International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

A review of Dataset Size Impacts the Effectiveness of Data Mining Methods

[1]Dammalapati Venu, B.Tech,III-Year, CSE Department, Vignan's Foundation for Science, Technology and Research (Deemed to be University), Guntur, Andhra Pradesh, India

Abstract:

This study delves into the critical relationship between dataset size and the efficiency of data mining algorithms. Data mining, also known as Knowledge Discovery in Databases (KDD), plays a crucial role in analyzing vast datasets to extract valuable insights. Data mining, a fundamental aspect of knowledge discovery, is pivotal in extracting valuable insights from vast datasets. However, the performance of data mining algorithms can vary significantly based on the size of the dataset they operate on. Understanding this influence is vital for optimizing algorithm selection and deployment in practical applications. Through systematic exploration and analysis, this research aims to elucidate the impact of dataset size on algorithm efficiency, shedding light on the optimal conditions for algorithmic performance. By examining various datasets across different scales and employing diverse data mining techniques, this study endeavors to uncover patterns, trends, and challenges associated with scaling data mining algorithms. The findings offer valuable insights for practitioners and researchers alike, guiding the development of more robust and scalable data mining solutions to meet the evolving demands of modern data-driven environments.

Keywords: Data Mining, Dataset Size, Knowledge Discovery.

I. Introduction

Knowledge Discovery in Databases (KDD) techniques encapsulate fundamental concepts crucial for comprehending Data Mining (DM) and Machine Learning (ML), serving as vital resources for professional training. However, these techniques often exhibit complexity, necessitating thorough practical analysis to discern their primary advantages and disadvantages. Such scrutiny facilitates a deeper understanding of their components and behavior across diverse

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

problem domains, aiding in selecting the most suitable approach for specific cases. Yet, a significant challenge arises as many of these techniques demand programming expertise and substantial time and effort to implement. Consequently, KDD instruction can devolve into mere programming exercises rather than focusing on analyzing the distinctive traits of each technique. This obstacle can be mitigated through the utilization of software tools that alleviate the burden of programming tasks, enabling students to concentrate on the intrinsic characteristics of KDD algorithms. However, on the web, there is a scarcity of implementations for these techniques, and the existing ones are often tailored for specific applications, making them challenging to apply in practical scenarios. This necessitates various modifications and adaptations to the original source code, posing risks to the accuracy of implementations and consequently, the reliability of conclusions drawn. Fortunately, in recent years, numerous software tools have emerged to address these challenges.

Data mining involves a meticulous examination of vast data sets to extract pertinent information. It can be described as the process of uncovering knowledge from extensive data repositories, hence earning the moniker "Knowledge Mining." Commonly referred to as Knowledge Discovery in Databases (KDD), data mining entails the extraction of implicit, previously unknown, and potentially valuable insights from database records. Preprocessing tasks encompass identifying and rectifying erroneous or absent data, as well as eliminating noise or outliers to enhance the quality of the dataset. The actual data mining process entails executing tasks to yield desired outcomes, while the integration and evaluation of results facilitate user comprehension. Various types of knowledge necessitate distinct representations, such as classification, clustering, and association rules, with this study focusing on classification and clustering techniques.

The pressing demand within the blood bank sector underscores the importance of efficiently leveraging stored data. Such data analysis serves as a pivotal tool for evaluating information collected by blood banks through their information systems. Specifically, this study aims to classify and predict the number of blood donors based on age and blood group information. Utilizing the J48 algorithm and the Weka tool, a data mining model was constructed to extract insights into blood donor classification, aiding clinical decision-making in blood bank centers. Multiple classification algorithms were evaluated to determine the most effective approach for classifying data accurately.

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

Identifying regular blood donors can enable blood banks and voluntary organizations to systematically plan and organize blood donation campaigns with greater efficiency.

II. Data Mining Models

The Supervised model predicts unknown data values by utilizing known values, while the Unsupervised model discerns patterns or relationships within data and investigates their properties. The figure below illustrates the data mining models and tasks employed in our study.

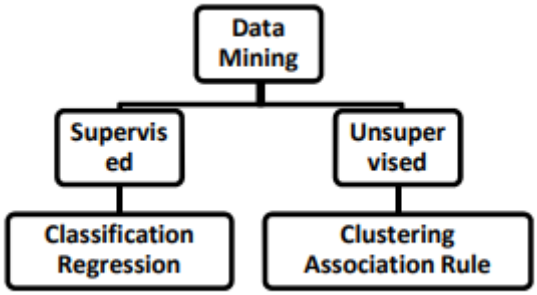

Figure.1 Data mining models

A. Supervised Learning

Supervised learning, often referred to as directed data mining, divides the variables under investigation into two categories: explanatory variables and one (or more) dependent variables. The goal is to establish a relationship between the explanatory variables and the dependent variable, much like in regression analysis. In applying directed data mining techniques, it is necessary to have known values for the dependent variable for a sufficiently large portion of the dataset. Training data comprises both the input and the desired results. In some cases, the correct results (targets) are known and provided to the model during the learning process. Creating an appropriate training, validation, and test set is crucial. These methods are typically fast and accurate, and they must be able to generalize, providing the correct results when new data is given as input without knowing the target beforehand.

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

B. Unsupervised Learning

Unsupervised learning aligns more closely with the exploratory spirit of Data Mining. In unsupervised learning situations, all variables are treated equally; there is no distinction between explanatory and dependent variables. However, despite being called undirected data mining, there is still a target to achieve. This target might be as general as data reduction or more specific, such as clustering. The model is not provided with correct results during training. Unsupervised learning can be used to cluster input data into classes based solely on their statistical properties. Cluster significance and labeling can be carried out, even if the labels are only available for a small number of objects representative of the desired classes. The boundary between supervised learning and unsupervised learning mirrors the distinction between discriminant analysis and cluster analysis. Supervised learning requires a well-defined target variable and a sufficient number of its values. For unsupervised learning, typically, either the target variable is unknown or has been recorded for too few cases.

C. Classification

Classification, a data mining function, assigns items in a collection to target categories or classes. Its aim is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify blood donors' availability based on blood group and location. A classification task begins with a dataset in which the class assignments are known. Classifications are discrete and do not imply order. Continuous, floating-point values would indicate a numerical, rather than a categorical, target. The input data, also known as the training set, consists of multiple records, each having multiple attributes or features. Each record is tagged with a class label, and the objective of classification is to analyze the input data and develop an accurate description or model for each class using the features present in the data.

D. Regression

Regression analysis encompasses a broader range of techniques than ordinarily appreciated. Statisticians commonly define regression as understanding "as far as possible with the available data how the conditional distribution of some response y varies across subpopulations determined by the possible values of the predictor or predictors." Regression is a data mining (machine learning) technique used to fit an equation to a dataset. The simplest form of regression, linear

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

regression, uses the formula of a straight line (y = mx + b) and determines the appropriate values for m and b to predict the value of y based on a given value of x. Advanced techniques, such as multiple regression, allow the use of more than one input variable and allow for the fitting of more complex models, such as a quadratic equation.

E. Clustering

A cluster is a subset of objects that are "similar." It is a subset of objects such that the distance between any two objects in the cluster is less than the distance between any object in the cluster and any object not located inside it, connected region of a multidimensional space containing a relatively high density of objects. Clustering is a process of partitioning a set of data (or objects) into a set of meaningful subclasses, called clusters. It helps users understand the natural grouping or structure in a dataset. Clustering is unsupervised classification, meaning there are no predefined classes. It is used either as a stand-alone tool to gain insight into data distribution or as a preprocessing step for other algorithms. Moreover, clustering can be utilized for data compression, outlier detection, and understanding human concept formation.

F. Association Rules

Association rules are if/then statements that help unveil relationships between seemingly unrelated data in a relational database or other information repository. Each association rule consists of two parts: an antecedent (if) and a consequent (then). The antecedent is an item found in the data, while the consequent is an item found in combination with the antecedent. These rules are created by analyzing data for frequent if/then patterns and using the criteria of support and confidence to identify the most important relationships.

Support indicates how frequently the items appear in the database, while confidence indicates the number of times the if/then statements have been found to be true. In data mining, association rules are valuable for analyzing and predicting customer behavior. They play a crucial role in shopping basket data analysis, product clustering, catalog design, and store layout. Programmers use association rules to develop programs capable of machine learning. Machine learning, a type of

artificial intelligence, aims to build programs with the ability to become more efficient without being explicitly programmed.

Conclusion

Data mining is the science of extracting information from large databases to derive knowledge from data and present it in a form that is easily understood by humans. Classification techniques are the primary tasks of data mining with broad applications for classifying various kinds of data. It is used to categorize items based on their features with respect to a predefined set of classes.

References

1. Han, J., Kamber, M., & Pei, J. (2011). Data mining: concepts and techniques. Elsevier.

2. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data mining: practical machine learning tools and techniques. Morgan Kaufmann.

3. Tan, P. N., Steinbach, M., & Kumar, V. (2019). Introduction to data mining. Pearson.

4. Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. Journal of Data Warehousing, 5(4), 13-22.

5. Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. ACM SIGMOD Record, 22(2), 207-216.

6. Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37-54.

7. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

8. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

9. Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

10. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

11. Hand, D. J., Mannila, H., & Smyth, P. (2001). Principles of data mining. MIT press.

12. Mitchell, T. M. (1997). Machine learning. McGraw Hill.

13. Thabtah, F. (2018). A review of associative classification mining. The Knowledge Engineering Review, 33, e3.

14. Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... & Steinbach, M. (2008). Top 10 algorithms in data mining. Knowledge and information systems, 14(1), 1-37.

15. Aggarwal, C. C., & Reddy, C. K. (2013). Data clustering: algorithms and applications. CRC Press.

16. Hand, D. J., & Yu, K. (2001). Idiot's Bayes—Not so stupid after all? International Statistical Review, 69(3), 385-398.

17. Lewis, D. D., Yang, Y., Rose, T. G., & Li, F. (2004). RCV1: A new benchmark collection for text categorization research. Journal of Machine Learning Research, 5(Apr), 361-397.

18. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

19. Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. Intelligent data analysis, 6(5), 429-449.

20. Quinlan, J. R. (1993). C4. 5: Programs for machine learning. Elsevier.