IMPROVING IMBALANCED QUESTION CLASSIFICATION USING STRUCTURED SMOTE BASED APPROACH

ALAA MOHASSEB[1], MOHAMED BADER-El-DEN[1], MIHAELA COCEA[1], HAN LIU[2]

[1]School of Computing, University of Portsmouth
[2]School of Computer Science and Informatics, Cardiff University
E-MAIL: alaa.mohasseb@port.ac.uk, mohamed.bader@port.ac.uk, mihaela.cocea@port.ac.uk, liuh48@cardiff.ac.uk

Abstract:

Questions Classification (QC) is one of the most popular text classification applications. QC plays an important role in question-answering systems. However, as in many real-world classification problems, QC may suffer from the problem of class imbalance. The classification of imbalanced data has been a key problem in machine learning and data mining. In this paper, we propose a framework that deals with the class imbalance using a hierarchical SMOTE algorithm for balancing different types of questions. The proposed framework is grammar-based, which involves using the grammatical pattern for each question and using machine learning algorithms to classify them. Experimental results imply that the proposed framework demonstrates a good level of accuracy in identifying different question types and handling class imbalance.

Keywords:

Information Retrieval, Text classification, Question classification, Machine Learning, Class Imbalance.

1 Introduction

Questions Classification (QC) is the main task in any question-answering systems. However, as in many real-world classification problems, QC may suffer from the problem of class imbalance [1]. The classification of imbalanced data has been a key problem in machine learning and data mining [2], [3]. Class imbalance occurs when one of the two classes having more instances than other classes in which the algorithm usually focuses on the classification of instances of the majority class, while ignoring or misclassifying instances of the minority class. The lack of information caused by a small sample size in the training set is one of the challenges in imbalance classification [4], [5] in which an insufficient number of instances will result in the difficulty for the algorithm to learn from similar patterns in the instances of the minority class.

The Synthetic Minority Over-sampling TEchnique (SMOTE) [6] is one of the most popularly used sampling technique to handle imbalance data [7], [8], [2], [9], [3]. SMOTE over-samples instances of the minority (abnormal) class which helps for achieving better classifier performance.

In this paper, we propose a framework which deals with the class imbalance issue using the hierarchical SMOTE algorithm when classifying different types of questions. The proposed framework is grammar-based, which involves using the structure of the question by identifying a grammatical pattern for each question and using machine learning algorithms to classify them. Experimental results imply that the proposed framework demonstrates a good level of accuracy in identifying different question types and handling class imbalance. The aim of the research presented in this paper is to: *"Evaluate the impact of handling class imbalance in the classification accuracy."*

The rest of the paper is organized as follows. Section 2 outlines previous work in question classification. Section 3 describes the proposed question classification framework. The experiments setup and results are presented in Section 4. Finally, section 5 concludes the paper and outlines directions for future work.

2 Question Classification

In many recent studies, users' question is classified using different features. Authors in [10] proposed head word features, which is one single word specifying the object that the question seeks. In [11], a framework has been proposed, which integrates a question classifier with a simple document/passage retriever, and proposed context-ranking models. In [12], a hybrid approach was proposed, named ATICM which is based

on dependency tree analysis by utilizing both syntactic and semantic analysis. In addition, authors in [13] proposed a method of using a feature selection algorithm to determine appropriate features using Support Vector Machine (SVM) for the classification algorithm. In [14], a statistical classifier has been proposed which is based on SVM. Furthermore, [15] proposed a SVM-based approach for question classification, a dependency relations and high-frequency words are incorporated into the baseline system.

Authors in [16] proposed an approach for question classification through using three different classifiers. Similarly, in [17] five machine learning algorithms were used such as SVM and Naive Bayes, with using two kinds of features bag-of-words and bag-of-ngrams. In [18] authors classified open-ended questions through training SVM to recognize the occurrence of certain keywords or phrases in a question class. Moreover, [19] proposed a neural network for a question answering system. In [20] a classification method was proposed for community question answering (CQA) system based on ensemble learning. Finally, authors in [21] proposed two trained recurrent neural networks to detect the entities in the question and to classify the question.

Unlike the previous approaches which ignore the problem of class imbalance in question classification, we propose a grammar-based framework which deal with the class imbalance using hierarchical SMOTE algorithm for questions classification. Details of the framework are presented in the next section.

3 Proposed Approach

We propose a framework, shown in Figure 1, which deal with the class imbalance issue when classifying different types of questions. The proposed framework transforms the given question to a pattern meaning that each term in the question is represented as its grammatical category. For example, the question "Who is Tim Berners-lee?" will be transformed to *question word who (QW_{Who}), linking verb (LV) and proper noun celebrity (PN_C)*, "QW_{Who} + LV + PN_C". The grammatical categories contain in addition to typical categories of English grammar, domain-related grammatical categories [22][23]. This new representation helps in the process of dealing with class imbalance and the final categorization and classification of the given question.

The three phases of the proposed framework for imbalance question categorization and classification are described as follow:

Phase 1: Question Analysis: this phase is executed using a simple version of the English grammar combined with domain-specific grammatical categories since domains such as question answering systems do not perceive the formal English grammar and natural language. First, the given question is analyzed, and this step is done by identifying each of the keywords and phrases in the question. Next, the grammar is generated. After this step a grammatical rule is generated, in which a question domain specific grammar will be created.

Phase 2: Pattern Formulation: this phase consists of three steps. (1) the system parses the given question to facilitate the tagging of each word to the right term category. (2) each term in the question will be tagged to its term category, (3) the question is transformed to a pattern. [24].

Phase 3: Question Classification: in this phase the patterns generated from the previous phase are used for machine learning, in this phase a model for automatic classification is built. The classification is done by splitting of the data set into (1) a training set which is used for building the model, and (2) a test set which is used to evaluate the performance of the model, before the model is used for classifying unseen instances, SMOTE Sampling technique is applied to over-sample instances of the minority class in the training set by constructing new instances of the minority class, which helps for achieving better classifier performance. One of SMOTE limitations is that it is designed for binary labels. When dealing with multiple labels, SMOTE will only over-sample for the label with the lowest number of instances. In order to overcome this limitation, in this study, SMOTE is applied several times in a hierarchical way to over-sample all the non-majority classes.

4 Experimental Study and Results

Naive Bayes (NB) was used as the machine learning algorithm for the automatic classification. The classification accuracy is obtained by using the implementation of the above algorithm from the Weka software. The effectiveness of the classification algorithm was evaluated in terms of Precision, Recall and F-Measure, i.e. typical metrics for the evaluation of classifiers, using 10-fold cross-validation and SMOTE with value of K= 5. To show the effectiveness of handling imbalance data on the classification performance, two experiments were conduct (1) using NB without applying SMOTE algorithm and (2) using NB with the implementation of SMOTE algorithm.

1,160 questions were randomly selected from three data sets (1) TREC 2007 Question Answering Data [1] and (2) a Wikipedia dataset [2] and (3) Yahoo Non-Factoid Question

[1] http://trec.nist.gov/data/qa/t2007_qadata.html
[2] https://www.cs.cmu.edu/~ark/QA-data

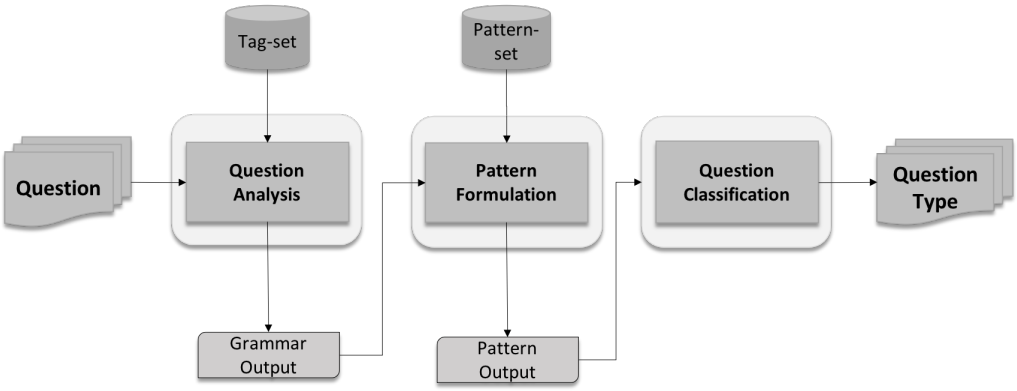

FIGURE 1. Framework

Dataset[3]. Their distribution is given in Table 1. Questions used in this experiment are labelled to six different categories,namely; causal, choice, confirmation (Yes-No Questions), factoid (Wh-Questions), hypothetical and list. These classifications were proposed by [25].

TABLE 1. Data distribution

Question type	Total
Causal	31
Choice	12
Confirmation	321
Factoid	688
Hypothetical	7
List	101

4.1 Results

Table 2 presents classification performance details (Precision, Recall and F-Measure) of the NB classifier and the performance details of the NB classifier with the use of the SMOTE algorithm. The results indicate that when handling imbalance classes the performance of the classifier is improved, as shown in Table 1. Choice, causal and hypothetical questions have much fewer instances, and without applying the SMOTE algorithm the classifier had poor performance especially with these three classes. However, when the SMOTE algorithm is applied, the performance of the classifier has been improved and the overall accuracy has increased.

Furthermore, these results show that NB is effective in the identification and classification of confirmation and factoid questions. In addition, NB could not distinguish between causal, choice, hypothetical and list types of questions and incorrectly classified most of them as confirmation and factoid questions. However, when applying SMOTE algorithm classification of most question types and the performance has been improved. For example, when the SMOTE algorithm is not applied, NB could correctly classified (Recall) less than 1% of the causal questions, and could not identify any of the choice questions. Furthermore, NB classified correctly 92.8% of the confirmation questions and 92.7% of the factoid questions. In addition, 28.6% of the hypothetical questions were correctly classified while the classification accuracy of the list questions were 27.7%.

On the contrary, when hierarchical SMOTE algorithm is applied, NB correctly classified 58.1% of the causal questions and 16.7% of the choice questions. In addition, classification of factoid, confirmation and hypothetical questions achieves a higher recall when handling imbalance classes, i.e. 95.5%, 94.1% and 71.4% accuracy respectively. Moreover, classification of list questions shows a lower recall (18.8%) with the implementation of SMOTE but higher precision. Overall, the results validate that the proposed approach is an effective method for question classification as well as for the distinction between different question types and handling the problem of imbalance classes.

5 Conclusion and Future Work

In this paper, we proposed a framework for question classification, which deals with the class imbalance issue using the hierarchical SMOTE algorithm by utilizing the structure of the question, based on the grammatical pattern of each question. The results show that our proposed solution led to a good

[3] https://ciir.cs.umass.edu/downloads/nfL6/

TABLE 2. NB classifier performance without/with the implementation of SMOTE algorithm

	Naive Bayes			Naive Bayes with (SMOTE)		
Question Types	**Precision**	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**
Causal	**0.231**	**0.097**	**0.136**	**0.621**	**0.581**	**0.600**
Choice	**0**	**0**	**0**	**0.154**	**0.167**	**0.160**
Confirmation	0.906	0.928	0.917	0.944	0.941	0.942
Factoid	0.85	0.927	0.887	0.870	0.955	0.911
Hypothetical	**0.133**	**0.286**	**0.182**	**0.417**	**0.714**	**0.526**
List	0.609	0.277	0.381	0.613	0.188	0.288
Overall	0.814	0.835	0.818	0.851	0.865	0.847

performance in classifying questions and handling class imbalance. As a future work, we will apply different imbalance algorithms e.g (cost-sensitive) and compare the performance of different classifiers, when different class imbalance methods are applied. In addition, we will test the proposed framework in other text classification domains with similar class imbalance problems.

References

[1] Y. Li, G. Sun, and Y. Zhu, "Data imbalance problem in text classification," in *Information Processing (ISIP), 2010 Third International Symposium on*. IEEE, 2010, pp. 301–305.

[2] Z. Zheng, Y. Cai, and Y. Li, "Oversampling method for imbalanced classification," *Computing and Informatics*, vol. 34, no. 5, pp. 1017–1037, 2016.

[3] K. Jiang, J. Lu, and K. Xia, "A novel algorithm for imbalance data classification based on genetic algorithm improved smote," *Arabian Journal for Science and Eng.*, vol. 41, no. 8, pp. 3255–3266, 2016.

[4] S. Visa and A. Ralescu, "The effect of imbalanced data class distribution on fuzzy classifiers-experimental study," in *Fuzzy Systems, 2005. FUZZ'05. The 14th IEEE International Conference on*. IEEE, 2005, pp. 749–754.

[5] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent data analysis*, vol. 6, no. 5, pp. 429–449, 2002.

[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[7] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, "Smoteboost: Improving prediction of the minority class in boosting," in *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, 2003, pp. 107–119.

[8] F. Chang, J. Guo, W. Xu, and K. Yao, "A feature selection method to handle imbalanced data in text classification," *J. Digit. Inf. Manage*, vol. 13, no. 3, pp. 169–175, 2015.

[9] J. Wang, M. Xu, H. Wang, and J. Zhang, "Classification of imbalanced data by using the smote algorithm and locally linear embedding," in *Signal Processing, 2006 8th International Conference on*, vol. 3. IEEE, 2006.

[10] Z. Huang, M. Thint, and Z. Qin, "Question classification using head words and their hypernyms," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2008, pp. 927–936.

[11] S.-J. Yen, Y.-C. Wu, J.-C. Yang, Y.-S. Lee, C.-J. Lee, and J.-J. Liu, "A support vector machine-based context-ranking model for question answering," *Information Sciences*, vol. 224, pp. 77–87, 2013.

[12] T. Hao, W. Xie, Q. Wu, H. Weng, and Y. Qu, "Leveraging question target word features through semantic relation expansion for answer type classification," *Knowledge-Based Systems*, vol. 133, pp. 43–52, 2017.

[13] N. Van-Tu and L. Anh-Cuong, "Improving question classification by feature extraction and selection," *Indian Journal of Science and Technology*, vol. 9, no. 17, 2016.

[14] D. Metzler and W. B. Croft, "Analysis of statistical question classification for fact-based questions," *Information Retrieval*, vol. 8, no. 3, pp. 481–504, 2005.

[15] S. Xu, G. Cheng, and F. Kong, "Research on question classification for automatic question answering," in *Asian Language Processing (IALP), 2016 International Conference on*. IEEE, 2016, pp. 218–221.

[16] M. Mishra, V. K. Mishra, and H. Sharma, "Question classification using semantic, syntactic and lexical features," *International Journal of Web & Semantic Technology*, vol. 4, no. 3, p. 39, 2013.

[17] D. Zhang and W. S. Lee, "Question classification using support vector machines," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 2003, pp. 26–32.

[18] J. Bullington, I. Endres, and M. Rahman, "Open ended question classification using support vector machines," *MAICS 2007*, 2007.

[19] T. Sagara and M. Hagiwara, "Natural language neural network and its application to question-answering system," *Neurocomputing*, vol. 142, pp. 201–208, 2014.

[20] Y. Li, L. Su, J. Chen, and L. Yuan, "Semi-supervised learning for question classification in cqa," *Natural Computing*, vol. 16, no. 4, pp. 567–577, 2017.

[21] F. Ture and O. Jojic, "Simple and effective question answering with recurrent neural networks," *arXiv preprint arXiv:1606.05029*, 2016.

[22] A. Mohasseb, M. Bader-El-Den, H. Liu, and M. Cocea, "Domain specific syntax based approach for text classification in machine learning context," in *2017 International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 2. IEEE Systems, Man and Cybernetics, 2017, pp. 658–663.

[23] A. Mohasseb, M. Bader-El-Den, A. Kanavos, and M. Cocea, "Web queries classification based on the syntactical patterns of search types," in *International Conference on Speech and Computer*. Springer, 2017, pp. 809–819.

[24] A. Mohasseb, M. El-Sayed, and K. Mahar, "Automated identification of web queries using search type patterns." in *WEBIST (2)*, 2014, pp. 295–304.

[25] A. Mohasseb, M. Bader-El-Den, and M. Cocea, "Question categorization and classification using grammar based approach," *Information Processing and Management*, 2018.