JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Sequence Clustering using Tree Structures

Dr. D. Mabuni

Assistant Professor
Dept. of Computer Science and Technology
Dravidian University, Kuppam, Andhra Pradesh, India

Abstract: Clustering is an important task in machine learning and data mining. Similarity measures are used for data clustering. Jaccard similarity measure considers only full length sequences during similarity computation discarding partial sequences even though similarity is present in partial sequences also. To overcome this problem a new sequence similarity finding measure is proposed in this paper and this proposed similarity measure is used for clustering tree structures. Particularly in the trajectory data and medical data representations and management full length sequences will be considered for finding similarity between any two sequences. In such cases Jaccard is inefficient and proposed new technique is more efficient and effective for full length as well as partial length sequence similarity finding before data clustering.

Index Terms: **partial length sequential branches, full length sequential branches, similarity between two trees, similarity between two structures, Jaccard similarity measure, similarity measures, Decision tree structures, Clustering.**

I. INTRODUCTION

Varieties of measures and metrics are available for finding similarity. Similarity between two structures particularly between two trees can be calculated using various metrics.Similarity based decision tree clustering is an important technique and it is predominantly used in trajectory data clustering, medical diagnosis and pattern recognition real world problems. Clustering of decision trees is one of the most important clustering techniques and frequently used in many real time applications such as medical diagnosis data clustering and trajectory data clustering. In day to day applications comparison of two data structures is inevitable for many applications correspondingly many researchers are continuously trying to find better and better state of the art techniques in the proper management of data and effective decision making situations.

Comparison of two Decision Trees:

Two relevant datasets constructed from the same domain that might have been taken at different times, generally, might result in two different decision trees. Then the question herethat may arise is how much similar these two selected decision trees are. If the two models are not similar then something has been changed in the dataset. A useful measure that gives an idea to judge the similarity value between two trees would be definitely useful in many areas especially in medical field and vehicle trajectory data management.

Advancements in intelligent medical diagnosis systems have made huge amounts of medical data available through many automatic and reliable data collection methods. A big part of this data is stored as sequences of collection of attributes and all these sequences belongs to a particular person can be stored in a tree data structure for efficient and effective management of medical data through intelligent data clustering and classification machine learning techniques. Automatic analysis and management of these individual and aggregated data with very minimal human supervision would both lower the costs and eliminate subjectivity of the data analysis. Medical diagnosis data clustering is an unsupervised task.

A new data clustering technique is proposed to automatically cluster the medical diagnosis data and trajectory data.

A central and very important part of medical diagnosis data clustering problems is choosing a correct and exactly suitable measure of similarity or distance.

II. RELATED WORK

Data clustering is one of the important data analysis techniques in machine learning, data mining and big data analytics. Large numbers of machine learning techniques are available for data clustering. Cluster data analysis techniques are mainly used for grouping data objects using cluster similarity finding measures. In supervised clustering number of clusters is predetermined and in unsupervised clustering number of clusters is not known at the beginning.

Various similarity measures that are used for finding similarity between two structures are:

1. Bhattacharyya distance
2. Euclidian distance
3. Longest common subsequence LCSS
4. Citi block distance
5. Cosine similarity
6. Jaccardsimilarity coefficient for similarity measure
7. Jaccard distance for dissimilarity measurement

Jaccard coefficient or Jaccard similarity measures the similarity between two sets of attributes or two sets of sequences, here, A and B are two sets of sequences of attributes.

$$Jaccard\_Similarity(A, B) = \frac{n(A \cap B)}{n(A \cup B)}$$

Jaccard coefficient value range lies between 0 and 1, 0 means no similarity and 1 means 100 percent similarity.The Jaccard distance measures the dissimilarity between two datasets and is calculated as:Jaccard distance = 1 – Jaccard Similarity.

In the literature of data clustering many similarity measures are proposed for clustering including categorical attributes data clustering, numerical attributes data clustering and combination of both attributes also. [1] Rezaie and Saunier [1] clearly explained clustering algorithms, comparison of similarity measures, trajectory performance evaluation measures and proposed a new trajectory data clustering algorithm. Various types of machine learning techniques are available for trajectory data clustering.Castin and Frenay [2] developed new criteria for node data splitting and proposed a new clustering algorithm. V. Estruch et al. [3] proposed a new technique for decision tree creation with center split similarity measure. Here distance based splits are defined for data division. Perner [4] proposed a new method for comparing two decision trees and these decision trees are represented with patterns and sequences.Berikov et al. [5] proposed a new algorithm for creating a decision tree with special similarity based approach and experimental results have shown that proposed method is far better than the many of the existing techniques.Hajjej et al. [6] experimentally verified performances of various tree algorithms to check whether tree algorithms are suitable for medical diagnosis or not. Here, the main intention is to find alternate tools for medical data diagnosis.

Kim et al. [7] proposed a decision tree structure for the purpose of learning accurate non-parametric spatiotemporal sequences. This approach has many advantages such as easy to train, scalable, fast performance, robust, high accurate, and ability to learn sequential data.A. Lukina et al. [8] proposed a new algorithm for creating optimal decision tree classifiers using optimality principles and also experimentally proved that the proposed algorithm’s performance is better than the existing decision tree classifiers.Xing and Keoh [9] was conducted a brief survey for sequence classification and gene data sequences, protein data sequences, query log sequences, web sequences, and heart data sequences are considered for experimental purpose.D. Alekseeva et al.[10] regarding traffic domain various machine learning algorithms in the communications domain are studied vigorously and prediction performance details of different algorithms are experimentally verified. Various experimental conclusions are listed after sufficient comparisons of different machine learning algorithms used for trajectory management.

I. Ntoutsi et al. [11] Different similarity measures, semantic and syntactic, between two decision trees are studied thoroughly. Different forms of similarity estimation techniques are used in this study and the elements used in this study are – training datasets, testing datasets, probability distributions, statistical techniques, probabilities. Zhang and Shasha [12] are used edit distance for comparing two different tree structures and the main parameter used is the count of number of edit operations to convert one tree into another tree.Beam and Kohane [13] have identified that machine learning algorithms are needed for efficient handling of communication networks.Sun et al. [14] have investigated thoroughly the effectiveness of various machine learning models in terms of the selected parameters such as prediction accuracy and the computational time cost.

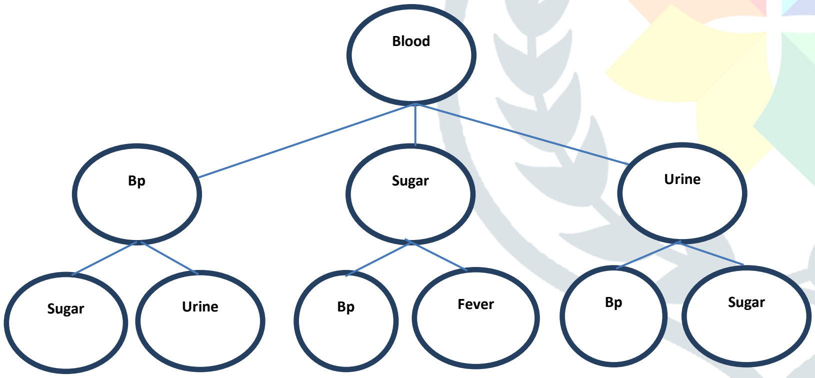
Authors [15] have created geometric decision trees before performing actual tree structure operations. Authors [16] have conducted in depth review on multivariate decision trees and many datasets are used during experimentation by considering different parameters such as number of attributes, number of tuples, and number of distinct classes.Authors [17] have tried for finding distances between numerical attributes also before data clustering.Decision tree creation method [18] was combined with distance based approach for data clustering and this technique follows first order logic. This idea is one type of hybrid technique useful for efficient data clustering.Two decision trees [19] are compared by comparing their rule sets and often this comparison gives an approximate measure about how good the two trees are. Clustering is an important data analysis task, B.Liu et al. [20] proposed a novel data clustering algorithm using supervised learning decision tree tool. Proposed algorithm produces true clusters and it is efficient in terms of handling high dimensional space.

III. PROBLEM DEFINITION

Decision trees or trees are used to represent sequences of attributes and then these trees are clustered using newly proposed cluster similarity measure. Initially, patient data or trajectory data is represented in treesbefore data clustering. In general,Jaccard similarity measure is used to find similarity between two sets of sequences of attributes and then data is clustered. Each branch of the tree represents a full sequence of attributes and Jaccard similarity measure is applicable to only these full sequences. The main disadvantage of the Jaccard similarity measure is that it does not measure partial sequences of branches. To overcome this problem a new technique is proposed and it can handle partial and as well as full branch sequences of attributes.

IV. MATHEMATICAL MODEL FOR CREATING CLUSTERS OF TREES

A tree is a nonlinear data structure that contains many branches and each branch represents one full sequence of attributes. To find similarity between two structures the full sequences must be compared and Jaccard similarity measure is exactly suitable for this type of comparisons but not suitable for partial sequence comparisons. The proposed method intelligently handles the partial sequence comparisons for tree clustering. The proposed cluster similarity finding measure is explained by taking simple examples of trees. Both full and partial sequences of patients, X, Y and Z are shown the respective figures.



Blood
Blood, Bp
Blood, Sugar
Blood, Urine
Blood, Bp, Sugar
Blood, Bp, Urine
Blood, Sugar, Bp
Blood, Sugar, Fever
Blood, Urine, Bp
Blood, Urine, Sugar

Figure-1 Medical tests of Person X

Blood
Blood, Sugar
Blood, Bp
Blood, Urine
Blood, Sugar, Fever
Blood, Sugar, Bp
Blood, Bp, Sugar
Blood, Bp, Urine
Blood, Urine, Bp
Blood, Urine, Sugar

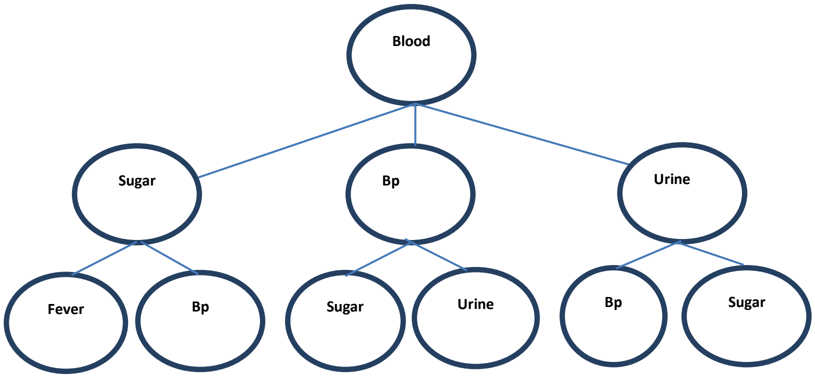
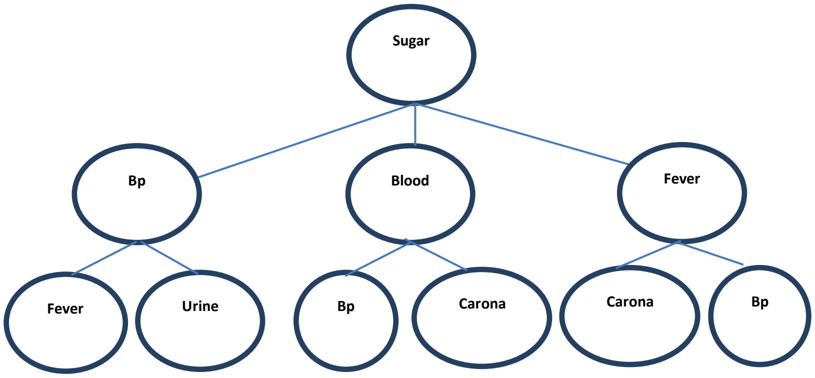


Figure-2 Medical tests of Person Y



Sugar
Sugar, Bp
Sugar, Blood
Sugar, Fever
Sugar, Bp, Fever
Sugar, Bp, Urine
Sugar, Blood, Bp
Sugar, Blood, Carona
Sugar, Fever, Carona
Sugar, Fever, Bp

Figure-3 Medical tests of Person Z

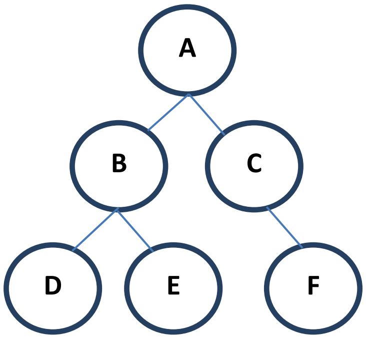
Full branch: Full branch is a tree branch consisting of all nodes from root node to a leaf node. In the Figure-1, {Blood, Bp, Sugar} is a full sequential branch.
Partial Branch: Part of any full sequential branch is called partial sequential branch. Here, {Blood} and {Blood, Bp} is a partial sequential branch.

Note that Jaccard similarity measures only similarity between full tree branches but not between partial tree branches. This is the main disadvantage of Jaccard similarity finding measure because sometimes sufficient similarity may present in partial branches also. In the present paper partial similarities between two different tree structure also taken into consideration in order to get full power of similarity measures.

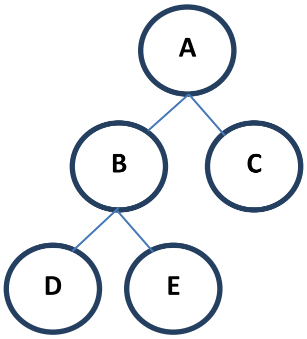
Consider three patients X, Y and Z and their tree structure representations of medical sequential tests. Figure-1, Figure-2 and Figure-3 represent medical diagnosis sequential records representations of patients X, Y and Z. Both full length and partial length sequences are shown right side of the trees.Before grouping patients into clusters their similarity measures are computed with the proposed new similarity finding formula. Similarity between X and Y = similarity (X, Y) = 1.0 and the similarity between X and Z = similarity (X, Z) = 0.1. Therefore, X and Y can be grouped into one cluster and X and Z cannot be grouped into one cluster. Proposed method similarity measures for X, Y and Z are computed as

$$Similarity(X,Y) = \frac{n\{X \cap Y\}}{n\{X \cup Y\}} = \frac{10}{10} = 1.0 \quad and \quad Similarity(X,Z) = \frac{n\{X \cap Y\}}{n\{X \cup Y\}} = \frac{2}{18} = 0.1$$

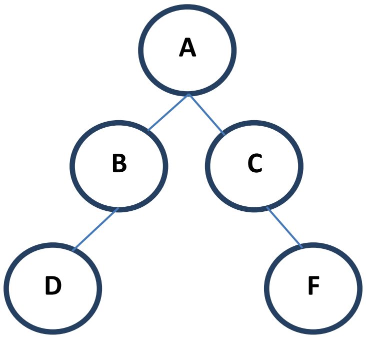
Now consider ten patients medical reports shown in T1 to T10 tree structures



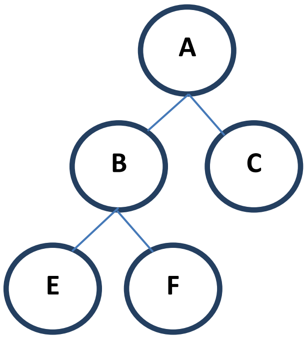
T1



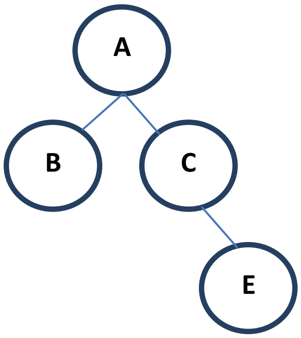
T2



T3



T4



T5

Jaccrd similarity measure computations for clustering tree structures are:				
T1	T2	T3	T4	T5
ABD ABE ACF	ABD ABE AC	ABD ACF	ABE ABF AC	AB ACE

Table-1 Full length sequences in Jaccard similarity

$$Jaccard\ similarity(T1,T2) = \frac{\{ABD, ABE, ACF\} \cap \{ABD, ACF\}}{\{ABD, ABE, ACF\} \cup \{ABD, ABE, AC\}} = \frac{2}{4} = 0.5$$

$$Jaccard\ similarity(T1,T3) = \frac{\{ABD, ABE, ACF\} \cap \{ABD, ABE, AC\}}{\{ABD, ABE, ACF\} \cup \{ABD, ACF\}} = \frac{2}{3} = 0.667$$

$$Jaccard\ similarity(T1,T4) = \frac{\{ABE\}}{\{ABD, ABE, ACF, ABF, AC\}} = \frac{1}{5} = 0.2$$

$Jaccard\ similarity(T1,T5) = 0$ $Jaccard\ similarity(T2,T3) = 0.25$

$Jaccard\ similarity(T2,T4) = 0.5$ $Jaccard\ similarity(T2,T5) = 0$

$Jaccard\ similarity(T3,T5) = 0$ $Jaccard\ similarity(T4,T5) = 0$

Proposed method similarity measure computations for clustering tree structures are:				
T1	T2	T3	T4	T5
A AB AC ABD ABE ACF	A AB AC ABD ABE	A AB AC ABD ACF	A AB AC ABE ABF	A AB AC ACE

Table-2 Full length and partial length sequences in the proposed similarity finding measure

$$\text{Proposed method } similarity(T1,T2) = \frac{\{A, AB, AC, ABD, ABE, ACF\} \cap \{A, AB, AC, ABD, ABE\}}{\{A, AB, AC, ABD, ABE, ACF\} \cup \{A, AB, AC, ABD, ABE\}}$$

$$= \frac{\{A, AB, AC, ABD, ABE\}}{\{A, AB, AC, ABD, ABE, ACF\}} = \frac{5}{6} = 0.83$$

$$\text{Proposed method } similarity(T1,T3) = \frac{\{A,AB,AC,ABD,ABE,ACF\} \cap \{A,AB,AC,ABD,ACF\}}{\{A,AB,AC,ABD,ABE,ACF\} \cup \{A,AB,AC,ABD,ACF\}}$$

$$= \frac{\{A,AB,AC,ABD,ACF\}}{\{A,AB,AC,ABD,ABE,ACF\}} = \frac{5}{6} = 0.83$$

$$\text{Proposed method } similarity(T1,T4) = \frac{\{A,AB,AC,ABE\}}{\{A,AB,AC,ABD,ABE,ACF,ABF\}} = \frac{4}{7} = 0.57$$

$$\text{Proposed method } similarity(T1,T5) = \frac{\{A,AB,AC\}}{\{A,AB,AC,ABD,ABE,ACF,ACE\}} = \frac{3}{7} = 0.43$$

$$\text{Proposed method } similarity(T2,T3) = \frac{\{A,AB,AC,ABD\}}{\{A,AB,AC,ABD,ABE,ACF\}} = \frac{4}{6} = 0.66$$

$$\text{Proposed method } similarity(T2,T4) = \frac{\{A,AB,AC,ABE\}}{\{A,AB,AC,ABD,ABE,ABF\}} = \frac{4}{6} = 0.66$$

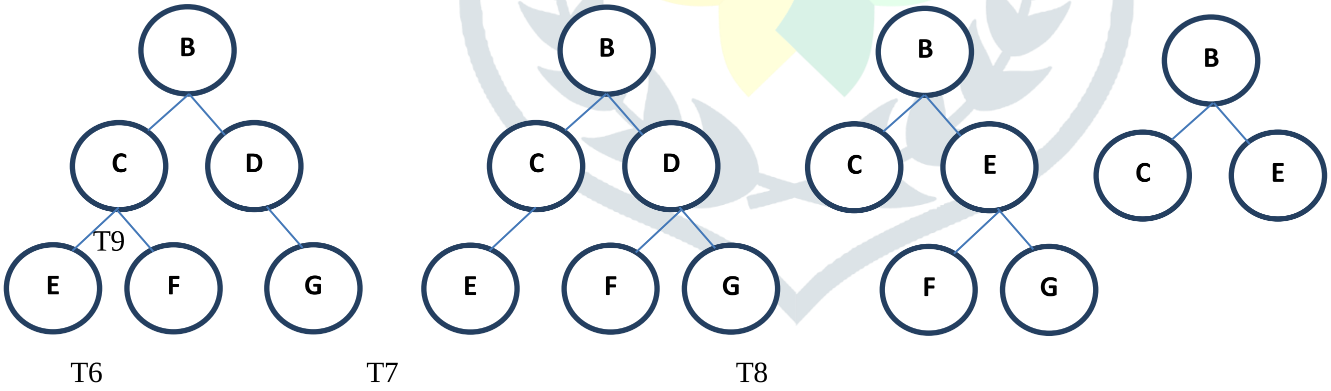
$$\text{Proposed method } similarity(T2,T5) = \frac{\{A,AB,AC\}}{\{A,AB,AC,ABD,ABE,ACE\}} = \frac{3}{6} = 0.5$$

$$\text{Proposed method } similarity(T3,T5) = \frac{\{A,AB,AC\}}{\{A,AB,AC,ABD,ACF,ACE\}} = \frac{3}{6} = 0.5$$

$$\text{Proposed method } similarity(T4,T5) = \frac{\{A,AB,AC\}}{\{A,AB,AC,ABE,ABF,ACE\}} = \frac{3}{6} = 0.5$$

When the specified threshold similarity measure is more than 0.7 then only trees T1, T2 and T3 are grouped into a single cluster but no other trees. Threshold similarity decides which trees must be clustered and which trees must not be clustered.

Now consider another set of trees with sequences representation



Jaccrd similarity measure computations for clustering tree structures are:			
T6	T7	T8	T9
BCE BCF BDG	BCE BDF BDG	BC BEF BEG	BC BE

Table-3 Full length sequences in Jaccard similarity

$$Jaccard\ similarity(T6,T7) = \frac{\{BCE,BCF,BDG\} \cap \{BCE,BDF,BDG\}}{\{BCE,BCF,BDG\} \cup \{BCE,BDF,BDG\}} = \frac{\{BCE,BDG\}}{\{BCE,BDG,BCF,BDF\}} = \frac{2}{4} = 0.5$$

$Jaccard\ similarity(T6,T8) = 0$ $Jaccard\ similarity(T6,T9) = 0$

$Jaccard\ similarity(T7, T8) = 0$ $Jaccard\ similarity(T7, T9) = 0$

$$Jaccard\ similarity(T8, T9) = \frac{1}{4} = 0.25$$

Clustering of trees is difficult because similarity measures are not high with respect to Jaccard method but trees T6 and T7 can be grouped into a single cluster with respect to proposed technique.

Proposed method similarity measure computations for clustering tree structures are:			
T6	T7	T8	T9
B BC BD BCE BCF BDG	B BC BD BCE BDF BDG	B BC BE BEF BEG	BC BE

Table-4 Full length and partial length sequences in the proposed similarity finding measure

$$\text{Proposed method } similarity(T6, T7) = \frac{\{B, BC, BD, BCE, BCF, BDG\} \cap \{B, BC, BD, BCE, BDF, BDG\}}{\{B, BC, BD, BCE, BCF, BDG\} \cup \{B, BC, BD, BCE, BDF, BDG\}}$$

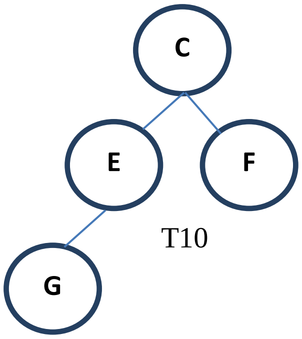
$$= \frac{\{B, BC, BD, BCE, BDG\}}{\{B, BC, BD, BCE, BCF, BDG, BDF\}} = \frac{5}{7} = 0.71$$

$$\text{Proposed method } similarity(T6, T8) = \frac{2}{9} = 0.22, \text{Proposed method } similarity(T6, T9) = \frac{1}{7} = 0.14$$

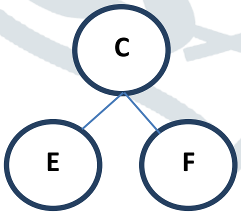
$$\text{Proposed method } similarity(T7, T8) = \tfrac{2}{9} = 0.22, \text{ Proposed method } similarity(T7, T9) = \tfrac{1}{7} = 0.14$$

$$\text{Proposed method } similarity(T8, T9) = \frac{2}{5} = 0.4$$

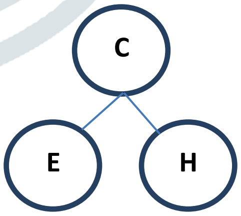
Consider third set of tree structures:



T10



T11



T12

Jaccrd similarity measure computations for clustering tree structures are:		
T10	T11	T12
CEG CF	CE CF	CE CH

Table-5 Full length sequences in Jaccard similarity

$$Jaccard\ similarity(T10, T11) = \frac{\{CEG, CF\} \cap \{CE, CF\}}{\{CEG, CF\} \cup \{CE, CF\}} = \frac{\{CF\}}{\{CEG, CF, CE\}} = \frac{1}{3} = 0.333$$

$Jaccard\ similarity(T10, T12) = 0$ $Jaccard\ similarity(T10, T11) = \tfrac{1}{3} = 0.333$

Jaccard similarity gives no chance for clustering because it considers only full branch sequences of attributes whereas the proposed method allows to group trees T10 and T11 into a cluster. So, proposed method is far better than Jaccard similarity.

Proposed method similarity measure computations for clustering tree structures are:		
T10	T11	T12
C CE CF CEG	C CE CF	CE CH

Table-6 Full length and partial length sequences in the proposed similarity finding measure

Proposed method $similarity(T10, T11) = \frac{3}{4} = 0.75$

Proposed method $similarity(T10, T12) = \frac{1}{4} = 0.25$

Proposed method $similarity(T11, T12) = \frac{1}{3} = 0.33$

CONCLUSIONS

There exist many similarity measures for data clustering. Jaccard similarity is one such measure for data clustering but its main disadvantage is that it takes into consideration only full sequences of attributes. To overcome this problem a new similarity finding measure is proposed which takes care of partial sequences of attributes also. In the future still better state of the art similarity measures will be investigated for data clustering.

REFERENCES

[1] M. Rezaie, N.Saunier, “Trajectory Clustering Performance Evaluation”, arXiv:2112.01570v1 [cs.LG] 2 Dec 2021.

[2] L. Castin, B. Frenay, “Clustring with Decision Trees”, ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.

[3] Estruch, V, Ferri C, Hernández-Orallo J,Ramírez-Quintana, M.J, “Similarity Functions for Structured Data”, de Inteligencia Artificial, vol. 10, núm. 29, primavera, 2006, pp. 109-121

[4] P. Perner, “How to Compare and Interpret Two Learnt Decision Trees from the Same Domain?”,Conference: Advanced Information Networking and Applications Workshops (WAINA), 2013 27th International Conference.

[5] V B Berikov, I APestunov, R M Kozinets and S A Rylov, “Similarity-based decision tree induction method and its application to cancer recognition on tomographic images”, Journal of Physics: Conference Series 1368 (2019) 052035 IOP Publishing doi:10.1088/1742-6596/1368/5/052035

[6]FahimaHajjej, Manal Abdullah Alohali, MalekBadr and Md Adnan Rahman, “A Comparison of Decision Tree Algorithms in the Assessment of Biomedical Data”, HindawiBioMed Research International Volume 2022, Article ID 9449497, 9 pages.

[7] T. Kim, Y. Yue, S. taylor, L.matthews, “A Decision Tree Framework for Spatiotemporal Sequence Prediction”, KDD’15, August 10-13, 2015, Sydney, NSW, Australia, ACM 978-1-4503-3664-2/15/08.

[8] A. Lukina, E. Hebrard, J. Chan, J.Bailey, J. Stukey, “MurTree: Optimal Decision Trees via Dynamic Programming and Search”, Journal of Machine Learning Research 23 (2022) 1-47 Submitted 5/20; Revised 10/21; Published 2/22

[9] Z. Xing, J Pei, E. Keogh, “A Brief Survey on Sequence Classification”, SIGKDD Explorations, Volume 12, Issue 1,

[10] D. Alekseeva, N. Stepanov, A. Veprev, “Comparison of Machine Learning Techniques Applied to Traffic Prediction of Real Wireless Network”, IEEE, Received October 19, 2021, accepted November 14, 2021, date of publication November 22, 2021,

[11] Irene Ntoutsi, AlexandrosKalousis and YannisTheodoridis, “A general framework for estimating similarity of datasets and decision trees: exploring semantic similarity of decision trees”.

[12] K. Zhang and D. Shasha, “Fast algorithms for the editing distance between trees and related problems”, SIAM, Journal of Computation, 18:1245–1262, 1989.

[13] [11] A. L. Beam and I. S. Kohane, ‘‘Big data and machine learning in health care,’’ Jama, vol. 319, no. 13, pp. 1317–1318, Apr. 2018.

[14] P. Sun, N. Aljeri, and A. Boukerche, ‘‘Machine learning-based models for real-time traffic flow prediction in vehicular networks,’’ IEEE Netw., vol. 34, no. 3, pp. 178–185, May/Jun. 2020.

[15] N. Manwani and P. S. Sastry, ‘‘Geometric decision tree,’’ IEEE Trans. Syst., Man, Cybern, B, Cybern., vol. 42, no. 1, pp. 181–192, Feb. 2012, doi: 10.1109/TSMCB.2011.2163392

[16] L. Sifu, R.Monroy, “Review and Experimental Comparison of Multivariate Decision Trees”, IEEE, Received July 18, 2021, accepted August 1, 2021, date of publication August 3, 2021, date of current version August 12, 2021.

[17] J. Ramon and M. Bruynooghe, “A polynomial time computable metric between point sets”, ActaInformatica, 37(10):765–780, August 2001.

[18] HendrikBlockeel, Luc De Raedt, and Jan Ramon.Top-down induction of clustering trees.CoRR, cs.LG/0011032, 2000.

[19] Georg G, Séroussi B, Bouaud J,“Does GEM Encoding Clinical Practice Guidelines Improve the Quality of Knowledge Bases? A Study with the Rule-Based Formalism”, AMIA AnnuSymp Proc. 2003, pp. 254–258 (2003)

[20] B. Liu, Y. Xia, PS. Yu, “Clustering Through Decision Tree Construction”, IBM Research Report RC 21695 (97737), 20 March 2000.