A Comparative Analysis of the Different Data Mining Tools by Using Supervised Learning Algorithms

Akarsh Goyal(✉), Ishan Khandelwal, Rahul Anand,
Anan Srivastava, and P. Swarnalatha

School of Computer Science and Engineering, VIT University, Vellore 632014, Tamil Nadu, India
akarsh.goyal15@gmail.com, ishankhandelwal23@gmail.com,
rahul.anand.2301@gmail.com, srivastava.anan@gmail.com,
pswarnalatha@vit.ac.in

Abstract. These days a lot of raw data is generated from various common sources. This large amount of data, which would appear useless at first glance, is very important for companies and researchers as could provide a lot of helpful information. The data could be mined to get useful knowledge that could be used to make fruitful decisions. A lot of online tools and proprietary toolkits are available to the users and it becomes all the more cumbersome for them to know which is the best tool among these for the supervised learning algorithm and datasets they are applying. In order to aid this process, the paper progresses in this direction by doing a comparison of various data mining tools on the basis of their classification finesse. The various tools used in the paper are weka, knime and tanagra. Rigorous work on this has given the result that the performance of the tools is affected by the kind of datasets used and the way in which the supervised learning is done.

Keywords: Weka · Tanagra · Knime · Mining tools · Supervised learning

1 Introduction

The amount of data available today is huge. This data often goes beyond the basic understanding and comprehending power of most humans. The situation makes it all the more necessary to come with tools that could automate and work on these datasets which could help in the derivation of some deep lying information engraved in them. Due to this notion, the advent of data mining took place.

The process of deriving important knowledge in order to culminate in good decision making is known as data mining. Data mining [1] lies at the intersection of various fields like machine learning, signal processing, graphics, artificial intelligence etc. Data mining has its usage in many areas like gaming, engineering, mathematics, biological sciences, analytics, and visualization. Currently, a lot of data mining tools are there in the market like Weka, Orange, Rapid Miner, Keel etc. The softwares specified provide a plethora of features and functionalities to the user to make an analysis on the data ranging from doing a regression analysis to performing fuzzy clustering.

A. Abraham et al. (eds.), *Proceedings of the Eighth International Conference on Soft Computing and Pattern Recognition (SoCPaR 2016)*, Advances in Intelligent Systems and Computing 614, DOI 10.1007/978-3-319-60618-7_11

The research presented in this paper has used some of the most common datasets and applied a host of classification techniques on them. Then the accuracy measure has been expounded by using validation techniques in order to reach to the main result.

The various sections in the paper are: Sect. 2 gives a brief overview of the literature done in this field, Sect. 3 describes the various tools used and the methodology followed in the paper and Sect. 4 elicits the experimental analysis and the accuracy measure obtained from the application of classification techniques [7, 11, 12]. At last, we have summarized our findings and displayed the result in the form of a clustered graph.

2 Literature Study

A lot of work has been done regarding the analysis on data mining tools and the comparison of the classification algorithms. We elucidate here a few of these.

In [1] a review on data mining and a survey has been done. In 2004, a study on k-fold cross validation is elicited in [2]. bin Othman et al. [3] did a research on using weka to perform different types of classification on breast cancer dataset. [4, 13] have performed a software type analysis on the different data mining tools present without proceeding in much depth. Whereas in [9, 10] an analysis of the data mining tools on the functionalities they offer to the user like classification, association and clustering has been given. In 2013 [5] did research on using weka and explained its advantages as compared to other tools when dealing with medical datasets. [6] is a primarily focused survey of weka tool. Different classification techniques have been comparatively analyzed in [7, 11, 12]. Jain [8] has focused on C4.5 decision tree and compared its working in the various mining tools. [14–16] give a detailed explanation on the use of weka, tanagra and knime respectively.

3 Overview of the Comparison

In this section, we have described the tools used for conducting the data analysis. In addition to this, we have discussed the data sets and the methods applied to use classification algorithms [7, 12] in these tools to get results.

3.1 Tools Used

The data mining tools that we have compared have been described below:

- **WEKA** toolkit [14] is developed by the University of Waikato which is situated in New Zeland. It is one of the easiest to use software and thus has a large user base. All the machine learning and data mining algorithms present in it have been written in Java. WEKA contains various functionalities like splitting, validation, regression and mining. It is widely used for research and didactic purposes.
- **Tanagra** [15] is a free data analysis tool. In addition to machine learning and data analysis, it also supports statistical learning algorithms. The primary function of Tanagra is to give students and teachers a free tool for mining. The secondary function

is to allow researchers to make their own classification rules and compare them with pre-existing methods. The tertiary purpose is to provide developers their open source code so that they could learn how the tool was made. Hence it could also be used to learn programming techniques in addition to the basic functions of a data mining tool.

- **KNIME** [16] is a comprehensive tool for analysis, exploration, and visualization of data. KNIME has been made using rigorous techniques and is used by a lot of people in the research academia. It is a modular tool that could be used to make data flows by connecting various functions, selectively run a part of it and then visualize the result obtained to get deeper insights.

3.2 Datasets

We have taken some of the common datasets from the UCI data repository for running the test. The datasets used are Sound, Cancer Breast, Evaluation Car, Country Honour, Alphabets, Plant Culture, Soybean Large, Spamming and Animal data. The datasets have been so chosen as they are very different from each other where the data objects range from 100 to 1000. In addition to this, the number of attributes vary as well and also some of the datasets are multi-attribute ones. So choosing like this make the analysis all the more comprehensive and credible.

3.3 Supervised Learning Techniques

The comparison of the tools [4, 9] is done by classification of the datasets listed in the above sub-section by using various classifiers [3] like naive bayes, zero r, one r, decision tree c4.5, support vector machine and k-nearest neighbor. First percentage split method is applied where 66% of data is used for training and 34% data for testing. Then k-fold cross-validation [2] is done where the data is split into k equal parts and one part is used for testing and the other parts for training the classifier.

4 Evaluation and Analysis

The evaluation of the tools [10] selected has been done on the back of the supervised learning algorithms. The results obtained by the tools on the application of the data mining algorithms to the datasets has been described in this section.

4.1 Preliminary Discussion

The methodology used to do the tests was as follows: percentage split of 66% training and 34% testing is there. And 10 fold cross-validation [2] has been applied subsequently. Later the accuracy measures were taken to compare the tools. Some capabilities of the data mining tools were found out on running the classifiers and using it for various datasets. In Weka all the algorithms ran successfully, the results were given by all the six classification techniques.

In Table 1 we have shown that some of the algorithms for different datasets were unable to run on KNIME [15] and Tanagra [16]. We observed that this is because of the following reasons; the dataset is a multi-class one but the algorithm on the tool is only able to process a binary class dataset this is referred in the table with entry as CM. The other reason being that the values in the dataset are discrete but the algorithm can only handle continuous values which have been referenced in the table below as DV. The final reason that the tool itself does not has the functionality of the algorithm present which is tagged as CH in the table.

Table 1. Knime and Tanagra's capability to handle the algorithms

	Sound	Cancer Breast	Evaluation Car	Country Honour	Alphabets	Plant Culture	Soybean Large	Spamming	Animal data
Naïve Bayes	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD
1R	CH	CH	CH	CH	CH	CH	CH	CH	CH
C4.5 (DT)	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD
Support VMachine	CM/DV	GOOD	CM/DV	CM	CM	CM/DV	CM/DV	GOOD	CM
K-Nearest	DV	GOOD	DV	GOOD	GOOD	DV	DV	GOOD	GOOD
0R	CH	CH	CH	CH	CH	CH	CH	CH	CH

GOOD: Algorithm Run Successfully. CH: Cannot Handle. DV: Discrete Value. CM: Classes are many

One Rule algorithm (1R) is not present in KNIME, and Tanagra. ZeroR is also not there in KNIME and Tanagra, and hence, both of them are presented in the table as CH. In addition to this KNN is not able to comprehend soybean, car, sound and plant culture datasets as they have discrete values. Only analysis for spamming and breast cancer is done by support v machine which is due to the fact that other datasets contain either multiple classes or discrete values.

4.2 Accuracy Measure Achieved by the Algorithms

The results obtained on running various algorithms on Weka are displayed in Table 2 while doing percentage split. The naive bayes and 1R range from 44%–97% and 5%–94% respectively. C4.5 [8] and SVM have almost the same range from 84% to 96%. Whereas KNN is in the upper echelons from 60%–99%. Finally, 0R ranges 4% to 70%.

Table 2. Percentage split on weka.

	Sound	Cancer Breast	Evaluation Car	Country Honour	Alphabets	Plant Culture	Soybean Large	Spamming	Animal data
Naïve Bayes	72.54%	95%	88.6%	44%	65.58%	91%	91.67%	78%	97%
1R	43.97%	93.13%	70%	5%	17.93%	70.5%	39%	76.82%	37%
C4.5 (DT)	84.23%	96.49%	91%	49.59%	86.58%	96%	91.67%	93.30%	95.31%
Support VMachine	85.53%	96.5%	93%	60.1%	82.24%	93%	94%	90.65%	95.31%
K-Nearest	59.55%	96.4%	91.76%	52.63%	94.68%	98.64%	90%	90.38%	77%
0R	28.38%	64.9%	70%	35%	4%	33.1%	13%	61.69%	38%

The result of running the various algorithms on KNIME is in Table 3. 1R and 0R did not run in KNIME. Naïve Bayes ranged from 43%–96%. SVM and KNN were not giving results for four datasets the reason for which has been explained before but they have ranges of 68%–99% and 26%–98% respectively. C4.5 was from 44% to 97%.

Table 3. Percentage split on knime.

	Sound	Cancer Breast	Evaluation Car	Country Honour	Alphabets	Plant Culture	Soybean Large	Spamming	Animal data
Naïve Bayes	54.21%	96.00%	87.21%	43.51%	63%	91%	85%	90%	83%
1R	CH	CH	CH	CH	CH	CH	CH	CH	CH
C4.5 (DT)	69%	96.11%	94.61%	44.21%	86.41%	97.81%	67.21%	92.21%	93.29%
Support VMachine	CM/DV	98.91%	CM/DV	CM	CM	CM/DV	CM/DV	68.00%	CM
K-Nearest	DV	97.51%	DV	26.92%	96.00%	DV	DV	81%	45.70%
0R	CH	CH	CH	CH	CH	CH	CH	CH	CH

At last, the Tanagra tool was tested by doing percentage split. NB and C4.5 ranged from 60%–97% and 40%–96% respectively. 1R and 0R have no implementation on Tanagra. Whereas SVM and KNN selectively run only for a few datasets where the range obtained is 90% to 96% and 29% to 99% respectively (Table 4).

Table 4. Percentage split on tanagra.

	Sound	Cancer Breast	Evaluation Car	Country Honour	Alphabets	Plant Culture	Soybean Large	Spamming	Animal data
Naïve Bayes	67.34%	96.91%	88.35%	64.75%	60.6%	91%	89%	88.55%	89%
1R	CH	CH	CH	CH	CH	CH	CH	CH	CH
C4.5 (DT)	82.92%	92%	91%	40.40%	86%	96%	91.67%	91.84%	89%
Support VMachine	CM/DV	95.53%	CM/DV	CM	CM	CM/DV	CM/DV	89.62%	CM
K-Nearest	DV	99%	DV	29%	95%	DV	DV	80.18%	88.97%
0R	CH	CH	CH	CH	CH	CH	CH	CH	CH

From Tables 5, 6 and 7 we have switched to 10 fold cross validation. In Table 5 the result obtained by weka [5, 6] has been displayed. SVM and KNN range equally from 60% to 97%. 1R and C4.5 have ranges 4%–94% and 57%–96% respectively. On the other hand, NB and 0R from 56% to 96% and 4% to 70% respectively.

Table 5. 10-Fold-CV using weka.

	Sound	Cancer Breast	Evaluation Car	Country Honour	Alphabets	Plant Culture	Soybean Large	Spamming	Animal data
Naïve Bayes	74.56%	96%	86.64%	56.26%	65.23%	91.43%	93.98%	80%	95%
1R	47.57%	93.81%	70%	4.45%	17.31%	70.32%	35.56%	79.51%	58.87%
C4.5 (DT)	77%	95.6%	92%	57%	88%	97%	92.62%	93%	92%
Support VMachine	82%	97.00%	94.6%	61%	82%	93%	94.2%	91.22%	97.04%
K-Nearest	63%	97.82%	94.63%	58.33%	96.63%	99.49%	90%	80.31%	95%
0R	26.33%	65.5%	70%	36.67%	4%	44.44%	13.5%	60.7%	41%

Table 6. 10-Fold-CV using knime.

	Sound	Cancer Breast	Evaluation Car	Country Honour	Alphabets	Plant Culture	Soybean Large	Spamming	Animal data
Naïve Bayes	59%	95.81%	86.91%	52.61%	62%	90.47%	91.30%	90%	88%
1R	CH	CH	CH	CH	CH	CH	CH	CH	CH
C4.5 (DT)	71%	94%	94.60%	55.61%	88.61%	98.32%	73.11%	92.41%	94.21%
Support VMachine	CM/DV	97.42%	CM/DV	CM	CM	CM/DV	CM/DV	68.41%	CM
K-Nearest	DV	98.61%	DV	34.11%	95%	DV	DV	81%	71%
0R	CH	CH	CH	CH	CH	CH	CH	CH	CH

Table 7. 10-Fold-CV using tanagra.

	Sound	Cancer Breast	Evaluation Car	Country Honour	Alphabets	Plant Culture	Soybean Large	Spamming	Animal data
Naïve Bayes	70.00%	95.90%	85.41%	63.74%	60%	90%	90%	88%	93.00%
1R	CH	CH	CH	CH	CH	CH	CH	CH	CH
C4.5 (DT)	72.47%	94.44%	87.54%	57.95%	86.95%	96.94%	91.35%	92.65%	89.11%
Support VMachine	CM/DV	97%	CM/DV	CM	CM	CM/DV	CM/DV	90%	CM
K-Nearest	DV	97.92%	DV	26.37%	96.87%	DV	DV	80.00%	93.00%
0R	CH	CH	CH	CH	CH	CH	CH	CH	CH

The accuracy measures for KNIME toolkit on using 10 fold is shown in Table 6. NB and C4.5 have their measures in range from 52%–96% and 55%–98% respectively. 1R and 0R are not implementable in it. SVM and KNN range from 68%–97% and 34%–99% respectively.

Finally, Table 7 shows accuracy measures for tanagra. 1R and 0R cannot be implemented in it. SVM [11] and KNN do not give readings for some datasets but for the one they give, it ranges from 90% to 97% and 26% to 98% respectively. Naive bayes ranged from 60%–96%. 58% to 97% is the range of C4.5.

4.3 Improvement in Performance

Here, we have measured the effect of using the various evaluation methods in the toolkits [13]. For that, we have displayed in the graph shown in Fig. 1 the increase in accuracy as we move from percentage split to cross-validation approach. Weka has shown the highest accuracy improvement which is of 32 accuracy measures. KNIME follows it with an increase of 12 accuracy measures. Whereas Tanagra falls back with only 8 accuracy measures.

Also, we have shown the number of accuracy measures decreased in the three kits. In KNIME only a decrease of 4 accuracy measures was there which is the best rate. Whereas in Weka there was a sharp reduction of 7 accuracy measures. Tanagra suffered from the worst rate of 8 measures decrease when switching from percentage split to cross-validation approach.

At last, the number of no results obtained by the tools has also been accounted for. Weka has shown all the results whereas KNIME and Tanagra have 29 each no results.

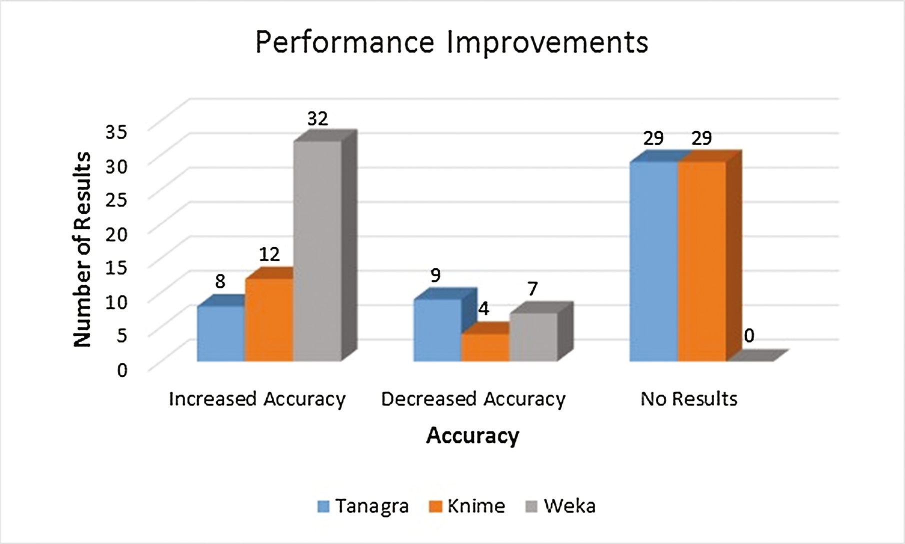

Fig. 1. The improve in performance when switched from % split to k-fold cross validation

These findings show that no tool is better than the other due to the fact that some of the datasets may not work in it or at other times the algorithm cannot be implemented properly in the mining tool. For instance, the KNN classifier could handle the multiclass attributes in Weka which is not the case in KNIME [9] and Tanagra.

When comparing the tools on the basis of applicability, Weka has the highest. This is because it was able to run all the algorithms successfully on all the datasets. Whereas KNIME and Tanagra [10] were able to run only two fully; namely Naive Bayes and C4.5. SVM and KNN ran partially on them and ZeroR and OneR had no implementation on the two tools.

The performance of the tools could be adjudged on the basis of the increase in accuracy and decrease in accuracy measure. Weka has the highest performance in which the increase and decrease were respectively 32 and 7. KNIME and Tanagra were low on performance as the increase in their accuracy measures was roughly 12 and 8 respectively whereas the decrease was by 4 and 9 accuracy measures.

5 Conclusion and Future Work

This research presided on conducting a comparison of three data mining toolkits using 9 datasets through the application of 6 algorithms namely which have been discussed before. The research concludes that no tool is better than the other in terms of the supervised learning task as it depends on the datasets used and also how the algorithms were implemented in the tool. However, when seeing them in terms of applicability, we get to know that Weka comes first as it was able to run all the algorithms followed by Tanagra and finally KNIME. At last, Weka has achieved the highest performance measure when moving from percentage split to 10 fold cross-validation approach. KNIME comes after it followed by Tanagra.

As a future research, other methods like association rules, ensemble learning, regression or clustering [9] could be done to get deeper insights on these tools and to be able to know in much detail about which is more applicable than the other.

References

1. Baker, R.S., Yacef, K.: The state of educational data mining in 2009: a review and future visions. JEDM-J. Educ. Data Min. **1**(1), 3–17 (2009)
2. Bengio, Y., Grandvalet, Y.: No unbiased estimator of the variance of k-fold cross-validation. J. Mach. Learn. Res. **5**, 1089–1105 (2004)
3. bin Othman, M.F., Yau, T.M.S.: Comparison of different classification techniques using WEKA for breast cancer. In: 3rd Kuala Lumpur International Conference on Biomedical Engineering 2006, pp. 520–523. Springer, Berlin, Heidelberg (2007)
4. Chauhan, N., Gautam, N.: Parametric comparison of data mining tools (2015)
5. David, S.K., Saeb, A.T., Al Rubeaan, K.: Comparative analysis of data mining tools and classification techniques using weka in medical bioinformatics. Comput. Eng. Intell. Syst. **4**(13), 28–38 (2013)
6. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. ACM SIGKDD Explor. Newslett. **11**(1), 10–18 (2009)
7. Iyer, A., Jeyalatha, S., Sumbaly, R.: Diagnosis of diabetes using classification mining techniques.arXiv preprint arXiv:1502.03774 (2015)
8. Jain, D.: A comparison of data mining tools using the implementation of C4.5 Algorithm. Int. J. Sci. Res. **3**(8), 33–37 (2014)
9. Patil, P.H., Thube, S., Ratnaparkhi, B., Rajeswari, K.: Analysis of different data mining tools using classification, clustering and association rule mining. Int. J. Comput. Appl. **93**(8), 35–39 (2014)
10. Solanki, H.: Comparative study of data mining tools and analysis with unified data mining theory. Int. J. Comput. Appl. **75**(16) (2013)
11. Tolan, G.M., Soliman, O.S.: An experimental study of classification algorithms for terrorism prediction (2015)
12. Vaithiyanathan, V., Rajeswari, K., Tajane, K., Pitale, R.: Comparison of different classification techniques using different datasets. Int. J. Adv. Eng. Technol. **6**(2), 764 (2013)
13. Wimmer, H., Powell, L.M.: A comparison of open source tools for data science. J. Inf. Syst. Appl. Res. **9**(2), 4 (2016)
14. WEKA, the University of Waikato. http://www.cs.waikato.ac.nz/ml/weka/
15. Tanagra – a Free Data Mining Software for Teaching and Research. http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html
16. KNIME (Konstanz Information Miner). http://www.knime.org/