Bulletin of Electrical Engineering and Informatics
Vol. 7, No. 3, September 2018, pp. 465~470
ISSN: 2302-9285, DOI: 10.11591/eei.v7i3.1272

Improving Classification Accuracy Using Clustering Technique

Norsyela Muhammad Noor Mathivanan, Nor Azura Md.Ghani, Roziah Mohd Janor
Center for Statistical and Decision Sciences Studies, Faculty of Computer & Mathematical Sciences,
Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia

Article Info

Article history:

Received May 08, 2018
Revised Jul 20, 2018
Accepted Aug 13, 2018

Keywords:

Classification
Clustering
Feature selection

ABSTRACT

Product classification is the key issue in e-commerce domains. Many products are released to the market rapidly and to select the correct category in taxonomy for each product has become a challenging task. The application of classification model is useful to precisely classify the products. The study proposed a method to apply clustering prior to classification. This study has used a large-scale real-world data set to identify the efficiency of clustering technique to improve the classification model. The conventional text classification procedures are used in the study such as preprocessing, feature extraction and feature selection before applying the clustering technique. Results show that clustering technique improves the accuracy of the classification model. The best classification model for all three approaches which are classification model only, classification with hierarchical clustering and classification with K-means clustering is K-Nearest Neighbor (KNN) model. Even though the accuracy of the KNN models are the same across different approaches but the KNN model with K-means clustering had the shortest time of execution. Hence, applying K-means clustering prior to KNN model helps in reducing the computation time.

Corresponding Author:

Nor Azura Md.Ghani,
Center for Statistical and Decision Sciences Studies, Faculty of Computer & Mathematical Sciences,
Universiti Teknologi MARA,
40450 Shah Alam, Selangor, Malaysia.
Email: azura@tmsk.uitm.edu.my

1. INTRODUCTION

Online commerce has rapidly grown since the past decade. The experience of purchasing goods not only from physical stores but also via online shopping. Consumers are provided with shopping ease and flexibility where they are able to search for products using specific keywords and know the product availability. There are millions of products on e-commerce websites such as Amazon, e-Bay, 11street, and Lazada sold by thousands of sellers. On top of that, many new products are registered in these websites every day. The ability of the websites to quickly and accurately retrieve the desired products for the consumers is the key component of being successful [1]. Each product is commonly represented by metadata such as its title, description, category, image, price and so on, where most of them are assigned manually by human sellers. Unlike the title and price, it is possible to automatically classify the product categories from the metadata. The automatic product categorization can reduce the time and economic costs as well as improves the accuracy of category assignment of the same product listed by different sellers [2]. Thus, precisely categorizing products emerged as a key issue in e-commerce domains.

Product classification typically is addressed as a text classification with a large product taxonomy [3]. It is a classic topic for natural language processing where predefined categories are assigned to text inputs using machine learning techniques. The classification is based on the basis of significant words or features extracted from the text document such as the title and description of the products. In particular, there are three main issue when dealing with product classification which are the products sparsely distributed in a large number of categories where the data distribution is quite skewed, the product titles and descriptions vary in length, and there is a possibility that the available pairs of current product title and assigned category are incorrect. Researchers took initiatives by conducting researches to face these problems and come out with good product classification model using different methodologies [1], [4]-[6].

Data pre-processing is a crucial step in dealing with product classification. The main objective in going through this step is to use suitable techniques for transforming original textual data into an understandable format. Pre-processing is important to maintain a good retrieval performance and increase the accuracy of the model. The space for storing the document and time required for processing the data can be efficiently decreased after undergoing this process. It is a complex process that leads to the representation of feature extracted from the textual input. The extraction of key features or key terms helps to enhance the relevancy of word towards the category and document used in the study. Therefore, pre-processing is important to prevent additional problems in classifying the products. The preprocessing step in text mining usually consists three tasks such as tokenization, stop word removal and stemming [6],[7]. Most of the previous studies applied these steps because they are important to make sure data are well-prepared before being used for text classification [8].

There is a need for finding the efficient technique to deal with the increasing amount of large text data sets. An optimal number of attributes or features is required to classify any text documents as the preliminary condition. There are two ways to fulfill this requirement which are feature extraction and selection. In feature extraction, researcher builds a new set of feature space that a more compact based on the original feature space [9]. Meanwhile, the researcher selects a subset of the original feature set in feature selection [10]. It selects features that are able to discriminate samples that belong to different categories. The usefulness of the techniques depends on the data used in a study. Both techniques provide significant impact in increasing the accuracy of the data and efficiency of the processing time.

The dimensionality of the data especially in machine learning and data mining tasks has increased dynamically. High dimensionality data provides difficult challenges to existing learning methods [11]. The presence of a large number of features tends to come out with an over fit learning model where the performance of the model decrease significantly during the prediction phase. Feature selection plays an important role in reducing the dimension of the dataset. The importance of features can be assessed with the availability of label information. If the feature and the class are highly correlated, then the feature is defined to be relevant to a class. It gives an idea to perform the feature selection after clustering the data with a purpose to prevent the curse of dimensionality. Then, the features are used to classify the products using existing classification models. The performance of a classifier depends on the early stage of pre-processing and features used in a study [12]. It is important to ensure that the classification model is based on convenient data pre-processing and reliable features. Hence, this paper aims to improve the classification accuracy with using clustering technique.

2. RELATED WORK

Researchers often face difficulty to apply the conventional procedures when working with product title classification [12]. Many combinations had been done by them to increase the accuracy of a classification model. Recent studies have shown that the combination of classification and clustering models can provide better classification result [13],[14]. Normally, a clustering algorithm is used to group unlabelled data into a homogeneous group based on selected features. It is an important tool to solve unsupervised learning problems. The purpose of the clustering algorithm is classifying the data into groups that share almost similar features.

The application of clustering technique can be seen through various study areas such as health [15], insurance [16] and marketing [17]. This technique has been applied to different fields especially in information retrievals such as text mining, image segmentation, data mining, and pattern recognition [18]. Recently, clustering technique plays an important role for proactive scheduling in cloud computing [19]. There are various clustering algorithms can be used by researchers, but the performance and execution time is different according to which one is used. There are some properties should be fulfilled to make sure the performance of a clustering algorithm is able to deal with noise and uncertainty, data with high dimension and different kind of attributes.

There are two commonly used clustering algorithms which are K-means and Hierarchical clustering techniques. Previous studies often utilized these techniques to enhance the performance of their classification models [13], [14],[20]-[22]. However, these previous studies often used one of the clustering techniques where there is not much comparison had been made between different clustering techniques to improve the performance of classification model. Alapati and Sindhu [13] stated that the K-means clustering over performed hierarchical clustering, but the computation time is not provided in the study. The accuracy of classification model should be improved along with the execution time. This is crucial when researcher deals with high-frequency data. Thus, an efficient method is needed to fulfill both requirements.

3. RESEARCH METHOD

In dealing with text classification, the data has to be pre-processed before applying a classification algorithm. This proposed method utilizes the use of feature selection in reducing the dimensionality of a dataset. The data used for this study have been collected from Tesco online stores using prototype web scrapers developed under STATSBDA project namely Price Intelligence (PI) by Department of Statistics Malaysia (DOSM). From the browse tree of the website, the study selected few leaf nodes where they represent categories. The data corpus used for this study is baby products data corpus. It contains data from four different categories which are baby food, baby toiletries, diapers and wipes, and milk powder. All these nodes altogether have 11419 items and the number of features in the product descriptions are 401. There are several steps involve before classifying the products in the research as shown in Figure 1.

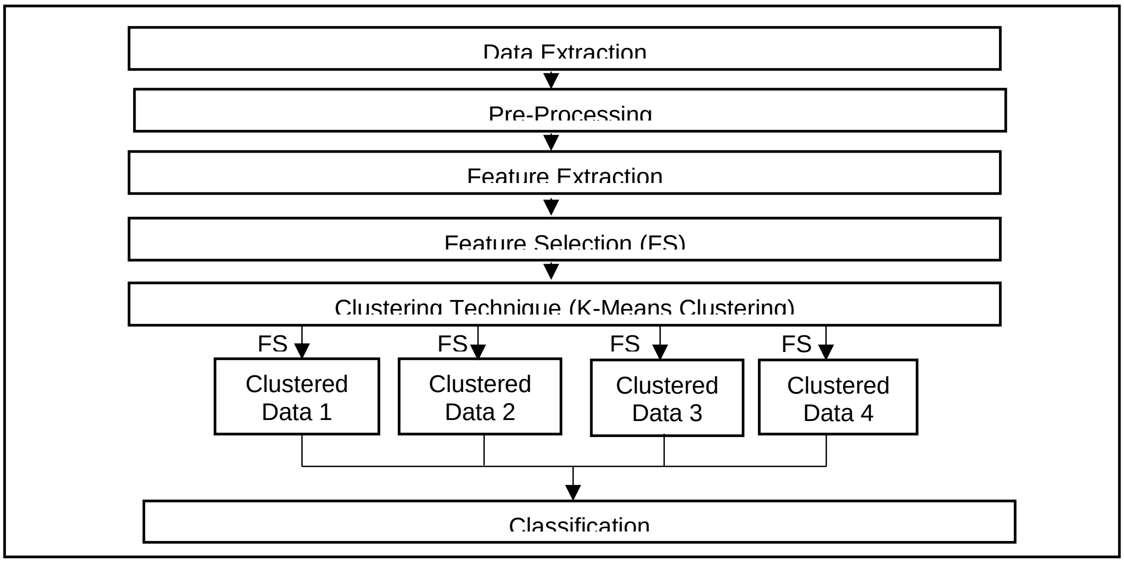

Figure 1. Research methodology

The steps involved are data extraction, data pre-processing, feature extraction, feature selection and clustering technique. After data extraction, there are several steps involve in pre-processing the data such as tokenization, word stop removal, and stemming process. Then, the key steps in this phase are the feature extraction and selection. This study has used bag-of-word to extract the features before performing the feature selection to reduce the dimensionality of the data. One of the commonly used technique for feature selection known as correlation feature selection (CFS) technique is used in this study.

All the steps involved are basic procedures in text mining before training the classifier except for applying the clustering technique to perform the reselection of features. The study used K-means and Hierarchical clustering to cluster the dataset before applying classification model. K-means clustering is based on the simple and understandable procedure to classify a given data according to a predefined number of clusters.

Let $X = \{x_1, x_2, x_3, \dots\dots, x_n\}$ be the set of data points and $V = \{v_1, v_2, v_3, \dots, v_4\}$ be the set of centers

The steps to perform K-means clustering are,

i. Provide the value for data point, n and cluster centers, c where in this study the n=11419 and c=4.
ii. Calculate the distance between each data point, n and cluster centers, c.
iii. Assign the data point, n to the cluster center, c whose distance from the cluster center is minimum compared to others cluster centers.
iv. Recalculate the new cluster using,
$f_i = (1/c_i)\sum_{j=1}^{c_i} x_i$, where c_i represents the number of data points in i^{th} cluster.

v. Recalculate the distance between each data point and new obtained cluster centers.
vi. If there is no reassigned of data point then the calculation is stopped or otherwise, it will be repeated from step (iii).

Let $X = \{x_1, x_2, x_3, \dots \dots ., x_n\}$ be the set of data points.

The steps to perform hierarchical clustering are,
i. The disjoint clustering consists of level $L(0) = 0$ and sequence number o is $m = 0$.
ii. The least distance pair of clusters in the current clustering is identified, such as a pair of (a) and(b), where $d[(a) , (b)] = \min of \ d[(n) , (m)]$ which is the minimum overall pair of clusters in the current clustering.
iii. Increment the sequence number, $m = m + 1$. Clusters (a) and(b) are merged into a single cluster to form the next clustering m. The level of the clustering is set to $L(m)$=$d[(n) , (m)]$.
iv. The distance matrix, D is updated by deleting both rows and columns according to clusters (a) and(b), and adding a row and column according to the new cluster. The distance between new cluster, denoted as (a, b) and old cluster (c) is defined by,
$d[(c) , (a, b)] = min(d[(c) , (a)], d[(c) , (b)])$
v. The calculation is stopped when all the data points are in one cluster, otherwise, it will be repeated from step (ii).

Clustering technique is used to enhance the performance of classification model. The proposed technique is done to overcome the complexity when dealing with high-frequency data. There are several classification algorithms used in the study which are Naive Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Random Forest (RF).

4. RESULTS AND ANALYSIS

Data with high-frequency may slow down the classification process and reduce the accuracy. Accuracy is the ratio of number of instances for which the outcome is correct to the total number of tests made. Table 1 shows the accuracy of a classifier without applying any clustering algorithms before classifying products in the dataset. The dataset is reduced from 401 features to 275 features through feature selection. The accuracy of KNN model is the best to classify the baby products compared to other three classification model.

Table 1. Accuracy of Different Classification Algorithms

Classification Model	Accuracy
NB	0.3992
SVM	0.9887
KNN	1
RF	0.9947

On the other hand, Table 2 shows the accuracy of a classifier after applying hierarchical and K-means clustering algorithms before classifying products in the dataset. When applying feature selection on the clustered dataset, the feature is reduced from 275 features to 257 and 260 for hierarchical and K-means clustering respectively. The accuracy of KNN model is still the best model to classify the baby products compared to other three classification models. As claimed by [23], the KNN model is a simple model but effective in many cases includes text mining. Similarly, Guo et.al [24] pointed out that the KNN model is one of the most effective classification models on a benchmark corpus in text categorization known as Reuters corpus of newswire stories.

Table 2. Accuracy of Different Classification Algorithms with Clustering

Classification Model	Accuracy	
	Classification model with Hierarchical Clustering	Classification model with K-means Clustering
NB	0.4982	0.4889
SVM	0.994	0.994
KNN	1	1
RF	0.9931	0.997

As shown in Table 3, the classification time for executing the best classification model which is KNN model to predict the baby products are different. When K-means clustering algorithm applied to the classification model, the execution time of KNN algorithm decrease compared to the classification model without applying any clustering algorithm. Meanwhile, the execution time sharply increases when Hierarchical clustering applied in the classification model.

Table 3. Performance of K-Nearest Neighbor (KNN) with Clustering

Classification Model	Execution Time (Second)
KNN	84.54
KNN with Hierarchical clustering	337.04
KNN with K-means clustering	60.09

5. CONCLUSION

This paper presents the results on classifying e-commerce products from online store website. The study has evaluated four classification models which are formed with two clustering algorithms known as Hierarchical and K-means clustering algorithms. It can be seen that both clustering algorithms help in improving the accuracy of the classification algorithms. However, the K-means clustering seems to provide more efficient computation time compared to Hierarchical clustering. Hence, the time for classifying the products can be reduced by using K-means clustering in selecting the important features. For future work, researchers plan to expand the size of the corpus and to include more categories for evaluation. It would be interesting to apply the technique towards semi-supervised learning model.

ACKNOWLEDGMENTS

We would like to express our appreciation to the University Teknologi MARA for financial support under the Grant Scheme 600-IRMI/PERDANA 5/3 BESTARI (062/2018). We gratefully thank the Department of Statistics Malaysia for providing knowledge and data supports.

REFERENCES

[1] Z. Kozareva, “Everyone Likes Shopping! Multi-class Product Categorization for e Commerce,” *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 168, pp. 1329–1333, 2015.

[2] A. Cevahir and P. A. South, “Large-scale Multi-class and Hierarchical Product Categorization for an E-commerce Giant,” *Proceedings of the 26th International Conference on Compu tational Linguistics (COLING-16)*, pp. 525–535, 2016.

[3] T. Zahavy, *et al.*, “Is a picture worth a thousand words? A Deep Multi-Modal Fusion Architecture for Product Classification in e-commerce,” pp. 1–10, 2016.

[4] V. Gupta, *et al.*, “Product Classification in e-Commer ce using Distributional Semantics,” 2016. http://arxiv.org/abs/1606.06083

[5] X. Qiu, *et al.*, “Hierarchical Text Classification with Latent Concepts,” *49th Annual Meeting of the Associatin of the Computational Linguistics*, pp. 598–602, 2011.

[6] S. Kumar and R. Karthika, “A survey on text mining process and techniques,” *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, vol/issue: 3(7), pp. 2279–2284, 2014.

[7] S. Vijayarani and R. Janani, “Text Mining: open Source Tokenization Tools – An Analysis,” *Advanced Computational Intelligence: An International Journal (ACII)*, vol/issue: 3(1), pp. 37–47, 2016.

[8] M. K. Dalal and M. A. Zaveri, “Automatic Text Classification: A Technical Review,” *International Journal of Computer Applications*, vol/issue: 28(2), pp. 37–40, 2011.

[9] J. C. Gomez, *et al.*, “Highly discriminative statistical features for email classification,” *Knowledge and Information Systems*, vol/issue: 31(1), pp. 23–53, 2012.

[10] D. Saxena, *et al.*, “Survey on Feature Extraction methods in Object,” *International Journal of Computer Applications*, vol/issue: 166(11), pp. 11–17, 2017.

[11] H. Liu and H. Motoda, “Computational methods of feature selection,” *Computer*, vol/issue: 198(1), pp. 2–13, 2008.

[12] H. Yu, *et al.*, “Product Title Classification versus Text Classification,” *Csie.Ntu.Edu.Tw*, pp. 1–25, 2012.

[13] Y. K. Alapati and K. Sindhu, “Combining Clustering with Classification : A Tech nique to Improve Classification Accuracy,” *Lung Cancer*, vol/issue: 32(57), pp. 3, 2016.

[14] A. Bansal, *et al.*, “Improved K-mean Clustering Algorithm for Prediction Analysis using Classification Technique in Data Mining,” *International Journal of Computer Applications*, vol/issue: 157(6), pp. 975–8887, 2017.

[15] A. Alsayat and H. El-Sayed, “Efficient genetic K-means clustering for health care knowledge discovery,” *2016 IEEE/ACIS 14th International Conference on Software Engineering Research, Management and Applications, SERA 2016*, 2016.

[16] F. Karamizadeh and S. A. Zolfagharifar, “Using the clustering algorithms and rule-based of data mining to identify affecting factors in the profit and loss of third party insurance, insurance company auto,” *Indian Journal of Science and Technology*, vol/issue: 9(7), 2016.

[17] P. Wankhade and R. Shelke, “Analysis Of Clustering Technique In Marketing Sector,” *International Journal For Research In Applied Science & Engineering Technology (Ijraset)*, vol/issue: 5(II), pp. 209–211, 2017.

[18] M. Alaqtash, *et al.*, “A Modified Overlapping Partitioning Clustering Algorithm for Categorical Data Clustering,” vol/issue: 7(1), 2018.

[19] R. Kaur and G. Kaur, “Proactive Scheduling in Cloud Computing,” vol/issue: 6(2), pp. 174–180, 2017.

[20] S. A. Ali, *et al.*, “K-means clustering to improve the accuracy of Decision tree response classification,” *Information Technology Journal*, vol/issue: 8(8), pp. 1256–1262, 2009.

[21] S. D. Sarkar, *et al.*, “A Novel Feature Selection Technique for Text Classification Using Naïve Bayes,” *International Scholarly Research Notices*, pp. 1–10, 2014.

[22] B. Madasamy and J. J. Tamilselvi, “Improving classification Accuracy of Neural Netw- ork through Clustering Algorithms,” vol/issue: 4(9), pp. 3242–3246, 2013.

[23] D. Hand, *et al.*, “Principles of data mining. Drug safety : an international journal of medical toxicology and drug experience,” vol. 30, 2001.

[24] G. Guo, *et al.*, “kNN Model-Based Approach in Classification,” *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, vol. 2888, pp. 986–996, 2003.