International Journal of Computing and Corporate Research
ISSN (Online) : 2249-054X
Volume 4 Issue 5 September 2014
International Manuscript ID : 2249054XV4I5092014-04

THE EFFECTIVE UNSUPERVISED ALGORITHM FOR DYNAMIC CLUSTERING IN MULTIPLE APPLICATIONS

Aditi Chawla
M.Tech. Research Scholar
Department of Computer Science and Engineering
Punjab Technical University
Jallandhar, Punjab, India

Navneet Kaur
Assistant Professor
Department of Computer Science and Engineering
RIMT Institute of Engineering and Technology
Mandi Gobindgarh, Punjab, India

ABSTRACT

Clustering is an important knowledge discovery technique in the domain of data mining with numerous applications, such as marketing and customer segmentation. Clustering typically groups data into sets in such a way that the intra-cluster similarity is maximized and while inter-cluster similarity is minimized. Clustering is an unsupervised learning. Clustering algorithms examines data to find groups of items that are similar. For example, an insurance company might group customers according to income, age, types of policy purchased, prior claims experience in a fault diagnosis application, electrical faults might be grouped according to the values of certain key variables. In this research work, we have proposed and implemented the algorithmic approach for dynamic unsupervised cluster formation that can be used in multiple domains. The test data sets are taken in the domains of shopping cart and network packets. The proposed algorithmic approach is making use of fuzzy based dynamic cluster formation and the research objectives are fulfilled as there is less complexity and effective output from the proposed approach.

KEYWORDS – Data Mining, Dynamic Clustering, Fuzzy Clustering

DATA MINING

Data mining refers to the analysis of the large quantities of data that are stored in computers. Data mining has been called exploratory data analysis, among other things. Masses of data generated from cash registers, from scanning, from topic specific databases throughout the company, are explored, analyzed, reduced, and reused. Searches are performed across different models proposed for predicting sales, marketing response, and profit. Classical statistical approaches are fundamental to data mining. Automated AI methods are also used. Data mining requires identification of a problem, along with collection of data that can lead to better understanding and computer models to provide statistical or other means of analysis.

Data comes in, possibly from many sources. It is integrated and placed in some common data store. Part of it is then taken and pre-processed into a standard format. This ‘prepared data’ is then passed to a data mining algorithm which produces an output in the form of rules or some other kind of ‘patterns’.

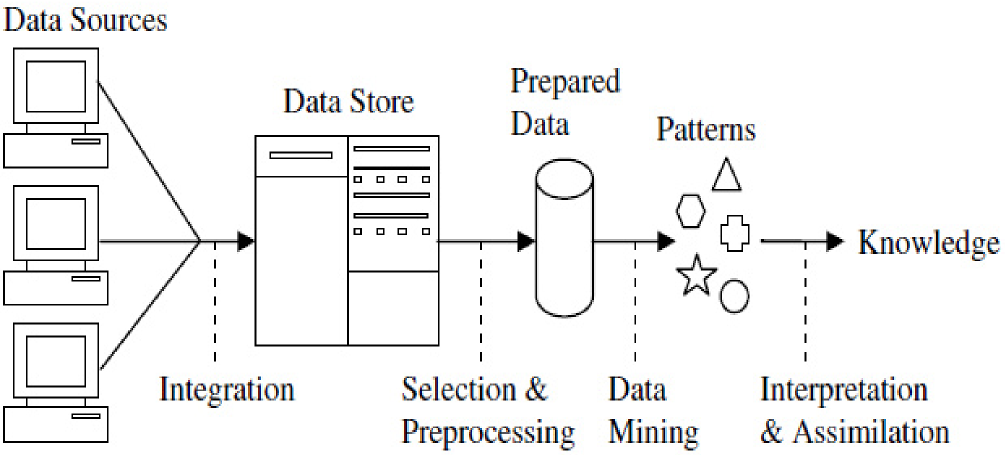

Figure 1: The Knowledge Discovery Process

A variety of analytic computer models have been used in data mining. The standard model types in data mining include regression (normal regression for prediction, logistic regression for classification), neural networks, and decision trees.

CLUSTERING PARADIGM

Clustering is an important KDD technique with numerous applications, such as marketing and customer segmentation. Clustering typically groups data into sets in such a way that the intra-cluster similarity is maximized and while inter-cluster similarity is minimized. Clustering is an unsupervised learning. Clustering algorithms examines data to find groups of items that are similar. For example, an insurance company might group customers according to income, age, types of policy purchased, prior claims experience in a fault diagnosis application, electrical faults might be grouped according to the values of certain key variables.

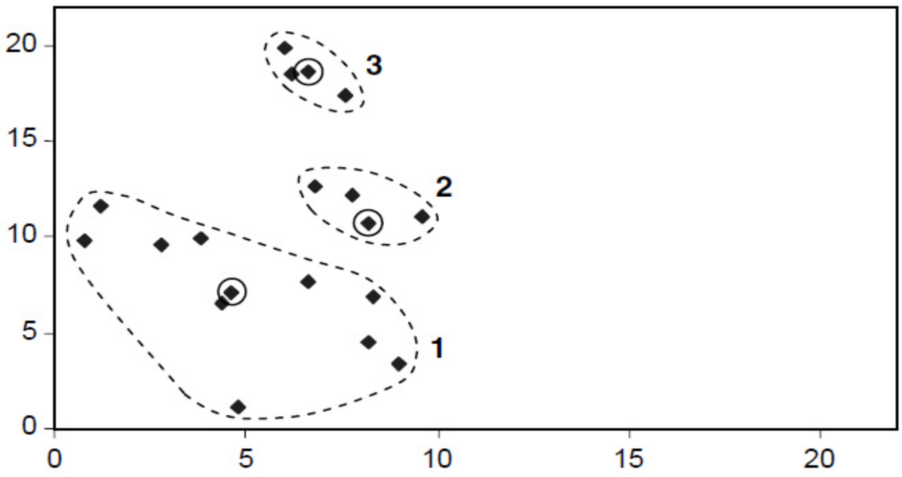

Figure 2: Clustering of Data

Most previous clustering algorithms focus on numerical data whose inherent geometric properties can be exploited naturally to define distance functions between data points. However, much of the data existed in the databases is categorical, where attribute values can't be naturally ordered as numerical values. Due to the special properties of categorical attributes, the clustering of categorical data seems more complicated than that of numerical data. To overcome this problem, several data-driven similarity measures have been proposed for categorical data. The behaviour of such measures directly depends on the data.

GOAL OF CLUSTERING

The objective of clustering is to figure out the innate amassing in a set of unlabeled information. Yet how to choose what constitutes a great grouping? It could be indicated that there is no supreme "best" rule which might be free of the last point of the bunching. Thus, it is the client which should supply this basis, in such a path, to the point that the aftereffect of the grouping will suit their necessities. Case in point, we could be fascinated by finding delegates for homogeneous gathers (information decrease), in finding "characteristic groups" and portray their obscure lands ("common" information sorts), in finding of service and suitable groupings ("helpful" information classes) or in finding irregular information objects (outlier location).

OBJECTIVES

- *To devise and implement a novel and efficient technique for dynamic as well as effective cluster formation.*
- *To apply and fetch the meaningful records in form of the aggregate values or clusters for intelligence.*
- *To analyze the proposed cluster formation algorithm with the existing technique and to prove the effectiveness of the proposed work.*

- *To devise a novel fuzzy fitness function to the transactional data so that the eligibility or relevance of the record can be analyzed*

IMPLEMENTATION ASPECTS

1. To devise and implement a novel and efficient technique for dynamic as well as effective cluster formation.
2. To apply and fetch the meaningful records in form of the aggregate values or clusters for intelligence and predictions.
3. To analyze the proposed cluster formation algorithm with the existing technique and to prove the effectiveness of the proposed work.
4. To devise a novel fitness function to the transactional data so that the eligibility or relevance of the record can be analyzed.

MATHEMATICAL FORMULATION

- At the very first level, there is a Data Items Repository (D) related to the Shopping Mart or Network Packet Analysis.
- In the Data Repository, there will be number of transactions or data sets or transactional data.
- Each transaction or record or data set is termed as R(Ti).
- The Data Set regardless of the value and associated parameters will be eligible and move forward the fitness function modelling.
- FLD refers to the fitness level of the data items.
- Once the fitness level of the data items is applied successfully in the unbiased and unsupervised learning based measurement, it will generate the novel criteria for the addition in the dynamic clusters.
- At the final level, the set of clusters are generated with effective results and optimal parameters in terms of time.

CLUSTERING ALGORITHMIC APPROACH

1. Generation of the Dataset (Sequentially / Randomly) / Tuple series (fetched from large and big data based warehouse)

   $|DT|= \{ DTm \mid m \in (1,N) \}$

2. Association of the Qualification Value (QVi) to each data item that is based on the Acceptability / Avoidance of the Data Item for joining the Dynamic Cluster Formation Modules

   $DT= \{ DTi[QVi] \mid i \in (1,N) \}$

3. Generation of the random clusters sets (if already exists)

   $DYC= \{ DYC\ i \mid i \in (1,N) \}$ and assign Threshold/Qualification Thrust

4. Compare DT with DYC based on Qualification value and Joined Parameters
5. If (DYCi == NULL) AND QTfitness==NULL GoTo Step 7

```
        Else
        If (DYCi= Initial Cluster)
                Assignment of the first thrust value based on the application
        Else
                GoTo Step 1
        Generate Final Results from both algorithms and calculate complexity
    6. End
```

TEST DATA SET

Dynamic Clustering Algorithm based on Multi-Pass Fuzzy Based Partitioned Clustering

Product ID	**Price**	**WeekDay**	**Date**	**Items**
001	1000	Thursday	21-February-2014	5
001	1000	Saturday	5-February-2014	43
001	1000	Friday	6-March-2014	78
001	1000	Thursday	23-April-2014	54
001	1000	Tuesday	14-February-2014	78
001	1000	Tuesday	9-March-2014	68
001	1000	Wednesday	13-April-2014	56
001	1000	Thursday	29-February-2014	90
001	1000	Tuesday	28-March-2014	57
001	1000	Saturday	30-April-2014	67
001	1000	Tuesday	18-April-2014	57
002	2000	Sunday	12-April-2014	18

002	2000	Monday	13-February-2014	46
002	2000	Tuesday	17-April-2014	8
002	2000	Sunday	1-January-2014	69
002	2000	Monday	15-February-2014	50

Aggregation and Analysis of the Sold Items based on the ProductType

```
ID (012) | Count 2487 | Pass-1 Similarity Measure : 14. => 0.0062439441680908 ms
ID (018) | Count 1540 | Pass-1 Similarity Measure : 9.1 => 0.0065557956695557 ms
ID (017) | Count 1314 | Pass-1 Similarity Measure : 7.7 => 0.0068409442901611 ms
ID (015) | Count 1126 | Pass-1 Similarity Measure : 6.6 => 0.0071499347686768 ms
ID (003) | Count 1089 | Pass-1 Similarity Measure : 6.4 => 0.0074269771575928 ms
ID (002) | Count 934 | Pass-1 Similarity Measure : 5.5 => 0.0077288150787354 ms
ID (016) | Count 875 | Pass-1 Similarity Measure : 5.1 => 0.0080068111419678 ms
ID (011) | Count 823 | Pass-1 Similarity Measure : 4.8 => 0.0083069801330566 ms
ID (013) | Count 799 | Pass-1 Similarity Measure : 4.7 => 0.0086228847503662 ms
ID (019) | Count 682 | Pass-1 Similarity Measure : 4.0 => 0.00895094871521 ms
ID (001) | Count 653 | Pass-1 Similarity Measure : 3.8 => 0.0092759132385254 ms
ID (005) | Count 642 | Pass-1 Similarity Measure : 3.8 => 0.0096218585968018 ms
ID (009) | Count 609 | Pass-1 Similarity Measure : 3.6 => 0.009929895401001 ms
ID (004) | Count 585 | Pass-1 Similarity Measure : 3.4 => 0.010228872299194 ms
ID (014) | Count 579 | Pass-1 Similarity Measure : 3.4 => 0.010524988174438 ms
ID (008) | Count 578 | Pass-1 Similarity Measure : 3.4 => 0.011129856109619 ms
ID (010) | Count 501 | Pass-1 Similarity Measure : 2.9 => 0.011451959609985 ms
ID (007) | Count 493 | Pass-1 Similarity Measure : 2.9 => 0.011785984039307 ms
ID (020) | Count 479 | Pass-1 Similarity Measure : 2.8 => 0.012058973312378 ms
ID (006) | Count 87 | Pass-1 Similarity Measure : 0.5 => 0.012462854385376 ms
```

Percentile Based Measurements for Membership of Data Items

```
ID (012) | | Pass-2 Fraction : 98. => 0.00058484077453613 ms
ID (018) | | Pass-2 Fraction : 60. => 0.00042104721069336 ms
ID (017) | | Pass-2 Fraction : 51. => 0.00083804130554199 ms
ID (015) | | Pass-2 Fraction : 44. => 0.00043296813964844 ms
ID (003) | | Pass-2 Fraction : 43. => 0.00038290023803711 ms
ID (002) | | Pass-2 Fraction : 36. => 0.00039410591125488 ms
ID (016) | | Pass-2 Fraction : 34. => 0.00041389465332031 ms
ID (011) | | Pass-2 Fraction : 32. => 0.00039792060852051 ms
ID (013) | | Pass-2 Fraction : 31. => 0.00039005279541016 ms
ID (019) | | Pass-2 Fraction : 26. => 0.00039196014404297 ms
ID (001) | | Pass-2 Fraction : 25. => 0.00043988227844238 ms
ID (005) | | Pass-2 Fraction : 25. => 0.00040006637573242 ms
ID (009) | | Pass-2 Fraction : 24. => 0.00039100646972656 ms
ID (004) | | Pass-2 Fraction : 23. => 0.00048089027404785 ms
ID (014) | | Pass-2 Fraction : 22. => 0.00045895576477051 ms
ID (008) | | Pass-2 Fraction : 22. => 0.00050520896911621 ms
ID (010) | | Pass-2 Fraction : 19. => 0.00049996376037598 ms
ID (007) | | Pass-2 Fraction : 19. => 0.0004730224609375 ms
ID (020) | | Pass-2 Fraction : 18. => 0.00039100646972656 ms
ID (006) | | Pass-2 Fraction : 3.4 => 0.00043988227844238 ms
```

Max. -> 98 | Min. -> 19 | Avg. -> 33

Dynamic Fuzzy Similarity Factor 6

CLUSTER FORMATION PROCESS BASED ON THE FITNESS FUNCTION VALUES AND THRESHOLD

The proposed approach is making use of Cut / Partitioned Clustering and it is evident from the results that out of assorted records there are very few Outlier Items. The effectiveness of the algorithm is measured from the minimum outlier items.

In the classical approach, all values are considered as the outlier because the existing algorithmic approach is not distributing the data items in equal and meaningful clusters

Super Set of Data Items

```
Array ( [0] => 012 [1] => 018 [2] => 017 [3] => 015 [4] => 003 [5] => 002 [6] => 016 [7] => 011 [8] =>
013 [9] => 019 [10] => 001 [11] => 005 [12] => 009 [13] => 004 [14] => 014 [15] => 008 [16] => 010
[17] => 007 [18] => 020 [19] => 006 )
```

Non Outlier Member Set of Data Items

```
Array ( [0] => 012 [1] => 002 [2] => 016 [3] => 011 [4] => 013 )
```

Outlier Data Items

Array ([1] => 018 [2] => 017 [3] => 015 [4] => 003 [9] => 019 [10] => 001 [11] => 005 [12] => 009 [13] => 004 [14] => 014 [15] => 008 [16] => 010 [17] => 007 [18] => 020 [19] => 006)

Number of Items in the Outlier - 15

It is evident from the simulated environment that the classical technique of the cluster formation is generating the clusters in very vague and very traditional method without intelligence and bound based measurement of each data set. The eligibility and the best fit parameter are nowhere being measured in the existing technique and moreover giving the turnaround time higher than the proposed technique. The classical technique is generating and classifying the data items in the cluster that may not be useful in the knowledge discovery.

In the proposed technique, an exclusive measurement is taken into consideration and implemented on the same transaction data for analysis of the results of the proposed as compared to the existing technique. In the proposed scenario, the results obtained are efficient in terms of the generation and inclusion in the clusters as well as the execution time.

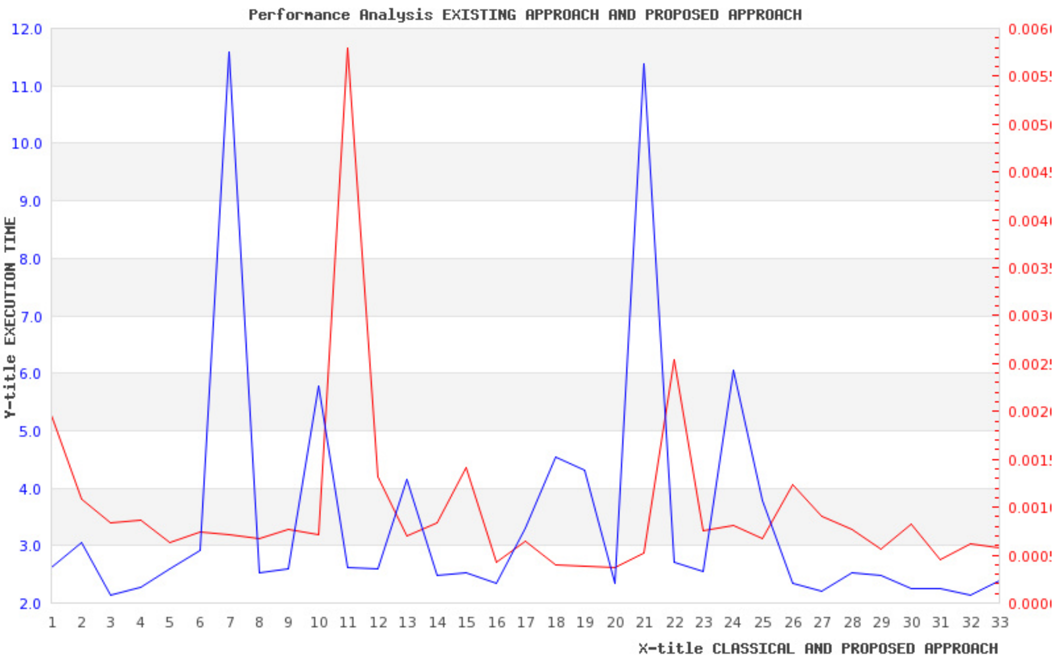

Figure 3 – Comparative Analysis

Using the exclusive and unique way to measure the eligibility as well as the inclusion of the relevant data items in the best fit cluster based on intelligence, the graph has been plotted to represent the pattern and behavior of the cluster formation process. Using the implementation of cluster formation, it is shown that the proposed technique is producing better results as compared to the classical technique. The graph has been plotted for the warehouse records with respect to the execution time. The investigation has been performed for the slabs or layers of the records for the efficiency analysis.

CONCLUSION

Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. A massive amount of research work is under process throughout the globe in assorted algorithms. In this research work, we have proposed and implemented a novel algorithm that makes use of the mathematical foundation and evolutionary approach for the formation of clusters in efficient and effective manners in terms of execution time and associated results. A sample data set of shopping mart has been implemented and the algorithm performs in excellent manner on the desired aspects. This area of implemented is not limited to the shopping and market survey. The presented and implemented approach can be used in multiple diversified areas including web log usage, forensic investigation, pattern analysis, biometric, astronomy and many other streams that require the efficient methods of aggregation simply called clustering. The future scope of the research work can extended to the hybrid approach. The hybrid approach makes use of two or more algorithmic approaches to be merged in single formulation to get the optimal results. The hybrid approach can make use of the ant colony optimization or genetic algorithm to get the optimal results. If the presented algorithm is executed to n iterations with genetic algorithmic approach, the best solution can be achieved. In the future work, the cluster formation can be integrated with best first search of the heuristic search methods for the removal of noise. In this research work, we have presented and implemented a novel and effective algorithmic approach for dynamic unsupervised clustering on multiple data sets in which the fuzzy implemented is done. The Rule Mining and Classification is a well known and decently explored technique for uncovering fascinating relations between variables in expansive databases for cyber security and network investigation. It is planned to recognize solid principles ran across in databases utilizing distinctive measures of interestingness Based on the idea of solid standards, Rakesh Agrawal et al. presented the association guidelines for finding regularities between items in substantial scale transaction information recorded by the points of sales frameworks in market basket analysis domain. Such associated data could be utilized as the premise for choices about showcasing exercises, for example, e.g., limited time evaluating or item positions. Notwithstanding the above illustration from business sector bushel dissection affiliation guidelines are utilized today in numerous requisition territories including As far as the future work is concerned, the individual patterns of each object can be analyzed on the network traffic and log files for deep analysis of the links, platform and behavior of the users.

For future scope of the work, following techniques can be used in hybrid approach to better and efficient results –

- Particle Swarm Optimization
- HoneyBee Algorithm
- Simulated Annealing
- Genetic Algorithmic Approaches

REFERENCES

[1] Achtert, E.; Böhm, C.; Kröger, P. (2006). "DeLi-Clu: Boosting Robustness, Completeness, Usability, and Efficiency of Hierarchical Clustering by a Closest Pair Ranking". LNCS: Advances in Knowledge Discovery and Data Mining. Lecture Notes in Computer Science 3918: 119–128. doi:10.1007/11731139_16. ISBN 978-3-540-33206-0.

[2] Aditya Desai, Himanshu Singh, Vikram Pudi, 2011. DISC: Data-Intensive Similarity Measure for Categorical Data, Pacific-Asia Conferences on Knowledge Discovery Data Mining

[3] Agrawal, R.; Imieliński, T.; Swami, A. (1993). "Mining association rules between sets of items in large databases". Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93. p. 207. doi:10.1145/170035.170072. ISBN 0897915925.

[4] Andre Baresel, Harmen Sthamer, Michael Schmidt,2002. Fitness Function Design to improve Evolutionary Structural Testing

[5] Andrew L.Nelson, Gregory J.Barlow, Lefteris Doitsidis,2008 .Fitness Functions in Evolutionary Robotics: A Survey and Analysis

[6] Barnett, V. and Lewis, T.: 1994, Outliers in Statistical Data. John Wiley & Sons., 3rd edition.

[7] Can, F.; Ozkarahan, E. A. (1990). "Concepts and effectiveness of the cover-coefficient-based clustering methodology for text databases". ACM Transactions on Database Systems 15 (4): 483. doi:10.1145/99935.99938

[8] David L. Olsen, Dursun Delen, Advances data mining techniques, Springer, 2008

[9] Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). "From Data Mining to Knowledge Discovery in Databases". Retrieved 17 December 2008

[10] Hans-Peter Kriegel, Peer Kröger, Jörg Sander, Arthur Zimek (2011). "Density-based Clustering". WIREs Data Mining and Knowledge Discovery 1 (3): 231–240. doi:10.1002/widm.30.

[11] He Zengyou, Xu Xiaofei, Deng Shenchun, 2002. Squeezer: An Efficient Algorithm for Clustering Categorical Data,Journal of Computer Science and Technology,Vol. 17, No. 5,pp 611-624

[12] He Zengyou, Xu Xiaofei, Deng Shenchun, 2003. Discovering Cluster Based Local Outliers,Article Published in Journal Pattern Recognition Letters, Volume 24. Issue 9-10,pp 1641-1650,01 June 2003

[13] He Zengyou, Xu Xiaofei, Deng Shenchun, 2006. Improving Categorical Data Clustering Algorithm by Weighting Uncommon Attribute Value Matches, ComSIS Vol.3,No.1

[14] Jerzy Stefanowski, 2009, Data Mining - Clustering, University of Technology, Poland

[15] Lloyd, S. (1982). "Least squares quantization in PCM". IEEE Transactions on Information Theory 28 (2): 129–137. doi:10.1109/TIT.1982.1056489.

[16] M.Davarynejad, M.-R.Akbarzadeh-T, N.Pariz,2007. A Novel Framework for Evolutionary Optimization: Adaptive Fuzzy Fitness Granulation, IEEE Conference on Evolutionary Computation, pp 951-956,2007

[17] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise". In Evangelos Simoudis, Jiawei Han, Usama M. Fayyad. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231. ISBN 1-57735-004-9.

[18] Max Bramer, Principles of data mining, Springer, 2007

[19] Microsoft academic search: most cited data mining articles: DBSCAN is on rank 24, when accessed on: 4/18/2010

[20] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, Jörg Sander (1999). "OPTICS: Ordering Points To Identify the Clustering Structure". ACM SIGMOD international conference on Management of data. ACM Press. pp. 49–60.

[21] “Outlier Detection in Clustering”
ftp://cs.joensuu.fi/pub/Theses/2005_MSc_Cherednichenko_Svethlena.pdf

[22] R. Ng and J. Han. "Efficient and effective clustering method for spatial data mining". In: Proceedings of the 20th VLDB Conference, pages 144-155, Santiago, Chile, 1994.

[23] R.Ranjini, S.Anitha Elavarasi, J.Akilandeswari.2012. Categorical Data Clustering Using Cosine Based Similarity for Enhancing the Accuracy of Squeezer Algorithm

[24] S Roy, D K Bhattacharyya (2005). "An Approach to find Embedded Clusters Using Density Based Techniques". LNCS Vol.3816. Springer Verlag. pp. 523–535.

[25] Shyam Boriah, Varun Chandola, Vipin Kumar, 2008. Similarity Measures for Categorical Data: A Comparative Evaluation, SIAM International Conference on Data Mining-SDM

[26] Tian Zhang, Raghu Ramakrishnan, Miron Livny. "An Efficient Data Clustering Method for Very Large Databases." In: Proc. Int'l Conf. on Management of Data, ACM SIGMOD, pp. 103–114.

[27] Varun Chandola, Arindam Banerjee, Vipin Kumar. Outlier Detection: A Survey

[28] Z. Huang. "Extensions to the k-means algorithm for clustering large data sets with cate