ESANN'2007 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2007, d-side publi., ISBN 2-930307-07-2.

A New Feature Selection Scheme Using Data Distribution Factor for Transactional Data

P. Y. Wang[1] and T.W.S. Chow[2]

1-City University of Hong Kong - Dept of Electronic Engineering
Tat Chee Avenue, Kowloon, Hong Kong SAR
2- City University of Hong Kong - Dept of Electronic Engineering
Tat Chee Avenue, Kowloon, Hong Kong SAR

Abstract. A new efficient unsupervised feature selection method is proposed to handle transactional data. The proposed feature selection method introduces a new Data Distribution Factor (*DDF*) to select appropriate clusters. This method combines the compactness and separation together with a newly introduced concept of singleton item. This new feature selection method is computationally inexpensive and is able to deliver very promising results. Four datasets from UCI machine learning repository are used in this studied. The obtained results show that the proposed method is very efficient and able to deliver very reliable results.

1. Introduction

Feature reduction is a generic term for a process that aims at reducing features with certain criteria. Feature selection is one of the most common approaches to achieve this goal. Each unsupervised feature selection inherits characteristics of its employed clustering algorithm, inclusive of the evaluation criteria [1]. Due to the evaluation criteria which involve distance calculation and no order information, unsupervised selection method is very rare.

In this paper, an efficient unsupervised feature selection scheme for transactional data is proposed. The proposed feature selection scheme, called UFSN, is able to directly process nominal dataset. It is to the best knowledge of us that only SUD [2] and the proposed scheme are able to perform unsupervised feature selection on transactional data. SUD uses entropy similarity measurement to determine the importance of features with respect to the underlying clusters. Features are ranked according to the entropy similarity measurement. In this paper, we propose a very computationally efficient approach for handling transactional data feature selection. In the obtained results, the proposed scheme is found to be up to 100 times more efficient than SUD because UFSN does not require the iterative calculation of entropy. As there is no class label provided, a clustering algorithm must be firstly used to generate a set of clustering results, which are called cluster descriptions in this paper. These cluster descriptions are generated by the clustering algorithm with different parameter settings. A Data Distribution Factor (*DDF*) is newly introduced to select an appropriate cluster description from the provided cluster descriptions for further measurement. *DDF* is the combination of both compactness and separation. Similar objects are grouped into the same cluster for obtaining a high compactness. For the separation part, another new idea of singleton item is introduced for handling transactional data. We show that the more singleton items there are in a separation

among clusters, the higher the entropy it will be. Thus, it indicates that separation and the number of singleton items have a very similar nature in terms of clustering information. In addition, the determination of singleton item is very computationally efficient. A relevance index using the concept of singleton item is then developed to evaluate features individually. Based on the selected cluster description, this study shows that the newly proposed scheme can deliver very promising results. This paper is organized as follows. Section 2 presents the proposed feature selection method. Section 3 shows our intensive simulations based on 4 UCI datasets. At last, a conclusion is drawn in Section 4.

2. Proposed Feature Selection Mechanism

In this study, a transactional dataset with *m* features and *N* nominal data instances is considered. The proposed feature selection scheme comprises three parts. First, a clustering algorithm that can handle transactional data is used to generate a set of cluster descriptions, which describe the data characteristics in clustering sense. Then, *DDF* is used to select an appropriate cluster description for further relevance ranking. Finally, a newly developed feature relevance index is applied for feature ranking.

a. Phase 1 Generation of Cluster Descriptions

Clustering is an unsupervised process aiming at grouping similar objects into the same cluster and separating dissimilar objects into different clusters. In this study, two different transactional clustering algorithms, CLOPE [3] and SLR [4], are used in this phase for comparison since they are the two most commonly-used techniques in transactional data clustering.

CLOPE and SLR have 1 and 3 user input parameters respectively. CLOPE uses the height-to-width ratio of the cluster histogram to determine clusters. A parameter, *r*, is used to control the tightness of the clusters. SLR is an enhanced version of LargeItem [5], which employs the large item idea from association rule. SLR introduces middle item, which is an item belong to neither large item nor small item. 3 user input parameters (minimum support, MinSup, damping factor, λ, and SLR threshold, α) are required.

b. Phase 2 Cluster Description Selection

In this phase, Cluster Description Selection, an appropriate cluster description is selected by Data Distribution Factor (*DDF*) and passed to Phase 3. The appropriate cluster description is a cluster description with highest *DDF* and its number of clusters is greater than 1. *DDF* is defined below.

$$DDF = \frac{\sum_{j=1}^{k}|C_j|\left[\frac{\sum_{fi\_v\in D(C_j)} Frequency(fi\_v, C_j)}{|D(C_j)|\times|C_j|}\right]}{\sum_{j=1}^{k}|C_j|} + \frac{\sum_{fi\_v\in Singleton} Frequency(fi\_v)}{\sum_{j=1}^{k}\sum_{fi\_v\in D(C_j)} Frequency(fi\_v, C_j)} = \frac{\sum_{j=1}^{k}\left[\frac{\sum_{fi\_v\in D(C_j)} f(F_i\_v, C_j)}{|D(C_j)|}\right]}{\sum_{j=1}^{k}|C_j|} + \frac{\sum_{fi\_v\in Singleton} f(F_i\_v)}{\sum_{j=1}^{k}\sum_{fi\_v\in D(C_j)} f(F_i\_v, C_j)}$$

Where $f(F_i\_v, C_j)$ is the frequency of value F_i_v in cluster Cj, $|D(Cj)|$ is the number of the distinct values in cluster j, $|Cj|$is the number of instances of cluster Cj.

The first part of *DDF* evaluates the compactness of the cluster description. As one of the purposes of clustering is to group similar objects into the same cluster, a high compactness within a cluster means the objects in the cluster exhibit higher similarity. The second part of *DDF* evaluates the separation of the cluster description. The second purpose of clustering is to group dissimilar objects into different clusters. In the following example, the concept of entropy is used to illustrate the relationship between the number of singleton items and its cluster separation.

Example A dataset consists of 7 transactions. t1={a, b, c}, t2={a, b, c, d}, t3={a, b, c, e}, t4={a, b, c}, t5={d, g, h}, t6={d, g, i}, t7={a, b, c}. Cluster description "A" is C1={t1, t2, t3, t4, t7} and C2={t5, t6}. Cluster description "B" is C1={t1, t4, t7}, C2={t2, t3} and C3={t5, t6}.

Obviously, for description "A", item "d" appears in both clusters 1 and 2. This means other items, i.e., "a", "b", "c", "e", "g", "h", "i" are all singleton items, which appear in only one cluster. Similarly, in description "B", only items "e", "g", "h", "i" are singleton items. Therefore, seven singleton items are found in description "A", while there are four singleton items in description "B". Entropy for cluster description "A" and cluster description "B" is 0.9597, and 0.53 respectively. Using the concept of entropy, this typical example shows that a clear separation among clusters has more singleton items.

A cluster description with the highest *DDF* and number of clusters greater than 1 is chosen for relevance rank in Phase 3

c. Phase 3 Relevance Rank

Based on the selected cluster description, relevance value of each feature is evaluated. A feature is a relevant feature when its relevance value is higher than or equal to the threshold, *IrrThreshold*. Relevance index, $REL(F_i)$, of feature i is defined.

$$REL(F_i) = \frac{|Singleton|}{|F_i|} \times \frac{N - Miss(F_i)}{N}$$

Where $|Singleton|$ is the number of singleton values, $|Fi|$ is the number of values in i^{th} feature, fi, N is the number of instances, and $Miss(F_i)$ is the number of instances with missing value in Fi.

A relevant feature is a feature grouping instances according to the cluster description closely. If all values of a particular feature are singleton values, this feature groups the instances exactly according to the cluster description. Hence, higher $|Singleton|/|F_i|$ means the feature groups the instances more closely to the cluster description. Since there may be missing values in features, $[N-Miss(F_i)]/N$ is used to weigh the singleton value percentage of a feature.

However, there are some cases that such definition seems too stringent. A parameter, *AccFreq* (Acceptable Frequency), is proposed to loosen the definition and satisfy those cases if required. If a value appears in more than one cluster and it mostly occurs in one cluster (i.e., its frequency in one cluster is greater than or equal to *AccFreq*), it is still regarded as a singleton value. *AccFreq* = 100% is used in all investigations. It is worth noting that the $REL(F_i)$, which is the label used for the

subsequent unsupervised clustering, is related to *IrrThreshold* and *AccFreq*. Apparently, the $REL(F_i)$ is just a synthetic class label.

d. Enhanced Version

To reduce the computational resources, an enhanced version of UFSN, called EUFSN, is designed. In some clustering algorithms, the number of clusters can be roughly estimated by their parameters. For example, the number of clusters tends to increase when *r* in CLOPE increases. Based on this property, the enhanced scheme changes the parameter automatically.

In the enhanced scheme, one of the stopping criteria is that the number of clusters is greater than $\sqrt{N}$, where *N* is the number of data instances. This criterion is used to prevent choosing a cluster description with the number of clusters close to the number of instances. In some datasets, the second part of *DDF* is very low for all cluster descriptions. In these cases, *DDF* increases when the number of clusters increases. Hence, it is suggested that only the cluster descriptions with the number of clusters between 2 and $\sqrt{N}$ [6] are evaluated.

Instead of generating all clustering descriptions in Phase 1 and evaluating them in Phase 2, the enhanced scheme combines Phase 1 and Phase 2 to save computational resources. First of all, the enhanced scheme generates a cluster description with parameter at the minimum and evaluates the description. Then, the parameter is raised by a user pre-defined step-up size and generates a cluster description with the new parameter setting, if necessary. The parameter step up repeats and goes on to generate and evaluate cluster description until the cluster description with the highest *DDF* value is determined. The cluster description with the highest *DDF* proceeds to Phase 3.

3. Results

In this study, we use four real datasets from UCI machine learning repository [7] to demonstrate the ability of the proposed method. And SUD [2] is compared with the proposed UFSN, where *AccFreq* is set at 1 for all comparisons. The scheme based on CLOPE with enhanced scheme in Phase 1 is called EUFSN-CLOPE. EUFSN-CLOPE starts at *r* of 0.1 and step size of 0.1 until the number of clusters is greater than $\sqrt{N}$. SLR used in Phase 1 with α at 0.5 is called UFSN-SLR-0.5. For UFSN-SLR-0.5, the cluster description is generated with minimum support at 0.6 and λ varying from 0.4 to 1. Table 1 shows the comparison of computational time among EUFSN-CLOPE, UFSN-SLR-0.5, and SUD. Compared with the proposed scheme, SUD is very computationally demanding. And EUFSN-CLOPE takes a bit more time than UFSN-SLR-0.5. However, both of them are much faster than SUD. The platform of this study is a desktop computer with 512MB of RAM, Intel P4 1.3 GHz CPU, Windows XP version 2002 Service Pack 2 and ActivePerl 5.8. All classification accuracies are obtained by J48 decision tree with 10-fold cross-validation in Weka.

Dataset	EUFSN-CLOPE	UFSN-SLR-0.5	SUD
Agaricus-lepiota	8685	1255	After 66 Days

Breast-Cancer	14	12	1216
Hepatitis	13	8	1753
Lung-Cancer	3	1	948

Table 1 Comparison of computational time (in second) of different methods

Fig. 1 shows the classification accuracy of different feature subsets conducted via EUFSN-CLOPE, UFSN-SLR-0.5, and SUD. Meanwhile, Fig. 1 presents a brief comparison between the proposed schemes (both EUFSN-CLOPE and UFSN-SLR-0.5) and SUD.

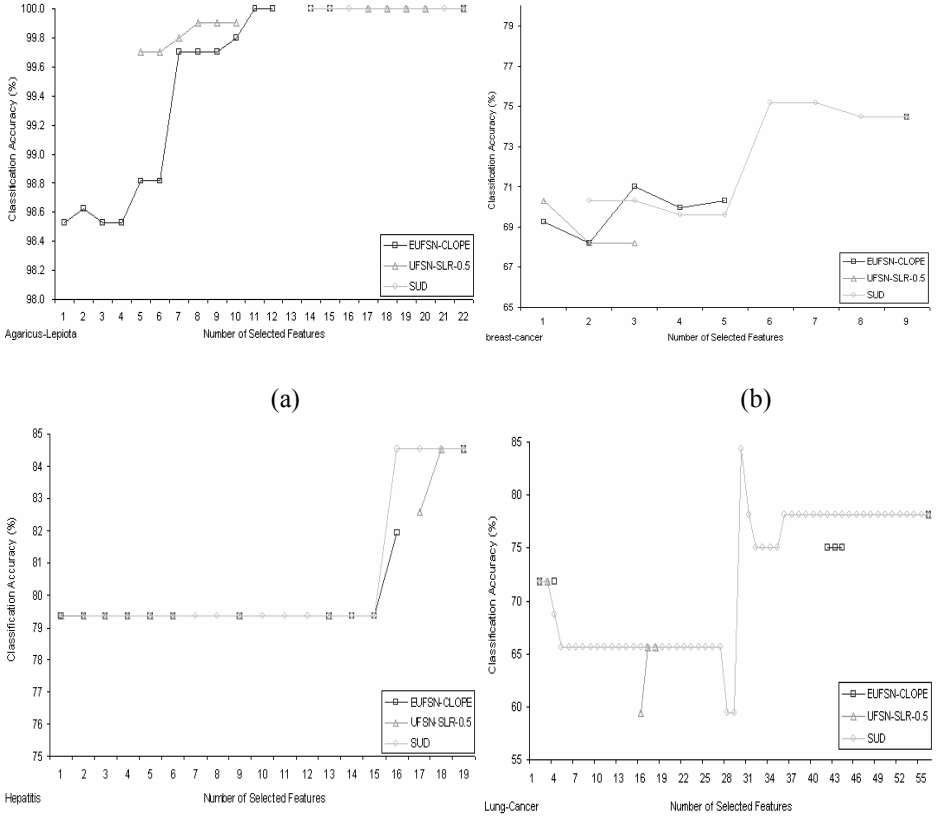

Fig. 1 Classification Accuracy for different number of features selected by EUFSN-CLOPE, UFSN-SLR-0.5, and SUD of dataset a) "Agaricus-Lepiota", b) "Breast-Cancer", c) "Hepatitis", and d) "Lung-Cancer".

Classification accuracies with respect to the dataset "Agaricus-Lepiota" is discussed in detail in the following paragraphs. The dataset "Agaricus-Lepiota" comprises two classes: "edible" and "poisonous". The features describe different fundamental characteristics such as odor and cap-color.

EUFSN-CLOPE reduces the dataset from 22 features to 11 features (50.0% reduced) without lowering the classification accuracy. When the number of features

is reduced to 1 (95.5% reduced), EUFSN-CLOPE is still able to maintain an accuracy of 98.5%, i.e., the most important feature is picked. In general, EUFSN-CLOPE outperforms others when there is only 1 feature left. EUFSN-CLOPE uses 2.5 hours to rank features respectively whereas UFSN-SLR-0.5 uses less than 0.5 hours to rank features of the same dataset. In addition, it takes more than 60 days to rank 8 features via SUD and the process is subsequently terminated on the 66^{th} day. As shown in the presented results, the proposed schemes are more efficient than SUD.

To sum up, the classification accuracy and the number of selected features by SUD is about the same as that of the proposed schemes. Nevertheless, the computational time of SUD, about 100 times on average, is substantially longer than the proposed scheme. It is clear that the proposed schemes select relevant features in a more efficient way compared with other methods.

4. Conclusion

An efficient unsupervised feature selection scheme is developed for performing transactional data feature selection. The proposed scheme can be used with different clustering algorithms, for instance, CLOPE and SLR. Data distribution factor (*DDF)* is introduced as a stopping criterion for selecting cluster description for relevance ranking. Singleton item, proved to be similar in nature of finding the higher entropy, is developed for efficient clustering. Based on the selected cluster description, the relevance of the features is measured by using the proposed relevance index. User is allowed to adjust the threshold, *IrrThreshold*, to control the number of features to be included. SUD is compared with the proposed scheme and the obtained results show that the proposed scheme is a reliable and efficient feature selection methodology.

References

[1] K. Z. Mao, "Identifying Critical Variables of Principal Components for Unsupervised Feature Selection", *IEEE Transactions on Systems, Man, and Cybernetics – Part B*, 2005, Vol. 35 (2), pp. 339-344.

[2] M. Dash, H. Liu, and J. Yao, "Dimensionality Reduction of Unsupervised Data." In *Proceedings of Ninth IEEE International Conference on Tools with Artificial Intelligence*, 1997, pp. 532-539.

[3] Y. Yang, X. Guan, J. You, "CLOPE: A Fast and Effective Clustering Algorithm for Transactional Data." In *Proceeding of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002, pp. 682-687.

[4] C. H. Yun, K. T. Chuang, and M. S. Chen, "An Efficient Clustering Algorithm for Market Basket Data Based on Small Large Ratios", In *Proceedings of 25th Annual International Computer Software and Applications Conference*, 2001, pp. 505-510.

[5] K. Wang, C. Xu, and B. Liu, "Clustering Transactions Using Large Items", In *Proceedings of 8th International Conference on Information and Knowledge Management*, 1999, pp. 483-490.

[6] J. C. Bezdek, and N. R. Pal, "Some new indexes for cluster validity", *IEEE Trans. Systems Man Cybernet – Part B*, 1998, Vol. 28, pp. 301-315.

[7] http://www.ics.uci.edu/~mlearn/MLRepository.html