Pri-Tri: An Innovative Algorithm for Clustering Categorical Data in Data Warehouse

S.Hari Ganesh
Assistant Professor
Department of Computer Applications
Bishop Heber College
Trichy, India

Dr.C.Chandrasekar
Associate Professor
Department of Computer Science
Periyar University
Salem, India

ABSTRACT
In the process of data mining to extract knowledge from large data set needs great potential to extract the hidden nuggets. To cluster the numerical data there are enormous clustering technique. Data mining for categorical data(qualitative and quantitative) the most frequently used algorithms are k-means, k-mediods and fuzzy rule all these methods needs a threshold value to overcome this problem. This paper propose an algorithm to optimize the number of clusters and it also uses novel way to construct the data mart using the concept of multiprocessing Pri-tri algorithm.

Keywords
Data mining, clustering, k-means, Multiprocessing.

1. INTRODUCTION
In today's world the data kept in computer files and databases are growing at a phenomenal rate. Though there are more sophisticated tools, still people find it very difficult to fulfill their needs. Today's world many use new tools that are incorporated with different new algorithms to analyze and predict the future information more accurately and quickly. To extract protein pattern from the protein data bank(PDB), which is a huge growing database. In the medical field they need many a tools and new algorithms that would predict the functionality of the protein[2]. This paper tries to solve the pattern matching in a huge growing database [3]. In Data mining for categorical data the most frequently used algorithms are k-means, k-mediods and fuzzy rule[12].This paper deals with data warehousing and data mining so as to fulfill the companies need with the best algorithm to construct the data mart that is subset of the data warehouse. It proposes Pri-Tri algorithm which uses the concept of multithreading to construct the data mart house of the integrated data which is cost effective and time saving. However the data warehouse contains traditional and operational data that is being converted into information data. This paper proposes one new way to construct the data mart[8].

2. METHODOLOGY
This paper proposes one new way of constructing the data mart. The fig.1 depicts the traditional way of constructing data warehouse. The fig.2 represents the new methodology that is being adopted for constructing the data mart. It uses the concept of priority based algorithm and the concept of multi-processing in Pri-Tri algorithm. The algorithm implemented for analyzing snake venom, comparing its categorical features of the snake. The snake venom is identified through its protein sequence from the Protein Data Bank. Priority Table.1 shows the process Multiprocessing concept related to Dataset(DS) Versus Time Series(T).

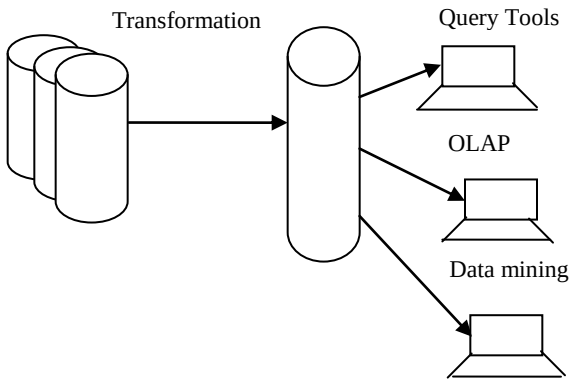
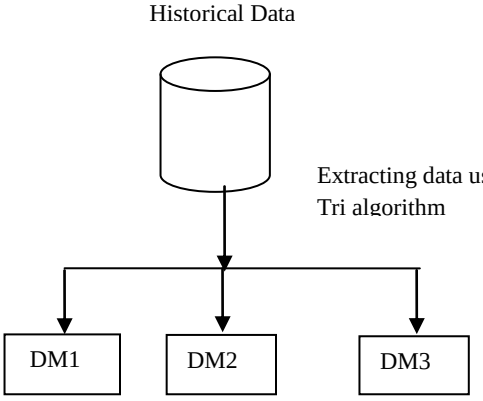


Fig 1: General Data warehouse



DM1-Data Mart1
DM2-Data Mart2
DM3-Data Mart3

Fig 2: Extracting informational data from the historical data through the data marts

The Generalized priority table of Pri-Tri data mining tool[3] is as follows

Table 1. Priority based on hits

	D1	D2	D3
T1	P1	P2	P3
T2	P3	P2	P1
T3	P1	P3	P2

T_i –Time series
D_i –Data records retrieving speed
P_i –Priority value of the thread

3. RELATED WORK

DSS (decision supporting system) is subject oriented, integrated time variant, non-volatile system. It need a single repository that contains the current and historical data set .Operational data represent the day to day data. Before going to the analysis of data mining, there is a need to analyze about data warehousing and its components like report query, EIS tools, OLAP tool [11], data mart, operational and external data, meta data, that could help us in future prediction and resolve the problems. In this section of the paper, it discusses the previous work on clustering categorical data. The EM (Expectation Maximization)[4][9] algorithm is a popular iterative clustering technique. Another algorithm is Sieving Through Iterated Relational reinforcement(STIRR,)[8] an iterative algorithm based on non-linear dynamical systems. This represent each attribute value as a weighted vertex in a graph. Starting with the set of weights on all vertices, the system is iterated until a fixed point is reached. Robust Hierarchical clustering with Links (ROCK)[6] an adaptation of an agglomerative hierarchical clustering algorithm, which heuristically optimizes a criterion function defined in terms of the number of links between tuples. Informally the number of links between two tuples is the number of common neighbours they have in the dataset Clustering CaTegorical data Using Summaries(CACTUS)[7] attempts to split the database vertically and tries to cluster the set of projections of these tuples to only a pair of attributes. In this paper a novel algorithm is proposed to cluster the categorical data, for constructing the data mart from the historical data through which Information can be gathered using the concept of Multi Threading.

3.1 Pri-Tri algorithm

The proposed data mining algorithm is as follows.
Step1:
From the historical data set d_i , where i=1,2,3..n extract the item set based on the sequential criteria
Step2:
Count the number of item set that is C_i, where i=1, 2 ,3..n
that is being fetched from the datasets for time series t_i ,where i=1 ,2 ,3..n
Step3:
Use the concept of multi-threading for processing of all the d_i , data sets.
Step4:
Create priority p_i, where i=1,2 ,3,..n concept multithreading for time series t_i which is dynamically changing for the period of time.
Step5:
Copy the highest priority item set count(c) dataset is taken for construction of the data mart.
Step6:
This dataset is stored in a data mart that contains the informational data.
Step7:This data mart checked for cost effectiveness and speed of retrieving records.
Step8: End.

3.2 Clustering in data mining:

The objectives of clusters are to uncover natural groupings, to initiate hypothesis about the data, find the consistent and valid organization of the data Clustering in a form of unsupervised then partitions the from observations are grouped into classes or clusters. The two broad categories and contain viz. hierarchical and partional. Partition clustering technique partition the data base into predefined number of clusters based on some criterion. Hierarchical clustering technique further divided in to agglometric and divisive. Basically principle of clustering things on concept of distance metric or similarity metric since the data are in real number for statistical tools and pattern recognition, a large class of metrical exists one can define one's own metric according to specific requirements. Clustering can be done for numerical and categorical data, numerical data the geometric properties like distance functions are used as criterion, for categorical data are connected to traditional data bases. The concept of similarity alone is not sufficient for clustering data along criterion based for the clustering of categorical data, attributes whose domain is totally ordered are called numeric, whereas attributes whose domain is not ordered are called categorical.

4. METHODOLOGY

4.1Pattern Match Calculations

The proposed similarity pattern value between the objects P_i and P_j is defined as follows:

Step 1**:** Find

$$Pat(P_i , P_j) = \sum_{k=1}^{m} \delta(X_k,Y_k) \text{ where } X_k \in P_i \text{ and } Y_k \in P \quad \rightarrow \text{(Equation 1)}$$

Step 2:
$\delta (X_k,Y_k) = 1$ if $X_k = Y_k$ otherwise 0 $\rightarrow$(Equation 2)

Step 3: To find the Threshold Value select the objects belonging to the same group, maximum value in each row is selected as a threshold value respectively.

Step 4: let there be 'k' Number of clusters .The unique categories from the domain of attributes are summed up as categorical count.

$\lambda = \sum C_{ij}/m$ →(Equation 3)
Step 5**:**End.

4.2 Proposed Algorithm

Assume the database with 'n' records and 'm' attributes. Construct a similarity matrix between objects using eqn. 1. Neighbours of object 1 are the objects with the similarity value equal to the threshold value of i^{th} object.

Input: Data set

Output: Clustered object

step1. Construct a similarity matrix using the definition of similarity measurement Pat(Pi, Pj) where $1 \leq i \leq n$, $1 \leq j \leq n$

step 2. Find maximum of each row of the similarity matrix, max., where $1 I \leq 1 \leq n$

step 3. Find neighbours of object i. near (I,1)= {x / x € pat(P_i, P_j) and x= max $_i$}

step 4.Group the objects in to clusters. Initially the first object is considered. Objects are selected from the list of objects until the list becomes empty. Select a neighbors of object "I" from nearmatrix, and the threshold value of the neighbour object 1 is compared with the threshold value of object "i" if the threshold value of 1 is less than or equal to the threshold value of object 'i'is placed in the cluster . This process is repeated for all the neighbours of object "i" .The objects which are placed in the cluster are removed from the objects list.

Cluster $_i$ ={ x / max_x, ≤ max_i and x € near(i) }

Step 5. Merge the clusters until the number of clusters are reduced to or less than or if the successive iteration results in the same number of clusters. The objects in the clusters are relocated to the cluster where the majority of its neighbors are clustered.

4.3 Similarity Check

Find the similarity matrix between object i and object j,

Pat (P_i , P_j) for all i and j.

step2. max_i = max.of Pat(P_i ,P_j) for all j.

step3. Find neighbors of object i. near(i,j) = {x / x Pat(P_i,P_j) and x = max_i}

step4. Group the objects in to clusters. $Cluster_k$ = {x / max_x <= max_i and x €neigh(i)}

step5. Merge the clusters until the number of clusters is reduced to 'λ' or less than 'λ' or less than k or if the successive iteration results in same number of clusters. The objects in the clusters are relocated to the cluster where the majority of its neighbors are clustered.

step6. Extract patterns

step7.end

The proposed algorithm is implemented using MatLab 6.1 and tested with the sample venom of snake data set available in Protein Data Bank(PDB)[10]using pubmed tool data repositories has been created. For comparison , this paper uses three protein extracting tool Swiss-prot, Blast and Fasta,

5. RESULTS AND DISCUSSIONS:

The snake venom dataset with 250 records samples and nine attributes resulted in 124 clusters first phase after iteration the number of cluster sample dataset reduced to 15 clusters in which 10(Table.2) are venomous snakes and 5 are non-venomous snakes(Table.3) Snake venom is taken for experiment which contains high protein and the snake can be identified by its physical structures(Fig.3) and classified into venomous and non-venomous snake. Snake venom can be broadly categorized into many types. They are as follows Hemotoxic, Neurotoxic , cytooxic,Myotoxic Venoms.

5.1 Features of venomous snakes can be categorically classified

The features of venomous and non- venomous snake can be observed by its body shape, head and neck shape, color, pattern, scale texture, eye pupil shape ,tail scales and anal plate division. This paper implements multi-threading[8] concepts in java environment, in net bean IDE and that will depict the process of the proposed algorithms

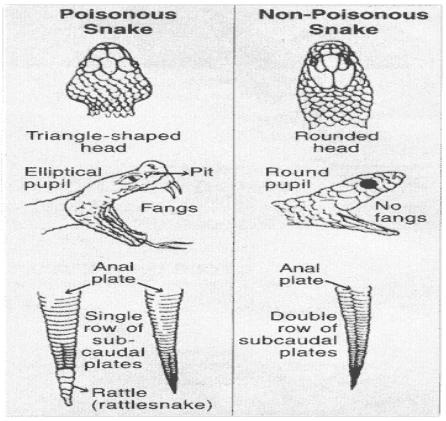


Fig 3: Features of venomous snakes

The Features are extracted from the venomous snakes with the protein sequence extracted (Table.4) so that a snake bitten person can get information related to the protein venom of the snake and its type in the absence of the medical expert at a critical condition, when the person is several miles away from the hospital, the protein analyses is made on the Snake venom related protein sequence which can also be used in pharmacology for preparing drugs.

Table 2 . Cluster details,features of venomous snakes

Cluster no	1	2	3	4	5	6	7	8
2	st	bh	br	pa	wk	el	ve	sn
3	st	bh	br	pa	wk	el	ve	sn

4	st	bh	br	pa	wk	el	ve	sn
6	st	bn	br	pa	wk	el	ve	sn
7	sl	bn	bl	np	wk	el	ve	sn
8	st	bh	gr	np	wk	el	ve	sn
9	sl	bn	mx	pa	wk	el	ve	sn
11	st	bh	br	pa	wk	el	ve	sn
12	st	vb	gr	np	wk	el	ve	sn
13	sl	lb	mx	pa	wk	el	ve	sn

Table 3.Cluster details, features of non-venomous snakes

Cluster No,	1	2	34	5	6	7	8	9
1	sl	sm	mx	pa	sm	rd	ve	db
5	sl	sm	mx	pa	sm	rd	ve	db
10	sl	sm	gr	np	sm	rd	ve	db
14	sl	sm	gw	np	sm	rd	ve	db
15	sl	sm	gr	np	sm	rd	ve	db

st-stout br-brown pa-pattern sn-single
sl-slender bl-black np-no pattern db-double
bh-broadhead gr-grey sm-smooth sm-small
vb-very broad gw-grey white el-elliptical mx-mixed
ve-vent

Table 4. Sample snake List

1) Borrelia burgdorferi haplotype	8)Naja Naja
2)Agkistrodon Acutus	9)black-banded coral snake
3)Bothrops Jararca	10)Drosophila melanogaster
4)Dobia russelli siamensis	11)Vipera Russelli Russelli
5)Green Mamba Snake Venom	12)King naja naja
6)Peruvian bothrops	13)Precursor
7)Dendroaspis Polylepis Polylepis	14)Bungarus candidus
	15)checked water snake

Table 5. Matching Protein Sequence

cluster id	Protein sequence
2	SLFELGKMIWQETGKNPVKNYGLYGCNCGV GGRGEPLDATDRCCFVHKCCYKKLTDCDSKK DRYSYKWKNKAIVCGKNQPCMQEMCECDK AFAICLRENLDTYNKSFRYHLKPSCKKTSEQC
3	QKYNPFRYVELFIVVDQGMVTKNNGDLDKIK ARMYELANIVNEILRYLYMHAALVGLEIWSN GDKITVKPDVDYTLNSFAEWRKTDLLTRKKH DNAQLLTAIDFNGPTIGYAYIGSMCHPKRSVA IVEDYSPINLVVAVIMAHEMGHNLGIHHDTDF CSCGDYPCIMGPTISNEPSKFFSNCSYIQCWDF IMKENPQCILNEPLGTDIVSPPVCGNELLEVGE ECDCGTPENCQNECCDAATCKLKSGSQCGHG DCCEQCKFSKSGTECRASMSECDPAEHCTGQ SSECPADVFHKNGQPCLDNYGYCYNGNCPIM YHQCYALFGADVYEAEDSCFKDNQKGNYYG YCRKENGKKIPCAPEDVKCGRLYCKDNSPGQ NNPCKMFYSNDDEHKGMVLPGTKCADGKVC SNGHCVDVATAY
4	MGRFISISFGLLVVFLSLSGTGAKQDCLSDWS FYEGYCYKVFNEKKTWEDAEKFCNEQVNGG YLVSFRSSEEMDFVIRMTFPIFRFDFFWIGLRD FWRDCYWRWSDGVNLDYKAWSREPNCFVS KTTDNQWLRWNCNDPRYFVCKSRVSC
6	TPEQQRYVDLFIVVDHGMFMKYNGNSDKIRR RIHQMVNIMKEAYSTMYIDILLTGVEIWSNKD LINVQPAAPQTLDSFGEWRXXXXXXKSHDNA QLLTDTDFDQVTINLAYTGSMCDLNKSTGVIQ DHSEQDLMVAITMAHELGHNLGISHDTGSCS CGGYSCIMSPVLSHEPSKYFSDCSYIQCWDFI MKENPQCILNKR
7	XAKYCKLPLRIGPCKRKIPSFYYKWKAKQCL PFDYSGCGGNANRFKTIEECRRTCVG
8	NLYQFKNMIKCTVPSRSWWDFADYGCYCGR GGSGTPVDDLDRCCQVHDNCYNEAEKISGC WPYFKTYSYECSQGTLTCKGDNNACAASVC DCDRLAAICFAGAPYNDNNYNIDLKARCQ
9	MKTLLLTLVVVTIVCLDFGYTIVCYKRHASD SQTTTCLSGICYKKITRGISRPEMGCGCPQSSR GVKVECCMRDKCNG
11	SLIQFETLIMKVAKKSGMFWYSNYGCYCGW GGQGRPQDATDRCCFVHDCCYGKVTGCDPK MDVYSFSEENGDIVCGGDDPCKKEICECDRA AAICFRDNLTLYNDKKYWAFGAKNCPQEESE PC
12	TKCYNHQSTTPETTEICPDSGYFCYKSSWIDG REGRIERGCTFTCPELTPNGKYVYCCRRDKCN Q
13	MKTLLLTLVVVTIVCLDLGYTRICFNHQSSQP QTTKTCSPGESSCYHKQWSDFRGTIIERGCGC PTVKPGIKLSCCESEVCNN

This paper proposes a novel Pri-Tri data mining tool that could useful in practical situations

The proteins sequences of snakes differ depend on the features and geographical location area in which inhabits. And its venom type which is generally classified in to four:

- Hemotoxic
- Neurotoxic.
- Cytotoxic
- Myotoxic

Accordingly the protein sequence can be obtained and similarity analysis can be done through pattern matching using various protein extracting tools like

- Swiss-prot
- Blast
- Fasta

This paper incorporates tool to parallel Check for pattern matching using Pri-tri algorithm to tailor to the needs of this problem. To implement the concept, this paper it uses java language with net bean IDE and multi threading concept can be achieved for parallelism .In Unix the demons can be applied for multi processing to achieve the same objective. After applying statistical tool which displays the proteins search time and score in Bar chart fig.(4,5)got from the data mining tools. Fasta is the fastest tool in extracting data, but contains less other informations, Blast tool speed much better than swiss_prot but lesser than Fasta.

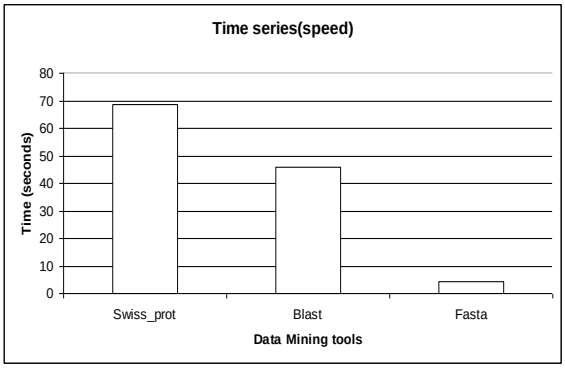


Fig 4: Bar Chart depicting the time

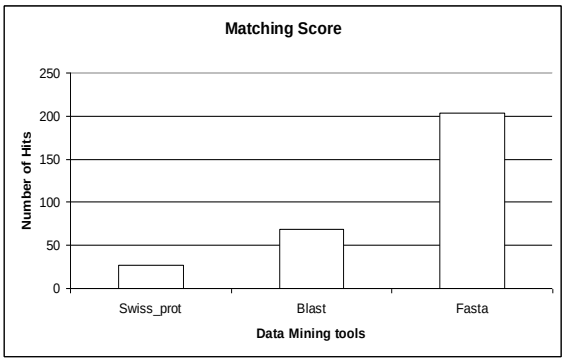


Fig 5: Bar Chart depicting the score

6. CONCLUSION

In some instances the data mart simply comprise with Relation OLAP, which is highly demoralized so the other option in to use data mining tool, so that the data mart can be incorporated with the data mining tools for best result[3]. This is more cost effective and time saving .In this paper deals with one such data mining tool. In many existing methods the data mart has to be constructed according to the customer needs with the help of the OLAP tool. Here this paper has to proposed a algorithm for building the data mart using a Pri-Tri algorithm for retrieving protein sequence from the PDB(Protein Data Bank). Similarly regarding to the clustering the number of cluster 'K' to construct is a input parameter the algorithm like ROCK user has to specify both K and threshold value but in the proposed algorithm the similarity measurement is used to cluster the categorical data and the objects are clustered without getting any input as 'K' or the threshold value. From the clusters, patterns can be extracted. These extracted patterns can be used to for the decision-rules. To be more effective workspace that can be partitioned, such that it contains only relevant attributes for the pattern needed.

7. REFERENCES

[1] Bambang Parmanto ,Mathew Scotch and Valerie Monaco, Usability Evaluation of the Spatial OLAP Visualization and Technique for Datasets using Analysis Tool (SOVAT), 2007, pp. 76- 95

[2] Bhaskar Sharma, Barkha Ratta, Gaurava Rai, Mayank Pokhariyal ,Meeta Saxena, K.P.Mishra ,Journal of Proteomics & Bioinformatics Predicting Secondary Structure of Oxidoreductase Protein Family Using Bayesian Regularization Feed-forward Backpropagation ANN Technique, 2010,pp179-182

[3] Dr.C.ChandraSekar,Hari Ganesh , A Parallel Computing Data Mining and Enhanced K-means Algorithm for Detecting Protein Sequence International Journal of Computing Technology and Information Security, 2011, Vol.1,No.1,pp.56-61.

[4] Edmonton, Alta., Canad, Popescu, C.A.; Wong, Y.S.; Grant MacEwan Coll., Knowledge and Data Engineering, IEEE Transactions , 2005, Volume: 17 Issue:12 , pp 1653 – 1663

[5] Feyyad, U.M.; Microsoft Res., Redmond, WA 2002 Data mining and knowledge discovery: making sense out of data ,IEEE XploreVolume: 11 Issue:5,pp 20 - 25 [6] Guha, S.; Rastogi, R.; Shim, K.; Stanford Univ., CA , ROCK: A RObust clustering algorithm for Categorical Attributes, 2002 pp512 – 521.

[7] Hua Yan, Lei Zhang and Yi Zhang, Clustering Categorical Data Using Coverage Density, Springer,2005,pp 10-15

[8] K.Rajendra Prasad et. al.A Survey On Clustering Efficient Graph Structures, Vol. 2 (7), 2010, pp2707-2714

[9] Margaret H.Dunham,Data Mining Introductory and advanced topics,Pearson education,2010, pp 50-52

[10] Popescu, C.A.; Wong, Y.S.;Grant MacEwan Coll., Edmonton, Alta., Canada ,Nested Monte Carlo EM algorithm for switching state-space models, IEEE Xplore,2005,pp 1653 - 1663

[11] Sushmita mitra,Tinku Acharya Data Mining, Multimedia, Soft computing and Bio informatics 2011,pp 39-40

[12] Yue Zhao; Yan-heng Liu; Xue-gang Yu; Hai-Yan Hu; FangMei; Coll. of Comput. Sci. & Technol., Jilin Univ., Changchun, A Method for Mobile Path Prediction Based on Data Mining, IEEE Xplore, 2007, Issue: 21-22,pp 691 – 695

[13] ZhexueHuang; Ng,M.K.; Manage. Inf. Principles Ltd., Melbourne, Vic ,2002 , A fuzzy k-modes algorithm for clustering categorical data, IEEE Xplore Volume: 7 Issue:4, pp446 - 452