Mining Student data by Ensemble Classification and Clustering for Profiling and Prediction of Student Academic Performance

Ashwin Satyanarayana
N-913, Dept. of Computer Systems Technology
New York City College of Technology (CUNY)
300 Jay St, Brooklyn, NY 11229.
{asatyanarayana@citytech.cuny.edu}

Gayathri Ravichandran
Dept. of Computer Science
M S Ramaiah Institute of Technology
MSR College Road, MSR Nagar,
Bengaluru, Karnataka 560054, India
{gayathrix7@gmail.com}

Abstract

Applying Data Mining (DM) in education is an emerging interdisciplinary research field also known as Educational Data Mining (EDM). Ensemble techniques have been successfully applied in the context of supervised learning to increase the accuracy and stability of prediction. In this paper, we present a hybrid procedure based on ensemble classification and clustering that enables academicians to firstly predict students' academic performance and then place each student in a well-defined cluster for further advising. Additionally, it endows instructors an anticipated estimation of their students' capabilities during team forming and in-class participation. For ensemble classification, we use multiple classifiers (Decision Trees-J48, Naïve Bayes and Random Forest) to improve the quality of student data by eliminating noisy instances, and hence improving predictive accuracy. We then use the approach of bootstrap (sampling with replacement) averaging, which consists of running k-means clustering algorithm to convergence of the training data and averaging similar cluster centroids to obtain a single model. We empirically compare our technique with other ensemble techniques on real world education datasets.

Keywords: *Educational Data Mining, Ensemble Classification, k-means Clustering, Bootstrap averaging, Student academic prediction.*

1. Introduction

The field of Data Mining (DM) is concerned with finding new patterns in large amounts of data. Data Mining (DM) techniques, allow a high level extraction of knowledge from raw data and offer interesting possibilities for the education domain. In particular, several studies have used DM methods to improve the quality of education and enhance school resource management by increasing student retention[1,2,3,14].

Educational Data Mining is defined as "an emerging discipline concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students and the settings they learn in"[13]. The process of tracking and mining student data in order to enhance teaching and learning is one of the goals of Educational Data Mining. Hence, the ability to predict students' academic performance is very important in educational environments. Predicting academic performance of students is challenging since the students' academic performance depends on diverse factors such as personal, socio-economic, psychological and other environmental

variables. Another way to enhance teaching is to identify groups of students with similar learning style and behavioral learning patterns.

The objective of this paper is three-fold: to improve the quality of student data, to predict student academic performance and cluster groups of students with similar learning styles, using data mining techniques such as ensemble classification, anomaly detection and clustering. Ensemble methods have been called the most influential development in data mining and machine learning in the past decade. They combine multiple models usually producing an accurate model than the best of its individual components.

The paper is organized as follows: section 2 surveys data mining techniques for clustering and evaluating student performance, section 3 mentions our contributions, section 4 describes our ensemble (filtering, ensemble classification and clustering) techniques in detail, and section 5 shows our experimental results using datasets from the UCI repository. Finally, in section 6 the conclusions are outlined.

2. Prior Work in this area

Various clustering algorithms have been applied to educational data sets in diverse studies[17]. Applied statistical clustering methods such as K-means clustering and Hierarchical clustering have been applied to student annotations[18]. Emotional intelligence of students is compared using k-means clustering on questionnaire data[19].

Alaa el-Halees[2] show that data mining can be used in educational settings to understand the learning process of identifying, extracting and evaluating variables related to the learning process of students. Han and Kamber[1] provide a good description of the different data mining tools and software on multidimensional data and their analysis. Bayes classification was used by Pandey and Pal[3] for student performance prediction based on 600 students from different colleges. They use attributes such as category, language and background qualification of students. Linear regression used by Hijazi and Naqvi[4] on the student performance prediction based on a sample of 300 students (225 males, 75 females) from different colleges. They consider attributes such as attendance, hours spent studying, family income, mothers age, mothers education. They found that the factors like mother's education and student's family income were highly correlated with the student academic performance.

Several other lines of research have explored data mining methods to predict student academic performance such as: Neural networks for giftedness identification[5], Predicting student performance using data mining with educational web-based system[6], Determination of factors influencing the achievement of the first year university students using data mining[7], Application of GMDH algorithm for modeling of student's quality[8], Predicting persistence of students using data mining methods[9] and Application of data mining methods to the student's dropout problem[10].

3. Our Contributions

Our contributions in this paper are as follows:

1. To use ensemble filtering on student data to improve the quality of the data, by eliminating mislabeled class noise.
2. To use ensemble classification to create a more accurate prediction of student performance in two different environments: high school and first year college data.

3. To use bootstrap averaged k-means clustering to identify groups of students with similar learning styles.

4. Methodology

4.1 Ensemble Noise Filtering

We propose an ensemble classifier framework for noise filtering and predicting student performance. We show that by having more than one classifier (or model) to evaluate the instances, we *extend* the model space as compared to a single classifier. Thus, by using multiple (in this paper, we use three) classifiers we perform *an approximation* of model averaging[15]. We focus on improving the quality of student academic training data by identifying and eliminating mislabeled instances by using multiple classifiers. An ensemble classifier detects noisy instances by constructing a set of classifiers (base level detectors). A *majority vote* filter tags an instance as mislabeled if more than half of the *m* classifiers classify it incorrectly. A *consensus* filter requires that all classifiers must fail to classify an instance as the class given by its training label.

Our filtering approach begins by performing *k*-fold cross validation. *k*-fold cross validation is a commonly used technique which takes a set of *n* examples and partitions them into *k* sets of size *n/k*. For each fold, multiple classifiers are trained on all the other folds and tested on the current fold. Thus *k* hypotheses θ_1, θ_2, ……θ_k are generated. This prediction is equivalent to outputting the average of *k*-hypotheses as shown in equation (1) below:

$$\Pr(y = t \mid x, \theta) = \frac{1}{k}\sum_{i=1}^{k}\delta(t, \theta_i) \tag{1}$$

where δ is a 0-1 loss function, which returns 1 if θ_i predicts the correct label *t*, else returns 0.

Our Ensemble Filtering algorithm (Fig 1) begins with *k* almost equal sized subsets of our dataset *E* (step 1) and an empty output set *A* of detected noisy examples (step 2). The main loop (steps 3-12) is repeated for each fold E_i. In step 4, we form a set E_y which includes all the examples from *E* except E_i. E_y is used as an input for the *k* inductive learning algorithms to generate models *k* models $\theta_{y,1}$, $\theta_{y,2}$…… $\theta_{y,j}$. The set E_i is evaluated by our *j* models in steps 8-11. If more than half of the models misclassify an instance, then it is treated as noise and eliminated.

Algorithm: EnsembleFiltering (*E*)
Input: *E* (training set)
Parameter: *k* (number of subsets of *E*, typically 10)
j (number of inductive learning algorithms, typically 3)
Output: A (a detected noisy subset of *E*)
(1) Form *k* almost equal sized subsets of E_i, where $\cup_i E_i = E$
(2) A ← Ø
(3) **for** i = 1,……,*k* do
(4) $E_y \leftarrow E \setminus E_i$
(5) **for** *m* = 1…….*j* do
(6) $\theta_{y,m}$ ← model built from bootstrap sample E_y and inductive algorithm *m*
(7) **end for**
(8) **for** every $e \in E_i$ do
(9) If *e* is misclassified by more than half the $\theta_{y,m}$ models built, then it is noisy and needs to be eliminated.
(10) A ← A U {*e*}
(11) **end** for
(12) **end for**

Fig 1. Ensemble Filtering Algorithm

4.2 Bootstrapped averaging using k-means clustering

In the area of clustering, we use k-means clustering with bootstrap averaging to identify groups of students with similar characteristics. K-means clustering is one of the most popular clustering algorithms used in data mining. The approach of bootstrap (sampling with replacement) averaging consists of running k-means clustering to convergence on small bootstrap samples of the training data and averaging similar cluster centroids to obtain a single model[16]. This approach is complimentary to other speed-up techniques such as parallelization. Our approach builds multiple models by creating small bootstrap samples of the training set and building a model from each, but rather than aggregating like bagging, we average similar cluster centers to produce a single model that contains *k* clusters. To test the effectiveness of bootstrap averaging, we apply clustering in finding representative clusters of the population.

The k-means clustering problem is to divide the *n* instances into *k* clusters with the clusters partitioning the instances $(x_1 \ldots x_n)$ into the subsets $Q_{1\ldots k}$. The subsets can be summarized as points $(C_{1\ldots k})$ in the *m* dimensional space, commonly known as centroids or cluster centers, whose co-ordinates are the average of all points belonging to the subset. K-means clustering can also be thought of as vector quantization with the aim being to minimize the vector quantization error (also known as the distortion) shown in equation (1).

$$VQ = \frac{1}{2}\sum_{i=1}^{n} D(x_i, C_{f(x_i)}), \text{where D is a distance function and f(x) returns the closet cluster index to instance } i \tag{1}$$

Typically the initial centroid locations are determined by assigning instances to a randomly chosen cluster. After initial cluster centroid placement the algorithm consists of the two following steps that are repeated until convergence. As the solution converged to is sensitive to the starting position the algorithm is typically restarted many times.

1) The assignment step: instances are placed in the closest cluster as defined by the distance function.

$$f(x_i) = \arg\min_j D(x_i, C_j)$$

2) The re-estimation step: the cluster centroids are recalculated from the instances assigned to the cluster.

$$C_j = \sum_{f(x_i)=j} \frac{x_i}{|Q_j|}$$

These two steps repeat until the re-estimation step leads to minimal changes in centroid values. Throughout this paper we use the version of the k-means clustering algorithm commonly found in data mining applications.

Algorithm: Bootstrap Averaging
Input: D: Training Data, T: Number of bags, K: Number of clusters
Output: A: The averaged centroids.
// Generate and cluster each bag
(1) For i = 1 to T
(2) X_i = BootStrap(D)
(3) C_i = k-means-Cluster(X_i,K) // Note C_i is the set of *k* cluster centroids and $C_i = \{c_{i1}, c_{i2} \ldots c_{iK}\}$
(4) EndFor
// Group similar clusters into bins with the bin averages stored in $B_1 \ldots B_k$ their sizes are $S_1 \ldots S_k$
(5) For i = 1 to T
(6) For j = 1 to K
(7) Index = AssignToBin(c_{ij}) //See section on signature based comparison
(8) B_{Index} += c_{ij}
(9) EndFor
(10) EndFor
(11) For i = 1 to K
(12) B_i /= S_i
(13) $A_i = B_i$
(14) EndFor

Fig 2. Bootstap Averaging K-means Clustering algorithm

For each cluster centroid we create a signature that can be generated quickly and group clusters according to the signature. We use the positions of the attributes and their values to create a signature of the form:

$Signature(c_{ij}) = \Sigma_l\, c_{ijl} * 2^l$, where c_{ijl} is the l^{th} attribute for the j^{th} cluster of the i^{th} model.

As each attribute is scaled to be between zero and one this creates a signature with the range 0 till 2^{m+1} as there are *m* attributes. After the signature from each cluster is derived we sort them in ascending order and divide them to form the final clusters. Throughout this paper we use this method.

5. Empirical Results

In this section, we discuss our experiments that demonstrate the improved predictive accuracy using our ensemble filtering approach as compared to single model filtering. We tested our approach on two datasets: (a) UCI Student Performance dataset[11] and (b) New York City College of Technology CST introductory course dataset. For each dataset, we compare the accuracies after filtering using the following techniques:

1. Single Model: We used decision trees (J48) as our single filtering base model.
2. Online Bagging: We implemented online bagging as illustrated by Oza[12] using Naïve Bayes as the base model.
3. Ensemble Filtering: Our algorithm (shown in Fig 1) uses the following classifiers: J48, RandomForest and Naïve Bayes. We use *consensus vote* for Student performance dataset and *majority vote* for the dataset from New York City College of Technology.

5.1 Student Performance Dataset (UCI):

This dataset is based on a study of data collected during the 2005-2006 school year from two public schools, from the Alentejo region of Portugal[11]. The database was built from two sources: school reports, based on paper sheets and including few attributes (i.e. the three period grades and number of school absences); and questionnaires, used to complement the previous information. The final version contained 37 questions in a single A4 sheet and it was answered in class by 788 students. Latter, 111 answers were discarded due to lack of identification details (necessary for merging with the school reports). Finally, the data was integrated into two datasets related to Mathematics (with 395 examples) and the Portuguese language (649 records) classes[11].

Attribute	Description (Domain)
sex	student's sex (binary: female or male)
age	student's age (numeric: from 15 to 22)
school	student's school (binary: *Gabriel Pereira* or *Mousinho da Silveira*)
address	student's home address type (binary: urban or rural)
Pstatus	parent's cohabitation status (binary: living together or apart)
Medu	mother's education (numeric: from 0 to 4[a])
Mjob	mother's job (nominal[b])
Fedu	father's education (numeric: from 0 to 4[a])
Fjob	father's job (nominal[b])
guardian	student's guardian (nominal: mother, father or other)
famsize	family size (binary: ≤ 3 or > 3)
famrel	quality of family relationships (numeric: from 1 – very bad to 5 – excellent)
reason	reason to choose this school (nominal: close to home, school reputation, course preference or other)
traveltime	home to school travel time (numeric: 1 – $<$ 15 min., 2 – 15 to 30 min., 3 – 30 min. to 1 hour or 4 – $>$ 1 hour).
studytime	weekly study time (numeric: 1 – $<$ 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – $>$ 10 hours)
failures	number of past class failures (numeric: n if $1 \leq n < 3$, else 4)
schoolsup	extra educational school support (binary: yes or no)
famsup	family educational support (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
paidclass	extra paid classes (binary: yes or no)
internet	Internet access at home (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
freetime	free time after school (numeric: from 1 – very low to 5 – very high)
goout	going out with friends (numeric: from 1 – very low to 5 – very high)
Walc	weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
Dalc	workday alcohol consumption (numeric: from 1 – very low to 5 – very high)
health	current health status (numeric: from 1 – very bad to 5 – very good)
absences	number of school absences (numeric: from 0 to 93)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20)

Table 1. Attributes of the UCI Student performance dataset.

In this work, the Mathematics and Portuguese grades (i.e. G3 of Table 1) will be modeled using 5-Level classification (Table 2) – based on the Erasmus (European exchange program) grade conversion system as used by Cortez[11]. The results are shown in Table 3.

16-20	14-15	12-13	10-11	0-9
A	B	C	D	F

Table 2. Five level classification of the final grade G3

Dataset	Predictive accuracy of student academic performance		
	Decision Tree (J48)	*Online Bagging*	*Ensemble Filtering*
Mathematics	0.78	0.82	**0.95**
Portugese	0.71	0.79	**0.94**

Table 3. Predictive accuracies after using the different classification techniques

As we can see in Table 3, ensemble filtering which uses multiple classifiers to vote and eliminate noisy instances in the training data produces higher statistically significant predictive accuracies on the test data, when compared to any single model (decision tree, the best single model on this dataset) or online bagging. We used *consensus voting* across classifiers on this dataset. We use *majority voting* on CUNY dataset in the next section. Our results prove what Quinlan demonstrated in [20] that as noise level increases, removing noise from the mislabeled training instances (class noise) increases the predictive accuracy of the resulting classifier.

After filtering the noisy data on the training set, we identify and cluster groups of students in the test set using k-means bootstrapped averaging. The results are shown in Fig 3, in which we see that clustering does not work well on noisy data (Fig 3(a)). But once the instances are filtered, well defined clusters can be identified as shown in Fig 3(b).

Fig 3 (a) Clustering student data without filtering. (b) Clustering student data after filtering on Mathematics Student data

5.2 First year college student performance dataset

First year Computer Systems Technology students from the New York City College of Technology (CUNY) enrolled in 6 different semesters (Fall 2013, Fall 2014, Fall 2015 Spring 2013, Spring 2014 and Spring 2015) taking an introductory computer systems course was used for this study. The same professor taught all the semesters. Data from students who dropped the class or stopped attending the class were excluded from the study. The class has two tests, a midterm and a final. We attempt to predict the final grade given the two test scores and the midterm score. The five level classification for the final grade is as shown in Table 4.

>=80	60-80	40-60	30-40	<30
A	B	C	D	F

Table 4. Five level classification of the final grade G3

As was done in the previous section, we used ensemble classifiers to firstly eliminate noisy instances and then to predict the final grade of the students on the test set. We use a *majority vote* (which requires at least two out of the three classifiers to mislabel the class value) amongst the classifiers in eliminating the noisy instances. The predictive accuracy numbers are as shown in Table 5.

Dataset	Predictive accuracy of student academic performance		
	Decision Tree (J48)	*Online Bagging*	*Ensemble Filtering*
CST Course	0.63	0.75	**0.91**

Table 5. Predictive accuracies after using the different classification techniques

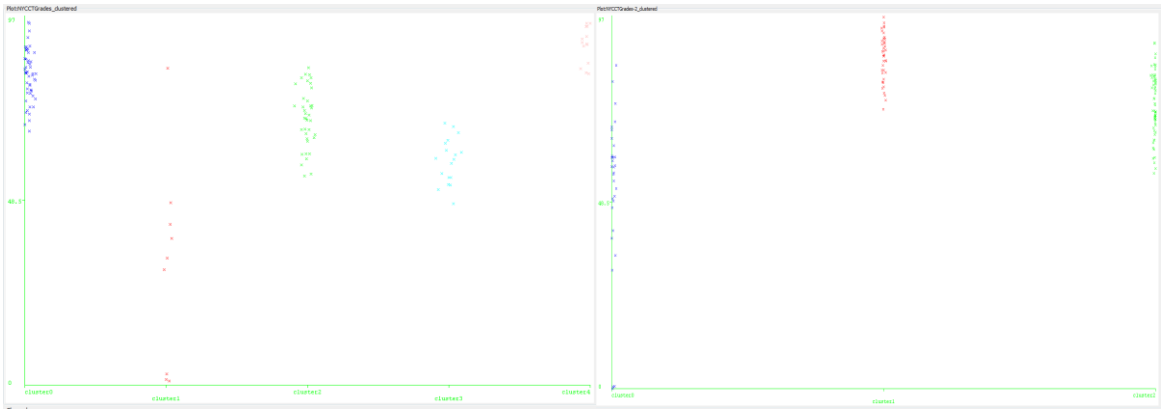

Fig 4 (a) .Clustering student data without filtering. (b) Clustering student data after using the filtered model on CUNY CST1100 data

6. Discussion and Conclusion

Resolving data quality issues in predicting student academic performance is often one of the biggest efforts in Educational Data Mining. Prior work in this area has focused on using single classifiers and no filtering on student data has been performed.

In this work, we show that student data when filtered can show a huge improvement in predictive accuracy. We compare using a single filter with ensemble filters and show that using ensemble filters works better for identifying and eliminating noisy instances. We show that both types of voting (majority and consensus) can show improvements. We have shown that this ensemble technique works for two different settings: high school data and first year college data. Although we have used decision trees, random forest and naïve bayes, other base classifier models can also be used.

In the area of clustering, we were able to show that student data can be clustered into well-defined groups based on their learning behavioral patterns. As we saw with classification, clustering after filtering noisy data on the training set produces good results. This kind of grouping of students endows instructors an anticipated estimation of their students' capabilities during team forming, in-class participation, advising and tutoring.

7. References:

1. J. Han and M. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann, 2000.
2. Alaa el-Halees, "Mining students data to analyze e-Learning behavior: A Case Study", 2009.
3. U . K. Pandey, and S. Pal, "Data Mining: A prediction of performer or underperformer using classification", (IJCSIT) International Journal of Computer Science and Information Technology, Vol. 2(2), pp.686-690, ISSN:0975-9646, 2011.
4. S. T. Hijazi, and R. S. M. M. Naqvi, "Factors affecting student‟s performance: A Case of Private Colleges", Bangladesh e-Journal of Sociology, Vol. 3, No. 1, 2006.
5. Hyuk Kwang, et al.,*Conceptual Modeling with Neural Network for Giftedness Identification and Education*, Lecture Notes in Computer Science, Volume 3611, pp. 560-538,2005.
6. Minaei-Bidgoli, B., et al., *Predicting student performance: an application of data mining methods with the educational web-based system LON-CAPA*,Proceedings of ASEE/IEEE Frontiers in Education Conference, Boulder, CO: IEEE, 2003.
7. Superby, J.F., Vandamme, J-P., Meskens, N., *Determination of factors influencing the achievement of the first-year university students using data mining methods*, Proceedings of the Workshop on Educational Data Mining at the 8th International Conference on Intelligent Tutoring Systems (ITS 2006), Jhongli, Taiwan, Pages 37-44, 2006.
8. Naplava, P. and Snorek N., *Modeling of student's quality by means of GMDH algorithms*, Modelling and Simulation 2001, 15th European Simulation Multiconference 2001, ESM'2001, Prague, Czech Republic, 2001.
9. Luan, J. and Serban, A. M., *Data Mining and Its Application in Higher Education*, Knowledge Management: Building a Competitive Advantage in Higher Education, New Directions for Institutional Research, Jossey-Bass, 2002.
10. Massa, S. and Puliafito P. P., *An application of data mining to the problem of the university students' dropout using Markov chains*, Principles of Data Mining and Knowledge Discovery, Third European Conference, PKDD'99, Prague, Czech Republic, 1999.
11. P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.
12. Oza, N. C. (2005, October). Online bagging and boosting. In Systems, man and cybernetics, 2005 IEEE international conference on (Vol. 3, pp. 2340-2345). IEEE.
13. R. S. J. D. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," Journal of Educational Data Mining, vol. 1, no. 1, 2009.
14. S. Lin, "Data mining for student retention management," J. Comput. Sci. Coll., vol. 27, no. 4, pp. 92-99, 2012.
15. Satyanarayana, A., & Chinchilla, R. (2016). Ensemble Noise Filtering for Streaming Data Using Poisson Bootstrap Model Filtering. In *Information Technology: New Generations* (pp. 869-879). Springer International Publishing.
16. Davidson, I., & Satyanarayana, A. (2003, November). Speeding up k-means clustering by bootstrap averaging. In *IEEE data mining workshop on clustering large data sets*.
17. Dutt, A., Aghabozrgi, S., Ismail, M. A. B., & Mahroeian, H. (2015). Clustering algorithms applied in educational data mining. *International Journal of Information and Electronics Engineering*, *5*(2), 112.

18. Wook, M., Yahaya, Y. H., Wahab, N., Isa, M. R. M., Awang, N. F., & Seong, H. Y. (2009, December). Predicting NDUM student's academic performance using Data mining techniques. In *Computer and Electrical Engineering, 2009. ICCEE'09. Second International Conference on* (Vol. 2, pp. 357-361). IEEE.
19. Etchells, T. A., Nebot, À., Vellido, A., Lisboa, P. J., & Mugica, F. (2006). Learning what is important: feature selection and rule extraction in a virtual course. In *ESANN* (pp. 401-406).
20. Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1), 81-106.

Ashwin Satyanarayana
Dr. Ashwin Satyanarayana is currently an Assistant Professor with the Department of Computer Systems Technology, New York City College of Technology (CUNY). Prior to this, Dr. Satyanarayana was a Research Scientist at Microsoft in Seattle from 2006 to 2012, where he worked on several Big Data problems including Query Reformulation on Microsoft's search engine Bing. He holds a PhD in Computer Science (Data Mining) from SUNY, with particular emphasis on Data Mining, Machine Learning and Applied Probability with applications in Real World Learning Problems. He is an author or co-author of over 20 peer reviewed journal and conference publications and co-authored a textbook – "Essential Aspects of Physical Design and Implementation of Relational Databases." He has four patents in the area of Search Engine research. He is also a recipient of the Indian National Math Olympiad Award, and is currently serving as Secretary/Treasurer of the ASEE (American Society of Engineering Education) Mid-Atlantic Conference.

Gayathri Ravichandran
Gayathri Ravichandran is currently pursuing her final year of undergraduate study in M S Ramaiah Institute of Technology, Bangalore, India. Her field of study is Computer Science, and she finds subjects like Data Mining and Artficial Intelligence intellectually simulating and satisfying. She has authored a paper titled – "Application of Genetic Algorithms for Traffic Light Control", and she is currently working on implementing an "Intelligent University Selection System" under the guidance of her professor. She is also a recipient of the Grace Hopper scholarship (GHCI) 2016.