IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

Survey and Taxonomy of Feature Selection Algorithms in IT log analytics

Mrs.S.Nithya Roopa[1], P.Karthika[2], S.Monisha[3]

Assistant Professor, Computer Science, Kumaraguru College of technology, Coimbatore, India[1]

Student, Computer Science, Kumaraguru College of Technology, Coimbatore, India[2,3]

Abstract: Feature selection is an important for high dimensional dataset. The best subset contains the least number of dimensions that highly contributes to the accuracy and so the remaining unimportant dimensions are ignored. Selecting relevant features from unlabelled data is a challenging task due to the absence of label information by which the feature relevance can be assessed. The unique characteristics of IT log further complicates the challenging problem of unsupervised feature selection, (e.g., part of IT log data is linked, which makes invalid the independent and identically distributed assumption), bringing about new challenges to traditional unsupervised feature selection algorithms. In this paper we compare the performance of Linked Unsupervised feature selection algorithm [1]and feature selection using feature similarity [2].We perform experiments with IT log dataset to evaluate the effectiveness of the both the frameworks.

Keywords: IT log analysis, high dimensional, unlabelled data, attribute-value.

I. INTRODUCTION

IT solutions and IT departments generate an enormous quantity of logs and trace data.. Trends in these databases can be identified using data mining practices, which sort and model the data in order to arrive at a conclusion. The data mining applications present the data in the form of data marts. In IT logs, however, the lack of standard vocabulary has hindered the process of data mining to a certain extent.

This could lead to unnecessary problems, during the process of data mining. The increase in the use of standardized terms will reduce the percentage of errors in the data mining process. The huge data poses new challenges to data mining tasks such as classification and clustering.

The effective approach to handle high-dimensional data is feature selection. According to whether the training data is labelled or unlabelled, feature selection algorithms can be roughly divided into supervised and unsupervised feature selection.

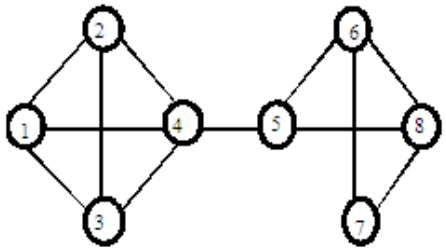
It is time-consuming and costly to obtain labelled data. Given the scale of IT log, we propose to study unsupervised feature selection.

Unsupervised feature selection is particularly difficult due to the absence of class labels for feature relevance assessment.

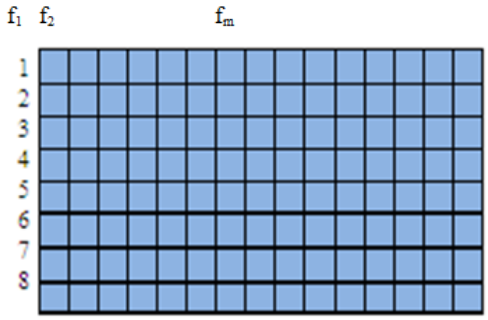
Most existing feature selection algorithms work with “flat” attribute-value data which is typically assumed to be independent and identically distributed. (i.i.d.).

However, the i.i.d. assumption does not hold for IT log since it is inherently linked.

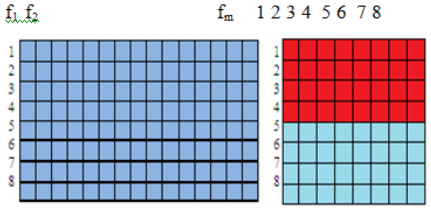
a) Linked users



b) Attribute – Value data



c) Attribute – Value and Linked Data



For linked data, except for the conventional representation, there is link information between instances. Linked data in the log file presents both challenges and opportunities for unsupervised feature selection. In this work, we investigate: (1) how to exploit and model the relations

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

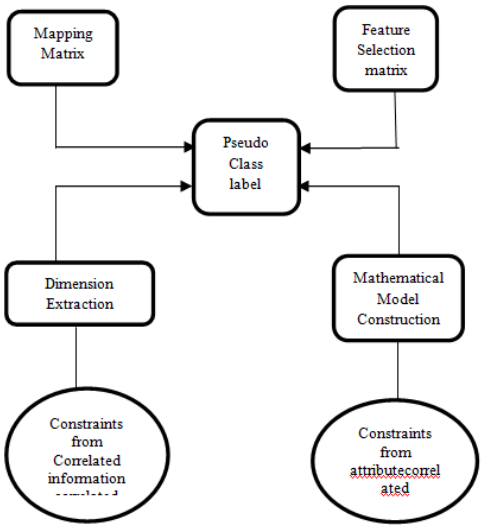
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

among data instances, and (2) how to take advantage of these relations for feature selection using unlabelled data. In supervised learning, label information plays the role of constraint. Without labels, other alternative constraints are proposed, such as data variance and separability.

IT industries will not have the manpower or resource to churn through all the information by hand, let alone in real time. Linked unsupervised feature selection algorithm and feature selection based on feature similarity is done to select best feature subset out of the log files for analysis LUFS-Framework.

By introducing the concept of pseudo-class labels, constraints from both link information and unlabelled attribute value data are ready for unsupervised feature selection. The framework of linked unsupervised feature selection, LUFS, with our solutions to the two challenges (need to take into account of linked data and lack of labels): extracting constraints from both linked and attribute-value data, and then constructing pseudo-class labels through social dimension extraction and spectral analysis

LUFS framework



Algorithm 1 Correlated Unsupervised Feature Selection
Input: {X, R, α, β, λ, c, K, k}
Output: k most relevant features

1: Obtain the social dimension indicator matrix H
2: Set $F = H(H^{\top}H)^{\frac{1}{2}}$
3: Construct S through Eq. (11)
4: Set L = D – S
5: Set $A = XLX^{\top} + \alpha X(I_n - FF^{\top})X^{\top}$
6: Set $B = XX^{\top} + \lambda I$
7: Set t = 0 and initialize D_0 as an identity matrix
8: while Not convergent do
9: Set $Ct = B^{-1}(A + \beta Dt)$
10:Set Wt = [q1, . . . , qc] where q1, . . . , qc are the eigenvectors of Ct corresponding to the first c smallest eigenvalues
11: Update the diagonal matrix D_{t+1}
12: Set t = t + 1
13: end while
14: Sort each feature according to W in descending order and select the top-k ranked ones;

In LUFS algorithm, dimension extraction and weighted dimension indicator construction are from line 1 to line 2. The iterative algorithm to optimize Eq. (19) is presented from line 8 to line 13

Feature selection using feature similarity
The task of feature selection involves two steps, namely, partitioning the original feature set into a number of homogeneous subsets and selecting a representative feature from each such cluster. Partitioning of the features is done based on the k-NN principle using one of the feature similarity measures. In doing so, we first compute the k nearest features of each feature. Among them the feature having the most compact subset is selected and its k neighbouring features are discarded. The process is repeated for the remaining features until all of them are either selected or discarded.

II.ALGORITHM

Let the original number of features be D, and the original feature set be $O=\{F_i, i=1,\ldots, D\}$.Represent the dissimilarity between features F_i and F_j by $S(F_i, F_j)$.Higher the value of S,the more dissimilar are the features. Let r_i^k represent the dissimilarity between feature F_i, and its kth nearest neighbour feature in R.

Then

Step 1: Choose an initial value of k≤D-1.Initialise the reduced feature subset R to the original feature set O,R O.←
Step2: For each feature $F_i \in R$, compute r_i^k
Step3: Find feature F_i,for which r_i^k is minimum.Retain this feature in R and discard k nearest features of F_i.
Step4: If k>cardinality(R)-1:k=cardinality(R)-1.
Step5: If k=1,: Go to step 8.
Step 6: While $r_i^k > \epsilon$ do:
(a) k=k-1
$r_i^k = \inf_{Fi \in R} r_i^k$
("k" is decremented by 1,until the "kth nearest neighbour " of atleast one of the features in R is less than ϵ-dissimilar with the feature)
(b) If k=1: Go to step 8.
(if no feature in R has less than ϵ-dissimilar "nearest-neighbor" select all the remaining features in R)
End While
Step 7: Go to Step 2 .
Step 8: Return feature set R as the reduced feature set.
The algorithm has low computational complexity with respect to both number of features and number of samples of the original data.

III. EXPERIMENTS AND DISCUSSION

In this section, we present experiment details to verify the effectiveness of both the framework, LUFS and feature similarity. After introducing real-world IT log, we first

DOI 10.17148/IJARCCE.2016.5330

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

evaluate the quality of selected features in terms of clustering performance, then study the effects of parameters on performance and finally further verify the constraint extracted from link information by all dimensions. We have collected two sample dataset from IT log files. Some statistics of the datasets are shown in Table 1.

Table 1: Statistics of the Datasets

	Dataset1	Dataset2
Size	6,816	5,553
# of Features	10,081	9,765
# of Classes	4	5
# of Links	20,567	34,441

LUFS selects features in batch mode by simultaneously exploiting linking information residing in the log files
Similarity Measure

Feature selection based on similarity measure works by evaluating the data set with maximum information compression index.

Following the existing evaluation practice for unsupervised feature selection, we assess LUFS in terms of clustering performance. We vary the numbers of selected features as {200, 300, 400, 500, 600, 700, 800, 900, and 1000}. Each feature selection algorithm is first performed to select features, then K-means clustering is performed based on the selected features.

The quality of features selected by both the algorithms using performance metrics is tabulated. We observe the performance change with the numbers of selected features: it increases, reaches the peak, and then decreases. For example, LUFS achieves its peak values when the number of selected features are 500 and 300 in both the log files respectively. The clustering performance with as few as 200 features is better than that with all features. For instances, LUFS obtains 10.51% and 18.68% relative improvement in terms of accuracy for both the log files, respectively. These results demonstrate that the number of features can be significantly reduced without performance deterioration

Table 1: Performance in logfile1

Accuracy	200	300	400	500	600	700	800	900	1000
LUFS	29.19	29.51	29.90	32.70	30.41	31.17	30.48	31.79	31.17
Similarity Measure	26.16	25.41	27.76	28.14	29.19	30.04	29.78	21.76	30.09

Table 2: Performance in logfile1

NMI	200	300	400	500	600	700	800	900	1000
LUFS	0.0951	0.1026	0.1489	0.1601	0.1582	0.1701	0.1614	0.1681	0.1596
Similarity Measure	0.1011	0.1223	0.1431	0.1566	0.1648	0.1309	0.1406	0.1346	0.1455

Table 1: Performance in logfile2

Accuracy	200	300	400	500	600	700	800	900	1000
LUFS	26.24	24.56	25.81	29.65	28.36	29.21	29.41	29.69	29.16
Similarity Measure	27.21	27.38	24.64	27.22	26.31	29.54	28.78	25.37	29.06

Table 2: Performance in logfile2

NMI	200	300	400	500	600	700	800	900	1000
LUFS	0.0843	0.1131	0.1245	0.1503	0.1673	0.1553	0.1460	0.1672	0.1358
Similarity Measure	0.1530	0.1241	0.1320	0.1441	0.1490	0.1176	0.1354	0.1230	0.1128

IV.CONCLUSION

This survey provides a comprehensive overview of two unsupervised feature selection algorithms for IT log analytics. The feature selection of audit data has adopted three main methods; wrapper, filter, and hybrid method. The hybrid approaches have been proposed to improve both filter and wrapper method. However, in some recent applications of feature selection, the dimensionality can be tens or hundreds of thousands. Such high dimensionality causes two major problems for feature selection. One is the so called "curse of dimensionality". As most existing feature selection algorithms have quadratic or higher time complexity about N, it is difficult to scale up with high dimensionality. Since algorithms in the filter model use evaluation criteria that are less computationally expensive than those of the wrapper model, the filter model is often preferred to the wrapper model in dealing with large dimensionality.

The quality of features selected by both the algorithms using performance metrics is tabulated. We observe the performance change with the numbers of selected features: it increases, reaches the peak, and then decreases. For example, LUFS achieves its peak values when the number of selected features are 500 and 300 in both the log files respectively. The clustering performance with as few as 200 features is better than that with all features. For instances, LUFS obtains 10.51% and 18.68% relative improvement in terms of accuracy for both the log files, respectively. These results demonstrate that the number of features can be significantly reduced without performance deterioration.

REFERENCES

[1]. J. G. Dy and C. E. Brodley. Feature subset selection and order identification for unsupervised learning. In Proceedings of the Seventeenth International Conference on Machine Learning, pages 247–254, 2000.

[2]. J. G. Dy and C. E. Brodley. Visualization and interactive feature selection for unsupervised data. In KDD, pages 360–364, 2000.

[3]. J. G. Dy, C. E. Brodley, A. C. Kak, L. S. Broderick, and A. M. Aisen. Unsupervised feature selection applied to content-based retrieval of lung images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(3):373–378, 2003.

[4]. Y. Kim, W. Street, and F. Menczer. Feature selection for unsupervised learning via evolutionary search. In KDD, pages 365–369, 2000

[5]. L. Wolf and A. Shashua. Feature selection for unsupervised and supervised inference: the emergence of sparsity in a weighted-based approach. Journal of Machine Learning Research, 6:1855–1887, 2005.

[6]. Z. Zhao and H. Liu. Spectral feature selection for supervised and unsupervised learning. In Proceedings of the 24th international conference on Machine learning, pages 1151–1157. ACM, 2007.

[7]. D. W. Aha and R. L. Bankert. Feature Selection for Case– Based Classification of Cloud Types. In Working Notes of the AAAI94, Workshop on Case–Based Reasoning, pages 106–112, Seattle, WA, 1994. AAAI Press.

[8]. H. Almuallim and T. G. Dietterich. Learning with Many Irrelevant Features. In Proc. of the 9th National Conf. on Artificial Intelligence, volume 2, pages 547–552, Anaheim, CA, 1991. AAAI Press

[9]. A. L. Blum and P. Langley. Selection of Relevant Features and Examples in Machine Learning. In R. Greiner and D. Subramanian, eds., Artificial Intelligence on Relevance, volume 97, pages 245–271. Artificial Intelligence, 1997.

[10]. G. H. John, R. Kohavi, and K. Pfleger. Irrelevant Features and the Subset Selection Problem. In Proc. of the 11th Int. Conf. on Machine Learning, pages 121–129, New Brunswick, NJ, 1994. Morgan Kaufmann

[11]. H. Liu and H. Motoda. Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic Publishers, London, GB, 1998.

[12]. Pabitra Mitra,C.A.Murthy,Sankar K.Pal.Unsupervised Feature Selection Using Feature Similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(3), 2002.