DIMENSIONALITY REDUCTION BY MATRIX FACTORIZATION USING CONCEPT LATTICE IN DATA MINING

Bhavana Jamalpur[1], S.S.V.N Sarma[2]

[1]*Asst.Prof, S.R.E.C, Hasanparthy, Warangal,* ***bhavana_j@yahoo.com***
[2]*Dean, Dept. CSE.Vaagdevi, Bollikunta, Warangal-A.P (INDIA),* ***ssvn.sarma@gmail.com***

Abstract

Concept lattices is the important technique that has become a standard in data analytics and knowledge presentation in many fields such as statistics, artificial intelligence, pattern recognition ,machine learning ,information theory ,social networks, information retrieval system and software engineering. Formal concepts are adopted as the primitive notion. A concept is jointly defined as a pair consisting of the intension and the extension. FCA can handle with huge amount of data it generates concepts and rules and data visualization. Matrix factorization methods have recently received greater exposure, mainly as an unsupervised learning method for latent variable decomposition. In this paper a novel method is proposed to decompose such concepts by using Boolean Matrix Factorization for dimensionality reduction. This paper focuses on finding all the concepts and the object intersections.

Keywords: *Data mining, formal concepts, lattice, matrix factorization dimensionality reduction.*

----------***----------

1. INTRODUCTION

Rapid advances in data collection, data mining has evolved as active research area challenges and practical implementations associated with the problem of extracting interesting patterns from heterogeneous databases and real world dataset. Data mining can be simply defined as the extraction / mining of knowledge from large repositories data. Various algorithms and techniques like Classification, Clustering, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method are used for knowledge discovery from huge databases. Now-a-days, lattice theory combined with data mining under the framework of Formal Concept Analysis (FCA) has made mathematical computational view for knowledge visualization representation and knowledge discovery.

Formal Concept Analysis is an unsupervised learning technique for conceptual clustering. Here in this paper the notion of concept lattices show the use in Knowledge Discovery. For dimensionality reduction can be made in two different ways: by only keeping the most relevant variables from the original dataset , this technique is called feature selection or by exploiting the redundancy of the input data and by finding a smaller set of new variables, each being a combination of the input variables .

2. PROBLEM STATEMENT

FCA reflects quite well for exploring hierarchies for classifying, clustering , establishing relationships. But the problem is that, concepts contain redundant information that exists in intent and extent can attribute reduce labeling helps in decomposition of a context and find the relationships .

3. BACKGROUND

3.1.1 Formal Concept Analysis

R. Wille has introduced Formal Concept Analysis(FCA) . FCA can be applied in many fields like mathematical computer science, biological science, psychology, sociology, anthropology ,medicine . Concept Data Analysis is a method of data analysis across heterogeneous domains, the entire data is expressed and is described by the relationship between a pair of elements or transactions and for a particular set of data items. The inherent features of FCA is the combining the three basic elements of conceptual processing of data and knowledge, discovery and reasoning with concepts, data discovery and understanding with reasoning and dependencies among data sets and presentation of the data concepts visually. Combination and visualization of these components makes FCA a powerful technique in data engineering which can be applied to various real-world data.

A concept consists of two parts extension and intension. The extension contains all objects belonging to the particular concept and the intension contains all attributes required and used for all the objects . Objects and attributes establish relations and sub-relation among the concepts hierarchical manner "Sub-concept- Super-Concept" relation between concepts the implication between attributes and the incidence relation among the "an object has an attribute".

3.1.2 Partial Order and Lattices

Let $(P^1, \leq)$ be an ordered set and S is the subset of P^1. A partial order on P^1 is a binary relation $\leq$, such that for all { x, y, z} belongs to P^1 the relation is which satisfies the following properties:

i) Reflexive ii) Anti-Symmetric iii) Transitive
Let (P^1, ≤) be an ordered set and S be a subset of P^1 . An element ub Є P^1 is an upper bound/Supremum is called **Join** of S and is denoted by Vs , and the greatest lower bound/Infimum is called the **Meet** of S and is denoted by ^S . Supremum and Infimum is used to denote the join and meet.

For Join is denoted by :

Join = x ∨ y (x' join y') instead of Sup{x,y} or Vs (join of S) implies Sup s
For Meet is denoted by :

Meet= x ∧ y (x' meet y') instead of Inf{x,y} or ∧s (meet of S) implies Inf s
If S={x, y} X V Y for join and X^Y for meet.
An ordered set (L, ≤) is a lattice , if for any pair of elements x and y in L the join(x ∨ y)and meet (x ∧ y) always exist.

L is called the Complete Lattice if ∨s ,^s exist for all S⊆ L . In case of complete lattice the topmost element is called the unit and bottom element is called the Zero-element .A lattice is complete , L is called a join semi-lattice , if only Supremum(join) lattice exists . L is called meet semi-lattice if only a Infimum(meet) exists. Let P^1 denoted the Power set of S. The ordered set (P^1(s), ⊆)is set of all possible elements (P^1(T), ⊆) are the set of all possible elements present in both complete lattices.

3.1.3 Contexts and Concepts

Formal context in FCA is a triplet (O,A,R) where O is a set of objects ,A is a set of attributes and R is the incidence relation the binary relation R ⊆ O × A shows which objects posses which attributes . Predicate gRm denotes object g having attribute m. For subsets of objects and attributes A ⊆ O and B ⊆ A galois operators are defined as
A^{attrib}={m ∈ A | ∀ g ∈ A (gRm)}
$B^{objects}$={g ∈ O | ∀ m ∈ B (gRm)}
The set of attributes for an object can be represented by the binary vector denoted by 0/1 depending on the availability of the attribute for the particular object.

Table1- Objects and Attributes

	a	b	c	d	e	f	g	h	i
chicken	x	x	x	0	0	x	x	x	x
crow	x	x	x	0	x	x	x	x	x
dove	x	x	x	0	0	x	x	x	x
duck	x	x	x	x	0	x	x	x	x
hawk	x	x	x	0	x	x	x	x	x
kiwi	x	x	0	0	x	x	x	x	x
ostrich	x	x	0	0	x	x	x	x	x

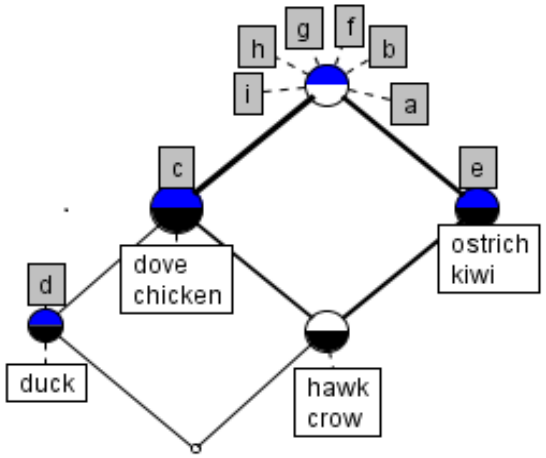


Fig1. Hasse Diagram

Fig.1 **Left** :Context K with O={chicken, crow, dove, duck ,hawk, kiwi ,ostrich } and A={a,b…,i}
A context is denoted as a table consisting of rows and columns with the objects corresponding to the rows of the table, the attributes corresponding to the columns of the table and a boolean value (in the example represented graphically as a check-mark) in cell (x, y) whenever object x has the attribute y.

3.2 Dimensionality Reduction

Dimensionality reduction is the process of reducing the number volume of data into small subset of the original data .The main objective of this method is to find a minimum set of attributes and the resulting can be categorized into feature selection and feature extraction. Attribute subset selection reduces the data size by removing or extracting irrelevant or redundant attributes. Data reduction play very important role in data mining .The main strategies for data reduction consists of data cube aggregation, attribute sub-set selection , numerosity reduction and data discretization.

3.3 Boolean Matrix Factorizations

To decompose the large data into smaller units of data Boolean Matrix factorizations is commonly used method used for knowledge extraction in data mining. When the input data is binary (0/1), replacing true or false values with the standard matrix multiplication with Boolean matrix multiplication produces more results. Finding a good Boolean decomposition is known to be computationally hard, with even many sub-problems being hard to approximate. Many real-world data sets are sparse and discreate , and it is often required that also the factor matrices are sparse. This requirement has motivated many new matrix decomposition methods and many modifications of the existing methods.

4. PROPOSED METHOD

Applying Boolean Matrix Factorization(BMF) based FCA method for Knowledge Discovery Process includes all the

preprocessing steps of Data Mining . Then apply FCA on the reduced context for a Object Intersection task.

5. COMPUTING THE INTERSECTIONS

A solution to the problem of finding all the concepts based on the fact that every concept extent is the intersection of attribute extents and every concept intent is the intersection of object intents then generates the possible set of concepts.

Algorithm for O-Intersect

Step-1:- Input the context (O,A,I).
Step-2:- Concepts(C) are drawn for each X and Y
Step-3: -Loop

If for a given set of Attributes (Y) for each Object(X)
I= {Y } ∩ *X *X={attributes of the object}

If I' is disjoint from any concept intent in Concept(C) then
C= C U {(I',I)}
Next Object(X)

6. OUTLINE OF OUR APPROACH USING MATRIX FACTORIZATION

A formal context may be depicted as |O| X |A| binary matrix ,where the objects of O form a row and the attributes A form a column which are represented by 1/0 whether the attribute is there or not.Dimensionality reduction methods are another mathematical models used.

For example consider a formal context in which,

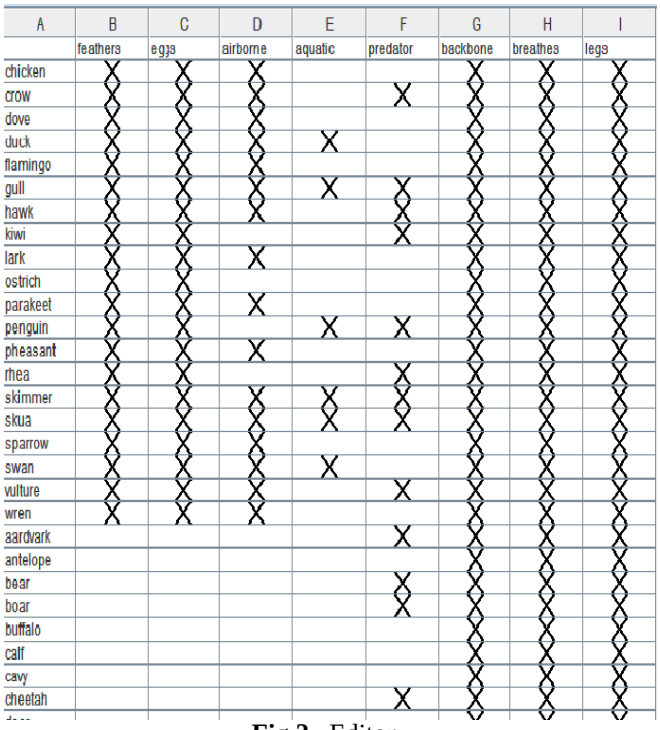
Objects (X)={ $x_1,x_2 \ldots x_n$} ∈ Animals and Birds

Attributes(Y)={ feathers, eggs, airborne, aquatic, predator, backbone, . breathes, legs}

By Boolean Matrix Factorization formal concepts R are obtained by a product of two or more matrices through matrix factorizing R means mapping the objects and attributes to a common latent factor.

For a real-time zoo dataset containing objects and attributes have been classified into different categories considering the animal objects(42) and birds objects(21) and 8 attributes the number of concepts that are generated are 17 as shown in Fig.2 and Fig.3

7. EXPERIMENTAL STUDY



	feathers	eggs	airborne	aquatic	predator	backbone	breathes	legs
chicken	X	X	X			X	X	X
crow	X	X	X		X	X	X	X
dove	X	X	X			X	X	X
duck	X	X	X	X		X	X	X
flamingo	X	X	X			X	X	X
gull	X	X	X	X	X	X	X	X
hawk	X	X	X		X	X	X	X
kiwi	X	X			X	X	X	X
lark	X	X	X			X	X	X
ostrich	X	X				X	X	X
parakeet	X	X	X			X	X	X
penguin	X	X		X	X	X	X	X
pheasant	X	X	X			X	X	X
rhea	X	X			X	X	X	X
skimmer	X	X	X	X	X	X	X	X
skua	X	X	X	X	X	X	X	X
sparrow	X	X	X			X	X	X
swan	X	X	X	X		X	X	X
vulture	X	X	X		X	X	X	X
wren	X	X	X			X	X	X
aardvark					X	X	X	X
antelope						X	X	X
bear					X	X	X	X
boar					X	X	X	X
buffalo						X	X	X
calf						X	X	X
cavy						X	X	X
cheetah					X	X	X	X

Fig.2 Editor

In the above figure shows the context of Birds and Animals where objects (Birds and Animals) and the attributes are the properties describing them . The Object Intersection in the ij position in the context editor indicates the object I is described by the attribute j. For instance , for object Chicken in ith row contains attributes such as feathers ,eggs, airborne ,backbone, breathes, legs etc. If the object has the attribute then it is denoted by 'X' and does not exhibit the property by 'Null' space.

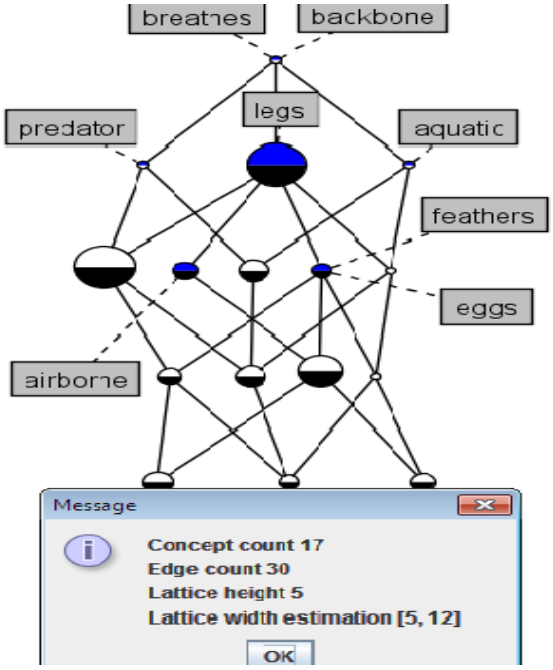


Fig.3 Lattice obtained from the above context

The fig.3 consists of 17 concepts, Edge count as 30 and the height of the lattice is 5 containing 8 attributes 61 objects.

	F1	airborne	predator	F2
chicken	X	X		X
crow	X	X	X	X
dove	X	X		X
duck	X	X		X
flamingo	X	X		X
gull	X	X	X	X
hawk	X	X	X	X
kiwi	X		X	X
lark	X	X		X
ostrich	X			X
parakeet	X	X		X
penguin	X		X	X
pheasant	X	X		X
rhea	X		X	X
skimmer	X	X	X	X
skua	X	X	X	X
sparrow	X	X		X
swan	X	X		X
vulture	X	X	X	X
wren	X	X		X
aardvark			X	X
antelope				X
bear			X	X
boar			X	X
buffalo				X
calf				X
cavy				X
cheetah			X	X
deer				X

Fig.4 Decomposed context by reduced labeling

Attributes are factored as factor1(F1) and factor2(F2) for the dimension reduction Where F1={feathers, eggs, aquatic} and F2={backbone, breathes, legs}

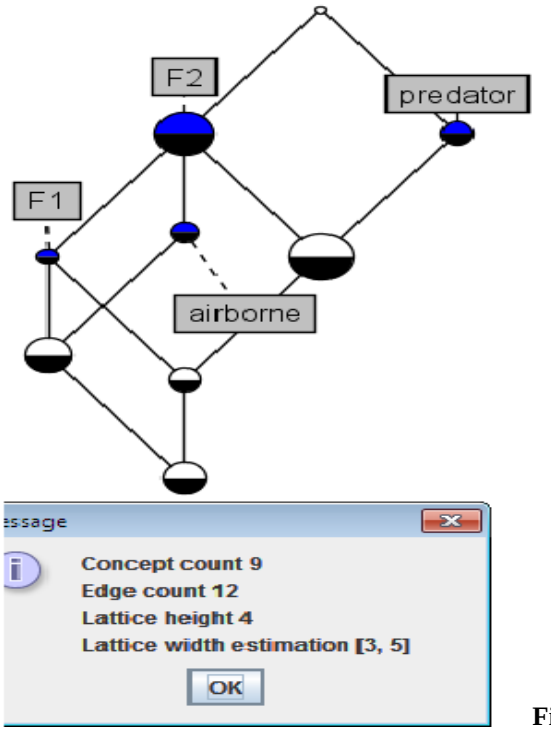


Fig.5 Lattice after factorization

Using the Factors F1,F2 which in turn consists of attributes generated a new lattice with number of concepts 6

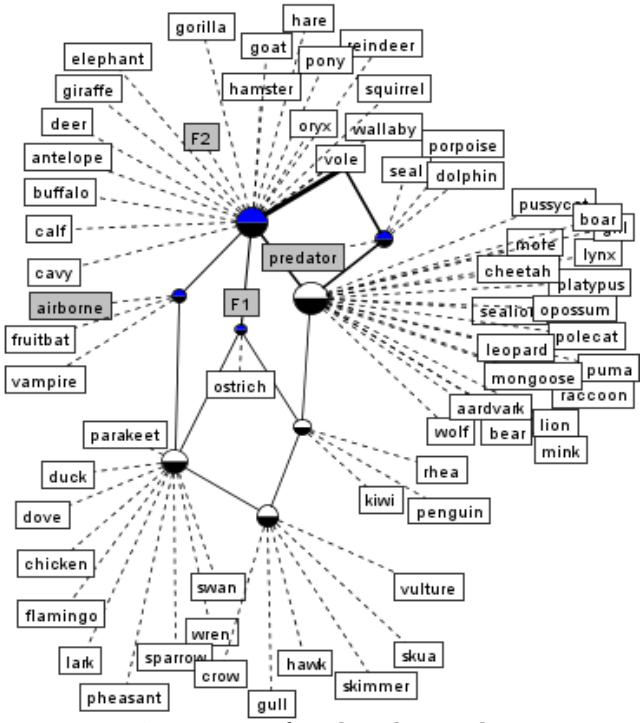


Fig.5 Lattice of Birds and Animals.

Above lattice obtained after the decomposition of the data by reduced labeling using the matrix factorization. The attributes of relevant are taken in consideration while leaving the irrelevant weak attributes. The attributes are factored by placing them into F1 and F2 factors in such a way that a new lattice is obtained contains all the information as the previous or prior lattice without any loss of information .Some of the attributes that common in both the objects share the attributes are strongly connected with each other .

The Object Intersection of these objects from the above lattice depicts that Ostrich lies in the group of Birds and Animals .As in the previous fig.2 lattice we obtained 17 contexts by BMF but after Factorization a new lattice is obtained containing 6 contexts by reduced labeling where knowledge is derived from the dataset

8. EXPERIMENTAL RESULTS

By plotting the Attributes verses Concepts after the reduced labeling it is observed that the when attributes are factored the number of concepts are also reduced which indicates symmetry in the data.

Table-2 Attributes and Concepts

Techniques	Attributes	Concepts
BMF	17	301
Reduced Labeling	8	17
Factorization	4	9

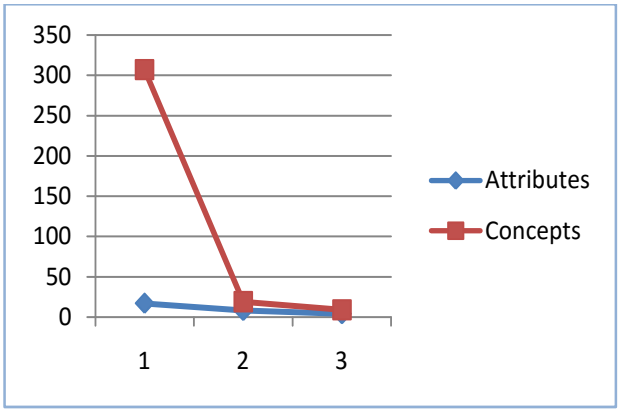


Fig.6 Attributes Vs Concepts

FUTURE WORK

- Cluster similarity and Dissimilarity
- Cluster Validity

CONCLUSION:

BMF is a strong data mining technique, when data is binary, consider BMF(Boolean matrix factorization) helps to decompose the data by reduced dimensions where factors tell something about the data and their association among them and knowledge derived from the lattice.
A summary of the findings is as follows:

- BMF based FCA is computationally easy by reduced dimensions
- Knowledge derived from the BMF-FCA based reduced context is able to extract all the object intersections

REFERENCES

[1]. David M Blei and J Lafferty. Topic models. Text mining: classification, clustering,and applications,2009
[2]. Ganter, B. and Wille, R. (1999).Formal Concept Analysis: Mathematical Foundations, Springer, Berlin http://www.math..tu-dresden.de/~ganter/fba.html
[3]. Ruslan Salakhutdinov and Andriy Mnih. Probabilistic matrix factorization. In NIPS, 2007.
[4]. Carpineto, C. and Romano, G. (2004).Concept Data Analysis:Theory and Applications, John Wiley, Chichester.
[5]. Poelmans, J., Elzinga, P., Viaene, S., Dedene., G. (2010). Formal concept analysis in knowledge discovery: A Surveyin R. Goebel (Ed.) Proceedings of the 18th International Conference on Conceptual Structures, Springer-Verlag, Berlin,pp. 139–153.
[6]. Priss, U. (2006). Formal concept analysis in information science,Annual Review of Information Science and Technology
[7]. Stumme, G. (2009). Formal concept analysis,inS. Staab and R.Studer (Eds.), Handbook on Ontologies,SpringerVerlag,Berlin, pp. 177–199.
[8]. Pattison, P.E. and Breiger, R.L. (2002). Lattices and dimensional representations: Matrix decompositions and order-ing structures,Social Networks 24(4): 423–444.
[9]. Snasel, V., Polovincak, M., Dahwa, H.M. and Horak, Z. (2008).On concept lattices and implication bases from reduced contexts, Proceedings of the ICCS Supplement, Toulouse,France.