University Recommender

A University recommendation system backed by data mining algorithms

Amitha Narasimha Murthy, Neeti Narayan, Vivekanandh Vel Rathinam
Dept. of Computer Science
University at Buffalo, The State University of New York
Buffalo, USA

Abstract — **In this project, we apply data clustering algorithms like K-means, DBSCAN and weighted-DBSCAN to develop a recommendation system to suggest universities that are sensitive to student's requirements. Query-based models that do the same job have no learning components to them. Such models have strict attributes based matching and do not consider the over-all nearness of data points to give results. We experimented with K-means and DBSCAN clustering algorithms. DBSCAN gave better results than K-means which led us to explore weighted-DBSCAN algorithm which further improved the results. The algorithm and results are explained in detail in the following sections.**

Keywords — data mining; clustering; k-means; DBSCAN; university guide; recommndation system;

I. Introduction

Recommendation systems help people in making decisions about things they are not completely knowledgeable about, based on their requirements or past preferences. There are a number of recommendation systems available in the fields of entertainment, social media, healthcare and crowdsourcing. We found that a recommendation system in the field of education is largely unexplored. Although there are a few systems that recommend about education-related decisions, they either suffer from the shortcomings of a query based model or are insufficiently implemented.

Getting a university degree is generally on every youth's mind. But during the initial stages, a potential student will have a lot of questions about how any university would suit his/her requirements. Given the fact that a high number of universities offer the same courses, it becomes increasingly difficult for students to actually apply to only a set of universities that are not only a good match for him/her but also have high chances of getting an admit from. Achieving this will reduce any unnecessary spending on applications to universities that would end up rejecting the student.

A. Novelty: A student may choose to attend a university based on factors like percentage of financial aid offered, location of the university, social life, quality of life, male-female ratio, student-teacher ratio, whether the university is private/public and a number of other factors. An important point to note here is that every student will weigh these attributes differently. So it's very important to keep this in mind while developing a university recommendation system for students. This is exactly what we are attempting to do with this project. Details about how this is achieved are explained in later sections.

B. Impact: Having a university recommendation system that takes in to account a student's need of various auxiliary factors like social life, quality of life, percentage of financial aid expected, male-female ratio, teacher-student ratio, expenses, etc., will match up the student with universities that have a real high probability of him/her getting in to. Also, arriving at the right set of universities to which the student can apply will minimize wastage of money on applications to universities that a student would end up getting rejects from or those that would not completely match his auxiliary requirements.

C. Challenges: The challenges identified during our initial phases were as follows: a) Implementation of a weighted clustering algorithm that would take in to account a student's rating of each auxiliary factor to suggest universities b) Choosing good epsilon and minimum points parameters for our implementation of weighted-DBSCAN algorithm. c) Unavailability of a dataset that captures university information in a required format d) Data Normalization and cleaning including duplicate elimination and dimensionality reduction.

D. Related work: There some internet sites/forums like Edulix[1] that claim to be a one stop resource for aspiring students with a database consisting of more than 1.25 million posts and 225,000 members. But as they say, the knowledge-sharing/recommendation of universities is mostly through word of mouth (electronically) where an admitted student comments about an aspiring student's chances of getting an admit at a particular university and also shares information about the percentage financial aid that he could expect, campus jobs available or the average expenses that are expected on an individual basis. It's quite tedious to go through hundreds of posts manually to gather information about a particular university or to wait until a senior student finds time to answer the questions that you have. Here, Universities are also rated with respect to availability of campus jobs, financial aid, expenses, housing, etc. But not much of this information is used to build a recommendation model for the student. Manual browsing is the only way this information can be used to the student's benefit.

There's also LinkedIn's University Finder [2] that works at a very high level by accepting the place where a student wants to work, the field of study and his/her dream company as inputs and suggests popular universities for that career goal.

We believe that such recommendation systems either do not consider the auxiliary factors we have already mentioned or even if they do collect this information, like in the case of Edulix, they do not make proper use of it in the process of recommendation.

II. Problem formulation

The university recommendation system can be formulated as a clustering problem. The idea is to take as input all the attributes used for clustering from the user and treating it as one of the data points that has to be clustered. The output would be the names of universities that fall in to the same cluster as the input data point. As part of the input, the user is also asked to rate certain attributes in the order of their importance to him/her. This input is later used to assign different weights to attributes during implementation that greatly influence the output.

III. Algorithm

In order to solve the problem mentioned above, we experiment with three different clustering algorithms. A) K-means B) DBSCAN C) Weighted-DBSCAN. The approach to each of these algorithms is explained below:

A. K-Means

This algorithm is the simplest of all available clustering algorithms involving a cluster assignment step followed by a centroid updating step that repeat until convergence. The output is very sensitive to initial centroids. If there are K ‘real’ clusters then the chance of selecting one centroid from each cluster is small. The application of this algorithm to solve the above stated problem resulted in different results with each run as the initial centroids would change at every run of the program.

The main issue with this approach was that a student whose interest was in Engineering would be suggested universities that offered a medical course or a mechanical course (or any other irrelevant course). This result does not make much sense. The reason for this according to our understanding was that all the input attributes were weighed equally in the implementation of K-Means. We used an existing implementation of Weka package to test our idea.

B. DBSCAN

As the results of K-Means were not convincing enough, we moved to a density based clustering method. DBSCAN uses two important parameters 1) Ɛ- neighborhood 2) Minimum points (MinPts). Ɛ-neighborhood is defined as objects that are within a radius of Ɛ from an object.

A point is a core point if it has more than a specified number of points (MinPts) within Epsilon. These are points that are at the interior of a cluster. A border point has fewer than MinPts within Epsilon, but is in the neighborhood of a core point. A noise point is any point that is not a core point or a border point.

The application of this algorithm to our dataset led to better results than K-Means. But the primary challenge we faced was to determine the correct epsilon and minPts parameters. A value of 1.3 for epsilon and 15 for minPts for our dataset performed convincingly well.

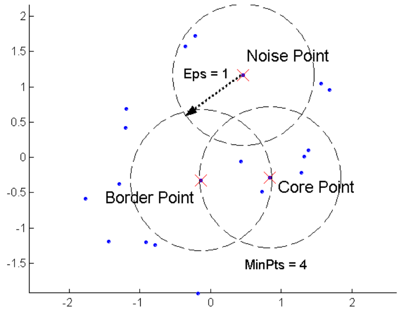


Fig: working of DBSCAN

C. Weighted-DBSCAN

As DBSCAN clustering performed better than K-means for our dataset, we thought a weighted implementation of DBSCAN would give more meaningful results. The ratings of users for attributes like financial aid, quality of life, social life, etc., from the input are translated in to weights for each of those attributes. This way the universities that are suggested to the user would be in accordance to his/her preferences for certain attributes.

We implemented the weighted DBSCAN algorithm on our own as we could not find any existing package. We used weighted Euclidean distance as a distance measure between the data points. As expected, the results were more meaningful than the results of un-weighted DBSCAN as the results now varied according to the rating (on a scale of 1-not important to 5- most important) the user gave for each attribute. Again, choosing the correct epsilon and MinPts parameters were a challenge and the values 1.5 for epsilon and 10 for MinPts worked well for us.

IV. Experimentation and Results

A. Dataset

The clustering methods discussed above are applied on the university dataset. The dataset consists of 10 features: State, Control (private or state), SAT verbal, SAT math, expenses (in thousands), percentage of financial aid, academics (a rating from 1 to 5), social (a rating from 1 to 5), quality of life (a rating from 1 to 5) and academic emphasis.

For our experimentation, we have used the university dataset from the UCI Machine Learning Repository [4]. The original dataset consists of 285 instances with missing values, duplicates and multivalued attributes.

Data Preprocessing: The original dataset in .data format has been converted to csv format. The information on missing values has been calculated and duplicates removed. Multivalued attributes are normalized by reducing them to atomic attributes for each observation. Each value in the nominal attribute is mapped to a single feature in the feature matrix. For transforming nominal attributes to numerical attributes, we have used the Dict Vectorizer implementation from the Scikit-learn package [3]. The preprocessed and cleaned dataset consists of 642 records, where each observation concerns one university with a given academic emphasis.

B. *Experimental Settings*

Huge manual effort is required to preprocess a large body of text and produce it in a machine readable format.

Machine used to run the experiments is configured with 4GB RAM, i5 processor, Windows 8 operating system.

User Interface: The user interface is designed such that the test scores and other attribute preferences of the user are taken as input. Based on user requirements, the weights are assigned dynamically.

Following are the experiments performed:

- Experiment 1

K-means clustering algorithm is used on the dataset to list the universities that best match user's requirements.

- Experiment 2

DBSCAN (Density-based) clustering algorithm is used on the dataset to list the universities that best match user's requirements.

- Experiment 3

Weighted DBSCAN (weighted Euclidean distance) clustering algorithm is used on the dataset to list the universities that best match user's requirements. Weights are assigned to attributes to reflect user preferences about the relative importance of each attribute.

C. *Results*

Each experiment is performed on a number of test samples. One such test sample is tabulated in Table 1.

The clustering results for the given test sample is shown in Fig. 1, Fig. 2, Fig. 3, Fig. 4, Fig. 5 and Fig. 6. The three clustering methods are compared using these results.

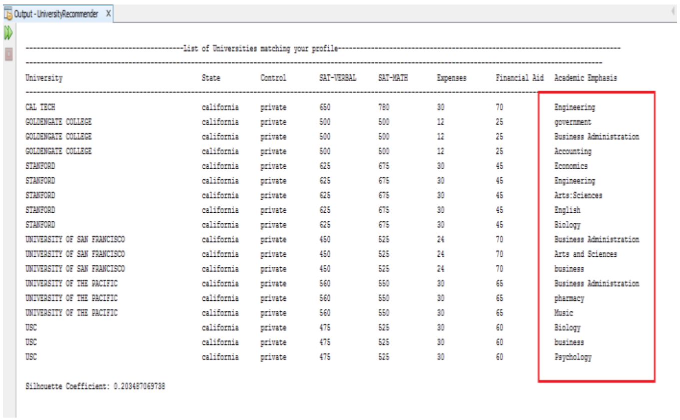
From Fig. 1 and Fig. 2, it can be seen that k-means clustering gives poor result. The user specified "Engineering" as his academic emphasis. But, k-means results show a mixture of subjects (Engineering, Accounting, Biology etc.) which is incorrect as the highest weight is assigned for the Academic Emphasis attribute. Also, it makes no sense if a university with academic emphasis on Biology is recommended to a student who wishes to pursue Engineering.

The following conclusions are made from Fig. 3 and Fig. 4: We obtain better result if the clustering algorithm used is DBSCAN compared to k-means.

From Fig. 5 and Fig. 6, it can be concluded that weighted DBSCAN gives the best result. Different students have different requirements for different attributes. The weighted DBSCAN takes these requirements into consideration and clusters accordingly. It can be seen that the results displayed are very close to the user's query. Hence, better than both k-means and unweight DBSCAN.

Table I. Tabulates a test sample

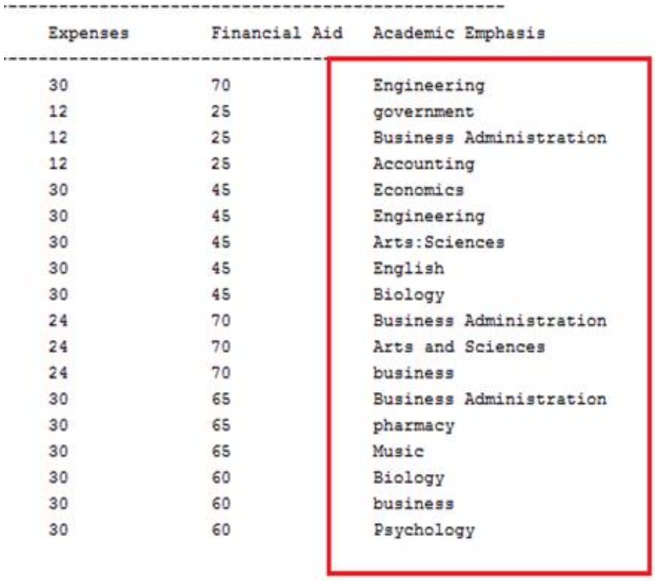
State	California
Control	Private
SAT Verbal	650
SAT Math	780
Expenses	30
Financial Aid	70
Academic	5
Social	1
Quality of Life	3
Academic Emphasis	Engineering



```
----------------------------------------List of Universities matching your profile----------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------------------------------------------------------
University                    State        Control      SAT-VERBAL    SAT-MATH     Expenses     Financial Aid  Academic Emphasis
------------------------------------------------------------------------------------------------------------------------------------
CAL TECH                      california   private      650           780          30           70             Engineering
GOLDENGATE COLLEGE            california   private      500           500          12           25             government
GOLDENGATE COLLEGE            california   private      500           500          12           25             Business Administration
GOLDENGATE COLLEGE            california   private      500           500          12           25             Accounting
STANFORD                      california   private      625           675          30           45             Economics
STANFORD                      california   private      625           675          30           45             Engineering
STANFORD                      california   private      625           675          30           45             Arts:Sciences
STANFORD                      california   private      625           675          30           45             English
STANFORD                      california   private      625           675          30           45             Biology
UNIVERSITY OF SAN FRANCISCO   california   private      450           525          24           70             Business Administration
UNIVERSITY OF SAN FRANCISCO   california   private      450           525          24           70             Arts and Sciences
UNIVERSITY OF SAN FRANCISCO   california   private      450           525          24           70             business
UNIVERSITY OF THE PACIFIC     california   private      560           550          30           65             Business Administration
UNIVERSITY OF THE PACIFIC     california   private      560           550          30           65             pharmacy
UNIVERSITY OF THE PACIFIC     california   private      560           550          30           65             Music
USC                           california   private      475           525          30           60             Biology
USC                           california   private      475           525          30           60             business
USC                           california   private      475           525          30           60             Psychology

Silhouette Coefficient: 0.203487069738
```

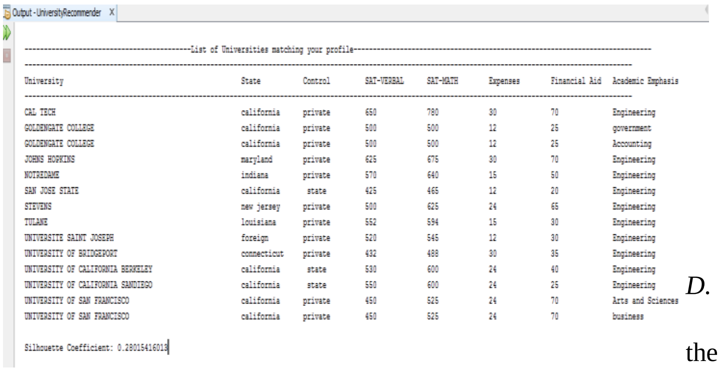
Fig. 1: k-means clustering result on the given test sample



```
-------------------------------------------------
    Expenses        Financial Aid    Academic Emphasis
-------------------------------------------------

    30              70               Engineering
    12              25               government
    12              25               Business Administration
    12              25               Accounting
    30              45               Economics
    30              45               Engineering
    30              45               Arts:Sciences
    30              45               English
    30              45               Biology
    24              70               Business Administration
    24              70               Arts and Sciences
    24              70               business
    30              65               Business Administration
    30              65               pharmacy
    30              65               Music
    30              60               Biology
    30              60               business
    30              60               Psychology
```

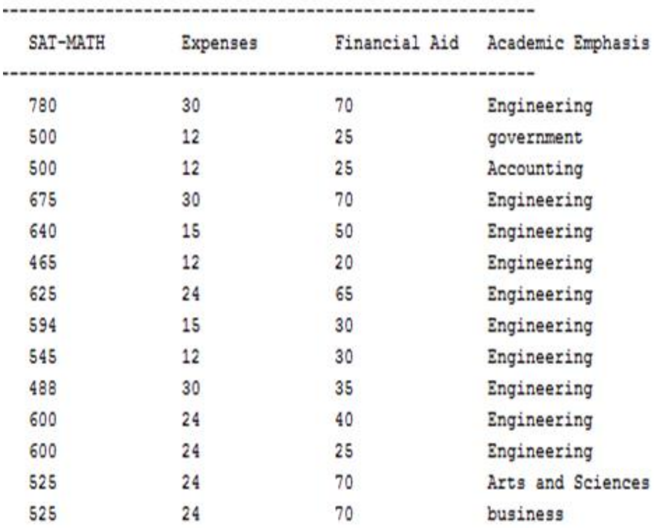
Fig. 2: Better view of Fig. 1



```
----------------------------------------List of Universities matching your profile------------------------------------------------
----------------------------------------------------------------------------------------------------------------------------------
University                          State        Control    SAT-VERBAL   SAT-MATH   Expenses   Financial Aid  Academic Emphasis
----------------------------------------------------------------------------------------------------------------------------------
CAL TECH                            california   private    650          780        30         70             Engineering
GOLDENGATE COLLEGE                  california   private    500          500        12         25             government
GOLDENGATE COLLEGE                  california   private    500          500        12         25             Accounting
JOHNS HOPKINS                       maryland     private    625          675        30         70             Engineering
NOTREDAME                           indiana      private    570          640        15         50             Engineering
SAN JOSE STATE                      california    state     425          465        12         20             Engineering
STEVENS                             new jersey   private    500          625        24         65             Engineering
TULANE                              louisiana    private    552          594        15         30             Engineering
UNIVERSITE SAINT JOSEPH             foreign      private    520          545        12         30             Engineering
UNIVERSITY OF BRIDGEPORT            connecticut  private    432          488        30         35             Engineering
UNIVERSITY OF CALIFORNIA BERKELEY   california    state     530          600        24         40             Engineering
UNIVERSITY OF CALIFORNIA SANDIEGO   california    state     550          600        24         25             Engineering
UNIVERSITY OF SAN FRANCISCO         california   private    450          525        24         70             Arts and Sciences
UNIVERSITY OF SAN FRANCISCO         california   private    450          525        24         70             business

Silhouette Coefficient: 0.28015416013
```

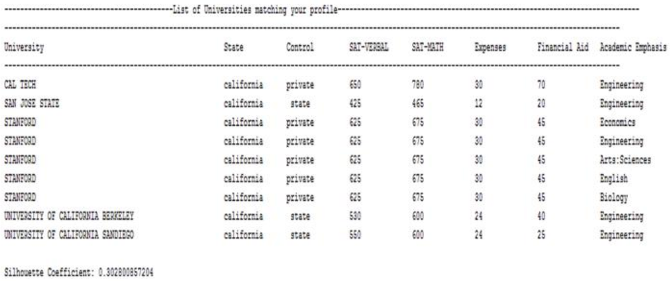
Fig. 3: Density-based clustering result on the given test sample



```
-------------------------------------------------------------
  SAT-MATH      Expenses      Financial Aid    Academic Emphasis
-------------------------------------------------------------

  780           30            70               Engineering
  500           12            25               government
  500           12            25               Accounting
  675           30            70               Engineering
  640           15            50               Engineering
  465           12            20               Engineering
  625           24            65               Engineering
  594           15            30               Engineering
  545           12            30               Engineering
  488           30            35               Engineering
  600           24            40               Engineering
  600           24            25               Engineering
  525           24            70               Arts and Sciences
  525           24            70               business
```

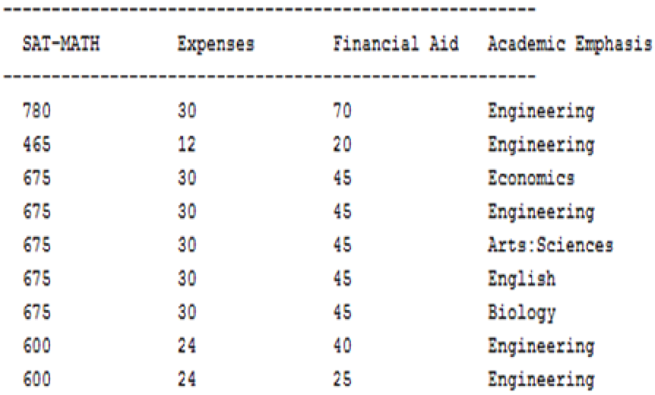
Fig. 4: Better view of Fig. 3



```
----------------------------------------List of Universities matching your profile------------------------------------------------
----------------------------------------------------------------------------------------------------------------------------------
University                          State        Control    SAT-VERBAL   SAT-MATH   Expenses   Financial Aid  Academic Emphasis
----------------------------------------------------------------------------------------------------------------------------------
CAL TECH                            california   private    650          780        30         70             Engineering
SAN JOSE STATE                      california    state     425          465        12         20             Engineering
STANFORD                            california   private    625          675        30         45             Economics
STANFORD                            california   private    625          675        30         45             Engineering
STANFORD                            california   private    625          675        30         45             Arts:Sciences
STANFORD                            california   private    625          675        30         45             English
STANFORD                            california   private    625          675        30         45             Biology
UNIVERSITY OF CALIFORNIA BERKELEY   california    state     530          600        24         40             Engineering
UNIVERSITY OF CALIFORNIA SANDIEGO   california    state     550          600        24         25             Engineering

Silhouette Coefficient: 0.302800857204
```

Fig. 5: Weighted Density-based clustering result on the given test sample



```
-------------------------------------------------------------
  SAT-MATH      Expenses      Financial Aid    Academic Emphasis
-------------------------------------------------------------

  780           30            70               Engineering
  465           12            20               Engineering
  675           30            45               Economics
  675           30            45               Engineering
  675           30            45               Arts:Sciences
  675           30            45               English
  675           30            45               Biology
  600           24            40               Engineering
  600           24            25               Engineering
```

Fig. 6: Better view of Fig. 5

D. *Validation*

Because of the unavailability of ground truth, we use only the data to measure cluster quality. Using Silhouette coefficient, an internal index, we calculate the correlation between clustering results and distance matrix. This is a way of validating the results i.e. evaluating the quality of clusters without reference to external information.

V. CONCLUSION

University recommender serves as a guide for aspiring bachelor students. We have applied three clustering algorithms (k-means, DBSCAN, weighted DBSCAN) to this problem. These algorithms were evaluated individually for many test samples.

Based on the best matches obtained, we can say that weighted-DBSCAN works better than k-means and un-weighted DBSCAN.

A. Future work: we intend to integrate Label Ranking algorithm to rank the relevance of the universities in the result of the given input. Further, we would also create an alumni/ current student dataset and use it to update the information in the existing university dataset; use it as training data for Label Ranking algorithm. The current application can be extended to other levels of education as well.

REFERENCES

[1] http://www.edulix.com/faq/

[2] https://www.linkedin.com/edu/university-finder?facets=

[3] http://scikit-learn.org/stable/modules/feature_extraction.html

[4] http://archive.ics.uci.edu/ml/datasets/University