International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Towards Identifying Feature Subset Selection for Mining High Dimensional Data

Punnana Sarath Kumar[1], Ganiya Rajendra Kumar[2]

[1]M.Tech (Software Engineering),Department of CSE, Sri Sivani College of Engineering, Chilakapalem, Etcherla Mandal, Srikakulam District, A.P, India

[2]M.Tech.,(Ph.D),Professor & HOD of CSE Department, Department of CSE, Sri Sivani College of Engineering, Chilakapalem, Etcherla Mandal, Srikakulam District, A.P, India

Abstract: ***High dimensional data is the data which has many features. Some features might have representative characteristics that can help in reducing search space in data mining activities. Therefore it is important to identify such features. The feature subset selection can improve the performance of data mining on high dimensional data. This will help in extracting business intelligence that can help in making expert decisions. However, it is challenging task to identify feature subset that is representative of all possible characteristics. Song et al., of late, proposed a framework that can be used to select feature subset from high dimensional data. Clustering is involved in their approach. Similarly in this paper we built a prototype system that demonstrates the feature subset selection. The application uses clustering and the results reveal that they are encouraging. The results are also compared with other algorithms like C4.5, Naïve Bayes, IB1 and RIPPER.***

Keywords: Data mining, feature subset selection, clustering

1. Introduction

Data mining with high dimensional data provides business intelligence required. In this context, feature subset selection is an active area of research as it can influence the results in data mining. In fact it can help in improving learning accuracy, removing unnecessary data for processing and reducing dimensionality [1], [2]. In literature many techniques came into existence for feature subset selection including machine learning techniques. These are classified into Filter, Wrapper, Embedded and Hybrid approaches. Embedded methods make use of the concept for feature selection as part of the training process. That is the reason; they are more efficient than other categories [3]. Examples of this kind of methods include artificial neural networks, decision trees and other machine learning algorithms [4]. Predictive accuracy is given importance in wrapper methods to know best features. However, they led to more computational overhead. The filter methods on the other hand have good generality and independent of learning algorithms.

The wrapper methods are also expensive computationally and cause overfit of training sets as explored in [5] and [6]. The filter methods are good choice especially when the features are more in the high dimensional data team. Pereira et al. [7], Baker and McCallum [8], and Dhillon et al. [9] followed distributed approach for feature subset selection. The results obtained by them are comparable to human observations [10]. The data mining scenario in the real world is presented in Figure 1.

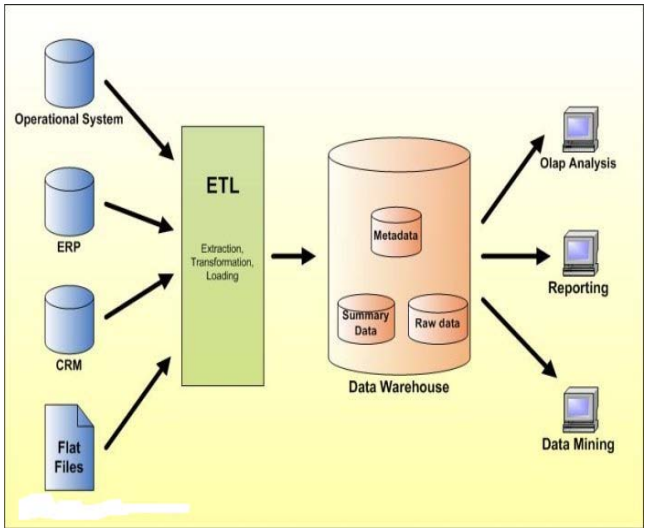

Figure 1: Data mining scenario in the real world

As can be seen in Figure 1, the data mining is used in the real world for extracting or discovering from data sources. Data mining is used along with other operations as shown and the high dimensional data used for mining in this paper. We focus on building a framework that lets us to obtain feature subset selection which will reduce dimensionality besides bestowing various advantages like reducing search space in data mining activities. 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.

2. Related Work

This section focus on the existing literature on the work done in similar lines. In [11] the problem of power exhaustion was fond and named it 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.

3. 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.

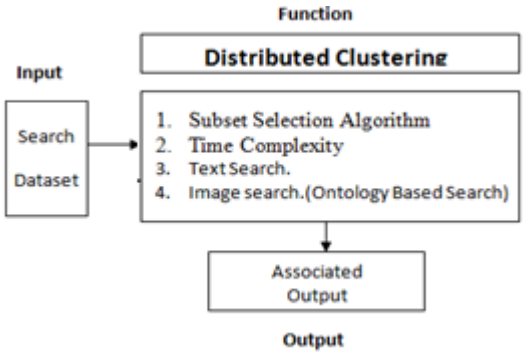

Figure 2: Overview of the proposed solution

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.

4. 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.

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.

5. Experimental Results

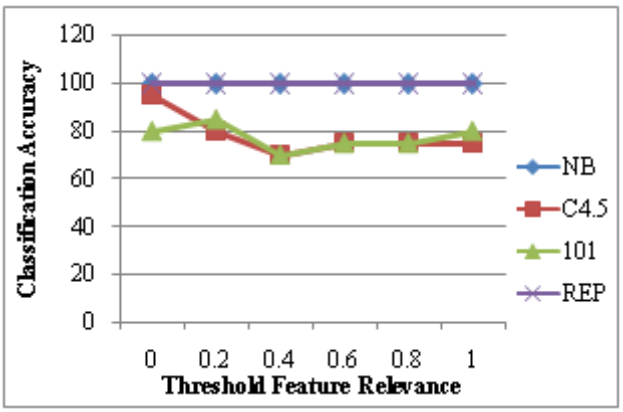

Figure 4: Accuracies of the four classification algorithms with different θ values

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.

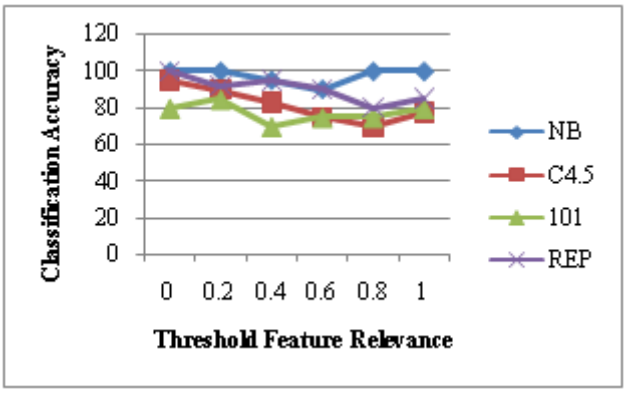

Figure 6: Accuracies of the four classification algorithms with different θ values

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

6. Conclusions and Future Work

In this paper, our research focus was on data mining on high dimensional data in terms of feature subset selection. Feature subset can be representative of the characteristics of underlying high dimensional data. Recently Song et al. [24] presented a solution in which all the features are identified and clustered before selecting the feature subsets. Similar kind of approach is implemented in this paper. A prototype application is built to demonstrate the proof of concept. The proposed solution is able to reduce dimensionality and provide other benefits of selecting representative features. The empirical results are encouraging. In future we further generalize our model to support feature subset selection so as to help in high dimensional data mining 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.

Author Profile

Punnana Sarath Kumar is currently pursuing his M. Tech (Software Engineering), Department of computer Science and Engineering, at Sri Sivani College of Engineering, Chilakapalem, Etcherla Mandal, Srikakulam District, A.P, India. His area of interest includes Data Mining.

Ganiya Rajendra Kumar is currently Working as a Professor & HOD of CSE Department, Department of computer Science and Engineering, at Sri Sivani College of Engineering, Chilakapalem, Etcherla Mandal, Srikakulam District, A.P, India. His area of interests includes Data Mining and Cloud Computing and its applications and Hack Tricks. He participated in A Two day AICTE sponsored National Seminar on Biologically Inspired Computing and its Applications (Sri Sivani College of engineering, srikakulam 22nd & 23rd Nov 2013). He also received awards/recognitions/scholarships such as: Seminar Grant sanctioned from AICTE for the academic year 2013-14 as co-ordinator; Best Care Taker Award in 'A' Flight 6(A) Air SQN NCC, Visakhapatnam; More than ten times 90% above student feedback for teaching.