International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 7 (2018) pp. 4614-4619

Similarity Search using Cluster based Ensemble Classification

M. Blessa Binolin Pepsi
Assistant Professor, Department of Information Technology
Mepco Schlenk Engineering College, Sivakasi, Tamilnadu, India.

Abstract:

An automatic retrieval system for similarity search using cluster based ensemble classifier is proposed in this work to help the doctors. The proposed system has designed with the modules (i) metric database construction (ii) cluster based ensemble classifier and (iii) similar image / test report retrieval. Metric database is constructed by first and second order features of lung tomography images available in the internet, heart, liver and diabetes datasets collected from UCI machine learning repository. Dataset is further reduced by association rule mining as a feature vector. The cluster based ensemble classifier technique includes two phases namely training and testing phase. The phase clusters the reduced feature vector into groups using clustering algorithms (k-means and fuzzy c-means). The defined cluster id is passed along with the metric data as an attribute value. These cluster vectors are further classified using ensemble classifier with the decision tree, SVM and bayesian classifiers. Accuracy obtained by cluster based ensemble classifier is 95.33% for lung images, 83.14% for diabetes and 86.97% for heart dataset and 93.32% for liver dataset which is higher than the accuracy obtained by clustering or by ensemble classifier (mixture of decision tree, SVM and Bayesian classifier) or by individual classifiers namely decision tree classifier, SVM or Bayesian classifier. Hence cluster based ensemble classifier is better for similarity search than other methods

Keywords: Association rule mining, clustering, Decision Tree classifier, SVM, Bayesian classifier, Ensemble Classifier

INTRODUCTION

Data mining task is the automatic or semi automatic analysis of large quantities of data to extract interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and association rule mining. This field can be successfully applied to various areas like information retrieval, similarity search, medical treatment planning, diagnosis, drug design, financial analysis, etc. Similarity search retrieves records which have closest match but not exact match [3]. This helps the doctors to plan for treatment by going through the similar records. In other areas also similarity search helps to explore the fast experiences.

Clustering is the process of grouping similar objects based on cluster parameters [8]. Classification predicts class labels and classifies objects based on the training set. Ensemble classification combines the prediction of the multiple base classifiers to assign the class label. Ensemble classifiers are mixture of experts [13]**.** An ensemble classifier produces more accurate results than its individual counterparts provided the base classifiers are uncorrelated [15] Two classifiers are diverse if they make different errors on different instances. Active research is going on in fusing clustering and ensemble classification.

Data is clustered with different cluster parameters into different segments and the training set is given to different base classifiers and decision for a test pattern is taken by majority voting. This approach [4] performs significantly well than ensemble classifiers. Data is clustered and given to ensemble classifiers to learn cluster confidence and fusion classifier is trained with class label. While testing cluster confidence is predicted by ensemble classifier and class label is identified by fusion classifier. This method [4] improves accuracy. Impact of clustering and number of clusters on ensemble classifier is demonstrated. A semi-supervised kernel learning framework is proposed for image retrieval [1] but the reliability is based on classification confidence which is not the best accurate for labeled images

An algorithm for deciding when to stop building classifiers is [16] developed using ensemble creation with decision tree classifier and the performance are analyzed for bagging and randomization based ensemble creation technique. Ensembling is done [12] by selecting a subset of instances from the training set achieving the same classification performance as a whole training set and combining boosting method. Pattern Mining based Ensemble pruning algorithm [14] takes the prediction of the base classifiers as a transaction and used fp-growth algorithm to reduce the number of base classifiers. Mining finds the possible ensemble size and the one which output best accuracy is selected. This ensemble pruning outperforms bagging. The features are selected [6] by two steps for classification. First data is clustered by graph theoretic clustering into clusters with subset of features. The relevant features to the target class are selected from each cluster ie. subset of features to form a feature vector. High dimensional data is classified by this approach which proves the accuracy improvement than other feature selection algorithms.

Computed Tomography brain images are classified as benign and malignant [15] using hybrid classifier built with decision tree classifier and association rule mining with recognition rate of 95%. For classifying mammograms, apriori algorithm with association rule mining was used and reported that the technique handles well even in imbalanced data set and outperforms neural network classifier back propagation with single hidden layer with a recognition rate of 70%.

In the proposed work, data is clustered by clustering algorithms and along with cluster confidence ensemble classifier is trained to predict the class label. Then class label and cluster confidence is used for retrieving similar images or patient records with id from the database. The structure of the paper is organized as follows: Section II discusses the background. Section III portraits the proposed system. Section IV deals with the experimental setup with dataset used. Section V discusses the results. Section VI concludes with future direction for further research.

BACKGROUND

Data set is a collection of feature vectors representing different objects / patterns and different categories / classes. When the data is clustered by partitioning methods each cluster will have objects belong to different classes. Each cluster contains patterns that are very close in Euclidean space. The clusters have well defined easy to learn boundaries. Hence, if the patterns are trained with cluster confidence, the classifiers learn them with high accuracy.

In a data set, if the classes are well separated then clusters may have same class data. This is termed as atomic cluster. But in reality, it is not possible because of overlapping patterns. Hence higher number of clusters which is higher than the number of classes will have atomic or near atomic cluster. This improves the learning thereby increases prediction accuracy.

If clustering is done within the class then it is termed as homogeneous clustering. It produces only atomic clusters. But small training set in a class produces more atomic clusters with little data, because 3 classes with 4 clusters produce 12 clusters, which leads to memorization not generalization. Non uniform number of clusters for classes depending on the size of the data set also can be adopted. This requires complete knowledge about the data set. Inadequate knowledge reduces accuracy. Too many clusters also misclassify patterns which reduces accuracy. Hence optimal number of clusters which is greater than the number of classes should be identified for the data set.

If clustering is done with all the patterns without the knowledge of class, then it is termed as heterogeneous clustering. For real time data, this type of clustering is suitable because of inadequate prior knowledge. This groups the similar objects in different classes which are very close in euclidean space. Hence predicting class label with cluster boundary requires powerful classifier also.

Since ensemble classification outperforms the base classifiers, for predicting class label, ensemble classifier is used. The base classifiers chosen are different type of classifiers because each classification algorithm behaves differently towards training set and test pattern. This satisfies diversity property.

Based on cluster confidence and class label, objects are retrieved from the database. This reduces the retrieval error because the distance metric as well as class label, both controls the retrieval.

PROPOSED SYSTEM DESIGN

The block diagram of the proposed system is given in Fig. 1.

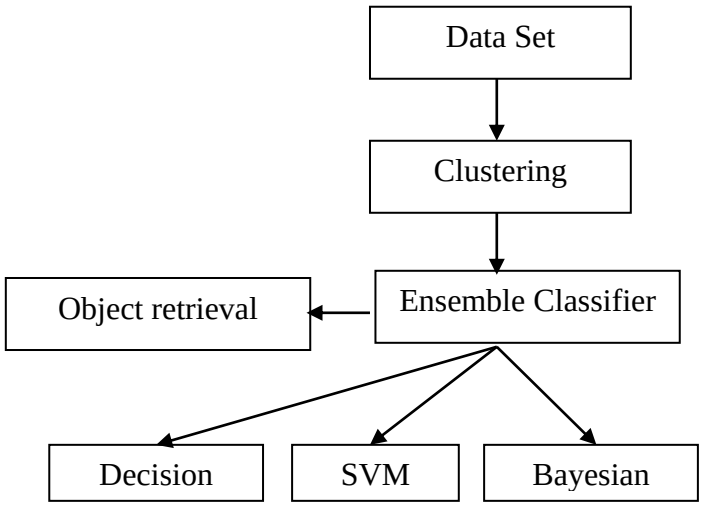

Figure 1. Similarity Search Architecture

The data set is a collection of patterns / objects represented by feature vectors. Let each pattern X is denoted by $\{x_1, x_2, x_3, \ldots, x_i\}$. The objective is to identify the patterns that are close to the given query pattern. The query pattern need not be available in the data set which justifies the requirement of closest match not exact match.

In training phase the data set is clustered by clustering algorithm into K clusters and cluster confidence C_k is added to the feature vector where k takes the value from 1 to K,. The modified data set is given as a training set to the ensemble classifier. Ensemble classifier is the set of base classifiers in which the base classifier is denoted by Bj where j takes the value from 1 to number of classifiers. It outputs class labels Tjc, where j denotes the base classifier and c denotes the class label predicted by B_j. They are ensemble to predict the class label Ec where c takes the value from 1 to number of classes.

In testing phase the test pattern X is assigned with the cluster confidence C_k based on the similarity measure with the cluster centroids. Then along with cluster confidence, modified data is given to ensemble classifer to predict target class label Ec. Then (C_k, Ec) is used to retrieve the patterns and object id is used to retrieve the relevant images / patient records.

This work also aims to investigate the performance of heterogeneous clustering and homogeneous clustering along with ensemble based classification in similarity search, whether the clustering based ensemble classifier outperforms base classifiers, whether the clustering based base classifiers outperform base classifier, to find the optimum number of clusters and cluster size because this contributes to the relevant number of images.

A. Feature Reduction :

Real world objects are represented by their features. But they contain redundant and irrelevant features. Hence

representative features are selected by feature reduction technique.

It is defined as a process of selecting n features among m from a set of features D = {d1, d2, d3… dm} features where n < m. In this work, association rule mining technique is used for selecting the relevant, useful and independent features which satisfies minimum support and confidence. Now X is a reduced subset representing the pattern.

B. Clustering based ensemble classification :

Data set is clustered by heterogeneous and homogeneous clustering. Cluster confidence is added to the data set as one feature / attribute. Then the data set along with cluster confidence is given to ensemble classifier for learning the class label. Clustering algorithm chosen are K means clustering and Fuzzy C Means Clustering. The base classifiers chosen are Decision tree classifier, Support Vector Machine (SVM) and Bayesian Classifier.

In testing phase, the test pattern is compared with the cluster centroids and cluster confidence is assigned. Then along with the cluster confidence, it is given to the individual base classifiers for predicting the class label. The prediction result is the ensemble of the prediction by the base classifiers. Here it is done by majority voting. Hence mode is used to identify the prediction.

C. Clustering

K-means clustering partition n vectors into k clusters in which each vector belongs to the cluster with the nearest mean. The value of k that is number of clusters must be predefined. The steps to perform k-means clustering is stated as the dataset is partitioned into K clusters and the data points are randomly assigned to the clusters resulting in clusters that have roughly the same number of data points.

In fuzzy clustering, each point has a degree belonging to clusters as in fuzzy logic. The steps followed to perform fuzzy c means clustering are the choice of number of clusters and assign it randomly to each point coefficients for being in the clusters. The step repeats until the coefficients change between two iterations not more than sensitivity threshold. The centroid is computed for each cluster using the equation **(1.1).**

$$C_k = \frac{\sum_x w_k(x)\, x}{\sum_x w_k(x)} \quad \text{-- Eq. (1.1)}$$

Where x is a point which has a set of coefficients and k is the cluster and $w_k(x)$ is the degree. For each point, the coefficients of clusters are computed.

D. Ensemble Classifier

An ensemble of classifiers **(Fig 2)** is a set of classifiers whose individual class result decisions are combined in a way to classify with new class labels. Ensemble classifier predicts class labels aggregating predictions made by multiple base classifiers. Combining predictions of an ensemble is more accurate than the individual classifiers and so we go for ensemble classification. Uncorrelated errors of individual classifiers can be eliminated. The combinations of various base classifiers are based on majority voting.

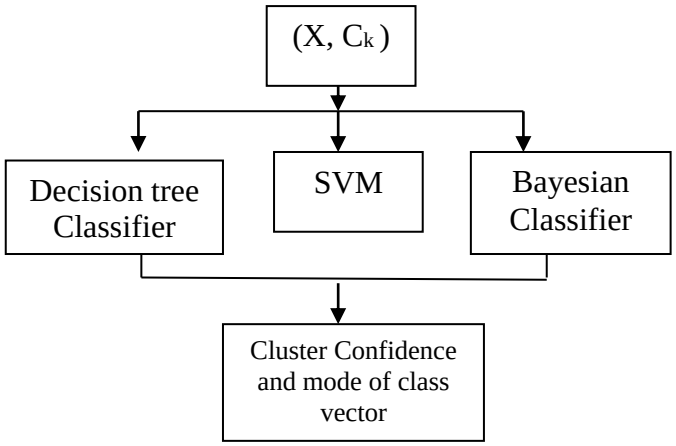

Figure 2. Ensemble Classifier

Decision tree is a flowchart like tree structure which is a predictive model to obtain a target value at conclusion. In these tree structures, leaves represent class labels and internal node denotes a test on an attribute and branches represent an outcome of the test. The use of decision tree is to test the attribute values of the sample against the decision tree and target class label is decided. SVM is a supervised learning model with algorithms that analyze data and recognize patterns that is used for classification. It constructs a hyper plane or set of hyper planes in a high or infinite-dimensional space which can be used for classification. A Bayes classifier is a simple probabilistic classifier based on Bayes' theorem with strong independent assumptions. Bayes' theorem is stated as,

$$P(T|E) = \frac{P(E|T)\times P(T)}{P(E|T)\times P(T)+P(E|\neg T)\times P(\neg T)} \quad \text{-- Eqn. (1.2)}$$

where T stands for a theory or hypothesis and E represents a new piece of evidence that seems to confirm or disconfirm the theory. A naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. The naive Bayes classifier requires a small amount of training data to estimate the parameters for classification.

E. Object Retrieval

The output of the ensemble classifier is the predicted class label along with cluster confidence. It is used to retrieve the objects from the data base. Hence for the test pattern the class label is predicted and the similar objects from the database are retrieved.

The objects are stored in the database with object id and

features. The object id is used to retrieve the hidden information there by maintaining privacy. This is very much essential in retrieval of medical data and military data

EXPERIMENTAL SETUP

The methodology is implemented in MATLAB. The number of clusters is varied from 2 to 10 for analysis. K Means clustering and Fuzzy C Means Clustering are used for clustering the data set. Before clustering, the features are reduced by association rule mining with support value of 5 and confidence value of 0.3. From the cluster top 10 relevant patterns are retrieved.

The data set used is bench mark datasets retrieved from UCI machine learning repository [20]. Experimented are liver, diabetes and heart dataset. Also publicly available CT lung image data is used for testing. The dataset of lung CT images includes the following lung diseases: sarcoidosis, atlectasis, pneumonia, cystic fibriosis, chronic obstructive pulmonary disease, interstitial lung disease and respiratory bronchiolitis.

They are stored as a collection of 110 images in the image database with 3 classes (upper, middle and lower lung affected). First order texture feature Skew and Kurtosis and second order texture feature are extracted from the images and stored in the database with object id. For the liver, heart and diabetes dataset the classes taken are 2 (normal and abnormal).

RESULTS AND DISCUSSION

The proposed system is tested by the overall percentage of object retrieval in accuracy measure. (How it differs from precision and recall) along with feature selection using Association rule mining result. The comparison of accuracy measure is made between clustering, cluster based base classifiers and cluster based ensemble classifier.

The clustering technique is performed to divide into K clusters. The accuracy measure comparison is made while using K means and Fuzzy C Means with the base classifiers and ensemble classifier. This is tabulated in Table 1 for liver, Table 2 for heart, Table 3 for diabetes and Table 4 for lung dataset.

Table 1: Performance Comparison – Liver dataset

Clustering Technique	Decision Tree	Bayesian	SVM	Ensemble
K Means	59.471	79.295	82.378	88.105
Fuzzy C means	84.67	85.12	89.51	93.86

Table 2: Performance Comparison – Heart dataset

Clustering Technique	Decision Tree	Bayesian	SVM	Ensemble
K - means	84.81	74.315	68.289	86.184
Fuzzy c means	85.52	74.67	65.83	87.10

Table 3: Performance Comparison – Diabetes dataset

Clustering Technique	Decision Tree	Bayesian	SVM	Ensemble
K - means	70	81.165	73.846	80
Fuzzy c means	72.57	83.14	77.89	83.10

Table 4: Performance Comparison – Lung dataset

Clustering Technique	Decision Tree	Bayesian	SVM	Ensemble
K - means	85.55	97.26	93.94	100
Fuzzy c means	95.32	98.28	99.28	100

Experimental results show that the number of clusters must be more than the number of classes for datasets. The optimum number of clusters for the lung dataset is assumed to be 10 because the database consists of images affected by ten diseases and classified into 3 classes (upper, middle and lower lung affected). While for other datasets, the optimum number of clusters is tested for 2, 4, 8 and 10 clusters and classified into 2 classes (normal or abnormal). These values are plotted in bar chart and given in Fig 3 for liver, Fig 4 for heart and Fig 5 for diabetes dataset.

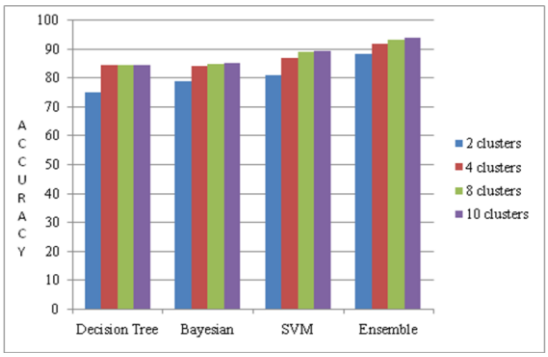

Figure 3: Performance Analysis – Number of cluster (liver)

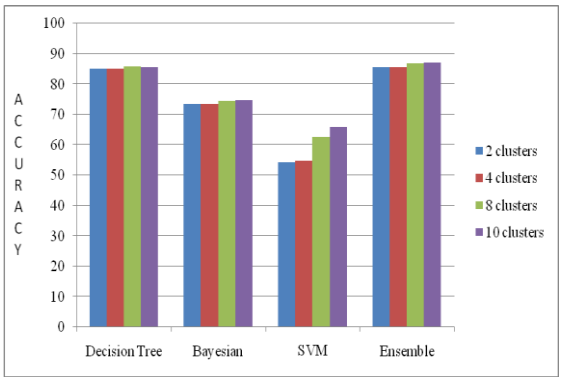

Figure 4: Performance Analysis – Number of cluster (heart)

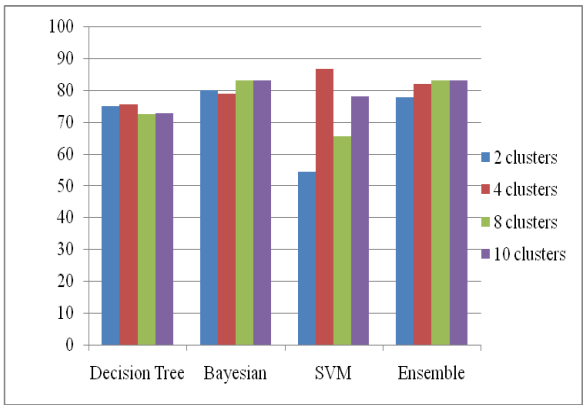

Figure 5: Performance Analysis–Number of cluster (diabetes)

To justify the prediction using association rule mining, results are taken with and without feature reduction and tabulated.

Table 5: Performance Evaluation – with & without association rule mining

Data	Without AR	With AR
Lung	83.93	90.09
Heart	73.36	85.68
Liver	83.45	92.01
Diabetes	76.67	81.88

CONCLUSION

The findings obtained from the proposed work are, the similarity retrieval using cluster based ensemble classifier is found to be good with high query accuracy compared to clustering and individual base classifiers. The optimum number of clusters is analyzed to be more than the number of clusters. The accuracy measure is high when features are reduced by association rule mining and retrieval is good using fuzzy c means clustering. The optimum number of clusters can be found and determined with more data. It can also include the concept of providing privacy to the sensitive data (i.e. medical record) against attacks from untrusted clients across the networks. The system can be also compared and analyzed with various incremental clustering methods and other classification methods.

REFERENCES

[1] Jianqing Liang, Qinghua Hu et. al, "*Semisupervised Online Multikernel Similarity Learning for Image Retrieval*", IEEE Transactions on Multimedia, Vol. 19, No. 5, May 2017

[2] Pengcheng Wu, Steven C. H. Hoi, Peilin Zhao, Chunyan Miao, and Zhi-Yong Liu, "*Online Multi-Modal Distance Metric Learning with Application to Image Retrieval*", IEEE Transactions on Knowledge and Data Engineering, Vol. 28, No. 2, February 2016

[3] Dr.K.Mala, M.Blessa Binolin Pepsi, "*Similarity Search on Metric Data of Outsourced Lung Images*", IEEE International Conference on Green High Performance Computing, 2013

[4] Ashfaqur Rahman, Brjjesh Verma, "*Novel Layered Clustering-Based Approach for Generating Ensemble of Classifiers*", IEEE Transactions on Neural Networks, vol 22, No. 5, 2011

[5] Brjjesh Verma, Ashfaqur Rahman, "*Cluster-Oriented Ensemble Classifier: Impact of Multicluster Characterization on Ensemble Classifier Learning*", IEEE Transactions On Knowledge and Data Engineering, Vol. 24, No. 4, 2012

[6] Qinbao Song, Jingjie Ni, Guangtao Wan, "*A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data*", IEEE Transactions on knowledge and Data Engineering, vol. 25, No. 1, 2013

[7] Jiawei Han, Hong Cheng,· Dong Xin, Xifeng Yan, "*Frequent pattern mining: current status and future Directions*" Data Mining and Knowledge Discovery. vol.15, pp. 55–86, 2007

[8] Johannes Grabmeier, Andreas Rudolph, "*Techniques of Cluster Algorithms in Data Mining*", Data Mining and Knowledge Discovery, vol. 6, pp. 303–360, 2002

[9] Kumar.U, Raja, S.K, Ramachandra, T.V., "*Hybrid Bayesian Classifier for Improved Classification Accuracy*", Geoscience and Remote Sensing Letters, IEEE vol. 8 , Issue: 3, 2011

[10] Man Lung Yiu, Ira Assent, Christian S. Jensen and Panos Kalnis, "Outsourced Similarity Search on Metric Data Assets", IEEE Transactions on Knowledge and Data Engineering, Vol. 24: pp.338-352, 2012

[11] Maria-Luiza Antonie, Osmar R. Zaiane, Alexandru Coman, "Application of Data Mining Techniques for Medical image classification", Proceedings of the Second International Workshop on Multimedia Data Mining (MDM/KDD'2001) in conjunction with ACM SIGKDD conference, San Francisco, USA, 2001

[12] Nicolas Garcia-Pedrajas, "Constructing Ensembles of Classifiers by Means of Weighted Instance Selection", IEEE Transactions on Neural Networks, vol 20, no. 2., 2009

[13] R.Polikar, "Ensemble Based Systems in Decision Making", IEEE Circuits and Systems Magazine, vol 6, no. 3, pp 21-45, 2006

[14] Qiang-Li Zhao, Yan-Huang Jiang, Ming Xu, "A fast ensemble pruning algorithm based on pattern mining process", Data Mining and Knowledge Discovery Vol. 19 pp. 277-292, 2009

[15] P.Rajendran, M.Madheswaran, "*Hybrid Image Classification using Association Rule Mining with Decision Tree Algorithm*", Journal of Computing, vol. 2, issue 1, pp. 127 – 136, 2010

[16] Robert E. Banfield, Lawrence O. Hall, Kevin Bowyer, "*A Comparision of Decision Tree Ensemble Creation Technique*", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, No. 1, 2007

[17] Rajeev Rastogi Kyuseok Shim, "*PUBLIC: A Decision Tree Classifier that Integrates Building and Pruning*", Data Mining and Knowledge Discovery, vol. 4, pp. 315–344, 2000

[18] Themis P. Exarchos, Costas Papaloukas, Dimitrios I. Fotiadis and Lampros K. Michalis , "*An Association Rule Mining-Based Methodology for Automated Detection of Ischemic ECG Beats*", IEEE Transactions on Biomedical Engineering, vol. 53, No. 8, 2006

[19] Yi Lin, "*Support Vector Machines and the Bayes Rule in Classification*", Data Mining and Knowledge Discovery, vol. 6, pp. 259–275, 2002

[20] www.archive.ics.uci.edu/ml/datasets.html

[21] www. radiopaedia.org

[22] www. mevis_research.de

[23] www. radiology.vcu.edu

[24] www. radiographics.rsna.org

[25] www. learningradiology.com