MICR: Multiple Instance Cluster Regression for Student Academic Performance in Higher Education

Sk Althaf Hussain Basha
Asst. Professor in CSE
Gokaraju Rangaraju Institute of Engineering & Technology
Bachupally
Hyderabad, India

A. Govardhan
Professor in CSE & Principal
JNTUH College of Engineering
Nachupally Kondagattu
Karimnagar, India

S Viswanadha Raju
Professor in CSE & Head
JNTUH College of Engineering
Nachupally
Kondagattu
Karimnagar, India

Nayeem Sultana
Lecturer in CSC
MNR Degree & PG College
Kukatpally
Hyderabad, India

ABSTRACT
There is an increasing need for the analysis and prediction of the student academic performance in higher education. The ability to predict the student academic performance is also most important in higher education system. There are many challenges in this regard. In this paper, we present a Multiple Instance regression algorithm that models the internal data structure of three data sets to identify items that are most relevant to the data set labels, which operates on three data sets with real valued labels, each containing a set of unlabeled items, in which the relevance of each item to its data set is unknown. The goal is to predict the labels of new data set from its contents. Unlike previous Multiple instance regression methods, Multiple instance cluster Regression can operate on datasets that are structured in which they contain items drawn from a number of distinct unknown distributions. Multiple instance cluster Regression provided predictions that were more accurate than those obtained with non multiple instance approaches or multiple instance regression methods that do not model the data set structures. This Paper makes an attempt towards the application of multiple instance regression algorithms to decide to which category a student belongs (i.e, low-risk students, medium-risk students, high-risk students).

Key Words: Data Mining in Higher Education, Multiple Instance Regression, Student Performance

1. INTRODUCTION
The growing volumes of data usually create an interesting challenge for the need of data analysis tools that discover regularities in these data. Data Mining has emerged as a discipline that contribute tools for data analysis, discovery of hidden knowledge and autonomous decision making in many application domains[1].One of these application domains in higher education system. The Student can take an admission with reference to rank into institution which gave clear scope for evaluation and comparison of predicted and real values[2].Who ever admitted through Rank Analysis that data have considered to deciding which is best one among the Non Linear Regression using XLSTAT [3].
This attempt is to use multiple instance regression algorithms to predict the new admitting student falls under which division or category (low-risk students, medium-risk students, high-risk students).

Classical Supervised learning methods operate on individual items, each represented by a feature vector and assigned a label, which is either categorical or real valued some learning problems do not fit this model. There are situations in which observations are instead data sets with a single label applied to the whole data set. When the data set have real valued labels, the goal is to construct a regression model that can predict a data set label from its contents primary instance regression [4,5] assumes that a single item in each dataset dictates the label. It can select primary instances for labeled dataset, but it cannot make predictions for a new data set unless it is known a priori which item is the primary one.

The core assumption of existing approaches is that the data set themselves are unstructured the items in the data set are drawn from a single, fixed distribution. For example, in drug activity prediction, all items in a data set are different conformations of the same molecule. However, in some domains, the datasets are structured; they contain items drawn from a number of distinct underlying data distributions.
Multiple instance cluster Regression leverages the within-data set Structure by modeling the distinct data components with a clustering step and then constructing local regression models for each Component. Multiple instance cluster Regress includes a final model selection steps that picks the best cluster / regression model.

The main challenges in this area as follow

a) Can the Multiple Linear Regression algorithm help us to predict the cluster to which the student is more relevant to?
b) Can the formed clusters using multiple linear Regressions help us to increase the student performance /graduation rate of the university?

The main objective of this paper to predict the new student who gains admission into the university falls under which group i.e., low-risk students, medium-risk students or high-risk students. We first try to predict and again try to form clusters of each group.

The rest of the paper is organized as follow: Section 2 briefly discusses related work on Multiple Instance Regression .Section 3 provides the Multiple Instance Algorithm. Section 4 Presents the Description of Methodology. Section 5 provides data set collection and data set preparation. Section 6 discusses experimental results. Section 7 presents the conclusions.

2. RELATED WORK
Multi-instance learning Problems originates from the research of drug activity prediction, where the multiple instance learning has been studied by many researchers.

Dietterich et al. (1997) presented three algorithms (i.e, standard algorithm, outside-in, inside-out) for learning axis parallel hyper rectangles (APRs) in the multiple-instance model. They presented three general designs for learning axis aligned boxes

in the multi-instance model. First, they considered the standard algorithm that forms the smallest box that bounds the positive examples. They also explored a noise tolerant version of this algorithm. Next they presented an algorithm they refer to as the outside-in algorithm. In this algorithm, first they construct the smallest box that bounds all of the positive examples, and then they shrink this box to exclude false positives. Finally, they presented a third algorithm, the inside-out algorithm, which starts with a set point in the feature space and "grows" a box with the goal of finding the smallest box that covers at least one example from each positive data set and no examples from any negative data set. They showed results that the inside-out algorithms perform much better than either of the others [6].

Auer (1997) [7] presented an algorithm that learns using simple statistics and hence avoids some potentially hard computational problems that were required by the heuristics used by Dietterich et al.

Maron et. al. presented a frame work called Diverse Density. When describing the shape of a molecule by n features, one can view each configuration of the molecule as a point in an n-dimensional feature space. As the molecule changes its shape, it traces out a manifold through this n-dimensional space.(To keep the size of the datasets manageable, only shapes of the molecule that have sufficiently low potential energy were considered). The diverse density at a point p in the feature space is a measure of both how many different positive data sets have an example near p, and how far the negative instances are from p. They use gradient ascent with multiple starting points (namely, starting from each point from a positive data set) to find the point that maximizes the diverse density [8].

Wang and Zucker (2000) proposed a lazy learning approach to multiple-instance learning by applying a variant of the k-nearest neighbor algorithm (k-NN). To compute the distance between data sets b1 and b2 they used the minimum distance between a point in b1 and a point in b2.While a standard k-NN approach did not work well, by also using citers of p (points who include p as one of its nearest neighbors) as well as p's nearest neighbors they reached [9].

S Andrew et. al presented two formulations of multiple instance learning as a maximum margin problem. They proposed extensions of the support vector machine learning approach lead to mixed integer quadratic programs that can be heuristically. They presented experimental results on a pharmaceutical data set and on applications in automated image indexing and document categorization [10].

Amar et al. (2001) extended the k nearest neighbor and diverse density approaches to Multi Instance Learning problems in which each data set has a real-valued label that indicates its proximity to the target concept [11]. Likewise, Goldman and Scott (2003) interpreted real valued labels to be "the degree to which the example satisfies the target concept" and used axis-aligned rectangles to learn the target concept .and used axis aligned rectangles to learn the target concept. While useful for identifying a target concept, none of these approaches are designed to model or learn general regression relationships [12].

In contrast, Multiple Instance Regression seeks to build a regression model that maps data sets to real-valued outputs; there is no notion of a target concept. Ray and Page [13] pioneered this area by developing a primary instance regression (PIR) method. The PIR approach assumes that the label of a data set is determined by exactly one primary instance and that the rest of the items in the data set are noisy observations of the primary instance.

PIR is an EM-based solution that alternately selects the most likely primary instance for each training data set and then to maximizes the fit of a linear regression through the primary instances. The learned model can only be applied to new data sets if the primary instance for each one is known. Cheung and Kwok [14] and Ray [15] identified problem domains in which it is possible to assume that the primary instance is the one with the largest output value. For other domains, min, average, or sums are appropriate combining functions, and it is possible to learn which of these four functions applies to a given data set [15]. However, none of these functions models per-item relevance to the data set label, so the presence of irrelevant items will skew the results. Methods that directly estimate item relevance include CH-FD for classification [16] and QPAP-Salience [17] for regression. These techniques use alternating optimization to iteratively estimate item relevance and coefficients for the learned models that predict labels. However, neither method can generalize to new data sets, where both the data set labels and item relevance are unknown. CH-FD was evaluated by applying the learned classifier only to individual items, not data sets. QPAP-Salience was not evaluated on new data.

Chen et al. (2006) proposed a method for multiple-instance classification that represents each data set by its similarity to each item in the data set, and then uses an Support Vector Machine (SVM) to select relevant features. Since each feature implicitly stands for an item, a subset of relevant items is also identified [18].

This paper advances the state of the art by proposing a method that addresses both goals: assigning per-item relevance and building regression models that can generate predictions for new data sets. These goals are achieved by explicitly with internal structure.

3. MULTIPLE INSTANCE REGRESSION

In the multiple instance regression problems, we seek a function that maps data sets to real values. In many cases data sets are also structured. The data set contents are drawn from a variety of different underlying distributions, not all of which are relevant to the dataset labels.

Definition of Multiple Instance Regression, MIR seeks to build a regression model that maps the data sets to real-valued outputs; there is no notion of a target concept .With this Multiple Instance Regression model we can predict the new incoming data sets to the already existing clusters in higher education.

Algorithm 1: The Multiple Instance Cluster Regression algorithm

Inputs: Data set data **DS** = $\{\mathbf{D}^{i}\}_{i=1\ldots m}$, labels *Y*,

Number of clusters *k*

1: Begin

2: **X:** = $\cup_{i=1\ldots m}$ $\mathbf{D}^{i}$ // Joining of all items into single set, ignoring data set structure//

3: $\theta_{i=1\ldots k}$, := Cluster (**X**, *k*) // Cluster all items into *k* clusters//

4: for *i* = 1 to m do

5. begin

6: for *j* = 1 to *k* do

7: begin

8: *R* : = Relevance ($\mathbf{D}^{i}$, θ_j) // Relevance vector for items in data set *i* with respect to cluster *j*//

9: $\hat{D}^{i}_{j}$:= $\mathbf{D}^{i}$ R // Sack for data set $\mathbf{D}^{i}$ in cluster *j*: weighted average of contents of $\mathbf{D}^{i}$ //

10: end

11: end

12: for $j = 1$ to k do

13: begin

14: ψ_j := Regression ($\{\hat{D}^i_j\}_{i=1...m}$, labels Y) // Regression model for cluster j//

16: end

17: $[\psi', \theta']$:= Select ($\{\psi_j, \theta_j\}_{j=1...k}$, $\{D^i\}_{i=1...m}$, Y) // Mapping the data by best local model//

18: end.

Outputs: regression parameters ψ' and cluster parameters θ' for the best local model

The Multiple instance cluster regression algorithm is mainly used to from cluster i.e., data very much similarity is formed as a group. The main assumption of the Multi instance Cluster Regression algorithm (algorithm 1) is that the individual items are drawn from a set of underlying clusters and that a data set's label is a function of one relevant cluster. Each data set is assumed to contain items drawn from one or more of these clusters. After clustering all items together (step2 and 3), we have soft assignments of each item in each data set to each cluster. Using these assignments, we construct per-data set sacks for each cluster (step 4–11). The sack for cluster j within data set i, $\hat{D}^i_j$ the average of all items in data set i weighted by their respective memberships in cluster j (i.e., "relevancies"), denoted by R. Given these sacks, we construct a regression model for each cluster j using the cluster j sacks from all data sets (step 12–17). Finally, a model selection step identifies the regression model (i.e., cluster) that best captures the relationship between data and data set labels (step18).

Algorithm 2: The Multiple Instance Cluster Prediction algorithm

Inputs: New data set **D**

Cluster parameters θ'

Regression model parameters ψ'

1: Begin

2: R: = Relevance (D, θ') // Per-item relevance//

3: $\hat{\hat{D}}$:= **D** R // Data set sack//

4: $\hat{y}$:= Regression Prediction ($\hat{D}$, ψ') // Regression prediction//

5: end

Output: Prediction for **D**: $\hat{y}$

The Multiple instance cluster prediction algorithm is used to find the relevancy of new incoming item with the existing cluster formed in previous algorithm.

The Multiple instances Cluster Prediction algorithm (Algorithm 2) assigns a predicted value, by to a previously unlabeled data set. It takes as input the new data set and the local model cluster and regression parameters, θ' and ψ', that were picked in the model-selection step of Multiple instance Cluster Regression. It computes the relevance of each item in the new data set to the local model cluster, and then constructs the corresponding weighted average sack for this data set. It employs a Regression Prediction routine, corresponding to the Regression learner from multiple instances Cluster Regression that uses previously learned regression parameters to predict a label for the sack.

Algorithm 3: The Relevance subroutine

Inputs: Single data set **D**

Parameters for one cluster θ_c

1: Begin

2: $r(i) = p(cl_i = c|D(i); \theta_c), \forall i$

3: $z := \sum_i r(i)$

4: $R(i) := r(i) / z, \forall i$ // Relevance of item i to cluster θ_c //

5**:** end

Output: Relevance column vector R

This algorithm assign the new item to the cluster which is more relevant to .This relevancy is predicted with the help of previous algorithm. Both Multiple Instance Cluster Regression and Multiple Instance Cluster Prediction calculate the relevance of a given item to a particular cluster (Algorithm 3). The generative model for a cluster c with parameter θc provides $p(D(i)|cl_i = c; \theta_c)$ where cl_i = c means that item i was generated by cluster c. Via Bayes's rule, we can calculate $p(cl_i = c|D(i); \theta_c)$. We renormalize these values across the data set so that the sum of all relevancies within a data set is 1. Thus, the relevance of an item, with respect to a cluster, is sensitive to its context, which is the rest of the data set's contents. The renormalization ensures that the sack for the cluster is a weighted average of the contents of the data set (i.e., it is an affine combination of the data set data).

4. BASIC IDEA OF THE PAPER'S METHODOLOGY

We randomly split the data set into clusters and Construct the label based on the average of the contents of the data set. When a new item enters we try to predict the relevancy of the item to the cluster with the help of the label of dataset which contains three data sets of students internally and they are

a)Low-Risk students

b)Medium-Risk students

c) High-Risk students

First combine all the datasets to form a single union of data set

(d) Cluster all the items by using Clustering algorithm (k-means algorithms)

(e)Identify the labels of each dataset by finding the weighted average of the items in the datasets

(f)When a new element comes the relevancy factor is to found out and the cluster to which the element is mostly relevant to must be assigned with that element.

5. DATA SET COLLECTION AND PREPARATION

In this session, We present the results of applying Multiple instance algorithm on student data set collected from Gokaraju Rangaraju Institute of Engineering and Technology (JNT University) from 2000-2008. The data set taken from MCA, MBA, M.Tech, B.Tech Courses. There are 4659 instances in this data set. Every instance contains five attributes such as previous semester marks; Practical knowledge, Assignment marks, internal marks, and Involvement of student in Extra-curricular activities and these attributes are numerical. This data set is pre-labeled in to three classes: Low-Risk students, Medium-Risk students and High-Risk students.

During in data collection, the relevant data is gathered and the quality of data must be verified. Usually, the assembled data contains of missing or incomplete attribute, noisy (containing errors, or outlier values that deviate from expected), and inconsistent of data are common. Therefore, the collected data must be cleaned and transformed before it can be utilized in data mining system since data mining should process cleaned data in order to come out with better and or quality results. Data cleaning involves several of processes such as filling in missing values; smoothing noisy data, identifying or removing outliers, and resolving inconsistencies. Then, the cleaned data are transformed into a form of table that is suitable for data mining model.

In the Data set preparation, Higher Education data is a real word benchmark test data for algorithm which contains the attributes with which we can predict the performance of students. So, for this we take up the attributes for predicting students performance and in order to assign them to a cluster are Previous semester marks, Practical knowledge, Assignment marks, Internal marks, Involvement of student in Extra-curricular activities and so on. Depending upon the predicted value we assign the student to a cluster he is very much relevant to.

6. EXPERIMENTAL RESULTS

We take the dataset or simply call them as cluster of low-risk students, medium-risk students, and high-risk students as input to this algorithm implementation and tries to predict a new incoming student into the university falls under which cluster. Using Multiple Linear Regression algorithm we are predicting more successfully to which cluster the new student is more relevant to and using this clusters and by using enhanced quality of education in the university we can easily increase the graduation rate/Student performance.

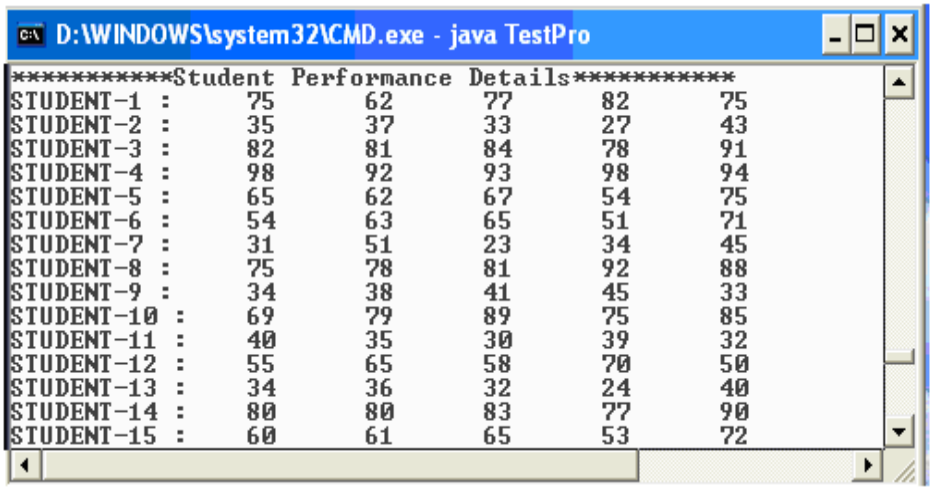

Figure 1. Student Performance with different activities

Fig. 1 shown the student's performance of previous semester marks, Practical knowledge, Assignment marks, Internal marks, and Involvement of student in Extra-curricular activities.

In first step, we take overall performance (average performance of student in different aspects) i.e The Overall Data set is as shown below (Fig.2).

Figure 2: overall or average performance of each student in different aspects.

Now the above data set is applied to Multiple instance Cluster Regression algorithm (Algorithm 1) is that the individual items are drawn from a set of underlying clusters and that a data set's label is a function of one relevant cluster. After the Clusters are formed we try to find out the labels of each clusters with the help of Arithmetic mean of data in each cluster.

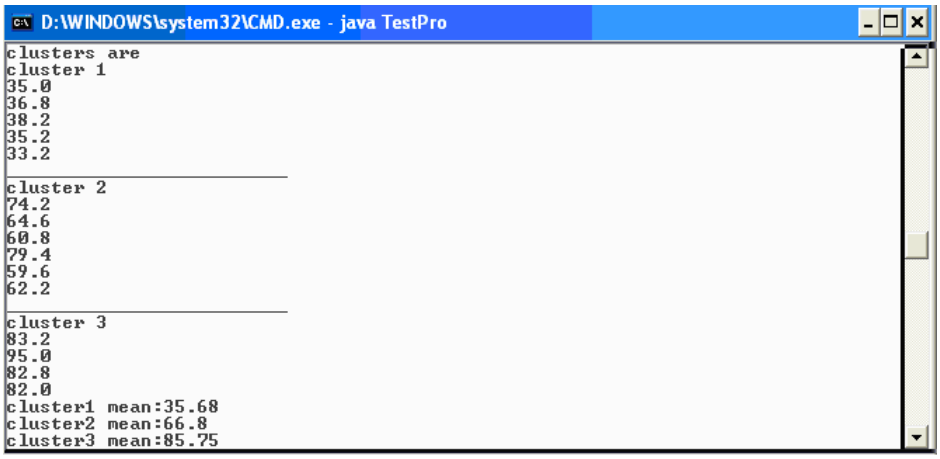

Figure 3 (a): Mean value of Cluster

Now we try to assign the data items to each cluster with the help of relevancy factor i.e., we try to compare the data item to each cluster label and to whichever label the data item is more relevant to we assign the respective data item to the relevant cluster.

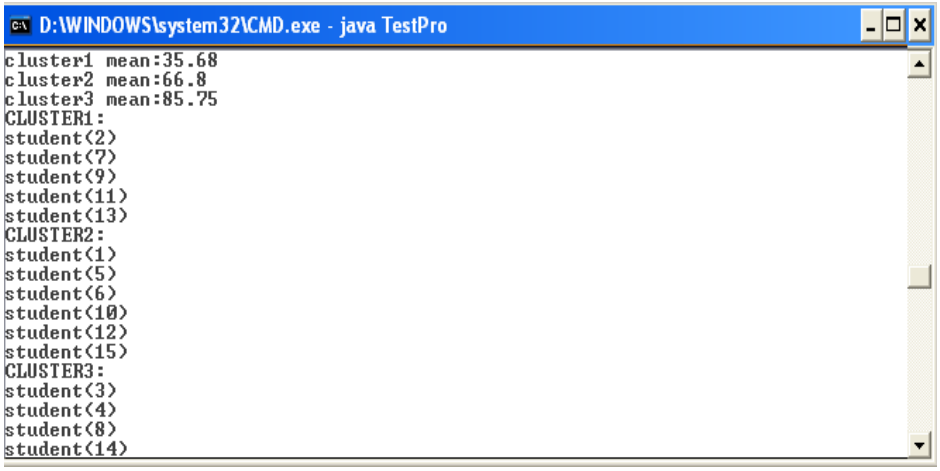

Figure 3(b): Clustering of data using figure 3(a).

After assigning the data items to the clusters we try to categorize them into three different items in figure 3(b) i.e, High risk students, medium risk students, and low risk students.

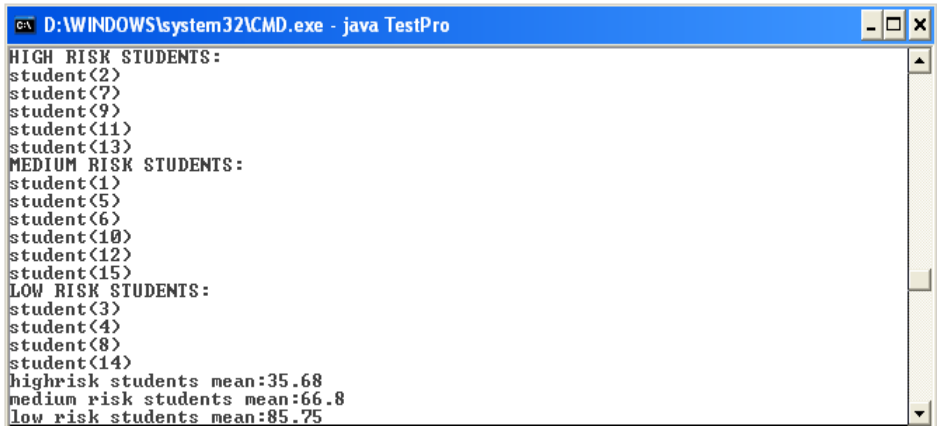

Figure 3 (c) : Clustering Labeling using Figure 3(b)

Now applying Multiple instances Cluster Regression and Multiple instances Cluster Prediction calculate the relevance of a given item to a particular cluster (Algorithm 3).

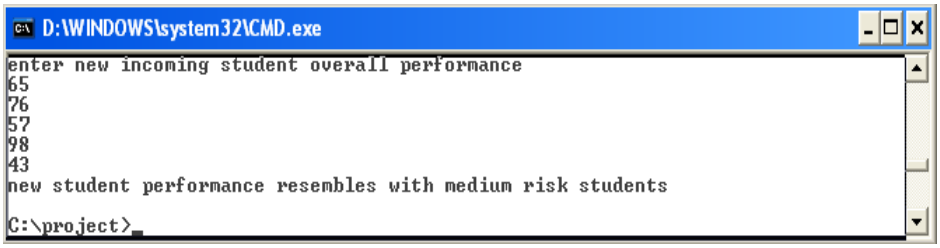

Figure 4: Identification of Cluster for new student.

Now we take new incoming new student overall performance is 67.8

This value matches or is relevant to label high risk student therefore it goes into cluster 2. i.e student will get Medium Risk Students category. Our Experimental results as shown below.

SNo	Data set	Size	Low Risk Students	Medium Risk Students	High Risk Student
1	MCA	1022	328	522	172
2	MBA	714	214	392	108
3	M Tech	512	256	204	52
4	B Tech	2411	603	1085	723
	Total	4659	1401	2203	1055

Table 1: Clustered data with three categories

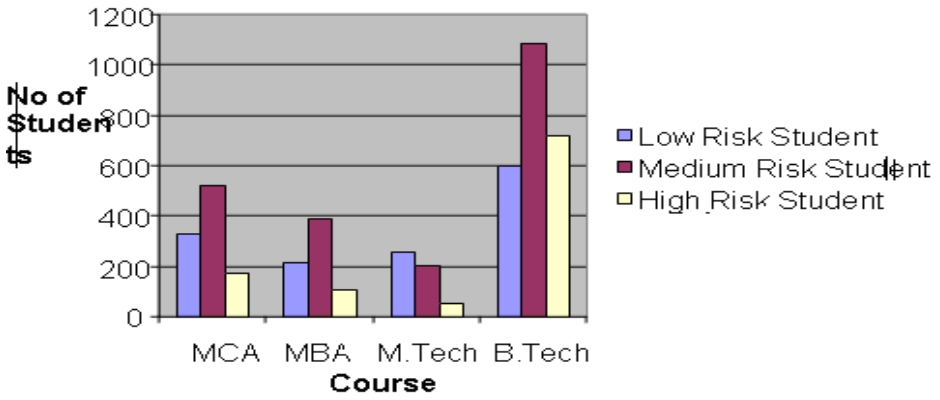

Figure 5 : Graphical Representation for Table 1.

7. CONCLUSION

Predicting students' academic performance is great concern to the higher education system. Recently data mining can be used in a higher educational system to predict the students' academic performance. In this paper, we have introduced the multiple instance regression algorithms for Student Performance in Higher Education system, to predict the relevancy of the incoming item from a new data set to the already existing data sets. All the dataset used in this experiments have the attributes are numerical which consist of Marks in previous semester, Practical Knowledge, Internal marks, Assignment marks, Extra-Curricular Activities. Our Experimental results on numerical data sets show that the multiple instance algorithms perform well. The proposed algorithm to find the cluster for new object at low computation cost.

The Future Research is the development of best model that incorporates domain knowledge and exploring other schemas for modifying the representation of multi instance prediction problems.

REFERENCES

[1]. Jiawei Han, Micheline Kamber, Data Mining: concepts and techniques, 2nd edition, Morgan Kaufmann publishers,2008.

[2]. Sk.Althaf.H.Basha , A.Govardhan, N. Sandhya, K. Anuradha, P. Premchand , Rank Analysis Through Polyanalyst using Linear Regression, IJCSNS(International Journal of Computer Science and Network Security) Vol.9 No.9,pp.290-293, 2009

[3]. Sk. Althaf Hussain Basha, A.Govardhan, S.Viswanadha Raju, Nayeem Sultana,A Comparative Analysis of Prediction Techniques for Predicting Graduate Rate of University, EJSR (European Journal of Scientific Research),Vol.46 No.2, pp.186-193,2010

[4]. S. Ray and D. Page. Multiple instance regression. In Proceedings of the 18th International Conference on Machine Learning, pp. 425–432, 2001.

[5]. P.-M. Cheung and J. T. Kowk. A regularization framework for multiple-instance learning. In Proceedings of the 23rd International Conference on Machine Learning, pages 193–200, 2006.

[6]. Dietterich, T. G., Lathrop, R. H. and Lozano-P´erez, T. Solving the multiple- instance problem with axis-parallel rectangles. Artificial Intelligence, 89, pp.31–71, 1997.

[7]. Auer,P. Online learning from multi-instance example: Empirical evaluation of a theoretical approach. In Proceedings of the 14th International Conference on Machine Learning, San Francisco:Morgan Kaufmann, pp. 21-29,1997.

[8]. Maron, O. and Lozano-P´erez, T. A framework for multiple-instance learning. Neural Information Processing Systems, 10, MIT Press, 1998.

[9]. Wang, J. and Zucker, J.-D. Solving the Multiple-Instance Learning Problem: A Lazy Learning Approach. Proceedings 17th International Conference on Machine Learning San Francisco: Morgan Kaufmann., pp. 1119–1125,2000.

[10]. S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In S. Becker, S. Thrun, and K. Obermayer, editors, Advances in Neural Information Processing Systems MIT Press, Cambridge, MA, 15, pp. 561–568. 2003.

[11]. R. A. Amar, D. R. Dooly, S. A. Goldman, and Q. Zhang. Multiple-instance learning of real-valued data. In Proceedings of the 18th International Conference on Machine Learning, pp. 3–10, 2001.

[12]. S. A. Goldman and S. D. Scott. Multiple-instance learning of real-valued geometric patterns. Annals of Mathematics and Artificial Intelligence, 39(3), pp.259–290, 2003.

[13]. S. Ray and D. Page. Multiple instance regression. In Proceedings of the 18th International Conference on Machine Learning, pp. 425–432, 2001.

[14]. P.-M. Cheung and J. T. Kowk. A regularization framework for multiple-instance learning. In Proceedings of the 23rd International Conference on Machine Learning, pp. 193–200, 2006.

[15]. S. Ray. Learning from Data with Complex Interactions and Ambiguous Labels. PhD thesis, University of Wisconsin, Madison, 2005.

[16]. G. Fung, M. Dundar, B. Krishnapuram, and R. B. Rao. Multiple instance learning for computer aided diagnosis. In Advances in Neural Information Processing Systems 19, pp. 425–432, 2006.

[17]. K. L. Wagstaff and T. Lane. Salience assignment for multiple-instance regression. In Proceedings of the ICML 2007Workshop on Constrained Optimization and Structured Output Spaces, 2007.

[18]. Chen, Y., Bi, J., & Wang, J. Z. (2006). MILES: Multiple-instance learning via embedded instance selection. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 28.