Dimensionality Reduction Using Clustering Technique

Snehal D.Borase
ME student
Department of Computer Engineering
K. K. W. I. E. E. R., Nasik
S. P. P. U.

Satish S.Banait
Assistant Professor
Department of Computer Engineering
K. K. W. I. E. E. R., Nasik
S. P. P. U.

ABSTRACT
Clustering is a method of finding homogeneous classes of the known objects. Clustering plays a major role in various applications in data mining such as, computational biology, medical diagnosis, information recovery, CRM, scientific data investigation, selling, and web analysis. Most of the researchers have a major interest in designing clustering algorithms. "Big data" involves terabytes and petabytes of data. Big data is challenging because of its five important characteristics such as volume, velocity, variety, variability and complexity. Therefore big data is difficult to handle using conventional tools and techniques. There are so many issues in clustering techniques, so some of the issues is how to process the data and big data is clustered in more compact format, Clustering algorithm suffer from stability problem, ensemble of single and multi level clustering. An important issue in clustering is that we do not have earlier knowledge regarding data. Also selection of input parameters such as number of nearest neighbours, number of clusters in these algorithms makes clustering a challenging task. The main objective is to study and analyze the existing clustering algorithms, impact of dimensionality reduction and dealing with outliers.

General Terms
Objects, Techniques, Dimensionality reduction

Keywords
Clustering algorithms, Big data, Nearest neighbours, Outliers

1. INTRODUCTION
Clustering is a task of an unsupervised classification of a known data items into equivalent classes. Ultimately, clustering is corresponding to classification. Clustering bulky data sets having large no of dimensions is a challenging thing for clustering algorithms. Many newly given clustering techniques tried to deal with data sets having extremely huge number of records. This work gives a brief idea about the compensations and restrictions of existing algorithms in literature when these datasets are operated on big data sets. Also this work overcomes the issues of scalability related to big data. There are some issues of clustering in data mining we are generally concerned with involves-

1) Variety of input parameter algorithm can handle
2) Minimal need for domain knowledge
3) Ability to work with data having large no of records or dimensions
4) Capability of finding clusters of uneven form
5) Robustness to noise
6) Time cost
7) Insensitivity to input order
8) Assignment or labeling (soft or fuzzy vs. hard or strict)
9) Confidence on user defined parameters and a priori knowledge
10) Ability to interpret results

Clustering has become challenging task in the literature of mining data and machine learning. Basically clustering is a method of finding out equivalent classes of the known objects. Most of the researchers have a major interest in designing clustering algorithms. A very important issue in case of clustering technique is that we do not have early knowledge about the data given.

Furthermore, the preference of parameters given as input such as the figure of nearest neighbors and clusters also with other important things in clustering algorithms that make the process of clustering more challenging. One of the crucial ways for dealing with these big data is to categorize or cluster these big data into groups of clusters. Clustering techniques have emerged as substitute dominant tool for meta-learning in order to correctly analyze the huge amount of data comes from many applications. The Big Data refers to data sets that are not only of big volume, but also high in range and velocity, which makes them tricky to handle using conventional tools and techniques. Because of the fast production of such big data, some efficient management is needed for extracting knowledge and value and also for handling this big amount of data. Therefore study of the various types of existing techniques of clustering for large datasets may provide significant and useful conclusions.

Basically a clustering technique that gives better clusters in which the matching inside of the cluster is less and the matching with outside clusters is more is good in performance. The superiority of the clustering algorithm generally evaluated according to the comparison measure used by both the technique and its implementation. Also it is measured by its ability to discover even arbitrary shapes of clusters. However, objective evaluation is problematic and usually done by human or an expert analysis.

In this work we are using two clustering algorithms Hierarchical Clustering Using Representative (CURE) Algorithm and DBSCAN clustering algorithm which is density based algorithm.

2. RELATED WORK
Nowadays, cluster investigation as well as improvement of clustering techniques has become the center of research. Variety of clustering techniques proposed for that purpose but not a single algorithm is appropriate for clustering and partitioning applications. J Mac QUEEN [1] described 'k-means' for clustering an N-dimensional population into k partitions based on a sample, which gives partitions so that they are practically more efficient for within-class variance

logic. After that Expectation Maximization algorithm [2] which overcomes the problems such as- Likelihood, State and Parameter evaluation. Some algorithm [3] uses diagonal, mahalonobis and Euclidean distance. Statistical clustering methods by Jain and Dubes [5] used similarity measures for partitioning objects while, conceptual clustering methods [4] used concepts that object carry for clustering objects.

In 1994 Ng & Han studied partition based algorithms [6]. CLARANS which is an enhanced k-medoid technique and it is based on random searching. Also it is partly motivated by two existing algorithms CLARA and PAM and the improvement to spatial algorithms for mining. Inside BIRCH [7] every clustering choice is done not including the scanning of all sample items BIRCH make use of similarity measures revealing the normal proximity of data, which is incrementally maintained throughout the clustering procedure. BIRCH and CLARANS works well when clusters are spherical or convex with even dimension, but not suitable when clusters are having various dimensions. Density based DBSCAN [8] proposed for clustering arbitrary shaped clusters. In DBSCAN there is requirement of single parameter for as input and it gives proper suggestion to the user for deciding input parameter.

After that in 1997 a basic version of the *k* prototypes [9] algorithm is described by Huang. Well known *k*-means algorithm is only applicable to numeric values. This limitation is overcome by the *k*-modes algorithm. DBCLASD [10] in which the data items are assigned to a cluster based on the data items taken till that point instead of taking the whole database or cluster. It works according to consideration that the data items within cluster are evenly spread. Basically, DBCLASD an incremental algorithm which performs an initial clustering by its neighbouring data items till the closest neighbour distance of the resultant cluster best fits into the distribution of required distance. CURE [11] Clustering by making use of Representative overcomes the limitations of DBSCAN [8] algorithm. CURE can deal with the occurrence of noisy data. Wave Cluster [12] is grid-based and density-based algorithm and also capable of detecting the clusters of non uniform shape. But it is only appropriate for data with low dimensionality.

After DENCLUE, scientists proposed another algorithm [13] created in which first of all a histogram for facts values for every aspect is generated and then level of noise for finding maxima of left and right side is determined. Another hierarchical based clustering algorithm by using dynamic modeling CHAMELEON [14] find outs the clusters within data set. Conventional techniques for clustering which uses distances among data items were inappropriate for attributes of Boolean and categorical type. In 2000 for measuring the likeness/closeness between a pair of data items with categorical attributes new method is proposed which uses links instead of distance [15].

Ng & Han [16] proposed new version of CLARANS that can work well with point objects as well as polygon objects. After that Echidna [17] algorithm is developed that is used for dealing with attributes of mixed type such as categorical, hierarchical and numerical type.

3. PROBLEM DEFINATION

With all the advancement in storage, digital sensors, communications and computation have got made huge collections of data. Because of considerable development and progress of the communication and computation technologies such as big and dominant data processing services, there is huge level of information and data production at very high rate from various servers and resources. These giant amounts of data come from offered variety of services and resources all over the world which works for serving their end users.

Various operations such as logical calculations, retrieval and process operations, which are very tricky and extremely time overwhelming, are performed on such massive level data. Also large storages are needed for such data. In order to cope with these troubles there is need to cluster big data in a compressed form. The well known data mining technique for that purpose is nothing but the clustering which divides data objects into groups or classes because of which the data points in one group are analogous to each other. Currently, various clustering techniques are proposed but not a single among all is appropriate.

4. PROPOSED SYSTEM

Following block diagram is the general architecture of the proposed system. First block is the data source layer which represents the training dataset. The next layer i.e. Pre-processing layer involves mainly two phases. In the first phase dataset is converted into XML patterns and the next phase is operational storage. Then there is intelligence layer in which basic input parameters for DBSCAN, Eps (Radius) and Min points and for CURE algorithm parameter K (Number of clusters) is set. Then finally there is processing layer in that clustering is performed using DBSCAN and CURE algorithms. In this last phase converted XML dataset is given as input data for testing and finally clusters are formed.

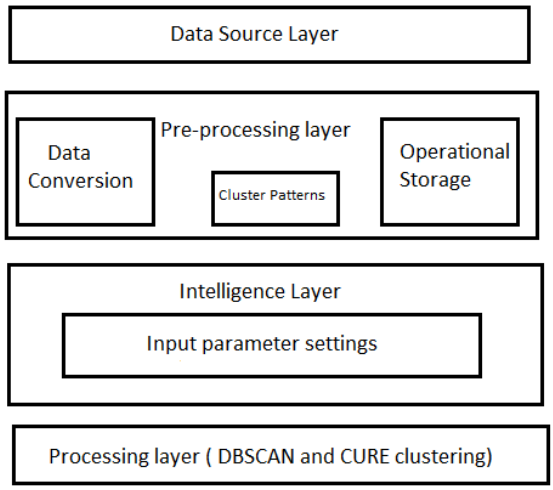

Fig. 1 System Overview (DBSCAN and CURE clustering)

5. ALGORITHMS

5.1 Density Based Spatial Clustering of Applications with Noise (DBSCAN) Algorithm

DBSCAN identifies clusters by analyzing the density of points and uses single parameter as input. Regions having more density of data points represent the existence of clusters while less dense regions of points specify noisy data. Density is like amount of points within a predefined radius (Eps). Basically cluster belongs to two types of points, border points and core points or centroids. If a point has greater than already defined number of points (Min Pts) in Eps neighborhood then it is called as core point and it lies internal to the cluster. And if the point is having less points than

predefined Min Pts inside Eps neighborhood but it is inside the neighborhood area of a centroid or centre point then it is called as border point. While noisy point is any point which is neither boundary point nor centroid. In the whole procedure user obtains an idea on, which parameter value that would be appropriate. And so there is minimal requirement of domain knowledge. There is need of three basic parameters as input for DBSCAN algorithm

1. K – the size of neighbor list
2. Eps - radius that restricts the neighborhood region of a point
3. Min Pts - least number of points that have to be present within Eps neighborhood.

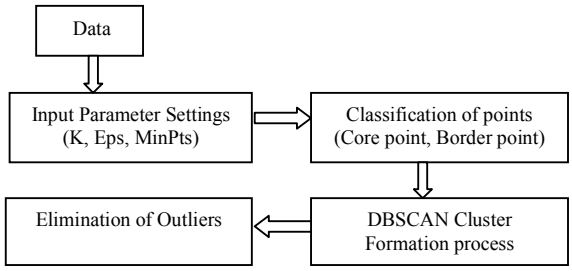

Fig. 2 Overview of DBSCAN algorithm

Procedure of DBSCAN clustering is basically nothing but the assignment of the data points including center points, boundary points and noisy data points. DBSCAN works according to the application of density associations among points (exactly density-attainable, density-connected and density-reachable) in order to perform clustering.

An algorithm starts with a random point P. After that it find outs all points that are density-attainable from P that follows already defined Eps and Min Pts conditions. If P is a centroid point, DBSCAN process gives cluster according to Eps and Min Pts and if P is a boundary point, there are no points that are density-attainable from point P then algorithm move towards next point in the database. Algorithm DBSCAN uses universal values for parameters Eps and Min Pts, and that's why it merges two clusters within single cluster, if two clusters of diverse density are near to each other. DBSCAN is basically developed to cope with datasets having large number of records, along with noise, and is able to find clusters having non-uniform shapes. Yet, clusters that are near to each other have a tendency to fit in the similar class.

5.2 Hierarchical Clustering Using Representative (CURE) Algorithm

CURE algorithm starts with, a steady figure of well scattered points (c) in the cluster. These spread points take the size and form of the cluster. Selected spread points are then reduced along the direction of the centroid of the cluster by portion α. After the process of reducing, these spread points are then taken as indicatives of the cluster. At each stage of CURE algorithm, clusters having the nearest set of two indicative points are combined. CURE is less perceptive to noisy data since noisy data are normally at a long distance from the mean point and because of which they are transferred at a longer distance for reducing purpose. Because of numerous spread CURE is able to find out non-spherical clusters.

For handling large data samples, an efficient method is necessary in order to sink the scale of the parameter given as input to CURE's clustering algorithm. For that purpose instead of pre-clustering total data points, an algorithm starts with representing an arbitrary sample taken from the database. For increasing the speed of clustering process, algorithm first makes clusters of the arbitrary sample and then partly classifies the data items in every cluster. In the last stage, after elimination of noisy points, pre-clustered data point in every cluster are again clustered in order to produce final clusters. Random sampling improves the worth of clustering. As the clustering of the arbitrary sample is finished, as a replacement for a individual centre point, numerous indicative points from every cluster are utilized for labelling of the residual data sample.

6. IMPLEMENTATION DETAILS

The proposed system will perform as a clustering mechanism for big data, which will use DBSCAN and CURE algorithm. The input to the system is the weather forecasting dataset. The whole work is implemented on 1.90 GHz, Intel Core 3 PC with 4 GB RAM with Windows Operating system using JAVA JDK 1.6, My SQL, Tomcat Apache server.

We are using "Weather Forecasting" dataset as a training data for the proposed work. This dataset includes various categories such as- Temp, Wind, Cloud, Max and Min Rain, Humidity, Snow etc. etc.

7. RESULTS

In the previous DBSCAN algorithm there is need to find out distance between every pair of data points. But, in proposed DBSCAN algorithm with boosting, suppose there are 3 points. Let us consider as A, B and C are three points and if we know the distance between points (A , B) and distance between points (B, C) then we directly get distance between points (A, C) and so there is no need to find out distance between (A, B) and (B, C) again. This facility results in low computation cost and so it also adds extension to the existing DBSCAN algorithm. Accuracy is measured in terms of computation Cost i.e. time requirement. And study shows that in both cases DBSCAN algorithm performs better than CURE algorithm.

Table 1 Computation Cost (Time in milliseconds): Existing DBSCAN Algorithm, Existing CURE Algorithm and Proposed DBSCAN Algorithm

XML Dataset Pattern	Existing DBSCAN Algorithm	Existing CURE Algorithm	Proposed DBSCAN Algorithm
XML pattern 1	35,298 ms	39,397 ms	31,874 ms
XML pattern 2	58,041 ms	80,796 ms	54,346 ms
XML pattern 3	55,631 ms	1,05,674 ms	49,296 ms
XML pattern 4	58,667 ms	1,32,429 ms	54,308 ms
XML pattern 5	53,238 ms	1,57,231 ms	51,454 ms
XML pattern 6	53,786 ms	1,59,374 ms	52,436 ms

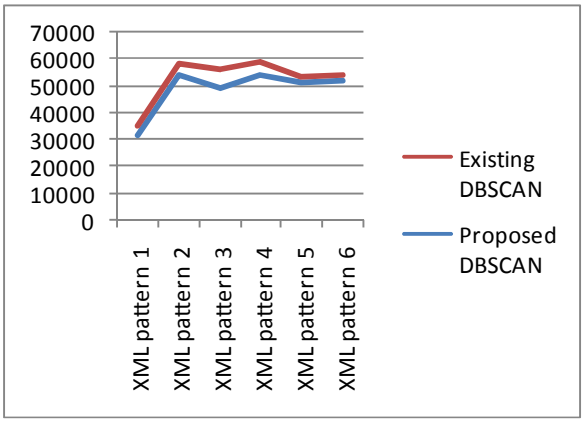

Fig. 3 Proposed DBSCAN algorithm Vs existing DBSCAN algorithm

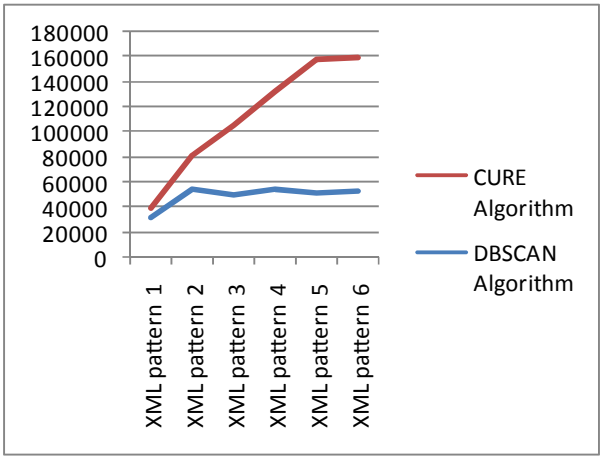

Fig.4 Proposed DBSCAN algorithm Vs Existing CURE algorithm

When DBSCAN and CURE algorithm are used for mining specific data or special data records during particular time period with particular features such as- Minimum/Maximum Humidity, Minimum/Maximum Temperature from the given dataset, results shows that the mining performed by using DBSCAN takes less time than mining performed using CURE algorithm.

Fig.5 Mining Analysis: DBSCAN Algorithm with CURE Algorithm

7.1 Mathematical Model

Let S be the system such that,

S= {s, e, I, O, A, E}

Where,

s- start state

e- end state

I- Input

O- Output

A- Algorithm

E- Entity

1) For DBSCAN clustering,

   I= {K, Eps, Min Pts}

   Where,

   K- size of neighbor list

   Eps- radius that restricts the neighborhood region of a point

   Min Pts- least number of points that have to be present within Eps neighborhood

2) For CURE clustering,

   I= {c, α }

   Where,

   c – Number of well scattered points

   α – shrinking parameter

O= {O_1, O_2}

Where,

O_1 – Output of DBSCAN clustering

O_2 – Output of CURE clustering

A= {A_1, A_2}

Where,

A_1 – DBSCAN Algorithm

A_2 – CURE Algorithm

E= {SU}

Where,

SU – System User

7.2 Validation criteria

Validation criteria refer to the procedure that evaluates the results of cluster analysis in a quantitative fashion. Many times clusters are justified by ad hoc methods. We can validate the cluster structure by comparing it with an a priori structure. The validity of cluster structure can be given in terms of three types of criteria

1) External criteria- It measures the performance of cluster structure by matching it with a priori information.

2) Internal criteria- It estimates the fix between the cluster structure and the data by using data only.

3) Relative criteria- Which of the two structures is better in terms of stability and which one is more appropriate for data is decided by relative criteria.

A criterion gives the strategy by which a cluster structure is to be validated. Stability of the results is considered as valuable thing for validation of an algorithm. While an index is statistic in terms of which the validity is to be tested. Following evaluation criteria are also used for cluster structure analysis.

1) Compactness (CP) is used to validate clusters by measuring the average value of distance among each pair of data items.

2) Separation (SP) expresses the level of separation among personal clusters

3) Davies- Bound in Index(DB) discover overlapping of cluster by calculating the fraction of the total of intra-cluster scatters to the inter-cluster separations

4) Dunn Validity Index (DVI) measures separation over compactness

5) Cluster Accuracy (CA) gives the percentage of accurately clustered data points in the clustering result as compared to

6) Adjusted Rand index (ARI) value lies between 0 - 1, larger value shows that all data points are correctly classified

7) Normalized Mutual Information (NMI) approximate the superiority of the clustering as compare to predefined labels of class.

8. ACKNOWLEDGMENT

I express sincere thanks to my project guide Prof. S. S. Banait for his contributions in this research work. The support from the K. K.Wagh Institute of Engineering, Education and research is gratefully acknowledged.

9. CONCLUSION

Clustering bulky databases is extremely tricky job and needs high computational cost. To use clustering technique is one way for that intention, and there are lots of techniques for clustering. On the other hand, they all have some troubles such as selection of the accurate input parameters, localizing non uniform clusters and the computational cost, efficiency on huge databases. No clustering technique provide key to these problems. Since data mining basically works with very bulky data sets, and firmness is the basic necessity for data mining techniques. Algorithm CURE is more accurate because it aadjusts properly to the geometry of non-spherical shapes. Also it is efficient with having space complexity O(n) and time complexity $O(n^2)$when data points are low dimensional and otherwise it is $O(n^2 \log n)$.

The DBSCAN and CURE algorithm gives solution to all these troubles as they are also capable of finding even non uniform shaped cluster very promptly. Outliers or noise data are also detected by DBSCAN and CURE algorithms. Existing DBSCAN executes directly on entire database but proposed DBSCAN algorithm performs clustering by executing on single data pattern. The proposed DBSCAN with boosting algorithm works well than the existing DBSCAN algorithm. Among both DBSCAN and CURE algorithm, DBSCAN outperforms than CURE and also DBSCAN takes less time than CURE algorithm to perform mining of weather dataset.

Future enhancements can be done by

1) Extending the proposed DBSCAN and CURE algorithm for Big data.

2) Extending the proposed system by making use of Distributed system such as Hadoop platform to improve the performance.

10. REFERENCES

[1] J. MacQueen, ‘‘Some methods for classification and analysis of multivariate observations,’’ in *Proc. 5th Berkeley Symp. Math. Statist. Probab*., Berkeley, CA, USA, 1967, pp. 281–297.

[2] A. P. Dempster; N. M. Laird; D. B. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society*. Series B (Methodological), Vol. 39, No. 1. (1977), pp. 1-38

[3] J. C.Bezdek, R.Ehrlich, and W.Full,‘‘FCM: Thefuzzy c-means clustering algorithm,’’ Comput. Geosci., vol. 10, nos. 2–3, pp. 191–203, 1984.

[4] D. H. Fisher, ‘‘Knowledge acquisition via incremental conceptual clustering,’’ Mach. Learn., vol. 2, no. 2, pp. 139–172, Sep. 1987.

[5] A.K.Jain and R.C.Dubes, Algorithms for Clustering Data.Upper Saddle River, NJ, USA: Prentice-Hall, 1988.

[6] R.T.Ng andJ.Han, ‘‘Efficient and effective clustering methods for spatial data mining,’’ in Proc. Int. Conf. Very Large Data Bases (VLDB), 1994, pp. 144–155

[7] T. Zhang, R. Ramakrishnan, and M. Livny, ‘‘BIRCH: An efficient data clustering method for very large databases,’’in Proc. *ACMSIGMOD Rec*., Jun. 1996, vol. 25, no. 2, pp. 103–114

[8] Ester M., Kriegel H.-P., Sander J., Xu X.: *"*A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", *Proc. 2cnd Int. Conf. On Knowledge Discovery and Data Mining*, Portland, Oregon, 1996, AAAI Press, 1996.

[9] Z.Huang,‘‘A fast clustering algorithm to cluster very large categorical datasets in data mining,’’in Proc. SIGMOD Workshop Res.Issues Data Mining Knowl. Discovery, 1997, pp. 1–8.

[10] X. Xu, M. Ester, H.-P. Kriegel, and J. Sander, ‘‘A distribution-based clustering algorithm fo rmining in large spatial ldatabases,’’inProc.14thIEEE Int. Conf. Data Eng. (ICDE), Feb. 1998, pp. 324–331.

[11] S. Guha, R. Rastogi, and K. Shim, ‘‘CURE: An efficient clustering algorithm For large databases,’’in *Proc. CMSIGMOD Rec*., Jun.1998,vol.27,no.2, pp. 73–84

[12] G. Sheikholeslami, S. Chatterjee, and A. Zhang, ‘‘Wavecluster: A multi resolution clustering approach for very large spatial databases,’’ in *Proc. Int. Conf. Very Large Data Bases (VLDB),* 1998, pp. 428–439.

[13] A. Hinneburg and D. A. Keim, ‘‘Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering,’’ in *Proc. 25th Int. Conf. Very Large Data Bases (VLDB),* 1999, pp. 506–517.

[14]G.Karypis,E.H.Han,andV.Kumar,‘‘Chameleon:Hierarchic alclustering using dynamic modelling,’’ *IEEE Comput*., vol. 32, no. 8, pp. 68–75, Aug. 1999.

[15] S. Guha, R. Rastogi, and K. Shim, ‘‘Rock: A robust clustering algorithm for categorical attributes,’’ *Inform.Syst.,* vol.25,no.5,pp.345–366,2000.

[16] R. T. Ng and J. Han, ‘‘CLARANS: A method for clustering objects for spatial data mining,’’*IEEE Trans. Knowl. Data Eng.(TKDE),*vol.14,no.5, pp. 1003–1016, Sep./Oct. 2002.

[17] A. N. Mahmood, C. Leckie, and P. Udaya, ‘‘ECHIDNA: Efficient clustering of hierarchical data for network traffic analysis,’’ in Proc. 5th *Int. IFIP-TC6 Conf. Netw. Technol., Services, Protocols Perform. Comput. Commun. Netw. Mobile Wireless Commun. Syst. (NETWORKING),* 2006, pp. 1092–1098.

[18] A. Fahad, N. alshatri, Z. Tari, A. Alamri, I. Khalil, A. Y. Zomaya, S. Foufou, and A. Bouras, "A survey of clustering algorithms for big data: taxonomy and empirical analysis", *IEEE Transactions on emerging topics in computing,* vol 2, no. 3, Sept 2014.