Dimensionality Reduction in Clustering Optimality for Next Generation Networks

A V V Satyanarayana Rao[1] **Dr. Jhansi Rani Singothu[2]**

1. M.Tech Scholar, Dept. of CS & SE, Andhra University College of Engineering(A), Visakhapatnam, India,
2. Assistant Professor, Dept. of CS & SE, Andhra University College of Engineering(A), Visakhapatnam, India.

Abstract :

Next-generation networks seem to be full of challenges. Network Management is one among them which plays a key role in making the network functional in terms of performance. It is the aim of operators or service providers to facilitate end-to-end monitoring and network optimization. With the modern machine learning techniques the networks by themselves incorporate self-healing (SH) frame work to provide a fully functional and optimized network management system. A SH framework for next-generation networks using dimensionality reduction in the means of Optimal clustering is proposed in this paper taking advantage of feature selection and dimensionality reduction techniques. The proposed framework performs the self- diagnosis and indicates the optimality factor.

Keywords : Feature selection, Next generation networks, Dimensionality Reduction, Self Healing.

Introduction:

Besides the speed and other positive aspects, Future cellular networks shall be quite complex. Hence the overall idea of networks aims at full automation and optimized Network management. The operators must design their Network architecture so that both operational expenditure (OPEX) and capital expenditure (CAPEX) shall be mitigated. Self Healing(SH) becomes the focus of future networks. Several network performance indicators decide the design of network architecture. There are several network performance indicators which range from alarms and event counters to more complex metrics. In the future networks predicting network failures and adopting counter measures will certainly be a tedious task and human biasing in such context causes a serious impact. Hence there is undoubted need of the tasks to be automatically done. This automatic diagnosis of fault cause and Self Healing in the network management may unnecessarily include several indicators. Hence to self heal the network from faults; it is the prime task to select relevant subset of performance indicators leading to minimized Diagnosis Error Rate. In this paper we presented a procedure for indicating the need of dimensionality reduction. We classified the transactions at a network base station which are based upon various parameters. Thereafter we used a dimensionality reduction technique and further clustered the same set of transactions. In both the cases i.e, before and after dimensionality reduction the clustering seems to be same which accomplishes our objective. The motivation behind our work is the concept which is based on [1]. Although in the reference paper, feature selection and feature extraction were applied we had done our work in the means of feature selection basing on Laplacian score. Besides we used Silhouette score to demonstrate our idea of our research work.

Related work:

In [1] David Palacios et.al proposed a self-healing framework which used both feature selection and feature extraction techniques. Dimensionality reduction also was done which enables reduction of network storage needs, as well as the eventual complexity of the self-healing mechanisms. Two types of tests were conducted, one on high dimensional data and the other on medium level dimensional data. In their work they had taken 359 sample dataset with 286 performance indicators and seven different situations were assessed so as to test different schemes for dimensionality reduction. Finally outliers were identified and they found that the OSS storage management costs could be reduced by 93%. They used LDA classification technique as well the dimensionality reduction is done using Principal Component Analysis.

In [2] Xiaofei He et.al used an unsupervised algorithm for Feature selection in which Laplacian score uses K-means clustering algorithm to select the top k features. It is like a filtering method which is based on the observation that, in many real world classification problems, and the importance of a feature is evaluated by its power of locality preserving, or, **Laplacian Score**. Laplacian Score (LS) is fundamentally based on Laplacian Eigen maps [3] and Locality Preserving Projection [4]. The basic idea of LS is to evaluate the features according to their locality preserving power. The proposed algorithm LS is compared with data variance for clustering where LS outperformed the latter algorithm in terms of Accuracy and Clustering.

Key performance indicators play a major role in assessing the performance of Cellular Networks. Also there exist several implicit associations between one another. This was addressed by Xingyu Guo et.al in [5]. They proposed an approach to figure out the relation between the indicators. The approach could well divide the data into clusters and the relationship between indicators in each cluster is effective. They worked out on various indicators like call drop rate, connection set up rate, density and other ones.

Darijo Raca et.al in [6] presented a production data set which contains Cellular Key performance Indicators collected from two different operators across different mobility patterns (Static, pedestrian, car, bus and train). The 4G trace dataset contains 135 traces, with an average duration of fifteen minutes per trace collected using G-NetTrack Pro. Also they supplemented with synthetic data set generated from simulations.

Proposed work:

In this paper our idea is to illustrate that, all the features may not be necessary to measure the performance of the network. We considered optimal clustering to exemplify our idea.

To demonstrate, we considered three data sets, one from kaggle and the other two from [6] and applied our proposed work. We had done the feature selection on the datasets using LS unsupervised algorithm. The Laplacian score is calculated for all features (performance indicators) in the data sets and applied K-means clustering algorithm. Thereafter the number of clusters is optimized. Average Silhouette score is calculated.

There are various dimensionality reduction techniques out of which we used backward feature elimination method and removed two features in each dataset. The above same process is followed for the pruned datasets and obtained the Silhouette score. It is observed that in both the cases viz. before and after dimensionality reduction the scores seems to be similar which affirms our idea of proposed work.

The architecture of proposed work could be seen in Fig. 1, followed by the algorithm. The indicators of the three datasets are mentioned in the table 1.

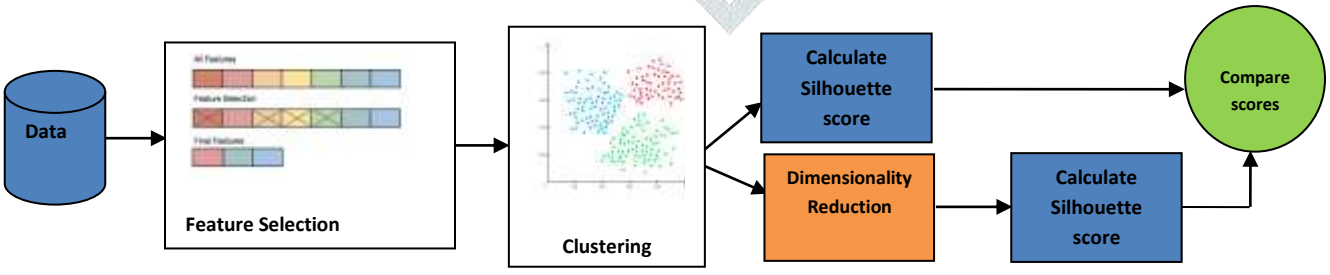

Figure 1: Architecture of our proposed system

Algorithm

1. Read the dataset
2. for all features
   Apply Laplacian score
3. Calculate the feature ranking
4. Input the number of clusters
5. Input the number of features
6. Apply K-Means clustering
7. Calculate Silhouette score
8. Reduce the features having low feature rank
9. Repeat steps 4 to 7
10. Output the scores

Dataset from Kaggle	Production dataset[6]	Supplement dataset[6]
Time	Timestamp	Time
CellName	Longitude	CCQI
PRBUsageUL	Latitude	NDI
PRBUsageDL	Speed	CQI
meanThr_DL	Operatorname	RSRP
meanThr_UL	CellID	CSINR
maxThr_DL	NetworkMode	TBSize
maxThr_UL	RSRP	PDCP Throughput
meanUE_DL	RSRQ	CTHR
meanUE_UL	SNR	SINR
maxUE_DL	CQI	THR
maxUE_UL	RSSI	Delay
maxUE_UL+DL	DL_bitrate	CellID_RSRP
	UL_bitrate	TBLER
	State	
	NRxRSRP	
	NRxRSRQ	
	ServingCell_Lon	
	ServingCell_Lat	
	ServingCell_Distance	

Table 1: Indicators in the datasets

Discussion of the algorithm:

1. The three datasets will be read one by one and the remaining process is continued for the respective dataset.
2. Laplacian score is calculated for each feature of the dataset using the below formula.

$$L_r = \frac{\tilde{\mathbf{f}}_r^T L \tilde{\mathbf{f}}_r}{\tilde{\mathbf{f}}_r^T D \tilde{\mathbf{f}}_r} \quad \text{Where} \quad \tilde{\mathbf{f}}_r = \mathbf{f}_r - \frac{\mathbf{f}_r^T D \mathbf{1}}{\mathbf{1}^T D \mathbf{1}} \mathbf{1}$$

and

$$\mathbf{f}_r = [f_{r1}, f_{r2}, \cdots, f_{rm}]^T, D = diag(S\mathbf{1}), \mathbf{1} = [1, \cdots, 1]^T, L = D - S$$

'S' is the Weight matrix of the graph model.

3. Depending on the score feature ranking will be obtained.

4 to 6: Input the number of clusters and features and apply K-Means clustering algorithm.

7. Average Silhouette score of all the clusters is calculated which represents the effective cluster division.

8 & 9: Dimensionality reduction is applied. The technique used here is "Backward Feature Elimination". Few features having low feature rank will be eliminated one by one. The obtained feature set will undergo steps 4 to 7 which gives the silhouette score.

10. The scores with respect to before and after elimination will be outputted.

The Laplacian scores of the data sets are shown in figures 2 to 4.

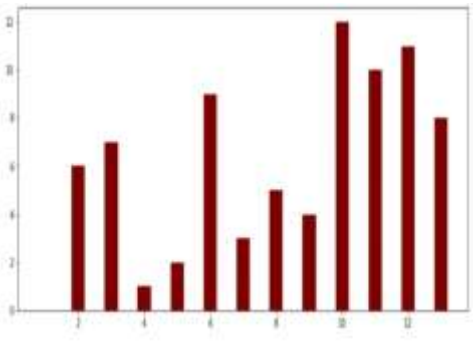

Fig 2: L scores for Kaggle dataset features

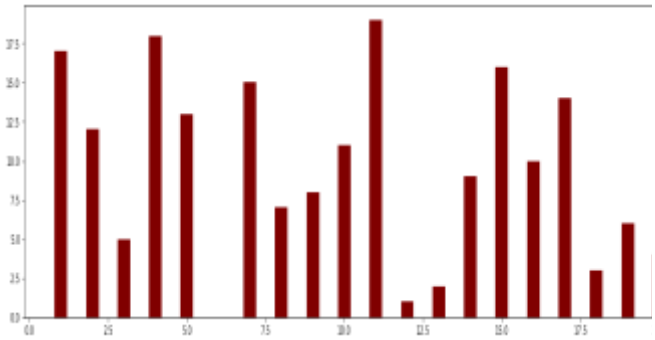

Fig 3: L scores for Production dataset features

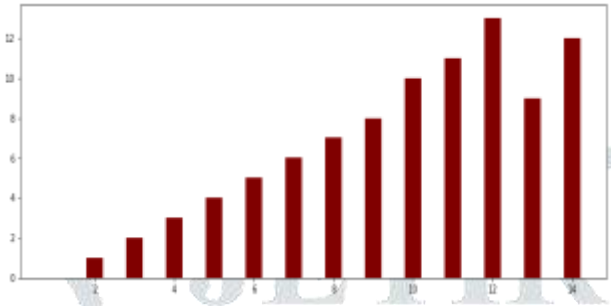

Fig 4: L scores for Synthetic dataset features

Experimental Results:

As discussed in [7], the entire clustering is displayed by combining the silhouettes into a single plot, allowing an appreciation of the relative quality of the clusters and an overview of the data configuration. The average silhouette width provides an evaluation of clustering validity, and might be used to select an 'appropriate' number of clusters.

The score for one silhouette could be calculated as follows and thereafter average score is calculated.

$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$, where a(i) = average dissimilarity of i to all other objects of A.

b(i) = minimum d(i, C) and
C≠A

d(i, C) = average dissimilarity of i to all objects of C.

The average Silhouette scores for the three datasets and the respective graphs are depicted in table 2 and figures 5 to 6 respectively. It is observed that the Silhouette scores before and after dimensionality reduction seems to be almost similar.

Dataset considered	Before Dimensionality reduction		After Dimensionality reduction	
	No. of Features	Score	No. of Features	Score
Kaggle Dataset	13	0.375	8	0.378
Production Dataset [6]	20	0.66	18	0.66
Synthetic Dataset [6]	14	0.5	8	0.5

Table 2 : Silhouette score for various datasets

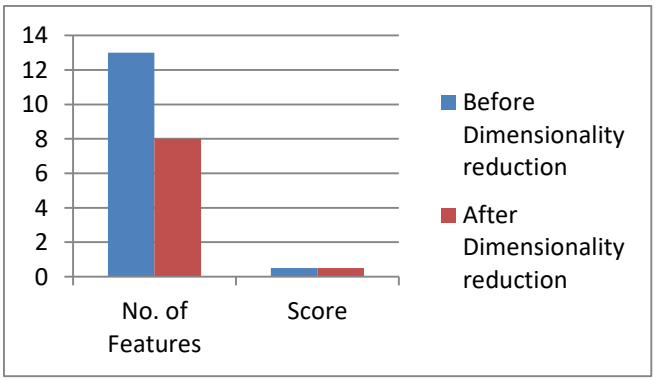

Figure 5 : Kaggle dataset

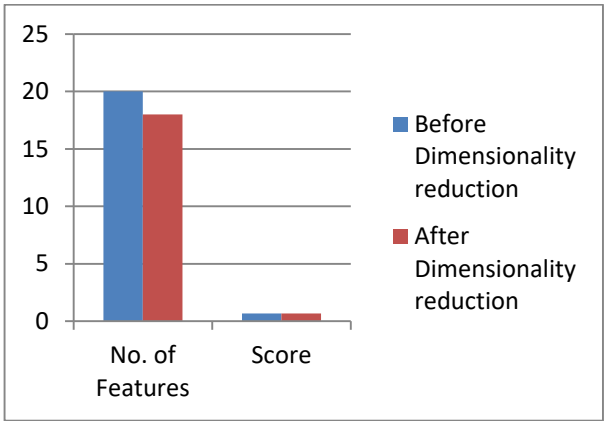

Figure 6 : Production dataset[6]

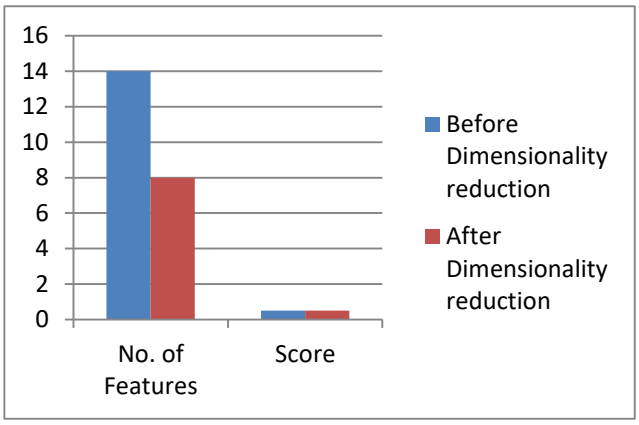

Figure 7 : Synthetic dataset [6]

Conclusion:

As indicated in our proposed work, we used Feature selection and Dimensionality reduction to address the need of the opted techniques. We applied the feature selection technique on three datasets and identified the features. Thereafter we applied the dimensionality reduction technique and clustered the transactions. In order to find the optimality of clustering we calculated the average score before and after applying dimensionality reduction. The results seem to be same in both the cases which signal the need of dimensionality reduction in self healing framework of next generation networks. However the technique to find the threshold number of features for achieving optimal clustering is not addressed which may be considered as future work.

References:

[1] David Palacios, Sergio Fortes, Isabel de-la-Bandera, and Raquel Barco, "Self-Healing Framework for Next-Generation Networks through Dimensionality Reduction," *IEEE Commun. Mag*., July 2018, pp. 170–176.
[2] Xiaofei He, Deng Cai and Partha Niyogi, "Laplacian Score for Feature Selection," Advan ces in Neural Information Processing Systems 2006
[3] M. Belkin and P. Niyogi, "Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering," Advances in Neural Information Processing Systems, Vol. 14, 2001.
[4] X. He and P. Niyogi, "Locality Preserving Projections," Advances in Neural Information Processing Systems, Vol. 16, 2003.
[5] Xingyu Guo, Peng Yu, Wenjing Li and Xuesong Qiu, "Clustering-based KPI Data Association Analysis Method in Cellular Networks," *IEEE/IFIP NOMS 2016 Workshop: International Workshop on Analytics for Network and Service Management (AnNet 2016)*, pp. 1101–1104.
[6] DarijoRaca, Jason J. Quinlan, Ahmed H.Zahran and Cormac J. Sreenan, "Beyond Throughput: a 4G LTE Dataset with Channel and Context Metrics," *MMSys'18: 9th ACM Multimedia Systems Conference, June 12–15, 2018, Amsterdam, Netherlands. ACM, New York, NY, USA, pp 460-465.*
[7] Peter J. ROUSSEEUW, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics 20 (1987)*, Elsevier Science Publishers, pp. 53–65.