Knowledge discovery from database Using an integration of clustering and classification

Varun Kumar
Department of Computer Science and Engineering
ITM University
Gurgaon, India
kumarvarun333@gmail.com

Nisha Rathee
Department of Computer Science and Engineering
ITM University
Gurgaon, India
nisharathee29@gmail.com

***Abstract*— Clustering and classification are two important techniques of data mining. Classification is a supervised learning problem of assigning an object to one of several pre-defined categories based upon the attributes of the object. While, clustering is an unsupervised learning problem that group objects based upon distance or similarity. Each group is known as a cluster. In this paper we make use of a large database 'Fisher's Iris Dataset' containing 5 attributes and 150 instances to perform an integration of clustering and classification techniques of data mining. We compared results of simple classification technique (using J48 classifier) with the results of integration of clustering and classification technique, based upon various parameters using WEKA (Waikato Environment for Knowledge Analysis), a Data Mining tool. The results of the experiment show that integration of clustering and classification gives promising results with utmost accuracy rate and robustness even when the data set is containing missing values**.

Keywords- Data Mining; J48; KMEANS; WEKA; Fisher's Iris dataset;

I. INTRODUCTION

Data mining is the process of automatic classification of cases based on data patterns obtained from a dataset. A number of algorithms have been developed and implemented to extract information and discover knowledge patterns that may be useful for decision support [2]. Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases [1]. Several data mining techniques are pattern recognition, clustering, association, classification and clustering [7]. The proposed work will focus on challenges related to integration of clustering and classification techniques. Classification has been identified as an important problem in the emerging field of data mining [5]. Given our goal of classifying large data sets, we focus mainly on decision tree classifiers [8] [9]. Decision tree classifiers are relatively fast as compared to other classification methods. A decision tree can be converted into simple and easy to understand classification rules [10].

Finally, tree classifiers obtained similar and sometimes better accuracy when compared with other classification methods [11]. Clustering is the unsupervised classification of patterns into clusters [6].The community of users has played lot emphasis on developing fast algorithms for clustering large datasets [14].It groups similar objects together in a cluster (or clusters) and dissimilar objects in other cluster (or clusters) [12]. In this paper WEKA (Waikato Environment for knowledge analysis) machine learning tool [13][18] is used for performing clustering and classification algorithms. The dataset used in this paper is Fisher's Iris dataset, consists of 50 samples from each of three species of Iris flowers (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample; they are the length and the width of sepal and petal, in centimeters. Based on the combination of the four features, Fisher developed a linear discriminant model to distinguish the species from each other.

A. Organisation of the paper

The paper is organized as follows: Section 2 defines problem statement. Section 3 describes the proposed classification method to identify the class of Iris flower as Iris-setosa, Iris-versicolor or Iris-virginica using data mining classification algorithm and an integration of clustering and classification technique of data mining. Experimental results and performance evaluation are presented in Section 4 and finally, Section 5 concludes the paper and points out some potential future work.

II. PROBLEM STATEMENT

The problem in particular is a comparative study of classification technique algorithm J48 with an integration of SimpleKMeans clusterer and J48 classifier on various parameters using Fisher's Iris Dataset containing 5 attributes and 150 instances.

III. PROPOSED METHOD

Classification is the process of finding a set of models that describe and distinguish data classes and concepts, for the purpose of being able to use the model to predict the class whose label is unknown. Clustering is different from classification as it builds the classes (which are not known in advance) based upon similarity between object features. Fig. 1 shows a general framework of an integration of clustering and classification process. Integration of clustering and classification technique is useful even when the dataset contains missing values. Fig. 2 shows the block diagram of

steps of evaluation and comparison. In this experiment, object corresponds to Iris flower, and object class label corresponds to species of Iris flower. Every Iris flower consists of length and width of petal and sepal, which are used to predict the species of Iris flower. Apply classification technique (J48 classifier) using WEKA tool. Classification is a two step process, first, it build classification model using training data. Every object of the dataset must be pre-classified i.e. its class label must be known, second the model generated in the preceding step is tested by assigning class labels to data objects in a test dataset.

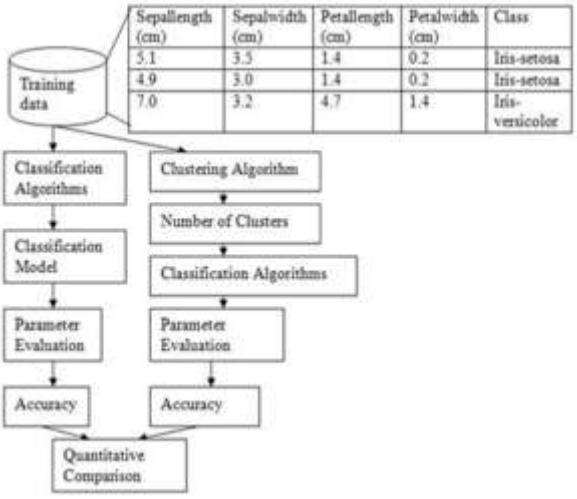


Figure 1. Proposed Classification Model

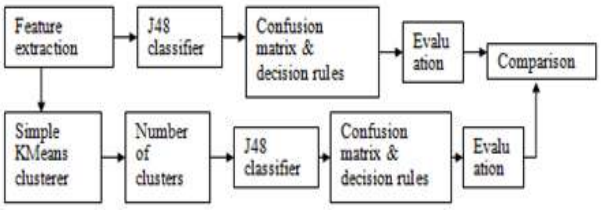


Figure 2. Block Diagram

The test data may be different from the training data. Every element of the test data is also preclassified in advance. The accuracy of the classification model is determined by comparing true class labels in the testing set with those assigned by the model. Apply clustering technique on the original data set using WEKA tool and now we are come up with a number of clusters. It also adds an attribute 'cluster' to the data set. Apply classification technique on the clustering result data set. Then compare the results of simple classification and an integration of clustering and classification. In this paper, we identified the finest classification rules through experimental study for the task of classifying Iris flower type in to Iris setosa, Iris versicolor, or Iris virginica species using Weka data mining tool.

A. Iris Dataset Preprocessing

We make use of a database 'Fisher's Iris dataset' containing 5 attributes and 150 instances to perform comparative study of data mining classification algorithm namely J48(C4.5) and an integration of Simple KMeans clustering algorithm and J48 classification algorithm. Prior to indexing and classification, a preprocessing step was performed. The Fisher's Iris Database is available on UCI Machine Learning Repository website http://archive.ics.uci.edu:80/ml/datasets.html in Excel Format i.e. .xls file. In order to perform experiment using WEKA [20], the file format for Iris database has been changed to .arff or .csv file.

The complete description of the of attribute value are presented in Table 1. A sample training data set is also given in Table 2 .During clustering technique we add an attribute i.e. 'cluster' to the data set and use filtered clusterer with SimpleKMeans algorithms which removes the use of 5,6 attribute during clustering and add the resulting cluster to which each instance belongs to, along with classes to the dataset.

TABLE I. COMPLETE DESCRIPTION OF VARIABLES

Variable/Attributes	Category	Possible Values
Sepallength	Numeric	4-8
Sepalwidth	Numeric	1-5
Petallength	Numeric	1-8
Petalwidth	Numeric	0-3
Class	Nominal	Three, Iris-setosa, Iris-virginica & Iris versicolor

TABLE II. SAMPLE OF INSTANCES FROM

sepal length	Sepal Width	Petal Length	Petal width	class
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3.0	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
4.6	3.1	1.5	0.2	iris-setosa
5.0	3.6	1.4	0.2	Iris-setosa
7.0	3.2	4.7	1.4	Iris-versicolor
6.4	3.2	4.5	1.5	iris-versicolor
7.6	3.0	6.6	2.1	Iris-virginica

B. Building Classifiers

1) J48(C4.5): J48 is an implementation of C4.5[17] that builds decision trees from a set of training data in the same way as ID3, using the concept of Information Entropy. . The training data is a set S = s1, s2... of already classified samples. Each sample si = x1, x2... is a vector where x1, x2… represent attributes or features of the sample. Decision tree are efficient to use and display good accuracy for large amount of data. At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other.

2) KMeans clusterer: Simple KMeans is one of the simplest clustering algorithms [4].KMeans algorithm is a classical clustering method that group large datasets in to clusters[15][16]. The procedure follows a simple way to

classify a given data set through a certain number of clusters. It select k points as initial centriods and find K clusters by assigning data instances to nearest centroids. Distance measure used to find centroids is Euclidean distance.

3) Measures for performance evaluation: To measure the performance, two concepts sensitivity and specificity are often used; these concepts are readily usable for the evaluation of any binary classifier. TP is true positive, FP is false positive, TN is true negative and FN is false negative. TPR is true positive rate, it is equivalent to Recall.

$$sensitivity = TPR = \frac{TP}{TP + FN} \quad (1)$$

$$specificity = \frac{TN}{FP + TN} \quad (2)$$

a) Confusion Matrix: Fig. 3 shows the confusion matrix of three class problem .If we evaluate a set of objects, we can count the outcomes and prepare a confusion matrix (also known as a contingency table), a three-three (as Iris dataset contain three classes) table that shows the classifier's correct decisions on a major diagonal and the errors off this diagonal.

		predicted		
		positive link	negative link	non-existent link
actual	positive link	TP	FP	FN
	negative link	FP	TP	FN
	non-existent link	FP	FP	TN

Figure 3. Confusion Matrix

The columns represent the predictions and the rows represent the actual class [3]. An edge is denoted as true positive (TP), if it is a positive or negative link and predicted also as a positive or negative link, respectively. False positives (FP) are all predicted positive or negative links which are not correctly predicted, i.e., either they are non-existent or they have another sign in the reference network. As true negatives (TN) we denote correctly predicted non-existent edges and as false negatives (FN) falsely predicted non-existent edges are defined i.e., an edge is predicted to be non-existent but it is a positive or a negative link in the reference network.

b) Precision: In information retrieval positive predictive value is called precision. It is calculated as number of correctly classified instances belongs to X divided by number of instances classified as belonging to class X; that is, it is the proportion of true positives out of all positive results. It can be defined as:

$$precision = \frac{TP}{TP + FP} \quad (3)$$

c) Accuracy: It is simply a ratio of ((no. of correctly classified instances) / (total no. of instances)) *100). Technically it can be defined as:

$$accuracy = \frac{TP + TN}{(TP + FN) + (FN + TN)} \quad (4)$$

d) False Positive Rate : It is simply the ratio of false positives to false positives plus true negatives. In an ideal world we want the FPR to be zero. It can be defined as:

$$FPR = \frac{FP}{FP + TN} \quad (5)$$

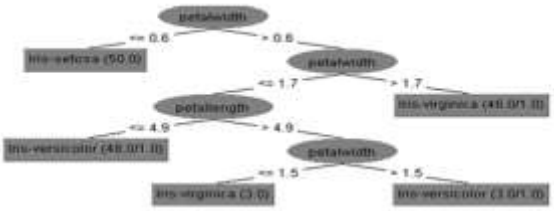
e) F-Meaure: F-measure is a way of combining recall and precision scores into a single measure of performance. The formula for it is:

$$\frac{2*recall*precision}{recall+precision} \quad (6)$$

IV. Experiment Results And Performance Evaluation

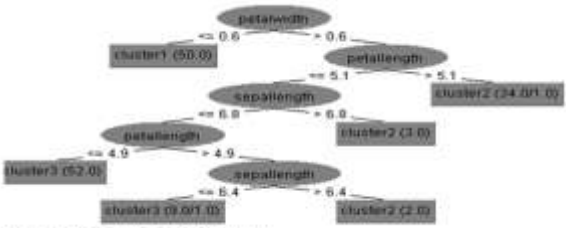
In this experiment we present a comparative study of classification technique of data mining with an integration of clustering and classification technique of data mining on various parameters using Fisher's Iris dataset containing 150 instances and 5 attributes. During simple classification, the training dataset is given as input to WEKA tool and the classification algorithm namely C4.5 (implemented in WEKA as J48) was implemented. During an integration of clustering and classification techniques of data mining first, Simple KMeans clustering algorithm was implemented on the training data set by removing the class attribute from the data set as clustering technique is unsupervised learning and then J48 classification algorithm was implemented on the resulting dataset.

The results of the experiment show that integration of clustering and classification technique gives a promising result with utmost accuracy rate and robustness among the classification and clustering algorithms (Table 3). An experiment measuring the accuracy of binary classifier based on true positives, false positives, false negatives, and true negatives (as per Equation 4), decision trees and decision tree rules are shown in Table 3 and Fig. 4 &5.



```
petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
|   petalwidth <= 1.7
|   |   petallength <= 4.9: Iris-versicolor (48.0/1.0)
|   |   petallength > 4.9
|   |   |   petalwidth <= 1.5: Iris-virginica (3.0)
|   |   |   petalwidth > 1.5: Iris-versicolor (3.0/1.0)
|   petalwidth > 1.7: Iris-virginica (46.0/1.0)
```

Figure 4. Decision tree and Rules during classification of Iris data



```
petalwidth <= 0.6: cluster1 (50.0)
petalwidth > 0.6
|   petallength <= 5.1
|   |   sepallength <= 6.8
|   |   |   petallength <= 4.9: cluster3 (52.0)
|   |   |   petallength > 4.9
|   |   |   |   sepallength <= 6.4: cluster3 (9.0/1.0)
|   |   |   |   sepallength > 6.4: cluster2 (2.0)
```

Figure 7. Decision tree and Rules of integration of clustering and classification technique

TABLE III. PERFORMANCE EVALUATION

Parameters	C4.5 (J48)	Simple KMeans+J48
TP	96	88
FP	4	1
TN	47	60
FN	3	1
Size of tree	9	11
Leaves in tree	5	6
Error Rate	0.1713	0.0909
Accuracy	95. 33%	98.6667%

A. *Observations and Analysis*

- It may be observed from Table 3 that the error rate of binary classifier J48 with Simple KMeans Clusterer is lowest i.e. 0.0909 in comparison with J48 classifier without clusterer i.e. 0.1713, which is most desirable.
- Accuracy of J48 classifier with KMeans clusterer is high i.e. 98.6667% (Table 3), which is highly required.
- Sensitivity (TPR) of clusters (results of integration of classification and clustering technique) is higher than that of classes (Table 4, 5 &6).
- In an ideal world we want the FPR to be zero. Considering results presented in Table 4, 5&6, FPR is lowest of integration of clustering and classification technique, in other words closet to the zero as compared with simple classification technique with J48 classifier.
- In an ideal world we want precision value to be 1. Precision value is the proportion of true positives out of all positive results. Precision value of integration of classification and clustering technique is higher than that of simple classification with J48 classifier (Table 4, 5&6).

TABLE IV. IRIS SETOSA CLASS AND CLUSTER 1

Parameters	J48 (iris-setosa)	SimpleKMeans+ J48 (cluster1)
Precision	1	1
Recall/Sensitivity (TPR)	0.98	1
Specificity (TNR)	0.9215	0.9836
F-measure	0.99	1
FPR	0	0

TABLE V. IRIS VERGINICA CLASS AND CLUSTER 2

Parameters	J48 (iris-setosa)	SimpleKMeans+ J48 (cluster1)
Precision	1	1
Recall/Sensitivity (TPR)	0.98	1
Specificity (TNR)	0.9215	0.9836
F-measure	0.99	1
FPR	0	0

TABLE VI. IRIS VERSICOLOR CLASS AND CLUSTER 3

Parameters	J48 (iris-versicolor)	SimpleKMeans+ J48 (cluster2)
Precision	0.922	0.987
Recall/Sensitivity (TPR)	0.94	0.987
F-measure	0.931	0.987
FPR	0.04	0.007

According to the experiments and result analysis presented in this paper, it is observed that an integration of classification and clustering technique is better to classify datasets with better accuracy.

V. CONCLUSION AND FUTURE WORK

A comparative study of data mining classification technique and an integration of clustering and classification technique helps in identifying large data sets. The presented experiments shows that integration of clustering and

classification technique gives more accurate results than simple classification technique to classify data sets whose attributes and classes are given to us. It can also be useful in developing rules when the data set is containing missing values. As clustering is an unsupervised learning technique therefore, it build the classes by forming a number of clusters to which instances belongs to, and then by applying classification technique to these clusters we get decision rules which are very useful in classifying unknown datasets. We can then assigns some class names to the clusters to which instance belongs to. This integrated technique of clustering and classification gives a promising classification results with utmost accuracy rate and robustness. In future we will perform experiments with other binary classifiers and try to find the results from the integration of classification, clustering and association technique of data mining.

REFERENCES

[1] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann, 2000.

[2] Desouza, K.C. (2001) ,Artificial intelligence for healthcare management In Proceedings of the First International Conference on Management of Healthcare and Medical Technology Enschede, Netherlands Institute for Healthcare Technology Management.

[3] http://www.comp.leeds.ac.uk/andyr

[4] I. K. Ravichandra Rao , "Data Mining and Clustering Techniques," DRTC Workshop on Semantic Web 8th – 10th December, 2003,DRTC, Bangalore ,Paper: K

[5] Rakesh Agrawal,Tomasz Imielinski and Arun Swami," Data mining : A Performance perspective ". IEEE Transactions on Knowledge and Data Engineering , 5(6):914-925, December 1993.

[6] Jain, A.K., Murty M.N., and Flynn P.J. (1999):" Data Clustering: A Review".

[7] Ritu Chauhan, Harleen Kaur, M.Afshar Alam, "Data Clustering Method for Discovering Clusters in Spatial Cancer Databases", International Journal of Computer Applications (0975 – 8887) Volume 10– No.6, November 2010.

[8] NASA Ames Res.Ctr. INTRO. IND Version 2.1, GA23-2475-02 edition, 1992

[9] J.R Quinlan and R..L. Rivest. Inferring decision tress using minium description length principle. Information and computation ,1989.

[10] J.R Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufman, 1993.

[11] D.Michie , D.J Spiegehalter, and C.C Taylor. Machine Learning, Neural and Statistical Classification. Ellis horword, 1994.

[12] Paul Agarwal, M.Afsar Alam, Ranjit Biswas. Analysing the agglomeratve hierarchical Clustering Algorithm for Categorical Attributes. International Journal of Innovation, Management and Technology , vol.1,No.2 , June 2010, ISSN: 2010- 0248.

[13] Weka 3- Data Mining with open source machine learning software available from :- http://www.cs.waikato. ac.nz/ml/ weka/

[14] U.M Fayyad and P. Smyth. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Presss, Menlo Park, CA, 1996.

[15] L.Kaufinan, and P.J Rousseeuw, Finding groups in Data: an Introduction to Cluster Analysis, John Wiley & Sons 1990.

[16] M.S Chen, J.Han, and P.S.Yu. Data mining: an overview from database perspective . IEEE Trans. On Knowledge and Data Engineering , 5(1) : 866-833, Dec 1996.

[17] Shuyan wang , Mingquan Zhou and guohua geng ," Application of Fuzzy cluster analysis for Medical Image Data Mining " proceedings of the IEEE International Conference on Mechatronics & Automation Niagra falls ,Canada, July 2005.

[18] Holmes, G.,Donkin,A.,Witten,I.H.: "WEKA a machine learning workbench". In: Proceeding second Australia and New Zeeland Conference on Intelligent Information System, Brisbane , Australia, pp.357-361 (1994).

[19] Kim, H. and Loh , W. –Y.2001, Classification trees with unbiased multiway splits, Journal of the American Stastical Association, vol. 96, pp. 589- 604.

[20] Garner S.R. (1995) "WEKA: The Waikato Environment for Knowledge Analysis Proc" New Zealand Computer Science Research Students Conference, University of Waikato, Hamilton, New Zealand, pp 57-64.

[21] The Result Oriented Process for Students Based On Distributed Data Mining. (2010). International Journal of Advanced Computer Science and Applications - IJACSA, 1(5), 22-25.

[22] Saritha, S. J., Govindarajulu, P. P., Prasad, K. R., Rao, S. C. V. R., Lakshmi, C., Prof, A., et al. (2010). Clustering Methods for Credit Card using Bayesian rules based on K-means classification. International Journal of Advanced Computer Science and Applications - IJACSA, 1(4), 2-5.

[23] Saritha, S. J., Govindarajulu, P. P., Prasad, K. R., Rao, S. C. V. R., Lakshmi, C., Prof, A., et al. (2010). Clustering Methods for Credit Card using Bayesian rules based on K-means classification. International Journal of Advanced Computer Science and Applications - IJACSA, 1(4), 2-5.

[24] Firdhous, M. F. M. (2010). Automating Legal Research through Data Mining. International Journal of Advanced Computer Science and Applications - IJACSA, 1(6), 9-16.

[25] Takale, S. A. (2010). Measuring Semantic Similarity between Words Using Web Documents. International Journal of Advanced Computer Science and Applications - IJACSA, 1(4), 78-85.

AUTHORS PROFILE

Dr. Varun Kumar, Professor & Head, Department of CSE, School of Engg. & Tech., ITM University, Gurgaon, completed his PhD in Computer Science. He received his M. Phil. in Computer Science and M. Tech. in Information Technology. He has 13 years of teaching experience. He is recipient of Gold Medal at his Master's degree. His area of interest includes Data Warehousing, Data Mining, and Object Oriented Languages like C++, JAVA, C# etc. He is an approved Supervisor for Ph.D., M. Phil., and M. Tech. Programme of various universities and currently he is guiding Ph.D. & M. Phil. scholars and M. Tech. students for their major project work in his area of research interest. He has published more than 35 research papers in Journals/Conferences/Seminars at international/national levels. He is working as an Editorial Board Member / Reviewer of various International Journals and Conferences.

Ms. Nisha Rathee is perusing her MTech in Computer Science and Engineering from ITM University Gurgaon, Haryana, India. Her area of interest includes Data Warehousing and Data Mining.