Clustering Techniques with Dimensionality Reduction for Data Mining

Dr. J. Thirumaran

Dean, Computer Science
Rathinam College of Arts and Science, Coimbatore

Abstract: Mining Data from large volumes of high-dimensional data, such as global climate patterns, stellar spectra, or human gene distributions, Content Management are regularly confront the problem of dimensionality reduction. The human brain confronts the same problem in everyday perception, extracting from its high-dimensional sensory inputs of 30,000 neurons. Here we describe an approach Bayesian Personalize to solve dimensionality reduction problems that uses easily measured local metric information to learn the underlying global geometry of a data set. Unlike classical techniques such as Filter Algorithm or Wrapper algorithm which are being used for nonlinear dimensionality reduction, ours efficiently computes a globally optimal solution. Our algorithm uses Open Directory Project (ODP) Taxonomy as data set for classification of pattern's and Radial Basis function to cluster the pattern which in turn avoids the problem of selecting distance and number of cluster's in case of K-means Clustering algorithm. The proposed Bayesian classifier identifies the user interest efficiently with less time complexity. Our classifier is well efficient than existing classifiers like svm and Ripper.

Keywords: Data mining, neuron, dimension reduction, K-means clustering algorithm

Introduction

With the fast growth of the Web, users often suffer from the problem of information overload since many Existing search engines response lots of non-relevant documents containing query terms based on the search Mechanism of keyword matching. In fact, it is eagerly expected by both users and search engine developers to reduce overloaded information by understanding user goals clearly. The information explosion makes it hard for users to obtain required information from the web searched results in a more personalized way. For the same input word, most search engines returns the same result to each user without taking into consideration user preference. It is crucial to analyze user's search and browsing behaviors based on searching keywords entered by users. To this end we have proposed a method to derive user searching profiles.

In this paper, we intend to utilize Web search results to identify user goals. We propose one novel probabilistic inference model, which effectively employs features to discover user goals and hence their interests.

Need for Clustering in Data Mining

Data Mining (DM) or Knowledge Discovery in Databases (KDD) as it is also known, is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data [1]. This encompasses a number of different technical approaches, such as clustering, data summarization, learning classification rules, finding dependency networks, analyzing changes, and detecting anomalies. Clustering is the task of identifying groups in a data set based upon some criteria of similarity [2]. Clustering aims to discover sensible organization of objects in a given dataset by identifying and quantifying similarities or dissimilarities between the objects [3]. Clustering is applied in various fields, including data mining, statistical data analysis, compression and vector quantization. In data mining, clustering is used especially as preprocess to another data mining application. A variety of clustering formulation exists. Mostly used

clustering approach is *k*-means and is one of the best for implementing the clustering process. The *k*-means algorithm implementing for cluster analysis as data mining approach has been discussed on several researches [4, 5].

Problem Definition

On web search providing exact result to the user is the most important task. The previous existing models and search engines are lagging with providing personalization in an exact manner. They provide results according to the ranking algorithm which its using. Identifying the user interest and providing search result according to that is still a challenging task. We identified the problem of identifying the way of user interest prediction, where the data set or the visiting history is huge. When the dimensionality of the data increases in the web search, how we going to identify the user interest in an efficient manner.

Machine Learning

Machine learning is the subfield of artificial intelligence that is concerned with the design and development of algorithms that allow computers to improve their performance over time based on data, such as from sensor data or databases. A major focus of machine learning research is to automatically produce (induce) models, such as rules and patterns, from data. Hence, machine learning is closely related to fields such as data mining, statistics, inductive reasoning, pattern recognition, and theoretical computer science.

Pattern Recognition

Pattern recognition is the process wanting to estimate the characteristics of an unknown model can rely only on the observable state. Pattern recognition is a subtopic of machine learning. It is "the act of taking in raw data and taking an action based on the category of the data. Most research in pattern recognition is about methods for supervised learning and unsupervised learning. Pattern recognition aims to classify data (patterns) based either on a-priori knowledge or on statistical information extracted from the patterns. The patterns to be classified are usually groups of measurements or observations, defining points in an appropriate multidimensional space. This is in contrast to pattern matching, where the pattern is rigidly specified.

Dimensionality Reduction

In statistics, dimension reduction is the process of reducing the number of random variables under consideration, and can be divided into feature selection and feature extraction. In physics, dimension reduction is a widely discussed phenomenon, whereby a physical system exists in three dimensions, but its properties behave like those of a lower-dimensional system.

The existing research on pattern classification has been mainly focused on devising new model development algorithms in order to improve generalizing performance. Among these algorithmic advances, support vector machines (SVMs) have been shown to be one of the most important developments, providing a sound methodological basis for constructing classification models with high generalizing ability. SVMs implement the structural risk minimization principle in order to build large-margin classifiers, i.e., models that maximize the class-separating margin, which is closely related to generalizing performance. SVM models are expressed in linear and nonlinear form. The construction, however, of nonlinear models requires increased computational resources for large data sets, and their interpretability/transparency is rather limited. This paper proposes a methodology to construct additive models using the SVM framework.

Radial Basis Function Networks

Radial Basis Functions (RBF) represents alternative approach to Multi-Layer Perceptrons (MLP) in universal function approximation. RBF's was first used in solving multivariate interpolation problems and numerical anaysis. Their prospect is similar in neural network applications, where the training and query targets are rather continuous. While MLP performs a global mapping (i.e., all inputs cause an output). RBF network performs a local mapping (i.e., only inputs near specific receptive fields will produce an activation). The units (in the hidden layer) receiving the direct input from a signal may see only a portion of the input pattern, which is further used in reconstructing a surface in a multidimensional space that furnishes the best fit to the training data. This ability of the RBF network to recognize whether an input is near the training set or outside the trained region provides a significant benefit over MLP's.

Implementation

We have implemented this paper for Document mining in the area of Data mining. Implementation contains the following steps:

1. Feature Extraction.
2. Feature Selection.
3. Categorization.
4. Clustering.
5. Dimensionality Reduction using Bayesian Personalizer Method.

Results and Discussion

In most existing text clustering algorithms, text documents are represented by using the vector space model. In this model, each document d is considered as a vector in the term-space and represented by the term-frequency (TF) vector: dtf = [tf1, tf2. . . tfh] . Where tfi is the frequency of the ith term in the document, and h is the dimension of the text database, which is the total number of unique terms. The cosine function measures the similarity between two documents as the correlation between the documents Vectors representing them. For two documents di and dj , the similarity between them can be calculated as: where • represents the vector dot product and denotes the length of vector di. The cosine value is 1 when two documents are identical, and 0 if there is nothing in common between them. The larger cosine value indicates that these two documents share more terms and are more similar.

The k-means algorithm is very popular for solving the problem of clustering a data set into k clusters. If the data set contains n documents, d1, d2, . . . , dn, then the clustering is the optimization process of grouping them into k clusters so that the global criterion function.

References

1. Hawley, W.J, Piatetsky-Shapiro, G., Mathews, C., J, "Knowledge Discovery in Databases: An Overview", AI Magazine, 13(3): 57-70, 1992
2. Dorigo, M., Maniezzo, V., Colorni, A., "The Ant System : Optimization by a colony of cooperating agents", IEEE Transactions on Systems, Man, and Cybernetics-Part B,Vol.26, No.1, pp.1-13, 1996
3. Shelokar, V.K., Jayaraman, V.K., Kulkarni, B.D., "An Ant Colony Approach for Clustering", Analytica Chimica Acta 509, 187–195, 2004
4. Kuo, R.J., Liao, J.L., TU, C., "Integration of ART2 neural network and genetic *K*-means algorithm for analyzing Web browsing paths in electronic commerce", Decision Support Systems 40, pp. 355–374,2005

5. Vrahatis, M.N., Boutsinas, B., Alevizos, P., Pavlides, G., "The new *k*-windows algorithm for improving the *k*-means clustering algorithm", Journal of Complexity 18, pp. 375–391,2002
6. Maurice Karnaugh, The Map for Synthesis of Combinational Logic Circuits, AIEE, Nov. 1953.
7. R. A. Fisher. "The use of multiple measurements in taxonomic problems", Eugen., 7, 1936.
8. T. Hastie, R. Tibshirani, and J. Friedman. "The elements of Statistical Learning - Data Mining, Inference and Prediction", Springer, 2001.
9. S. Lloyd, "Least squares quantization in pcm", Technical report, Bell Laboratories, 1957.