ISSN 2321-743X

International Journal of Research in

Engineering and Bioscience

Volume 2 (Issue 4) Pages (209-218)

IJREB

Journal home page: www.ijreb.org

CLUSTERING SENTENCE - LEVEL TEXT USING FUZZY RELATIONAL CLUSTERING ALGORITHM

Ajini. A and V. Kathiresan

Department of Computer Science and Engineering, Narayanaguru College of Engineering, Manjalumoodu, K.K Dist, Tamil Nadu.

ABSTRACT

Fuzzy clustering algorithms allow patterns to belong to all clusters with differing degrees of membership. This is important in domains such as sentence clustering. It is a novel fuzzy clustering algorithm that operates on relational input data; i.e., data in the form of a square matrix of pair wise similarities between data objects. The algorithm uses a graph representation of the data, and operates in an Expectation-Maximization framework in which the graph centrality of an object in the graph is interpreted as likelihood. sentence clustering tasks demonstrate that the algorithm is capable of identifying overlapping clusters of semantically related sentences, and that it is therefore of potential use in a variety of text mining tasks. Also include results of applying the algorithm to benchmark data sets in several other domains. A detailed study about the Fuzzy Clustering Algorithm has been discussed in this review journal.

KEYWORDS: Fuzzy relational clustering, natural language processing, graph centrality.

IJREB

ISSN 2321-743X

International Journal of Research in

Engineering and Bioscience

Volume 2 (Issue 4) Pages (209-218)

Journal home page: www.ijreb.org

INTRODUCTION

The main objective is to develop a fuzzy relational clustering algorithm, it allow patterns to belong to all clusters with differing degrees of membership. Sentence clustering plays an important role in many text-processing activities. For example, various authors have argued that incorporating sentence clustering into extractive multi document summarization helps avoid problems of content overlap, leading to better coverage. However, sentence clustering can also be used within more general text mining tasks.

For example, consider web mining, where the specific objective might be to discover some novel information from a set of documents initially retrieved in response to some query. By clustering the sentences of those documents we would intuitively expect at least one of the clusters to be closely related to the concepts described by the query terms; however, other clusters may contain information pertaining to the query in some way hitherto unknown to us, and in such a case we would have successfully mined new information.

However, clustering text at the sentence level poses specific challenges not present when clustering larger segments of text, such as documents. Now highlight some important differences between clustering at these two levels, and examine some existing approaches to fuzzy clustering.

It is important in domains such as sentence clustering, since a sentence is likely to be related to more than one theme or topic present within a document or set of documents. Conventional fuzzy clustering approaches based on prototypes or mixtures of Gaussians are generally not applicable to sentence clustering. This novel fuzzy clustering algorithm that operates on relational input data; i.e., data in the form of a square matrix of pair wise similarities between data objects. The algorithm uses a graph representation of the data, and operates in an Expectation-Maximization framework. The algorithm is capable of identifying overlapping clusters of semantically related sentences. It can be use in a variety of text mining tasks.

This type of data, which refer to as attribute data, is amenable to clustering by a large range of algorithms. Since data points lie in a metric space, it can readily apply prototype-based algorithms such as k-Means, Isodata, Fuzzy c-Means (FCM) and the closely

IJREB

ISSN 2321-743X

International Journal of Research in

Engineering and Bioscience

Volume 2 (Issue 4) Pages (209-218)

Journal home page: www.ijreb.org

related mixture model approach, all of which represent clusters in terms of parameters such as means and covariance's, and therefore assume a common metric input space.

The vector space model has been successful in IR because it is able to adequately capture much of the semantic content of document-level text. This is because documents that are semantically related are likely to contain many words in common, and thus are found to be similar according to popular vector space measures such as cosine similarity, which are based on word co-occurrence.

RELATED WORK

Sentence clustering plays an important role in many text-processing activities. The section also includes a description of the sentence similarity measure used throughout the paper, as well as a description of several cluster evaluation criteria. The reports on the application of the algorithm to clustering sentences from a recent newswire article, demonstrating that the algorithm is capable of identifying overlapping clusters of semantically related sentences, and that it is therefore of potential use in a variety of text processing tasks including document summarization and text mining of a more general nature. Cluster membership values for each node represent the degree to which the object represented by that node belongs to each of the respective clusters, and mixing coefficients represent the probability of an object having been generated from that component. By applying the Page Rank algorithm to each cluster, and interpreting the Page-Rank score of an object within some cluster as likelihood, we can then use the Expectation- Maximization (EM) framework to determine the model parameters (i.e., cluster membership values and mixing coefficients). The result is a fuzzy relational clustering algorithm which is generic in nature, and can be applied to any domain in which the relationship between objects is expressed in terms of pair wise similarities. The input data in the form of a square matrix W=wij (often referred to as the affinity matrix), where wij is the (pair wise) relationship between the ith and jth data object. To distinguish it from attribute data,refer to such data as relational data.

IJREB

ISSN 2321-743X

International Journal of Research in

Engineering and Bioscience

Volume 2 (Issue 4) Pages (209-218)

Journal home page: www.ijreb.org

First describe the use of PageRank as a general graph centrality measure, and review the Gaussian mixture model approach. Then describe how PageRank can be used within an Expectation-Maximization framework to construct a complete relational fuzzy clustering algorithm.

In both TextRank and LexRank, each sentence in a document or documents is represented by a node on a graph. However, unlike a web graph, in which edges are unweighted, edges on a document graph are weighted with a value representing the similarity between sentences. The PageRank algorithm can easily be modified to deal with weighted undirected edges, resulting in,

$$PR(V_i) = (1-d) + d \times \sum_{j=1}^{N} w_{ji} \frac{PR(V_j)}{\sum_{k=1}^{N} w_{jk}}$$

Where W_{ij} is the similarity between Vj and Vi, and we assume that these weights are stored in a matrix W=$\{W_{ij}\}$ which refer to as the "affinity matrix." Note that the summations are now over all objects in the graph.

PROPOSED METHOD

This section describes a novel fuzzy clustering algorithm that operates on relational input data. Data in the form of a square matrix of pairwise similarities between data objects. The algorithm uses a graph representation of the data, and operates in an Expectation-Maximization framework in which the graph centrality of an object in the graph is interpreted as likelihood. Results of applying the algorithm to sentence clustering tasks demonstrate that the algorithm is capable of identifying overlapping clusters of semantically related sentences, and that it is therefore of potential use in a variety of text mining tasks. a fuzzy relational clustering algorithm which is generic in nature, and can be applied to any domain in which the relationship between objects is expressed in terms of pair wise similarities.

A. PageRank

PageRank is a general graph centrality measure and used to determine the importance of an object in a network. PageRank assigns to every node in a directed graph a numerical score between 0 and 1, known as its PageRank score (PR), and defined as;

ISSN 2321-743X
International Journal of Research in
Engineering and Bioscience
Volume 2 (Issue 4) Pages (209-218)
IJREB
Journal home page: www.ijreb.org

$$PR(V_i) = (1-d) + d \times \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} PR(V_j)$$

Where $\text{In}(V_i)$ is the set of vertices that point to $V_i, Out(V_j)$ is the set of vertices pointed to by V_j, and d is a damping factor.

The Expectation- Maximization algorithm is motivated by the mixture model approach, in which a density is modeled as a linear combination of C component densities p (x|m) in the form ∑∏ mp (m|p), where they $\prod_m$ represent the prior probability of data point x having been generated from component m of the mixture. Assuming that the parameters of each component are represented by a parameter vector , the problem is to determine the values of the components of this vector, and this can be achieved using the Expectation- Maximization algorithm. Following random initialization of the parameter vectors, m=0,1,…C can Expectation step (E-step), followed by a Maximization step (M-step), and are iterated until convergence.

C. Fuzzy Relational Clustering

Fuzzy clustering algorithms allow patterns to belong to all clusters with differing degrees of membership. This is important in domains such as sentence clustering. The data in the form of a square matrix of pairwise similarities between data objects. The algorithm uses a graph representation of the data, and operates in an Expectation-Maximization framework

D. Purity and Entropy

Two widely used external clustering evaluation criteria are purity and entropy fraction of the cluster size that the largest class of objects assigned to that cluster. Entropy of a cluster is defined as the measure of how the objects mixed within a cluster. Good clustering is thus characterized by a high purity and low entropy.

E. Rand Index and F-Measure

RI and F-Measure are based on statistics, Rand Index and F-measure are based on a combinatorial approach which considers each possible pair of objects.

Each pair can fall into one of four groups: if both objects belong to the same class and same cluster then the pair is a true

IJREB

ISSN 2321-743X

International Journal of Research in

Engineering and Bioscience

Volume 2 (Issue 4) Pages (209-218)

Journal home page: www.ijreb.org

positive (TP); if objects belong to the same cluster but different classes the pair is a false positive (FP); if objects belong to the same class but different clusters the pair is a false negative (FN); otherwise the objects belong to different classes and different clusters, and the pair is a true negative (TN). The Rand index is simply the accuracy; i.e.,

RI = (TP+TN) / (TP+FP+FN+TN)

The F-measure is another measure commonly used in the IR literature, and is defined as the harmonic mean of precision and recall; i.e., F-measure = 2PR/(P+R), where P = TP/(TP+FP) and R = TP/(TP+FN).

EXPERIMENTS AND PERFORMANCE

Fuzzy clustering algorithms allow patterns to belong to all clusters with differing degrees of membership. This is important in domains such as sentence clustering. The data in the form of a square matrix of pairwise similarities between data objects. The algorithm uses a graph representation of the data, and operates in an Expectation-Maximization framework. It can be use in a variety of text mining tasks. Unlike Gaussian mixture models, which use a likelihood function parameterized by the means and covariances of the mixture components, the proposed algorithm uses the PageRank score of an object within a cluster as a measure of its centrality to that cluster. These PageRank values are then treated as likelihoods. Since there is no parameterized likelihood function as such, the only parameters that need to be determined are the cluster membership values and mixing coefficients.

A. Convergence and Complexity

With regard to space complexity, the FRECCA algorithm is no more expensive than either the Spectral Clustering or k- Means families of algorithms, since all require the storage of the same, potentially large, similarity matrix. However, the time complexity of FRECCA far exceeds that of both Spectral Clustering and k-Means. Whereas Spectral Clustering performs single eigen value decomposition, FRECCA calls PageRank on each cluster during each Expectation step.

Moreover, membership values must be normalized following the computation of likelihoods. This will perturb the within-cluster PageRank values, and means that the outer repeat loop commencing may be slow to

IJREB

ISSN 2321-743X

International Journal of Research in

Engineering and Bioscience

Volume 2 (Issue 4) Pages (209-218)

Journal home page: www.ijreb.org

converge, particularly if cluster membership values are initialized randomly. Gaussian mixture models do not have this problem because they represent clusters using means and covariances, which are more stable under changes in cluster membership.

B. Duplicate Clusters

The number of initial clusters must be specified as input to the algorithm. If this number is too high, then duplicate clusters (i.e,clusters with identical membership values across all objects) will be found. While it might appear at first sight that duplicate clusters can simply be removed after the algorithm has converged, and membership subsequently renormalized to sum to one, this is not possible because of the coupling between membership values and PageRank values. That is, we cannot assume that the current PageRank values will be correct under a renormalization of membership values. The solution is to perform a check for duplicate clusters at the completion of each Maximization step. If duplicate clusters are found, membership values are renormalized, and the algorithm is allowed to proceed until a stage at which convergence has been achieved and no duplicate clusters exist.

C. Hard Clustering

The algorithm outputs cluster membership values which represent the degree of membership of object i to cluster m. If hard clustering is required, this can be trivially achieved by assigning a sentence to the cluster m for which membership is highest; i.e.., * +

D. Effect of Damping Factor

The damping factor d that appears in the PageRank calculation affects the fuzziness of the clustering, but generally does not affect the number of clusters, provided that the value is above approximately 0.8.

In general, the higher the value of d, the harder is the clustering, with cluster membership values being close to either zero or one. We have used a value of 0.85 in all of the experiments described.

Thresholding of Similarity Values Depending on the domain, the graph representing the relation between objects may be heavily or sparsely connected. In the case of sentence clustering, we have found that many of the similarities sij between sentences are very small, and are likely to be spurious,

IJREB

ISSN 2321-743X

International Journal of Research in

Engineering and Bioscience

Volume 2 (Issue 4) Pages (209-218)

Journal home page: www.ijreb.org

arising from incidental similarities between words in sentences which are in fact not semantically related. In practice, we have found that the clustering performance of the algorithm can be improved by thresholding these similarity values such that all values below the threshold are converted to zero. All sentence clustering results reported in this paper are based on thresholding similarity values such that 50 percent of the values in the affinity matrix are zero.

E. Measuring Sentence Similarity

To calculate similarity values sij for the affinity matrix we use a modified version of the measure proposed by Li et al. This approach is similar to that used to calculate document similarity in the IR literature; however, rather than using a common vector space representation for all sentences, the two sentences being compared are represented in a reduced vector space of dimension n, where n is the number of distinct nonstop words appearing in the two sentences.

F. Cluster Evaluation Criteria

Cluster evaluation may be either supervised, in which case external information (usually known class labels associated with the instances) is used to measure the goodness of the clustering; or unsupervised, in which case no external information is used. In the following L= {W1, W2….} is the set of clusters, C={C1, C2…..}is the set of classes (for supervised evaluation) and N is the number of objects.

Once PageRank scores have been determined, these are treated as likelihoods and used to calculate cluster membership values. Maximization step. Since there is no parameterized likelihood function, the maximization step involves only the single step of updating the mixing coefficients based on membership values calculated in the Expectation Step. Assumes here that cluster membership values are initialized randomly, and normalized such that cluster membership for an object sums to unity over all clusters. Mixing coefficients are initialized such that priors for all clusters are equal. The algorithm uses Expectation Maximization to optimize these parameters. We assume in the following that the similarities between objects are stored in a similarity matrix S= { }.

IJREB

ISSN 2321-743X
International Journal of Research in
Engineering and Bioscience
Volume 2 (Issue 4) Pages (209-218)
Journal home page: www.ijreb.org

CONCLUSION

In comparison with hard clustering methods, in which a pattern belongs to a single cluster, fuzzy clustering algorithms allow patterns to belong to all clusters with differing degrees of membership. This is important in domains such as sentence clustering, since a sentence is likely to be related to more than one theme or topic present within a document or set of documents. However, because most sentence similarity measures do not represent sentences in a common metric space. A novel fuzzy clustering algorithm that operates on relational input data; i.e., data in the form of a square matrix of pair wise similarities between data objects. The algorithm uses a graph representation of the data, and operates in an Expectation-Maximization framework in which the graph centrality of an object in the graph is interpreted as likelihood. Results of applying the algorithm to sentence clustering tasks demonstrate that the algorithm is capable of identifying overlapping clusters of semantically related sentences, and that it is therefore of potential use in a variety of text mining tasks. The result is a fuzzy relational clustering algorithm which is generic in nature, and can be applied to any domain in which the relationship between objects is expressed in terms of pair wise similarities.

REFERENCES

1. B.J. Frey and D. Dueck, "Clustering by Passing Messages between Data Points," Science, vol. 315, pp. 972-976, 2007.

2. H. Zha, "Generic Summarization and Keyphrase Extraction Using Mutual Reinforcement Principle and Sentence Clustering," Proc. 25th Ann. Int'l ACM SIGIR Conf. Research and Development In Information Retrieval, pp. 113- 120, 2002.

3.D.R. Radev, H. Jing, M. Stys, and D. Tam, "Centroid-Based Summarization of Multiple Documents," Information Processing and Management: An Int'l J., vol. 40, pp. 919-938,

2004.

4. R.M. Aliguyev, "A New Sentence Similarity Measure and Sentence Based Extractive Technique for Automatic Text Summarization," Expert Systems with Applications, vol. 36, pp. 7764-7772, 2009.

5. R. Kosala and H. Blockeel, "Web Mining Research: A Survey," ACM SIGKDD Explorations Newsletter, vol. 2, no. 1, pp. 1-15, 2000.

6. G. Salton, Automatic Text Processing:

ISSN 2321-743X
International Journal of Research in
Engineering and Bioscience
Volume 2 (Issue 4) Pages (209-218)
IJREB
Journal home page: www.ijreb.org

The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley, 1989.

7. J.B MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," Proc. Fifth Berkeley Symp. Math.Statistics and Probability, pp. 281-297, 1967.

8. G. Ball and D. Hall, "A Clustering Technique for Summarizing Multivariate Data," Behavioural Science, vol. 12, pp. 153-155, 1967.

9. J.C. Dunn, "A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact Well- Separated Clusters," J. Cybernetics,vol. 3, no. 3, pp. 32-57, 1973.

10. J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, 1981.

11. R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification, second ed. John Wiley & Sons, 2001.