International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

Performance Evaluation of Cluster Based Algorithm used for Text Document Classification

Rohit S. Patil[1], Manish Bhardwaj[2]

[1, 2]Department Of Computer Technology, Kavikulguru Institute of Technology & Science, Ramtek, Nagpur, Maharashtra, India- 441 106

Abstract: ***In this paper we develop a complete methodology for document classification and clustering. We start by investigating how the choice of document features influences the performance of a document classifier and then use our findings to develop a clustering method suitable for document collections. From our study of the effect of frequency transformation, term weighting and dimensionality reduction through principal components analysis on the performance of a simple K-nearest-neighbors classifier, we conclude that: (a) applying a logarithm or square-root transformation to the term frequencies reduces error rates; (b) term weighting of the transformed frequencies does not appear to help much; and (c) increasing the feature space dimension beyond 50 does not improve performance. We use these findings in the construction of a Gaussian Mixture Document Clustering (GMDC) algorithm. This algorithm models the data as a sample from a Gaussian mixture. The model is used to build clusters based on the likelihood of the data, and to classify documents according to Bayes rule. Finally we will build our own classifier which will have ability to automatically select the number of clusters present in the document collection and do classification more efficiently then above two classifier.***

Keywords: clustering, classification, text mining, dimensionality reduction, Gaussian mixture

1. Introduction

Document detection and tracking is different from document retrieval. The goal of document retrieval is to find the documents in a collection that best match some query. The query might be considered a very short document consisting of a few keywords, and the goal then is to find the documents in the collection that are most similar to the query document.

In Statistics terminology, topic detection is a clustering problem: we want to partition C into groups such that documents in each group are similar to each other, and dissimilar from documents in other groups.

In its simplest form, topic tracking is a classification problem. We have a collection C of documents, each labeled with a topic, and we want to assign a label to a new document. The unusual aspect of the problem is that our answer could be “none”, in which case the document is taken to represent a new topic.

Clustering and classification methods play a central role in the reduction of both the number of operations needed for document classification, and the retrieval time. Also, they can be designed to make accurate decisions on whether or not a document represents a new topic.

The problem of clustering has been studied widely in the database and statistics literature in the context of a wide variety of data mining tasks. The clustering problem is defined to be that of finding groups of similar objects in the data. The similarity between the objects is measured with the use of a similarity function. The problem of clustering can be very useful in the text domain, where the objects to be clusters can be of different granularities such as documents, paragraphs, sentences or terms. Clustering is especially useful for organizing documents to improve retrieval and support browsing

The study of the clustering problem precedes its applicability to the text domain. Traditional methods for clustering have generally focused on the case of quantitative data in which the attributes of the data are numeric. The problem has also been studied for the case of categorical data in which the attributes may take on nominal values. A broad overview of clustering (as it relates to generic numerical and categorical data) may be found in . A number of implementations of common text clustering algorithms, as applied to text data, may be found in several toolkits such as Lemur and BOW toolkit. The problem of clustering finds applicability for a number of tasks:

Document Organization and Browsing: The hierarchical organization of documents into coherent categories can be very useful for systematic browsing of the document collection. A classical example of this is the Scatter/Gather method, which provides a systematic browsing technique with the use of clustered organization of the document collection.

Corpus Summarization: Clustering techniques provide a coherent summary of the collection in the form of cluster-digests or word-clusters, which can be used in order to provide summary insights into the overall content of the underlying corpus. Variants of such methods, especially sentence clustering, can also be used for document summarization, a topic. The problem of clustering is also closely related to that of dimensionality reduction and topic modeling. Such dimensionality reduction methods are all different ways of summarizing a corpus of documents.

Document Classification: While clustering is inherently an unsupervised learning method, it can be leveraged in order to improve the quality of the results in its supervised variant. In particular, word-clusters and co-training methods can be used in order to improve the classification accuracy of supervised applications with the use of clustering techniques.

We note that many classes of algorithms such as the k-means algorithm or hierarchical algorithms are general-purpose

methods, which can be extended to any kind of data, including text data. A text document can be represented either in the form of binary data, when we use the presence or absence of a word in the document in order to create a binary vector. In such cases, it is possible to directly use a variety of categorical data clustering algorithms [10, 4] on the binary representation.

A more enhanced representation would include refined weighting methods based on the frequencies of the individual words in the document as well as frequencies of words in an entire collection Quantitative data clustering algorithms can be used in conjunction with these frequencies in order to determine the most relevant groups of objects in the data.

2. Literature Survey

Our document clustering method borrows ideas from the model based clustering literature (Banfi eld and Raftery, 1993; Celeux and Govaert, 1995). It explicitly models the data as a sample from a Gaussian mixture. Each of the components in the mixture distribution is assumed to be a multivariate Gaussian distribution with uncorrelated components. This assumption fits the data well and greatly simplifies the computations involved, including the estimation of the parameters.

These are efficiently estimated through the Expectation-Maximization (EM) algorithm (Dempster, Laird and Rubin, 1977). We propose several ways of initializing the EM algorithm; these include efficient and accurate variations of the K-means algorithm (Ward, 1963), as well as the more popular agglomerative hierarchical clustering techniques. The model is used to build clusters based on the likelihood of the data, and to classify documents according to Bayes rule (Sec. We call this approach to document clustering Gaussian Mixture Document Clustering (GMDC).

One main advantage of our approach is the ability to automatically estimate the number of clusters (topics) present in the document collection via Bayes factors (Raftery, 1995)., with the TDT Corpus (Allan, Carbonell, Doddington, Yamron and Yang, 1998), are extremely encouraging, demonstrating the ability of GMDC to choose a reasonable number of clusters as well as to generate meaningful partitions of the data.

Our ideas have been successfully applied to large collections of documents and in general to large data sets, through a simple procedure that combines “fractionation” (Cutting, Karger, Pedersen and Tukey, 1992) with Gaussian mixture document clustering. The study of this extension has been published elsewhere (Tantrum, Murua and Stuetzle, 2004; Tantrum, Murua and Stuetzle, 2002). The present work focuses on the foundations of our methodology. Similar ideas to deal with large datasets, but not necessarily documents, have been reported in the literature recently. But unlike our comprehensive treatment of the problem for document collections, they focus on scalability of the original EM algorithm used to fit Gaussian mixtures (Jin, Wong and Leung, 2005), or on feature extraction procedures for general data sets (Hsieh, Wang and Hsu, 2006).

Text clustering algorithms are divided into a wide variety of different types such as agglomerative clustering algorithms, partitioning algorithms, and standard parametric modeling based methods such as the EM-algorithm. Furthermore, text representations may also be treated as strings (rather than bags of words). These different representations necessitate the design of different classes of clustering algorithms. Different clustering algorithms have different tradeoffs in terms of effectiveness and efficiency. An experimental comparison of different clustering algorithms may be found in [9, 11].

Many clustering techniques have been proposed in the literature. Clustering algorithms are mainly categorized into hierarchical and partitioning methods [2, 3, 4 5]. A hierarchical clustering method works by grouping data objects into a tree of clusters [6]. These methods can further be classified into agglomerative and divisive hierarchical clustering depending on whether the hierarchical decomposition is formed in a bottom-up or top-down fashion. Kmeans and its variants [7, 8, 9] are the most well-known partitioning methods [10]. Lexical chains have been proposed in [11] that are constructed from the occurrence of terms in a document.

Problem to improve the clustering quality is addressed in [10] where the cluster size varies by a large scale. They have stated that variation of cluster size reduces the clustering accuracy for some of the state-of-the-art algorithms. An algorithm called frequent Itemset based Hierarchical clustering (FIHC) has been proposed, where frequent items i.e. minimum fraction of documents have used to reduce the high dimensionality and meaningful cluster description. However, it ignores the important relationship between words.

3. Proposed Methodolgy

Clustering and classification methods play a central role in the reduction of both the number of operations needed for document classification, and the retrieval time. Also, they can be designed to make accurate decisions on whether or not a document represents a new topic.

In order to apply clustering and classification methods, we fi first map documents to vectors in some p-dimensional space. This is not strictly speaking necessary. Most clustering methods and some classification methods (for example K-nearest-neighbor classification) only require similarities or dissimilarities between documents. However, the distinction is not as important as it might seem at first glance. Given a representation of documents as p-dimensional points we can always define dissimilarity as interpoint distance. Given a dissimilarity matrix, on the other hand, we can use multi-dimensional scaling to find points in p-dimensional Euclidean space such that the interpoint distances approximately or exactly match the dissimilarities.

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses.

Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems.

Data mining tools can answer business questions that traditionally were too time consuming to resolve. They search databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. Different types of data mining tools are available in the marketplace, each with their own strengths and weaknesses. Internal auditors need to be aware of the different kinds of data mining tools available and recommend the purchase of a tool that matches the organization's current detective needs. This should be considered as early as possible in the project's lifecycle, perhaps even in the feasibility study.

Data mining commonly involves four classes of tasks.[1]

Classification- Arranges the data into predefined groups. For example an email program might attempt to classify an email as legitimate or spam. Common algorithms include Decision Tree Learning, Nearest neighbor, naive Bayesian classification and Neural network.

Clustering- Is like classification but the groups are not predefined, so the algorithm will try to group similar items together.

Regression- Attempts to find a function which models the data with the least error.

Association rule learning- Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as "market basket analysis". Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line. When implemented on high performance client/server or parallel processing computers, data mining tools can analyze massive databases to deliver answers to questions such as, "Which clients are most likely to respond to my next promotional mailing, and why?" As shown in Fig1, Data mining process consist of three main phases:-

1. Data preprocessing
2. Applying data mining techniques
3. Interpretation of Results

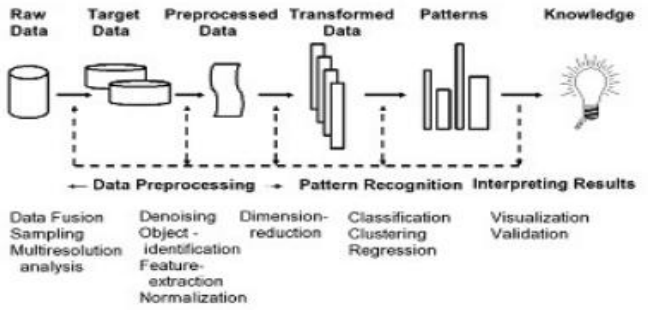

Figure 1: Data Flow of the Proposed System

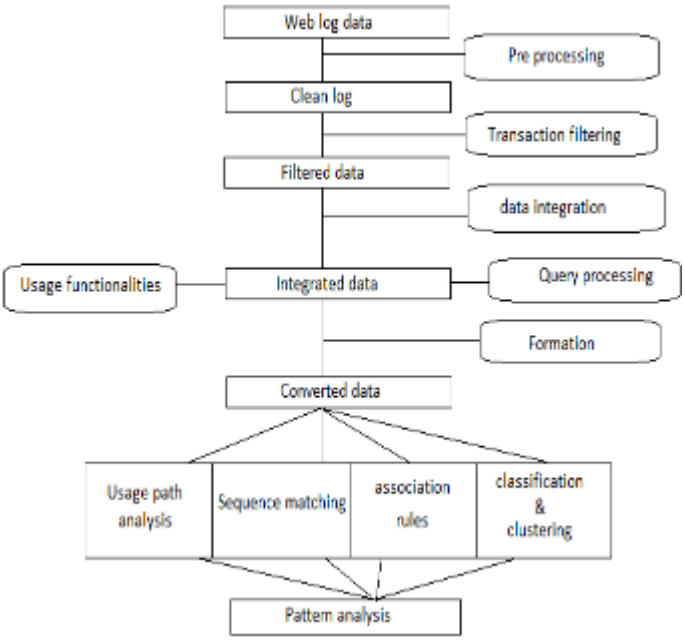

Figure 2: General Architecture of Text Clustering and Classification

According to the figure the processing steps and various techniques are similar to the data mining process.

References

[1] J. Han and M. Kimber. 2000. Data Mining: Concepts and Techniques. Morgan Kaufmann.
[2] Jain, A.K, Murty, M.N., and Flynn P.J. 1999. Data clustering: a review. ACM Computing Surveys, pp. 31, 3, 264-323.
[3] M. Steinbach, G. Karypis, and V. Kumar. 2000. A comparison of document clustering techniques. KDD Workshop on Text Mining"00.
[4] P. Berkhin. 2004. Survey of clustering data mining techniques [Online]. Available: http://www.accrue.com/products/rp_cluster_review.pdf.
[5] Xu Rui. 2005. Survey of Clustering Algorithms. IEEE Transactions on Neural Networks, 16(3):pp. 634-678.
[6] Miller G. 1995. Wordnet: A lexical database for English. CACM, 38(11), pp. 39–41.
[7] L. Zhuang, and H. Dai. 2004. A Maximal Frequent Itemset Approach for Document Clustering. Computer and Information Technology, CIT. The Fourth International Conference, pp. 970 – 977.
[8] R. C. Dubes and A. K. Jain. 1998. Algorithms for clustering Data. Prentice Hall college Div, Englewood Cliffs, NJ, March.

[9] D. Koller and M. Sahami. 1997. Hierarchically classifying documents using very few words. In Proceedings of ICML) 97, 14th International Conference on Machine Learning, pp. 170–178, Nashville, US.
[10] B.C.M.Fung, K.Wan, M.Ester. 2003. Hierarchical Document Clustering Using Frequent Itemsets", SDM"03.
[11] Green, S. J. 1999. Building hypertext links by computing
[12] Allan, J., Carbonell, J., Doddington, G., Yamron, J. and Yang, Y. 1998. Topic detection and tracking pilot study final report.
[13] Banfield, J. D. and Raftery, A. 1993. Model-based Gaussian and non-Gaussian clustering, Biometrics 49: 803–821.
[14] Berry, M., Drmac, Z. and Jessup, E. 1999. Matrices, vector spaces, and information retrieval, SIAM Review 41(2): 335–362.
[15] Berry, M., Dumais, S. and O'Brien, G. 1995. Using linear algebra for intelligent information retrieval, SIAM Review 37(4): 573–595.
[16] Celeux, G. and Govaert, G. 1995. Gaussian parsimonious clustering models, Pattern Recognition 28: 781–793.
[17] Javed Aslam, Katya Pelekhov, and Daniela Rus, A Practical Clustering Algorithm for Static and Dynamic Information Organization, Proceedings of the 1998 ACM CIKM International Conference on Information and Knowledge Management, Bethesda, Maryland, USA, Pages 208-217, November3-7, 1998