International Journal of Scientific Engineering and Research (IJSER)
ISSN (Online): 2347-3878
Index Copernicus Value (2015): 56.67 | Impact Factor (2017): 5.156

Performance Analysis of K-Nearest Neighbour Classifier and Cosine Similarity Measure in Generating Weighted Scores for an Automated Essay Grading System

O.E. Oduntan Ph.D, BUOYE, P. A.

Department of Computer Science, The Federal Polytechnic, Ilaro, Nigeria
estherbest2000@yahoo.com, odunayo.oduntan@federalpolyilaro.edu.ng

Abstract: ***Grading of student academic performance in any examination is an inevitable exercise in educational assessment. This has become a major challenge in institutions where students' enrolment is enormous. Efficiency of the grading techniques determines the fate of the examinee. This study focuses on analyzing the performance of grading techniques used in automated essay grading system (AES): Cosine Similarity and K-Nearest Neighbor Classifier. These two techniques among others are used to compare documents and allocate similarity score, which is used to determine weighted score of a student. In this research, electronic copy of marking scheme (MS) and students' response (SR) were acquired in .txt format, preprocessed to remove stop words, vector space model was used to derive the document vectors of MS and SR. The document vectors were compared using the cosine similarity measure and the k-nearest neighbor classifier to derive the similarity score. The machine generated student score was computed as a weighted aggregate of similarity scores, where the weight is the mark assigned for each question in the marking scheme. Performance evaluation was carried out on two datasets: CMP 401(Organization of Programming Languages) and CMP 201(Operating System) courses by comparing the effectiveness of the k-nearest neighbor classifier and cosine similarity measure on weighted scores using coefficient of determination. (R^2). Results shows that Cosine Similarity Measure is more efficient in comparing document vectors for an automated essay grading system.***

Keywords: Grading, Weighted Scores, Cosine Similarity Measure, K-nearest Neighbour Classifier

1. Introduction

Assessment is the activity of measuring student learning and it could be limited to test some learning outcome or, in a broader sense, the learning process as well. Its aim might be as simple as certifying a minimum competence or as ambitious as discovering the quality of the writing and what cognitive strategies or patterns are used by the student.

According to Burstein, Leacock and Swartz (2001), these researchers were of the opinion that assessment tasks influence the direction and quality of student learning; hence it's a crucial part of the learning process, which help examiners to judge the effectiveness of teaching and as a tool to help individuals improve on their performances in the future.

Recently, technology was applied to academic assessment in order to ease the stress of evaluating student performances. The Computer Based Test which is used for conducting multiple choice questions examination has been introduced to address the challenges faced by manual methods of assessment; there are concerns for the possibility of using computer to conduct and grade essay examination.

The development of automated essay grading (AEG) system has captured the interests of researchers due to the need to assess students based on short or free text essay. The invention of Electronic Learning (E-Learning) which has greatly increased the level of enrollment of students in various schools of learning is a major challenge to academic assessment (Bresciani, 2006). The distance learning students needed to be assessed by their examiners and the manual method of assessment may not be applicable. Therefore, there was a need for an automated method of assessment.

Weighted score is key to Automated Essay Grading, The examiner assigns mark to each keyword or n-gram in the corpus to be assessed, then similarity score will be derived using an algorithm, the similarity score will then be multiplied with the mark allocated by the examiner to derive the weighted score per question.

The basic functions for an AES system are to select relevant features and to perform classification tasks to assign "scores" to essays. The classification task is based on supervised machine learning principles. Features that best fit to be assign scores in essays will be identified and evaluated. In this study the K-Nearest-Neighbor (KNN) Algorithm Classifier and the Cosine Similarity Measure were used. The KNN Classifier captures information of all training cases and classifies new cases based on a similarity. Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0° is 1, and it is less than 1 for any other angle. It is thus a judgment of orientation and not magnitude: two vectors with the same orientation have a cosine similarity of 1, two vectors at 90° have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude. Cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0,1].

2. Related Works

Streeter, 2004, gave a description of what an Automated Essay Grading system entails as shown in Figure 1, essays from the students are received in digital form as an input and it outputs a score. The AES system consists of software functions. There are functions to pre-process the essays into an internal format that is accepted by the other functions, there are functions to extract the required features from the essays and there are functions to perform the classification task to decide which score should be assigned to an essay. The classifier is based on supervised machine learning where the classes are the scores and each essay is represented by a set of features.

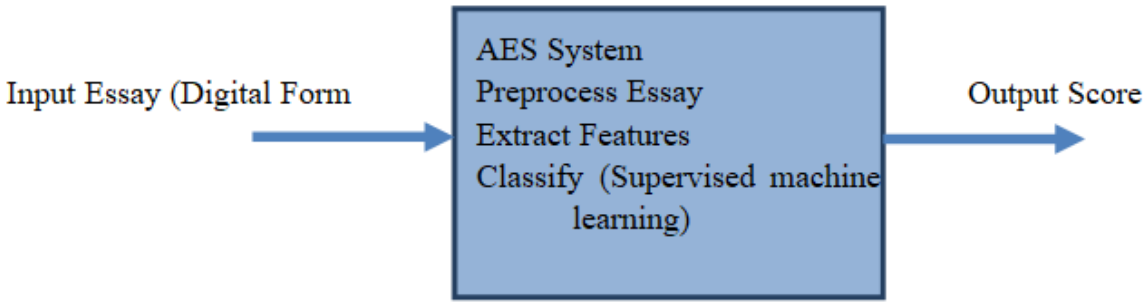

Figure 1: Features of an AES System (Streeter, 2004)

Many researchers have applied various approaches to the development of Automated Essay Grading System, majority of these systems worked by extracting a set of attributes (system-specific) and using some machine learning algorithm to model and predict the final score. Some of the works done by these researchers are summarized below;

Page's PEG used three steps to generate scores (Yang, Buckendahl, Juszkiewicz, and Bhola, 2002). First, it identifies a set of measurable features that are approximations or correlates of the intrinsic variables of writing quality (proxes); second, a statistical procedure (linear multiple regression) was used to find out the "optimal combination" of these proxes that can "best predict the ratings of human experts" (Yang et al., 2002); third, the proxes and their optimal combination are then programmed into the computer to score new essays.

Building on the strategies utilized by PEG, IEA, and e-rater, IntelliMetric™, developed by Vantage Learning, incorporates the technologies of artificial intelligence and natural language processing, as well as statistical technologies. These combined approaches are treated as a "committee of judges," and "potential scores" from these judges are calculated by using proprietary algorithms to achieve the most accurate score possible (Vantage Learning, 2003). The algorithm was capable of analyzing more than 300 semantic, syntactic, and discourse level features. IntelliMetric system functions by building an essay scoring model firstly by using samples of essays with scores already assigned by human expert raters are processed into the machine, which would then extract features that distinguish essays at different score levels. Once the model is established, it is validated by another set of essays. Finally, it was used to score new essays (Elliot, 2003).

Ade-Ibijola, Wakama and Amadi (2012) developed an expert system for automated essay scoring , they built a knowledge base system and populated it with answer templates from Lecturers on a specific course, designed an inference engine using Information Extraction (a shallow NLP technique which is an hybrid of Statistical Keyword Analysis technique and Pattern Matching with Domain Specific Dictionary), attached a Fuzzy-Module for correctness evaluation and developed two-web applications; one as a user-interface for lecturers to set their test questions and supply answer templates, and the other for Students' to write open-ended tests online and obtain an instantaneous feedback of their performance. IE was used to extract dependencies between concepts; the dependencies found are compared against the human experts to give the student's score.

Islam and Hoque (2012) developed a generalized latent semantic analysis (GLSA) based automated essay scoring system in which n-gram by document was created instead of word by document matrix of LSA, GLSA system involves two stages: The generation of training essay set and the evaluation of submitted essay using training essay set. Essays were graded first by human grader; the average value of human score is treated as training score for a particular essay. The first stage involves: pre-processing the training essay set which is done in three steps: removal of stopwords, word stemming, selecting of the n-gram index terms, computing of the SVD of n-gram by document, the n-gram by document matrix contains orthogonal, diagonal and orthogonal matrices, reduce dimensionality and determine the document similarity using the cosine formula.

Oduntan, 2016; developed an Automated Essay-type grading system using Modified Principal Component Analysis Algorithm, in the research; hardcopies of examiners marking schemes and softcopies of students' answers for two courses, Management Information System (COM 317) and Research Methodology (COM 325), offered at the Department of Computer Science, Federal Polytechnic Ilaro during 2013/2014 academic session were used as case studies. Contents of the marking schemes were transcripted into electronic form to derive a *.txt* file extension using text editor while students' answers assumed *.txt* format. The inherent stopwords and

stemming in the *.txt* document were pre-processed to address morphological variations using standard stopwords list and porters stemmer algorithm, respectively. N-gram terms were derived for each student's response and the marking schemes (MS) using the vector space model. A Document Term Matrix (DTM) was generated with n-gram terms of MS and students response representing columns and rows, respectively. Principal Component Analysis (PCA) algorithm was modified by integrating n-gram terms as input into existing PCA to derive Modified Principal Component Analysis (MPCA) algorithm. The MPCA was used to reduce the sparseness of the DTM to obtain a vector representation of the students' answers and the marking scheme. The reduced vector representation of the students' answers was graded according to the mark assigned to each question in the marking scheme using cosine similarity measure. The developed Automated Essay-Type Grading System (AETGS) was implemented in Matrix Laboratory 8.1 (R2013a). Performance of the MPCA was compared with existing PCA to determine the effectiveness of AETGS on the grading of students' answers in COM 317 and COM 325 courses, in terms of Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Pearson correlation coefficient(r) and coefficient of determination (R^2).

It was observed that although these systems perform a similar task, each of them uses different combination of methodologies for attribute extraction and model building. The prevailing methodology used in described systems is Natural Language Processing (NLP).This is consistent with our argument that NLP strongly influenced the development of AES systems in the last 20years.

3. Research Method

In this study, a graphics user interface was designed to enable students to attempt exam and supply theoretical based answers. Questions to be attempted were pulled from the database while the user is expected to supply answer in the input box, at the end of the exam, grading is been done by the system. The new system is developed with the PHP (Hypertext Preprocessor) and other web client and server side languages including JavaScript, HTML 5.0, and MySQL as the database. The weighted score was derived using K Nearest Neighbor Classifier (KNN) and Cosine Similarity Measure. Performance evaluation of the KNN Classifier and the Cosine Similarity Measure on Automated Essay Grading (AEG) was determined using coefficient of determination (R^2). The implementation was performed using MATLAB R2013a development tool.

3.1 K-Nearest-Neighbor-Algorithm Classifier

KNN is a non-parametric supervised learning technique in which the data point classified to a given category with the help of training set. In simple words, it captures information of all training cases and classifies new cases based on a similarity. In this study, the response of the student to each question is being compared to determine the similarity measure. Predictions are made for a new instance (x) by searching through the entire training set for the K most similar cases (neighbors) and summarizing the output variable for those K cases.

In the k-nearest-neighbor classification we find the k essays in the training collection that are most similar to the test essay then receives a score which is a similarity-weighted average of the grades that were manually assigned to the k retrieved training essays. During the implementation, the similarity between a test essay and the training set was measured by in query retrieval system using a probabilistic retrieval system. The entire test document was submitted as a query against a database of training documents. This resulting ranking score or belief score was used as similarity metric. The parameter k, the number of top-ranked documents over which to average was tuned on the training set. We chose the value that yielded the highest correlation with the manual ratings (Gujarati and Porter 2009).

3.2 Cosine Similarity Measure

Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0° is 1, and it is less than 1 for any other angle. It is thus a judgment of orientation and not magnitude: two vectors with the same orientation have a cosine similarity of 1, two vectors at 90° have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude. Cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0,1].

In this study, documents to be compared are the generalised document vector of the marking scheme and the students' response. The generalised document vector was compared using the cosine similarity to generate the similarity score. The similarity score was multiplied with the mark allocated by the examiner to derive the weighted score per question. A summation of the weighted score was used to determine the machine score. Equation (1) shows the cosine similarity formula (Singhal, 2001).

$$CosSim(Lj, M) = \frac{\sum_{i=1}^{t}(l_{ij} * m_i)}{\sqrt{\sum_{i=1}^{t} l_{ij}^{\,2} * \sum_{i=1}^{t} m_i^{\,2}}} \qquad 1$$

where l_{ij} denotes the weight of the ith term in the essay-type marking scheme document term matrix (Lj) and m_i denotes the weight of the ith term in the essay-type student script document term matrix(M).

Cosine similarity gives a useful measure of how similar two documents are likely to be in terms of their subject matter. One of the reasons for the popularity of cosine similarity is that it is very efficient to evaluate, especially for sparse vectors, as only the non-zero dimensions need to be considered. (Singhal, 2001)

3.3 Research Procedure

The approach used in this study entails; acquit ion of data; preprocessing of text, document representation and feature extraction, document similarities and generating of weighted scores. These steps are further explained below:

3.3.1 Data Acquisition

The data used for this study comprises of electronic copy of students' responses and electronic copy of examiners' marking scheme which were captured in .txt. In this research CMP 401: Organization of Programming Languages and CMP 201: Operating System; was used a case study.

3.3.2 Text Pre-processing

The Preprocessing operation performed includes the removal of stopwords and stemming. Stopwords refers to frequently occurring words or text in a sentence; it includes words such as "is", "which", "to", "in" etc. Stemming, is also called conflation, is a component of text processing that captures the relationships between different variations of a word. More precisely, stemming reduces the different forms of a word that occur because of inflection (e.g., plurals, tenses) or derivation (e.g., making a verb to a noun by adding the suffix -ation) to a common stem. In general, using a stemmer for search applications with English text produces a small but noticeable improvement in the quality of results. In applications involving highly inflected languages, such as Arabic or Russian, stemming is a crucial part of effective search. In this study, the algorithmic stemmer is the suffix-s stemmer. This kind of stemmer assumes that any word ending in the letter 's' is plural, so cakes → cake, dogs → dog.

3.3.3 Document Representation and Feature Extraction

Vector space model otherwise known as term vector model is an algebraic model for representing text documents as vectors of identifiers, such as index terms. It is used in information filtering, information retrieval, indexing and relevancy rankings. In this study, unique words were extracted to possess 'coordinates' for vector space model. N-gram terms were derived for each student's response and the marking schemes (MS) using the vector space model. Document term matrix is a representation of the certain text in the space which is built in according to Vector Space Model.

Feature extraction involves the transformation of input data into a set of features. It can be done by the process of dimensionality reduction. In text processing, feature extraction is performed by reducing sparseness of a document vector. Sparseness is the sequencing out of any zero elements in a matrix from a full matrix. In this study, the modified principal component analysis algorithm (MPCA)(oduntan et. al., 2018) was used.

3.3.4 Document Similarities and Generating Weighted Score

Documents similarities deals with the comparison of two separate documents to examine the level at which the items of one document matches the other. Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them.

K-Nearest Neighbour Classifier is used in search applications when looking for "similar" items; it measure similarity by creating a vector representation of the items, and then compare the vectors using an appropriate distance metric.

In this study, the Cosine Similarity measure and KNN were used on document vector derived using the vector space model to generate a similarity score. Automated Weighted Score is generated according to the mark assigned to each question in the marking scheme using cosine similarity measure and the k-nearest neighbor.

3.4 Performance Evaluation

The effectiveness of KNN Classifier and Cosine Similarity Measure compared with the human assessor was determined making use of the coefficient of determination (R^2).

3.4.1 The coefficient of determination (R^2)

This indicates how well data points fit a statistical model. It is a measure of the model's predictive power and it can be useful for evaluating the statistical importance of the independent predictor variables. It also provides a measure of how the observed outcomes are replicated by the model. R^2 is derived by the formula in Equation (3). The coefficient of determination value ranges from 0 to 1. The better the linear regression fits the data in comparison to the simple average, the closer the values of R^2 is to 1, if $R^2 = 1$, it indicates that the regression line perfectly fits the data. It gives information about the goodness of fit of a model (Gujarati and Porter, 2009).

$$R^2 = \left(\frac{\sum_{i=1}^{n} (O_i - \bar{O})(P_i - \bar{P})}{\sqrt{\sum_{i=1}^{n} (O_i - \bar{O})^2} \sqrt{(P_i - \bar{P})^2}} \right) \quad 3$$

where O_i is the actual estimate or value of the original assessor, P_i is the predicted value generated by the system, $\bar{O}$, $\bar{P}$ are their respective mean and n is the number of data.

4. Results and Discussion

In this study, the weighted score for Automated Essay Grading System was implemented in Matrix Laboratory 8.1 (R2013a). Performance of K-Nearest Neighbour

Clasifier compared with Cosine Similarity Measure in determining Weighted Score for Automated Essay Grading System was determined using coefficient of determination (R^2). The results are presented as follows:

4.1 Automated Weighted Score Generated using K-Nearest Neighbour and Cosine Similarity Measure:

Table 1 represents the automated weighted score generated using k nearest neighbour and cosine similarity measure for thirty five students of CMP 401: Organization of Programming Languages.

Table 1: Weighted Score

std id	Human Assessor's score(X)	KNN Classifier	Cosine Similarity Measure
std 1	25	18.7706	28.2227
std 2	24	18.9724	28.2227
std 3	23	19.5618	28.3704
std 4	27	18.5527	27.7114
std 5	26	18.6027	27.7114
std 6	25	18.6027	27.7114
std 7	24	20.9833	29.4594
std 8	23	18.9724	27.7968
std 9	25	18.2284	27.653
std 10	24	18.6027	27.5711
std 11	22	18.3803	27.653
std 12	23	17.0361	27.7968
std 13	25	17.0361	27.7968
std 14	24	18.5821	27.7249
std 15	24	18.7728	27.643
std 16	25	18.2284	27.653
std 17	25	17.0361	27.7968
std 18	25	17.0361	27.7968
std 19	25	18.3803	27.7968
std 20	23	18.2284	27.653
std 21	23	18.7545	27.5711
std 22	24	17.0361	27.7968
std 23	25	18.7545	27.5711
std 24	25	20.9833	29.4594
std 25	26	18.9025	27.8632
std 26	26	18.6027	27.7114
std 27	25	18.6027	27.7114
std 28	24	20.9833	29.4594
std 29	23	18.9724	27.7968
std 30	26	18.2284	27.653
std 31	25	18.6027	27.5711
std 32	25	18.3803	27.653
std 33	23	17.0361	27.7968
std 34	25	18.2284	27.653
std 35	25	17.0361	27.7968

4.2 Coefficient of Determination (R^2) Results,

Table 2 shows that R^2 = 0.45 and 0.5 for CMP 401 and CMP 201 for Automated Weighted Score generated using K-nearest neighbor classifier respectively and R^2 = 0.60 and 0.75 for Weighted Score using Cosine Similarity Measure.

Table 2: Coefficient of Determination Results

Dataset	KNN	Cosine Similarity
CMP 401	0.45	0.6
CMP 201	0.5	0.75

5. Conclusion

Table 1 and 2 shows that Cosine Similarity measure is efficient in comparing two document vectors especially in text processing. Cosine Similarity is a *measure* of how similar two documents are likely to be in terms of their subject matter. One *advantage of cosine similarity* is its low-complexity, especially for sparse vectors: only the non-zero dimensions need to be considered.

In determining weighted score for Automated Essay Grading System, it is however recommended that the Cosine Similarity Measure is used.

References

[1] Ade-Ibijola, A.O., Wakama, I. and Amadi, J.C. (2012): An Expert System for Automated Essay Scoring(AES) in Computing using shallow NLP Techniques for Inferencing. International Journal of Computer Applications Vol. 51, pp. 37-45.

[2] Attali Y. and Burstein, J. (2006): Automated essay scoring with e-raterR v. 2. *The Journal of Technology, Learning and Assessment*, 4(3).

[3] Bresciani, M. J. (2006): Outcomes-Based Academic and Co-curricular Program Review: A compilation of Intitutional Good Practices Sterling, VA Stylus, pp.50-60.

[4] Burstein, J., Leacock, C., Swartz, R. (2001): Automated evaluation of essay and short answers. In M. Danson (Ed.), *Proceedings of the Sixth International Computer Assisted Assessment Conference*, Loughborough University, Loughborough, UK, pp. 55

[5] Elliot, S. (2003): Intellimetric™: From here to validity. In M.D. Shermis & J.C. Burstein (Eds.), *Automatic essay scoring: A cross-disciplinary perspective* (pp. 71-86). Mahwah, NJ: Lawrence Erlbaum Associates.

[6] Gujarati, D.N., Porter, D.C. (2009): Basic Econometrics (Fifth ed.). New York; McGraw-Hill/Irwin, pp. 73-78. ISBN 978-0-07-337577-9.

[7] Islam, M.M., Hogue, A.S.M.L.,(2012): "Automated Essay Scoring Using Generalized Latent Semantic Analysis" *Journal of Computers vol 7 no 3 pp.616-626.*

[8] *Oduntan, O. E., Adeyanju, I.A., Falohun, A.S. and Obe, O.O.(2018)* A Comparative Analysis of Euclidean Distance and Cosine Similarity Measure for Automated Essay-Type Grading System *Journal of Engineering and Applied Science 13(11) 4198-4198-4204 ISSN: 1816-949X c Medwell Journals.*

[9] Oduntan, O. E., et. al., (2016) A Modified Principal Component Analysis Approach to Automated Essay-Type Grading System, Book of Proceedings of Future Technology Conference 2016, ISBN: 978-1-5090-4171-8/16/$31.00,IEEE

[10] Page, E.B. (1996): Grading essay by computer: Why the controversy? Handout for *NCME Invited Symposium*.
[11] Singhal, A (2001): "Modern Information Retrieval: A Brief Overview". Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24(4): pp.35-43.
[12] Streeter, L., Psotka, J., Laham, D., and MacCuish, D. (2004). The credible grading machine: Essay scoring in the DOD [Department of Defense]. Retrieved from http://www.k-a-t.com/papers/ essayscoring.pdf
[13] Yang, Y., Buckendahl, C. W., Juszkiewicz, P. J., and Bhola, D. S. (2002): A review of strategies for validating computer automated scoring. Applied Measurement in Education, 15(4), pp.391-412