International Conference on Latest Advancements & Future Trends In Engineering, Science & Management
Sambhram Institute of Technology, Bangalore (India)
03rd - 04th May 2019 www.conferenceworld.in
SAMBHRAM INSTITUTE OF TECHNOLOGY
ISBN : 978-93-87793-86-6

Self-Tuned Descriptive Document Clustering

[1]Kailash Vashisth, [2]Lumeez Fathima, [3]Shreeja, [4]Mallamma C.G.
[1,2,3]Department of CSE ,Sambhram Institute of Technology, Bangalore
[4]Assistant Professor, Department of CSE, Sambhram Institute of Technology, Bangalore

Abstract

Descriptive clustering is a problem of discovering diverse groups of semantically related documents described with significant and close-packed text labels. It comprises of naturally systematize information into clusters and producing a illustrative synopsis for every cluster. The thought displayed endeavors moving the importance to cluster summary. The depiction ought to illuminate about the substance regarding each cluster moving along without any more examination of the specific cases, allowing a user to swiftly search for appropriate clusters.

We model descriptive clustering as an auto-encoder network that predicts features from cluster assignments and predicts cluster assignments from a subset of features. The subset of features used for predicting a cluster serves as its description.

In the proposed network, cluster predictions are made using logistic regression models, and feature predictions rely on logistic or multinomial regression models. Optimizing these models leads to a completely self-tuned descriptive clustering approach that automatically selects the number of clusters and the number of feature for each cluster.

Keywords*– Descriptive Clustering, *Predictive clustering, Regression models.

1. INTRODUCTION

Data is at the heart of technical innovations, achieving any result is now possible using predictive models.Clustering is the errand of collection a lot of items so that objects in a similar gathering (called a cluster) are progressively comparative (in some sense) to one another than to those in different gatherings (clusters). Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups Clustering is an automatic learning technique aimed at grouping a set of objects into subsets or clusters. The aim is to generate clusters that are coherent internally, but substantially different from each other. In plain words, objects in the same cluster should be as similar as possible, whereas objects in one cluster should be as dissimilar as possible from objects in the other clusters.

Automatic document cluster has compete a very important role in several fields like info retrieval, data processing, etc .Automatic document cluster has compete a very important role in several fields like info retrieval, data processing, etc . In particular, descriptive clustering consists of automatically grouping sets of similar instances into clusters and automatically generating a human-interpretable description or summary for each cluster. Each cluster's description allows a user to ascertain the cluster's relevance without having to examine its contents.The aim of this thesis is to improve the efficiency and accuracy of document clustering.

With the aim of defining an objective criterion, we consider a direct correspondence between description and prediction. We assume each instance is represented with sparse features (such as a bag of words), and each cluster will be described by a subset of features. A cluster's description should summarize its contents, such that the description alone should enable a user to predict whether an arbitrary instance belongs to a particular cluster. Likewise, a machine classifier trained using the features subset should also be predictive of the cluster membership. The classification accuracy provides an objective and quantitative criterion to compare among different feature subsets. To serve as a concise description, the number of features used by the classifier must be limited (e.g., a linear classifier that uses all features is not easily interpretable). A relatively small set of predictive features can be identified using various feature selection methods. In particular, we identify features subsets by various statistical and information-theoretic criteria and by training linear classifiers with additional sparsity-inducing regularizations , e.g., the `1-norm for the Lasso or a combination of `1 and `2-norms for the Elastic Net, such that only a small set of features have non-zero coefficients. In a similar spirit, Lasso has been used for selecting predictive features for explaining classification models. In addition to the cardinality constraint on the number of features, we only permit features that are positively correlated with a given cluster, i.e., features whose presence are indicative of the cluster. This constraint ensures that no cluster is described by the absence of features, which are present in other clusters. For instance, given a corpus of book and movie reviews, the positivity constraint avoids a cluster consisting of mainly of book reviews from being described as ¬movie, i.e., the absence of the word feature movie. In general, this constraint can be enforced by admitting only features that are positively correlated with a particular cluster; for linear classifiers, this can be done by enforcing the constraint that the coefficients are nonnegative.

2. PROPOSED SYSTEM

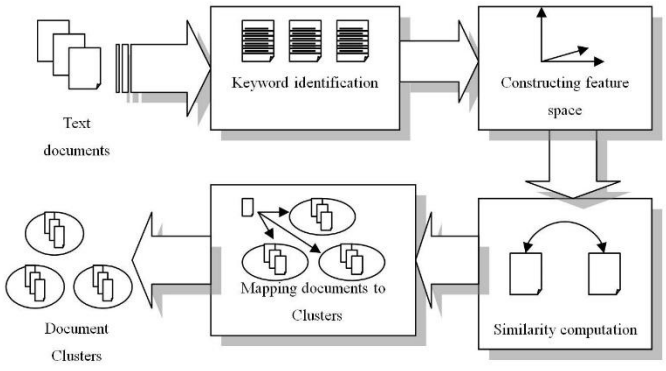


Figure 1: System architecture

The PDC framework consists of two prediction tasks: predicting the original feature occurrences based on the cluster assignments and predicting whether an instance belongs to a specific cluster using only a small set of feature dimensions that serve as the description for the cluster. The first task provides a quantitative objective to automatically select from clusterings with different numbers of clusters. Each cluster is associated with a certain distribution of the features, with some features occurring more (or less) frequently for instances within that

ISBN : 978-93-87793-86-6

cluster. The second task of predicting the cluster membership is clearly dependent on the chosen clustering.The amount of information carried by the clustering increases with more clusters.

• *Create candidate clustering*: A set of possible clusterings are formed.

• *Select the most predictive clustering*. For each clustering, a model is trained to predict feature occurrences from the cluster assignments. The clustering associated with the most predictive model is selected.

• *Create descriptive feature subsets*. For each cluster, different subsets of features are chosen by a feature selection mechanism. In particular, candidate feature subsets are identified using logistic regression with positivity constraints and sparsity-inducing regularization by varying the amount of regularization.

• *Select the most informative feature subset:* A logistic regression model is trained for each candidate feature subset. The best model is selected by a model order selection criterion that balances cardinality with predictive performance.

3. METHODOLOGY

3.1) LINEAR REGRESSION

Linear regression is the most simple method for prediction. It uses two things as variables which are the predictor variable and the variable which is the most crucial one first whether the predictor variable and su. These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The equation of the regression equation with one dependent and one independent variable is defined by the formula.

$$B = Y + X*A$$

where b = estimated dependent variable score, y = constant, x = regression coefficient, and a = score on the independent variable.

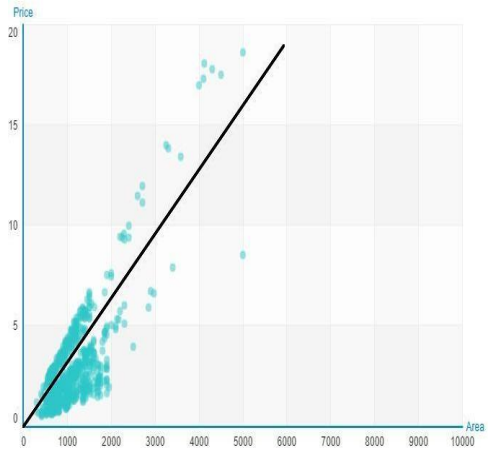


Figure 2: Linear regression plot.

ISBN : 978-93-87793-86-6

3.2) Forest Regression

Forest regression uses the technique called as Bagging of trees. The main idea here is to decorrelate the several trees. We then reduce the Variance in the Trees by averaging them. Using this approach, a large number of decision trees are created.

Random forest training algorithm applies the technique of bootstrap aggregating, or bagging, to tree learners.

Given a training set $X = x1, ..., xn$ with responses $Y = y1, ..., yn$, bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples:

For $b = 1, ..., B$:

1. Sample, with replacement, n trainingexamples from X, Y; call these Xb,Yb.
2. Train a classification or regression tree fb on Xb,Yb.

After training, predictions for unseen samples a' can be made by averaging the predictions from all the individual regression tress on a':

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

Additionally, an estimate of the uncertainty of the prediction can be made as the standard deviation of the predictions from all the individual regression trees ona':

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B}(f_b(x') - \hat{f})^2}{B-1}}.$$

Representation of it is as follows-

ISBN : 978-93-87793-86-6

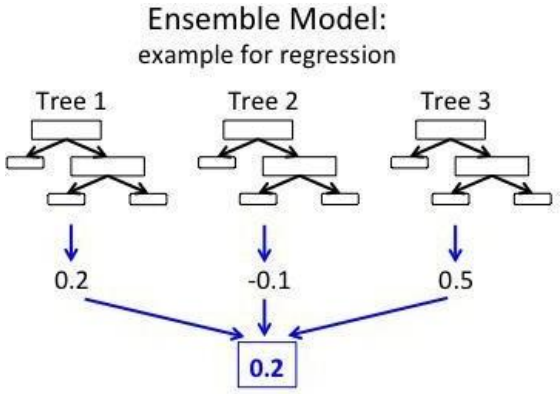


Figure 3: Boosted regression model

3.3) Boosted Regression

Boosted regression is a type of learning technique which produces prediction with the help of decision trees that usually ensemble a number of weak prediction models .

This Boosting algorithm assumes a real life value *y* and seeks an approximation F(x) in the form of a weighted sum of hi(x) from class H called weak learners:

$$F(x) = \sum_{i=1}^{M} \gamma_i h_i(x) + \text{const}.$$

3.4) Partitioning Algorithms

- Partitioning method: Usually constructs a partition of a given database D of n objects into a set of k clusters •
- Given a k, find a partition of k clusters that optimizes the chosen partitioning criterion

  – Global optimal: exhaustively enumerate all partitions

  – Heuristic methods: k-means and k-medoids algorithms

  – k-means (MacQueen'67): Each cluster is represented by the center of the cluster

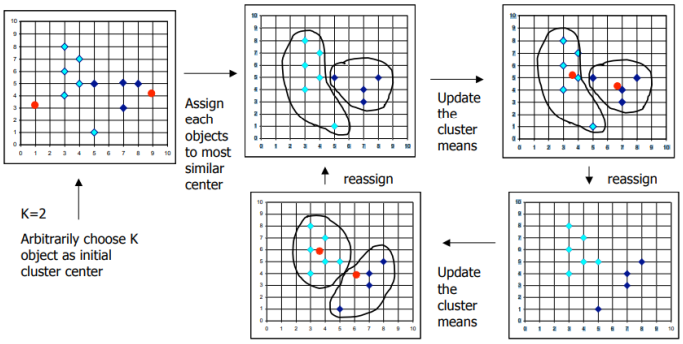
  – k-medoids or PAM (Partition around medoids) (Kaufman & Rousseeuw'87): Each cluster is represented by one of the objects in the clusters.

ISBN : 978-93-87793-86-6

3.5) K means algorithm

```
randomly chose k examples as initial centroids
while true:
   create k clusters by assigning each
       example to closest centroid
   compute k new centroids by averaging
       examples in each cluster
   if centroids don't change:
       break
```

Example:-



The k means algorithm attempts to minimize the within-cluster point scatter

$$W(C)=\frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(i')=k}d(x_i,x_{i'})$$

Since the total point scatter T is constant. Hence minimizing W(C) is equal to maximizing B(C)

$$B(C)=\frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(i')\neq k}d(x_i,x_{i'})$$

Since T=W(C) + B(C)

4. CONCLUSION

There are two prediction tasks: 1) selecting a cluster which is predicts thefeature. 2)Predicting the assignment of cluster's feature. Hence, by using Predictive nature as the objective criterion i.e. the parameter of descriptiveclustering, the number of clusters and its features are selected based on Modal Selection. The resulting solution has minimum subset which helps in identifying a cluster by its special features. The hypothesis demonstrates how the user will be able to predict the cluster(i.e. descriptive features) based on algorithm. When provided with adequate requirements it is very likely to locate the clusters. This is tested on datasets count-valued features. Hence for the feature selection information theoretic feature selection and linear

classifier is used. These tests prove to be successful as it uses meaningful terms of clusters and provide subsets which are indicative about the cluster content. The proposed PDC Framework distinguishes from the previous clustering techniques as they lacked a proper description for the clusters. The proposed framework helps the users to screen large collections of abstracts to support the development of systematic reviews, mainly in the public health domain.

5. REFERENCES

[1] D. Cai, C. Zhang, X. He, "Unsupervised feature selection for multi-cluster data", Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 333-342, 2010.

[2] O.Zami O. Etzioni "Web document clustering: a feasibility demonstration". In ACM SIGIR 1997.

[3] P. S. Bradley and O. L. Mangasarian, "Feature selection via concave minimization and support vector machines," in Proc. Int.Conf. Machine Learning 1998.

[4] R. Kohavi and G. H. John, "Wrappers for feature subset selection," Artificial Intelligence., vol. 97, no. 1, pp. 273–324, 1997.

[5] J. Kittler, Feature Selection and Extraction (Academic Press, New York, 1986) Chapter 3, 59-83.

[6] D.W. Aha, Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms, Infernat. J. Man-Machine Studies 36 (1992) 267-287.

[7] M. A. Hearst, The Use of Categories and Clusters for Organizing Retrieval Results. Springer Netherlands, 1999, pp. 333–374.

[8] F. Pereira, N. Tishby, and L. Lee, "Distributional clustering of English words," in Proc. Annu. Meeting Assoc. Comput. Ling., 1993, pp. 183–190.