Journal of Theoretical and Applied Information Technology
15th July 2016. Vol.89. No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

MINING EDUCATIONAL DATA TO ANALYZE TEACHING EFFECTIVENESS

[1,2]ANWAR ALI YAHYA, [2]ADDIN OSMAN, and [2]MOHAMED KHAIRI

[1] Faculty of Computer Science and Information Systems, Thamar University, Thamar, Yemen

[2] College of Computer Science and Information Systems, Najran University, Najran, Sauidi Arabia

E-mail: {aaesmail, aomaddin, makhairi}@nu.edu.sa

ABSTRACT

Teaching effectiveness is a multidimensional construct in which teacher questioning skill is one of its key indicators. This paper explores the feasibility of applying data mining techniques to analyze teaching effectiveness using a data set of teachers' questions. More specifically, the performance of nine data mining techniques is investigated for the classification of teachers' classroom questions into the six Bloom's cognitive levels. To this end, a data set has been collected, annotated with Bloom's cognitive levels, transformed into a suitable representation, and the data mining techniques have been applied. The results confirm the feasibility of applying data mining techniques to analyze teaching effectiveness. Moreover, the results show that the performances of these techniques vary, depending on the sensitivity of each technique to the curse of dimensionality problem. Most remarkably, Support Vector Machine and Random Forest techniques show a striking performance, whereas Adabost and J48 show a sharp deterioration in their performances as the dimensionality grows.

Keywords: *Data Mining, Teaching Effectiveness, Machine Learning, Curse of Dimensionality, Learning Analytics*.

1. INTRODUCTION

In the field of education, teaching is a multidimensional process involving a number of separable dimensions that are difficult to evaluate [1, 2]. Nevertheless, the concept of teaching effectiveness, defined as the ability of a teacher to inculcate knowledge and skills in students, as well as changing their behavior [3], is commonly used to evaluate the quality of teaching using several indicators. In this regard, it is widely acknowledged that teacher's questioning skill is a key indicator of teaching effectiveness. Formerly, Hamilton was quoted as saying "questions are the core of effective teaching" [4] and Ornstein stated that "the essence of good teaching is related to good questioning" [5]. Presently, questioning is still the most frequent instructional strategy used for variety of purposes: to develop interest and motivate students, to evaluate students, to develop critical thinking skills and, to review and summarize previous lessons.

Realizing the key role of teachers' questioning skill, it has been used in many practical protocols developed to evaluate teaching effectiveness [6]. For example, in the framework of teaching, one of the most widely used protocols which consists of four domains broken down into 22 components, teachers' questioning skill is one of its main components.

The evaluation of teacher's questioning skills can be performed by analyzing the classroom questions using taxonomy-based analysis [7]. It applies a specific question taxonomy to classify the questions into different types to elicit information such as the level of thinking they invoke, how the questions align with the goals of the lesson, and comparing between teachers, teaching the same lesson, with different levels of experience … etc. For this purpose, many question taxonomies have been developed, among which Bloom's Taxonomy [8] is the most salient one. It was developed by Benjamin Bloom in his efforts to classify thinking behaviors into three domains: cognitive (mental skills), affective (growth in feelings or emotional areas) and psychomotor (manual or physical skills). The cognitive domain has gained much attention because of its applicability in secondary and postsecondary education. Under the cognitive domain, Bloom identified six different levels of learning known as Bloom's Cognitive Levels (BCLs) as shown in Figure 1. The levels were arranged in a hierarchical form, starts with

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

knowledge (recalling information), which is the lowest and the simplest level of cognition, and moves to comprehension, application, analysis, synthesis and ends up with evaluation (making judgment about something), which is the highest and most complex level of cognition. The levels were cumulative; to master any level a learner needs to master the earlier levels.

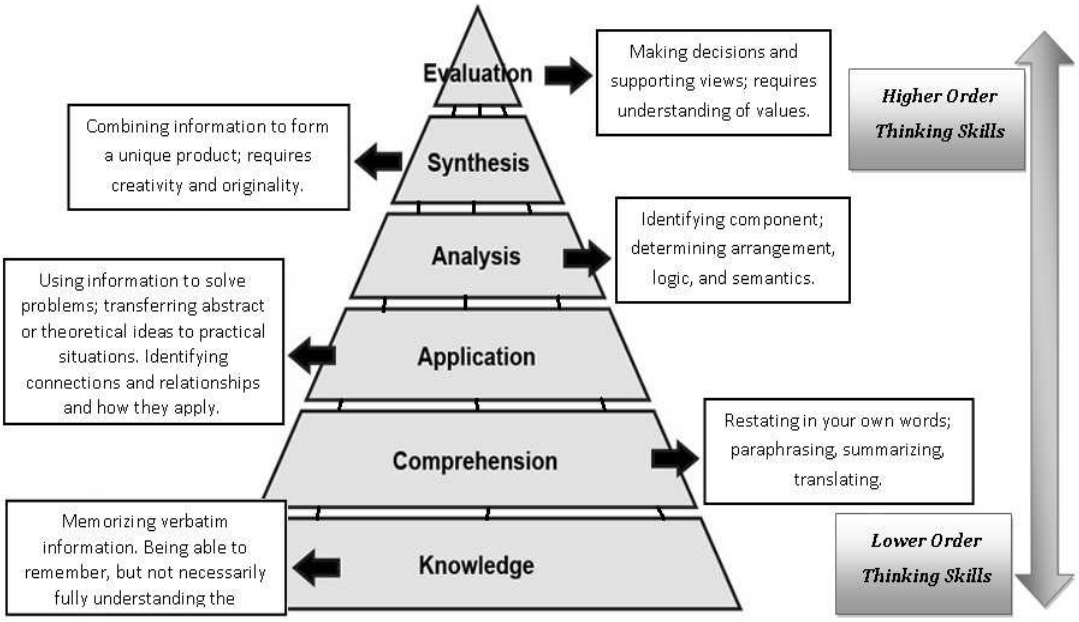

Figure 1 : The Six Cumulative Cognitive Levels of Bloom's Taxonomy

Motivated by the importance of analyzing teachers' classroom questions for evaluating teaching effectiveness, this paper investigates the feasibility of using data mining (DM) techniques to automate this task. The rest of the paper is organized as follows: Section 2 presents the current state of the art, Section 3 describes the general methodology of applying DM techniques to questions classification, Section 4 presents the results, Section 5 discusses them, and finally Section 6 concludes this work.

2. LITERATURE REVIEW

In the literature of education, teachers' questions have received a great deal of researcher's consideration as they provide a valuable source of information to study various aspects of teaching and learning. In fact, there is ample research works focusing on analyzing teachers' classroom questions for different purposes. As researches have indicated that teachers' questioning skill is typically less effective than it could be [9-12], measuring teaching effectiveness is a key purpose of analyzing teachers' classroom questions. Other purposes of can be found in [13], which provides a comprehensive survey of a large number of works focusing on different purposes. According to this survey, a good number of researches are concerned with the relative effects on student learning produced by questions at higher and lower cognitive levels, while other researches focus on the relationship between teacher wait-time and learning outcomes. Other purposes includes manipulating the placement and timing of questions during lessons, using probing, redirection and reinforcement strategies, training students in responding to higher cognitive questions and making inferences.

As it has been pointed out earlier, the taxonomy-based analysis approach analyzes teacher's questions by classifying them into different categories, identified by question taxonomy (e.g. the six cognitive levels of Bloom's Taxonomy). It is at this particular point where DM techniques can be applied to automate the assignment of a specific BCL to the question based on its content. In DM, the classification of text into pre-specified classes based on its content can be found in many domains [14]. Although in all domains the same techniques are applied, it is widely accepted fact that the specific characteristics of the domain (e.g. corpus type, structure, size, language, etc.) highly affect their application and lead to different results and conclusions [14, 15]. For instance, in the traditional text classification problem, document classification, the corpus is a set of textual documents, where each document consists of hundred words and the task is to classify a document based on its content. In this

Journal of Theoretical and Applied Information Technology
15th July 2016. Vol.89. No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

corpus, each class is often associated with a number of words that are indicative of the class. Since text documents often contain at least a few hundred words, a number of indicative words will likely appear in each document. It is thus relatively easy for the standard DM techniques to find most of these words even with a small amount of training data [16]. A weighted average of the words will give a good estimate of whether a document belongs to a certain class or not. Another example is spam filtering in email management systems. According to [17], spam filtering is a text classification problem with several distinguishing characteristics including skewed and changing class distributions, unequal and uncertain misclassification costs of spam and legitimate messages, complex text patterns and concept drift (a change in a target concept, such as terms indicative of spam messages).

With regard to the current question classification problem, it has several distinguishing characteristics that make it a new domain for text mining application. It is a form of short text classification [15, 18], where a corpus consists of a set of questions and each question often contains only a very small number of words. It is therefore very difficult for the standard DM techniques to find many indicative words for a class from the training questions[19]. In fact, many terms that appear in the test questions do not occur in the training questions at all, which cause a data-sparsity problem, a well-known problem in natural language processing.

Although question classification in education-related context, has appeared in several works [20-23], these works specifically focus on written exam questions and the classification of questions is based on the levels of difficulty. Conversely, the current work has several distinguishing aspects: First, the classification of the questions is based on the cognitive levels of Bloom's taxonomy. Second, the current work focuses on oral questions asked by teacher during classrooms. Third, the current work uses different questions representation methods. Fourth, none of the previous works investigated the DM techniques experimented in the current work.

In summary the contribution of this work is twofold: For education, it represents an original attempt to automate an important educational practice of analyzing teacher's classroom questions and for DM, it provides a new domain of DM application.

3. METHODOLOGY

This section describes the methodology of applying DM techniques to analyze teaching effectiveness.

3.1 Data Collection

A data set of teachers' classroom questions has been collected from a set of lectures of a number of courses in computer science program at Najran University. The procedure of questions collection was based on lecturers, who were asked to keep records of questions they are intending to ask their students in classrooms. The total number of collected questions is 7348 questions in English language, which is the adopted medium of instruction in the program. After the collection process, annotation of questions using the six levels of BCLs was carried out by the lectures. Besides that another cycle of annotation was carried out with a help of pedagogical expert. In this process a kappa statistic is used to measure the agreement between the two cycles of annotation and the obtained kappa, 0.82, indicates a very good agreement. The distribution of the questions data set over BCL's is shown in Figure 2, in which it can be observed that the distribution varies among BCLs with knowledge has the highest number of questions and analysis has the lowest number. To avoid the potential effect of skewed data, an equal number of questions for each BCL is selected, where each BCL has 1000 questions.

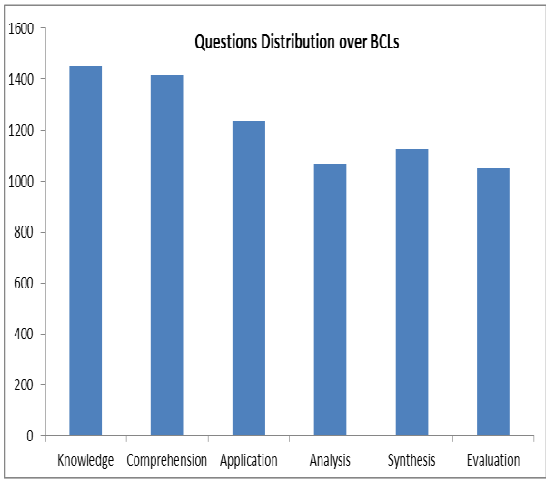

Figure 2 : Questions Distribution over BCLs

It worth mentioning that, this data set has been used in several researches [24, 25]. Table 1 shows samples of questions and their corresponding BCLs classes.

Journal of Theoretical and Applied Information Technology
15th July 2016. Vol.89. No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1: Questions Data Set Examples

BCL	Question Example
Knowledge	Can you define the fourth normal form?
Comprehension	Can you explain how we can implement these operations on an array?
Application	How can we use the table generated by the dynamic programming algorithm?
Analysis	Can this algorithm be classified as a stable algorithm?
Synthesis	How to develop a bottom-up version of merge-sort algorithm?
Evaluation	Can you decide which grammar should be used?

3.2 Data Preprocessing

Since the textual data in its raw form is not suitable for most DM techniques, the questions of the data set are transformed into a vector space representation, which is suitable for most DM techniques. Figure 3 shows the preprocessing steps applied to the questions data set.

3.2.1 Tokenization (term extraction)

Tokenization involves breaking text stream into meaningful tokens, also called terms, such as symbols, phrases, or words. For the questions data set, the token is defined as a maximal sequence of nonblank characters, where all letters are in lower case form. Therefore, the tokenization involves reducing of the question text to lower case characters and generating the set of terms of each question.

3.2.2 Useless term removal

Useless terms are groups of words, which are not informative for text classification. A well-known group of these words is stop words (the most frequently used words) such as pronouns, prepositions, conjunctions, etc. In this work, the stop words as defined in [26], have been removed. Besides that, the following three groups of useless terms have also been removed.

- Punctuations: all types of punctuations.
- Numbers : terms consisting purely of digits.
- Low frequency terms: terms with frequency less than three.

3.2.3 Stemming

Stemming is the process of reducing the inflected words to their root or base form, known as stem. It is usually performed by removing any attached suffixes and prefixes (affixes) from terms to improve the classification accuracy. For the questions data set, a porter stemmer has been used [27]. It has five steps, and within each step, rules are applied until one of them passes the conditions. If a rule is accepted, the suffix is removed accordingly, and the next step is performed. The resultant stem at the end of the fifth step is returned.

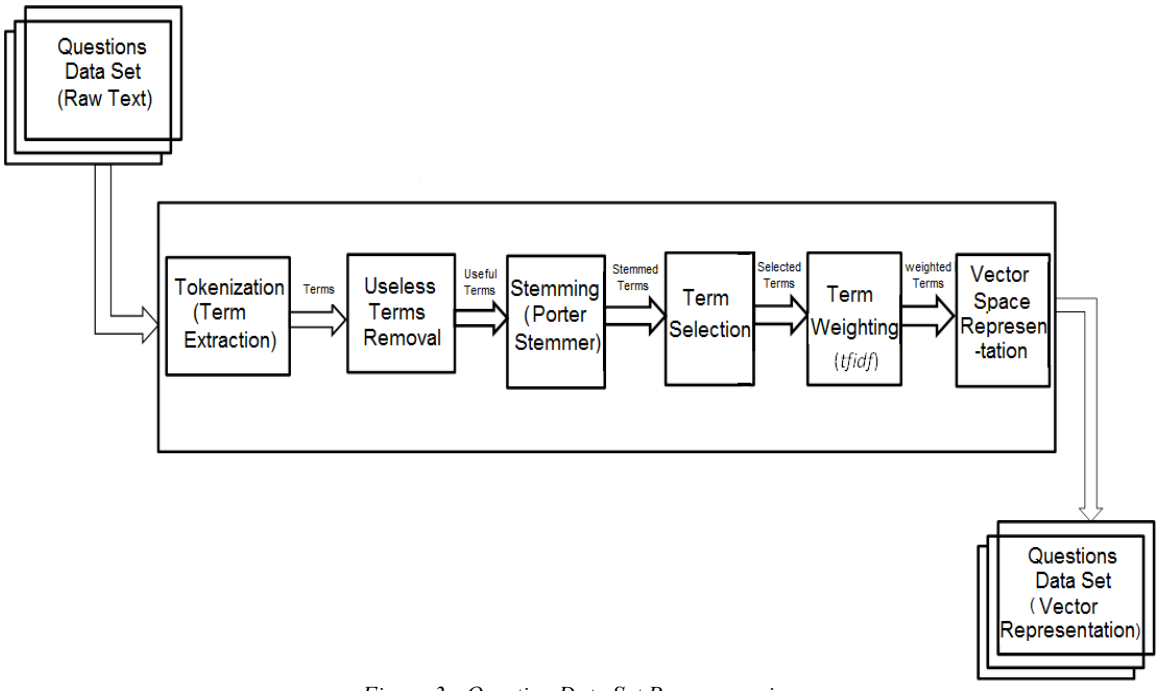

Figure 3 : Question Data Set Ppreprocessing

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.2.4 Term selection

In this step, a feature selection approach is applied to select from the original term set (a set containing all the terms from questions) a subset such that only the most representative terms are used. In this work, a filter approach based on Term Frequency (*TF*) has been applied, due to its ability to take into account the multiple appearance of a term in questions [28].

3.2.5 Term weighting

Term weighting is the process of assigning to each term a numerical value based on its contribution in distinguishing classes. In its simplest form, it can be a binary weight, where 1 denotes presence and 0 absence of the term. However, non-binary weight forms are most often used. In this work, a non-binary weight, in the form of the standard *tfidf* (term frequency inverse document frequency) [14] has been applied. Frist the *tfidf* of each term t_k in a question q_j is computed as follows:-

$$tfidf(t_k,q_j) = tf(t_k,q_j) \times \log\left(\frac{N(Tr)}{N(q_{t_k},Tr)}\right) \quad (1)$$

where $tf(t_k, q_j)$ denotes the number of times t_k occurs in q_j , $N(Tr)$ denotes the number of questions in the training set Tr and $N(q_{t_k},Tr)$ denotes the number of q_{t_k} questions, in Tr , where the term, t_k , occurs. The term weight is then computed as follows

$$w(t_k,q_j) = \frac{tfidf(t_k,q_j)}{\sqrt{\sum_{k=1}^{T}\left(tfidf(t_k,q_j)\right)^2}} \quad (2)$$

where T is the number of unique terms in Tr.

3.2.6 Vector space representation

In this step each question q_j is represented as a vector of terms weights $<w_{1j}, ..., w_{Tj}>$, where $0 \leq w_{kj} \leq 1$, represents the weight of term t_k in q_j.

3.3 Classifiers Learning

In this phase, DM technique is applied to learn a classifier of a given BCL class from its training set. The main idea is that given N-dimensional data instances in the training set divided into instances labeled with the given BCL and instances labeled with other BCLs, DM technique is applied to learn a binary classifier for that BCL.

3.4 Classifiers Evaluation

The performance of the learnt BCL classifier can be evaluated using several measures computed from contingency table. The contingency table of a given BCL classifier consists mainly of the following values:

- True Positive (TP): number of questions a classifier correctly assigns to the BCL class.
- False Positive (FP): number of questions a classifier incorrectly assigns to the BCL class
- False Negative (FN): number of questions that belong to the class but the classifier does not assign to the BCL class.
- True Negative (TN): number of questions a classifier correctly does not assign to the BCL class.

From the above values, the following are the common measures used to evaluate the performance of a given BCL classifier.

- Precision (P): the probability that if a question is classified under BCL, the decision is correct. It can be viewed as the degree of soundness of the classifier with respect to the BCL. That is

$$P = \frac{TP}{TP + FP} \quad (3)$$

- Recall (R): the probability that if a random question ought to be classified under BCL, this decision is taken. It can be viewed as the degree of completeness of the classifier with respect to the class. That is

$$R = \frac{TP}{TP + FN} \quad (4)$$

Normally, the above P and R measures are combined into the so called F_β measure, which is the harmonic mean of recall and precision that is defined, for β=1.0, as follows

$$F_{1.0} = \frac{2RP}{R + P} \quad (5)$$

Based on the above measure, the performance across a set of BCLs classifiers can be measured by Macro-Average (unweighted mean of performance across all classes) and Micro-Average (performance computed from the sum of per-class contingency tables). In this work, Macro-Average of F_1 is used.

4. RESULTS

This section presents the results of applying the following DM techniques: k-Nearest Neighbor (kNN), Naïve Bayes (NB), Support Vector

Journal of Theoretical and Applied Information Technology
15th July 2016. Vol.89. No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Machine (SVM), Rochio Algorithm (RA), C4.5 decision tree algorithm (J48), a rule based DM method (JRip), Adaptive Boosting (AdaBoost), Bayesian Networks method (BNs), and Random Forest using Weka DM tool [29]. For each technique, several experimental cases have been carried out, where in each case different numbers of terms have been selected. Figure 4 depicts the performance of DM techniques in terms of Macro-Average F_1 obtained for all experimental cases. It is obvious that most of the DM techniques except J48, AdaBoost, and JRip have similar curve pattern. For J48 and AdaBoost, their performances decline rapidly as the number of selected terms increases, whereas, JRip curve tends to be unstable with higher performance. Among DM techniques, the performances of RA and SVM techniques is comparable when the number of selected terms falls in the range between 260 and 370. However, for most of the remaining experimental cases, RA outperforms SVM. On the other hand, the performances of NB and kNN are comparable and lower than RA and SVM, yet higher than BNs.

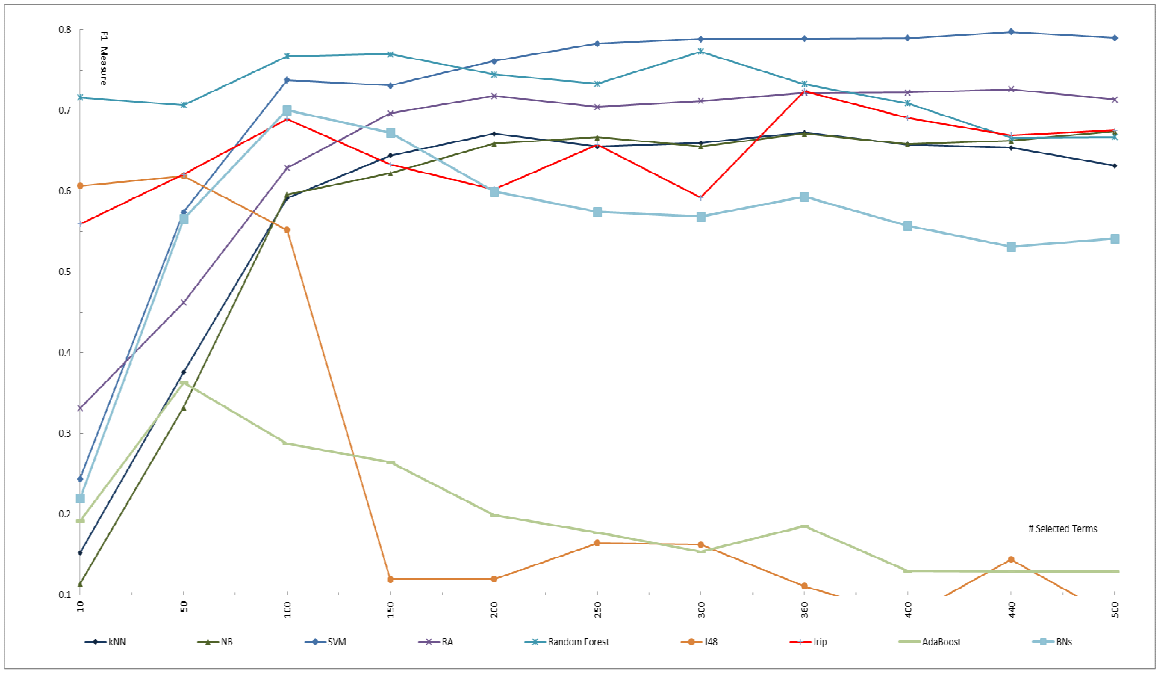

Figure 4 : Performances of DM Techniques

In Table 2, a comparison between DM techniques in terms of Average F_1 over all experimental cases for each BCL is given. The Macro-Average F_1 given in the last row of the table indicates that the average ability of Random Forest technique to build a question classification system is the highest and SVM technique shows a relatively close ability. Concerning the remaining DM techniques, the ability of J48 and AdaBoost are the lowest, the ability of kNN, NB, and BNs are higher, where kNN and NB are comparable, and the ability of RA and JRip are higher and comparable.

Table 2 : Performances of DM Techniques (Average Performance)

ML / BCL	kNN	NB	SVM	RA	Random Forest	J48	Jrip	AdaBoost	BNs
Knowledge	0.591	0.437	0.792	0.650	0.841	0.214	0.743	0.195	0.480
Comprehension	0.641	0.633	0.750	0.675	0.766	0.233	0.680	0.264	0.588
Application	0.452	0.455	0.663	0.545	0.611	0.275	0.502	0.233	0.630
Analysis	0.709	0.771	0.800	0.772	0.823	0.420	0.756	0.267	0.739
Synthesis	0.539	0.550	0.638	0.622	0.681	0.123	0.584	0.198	0.584
Evaluation	0.540	0.598	0.604	0.630	0.660	0.233	0.617	0.048	0.322
Macro- Average F_1	0.579	0.574	0.708	0.649	0.726	0.250	0.647	0.201	0.557

Journal of Theoretical and Applied Information Technology
15th July 2016. Vol.89. No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

On the other hand, Table 3 presents a comparison between the DM techniques in terms of the best F_1 obtained using specific number of term for all BCLs classifiers. It is obvious that Random Forest and SVM outperform the remaining techniques. The performances of SVM, RA, and JRip are comparable, BNs is lower than them, NB and kNN are comparable and in between BNs and J48, and AdaBoost performs the lowest.

Table 3 : Performances of DM Techniques (Best Macro-Average F_1)

ML / BCL	kNN @ 360	NB @ 500	SVM @ 440	RA @ 440	Random Forest @ 150	J48 @50	JRip @360	AdaBoost @50	BNs @100
Knowledge	0.612	0.533	0.852	0.690	0.857	0.723	0.800	0.389	0.776
Comprehension	0.741	0.742	0.847	0.769	0.724	0.754	0.780	0.667	0.742
Application	0.578	0.558	0.694	0.615	0.756	0.619	0.700	0.471	0.595
Analysis	0.780	0.852	0.847	0.828	0.786	0.678	0.769	0.326	0.821
Synthesis	0.625	0.640	0.778	0.698	0.792	0.439	0.642	0.125	0.667
Evaluation	0.702	0.720	0.767	0.759	0.704	0.500	0.655	0.200	0.604
Macro-Average F_1	0.673	0.674	0.798	0.726	0.770	0.619	0.724	0.363	0.701

Finally, Table 4 illustrates a comparison between the techniques in terms of the best F_1 obtained at different number of terms for each BCL classifiers. Obviously, the performances of SVM and Random Forest are the highest. Interestingly, BNs performs slightly better than JRip and RA, which show comparable performances. The performances of kNN and NB are comparable as well, and J48 performs lower than them, but better than AdaBoost, which performs the lowest.

Table 4 : Performances of DM Techniques (Best F_1 for each BCL)

ML / BCL	kNN		NB		SVM		RA		Random Forest		J48		JRip		AdaBoost		BNs	
	Best F_1	*No. of Terms*	*Best F_1*	*No. of Terms*	*Best F_1*	*No. of Terms*	*Best F_1*	*No. of Terms*	*Best F_1*	*No. of Terms*	*Best F_1*	*No. of Terms*	*Best F_1*	*No. of Terms*	*Best F_1*	*No. of Terms*	*Best F_1*	*No. of Terms*
Know.	0.67	150	0.57	100	0.87	150	0.74	150	0.90	50	0.73	50	0.80	200	0.46	150	0.78	100
Comp.	0.79	500	0.74	440	0.86	300	0.77	440	0.82	50	0.76	50	0.78	360	0.67	50	0.76	50
Appl.	0.58	360	0.56	300	0.78	200	0.64	150	0.76	150	0.65	10	0.70	10	0.56	10	0.74	360
Anls.	0.87	200	0.87	300	0.88	100	0.85	360	0.92	300	0.76	10 0	0.84	100	0.59	10	0.83	360
Synt.	0.73	150	0.69	150	0.79	250	0.71	360	0.80	150	0.50	10 0	0.64	360	0.27	100	0.68	250
Eval.	0.71	300	0.78	200	0.77	440	0.77	400	0.79	10	0.64	10	0.71	150	0.22	150	0.76	150
Avg.	0.72	277	0.70	248	0.82	240	0.75	310	0.83	118	0.67	53	0.74	196	0.46	78	0.76	211

5. DISCUSSION

In conclusion, the above results provide experimental evidences on the feasibility of using DM techniques for analyzing teaching effectiveness. Moreover, the comparison between the techniques shows a variation in their performances, which is attributed to the level of sensitivity of each technique to the curse of dimensionality problem. The performance of k-NN, which classifies a new object by examining the class values of the k most similar data points, is affected in a high dimensional data classification by two main aspects of the high dimensional data, distance concentration and hubness of the search space. The distance concentration problem refers to the tendency of distances between all pairs of points in high-dimensional data to become almost equal and the meaningfulness of finding nearest neighbors in high dimensional spaces [30]. Hubness of the search space refers to the skewness of the number of times a point appears among the k-NNs of other points in a data set [31]. As the dimensionality increases, this distribution becomes considerably skewed and hub points emerge.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Although NB technique mitigates the effect of the curse of dimensionality by making a conditional independence assumption that dramatically reduces the number of parameters to be estimated, its performance is affected, because in practice this assumption is rarely likely to hold. In fact, it has been shown that NB assumption is only a sufficient but not a necessary condition for the optimality of the NB [32]. In contrast, BNs relaxes the conditional independence assumption, however as reported in [33], in a high dimensional data applications, practically, its performance is affected by requiring initial knowledge of many probabilities.

With regard to the SVM, theoretically it can bypass the curse of dimensionality effects of increasing the dimensionality of the data set by providing a way to control model complexity independent of the dimensionality and offers the possibility to construct generalized, non-linear classifiers in high-dimensional spaces; however, increasing data dimensionality affects its performance in many practical cases. For example, the performance is affected by the characteristics of the data set, i.e., if the number of dimensions is much greater than the number of data samples, it is likely to give poor performances. Also the selection of SVM parameters (kernel function and its parameters, and the margin parameter C) becomes a very serious problem in high dimensional data. Moreover, the hubness of the search space caused by increasing the dimensionality of the data set also affects the performance of SVM. RA is conceptually simple and showed underperformance compared to some of the techniques, however its outperformance over some of the techniques for high dimensional data classification such boosting [34] has been reported.

On the other hand, the poor performance of AdaBoost for question classification is expected. In fact, it was previously reported to perform poorly with high-dimensional data [35, 36]. As reported in [36], when it is easy to overfit the training data with the base classifier, AdaBoost.M1 perform exactly as their base classifiers, which can explain the poor performance of AdaBoost in high dimensional data classification. The proneness to overfitting data is related to the number of variables, the number of samples. Finally, the poor performance of J48 is attributed to the strict hierarchical partitioning of the data it uses as a decision tree algorithm, which causes disproportionate importance to some features, and a corresponding inability to effectively leverage all the available features [37].

6. CONCLUSION

This paper presents empirical evidences on the feasibility of using DM techniques to analyze teaching effectiveness by classifying teacher's questions into the cognitive level of Bloom taxonomy. The obtained results show a variation in the level of performance between the techniques according to the level of sensitivity to the curse of dimensionality problem. In this respect, Random Forest and SVM show striking performance, whereas J48 and AdaBooost show a sharp deterioration as the data dimensionality grows.

Finally, this research can be extended in several directions: First, in this research a simple term selection method has been implemented; however, there is a wide range of term selection methods that can be experimented to evaluate their role in the performance of DM techniques. Second, the effect of different question representation methods on the performance of DM techniques could be investigated.

AKNOWLEDGMENT

This work is supported by the Scientific Research Deanship in Najran University, Kingdom of Saudi Arabia under research project number NU 21/11.

REFRENCES:

[1] R. A. Arreola, "Developing a Comprehensive Faculty Evaluation System". *Bolton, Mass: Anker*, 1995.

[2] J. F. Boex, "Identifying the attributes of effective economics instructors: An analysis of student evaluation of instructor data". *Journal of Economic Education*, Vol 31, pp. 211-26, 2000.

[3] S. O. Popoola, and Y. Haliso, "Use of library information resources and services as predator of teaching effectiveness of social scientists in Nigerian universities". *AJLAIS,* Vol, No. 19, pp. 65-77, 2009.

[4] S. Ramsey, C. Gabbard, K. Clawson, L. Lee, and K. T. Henson, "Questioning: An effective teaching method", *Clearing House*, Vol. 63, pp. 420-422, 1990.

[5] A. C. Ornstein, "Questioning: The Essence of Good Teaching" *NASSP Bulletin*, pp. 72-80, 1987.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[6] L. Goe, C. Bell, and O. Little, "Approaches to evaluating teacher effectiveness: A research synthesis".*Washington, DC: National Comprehensive Center for Teacher Quality*, http://www.gtlcenter.org/sites/default/files/docs/EvaluatingTeachEffectiveness.pdf, 2008.

[7] F. P. Hunkins, "The influence of analysis and evaluation questions on achievement in the sixth grade social studies", *Educational Leadership*, vol. 1, pp. 326-332, 1968.

[8] M. D. Gall, "The use of questions in teaching". *Review of Educational Research,* Vol. 40, pp. 707-720, 1970.

[9] L. Anderson, and R. Burns, "Research in classrooms: The study of teachers, teaching and instruction*". New York: Pergamon*, 1989.

[10] M. Dantonio, "How can we create thinkers? Questioning strategies that workfor teachers". Bloomington, in : *National Education Service*, 1990.

[11] A. C. Graesser, and N. K. Person, "Question asking during tutoring". *American Educational Research Journal*, Vol 31, No. 2, pp 104-37, 1994.

[12] J. R. Seymour, and H. P. Osana, "Reciprocal teaching procedures and principles: Two teachers' developing understanding*", Teaching and Teacher Education*, Vol. 19, No. 3, pp 325-44, 2003.

[13] K. Cotton, "Classroom Questioning". *School Improvement Research Series*, http://www.nwrel.org/scpd/sirs/3/cu5.html, 2001.

[14] F. Sebastiani, "Machine Learning in Automated Text Categorization". *ACM Computing Surveys*, Vol 34, No 1, pp 1–47, 2002.

[15] F. P. Romero, P. Julián-Iranzo, M. Ferreira-Satler, and J. Gallardo-Casero: "Classifying unlabeled short texts using a fuzzy declarative approach", *Language Resources and Evaluation*, Vol. 47, No. 1, pp 151-178 , 2013.

[16] E. Gabrilovich, A. Z. Broder, M. Fontoura, A. Joshi, V. Josifovski, L. Riedel, and T. Zhang, "Classifying search queries using the web as a source of knowledge", *ACM Transactions on the Web*, Vol. 3, No 2, pp1-28, 2009.

[17] T. S. Guzella, and T. M. Caminhas, "A review of machine learning approaches to Spam filtering", *Expert Systems with Application*, Vol. 36, No. 7, pp. 10206-10222, 2009.

[18] I. Taksa, S. Zelikovitz, and A. Spink."Non-Topical Classification of Query Logs Using Background Knowledge", In *Machine Learning Techniques for Adaptive Multimedia Retrieval: Technologies Applications and Perspectives*, ed. Chia-Hung Wei and Yue Li, pp. 194-212, 2011.

[19] A. Nuntiyagul, K. Naruedomkul, N. Cercone N., and D. Wongsawang, "Adaptable Learning Assistant for Item Bank Management", *Computers & Education*, Vol. 50, No. 1, pp. 357-370, 2008.

[20] F. Ting, J. H. Wei, C. T. Kim, Q., Tian, "Question Classification for E-learning by Artificial Neural Network". *In Proceedings of the Joint Conference of the Fourth International Conference*, pp. 1757- 1761, 2003.

[21] S.-C.Cheng,Y.-M.Huang, J.-N.Chen, andY.-T. Lin, "Automatic leveling system for e-learning examination pool using entropybased decision tree," in *Advances in Web-Based Learning-ICWL 2005*, vol. 3583 of Lecture Notes in Computer Science, pp. 273–278, 2005.

[22] B. C., Chein and S. T., Liau, "An Automatic Classifier for Chinese Items Categorization", *In Proceedings of National Conference on Artificial Intelligence and Its Application*, Taiwan, 2004.

[23] R. Kavitha, A., Vijaya, D., Saraswathi, "A Two-Phase Item Assigning in Adaptive Testing Using Norm Referencing and Bayesian Classification". *Advances in Computer Science, Engineering & Applications*, Vol. 166, pp. 809-816, 2012.

[24] A. A., Yahya, A., Osman, and A., Alatab, " Educational Data Mining: A case Study of Teacher's Classroom Questions", ISDA'13, UPM, Malaysia , pp.8 -10 , 2013.

[25] A. A.,Yahya, A., Osman, A., Taleb, and A., Al-attab, "Analyzing the Cognitive Level of Classroom Questions Using Machine Learning Techniques", *Procedia Social and Behavioral Sciences 97*, pp. 587–59, 2013.

[26] G. Salton,"Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer", *Addison-Wesley Reading*, 1989.

[27] M. F. Porter, "An algorithm for Suffix Stripping. *Program*, Vol. 14, No. 3, pp. 130–137, 1980.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[28] Y. Xu, and L. Chen, "Term-frequency based feature selection methods for text categorization", *In proceedings of the 4th International Conference on Genetic and Evolutionary Computing*, pp. 280-283, 2010.
[29] *The University of Waikato*. Weka 3: Data Mining Software in Java http://www. cs.waikato.ac.nz/ml/weka/.
[30] C. C., Aggarwal, A., Hinneburg, and D., A., Keim, "On the Surprising Behavior of Distance Metrics in High Dimensional Spaces", *In Proceedings of the International Conference on Database Theory*, pp. 420–434, 2002.
[31] M. Radovanović , A. Nanopoulos , M. Ivanović, "Nearest neighbors in high-dimensional data: the emergence and influence of hubs", *In Proceedings of the 26th Annual International Conference on Machine Learning*, p.865-872, Montreal, Quebec, Canada, June 14-18, 2009.
[32] P. Domingos, and M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss". *Machine Learning,* Vol. 29, pp. 103–130, 1997.
[33] A. Mittal, L. F. Cheong, "Addressing the Problems of Bayesian Network Classification of Video Using High-Dimensional Features", *IEEE Transaction on Knowledge Data Engineering*. Vol. 16, No. 2, pp. 230-244, 2004.
[34] R. E., Schapire, Y., Singer, and A., Singhal, "Boosting and Rocchio applied to text filtering", *In Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval*, pp. 215–223, Australia, 1998.
[35] S. Dudoit, J., Fridlyand, and T. P. Speed, "Comparison of Discrimination Methods for the Classification of Tumors using Gene Expression Data". *Journal of the American Statistical Association,* Vol. 97, pp. 77-87, 2002.
[36] R. Blagus, and L. Lusa, "Boosting for high-dimensional two-class prediction". *BMC Bioinformatics*, Vol. 16, pp. 1-17, 2015.
[37] H. Tong., "Big data classification", *Data Classification: Algorithms and Applications*. *Chapter 10*, Eds. C.C. Aggarwal, Taylor and Francis Group, LLC. pp. 275–286. 2015.