Research Journal of Computer and Information Technology Sciences _____ISSN 2320 – 6527
Vol. **3(1),** 1-4, March **(2015)** *Res. J. Computer and IT Sci.*

Self Organized Map Network for Classification of Multilevel Data

Varsha Namdeo, R.S. Thakur and G.S. Thakur
Department of Computer Applications, Maulana Azad National Institute of Technology, Bhopal, 462-052, INDIA

Available online at: www.isca.in
Received 19rd October 2014, revised 9th February 2015, accepted 3rd March 2015

Abstract

Data Mining is extraction of hidden knowledge from large set of data. Classification is a very important technique of data mining. Associative classification is a form of classification in which classification and association techniques are merged. From associative classification we can find more important rules and more accurate classification results. This paper presents an approach for classification of multilevel data using Self Organized Map Network. It is a kind of Artificial Neural Network. This approach gives a new kind of classification. It's importance can be felt in such applications in which data is sparsed and stored in different abstraction levels.

Keywords: Multilevel abstraction, Multilevel Data, Associative Classification, Self Organized Map Network, Multilevel Classification.

Introduction

Data Mining[1-3], is the extraction of hidden predictive information from large data bases. Classification[2] includes partition of the data into disjoint groups by finding rules. Labels of pre classified patterns are identified by this supervised learning. And the pre classified pattern is called training data set. A class attribute from the training data set will be used for labeling a newly encountered unlabelled pattern[1]. It involves the process as shown in the following figure 1. There are two very important data mining techniques Association Rule Mining and Classification. When both techniques are merged, it gives more accurate results. And it is known as Associative Classification technique. These techniques can be applied on multilevel data.

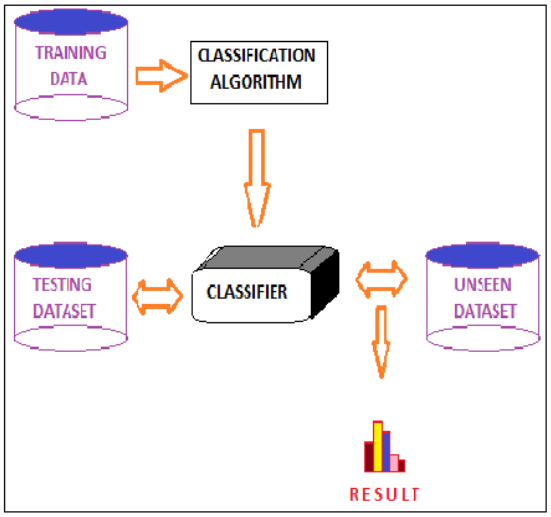

Figure-1
Classification Process

Multilevel data: Multilevel data[1] is clustered, nested and hierarchically structured. Data taken from organized units at higher level (from which data are also obtained). In following figure 2 example of multilevel data is shown.

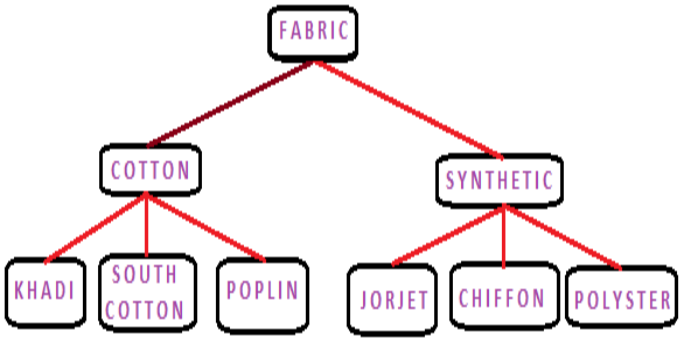

Figure-2
Example of Multilevel Data

Self Organized Map network (SOM): The learning is dependent on the input data; For training a network no error is calculated. Here Net can react to various output categories on training and only one from the various neurons has to react[4]. Hence extra structure can be built-in the network so that the net is forced to make a decision as to which one unit will respond. The method by which only one unit is chosen to react is called competition. The frequently used competition among cluster of neurons is Winner-Takes-all. Here only one neuron in the competing group will have a non-zero output signal when the competition is completed. This form of learning depends on the purpose for which the net is being trained. Kohonen worked in the development of the theory of competition as a result the competitive processing elements are referred to as Kohonen unit. This SOM[5] can be termed as topology-preserving maps. SOM by Kohonen[6], makes the clusters of the input data that are normally used for unsupervised training. SOM is the only artificial neural network which is having topological structure

property like human brain, which has 'm' cluster units, in one or two dimensional array and the signals are n-tuples.

Previous Work: In their paper R. P. Gangwar[7] present the approach of classification which is based on MrCAR (Multi-relational Classification Algorithm) and Kohonen's Self-Organizing Maps (SOM) approach. They claim that they optimize the classification rate of Mr CAR with SOM network and also improve the efficiency of classification.

In their paper Del-Hoyo[8] proposed an extension for the Self Organizing Map model which is called Associative SOM. Incomplete data patterns and incomplete class labels can be easily dealt by ASOM model. They also compared ASOM with Multilayer Perceptrons in the incremental classification.

These two approaches are used for classification of data at only single level. And we are proposing classification approach for multilevel data.

Methodology

Proposed Approach for Classification of Multilevel Data: The proposed approach contains following steps: Step-1: Taxonomy can be represented by a sequence of numbers and the symbol ''*''. Step-2: Take a variable h for storing level number of abstraction. Here h may take values from 1 to 3 (as we considered up to 3-levels of hierarchies). Firstly initialize h=1. Step-3: Put the min_support and min_confidence for each level using uniform minimum support scheme across levels. Step-4: Generate Class Association Rule using CBA_RG Algorithm [9]. Step-5: Make classifier model from dataset using SOM network classifier algorithm. Step-6: Repeat the above process from step-2 to step-6 for next levels (i.e. h= 2 and h= 3).

Working of Proposed Approach: Proposed algorithm is explained with example. In this approach we have used transactional database which is shown in Table-1, having 6 attributes C_id, *I1*, *I2*,*I3*, *I4* and one class label attribute C and it having 10 records. Table-2, shows the codes for each item name with their description.

Table-1
Database: D1

C_id	I1	I2	I3	I4	C
1	AAA	BAB	CBB	DAA	X
2	AAA	BBB	CBA	DBA	Y
3	AAB	BAA	CAB	DCB	Y
4	AAA	BBA	CCA	DAA	X
5	ABB	BAB	CAB	DBB	Y
6	ABA	BAA	CCA	DAA	X
7	ABB	BAB	CAB	DBB	Y
8	AAA	BAB	CBB	DAA	X
9	ABC	BDB	CBA	DBA	X
10	ACA	BAC	CBB	DAA	X

Table-2
Codes of item name

Code	Description
A**	work class
B**	education
C**	age
D**	credit rating
AA*	work class self employed
AB*	work class govt. employee
AC*	work class private
BA*	education PG
BB*	education graduate
BC*	education school
BD*	education doctorate
CA*	age Young adult
CB*	age Mid life adult
CC*	age Mature adult
DA*	credit rating Excellent
DB*	credit rating Fair
DC*	credit rating Poor
AAA	work class self employed single handed
AAB	work class self employed partnership
ABA	work class govt. employee class 1
ABB	work class govt. employee class 2
ABC	work class govt. employee class 3
ACA	work class private executive
BAA	education PG medical
BAB	education PG engg.
BAC	education PG other
BBA	education graduate technical
BBB	education graduate nontechnical
BCB	education school HSC
BDA	education doctorate research
BDB	education doctorate professional
CAB	age Young adult 25yrs-30yrs
CBA	age Mid life adult 30yrs-40yrs
CBB	age Mid life adult 40yrs-50yrs
CCA	age Mature adult 50yrs-58yrs
CCB	age Mature adult 58yrs-65yrs
DAA	credit rating Excellent High
DBA	credit rating Fair High
DBB	credit rating Fair Average
DCB	credit rating Poor Average

For explanation purpose proposed classification approach uses transactional database D1. The dataset D1 have items with only three levels. So, the proposed approach is applying on three levels. Table 3 represents Database D1 at abstraction level 1 of Hierarchy, where A** shows the generalization of work class group at level 1, similarly other encoding shows generalization of other groups.

Table-4 represents database D1 at abstraction level 2 of Hierarchy, where AA* shows self employed work class

generalization which resides at level 2 of hierarchy, similarly other encoding shows generalization of other groups at level 2. Table-5 shows the database D1 at level 3 which has most detailed information.

Table-3
Database at level 1

C_id	I1	I2	I3	I4	C
1	A**	B**	C**	D**	X
2	A**	B**	C**	D**	Y
3	A**	B**	C**	D**	Y
4	A**	B**	C**	D**	X
5	A**	B**	C**	D**	Y
6	A**	B**	C**	D**	X
7	A**	B**	C**	D**	Y
8	A**	B**	C**	D**	X
9	A**	B**	C**	D**	X
10	A**	B**	C**	D**	X

Table-4
Database at level 2

C_id	I1	I2	I3	I4	C
1	AA*	BA*	CB*	DA*	X
2	AA*	BB*	CB*	DB*	Y
3	AA*	BA*	CA*	DC*	Y
4	AA*	BB*	CC*	DA*	X
5	AB*	BA*	CA*	DB*	Y
6	AB*	BA*	CC*	DA*	X
7	AB*	BA*	CA*	DB*	Y
8	AA*	BA*	CB*	DA*	X
9	AB*	BD*	CB*	DB*	X
10	AC*	BA*	CB*	DA*	X

Table-5
Database at level 3

C_id	I1	I2	I3	I4	C
1	AAA	BAB	CBB	DAA	X
2	AAA	BBB	CBA	DBA	Y
3	AAB	BAA	CAB	DCB	Y
4	AAA	BBA	CCA	DAA	X
5	ABB	BAB	CAB	DBB	Y
6	ABA	BAA	CCA	DAA	X
7	ABB	BAB	CAB	DBB	Y
8	AAA	BAB	CBB	DAA	X
9	ABC	BDB	CBA	DBA	X
10	ACA	BAC	CBB	DAA	X

Ability of a classifier to correctly assign the class is called its accuracy. The result of proposed approach is given below at each level for database D1. On Min. Support= .4, Min. conf=.6 and iteration=50 we found the results as follows-

Table-4
The results of proposed approach

At Level-1	At Level-2	At Level-3
Accuracy =86.980000	Accuracy =88.450000	Accuracy =90.120000
Precision =85.960000	Precision =88.900000	Precision =92.240000
Recall =86.540000	Recall =90.950000	Recall =91.120000
F-Measure =88.700000	F-Measure =91.640000	F-Measure =94.980000

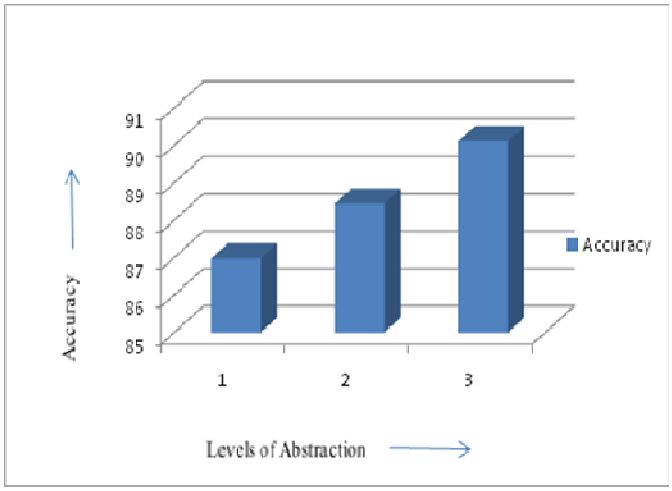

Figure-3
Accuracy of proposed approach at different Abstraction levels on example database

Results and Discussion

This section presents experiments made to show the performance of the proposed approach. They were implemented in Matlab 7.8 on core 2 duo processor personal computer.

We have tested our proposed algorithm on the dataset from real world. The dataset are taken from a Super Market of Bhopal, India. It consists of daily transaction records. A transaction record contains the item name that customer purchased at the Super Market each time. During the preprocessing phase we consider only readymade garments section related transaction records. The numbers of levels are set as 3.Optimum number of levels depends on the nature of the data and our requirement of abstraction level upto which the data is considered for classification.

There are 58 items (for example man T-shirt medium size) at primitive level i.e. level 3 which is generalized into 14 items according to their brands, uses and sizes (because same item might be sold in multiple forms, brands, and sizes.) i.e. level 2. The highest level (i.e. level 1) consists of 4 items or categories i.e. Men, Women, Kids, Brand new apparel.

The dataset consists of 1500 transactions in a week. From which we use 70% data as training set and 30% data for testing. Experiments performed to find accuracy of classification at different abstraction level on different minimum support value.

The result of proposed approach is given below at each level on dataset. On Min. Support= .4, Min. conf=.6 and iteration=50 we found the classification accuracy as follows-

At Level1, Accuracy =88, at Level2, Accuracy =91 and at Level3 Accuracy =92

As level increases (from level 1 to level 3), the accuracy of proposed approach is also increases because of level 1 contains generalized information about item sold but level 2 and level 3 contains more detailed information about item sold.

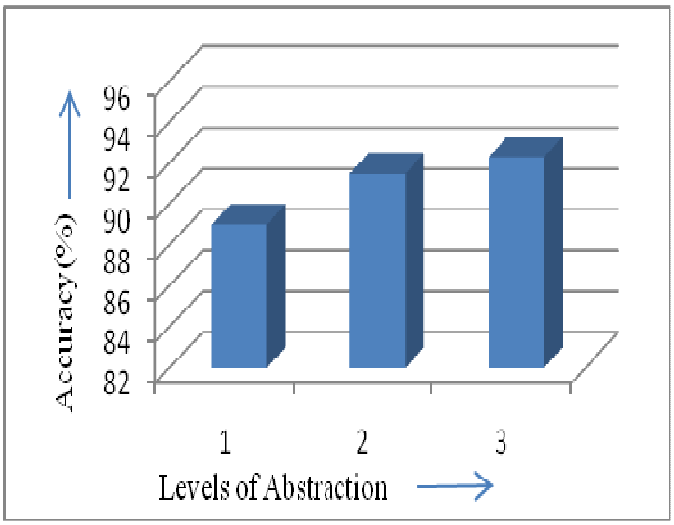

Figure-4
Accuracy of proposed approach at different Abstraction levels on real dataset

Conclusion

Associative classification method can be applied to multilevel data for classification. We proposed an Associative classification method in this paper by which data can be classified at multiple level of abstraction in a concept hierarchy. The proposed algorithm follows the six steps for all level of abstraction and builds a classifier based on the Self Organized Map Network Classifier. The result shows that at lower level of abstraction there is more accuracy in classification results because lower level contains more detailed information. Transactional data and relational data both can be handled by it. It gives a new way of classification.

References

1. Han J. and Kamber M., Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, **(2006)**

2. Pujari A.K., Data Mining Techniques, Universities Press (India) Private Limited, **(2001)**

3. Stonebraker M., Agrawal R., Dayal U., Neuhold E.J. and Reuter A., DBMS Research at a Crossroads: The Vienna Update, Proc. 19th very large data bases conf., 688-692 **(1993)**

4. Kothari A., Keskar A., Chalasani R., Srinath S., Rough Neuron Based Neural Classifier, Emerging Trends in Engineering and Technology, 2008, ICETET '08, 624-628, 16-18 July **(2008)**

5. Mishra M. and Behera H.S., Kohonen Self Organizing Map with Modified K-means clustering For High Dimensional Data Set, *International Journal of Applied Information Systems (IJAIS),* Foundation of Computer Science FCS, New York, USA, **2(3),** May, **(2012)**

6. Paigwar S. and Shukla S., Neural Network Based Offline Signature Recognition and Verification System, *Research Journal of Engineering Sciences,* **2(2),** 11-15, February, **(2013)**

7. Gangwar R.P., Agrawal J. and Sharma V., Auto-Label Threshold Generation for Multiple Relational Classifications based on SOM Network, IJCA, **40(7), (2012)**

8. Del-Hoyo R., Buldain D. and Marco A., Supervised Classification with Associative SOM, Computational Methods in Neural Modeling, *Lecture Notes in Computer Science,* **2686,** 334-341, **(2003)**

9. Liu B., Hsu W. and Ma Y., Integrating classification and association rule mining, In Knowledge Discovery and Data Mining, 80–86, **(1998)**