International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0020
ISSN (Online): 2279-0039

International Journal of Engineering, Business and Enterprise Applications (IJEBEA)

www.iasir.net

Cluster based Feature Selection for Dermatology

Prafulla Bafna
Symbiosis Institute of Computer Studies and Research (SICSR),
Symbiosis International University (SIU), Atur Centre, Gokhale Cross Road,
Model Colony, Pune – 411 016, Maharashtra State, INDIA

Abstract: *To process applications from target marketing to weather report analysis, clustering is an effective data mining technique. It is also used for preprocessing including feature selection. Various clustering algorithms are stated in the literature. The cluster quality gets affected by irrelevant data and domain characteristics. The first task is therefore to select an appropriate clustering algorithm suitable for the domain and remove noisy features. To choose the algorithm we have used different clustering evaluation parameters.*
Feature Selection (FS) is the fundamental process for identifying the features that play an important role in decision making process. Data mining techniques are widely used for data driven decision support. Feature selection is an important part of preprocessing that is carried out prior to data mining. FS algorithm or technique helps to reduce dimensions by removing unwanted or noisy data and also the data which is not having any influence on the outcome. It decreases performance and efficiency of the algorithm which is used to run on the given dataset .FS algorithm chooses optimum feature set from the input data set.
In this paper we present a novel approach of using clustering technique for FS and vice a versa. Data sets in online repositories contain multiple irrelevant attributes. It's a need to choose only the significant parameters which affect the result. As a first step, features are selected using domain knowledge. On these selected features 3 clustering algorithms are applied which are from hierarchical and partitioning family and quality of clusters thus obtained is used to choose the algorithm considering different evaluation parameters. The chosen algorithm is to be used in the next step. It is in turn applied to all features in forward feature selection way to get the clusters. Cluster quality is used as decision support for the optimum feature set. This approach will save resources in the next round of data collection and the use of Clustering on complete data set for effective decision support. The data set used is related to dermatology data of UCI repository.
Keywords: *Domain based, hierarchical and partitioning, continuous-k-means clustering, pairwise agglomerative clustering, dimension reduction, optimal features, dermatology.*

I. Introduction

Data mining is a form of knowledge discovery essential for solving problems in specific domain. It is used in financial banking for credit risk covering, market segmenting, health care etc. Clustering is one of the data mining techniques, used to discover hidden patterns when relation between the data is unknown. Clustering is to be used on complete data. There are different methods of clustering viz. partitioning, hierarchical, density based, model based and grid based.[14] The decision making process heavily depends on the quality of clusters and the algorithm should be chosen appropriately. The suitability of the algorithm depends on the domain and usually several algorithms need to be tried before choosing the best which is expensive. For most applications not only data is of dynamic nature that is continuously growing in size, but also is multidimensional in nature leading to scalability problems. Data size cannot be reduced but irrelevant dimensions can be definitely reduced. So before applying any algorithm, it is essential to use some preprocessing technique to reduce unnecessary dimensions/features and select the most relevant features. Feature selection is one of the approaches for dimension reduction. Dimension reduction limits storage requirements. It speeds up the running time of the learning algorithms and improves its accuracy. Domain knowledge can be used to select the features [11], but there is also possibility of missing relevant features in the absence of domain expertise. Clustering based technique could be one of the ways to achieve feature selection.
We propose a novel approach in which cluster accuracy is used as an indicator to select the set of features. We have carried out experiment on a live data set of UCI repository. All these datasets are large in size and also with multiple attributes. Some of the attributes are really not significant in the decision making process. We have identified the significant attributes from dermatology data set. However as a first step we had to select appropriate Clustering algorithm amongst the wide variety of clustering algorithm available in the Literature. We filtered out three best clustering algorithms. Suitability of an algorithm is domain specific. To finalize on the algorithm suitable for current dataset, we have done experimental evaluation of these three clustering algorithms on basic set of features that were selected using domain knowledge[14].The selection criteria broadly covered

all the cluster quality parameters such as cluster types(overlapped /non overlapped),accuracy of resultant clusters, consistency of clusters in successive iterations, centroid selection, time required to converge to a solution , scalability and the graph . The selected algorithm was then used to identify optimal features by extending the basic set using forward selection. Each feature set was evaluated based on the cluster quality. The parameters used were database indexing, entropy, purity, error rate and overlapping nature. The optimal feature set thus obtained will save resources in the next round of data collection and will increase the efficiency of the algorithms.[31]

II. Background

Technology advancement has led to exponential growth in the data with respect to dimensionality and sample size. Manual processing for these datasets is really impractical. Data mining and machine learning tools were proposed to automate pattern recognition and knowledge discovery process. Knowledge discovery from this data really requires storage and processing, which is a great challenge in the field of pattern recognition, statistics, and data mining. However, using data mining techniques on the collected data is mostly useless due to the high level of noise, missing, irrelevant attributes, variations etc. associated with collected samples[9]. The reason is either imperfection in the technologies that collected the data or the nature of the source of this data itself e.g. datasets crawled from the internet, are noisy by nature because it generally contains grammatical mistakes, misspelling, and improper punctuation. So extracting useful knowledge from such huge and noisy datasets is need of decision makers and challenge in front of researchers.[16]

Dimensionality reduction is one popular technique to remove noisy (i.e. irrelevant) and redundant attributes. This preprocessing technique can be categorized mainly into feature extraction and feature selection. Principle Component Analysis (PCA), Linear Discriminant Analysis (LDA), Singular Value Decomposition (SVD), are few of widely used feature extraction techniques. On the other hand, the feature selection approach aims to select a small subset of features that minimize redundancy and maximize relevance to the target (i.e. class label). Popular feature selection techniques include Relief, Fisher Score etc[4]. Information Gain, Chi Squares are test to evaluate selected features [1]. There are two main models of FS i.e Wrapper and Filter .Wrapper model uses classifier and a strategy. Wrapper is more superior to filter model in terms of classification accuracy but it is expensive and biased to the chosen classifier, so filter model is used for large datasets, which uses certain criteria to filter out the features.

A cluster is a collection data objects whose properties are similar in inter group and dissimilar in intra-group. Cluster analysis is finding this similarity between the data according to the characteristics found in data .There are five significant classes in which clustering algorithms can be grouped i.e. partitioning, hierarchical, density based , model based and grid based[14]. The wide variety is also due to wide range of applications of clustering approach. It is used to identify the Data distribution, Outlier detection, Data reduction, Summarization, Preprocessing for regression, PCA, Classification, finding characteristics/patterns for each group..Thus clustering is used for data reduction which is a preprocessing step for other algorithms. [10]

There are several issues that need to be addressed before using the clustering approach for a given Dataset. The interpretability and usability of results heavily depends on the cluster quality and several measures need to be taken to improve cluster quality. The noisy data, high dimensionality, input order, data characteristics such as types of attributes are some of the factors that affect quality of clusters and choosing the right algorithm and resolving some of these issues plays an important role in the success of clustering technique for decision making.[6] Clustering methods can be developed to learn more accurate cluster labels of the input samples, which guide feature selection simultaneously. Meanwhile, the cluster labels are also predicted by exploiting the hidden structure shared by different features, which can uncover feature correlations to make the results more reliable [19]. Choosing the initial centroids is always issue for all types of clustering. A careful and comprehensive study of data is required for the same. Also, if the initial clusters are not properly chosen, then after a few iterations it is found that clusters may even be left empty.[20]

If Feature selection technique is applied before applying any algorithm on sample data, then it reduces the cost and improves performance of learning of an algorithm by reducing the number of attributes. K-meansClustering is an unsupervised learning data mining technique that groups objects based upon distance or similarity[1][5]. Feature selection is used with k-means clustering algorithm and applied on dataset having 14 attributes in [8]. The result obtained when K-means clustering (Unsupervised grouping algorithm) is applied with feature selection is 50 % more accurate than just applying K-means clustering algorithm without using feature selection. Different algorithms are studied with respect to different parameters in literature and it is observed that improving stability of an algorithm by reducing noise remains a challenge. Applying iterative algorithm to identify maximum relevant feature set is used in[2]. Feature subset selection method iteratively selects subset of features from the database. Results based on 140,000 records show that the performance of the method exceeds by 70% than those of the other methods. The feature selection algorithm works in the iterations to find out feature subset. Error gets reduced at each iteration. Error calculation at each iteration indicates that, it is being minimized to get accuracy in the result.[3][7][12]. Feature selection method is mainly used to improve quality of clusters. Initialization parameters for identifying different clusters are an unresolved issue. Generally

many search strategies are being used for feature selection. Existing feature selection algorithms involve search strategies which reduces degree of optimality of final feature subset. [31]

There are different metrics for evaluating clustering algorithms. Database indexing is an internal evaluation scheme, which uses quantities and features inherent to the dataset. Cluster quality also can be measured using F-measure, entropy and precision. Entropy measures the uniformity or purity of a cluster and Precision directly reflects the performance of clustering .Entropy is used with various preprocessing methods such as wrapper, filter for feature elimination, reduction and selection [23,24,25,26,27,28].

III. Data Collection and Pre-processing:

Collected dataset is of dermatology from UCI online repository.[29]. If the entire data set is given as an input it not only needs more time to run the experiment but also produces inaccurate results. Decision support system can be fast, if representative and important data from input dataset is being chosen. Data consists of 366 X 34 dimensions and selected dimensions are 21. Base set attributes selected from domain knowledge.

IV. Selection of clustering algorithm

We carried out extensive literature survey of different clustering algorithms. Clustering algorithms are classified into partitioning, hierarchical, density based , model based and grid based ,each having pros and cons. Our first step involved selecting the clustering algorithm from two significant families i.e. hierarchical and partitioning that will be used for feature selection in the next step. Clustering algorithms depends on nature of data .So in the first step these algorithms are applied .We chose three algorithms and used the experimental setup to finalize the algorithm. The data set used was the reduced feature set based on domain knowledge. The cluster quality obtained was the criteria for selection of the algorithm.

In this section we present results for three clustering algorithms. Table 1 shows number of points present in each cluster. Table 2 indicates the different performance metrics to conclude the best algorithm. Lower the value of Database Index ,number of overlapped points and total withinness[27,28] indicates better performance of the algorithm.

1. K-means 2. Continuous K-means 3. Pairwise agglomerative algorithm.

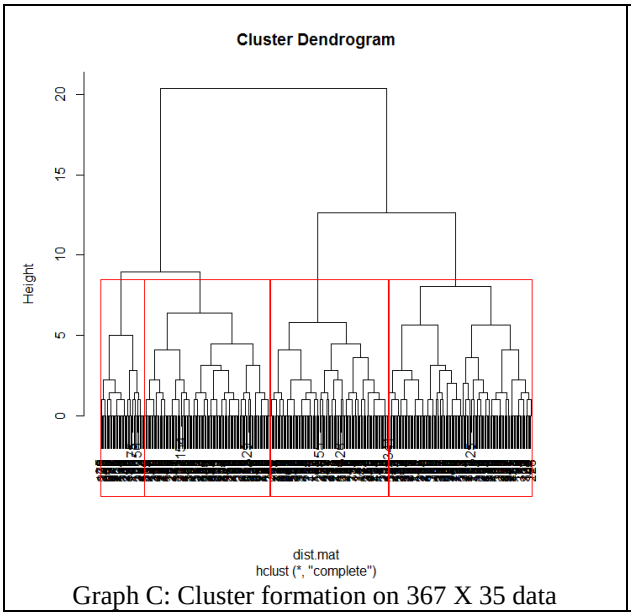
Table 1: Cluster size Table

Algorithm	C1	C2	C3	C4
K-means	150	100	95	71
continuous K-means	100	100	95	90
Pairwise Agglomerative	75	200	50	134

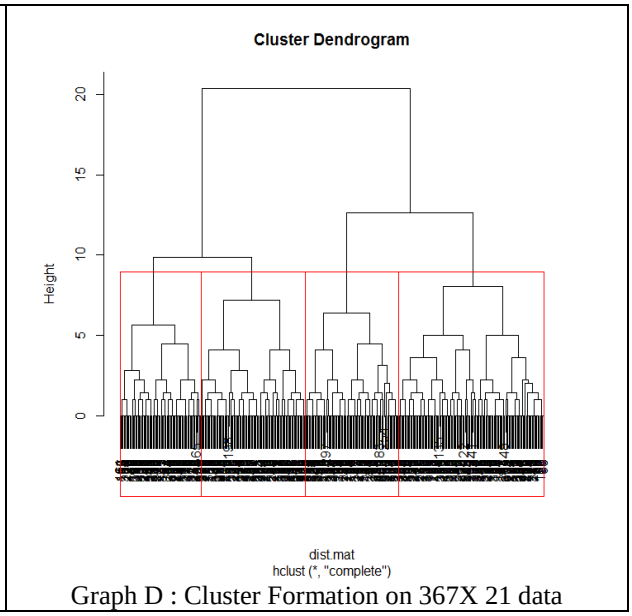
Table 2: Cluster Validation Table

Algorithm	Database Index	Overlapping Points	Total With Innes
K-means	.389	50	4256.3
continuous K-means	0.61	19	1432.5
Pairwise Agglomerative	.812	92	15564.

We can conclude that continuous k-means is the best algorithm which is used on more extended features in the next step



Graph C: Cluster formation on 367 X 35 data



Graph D : Cluster Formation on 367X 21 data

VII. Selection of optimal feature set

Complexity of algorithm will increase exponentially, if different combinations of existing features are considered. So continuous k-means is applied on domain specific enhanced feature set by forward FS approach. Graph A,B,C,D represent the variations of clusters wrt. number of parameters. Continuous K means and hierarchical agglomerative clustering algorithm is applied on enhanced features gives best clusters wrt to error rate, Database index, entropy, purity. It measures the average of the squares of the distances of each point and the centroid of each cluster. More the distance more is the error for that cluster. Better clusters are always with minimum entropy, less database index[27,28] and maximum purity. We can easily detect optimal number of features from table. The study shows that 21 features are selected by forward selection method [17]from given features for further analysis. The features are erythema, polygonal papules, follicular papules, clubbing, eosinophils, infiltrate, scalp, melanin, eosinophils, munlo, focal , granular ,horn plug, parakeratosis, damage, spongiosis, hyper, para, disappearance, saw, Para ketosis.

Table 3: Best Feature detection

No of features	DataBase Index	SMSE (sum mean square error)	entropy	purity
10	0.8769	46.1783	0.6812	.42
14	0.4831	34.5623	0.6214	.7136
21	0.2468	8.26	0.0901	.8981
26	0.4231	28.4213	0.3981	.7621
30	0.4293	14.1342	0.2682	.7845
34	0.6541	31.3431	0.3216	0.5928

VIII. Conclusion

Any application where clustering is to be used on input dataset can adopt this generic approach. There are different online data set repositories. These repositories are set of tables having multiple irrelevant attributes. These attributes decreases efficiency of the algorithm. To avoid so, the clustering algorithm can be chosen by applying the algorithms on a reduced feature set. Once the algorithm is finalized by considering different cluster evaluation parameters it can be used for feature selection. Thus Instead of applying clustering algorithm on the whole data set, if Feature selection technique is used for irrelevant dimension reduction, then algorithm will converge to solution with more speed and less error in turn will reduce the efforts of future data collection. So this two step process of selecting clustering algorithm and selecting optimal feature set can improve the efficiency of decision process.

References

[1] S Alelyani, J Tang, H Liu,Feature Selection for Clustering: A Review. public.asu.edu 2013.
[2] Ritu Ganda ,Vijay Chahar ,A Comparative Study on Feature Selection Using Data Mining Tools, International Journal of Advanced Research in Computer Science and Software Engineering, 2013 .
[3] Ren Diao , Qiang Shen ,Feature Selection With Harmony Search ,IEEE Systems, Man, and Cybernetics, 2012.
[4] Salem Alelyani, Thesis On Feature Selection Stability: A Data Perspective,2012.
[5] Sunita Beniwal, Jitender Arora Classification and Feature Selection Techniques in Data Mining, International Journal of Engineering Research & Technology (IJERT),Vol. 1 Issue 6, August – 2012.

[6] Gauthier Doquire and Michel Verleysen,s,Mutual information for feature selection with missing datacomputational Intelligence and Machine Learning,2011.
[7] Farahat, A.K. ,Data Mining (ICDM), An Efficient Greedy Method for Unsupervised Feature Selection, IEEE,2005.
[8] Ren Diao ,Qiang Shen, Two New Approaches to Feature Selection with Harmony Search, IEEE World Congress on Computational Intelligence ,2010.
[9] M. Ramaswami and R. Bhaskaran,A Study on Feature Selection Techniques in Educational Data Mining,journal of computing,2009.
[10] Liu, H. Torkkola, K. ,Evolving feature selection ,Intelligent Systems, IEEE,2005.
[11] Huan Liu and Lei Yu, Toward Integrating Feature Selection Algorithms for Classification and Clustering,
[12] Hiroshi Mamitsuka ,Principles of Data Mining and Knowledge Discovery ,Lecture Notes in Computer Science springer 2002.
[13] Lesh, N. MERL, Zaki, M.J.,Scalable feature mining for sequential data, Intelligent Systems and their Applications, IEEE (Volume:15 , Issue: 2) ,2000.
[14] Xiaofei He, Ming Ji, Chiyuan Zhang, and Hujun Bao,A Variance Minimization Criterion to Feature Selection Using Laplacian Regularization, IEEE Transaction on Pattern Analysis and machine Intelligence, VOL. 33, NO. 10, 2011.
[15] L.V. Bijuraj,Clustering and its Applications ,Proceedings of National Conference on New Horizons in IT - NCNHIT ,2013.
[16] Lan H. Witten, Eibe Frank, Data miming:practical machine tools and techniques.
[17] Huan Liu,Evolving Feature Selection,Intelligent Systems, IEEE (Volume:20 , Issue: 6) ,2005.
[18] Jiawei Han, Micheline Kamber ,Data Mining: Concepts and Techniques .
[19] Zechao Li et al , Clustering-Guided Sparse Structural Learning for Unsupervised Feature Selection ,IEEE Transactions on Knowledge and Data Engineering (TKDE),2014.
[20] Parul Agarwal ,M. Afshar Alam,Ranjit Biswas,Issues, Challenges and Tools of Clustering Algorithms , IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 2, May 2011.
[21] Prafulla Bafna1 , Hema Gaikwad , A hybrid approach to measure design improvement factor of website., American International Journal of Research in Science, Technology, Engineering & Mathematics" (ISSN (Print): 2328-3491, ISSN (Online): 2328-3580, ISSN (CD-ROM): 2328-362) December-2014 to February-2015.
[22] A novel clustering approach to select optimal usability principles for educational websites ,,Prafulla Bafna, International Journal of Software and Web Sciences , ISSN (Print): 2279-0063, ISSN (Online): 2279-0071, Issue 11, December-2014 to February-2015 Sciences(IJSWS).
[23] Zitao Liu et al, A feature selection method for document clustering based on part-of-speech and word co- occurrence, IEEE Conference on Fuzzy Systems and Knowledge Discovery, volume: 5 , pg 2331 - 2334 ,2010
[24] Mohammad-Amin Jashki et al ,An iterative hybrid filter-wrapper approach to feature selection for document clustering, Proceedings of the 22nd Canadian Conference on Advances in Artificial Intelligence Volume 5549, 2009, pp 74-85.
[25] Huan Liu, Evolving feature selection, intelligent systems, IEEE (Volume:20 , Issue: 6) , pp 64 - 76 ,2005
[26] Xiaofei He, Ming Ji, Chiyuan Zhang, and Hujun Bao,A variance minimization criterion to feature selection using laplacian regularization,IEEE Transaction on Pattern Analysis and machine Intillegence, VOL. 33, NO. 10, 2011
[27] Sriparna Saha , Sanghmitra Bandyopadya,Some connectivity based cluster validity indices, ACM Applied Soft Computing Volume 12 Issue 5, May, 2012 Pages 1555-1565
[28] Mamta Mittal,R.K.Sharma,V.P.Singh Validation of K-means and Threshold based Clustering Method , International Journal of Advancements in Technology,pp153-160.
[29] https://archive.ics.uci.edu/ml/datasets/Dermatology available on 01/03/2015
[30] Prafulla Bafna,, Pravin Metkewar, Shailaja Shirwaikar, ,Novel Clustering approach for Feature Selection , American International Journal of Research in Science, Technology, Engineering & Mathematics,pp62-67,2014