Indian Journal of Science and Technology, *Vol 8(S7), 135–140, April 2015*

ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

A Novel Clustering based Feature Selection for Classifying Student Performance

R. Sasi Regha[1*] and R. Uma Rani[2]

[1]Department of Computer Science, SSM College of Arts & Science, PinCode-638 183, Tamilnadu, India; sasiresearchphd@gmail.com

[2]Department of Computer Science, Sri Sarada College for Women, PinCode-638 016, Tamilnadu, India; umaraniresearch@gmail.com

Abstract

The main intent of this work is to identify and eradicate the irrelevant as well as redundant features that are used to improve the accuracy of student performance classification.

In this article, a novel technique is introduced for feature or attributes selection purpose called as Non-negative Matrix Factorization Clustering based Feature Selection (NMFCFS). NMFCFS uses symmetric uncertainty (SU) estimation.

The performance of this work is evaluated by using the student dataset that includes collection of students' information from various colleges. For analyzing the performance of this work, comparative evaluation is performed between the classifiers (in the experiment, Prism and J48 is taken) without feature selection and classifiers with the NMFCFS. The experimental result shows that NMFCFS approach is attaining higher accuracy rate i.e., 97.8 %. The proposed feature selection method is highly efficient when compared to other schemes.

The findings demonstrate that the proposed method has high performance of the students' failure and dropout prediction. In other words, this can improve the accuracy of the classification result.

Keywords: Educational Data Mining (EDM), Feature Selection, Non-negative Matrix Factorization, Symmetric Uncertainty (SU) and Classification

1. Introduction

Educational Data Mining (EDM) helps to understand productively the students and the settings which they learn in, and it is an independent investigation area in recent years[1]. One of the main areas in EDM is development of student models that would evaluate student's individuality or academic performances in educational institutions.

Prediction methods can be used to examine the features of a model that are significant for the prediction and also to provide important information. Various prediction methods are suggested for deciding the performance of the students and different classifiers performance is analyzed like Decision Tree, Bayesian Network, and other classification methods are also investigated in[2-5]. But these methods only provide the classifier accuracy without using any feature methods. The performance of the prediction method is tremendously depends on the option of selection of most pertinent features from the list which is used in student data set. This can be acquired by using the different feature selection methods on data set. The percentage of accuracy is typically not selected for classification, as values of accuracy are extremely according to the base rates of various classes. In addition to that, different factors affect the performance of data mining algorithms for a given task. The value of the data is defined as the parameter of analyzing the information is irrelevant or superfluous, or the data is noisy and undependable, then knowledge discovery during training is more complicated. Generally, attribute subset selection is the process of identifying and removing both irrelevant and redundant features as hard as possible.

**Author for correspondence*

Learning approaches vary for emphasis they place on attribute selection. These approaches are used to categorize the novel examples by acquiring the nearest accumulated training example, using all the presented features in its distance computations. In the other extreme, the algorithms are concentrated on the relevant features without considering the irrelevant features. So, attribute selection is a main concept for improving the accuracy. The reduction in the element of the data diminishes the size of the hypothesis space and allows techniques to operate more quickly and efficiently.

Feature selection is an important research in the pattern recognition, machine learning, statistics and data mining communities[6,7]. The importance of an individual feature is identified by ranking the features and removing the probable associations. The ranking methods are based on the statistics, information theory, or on some functions of output result of classifier[8]. Different justifications for the use of filters for subset selection have been presented[9,10] and filters are reasonably faster than wrappers. Furthermore, the above mentioned methods are only concentrated on the relevant features and do not considering the redundant features. To overcome these issues, the NMFCFS is presented which is considering both removal of irrelevant features as well as redundant features in the dataset. Additionally, NMFCFS is not only gives the smaller subsets of features and also used to enhance the performance of the classifiers. Furthermore, the proposed NMFCFS does not limit to some specific types of data.

Rozita Jamili Oskouei et. al,[11] intends to study the discrepancy effect in the Web technology in various scopes of students' academic performance co-curricular, extra-curricular and non-academic activities. The average time spent for an internet for a particular day and class of visited Websites by them with the impacts of the internet usage behavior in the academic performance (CPI) and other co-curricular and extra-curricular activities. Manickasankari[12] suggested a class concept hierarchy method for developing and recovering information regarding to the e-learning domain via PROTÉGÉ tools and also shows RDFS and OWL based samples method. This creates the result such as number of course material arranged by instructor, registered students and number of exam, quiz, and assignment conduct by particular instructor by using SPARQL query language.

2. Non-negative Matrix Factorization Clustering based Feature Selection Technique

The proposed system introduces a novel technique for feature or attributes selection purpose called as Non-negative matrix factorization Clustering based Feature Selection. This NMFCFS is defined as a technique for feature selection based on the Non-negative matrix factorization Clustering process. A novel technique is introduced for feature or attributes selection purpose called as NMFCFS `that compiled of the two related components of irrelevant and also redundant feature removal. NMFCFS uses symmetric uncertainty (SU) estimation for removing the irrelevant features or in other words obtaining the relevant features. After finding the relevant features, determine redundant features present in the relevant features. For this purpose, a novel Non-negative matrix factorization based clustering approach is used. After feature selecting process, two effective classification techniques i.e., Prism and J48 are used for predicting the student performance.

2.1 Irrelevant Feature Removal

Relevant features are highly associated with the target concepts so that this is significant for best subset. But the redundant features are not suitable for best subset because their values are completely correlated with each other. The notion of feature relevance is typically in terms of feature-target concept association. Mutual information calculates how much the distribution of the feature values and the target classes that is different from statistical independence. It is a non-linear evaluation of association that is computed between the feature values and target classes. The symmetric uncertainty is computed as the mutual information by standardizing it to the entropies of feature values and target classes, and this method is valuable for examine the prosperous of features for classification. As a result, symmetric uncertainty is used as the measure of the association between the feature values and target concepts. The symmetric uncertainty is defined in the equation (1):

$$SU(X,Y) = \frac{2 \times Gain(X|Y)}{H(X) + H(Y)} \quad (1)$$

In the equation (1), H(X) denotes the entropy of a discrete random variable X. Gain (X|Y) is defined as the amount by which the entropy of Y decreases. It reproduces

the extra information about Y provided by X and it is represented as information gain.

After finding this SU measured value between the feature and target class, the comparison of the SU value with predefined threshold value takes place. Those features are higher SU measured value with target class when compared to the predefined threshold value which is considered as the relevant features. Also, those features having lesser SU measured value with the target class when compared to the predefined value is considered as the irrelevant features and these features are eliminated. In the end of this process, the relevant feature from the collection of features in the data is attained.

2.2 Redundant Features Removal

After acquiring the relevant features, the redundant features are also presented in the relevant features. So that, the NMF methods are used for clustering the features. The removal of redundant features from the relevant features is accomplished by choosing representative feature from clusters, and the final feature subset is created.

NMF is a matrix factorization method that determines the positive factorization for a given positive matrix. The specified feature contains k clusters. The main objective is to factorize X into the non-negative $m \times k$ matrix U and the non-negative $k \times n$ matrix V^T so that there is diminution in the succeeding objective function Equation 2:

$$J = \frac{1}{2} || X - UV^T || \tag{2}$$

In this equation (2), ||.|| denotes the squared sum of all the elements in the matrix. Let U = [u_{ij}], V = [v_{ij}], U = [U_1, U_2, . . . , U_k]. The above minimization trouble can be redefined as follows: diminish J with respect to U and V under the restraints of $u_{ij} \geq 0$, $v_{xy} \geq 0$, where $0 \leq i \leq m$, $0 \leq j \leq k$, $0 \leq x \leq n$, and $0 \leq y \leq k$. This is a distinctive constrained optimization problem, and can be solved by utilizing the Lagrange multiplier method. Let α_{ij} and β_{ij} be the Lagrange multiplier for restraint $u_{ij} \geq 0$ and $v_{ij} \geq 0$, correspondingly, and $\alpha = [\alpha_{ij}]$, $\beta = [\beta_{ij}]$, the Lagrange L is defined in Equation 3,

$$L = J + \mathrm{tr}(\alpha U^T) + \mathrm{tr}(\beta V^T) \tag{3}$$

The derivatives of L with respect to U and V are as Equation 4 & 5:

$$\frac{\partial L}{\partial U} = -XV + UV^T V + \alpha \tag{4}$$

$$\frac{\partial L}{\partial V} = -X^T U + VU^T U + \beta \tag{5}$$

Using the Kuhn-Tucker condition $\alpha_{ij} u_{ij} = 0$ and $\beta_{ij} v_{ij} = 0$, get the following Equation 6 & 7 for u_{ij} and v_{ij} respectively:

$$(XV)_{ij} u_{ij} - (UV^T V)_{ij} u_{ij} = 0 \tag{6}$$

$$(X^T U)_{ij} v_{ij} - (VU^T U)_{ij} v_{ij} = 0 \tag{7}$$

These equations lead to the following updating formulas as Equation 8 & 9,

$$u_{ij} \leftarrow u_{ij} \frac{(XV)_{ij}}{(UV^T V)_{ij}} \tag{8}$$

$$v_{ij} \leftarrow v_{ij} \frac{(X^T V)_{ij}}{(VU^T U)_{ij}} \tag{9}$$

To make the solution unique, further require that the Euclidean length of the column vector in matrix U is one. This requirement of normalizing U can be achieved by Equation 10 & 11,

$$v_{ij} \leftarrow v_{ij} \sqrt{\sum_i u_{ij}^2} \tag{10}$$

$$u_{ij} \leftarrow \frac{u_{ij}}{\sqrt{\sum_i u_{ij}^2}} \tag{11}$$

Every element u_{ij} of matrix U denotes the degree to which instance $f_i \in W$ belongs to cluster j, whereas every element v_{ij} of matrix V denotes to which degree feature i is related with cluster j. Redundant features are collected in a cluster that is formed by the row of the V^T matrix and a representative feature can be remove from the cluster. These attaining features are the final feature set that is called effective feature in the data.

Algorithm 1: Non-negative matrix factorization Clustering based Feature Selection algorithm

Input: X ($F_1 F_2, \ldots\ldots\ldots F_{M,C}$- given dataset, θ - T-relevance threshold

Output: S-selected feature subset

//irrelevant feature removal

1. For i=1 to m do
2. T-relevance = S U ($F_{i,C}$)
3. If T-relevance > θ then
4. S = S$\cup${F_i};

// redundant feature removal

5. Factorize X into the non-negative $m \times k$ matrix U and the non-negative $k \times n$ matrix V^T that minimize the objective function Equation 2,
6. Kuhn-Tucker condition $\alpha_{ij} u_{ij} = 0$ and $\beta_{ij} v_{ij} = 0$
7. Obtaining the two non-negative matrices U and V using Equation 8 & 9 respectively
8. Normalize U and V using Equation 10 & 11 correspondingly
9. Taking transpose of obtaining matrix of V and get V^T

$$V^T_{k \times n} = V_{n \times k}$$

10. Rows in the V^T matrix formed the clusters i.e., n clusters are created
11. From each cluster, choose a representative feature and include in S
12. Return S

In the above algorithm, firstly the irrelevant features are removed from the dataset. For this purpose, the SU measure is used. In this measure, T-relevance of the features is estimated and which is compared to the threshold value. If the T-relevance of the particular feature is above the given threshold then this feature is added to the relevant features set. From the first part of the algorithm, the relevant feature set is obtained. The redundant features are eliminated in the second part of this method that is present in the relevant features. For this, the effective Non-negative matrix factorization based clustering approach is used. Obtained relevant features are given to the NMF based clustering for eliminating the redundant features. The main objective is to factorize X into the non-negative $m \times k$ matrix U and the non-negative $k \times n$ matrix V^T. After normalizing these two matrices, taking transpose of obtaining matrix of V and clustering the features is performed. From the clusters, the representative feature is selected as it known as final selected features without irrelevant features as well as redundant features. Finally, the effective features are obtained which can improve the classification performance.

3. Experimental Result

In this section, the performance of the classification is compared with proposed system such as NMFCFS with classification result without feature selection method in terms of True Positive rate (TP rate), True Negative rate (TN rate) and classification accuracy. To assess efficiency, these comparison parameters for proposed system are measured. From the end of this experimentation section, conclude that proposed system has higher efficiency than the other techniques.

3.1 Dataset Description

The students' dataset contains 297 data instance which is collected from various colleges are considered. In the dataset 40 attributes are there which includes students' name, course, age, gender, and nature of college such as medical/engineering, college type like government, self-financed, location feature, family belong to nuclear family or joint family, family factors such that occupation & educational qualification of family members, economic factors, college factors, social factors and spending time in television, mobile, computer, personal factors, academic factors etc.,. For instance, location features defined as the location in which students' home, school and college placed such as rural area, urban area and semi-urban area. College factors is one of the attributes which provides the information about whether student refer lecturer notes which is given by lecturer or books, method of teaching such as lecturer method/black board, number of students in class, whether college allowed mobile phones or not, etc. Social factors such that guidance of relatives for studies, number of friends and academic performance of friends.

The student's performance is assessed such that good /poor in academy according to the features present in the data. Data instance with these features are given to the feature selection technique after that attain selected features. These selected features are given to the classifiers for performance evaluation. The Prism and J48 classifier is used for the experimentation. Prism is a classification algorithm for inducing modular rules and J48 is also one of the classification algorithms for building apruned or unpruned C4.5 decision tree. The 150 data instance is given as training data (with class label) to classifier for learning process. Remaining data are considered as test data (without class label) which is given to classifier with the intention of finding the class label. Finally, the output variable or attribute or class is to be decided in the dilemma is the academic status or student performance that has two possible values: PASS (student who pass the course) or FAIL (student who has to repeat the course).

3.2 True Positive Rate

True Positive rate (TP rate) is also called as sensitivity or recall, is the proportion of actual positives which are

predicted to be positive and is calculated as Equation 12. According to this work, if the outcome class label from a prediction is PASS and the actual class label is also PASS, then it is called a TP rate.

$$TP = \frac{\text{True positive}}{\text{True positive} + \text{False negative}} \quad (12)$$

Figure 1 showed that comparison of the TP parameter between classification without feature selection and classification with proposed NMFCFS feature selection. TP rate is mathematically calculated by using formula. As usual in the graph X-axis will be classification methods such as Prism, J48 and Y-axis will be TP rate. This graph shows that the proposed NMFCFS algorithm has more effective in terms of TP performance.

3.3 True Negative Rate

True Negative rate (TN rate) is also called as specificity which is the proportion of actual negatives which are predicted to be negative and is calculated as Equation 13. According to this work, if the outcome class label from a prediction is FAIL and the actual class label is also FAIL, then it is called a TN rate.

$$TN = \frac{\text{True negative}}{\text{True negative} + \text{False positive}} \quad (13)$$

Figure 2 showed that comparison of the TN parameter between classification without feature selection and classification with the proposed NMFCFS feature selection. TN rate is mathematically calculated by using formula. As usual in the graph X-axis will be classification methods such as Prism, J48 and Y-axis will be TN rate. This graph

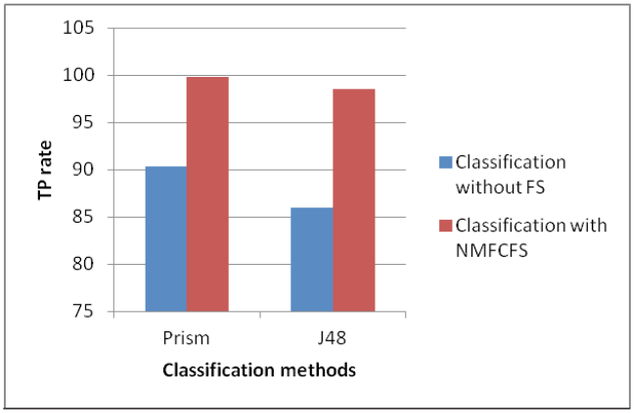

Figure 1. TP rate comparison.

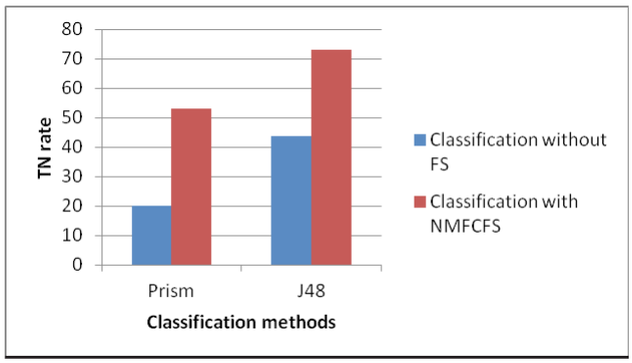

Figure 2. TN rate comparison.

shows that the proposed NMFCFS algorithm has more effective in terms of TN rate performance.

3.4. Accuracy rate

Accuracy is defined as the overall accuracy rate or classification accuracy and is calculated as Equation 14. The details about TP rate and TN rate are given in section 3.2 and 3.3 correspondingly. False Positive rate (FP rate) is defined as the proportion of actual negatives which are predicted to be positive. According to this work, if the outcome class label from a prediction is PASS and the actual class label is FAIL, then it is called a FP rate. Similarly, False Negative rate (FN rate) is defined as the proportion of actual positives which are predicted to be Negative. According to this work, if the outcome class label from a prediction is FAIL and the actual class label is PASS, then it is called a FN rate.

$$\text{Accuracy} = \frac{\text{True positive} + \text{True negative}}{\text{True positive} + \text{True negative} + \text{False positive} + \text{False negative}} \quad (14)$$

Figure 3 showed that comparison of the accuracy parameter between classification without feature selection and classification with proposed NMFCFS feature selection. Accuracy rate is mathematically calculated by using formula. As usual in the graph X-axis will be classification methods such as Prism, J48 and Y-axis will be accuracy rate. This graph shows that the proposed NMFCFS algorithm has high accuracy when compared to the existing method.

4. Conclusion

The proposed system introduces a new feature selection method for classifying the performance of the students

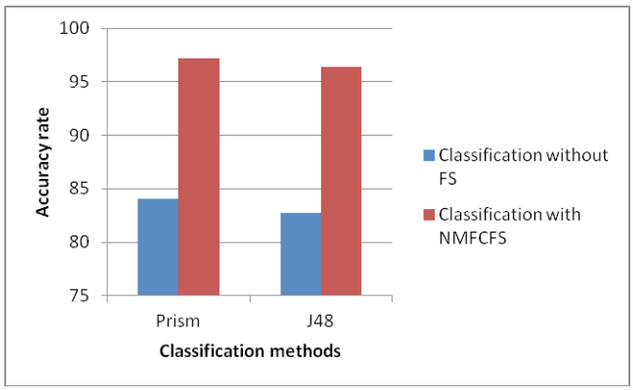

Figure 3. Accuracy comparison.

in the educational institutions. This technique has several benefits such as removal of irrelevant features and also eliminating the redundant features present in the relevant features as much as possible. This will improve the accuracy of the classification result. The classification technique such that prism and J48 is used for classifying students' performance of various colleges. The accuracy level, TP rate and TN rate of the NMFCFS system is higher than the classifier without feature selection. In addition, this novel NMFCFS method is reduce the time complexity as well as memory complexity of the classification process. By using the proposed system, the results can be improved on the performance of the students' failure and dropout prediction.

5. References

1. Baker E. International encyclopedia of education. 3rd edition. Oxford, UK: Elsevier: (In Press), 2010.
2. Ramaswami M, Bhaskaran R. Student performance forecasting: a study of classification models. intelligent computting models. Narosa Publishing House; New Delhi: 2009. p. 38–45.
3. Cope MK, Baker HH, Fisk R, Gorby JN, Foster RW. Prediction of student performance on the comprehensive osteopathic medical examination level based on admission data and course performance. J Am Osteopath Assoc. 2001; 101(2):84–90.
4. Nghe NT, Janecek P, Haddawy P. A comparative analysis of techniques for predicting academic performance. Paper presented at: 37th ASEE/IEEE Frontiers in Education Conference; Milwaukee, WI: 2007 Oct 10–13.
5. Veitch WR. Identifying characteristics of high school dropouts: data mining with a decision tree mode. Paper Presented at: Annual Meeting of the American Educational Research Association, San Diego, CA: 2004. (ERIC Document No. ED490086).
6. Mitra P, Murthy CA, Pal SK. Unsupervised feature selection using feature similarity. IEEE Trans Pattern Anal Mach Intell. , 2002; 24(3):301–12.
7. Miller. Subset Selection in Regression. 2nd edition. Chapman & Hall/CRC; 2002.
8. Duch W, Winiarski T, Biesiada J, A. Kachel A. Feature ranking, selection and discretization. Int Conf on Artificial Neural Networks (ICANN) and Int Conf on Neural Information Processing (ICONIP); 2003; p. 251–54.
9. Zhu Z, Ong Y et. al. Wrapper filter feature selection algorithm using a memetic framework. IEEE Transactions on Systems, Man and Cybernetics. 2007; 37(1):70–6.
10. Guyon I, Elisseeff A. An Introduction to Variable and Feature selection. J Mach Learn Res. 3:1157–82.
11. Oskouei RJ. Askari M, Sajja PRP. Differential impact of web technology on various dimensions of students' academic and non-academic activities (a case study). Indian Sci Technol (INDJST). 2003 Jul; 6(7): 4912–4922.
12. Manickasankari N, Arivazhagan D, Vennila G. Ontology based Semantic Web Technologies in E-learning Environment using Protege. Indian Sci Technol (INDJST). 2014 Oct; 7(S6):64–7.