The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, Special Issue:
The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE'13],
San Francisco State University, CA, U.S.A., March 2013
Doi: 10.7321/jscse.v3.n3.34 e-ISSN: 2251-7545

EFFICIENT KNOWLEDGE DISCOVERY FOR AN INTELLIGENT DATAMINER

Dr Rajalakshmi Selvaraj
Department of Computing
Botho College
Gaborone, Botswana
E-mail: rajalakshmi.selvaraj@bothocollege.ac.bw

Dr Venu Madhav Kuthadi
Department of AIS
University of Johannesburg
Johannesburg, South Africa
E-mail:vkuthadi@uj.ac.za

Abstract—***In each and every field, data is the most significant property, which is been concealed and exposed in an open environment. Processing of non-trivial and removal of data is done by a novel approach, which is been actionable and have implicit information retrieval from a huge amount of data .The proposed approach is used for an effective investigation on data which is to find out the most relevant patterns and information, which are not unpredictable by the users. It also computes the patterns and relationships in an infrequent data and distributes the outcome which can be assessed by a human analyst or to make use by an automatic decision support system. Also it has multiple stages that are available for processing the data. The proposed approach is a very simple process that involves in post processing, and which is used in getting and extracting mined result, also the approach uses a pre-processing step for data processing and finally selecting a suitable data mining algorithm. The proposed system is operated by selecting a suitable Data mining algorithm for the user's needs .The proposed system first performs the pre-processing of the data that converts the data into most appropriate form , then selection of the algorithm will take place. Finally the pre-processed mined data is post-processed and acquired frequent pattern. The main aim of this paper is to deploy some data mining algorithm and to develop patterns. This general structure is used for any type of data set and it is mostly used in creating intelligence for a huge quantity of data, which needs in processing of knowledgeable data. The performance evolution is shown and which clearly explains about the proposed system.***

Keywords- Data mining, A-priori algorithm, Pincer search algorithm, Database, Artificial Intelligence, Pattern Recognition, Neural Network

I. INTRODUCTION

In data mining technique, the information's are collected from large database which are been concealed for information extraction. This technique is more powerful technology with huge ability to assist the association and to concentrate on the majority significant data in their information warehouse. Data mining tool calculates the potential behaviors and development by permitting the organizations and business to create more positive information-driven decisions. Demonstrative tools usual of decision support system give the automatic prospective analysis, which is accessed by data mining, shift away from the analysis of earlier proceedings.

Data mining tools are able to reply business questions, which conditionally are huge amount of time consuming to decide. Applications like sales, marketing, and promotions use the data mining. The middle idea of data mining is knowledge discovery process. Knowledge finding from data is the effect an investigation process linking the application of different algorithm procedures for manipulating models, manipulating the data, constructing the models from data. Multiple algorithmic components are involved in the Knowledge Discovery process. The components are related in nontrivial ways. Data mining supports three main technologies, which are now adequately mature, they are mentioned below

- Data mining algorithm
- Enormous Data collection
- Authoritative Multiprocessor System.

It is nothing but which is the Computer assisted process of data analysis. The data analysis can be executed using either the bottom-up approach or top-down data mining analysis rare information in an effort to find out hidden groups and trends, whereas the main objective of top-down data mining is to analysis a particular hypothesis. By using a selection of technologies, data mining will be performed effectively, which includes multi-dimensional analysis tools, including intelligent agents, powerful database queries. In this paper, the proposed tool is designed to find out best frequent pattern algorithm to eliminate frequent data item sets. The tool is designed with assist of back propagation in neural network method.

II. RELATED WORK

A number of research investigators have been creates the more mechanisms for retrieve the data from the database in the data mining area. In William F. Punch [1],

The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, Special Issue:
The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE'13],
San Francisco State University, CA, U.S.A., March 2013
Doi: 10.7321/jscse.v3.n3.34 e-ISSN: 2251-7545

make use of Genetic Algorithm for classifying the raw information through weighting the extracted features vectors in the web contents. This method is optimizing the raw data classification. However, when increase the feature of web contents the selection process is does not work very well. In Christian Borgelt [2], using FP-Growth Algorithm, frequent itemset is eliminated in the transactions. In this method, it is possible to delete the most valuable data in the transaction. Because, the user does not specify frequently occurred data items in the transactions. In [3], a designed system is combinations of TDB and RSTDB approaches, which are used to increase the efficiency of the overall frequent pattern. Cache Coherence Technique like as Prefix tree for store up the TDB, which increase the effectiveness of frequent pattern and RSTDB with Heuristic purpose utilize both downward and upward conclusion properties. In this approach, does not provide efficient closure properties to select candidate item sets from the frequent data items. In [5], the frequent mining process is based on time sensitive mining patterns. In this approach frequent pattern historical information is maintained, these historical information will be effectively used for Time sensitive Queries. When historical information increased, it is a possible to decrease the performance of the approach.

III. FREQUENT PATTREN

Frequent Pattern is nothing but which is the number of data items sets are frequently occurred in the database. In Database, frequent pattern mining stores the numbers of same data, which make more difficult incase of web oriented system. Many of the association spend the plenty of saving to reduce the frequent data's in the database. Different methods are used to reduce the frequent itemset in the database. In order to avoid the frequent itemsets, some methods are discussed as shown in below.

A. Apriori algorithm

Apriori algorithm make use of Level-wise search, where N-itemsets (An itemset which consist of N items of N-itemset) are use to discover (N+1)-itemsets, from the transactional database have to mine the frequent itemset for Boolean Associated Policy. Initially, the frequent itemset is established. This itemset is indicated as M1. M1 is mostly needed to discover M2. The 2-itemsets, which is mostly required to discover M3, and so on, until the final frequent N-itemset found. Based on association rules for knowledge mining Apriori is an efficient algorithm. In order to makes the document categorization, Apriori is most convenient. The Apriori algorithm mostly useful to a transactional database of market baskets. With the particular terms occur in text document and the transactional database is in set of documents. Constantly, with the actual terminology let us indicate terms as items and basket of terms as an itemset.Consider Li is an item L={L1,L2……Lm} is an itemset and L is database of documents. Frequency of an itemset should exceed this threshold value. It is also selected as frequent itemset. The main transaction maintain up in our proposed system, which communicate to the frequency of an itemset in the database. Data mining technique is used in digital library to characterize the individual topics. Frequent itemset is the iterative searching process. At the start, frequent 1-itemsets are established, which are used to produce frequent 2-itemset etc. Let assume that TDS distinct important algorithm we have in the database and produce candidates of frequent itemset-1. In Document frequency database, frequent itemset candidates are contained in the application directly. As a result, frequent itemset-1 is computed. In the consequent steps, from 1-itemsets have to compute 2-itemset. Regarding Apriori property frequent itemsets search terminated to create a subsequent itemset and frequent n-itemsets. While using this type of method, make use of technique, which is similar to transaction method. In our proposed consideration, a document which is not contains the N-itemset, which will be left out. In view of the fact that it couldn't contain any of (N+1)-itemsets. Consider Gk indicates set of candidates of N-itemsets and Hk-1 as a set of frequent (N-1)-itemsets.

Algorithm: Apriori algorithm
Input: a user-defined minimum support and a database
Output: all frequent itemsets

1. $I0 := \emptyset$; $N := 1$;
2. $G1 := \{\{l\} \mid l \in L\}$
3. response: $= \emptyset$
4. Whereas GN
_$= \emptyset$
5. read database and count supports for *GN*
6. *MN:* = {frequent itemsets in GN}
7. *GN*+1 := *Apriori-gen*(*MN*)
8. $N := N + 1$
9. response := response $\cup$ M*N*
10. return response

Figure 1: Apriori Algorithm

Apriori Property

Apriori property reduce the searching space in order to avoid discovering each of IN needs one complete of the examined database.

When the itemset do not support the minimum threshold value, min-sup is the L is not frequent, which is P (L) <min-sup.

When the itemset A is added to the itemset L, resulting itemset will not occur frequently than L. hence, IA is not frequent that is P (LA) <min-sup.

The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, Special Issue:
The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE'13],
San Francisco State University, CA, U.S.A., March 2013
Doi: 10.7321/jscse.v3.n3.34 e-ISSN: 2251-7545

B. pincer search algorithm

For Association Rule Mining, two approaches such as bottom up approach and top-down approach is used by Pincer Search Algorithm. The Pincer Search Algorithm is minor alternation of the Apriori Algorithm by R.Aggarwal & Srikant In this algorithm, main explore direction is bottom up like as Apriori Algorithm, but it except from Apriori concept such that it conducts concurrently a limited top down search. It maintains the one more data structure called as MFCDS (Maximum Frequent Candidate Data Set). It means the data sets including whole maximal frequent item sets, which consequently indicates immediately whole frequent itemsets. This algorithm focuses in handling with huge frequent itemsets of big length.

Theory:

Let L= (l_1, l_2,… l_m)

Here (l_1, l_2,… l_m) is nothing but m different items.

Transaction:

Here transaction T is described because several subsets of item in set of different items L.

Database:

A set of transactions is called a database, which is denoted as D

Itemset: Itemset is a set of items

Maximum Frequent Candidate Set (MFCS):
It is candidate set, which does not have any infrequent itemset .MFCS consist of itemset, which is the grouping of whole subsets of constituent consisting of all frequent itemset. i.e. MFCS is a minimum cardinality set fulfilling the following situation:

$$\text{INFREQUENT} \subseteq \{ 2^y \mid Y \in \text{MFCS}\}$$
$$\text{FREQUENT} \subseteq \{ 2^y \mid Y \in \text{MFCS}\}$$

Where INFREQUENT and FREQUENT indicate correspondingly for whole frequent and infrequent items (categorized as such as faraway).

Pincer-Search Methods
Combination of top down and bottom up approach is called Pincer Search Algorithm:
Top-Down
In top down approach, first generating the parent sets after that only generating subsets.
.***Bottom-up***
In Bottom up approach, first generates subsets and then go on to producing parent-set, candidate sets utilizes their frequent subsets.

The Pincer Search algorithm uses two particular properties. They are mentioned below,

Downward Closure Property
If an itemset is regular, then a whole its must be regular.

Upward Closure Property
If an itemset is irregular, whole its supersets must be irregular.

C. FP-Tree Algorithm

The association rule learning of data mining is a well examined and popular method for finding exciting relations among the variables in huge data storage. Piatetsky-Shapiro explains analyzing and presenting strong rules find in databases proposing various measures of interestingness. In FP-Tree algorithm, association rules are introduced, in order to discover the similarity among the products in a wide level business data, which are evidence by (Point-of-scale) POS in supermarket. Based on the concept of strong rules, supermarkets are running. For an example the sales data of a supermarket would specify with the purpose of customers buy potatoes and onions together, and also customers purchase a beef. These information can be proposed as basis information of decisions concerns marketing actions such as (e.g.) product assignment or promotional price .The market basket study association rules are used nowadays in lots of purpose, which used in many application area like as intrusion discovery, bioinformatics and Web usage mining are mentioned in the above example.

IV. PROPOSED SYSTEM

BACK PROPAGATION ALGORITHM

Construction

In this research, artificial intelligence of neural network method is used. From the neural network, back propagation mechanism is used to take the decision. Mainly in data mining domain problem modeling of ANN training method is most worked in the Back propagation algorithm. The difficult nonlinear mapping network model can understand by the multilayer feed backward neural network.

This part explains the construction of the neural network model, which is generally needed in the back propagation algorithm with the help of multilayer feed backward network.

The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, Special Issue:
The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE'13],
San Francisco State University, CA, U.S.A., March 2013
Doi: 10.7321/jscse.v3.n3.34 e-ISSN: 2251-7545

Neuron Model

As shown in below figure, *R*-inputs with an elementary neuron. Every input value is weight with the suitable *w* value. For the transfer function *f* input is given as bias and sum of the weighted inputs. To generate neuron output is able to make use of any differentiable transfer function f.

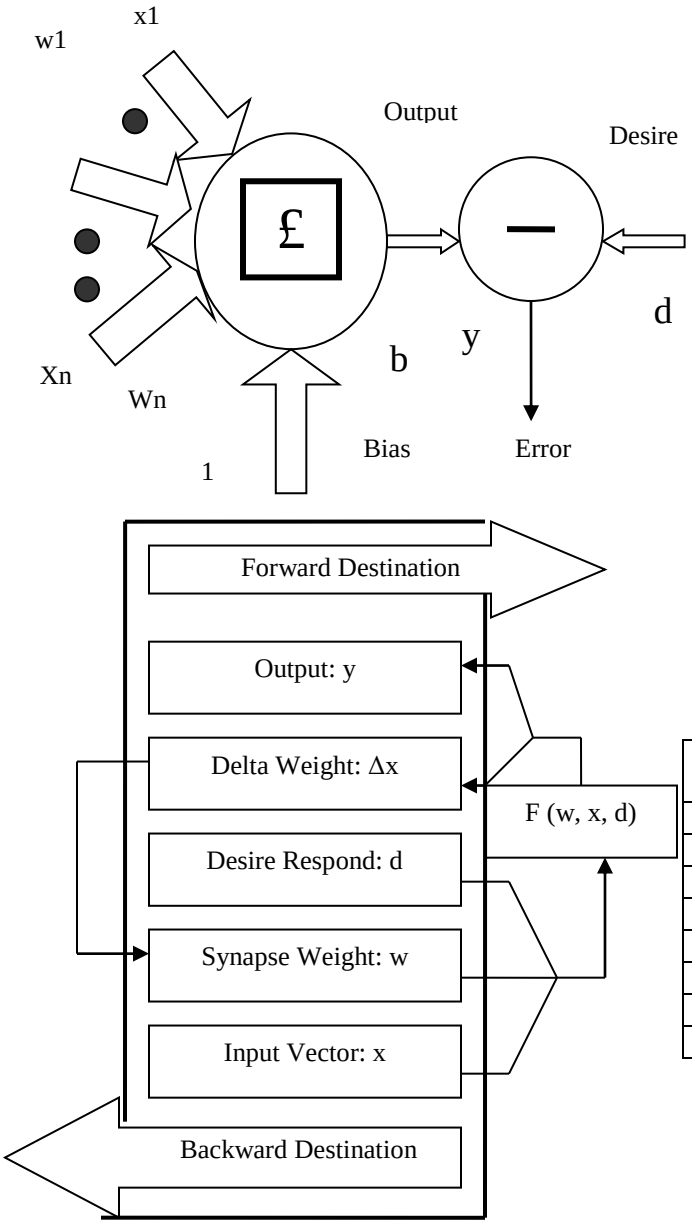

Fig: 2-Architecture of Neural Network
The Sigmoid function makes use of in neuron model

$$f(y) = 1/1+e^{\lambda y}$$

$$f(y) = \{ y, \text{if } y \geq \Theta$$
$$0, \text{if } y < \Theta$$

Neuron error signal cpj output is shown in below

$$Cpj=(Wpj-Qpj)\ Qpj\ (1-Qpj)$$

- Cpj is the aimed value , which is the output neuron m for pattern p*Opj* is the actual output value of output m for example m (m-output of neuron model)
- The concealed neuron fault signal cpj is specified by where *dpk* is the error signal of a postsynaptic neuron n and Wkj is the weight of the connection from hidden neuron j to the post-synaptic neuron k Compute weight adjustments DWji at time t by DWji(t)= η dpj Opi. Apply weight adjustments according to Wji(t+1) = Wji(t) + DWji(t)

Method

In this method, we give discussed algorithm feature values as an input to back propagation. These attributes are obtained from above discussed algorithms. The above algorithms are taking the frequent data values from, when users logged into number of times in the web server or system. The frequent data items are given to appropriate algorithm and collect the output from the particular algorithm. Each pattern mining process algorithms are processed by this same frequent data item sets. Based on the output the back propagation input values are desired.

Table 1: Student Frequent Data Item Sets

S.No	Name	Data of Birth	Username	Time
1	Raj	15/05/1990	raj	10 am
2	Viki	15/05/1991	viki	10am
3	Ram	15/05/1990	ram	11am
4	Raj	15/05/1991	raj	11am
5	Raj	15/05/1990	raj	11am
6	Viki	15/05/1991	raj	11am
7	Arun	15/05/1990	arun	10am
8	Ram	15/05/1991	ram	12am

Above mentioned frequent data item sets are given input to the above discussed data mining algorithms. After that, we justify the algorithm performance i.e. efficiency of the proposed algorithms.

Table 2: Obtain Data Values of each Algorithm

Attributes	Apriori algorithm	Pincer search algorithm	FP-Tree Algorithm

The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, Special Issue:
The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE'13],
San Francisco State University, CA, U.S.A., March 2013
Doi: 10.7321/jscse.v3.n3.34 e-ISSN: 2251-7545

Speed (per second	5	30	10
Frequent Pattern Generation(per 100 data values)	90	80	70
Quality	99	95	198

In above Table 2 data are generated, based on frequent pattern data items detection and elimination of frequent data sets by the discussed algorithms. These numerical values are submitted into the input of back propagation method. The back propagation method is identifying the best algorithm based on these data values.

IV. Performance Evaluation

The best algorithm is selected depend upon the selected attributes. In which, the input key to the neural network is the fragment equivalent to each attribute. The input will be a row, which includes significance of the attribute. The values of some parameters may be kept to levels. The different points are concerned here,

1. From the criteria chosen module, the values are taken to parameter and then passed.
2. Based on these values a matrix is formed, which is input values to a training function
3. The input parameters are mapping. The input mapping result values are previously stored.
4. The result in the outline of decimal values, which match completely three algorithms.
   i. Pincer search : .0001
   ii. A-priori : .0003
   iii. FP-tree : .0002
5. The top value has to prefer the equivalent algorithm.

For the above particularly mentioned result, the selected method is "A-Priori". The result is printed in a file. This is provided as input to the Java program, which implements the matching method.

IV.RESULTS

With the help of Java program, the back propagation algorithm is executed with the given sample datasets. The proposed models are trained with that of the hidden layers with threshold value of 0.001. This modified weight is scheduled with specifications of the specified algorithm.

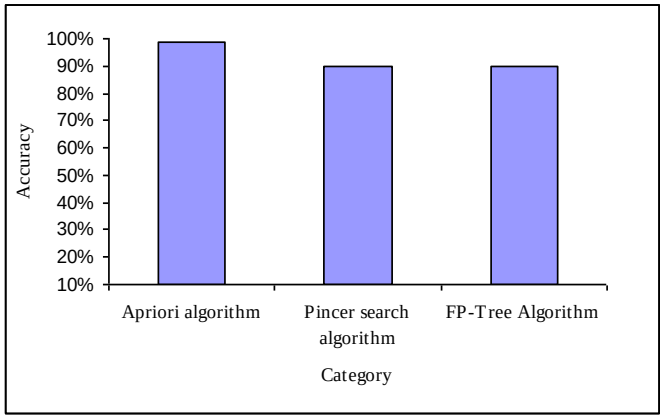

Fig.3 Performance Accuracy of Algorithms

With the describe requirement, the three algorithms are experimentally done and established, that which is best algorithm to get the frequent data items by eliminating the duplicate record sets.

Fig .3 shows the Apriori methods, which provide the best accuracy values to number of frequent pattern item sets present in the database.

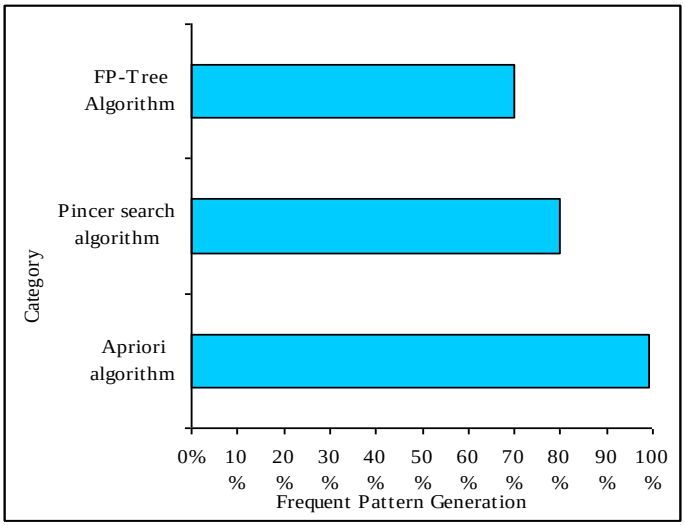

Fig .4 Frequent Pattern Generations

Fig .4 shows the Apriori methods, which provide the best frequent pattern generation in number of frequent item sets present in the data base with less number of duplicate record sets.

The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, Special Issue:
The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE'13],
San Francisco State University, CA, U.S.A., March 2013
Doi: 10.7321/jscse.v3.n3.34 e-ISSN: 2251-7545

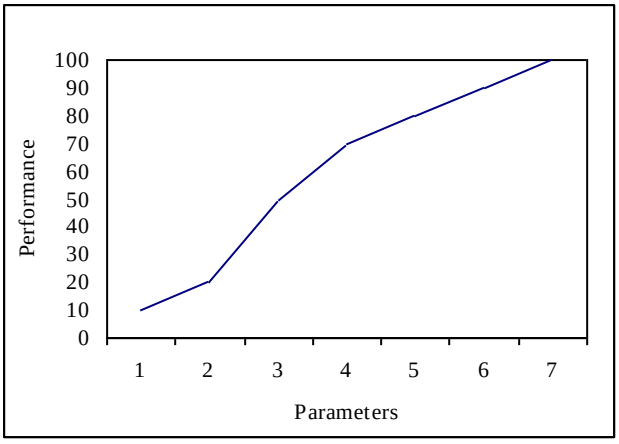

Fig.5 Proposed Tool Performance

Fig .5, shows the proposed neural network method which provides the best performance when number of parameters that are used in the selection of algorithm.

V. CONCLUSION

Data mining plays a vital role in large data sets, which brings out the most effective information retrieval for hidden patterns from the large data set. The proposed system uses a data miner approach in many ways to process the data and make it to appear as an underlying association between attributes. Also, the proposed system can facilitate to mine any type of dataset supplied by the user. The resultant pattern, which we obtained in the proposed system, can help the data miners to create informed decisions. In order to identify the best algorithm for the frequent pattern mining, the neural network concept is used for selecting the best algorithm. The proposed system uses back propagation method in a neural network. Finally, our the results are shown by proposed tool i.e. back propagation which select a good data mining algorithm to data miner for frequent data set item deleting in the database.

REFERENCE

[1] William F. Punch, "Using Genetic Algorithms for Data Mining Optimization in an Educational Web-based System". Proceedings of springer GECCO-2003,pages 2252-2263.

[2]Y. Li, W. Yang "Multi-Tier Granule Mining for Representations of Multidimensional Association Rules" proceedings of IEEE international conference on data mining , IEEE computer society press, California, USA, 2006..

[3] Y. Li, C. Zhang, and J.R. Swan, "An information Filtering Model on the Web and Its Application in Job agent," Knowledge-Based Systems, IOS press 2006.

[4] Y. Li, X. Zhou, P. Bruza, Y. Xu, and R.Y. Lau, "A Two-Stage Text Mining Model for Information Filtering," IEEE computer society press 2008

[5] Y. Li, X. Zhou, P. Bruza, Y. Xu, and R.Y. Lau, "A Two-Stage Text Mining Model for Information Filtering," . conference proceedings of the 17th ACM conference on information and knowledge management, pages 1023-1032,2008

[6] Christian Borgelt, "An Implementation of the FP-growth Algorithm". Proceedings of the 1st international workshop on open source data mining implementations, pages 1-5,2005.

[7] Singh Chauhan et al, "neural networks in data mining" JTAIT vol-5,no-1 2009.

[8] R. Gavald`a and O. Watanabe. Sequential sampling algorithms: Unified analysis and lower bounds. In SAGA, pages 173–188, 2001.

[9]J. Kivinen and H. Mannila. The power of sampling in knowledge discovery. In PODS, pages 77–85, 1994.

[10] T. Sche!er and S. Wrobel. Finding the most interesting patterns in a database quickly by using sequential sampling. Journal of Machine Learning Research, 3:833–862, 2002.

[11] N.Zong and L.Zhou. methodologies for knowledge discovery and data mining, lecture notes in artificial intelligence 1574, springer 1999.