Optimization and scalable constrained clustering performances

R. Nagaraj[1], Dr.V. Thiagarasu[2], B. Jeevithapriya[3]

[1]*Research Scholar, Department of Computer Science, Karpagam University, Coimbatore, India*

[2]*Associate Professor, Department of Computer Science, Gobi Arts and Science College, Gobichettipalyam, India*

[3] *Research Scholar, Department of Computer Science, kamadhenu College of Arts and Science, Sathyamangalam, India*

nagarajresearch123@gmail.com[1], thiagarasuresearch123@gmail.com[2], jeevithapriyamphil123@gmail.com[3].

Abstract

Objectives: To achieve the accuracy of clustering performances higher and to optimize the scalable approaches.
Methods: Constrained spectral clustering and optimization algorithms are used to analyze and evaluate the large dataset. It is used to produce quality of clustering results.
Findings: The proposed method achieves high performance in terms of precision, recall and accuracy.
Application/Improvements: The proposed system is done by using optimization algorithm and pairwise constraints concepts. The optimization algorithm is used to increase the clustering accuracy and produce more optimal performances.

Keyword: Constrained spectral clustering, scalability and optimization

1. Introduction

Data mining is the process of extracting the important information from the large repository. Nowadays, data in an extensive diversity of fields tend to huge scales. In several conventional learning based data mining techniques, this is major issue to effectively extract information from the speedy rising data. It is such as information systems, images and even videos. To avoid the above mentioned issue, it is prominent to progress scalable learning approaches. Clustering is used to collect similar kinds of unlabeled data in one group and dissimilar data into another group separately. Similarity between data is calculated through, for instance Euclidean distance. Clustering is used to optimize an objective function repeatedly to create the clusters within a dataset [1].

There are several types of clustering techniques recommended in the data mining applications. The clustering is categorized into two major types such as hierarchical clustering and partitional clustering. Constrained clustering is a significant field in the machine learning applications to handle the scalable approaches [2] [3]. New algorithms are focused on the improvements of clustering accuracy in terms of encoding side information into unsupervised algorithms. Kiri Wagstaff and Claire Cardie [4] introduced instance level clustering constraints for grouping identical example collectively and maintain different example separately. This is used to discover the comparative effects of every type of constraint and also determine that the type to supply for improving the clustering algorithm precision values without constraints.

In [5], the constrained k-means clustering algorithm is suggested to progress the clustering similarity as well as clustering performance in higher. This method is generally used to partition the data set into k groups and selects only the best constraints. However it has issue with time complexity in this scenario. In [6], the distance metric knowledge approaches are recommended for increasing clustering accuracy significantly. The distance metrics are calculated based on the relationships among instances. The metrics are computed until good clusters are discovered. However it has problem with huge dimensional dataset.

In [7] the method suggested named as hidden markov casual fields which produce a principled structure to include management into model based clustering. The prototype creates and joints constraints as well as Euclidean distance model then permits the usage of a wide range of cluster measurements. However still it has issue with providing distance more accurately in few cases. In [8] the author discussed Gaussian mixture model along with constraints concepts. The equivalence constraints are described on pairs of data points, identifying the points occur from positive constraints which means similar source or from negative constraints which implies dissimilar source.

In [9] semi supervised graph clustering methods are introduced to develop the pairwise constraints along with kernel approach. This algorithm is motivated to enhance the quality of clustering results by using restricted supervision. The kernel approach is used to allow clusters with non linear restrictions in the input data space. However it has issue with inaccurate results in some cases. In [10] the methods and algorithms are recommended to produce optimized results in the clustering scenario. The algorithm is named as semi supervised algorithm which is used for handling both the labeled data and unlabeled data efficiently. The technique is call as maximum margin clustering method which is used for extending the maximum margin structure in the supervised approach. This scenario is also introduced pairwise constraints which are motivated to progress the performance of maximum margin algorithm. However it has issue with selection of numerous clusters is a critical problem.

In [11] the technique is introduced named as spectral clustering approach along with pairwise constraints. This scenario is used to indicate the two objects are belonging to similar cluster or not. Unlike preceding techniques that change the similarity matrix along with pairwise constraints. We can adjust the spectral clustering towards a model embedding as reliable along with the pairwise constrictions probable. This approach directs to a small semi specific program whose complexity is autonomous of the number of pairwise constraints, creating it scalable to huge scale problems. Hence it is suitable for multi class problems and handle with must link as well as cannot link constraints to propagate pairwise constraints more effectively. However it has issue with large scale problems still in a few cases.

In [12] the author suggested huge scale spectral clustering for many popular clustering applications. In preceding research, many of the algorithms are failed to provide the solutions for time complexity issues. In this scenario, landmark based spectral clustering is introduced which is focused on the effectiveness and efficiency of clustering. In particular, we have to choose the illustration points as the landmark and illustrate the real data points as linear combinations of these landmarks.

2. Materials and Methods

2.1. Constrained normalized cuts

Consider a vector dataset X=$\{x_i\}_{i=1}^n$ where $x_i \in R^d$ and a constraint set $\{C_=, C_{\neq}\}$ where $(\mathrm{x_i}, \mathrm{x_j}) \in C_=$ if the patterns of $\mathrm{x_i}$ and $\mathrm{x_j}$, are similar and $(\mathrm{x_i}, \mathrm{x_j}) \in C_{\neq}$) otherwise, the aim of the partition X into k clusters biased by the constraint set. Let W be the similarity matrix over X where W_{ij} represents the similarity between instances $\mathrm{x_i}$ and $\mathrm{x_j}$. Let D be the degree matrix over X which is a diagonal matrix with elements D_{ij}=$\sum_j W_{ij}$. Let $\bar{L} = I - D^{-1/2}WD^{-1/2}$ be the normalized graph Laplacian where I denotes the identity matrix. Let Q denote the constraint matrix where Q_{ij}=1 expresses $(\mathrm{x_i}, \mathrm{x_j}) \in C_=$ and Q_{ij}=-1 expresses$(\mathrm{x_i}, \mathrm{x_j}) \in C_{\neq}$) and Q_{ij}=0 expresses no available side information. Let $\bar{Q} = D^{-1/2}QD^{-1/2}$ be the normalized constraint matrix. The constrained normalized cuts can be rewritten as

$$argmin_{v\in\left\{+\frac{1}{\sqrt{n}},-\frac{1}{\sqrt{n}}\right\}n} V^T\overline{L_V} \qquad (1)$$

Here the parameter α controls what degree the input side information is respected. The problem is NP hard and the feasible way is to allow v to take any real values. It requires memory cost as well as time cost in this scenario.

2.2. Sparse coding based graph construction

Graph construction amounts to computing a similarity matrix. In this scenario, the sparse coding based graph construction. For the specified dataset X, of the form d by n matrix, X, sparse coding aims to find a pair of matrices, U$\in R^{d\times p}$ and $Z \in R^{p\times n}$ such that UZ could best approximate X where U's columns represent the desired base vectors and Z's columns represent sparse coefficient vectors has few non zero components. The cost function to be minimized is

$$f(U, Z) = ||X - UZ||_F^2 \qquad (2)$$

Unfortunately, it is costly to precisely solve for U^* and Z^*. To estimate each column vector of Z according to

$$Z_{ij} = \frac{K_\sigma(x_j,u_i)}{\sum_{i\epsilon rNB(j)} K_\sigma(x_j,u_i)} \qquad (3)$$

Where σ denotes the bandwidth of Gaussian kernel $K_\sigma(.,.)$ and $i \epsilon r NB(j)$ means that the base vector u_i is among the r nearest base (rNB) vector of instances x_j. The sparseness of the matrix Z is controlled by the parameter r. After obtained Z, one can construct two forms of graph matrices. G = $Z^T Z$ and S=ZZ^T. Choose $\hat{Z} = D^{-1/2} Z \; where \; D_{ii} = \sum_j Z_{ij}$. We can obtain the normalized graph matrices and it is easy to check that the normalized graph Laplacian over X is (1-$\hat{G}$).

2.3. Scalable constrained normalized cuts

In this scenario, we consider $\hat{G}$ be the similarity matrix over X. The relaxed constrained normalized cuts problem can be formulated as

$$min_{v \in R^n} v^T L v, such \; that \; v^T Q v \geq \alpha, \qquad v^T v = 1, \quad v \perp 1 \; (4)$$

Where L=1-$\hat{G}$ is the normalized graph Laplacian
Generalized eigen value is represented as

$$Lv = \lambda (Q - \beta I) v \qquad (5)$$

Where β is a lower bound of α. The time cost for solving this problem is O(n^3), infeasible for handling the large datasets. We can refine the above mentioned problem as following method.
Now we consider

$$min_{u \in R^p} u^T A u, s.t. u^T \hat{Q} u \geq \alpha, u^T \hat{S} u = 1, 1^T \hat{S} u = 0 \qquad (6)$$

Here,

$$A = \hat{S} - \hat{S}\hat{S}, \hat{Q} = \hat{S} Q \widehat{Z^T} \qquad (7)$$

It is used to efficiently recover the solution by using above mentioned procedure.

2.4. Scalable constrained spectral clustering

1. Consider the input as dataset X$\epsilon R^{d \times n}$, the base vector number p, the n-by-n constraint matrix Q, β and cluster number k

2. Select p vector data among the input dataset at random, and stack them in the columns of matrix U$\in R^{d \times p}$.

3. Compute Z$\in R^{d \times n}$ using equation 3 and then compute

4. Compute $\hat{S} = \hat{Z}\hat{Z}^T$ and $\hat{Q} = \hat{Z} Q \hat{Z}^T$

5. Find the largest eigen value $\gamma_{max}$of the generalized eigen system $\hat{Q}x = \gamma \hat{S} x$.

6. If $\beta \geq \gamma_{max}, return \; \{v^*\} = \emptyset$ otherwise find all the eigen vectors $\{u_i\}$ by solving generalized eigen system.

7. Find among $\{u_i\}$ the eigen vectors $\{u_i\}^+$ associated with positive eigen values

8. Normalize each $u_i \in \{u_i\}^+$ by multiplying a factor $\sqrt{\frac{1}{u_i^T S u_i}}$.

9. Remove the eigen vectors from $\{u_i\}^+$ that are not orthogonal to the vector $1^T \hat{S}$

10. Find among $\{u_i\}^+$ the m eigen vectors that lead to the smallest values of $u_i^T A u_i$

11. Compute $V^{(r)} = \hat{Z}^T V(I - V^T AV)$

12. Normalize $V^{(r)}$'s rows to have unit length, then feed it to the k-means algorithm.

13. Output: the grouping indicator

This scenario algorithm indicates a binary constrained spectral clustering problem where the solution vector V^* plays the role of grouping indicator. Without loss of generality, here we directly derive an algorithm for k-class problems (k≥2). We call the algorithm scalable constrained spectral clustering and list 13 steps in Algorithm 1.

Several key steps are interpreted as follows: (1) Step 7 aims to satisfy the condition $\lambda > 0$; (2) Step 8 aims to scale each eigenvectors for satisfying the condition of Eq. (11); (3) Step 9 aims to satisfy the condition of Eq. (4) In Step 11, we recover the solution vectors by the linear transformation $\hat{Z}^T u$ and we weight each solution vector by one minus the associated value of the objective function. It is worth mentioning that the input parameter b is tunable, making the algorithm flexible to noisy side information or inappropriate mathematical expressions for side information. Usually, the larger b is given, the more side information is respected.

2.5. Optimization algorithm

In this section, we consider the optimization algorithm to improve the scalable clustering performances.

Input: $\{x_i\}_{i\in u,}\{(x_{j1}, x_{j2,} l_j)_{j=1}^{L} \searrow, \delta_0$

Repeat

If t=0, 1, 2 then

$\delta = 0;$

Else

$\delta = \delta_0$

End

Find $y_i^{(t)}$ cluster optimally

Find $z_{j1}^{M(t)}, z_{j2}^{M(t)}$ must link constraints

Find $z_{j1}^{C(t)}, z_{j2}^{C(t)}$ cannot link constraints

Obtain optimal cluster performance

This algorithm is used to improve the clustering performance in higher rather than existing algorithm. The label is created more effectively based on the most informative cluster and also it can be able to handle the high dimensional dataset more effectively.

3. Results and Discussion

In this section the existing and proposed methodologies are compared using scalable clustering and optimization methods. The performance metrics are such as accuracy, precision and recall values which are evaluated by using efficient methods. In existing scenario the performance values are lower and the proposed method shows highest performance using optimization method. An experimental result shows that the proposed method achieves high performance in terms of precision, recall, and accuracy.

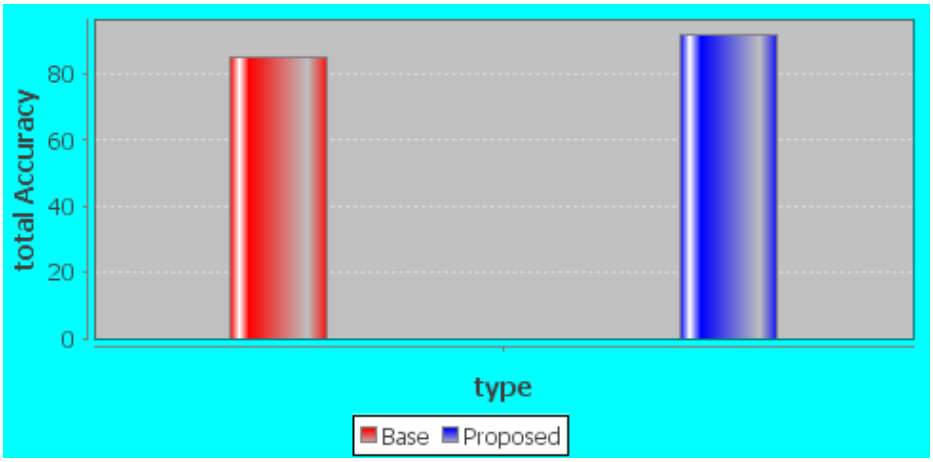
3.1. Accuracy

Accuracy is defined as the degree of generating the experimental output that is matches with the expected output. The accuracy is calculated by using the following equation

$$\text{Accuracy} = \frac{\text{True Positive+False Negative}}{\text{True Positive+True Negative+False Positive+False Negative}}$$

In this graph, x axis is taken for two methods of and y axis is taken for accuracy. From the Figure.1 the proposed scenario shows the highest accuracy rather than existing method. The proposed optimization method provides superior clustering performances in terms of accuracy values.

Figure 1. Accuracy comparison



3.2. Precision

Precision is defined as the Percentage of correct predicted results from the set of input terms. The precision value should be more in the proposed methodology than the existing approach for the better system performance.

Precision is calculated by using following equation

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive+False Positive}}$$

In this graph, x axis is taken for two methods of and y axis is taken for precision. From the Figure.2 the proposed scenario shows the highest precision rather than existing method. The proposed optimization method provides superior clustering performances in terms of precision values.

3.3. Recall

The recall or true positive rate (TP) is the proportion of positive cases that were correctly identified, as calculated using the equation:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive+True Negative}}$$

In this graph, x axis is taken for two methods of and y axis is taken for recall. From the Figure.3 the proposed scenario shows the highest recall rather than existing method. The proposed optimization method provides superior clustering performances in terms of recall values.

Figure 2. Precision comparison

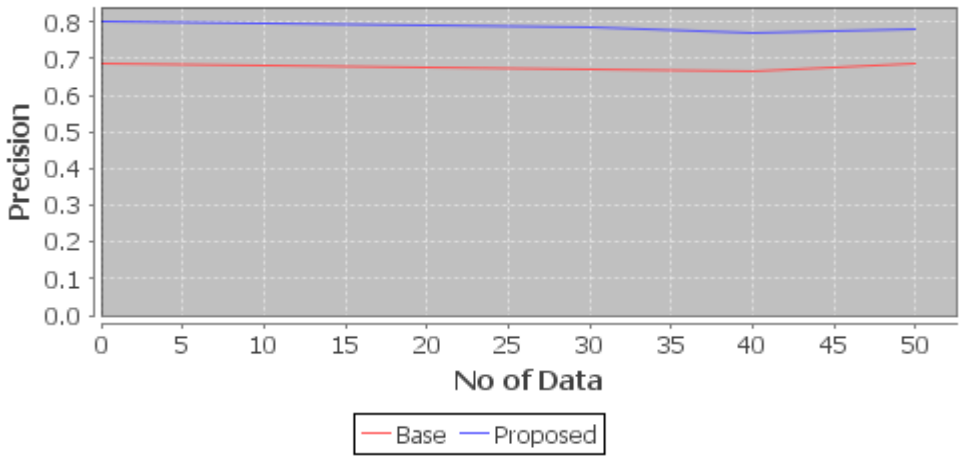
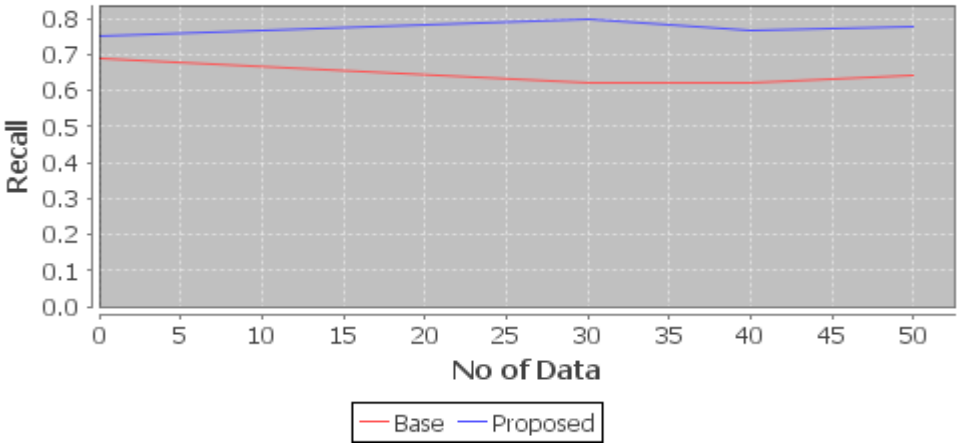


Figure 3. Recall comparison



4. Conclusion

Our proposed system yields greater performance by using efficient methods. We have developed a new k-way scalable constrained spectral clustering algorithm based on a closed-form integration of the constrained normalized cuts and the sparse coding based graph construction. The efficient methods are used to produce more accurate and scalable spectral clustering. It is focusing on the selection of most informative nodes in the clustering network. In real-world applications continuously and efficiently updates are required, over the data sets evolution. However clustering accuracy and optimization concept is still an issue in the existing scenario. To overcome this problem, we go for proposed scenario. In proposed system, we introduced the technique named as optimization algorithm which is used for improving the clustering performance more significantly. Pairwise constraints are introduced to estimate the higher similarity in the scalable clustering approaches. From the experimental result, we can say that the proposed system performed better than the existing system.

5. Acknowledgement

Authors express no conflict of interest.

6. References

1. P. More, O. Lawrence hall. Scalable clustering: a distributed approach. *Fuzzy Systems, Proceedings IEEE International Conference on*. 2004; 1.
2. Yingjie Xia, Zhenyu Shan, Yuncai Liu. Scalable constrained spectral clustering. *Knowledge and Data Engineering, IEEE Transactions on.* 2015; 27(2), 589-593.
3. Hendry Lin. Clustering. http://www.cs.cmu.edu/afs/andrew/course/15/381-f08/www/lectures/clustering.pdf.
4. Wagstaff Kiri, Claire Cardie. Clustering with instance-level constraints, *AAAI/IAAI* 1097.2000.
5. Wagstaff Kiri, Claire Cardie, Seth Rogers, Stefan Schroedl. Constrained k-means clustering with background knowledge, Proceedings of the Eighteenth International Conference on Machine Learning. 2001; 577-584.
6. P.Xing Eric, Andrew Y. Ng, Michael I. Jordan, Stuart Russell. Distance metric learning with application to clustering with side-information. *Advances in neural information processing systems*. 2002.
7. Basu Sugato, Mikhail Bilenko, J.Raymond Mooney. A probabilistic framework for semi-supervised clustering, *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2004; 59-68.
8. Shental Noam, Tomer Hertz, Aharon Bar-Hillel, Daphna Weinshall . Computing gaussian mixture models with EM using equivalence constraints, Advances *in neural information processing systems.*16.8 2004; 465-472.
9. Kulis Brian, Sugato Basu, Inderjit Dhillon, Raymond Mooney. Semi-supervised graph clustering: a kernel approach, *Machine learning.* 2009; 74(1), 1-22.
10. Zeng Hong, Yiu-ming Cheung. Semi-supervised maximum margin clustering with pairwise constraints, *Knowledge and Data Engineering, IEEE Transactions on*. 2012; 24(5), 926-939.
11. Li Zhenguo, Jianzhuang Liu, Xiaoou Tang. Constrained clustering via spectral regularization, *Computer Vision and Pattern Recognition, 2009, CVPR 2009. IEEE Conference on*. IEEE, 2009; 421-428.
12. Chen Xinlei, Deng Cai. Large scale spectral clustering with landmark-based representation. *AAAI*. 2011.

The Publication fee is defrayed by Indian Society for Education and Environment (iSee). www.iseeadyar.org

Citation:

Nagaraj, Thiagarasu, Jeevithapriya. Optimization and scalable constrained clustering performances. *Indian Journal of Innovations and Developments.* 2015; 4 (7), November.