The Scientific Temper (2025) Vol. 16 (5): 4215-4219
E-ISSN: 2231-6396, ISSN: 0976-8653
Doi: 10.58414/SCIENTIFICTEMPER.2025.16.5.06
https://scientifictemper.com/

RESEARCH ARTICLE

Clustering of cancer text documents in the medical field using machine learning heuristics

C. Premila Rosy*

Abstract

The data clustering over medical text documents plays a major role in extracting relevant information from the documents. However, most of the methods fail to find the accurate solution for finding the relevant cancer type due to the presence of redundant data items. It is, hence necessary to develop a clustering framework that strictly eliminates the redundant data items. In this paper, we present a clustering framework that tends to accurately cluster the cancer text documents to predict what type of cancer is present in a patient. A large database is tested and clustering using the machine learning model. The clustering framework consists of pre-processing the text documents, feature extraction, feature selection and clustering. The clustering using a multi-support vector machine enables optimal clustering of text documents. The cancer datasets are used to validate the models over various medline cancer document datasets. The experimental validation shows improved clustering of documents using the proposed models than other methods.

Keywords: Machine learning, Soft computing paradigm, Cancer text documents, Redundancy reduction

Introduction

The clustering strategy is generally used to organize a limited collection of clusters, which mostly consist of predefined numbers of clusters that reflect a dataset based on correlations between its objects (Baker *et al*., 2017). As isolation of overlapping clusters is performed, clustering strategies are of low efficiency. In the partitioning clustering process, meta-heuristic optimisation algorithms are used. The partial clustering partitions a dataset into a subset of classes dependent on a specific fitness function (Chen *et al*., 2018). The exercise mechanism directly affects the type of group forming. The partitioning method becomes an optimization problem until an adequate fitness function is selected (Diab and El Hindi, 2017).

PG and Research Department of Computer Science, Idhaya College for Women (Autonomous), Affiliated to Bharathidasan University, Kumbakonam.

***Corresponding Author:** C. Premila Rosy , PG and Research Department of Computer Science, Idhaya College for Women (Autonomous), Affiliated to Bharathidasan University, Kumbakonam, E-Mail: premilarosy@yahoo.com

How to cite this article: Rosy, C.P. (2025). Clustering of cancer text documents in the medical field using machine learning heuristics. The Scientific Temper, **16**(5):4215-4219.

Doi: 10.58414/SCIENTIFICTEMPER.2025.16.5.06

Source of support: Nil

Conflict of interest: None.

Clustering can also be officially viewed from an optimization perspective as a sort of hard-optimization problem (NP) (Dwivedi,2018). The need for functional algorithms of optimization has been promoted, which not only provides ad hoc learning for some groups of problems but also provides the usability of methods of general optimization (Gao *et al*., 2018). Unusually, meta-heuristical and evolutionary optimizing approaches are generally found useful in resolving NP-hard issues, which allows the clustering of problems in a reasonable time to produce almost optimal solutions (optimal clusters)(Gowrishanka*et al*., 2020). According to this assumption, several meta-heuristic algorithms have been implemented in the literature to solve problem clusters, in particular text clusters. These algorithms of optimization are used for optimal purposes (i.e. fitness function) for the control of search improvement (Ho *et al*., 2020).

This paper presents a clustering framework that tends to accurately cluster the cancer text documents to predict what type of cancer is present in a patient. A large database is tested and clustering using the machine learning model. The clustering framework consists of pre-processing the text documents, feature extraction, feature selection and clustering. The clustering using a multi-support vector machine enables optimal clustering of text documents. The cancer datasets are used to validate the models over various medline cancer document datasets.

Clusters are a mathematical tool that tries to identify structures or certain patterns in a dataset where there is

Received: 20/03/2025 **Accepted:** 22/04/2025 **Published** : 31/05/2025

a certain resemblance to the object inside each cluster (Hughes *et al*.,2017). Clustering is a collection of data structures in the same group that are identical to each other and distinct from objects in the other group (Ibrahim *et al*.,2021). Contrary to the classification, class labels are in classification unknown and groups must be determined by the clustering algorithm (Jang *et al*., 2020). Clustering is often referred to as unmonitored learning because the clustering does not rely on established classes (Jasmir *et al*., 2021). The clusters are usually focused on the rules of maximization of similarities between objects of the same classes and minimization of similarity between objects of different classes (interclass proximity).

Massive amounts of data are collected in the modern era is supposed to double the amount of data received. Gaining data volume information is one of the preferred data mining attributes (Jensen *et al*., 2017; Kannan *et al*., 2019; Karthick *et al*., 2021; Kečo *et al*., 2018). There is usually a big difference between the data stored and the information that the data can interpret. These gaps cannot be bridged wherever data extraction enters into the picture immediately. Preliminary information is derived from data in exploratory data studies, but data mining may be helpful by delivering more comprehensive data knowledge. The rapid advancement of IT, engineering and methodology introduces new requirements for the IT paradigm. For some time now, manual data processing has been around, but it causes a significant data analysis bottleneck (Khadidos *et al*., 2020; Liao *et al*., 2021; Nguyen *et al*., 2020; Raja and Kousik, 2021).

Complex natural world definitions also have a vast range of characteristics. The result of the slowly filled space cannot exponentially increase the amount of points accessible by dimensionality, and the discrimination between clusters is often very weak in full-dimensional space (Raja *et al*., 2020; Viloria *et al*., 2021; Yao *et al*., 2019; Yoon and Alawad *et al*., 2018).

Not all dimensions are usually linked in high-dimension spaces: data is combined into a specific cluster around those dimensions. Some clusters in the data decrease the frequency of insignificant features. For a clustering process, clusters that are embedded in sub-spaces that are potentially made of various dimensional mixtures in different data locations are determined.

The large sizes and dimensional databases related to the application of data mining need very scalable clustering algorithms. Sampling and parallelization can be used to enhance scalability. Clustering algorithms at a distance prefer to find clusters of the same size and density. Clustering algorithms to detect groups of arbitrary form, density, data coverage, and size are essential. This would help to develop a better understanding of the various associations between functions, thereby facilitating additional phases of KDD, for example, decision-making (Yoon *et al*., 2018; Yue *et al*., 2018; Yuvaraj *et al*., 2020; Yuvaraj and Srihari *et al*., 2021).

Methodology

The following steps for the extraction and clustering model in this section are presented.

Pre-processing

The main aim of clustering is to generate classes according to the intrinsic contents of the objects. If clusters are formed, text requires pre-processing of model steps.

Tokenization

Tokenization is the act of splitting words into pieces (words), which are called tokens, possibly lacking characters, such as punctuation concurrently. These tokens are typically linked to terms/words, but distinguishing between type/tokens is critical.

Stop Words Removal

Standard vocabulary and common phrases and other popular sentences in the language, which are commonly used and useful short words. These words should be omitted from the document (text) provided because they are extremely redundant in general and hence, reduce the effectiveness of the clustering techniques.

Stemming

Stemming is the period where a common language is shortened to its source. The stem form is not the same as the morphological root method; it is usually the same stem as describing words, but it is not a true root alone.

Document Representation

The vector space model (VSM) is an important model for defining the contents of documents in an official format. It appeared in the early 1970s. Any paper is designed to support the estimation of similarities as a term weight vector. Every word in the text transmits a measure of the weighted

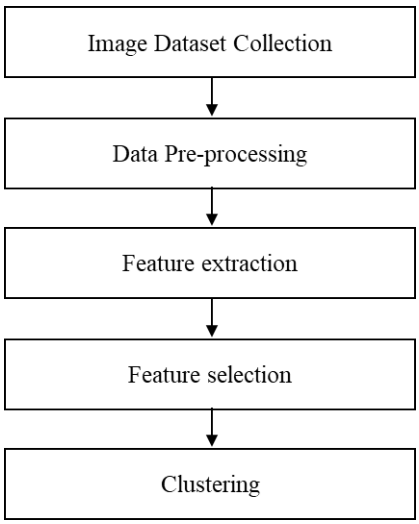

Figure 1: Proposed Classification Model

output in order to increase the efficiency of the clustering algorithm and to reduce time costs.

Feature Extraction

The collection of features is one of the most critical classification system measures. Application filtering is usually used to decrease data size for tens to hundreds of thousands of features that cannot be processed further. The high dimensionality of the function area is a big challenge in web page classification. The best subset of features includes the least number of features that add to the precision and reliability of classification.

MSVM Classification

The MSVM solves the hyper-plane separation between the groups, and it provides the optimal separation between the features in a cancer dataset. If the classification is not linear from the data points, the dataset is divided into multi-distinct classes. The data points are transformed by the nonlinear mapping $\varphi(x)$ over a high dimensional space that resolves the nonlinear problem between the classes x_i. This helps in the separation of points in the solution space using a label y_i. Thus MSVM solves the resultant problem with the feature extracted from the N-point data x_i and it is represented as below:

$$\min_{w,b,\xi_i} 0.5w^T w + c\sum_{i=1}^{N}\xi_i \quad (1)$$

$$\text{s.t. } y_i\left(w^T\phi(x)+b\right) \geq 1-\xi_i \quad (2)$$

where,

ξ_i- slack variables

$c \geq 0$ that defines the tradeoff existing between the computational complexity and training error.

A dual Lagrangianfunction optimizes the MSVM class and it is expressed as below:

$$\min_{\alpha} 0.5\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j K\left(x_i,x_j\right)\alpha_i\alpha_j - \sum_{j=1}^{N}\alpha_i \qquad 0 \leq \alpha_i \leq C \quad (3)$$

$$\text{s.t. } \sum_{j=1}^{N}\alpha_i y_i = 0 \quad (4)$$

where

K - kernel vector and it is expressed as below between two different variable x and y:

$K(x_i,x_j) = <\varphi(x_i), \varphi(x_j)>$ with $<\varphi(x_i), \varphi(x_j)>$ (5)

Hence for a data point x, the predicted class from the text document is given as below:

$$sign\left(\sum_{j=1}^{N}\alpha_i y_i K\left(x,x_i\right)+b\right) \quad (6)$$

The unwanted data samples are removed at the end of the prediction using MSVM, removing the irrelevant support vectors. The prediction limits are hence classified with relevance to the clear and most relevant feature set.

Results and Discussions

In this section, 85% of the data is used for training the clustering algorithm and 15% of the data is used for testing the algorithm. The performance of the model is tested against medline cancer datasets over different performance metrics that include accuracy. The proposed model is compared with the existing three different classifiers, including Random Forest, Naïve Bayes, K-nearest neighbor, and SVM. The M-SVM classifies the multi-data point into relevant classes and a hyperplane provides the separation between the groups that helps in the estimation of relevant clusters.

The accuracy for clustering is expressed as below:

Accuracy = $(TP+TN)$ / $(TP+TN+FP+FN)$ (7)

where,

TP is True Positive rate

TN is True Negative rate

FP is False Positive rate

FN is False Negative rate

FN is False Negative rate

Figure 2 displays the clustering accuracy of various algorithms on the breast cancer Wisconsin (Diagnostic) dataset. It compares the performance of algorithms such as K-NN, naïve bayes (NB), random forest (RF), support vector machine (SVM), and multi-support vector machine (MSVM). The results indicate the effectiveness of each algorithm in accurately clustering the data.

Figure 3 illustrates the clustering accuracy of different algorithms on the breast cancer dataset. It provides a comparative analysis of the performance of K-NN, naïve bayes (NB), random forest (RF), SVM, and MSVM. The figure justifies the effectiveness of these algorithms in accurately

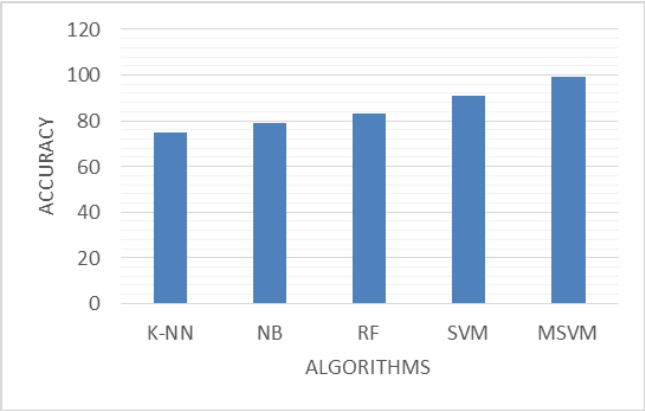

Figure 2: Clustering accuracy on breast cancer Wisconsin (Diagnostic)

clustering the dataset, highlighting their respective accuracy rates.

Figure 2 shows the clustering accuracy of breast cancer Wisconsin (Diagnostic). Figure 3 shows the clustering accuracy on the breast cancer dataset. Figure 4 shows the clustering accuracy on the medline cancer dataset. Figure 5 shows the clustering accuracy on the Haberman breast cancer survival dataset. Figure 6 shows the clustering accuracy on the lung cancer dataset. From the results, it is seen that the proposed method achieves a higher rate of accuracy on all datasets than other methods.

Conclusion

The conclusion emphasizes the significance of effective clustering algorithms in the context of data mining, particularly for complex datasets like those related to breast cancer. It highlights that the proposed methods enhance the understanding of data relationships, which is crucial for informed decision-making in healthcare. Furthermore, the study underscores the necessity for scalable algorithms that can handle high-dimensional data, ensuring that insights derived from such data can bridge the gap between raw data and actionable information, ultimately improving outcomes in medical diagnostics and treatment strategies.

Acknowledgments

The authors extend their thanks and gratitude to the journal editors and reviewers for their valuable comments.

References

Baker, S., Ali, I., Silins, I., Pyysalo, S., Guo, Y., Högberg, J. and Korhonen, A. (2017). Cancer Hallmarks Analytics Tool (CHAT): A text-mining approach to organize and evaluate scientific literature on cancer. Bioinformatics, 33(24), 3973-3981.

Chen, D., Qian, G., and Pan, Q. (2018). Breast cancer classification with electronic medical records using hierarchical attention bidirectional networks. 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 983-988. IEEE.

Diab, D. M., and El Hindi, K. M. (2017). Using differential evolution for fine tuning naïve Bayesian classifiers and its application for text classification. Applied Soft Computing, 54, 183-199.

Dwivedi, A. K. (2018). Artificial neural network model for effective cancer classification using microarray gene expression data. Neural Computing and Applications, 29(12), 1545-1554.

Gao, L., Ye, M., and Wu, C. (2017). Cancer classification based on support vector machine optimized by particle swarm optimization and artificial bee colony. Molecules, 22(12), 2086.

Gowrishankar, J., Narmadha, T., Ramkumar, M., and Yuvaraj, N. (2020). Convolutional Neural Network Classification On 2d Craniofacial Images. International Journal of Grid and Distributed Computing, 13(1), 1026-1032.

Ho, C. Y., Syamsudin, M., and Shen, Y. C. (2020). Cancer literature classification methods performance. 2020 International Conference on Decision Aid Sciences and Application (DASA), 801-805. IEEE.

Hughes, M., Li, I., Kotoulas, S., and Suzumura, T. (2017). Medical text classification using convolutional neural networks. Studies in Health Technology and Informatics, 235, 246-250.

Ibrahim, M. A., Khan, M. U. G., Mehmood, F., Asim, M. N., and Mahmood, W. (2021). GHS-NET a generic hybridized shallow neural network for multi-label biomedical text classification. Journal of Biomedical Informatics, 116, 103699.

Jang, B., Kim, M., Harerimana, G., Kang, S. U., and Kim, J. W. (2020). Bi-LSTM model to increase accuracy in text classification: Combining Word2vec CNN and attention mechanism. Applied Sciences, 10(17), 5841.

Jasmir, J., Nurmaini, S., Malik, R. F., and Tutuko, B. (2021). Bigram feature extraction and conditional random fields model to improve text classification clinical trial document. Telkomnika, 19(3),15-25.

Jensen, K., Soguero-Ruiz, C., Mikalsen, K. O., Lindsetmo, R. O., Kouskoumvekaki, I., Girolami, M., and Augestad, K. M. (2017). Analysis of free text in electronic health records for identification of cancer patient trajectories. Scientific Reports, 7(1), 1-12.

Kannan, S., Dhiman, G., Natarajan, Y., Sharma, A., Mohanty, S. N., Soni, M., and Gheisari, M. (2021). Ubiquitous vehicular ad-hoc network computing using deep neural network with IoT-based bat agents for traffic management. Electronics, 10(7), 785.

Karthick, S., Rajakumari, P. A., and Raja, R. A. (2021). Ensemble similarity clustering framework for categorical dataset clustering using swarm intelligence. In Intelligent Computing and Applications (pp. 549-557). Springer.

Kečo, D., Subasi, A., and Kevric, J. (2018). Cloud computing-based parallel genetic algorithm for gene selection in cancer classification. Neural Computing and Applications, 30(5), 1601-1610.

Khadidos, A., Khadidos, A. O., Kannan, S., Natarajan, Y., Mohanty, S. N., and Tsaramirsis, G. (2020). Analysis of COVID-19 infections on a CT image using DeepSense model. Frontiers in Public Health, 8.

Liao, Y., Peng, Y., Liu, D., and Liu, J. (2021). Intelligent classification of breast cancer based on deep learning. Journal of Physics: Conference Series, 1827(1), 012171. IOP Publishing.

Nguyen, E., Theodorakopoulos, D., Pathak, S., Geerdink, J., Vijlbrief, O., Van Keulen, M., and Seifert, C. (2020). A hybrid text classification and language generation model for automated summarization of Dutch breast cancer radiology reports. 2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI), 72-81. IEEE.

Raja, R. A., and Kousik, N. V. (2021). Privacy preservation between privacy and utility using ECC-based PSO algorithm. In Intelligent Computing and Applications, 2021(1), 567-573)..

Raja, R. A., Karthikeyan, T., and Kousik, N. V. (2020). Improved privacy preservation framework for cloud-based Internet of Things. In Internet of Things: Integration and Security Challenges 2020(1). 155-165.

Viloria, A., Alberto, N., and Pinillos-Patiño, Y. (2021). Classification of clinical reports for supporting cancer diagnosis. In Proceedings of International Conference on Intelligent Computing, Information and Control Systems (pp. 421-428). Springer.

Yao, L., Mao, C., and Luo, Y. (2019). Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. BMC Medical Informatics and Decision Making, 19(3), 71.

Yoon, H. J., Alawad, M. M., and Tourassi, G. (2018). Multi-task convolutional neural networks for natural text classification (Technical Report No. 005841WKSTN00). Oak Ridge National Laboratory.

Yoon, H. J., Robinson, S., Christian, J. B., Qiu, J. X., and Tourassi, G. D. (2018). Filter pruning of convolutional neural networks for text classification: A case study of cancer pathology report comprehension. 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), 345-348. IEEE

Yue, W., Wang, Z., Chen, H., Payne, A., and Liu, X. (2018). Machine learning with applications in breast cancer diagnosis and prognosis. Designs, 2(2), 13.

Yuvaraj, N., Raja, R. A., Kousik, N. V., Johri, P., and Diván, M. J. (2020). Analysis on the prediction of central line-associated bloodstream infections (CLABSI) using deep neural network classification. In Computational Intelligence and Its Applications in Healthcare (pp. 229-244). Academic Press.

Yuvaraj, N., Srihari, K., Dhiman, G., Somasundaram, K., Sharma, A., Rajeskannan, S., and Masud, M. (2021). Nature-inspired-based approach for automated cyberbullying classification on multimedia social networking. Mathematical Problems in Engineering, 2021.