ISSN:2249-5789
Simmi John et al , International Journal of Computer Science & Communication Networks,Vol 2(4), 464-467

High Dimensional Hierarchical Data Clustering using SVM with Kernel Region Approximation Indexing

Ms. Simmi John
Research Scholar,
Dr. G. R. Damodaran college of Science,
Coimbatore, Tamilnadu, India.
simmishibu_john@yahoo.com

Mr. R. Boobathiraj
Asst. Professor,
Dr. G. R. Damodaran college of Science,
Coimbatore, Tamilnadu, India.
boobathiraj.r@grd.edu.in

Abstract

Text clustering is the process of partitioning a particular collection of texts into subgroups including content based similar ones. The importance of text clustering is to meet human interests in information searching and understanding. This paper proposes a scalable SVM classification method called CB-SVM (Cluster Based SVM). This applies an agglomerative hierarchical clustering method that provides an SVM with high quality samples that carry the statistical summaries of the data such that the summaries maximize the benefit of learning the SVM. However the indexing structure used in SVM is not suitable for high dimensional datasets. To overcome this, we propose a high dimensional indexing method (Kernel Region Approximation Blocks) which is an extension of the region approximation approach of the kernel space. This paper also compares the system with hierarchical text clustering algorithm HBSH (Hash-Based Structure Hierarchical Clustering), which is suitable for massive text clustering. The average time of CB-SVM is faster than that of traditional text clustering algorithms.

Keywords: *– HBSH, high dimensional, kernel trick, SVM, text clustering.*

1. Introduction

Text clustering is used in an effective way for sorting large or massive documents, which help the users to access, summarize, and organize text documents. Document clustering is one of the most crucial techniques for organizing documents in an unsupervised manner. In recent years text mining has become an important research area, as most of the information over 80% is stored as text. Document clustering is a subset of the larger field of data clustering; it includes the concepts of information retrieval (IR), natural language processing (NLP), machine language (ML) and others [8]. Now text mining is believed to have a high commercial potential value. A collection of text can be organized in two ways [1]. The first one is text categorization and the second one is text clustering. The first one is based on the prior knowledge about the collection of text. It catalogs data, according to a predetermined taxonomy or organization such as color, weight and so on. In the absence of the prior knowledge we go for the second one. Text clustering has greater complexity in its algorithm while comparing with text categorization [2]. A good clustering can be viewed as one that organizes a collection into groups such that the documents within each group are both similar to each other and dissimilar to those in other groups. According to this paper, based on the user's query, the data are grouped. Most of the existing text clustering algorithms, documents are represented using the vector space model. One of the major problems of this representation is the high dimensionality of the vector space, which cause a great challenge to the performance of clustering algorithms. To overcome this we go for Support Vector Machine (SVM) [3].

SVMs are highly used for classification and regression purposes. The machine learning capability of SVM is very useful in areas like text mining. The training complexity of SVM is highly dependent on the size of the data set. The CB-SVM constructs cluster using kernel trick method. Kernel trick is an SVM based method for nonlinear data analysis. Using this we can implicitly map data from the original space to a high dimensional feature space via a nonlinear mapping. The CB-SVM can be specifically used for clustering massive data sets [3].

2. Existing Methodology

In this paper we have used the existing approaches to give more weight to our work. The below mentioned approaches are stated as under:

A. Hierarchical clustering

B. Support Vector Machine
C. Kernel Region Approximation Block

A. Hierarchical Clustering

In agglomerative (bottom-up) hierarchical clustering, the clusters are merged together according to some criteria. This method starts with an example in its own cluster and iteratively combines them to form larger and larger clusters. It assumes a similarity function for determining the similarity of two instances. It starts from the bottom of a tree, combining instances in separate cluster and then repeatedly joining the two clusters that are most similar until there is only one cluster. In single linkage clustering, the distance between two clusters is defined as the minimum distance from any member of one cluster to any member of the other cluster. So it is sometimes called a nearest-neighbor clustering algorithm [4]. In the single linkage algorithm the clustering process is terminated when the distance between the nearest clusters exceeds a particular threshold. An algorithm that uses the minimum distance measure is called a minimal spanning tree algorithm [4]. The problem with hierarchical cluster structure depends on the criterion of choosing the clusters to merge or spilt.

B. Support Vector Machine (SVM)

SVM is a classifier used for both linear and non linear data. The dataset is said to be linearly separable, if the patterns can be separated either by a straight line or by a hyper-plane. SVM classification also supports both binary and multiclass targets. This can be used for prediction as well as for classification. It uses a nonlinear mapping to transform the original training data into a higher dimension. Within this new dimension, it searches for the linear optimal separating hyper plane. The hyper plane indicating the "decision boundary" separates the tuples of one class from another. The SVM supports the maximum margin hyper plane (MMH) [4]. In the case of maximum marginal hyper plane, it gives the largest separation between the classes. The support vectors are the instances that are closest to the MMH. There must be atleast one support vector for each class. The greatest marginal hyper plane should have greater accuracy.

This paper is based on a reliable and scalable method called Clustering Based SVM. It uses the same idea of hierarchical clustering algorithm. CB-SVM effectively approximates an SVM classification function using a single scan of data given in a fixed size of memory. But it is not suited for high dimensional feature space. So CB-SVM needs effective indexing structures. To overcome this we introduce a new indexing method called Kernel Region Approximation Block.

The processing details are represented in figure 1.

section1. There is a large collection of documents. From the documents compute tf * idf, rank the documents according to the word frequency. The documents are then clustered using hierarchical clustering method.

section2. As per the user's requirements, SVM learns and classifies samples as positive or negative. The result is shown based on the query. The accuracy &precision is highly ensured in SVM.

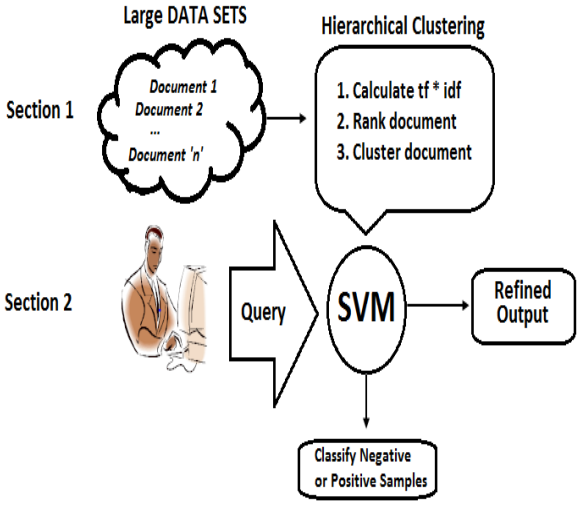

Figure 1: Users query processed by SVM.

C. Kernel Region Approximation Block

It is an efficient indexing method for high dimensional vector space using region approximation approach. It determines nonlinear relationship between features so that more accurate similar comparisons between vectors are supported. Kernel Region approximation Blocks is an extension of the improved RA-Blocks in the kernel space [6].It is designed to support kernel distance by building the KRA+- Blocks index structure in a reduced feature space. The kernel trick is an efficient method for nonlinear data analysis early used by SVM. This techniques use K-Nearest Neighbor search when using heterogeneous features. The region approximation blocks are differentiated according to

the different partitioning strategy used. The high dimensional indexing method has been proved to be quite useful for the improvement of the response time and in precision while dealing with high dimensional and heterogeneous vectors.

3. Text Classification

Document classification consists of data preprocessing such as term extraction, dimensionality reduction, feature-selection and so on. It includes training of data sets and creation of the classification model using the classification algorithm. Based on the algorithm it classifies the new documents. Using Term Frequency- Inverse Document Frequency (TF-IDF) we calculate the word weight [5]. It is a numerical statistic which reflects how important a word is in a collection of documents. In tf * idf calculation, word count C can be calculated using the formula:

$$C_{tf*idf} = tf * \log \frac{D}{df}$$

C - Word count
tf - term frequency
df- document frequency
D - Total no of documents

The number of times a word occurs in a document is termed as its term frequency. An inverse document factor (idf) which diminishes the weight of the terms that occur very frequently in the collection and increases the weight of the terms that occur rarely. It is often used as a weighting factor in information retrieval and text mining techniques. The search engines use the variations of the tf * idf weighting scheme as a major tool in scoring and ranking a document according to the user's query. The tf_idf can also be successfully used for stop-words filtering, in various subject fields like text classification and summarization.

4. Feature Selection by Hierarchical Clustering

Feature selection is very important in the case of feature space problems like text classification.
This hierarchical clustering method can be used to derive feature selection problem. It starts with all instances as separate cluster and then repeatedly joins the two clusters that are most similar until there is only one cluster. The algorithm is shown below:

1. Start with all instances in their own cluster.
2. Among the current clusters, determine the two clusters, c_i and c_j that are most similar.
3. Replace c_i and c_j with a single cluster $c_i \cup c_j$
4. Until there is only one cluster.

After getting the clusters, we can combine all the clusters using cosine similarity measure:

$$sim(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$$

The aggregation criteria is defined as:

$$\max \sum_t \sqrt{\sum_{u,v \in S_t} sim(u, v)}$$

S_t - The set of current clusters
u,v – Elements in one cluster.

The sum $\sum_{u,v \in S_t} sim(u, v)$ calculates the with-in cluster similarities on S_t. The similarity is computed for all cluster. And the clusters with high similarities are combined together.

5. Conclusion

This paper proposes a new method called CB-SVM that integrates a scalable clustering method with an SVM classifier and effectively runs SVMs for massive data sets. Today, the rapid growth of internet has made the web as a popular place for collecting large amount of information. With the search engines the internet user accesses billions of web pages online [7]. Using the learning behavior of SVM classifier we can improve the accuracy and precision in searching. The conventional indexing methods like SS-Tree, R-Tree, and K-D-B-Tree are more complex. So here we introduce an efficient indexing method called KRA+-Block. It supports heterogeneous feature vectors in high dimensional multimedia databases. HBSH clustering is more suitable for massive text clustering and uses hash table as its input data. The proposed system which is SVM based uses numerical vectors as its input data and proves to be more accurate. It has minor complexity than compared to the HBSH.

6. References

[1] Zamir O., Etzioni O.: Web Document Clustering: A Feasability Demonstration, *Proc. ACM SIGIR 98*, 1998,pp. 46-54.
[2] A. K. Jain, R. C. Dubes, Algorithms for Clustering Data,Prentice Hall,EnglewoodCliffs,NJ,1988.
[3] Hwanjo Yu, Jiong Yang, Jiawei Han "Classifying Large Data Sets Using SVMs with Hierarchical Clusters", *SIGKDD* '03. August 24-27. 2003.
[4] Jiawei Han, Micheline Kamber, "Data Mining Concepts and Techniques" Second Edition,Morgan Kaufmann Publishers,2006, Elsevier Inc.
[5] Yin Luo,Yan Fu, "A Hash-based Hierarchical Algorithm for Massive Text Clustering," *In Proceedings of the 2009 International Symposium on Web Information Systems and Applications(WISA'09)*, Nanchang, P.R.China, May 22-24,2009,pp.140-143.
[6] Daoudip, K.Idrissi, S.E.Ouatik, "Kernal Region Approximation Blocks for Indexing Heterogonous Databases".
[7] Rita. S. Shelke , Devendra Singh Thakore, "Cluster Based Web Search Using Support Vector Machine" International Journal of Engineering (IJE), Volume (5) : Issue (1) : 2011.
[8] Param Deep Singh, Jitendra Raghuvanshi, and Rising of Text Mining Technique: As Unforeseen-part of Data Mining, *ISSN: 2277 – 9043 International Journal of Advanced Research in Computer Science and Electronics Engineering Volume 1, Issue 3, May2012.*