SCITECH
RESEARCH ORGANISATION|

Volume 6, Issue 1
November 28, 2016

Journal of Information Sciences and Computing Technologies
www.scitecresearch.com/journals

A Performance Assessment on Various Data mining Tool Using Support Vector Machine

Mr. G.Karthikeyan[1], Mrs. K.Saroja[2], Dr. S.Prasath [3]
[1] Asst. Prof. of BCA, Nandha Arts & Science College, Erode, TN. E-mail:gykarthikeyan@gmail.com
[2] Asst. Prof. of CS, Nandha Arts & Science College, Erode, TN. E-mail:sarojak1983@gmail.com
[3] Asst. Prof. of CS, Nandha Arts & Science College, Erode, TN. E-mail:softprasaths@gmail.com

Abstract

Data mining is essentially the discovery of valuable information and patterns from huge chunks of available data. Two indispensable 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 clustered 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. From the four tools used, KNIME 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: SVM; WEKA; KDD; DM; KNIME; KNN.

1. Introduction

Data mining centers on the automated discovery of new facts and relationships in already existing data. 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, Tanagra, orange and WEKA (Waikato Environment for Knowledge Learning). The various classifier used for this purpose are Naïve Bayes, Support Vector machine, K Nearest Neighbour, Decision tree.

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. Data mining also known as KDD. Data preprocessing, pattern recognition, clustering, classification are the popular technologies in data mining.

2. Knowledge Discovery

The terms Knowledge Discovery in Databases (KDD) and Data Mining are often used interchangeably. KDD is the process of turning the low-level data into high-level knowledge. Hence, KDD refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and KDD are often treated as equivalent words but in real data mining is an important step in the KDD process.

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge [2]. The iterative process consists of the following steps:

Data cleaning: also known as data cleansing it is a phase in which noise data and irrelevant data are removed from the collection.

Data integration: at this stage, multiple data sources, often heterogeneous, may be combined in a common source.

Data selection: at this step, the data relevant to the analysis is decided on and retrieved from the data collection.

Data transformation: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.

Data mining: it is the crucial step in which clever techniques are applied to extract patterns potentially useful.

Pattern evaluation: this step, strictly interesting patterns representing knowledge are identified based on given measures.

Knowledge representation: is the final phase in which the discovered knowledge is visually represented to the user. In this step visualization techniques are used to help users understand and interpret the data mining results.

3. Data Mining Methods

Classification: Supervised Learning. The classes are known

Clustering: Unsupervised Learning. The classes are unknown

Association Rule Mining: Identifying the hidden, previously unknown relation between the entities.

Temporal mining: Use with temporal data, modeling temporal events, time series, pattern detection, sequences and temporal association rules are some tasks.

Time Series Analysis: Describe the trend, nature and behavior of time series data. Predict the future trend and behavior of the data.

Web Mining: Mining web data, Web content mining, Web structure mining and Web usage mining.

Spatial Mining: Use with GIS for mining knowledge from spatial database. Spatial classification and clustering and rule generation are some task under this mining [4] [5].

4. Data Mining Classification

The different classification algorithms available are

Naïve Bayes (NB): An independent feature probability model, it is based on the Bayes theorem and is thus a probabilistic classifier.

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 Neighbour (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.

5 Clustering

This pattern divides the records in database into different groups. In the same group, the groups have the similar properties. Between groups the differences should be as bigger as possible and in the same group, the differences should be as smaller as possible. There is no predefined class that's why it comes under the unsupervised learning. some examples of cluster applications are seen as in marketing, land use, insurance, earthquake studies and in city planning Methods [5] involve in cluster analysis are portioning methods, hierarchical Methods, density-Based Methods, grid-Based Methods, model-Based Methods, clustering high-dimensional data, constraint-based clustering [8] and Outlier analysis. Many algorithms exist for clustering. The three major clustering methods and their approach for clustering.

K-means Clustering

The term "k-means" was first used by James Mac Queen in 1967. The standard algorithm was first proposed by Stuart Lloyd in 1957 as a technique for pulse-code modulation, though it wasn't published until 1982. K-means is a widely used partitioned clustering method in the industries. The K-means algorithm is the most commonly used partitioned clustering algorithm because it can be easily implemented and is the most efficient one in terms of the execution time.

Hierarchical Clustering

Hierarchical clustering builds a cluster hierarchy or in other words a tree of clusters also known as a dendrogram.

Density Based Clustering

Density-based clustering algorithms try to find clusters based on density of data points in a region. The key idea of density-based clustering is that for each instance of a cluster the neighborhood of a given radius (Eps) has to contain at least a minimum number of instances (Min Pts). One of the most well-known density-based clustering algorithms is the DBSCAN.

6. Data Mining Tools

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

WEKA

WEKA is Waikato Environment for Knowledge Analysis, data mining/machine learning tool developed by Department of Computer Science, University of Waikato, New Zealand [6]. It is a collection of open source of many data mining and machine learning algorithms, including pre-processing on data, Classification and regression, clustering, association rule extraction, feature selection. It supports .arff (attribute relation file format) file format.

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.

KNIME

KNIME, the Konstanz Information Miner, is an open source data analytics, reporting and integration platform. KNIME integrates various components for machine learning and data mining through its modular data pipelining concept. A graphical user interface allows assembly of nodes for data preprocessing (ETL: Extraction, Transformation, Loading), for modeling and data analysis and visualization.

Orange

Orange is a component-based data mining and machine learning software suite, featuring a visual programming front-end for explorative data analysis and visualization, and Python bindings and libraries for scripting. It includes a set of components for data preprocessing, feature scoring and filtering, modeling, model evaluation and exploration techniques. It is implemented in C++ and Python.

7. Experimental Results

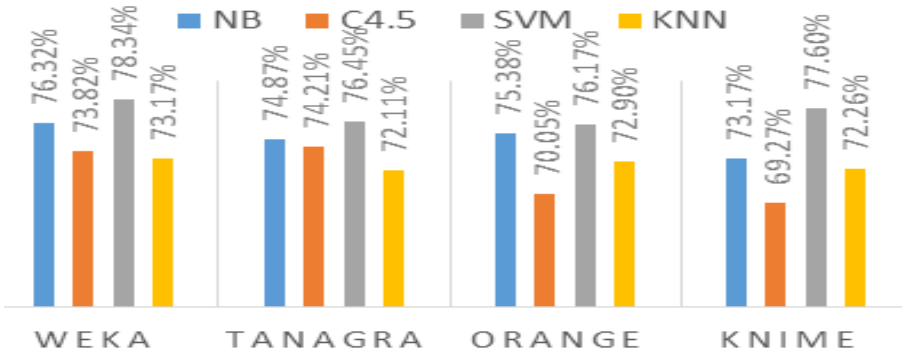
The comparative study includes the use of the dataset "Pima Indian Diabetes" and use of the Kmeans, Hierarchical, Density based clustering technique, making use of the different classification algorithms available on different data mining tools. The pima Indian diabetes dataset is available on UCI machine learning repository website http://archive.ics.uci.edu/ml/datasets/pima+Indians+Diabetes.

Table I shows the accuracy measure of the Kmeans clustering 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 76-78%, 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%.

Table I : Accuracy for K-means clustering

Classifier	Weka	Tanagra	Orange	KNIME
NB	76.32 %	74.87%	75.38%	73.17%
C4.5	73.82 %	74.21%	70.05%	69.27%
SVM	78.34 %	76.45%	76.17%	77.60%
KNN	73.17 %	72.11%	72.90%	72.26%

Figure: 1 Accuracy for KMeans Clustering

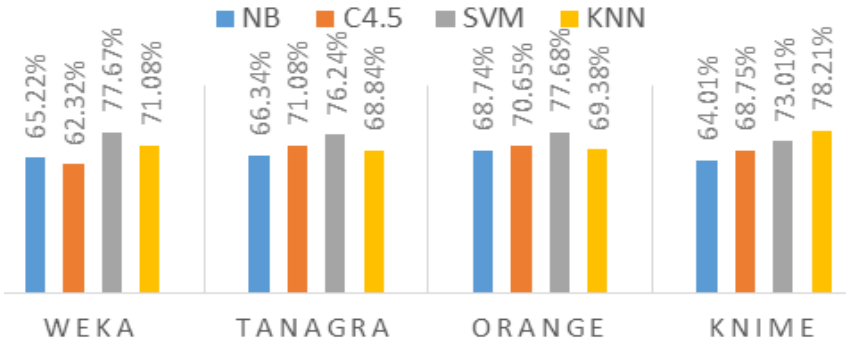


From Table II, it can be seen that the accuracy measure of the Hierarchical Clustering technique for different classifiers used. The SVM classifier gives the accuracy measure between 73-77%. This is followed by Naïve Bayes, with accuracy between 64-68%. Third and fourth are KNN and C 4.5, respectively, where the former has accuracy between 62-71% and for the latter it is between 89-99%.

Table II: Accuracy for Hierarchical Clustering

Classifier	Weka	Tanagra	Orange	KNIME
NB	65.22 %	66.34%	68.74%	64.01%
C4.5	62.32 %	71.08%	70.65%	68.75%
SVM	77.67 %	76.24%	77.68%	73.01%
KNN	71.08 %	68.84%	69.38%	78.21%

Figure 2: Accuracy for Hierarchical Clustering

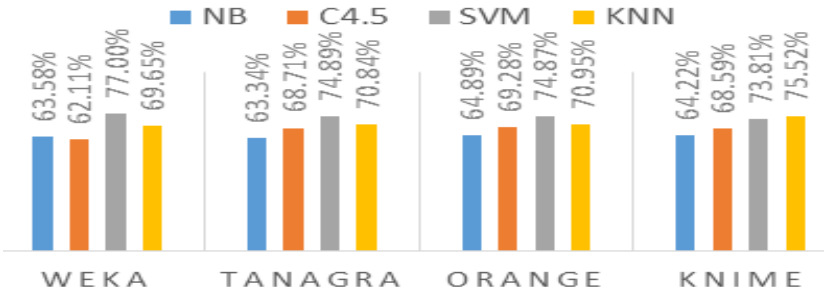


From Table III, it can be seen that the accuracy measure of the Hierarchical Clustering technique for different classifiers used. The SVM classifier gives the accuracy measure between 73-77%. This is followed by Naïve Bayes, with accuracy between 64-68%. Third and fourth are KNN and C 4.5, respectively, where the former has accuracy between 62-71% and for the latter it is between 89-99%.

Table III : Accuracy for Density based clustering

Classifier	Weka	Tanagra	Orange	KNIME
NB	63.58 %	63.34%	64.89%	64.22%
C4.5	62.11 %	68.71%	69.28%	68.59%
SVM	77.00 %	74.89%	74.87%	73.81%
KNN	69.65 %	70.84%	70.95%	75.52%

Figure III: Accuracy for Density Based Clustering



Comparing the data in Table 1 and 2, the SVM classifier is the best for the K-means, Hierarchical and Density based clustering clustering-classification techniques. However, the percentage accuracy for the latter using SVM classifier is in the range of 76-78%, and that for the former is in the range of 76-77%. From the comparison of the three tables, it can be said that the results of the accuracy of K-means clustering technique is more accurate than the other classification data mining technique. Overall, the K-means clustering technique is about 2-12% greater than the other Clustering technique, over a range of tools and algorithms used.

Conclusion

Data mining is the extraction of useful patterns and relationships from data sources, such as databases, texts, the web etc. This research has conducted a comparison between different data mining tool focuses the usefulness and importance of these tools by considering various aspects. Analysis presents various benefits of these data mining tools with respect to functionalities, advantages and disadvantages when compared them accordingly. The experimental results shows that compared with existing techniques such as clustering and classification gives better results to improve the accuracy of algorithm shows in table mentioned above section 5 gives SVM is the best compare to other method.

REFERENCES

[1] David Heckerman. Bayesian Network for Data Mining. Data Mining and Knowledge Discovery, 1997:79-119..

[2] David Hand, Heikki Mannila and Padhraic Smyth. Principles of Data Mining, the MIT Press, 2001:1-5...

[3] A Short Introduction to Data Mining and Its Applications Zhang Haiyang

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

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

[6] S.Kotsiantis, D.Kanellopoulos, P.Pintelas, "Data Preprocessing for Supervised Leaning", International Journal of Computer Science, 2006, Vol 1 N. 2, pp 111–117.

[7] MacQueen.J.B., "Some Methods for classification and Analysis of Multivariate Observations",Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability.University of California Press. 1967, pp. 281–297.

[8] Lloyd S.P."Least square quantization in PCM". IEEE Transactions on Information Theory 28,1982,pp.129-137.

[9] Manish Verma, MaulySrivastava, NehaChack, Atul Kumar Diswar and Nidhi Gupta, ―A Comparative Study of Various Clustering Algorithms in Data Mining, International Journal of Engineering Research and Applications (IJERA) Vol. 2, Issue 3, May-Jun 2012, pp.1379-1384.

Authors' Biography

Mr.G.KARTHIKEYAN is currently working as Assistant Professor at Department of Computer Applications, Nandha Arts & Science College, Erode, Tamil Nadu, India. He has obtained his Master Degree in Computer Applications from Sengunthar Arts and Science College, Tiruchengode and M.Phil Degree in Computer Science, Erode Arts & Science College (Autonomous), Erode Affiliated to Bharathiar University, Coimbatore. His area of interests includes, Image Processing and Data Mining. He has presented 5 papers in National and 1 International level conference.

Mrs.K.SAROJA is currently working as Assistant Professor at Department of Computer Science, Nandha Arts & Science College, Erode, Tamil Nadu, India. She has obtained her Master Degree in Computer Science from JKKN College of Arts and Science, Komarapalayam and M.Phil degree in Computer Science from Mother Teresa Women's University, Kodaikanal Her area of interests includes, Networking and Data Mining. She has presented 5 papers in National level conferences. She has published 4 papers in International journals.

Dr.S.PRASATH is currently working as Assistant Professor at Department of Computer Science, Nandha Arts & Science College, Erode, Tamil Nadu, India. He has obtained his Master Degree in Software Engineering from M.Kumarasamy College of Engineering, Karur under Anna University, Chennai and M.Phil degree in Computer Science and he obtained Doctorate in Computer Science from Erode Arts & Science College (Autonomous), Erode under Bharathiar University, Coimbatore. His area of interests includes, Networking, Image Processing and Data Mining. He has presented 8 papers in National and 3 International level conferences. He has published 20 papers in International journals.