ISSN: 2442-6571

K-means clustering based filter feature selection on high dimensional data

Dewi Pramudi Ismi[a,1]*, Shireen Panchoo[b,2], Murinto[c,3]

[a] Department of Informatics, Faculty of Industrial Technology, Universitas Ahmad Dahlan, Yogyakarta 55164, Indonesia
[b] Department of Business Informatics and Software Engineering, University of Technology, Mauritius
[c] Department of Informatics, Faculty of Industrial Technology, Universitas Ahmad Dahlan, Yogyakarta 55164, Indonesia
[1] dewi.ismi@tif.uad.ac.id *; [2] s.panchoo@umail.utm.ac.mu; [3] murintokusno@tif.uad.ac.id
* corresponding author

ARTICLE INFO

Article history:
Received February 20, 2016
Revised March 20, 2016
Accepted March 21, 2016

Keywords:
feature selection,
dimensionality reduction,
clustering,
k-means clustering,
classification,
high dimensional data

ABSTRACT

With hundreds or thousands of features in high dimensional data, computational workload is challenging. In classification process, features which do not contribute significantly to prediction of classes, add to the computational workload. Therefore the aim of this paper is to use feature selection to decrease the computation load by reducing the size of high dimensional data. Selecting subsets of features which represent all features were used. Hence the process is two-fold; discarding irrelevant data and choosing one feature that representing a number of redundant features. There have been many studies regarding feature selection, for example backward feature selection and forward feature selection. In this study, a k-means clustering based feature selection is proposed. It is assumed that redundant features are located in the same cluster, whereas irrelevant features do not belong to any clusters. In this research, two different high dimensional datasets are used: 1) the Human Activity Recognition Using Smartphones (HAR) Dataset, containing 7352 data points each of 561 features and 2) the National Classification of Economic Activities Dataset, which contains 1080 data points each of 857 features. Both datasets provide class label information of each data point. Our experiment shows that k-means clustering based feature selection can be performed to produce subset of features. The latter returns more than 80% accuracy of classification result.

I. Introduction

With the opportunities that Machine learning offers, research is extending rapidly to all major areas where processing manual data used to be a constraint. Machine learning studies how computers can extract knowledge from data, and then is able to make predictions on the new data based on the knowledge that has been owned. Technically, the knowledge obtained from prior data is used to formulate objective function. The objective function is usually called as model. In supervised learning, the objective function associates each input data to the corresponding class. Mapping process of the input data to the appropriate class label (target class) is called as classification. Moreover, the objective function that is used for this mapping called as a classifier. Input of the objective function is data that has one or more features or variables. In high-dimensional data, there are hundreds or thousands of features. In fact, not all features on high-dimensional data are relevant or important for the objective function. That is, if these irrelevant features are removed from the input data, it will not affect the results of the objective function. Besides irrelevant features, the high dimensional can also contain redundant features, meaning that several features have same effect on the results of the objective function. To reduce the computation workload of high dimensional data, such features can be represented by just one feature.

Feature selection algorithm is an algorithm that search for most essential features of high dimensional data. Features that are not considered as essential features (irrelevant, redundant) will be deleted. This deletion aims at reducing computational workload of high dimensional data, so as to

DOI: http://dx.doi.org/10.12928/ijain.v2i1.54

fasten the computation of objective function in the classification process. The most essential features are expected to represent the whole feature set of high dimensional data and are sufficient to be used in predicting class labels (an output of the objective function) accurately. The advantage of applying feature selection algorithm is to reduce the complexity of high-dimensional data computation without changes in the resulted class labels. Feature selection algorithms that have existed in the current literature, divided into three groups; feature selection algorithms that are focused on removing irrelevant features; feature selection algorithms that are focused on the removal of redundant features; feature selection algorithm that focuses on removing features both irrelevant and redundant features.

Clustering is one of the most efficient methods and k-means is commonly used because of its simplicity and also it is relevant to be used for large number of variables [1]. In this work, k-means clustering method is used to perform feature selection, aiming at producing reduced dataset which only consists of essential features, diminishing the size of high dimensional data. Therefore the feature selection algorithm accelerates the classification process.

II. Literature Review

A. *Feature Selection*

Feature selection is a commonly used technique to reduce computational efforts in the processing of high dimensional data. It involves the searching and evaluation of various feature subsets using defined criterion [2]. The process of learning in high-dimensional data has high computational load because of the large dimensions of the dataset. Feature selection is an attempt to identify the features that are most significant producing the result of objective function that is similar to which produced by the original feature set [3] [4]. Feature selection aims to take some of the features which are considered as the most important thus reducing the computational load on the process of learning. According to[3][4] feature selection process focuses on two features namely:

1. Features which are not relevant to the outcome of the objective function, meaning any value that of the feature does not alter the target class.
2. Features that are redundant, that is some features that have similar influence towards results of the objective function.

The evaluation criteria are commonly based on the wrapper model and the filter model [5] [6]. And, according to [7] [8], there are three types of feature selection algorithms:

1. Filter Feature Selection
   This feature selection algorithm assesses relevancy of a feature based on the intrinsic properties of the data. In this case, the feature selection algorithm stands alone and is apart from classifier or objective function. Feature selection algorithm is executed as a pre-processing step of the whole learning process.

2. Wrapper Feature Selection
   This feature selection algorithm does not only assess the relevancy of features based on the intrinsic properties of the data, but also directly evaluate the results of the objective function (classifier) for each of the relevant features.

3. Embedded Feature Selection
   This feature selection algorithm is integrated into the objective function (classifier) such that feature selection algorithm does not stand alone.

Each of the above types of feature selection has shortcomings [7][9]. Disadvantage of filter feature selection is that it does not involve the object function (classifier) to determine which features are most important hence it is possible that the results of the objective function on selected features are not same as the results of the objective function on original features of the dataset. Disadvantage of wrapper feature selection is high computational complexity such that the execution time is long. Whereas the disadvantage of embedded feature selection is that it depends greatly on the type of classification model, thus it cannot be generalized for different type of classification model.

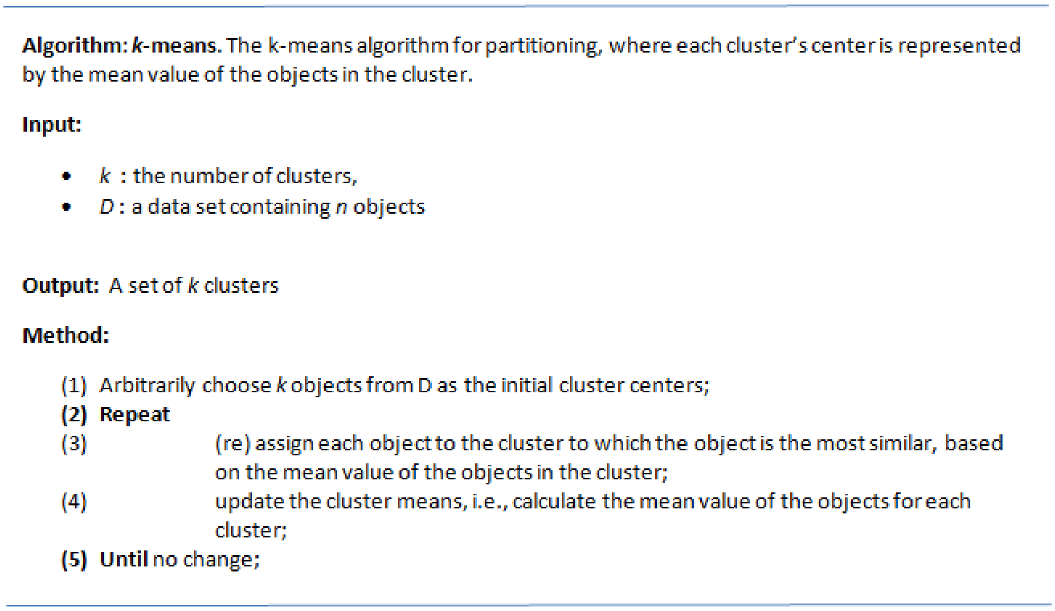
B. *Clustering Algorithm*

Clustering has proved to be commonly used in exploring data, creating predictions and in overcoming the anomalies in the data [8] [10]. Clustering is a process of grouping similar data into a cluster such that dissimilar data belongs to different clusters [4]. Clustering is intuitively used by human to differentiate objects. This is done by observing the characteristics owned by each object. Similarity between objects can be calculated using various formulas, for example Euclidean distance, Manhattan distance, cosine similarity, etc. Clustering method can be classified into several categories based on the approaches applied in the clustering process [3][11].

1. Partitioning Method
   In the Partitioning method, *n* number of objects are divided into k groups where k < n. Each cluster has to possess one object at minimum. Each object has to be associated with one cluster only. K-means clustering and k-medoids clustering apply partitioning method.
2. Hierarchical Method
   In this method objects are divided hierarchically. This is done by constructing a tree structure and then grouping the objects into clusters. There are two approaches in hierarchical method. First approach is agglomerative (bottom up), second approach is divisive (top down). In agglomerative fashion, each object initially creates its own cluster and furthermore two nearby clusters make up a new cluster. Eventually all objects belong to one huge cluster. Divisive approach works the other way round.
3. Density based Method
   The method density is used to measure a cluster whereby a cluster is grown until it reaches a certain density. Objects in a certain cluster have to possess a number of neighbors. DBSCAN clustering applies density based method.
4. Model based Method
   In this method initially a model of each cluster is formulated. Data which fit into a model is grouped into a cluster. Data which fit into different models will be located in different clusters. Models are generated using density function or spatial distribution. Expectation Maximization clustering algorithm is one the most popular model based clustering algorithm.

C. *K-means Clustering*

K-means is the one of most widely used clustering techniques and its performance is influenced by the significant feature subset [12][13] [14][15]. K-means clustering divides input data into k clusters and determines centroid of a certain cluster by calculating the mean of data located in that cluster. K-means clustering is initiated using random cluster centroids, and then each input data is associated to a cluster based on its distance to cluster centroids. An input data belongs to a cluster where the centroid is closest to it. K-means clustering includes repetition of two steps, namely associating an input data to a cluster with nearest cluster centroid, and updating centroid of a cluster with the mean of all data located in that cluster [9][16]. This repetition is done until a convergence criterion is met. Fig. 1 shows the detail of k-means clustering algorithm.



Algorithm: ***k*-means.** The k-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input:

- *k* : the number of clusters,
- *D* : a data set containing *n* objects

Output: A set of *k* clusters

Method:

```
Arbitrarily choose k objects from D as the initial cluster centers;
Repeat
        (re) assign each object to the cluster to which the object is the most similar, based
        on the mean value of the objects in the cluster;
        update the cluster means, i.e., calculate the mean value of the objects for each
        cluster;
Until no change;
```

Fig. 1 k-means clustering [3]

Research done shows that k-means clustering algorithm is more accurate for dense dataset compared to sparse dataset [10] [17]. The same algorithm was used to demonstrate that time complexity can be reduced without sacrificing the accuracy of the clusters [18] [19].

III. Methodology

To address the problem of the curse of dimensionality, feature selection technique is widely used [20] [21]. Our methodology addresses this aspect and our experiment was conducted in this regard. Several steps were performed namely, dataset collection, performing k-means clustering for feature selection, building classification model using the reduced feature set, and evaluation of the generated classification model.

A. Dataset Collection

In this work high dimensional data is used. Two different datasets are used in the experiments, both are taken from UCI Machine Learning Dataset Repository [11][6]. The first dataset is Human Activity Recognition Using Smartphones (HAR) Dataset, which contains 7352 data points each of 561 features [5][14]. The second dataset is National Classification of Economic Activities (Classificação Nacional de Atividade Econômicas - CNAE) Dataset, which contains 1080 data points each of 857 features [2][15]. Both datasets provide class label information of each data point.

B. Performing K-means Clustering for Feature Selection

In this step, k-means clustering algorithm is performed to divide features of the original dataset into predefined number of clusters. The original dataset and the number of intended clusters are taken as input parameters of the clustering algorithm. The objective of applying k-means clustering in this work is to group features, in which data points possess similar values, into a cluster. Assume that the dataset used has *D* features and size of the dataset used is *N*. Performing k-means clustering over this dataset aims at producing feature set of size *k*, where $k < D$, which is representative with regard to the whole feature set. These *k* features consist of the centroids of clusters resulted after convergence criteria are met.

The clustering process divides the features of original data into clusters, hence partitioning is done horizontally. The original dataset is transformed into its transpose matrix to be used as input of the clustering algorithm. Fig. 2 shows the clustering processing carried out in this step.

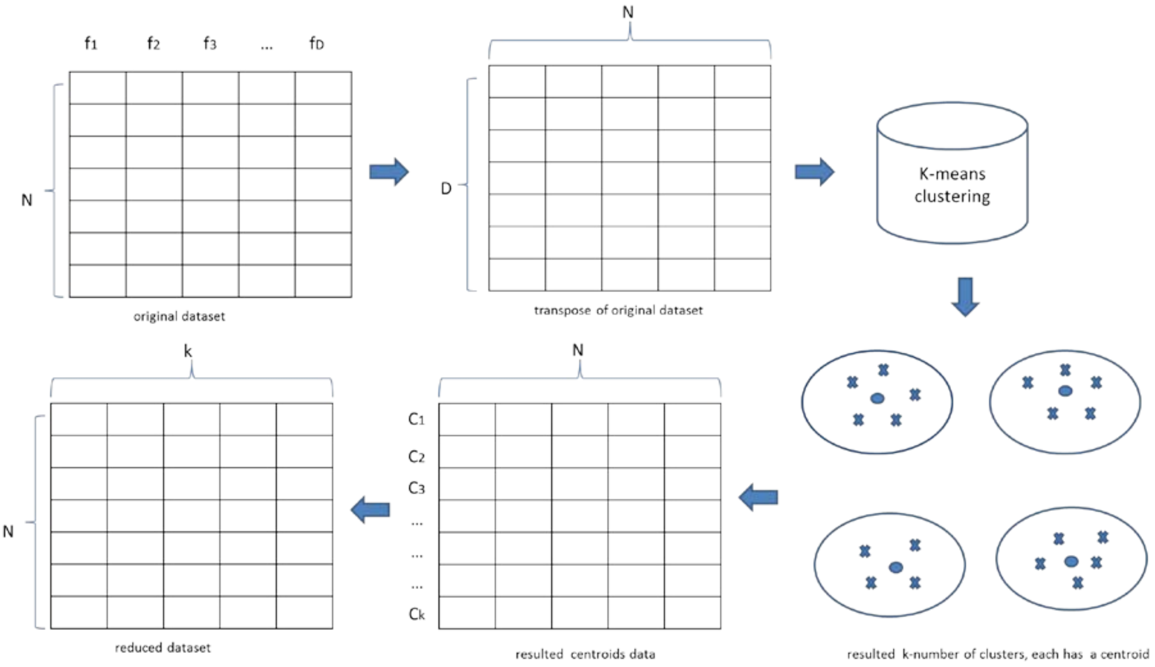


Fig. 2 Feature Selection Process

The output of this step is to reduce data of which the size is *N* x *k*. These *k* features are then employed to build classification model in the next step.

C. *Building Classification Model using Reduced Feature Set*

The reduced dataset resulted from previous step is employed to build the classification model. In this step, class labels provided in the original dataset are used to build the model in supervised manner. Naïve Bayes classifier is the classification model built in this work.

D. *Evaluation of The Generated Classification Model*

The objective of this step is to verify how valid the generated classification model is, so as to compare the performance of classification model generated using original dataset and the performance of classification model generated using the reduced dataset. 10 folds cross validation is done to validate performance of the model. Performance of the model is measured by the number of correctly classified instances and the number of incorrectly classified instances.

IV. Result and Discussion

The endeavor of this work is based on the machine learning research which aims at improving learning algorithms that tackles the problems of high dimension dataset. Before the performance of generated models is analyzed, the performance of models generated using original datasets and class labels are measured. It is done to gain references for the performance analysis of models generated using reduced datasets.

A. *Performance of The Generated Model*

Prior to performance evaluation of the generated model, the performance of Naïve Bayes classifier generated using original dataset (including the whole feature set) is measured. This performance is used as reference to assess the performance of classifier built using the reduced dataset. Our experiment shows that using original HAR dataset to build Naïve Bayes classifier resulted in 76.77% of correctly classified instances and 23.23% of incorrectly classified instances. Moreover, the original CNAE dataset used to build Naïve Bayes classifier resulted in 93.15% of correctly classified instances and 6.85% of incorrectly classified instances.

With the increasing number of data in all sectors and the constraint of time, it is imperative to focus on optimization techniques. Digital data is growing at an incredible rate [12][17]. The objective of our experiments was to minimize the size of feature set and to maximize the classification accuracy using that feature set. Thus in our experiment several number of clusters (*k*) were tried to identify which *k* will produce the highest accuracy of classification. Several different *k* had been used for this exercise where $k < 100$. As shown in Table 1 below, as *k* increases the incorrect classification (%) decreases. When *k* equals to 70 which is the highest classification accuracy of HAR dataset, the result gives 83.35% of correctly classified instances and a percentage of 16.65 incorrect classification.

Table 1 Performance of Naive Bayes Classifier Built Using Reduced HAR Dataset

	Number of features	**Correct Classification (%)**	**Incorrect Classification (%)**
HAR Dataset	10	69.40	30.60
	20	76.81	23.19
	30	79.56	20.44
	40	79.33	20.67
	50	81.38	18.62
	60	79.84	20.16
	70	**83.35**	**16.65**

The same process was done with the CNAE dataset. The result is described in Table 2. It shows the results over several different *k* that had been used where $k < 100$. As the number of features increases, the percentage of correction classification also increases. However, as shown the number *k* equals to 60 gives 88.33% of correctly classified instances, which is the highest classification accuracy of CNAE dataset.

Table 2 Performance of Naive Bayes Classifier Built Using Reduced CNAE Dataset

	Number of features	**Correct Classification (%)**	**Incorrect Classification (%)**
CNAE Dataset	20	79.35	20.65
	30	78.61	21.39
	40	84.17	15.83
	50	82.78	17.22
	60	**88.33**	**11.67**
	70	86.94	13.06

Our experiments show that in general the more features are selected the higher classification accuracy of the generated model. However as the objective is to reduce the dimensionality of high dimensional data such that computational workload is decreased, minimum *k* with decent classification accuracy has to be selected.

B. *Redundant Features Removal*

In this work, redundant features were automatically represented by the most essential feature that is the centroids of clusters produced by k-means clustering algorithm. This result goes in line with the fact that the performance of Naive-Bayes does improve with the removal of relevant features [13][19].

C. *Irrelevant Features Removal*

Our experiment shows that irrelevant features of the original datasets have not been taken into account in the reduced feature sets. This is because irrelevant features are located far from other features in the clusters. Moreover only centroids of clusters are eventually taken for building the classification model, hence there are no irrelevant features used for further classification process.

V. Conclusion

Due to the high computational workload needed to process high dimensional data, much efforts are being used to cater for this problem. In our paper, it was found that K-means clustering can be utilized for reduction of redundant features, as it attempts to group similar features together into one cluster. Similar features could be represented by one representative feature. Furthermore, the value of those features located in the same cluster can be represented by the value of cluster centroid since the centroid is measured by the mean of all features in the cluster. Interestingly, using cluster centroids as representatives of the whole feature set, results in high accuracy of class prediction. Features which are irrelevant to class prediction are automatically not taken into account in class prediction as they are located at the edge of clusters and they are not cluster centroids. This experiment can be extended to various datasets with the aim of understanding the sensitivity and the nature of data analyzed. However with the increasing growth of data and the need for online classifiers, feature selection should be given more emphasis. Analyzing trends and patterns derived from big data will be one of the most important tools in all spheres of life.

Acknowledgment

This work was fully funded by Lembaga Penelitian dan Pengembangan (LPP) of Universitas Ahmad Dahlan, Indonesia.

References

[1] Dhanachandra N., Manglem K. and Chanu Y., Image Segmentation using K-Means clustering algorithm and subtractive clustering algorithm, Eleventh International Multi-Conference on Information Processing 2015, Procedia Computer Science 54, 2015, pp 764-771.

[2] L. Yu and H. Liu. Feature selection for high-dimensional data: a fast correlation-based filter solution. In Proceedings of the twentieth International Conference on Machine Learning, 2003, pp 856–863.

[3] Song, Q., J.Ni, dan G.Wang. A Fast Clustering Based Feature Subset Selection Algorithm for High Dimensional Data. *IEEE Transaction on Knowledge and Data Engineering*. Vol. 25, No. 1, 2013.

[4] Alpaydin, E. 2010. Introduction to Machine Learning, second Edition. The MIT Press. ISBN:978-0-262-01243-0 2.

[5] R. Kohavi and G. John. Wrappers for feature subset selection. Artificial Intelligence, 97(1-2), pp 273–324, 1997.

[6] UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets.html.

[7] Saeys, Y., I.Inza, dan P.Larranaga. A Review of Feature Selection Techniques in Bioinformatics. *Bioinformatics.* Vol. 23, No. 19, pp 2507-2517, 2007.

[8] Arora P., Deepali, Varshney S., 2016. Analysis of K-Means and K-Medoids Algorithm for Big Data, International Conference on Information Security & Privacy, 11-12 Dec. 2015, India, published in Science Direct, Procedia Computer Science 78, 2016, pp. 507-512.

[9] Liu, H. dan L.Yu. Toward Integrating Feature Selection Algorithms for Classification and Clustering. *IEEE Transaction on Knowledge and Data Engineering.* Vol. 17, No. 4, 2005.

[10] Pallavi Purohit and Ritesh Joshi, A New Efficient Approach towards k-means Clustering Algorithm, In International Journal of Computer Applications, (0975-8887), vol. 65, no. 11, March 2013.

[11] Han, J. dan M. Kamber. 2006. Data Mining Concepts and Techniques, *Second Edition*. Morgan Kaufman. ISBN-13:978-1-55860-901-3.

[12] Pavel Berkhin, “Survey of Clustering Data Mining Techniques”, Accrue Software, Inc, 2002.

[13] Shah M. and Nair S., A survey of data mining clustering algorithms, International Journal of Computer Applications, Vol. 128 No. 1, 2015.

[14] Human Activity Recognition Using Smartphones (HAR) Dataset. https://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones.

[15] CNAE-9 Dataset. https://archive.ics.uci.edu/ml/datasets/CNAE-9.

[16] MacKay, D. 2003. Information Theory, Inference and Learning Algorithms. Cambridge University Press. Cambrige – England. pp. 284–292. ISBN 0-521-64298-1.

[17] Qiu et al., 2016, A survey of machine learning for big data processing, EURASIP Journal on Advances in Signal Processing, 2016. doi:10.1186/s13634-016-0355-x.

[18] K. A. Abdul Nazeer and M. P. Sebastian, Improving the Accuracy and Efficiency of the k-means Clustering Algorithm, In Proceedings of the World Congress on Engineering, London, WCE, vol. 1, July 2001.

[19] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. The Journal of Machine Learning Research, Vol. 3, pp. 1157–1182, 2003.

[20] Parsons, L., Haque, E., Liu, H.: Subspace clustering for high dimensional data: a review. SIGKDD Explorations Newsletters. Vol. 6, Issue 1, pp. 90–105, 2004.

[21] Jiliang Tang, Salem Alelyani and Huan Liu, Feature Selection for Classification: A Review, Data Classification: Algorithms and Applications, CRC Press pp. 37, 2014.