EVOLVING EFFICIENT CLUSTERING AND CLASSIFICATION PATTERNS IN LYMPHOGRAPHY DATA THROUGH DATA MINING TECHNIQUES

Shomona Gracia Jacob[1] and R.Geetha Ramani[2]

[1]Department of Computer Science and Engineering, Rajalakshmi Engineering College (Affiliated to Anna University, Chennai)
graciarun@gmail.com
[2]Department of Information Science and Technology, College of Engineering, Guindy, Anna University, Chennai.
rgeetha@yahoo.com

ABSTRACT

Data mining refers to the process of retrieving knowledge by discovering novel and relative patterns from large datasets. Clustering and Classification are two distinct phases in data mining that work to provide an established, proven structure from a voluminous collection of facts. A dominant area of modern-day research in the field of medical investigations includes disease prediction and malady categorization. In this paper, our focus is to analyze clusters of patient records obtained via unsupervised clustering techniques and compare the performance of classification algorithms on the clinical data. Feature selection is a supervised method that attempts to select a subset of the predictor features based on the information gain. The Lymphography dataset comprises of 18 predictor attributes and 148 instances with the class label having four distinct values. This paper highlights the accuracy of eight clustering algorithms in detecting clusters of patient records and predictor attributes and highlights the performance of sixteen classification algorithms on the Lymphography dataset that enables the classifier to accurately perform multi-class categorization of medical data. Our work asserts the fact that the Random Tree algorithm and the Quinlan's C4.5 algorithm give 100 percent classification accuracy with all the predictor features and also with the feature subset selected by the Fisher Filtering feature selection algorithm.. It is also stated here that the Density Based Spatial Clustering of Applications with Noise (DBSCAN) clustering algorithm offers increased clustering accuracy in less computation time.

KEYWORDS

Data mining, Clustering, Feature Selection, Classification, Lymphography Data

1. INTRODUCTION

Data mining [1-2] is the process of discovering and distinguishing related, credential and critical information from a prolific database. Machine learning, [2,3] is concerned with the design and development of algorithms that enable a system to automatically learn to identify complicated patterns and make intelligent decisions based on the available data. However the enormous size of available data poses a major impediment in recognizing patterns. To handle the vast collection of data, a clustering process [4-5] was proposed to detect groups in a collection of records. A clustering algorithm [6-8] partitions a data set into several groups such that the similarity within a

group is larger than between the groups. Feature Selection attempts to select a subset of attributes based on the information gain. Classification is a supervised technique that designates items in a collection to target categories or classes [9-10]. The main aim of classification is to precisely predict the target class for each unknown case in the data. Multi class Classification, also called Multinomial classification [11-13] assigns the given set of input data to one of many categories.
A lymph node [14] is an oval-shaped organ of the immune system, distributed widely throughout the body. They tend to expand in size for diverse reasons, indicating health complications that scale from trivial, to life-threatening ailments such as cancers. In the latter, the condition of lymph nodes is so significant that it is used to accurately sense the stage in Cancer progression, which decides the treatment to be adopted. Lymphography [15]is a medical imaging technique in which a radio contrast agent is injected, and then an X-ray picture is taken to visualize structures of the lymphatic system, including lymph nodes, lymph ducts, lymphatic tissues, lymph capillaries and lymph vessels. This data is necessary to decide on whether the clinical details acquired from a Lymphograph pertains to a normal or abnormal finding. Additionally the existing state of the lymph nodes could also suggest the possibility of occurrence of cancer [14-15]. Though the procedure for performing Lymphography involves potential hurdles, the data from the images facilitate accurate and precise determination of the state of the lymph nodes, ducts and capillaries. Hence proper classification and determination of credential attributes could simplify the process of disease prediction and evoke deterrent measures. Since cancer is a leading cause of death round the globe, crafting an efficient classifier for an oncogenic database has been the rationale for our research.

Our research work mainly focuses on recognizing a suitable clustering and classification algorithm for the Lymphography dataset from the UCI Machine Learning repository. We realize this by executing eight clustering algorithms namely Expectation-Maximization Clustering (EM-Clustering) algorithm, Hierarchical Clustering (HAC) algorithm, FilteredFirst Algorithm, FarthestFirst algorithm, Density Based Spatial Clustering of Application with Noise (DBSCAN) algorithm, K-Means Clustering, and CobWeb Algorithm. Sixteen multi-class classification algorithms viz, Quinlan's C4.5 decision tree algorithm (C4.5) , Classification Tree(C-RT), Cost-Sensitive Classification Tree(CS-CRT), Cost-sensitive Decision Tree algorithm(CS-MC4), SVM for classification(C-SVC), Iterative Dichomotiser(ID3), K-Nearest Neighbor(K-NN), Linear Discriminant Analysis (LDA), Multilayer Perceptron(MP), Multinomial Logistic Regression(MLR), Naïve Bayes Continuous(NBC), Partial Least Squares -Discriminant/Linear Discriminant Analysis(PLS-DA/LDA), Prototype-Nearest Neighbor(P-NN), Radial Basis Function (RBF), and Random Tree (Rnd Tree), classification algorithms executed on the dataset for a comparative analysis. We also investigate the effect of feature selection using Fisher Filtering (FF), ReliefF, Runs Filtering, and Stepwise Discriminant (Step Disc/SD) Analysis algorithms to enhance the classifier accuracy and reduce the feature subset size.
The following section reviews the past and current state of research in related areas of data mining.

2. BACKGROUND OF THE STUDY

Previous research on application of feature selection and classification techniques of data mining in the field of medical research is briefly reviewed in the following paragraphs.

Hammouda et.al [16] made a study of four clustering techniques and reviewed the most representative off-line clustering techniques: K-means clustering, Fuzzy Cmeans clustering, Mountain clustering, and Subtractive clustering. The techniques are implemented and tested against a medical problem of heart disease diagnosis. Performance and accuracy of the four techniques were presented and compared-Means achieved a clustering accuracy of 80% while

Fuzzy Cmeans and Mountain clustering achieved an accuracy of 78%. Subtractive clustering offered the least accuracy of 75%.

Hirano et.al, [17] presented a cluster analysis method for multidimensional time-series data on clinical laboratory examinations. Their method represented the time series of test results as trajectories in multidimensional space, and compared their structural similarity by using the multiscale comparison technique. It enabled identification of the part-to-part correspondences between two trajectories, taking into account the relationships between different tests. The resultant dissimilarity could be further used with clustering algorithms for finding the groups of similar cases. The method was applied to the cluster analysis of Albumin-Platelet data in the chronic hepatitis dataset. The results demonstrated that it could form interesting groups of cases that had high correspondence to the fibrotic stages.

Chuang et. al [18] revealed the means of effectively using a number of validation sets obtained from the original training data to improve the performance of a classifier. The proposed validation boosting algorithm was illustrated with a support vector machine (SVM) in Lymphography classification. A number of runs with the algorithm was generated to show its robustness as well as to generate consensus results. At each run, a number of validation datasets were generated by randomly picking a portion of the original training dataset. At each iteration, the trained classifier was used to classify the current validation dataset. Experimental results on the Lymphography dataset showed that the proposed method with validation boosting could achieve much better generalization performance (on repeated iterations) with a testing set than the case without validation boosting.

Polat et. al,[19] proposed a novel hybrid classification system based on C4.5 decision tree classifier and one-against-all approach to classify the multi-class problems including dermatology, image segmentation, and Lymphography datasets taken from UCI (University of California Irvine) machine learning database. In their work, initially C4.5 decision tree was executed for all the classes of datasets and they reported 84.48%, 88.79%, and 80.11% classification accuracy for dermatology, image segmentation, and Lymphography datasets using 10-fold cross validation, respectively. The proposed method based on C4.5 decision tree classifier and one-against-all approach obtained 96.71%, 95.18%, and 87.95% for the above datasets, respectively.

Holte [20] presented the results of an investigation into the classification accuracy of very simple rules called "1−rules", or 1-level decision trees, ones that classify instances based on a single attribute. A program, called 1R, learns 1−rules from examples on 16 datasets commonly used in machine learning research.

Mc.Sherry et.al, [21] presented an algorithm for Conversational Case Based Reasoning (CCBR) called iNN(k) in which feature selection was motivated by the goal of confirming a target class and informed by a measure of feature's discriminating power that supported the target class. The performance of iNN (k) on a given dataset was shown to depend on the value of ‘k’ and on whether local or global feature selection was used in the algorithm. Only 42% and 51% on an average of features in a complete problem description were needed by iNN (k) to provide accuracy levels of 86.5% and 84.3% respectively on the Lymphography and SPECT heart datasets from the UCI machine learning repository.

2.1. Paper Organization

This paper is organized in the following manner. Section 3 portrays the data mining framework for clustering and classification, clearly explaining each phase employed in the data mining

process. In Section 4, we discuss the performance of the system with respect to the various algorithms employed while Section 5 concludes the paper.

3. DATA MINING FRAMEWORK

The data mining framework is viewed to consist of two distinct phases namely clustering and classification. The clustering framework [2] attempts to discover groups in the patient records and detect clusters of attributes that contribute to the identification of the target class of a patient record. Feature selection [4] followed by classification enables the recognition of an appropriate target class for the clinical findings. The proposed system design of a data mining framework to detect clusters and classes of a given patient record is given in Figure 1.

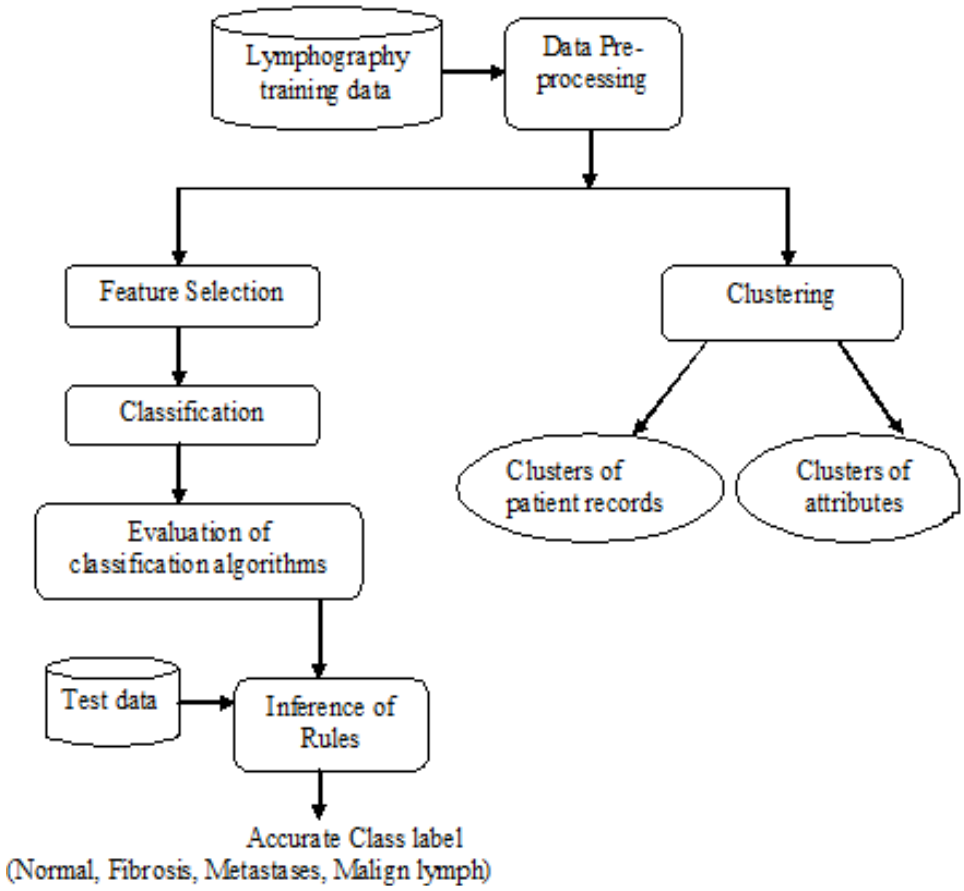

Fig. 1. Proposed Data Mining Framework to cluster and classify Lymphography patient records

The evaluation criteria for the sixteen classification algorithms are taken to be the accuracy in classification, decision tree size and the computation time for execution. The best classifier is selected by recording the misclassification rate and comparing the tree size and computation time used by the algorithms for classification. The detailed description of the training dataset and the phases in the design of the system is given below.

3.1 Lymphography Dataset

This Lymphography dataset was obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia [22]. The dataset comprises of a target class that can have four distinct values and the number of predictor attributes sums up to eighteen. This data provides 148 cases to train the classifier. The details of the attributes, their possible values and the associated Attribute ID are clearly listed in Table 1.

Table 1. Description of the Attributes in the Lymphography dataset

Attribute	**Possible Values**	**Assigned values**	**Attribute ID**
Lymphatics	Normal, arched, deformed, displaced	1-4	1
Block of afferent	No, Yes	1,2	2
Block of lymph.c (superior and inferior flaps)	No, Yes	1,2	3
Block of lymph.s (lazy incision)	No, Yes	1,2	4
Bypass	No, Yes	1,2	5
Extravasates (force out of lymph)	No, Yes	1,2	6
Regeneration	No, Yes	1,2	7
Early uptake in	No, Yes	1,2	8
Lymph nodes diminish	0-3	0-3	9
Lymph nodes enlarge	1-4	1-4	10
Changes in lymph	Bean, oval, round	1-3	11
Defect in node	No, lacunar, lacunar marginal, lacunar central	1-4	12
Changes in node	No, lacunar, lacunar marginal, lacunar central	1-4	13
Changes in structure	no, grainy, drop-like, coarse, diluted, reticular, stripped, faint	1-8	14
Special forms	No, Chalices, vesicles	1-3	15
Dislocation	No, Yes	1-2	16
Exclusion of no.	No, Yes	1-2	17
Number .of nodes in	0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, >=70	1-8	18
Target Class	Normal , metastases, malign lymph, fibrosis	1-4	19

3.2 Data Pre-processing

The Lymphography dataset was obtained from the UCI Machine Learning Repository website (UCI, SGI MLC++) [22] and saved as a text file. This file is then imported into Excel spreadsheet and the values are saved with the corresponding attributes as column headers. The Excel file is then uploaded into TANAGRA (Tanagra data mining) [23], a data mining tool and the uploaded data is visualized to ensure that the precise values are inserted. The predictor and the target attributes are specified. In order to apply clustering algorithms, we make use of the WEKA data mining tool [24]. The textual data needs to be stored as a Comma Separated Version (.CSV) file and the attribute selection must be categorical.

The algorithmic techniques applied for clustering, feature relevance analysis and classification are elaborately presented in the following sections.

3.3 Clustering Algorithm

The possibility of having to encounter enormous data motivates human thinking to categorize this huge data into smaller groups or categories to further facilitate its analysis[25]. Yet another important reason for clustering [26] is to discover relevance knowledge in data. We present the clustering algorithm that clusters with more than 90% accuracy in the following sub-section.

3.3.1 Density Based Spatial Clustering of Applications with Noise (DBSCAN) Algorithm

The DBSCAN algorithm was first introduced by Ester, et al. [5], and attempts to cluster based on density. Clusters [6] are identified by looking at the density of points. Regions with larger density depict the existence of clusters whereas regions with a lower density of points indicate clusters of noise or clusters of outliers. This algorithm [7] is particularly suited to deal with large datasets, with noise, and is able to identify clusters of different sizes and shapes. The key idea of the DBSCAN algorithm [8] is that, for each point of a cluster, the neighbourhood of a given radius has to contain at least a minimum number of points, that is, the density in the neighbourhood has to exceed a predefined threshold. The Pseudo code of the DBScan algorithm is given in Figure 2.

```
DBSCAN (Data, Eps, MinPts)
Cluster = 0
for each unvisited point X in dataset Data
mark X as visited
NeighborPts = regionQuery(X, eps)
if sizeof(NeighborPts) < MinPts
mark X as NOISE
else
Cluster = next cluster
expandCluster(X, NeighborPts, Cluster, eps, MinPts)

expandCluster(X, NeighborPts, Cluster, eps, MinPts)
add X to Cluster
for each point X' in NeighborPts
if X' is not visited
mark X' as visited
NeighborPts' = regionQuery(X', eps)
if sizeof(NeighborPts') >= MinPts
NeighborPts = NeighborPts joined with NeighborPts'
```

```
if X' is not yet member of any cluster
add X' to Cluster

regionQuery(X, eps)
return all points within X's eps-neighborhood
```

Figure 2. Pseudocode of DBSCAN algorithm

This algorithm[25] needs three input parameters namely data that defines the neighbour list size, Eps, the radius that delimitate the neighbourhood area of a point (Epsneighbourhood) and MinPts, the minimum number of points that must exist in the Eps-neighbourhood. The clustering process is based on the classification of the points in the dataset as core points, border points and noise points, and on the use of density relations between points (directly density-reachable, density-reachable, density-connected [5][25][26][27] to form the clusters

3.4 Feature selection and Classification algorithms

The Feature selection, as a pre-processing step to application of supervised learning techniques, is effective in reducing dimensionality, eliminating irrelevant data, improving learning accuracy, and improving result interpretability. Feature selection methods [28] can be broadly classified into two broad categories viz, Filter and Wrapper methods. Filter methods select features based on discriminating criteria that are relatively independent of classification whereas Wrapper methods utilize the classifier as a black box to score the subsets of features based on their predictive power [29][30-32]. We make use of wrapper methods for feature selection and the Fisher Filtering algorithm selects the most predominant features that enhance classifier accuracy as is seen in the case of Random Tree and C4.5 algorithm.

3.4.1 Fisher Filtering Feature Selection Algorithm

It is also known as Univariate Fisher's ANOVA ranking (Tanagra tutorials). It is a supervised feature selection algorithm that processes the selection independently from the learning algorithm. It follows a filtering approach that ranks the input attributes according to their relevance [23] [28-29]. A cutting rule enables the selection of a subset of these attributes. The classification algorithms that generated 100 percent accurate classification on the Lymphography data are described below.

3.4.2 Quinlan's C4.5 Decision Tree Classification Algorithm

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan [33]. C4.5 relies on greedy search, selecting the candidate test that maximizes a heuristic splitting criterion [34]. C4.5 operates on two criteria viz, information gain and gain ratio.

A sample rule generated by the Quinlan's C4.5 classification algorithm is given in Figure.2.

changes in struc < 5.5000
    changes in lymph < 2.5000 then Class = **Malign lymph** (100 % of 2 examples)
    changes in lymph >= 2.5000 then Class = **Metastases** (100 % of 1 examples)
changes in struc >= 5.5000 then Class = **Metastases** (100.00 % of 5 examples)

Fig. 3 Sample Rule from C4.5 Algorithm for Lymphography Dataset

3.4.3 Random Tree Classification Algorithm

Random trees [35] [36] have been introduced by Leo Breiman and Adele Cutler. Random trees are a collection of tree predictors that is called forest. A sample rule generated by the Random Tree classification algorithm with Fisher Filtering feature selection algorithm is given in Figure. 4.

```
changes in node < 2.5
exclusion of no. < 1.5 then Class = Metastases (100 % of 5 examples)
exclusion of no. >= 1.5
early uptake in < 1.5 then Class = Metastases (100 % of 1 examples)
early uptake in >= 1.5
lymph nodes enlarge < 3.5 then Class = Malign lymph (100 % of 4 examples)
```

Fig. 4 Sample Rules from Random Tree Classification Algorithm

In majority of the machine learning algorithms, the best estimate of the target function is presupposed to be the uncomplicated, simple classifier [20] that fits the given data, since more complex models tend to over fit the training data and generalize poorly [25].

4. PERFORMANCE EVALUATION

The clustering and classification algorithms are ranked based on their accuracy and less computational complexity. Accuracy [2] of a classifier is measured in terms of how precisely the classifier places the input datasets under the correct category [2] [3]. This is denoted as the Misclassification rate which is computed as 1- Accuracy(C) where C denotes Classifier.

4.1 Experimental Results

The results of the eight clustering algorithms that have been considered for analysis on the Lymphography data have been depicted in Table 2. The results portrayed exhibit the accuracy and the computation time.

Table 2. Performance Comparison of Clustering Algorithms

S.No	Clustering Algorithms	Clustering Accuracy (%)	Computation time(s)
1.	EM-Clustering	69.5	3.2
2.	Hierarchical Clustering	56.1	0.06
3.	K-Means	60.82	0.02
4.	CobWeb	42.68	0.02
5.	DBScan	91.9	0.03
6.	Farthest First	57.44	0.01
7.	FilteredFirst	60.82	0.02
8.	MakeDensityBased	61.5	0.03

The feature subset size selected by the feature relevance algorithms is given in Table 3.

Table 3. Feature Subset Selected on the Lymphography dataset

S.No	Feature Selection Algorithms	Attribute ID of Selected Features (Referring Table 1)
1	Fisher Filtering(FF)	9,7,18,2,10,8,15,11,13,4,14
2	Stepwise Discriminant Analysis (Step Disc)	2,13,14,7,11,8,10,18
3	Runs Filtering (RF)	2,13
4	ReliefF Filtering (RF)	9,18,2,14,7,1,8,15,11,4,6

The graphical representation of the performance of the clustering algorithms is given in Figure 5.

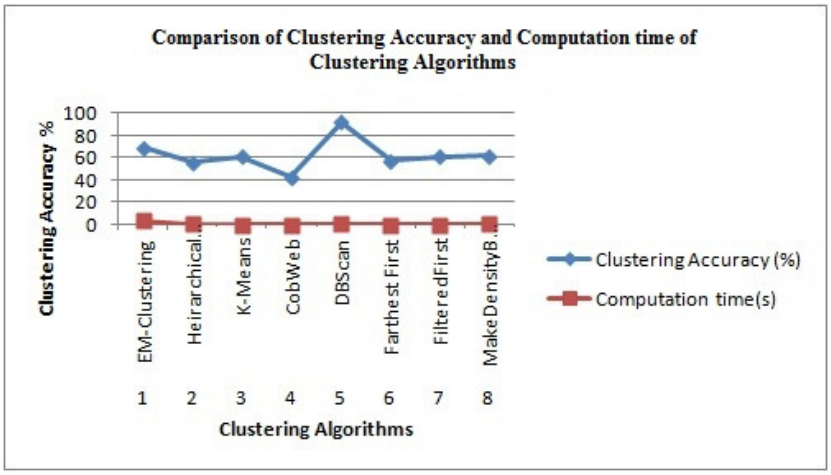

Figure 5. Graphical Representation of Clustering Algorithms 'Performance

The graphical representation of the filtered feature subset size is presented in Figure 6.

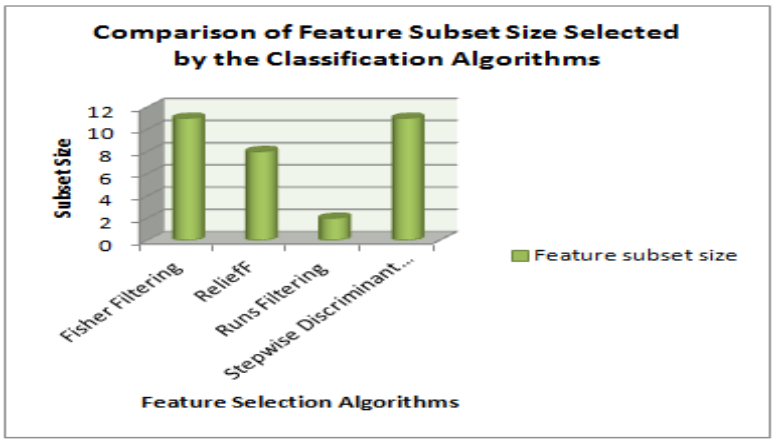

Fig. 6 Feature Subset Size filtered by Feature Selection Algorithms

The comparative classifier results for the feature selection algorithms are tabulated in Table 4.

Table 4. Performance Comparison of Classification Algorithms

S.No	Classification Algorithms	Accuracy % (Before Feature Selection)	Feature Selection Algorithms		
			Fisher Filtering	ReliefF	SD Analysis
1	Radial Basis Function	85.14	85.14	86.49	81.08
2	Random Tree	100	100	99.32	98.65
3	C4.5	100	100	99.32	98.65

Sixteen classification algorithms are applied on the Lymphography dataset after it is pre-processed and the feature subsets are selected.The size of the feature set to be considered for classification is reduced and hence less storage space is required for the execution of the algorithms. The Fisher filtering algorithm reduces the feature subset size to 11 and also provides 100 percent accuracy for C4.5 and Random Tree classification algorithm. The accuracy of the classification algorithms is tabulated in Table 5.

Table 5. Comparison of Performance of Classification Algorithms on the Lymphography dataset

S.No	Classification Algorithms	Accuracy (%) before feature selection	Accuracy (%) After Fisher Filtering
1	C4.5	100	100
2	C-RT	86.49	68.92
3	CS-CRT	86.49	68.92
4	CS-MC4	86.49	83.11
5	C-SVC	91.89	88.51
6	ID3	54.73	54.73
7	KNN	89.86	85.81
8	LDA	91.22	87.16
9	MP	89.86	86.49
10	MLR	8.11	5.41
11	NBC	87.16	83.78
12	PLS-DA	86.49	86.49
13	PLS-LDA	89.19	85.81
14	PROTOTYPE-NN (Local variance)	84.46	81.76
15	RBF	85.14	85.14
16	RND TREE	100	100

However the number of attributes for split need to be set to 11 or more in the Random tree algorithm and the minimum size of the leaves and the confidence level need to be set to 1 in C4.5 algorithm to achieve 100 percent accuracy. The number of attributes chosen for a split on Random Tree should be specified according to the number of features considered for classification. The Runs filtering algorithm has filtered only two attributes, hence the results are not considered for classification. The size of the tree identified by the number of nodes and

number of leaves in the decision tree generated by the classification algorithms is given in Table 6.

Table 6. Evaluation Parameters of Classification Algorithms

S.No	Classification Algorithms	Number of nodes	Computation time (ms)
1	Quinlan's C4.5 algorithm	61	15
2	Random Tree algorithm	61	15

The graphical representation of the classifier result is depicted in Figure 7.

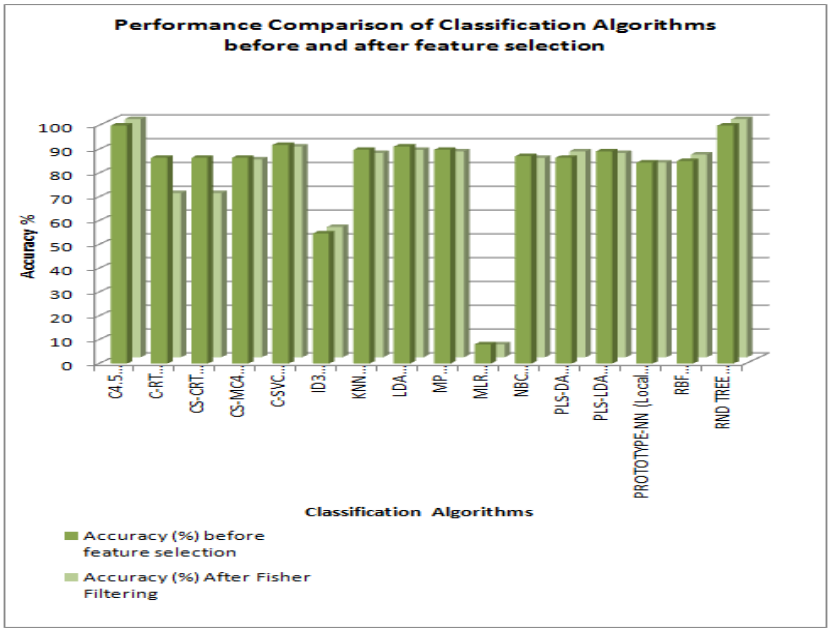

Fig. 7 Performance Comparison of classification Algorithms

The graphical representation of the evaluation criteria of classification algorithms on the Lymphography data is given in Figure 8.

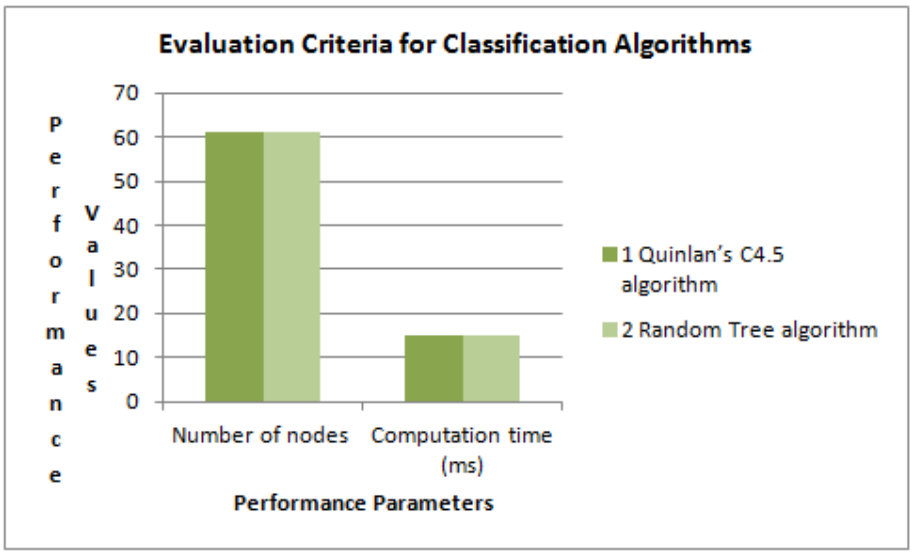

Fig. 8 Performance Evaluation of classification Algorithms

5. CONCLUSION

In this paper we have analyzed the impact of clustering clinical data and proposed the design of a classifier that is trained on the Lymphography dataset from the UCI Machine Learning Repository to perform multi-class categorization of clinical data. We have evaluated the performance of eight clustering algorithms and sixteen classification algorithms on the dataset and compared the clustering accuracy followed by the accuracy given by the classification algorithms before and after feature selection. Moreover the size of the decision tree generated and the computation time is recorded to bring out the efficiency of the classifier. Our findings suggest with necessary results that the DBSCAN algorithms generates ~90% accurate clusters while the Random Tree and Quinlan's C4.5 algorithm give 100 percent accuracy in classification with least computational complexity. This research will aid in enhancing the current state of ailment prediction and classification in the field of clinical research.

ACKNOWLEDGMENT

This research work is a part of the All India Council for Technical Education(AICTE), India funded Research Promotion Scheme project titled "Efficient Classifier for clinical life data (Parkinson, Breast Cancer and P53 mutants) through feature relevance analysis and classification" with Reference No:8023/RID/RPS-56/2010-11, No:200-62/FIN/04/05/1624.

REFERENCES

[1] S.B.Kotsiantis (2007), Supervised Machine Learning: A Review of Classification Techniques, Informatica (31), 249-268.
[2] J. Han and M. Kamber (2000), Data Mining; Concepts and Techniques, Morgan Kaufmann Publishers.
[3] Mitchell, Tom M (1997), Machine Learning. The Mc-Graw-Hill Companies, Inc.
[4] Hans-Peter Kriegel, Peer Kröger, Jörg Sander, Arthur Zimek (2011). "Density-based Clustering". WIREs Data Mining and Knowledge Discovery 1 (3): 231–240. DOI: 10.1002/widm.30. 5.

[5] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise". In Evangelos Simoudis, Jiawei Han, Usama M. Fayyad. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231. ISBN 1-57735-004-9. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.71.1980.

[6] Sander, Jörg; Ester, Martin; Kriegel, Hans-Peter; Xu, Xiaowei (1998). "Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications". Data Mining and Knowledge Discovery (Berlin: Springer-Verlag) 2 (2): 169–194. DOI:10.1023/a:1009745219419. http://www.springerlink.com/content/n22065n21n1574k6.

[7] Sander, Jörg (1998). Generalized Density-Based Clustering for Spatial Data Mining. München: Herbert Utz Verlag. ISBN 3-89675-469-6.

[8] Agrawal, R.; Gehrke, J.; Gunopulos, D.; Raghavan, P. (2005). "Automatic Subspace Clustering of High Dimensional Data". Data Mining and Knowledge Discovery 11: 5. DOI: 10.1007/s10618-005-1396-1.

[9] Nancy.P, Dr.R.Geetha Ramani, Shomona Gracia Jacob (2011a), "Discovery of Gender Classification Rules for Social Network Data using Data Mining Algorithms", Proceedings of the IEEE International Conference on Computational Intelligence and Computing Research(ICCIC'2011), Kanyakumari, India, , IEEE Catalog Number:CFP1120J-PRT, ISBN:978-1-61284-766-5, pp 808-812.

[10] P.Nancy and Dr.R.Geetha Ramani (2011b) Article: "A Comparison on Performance of Data Mining Algorithms in Classification of Social Network Data". International Journal of Computer Applications 32(8):47-54, DOI: 10.5120/3927-5555.Published by Foundation of Computer Science, New York, USA

[11] Tan, Steinbach (2004), Kumar, Introduction to Data Mining.

[12] Vapnik, V. N. (2000) The Nature of Statistical Learning Theory (2nd Ed.), Springer, Verlag,

[13] Mrs.Shomona Gracia Jacob and Dr. R.Geetha Ramani (2011a),"Discovery of Knowledge Patterns in Clinical Data through Data Mining Algorithms: Multi-class Categorization of Breast Tissue Data", International Journal of Computer Applications (IJCA), 32(7): 46-53,DOI: 10.5120/3920-5521. Published by Foundation of Computer Science, New York, USA.

[14] Warwick, Roger; Peter L. Williams (1973) [1858]. "Angiology (Chapter 6)". Gray's anatomy. Illustrated by Richard E. M. Moore (Thirty-fifth Ed.). London: Longman. pp. 588–785.

[15] Guermazi A, Brice P, Hennequin C, Sarfati E (2003). "Lymphography: an old technique retains its usefulness". Radiographics 23 (6): 1541–58; discussion 1559–60.

[16] K. Hammouda, and F. Karay, A Comparative Study of Data Clustering Techniques, Course Project, 2000.

[17] Shoji Hirano, Shusaku Tsumoto, "Cluster Analysis of Time-series Medical Data Based on the Trajectory Representation and Multiscale Comparison Techniques, Proceedings of the Sixth International Conference on Data Mining (ICDM'06)0-7695-2701-9/06 $20.00 © 2006

[18] Tzu-cheng Chuang, Okan K. Ersoy, Saul B. Gelfand, Boosting Classification Accuracy With Samples Chosen From A Validation Set, ANNIE (2007), Intelligent ` Engineering systems through artificial neural networks, St. Louis, MO, pp. 455-461.

[19] Kemal Polat, Salih Gunes (2009), A novel hybrid intelligent method based on C4.5 decision tree classifier and one-against-all approach for multi-class classification problems", Expert Systems with Applications: An International Journal,Volume 36, Issue 2, Pergamon Press, Inc. Tarrytown, NY, USA

[20] Robert C. Holte (1993), Very Simple Classification Rules Perform Well on Most Commonly Used Datasets, Machine Learning, 11:63-91

[21] McSherry D (2011), Conversational case-based reasoning in medical decision making, Artificial Jun; 52(2):59-66. Epub

[22] SGI - MLC++: Datasets from UCI

[23] Tanagra Data Mining tutorials, http://data-mining-tutorials.blogspot.com/

[24] Weka Tool http://www.cs.waikato.ac.nz/ml/weka/arff.html

[25] Achtert, E.; Böhm, C.; Kröger, P. (2006). "DeLi-Clu: Boosting Robustness, Completeness, Usability, and Efficiency of Hierarchical Clustering by a Closest Pair Ranking". LNCS: Advances in Knowledge Discovery and Data Mining. Lecture Notes in Computer Science 3918: 119–128. DOI: 10.1007/11731139_16. ISBN 978-3-540-33206-0.

[26] S Roy, D K Bhattacharyya (2005). "An Approach to find Embedded Clusters Using Density Based Techniques". LNCS Vol.3816. Springer Verlag. pp. 523-535.

[27] Fisher, Douglas H. (1987). "Knowledge acquisition via incremental conceptual clustering". Machine Learning2(2):139172.DOI:10.1007/BF00114265.http://www.springerlink.com/content/qj16212n7537n6p3/fulltext.pdf.
[28] F.C. Garcia-Lopez, M. Garcia-Torres, B. Melian, J.A. Moreno-Perez, J.M. Moreno-Vega (2006). Solving feature subset selection problem by a Parallel Scatter Search, European Journal of Operational Research, vol. 169, no. 2, pp. 477-489.
[29] Hai Nguyen, Katrin Franke, and Slobodan Petrovic (2009), Optimizing a class of feature selection measures, Proceedings of the NIPS 2009 Workshop on Discrete Optimization in Machine Learning: Sub modularity, Sparsity & Polyhedra (DISCML), Vancouver, Canada.
[30] Shomona Gracia Jacob, Dr.R.Geetha Ramani, P.Nancy (2011 b), "Feature Selection and Classification in Breast Cancer Datasets through Data Mining Algorithms", Proceedings of the IEEE International Conference on Computational Intelligence and Computing Research (ICCIC'2011), Kanyakumari, India,, IEEE Catalog Number: CFP1120J-PRT, ISBN: 978-1-61284-766-5. Pp. 661-667
[31] Shomona Gracia Jacob, Dr.R. Geetha Ramani, Nancy .P (2012), "Efficient Classifier for Classification of Hepatitis C Virus Clinical Data through Data Mining Algorithms and Techniques", Proceedings of the International Conference on Computer Applications,Pondicherry, India, January 27-31, 2012,Techno Forum Group, India. ISBN: 978-81-920575-8-3: DOI: 10.73445/ISBN_0768, ACM#.dber.imera.10.73445.
[32] Tran Huy Dat, Cuntai Guan (2007), Feature Selection Based on Fisher Ratio and Mutual Information Analyses for Robust Brain Computer Interface, IEEE International Conference on Acoustics, Speech and Signal Processing.
[33] Ron Kohavi and Ross Quinlan (2009), Decision Tree Discovery.
[34] Thales Sehn Korting(2006), C4.5 algorithm and Multivariate Decision Trees, Image Processing Division, National Institute for Space Research – INPES˜ao José dos Campos– SP, Brazil
[35] LeoBreiman,AdeleCuttler,RandomTrees,http://www.stat.berkeley.edu/users/breiman/RandomForests/
[36] Chandra.B, Basker.S (2011), A new approach for classification of patterns having categorical attributes, Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference 9-12-10-2011,pp.960-964 Anchorage, AK ISSN:1062-922XISBN:978-1-4577-0652-3 INSPEC Accession Number: 12387415 DOI10.1109/ICSMC.2011.6083793.

Authors

Dr.R. Geetha Ramani is Associate Professor, Department of Information Science and Technology, College of Engineering, Guindy, Anna University, Chennai, India. She has more than 15 years of teaching and research experience. Her areas of specialization include Data mining, Evolutionary Algorithms and Network Security. She has over 50 publications in International Conferences and Journals to her credit. She has also published a couple of books in the field of Data Mining and Evolutionary Algorithms. She has completed an External Agency Project in the field of Robotic Soccer and is currently working on projects in the field of Data Mining. She has served as a Member in the Board of Studies of Pondicherry Central University. She is presently a member in the Editorial Board of various reputed International Journals.

Mrs.Shomona Gracia Jacob completed her M.E. in Computer Science and Engineering at Jerusalem College of Engineering, affiliated to Anna University, Chennai, India. She has more than 3 years of teaching experience. Presently she is pursuing her Ph.D in Computer Science and Engineering at Rajalakshmi Engineering College, affiliated to Anna University, Chennai. Her areas of interest include Data Mining, Artificial Intelligence and Software Engineering. She has attended and presented papers at National and International Conferences.