and make recommender systems applicable to an even broader range of applications. Here the drawback is they have the overspecialization is not only that the content-based systems cannot recommend items that are different from anything the user has seen before. They would not be able to get accurate recommendations

III. EXISTING SYSTEM

In the existing system, they propose the SMV classifier it means the support vector machine that was mainly used to train to differentiate between the matching and non-matching records. To know the record pair for matching they used to calculate the expectation maximization process to calculate the probability of a record pair of the entities. In the other prior work they used to analyze based on the behavior between the matching records. But these methods assume that the data have same entities need to be appear in the two datasets to link that and they try to link the data based on the same entities on the two datasets. But they are not applicable to the data linkage of the data entities of the different types.

Existing system Disadvantages:

- The system did not support the process with the non- similar attribute values.
- They have time complexity.
- They have less accuracy.

IV. PROPOSED SYSTEM

In the proposed system, they are considering the clustering tree which is in particularly one class clustering tree. In one class clustering tree is highly preferable because it supports more efficiently the decision model. In this paper they propose the four splitting criteria and two pruning methods to obtain the efficient decision model. Here they considered the one-to-many linkage method which is obtained based on the one class clustering tree. In this method each node is considered as the clusters. Then the tree describes the hierarchy. Here we are applying the maximum likelihood estimation score based method as the splitting criteria and then have to apply the MLE pruning method to prune the data and to provide the efficient data linkage process in the non-presence of the same entities in the data.

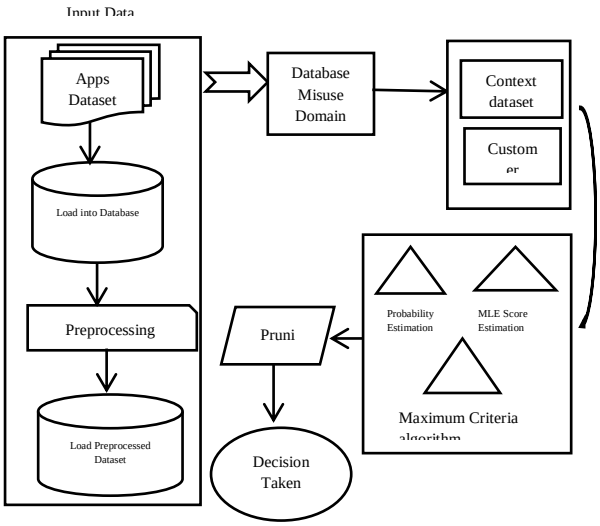

Figure 1 Our Proposed Architecture

Proposed System advantages:

- They can support the data linkage process with the non- similar attribute values.
- They have reduced time complexity.
- They have better accuracy.

V. IMPLEMENTATION

A. Dataset loading:

The dataset should be loaded into the database. Here we are going to implement is database misuse domain they are called as context dataset and the customer dataset. The context dataset contains the following attributes they are time_of_execution, day of execution, geographical location of the request, user's role, and type of request. The customer dataset contains the following attributes they are customer id, customers' fname, customers' lname, address, zip code, place of work, customer type. This dataset has to be prepared by our self by referring the reference papers. Then have to load this data into the database as to make the raw data to enter into the process.

B. Data removal:

The loaded datasets from the database which has to be considered for the further process by entering into the pre-processing process. The pre-processing process is to make the raw data to convert into the useable format. In basic it contains three ways of pre-processing they are data conversion, data cleaning and the data removing. In those three ways we are going to implement the data removal process to reduce the number process to enter into the decision taking model.

C. Probability and entropy:

The user will select the splitting criteria method which should be applied in the decision making process by linking the data. Here we use Maximum-Likelihood algorithm for splitting criteria. Then the attributes are split up by the similarity values between the clusters. Finally tree will be constructed.

D. Distribution tree:

The distribution tree is constructed based on the entropy value and probability which is found using maximum likelihood algorithm. The algorithm can be expressed by assuming X as a set of feature vectors, also called feature space and C as a set of classes. C: X! C is the ideal classifier for X. Create a node t for the tree. If all examples in D are positive, return the single-node tree t with label "+". If all examples in D are negative, return the single-node tree t, with label "–". Label t with the most common value of Target in D. If Attributes is empty, return the single-node tree t.

E. Double pruning:

In this module we are going to prune the constructed distributed tree based on the two proposed method with basic pruning and the end point sampling. Both are used to improve the pruning process in the constructed tree to improve the result of the classification. Both the pruning is used to reduce the error on the pruning process because the pruning is the main process to improve the classification result. Finally have to evaluate the process in terms of the time complexity and the space consumption.

VI. CONCLUSION

In this paper, we are proposing the novel method to link the data which does not have the common entity and also based on the clustering tree. Based on this we can match the two different cluster of data from the different dataset. That was the main challenge in the data linkage here we are applying the one to many data linkage technique with the one class clustering tree in particularly database misuse domain. That was takes place by the decision tree technique. Here each node will be considered as the cluster of nodes and the whole data on the different dataset will be matched as the result. Here we attain the improved efficiency of the data linkage process.

© 2015, IJCSMC All Rights Reserved

REFERENCES

[1] Josep Domingo, Ferrer, and Vicenc Torra Disclosure Risk Assessment in Statistical Micro Data Protection via Advanced Record Linkage, 2003.
[2] S. Dhillon, Subramanyam Mallela, Dharmendra S. Modha Information-Theoretic Co-Clustering,2003
[3] Cesar Ferri, Peter Flach, Jose Hernandez-Orallo, "Learning Decision Trees Using the Area under the Roc Curve", 2002.
[4] Gediminas Adomavicius and Alexander Tuzhilin, "Towards the Next Generation of Recommender Systems: A Survey of the State-Of-The-Art and Possible Extensions", 2005.
[5] Hendrik Blockeel, Luc De Raedt, Jan Ramon, "Top-Down Induction of Clustering Trees", 1998.
[6] S.Mathew, M. Petropoulos, H.Ngo, and S.Upadhyaya, "A Data-Centric Approach To Insider Attack Detection in Database Systems", 2009.
[7] N.Golbandi, Y.Koren And R. Lembel, "Adaptive Bootstrapping Of Recommender Systems Using Decision Trees", 2011.
[8] A.Kamra, E.Terzi and E. Bertino, "Detecting Anomalous Access Patterns in Relational Databases", 2008.
[9] M.B.Salem and S.J.Stolfo, "Modeling User Search Behavior for Masquerade Detection", 2011.
[10] Adomavicius and A. Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the State-Of-The-Art and Possible Extensions", 2005.
[11] Ma'ayan Dror, Asaf Shabtai, Lior Rokach, and Yuval Elovici, "OCCT: A One-Class Clustering Tree for Implementing One-to-Many Data Linkage", March 2014.

© 2015, IJCSMC All Rights Reserved