Connecting the System to Enhance the Practitioner and Consumer Experience in Healthcare
E. Cummings et al. (Eds.)

doi:10.3233/978-1-61499-890-7-73

Comparative Analysis of Algorithmic Approaches for Auto-Coding with ICD-10-AM and ACHI

Rajvir KAUR[a,1] and Jeewani Anupama GINIGE[a,2]
[a]*School of Computing, Engineering & Mathematics, Western Sydney University, Australia*

Abstract. Clinical coding is done using ICD-10-AM (International Classification of Diseases, version 10, Australian Modification) and ACHI (Australian Classification of Health Interventions) in acute and sub-acute hospitals in Australia for funding, insurance claims processing and research. The task of assigning a code to an episode of care is a manual process. This has posed challenges due to increase set of codes, the complexity of care episodes, and large training and recruitment costs of clinical coders. Use of Natural Language Processing (NLP) and Machine Learning (ML) techniques is considered as a solution to this problem. This paper carries out a comparative analysis on a selected set of NLP and ML techniques to identify the most efficient algorithm for clinical coding based on a set of standard metrics: precision, recall, F-score, accuracy, Hamming loss and Jaccard similarity.

Keywords. Classifiers, Machine Learning, Natural Language Processing, Precision, Recall, F-score, Hamming Loss, Jaccard Similarity.

Introduction

Clinical coding is the process of assigning alphanumeric codes, based on a set of clinical coding standards to a patient's episode of care details presented in the hospital discharge summary [1]. These clinical codes are assigned by trained clinical coders who have good knowledge of coding rules and acquainted with latest medical classification systems such as ICD-10 (International Classification of Diseases, version 10). The assignment of clinical codes serves as a justification for funding, insurance claim processing and research [2]. Insurance companies make reimbursement based on the clinical codes assigned to each report after clinical treatment. Moreover, government and policy makers use coded data to: analyse the healthcare system to reveal most diseases prone geographical areas, justify investment done by government in the healthcare industry and make future investments based on these statistics [3]. Any inaccurate assignment of codes may lead to various issues such as reviewing the whole process of assigning codes, delay in payment processes, increased labour costs as well as financial losses.

The World Health Organisation (WHO) maintained the Bertillon classification in 1948 and named it as International Statistical Classification of Diseases, Injuries and

Corresponding Author: [1]Rajvir Kaur is student of Master of Research at Western Sydney University, Australia. Email: 18531738@student.westernsydney.edu.au.
[2]Jeewani Anupama Ginige is senior lecturer in Health Informatics at Western Sydney University, Australia. Email: j.Ginige@westernsydney.edu.au

Causes of Death [4]. Since then, roughly every ten years, this classification system has been revised and in 1992, WHO published ICD-10 version. The next generation of ICD-11 is currently under development at WHO [5]. Twenty-six (26) years after the introduction of ICD-10, this year (2018), the ICD-11 will be put forward to WHO general assembly for approval.

Since its introduction, ICD-10 is widely used all over the world. Many countries extended ICD-10 classification system to make it suitable for their country specific reporting purposes. For example, ICD-10-CM (Clinical Modification) is used by the USA, ICD-10-CA (Canadian Modification), and ICD-10-GM (German Modification) [4]. The Australian Modification (ICD-10-AM) is used in Australia along with 15 other countries including, Ireland, Singapore and Saudi Arabia[6]. Currently, in Australia, the Australian Consortium for Classification Development (ACCD) is responsible for updating the classification system every two years, on behalf of the Independent Hospital Pricing Authority (IHPA). The health classification systems used in Australia include: ICD-10-AM, the Australian Classification of Health Interventions (ACHI) and the associated Australian Coding Standard (ACS) [7]. ICD-10-AM contains Alphabetic Index and Tabular List for disease classification, ACHI contains intervention classification in conjunction with ICD-10-AM. The major difference between ICD-10 and ICD-10-AM is that ICD-10-AM provides more specificity of the disease codes.

With the transition from ICD-9 to ICD-10, the manual assignment of clinical codes has become a non-trivial task, due to the increased number of codes. On an average, a clinical coder codes 4 to 5 discharge summaries per hour. This results in 15 to 42 records per day depending upon the experience and efficiency of the clinical coder [8]. A study [2], estimates that the US spends about 25 billion dollars per year for assigning clinical codes and their follow-up corrections. To reduce these errors, research is being conducted to develop methods for computer based coding or auto-coding [2], [9]. The concept of auto-coding is at the inception level despite the much advancement in Artificial Intelligence (AI) and Machine Learning (ML). This is mainly due to the continuous use of paper based records rather than electronic, inconsistent document structures and content variations across the healthcare organisations.

The research studies that concentrate on auto-coding are focused on ICD-9, ICD-9-CM or ICD-10-CM [2], [10]. In addition, there has not been any previous research that has ICD-10-AM and ACHI for auto-coding purposes using discharge summary data. Therefore, this paper aims to carry out a comparative analysis of pattern matching, rule based and machine learning approach to gauge the most suitable computational approach to auto-coding with ICD-10-AM and ACHI. The work presented here only concentrate on two ICD-10-AM chapters that represent the code sets associated with respiratory and gastrointestinal diseases.

1. Literature Review

Since 1990's, various attempts have been made by many researchers to create automated systems for assigning codes to patient's episode of care [2], [10]. Depending upon the applications, different methods and techniques ranging from pattern matching to machine learning approaches are being applied to lower the healthcare cost and improve the quality. Several measures have been proposed for evaluating the efficiency of text classification outcomes but the standard evaluation criteria are followed by calculating

Precision, Recall, F-score, Accuracy, Hamming Loss, Jaccard Similarity and Zero-One Loss [11].

1.1. Pattern Matching Approach

Pattern matching approach is the simplest and fundamental technique that searches a text-string within the text. Text-string is matched character for character against the given text with the use of regular expressions [12]. Though, pattern matching is the simplest approach but inevitably introduces errors [13]. For example, consider a report having the text *"A 51 year old patient has serious cough but no sign of pneumonia."* In pattern matching, "cough" and "pneumonia" are the keywords identified from the text. Therefore, treating "pneumonia" as a match and assigning its relevant ICD code, in this case, it is a mistake. Moreover, in natural language, a word or phrase can have multiple meanings which do not mean that every extracted keyword does necessarily mean the same thing. To overcome this problem, a set of rules are defined to avoid unnecessary coding of the wrong patterns.

1.2. Rule-based Approach

In early 80's, group of experts manually defined set of rules and categories for text classification using some logical expressions and Boolean operations to implement the mapping and code assignment [12], [14]. For example, a rule can be in the form of "if **(logical expression)** then **(category)**". The text is classified under **category** if it satisfies the **logical expression** as shown in Table 1.

Table 1. Generating rules from ICD-10 [2].

ICD-10 Codes	**Generating Rules**
K05.3	**If document contains**
Acute periodontitis	acute periodontitis **OR**
Acute pericoronitis	acute pericoronitis **OR**
Paradontal abscess	paradontal abscess **OR**
Peridontal abcess	peridontal abcess **OR**
Excludes	**AND document NOTcontains**
acute apical periodontitis (K04.4)	acute apical periodontitis **AND**
periapical abscess (K04.7)	periapical abscess **AND**
periapical abcess with sinus (K04.6)	periapical abcess with sinus
	assign code K05.3

This approach is usually very accurate as it is based on expert's knowledge and experience but is time-consuming. A literature survey,[15], found that about 65% of the research is based on rule-based approaches. The main drawback of rule-based approach is the knowledge acquisition bottleneck, which means that rules are manually defined by domain experts and if there is any up-gradation in the codes or categories then rules need to be revised [14].

1.3 Machine Learning Approach

Machine learning, also known as statistical approach often utilise different linguistic principles and features for statistical measurements to extract semantic information. Hasan et al.[9], performed classification on clinical interview transcripts using 8 classifiers: Naïve Bayes [16], Support Vector Machine [17], Decision Tree, Conditional

Random Fields, Adaboost, Random forest, DisCLDA and Convolutional Neural Network in conjunction with lexical, contextual and semantic features. In the analysis with the removal of stop-words the performance of Naïve Bayes and SVM decreased by 19.9% and 15.5% respectively.

2. Proposed Methodology

There are various methods and techniques for handling and processing unstructured clinical text but the methodology used in this work is associated with the steps shown in Table 2.

Table 2. Methodology used in the work.

Step		Description
1.	*Data Acquisition*	For this research, a collection of medical records from hospitals all over Australia held by the National Centre for Classification in Health (NCCH) was used under Western Sydney University ethics approval with number *H12628190*. The dataset contains 190 anonymised clinical records associated with respiratory and gastrointestinal diseases and interventions. As most of the clinical records were paper-based records, therefore, with the help of clinical coder, a gold standard data was organised in a structured manner. On the basis of structured data, 190 text narratives were created. Due to the limited number of clinical records, an additional 45 clinical records similar to 190 records for respiratory and gastrointestinal diseases were created by mixing and matching certain diagnosis and interventions. The dataset with the original number of clinical records is referred to as *Data190* and the dataset with 235 clinical records is referred as *Data235*.
2.	*Data Pre-processing*	In data pre-processing, the clinical records were cleaned up to extract useful information which includes principal diagnosis, additional diagnosis, smoke related diagnosis, diabetes condition, supplementary conditions, principal procedure, additional procedure, anesthesia type, ventilation details and allied health care interventions. Tokenisation is done to split the clinical documents into sentences and words. The abbreviations were replaced by creating a dictionary of abbreviations and replaced with full forms. In terms of British English and American English, some medical terms are spelled differently. Therefore, spell check is performed by detecting, suggesting and replacing the words with correct spellings using PyEnchant[2] python library. The data is filtered out to remove unwanted words by creating a list of stop-words. Negated terms were not removed during processing stage as these terms are very important for providing a clue for negated and uncertain findings.
3.	*Feature Extraction*	Once the clinical records were cleaned up by removing unwanted information, the next step was to represent the text data in numeric form for feature extraction. We used Bag-of-Words model to create a list of unique words and 1-gram (uni-gram), 2-gram (bi-gram), 3-gram (tri-gram) and 4-gram as feature set.
4.	*Classification*	Comparative analysis of 7 classifiers namely, SVM, Naïve Bayes, Decision Tree, k-Nearest Neighbor, Random Forest, AdaBoost, and Multi Layer Perceptron (MLP) is done.
5.	*Evaluation Metrics*	All the machine learning experiments were conducted using 80-20 ratio which means 80% of the data was used for training and the remaining 20% for testing. Evaluation metrics was performed by calculating Precision, Recall, F-score, Accuracy, Hamming Loss and Jaccard similarity [11].

[2] https://pypi.python.org/pypi/pyenchant/

3. Experimental Results

This section presents the results of the experiments that were performed on Data190 and Data235 datasets. The experimental work was divided into two tasks namely Task-1 and Task-2.

3.1 Task 1: ICD-10-AM/ACHI Chapter Classification

The clinical records containing diseases and interventions belonging to respiratory system are labelled as ***Respiratory*** class and the clinical records belonging to digestive system are labelled as ***Gastrointestinal*** class. In Data190 chapter classification results, SVM outperforms in comparison to all other classifiers achieving 0.95 F-score with 0.05263 and 0.94736 Hamming loss and Jaccard similarity respectively. Similarly, in Data235 results, Naïve Bayes classifier gives better results than SVM with 0.02127 hamming loss and 0.97872 Jaccard similarity score.

3.2. Task 2: ICD-10-AM/ACHI Code Assignment

In Task-2, three different approaches were applied on both the datasets to perform comparative analysis. The pattern matching and rule-based approach do not require any training and testing data. As observed in Table 3, rule based approach gives better results than pattern matching approach because in rule based approach certain rules are defined for diseases having different synonyms for same set of codes. For example, "A09.9" is the ICD-10-AM code for "Gastroenteritis and colitis of unspecified origin", "Loose stool" and "Diarrhoea".

Table 3. Comparison of Pattern Matching and Rule based approach results for Data190 and Data235.

Approach	Dataset	Precision	Recall	F-score	Accuracy	Hamming Loss	Jaccard Similarity
Pattern Matching	**Data190**	0.7953	0.4184	0.5277	0.4027	0.0430	0.4365
	Data235	0.8029	0.4090	0.5201	0.3945	0.0405	0.4255
Rule based Approach	**Data190**	0.7913	0.6916	0.7257	0.6053	0.1728	0.5803
	Data235	0.7920	0.6872	0.7222	0.6011	0.1745	0.5768

Machine learning results using 2-gram and 4-gram feature set are shown in Table 4. Though, we have applied 1-gram and 3-gram also but 4-gram feature set gave better results in Data190 and 2-gram in Data235. Decision Tree outperformed among other classifiers with precision of 0.9206, recall of 0.8501 and 0.8730 F-score. Though, its Hamming loss is 0.08776 which is far less than 0.41034 (unigram feature set) but gives 79.20% overall accuracy. On the other hand, in Data235, results are improved using 2-gram features and AdaBoost provides better results than Decision Tree.

Table 4. Comparative Analysis of Data190 using 4-gram feature set Data235 using 2-gram feature set.

Classifier	Dataset	Precision	Recall	F- score	Accuracy	Hamming Loss	Jaccard Similarity
SVM	**Data190**	0.76798	0.45175	0.54361	0.44051	0.03706	0.44776
	Data235	0.89308	0.55191	0.65373	0.54143	0.01955	0.52697
Naïve Bayes	**Data190**	0.62534	0.63168	0.57465	0.44051	0.67841	0.42014
	Data235	0.72891	0.61722	0.61821	0.49643	0.35158	0.48805
Random Forest	**Data190**	0.58333	0.25586	0.33523	0.25389	0.01392	0.27135
	Data235	0.66666	0.30773	0.39793	0.29717	0.02453	0.32365
AdaBoost	**Data190**	0.81421	0.81329	0.79115	0.66831	0.23514	0.65517
	Data235	**0.92392**	**0.92019**	**0.91412**	**0.86118**	0.09458	**0.82945**

Decision Tree	**Data190**	**0.92062**	0.85015	**0.87305**	**0.79201**	0.08776	**0.74537**
	Data235	0.91407	0.91295	0.90351	0.84462	0.11271	0.79245
MLP	**Data190**	0.62938	0.29488	0.37559	0.29073	0.02192	0.29000
	Data235	0.63475	0.34756	0.38689	0.34537	0.00942	0.33055
kNN	**Data190**	0.68001	0.46388	0.51485	0.38567	0.34411	0.36667
	Data235	0.57679	0.46974	0.40582	0.40993	0.24057	0.39130

4. Conclusion and Future Work

The clinical records in hospitals contain patient's medical information which is used for assigning alphanumeric codes based on ICD-10-AM and ACHI, in Australia. This paper, reports a comparative analysis done on two datasets using three different approaches pattern matching, rule based approach and machine learning. In machine learning, 7 classifiers with unigram, bigram, trigram and 4-gram feature sets were used. The Decision Tree and AdaBoost give better results as compared to other classifiers. The reason behind low performance of the other classifiers namely MLP and Random Forest is that some diseases names have occurred only once or twice in the clinical records which lowers the learning rate and performance of the system. Therefore, in future, we will expand the dataset size and work upon other chapters in ICD-10-AM and ACHI. Further exploration will be carried out on applying a hybrid approach and deep learning techniques to improve the overall system performance.

References

[1] (HETI), H.E.a.T.I. *Clinical Coding*. Available from: http://www.heti.nsw.gov.au/Programs/Clinical-Coding-Workforce-Enhancement-Project.

[2] Farkas, R. and G. Szarvas, *Automatic construction of rule-based ICD-9-CM coding systems.* BMC Bioinformatics, 2008. **9**(Suppl 3): p. S10.

[3] (WHO), W.H.O. *Uses of Coded Clinical Data* September 2012 Available from: https://www.cdc.gov/nchs/data/icd/uses_coded_clinicalinfosheet.pdf

[4] Cumerlato, M., et al., *Fundamentals of morbidity coding using ICD-10-AM, ACHI, and ACS eighth edition*. 2013.

[5] Organization, W.H. *The 11th Revision of the International Classification of Diseases (ICD-11)*. 26 February 2018; Available from: http://www.who.int/classifications/icd/revision/en/.

[6] (IHPA), I.H.P.A.; Available from: https://www.ihpa.gov.au/what-we-do/products/AR-DRG-classification-system/country-licence-agreement.

[7] *Australian Consortium for Classification Development (ACCD)*. 2018; Available from: https://www.accd.net.au/.

[8] Santos, S., et al., *Organisational Factors Affecting the Quality of Hospital Clinical Coding.* Health Information Management Journal, 2008. **37**(1): p. 25-37.

[9] Hasan, M., et al., *A study of the effectiveness of machine learning methods for classification of clinical interview fragments into a large number of categories.* Journal of Biomedical Informatics, 2016. **62**: p. 21-31.

[10] Kavuluru, R., A. Rios, and Y. Lu, *An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records.* Artificial Intelligence in Medicine, 2015. **65**(2): p. 155-166.

[11] Aldrees, A. and A. Chikh, *Comparative evaluation of four multi-label classification algorithms in classifying learning objects.* Computer Applications in Engineering Education, 2016. **24**(4): p. 651-660.

[12] Cai, T., et al., *Natural Language Processing Technologies in Radiology Research and Clinical Applications.* RadioGraphics, 2016. **36**(1): p. 176-191.

[13] Chen, P., A. Barrera, and C. Rhodes. *Semantic analysis of free text and its application on automatically assigning ICD-9-CM codes to patient records*. in *Cognitive Informatics (ICCI), 2010 9th IEEE International Conference on*. 2010.

[14] Sebastiani, F., *Machine learning in automated text categorization.* ACM Comput. Surv., 2002. **34**(1): p. 1-47.
[15] Wang, Y., et al., *Clinical information extraction applications: A literature review.* Journal of Biomedical Informatics, 2018. **77**: p. 34-49.
[16] McCallum, A. and K. Nigam. *A comparison of event models for naive bayes text classification*. in *AAAI-98 workshop on learning for text categorization*. 1998. Citeseer.
[17] Cortes, C. and V. Vapnik, *Support-Vector Networks.* Machine Learning, 1995. **20**(3): p. 273-297.