International Journal of Engineering, Science and Mathematics
Vol. 6 Issue 3, July 2017,
ISSN: 2320-0294 Impact Factor: 6.765
Journal Homepage: http://www.ijesm.co.in, Email: ijesmj@gmail.com

A NOVEL APPROACH TO PREDICT ACADEMIC PERFORMANCE USING DATA MINING

Jashanpreet Kaur[a]*, Yogesh Kumar[b],
a Computer Engineering, Gurukul Vidyapeeth Group of Institutions, Banur
b Computer Engg, Gurukul Vidyapeeth Group of Institutions, Banur

ABSTRACT

Management of huge amount of data has always been a matter of concern. With the increase in awareness towards education, the amount of data in educational institutes is also increasing. The increasing growth of educational databases, have given rise to a new field of data mining, known as Educational Data Mining. The present paper shows the performance of GPSO algorithm on the data mining. In the present paper review of literature on the data mining gives results of experimental work on basics of academic student performance. It has been observed that the GPSO algorithm gives better results than existing algorithm. Future scope of the data mining is also discussed in this paper.

Keywords:* *Data Mining, Academic Performance, Genetic, PSO and Education.

1. Introduction

Text mining is the process to find the valuable information and related stuff from the messages, threads, discussion forums and other sources attached to the social media. The text analytics is the branch of data science specialization, where the text data is analyzed by using the various text processing methods. The text mining methods require the 'High Quality' combinations for the various techniques altogether for the discovery of text data using connections between the keywords & phrases, estimating their novelty and dynamic methodology.

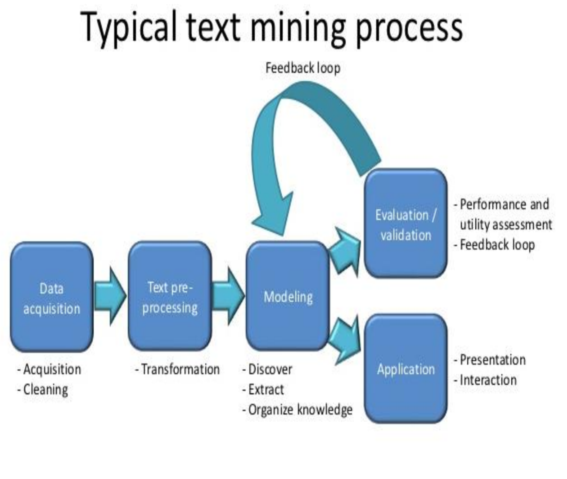

Figure 1.1: Typical models for the general text mining algorithms [1]

Organizations have been approaching servitisation in an unstructured fashion. This is partially because there is insufficient understanding of the different types of Product-Service offerings. Therefore, a more detailed understanding of Product-Service types might advance the collective knowledge and assist organizations that are considering a servitisation strategy [2].

The statistical methods based upon the feature description, dimensionality reduction and pattern discovery, which has been described in the various computing methods, which involves the classification methods such as artificial neural networks (ANN), support vector machine (SVM), Naïve Bayes (NB), Random Forest (RF), Co-Forest (CF), Simple Regression, Linear Regression, Logistic Regression, etc. [3].

News Classification

News classification is method of mechanically classifying the news knowledge into the varied classes on the premise of knowledge patterns, associations, changes, anomalies and important structures, from great deal of knowledge keep in news information or different information repositories.

The news classification is that the branch of the info mining techniques and listed beneath the text mining and opinion mining classes. The info mining has popularly treated as equivalent word of data discovery in information, though some researchers read data processing as a vital step of data discovery.

A data discovery method for the net new classification consists of the repetitious sequence of following step:

- Data improvement that handles blatant, erroneous, missing or inapplicable data.
- Data integration, wherever multiple, heterogeneous information supply is also integrated into one.
- Data choice, wherever information relevant to analysis task are retrieved from information.
- Data transformation, wherever information are remodeled or consolidated into from acceptable for mining by acting combination operations.
- Data mining, that is crucial method wherever intelligent strategies are applied so as to extract information patterns.
- Pattern analysis, that is to spot the actually attention-grabbing pattern a represent data supported some powerfulness live.
- Knowledge presentation, wherever data and data illustration techniques are accustomed gift the mined data to the user.

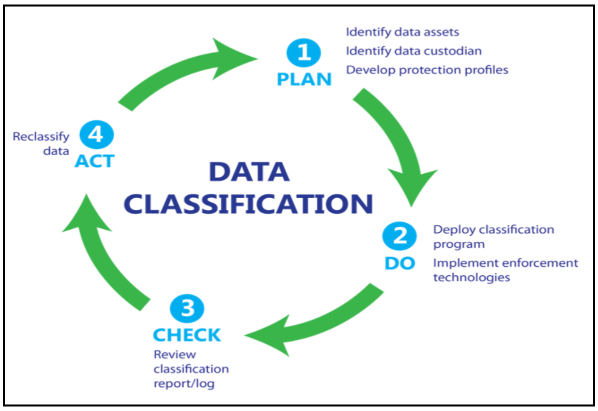
1 PLAN
Identify data assets
Identify data custodian
Develop protection profiles
2 DO
Deploy classification program
Implement enforcement technologies
3 CHECK
Review classification report/log
4 ACT
Reclassify data
DATA CLASSIFICATION

Figure 1.6: A Best Fit Data Classification model for Text Classification [7]

2. Literature Review

Aktepe, Adnan et. al.(2015) has worked on the client satisfaction and loyalty analysis with classification algorithms and structural equation modeling. Businesses will maintain their effectiveness as long as they need glad and constant customers. Client relationship management provides provides vital blessings for corporations particularly i n gaining fight. So as to succeed in these objectives primarily corporations got to establish and analyze their customers. In this respect, effective communication and commitment to customers and dynamical market conditions is of nice importance to extend the amount of satisfaction and loyalty. To guage this example, level of client satisfaction and loyalty ought to be measured properly with a comprehensive approach. In this study, customers are investigated in four main teams in keeping with their level of satisfaction and loyalty with a criteria and cluster primarily based analysis with a brand new methodology. The authors have used classification algorithms in programming code and Structural Equation Modeling (SEM) with LISREL tools along to investigate the impact of every satisfaction and loyalty criteria in an exceedingly satisfaction–loyalty matrix and extend the client satisfaction and loyalty post-analysis analysis bridging the gap in this field of research [6].

Gaiardelli, Paolo et. al. (2014) has worked towards A classification model for product-service offerings. in this paper, the authors have developed a comprehensive model for classifying ancient and inexperienced Product-Service offerings, therefore combining business and inexperienced offspring in an exceedingly single model. They need conjointly represented the model building method and its utilization in an exceedingly case study. The model reveals the assorted ancient and inexperienced choices out there to corporations and identifies the way to vie between services; it permits servitisation positions to be known specified a corporation might track its journey over time. Finally it fosters the introduction of innovative Product-Service Systems as promising business models to deal with environmental and social challenges [2].

Lu, Ning et. al. (2014) has developed the customer churn prediction model in telecom industry using boosting. This research conducts a real-world study on customer churn prediction and proposes the use of boosting to enhance a customer churn prediction model. Unlike most research, that uses boosting as a method to boost the accuracy of a given basis learner, this paper tries to separate customers into two clusters based on the weight assigned by the boosting algorithm. As a result, a higher risk customer cluster has been identified. Logistic regression is used in this research as a basis learner, and a churn prediction model is built on each cluster respectively [3].

N. Kim et. al. (2012) has worked towards the development of uniformly subsampled ensemble (use) for churn management. The present paper explores the possible application of a new ensemble model. The model, which is based on multiple SVM classifiers, is employed to address churner identification problems in the mobile telecommunication industry, a sector in which the role of customer retention program becomes increasingly important due to its very competitive business environment. In particular, the current study introduces a uniformly subsampled ensemble (USE) model of SVM classifiers, not only to reduce the computational complexity of large-scale data, but also to boost the reliability and accuracy of calibrated models on data sets with highly skewed class distributions [11].

W. Verbeke et. al. (2012) has focused on the new insights into churn prediction in the telecommunication sector: a profit driven data mining approach. in the first part of this paper, a novel, profit centric performance measure is developed, by calculating the maximum profit that can be generated by including the optimal fraction of customers with the highest predicted probabilities to attrite in a retention campaign. The novel measure selects the optimal model and fraction of customers to include, yielding a significant increase in profits compared to statistical measures. In the second part an extensive benchmarking experiment is conducted, evaluating

various classification techniques applied on eleven real-life data sets from telecom operators worldwide by using both the profit centric and statistically based performance measures [4].

M. Owczarczuk et. al. (2010) has worked on the churn models for prepaid customers in the cellular telecommunication industry using large data marts. In this article, the authors have tested the usefulness of the popular data mining models to predict churn of the clients of the Polish cellular telecommunication company. When comparing to previous studies on this topic, this research is novel in the following areas: (1) the authors dealt with prepaid clients (previous studies dealt with postpaid clients) who are far more likely to churn, are less stable and much less is known about them (no application, demographical or personal data), (2) they have 1381 potential variables derived from the clients' usage (previous studies dealt with data with at least tens of variables) and (3) they have tested the stability of models across time for all the percentiles of the lift curve – their test sample is collected six months after the estimation of the model [12].

J. Burez and D. V. Poel (2009) has worked towards handling class imbalance in customer churn prediction. In this paper, the authors have investigated the increase in performance of sampling (both random and advanced under-sampling) and two specific modeling techniques (gradient boosting and weighted random forests) compared to some standard modeling techniques. AUC and lift prove to be good evaluation metrics. AUC does not depend on a threshold, and is therefore a better overall evaluation metric compared to accuracy. Lift is very much related to accuracy, but has the advantage of being well used in marketing practice [13].

K. Coussement and D.V. Poel (2008) has worked on the churn prediction in subscription services: an application of support vector machines while comparing two parameter-selection techniques. This study applies support vector machines in a newspaper subscription context in order to construct a churn model with a higher predictive performance. Moreover, a comparison is made between two parameter-selection techniques, needed to implement support vector machines. Both techniques are based on grid search and cross-validation. Afterwards, the predictive performance of both kinds of support vector machine models is benchmarked to logistic regression and random forests. This study shows that support vector machines show good generalization performance when applied to noisy marketing data. Nevertheless, the parameter optimization procedure plays an important role in the predictive performance. The authors have shown that only when the optimal parameter selection procedure is applied, support vector machines outperform traditional logistic regression, whereas random forests outperform both kinds of support vector machines [9].

3. Conclusions

Following points can be concluded from the literature review.

1. The results show that the Hybrid GPSO approach delivers a significant performance for the data sets.
2. Proposed Hybrid (GPSO) approach improves the performance by featuring the faster convergence and high computational speed than the individual comparison.
3. The Hybrid GPSO algorithm merges the capability of fast convergence of the PSO algorithm with the competency of ease to exploit preceding solution of GA for eliminating the early convergence.
4. PSO works proficiently on large datasets by minimizing the time, utilizing the less parameter and gives the better performance than the GA.

4. Research Scope

For future work, we would like to refine our work by taking more number of examples set and come up with more accuracy and other techniques to help students in their educational careers.

REFERENCES

[1] Kumar, B. Shravan, and Vadlamani Ravi. "A survey of the applications of text mining in financial domain." *Knowledge-Based Systems* 114: 128-147, (2016).

[2] Gaiardelli, Paolo, Barbara Resta, Veronica Martinez, Roberto Pinto, and Pavel Albores. "A classification model for product-service offerings." *Journal of cleaner production* 66: 507-519, (2014).

[3] Lu, Ning, Hua Lin, Jie Lu, and Guangquan Zhang. "A customer churn prediction model in telecom industry using boosting." *Industrial Informatics, IEEE Transactions on* 10, no. 2n: 1659-1665, (2014).

[4] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens, "New Insights into Churn Prediction in the Telecommunication Sector: A Profit Driven Data Mining Approach," *European Journal of Operational Research*, Vol. 218, No. 1, Apr. , pp. 211-229, (2012).

[5] Jurafsky, Dan, and James H. Martin. *Speech and language processing*. Vol. 3. Pearson, (2014).

[6] Aktepe, Adnan, Süleyman Ersöz, and Bilal Toklu. "Customer satisfaction and loyalty analysis with classification algorithms and structural equation modeling." *Computers & Industrial Engineering* 86 : 95-106, (2015).

[7] Taylor, Michael J., Chris McNicholas, Chris Nicolay, Ara Darzi, Derek Bell, and Julie E. Reed. "Systematic review of the application of the plan–do–study–act method to improve quality in healthcare." *BMJ quality & safety* : bmjqs, (2013).

[8] Kaisler, Stephen, Frank Armour, J. Alberto Espinosa, and William Money. "Big data: Issues and challenges moving forward." In *System sciences (HICSS), 2013 46th Hawaii international conference on*, pp. 995-1004. IEEE, (2013).

[9] K. Coussement and D. V. Poel, "Churn Prediction in Subscription Services: An Application of Support Vector Machines while Comparing Two Parameter-Selection Techniques," *Expert Systems with Applications*, Vol. 34, No. 1, pp. 313-327, (Jan. 2008).