Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 12, December 2014, pg.50 – 55

RESEARCH ARTICLE

A Framework for Mining High Dimensional Data for Feature Subset Selection

Venu Pullela[1], V.Santosh Kumar[2], Ch.Ravindranath Yadav[3]

Student, Department of CSE, Sreyas Institute of Engineering and Technology, Hyderabad, India[1]
Associate Professor, Department of CSE, Sreyas Institute of Engineering and Technology, Hyderabad, India[2]
Asst.Professor, Dept. of CSE, Keshav Memorial Institute of Technology, Hyderabad, India[3]

Pullelavenu558@gmail.com [1], Vennu.santosh@gmail.com [2], ravi_chintala@hotmail.com [3]

Abstract-- Features are representative characteristics of data sets. Identifying such fetures in a high dimensional dataset play an important role in real world applications. Data mining is best used to determine important features. Selecting important features from a subject of identified features can help in making expert decisions. However, efficient identification of such feature subset and selection is a challenging problem. Recently Song et al. proposed a solution that is capable of selecting subset of features with good quality. They used clustering approach before selecting representative features for final selection. Similar work is carried out in this paper which demonstrates the proof of concept. The proposed solution makes use of clustering for achieving the goal of the system. The empirical results reveal that the application is useful. The results are compared with many existing algorithms like C4.5, Naïve Bayes, IB1 and RIPPER.

Index Terms – Data mining, feature subset selection, clustering

I. INTRODUCTION

Feature subset selection with respect to the target concepts, feature subset selection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility [1], [2]. 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 [3]. Traditional machine learning algorithms like decision trees or artificial neural networks are examples of embedded approaches [4]. 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.

The wrapper methods are computationally expensive and tend to overfit on small training sets [5], [6]. 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. Pereira et al. [7], Baker and McCallum [8], and Dhillon et al. [9] employed 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. Their results have, sometimes, the best agreement with human performance [10]. 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 our study, we apply graph-theoretic clustering methods to features. In particular, we adopt the minimum spanning tree (MST)-based clustering algorithms, because they 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, we propose a Proposed clustering based feature Selection algorithm.

The remainder of this paper is structured as follows. Section II provides review of literature. Section III presents proposed work. Section IV focuses on the prototype application and experimental results while section V concludes the paper.

II. RELATED WORK

This section provides review of literature on prior works. A very early mention of power exhaustion can be found in [11], as “sleep deprivation torture.” As per the name, the proposed attack prevents nodes from entering a low-power sleep cycle, and thus depletes their batteries Proposed. Newer research on “denial-of-sleep” only considers attacks at the MAC layer [12]. Additional work mentions resource exhaustion at the MAC and transport layers [13], [8] but only offers rate limiting and elimination of insider adversaries as potential solutions. Malicious cycles (routing loops) have been briefly mentioned [14], [15], but no effective defenses are discussed other than increasing efficiency of the underlying MAC and routing protocols or switching away from source routing.

Even in non-power-constrained systems, depletion of resources such as memory, CPU time, and bandwidth may easily cause problems. A popular example is the SYN flood attack, wherein adversaries make multiple connection requests to a server, which will allocate resources for each connection request, eventually running out of resources, while the adversary, who allocates minimal resources, remains operational (since he does not intend to ever complete the connection handshake). Such attacks can be defeated or attenuated by putting greater burden on the connecting entity (e.g., SYN cookies [16], which offload the initial connection state onto the client, or cryptographic puzzles [17], [18], and [19]). These solutions place minimal load on legitimate clients who only initiate a small number of connections, but deter malicious entities who will attempt a large number. Note that this is actually a form of rate limiting and not always desirable as it punishes nodes that produce bursty traffic but may not send much total data over the lifetime of the network. Since Vampire attacks rely on amplification, such solutions may not be sufficiently effective to justify the excess load on legitimate nodes.

Other work on denial of service in ad hoc wireless networks has primarily dealt with adversaries who prevent route setup, disrupt communication, or preferentially establish routes through themselves to drop, manipulate, or monitor packets [20], [3], [21], [22], and [23]. The effect of denial or degradation of service on battery life and other finite node resources has not generally been a security consideration, making our work tangential to the research mentioned above. Protocols that define security in terms of path discovery success, ensuring that only valid network paths are found, cannot protect against Vampire attacks, since Vampires do not use or return illegal routes or prevent communication in the short term.

III. PROPOSED SOLUTION

The proposed solution is based on the distributed clustering approach that is meant for feature selection effectively and efficiently. The feature subset selection from high dimensional data involves removal of irrelevant features, construction of minimum spanning tree, partitioning tree, and selection of representative features and finally the selected features are used in further processing. In the process of feature selection there is an important consideration for elimination of redundant features. The proposed distributed clustering has many activities such as subset selection algorithm [24], time complexity, and text search and output representations.

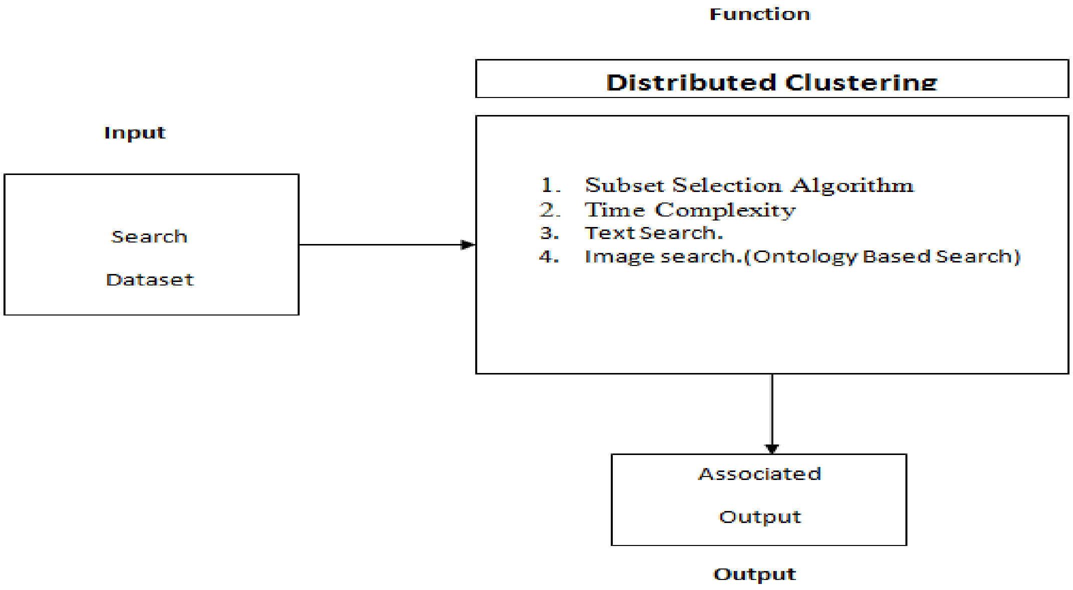
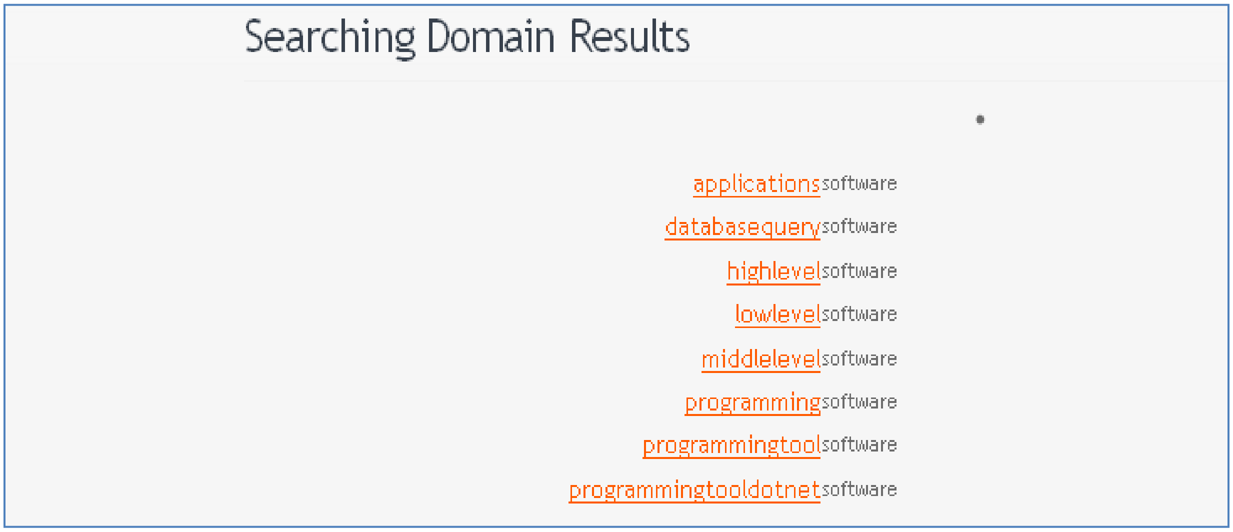


Figure 2 – Overview of the proposed solution

As can be seen in Figure 2, it is evident that the proposed solution takes high dimensional dataset as input and performs feature subset selection which gets subset of features that are representatives of all clusters. This will improve the performance of selection process. Once features are selected, further processing can be possible that is based on the application requirements.

IV. PROTOTYPE IMPLEMENTATION AND RESULTS

We built a prototype application that demonstrates the proof of concept. The application is built using Java technologies like Servlets and JSP. JDBC is used for interacting with dataset. The prototype is able to perform the intended operations by taking high dimensional datasets as inputs. One of the sample screens of the prototype application is presented in Figure 3.


Searching Domain Results

applicationssoftware

databasequerysoftware

highlevelsoftware

lowlevelsoftware

middlelevelsoftware

programmingsoftware

programmingtoolsoftware

programmingtooldotnetsoftware

Figure 3 - Searching Domain Results

As can be seen in Figure 3, it is evident that the prototype application is able to produce various fields as subset of features using the clustering approach. The clusters are represented by the underlying features of the dataset.

V. EXPERIMENTAL RESULTS

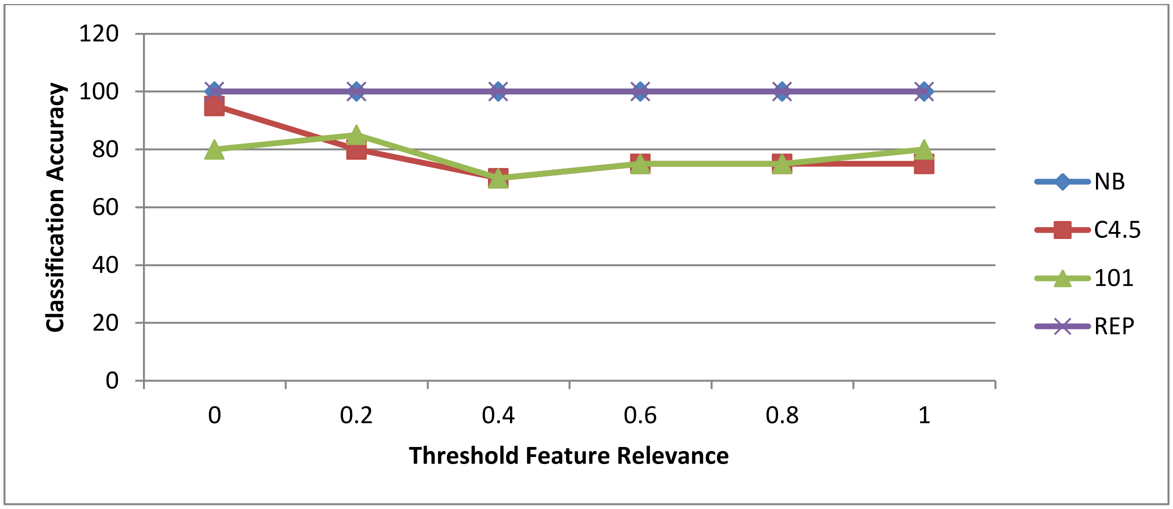


Figure 4 - Accuracies of the four classification algorithms with different θ values

As can be seen in Figure 4, it is evident that the experiment results reveal that the performance of the proposed solution is better than the other algorithms.

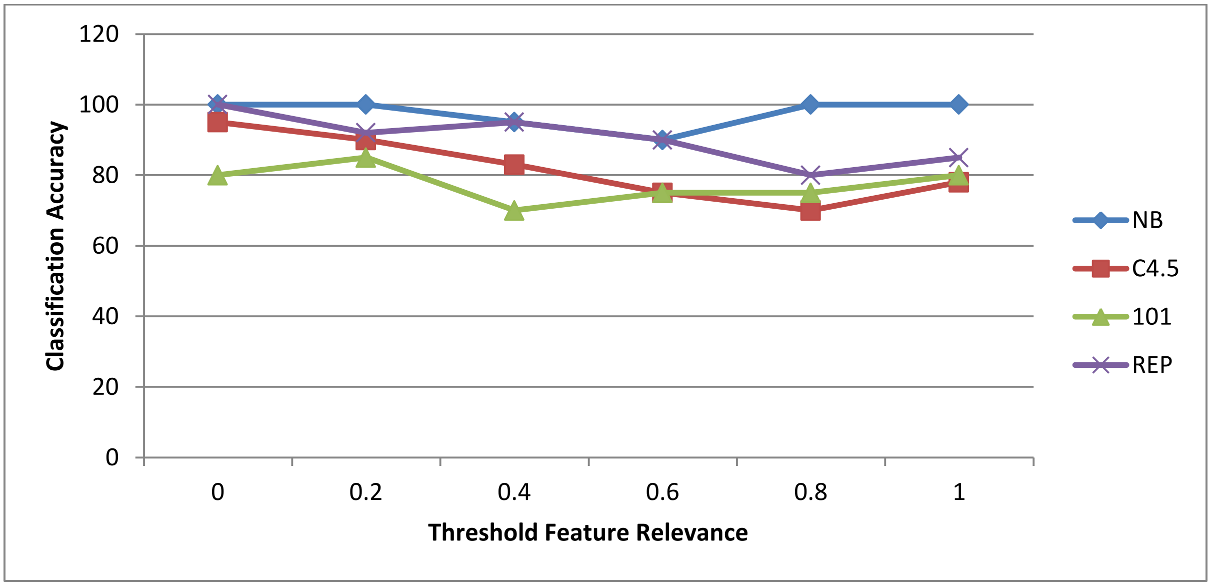


Figure 6 - Accuracies of the four classification algorithms with different θ values

As can be seen in Figure 6, it is evident that the experiment results reveal that the performance of the proposed solution is better than the other algorithms

VI. CONCLUSIONS AND FUTURE WORK

In this paper we studied the problem of feature subset selection for high dimensional data. A Proposed clustering based approach was proposed recently by Song et al. [50] which has two steps in the processing. In the first step, features are clustered and in the second step the most representative features are taken out in order to have decision making and further use in data mining purposes. High dimensional data is used as input and clustering is performed to make number of clusters. Each cluster can have many features and the best features that are representative of the features are selected as subset of features. This is very important as it has many real time utilities in the real world. The selected features can help in finding new avenues for further data mining purposes. It involves in the generation of such feature subset so as to be very useful for many applications in the enterprises. We built a prototype application that demonstrates the proof of concept. The empirical results are encouraging. An important direction for future work is building a model that will have generalized features which can help to adapt to various high dimensional datasets of different domains.

REFERENCES

[1] H. Liu, H. Motoda, and L. Yu, “Selective Sampling Approach to Active Feature Selection,” Artificial Intelligence, vol. 159, nos. 1/2, pp. 49-74, 2004.

[2] L.C. Molina, L. Belanche, and A. Nebot, “Feature Selection Algorithms: A Survey and Experimental Evaluation,” Proc. IEEE Int’l Conf. Data Mining, pp. 306-313, 2002.

[3] I. Guyon and A. Elisseeff, “An Introduction to Variable and Feature Selection,” J. Machine Learning Research, vol 3, pp. 1157-1182, 2003.

[4] T.M. Mitchell, “Generalization as Search,” Artificial Intelligence, vol. 18, no. 2, pp. 203-226, 1982.

[5] M. Dash and H. Liu, “Feature Selection for Classification,” Intelligent Data Analysis, vol. 1, no. 3, pp. 131-156, 1997.

[6] S. Das, “Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection,” Proc. 18th Int’l Conf. Machine Learning, pp. 74- 81, 2001.

[7] F. Pereira, N. Tishby, and L. Lee, “Distributional Clustering of English Words,” Proc. 31st Ann. Meeting on Assoc. for Computational Linguistics, pp. 183-190, 1993.

[8] L.D. Baker and A.K. McCallum, “Distributional Clustering of Words for Text Classification,” Proc. 21st Ann. Int’l ACM SIGIR Conf. Research and evelopment in information Retrieval, pp. 96-103, 1998.

[9] I.S. Dhillon, S. Mallela, and R. Kumar, “A Divisive Information Theoretic Feature Clustering Algorithm for Text Classification,” J. Machine Learning Research, vol. 3, pp. 1265-1287, 2003.

[10] J.W. Jaromczyk and G.T. Toussaint, “Relative Neighborhood Graphs and Their Relatives,” Proc. IEEE, vol. 80, no. 9, pp. 1502-1517, Sept. 1992.

[11] H. Almuallim and T.G. Dietterich, “Algorithms for Identifying Relevant Features,” Proc. Ninth Canadian Conf. Artificial Intelligence, pp. 38-45, 1992.

[12] H. Almuallim and T.G. Dietterich, “Learning Boolean Concepts in the Presence of Many Irrelevant Features,” Artificial Intelligence, vol. 69, nos. 1/2, pp. 279-305, 1994.

[13] A. Arauzo-Azofra, J.M. Benitez, and J.L. Castro, “A Feature Set Measure Based on Relief,” Proc. Fifth Int’l Conf. Recent Advances in Soft Computing, pp. 104-109, 2004.

[14] R. Battiti, “Using Mutual Information for Selecting Features in Supervised Neural Net Learning,” IEEE Trans. Neural Networks, vol. 5, no. 4, pp. 537-550, July 1994.

[15] D.A. Bell and H. Wang, “A Formalism for Relevance and Its Application in Feature Subset Selection,” Machine Learning, vol. 41, no. 2, pp. 175-195, 2000.

[16] J. Biesiada and W. Duch, “Features Election for High-Dimensional data a Pearson Redundancy Based Filter,” Advances in Soft Computing, vol. 45, pp. 242-249, 2008.

[17] R. Butterworth, G. Piatetsky-Shapiro, and D.A. Simovici, “On Feature Selection through Clustering,” Proc. IEEE Fifth Int’l Conf. Data Mining, pp. 581-584, 2005.

[18] C. Cardie, “Using Decision Trees to Improve Case-Based Learning,” Proc. 10th Int’l Conf. Machine Learning, pp. 25-32, 1993.

[19] P. Chanda, Y. Cho, A. Zhang, and M. Ramanathan, “Mining of Attribute Interactions Using Information Theoretic Metrics,” Proc. IEEE Int’l Conf. Data Mining Workshops, pp. 350-355, 2009.

[20] M. Dash, H. Liu, and H. Motoda, “Consistency Based Feature Selection,” Proc. Fourth Pacific Asia Conf. Knowledge Discovery and Data Mining, pp. 98-109, 2000.

[21] M.A. Hall, “Correlation-Based Feature Subset Selection for Machine Learning,” PhD dissertation, Univ. of Waikato, 1999.

[22] D. Koller and M. Sahami, “Toward Optimal Feature Selection,” Proc. Int’l Conf. Machine Learning, pp. 284-292, 1996.

[23] J. Demsar, “Statistical Comparison of Classifiers over Multiple Data Sets,” J. Machine Learning Res., vol. 7, pp. 1-30, 2006.

[24] Qinbao Song, Jingjie Ni and Guangtao Wang. (2013). A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data.*IEEE*. 25 (1), p1-14.

AUTHORS

Venu Pullela received the BE degree in computer science and technology from Jawaharlal Nehru Technological University, Hyderabad, India in 2010. He is currently working towards his M.Tech degree in Sreyas Institute of Engineering and Technology, Hyderabad, India. His research interests include data mining.

Vennu Santosh Kumar received the Masters degree in Computer Science and Engineering in the year 2010. He is Microsoft Certified System Engineer & CISCO Certified Network Administrator, he worked as a System Engineer in WIPRO Technologies(INDIA). In 2011 he joined as an Associate Professor at Sreyas Institute of Engineering and Technology in Computer Science Department. He has been involved in several tutorials, workshops, technical paper presentations .His research interests are focused on Computer Networks, Network Security & Mobile Computing.

Chintala Ravindranath Yadav is a perceptive academician, with an M.B.A to his credit, and pursued his Post-masters in Business Administration from University of North Carolina at Greensboro, U.S.A. He worked in leading edge organizations for several years in U.S.