

**APPLYING COMBINATORIAL  
TESTING TO DATA MINING  
ALGORITHMS**

**MICANS INFOTECH**



# ABSTRACT

- Data mining algorithms are used to analyze and discover useful information from data.
- paper presents an experiment that applies Combinatorial Testing (CT) to five data mining algorithms implemented in an open-source data mining software called WEKA
- For each algorithm, we first run the algorithm with 51 datasets to study the impact different datasets have on the test coverage.
- We select one dataset that achieves the highest branch coverage.
- Test effectiveness is measured using branch and mutation coverage.



# EXISTING SYSTEM

- Big data applications are becoming more popular as large amounts of data are generated and collected in virtually every domain, e.g., e-commerce, social networking, and scientific computing
- t-way testing strategy for testing concurrent programs
- our knowledge, our work is the first one that applies CT to data mining algorithms.

MICANS INFOTECH



# DISADVANTAGES

- random testing to detect deadlocks
- effectiveness of pairwise testing

**MICANS INFOTECH**



# PROPOSED SYSTEM

- Combinatorial Testing (CT) to five data mining algorithms implemented in an open-source data mining software called WEKA
- For each algorithm, we first run the algorithm with 51 datasets to study the impact different datasets have on the test coverage
- Test effectiveness is measured using branch and mutation coverage
- Our results suggest that when testing data mining algorithms:  
(1) larger datasets do not necessarily achieve higher coverage than smaller datasets;
- (2) test coverage increases progressively slower as test strength increases
- (3) branch coverage correlates well with mutation



# ADVANTAGES

- detailed code analysis
- create representative datasets
- effectiveness of CT on data mining algorithms

**MICANS INFOTECH**



# REFERENCES

- A. Frank, and A. Asuncion, A.UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California. School of Information and Computer Science, 213, 2010
- A. Gargantini, J. Petke, M. Radavelli and P. Vavassori. Validation of Constraints Among Configuration Parameters Using Search-Based Combinatorial Interaction Testing. Proceedings of the International Symposium on Search Based Software Engineering, 49-63, 2016.
- M. Grindal and J. Offutt. Input parameter modeling for combination strategies. Proceedings of the 25th conference on IASTED International Multi-Conference Software Engineering, 255-260, 2007.
- M. Hoffmann, B. Janiczak, E. Mandrikov and M. Friedenhagen. Jacoco code coverage tool. Online , 2016

