Incorporating Feature Selection and Clustering Approaches for High-Dimensional Data Reduction

Been-Chian Chien
Department of Computer Science and Information Engineering
National University of Tainan, Tainan, Taiwan, China

Keywords: High-dimensional Data, Data Reduction, Feature Selection, Clustering, Document Categorization.

Abstract: Data reduction is an important research topic for analyzing mass data efficiently and effectively in the era of big data. The task of dimension reduction is usually accomplished by technologies of feature selection, feature clustering or algebraic transformation. A novel approach for reducing high-dimensional data is initiated in this paper. The main idea of the proposed scheme is to incorporate data clustering and feature selection to transform high-dimensional data into lower dimensions. The incremental clustering algorithm in the scheme is used to handle the number of dimensions, and the relative discriminant variable is design for selecting significant features. Finally, a simple inner product operation is applied to transform original high-dimensional data into a low one. Evaluations are conducted by testing the reduction approach on the problem of document categorization. The experimental results show that the reduced data have high classification accuracy for most of datasets. For some special datasets, the reduced data can get higher classification accuracy in comparison with original data.

1 INTRODUCTION

Handling a huge number of data records and high-dimensional data features efficiently and effectively is the main challenge in the era of big data. For example, a large number of digital documents such as blogs, e-news, e-papers, and on-line reports are produced by persons and enterprises on the Internet everyday. The numerous documents will derive the problems of textual analysis and high-dimensional feature space. However, it is time consuming to process large amount of text and high-dimensional data. Especially, the curse of dimensionality may become a serious obstacle while machine learning and data mining technologies are employed in some applications, e. g. data classification, regression, etc. A practical task is automatic text categorization which uses bag-of-words model (Salton, 1983) based on a set of feature keywords extracted from numerous documents. The set of keywords thus forms a large sparse matrix with high-dimensional frequencies of terms and it is difficult for general tools to process such a huge matrix.

To reduce the number of attributes and reserve meaningful information in high-dimensional data, many feature reduction methods were proposed in the past. Generally, feature selection (Liu 2005) and feature clustering (Kriegel et al., 2009) are the two main categories of methods to reduce dimension space of features. An alternative class of transformation method, like Principle Component Analysis (Jolliffe, 2002), uses projecting process of algebraic operation to convert a high-dimensional dataset into a lower-dimensional dataset. Although the transformation method can provide effective results of reducing dimensions, the computational cost is expensive. Furthermore, the conversion of a high-dimensional matrix in big data is impossible since the number of data or the dimension of features may be very large.

The idea of incorporating the strategies of feature selection and data clustering approach to transform high-dimensional data into low-dimensional data is proposed in this paper. The proposed approach first gives a simple incremental clustering method to agglomerate data with a proper similarity function for a specific application. The clustering results are then used to analyze the relative discriminant variables which represent the discerning ability of a feature on different clusters. Through the matrix of relative discriminant variables, the original dataset with high dimensions can be transformed into a new one with lower dimensions by inner product

Chien B..
Incorporating Feature Selection and Clustering Approaches for High-Dimensional Data Reduction.
DOI: 10.5220/0005093300720077
In *Proceedings of 3rd International Conference on Data Management Technologies and Applications* (DATA-2014), pages 72-77
ISBN: 978-989-758-035-2

operation. The number of dimensions will be the number of clusters after the transformation.

One of the high-dimensional data applications, document categorization, is adopted to verify the performance of the proposed scheme. Three well-known large text datasets, 20 Newsgroups, Cade12, and RCV1, are used to evaluate whether the data reduction will degrade the accuracy of classification or not. The experimental results illustrate the fact that some of the reduced datasets produced by the proposed scheme even have better classification accuracy than original datasets. Further, most of the datasets still maintain effectiveness in a very low data dimensions after the processing of reduction.

This paper is organized as follows. Section 2 introduces the related work on feature reduction. The idea of combining feature selection and data clustering is revealed in Section 3. The experiments of employing the two techniques are evaluated in Section 4 to demonstrate the feasibility of the work. Finally, summary and discussion are depicted.

2 REVIEW ON DATA REDUCTION

The previous researches on feature reduction are briefly reviewed and summarized as follows.

2.1 Feature Selection

Selecting informative features is the simplest and direct way to reduce data dimensions. The objective of feature selection is to find a subset of significant features from a large number of high dimensional features according to specific task of measurement on a dataset. For instance, information gain (IG) (Yang & Pedersen, 1997) is the most popular feature selection method which is frequently used on data classification. Many earlier researches on feature selection were proposed and designed for machine learning, such as (Daphne & Sahami, 1996), (Blum & Langley, 1997), and (Combarro et al., 2005). The recent work in (Hsu & Hsieh 2010) uses correlation coefficients to select the class-dependent features. Generally, most of the feature selection methods are efficient in computation time.

2.2 Feature Clustering

The method of clustering features was initiated by (Baker & McCallum, 1998). The main technique is to aggregate similar features together first and partition features into distinct clusters. Then, the representative features for clusters are extracted to be the features of a dataset. Many related works and improvement were proposed in the past, like distributional clustering of features (Slonim & Tishby 2001) and clustering features based on the distribution of class labels associated with each feature (Bekkerman et al., 2003). Recently, an efficient self-constructing fuzzy feature clustering algorithm (Jiang et al., 2010) is proposed to extract features by clustering data records instead of features. An extracted feature is a fuzzy weighted combination of original features on all clusters.

2.3 Other Methods

Feature transformation is the other type of feature extraction which transforms high-dimensional data into new subspace with lower dimensions. Principal Component Analysis (PCA) (Jolliffe, 2002) is a well-known method of feature transformation. PCA transforms original data into new coordinate systems such that the projection on the first coordinate has the greatest variance among all possible projections, and the projection on the second coordinate has the second greatest variances, etc. The similar methods include LDA (Martinez & Kak, 2001) and IOC (Park, 2003). The incremental orthogonal centroid method (IOC) is a feature extraction method that tries to find an optimal transformation matrix to convert an original matrix $|D| \times n$ into a $|D| \times k$ matrix, where k is much less than n.

3 DATA REDUCTION SCHEME

Given a dataset D, d_i is a data row and $d_i \in D$. $F = \{f_1, f_2, \ldots, f_n\}$ represents the set of features with n dimensions in d_i. Let d_{ij} be the value of jth feature for the datum d_i, where $1 \le i \le |D|$, $1 \le j \le n$, and $|D|$ is the number of data in D.

To reduce feature dimensions, the proposed feature reduction scheme combines a clustering algorithm and feature selection methods. The procedures are described in the following sub-sections.

3.1 Data Clustering

First, the similar data in the dataset D are grouped together by their original features F. However, we know that data clusters are dependent on not only the steps of the clustering algorithm but also the similarity function they applied. Since there are different measures of similarity for various applications, the used similarity function will reflect selecting results of significant features.

Let $Sim(d_i, d_j)$ be a general form of specific similarity functions that measures the similarity degree between two data d_i and d_j. The mean of data for each feature dimension belonging to the cluster G_l can be used to represent the centre of cluster. A primitive incremental clustering algorithm based on a similarity function $Sim(d_i, d_j)$ is given as follows.

Algorithm: Primitive incremental clustering.

Input: Data set D, a threshold ρ.

Output: Clusters $\mathbf{G} = \{G_1, G_2, \ldots, G_k\}$.

{

  $\mathbf{G} = \{G_1\}$; // the set of clusters

  $k = 1$; // the number of clusters

  $G_1 = \{d_1\}$;

  for all $d_i \in D$

    if (for all $G_l \in \mathbf{G}$, $Sim(G_l, d_i) < \rho$)

      $k = k + 1$;

      $G_k = \{d_i\}$;

      $\mathbf{G} = \mathbf{G} \cup \{G_k\}$;

    else

      $t = \arg_l \max_{G_l \in \mathbf{G}} \{Sim(G_l, d_i)\}$;

      $G_t = d_i \cup G_t$;

    endif

  endfor

}

The above clustering algorithm is an incremental based scheme. The ρ is a threshold to determine the mutual difference between the clusters. The first cluster G_1 is initiated by the first data d_1. The latter joined data d_i has two possible cases: the first one is to merge the data d_i into the existing cluster G_l having the maximal similarity if $Sim(G_l, d_i)$ is larger than or equal to ρ. The other case is to generate a new cluster when $Sim(G_l, d_i)$ is less than ρ for all current clusters G_l in $\mathbf{G}$. A lower threshold ρ will generate more clusters than a higher threshold. The threshold ρ and the similarity function $Sim(\cdot)$ can be set and defined, respectively, by a user according to the requirement of an application.

3.2 Feature Selection

After clustering the data, all of the generated clusters are used to analyze the importance of features. The basic procedure of feature analysis is described as follows.

Let G_l be one of the clusters generated by the primitive incremental clustering algorithm, $1 \le l \le |\mathbf{G}|$, $|\mathbf{G}|$ is the number of total clusters. Assume that d_{ij} is the value of jth feature for the datum d_i and $\mathbf{d} = [d_{ij}]_{|D|\times n}$ is the matrix of the original dataset D. First, the feature weight of each cluster, w_{lj}, is obtained by averaging d_{lj} in each cluster G_l, and $\mathbf{w} = [w_{lj}]_{|\mathbf{G}|\times n}$ is defined as follows.

$$w_{lj} = \frac{1}{|G_l|} \sum_{d_i \in G_l} d_{ij}, \tag{1}$$

where $G_l \in \mathbf{G}$ and $1 \le j \le n$. Then, each weight w_{lj} is normalized by the maximum value of the jth feature, as follows.

$$\tilde{w}_{lj} = \frac{w_{lj}}{\max_{1\le i\le|\mathbf{G}|}\{w_{ij}\}}, \tag{2}$$

where $1 \le l \le |\mathbf{G}|$ and $1 \le j \le n$.

Let z_{lj} be the relative discriminant variable of the jth feature between the cluster G_l and other clusters $G_i \in \mathbf{G}$. The discriminative degree is considered as the product of relative differences of normalized weights for the corresponding cluster. The formal definition is shown as:

$$z_{lj} = \begin{cases} \prod_{\substack{1\le i\le|\mathbf{G}| \\ i\ne k}} |\tilde{w}_{lk} - \tilde{w}_{li}| & \text{if} \quad w_{lj} \ne 0, \\ 0 & \text{if} \quad w_{lj} = 0; \end{cases} \tag{3}$$

where $1 \le l \le |\mathbf{G}|$ and $1 \le j \le n$. The normalized relative discriminant variable is defined as

$$\tilde{z}_{lj} = \begin{cases} 1 + \dfrac{\log z_{lj}}{z_{\max}} & \text{if} \quad z_{lj} > 0 \text{ and } \log z_{lj} \le z_{\max}, \\ 0 & \text{if} \quad z_{lj} = 0 \text{ or } \log z_{lj} > z_{\max}; \end{cases} \tag{4}$$

where $z_{\max}$ is a presetting constant which describes the maximum of computational precision. The range of $\tilde{z}_{lj}$ is between 0 and 1.

3.3 Feature Reduction

Feature reduction for the dataset D is to find a reduced matrix such that the dimension of features is smaller than the dimension of original data. The reduction step simply uses the original data matrix $\mathbf{d}$ and the normalized relative discriminant variable matrix $\tilde{z}$ to get the reduced feature matrix $\mathbf{r}$.

$$\mathbf{r} = \mathbf{d} \cdot \tilde{z}^{\mathrm{T}}, \tag{5}$$

where $\mathbf{d} = [d_{ij}]_{|D|\times n}$ is the original data set matrix with dimensions $|D|\times n$, and $\tilde{z}^{\mathrm{T}}$ is the transpose of the matrix $\tilde{z}$ with dimensions $n\times|\mathbf{G}|$. The reduced feature matrix $\mathbf{r}$ results a $|D|\times|\mathbf{G}|$ matrix, where $|\mathbf{G}|$ is the number of total clusters. The n dimensions of original data features thus are reduced to |**G**| dimensions.

4 EVALUATION

To validate the feasibility of the proposed data reduction method, a popular high-dimensional application, document classification, was considered. Three well-known document sets, 20Newsgroups (20Newsgroup, 2013), Cade12 (Cade, 2014), and the Reuters Corpus Volume 1 (RCV1, 2004) were used to evaluate the effectiveness of the proposed scheme. The experiments were conducted and evaluated on a computer with Intel Core i7-2600 3.40GHz CPU and 16GB RAM. The programming tool is MATLAB7. 13.0 (R2011b).

The setup of experiments was designed and built by the following steps. Given a set of documents, the keywords first are extracted and analyzed using the textual processing tool - WVtools (VMtools, 2013). Then, the number of keywords in each document is counted to form the original dataset D. The total number of distinct keywords for all documents, n, is the dimensions of D. To accomplish the objective of text categorization, the cosine similarity function is used in the primitive cluster incremental algorithm to measure the similarity degree between two documents, as follows.

$$Sim(d_i, d_j) = \frac{\sum_{k=1}^{n}(d_{ik} \times d_{jk})}{\sqrt{\sum_{j=1}^{n}(d_{ik})^2} \times \sqrt{\sum_{j=1}^{n}(d_{jk})^2}}, \qquad (6)$$

The resulting set of clusters for each document class was obtained after setting a specific threshold ρ to the clustering algorithm. The total clusters are used to compute the normalized relative discriminant variable and generate the reduced matrix **r**. The **r** was taken as the training data to build multiple classifiers by one-against-all strategy using support vector machines (LIBSVM, 2013). To classify K document categories, two types of classification models were built. The first type is to learn one classifier for each categories. Totally K classifiers are learned in the model. The second type is to build a classifier for every cluster we got. The total number of classifiers is |**G**|.

While classifying an unknown document, we first extract its keywords to get the matrix $\mathbf{t} = [t]_{1\times n}$. The reduction process is then applied to **t**, such that

$$\mathbf{t}' = \mathbf{t} \cdot \tilde{z}^{\mathrm{T}}, \qquad (7)$$

The classification model will use **t′** to determine the category of the unknown document.

The effectiveness of document categorization for multiple classifiers is evaluated by the measures of microaveraged precision (*MicroP*), microaveraged recall (*MicroR*), microaveraged F1 (*MicroF*1), and microaveraged accuracy (*MiacroAcc*). (Jiang, 2010) In order to observe the effectiveness of the proposed data reduction method, the classification results using the original full keywords taken from (Jiang, 2010) are shown in Table 1 as baseline.

Table 1: The results using original features (in %).

Datasets	20Newgroups	Cade12	RCV1
Features #	25,718	122,607	47,152
MicroP	94.53	69.57	86.66
MicroR	73.18	40.11	75.03
*MicroF*1	82.50	50.88	80.43
MicroAcc	98.45	93.55	98.83

Experiment 1: 20Newsgroups dataset

This data set consists of 20,000 news messages. The original document set is partitioned evenly across 20 different categories of newsgroups. Two-thirds of the dataset were selected as training set. The others are testing documents. The version here got 25,828 features after the pre-processing of WVtools. That is to say, the original data matrix is a 20,000 × 25,828 matrix. The number of reduced features was determined by the number of clusters which is obtained by setting threshold ρ. Generally, the larger ρ is, the more number of clusters will be generated in the proposed clustering algorithm.

The experimental results of 20Newsgroups are shown in Table 2 and Table 3. The second column in the table lists the results of taking original classes as document clusters directly without further clustering.

The tables illustrates that the *MicroP* values decrease as the number of clusters increases. However, the *MicroR* and *MicroF*1 values show that the reduced dataset with 56 features gets the best results. In comparison with the results of Table 1 using full features, The proposed data reduction method gets an excellent performance in recall and F-measure. The difference on the measures between Jiang's and this paper should be the setting of parameters of SVM learners. Generally, the results of using reduction dataset is even better than the original full features. The main reason is that the unique set of keywords in document categories can be extracted effectively.

Table 2: 20Newsgroups dataset with 20 classifiers (in %).

Features #	20	56	94	195	297
ρ	-	0.100	0.120	0.140	0.150
MicroP	89.15	88.84	88.28	86.87	85.54
MicroR	79.95	81.25	80.50	80.42	80.42
*MicroF*1	84.30	84.87	84.21	83.52	82.95
MicroAcc	98.55	98.52	98.49	98.41	98.34

Table 3: 20Newsgroups dataset with |**G**| classifiers (in %).

Features #	20	56	94	195	297
ρ	-	0.100	0.120	0.140	0.150
MicroP	89.15	88.87	88.15	87.21	86.26
MicroR	79.95	80.53	80.07	79.41	78.38
*MicroF*1	84.30	84.50	83.91	83.13	82.13
MicroAcc	98.55	98.52	98.46	98.33	98.29

Experiment 2: Cade 12 Dataset

The Cade12 is a set of classified web pages. This dataset is classified into 12 categories. There are totally 40,983 documents in this dataset. This benchmark selects 27,322 documents as the training set, and 13,661 documents are used for testing. The distribution of documents in the 12 categories is not as uniform as the 20Newsgroups dataset and the numbers of documents for the 12 categories are very different in quantity. After textual pre-processing, 157,483 features were got totally from the Cade12 dataset.

Table 4: Cade12 dataset with 12 classifiers (in %).

Features #	12	190	236	316	652
ρ	-	0.0005	0.0010	0.0050	0.0100
MicroP	71.66	68.53	68.50	67.75	65.51
MicroR	45.92	52.62	52.88	53.25	54.74
*MicroF*1	55.97	59.53	59.69	59.63	59.64
MicroAcc	93.98	94.04	94.05	93.99	93.82

Table 5: Cade12 dataset with |**G**| classifiers (in %).

Features #	12	190	236	316	652
ρ	-	0.0005	0.0010	0.0050	0.0100
MicroP	71.66	75.82	67.08	66.29	64.06
MicroR	45.92	32.66	45.11	44.67	43.93
*MicroF*1	55.97	45.66	53.49	53.37	52,12
MicroAcc	93.98	93.59	93.58	93.46	93.27

Table 4 and Table 5 list the experimental results of Cade12. The results show that the *MicroR* values increase rapidly in this dataset as the number of features is increasing. On the contrary, the *MicroP* values decrease slowly. Hence, the *MicroF*1 measure was improved in the larger number of clusters. While comparing with the result of full features in Table 1, the reduced dataset can improve recall and F-measure significantly. Generally, the results of using reduction dataset is more effective than using original full features in Cade12 dataset.

Experiment 3: Reuters Corpus Volume 1 Dataset

The Reuters Corpus Volume 1 (RCV1) dataset consists of 804,414 news stories produced by Reuters from 20 Aug. 1996 to 19 Aug. 1997. The set of documents are divided into 23,149 training documents and 781,265 testing documents. The characteristics of this dataset are large number of categories and multi-label for documents. There are 101 non-empty categories totally. All the documents are categorized into one or more classes. There are 47,152 features for this dataset.

Table 6: RCV1 dataset with 101 classifiers (in %).

Features #	101	120	169	213	315
ρ	-	0.0300	0.0500	0.0600	0.0750
MicroP	86.77	86.54	85.78	85.07	83.92
MicroR	68.78	68.99	69.70	69.96	70.67
*MicroF*1	76.74	76.77	76.91	76.78	76.72
MicroAcc	98.66	98.66	98.66	98.64	98.62

Table 7: RCV1 dataset with |**G**| classifiers (in %).

Features #	101	120	169	213	315
ρ	-	0.0300	0.0500	0.0600	0.0750
MicroP	86.77	86.69	86.03	84.71	81.94
MicroR	68.78	68.75	68.82	69.30	70.79
*MicroF*1	76.74	76.69	76.47	76.23	75.96
MicroAcc	98.66	98.66	98.64	98.61	98.65

The experimental results of RCV1 are shown in Table 6 and Table 7. The results in this dataset are not so ideal like the previous two datasets. It is similar to previous two datasets, the *MicroP* values decrease as the number of features increases; on the contrary, the *MicroR* values increase reversely. However, all the measures are not as well as the results of using full features in Table 1. Such an outcome may be caused by several possible reasons. First, since the number of document categories is large, the one-against-all learning strategy will lead to the problem that the number of positive examples is much less than negative ones. The classification model will be dominated by negative examples. Second, the same data appear at different categories simultaneously due to the documents are multi-label. The feature selection using the relative discriminant variables cannot handle the recognition of multi-class well at this moment.

5 SUMMARY AND DISCUSSION

The problem of high-dimensional data not only increase the computation time of process but also degrade the effectiveness of utilization. This paper proposes a novel scheme of data reduction by incorporating of the data clustering approach and feature selection techniques. The proposed scheme includes a primitive incremental clustering algorithm and a discerning method of selecting features based on relative difference. The evaluation has shown that the proposed method is effective for different types of single-label dataset. However, it still needs more investigation on discerning the distinction among the features for multi-label problem.

The advantages of the proposed scheme are discussed as follows. First, the number of reduced dimensions can be controlled by the threshold ρ in the incremental clustering algorithm easily. Second, the scheme is scalable since the relative discriminant variable for each feature can be calculated independently. The computation will not be limited by the size of memory space or software tools. Third, unlike conventional feature selection methods, the final reduced features are the combinations of all possible significant features instead of a set of single features from original datasets.

The process of high-dimensional features is the key problem for many modern applications, such as text classification, information retrieval, social network, and web analysis. The increment of data including data rows and feature columns is a common characteristic in applications of big data. It is worthy to make further investigation on extending the proposed scheme to keep effective data reduction and efficient adaptation along with the increase of data. Developing effective dynamic data reduction solution should be considered as an important issue in the future.

ACKNOWLEDGEMENTS

This research was supported in part by National Science Council of Taiwan, R. O. C. under contract NSC 102-2221-E-024-016.

REFERENCES

20Newsgroups, 2013. http://people.csail.mit.edu/jrennie /20Newsgroups/

VMtools, 2013. http://sourceforge.net/projects/wvtool/

Cade, 2014. http://web.ist.utl.pt/acardoso/datasets/

LIBSVM, 2013. http://www.csie.ntu.edu.tw/~cjlin/libsvm/

RCV1, 2004. http://jmlr.org/papers/volume5/lewis04a/ lewis04a.pdf.

Baker, L. D., McCallum, A., 1998. Distributional clustering of words for text classification. In *ACM SIGIR98 the 21st Annual International*, pp. 96-103.

Bekkerman, R., El-Yaniv R., Tishby N., Winter Y., 2003. Distributional word clusters vs. words for text categorization. *Journal of Machine Learning Research*, vol. 3, 1183-1208.

Blum, A. L., Langley, P., 1997. Selection of relevant features and examples in machine learning. *Aritficial Intelligence*, vol. 97, no.1-2, 245-271.

Combarro, E. F., Montâ*n*és, E., Díaz, I., Ranilla, J., Mones, R., 2005. Introducing a family of linear measures for feature selection in text categorization. *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 9, 1223-1232.

Daphne, K., Sahami, M., 1996. Toward optimal feature selection. In *the 13th International Conference on Machine Learning*, pp. 284-292.

Hsu, H. H., Hsieh, C. W., 2010, Feature selection via correlation coefficient clustering. *Journal of Software*, vol. 5, no. 12, 1371-1377.

Jiang, J. Y., Liou, R. J., Lee, S. J., 2011. A fuzzy self-constructing feature clustering algorithm for text classification. *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 3, 335-349.

Jolliffe, I. T., 2002. *Principal Component Analysis*, 2nd edition, Springer, 2002.

Kriegel, H. P., Kröger, P., Zimek, A., 2009. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data*, vol. 3, no. 1, 1-58.

Liu, H., Yu, L., 2005. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, , 17(4), 491-502.

Martinez, A. M., Kak, A. C., 2001. PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, 228-233.

Park, H., Jeon, M., Rosen, J., 2003. Lower dimensional representation of text data based on centroids and least squares. *BIT Numberical Math*, vol. 43, 427-448.

Salton, G., McGill, M. J., 1983. *Introduction to modern retrieval*. McGraw-Hill Book Company.

Slonim, N., Tishby, N. 2001. The power of word clusters for text classification. In *23rd European Colloquium on Information Retrieval Research* (Vol. 1).

Yang, Y., Pedersen, J. O., 1997. A comparative study on feature selection in text categorization. In *the 14th International Conference on Machine Learning*, pp. 412-420.