International Journal Of Core Engineering & Management

Volume-5, Issue-12, March-2019, ISSN No: 2348-9510

MAP REDUCED SERVICE RECOMMENDATION MABFAST FOR BIGDATA

M.H.Mohamed Aashik
Tamil Nadu

Abstract

Feature selection process is the vital one in the architecture of data retrieval process in web mining. It involves identifying a subset of the most useful features that produces compatible results as the original entire set of features. A feature selection algorithm may be evaluated from both the efficiency and effectiveness points of view. While the efficiency concerns the time required to find the multiple attribute based feature selection, the effectiveness is related to the quality of the mechanism designed to perform the feature selection.

Based on the proposed idea, Mulit-attribute based fast clustering-based feature selection algorithm (MABFAST) is proposed and going to experiments with different parameter set. The MABFAST algorithm works in three steps. In the first step, features are divided into clusters by using graph-theoretic clustering methods. In the second step, the most representative feature that is strongly related to target cluster classes is selected from each cluster to form an attribute based classes. Features in different clusters are relatively either dependent or independent, the clustering based strategy of MABFAST has a high probability of producing a subset of useful and independent features.

Keywords : Clustering,Relevance,MABFAST,hadoop,Mapreduce

I. INTRODUCTION

The feature subset selection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility. Many feature subset selection methods have been proposed and studied for machine learning applications. They can be divided into four broad categories: the Embedded, Wrapper, Filter, and Hybrid approaches. The embedded methods incorporate feature selection as a part of the training process and are usually specific to given learning algorithms, and therefore may be more efficient than the other three categories.

Traditional machine learning algorithms like decision trees or artificial neural networks are examples of embedded approaches. The wrapper methods use the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subsets, the accuracy of the learning algorithms is usually high. However, the generality of the selected features is limited and the computational complexity is large. The filter methods are independent of learning algorithms, with good generality. Their computational complexity is low, but the accuracy of the learning algorithms is not guaranteed. The hybrid methods are a combination of filter and wrapper methods by using a filter method to reduce search space that will be considered by the subsequent wrapper.

It mainly focuses on combining filter and wrapper methods to achieve the best possible performance with a particular learning algorithm with similar time complexity of the filter methods. The wrapper methods are computationally expensive and tend to over fit on small training sets. The filter methods, in addition to their generality, are usually a good choice when the number of features is very large. Thus, we will focus on the filter method in this paper. With respect to the filter feature selection methods, the application of cluster analysis has been demonstrated to be more effective than traditional feature selection algorithms are applied in the distributional clustering of words to reduce the dimensionality of text data. In cluster analysis, graph theoretic methods have been well studied and used in many applications. The results have, sometimes, the best agreement with human performance. The general graph theoretic clustering is simple: Compute a neighborhood graph of instances, then delete any edge in the graph that is much longer/shorter (according to some criterion) than its neighbors. The result is a forest and each tree in the forest represents a cluster. In this study, apply graph theoretic clustering methods to features.

In particular, it adopts the minimum spanning tree (MST) based clustering algorithms, because it do not assume that data points are grouped around centers or separated by a regular geometric curve and have been widely used in practice. Based on the MST method, this propose a attribute based Fast clustering based feature Selection algorithm (MABFAST).The MABFAST algorithm works in two steps. In the first step, features are divided into clusters by using feedback verification clustering methods. In the second step, the most representative feature that is strongly related to target classes is selected from each cluster to form the final subset of features.

Features in different clusters are relatively independent, the clustering based strategy of MABFAST has a high probability of producing a subset of useful and independent features. The proposed feature subset selection algorithm MABFAST was tested upon 35 publicly available image, microarray, and text data sets. The experimental results show that, compared with other five different types of feature subset selection algorithms, the proposed algorithm not only reduces the number of features, but also improves the performances of the four well known different types of classifiers.

II. LITERATURE REVIEW

This chapter explores closely related literature and the placement of this dissertation research in the areas of the selection of scholarly materials, data mining techniques and software agents.

Data Collection

Finding scholarly information on the World Wide Web can be very frustrating. There is no way to search through a large selection of only scholarly sites with the current Web search tools. The existing search tools provide search algorithms that sift through millions of Web pages with no way to limit the search to a category of Web sites. Nobody seems to know how to do any automatic filtering for quality of Web sites[1]. However, librarians have been doing quality filtering of materials for many years, but “no one seems conscious of the standards carefully developed by information professionals over the past century” (Collins 1996, 122).

In the print world, the academic library performs this filtering function by providing patrons with a subset of print works pertaining to academia. This selection role is filled by library staff members using either explicit or tacit criteria to select individual works. Some sites, such as the Internet Public Library (http://www.ipl.org), attempt to select scholarly sites. However, because of the rapid introduction of new documents on the World Wide Web, a human cannot keep up and the resource is quickly outdated.

In order to handle the vast number of documents on the Web, an automated selection system is needed. First, the criteria used by academic librarians to select print works will be examined. These criteria can be translated into equivalent criteria for Web pages. A Web robot can then be designed to determine these criteria for a page. After creating a training set of examined Web pages with their selection decisions, data mining techniques can be used to create a classification model that will be a quality filter for Web pages.

Most of the existing works are motivated by a commonly performed task in the biomedical domain, that of constructing a systematic review. Authors of systematic reviews seek to identify as much as possible of the relevant literature in connection with some aspect of medical practice, typically a highly specific clinical question. The review's authors assess, select, and synthesize the evidence contained in a set of identified documents, to provide a "best currently known" summary of knowledge and practice in that field.

A variety of organizations provide central points of call for systematic reviews, including the Cochrane Collaboration,2 the largest of these efforts, and the Agency for Healthcare Research and Quality, AHRQ.3 The collections used as the source material are already large, and continue to grow. For example, as at end of 2009[2], MEDLINE, the largest of the available collections, contained more than 19 million entries, with more than 700,000 citations having been added during the year. To construct each systematic review, a complex Boolean search query is used to retrieve a set of possibly relevant documents (typically in the order of one to three thousand), which are then comprehensively triaged by multiple assessors.

Recently, hierarchical clustering has been adopted in word selection in the context of text classification. Distributional clustering has been used to cluster words into groups based either on their participation in particular grammatical relations with other words by Pereira or on the distribution of class labels associated with each word by Baker and McCallum[3][5]. As distributional clustering of words is agglomerative in nature, and result in sub-optimal word clusters and high computational cost, it shows a new information-theoretic divisive algorithm for word clustering and applied it to text classification. It proposed to cluster features using a special metric of Barthelemy distance[8][9], and then makes use of the dendrogram of the resulting cluster hierarchy to choose the most relevant attributes. Unfortunately, the cluster evaluation measure based on Barthelemy distance does not identify a feature subset that allows the classifiers to improve their original performance accuracy. Furthermore, even compared with other feature selection methods, the obtained accuracy is lower.

The Boolean query might consist of as many as several dozen query lines, each describing a concept with fielded keywords, MESH headings[10][11] (a hierarchical taxonomy of medical terms), metadata, free-text term expansions, and Boolean operators aggregating and balancing the concepts shows the structure of one such query; this expression would be one small component of a typical complex query of (say) 50 clauses.

Proposed Methodology

Feature selection process is the vital one in the architecture of data retrieval process in web mining. It involves identifying a subset of the most useful features that produces compatible results as the original entire set of features. A feature selection algorithm may be evaluated from both the efficiency and effectiveness points of view. While the efficiency concerns the time required to find the multiple attribute based feature selection, the effectiveness is related to the quality of the mechanism designed to perform the feature selection. Based on the proposed idea, attribute based fast clustering-based feature selection algorithm (MABFAST) is proposed and going to experiments with different parameter set.

Similarity cluster identification in Mining Servers

In this module we extract the similar document from the data set based on the given Boolean query. The similar document is extracted using TF-IDF values. Compute the similarity score for the given query and the data set. Get the highest similarity score document.

Score Computation using MST

In this module shows compute the score for each document in a data set from various database servers. The recursive nature of ABFAST queries makes it necessary to calculate the scores on lower levels in the query tree first. One obvious possibility would be to try and add processing logic to each query node as it acts on its clauses. But optimizations such as max-score could only be employed at the query root node, as a threshold is only available for the overall query score. Instead, It follow a holistic approach and prefer to be able to calculate the document score given a set of query terms S $\subseteq$ T present in a document, no matter where they appear in the query tree.

Ranking Cluster s

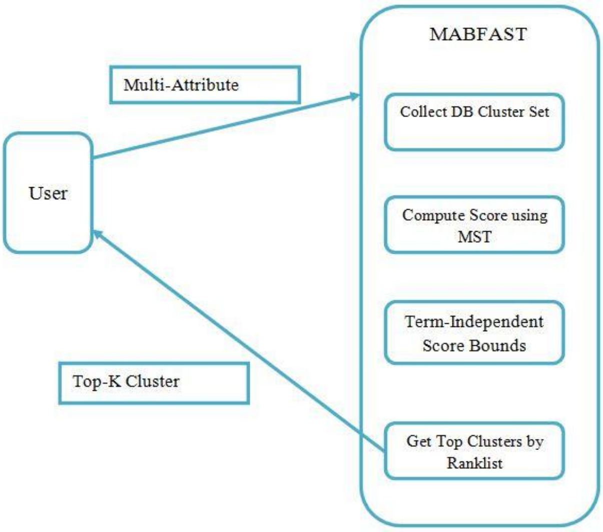
To provide early termination of document scoring, It also propose the use of term independent score bounds that represent the maximum attainable score for a given number of terms. A lookup table of score bounds M_r is created, indexed by r that is consulted to check if it is possible for a candidate document containing r of the terms to achieve a score greater than the current entry threshold. That is, for each r = 1. . . n , we seek to determine

1. $$M_r = \max\{CalcScore(S, B) \mid S \subseteq T \text{ and } |S| = r\}$$

The number of possible term combinations that could occur in documents is $O(\binom{n}{r})$ for each r, which is $O(2^n)$ in total, and a demanding computation. However, the scoring functions only depend on the clause scores, that is, the overall score of each sub-tree, meaning that the problem can be broken down into subparts, solved for each sub-tree separately, and then aggregated. The simplest (sub) tree consists of one term, for which the solution is trivial. For a particular operator node with clauses C, let n_c denote the number of terms in the sub-tree of clause c ε C. A table with $r = 0, \ldots, \sum_{c \in C} n_c$ possible terms present is then computed; and to compute each M_r, all possibilities to decompose r into a sum over the clauses r ¼ P c2C rc $r = \sum_{c \in C} r_c$ have to be considered.

Top Cluster using MABFAST

In this module it gets the top k document servers for the give correlated clustering query. Our objective is to construct a query sequence q1, q2... qv of ABFAST return data queries that can be submitted to the database, retrieve as few data as possible, and still contain all the documents that would be in the top-k results.



III. CONCLUSION

This proposed technique MABFAST had been successfully done the novel clustering based feature subset selection algorithm for high dimensional data. In the proposed algorithm, a cluster consists of features. Each cluster is treated as a single feature and thus dimensionality is drastically reduced. It also have compared the performance of the proposed algorithm with those of the five well known feature selection algorithms FCBF, CFD, Fuzzy sets based clustering on the 35 publicly available image, microarray, and text data from the four different aspects of the proportion of selected features, runtime, classification accuracy of a given classifier, and the Win/Draw/Loss record. Generally, the proposed algorithm obtained the best proportion of selected features comparing to the existing systems.

IV. FUTURE ENHANCEMENT

In future this model for feature selection and ranking from the high dimensional database systems will have been implemented and tested with the different set of parameters. From the analysis above we can know that FAST performs very well on the microarray data. The reason lies in both the characteristics of the data set itself and the property of the proposed algorithm. For the purpose of exploring the relationship between feature selection algorithms with high intensity of data volume, in which algorithms are more suitable for which types of data, it ranks the six feature selection algorithms according to the classification accuracy of a given classifier on a specific type of data after the feature selection algorithms are performed.

REFERENCES

[1] M. E. Newman and M. Girvan, “Finding and evaluating community structure in networks,” Physical Review E, vol. 69, no. 2, p. 026113, 2004.

[2] J. Dean and S. Ghemawat, “MapReduce: simplified data processing on large clusters,” Communications of the ACM, vol. 51, no. 1, pp. 107–113, 2008.

[3] R. Chaiken, B. Jenkins, P.-A° . Larson, B. Ramsey, D. Shakib, S. Weaver, and J. Zhou, “Scope: easy and efficient parallel processing of massive data sets,” Proceedings of the VLDB Endowment, vol. 1, no. 2, pp.1265–1276, 2008.

[4] J. Cohen, B. Dolan, M. Dunlap, J. M. Hellerstein, and C. Welton, “Mad skills: new analysis practices for big data,” Proceedings of the VLDB Endowment, vol. 2, no. 2, pp. 1481–1492, 2009.

[5] A. F. Gates, O. Natkovich, S. Chopra, P. Kamath, S. M. Narayanamurthy, C. Olston, B. Reed, S. Srinivasan, and U. Srivastava, “Building a high-level dataflow system on top of Map-Reduce: the Pig experience,” Proceedings of the VLDB Endowment, vol. 2, no. 2, pp. 1414–1425,2009.

[6] H.-c. Yang, A. Dasdan, R.-L. Hsiao, and D. S. Parker, “Map-Reduce-Merge: simplified relational data processing on large clusters,” in Proceedings of 2007 ACM SIGMOD International Conference on Management of Data, 2007, pp. 1029–1040.

[7] “Apache Hadoop,” http://hadoop.apache.org/, accessed:August 1, 2013.

[8] Z. F. Zeng, B. Wu, and T. T. Zhang, “A multi-source message passing model to improve the parallelism efficiency of graph mining on MapReduce,” in Proceedings of 2012 IEEE International Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW), 2012,pp. 2019–2025.

[9] “Stanford large network dataset collection,” http://snap. stanford.edu/data/, accessed: August 1, 2013.

[10] J. Han, M. Kamber, and J. Pei, Data mining: concepts and techniques, 2nd ed. Morgan Kaufmann, 2006.

[11] Chikhi S. and Benhammada S., ReliefMSS: a variation on a feature ranking ReliefF algorithm. Int. J. Bus. Intell. Data Min. 4(3/4), pp 375-390, 2009.

[12] Cohen W., Fast Effective Rule Induction, In Proc. 12th international Conf.Machine Learning (ICML’95), pp 115-123, 1995.

[13] Dash M. and Liu H., Feature Selection for Classification, Intelligent Data Analysis, 1(3), pp 131-156, 1997.

[14] Dash M., Liu H. and Motoda H., Consistency based feature Selection, In Proceedings of the Fourth Pacific Asia Conference on Knowledge Discovery and Data Mining, pp 98-109, 2000.

[15] Das S., Filters, wrappers and a boosting-based hybrid for feature Selection, In Proceedings of the Eighteenth International Conference on Machine Learning, pp 74-81, 2001.

[16] Dash M. and Liu H., Consistency-based search in feature selection. Artificial Intelligence, 151(1-2), pp 155-176, 2003.

[17] Demsar J., Statistical comparison of classifiers over multiple data sets, J.Mach. Learn. Res., 7, pp 1-30, 2006.
[18] Dhillon I.S., Mallela S. and Kumar R., A divisive information theoretic feature clustering algorithm for text classification, J. Mach. Learn. Res., 3,pp 1265-1287, 2003.
[19] Dougherty, E. R., Small sample issues for microarray-based classification. Comparative and Functional Genomics, 2(1), pp 28-34, 2001.
[20] Fayyad U. and Irani K., Multi-interval discretization of continuous-valued attributes for classification learning, In Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence, pp 1022-1027, 1993.
[21] Fisher D.H., Xu L. and Zard N., Ordering Effects in Clustering, In Proceedings of the Ninth international Workshop on Machine Learning, pp 162-168, 1992.
[22] Fleuret F., Fast binary feature selection with conditional mutual Information,Journal of Machine Learning Research, 5, pp 1531-1555, 2004.
[23] Forman G., An extensive empirical study of feature selection metrics for text classification, Journal of Machine Learning Research, 3, pp 1289-1305,2003.
[24] Friedman M., A comparison of alternative tests of significance for the problem of m ranking, Ann. Math. Statist., 11, pp 86-92, 1940.
[25] Garcia S and Herrera F., An extension on “Statistical Comparisons of Classifiers over Multiple Data Sets” for all pairwise comparisons, J. Mach. Learn. Res., 9, pp 2677-2694, 2008.
[26] Garey M.R. and Johnson D.S., Computers and Intractability: a Guide to the Theory of Np-Completeness. W. H. Freeman & Co, 1979.
[27] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov,J.P., Coller, H., Loh, M.L., Downing, J.R., and Caligiuri, M. A., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 286(5439), pp 531-537, 1999.
[28] Guyon I. and Elisseeff A., An introduction to variable and feature selection,Journal of Machine Learning Research, 3, pp 1157-1182, 2003.
[29] Hall M.A., Correlation-Based Feature Subset Selection for Machine Learning,Ph.D. dissertation Waikato, New Zealand: Univ. Waikato, 1999.
[30] Hall M.A. and Smith L.A., Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper, In Proceedings of the Twelfth international Florida Artificial intelligence Research Society Conference, pp 235-239, 1999.
[31] Hall M.A., Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning, In Proceedings of 17th International Conference on Machine Learning, pp 359-366, 2000.