© 2016 IJSRSET | Volume 2 | Issue 2 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

Tree Dataset Extraction Using HAC Based Algorithm

Satyanarayanan K S, Srikanth B, Murugesan M

Computer Science and Engineering, Dhanalakshmi College of Engineering, Tambaram, Chennai, Tamil Nadu, India

ABSTRACT

The main objective of this work is to formulate a trouble-free ways of fetching data and finding appropriate values and here we are using apparent concepts of data mining, which is an analytical process designed to explore enormous amounts of data typically business or market related. The existing system of Hierarchical Archimedean Copulas (HAC) has been exhaustively used in this project.

Keywords: HAC, Visualization, KDD, GOF

I. INTRODUCTION

Data mining is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data use. A side from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post processing of discovered structures, visualization, and online updating.

The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records. The data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation, nor result interpretation and reporting are part of the data mining step, but do belong to the overall KDD process as additional steps. KDD field is concerned with the development of methods and techniques for mining process is to extract information from a data set and transform it into an understandable structure for further making sense of data. At the core of the process is the application of specific data-mining methods for pattern discovery and extraction. Data mining is the application of specific algorithms for extracting patterns from data.

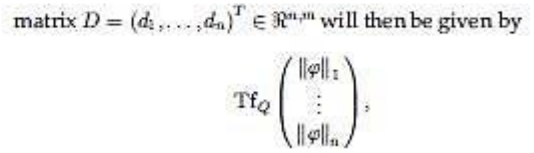
The work in this paper is to formulate a trouble-free ways of fetching data and finding appropriate values and here we are using apparent concepts of data mining, which is an analytical process designed to explore enormous amounts of data typically business or market related. Copula has been undeniably proven that is the sufficient condition for defining a HAC. These conditions allow us to define new generalized quantifiers, which are then intensively validated on all standard data set and one data set for real world application.

The paper concludes by comparing the proposed quantifiers to a more traditional approach-minimum spanning tree which is proficient to find the nodes within the best case. So there will be no need of long traverse and also HAC helps to cluster and embed the nodes in different categories. Copula modelling has become an increasingly popular tool in finance to model asset returns dependency.

In essence, copulas enable us to extract the dependence structure of the joint distribution function of a set of random variables and, at the same time, to separate the dependence structure from the univariate marginal behaviour. And also enables the user to have a wide view of considering the dependency of the distribution function of random process and makes easier ways to predict the result of the destination while before starting of indiscriminate concepts.

II. METHODS AND MATERIAL

In this section, we want the basic way of defining generalized quantifiers, which was recalled and exemplified in Section 2 and which consists in assessing the results of operations with the data, to be applied to the context of HACs. In particular, it will be applied to the description of data fulfilling some open formula' with a homogeneous binary HAC that is in a sufficient agreement with that data. The fact that the considered data should be determined by a single formula implies that we need to define unary generalized quantifiers to this end According to (2), the evaluation of Q' based on a data.



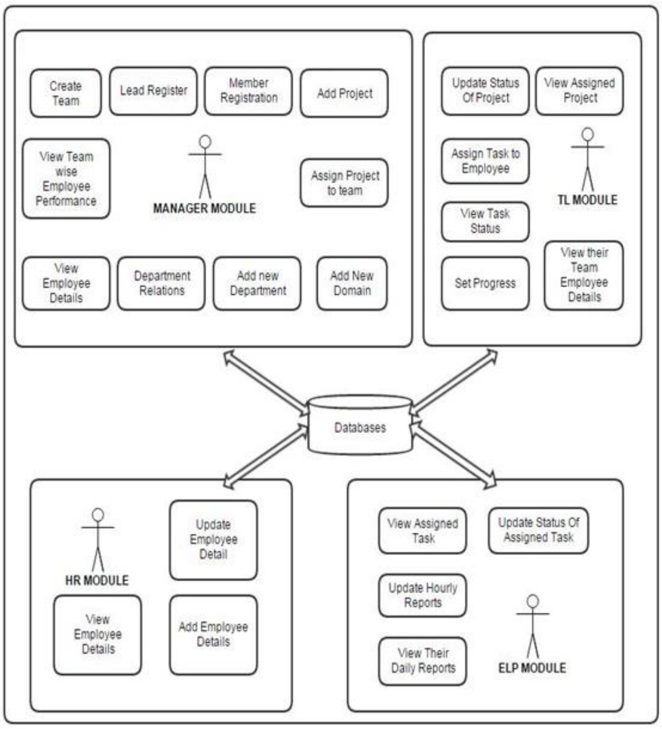
matrix $D = (d_1, \ldots, d_n)^T \in \Re^{n,m}$ will then be given by

$$\mathrm{Tf}_Q \begin{pmatrix} \|\varphi\|_1 \\ \vdots \\ \|\varphi\|_n \end{pmatrix},$$

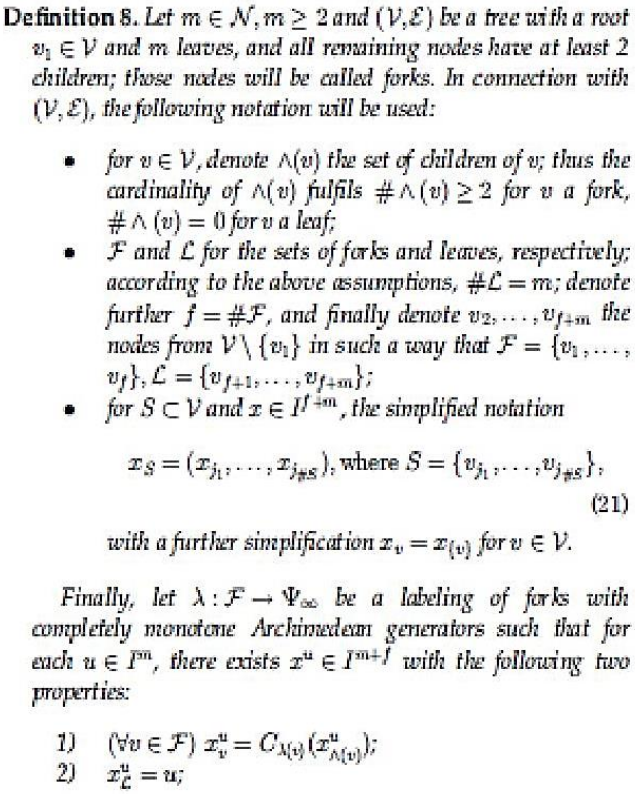
We restrict our attention to the case when the Archimedean copulas assigned to forks belong to a family fulfilling (32), and we are particularly interested in the following properties of the HAC describing the data:

i) The generators of the copulas in the forks belong to a given family;
ii) A given goodness of fit (GOF) test doesn't reject, at a given significance level, the agreement of the estimated HAC with the data;
iii) The leaves corresponding to particular random variables Xi; Xj are at most a particular distance L apart;
iv) Along the path between leaves corresponding to particular random variables, the Kendall's rank correlation coefficient achieves at least particular nonnegative threshold #.

A. System Architecture



B. Hierarchical Archimedean Copulas



Definition 8. *Let $m \in \mathcal{N}, m \geq 2$ and $(\mathcal{V}, \mathcal{E})$ be a tree with a root $v_1 \in \mathcal{V}$ and m leaves, and all remaining nodes have at least 2 children; those nodes will be called forks. In connection with $(\mathcal{V}, \mathcal{E})$, the following notation will be used:*

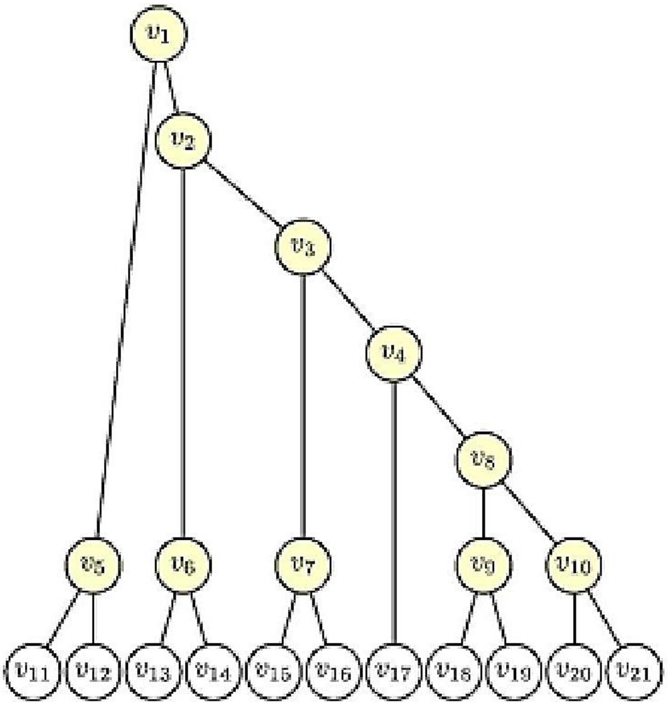
- *for $v \in \mathcal{V}$, denote $\wedge(v)$ the set of children of v; thus the cardinality of $\wedge(v)$ fulfils $\# \wedge (v) \geq 2$ for v a fork, $\# \wedge (v) = 0$ for v a leaf;*
- *$\mathcal{F}$ and $\mathcal{L}$ for the sets of forks and leaves, respectively; according to the above assumptions, $\#\mathcal{L} = m$; denote further $f = \#\mathcal{F}$, and finally denote $v_2, \ldots, v_{f+m}$ the nodes from $\mathcal{V} \setminus \{v_1\}$ in such a way that $\mathcal{F} = \{v_1, \ldots, v_f\}, \mathcal{L} = \{v_{f+1}, \ldots, v_{f+m}\}$;*
- *for $S \subset \mathcal{V}$ and $x \in I^{f+m}$, the simplified notation*

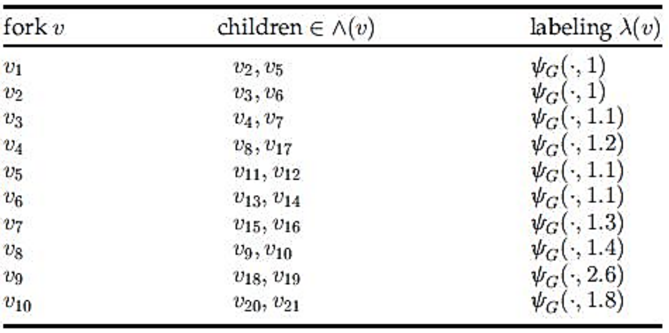
$$x_S = (x_{j_1}, \ldots, x_{j_{\#S}}), \text{ where } S = \{v_{j_1}, \ldots, v_{j_{\#S}}\}, \tag{21}$$

with a further simplification $x_v = x_{\{v\}}$ for $v \in \mathcal{V}$.

Finally, let $\lambda : \mathcal{F} \to \Psi_\infty$ be a labeling of forks with completely monotone Archimedean generators such that for each $u \in I^m$, there exists $x^u \in I^{m+f}$ with the following two properties:

1) $(\forall v \in \mathcal{F})\ x^u_v = C_{\lambda(v)}(x^u_{\wedge(v)})$;
2) $x^u_{\mathcal{L}} = u$;





fork v	children $\in \wedge(v)$	labeling $\lambda(v)$
v_1	v_2, v_5	$\psi_G(\cdot, 1)$
v_2	v_3, v_6	$\psi_G(\cdot, 1)$
v_3	v_4, v_7	$\psi_G(\cdot, 1.1)$
v_4	v_8, v_{17}	$\psi_G(\cdot, 1.2)$
v_5	v_{11}, v_{12}	$\psi_G(\cdot, 1.1)$
v_6	v_{13}, v_{14}	$\psi_G(\cdot, 1.1)$
v_7	v_{15}, v_{16}	$\psi_G(\cdot, 1.3)$
v_8	v_9, v_{10}	$\psi_G(\cdot, 1.4)$
v_9	v_{18}, v_{19}	$\psi_G(\cdot, 2.6)$
v_{10}	v_{20}, v_{21}	$\psi_G(\cdot, 1.8)$

C. Common Terminologies

Copula:

Copula modelling has become an increasingly popular tool in finance to model assets returns dependency. In essence, copulas enable us to extract the dependence structure from the joint distribution function of a set of random variables and, at the same time, to separate the dependence structure from the univariate marginal behaviour. And also enables the user to have wide view of considering the dependency of the distribution function of random process and makes easier ways to predict the result of the destination while before starting of indiscriminate concepts.

Spanning Tree:

A Spanning tree T of a undirected graph G is a sub graph that is a tree which includes all vertices of G. A Spanning tree of a connected graph can also be defined as maximal set of edges of G that contains no cycles or as a minimal set of edges that connect all vertices.

Minimum Spanning Tree:

A Minimum Spanning tree is a spanning tree of a connected, undirected graph that connects all the vertices together with the minimal total weighting for its edges.

III. RESULTS AND DISCUSSION

A. Proposed System

In this paper we have exhaustively used the concepts of Hierarchical Archimedean copulas (HAC) algorithm.

Here we have using three generalized quantifiers

HAC distance, HAC Kendall's, HAC Distance + Kendall's.

HAC distance quantifier is used to find the shortest distance to reach the required nodes.

HAC Kendall's quantifier is to find threshold for the value of Kendall's rank correlation coefficient.

Here we are using minimum spanning tree approach which helps to find the data rapidly even in a best case itself ,which means there will be no need long traverse and also less time consumption.

Therefore, reduces time delay, no need of long traverse to find the nodes etc…

Spanning trees a very important in designing efficient routing algorithms.

Spanning trees have wide applications in many areas, such as network design, etc.

Each time a step of the algorithm is performed, one edge is examined. If there are only a finite number of edges in the graph, the algorithm must halt after a finite number of steps. Thus, the time complexity of this algorithm is clearly O (n), where n is the number of edges in the graph.

B. Advantages

- Any non-trivial incidents can be analyzed easily.
- The type of incidents can be identified earlier based on the HAC.
- Minimum weight spanning techniques helps us to find the nodes within the shortest path.
- Using quantifiers fetching analysis structure.
- 1of data is trouble-free.
- Analyzing the output of organization is more efficient than previous system.
- HAC algorithm gives us efficient ways of predicting
- No apriority information about the number of clusters required.
- Easy to implement and gives best result in some cases.

IV. CONCLUSION

Efficient retrieval of records from a database has been an active research field for many years. We approach the problem from a real-world application point of view, in which the order of records according to some similarity function on an attribute is not unique.

We experimentally show that our method outperforms Hierarchical Archimedean Copulas (HAC) for the retrieval of those very common cases where we used with minimum spanning tree technologies.

Here in this Paper we are going to use tree data set extraction technique by which we can easily and efficiently access the data and the entire process is analysed with the minimum spanning tree technology in order to find the variations in the speed rate of data which has been fetched from the database.

V. REFERENCES

[1] O. Okhrin, Y. Okhrin, and W. Schmid, "Properties of hierarchical Archimedean copulas," Statist. Risk Model., vol. 30, pp. 21–54, 2013.

[2] O. Okhrin, Y. Okhrin, and W. Schmid, "On the structure and estimation of hierarchical Archimedean copulas," J. Econometrics, vol. 173, pp. 189–204, 2013.

[3] N. Whelan, "Sampling from Archimedean copulas," Quantitative Finance, vol. 4, pp. 339–352, 2004.

[4] A. McNeil, R. Frey, and P. Embrechts, Quantitative Risk Management: Concepts, Techniques, and Tools. Princeton, NJ, USA: Princeton Univ. Press, 2005.

[5] B. Choro_s, W. H€ardle, and O. Okhrin, "CDO pricing with multifactor and copulae models," in Proc. 57th Biennial Session Int. Statist. Inst., 2009, pp. 1–26.

[6] M. Hofert and M. Scherer, "CDO pricing with nested Archimedean copulas," Quantitative Finance, vol. 11, pp. 775–787, 2011.

[7] M. Hole_na and M. _S_cavnick_y, "Application of copulas to data mining based on observational logic," in Proc. Inf. Techonol. Appl. Theory Workshops, Posters, Tuts., 2013, pp. 77–85.

[8] J. G_orecki, M. Hofert, and M. Hole_na, "An approach to structure determination and estimation of hierarchical archimedean copulas and its application to Bayesian classification," J. Intell. Inf. Syst., vol. 44, 2015, Doi: 10.1007/s10844-014-0350-3.

[9] P. H_ajek, and M. Hole_na, "Formal logics of discovery and hypothesis formation by machine," Theor. Comput. Sci., vol. 292, pp. 345–357, 2003.

[10] A. McNeil and J. Ne_slehov_a, "Multivariate Archimedean copulas, d-monotone functions and l1-norm symmetric distributions," Ann. Statist., vol. 37, pp. 3059–3097, 2009