Volume 3, Issue 4, April 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Discovering Knowledge Patterns from Integration of Clustering and Classification Techniques

Raman Pathrey*
Department of Information Technology
Maharaja Surajmal Institute of Technology
New Delhi, India

Yogesh Kumar
Department of Information Technology
Maharaja Surajmal Institute of Technology
New Delhi, India

Nitin
Department of Information Technology
Maharaja Surajmal Institute of Technology
New Delhi, India

Nisha Rathee
Asst. Professor, Department of CSE
Maharaja Surajmal Institute of Technology
New Delhi, India

Abstract— Data mining is essentially the discovery of valuable information and patterns from huge chunks of available data. Two indispensible techniques of data mining are clustering and classification, where the latter employs a set of pre-classified examples to develop a model that can classify the population of records at large, and the former divides the data into groups of similar objects. In this paper we have proposed a new method for data classification by integrating two data mining techniques, viz. clustering and classification. Then a comparative study has been carried out between the simple classification and new proposed integrated clustering-classification technique. Four popular data mining tools were used for both the techniques by using six different classifiers and one clusterer for all sets. It was found that across all the tools used, the integrated clustering-classification technique was better than the simple classification technique. This result was consistent for all the six classifiers used. For both of the techniques, the best classifier was found to be SVM. Out of the four tools used, WEKA was found to be the best in terms of flexibility of algorithm. All comparisons were drawn by comparing the percentage accuracy of each classifier used.

Keywords— data mining, classification, integrated clustering-classification, data mining tools, pima Indians Diabetes dataset, hybrid.

I. INTRODUCTION

In this day and age of data overload it has become increasingly important to be able to sort "meaningful" information from a huge chunk of data available in countless databases and then be able to make decisions using the results obtained. The task is not so simple as to be achieved manually, which is why data mining and its applications are useful. Data mining centres on the automated discovery of new facts and relationships in already existing data [1]. Data mining, in this regard, has not only given us means to shift through piles of data quickly, but also a variety of methods to do so. The various techniques of data mining include association, regression, prediction, clustering and classification [3]. Clustering is the division of data into groups of similar objects. Clustering is an example of unsupervised learning as it learns by observation rather than example [7]. Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data [8]. This paper deals with the use of the integrated clustering-classification technique on some of the free data mining tools available these days. Tools on which integrated clustering-classification technique has been implemented are KNIME (Konstanz Information Miner) [19], Tanagra [11], orange [20] and WEKA (Waikato Environment for Knowledge Learning) [12]. The various classifier used for this purpose are Naïve Bayes, Support Vector machine, K Nearest Neighbor, Zero Rule, Decision tree and One Rule. The paper is classified as follows: Section 2 describes the problem statement; Section 3 is a proposed method of how the integrated clustering-classification technique works; Section 4 is a description of the methodology followed; Section 5 gives detailed results of the experiment and the comparative results of the tools used, and finally Section 6 gives the conclusion and future work.

.

II. PROBLEM STATEMENT

The problem in particular is the comparison between the simple classification and the new proposed hybrid technique (Integration of clustering and classification). This has been done using the large dataset "pima Indians Diabetes" and the two techniques have been applied using four data mining tools viz. WEKA, Tanagra, Orange and KNIME.

III. PROPOSED METHOD

Classification is the process which finds the common properties among a set of objects in a database and classifies them into different classes, according to a classification model. Clustering is the task of segmenting a diverse group into a number of similar subgroups or clusters. Figure I gives an insight into the working of the integration of the above two techniques. In this proposed technique first the clustering algorithm is applied on the dataset with the help of any clustering algorithm such as Simple K-Mean. Clustering algorithm adds the attribute 'cluster' on the dataset. After that, classification algorithm is applied on this clustered dataset. This approach gives results with a better accuracy and in a shorter time than the simple classification technique.

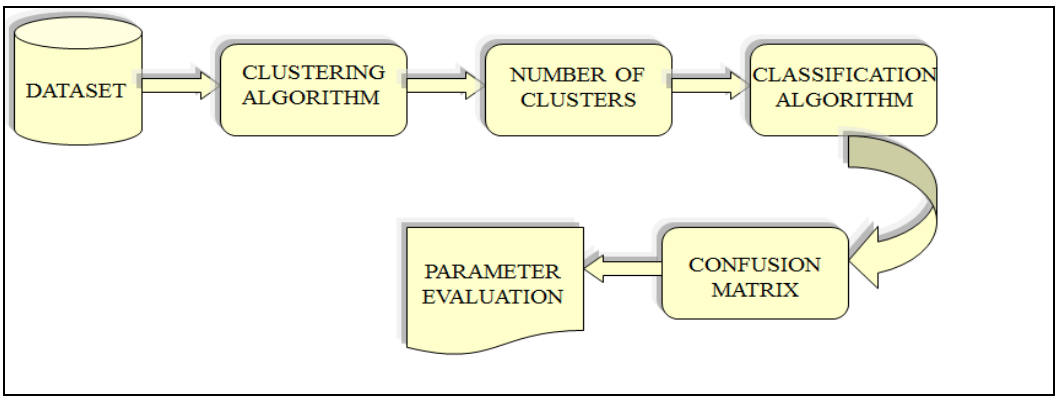


Fig. 1 Proposed Integrated Clustering-Classification Technique

IV. METHODOLOGY OF WORK

The comparative study includes the use of the dataset "pima Indian Diabetes" and use of the integrated clustering-classification technique as well as the simple classification technique, making use of the different classification algorithms available on different data mining tools. The pima Indian diabetes dataset is available on UCI machine learning repository [21] website http://archive.ics.uci.edu/ml/datasets/pima+Indians+Diabetes.

I. Data mining tools description:

The data mining tools on which the integrated clustering-classification technique has been implemented are:

1. *WEKA:* WEKA, formally called Waikato Environment for Knowledge Learning, is a computer program that was developed at the University of Waikato in New Zealand for the purpose of identifying information from raw data gathered from agricultural domains. WEKA supports many different standard data mining tasks such as data preprocessing, classification, clustering, regression, visualization and feature selection.
2. *Tanagra:* Tanagra was written as an aid to education and research on data mining by Ricco Rakotomalala. The entire user operation of Tanagra is based on the stream diagram paradigm. Under the stream diagram paradigm, a user builds a graph specifying the data sources, and operations on the data. Paths through the graph can describe the flow of data through manipulations and analyses. Tanagra simplifies this paradigm by restricting the graph to be a tree. This means that there can only be one parent to each node, and therefore only one data source for each operation.
3. *Orange:* It is a data mining tool based on C++ and Python. Analysis of data is made simple through the visual-programing front-end; thus the user can actually visualise the data flow. Its components-based features allow for set components for data pre-processing, feature scoring and filtering, modelling and evaluation, as well as explorative procedures.

4. *KNIME:* Short for Konstanz Information Miner, KNIME is a freely available data analytics, reporting and integration software. Based on Java, it is one of the most widely used software which provides easy data loading, processing, transformation and analysis. Data flows can be viewed through models and interactive display of results.

II. Classification Algorithms:

In this paper six different classification algorithms have been used, which have been listed below:

- *Naïve Bayes (NB)*: An independent feature probability model, it is based on the Bayes theorem and is thus a probabilistic classifier.
- *One Rule (OneR)*: This classifier produces one rule in each predictor of the dataset, and then selects the rule with the smallest total error as the *one rule*.
- *Decision tree (C4.5)*: This is a statistical classifier developed by Ross Quinlan, and classifies data by generating decision trees.
- *Support Vector Machine (SVM)*: It is an example of non-probabilistic binary linear classifier and from the set of input data predicts which of the two possible classes forms the output.
- *K Nearest Neighbor (KNN)*: An example of instance-based learning, KNN is sensitive to the local structure of the data; thus the function is approximated locally and computation is done after classification is complete.
- *Zero Rule (ZeroR)*: The classifier simply predicts the majority class, ignoring all predictors. It is the simplest of all classifiers.

III. Measure for performance evaluation:

Comparisons were drawn between the techniques primarily on the basis of accuracy. Accuracy is defined as the ratio of the number of correctly defined instances to the total number of instances, i.e.

Accuracy = (number of correctly defined instances)/(total number of instances)

IV. Experimental procedure:

Initially the dataset is loaded in different data mining tools. For simple classification, the above mentioned classifiers were simply applied one by one on the dataset. The results of classification were produced on the basis of accuracy.

For the integrated clustering-classification technique, the K-means clusterer was applied on the dataset. On the obtained clusters, different classifiers were applied. To determine which of the six classifiers used was the most suitable to be integrated with K-Means clusterer the 10-Fold Cross Validation (10-Fold CV) test mode was carried out. The results obtained from this technique were compared with ones obtained from simple classification technique on the basis of % accuracy.

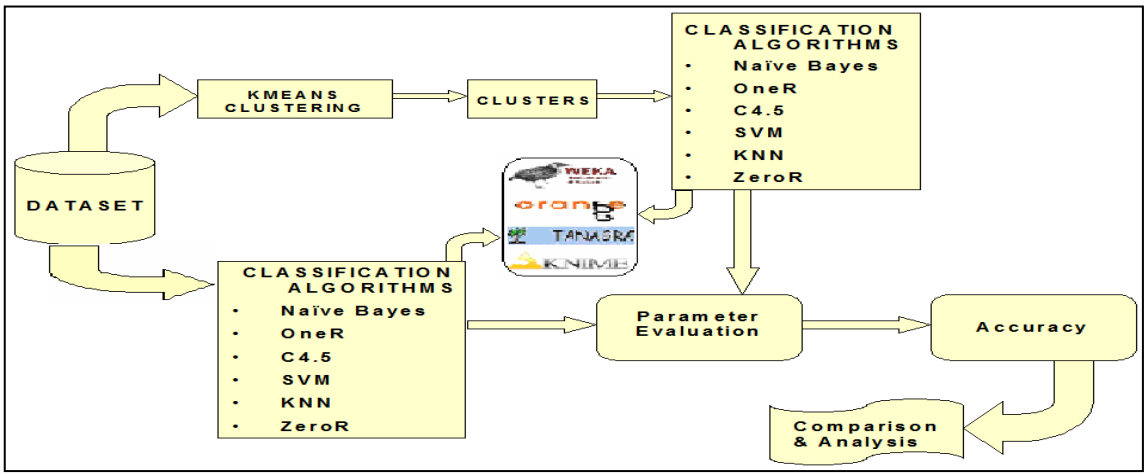


Fig. 2 Methodology of Work

V. EXPERIMENTAL RESULTS AND PERFORMANCE EVALUATION

A: Result Evaluation: The experiments performed on the dataset gave the results as listed below:

Table I shows the accuracy measure of the simple classification technique for different classifiers used. It was found that in all the tools, SVM algorithm gave results with the highest accuracy in the range of 76-77%, followed by Naïve Bayes with accuracy in the range of 73-76%. KNN comes third

with accuracy ranging between 72-73%, followed closely by C4.5 with accuracy in the range 69-74%. OneR has about 73% accuracy, which is larger than the accuracy of ZeroR which is approximately 65%.

TABLE I
ACCURACY FOR SIMPLE CLASSIFICATION

Classifier	Weka	Tanagra	Orange	KNIME
NB	**76.32 %**	**74.87%**	**75.38%**	**73.17%**
OneR	**72.78 %**	**N/A**	**N/A**	**N/A**
C4.5	**73.82 %**	**74.21%**	**70.05%**	**69.27%**
SVM	**77.34 %**	**76.45%**	**76.17%**	**77.60%**
KNN	**73.17 %**	**72.11%**	**72.90%**	**72.26%**
ZeroR	**65.14 %**	**N/A**	**65.11%**	**N/A**

***N/A: Algorithm not implemented.**

From Table II, it can be seen that the best classifier for the integrated clustering-classification technique is SVM, with an accuracy measure between 98-100%. This is followed by Naïve Bayes, with accuracy between 94-97%. Third and fourth are KNN and C 4.5, respectively, where the former has accuracy between 93-99% and for the latter it is between 89-99%. OneR follows these with an accuracy measure of approximately 90%, and the last of all is ZeroR, having accuracy between 67-68%.

TABLE II
ACCURACY FOR INTEGRATED CLUSTERING-CLASSIFICATION

Classifier	Weka	Tanagra	Orange	KNIME
NB	96.22 %	94.34%	96.74%	94.01%
OneR	89.97 %	N/A	N/A	N/A
C4.5	95.31 %	89.08%	96.74%	99.47%
SVM	98.57 %	97.24%	99.22%	99.87
KNN	94.27 %	93.95%	95.83%	99.21%
ZeroR	67.0 %	N/A	68.36%	N/A

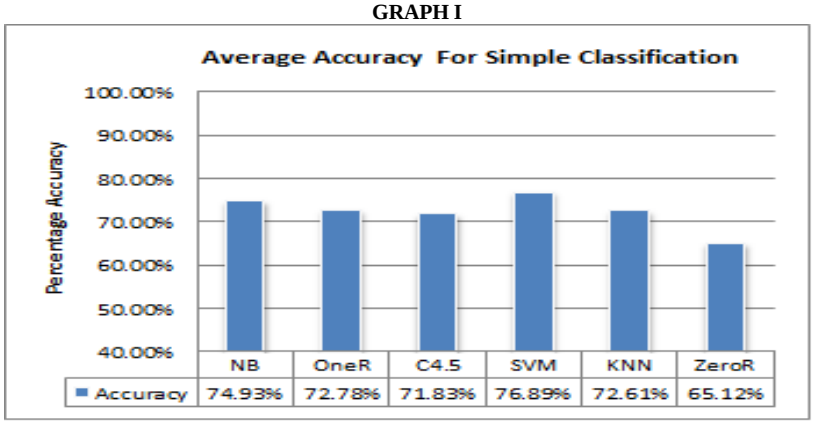
***N/A: Algorithm not implemented**

Comparing the data in Table 1 and 2, the SVM classifier is the best for both the simple classification and the integrated clustering-classification techniques. However, the percentage accuracy for the latter using SVM classifier is in the range of 97-99%, and that for the former is in the range of 76-77%.

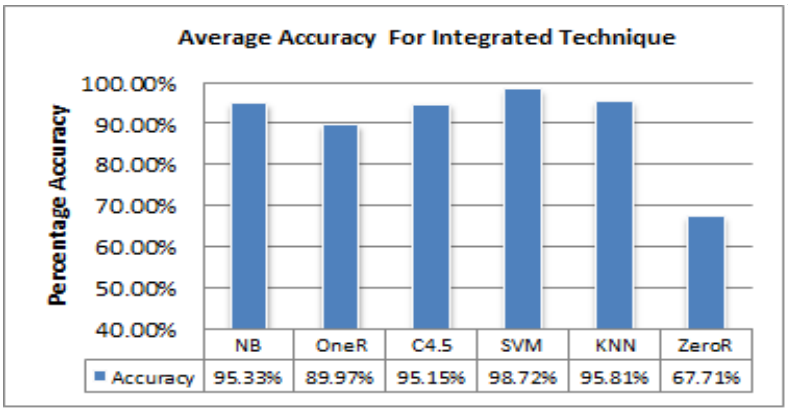
From the comparison of both the tables, it can be said that the results of the proposed integrated technique are more accurate than the simple classification data mining technique. Overall, the accuracy of the integrated technique is about 2-22% greater than the simple technique, over a range of tools and algorithms used.

The following graphs better illustrate the tabulated results shown above:

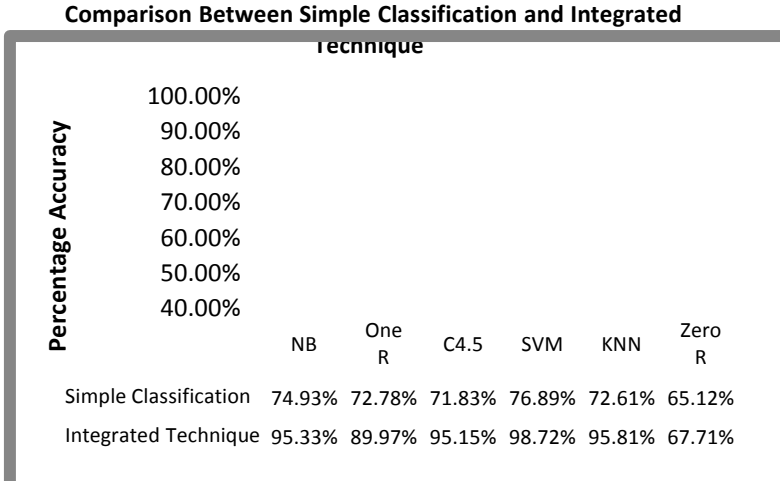
GRAPH I



GRAPH II



GRAPH III



*B: **Observation and Analysis:***

The results obtained from the experiments mentioned in Section 4 were used to carry out the following comparative evaluation:-

- It was found that out of the two data mining techniques used, the proposed integration of clustering and classification techniques gave results which were more accurate than the simple classification technique. This observation was found to be consistent over all the tools and algorithms used.
- For both of the techniques, the best classifier was found to be SVM, and it achieved significantly higher accuracy than the other classifiers used.

The comparison of the four tools, viz. WEKA, Tanagra, Orange and KNIME, lead to the conclusion that WEKA was best of all with respect to the applicability of the tool, which is defined as the ability to run a specific algorithm on a desired tool. As such, the accuracy of none of the tools can be said to be the best, as the results of all differ when different classifiers were used.

VI. Conclusions And Future Work

According to the experiments and the analysis of results presented in this paper, we conclude that the integration of the two pre-existing techniques, viz. clustering and classification, gives invariably better results over a range of different tools and algorithms. Large datasets can thus be organized and classified with an improved accuracy. The algorithm which gives the best results in terms of accuracy for both the techniques was found to be SVM. This observation was consistent over all the four data mining tools used. This integrated technique is advantageous over the simple techniques, as it gives the user numerous possibilities of generating better classified data, by changing the clusterer and classifier as desired. Thus, this technique allows flexibility of use of the algorithm. For future research we plan to use this integrated technique with other clusterers and classifiers.

References

[1] David J. Hand, Heikki Mannila and Padhraic Smyth, Principles of Data Mining, Prentice Hall of India, 2001

[2] Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, 2006.
[3] Paulraj Ponniah, Data warehouse fundamentals, 2001.
[4] Bing Liu, Wynne Hsu and Yiming Ma, Integrating Classification and Association Rule Mining, 1998.
[5] Ming-Syan Chen, Jiawei Han, and Philip S. Yu, Data mining: An overview from database perspective, 1996.
[6] Varun Kumar and Nisha Rathee, Knowledge discovery from database using an integration of clustering and classification, 2011.
[7] A.K. Jain, M.N. Murty and P.J. Flynn, Data Clustering: A Review, 1999.
[8] J. R. Quinlan, Induction of decision trees," Machine Learning, Vol. 1, No. 1, 1996.
[9] J. R. Quinlan, C4.5: Programs for machine learning, Morgan Kaufmann, San Mateo, CA, 1993.
[10] Tanagra – Data mining tutorials, available from http://data-mining-tutorials.blogspot.in/.
[11] Tanagra 1.4 –Data mining software available from:- http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html.
[12] Weka 3.6- Data Mining with open source machine learning software available from: - http://www.cs.waikato. ac.nz/ml/ weka.
[13] Rakesh Agrawal, Tomasz Imielinski and Arun Swami, Data mining : A Performance perspective ,1993.
[14] G. Piatetsky-Shapiro, U. Fayyad, and P. Smith, "From Data Mining to Knowledge Discovery: An Overview," U.M. Fayyad,G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, eds., Advances in Knowledge Discovery and Data Mining, 1996.
[15] G. Piatetsky-Shapiro and W.J. Frawley, Knowledge Discovery in Databases, 1991.
[16] J. R. Quinlan, Decision trees and decision making, 1990.
[17] L.Kaufinan, and P.J Rousseeuw, Finding groups in Data: an Introduction to Cluster Analysis, John Wiley & Sons 1990.
[18] Jessica Enright and Jonathan Klippenstein, Tanagra: An Evaluation, 2004.
[19] KNIME(Konstanz Information Miner)- Available at: http://www.knime.org.
[20] Orange–Data Mining Fruitful and Fun, Available at: http:// http://orange.biolab.si.
[21] UCI Machine Learning Repository, Available at: http:// http://archive.ics.uci.edu/ml
[22] Flach, P., A., Lachiche, N., Naive Bayesian Classification of Structured Data, Machine Learning, v.57 n.3, p.233-269, December 2004.
[23] Li, Y., Bontcheva, K., dapting Support Vector Machines for F-term-based Classification of Patents, Journal ACM Transactions on Asian Language Information Processing, Volume 7 Issue 2, June 2008.
[24] Goebel, M., Gruenwald, L., A survey of data mining and knowledge discovery software tools, ACM SIGKDD Explorations Newsletter, v.1 n.1, p.20-33, June 1999.S. M. Metev and V. P. Veiko, *Laser Assisted Microtechnology*, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.

