RECENT ADVANCES AND EMERGING APPLICATION CHALLENGES OF FEATURE SELECTION FOR VIRTUAL LEARNING

[a]DINESH MAVALURU, [b]R MYTHILY, [c]AZATH MUBARAKALI

[a] *College of Computing and informatics, Saudi Electronic University, Abha, Saudi Arabia*
[b] *Department of Information Technology, B.S. Abdur Rahman Crescent University, Chennai*
[c] *College of Computer Science, King Khalid university, Abha, Saudi Arabia*
Email: [a]d.mavaluru@seu.edu.sa, [b]mythily@bsauniv.ac.in, [c]mailmeazath@gmail.com

Abstract: The use of data mining procedures for decision making has been increasing in the past decades in identifying students' behavior and use of knowledge available in university is a great concern to the e-learning or virtual learning higher education system is increased and the big data phenomenon is also growing from the dawn of time up to 2003 people produced a total of 5 exabytes of data and by 2008 this figure had increased to 14.7 exabytes. These days 5 exabytes of data is produced every 2 days. Its acceptance can be credited to improved applying data mining algorithms, increased performance in different educational systems, and results which can be measured and applied for decision making. The current use of data mining methods to analyze various types of data has shown great advantages in various application domains and particularly in education domain. While some data sets need little preparation to be mined, whereas others, high-dimensional data sets, need to be preprocessed in order to be mined due to the complexity and inefficiency in mining high dimensional data processing. Feature selection or attribute selection is one of the techniques used for dimensionality reduction. Previous research has shown that data mining results can be improved in terms of accuracy and efficacy by selecting the attributes with most significance. A teacher or an artificial instructor, in an education system is embedded with an intelligent tutoring system, is interested in predicting the performance of their students to better adjust the educational materials and strategies throughout the learning process. This research work analyzes different student's performance in different categories of their entire learning process of the course. The purpose of this research work is to find a model that better classifies existing students to improve their grades and the teachers to improve their teaching methodology process to get better results in virtual learning or e-learning. In this research work few different feature selection methods such as Correlation Based Feature Selection, Information Gain, Relief-F, Wrapper, and Hybrid methods were used to reduce the number of features or attributes in the data sets are compared. The data sets with the attributes selected were run through the popular classification algorithms like Decision Trees, k-Nearest Neighbor, and Support Vector Machines, and the results are compared and analyzed.

Key words: virtual learning, Support Vector Machines, e-learning

1 Introduction

Nowadays, as computing power rises and changes cheaper, a new trend playing a significant role is to mine the data for unknown patterns and to extract data that is formerly unknown that may be useful to improve the decision making process (H. Liu, H. Motoda, 2007).

Knowledge Discovery in Data (KDD) - The field of Knowledge Discovery in Databases (KDD) has grown in the past several decades as more industries find a need to find valuable information in their databases. The KDD process (Fayyad, Piatetsky-Shapiro, & Smyth, 1996) is broken down into five phases:

- Selection - The first stage consists of collecting data from existing sources to be used in the discovery process. The data may come from single or multiple sources. This may be the most important stage since the data mining algorithms will learn and discover from this data.
- Preprocessing - The main goal of this stage is to make the data more reliable. Methods used to account for missing data are analyzed and implemented. Dealing with noisy data or outliers is also part of this stage.
- Transformation - Now that we have reliable data we can make it more efficient. The uses of feature selection methods to reduce dimensionality and feature extraction to combine features into new ones are implemented at this point. Discretization of numerical attributes and sampling of data are also common tasks performed in this stage.
- Data Mining - Before the data is mined, an appropriate data mining task such as classification, clustering, or regression needs to be chosen. Next, one or several algorithms specific to the task, such as decision trees for classification, must be properly configured and used in the discovery of knowledge. This process is repeated until satisfying results are obtained.
- Evaluation - The last step is the interpretation of results in respect to pre-defined goals. A determination is made if the appropriate data mining model was chosen. All steps of the process are reviewed and analyzed in terms of the final results. This study concentrates in two critical areas of the KDD process; transformation by reducing the feature set and the data mining process.

2 Literature Review

In recent years, most creativities and establishments have stored large amounts of data in a regular way, but without a clear idea of its potential helpfulness. In addition, the growing acceptance of the Internet has produced data in many different presentations (text, multimedia, etc.) and from many different foundations (systems, sensors, mobile devices, etc.). To be able to extract useful information from all these data, we require new study and processing tools. Most of these data have been produced in the last few years as we continue to produce quintillions of bytes daily (Z.A. Zhao, H. Liu, 2011). Big data large volumes and ultrahigh dimensionality is now a recurring feature of many machine learning application fields, such as text mining and information retrieval (L. Yu, H. Liu, 2004).

In (C. Boutsidis, P. Drineas, M.W. Mahoney, 2009.), for example, conducted a study of a collective email junk riddling task with loads of unique features, whereas the study in (V. Roth, T. Lange, 2003) was based on a wide range of artificial and real world datasets of tens of million data points with features. The growing size of datasets raises an interesting challenge for the research community. In (R. Leardi, A. Lupiáñez González, 1998) the research ''our task is to find a needle in a haystack, teasing the relevant information out of a vast pile of glut''. Ultrahigh dimensionality implies massive memory requirements and a high computational cost for training. Generalization capacities are also undermined by what is known as the ''curse of dimensionality''. According to research in (D. Paul, E. Bair, T. Hastie, R. Tibshirani, 1595), Bellman coined this colorful term in 1957 to describe the difficulty of optimization by exhaustive enumeration on product spaces. This term refers to various phenomena that arise when analyzing and organizing data in high-dimensional spaces (with hundreds or thousands of dimensions) that do not occur in low-dimensional settings. A dataset is usually represented by a matrix where the rows are the recorded instances (or samples) and the columns are the attributes (or features) that represent the problem at hand. In (M. Pal, G.M. Foody, 2010) order to tackle the dimensionality problem, the dataset can be summarized by finding ''narrower'' matrices that in some sense are close to the original. Since these narrower matrices have a smaller number of samples and/or features, they can be used much more efficiently than the original matrix. The process of finding these narrow matrices is called dimensionality reduction.

Feature extraction is a dimensionality reduction technique that addresses the problem of finding the most compact and informative set of features for a given problem so as to improve data storage and processing efficiency. Feature extraction (H. Liu, R. Setiono, 1995) is decomposed into the steps of construction and selection. Feature construction methods complement human expertise in converting ''raw'' data into a set of useful features using preprocessing transformation procedures such as standardization, normalization, discretization, signal enhancement, and local feature extraction. Some

construction methods do not alter space dimensionality, while others enlarge it, reduce it or both.

It is crucial not to lose information during the feature construction stage. In (R.O. Duda, P.E. Hart, D.G. Stork, 1999) recommend that it is always better to err on the side of being too inclusive rather than run the risk of discarding useful information. Adding features may seem reasonable but it comes at a price: the increase in the dimensionality of patterns incurs the risk of losing relevant information in a sea of possibly irrelevant, noisy or redundant features. The goal of (J.R. Quinlan, 1986) feature selection methods is to reduce the number of initial features so as to select a subset that retains enough information to obtain satisfactory results. In a society that needs to deal with vast quantities of data and features in all kinds of disciplines, there is an urgent need for solutions to the indispensable issue of feature selection. To understand the challenges that researchers face, the next section will briefly describe the origins of feature selection and recent contributions.

Feature selection is defined as the process of detecting relevant features and discarding irrelevant and redundant features with the goal of obtaining a subset of features that accurately describe a given problem with a minimum degradation of performance (Kononenko, 1994). Theoretically, having a large number of input features might seem desirable, but the curse of dimensionality is not only an intrinsic problem of high-dimensionality data, but more a joint problem of the data and the algorithm being applied. For this reason, researchers began to select feature in a pre-processing phase in an attempt to convert their data into a lower-dimensional form.

The first research into feature selection dates back to the 1960s (Guyon, 2006). In (Y. Zhai, Y. Ong, I. Tsang, 2014) used a general parametric model to study the accuracy of a Bayesian classifier as a function of the number of features, concluding as follows: ‘‘Dimension selection, reduction and grouping are not proposed as developed techniques. Rather, they are illustrative of a agenda for further investigation’’. Since then, research into feature selection has posed many challenges, with some researchers highly skeptical of progress. In (M. Tan, I.W. Tsang, L. Wang, 2014), stated: ‘‘If variable elimination has not been sorted out after two decades of work assisted by high-speed computing, then perhaps the time has come to move on to other problems’’. In the 1990s, notable advances were made in feature selection used to solve machine learning problems (K. Weinberger, A. Dasgupta, J. Langford, A. Smola, J. Attenberg,2009). Nowadays, feature selection is acknowledged to play a crucial role in reducing the dimensionality of real problems, as evidenced by the growing number of publications on the issue (D.L. Donoho et al., 2000).

The new feature selection methods developed in the last few decades classified as filter, wrapper or embedded methods are based upon the relationship between the feature selection algorithm and the inductive learning method used to infer the model (R. Bellman, 1957). Feature selection methods can also be classified according to individual evaluation and subset evaluation approaches (Guyon, A. Elisseeff, 2003) the former also known as feature ranking assesses individual features by assigning them weights according to relevance, whereas the latter produces candidate feature subsets based on a specific search strategy which are subsequently evaluated by some measure.

Given its ability to enhance the performance of learning algorithms, feature selection has attracted increasing interest in the field of machine learning, in processes such as clustering (B. Bonev, 2010), regression and classification, whether supervised or unsupervised. Of the numerous feature selection algorithms available, several have become particularly popular among researchers. These widely used methods are part of the state of the art in feature selection. Multivariate methods generally tend to obtain better results than univariate approaches, but at a greater computational cost. There is no one-size-fits all method, as each is more suitable for particular kinds of problems. In a previous work (G. Hughes, 1968), we reviewed the performance of some of these state-of-the-art algorithms in an artificial controlled scenario, checking their efficiency in tackling problems such as redundancy between features, non-linearity, noise in inputs and in output and a higher number of features than samples (as happens with DNA microarray classification).

In (A.J. Miller, 1984) DNA microarrays are used to collect information on gene expression differences in tissue and cell samples that could be useful for disease diagnosis or for distinguishing specific types of tumors. The sample size is usually small (often less than 100 patients) but the raw data measuring the gene expression enmasse may have from 6000 to 60000 features. In this scenario, feature selection inevitably became an indispensable preprocessing step.

Image classification has become a popular research field given the demand for efficient ways to classify images into categories. The numerical properties (A.L. Blum, P. Langley, 1997) of image features are usually analyzed to determine to which category they belong. With recent advances in image capture and storage and Internet technologies, a vast amount of image data has become available to the public, from smartphone photo collections to websites and even video databases. Since image processing usually requires a large amount of computer memory and power, feature selection can help reduce the number of features needed in order to be able to correctly classify the image.

The filter method of feature selection is a common choice, mainly due to its low computational cost compared to the wrapper or embedded methods. In (M. Dash, H. Liu, 1997) presented a method based on the physical meaning of the generalized Fisher criterion in order to choose the most discriminative features for recognition. In (R. Kohavi, G.H. John, 1997) proposed a novel method for choosing a subset of original features containing the most essential information; called principal feature analysis (PFA), it is similar to principal component analysis (PCA) methods. Even after this preprocessing step, the number of possible words in a document may still be high, so feature selection is paramount.

A number of techniques have been developed and applied to this problem in recent years. In (V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, J.M. Benítez, F. Herrera, 2014) proposed a novel feature selection metric, called bi-normal separation (BNS), which is a useful heuristic for increased scalability when used with wrapper techniques of text classification. In (P.P. Ohanian, R.C. Dubes, 1992) applied several novel feature selection methods to clustered data, while in proposed an unsupervised feature selection strategy that theoretically guarantees the generalization power of the resulting classification function with respect to the classification function based on all the features.

As stated at the opening of this research work, constant advances in computer based technologies have enabled scholars and engineers to collect data at an increasingly fast pace. To address the challenge of analyzing these data, feature selection becomes an imperative preprocessing step that needs to be adapted and improved to be able to handle high dimensional data. We have highlighted the need for feature selection and discussed recent contributions in several different application areas. However, in the new big data scenario, an important number of challenges are emerging, representing current research trends in virtual learning also. The feature selection process is shown in figure 1.

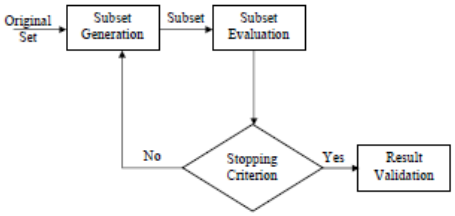

Figure 1: Feature selection process

3 Methodology

This study is a comparative analysis of feature selection methods on classification systems in a real domain setting. As with any data mining exercise, before the data are mined, several key steps need to be performed. These steps, referred to as the preprocessing stage, will account for dealing with missing values, balancing data, discretizing or normalizing attributes depending on which algorithm is used, and finally minimizing the dimensionality of the data set by reducing the number of features with different feature selection methods. The CRISP-DM framework breaks down the data mining process into six phases:

- Understanding the business process and determining the ultimate data mining goals
- Identifying, collecting, and understanding key data sources
- Preparing data for data mining
- Selecting which modeling technique to use
- Evaluating and comparing results of different models against the initial goals
- Deploying Model

Big Data provides teachers with highly relevant information, but we must not forget that it also brings benefits to students. For instance, if one of the benefits is that educators are able to produce better teaching materials to meet their learning patterns, the student will benefit from it as well, that is, if a student is presented with the information in a meaningful way he or she is going to be more motivated in their learning process. Students and participants in eLearning courses have much to gain from the benefits that the information from Big Data provides us. Next, a case study is presented based on the large amount of data collected using a virtual learning environment, where it is shown how eLearning and Big Data form a binomial that must be considered in the future to improve the knowledge society.

The use of these LMS in education eLearning involves generation of a large and complex set of data. These data obtained from the use by many users of the different technological tools in a virtual learning platform can be drawn great benefits to improve eLearning education. How can one profit in the learning context from Big Data? Everyone interested in eLearning education wants to find the answer to this question. Ambrose pointed to the company in collaboration with IBM Skillsoft how big data can help create a learning experience more personalized and adaptive based on real information about each student. So, the study on the use of personal data can be applied perfectly to eLearning. It offers us the opportunity to learn more about our students and their behavior patterns in a way not known before. Moreover, we can use this knowledge to develop eLearning courses really geared to the needs of our students through scenarios that meet their real situations.

Big Data offers us the opportunity to provide students with more efficient courses and more effective on-line learning modules which are attractive and informative. The reasons why large amounts of data can revolutionize the industry of eLearning are:

It allows eLearning professionals design more customized eLearning courses. If you give eLearning professionals the opportunity to learn what works best for their students, in terms of content and delivery, this will allow create more personalized and attractive eLearning courses, thus providing high quality and meaningful learning experience.

It provides counseling on effective online strategies. ELearning big data can give us visions of which eLearning strategies work and which do not. It allows the monitoring of students´ patterns. With large data eLearning, educators gain the ability to track the students throughout the learning process. This helps them to find out patterns that not only will allow them to learn more about the behavior of each pupil, but also of the group of students as a whole.

It enables the possibility to expand our understanding of the process of virtual learning. It is essential that eLearning professionals get to know how students learn and acquire knowledge. Big Data gives us the opportunity to gain a deeper understanding of the process of eLearning and how students are responding to the eLearning courses. This information can be used to design new learning methods. Big Data provides teachers with highly relevant information, but we must not forget that it also brings benefits to students. For instance, if one of the benefits is that educators are able to produce better teaching materials to meet their learning patterns, the student will benefit from it as well, that is, if a student is presented with the information in a meaningful way he or she is going to be more motivated in their learning process. Students and participants in eLearning courses have much to gain from the benefits that the information from Big Data provides us. Next, a case study is presented based on the large amount of data collected using a virtual learning environment, where it is shown how eLearning and Big Data form a binomial that must be considered in the future to improve the knowledge society.

This researched used this framework and provided a structured way to conduct the experiments used in this comparative study. Therefore, it improved the validity and reliability of the final results.

Big Data offers us the opportunity to provide students with more efficient courses and more effective on-line learning modules which are attractive and informative. The reasons why large amounts of data can revolutionize the industry of eLearning are:

- It allows eLearning professionals design more customized eLearning courses. If you give eLearning professionals the opportunity to learn what works best for their students, in terms of content and delivery, this will allow create more personalized and attractive eLearning courses, thus providing high quality and meaningful learning experience.
- It provides counseling on effective online strategies. E-learning big data can give us visions of which eLearning strategies work and which do not. It allows the monitoring of students´ patterns. With large data eLearning, educators gain the ability to track the students throughout the learning process. This helps them to find out patterns that not only

will allow them to learn more about the behavior of each pupil, but also of the group of students as a whole.
- It enables the possibility to expand our understanding of the process of virtual learning. It is essential that eLearning professionals get to know how students learn and acquire knowledge. Big Data gives us the opportunity to gain a deeper understanding of the process of eLearning and how students are responding to the eLearning courses. This information can be used to design new learning methods.

4 Conclusion

Feature selection has been usually used as a preprocessing step that condenses the extents of a problematic and advances classification precision. The need for this kind of technique has improved intensely in recent years in order to cope with situations categorized by both a high number of input features and/or of models. In other words, the big data bang now has the added problem of big dimensionality. This research work assessed the main need for feature selection and momentarily revised the most popular feature selection methods and some typical applications that are used for virtual learning in higher education. While feature selection may well be one of the improved preprocessing methods, it is vital not to overlook the factors affecting feature selection choices. For illustration, it is important to choose a satisfactory discretization method, given that some feature selection approaches especially those from the information theory field were developed to work with separate data. Certainly, it has been established that the choice of method affects the results of the feature selection process in virtual learning.

Literature:

1. A.J. Miller, *Selection of subsets of regression variables*, J. Roy. Stat. Soc. Ser. A (Gener.) (1984) 389–425.
2. A.L. Blum, P. Langley, *Selection of relevant features and examples in machine learning*, Artif. Intell. 97 (1) (1997) 245–271.
3. B. Bonev, *Feature Selection based on Information Theory*, Universidad de Alicante, 2010.
4. C. Boutsidis, P. Drineas, M.W. Mahoney, *Unsupervised feature selection for the k-means clustering problem*, Adv. Neural Inform. Process. Syst (2009) 153–161.
5. D.L. Donoho et al., *High-dimensional data analysis: the curses and blessings of dimensionality*, AMS Math. Challenges Lect. (2000) 1–32.
6. D. Paul, E. Bair, T. Hastie, R. Tibshirani, *''Preconditioning'' for feature selection and regression in high-dimensional problems*, Ann. Stat. (2008) 1595–1618.
7. Guyon, *Feature Extraction: Foundations and Applications*, vol. 207, Springer, 2006.
8. Guyon, A. Elisseeff, *An introduction to variable and feature selection*, J. Mach. Learn. Res. 3 (2003) 1157–1182.
9. G. Hughes, *On the mean accuracy of statistical pattern recognizers*, IEEE Trans. Inform. Theory 14 (1) (1968) 55–63.
10. H. Liu, H. Motoda, *Computational Methods of Feature Selection*, CRC Press, 2007.
11. H. Liu, R. Setiono, Chi2: *feature selection and discretization of numeric attributes, in: Tools with Artificial Intelligence*, 1995. Proceedings., Seventh International Conference on, IEEE, 1995, pp. 388–391.
12. J.R. Quinlan, *Induction of decision trees*, Mach. Learn. 1 (1) (1986) 81–106.
13. Kononenko, *Estimating attributes: analysis and extensions of relief, in: Machine Learning: ECML-94*, Springer, 1994, pp. 171–182.
14. K. Weinberger, A. Dasgupta, J. Langford, A. Smola, J. *Attenberg, Feature hashing for large scale multitask learning*, in: Proceedings of the 26th Annual International Conference on Machine Learning, ACM, 2009, pp. 1113–1120.
15. L. Yu, H. Liu, *Efficient feature selection via analysis of relevance and redundancy*, J. Mach. Learn. Res. 5 (2004) 1205–1224.
16. M. Pal, G.M. Foody, *Feature selection for classification of hyperspectral data by svm, IEEE Trans.* Geosci. Remote Sens. 48 (5) (2010) 2297–2307.
17. M. Tan, I.W. Tsang, L. Wang, *Towards ultrahigh dimensional feature selection for big data*, J. Mach. Learn. Res. 15 (1) (2014) 1371–1429.
18. M. Dash, H. Liu, *Feature selection for classification*, Intell. Data Anal. 1 (3) (1997) 131–156.
19. P.P. Ohanian, R.C. Dubes, *Performance evaluation for four classes of textural features*, Pattern Recognit. 25 (8) (1992) 819–833.
20. R. Leardi, A. Lupiáñez González, *Genetic algorithms applied to feature selection in pls regression: how and when to use them*, Chemomet. Intell. Lab. Syst. 41 (2) (1998) 195–207.
21. R.O. Duda, P.E. Hart, D.G. Stork, *Pattern Classification*, John Wiley & Sons, 1999.
22. R. Bellman, *Dynamic Programming*, vol. 18, Princeton UP, Princeton, NJ, 1957.
23. R. Kohavi, G.H. John, *Wrappers for feature subset selection*, Artif. Intell. 97 (1) (1997) 273–324.
24. V. Roth, T. *Lange, Feature selection in clustering problems*, Adv. Neural Inform. Process. Syst. (2003).
25. V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, J.M. Benítez, F. Herrera, *A review of microarray datasets and applied feature selection methods*, Inform. Sci. 282 (2014) 111–135.
26. Y. Zhai, Y. Ong, I. Tsang, *The emerging 'big dimensionality'*, Comput. Intell. Mag., IEEE 9 (3) (2014) 14–26.
27. Z.A. Zhao, H. Liu, *Spectral Feature Selection for Data Mining*, Chapman & Hall/CRC, 2011.

Primary Paper Section: A

Secondary Paper Section: AO, AM