Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 9, September 2015, pg.8 – 15

RESEARCH ARTICLE

ARTIFICIAL INTELLIGENCE BASED CLUSTER OPTIMIZATION FOR TEXT DATA MINING

Er.Poonam, Dr. Rajeev Dhaiya

Computer Science & Punjab Technical University, India

er.poonam.cse88@gmail.com; raj8878@gmail.com

***Abstract*— The relationship between data mining and evaluation system disciplines need to be emerged and focused for better decision making. Studying the current environment to apply data mining to search, sort, group and categorize the relevant information to perform DM(Data Mining).An Academic Evaluation Support System that utilizes both quantitative and qualitative information is needed in the current scenario. There is need to apply mining algorithms for clusters in academic evaluation system. There is also need to develop effective means for data co-relation and data reduction. New mining and search algorithms capable of extracting more complex relationships between fields and able to account for structure over the fields. In our research we are doing clustering using k-mean and optimize it using HBO and ACO and then comparison is done by using the parameters cohesion, variance, precision and recall.**

Keywords*— *"ACO", "HBO" , "clustering ", "Data mining" ,"k-mean"

I. INTRODUCTION

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), an interdisciplinary subfield of computer science, 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 mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside 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.

Data mining involves six common classes of tasks:

- Anomaly detection (Outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation.

- Association rule learning (Dependency modelling) – Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits.

- Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.

- Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".

- Regression – Attempts to find a function which models the data with the least error.

- Summarization – providing a more compact representation of the data set, including visualization and report generation.

Proposed Work

Cluster is a group of objects that belong to the same class. In other words the similar object are grouped in one cluster and dissimilar are grouped in other cluster. Clustering is the process of making group of abstract objects into classes of similar objects. A cluster of data objects can be treated as a one group. While doing the cluster analysis, we first partition the set of data into groups based on data similarity and then assign the label to the groups.

In previous work, they propose a new rule-based classification method for text data based on an efficient closed itemset mining algorithm called Linear time Closed itemset Miner (LCM). In our proposed work, we will propose an efficient mining based hybrid optimization technique for cluster optimisation. By using k-mean algorithm we will find the clusters. Then we apply HBO & ACO for optimizing the clusters.

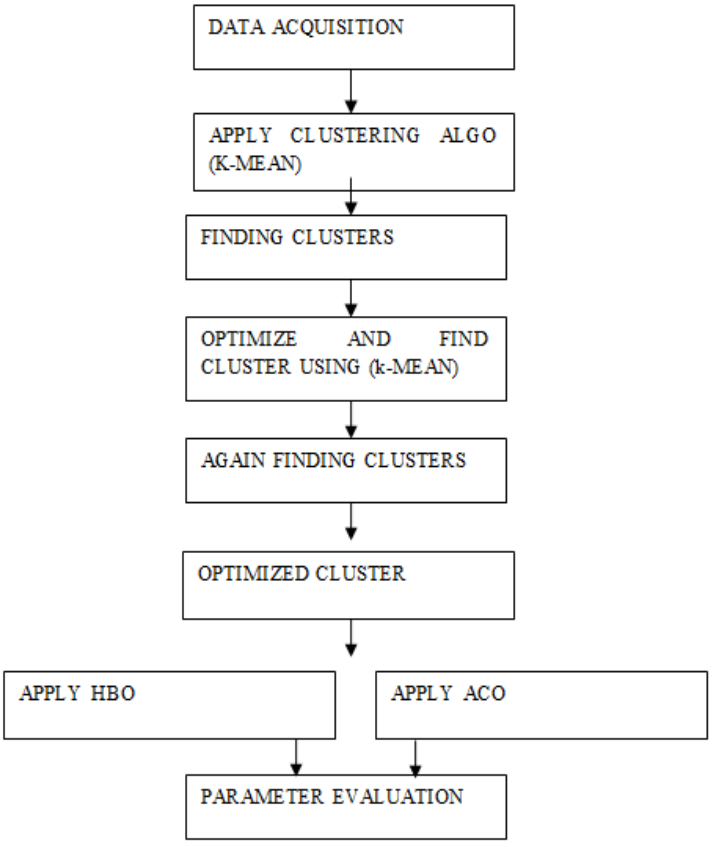

Fig 1: FLOW OF PROPOSED WORK

a. Data Acquisition – It converts the sample into digital form that are manipulated by a computer.

b. Applying algo – Applying clustering algorithm k-mean.

c. Finding clusters - Finding clusters on the basis of algorithm and distance.

d. HBO and ACO Optimization and cluster using k-mean.

e. Parameter evaluation - Now parameters are evaluated and comparison is done on the basis of precision, recall and e-measure.

HONEYBEE ALGORITHM USED FOR DATA CLUSTERING

Honeybee algorithm starts with the generating initial clusters by randomly search in the data points. Then, the fitness of the clusters selected is evaluated. The best *m* sites will be selected from *n*. The evaluation of fitness leads to selecting m clusters from n. In this m selected sites, e sites are introduced as good selected sites and other (*m*-*e*) sites are introduced as bad selected ones.

A neighborhood search sites of size *ngh* is selected for each m sites. In this step a neighborhood size *ngh* is determined which will be used to update the m clusters declared in the previous step. Number of bees (*n2*) will be selected randomly to be sent to e sites and choosing n1 bees randomly which their number is less than n2, to be sent to (*m*-*e*) sites. In this step, recruit bees for the selected sites and evaluate the fitness of the sites.

Finally, the best cluster from each site (the highest fitness) is chosen to form the next bee population. The remaining bees for initialing new population will be assigned randomly around the search space. This algorithm is repeated until the stopping criterion is met. Usually stopping criteria is the number of the repetitions. Algorithm for Honeybee Optimization for data clustering.

- 0. Begin
- 1. Generate initial cluster centroids randomly from data points.
- 2. Evaluate fitness of the clusters.
- 3. While (stopping criterion not met)
- 3.1. Forming new clusters.
- 3.2. Select sites for neighborhood search.
- 3.3. Recruit bees for selected sites (more bees for best e sites) and evaluate fitness.
- 3.4. Select the fittest bee from each patch.
- 3.5 Sort the result based on fitness.
- 3.6. Allocate the remaining clusters to search randomly and evaluate their fitness.
- 9. End While.
- 10. End.

RESULTS AND DISCUSSIONS

The interface used for regionalization of spatial object is shown in Figure. GUI works as follows:

The main central GUI is linked to three windows.

a. SELECT A DATASET
b. K-MEAN ALGORITHM
c. OPTIMIZATION TECHNIQUES

On clicking on the button SELECT DATASET as shown in Figure 2 a new pop-up window will be opened as shown in Figure 3. User can select different spatial dataset according to the choice.

On clicking on the button K-MEAN ALGORITHM a new window will be opened as shown in Figure

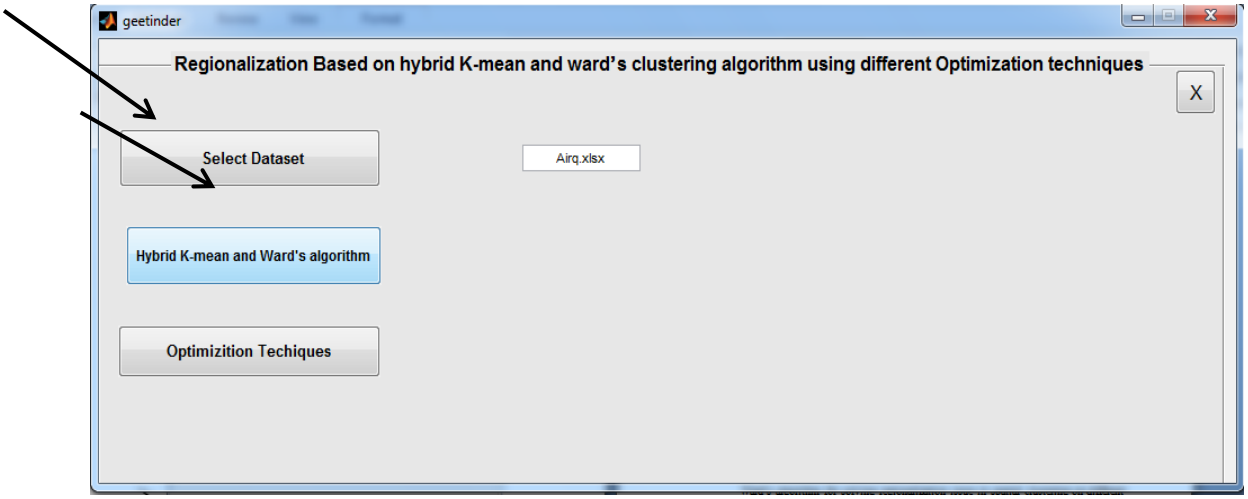

Fig 2: Snapshot of GUI used for Applying K-mean Algorithm for Regionalization

User when press the button the results after applying K-mean algorithm for solving regionalization issue in spatial clustering on different parameters i.e. Cohesion, Variance, Precision, Recall, F-measure and H-measure are shown.

After calculating the value for hybrid algorithm, click on BACK button. When we click on BACK Button ,it shows main window.

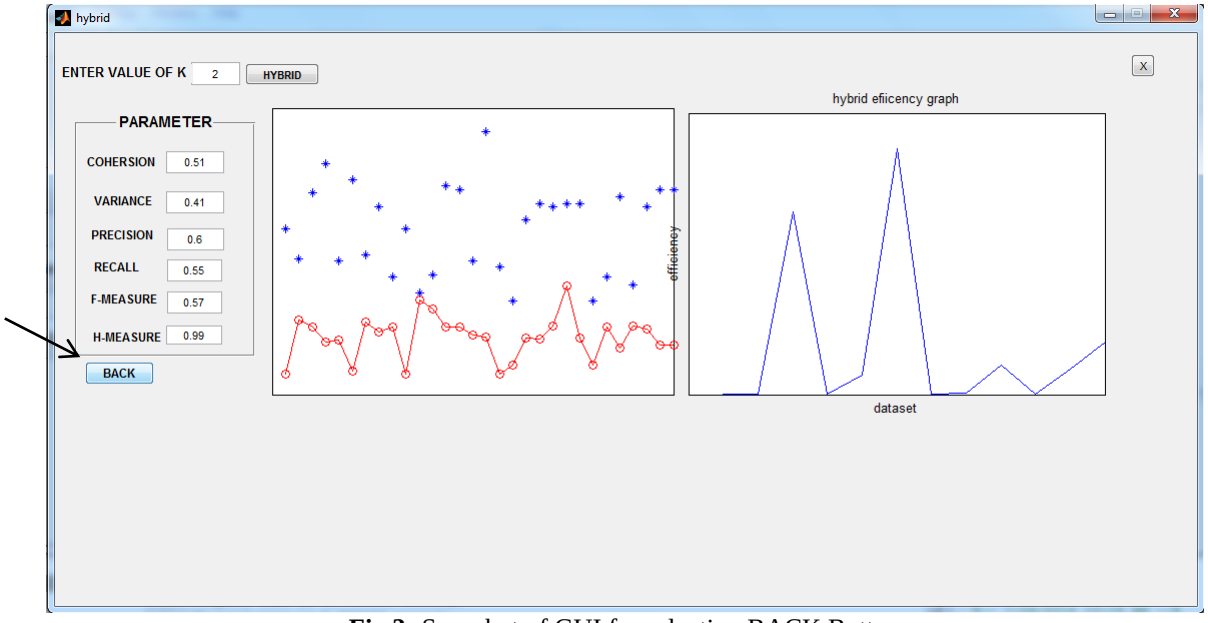

Fig 3: Snapshot of GUI for selecting BACK Button

After applying Hybrid K-mean and Ward's algorithm for solving regionalization problem, now we apply different optimization techniques on the result of hybrid algorithm to improve the efficient of clustering spatial objects in Figure.

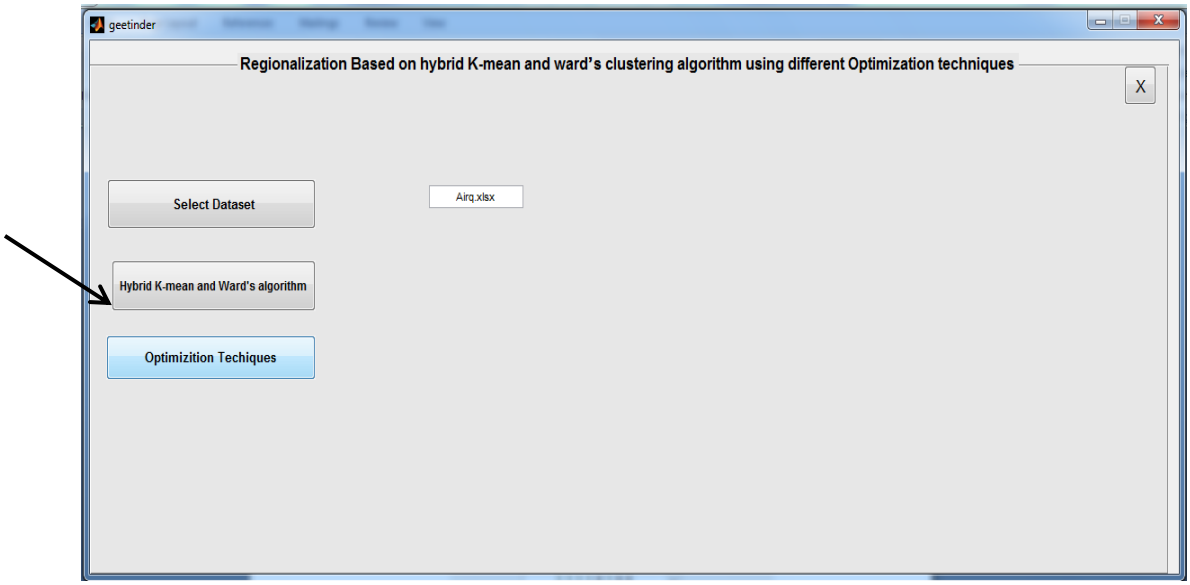

Fig 4: Snapshot of GUI showing selecting of OPTIMIZATION TECHNIQUES Button

After opening OPTIMIZATION TECHNIQUES window we have two options to get efficient and homogenous cluster for Regionalization i.e. using HBO Algorithm or ACO Algorithm as shown in Figure. When we click on "OPTIMIZATIOM TECHNIQUES" the window shown in the Figure is Open

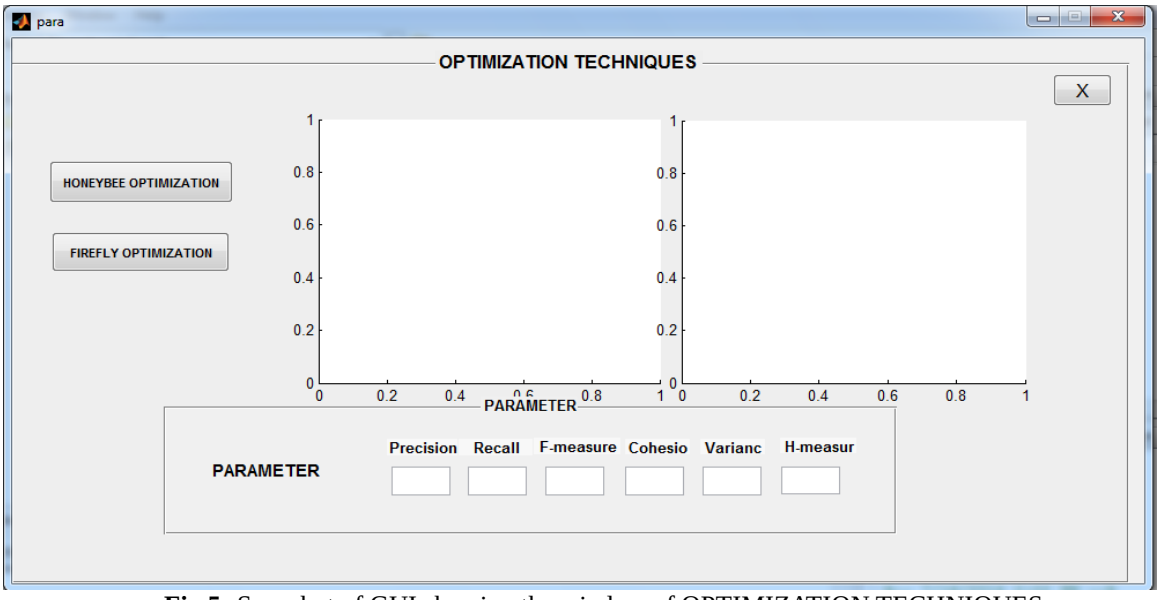

Fig 5: Snapshot of GUI showing the window of OPTIMIZATION TECHNIQUES

When we press the button HONEYBEE OPTIMIZATION the results for solving regionalization issue in spatial clustering on different parameters i.e. Cohesion, Variance, Precision, Recall, F-measure and H-measure are shown and also figure of Honeybee optimization and clusters comes after optimizing data.

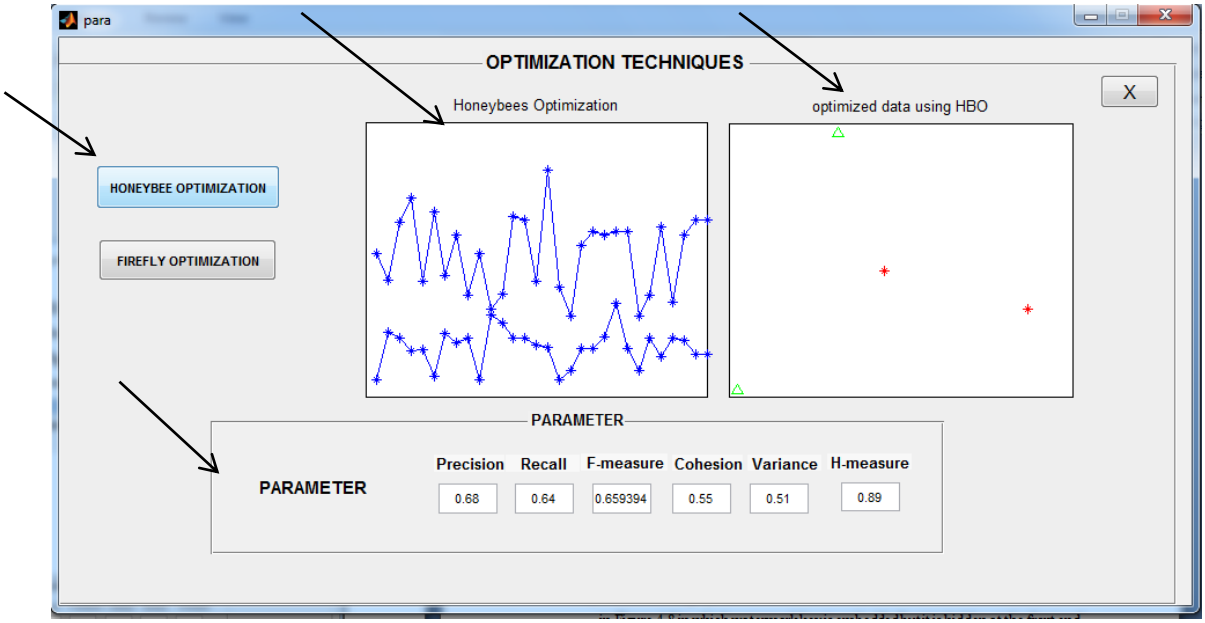

Fig 6: Snapshot of GUI showing the results of Honeybee Optimization algorithm for doing Regionalization on different Parameters.

When we press the button ACO OPTIMIZATION the results for solving regionalization issue in spatial clustering on different parameters i.e. Cohesion, Variance, Precision, Recall, F-measure and H-measure are shown and also figure of Honeybee optimization and clusters comes after optimizing data.

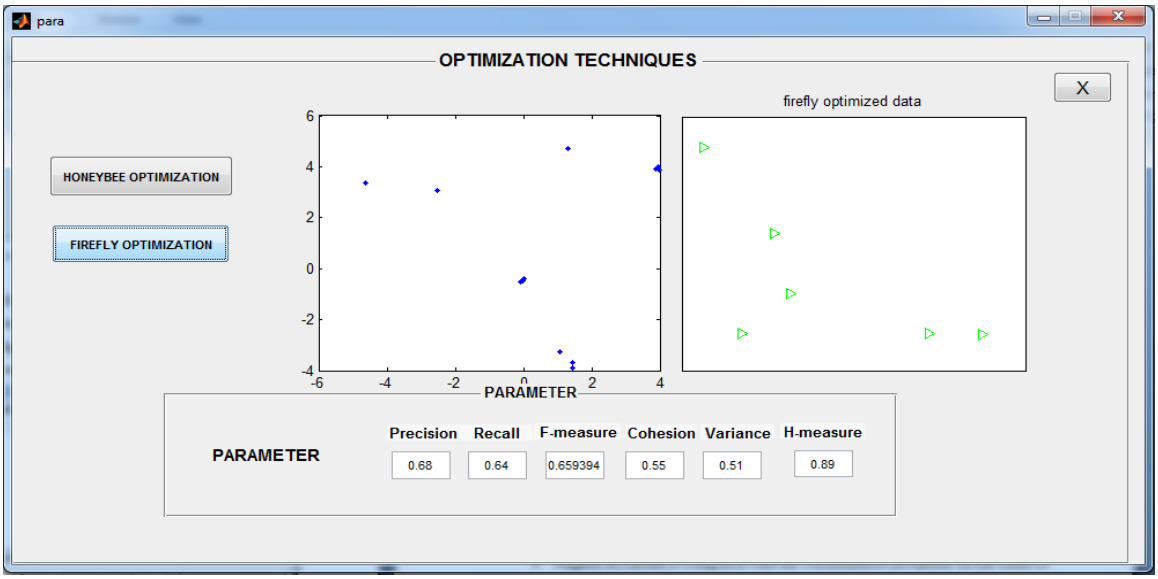

Fig 7: Snapshot of GUI showing the results of ACO Optimization algorithm for doing Regionalization on different Parameter

No of cluster		k-means	HBO	ACO
2	cohesion	.3079	.7850	.79
	variance	.1634	.2910	.2930
	Precision	.20	.32	.33
	recall	.29	.31	.30
4	cohesion	.4008	.6221	.6431
	variance	.2010	.2301	.2120
	Precision	.32	.36	.37
	recall	.26	.34	.32

6	cohesion	.5087	.8012	.8120
	variance	.2816	.3810	.3712
	Precision	.49	.52	.51
	recall	.31	.35	.34
8	cohesion	.3201	.4903	.49
	variance	.2034	.2506	.2101
	Precision	.38	.42	.41
	recall	.25	.30	.30

By above evaluation of parameter shows , artificial intelligence are better optimizer and can give efficient results for text mining.

CONCLUSION

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining software is one of a number of analytical tools for analyzing data allowing users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, text mining is the process of finding correlations or patterns among dozens of fields in large relational Text Databases. In our research we applied clustering technique using k-mean and optimize it using HBO and ACO which gives us optimized clusters and then parameter evaluation is done. Comparison is also done on the basis of some parameters like cohesion, variance, precision and recall. In our work evaluation of parameter shows , artificial intelligence are better optimizer and can give efficient results for text mining.

FUTURE WORK

In future we can apply same algorithms in video mining to find better clusters.

REFERENCES

[1]Qi Luo "Advancing Knowledge Discovery and Data Mining"-Workshop on Knowledge Discovery and Data Mining 2008.

[2]Ji-weiZhu,Yu-guo Tang "A Dynamic Data Mining Model for Engineering Management"-ISECS International Colloquium on Computing, Communication, Control, and Management 2009.

[3]Linna Li, Bingru Yang, Faguo Zhou "A Framework for Object-Oriented Data Mining"-Fifth International Conference on Fuzzy Systems and Knowledge Discovery2010.

[4]DezhenFeng,ZaimeiZhang,FangZhou,JianhengJi[4] "Application Study of Data Ming on Customer Relationship Management in E-commerce".

[5]BhavaniThuraisingham "Data Mining for Malicious Code Detection and Security Applications"-IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology – Workshops 2012.

[6] Longbing Cao " Domain Driven Data Mining (D^3M)" - IEEE International Conference on Data Mining Workshops 2008.

[7] Hidenao Abe "Developing an Integrated Time-Series Data Mining Environment for Medical Data Mining"-Seventh IEEE International Conference on Data Mining Workshops

[8] Fasong Wang, HongweiLi,Rui Li "Data Mining with Independent Component Analysis"-Proceedings of the 6th World Congress on Intelligent Control and Automation, June 21 - 23, 2006, Dalian, China

[9] Dr. JaideepSrivastava "Data Mining for Social Network Analysis"- IEEE ISI 2008 Invited Talk (III)

[10] He YueShun, Ding QiuLin "Application research of data mining architecture for intelligent decision" -Asia-Pacific Conference on Information Processing 2009

[11] Aihua Li, Lingling Zhang "A Study of the Gap from Data Mining to its Application with Cases"-International Conference on Business Intelligence and Financial Engineering 2009

[12] Luo Fang, QiuQizhi "The Study on the Application of Data Mining Based on Association Rules" -International Conference on Communication Systems and Network Technologies (2012).

[13] Xindong Wu "Data Mining: Artificial Intelligence in Data Analysis" -Proceedings of the IEEE/WIC/ACM International Conference on Intelligent Agent Technology 2004

[14] Riccardo Mazza, VaniaDimitrova "CourseVis: Externalising Student Information to Facilitate Instructors in Distance Learning " Proceedings of the International conference in Artificial Intelligence in Education.2004

[15] BehrouzMinaei-Bidgoli , Deborah A. Kashy , GerdKortemeyer, William F. Punch "Predicting Student Performance: An Application Of Data Mining Methods With The Educational Web-Based System Lon-Capa" 33rd ASEE/IEEE Frontiers in Education Conference 2003

[16] KalinaYacef "The Logic-ITA in the classroom: a medium scale experiment" - International Journal of Artificial Intelligence in Education 2005

[17] Cristóbal Romero , Sebastián Ventura , Carlos de Castro , Wendy Hall , and Muan Hong Ng "Using Genetic Algorithms for Data Mining in Web-based Educational Hypermedia Systems "
[18] Heiner, C., Beck, J. E., &Mostow,J.(2004, June 17-19). Improvingthe Help Selection Policy ina Reading Tutor that Listens. Proceedings of theInSTIL/ICALL Symposium on NLPand Speech Technologies in AdvancedLanguageLearningSystems
[19]http://www.theiia.org/intAuditor/itaudit/archives/2006/august/data-mining-101-tools-and techniques
[20] http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html
[21] http://www.laits.utexas.edu/~anorman/BUS.FOR/course.mat/Alex/

