ISSN: 2277-9655
Impact Factor: 1.852

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Performing Age Group Clustering in Breast Cancer Datasets Using FCM Algorithm

B.Durgadevi[*1], Dr.S.Rajalakshmi[2]

[*1] Student, Department of Computer Science and Engineering, Jay Shriram Group of Institutions, Tirupur, India

[2] Associate Professor, Department of Computer Science and Engineering, Jay Shriram Group of Institutions, India

bdurgadevi028@gmail.com

Abstract

The goal of clustering is to group and distinguish comparable units and to separate them from differing units. Classifying patients into groups is not a new phenomenon. It is a concept that, although not always recognized as such, dates back to the beginning of medical science .In fact, it can be said that the idea is based on the notion of a search for a natural ordering of things, which is a basic characteristic of human beings. Fairly recent additions to this concept, however, are 1) the wide-scale application of clustering and classification techniques to patients intra- and inter institutionally for determining medical resource utilization and 2) the growing importance being attached to the reliability and validity aspects of classification procedures and the resulting schemes in general.3) Certain critical decisions must be made in order to properly utilize cluster analysis. Towards this end, cluster analysis encompasses a wide range of statistical techniques.

In this paper from the large number of database the retrieved information can have the details of person who are affected by breast cancer. Using those breast cancer datasets performs 1) information retrieval is based on the specified region. That retrieving breast cancer details from the selected database. 2) Clustering the breast cancer details based on age group.

Keywords: medical data mining, Fuzzy c-means clustering, information retrieval, porter stemming.

Introduction

Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. We can show this with a simple graphical example:

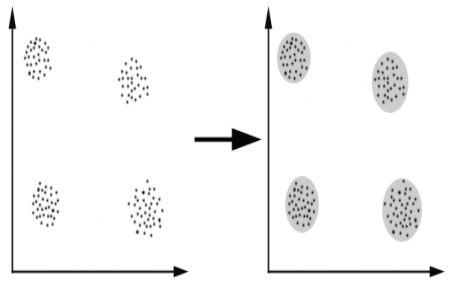

Fig1: simple clustering

The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering? It can be shown that there is no absolute "best" criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs. For instance, we could be interested in finding representatives for homogeneous groups (*data reduction*), in finding "natural clusters" and describe their unknown properties (*"natural" data types*), in finding useful and suitable groupings (*"useful" data classes)* or in finding unusual data objects (*outlier detection*).

Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text (or other content-based) indexing.

For clustering based on age group fuzzy c-means algorithm is used. Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. And for retrieving

clustered information porter stemming algorithm is used. Stemming algorithms--programs that relate morphologically similar indexing and search terms. Stemming is used to improve retrieval effectiveness and to reduce the size of indexing files. Several approaches to stemming are described--table lookup, affix removal, successor variety, and n-gram. Empirical studies of stemming are summarized.

Healthcare industry today generates large amounts of complex data about patients, hospitals resources, disease diagnosis, electronic patient records, medical devices etc. The large amount of data is a key resource to be processed and analyzed for knowledge extraction that enables support for cost-savings and decision making. Medical Data mining brings a set of tools and techniques that can be applied to this processed data to discover hidden patterns**.**

Breast cancer is a type of cancer originating from breast tissue. And there are several symptoms for this cancer. In this paper those symptom details are used as datasets.

System Model

The Preprocessing of the datasets

The datasets is selected for the project. Its relevant and irrelevant dataset attributes are refined and sorted out. The breast cancer dataset is taken into consideration. The data separation methods for the classification of the datasets are to be applied on this datasets. The irrelevant attributes (noise) are removed and datasets made ready for the implementation of the group wise clustering and retrieval methods. The attributes in the database are Class,Age,Menopause,Tumor size,Inv nodes,Node caps,Deg-Malig,Breast-Quad,Irradiat.

Region Selection and Indexing

A database indexing is a data structure that improves the speed of data retrieval operations on a database table at the cost of additional writes and the use of more storage space to maintain the extra copy of data. Indexes are used to quickly locate data without having to search every row in a database table every time a database table is accessed. Indexes can be created using one or more columns of a database table, providing the basis for both rapid random lookups and efficient access of ordered records. Region selection means selecting a specified region among the database which consists of several regions.

After loading the dataset we need to sort out the details according to many categories. From the data set we can track a specific region which facilitates the performance comparison of different information retrieval systems for finding relevant breast cancer details. So the first category is region selection. There are patients from different regions, for the survey we can sort region wise details. And then we have to give the index according to the region in the alphabetical order.

Age Wise Clustering

Grouping of breast cancer information is based on the age. Cluster the details of age between X and Y. Now we have the details of patients which are based on region. The next process is, we have to retrieve the patients in the particular region based on age. So that we can get sorted list of peoples who are categorized on region wise and age wise. For example we can retrieve the details based on peoples from their specified region and their age is between (25 to 30). The selection is based on fuzzy c-means algorithms.

This algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point. More the data is near to the cluster center more is its membership towards the particular cluster center. Clearly, summation of membership of each data point should be equal to one. After each iteration membership and cluster centers are updated according to the formula.

Algorithmic steps for Fuzzy c-means clustering (FCM)

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

1) Randomly select '*c*' cluster centers.

2) Calculate the fuzzy membership 'μ_{ij}' using:

$$\mu_{ij} = 1/\sum_{k=1}^{c} (d_{ij}/d_{ik})^{(2/m-1)}$$

3) compute the fuzzy centers 'v_j' using:

$$v_j = \frac{\sum_{i=1}^{n} (\mu_{ij})^m x_i}{\sum_{i=1}^{n} (\mu_{ij})^m}, \forall j = 1, 2, ... c$$

Where, '*n*' is the number of data points.'*vj*' represents the j^{th} cluster center.'*m*' is the fuzziness index m € [1, ∞]. *c*' represents the number of cluster center.'μij' represents to membership of i^{th} data to j^{th} cluster center. '*dij*' represents the Euclidean distance between i^{th} data and j^{th} cluster center.Main objective of fuzzy c-means algorithm is to minimize:

$$J(U, V) = \sum_{i=1}^{n} \sum_{j=1}^{c} (\mu_{ij})^m ||x_i - v_j||^2$$

Where,'$||x_i - v_i||$' is the Euclidean distance between i^{th} data and j^{th} cluster center.

4) Repeat step 2 and 3 until the minimum '*J*' value is achieved or $||U^{(k+1)} - U^{(k)}|| < \beta$ where,'*k*' is the iteration step. 'β' is the termination criterion between [0, 1]. '$U = (\mu_{ij})_{n*c}$' is the fuzzy membership matrix.'*J*' is the objective function.

ISSN: 2277-9655
Impact Factor: 1.852

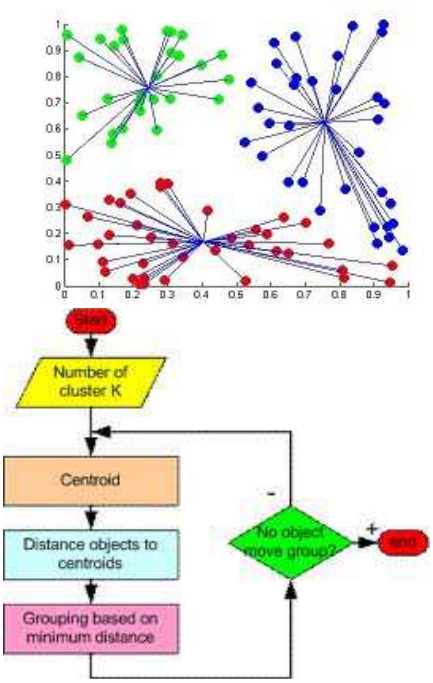

Fig 2: Fuzzy c-means clustering

Stemming Algorithm

Stemming is used to improve retrieval effectiveness and to reduce the size of indexing files. One technique for improving IR performance is to provide searchers with ways of finding morphological variants of search terms. There are several criteria for judging stemmers: correctness, retrieval effectiveness, and compression performance. Stemmers can also be judged on their retrieval effectiveness--usually measured with recall and precision, and on their speed, size, and so on. Finally, they can be rated on their compression performance. Porter stemming algorithm is used for retrieving clustered breast cancer details. This algorithm improves the retrieval effectiveness of dataset.

Fig 3: porter stemming

Conclusion and Future Work

In this paper, fuzzy c-means algorithm is used for effective clustering of breast cancer datasets. That gives best result for data set and comparatively better than other clustering algorithms like k-means clustering. Here data point is assigned membership to each cluster center as a result of which data point may belong to more than one cluster center. Grouping patients based on age is effectively implemented. And retrieval of clustered details are done using porter stemming algorithm. It improves the information retrieval effectiveness. Porter stemming algorithm is an error less one.

By using the same breast cancer database we can implement additional features such as analyzing the disease.

References

[1] Balaji K and Juby N Zacharias Fuzzy c-means .

[2] Weiling Cai, Songcan Chen and Daoqiang Zhang Fast and Robust Fuzzy C-Means Clustering Algorithms Incorporating Local Information for Image Segmentation .

[3] James C. Bezdek, James Keller, Raghu Krishnapuram and Nikhil R. Pal, Fuzzy Models and Algorithms for Pattern Recognition and Image Processing, Kluwer Academic Publishers, TA 1650.F89, 1999.

[4] N. R. Pal, K. Pal and J. C. Bezdek, “A mixed c-means clustering model,” Proceedings of the Sixth IEEE International Conference on Fuzzy Systems, Vol. 1, pp. 11-21, Jul. 1997.

[5] M.F. Porter. 1980. An algorithm for suffix stripping. Program, vol 14, no 3, pp 130-130. Xu and W. B. Croft. 1998. Corpus-based Stemming using Co-occurrence of Word Variants. ACM Transactions on Information Systems, Volume 16, Number 1, pp 61-81, January 1998.

[6] Jun Gu, Wei Feng, Member, IEEE, Jia Zeng, Hiroshi Mamitsuka, and Shanfeng Zhu,

Efficient Semisupervised MEDLINE Document Clustering With MeSH-Semantic and Global-Content Constraints, IEEE TRANSACTIONS ON CYBERNETICS, VOL. 43, NO. 4, AUGUST 2013

[7] W. Kraaij and R.Pohlmann. 1994. Porter's stemming algorithm for Dutch. In L.G.M. Noordman and W.A.M. de Vroomen,editors, Informatie wetenschap 1994: Wetenschappelijke bijdragen aan de derde STINFON Conferentie, pp. 167-180.

[8] Jun Yan, Michael Ryan and James Power, Using fuzzy logic Towards intelligent systems, Prentice Hall, 1994.

[9] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum, NY, 1981.

[10] M. Krallinger, A. Valencia, and L. Hirschman, “Linking genes to literature: Text mining, information extraction, and retrieval applications for biology,” Genome Biol., vol. 9, no. S2, pp. S8–S14, Sep. 2008.