IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

A STUDY ON SEQUENCE MINING ALGORITHMS IN DATA RELATED TO EMM AND MARKOV PROCESS

Neethu John
Research Scholar, Department of Computer Applications,
Noorul Islam Centre for Higher Education, Kumaracoil, Tamil Nadu, India

Dr. J.R Jeba
Associative Professor and Head, Department of Computer Applications,
Noorul Islam Centre for Higher Education, Kumaracoil, Tamil Nadu, India.

Abstract

To analyze the behavior of various domains of data, sequence mining is an important task. The random forest algorithm, for example, can be used to extract patterns from data. Our main purpose in this paper is to compare the accuracy of various algorithms used in sequence analysis so far, such as SPADE, GPS, KNN, and Naive Bayesian. This paper is based on the research title: "Discover sequences using exceptional model mining and transition behavior". Even if the selection of data mining algorithm is domain specific the algorithms described must be compared to the accurate selection of algorithm that is suitable for the research.
This paper focuses on work done by various researchers.

Keywords: Sequence mining, Pattern analysis, Naive bayesian,EMM,Markov Process

Introduction

Sequence mining algorithms are an important tool for analyzing sequential data, such as those found in genetic sequences, social media interactions, and financial transactions. Numerous studies have been undertaken on the various types of sequence mining algorithms to enhance their effectiveness and efficiency. Sequence mining algorithms refer to a set of data mining techniques that aim to discover knowledge in sequential datasets. These algorithms analyze data that is represented as sequences and can identify patterns or trends in the order in which events occurred. This is particularly useful in situations where events occur over time, such as website clickstreams, retail transactions, or medical procedures. By analyzing the order in which events occurred,

Sequence mining algorithms can identify frequent patterns and subsequences to help improve data-driven decision-making. Some commonly used sequence mining algorithms include prefix span, SPAM (Sequential Pattern Mining), and GSP (Generalized Sequential Patterns).One important concept in sequence mining algorithms is the notion of transition behavior,

which refers to the likelihood or probability of a particular event following another event. By analyzing transition behavior, sequence mining algorithms can identify which events are more likely to occur based on prior occurrences. Naive Bayesian. In addition to sequence mining algorithms, there are also translation algorithms that can be used in combination with them. Translation algorithms in sequence mining aim to transform the original data format into a more suitable one for analysis

EMM

Exceptional model mining is a data mining technique used to identify exceptional or anomalous patterns within a dataset. It is a field within data mining and machine learning that focuses on discovering patterns that deviate significantly from the norm or exhibit unusual behavior. Exceptional model mining is also commonly referred to as anomaly detection or outlier detection. The primary goal of exceptional model mining is to identify data points or patterns that are significantly different from the majority of the data. These exceptional patterns are often considered outliers, anomalies, or rare events.

Applications: Exceptional model mining has a wide range of applications across various domains. It is used in fraud detection, network security, quality control in manufacturing, healthcare (for detecting unusual patient conditions), and many other fields where detecting rare and potentially important events is crucial.

Techniques: Exceptional model mining employs various statistical, machine learning, and data mining techniques to identify anomalies. Common approaches include clustering, density estimation, distance-based methods, and machine learning algorithms like isolation forests, one-class SVMs (Support Vector Machines), and neural networks.

Unsupervised vs. Supervised: Anomaly detection can be carried out in both unsupervised and supervised manners. In unsupervised anomaly detection, the algorithm identifies anomalies without using labeled data, while supervised approaches rely on labeled examples of normal and anomalous behavior for training.

Challenges: Exceptional model mining faces challenges such as defining what constitutes an anomaly, dealing with imbalanced datasets (where normal instances vastly outnumber anomalies), and adapting to evolving data distributions in dynamic environments.

Evaluation: The performance of exceptional model mining algorithms is typically evaluated using metrics like precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). These metrics measure how well the algorithm identifies anomalies while minimizing false positives. Real-time Detection: In some applications, it is essential to detect anomalies in real-time, such as identifying fraudulent transactions as they occur. Exceptional model mining algorithms can be adapted to work in real-time or near real-time environments. Interpretability: Understanding why a particular data point is flagged as an anomaly is crucial in some applications. Ensuring the interpretability of the model's results can be a significant challenge.

Scalability: Scalability is a concern when dealing with large datasets. Developing algorithms that can efficiently process massive amounts of data is an ongoing research area. Continuous Learning: In dynamic environments, exceptional model mining techniques may need to adapt to changes in data distributions over time, requiring continuous learning and updating of models. Hidden Markov.Models (HMMs) are a powerful statistical tool used in various fields, including pattern recognition, speech recognition, bioinformatics, and natural language processing. They are particularly useful when dealing with sequential data, where the underlying system has some hidden states that generate observable outcomes or emissions. HMMs are employed for various pattern recognition tasks because of their ability to model complex dependencies in sequential data.

Basic Components:

States: An HMM represents a system that undergoes a sequence of transitions between a finite number of hidden states. These states are not directly observable and are often used to model underlying patterns or behaviors.

Observations or Emissions: At each state, the system emits an observable outcome. These emissions are associated with probabilities and depend on the current hidden state.

State Transition Probabilities: HMMs are characterized by transition probabilities that define the likelihood of moving from one hidden state to another. These transition probabilities are usually represented by a transition matrix. The transitions are typically modeled as a Markov process, meaning that the probability of moving to the next state depends only on the current state and not on the previous states.

Emission Probabilities: Each hidden state has associated emission probabilities for different observable outcomes. These probabilities are often represented by an emission matrix. The emission probabilities describe the likelihood of generating a specific observation when in a particular hidden state.

Initialization: HMMs typically start in one of the hidden states, which is determined by an initial probability distribution.

Three Fundamental Problems: HMMs are used to address three primary problems: Evaluation: Calculate the probability of observing a given sequence of observations given the model. This problem is solved using the forward algorithm.

Decoding: Determine the most likely sequence of hidden states that generated a given sequence of observations. This problem is solved using the Viterbi algorithm.

Learning: Adjust the model's parameters (transition and emission probabilities) based on observed data. This problem is often addressed using the Baum-Welch algorithm, which is a form of the Expectation-Maximization (EM) algorithm.

HMMs assume the Markov property, meaning that the future state depends only on the current state. This assumption may not hold in all real-world scenarios.

Estimating the model's parameters (training) can be computationally intensive, especially for large datasets.

HMMs may struggle with very long sequences due to the vanishing gradient problem.

Related Work

In[1] A work by Stefan Bloemheuvel,Benjamin Kloepper,Jurgen Van Den Hoogen,Martin AtzmuellerPortrayed that enhanced sequential pattern algorithm with Marky chain probabilities .Markov chain probabilities are used to find sequences of equal support.

In[[2] Improved naive bayesian classification algorithm for traffic risk management by Hong chen ,Songhua HU,Ruihua,Xiuju Zhao used the technique of Laplace collaboration to improve the short coming of naive bayesian algorithm.A 95% accuracy is obtained with a study on 400 samples and 24 attribute categories.

In[3]A study by Sadri Sa'di1, Amanj Maleki1, Ramin Hashemi2, Zahra Panbechi1 and Kamal entitled Comparative study of various data mining algorithms in diagnosis of type 2 diabetics. in in this study the data set used was Pima Indian data set The dataset includes 768 samples from diabetic patients; the algorithm used for comparesis includes naive bayesian,RBF network, J 48 the use of the technique called Weka. The result obtained was almost 77% accuracy with Limited number of samples.

In[4]An exploring through Exceptional Model Mining for Repeated Cross-Sectional Data (EMM-RCS) by Rianne Margareta Schouten Wouter Duivesteijn Mykola Pechenizkiy*The above study depicts the importance of exceptional model mining instance design to find sub groups displaying exceptional trend behavior in RCS data.The Emma model developed by the scientists showed expressive quality measures built on standard error of trend estimates to find a variety of exceptionality .

In[5] A Naïve Bayesian Classifier for Educational Qualification,by S. Karthika* and N. SairamThis work obtained 90% of accuracy by correctly classifying the data with high Kappa value into 3 classes and this research mainly focuses on the application of naive Bayesian classifiers in the field of education to classify the data with high accuracy.

In[6]J.G.Lianga*, X.F.Zhoua, P.Liua, L.Guoa, S.Baia, discussed An EMM-based Approach for Text ClassificationExplicit Marco model is applied for text classification based on hidden Marco model text classification is previously executed by variety of algorithms such as SVM decision free etc but mm explicitly Marco model to select effective features from the target set. The study showed the result that Emm is highly competitive with SVM.

In[7]A universally applicable EMM framework by Wouter de Ruiter EMM model dynamically select Components ,a quality metric is used for subgroup quality and a tree vice algorithm which is a numerical indicator Used as a search strategy

In[8]Exceptionally monoton models-The Rank correlation modelClass for exceptional model mining this paper by Lennart Downar Wouter Duivestejin Used the concept of EMM with the Pearson rank correlation, spearman rank correlation and Kendall's Correlation between two target variables. The resultant model class obtained an exceptionally monotone relation between targets.

In[9] Detecting individual content-structure patterns in time series data,A work by Lu Feng, Xianyang Xu, Hua Yuan and Qian Zhangt.This paper mainly focused for an efficient method to handle Spatio temporal dataWhich is done by the technique called CSM content based and structure based mapping and this method is better than SWM which stands for split whole mapping. Time Sequence based user behavior pattern describing method in which time related events are mapped into time streams respectively.

In [10]An Implementation of Naive Bayes Classifier,Feng-Jen Yang were fully based on bayesian theory.python is used for implementing naive bayesian concepts without falling for an intensive coding strategy.conditional probability is the underlying system.

In[11]Prediction of student Assessments readiness in Online learning Environments :The sequence matters a detailed study by Donia malekian James bailey and Gregor kennedy focused on student modeling structure and relationship discovery.Comparison of LR,MLP and LSTM algorithms performance based AUC and accuracy.

In[12]Mining sequences with exceptional transition behavior of varying order using quality measures based on information-theoretic scoring functions Rianne M. Schouten1 · Marcos L. P. Bueno ,Wouter Duivesteijn , Mykola Pechenizkiy.This work

In[14]Exceptional model mining for behavioral data analysis Adnene Belfodil ,This work portrayed two approaches on for analyzing behavioral data and another one for discovering pattern(exceptional) from voting data.The study was completed by SD (subgroup discovery) and EMM(Exceptional model mining)

In [16]A study by Kavitha and Natarajan (2018) evaluated the performance of several sequence mining algorithms, including PrefixSpan, SPADE, and GSP, on different datasets. Their findings indicated that PrefixSpan and SPADE were superior to GSP in terms of execution time, with PrefixSpan being the most efficient overall. In another study by Zhang et al. (2018), a novel sequence mining algorithm called HEPT was proposed, which used heuristic optimization techniques to improve the accuracy and efficiency of sequence mining. HEPT achieved better results compared to other existing sequence mining algorithms, such as PrefixSpan and GSP.

In [17] Additionally, a study by Li et al. (2020) proposed a hybrid sequence mining algorithm that combined the strength of both GSP and SPADE to enhance the performance of sequence mining for healthcare data.However, to the best of our knowledge, no study has directly addressed the performance of a sequence mining algorithm that combines heuristics with probabilistic modeling such as the naive Bayesian algorithm.

CONCLUSION

In conclusion, sequence mining and transition behavior algorithms are important techniques in data mining that have been extensively used in various applications**.** Sequence mining and transition behavior algorithms are widely used techniques in data mining that aim to extract valuable patterns from sequential data.

These techniques have been applied in numerous fields such as healthcare, finance, and retail to identify trends and patterns. Moreover, studies have shown that sequence mining and transition behavior algorithms can significantly improve decision-making processes and help organizations enhance their performance. Overall, the literature review highlights the importance of sequence mining and transition behavior algorithms in data mining. Their effectiveness in identifying patterns and relationships between sequential data results in better decision-making and can lead to significant improvements in organizational performance. Furthermore, the review reveals that while these techniques have several advantages, they also face some challenges. These challenges include difficulties in handling large datasets and incorporating real-time data, as well as ensuring proper privacy protection. In order to overcome these challenges, further research is needed to develop more efficient algorithms and techniques that can handle real-time data processing with a focus on ensuring privacy protection.As a comparative study, it can be stated that sequence mining and transition behavior algorithms are constantly evolving, with new advancements being made every year to improve their efficiency and accuracy

REFERENCES

1.Information Technology and Quantitative Management (ITQM2013) An EMM-based Approach for Text Classification J.G.Lianga , X.F.Zhoua , P.Liua , L.Guoa , S.Baia,bProcedia Computer Science 17 (2013) 506 – 513

2. Mouratis, T., Kotsiantis, S. Increasing the Accuracy of Discriminative Multinomial Bayesian Classifier in Text Classification. 4thInternational Conference on Computer Sciences and Convergence Information Technology. 2009, p1246-1251.

3. Research Article EMM-CLOUDS: An Effective Microcluster MinimalPruningCLustering-BasedTechnique for Detecting Outliers inData StreamsMohamedJ.G.Lianga, X.F.Zhoua, P.Liua, L.Guoa, S.Baia,bProcedia Computer Science 17 (2013) 506 – 513

4.International Journal on Computational Science & Applications (IJCSA) Vol.5, No.5,October 2015 EMM-CLOUDS: An Effective Microcluster and Minimal Pruning CLustering-Based Technique for Detecting Outliers in Data StreamsSadri Sa'di1, Amanj Maleki1, Ramin Hashemi2, Zahra Panbechi1 and KamalSadri Sa'di1, Amanj Maleki1, Ramin Hashemi2, Zahra Panbechi1 and Kamal

5. A Naïve Bayesian Classifier for Educational Qualification S. Karthika* and N. SairamIndian Journal of Science and Technology, Vol 8(16), DOI: 10.17485/ijst/2015/v8i16/62055, July 2015

6. An Implementation of Naive Bayes Classifier Feng-International Journal on Computational Science & Applications (IJCSA) Vol.5, No.5,October 2015

7. Yang2018 International Conference on Computational Science and Computational Intelligence (CSCI).

8. A universally applicable EMM framework/Wout de Ruite

9. Applying Sequence Mining for Outlier Detection in Process Mining Mohammad Reza Fani Sani1 , Sebastiaan J. van Zelst2 , Wil M.P. van der Aalst1,2

10.Comparative study of various data mining algorithms in diagnosis of type 2 diabeticsSadri Sa'di 1, Amanj Maleki1, Ramin Hashemi2, Zahra Panbechi1 and KamalInternational Journal on Computational Science & Applications (IJCSA) Vol.5, No.5,October 2015.

11 Improved naive Bayes classification algorithm for traffic risk management Hong Chen1, Songhua Hu2, Rui Hua3 and Xiujun Zhao4